التحليل الإحصائي باستخدام لغة R

(Statistical Analysis using R)

تأليف

د . رامي صلاح جبريل

الطبعة الأولى

2016

التحليل الإحصائي باستخدام لغة R

(Statistical Analysis using R)

تأليف

د. رامي صلاح محمد جبريل

أستاذ مشارك في قسم الإحصاء – كلية العلوم
جامعة بنغازي – ليبيا

الطبعة الأولى – 2016

الطبعة الأولى – 2016

اسم الكتاب: التحليل الإحصائي باستخدام لغة R

اسم المؤلف: د. رامي صلاح محمد جبريل

الوكالة الليبية للترقيم الدولي الموحد للكتاب

دار الكتب الوطنية

بنغازي – ليبيا

هاتف: 9097074 – 9096379 – 9090509

بريد مصور: 9097073

البريد الإلكتروني: nat_lib_libya@hotmail.com

رقم الإيداع القانوني 2016 / 303

ردمك ISBN 978-9959-1-1656-7

تمهيد

بسم الله الرحمن الرحيم

وبه نستعين

الحمد لله رب العالمين، والصلاة والسلام على أشرف المرسلين سيدنا محمد وعلى آله وصحبه أجمعين.

يُعد **التحليل الإحصائي** على مر العقود الأخيرة أداة هامة تُستخدَم في استكشاف وفهم ما يُحيط بنا من ظواهر وحل الكثير من المشاكل واتخاذ القرارات الحيوية بناء على ما يتوفر من بيانات، وذلك في معظم، إن لم يكن في كل مجالات الحياة المعاصرة.

حيث يرتكز التحليل الإحصائي على استراتيجية جمع وتنظيم واستكشاف سلوك البيانات ثم تحليلها باستخدام النموذج أو النماذج المناسبة بهدف الوصول للمعلومات "الكامنة" بين طيات هذه البيانات، ومن ثمة تقديم هذه المعلومات بالطرق البسيطة المناسبة لأصحاب القرار.

ولتحقيق هذا الهدف، لابد من استخدام البرامج الإحصائية المناسبة التي توفر للباحث ما يلزمه من الأساليب والنماذج الإحصائية والرياضية لتنفيذ التحليل الإحصائي الاستكشافي والمتقدم بحسب ما تتطلب الدراسة. ويُعد برنامج R من أهم وأقوى هذه البرامج لما يتمتع به من مميزات قد لا تتوفر كلها في الكثير من البرامج الإحصائية الأخرى.

- **ما هو برنامج R ؟**

برنامج R هو نظام يحتوي على لغة برمجة بسيطة في تركيبتها، هي لغة R، والتي توفر للمستخدم مجموعة هائلة من الأوامر والدوال التي يستطيع من خلالها توظيف كافة الأساليب الإحصائية من مقاييس حسابية وجداول ورسومات بيانية ونماذج متقدمة في التعامل مع البيانات بمختلف أنواعها وأحجامها. وأهم ما يميز برنامج R هو بساطة برمجياته وتنوع دواله ومرونة رسوماته وقوة وسرعة أدائه، إضافة إلى أنه مجاني ومتوفر عبر الانترنت تحت ما يعرف بالرخصة العمومية العامة ((General Public License, (GPL)، والتي تعطي الحق للجميع بتحميله واستخدامه وإعادة توزيعه للآخرين بصورة مجانية.

- **من المستفيد من هذا الكتاب؟**

هذا الكتاب لا يستهدف المتخصصين في علم الإحصاء أو التحليل الإحصائي فقط، وإنما يخاطب بأسلوبه التدرّجي المُبسط غير المتخصصين أيضا من طلاب المرحلة الثانوية والجامعية والباحثين الأكاديميين وغيرهم من المهتمين بالتحليل الإحصائي والرياضي، ويمكن اعتباره مرجعا أساسيا في تعلم لغة R. والكتاب في العموم لا يشترط وجود أية خبرة سابقة في التعامل مع برنامج R أو أي برنامج إحصائي آخر، إلا أنه يفترض أن القارئ لديه إلمام بأساسيات علم الإحصاء بشقّيه الاستكشافي (أو الوصفي) والاستدلالي، علما بأنه يوفر

الخلفية النظرية للكثير من الأساليب الإحصائية إضافة للتطبيق العملي على البيانات. كما أنه يقدم في بعض الفصول مواضيع وأساليب إحصائية متقدمة للراغبين في استخدام البرنامج بشكل أكثر تخصصا، مثل طلبة الدراسات العليا والباحثين المتخصصين في علوم الإحصاء والرياضيات التطبيقية والاقتصاد والهندسة والطب وغيرها من التخصصات العلمية المختلفة.

- **ما هي التقنية المطلوبة إضافة للكتاب؟**

للاستفادة المُثلى من مواضيع الكتاب يُفضل تطبيق الأوامر والدوال الخاصة بكل فصل باستخدام جهاز حاسوب يعمل تحت نظام تشغيل مايكروسوفت ويندوز (MS Windows) لأي إصدار من ويندوز 7، 8، 8.1 و 10، (علما بأن معظم الأوامر المُستخدمة تعمل على أنظمة التشغيل الأخرى (مثل آبل ماكنتوش (Apple Macintosh OS))، ويتوجّب وجود اتصال بالإنترنت فقط فيما يتعلق بتحميل البرنامج لأول مرة وتحميل الحِزم الإضافية له، في حال عدم توفر ذلك لدى المُستخدم من مصدر آخر.

- **كيفية تنظيم مواضيع الكتاب:**

إن عرض المواضيع ضمن فصول الكتاب يأخذ الأسلوب التدرجي حيث نبدأ في **الفصل الأول** بتوضيح كيفية تحميل برنامج R من الانترنت أو من مصدر آخر وتنصيبه على جهاز الحاسوب والتعرف على مكوناته وخصائصه وتوضيح كيفية تحميل واستدعاء الحِزم الإضافية التابعة له.

في **الفصل الثاني** يتم تناول التعيينات وتعريف الدوال الخاصة بإنشاء المتجهات والمصفوفات في نظام R، وكيفية دمج واستدعاء واستبدال القيم في تلك المصفوفات، وكذلك تنفيذ العمليات الحسابية الأساسية عليها، وتعريف نظام القوائم.

يستمر التدرج في توضيح كيفية التعامل مع البيانات في **الفصل الثالث** عن طريق تعريف أُطر البيانات باستخدام المتجهات والمصفوفات ومحرر بيانات R. يلي ذلك استخدام الدوال الشرطية مع أُطر البيانات وتكوين أُطر البيانات الفرعية. وينتهي هذا الفصل بعرض طرق استيراد وتصدير ملفات البيانات النصية وملفات اكسل (MS Excel).

ونبدأ في **الفصل الرابع** بالتعامل الإحصائي الفعلي مع البيانات داخل نظام R، حيث نتناول أنواع البيانات وتطبيق التحليل الاستكشافي الأحادي للمتغيرات الكمية والنوعية والتمثيل البياني لها، ومن ثم التحليل الاستكشافي متعدد المتغيرات وتكوين الجداول ذات الاتجاهين. ونختتم هذا الفصل بدراسة حالة على بيانات افتراضية كتمرين عملي للقارئ.

في **الفصل الخامس** يتم تناول دوال الاحتمال والتوزيعات الاحتمالية في لغة R، حيث يتم التعرض لتكوين فراغ العينة والأحداث وحساب الاحتمالات. يلي ذلك تعريف أهم دوال التوزيعات الاحتمالية المنفصلة مثل التوزيع

المنتظم وتوزيع ذي الحدين والتوزيع الهندسي وغيرها، وكذلك أهم التوزيعات الاحتمالية المتصلة في R مثل التوزيع الطبيعي وتوزيع جاما والتوزيع الأسي وغيرها. بعد ذلك يتم تعريف دوال العزوم والدوال المولدة للعزوم للتوزيعات الاحتمالية المنفصلة والمتصلة الخاصة.

أما طرق الاستدلال الإحصائي في R فموقعها في **الفصل السادس**، والذي يتناول الاستدلالات لمعالم المجتمع الإحصائي الواحد وللمجتمعين، حيث يتم التركيز على المعالم مثل الوسط الحسابي والتباين والنسبة، واختبار تبعية العينات لنفس التوزيع. يلي ذلك تطبيق اختبارات مربع كاي لبيرسون، والاختبارات الخاصة بتحليل التباين، ثم طرق تنفيذ تحليل الارتباط والانحدار الخطي والتمثيل البياني لها. مع تخصيص البند الأخير في هذا الفصل لعرض كيفية توفيق النماذج الإحصائية، الخطية وغير الخطية بصورة عامة.

في **الفصل السابع** والأخير، يتم التطرق لبعض المواضيع المتقدمة في R والتي قد تهم بعض الباحثين المتخصصين، حيث يتم عرض الدوال الشرطية الإضافية المتخصصة، وتوضيح كيفية كتابة دوال المُستخدِم لمتغير أو أكثر حيث تُعد هذه الدوال السمة المميزة للغة R، وكذلك تناول أوامر المحاكاة أو توليد البيانات العشوائية، إضافة إلى بعض الدوال الإضافية مثل دوال الحلقات ودوال إعادة المعاينة وغيرها.

أما ملاحق الكتاب فهي مقسمة كالتالي:

الملحق (1) يحتوي على أهم جداول البيانات التي يتم التعامل مع متغيراتها ضمن فصول الكتاب.
الملحق (2) يعرض كيفية تطبيق خيارات الرسم الإضافية والتحكم بنوع وطبيعة الخطوط والألوان، وعناوين الرسومات الأساسية والهامشية، وتنسيق المحاور الأساسية والفرعية، وكيفية دمج أكثر من تمثيل بياني في شكل واحد، وكيفية التعامل مع الرسم التفاعلي وغير ذلك من الخيارات.
الملحق (3) يشتمل على كافة الأوامر والدوال المستخدمة في الكتاب بحسب الترتيب الأبجدي لها.
الملحق (4) يشتمل على حِزم R الإضافية والتي يحتاجها المُستخدم لتنفيذ بعض الأوامر الخاصة.

- **تقدير واعتزاز:**

دائما ما يتملكني، عند التعامل مع علم الإحصاء، شعور بالاعتزاز بهذا العلم الجميل الراقي الذي يكفيه فخرا أنه الأكثر ذكرا في القرآن الكريم وذلك في أكثر من سورة؛ (... مال هذا الكتاب لا يغادر صغيرة ولا كبيرة إلا **أحصاها** ...، الكهف (49))، (لقد **أحصاهم** وعدهم عدا، مريم (94))، (... **أحصاه** الله ونسوه ...، المجادلة (6))، وغيرها من الآيات الكريمة.

المؤلف

المحتويات

الفصل الأول

مقدمة: تنصيب وتشغيل نظام R

(Introduction: Installing and Operating the R Environment)

1.1 تعريف نظام R (Definition of the R Environment)

1.1.1 مميزات نظام R (Advantages of R Environment)

2.1 تحميل وتنصيب نظام R (Downloading and Installing R)

1.2.1 تحميل النظام من موقع مشروع R على الانترنت
(Downloading R from the R Project Website)

2.2.1 تنصيب نظام R على الحاسوب (Installing R on PC)

3.1 التعرّف على لوحة مراقبة R (Exploring the R Console)

1.3.1 مكونات لوحة مراقبة R (Components of R Console)

2.3.1 بدء التعامل مع لوحة مراقبة R (Start Working with R Console)

4.1 تحميل واستدعاء حِزم R (Downloading and Calling the Packages of R)

1.1 تعريف نظام R (Definition of the R Environment)

إن نظام أو بيئة R يعد تركيبة متكاملة من الوسائل البرمجية التي تستخدم لتنفيذ الأوامر الرياضية والتحليلات الإحصائية، بما في ذلك الرسومات البيانية، من خلال التعامل مع البيانات. فهو ببساطة برنامج رياضي إحصائي متوفر عبر الانترنت تحت ما يعرف بالرخصة العمومية العامة ((General Public License, (GPL)، والتي تعطي الحق بتحميله واستخدامه مجانا وكذلك إعادة توزيعه للآخرين.

وترجع القصة التاريخية لظهور نظام R باختصار إلى أن ريك بيكر (Rick Becker) وجون شامبرز (John Chambers) قاما معا في بداية الثمانينات بابتكار لغة برمجة إحصائية أطلقا عليها اسم لغة S[1]. بعد ذلك قام كلا من الباحثين روس إيهاكا (Ross Ihaka) وروبرت جنتلمان (Robert Gentlman) من جامعة أوكلاند في نيوزيلاندا بكتابة نسخة مختصرة من لغة S لأغراض تعليمية، ويُعتقد أن التسمية الجديدة R كانت مستوحاه من الأحرف الأولى لإسميهما.

في عام 1995 قام مارتن ماكلر (Martin Maechler) بإقناع كلا من روس وروبرت لجعل لغة R متاحة لجميع المستخدِمين بصورة مجانية تحت الرخصة العمومية العامة. وفي التاسع والعشرين من شهر فبراير في العام 2000 تم إصدار النسخة رقم 1.0.0 من لغة R رسميا، ولا تزال النسخ تتوالى حتى يومنا هذا[2]. وتُشرف مختبرات بِل (AT&T Bell Laboratories) حاليا على تطوير نظام R وبرمجياته الملحقة بإدارة ريك بيكر، جون شامبرز، وآلان ويلكسِ (Allan Wilks).

1.1.1 مميزات نظام R (Advantages of R Environment)

يتمتع نظام R بعدة خصائص ومزايا تجعله في الواقع من أفضل البرامج الإحصائية المتوفرة في مجال تحليل البيانات مقارنة بالبرامج الأخرى. ومن أهم هذه المزايا ما سنعرضه في النقاط التالية:

1. برنامج R هو برنامج مجاني ومتاح لجميع المستخدمين حول العالم عبر الانترنت.
2. يعمل R على كل أنظمة التشغيل المعروفة؛ مايكروسوفت ويندوز[3] (MS Windows)، لينكس (Linux)، وآبل ماكنتوش (Apple Macintosh OS).

---

[1] تم تطوير لغة S لاحقا لتصبح الحزمة الإحصائية المتكاملة المعروفة باسم S-Plus.

[2] تم اعتماد أحدث نسخة متوفرة من نظام R وقت إعداد هذا الكتاب وهي النسخة 3.2.3 .

[3] نشير هنا إلى أن نظام R المستخدم في هذا الكتاب هو ذلك الخاص بنظام ويندوز، علما بأن الأوامر والدوال المُعرفة في لغة R هي نفسها في كل أنظمة التشغيل الأخرى.

3. يحتوي نظام R على تركيبة واسعة جدا من الأدوات والدوال الرياضية التي تستخدم لإجراء الحسابات على معظم الأنظمة الرياضية مثل المصفوفات والمتجهات وغيرها.
4. يضم R حِزَم (Packages) خاصة بالتحليلات الرياضية والإحصائية المتقدمة والمركّبة، والتي يتم تحديثها بصورة دورية، إضافة إلى استحداث حزم جديدة مواكبة للتطور المستمر في الأساليب الإحصائية بشكل دائم.
5. إمكانية إجراء حسابات إضافية على نتائج التحليلات الإحصائية المُتحصل عليها، وكذلك تعديل الرسومات البيانية بطرق متعددة ومرنة.
6. لغة البرمجة المستخدمة في بيئة R هي لغة بسيطة وفعالة في نفس الوقت وتضم دوال شرطية (Conditional Functions)، حلقات تكرارية (Loops)، ودوال يمكن تجميعها أو تعريفها من قِبل المستخدم (User-defined Functions).
7. إمكانية استيراد وتصدير ملفات البيانات (وأحيانا النتائج) من وإلى البرامج الإحصائية الأخرى مثل اكسل (Excel)، (S-Plus)، (SPSS)، (SAS)، وغيرها.
8. يتيح نظام R استخدام مساهمات الآخرين من دوال مُعرّفة أو قواعد بيانات مثلما يتيح لك المشاركة في تطوير الحِزم المتوفرة أو المساهمة بإصدار حِزم جديدة، وهذا يندرج تحت طبيعة ما يُعرف ببرامج المصدر المفتوح (Open-source Software's).

2.1 تحميل وتنصيب نظام R (Downloading and Installing R)

إن التعليمات المتعلقة بتحميل وتنصيب نظام R تعتمد بصورة كبيرة على الكيان المادي ونظام التشغيل الخاص بالمستخدِم. ويمكن للقارئ أن يجد كل المعلومات الإضافية أو المتقدمة المتعلقة بأنظمة تشغيل الحاسوب وما يتعلق بها من خيارات حول النسخة المناسبة له في موقع مشروع R (R Project) وهو[1] "**http://cran.r-project.org**". إلا أننا سوف نقوم بشرح خطوات تحميل وتنصيب نظام R الخاص بنظام تشغيل ويندوز[2] بصورة مبسطة لا تتطلب خبرة كبيرة في استخدام الانترنت أو برمجيات الحاسوب من قِبل المستخدِم. وننوه هنا إلى أن برنامج R لا يتطلب اتصالا بالإنترنت لكي يعمل على الحاسوب، لكن تحميله، (إن لم يتحصل عليه القارئ مسبقا من مصدر آخر كملف تنفيذي جاهز)، يتطلب ذلك.

1.2.1 تحميل النظام من موقع مشروع R على الانترنت

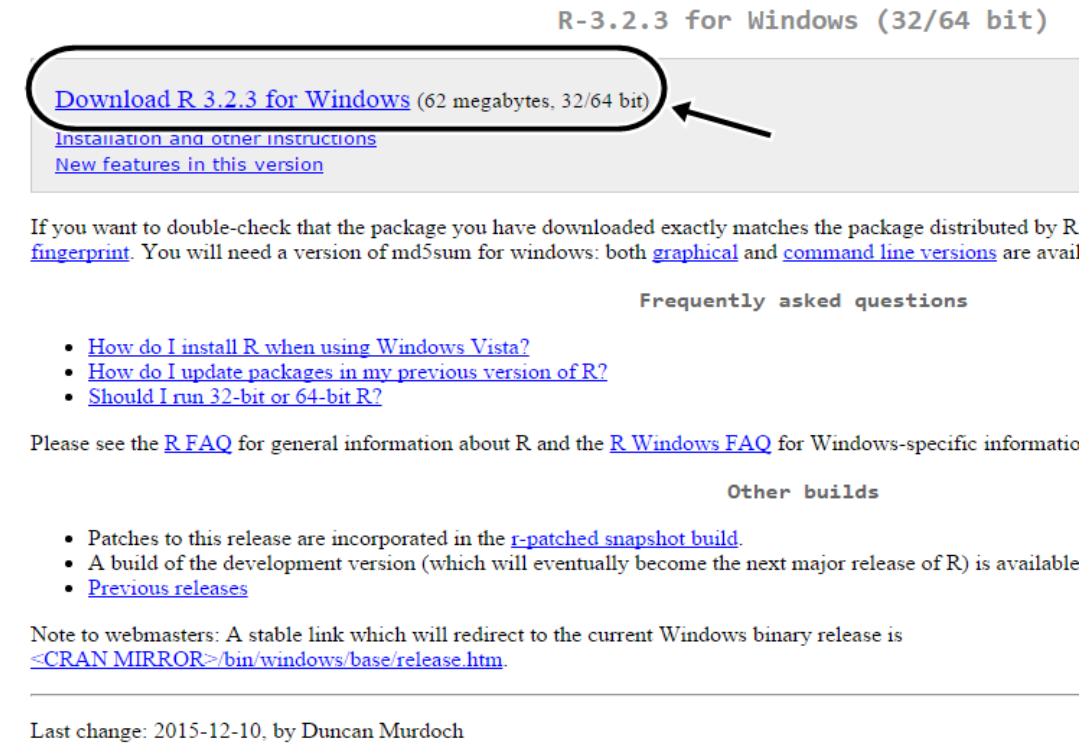
(Downloading R from the R Project Website)

يمكن تحميل النسخة الأحدث من نظام R والخاصة بنظام ويندوز من خلال الموقع الرسمي لمشروع R عبر الرابط: "**http://cran.r-project.org/bin/windows/base**". وشكل (1.1) يمثل صفحة الموقع

[1] **CRAN** هو اختصار (Comprehensive R Archive Network) ويعني شبكة محفوظات R الشاملة.

[2] هذه الخطوات تصلح لأنظمة تشغيل ويندوز 7 (ذات 32-bit أو 64-bit)، ويندوز 8، وويندوز 10.

التي تحتوي على أمر تحميل الملف التنفيذي لنسخة R الخاصة بنظام ويندوز، ونُذكّر هنا مجددا أن النسخة الأحدث من R وقت إعداد الكتاب كانت **R 3.2.3**. أما إذا توفرت لديك هذه النسخة أو نسخة أحدث منها فيمكنك تجاوز هذا البند إلى البند (3.1).

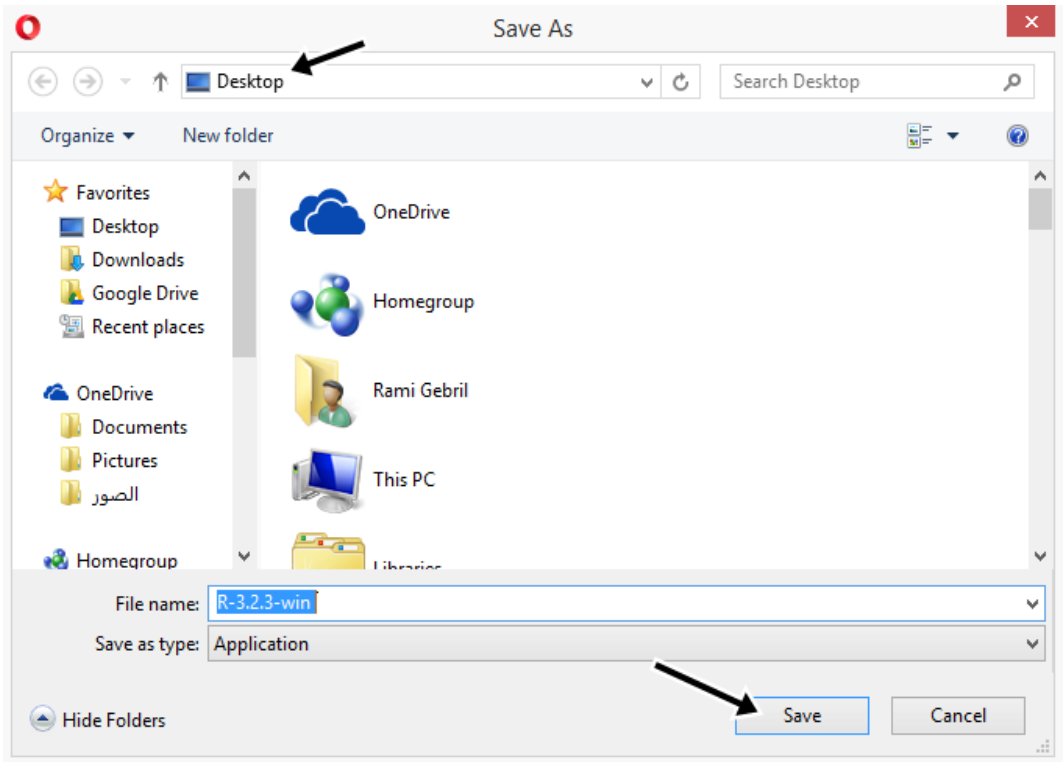


شكل 1.1: الصفحة الالكترونية لموقع مشروع R لتحميل البرنامج لنظام ويندوز

ولتحميل نظام R نقوم باختيار الأمر المُشار إليه بالسهم في شكل (1.1) فتظهر نافذة جديدة بها أمر التحميل[1] كما هو موضح في شكل (2.1). يتم بعد ذلك إعطاء أمر التحميل لتخزين البرنامج التنفيذي الذي سيحمل اسم النسخة الأخيرة (وهي هنا **R-3.2.3-win.exe**) على جهازك[2].

[1] قد تختلف نافذة أمر التحميل باختلاف كُلاً من نوع متصفح الانترنت ونوع برنامج التحميل المُستخدَم في جهازك.

[2] يُفضل تخزين البرنامج على مسار سطح المكتب (Desktop) في جهازك لمواكبة شرح تنصيب البرنامج في هذا البند.



شكل 2.1: تحميل برنامج R إلى جهاز الحاسوب

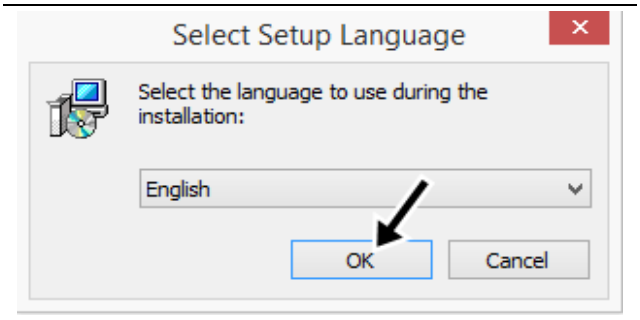
وإذا ما تم تحميل البرنامج التنفيذي على سطح المكتب ستجد بعد الانهاء الأيقونة التالية موجودة عليه (الشكل (3.1))؛

شكل 3.1: أيقونة الملف التنفيذي لنظام R على سطح المكتب

2.2.1 تنصيب نظام R على الحاسوب (Installing R on PC)

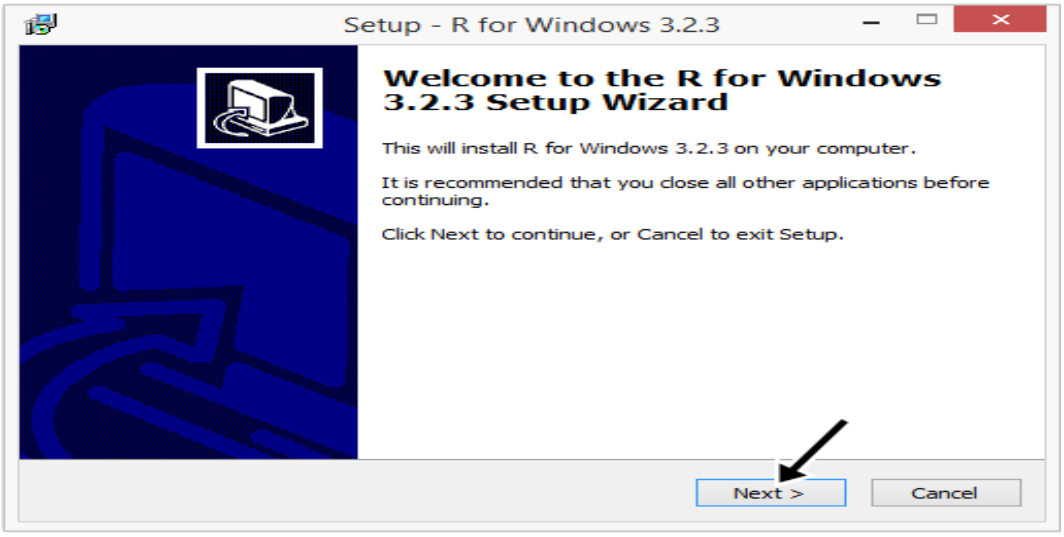
في هذا البند، سنقوم بشرح خطوات تنصيب نظام R على جهاز الحاسوب خطوة بخطوة من خلال الاستعانة بالأشكال التوضيحية (النوافذ) العشرة التالية التي تبدأ بالشكل (4.1 أ) وتنتهي بالشكل (4.1 ي).

بداية نقوم بفتح الملف التنفيذي على سطح المكتب فتظهر النافذة التالية (شكل (4.1 أ)):



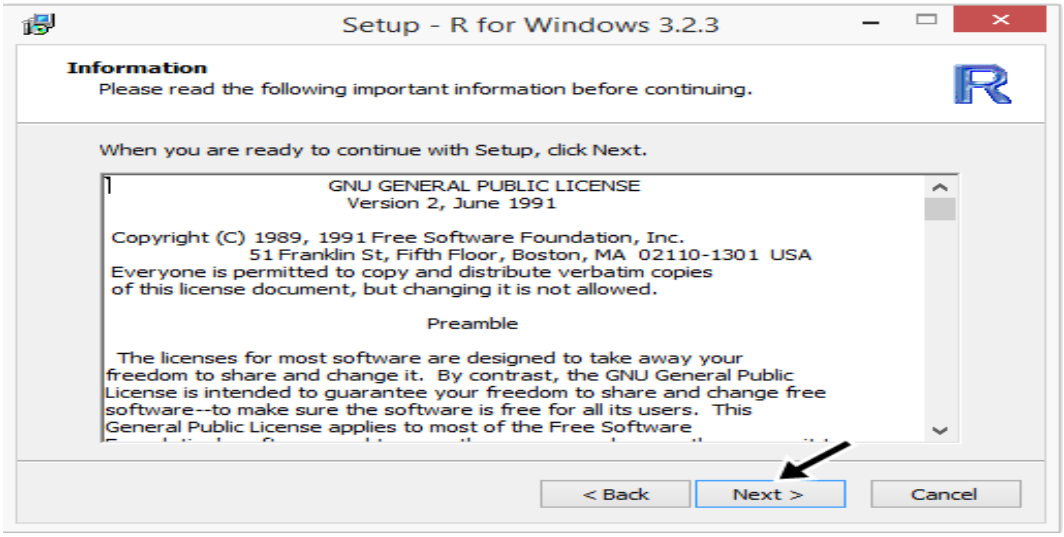
شكل 4.1 أ: النافذة الأولى في إعدادات تنصيب R

وحيث أن اللغة العربية غير متوفرة في إعدادات التنصيب، (وغير متوفرة أيضا في تشغيل نظام R حتى الآن)، سيتم متابعة الإعدادات باللغة الإنجليزية، ويتم في الشكل (4.1 أ) اختيار (OK). نختار (Next) في الشكل (4.1 ب)، والذي يمثل النافذة الترحيبية لتنصيب البرنامج، للمتابعة.



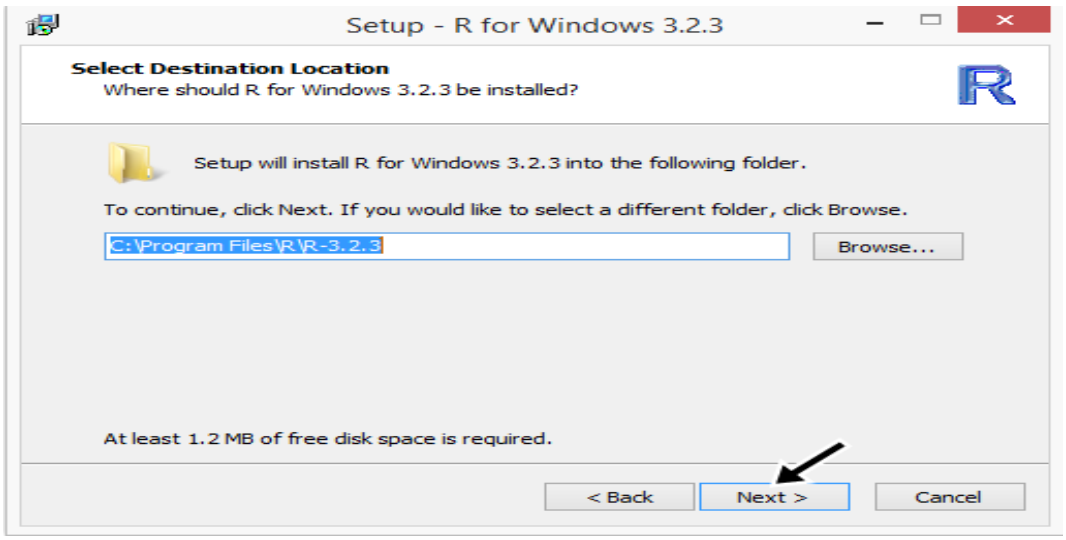
شكل 4.1 ب: النافذة الترحيبية في إعدادات تنصيب R

في النافذة الثالثة، (الشكل (4.1 ج))، ستجد شروط الرخصة العمومية العامة (GPL)، والتي بعد الانتهاء من قراءتها يمكنك اختيار (Next).



شكل 4.1 ج: النافذة الثالثة في إعدادات تنصيب R

يتم في الخطوة التالية، (الشكل (4.1 د))، اختيار موقع الحافظة التي سيتم فيها تخزين الملفات الداخلية لبرنامج R، وهو عادة ما يكون على تجزئ القرص الصلب الذي يعمل عليه نظام ويندوز على جهازك، (والذي غالبا ما يكون C:\)، وننصح المستخدمين المبتدئين بعدم تغييره. نختار (Next) في نفس الشكل ونمضي قدما.

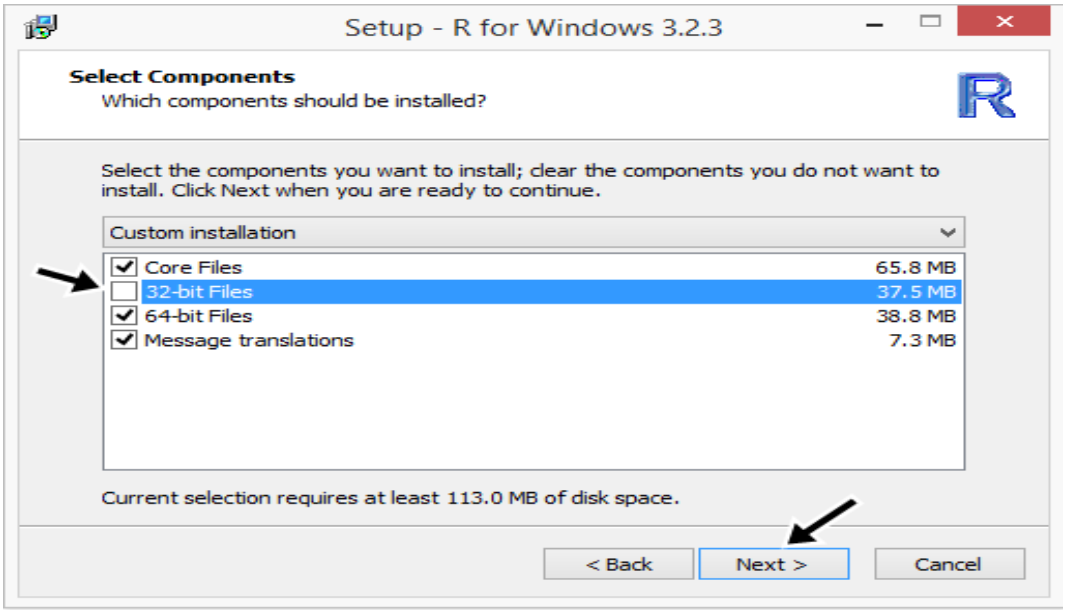


شكل 4.1 د: النافذة الرابعة في إعدادات تنصيب R

في الشكل (4.1 هـ) ستجد قائمة بأنواع الملفات التي سيقوم نظام R بتنصيبها على الحاسوب، فإذا كان جهازك يعمل على نظام ويندوز 64-bit ستكون الاختيارات كما هو معروض في الشكل[1]، وأما إذا كان جهازك يعمل

[1] إذا ما تم الإبقاء على اختيار تحميل ملفات (32-bit Files) بالإضافة لملفات (64-bit Files)، فسيظهر لك في نهاية عملية التنصيب أيقونتين للبرنامج؛ إحداهما لتشغيل البرنامج تحت نظام 64-bit والأخرى لتشغيله بنظام 32-bit، ويمكنك

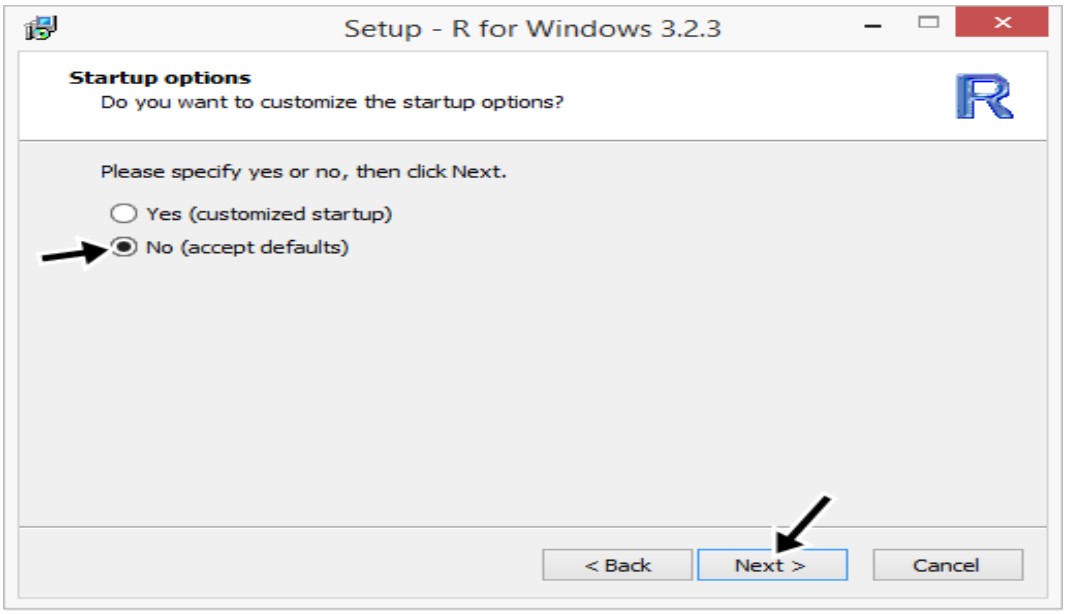
على ويندوز 32-bit فستجد أنه تم اختيار (32-bit Files) تلقائيا. وحيث أن الجهاز الذي نستخدمه في التطبيق يعمل على نظام 64-bit، فإننا سنختار ملفات 64-bit من ضمن الخيارات المتاحة ونلغي اختيار ملفات 32-bit. بعد ذلك يتم اختيار (Next) كما هو موضح.



شكل 4.1 هـ: النافذة الخامسة في إعدادات تنصيب R

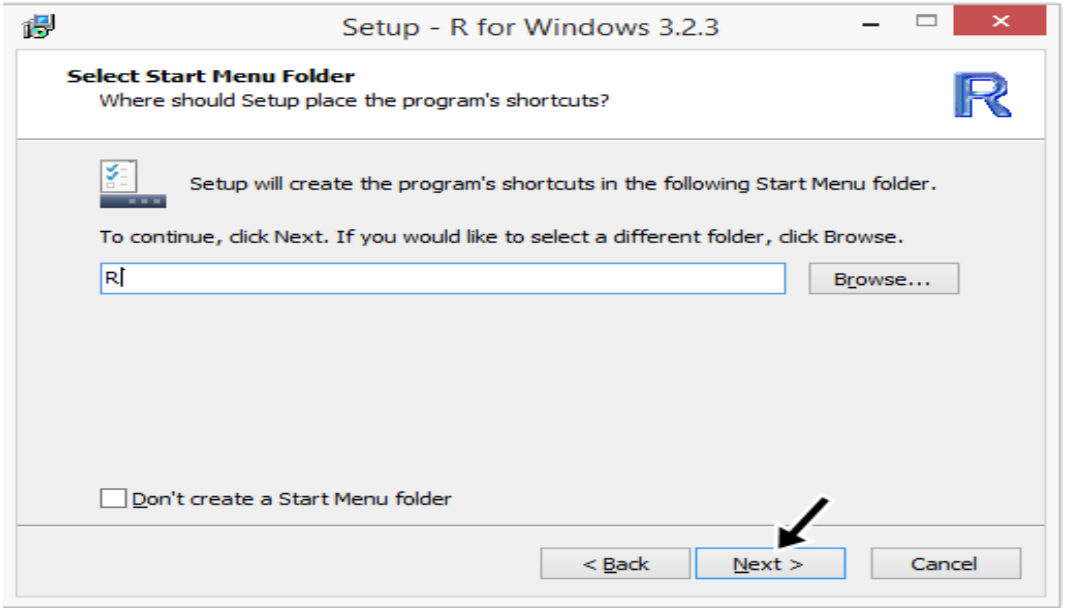
في النافذة السادسة، (الشكل (4.1 و))، نجد الخيارات الخاصة بتعديل تشغيل البرنامج ضمن لوحة البدء في ويندوز (Start up) موجودة في الاختيار الأول، وأيضا ننصح المبتدئين بعدم اختيارها واستخدام الخيار الثاني الموجود افتراضيا، ثم اختيار (Next).

في جميع الأحوال اختيار العمل بأحد النظامين، إلا أنه في العموم يفضل استخدام البرنامج من خلال أيقونة 64-bit عند توفر الخيارين.



شكل 4.1 و: النافذة السادسة في إعدادات تنصيب R

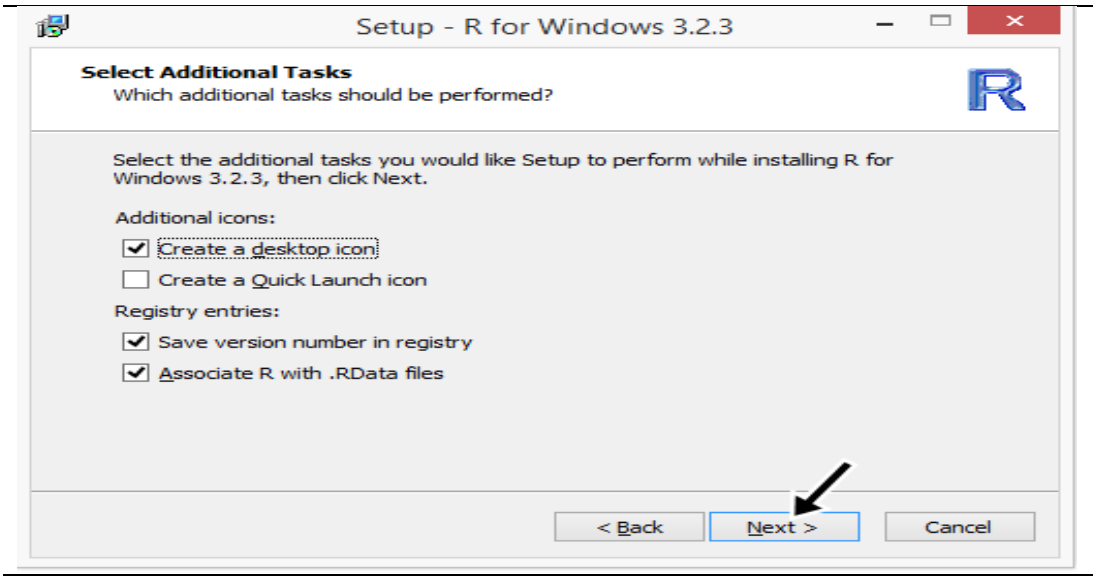
النافذة السابعة، (الشكل (4.1 ز))، ستكون لتحديد الحافظة التي سيقوم نظام R بتخزين الملفات المسقبلية فيها بصورة تلقائية، حيث يُنشِئ البرنامج التنفيذي حافظة افتراضية لقائمة البدء باسم R، وننصح المستخدم أيضا بإبقائها كما هي واختيار (Next).



شكل 4.1 ز: النافذة السابعة في إعدادات تنصيب R

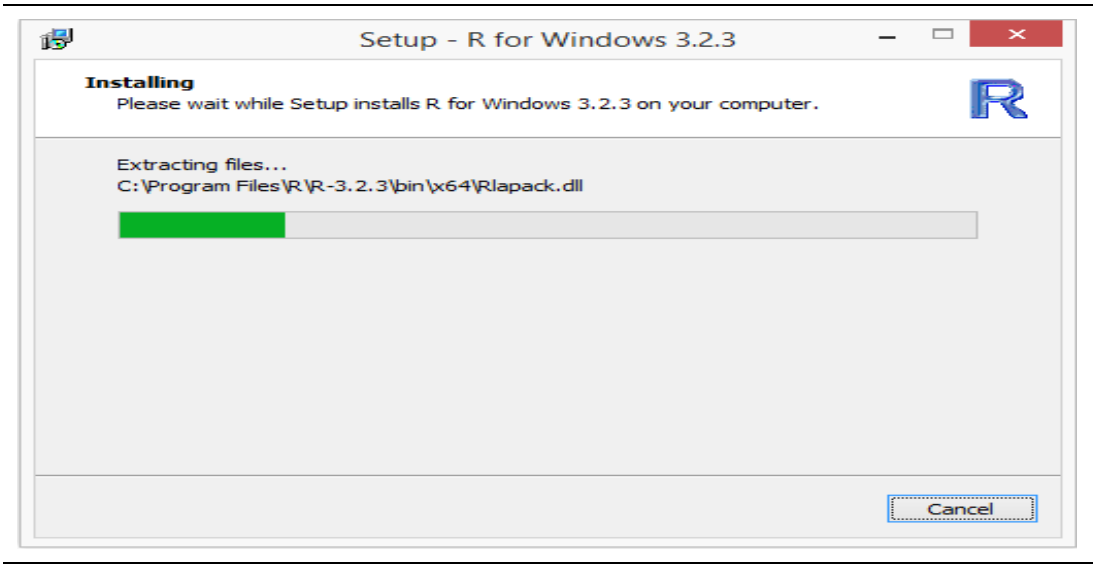
في الشكل (4.1 ح) والذي يمثل النافذة الثامنة، يمكنك الاختيار من بعض خيارات الإعداد الإضافية التي تشمل خيار إنشاء أيقونة تشغيل لنظام R على سطح المكتب (ننصح بإبقائها) وخيار إنشاء أيقونة إطلاق سريع

(Quick Launch)، والتي يمكن للمستخدِم اختيار أو عدم اختيار إنشائها حيث أن الأيقونة على سطح المكتب تفي بالغرض. ثم نضغط (Next) بعد ذلك.



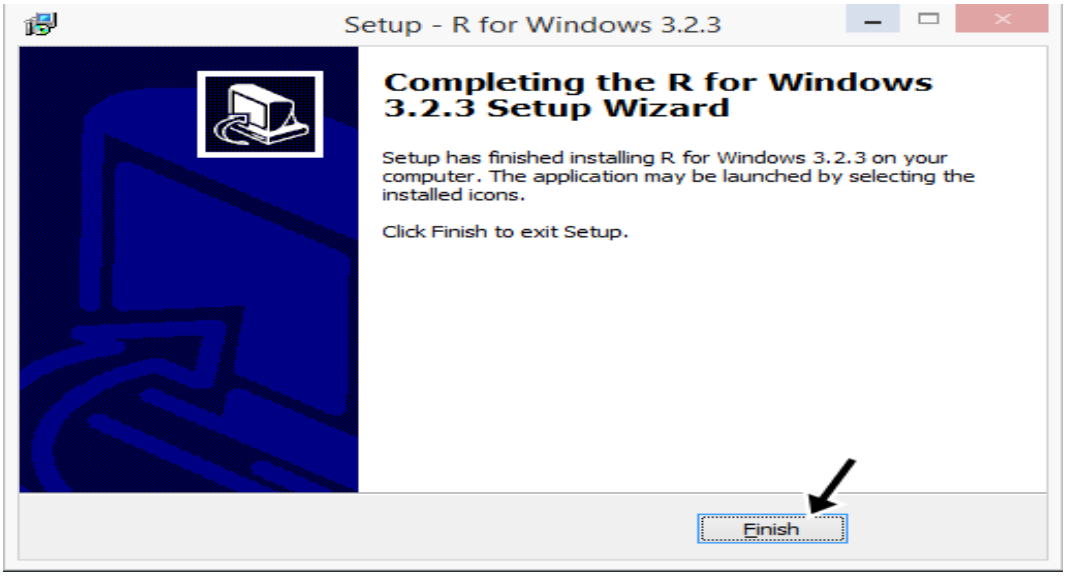
شكل 4.1 ح: النافذة الثامنة في إعدادات تنصيب R

بعد اختيار (Next) في النافذة السابقة، ستظهر نافذة استخراج ملفات نظام R وتنصيبها في المسار المحدد كما هو ظاهر في النافذة التاسعة؛ الشكل (4.1 ط)؛



شكل 4.1 ط: النافذة التاسعة في إعدادات تنصيب R

ولن يستغرق ذلك طويلا، (بغض النظر عن سرعة المعالج في جهازك)، علما بأنه لا توجد اختيارات في تلك النافذة، وسرعان ما ستظهر النافذة العاشرة والأخيرة، (الشكل (4.1 ي))، التي تفيد بانتهاء تنصيب نظام R على حاسوبك.



شكل 4.1 ي: النافذة العاشرة في إعدادات تنصيب R

قم باختيار إنهاء (Finish) وستجد أنه قد تم إنشاء أيقونة تشغيل لنظام R تحت نظام[1] 64-bit على سطح المكتب[2] كما هو ظاهر في الشكل التالي، (شكل (5.1))؛

شكل 5.1: أيقونة تشغيل نظام R على سطح المكتب

3.1 التعرّف على لوحة مراقبة R (Exploring the R Console)

بعد أن تم تنصيب نظام R على الحاسوب، يمكننا الآن الانتقال للتعرّف على مكونات **لوحة مراقبة R** (R Console)، والتي تمثل ببساطة النافذة التي سيتم التعامل مع نظام R من خلالها، ونقصد بكلمة "التعامل" هنا كل ما يتعلق بإدخال الأوامر وإظهار وعرض النتائج وتحميل الحزم الإضافية وإدارة الملفات وطلب المساعدة وغيرها من الأدوات اللازمة في استخدام لغة R.

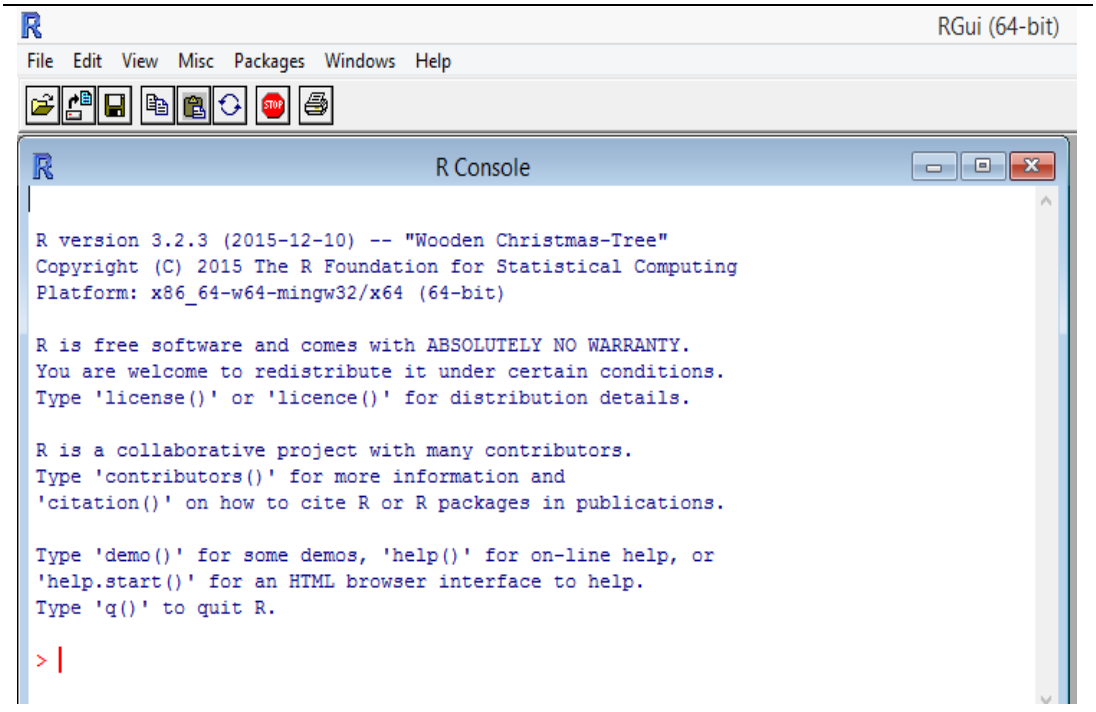
[1] أو تحت نظام 32-bit إذا كان هو النظام الذي تم اختياره خلال عملية التنصيب.

[2] يمكن للقارئ، كإجراء اختياري، نقل الملف التنفيذي لبرنامج R (الظاهر في الشكل (3.1)) من سطح المكتب إلى إي مسار آخر، كما ننصح بتخزينه في جهاز الذاكرة الوميضية أو الفلاش (Flash Memory) أو اسطوانة مدمجة (CD) أو أي وسيلة تخزين أخرى لإعادة استخدامه أو نسخه لجهاز حاسوب آخر عند الضرورة.

1.3.1 مكونات لوحة مراقبة R (Components of R Console)

سنقوم في هذا البند بعرض أهم المكونات الأساسية للوحة مراقبة R بشكل مختصر مراعاة للقارئ المبتدئ في بداية استخدامه لنظام R، وسيتم التدرج في عرض باقي المكونات بصورة أكثر توسعا بحسب الحاجة لها في الفصول القادمة. بعد فتح أيقونة تشغيل نظام R الخاصة بنظام 64-bit أو 32-bit، (المبينة في الشكل (5.1))، ستُفتح نافذة البرنامج كما هو موضح في الشكل (6.1)، وسيلاحظ القارئ أن الإطار الخارجي للنافذة يشبه إلى حد كبير الأُطر التقليدية لمعظم البرامج التي تعمل تحت بيئة ويندوز، حيث أن إطار النافذة يحتوي على شريط العنوان، (الذي يظهر عليه شعار واسم البرنامج[1] **RGui**)، وشريط الأدوات أسفل منه (ويظهر عليه قوائم File، Edit، View، وغيرها). ويظهر بداخل الإطار الخارجي النافذة الداخلية المخصصة للكتابة بلغة R وهي لوحة مراقبة R، ويظهر عليها شعار البرنامج أيضا ومصطلح "R Console".

داخل هذه النافذة ستجد رقم الإصدار الخاص بنظام R (وهي النسخة 3.2.3 كما أسلفنا)، وتاريخ الإصدار، والاسم المستعار (وهو "Wooden Christmas-Tree" لهذه النسخة)، إضافة إلى وصف بسيط لنظام R، والذي يُعطى المستخدم حق إعادة توزيع البرنامج تحت بعض الشروط، كما يوضح للمستخدم بعض الأوامر الخاصة البسيطة، (مثل `license()`، `demo()`، `help()`، وغيرها)، والتي يمكن تنفيذها عند الحاجة[2].



شكل 6.1: نافذة لوحة مراقبة R الاعتيادية

[1] **RGui** هو اختصار (R Graphical User Interface) وهو مصطلح يعني الواجهة التي يتعامل معها المستخدم مع R بطريقة الإشارة والنقر (Points-and-Clicks) على نظام القائمة (Menu) باستخدام الفأرة.

[2] سيتم توضيح طريقة استخدام هذه الأوامر بحسب التدرّج المُتبع في الكتاب.

2.3.1 بدء التعامل مع لوحة مراقبة R (Start Working with R Console)

سنبدأ في هذا البند التعامل الفعلي[1] مع لوحة مراقبة R بحيث نوضح كيفية تنظيم جلسة التعامل مع لغة R خطوة بخطوة عن طريق إدخال الأوامر وقراءة المخرجات بصورة متدرجة. سنلاحظ في لوحة مراقبة R في السطر الأخير وجود رمز "أكبر من" (>)، الذي يُقرأ من اليسار، متبوعا بالمؤشر الوميض (|) باللون الأحمر. هذا المؤشر سيشير دائما إلى السطر الذي يتم فيه إدخال الأوامر أو تعديلها بعد كتابتها، وهو ما يُعرف **بسطر الأوامر** (Command line).

نقوم في البداية بإنشاء حافظة جديدة (New Folder)، وليكن موقعها على سطح المكتب لتسهيل المتابعة، حيث نعطيها اسم معين وليكن "myR" مثلا. هذه الحافظة ستشتمل لاحقا على كل الملفات التي سنتعامل من خلالها مع نظام R في هذا الكتاب بالتدريج. بعد ذلك سنقوم بتحديد **مسار عمل نظام R** (R Working Directory) على الحاسوب وتخزين ملف العمل للجلسة الحالية، وذلك باتباع الخطوات التالية داخل لوحة مراقبة R:

1. نقوم أولاً بكتابة الأمر[2]

```
> getwd()
```

للتعرف على مسار العمل الحالي في نظام R، ثم الضغط على زر الإدخال (Enter) في لوحة المفاتيح لتنفيذ الأمر.

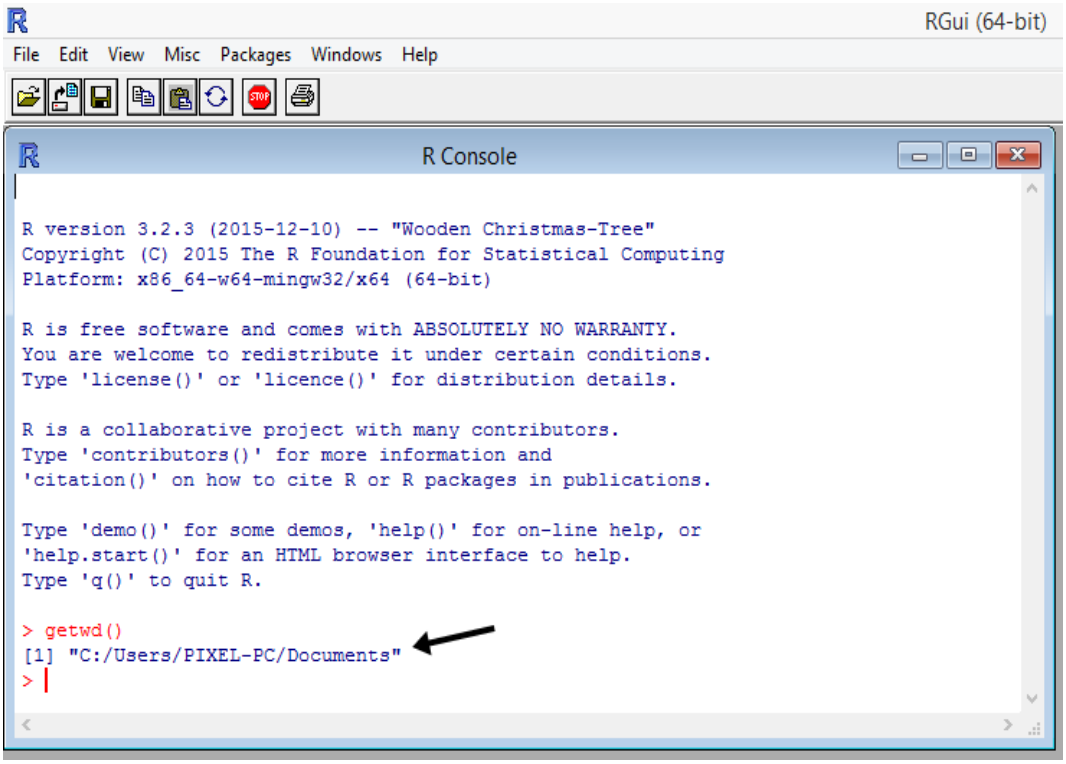
وغالبا ما يختلف مسار عمل R الافتراضي من جهاز حاسوب لآخر باختلاف إعدادات نظام ويندوز واختلاف أسماء المسارات المعرّفة على القرص الصلب، إلا إن ذلك لا يمثل مشكلة على الإطلاق لأننا سنقوم بتغيير مسار عمل R إلى مسار سطح المكتب في الخطوة التالية، والشكل (7.1) يوضح تنفيذ الخطوة (1).

لاحظ أن مسار العمل الحالي المُوضح في الشكل (1.7) سيختلف بالتأكيد في الجزء الأوسط عن ذلك الذي سيظهر في جهازك كما تم التنويه إليه.

---

[1] ننصح القارئ بداية من هذه المرحلة بضرورة تشغيل برنامج R، إن لم يكن قد بدأ بذلك، ومتابعة تنفيذ إرشادات الكتاب أولاً بأول.

[2] لتمييز أوامر لغة R التنفيذية عن باقي المصطلحات المكتوبة باللغة الإنجليزية أو العربية في الكتاب تم استخدام نوع الخط "Courier New" لتلك الأوامر، وهو نوع الخط الافتراضي في لوحة مراقبة R، وتم استخدام نوع الخط " Times New Roman" لباقي المصطلحات باللغة الإنجليزية، والخط "Simplified Arabic" للغة العربية.



شكل 7.1: التعرف على مسار عمل R الحالي

2. لتغيير مسار ملفات R من المسار الافتراضي إلى مسار الحافظة "myR" على سطح المكتب نقوم بكتابة الأمر التالي[1] في نافذة لوحة مراقبة R:

```
> setwd(dir="C:/Users/PIXEL-PC/Desktop/myR")
```

ثم نضغط زر الإدخال. ستلاحظ عدم ظهور أية مخرجات في النافذة، وذلك لأن هذا الأمر تم تنفيذه في خلفية البرنامج ولم يَعرض نتائج مُشاهَدة. ويمكنك التأكد من تغيير مسار عمل R من المسار القديم إلى المسار الجديد وذلك بتنفيذ الأمر `getwd` من جديد حيث سيظهر مسار الحافظة "myR" على سطح المكتب كمسار العمل الحالي كما يلي:

```
> getwd()
[1] "C:/Users/PIXEL-PC/Desktop/myR"
```

3. الآن نقوم بحفظ ما يُعرف في نظام R **بمساحة** أو **ملف العمل** (R Workspace) للجلسة الحالية، والذي سيحتوي دائما على كل الإعدادات والبيانات التي تم التعامل معها حتى وقت حفظ الجلسة.

[1] لاحظ حيث أنه لن يظهر المقطع الأوسط (`Users/PIXEL-PC`) كما هو في جهازك، فإنك فعليا ستقوم بتغيير الأمر فقط فيما يتعلق بالمسار الأخير الذي يشمل الحافظة "`Documents`".

ولنعطي الاسم "work1" كاسم لملف العمل الحالي مثلا، فنقوم باستخدام **أمر الحفظ** `save.image` بالصورة التالية:

```
> save.image("C:/Users/PIXEL-PC/Desktop/myR/
work1.RData")
```

ولاحظ ضرورة إدراج الامتداد[1] "`.RData`" للملف "work1"، حيث يتم إدراج هذا الامتداد مع معظم الملفات التنفيذية وملفات البيانات التي يتم تعريفها في نظام R.

بعد تنفيذ الأمر الأخير، يمكنك ملاحظة وجود ملف العمل الذي أعطيناه الاسم "work1" داخل الحافظة "myR". وحيث أننا لم نقم بإدخال أية أوامر تُذكر، (غير تغيير مسار العمل)، فإن ملف العمل الحالي "work1" يعتبر فارغا، وقد تم تنظيم جلسة العمل مع R بهذه الطريقة منذ البداية بغرض توجيه انتباه القارئ إلى ما سيتم عرضه من أوامر لاحقا. وقبل البدء بعرض هذه الأوامر سنستعرض فيما يلي الملاحظات التالية التي تحتوي ما قد يغيب عن بعض المستخدمين من أمور تقنية وعملية أثناء التعامل مع نظام R.

ملاحظات هامة:

- لا يتم تنفيذ أي أمر مكتوب في لوحة مراقبة R إلا بعد الضغط على زر الإدخال في لوحة المفاتيح.
- يمكن اختيار أي مسار عمل آخر غير مسار سطح المكتب، إلا أن الأخير يكون أكثر ملائمة للعمل من حيث البساطة وسرعة الوصول للملفات.
- توجد أوامر تنفيذية في لغة R لها مخرجات مُشاهَدة، (مثل الأمر `getwd()`)، وأخرى يتم تنفيذها بدون ظهور مخرجات على النافذة (مثل الأمر `setwd(dir="*")`، حيث يتم كتابة مسار العمل المُراد في موضع الرمز *).
- يجب الالتزام بطريقة كتابة أوامر لغة R كما هي موضحة ضمن الكتاب وعدم استخدام مسافات، نقاط، أقواس، أو أية رموز أخرى غير المنصوص عليها في الأمر نفسه لأن ذلك سيؤدي لحدوث خطأ في التنفيذ وستظهر رسالة في نافذة لوحة مراقبة R للتأكيد على وجود ذلك الخطأ[2]، وكذلك يجب مراعاة استخدام أحرف اللغة الإنجليزية الصغيرة أو الكبيرة كماهي في نص الأمر.
- يمكن استخدام الفأرة لتحديد مكان المؤشر الوميض على السطر الذي يتم فيه كتابة الأمر، كما يمكن استخدامها لتظليل الأوامر أو أجزاء منها ونسخها ومن ثمة إعادة استخدامها سواء في نفس السطر أو في سطر جديد.

---

[1] امتداد الملف هي الأحرف المختصرة التي تظهر بعد النقطة التي تلي اسم الملف، وتعبر عن الصيغة الذي يمثلها هذا الملف.

[2] عند حدوث ذلك معك يمكنك ببساطة مراجعة الأمر الذي قمت بإدخاله وكتابته من جديد بصورته الصحيحة في سطر جديد عن طريق الضغط في لوحة المفاتيح على زر الخروج (ESC) ثم الضغط على زر الإدخال.

- من ضمن الأدوات المساعدة إضافة للفأرة، الأسهم الأربعة الموجودة في لوحة المفاتيح، حيث يمكن استخدام السهمين (→) و (←) للتنقل يمينا ويساراً في نفس السطر. أما السهم العلوي (↑) فيقوم باسترجاع الأوامر السابقة التي تمت كتابتها كلما ضغطت عليه، ويقوم السهم السفلي (↓) بالعملية العكسية للسهم العلوي.
- إذا ما تم كتابة الرمز (#) في أي مكان في سطر الأوامر فإن كل ما يتم كتابته بعد هذا الرمز لن يتم تنفيذه بعد الضغط على زر الإدخال، ويُسمى ذلك في نظام R بإدراج **الملاحظات** (Comments). جرّب أن تُنفذ الأمر `#getwd()` على سبيل المثال وستلاحظ عدم ظهور نتيجة.
- كلما تم كتابة أمر أو ملاحظة أو ظهور نتيجة (أو حتى إذا قمت بضغط زر الإدخال بدون كتابة أي شيء) فإن المؤشر سينتقل للسطر التالي في الأسفل، إلا أنه يمكنك متى أردت استخدام أمر **المسح** (CTRL+L) في لوحة المفاتيح لمسح كل الموجود داخل النافذة، علما بأن ذلك الأمر لا يقوم بإلغاء ما تم إدخاله أو تنفيذه من أوامر من ذاكرة نظام R خلال الجلسة الحالية، بل يقوم "بتنظيف الشاشة" في لوحة مراقبة R.

سنقوم الآن باستخدام لوحة مراقبة R لإجراء بعض الحسابات البسيطة لتوضيح الطريقة النمطية في التعامل مع **المُدخلات** (Inputs)، **والمُخرجات** (Outputs)، والملاحظات. إن نظام R يُعد، إلى جانب كونه برنامج تحليل رياضي وإحصائي متقدم، آلة حاسبة متقدمة يمكنك من خلالها تنفيذ كل العمليات الرياضية البسيطة والمُركبة، ولنقم بتنفيذ الآتي:

كتابة ملاحظة[1] باللغة العربية[2] بأن هذه هي عملية جمع الرقمين 5 و2 (في السطر الأول)، ثم تنفيذ العملية الحسابية لجمع العددين 2 و5 (في السطر الثاني)، وعرض الناتج (في السطر الثالث)، ونُذكّر المُستخدِم دائما بالانتباه لتغيير لغة الكتابة في لوحة المفاتيح من اللغة العربية إلى الإنجليزية وبالعكس بحسب الحاجة عند إدخال الأوامر:

```
> # هذه عملية جمع 5 و 2
> 5+2
[1] 7
```

ولاحظ وجود القوسين [ ] في السطر الثالث على يسار نتيجة عملية الجمع، (والذي يمثل المخرجات أو الناتج)، وبداخلهما الرقم 1، وهذه طريقة لغة R بإخبارنا بأن الناتج موجود في خلية واحدة فقط[3].

سنستمر فيما يلي بتنفيذ بعض العمليات الحسابية مع كتابة ملاحظات مختصرة في نفس سطر الأمر للتوضيح:

---

[1] الملاحظات تستخدم في لغة R للتوضيح فقط وليست جزءاً أساسيا من مكونات العملية الحسابية.

[2] يمكن استخدام اللغة العربية فقط في كتابة الملاحظات في نظام R ولا يمكن استخدامها كمصطلحات في لغة البرمجة.

[3] سنتطرق لعدد خلايا المخرجات لاحقا كلما تقدم استخدامنا للأوامر المُركبة.

```
> (4*7)/3 # عملية ضرب وقسمة
[1] 9.333333

> ((3-5)/(9+2))*(-4) # عملية طرح، قسمة، جمع، وضرب
[1] 0.7272727

> 8^5 # مرفوع للأس
[1] 32768

> sqrt(81) # الجذر التربيعي
[1] 9

> (27)^1/3 # الجذر التكعيبي
[1] 9

> log10(10) # اللوغاريتم للأساس 10
[1] 1

> log(10) # (Ln) اللوغاريتم الطبيعي
[1] 2.302585

> exp(0) # دالة الأس
[1] 1
```

لاحظ مما سبق أن الرموز (+، -، *، /، ^) تستخدم في لغة R بصورة أساسية للدلالة على الجمع، الطرح، الضرب، القسمة، والرفع للأس على الترتيب إلى جانب بعض الاستخدامات الإضافية الأخرى التي سنتناولها لاحقا. وأن كل الدوال[1] الرياضية الحسابية مثل دالة الجذر التربيعي، دوال اللوغاريتم، ودالة الأس يجب كتابتها بصيغتها المحددة في لغة R.

أيضا لابد من الإشارة هنا أن الأقواس ( ) يمكن استخدامها دائما لتحديد العملية أو العمليات التي ترغب في تنفيذها أولاً، فمثلا لاحظ في الأمر الثاني أعلاه أنه بناءً على ترتيب الأقواس سيتم تنفيذ عملية طرح 5 من 3 وتنفيذ عملية جمع 9 و2 ثم قسمة ناتج العملية الأولى على ناتج الثانية وأخيرا ضرب ناتج ذلك في الرقم 4-.

الآن وفي نهاية هذا البند، يمكن للقارئ التدرّب على تنفيذ العمليات الرياضية المتنوعة بغية اكتساب المرونة في إدخال الأوامر وتنفيذها ثم استدعاؤها من جديد، (باستخدام السهمين العلوي والسفلي)، وتعديلها، (بإضافة عمليات حسابية أو حذفها أو تصحيح الأخطاء التي حدثت)، وقراءة النواتج وكتابة الملاحظات متى أراد ذلك.

[1] سنوضح للقارئ لاحقا كيفية البحث عن صيغ الدوال الرياضية والإحصائية المختلفة التي لم يتم إدراجها في الكتاب.

ويمكن، كخيار إضافي، إنشاء ملف خاص لحفظ سطور الأوامر[1] التي تم إدخالها خلال جلسة العمل، حيث سيمكنك هذا الملف من نسخ سطور الأوامر، ثم لصقها في لوحة تحكم R ثم إعادة تنفيذها من جديد متى أردت.

ولعمل ذلك يجب إعطاء اسم لهذا الملف، وليكن "his1"، وتحديد مسار الحفظ، وليكن نفس مسار حفظ ملف العمل "work1"، ثم كتابة أمر **الحفظ** `savehistory` بالصورة التالية:

```
> savehistory("C:/Users/PIXEL-PC/Desktop/myR/his1.txt")
```

ولاحظ أن امتداد الملف النصي his1 هو ".txt" وليس ".RData".

بعد كتابة ذلك الأمر، سيتم حفظ كل سطور الأوامر في الملف "his1"، ويمكنك بعد ذلك استدعاء سطور الأوامر المحفوظة فيه إلى ذاكرة مسار العمل الحالي في أي وقت آخر بكتابة **أمر استدعاء الأوامر** `loadhistory` كالتالي:

```
> loadhistory("C:/Users/PIXEL-PC/Desktop/myR/his1.txt")
```

عندها يمكنك التنقل بين الأوامر التي تم كتابتها في الملف "his1" باستخدام السهمين العلوي والسفلي في لوحة المفاتيح (↑) و(↓). أو يمكنك ببساطة فتح الملف "his1" كملف نصي[2] من الحافظة "myR" أو يمكن فتحه من داخل[3] برنامج R.

وعموما، يمكن مشاهدة كل الأوامر التي تم تنفيذها في الجلسة الحالية في نافذة فرعية باستخدام الأمر التالي؛

```
> history()
```

فتظهر نافذة فرعية تضم كل ما تم تنفيذه من أوامر، (صحيحة وغير صحيحة)، خلال الجلسة الحالية. ولاحظ أنه إذا ما رغبت في حفظ كل سطور الأوامر التي تستخدمها فعليك استخدام الأمر `savehistory` للملف الذي ترغب فيه، (وهو "his1" حاليا)، في نهاية كل جلسة عمل ضمن نظام R، ويمكنك إعادة استدعاؤه في بداية أي جلسة أخرى باستخدام الأمر `loadhistory` من نفس مسار العمل وبنفس اسم الملف.

عند الانتهاء قم بحفظ ملف العمل "work1" مستخدما نفس صيغة[1] الأمر السابق `save.image`، ثم بعد ذلك يمكنك إغلاق الإطار الخارجي وعندها سيظهر السؤال التالي من خلال نافذة صغيرة منبثقة؛ " Save Workspace image?" فتقوم بالضغط على "Yes" لحفظ ما تم كتابته في ملف الحفظ العام وتنهي الجلسة[2].

---

[1] سيشمل هذا الملف كل ما تم كتابته كسطور أوامر تم تنفيذها، (سواء الصحيحة منها أو التي بها أخطاء في الصياغة)، ولا يشمل النتائج التي ظهرت بعد استخدام تلك الأوامر.

[2] وذلك بفتح الملف مباشرة من الحافظة "myR" باستخدام **برنامج النصوص** (Notepad) مثلا المتوفر افتراضيا في نظام ويندوز.

[3] يتم ذلك عن طريق فتح (File >Open script) وتغيير نوع الملفات المطلوب فتحها (Files of Type) إلى (All files(*.*)) ثم اختيار الملف "his1" وفتحه فتظهر نافذته الفرعية داخل برنامج R.

4.1 تحميل واستدعاء حِزم R (Downloading and Attaching the Packages of R)

إن الدوال والأوامر الخاصة بلغة R، وكذلك مجموعات البيانات، تكون مخزنة كلها ضمن كيانات تسمى **حِزم R** (Packages of R)، ولا يتم التعرّف على أي دالة أو مجموعة بيانات إلا عندما يتم استدعاء الحزمة التي تحتويها. ويتبع نظام R هذا الأسلوب لتحرير أكبر مساحة ممكنة من ذاكرة الحاسوب، وأيضا لحماية خصوصية مطوري هذا الحزم، وعدم تعارض الدوال بين الحزم.

وعند تنصيب نظام R لأول مرة، يتم تنصيب واستدعاء الحزم الأساسية الافتراضية فقط. لكن في كثير من الأحيان قد يحتاج المُستخدم لتطبيق دوال معينة متقدمة أو متخصصة غير متوفرة في تلك الحزم الأساسية، عندها يمكنه تحميل واستدعاء الحزم الإضافية التي يرغب بها بالطرق التي سيتم تناولها في هذا البند.
ويمكن الحصول على حزم R عادة بطريقتين رئيسيتين؛ الأولى هي عن طريق تحميلها[3] من موقع مشروع R عن طريق طلبها من داخل لوحة مراقبة R كما سنوضح تاليا، وأما الطريقة الثانية فهي تُستخدم عند توفر هذه الحزم مسبقا لدى المُستخدم، (في أي وسيلة تخزين بيانات)، حيث يمكن نسخها إلى المسار المحدد للحزم مباشرة واستدعاؤها بعد ذلك من لوحة المراقبة.

ونبدأ بشرح الطريقة الأولى:

أولا: يمكنك دائما معرفة حزم R المتوفرة في جهازك عن طريق استخدام أمر **عرض مكتبة الحِزم** `library` الذي يقوم بفتح نافذة فرعية تحتوي على أسماء تلك الحزم مرتبة أبجديا إضافة إلى شرح موجز عن ما تحتويه أو تمثله هذه الحزم، كما نشاهد في الشكل (8.1)؛

```
> library()
```

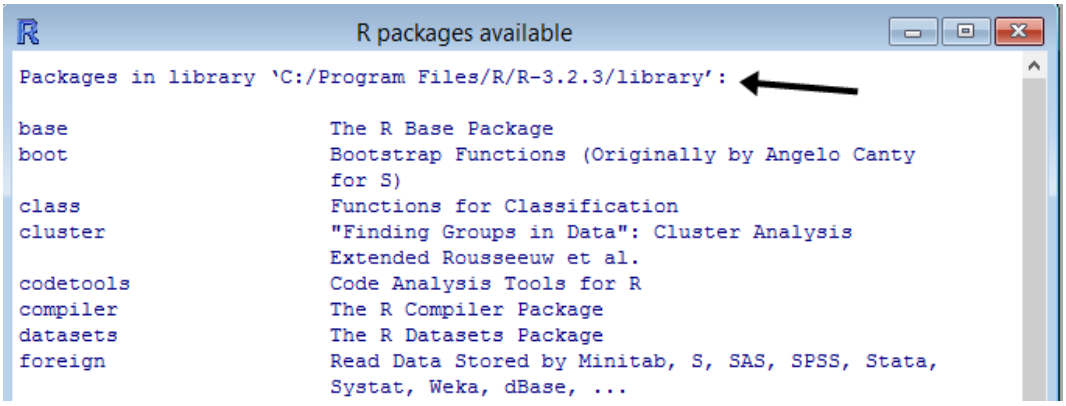
ولاحظ أن مسار حفظ حِزم R الافتراضي على الجهاز هو "C:/Program Files/R/R-3.2.3/library".

---

[1] سيتم من الآن فصاعدا عرض الأوامر والدوال بدون استخدام الأقواس بعدها وذلك للاختصار.

[2] اختيار "Yes" سيؤدي لإنشاء (أو تحديث في الجلسات اللاحقة) ملفين داخل الحافظة "myR" الأول بدون اسم ويحمل شعار البرنامج، وهو يمثل ملف عمل عام للجلسات، والملف الثاني هو باسم (.Rhistory) ويضم كل الأوامر التي تم تنفيذها حتى وقت الحفظ، أما اختيار "No" فيؤدي إلى عدم إنشاء الملفين السابقين أو عدم تحديثهما إذا كانا موجودين في الأصل.

[3] الطريقة الأولى تتطلب اتصالا بشبكة الانترنت لتحميل الحزم.

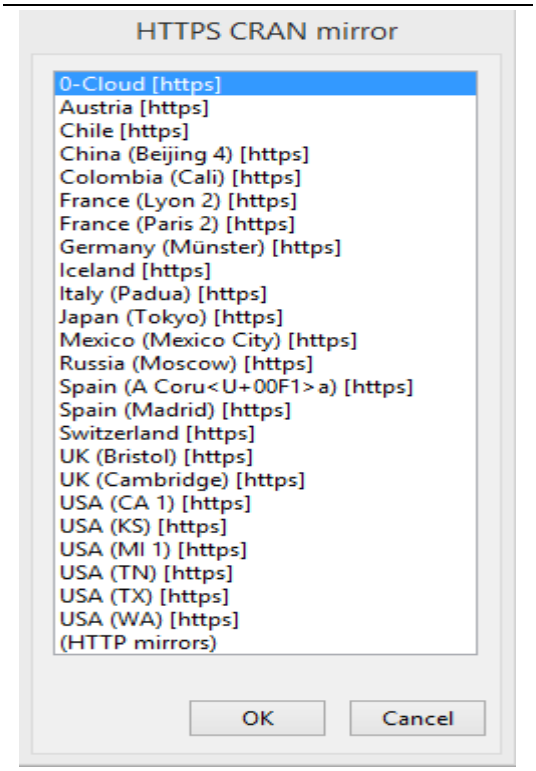


شكل 8.1: نافذة عرض حزم R الموجودة في الحاسوب

ثانيا: يمكن تحميل أي حزمة باستخدام أمر **تحميل الحِزم** `install.packages`، حيث يتم فتح نافذة فرعية لاختيار الدولة ومنفذ الانترنت الذي يرغب المُستخدم بتعيينه للتحميل. وإذا لم تتمكن من إيجاد المنفذ المناسب لدولتك من ضمن الخيارات المتاحة، يمكنك دائما اختيار الخيار الأول في تلك النافذة وهو؛ " 0-Cloud [https]". وبتنفيذ الأمر:

```
> install.packages()
---Please select a CRAN mirror for use in this session---
```

تظهر رسالة تفيد بضرورة اختيار المنفذ وتظهر النافذة كما في الشكل (9.1)؛



شكل 9.1: نافذة اختيار منفذ الانترنت حسب الدولة

بعد اختيار المنفذ المناسب، أو اختيار الخيار الأول، والضغط على أمر "OK" تظهر نافذة ثانية، (كما في الشكل (10.1))، وهي قائمة بها أسماء الحزم الموجودة في مشروع R مرتبة أبجديا، والتي يتم تحديثها دوريا.

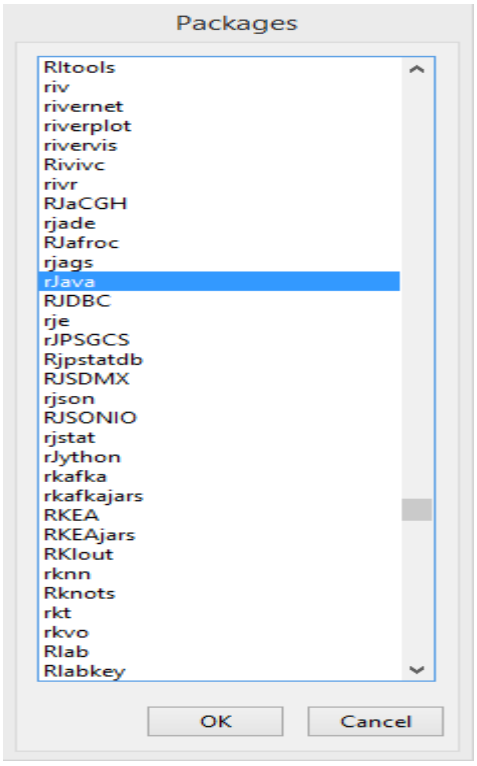
سنقوم الآن، على سبيل المثال، باختيار ثلاثة حزم لتحميلها، حيث أن تلك الحزم ستكون ضرورية للتعامل مع بعض الدوال الهامة في الفصل القادم.

تلك الحزم هي `rJava`، `XLConnect`، و`XLConnectJars`. يتم اختيار أول حزمة بالضغط عليها بالفأرة، ثم يتم البحث عن الحزمتين المتبقيتين في القائمة واختيارهما بالضغط على زر التحكم (CTRL) في لوحة المفاتيح والزر الأيسر في

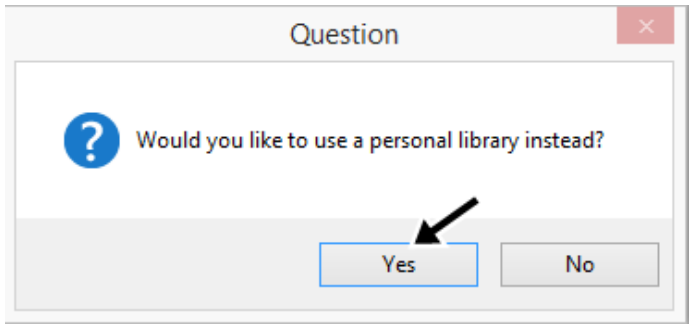
الفأرة في نفس الوقت، بعدها يتم اختيار (OK).

وننوه هنا أنه عند تحميل حزم إضافية للمرة الأولى، قد تكون هنالك إعدادات خاصة بمنع تخزين أو إدراج الحزم في المسار الأصلي لنظام R وخاصة في معظم إصدارات نظام ويندوز الأخيرة، وعندها ستظهر نافذة في هذه المرحلة بها رسالة من النظام لطلب الموافقة على استخدام مكتبة حِزم إضافية (شخصية) جديدة في مسار آخر، (الشكل (11.1))؛

عندها قم باختيار نعم (Yes)، لكي يتم حفظ الحِزم الإضافية في حافظة مختلفة عن حافظة الحِزم الافتراضية حتى تتمكن من التعرف دائما على ما لديك من حِزم إضافية في جهازك وكذلك يمكنك تخزينها ونقلها إلى حاسوب آخر عند الضرورة.



شكل 10.1: قائمة الحزم الموجودة في مشروع R

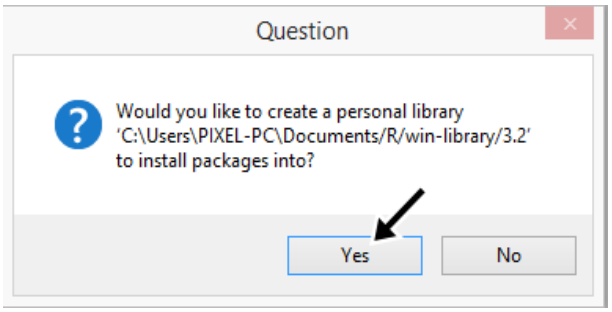


شكل 11.1: نافذة اختيار استخدام مكتبة حِزم شخصية جديدة

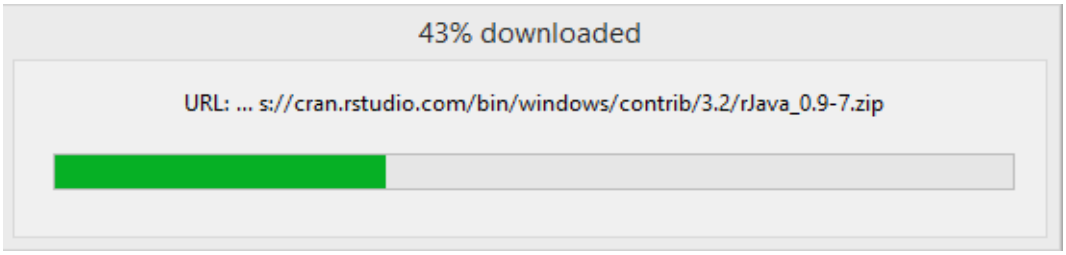
بعد ذلك ستظهر نافذة فرعية أخرى، (كما في الشكل (12.1))، ولتأكيد إنشاء مكتبة حِزم شخصية جديدة في المسار الجديد الموضح في الشكل.

قم باختيار نعم و ستظهر بعد ذلك نافذة تحميل الحزم المطلوبة، (كما في الشكل (13.1)).

وقد يستغرق ذلك بعض الوقت بحسب حجم الحزمة وسرعة الانترنت لديك.



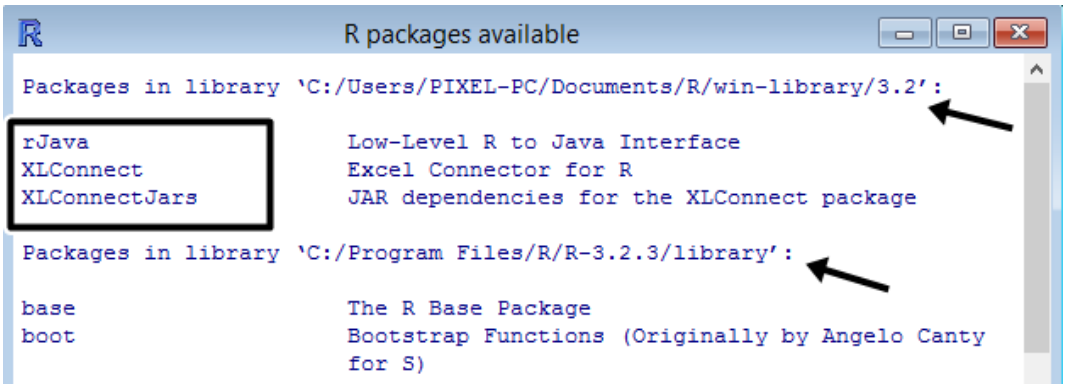
شكل 12.1: نافذة اختيار إنشاء مكتبة حِزم شخصية في مسار جديد



شكل 13.1: نافذة تحميل الحِزمة rJava من مشروع R

بعد اكتمال التحميل، سيتم تنصيب هذه الحزم في نظام R بشكل تلقائي، وبعدها ستظهر المعلومات الخاصة بعملية التحميل والتنصيب في لوحة مراقبة R، والتي تشمل؛ المسار الذي توجد به كل حزمة، حجمها (بالكيلوبايت أو الميغابايت)، وإخطار بنجاح عملية التنصيب، إضافة إلى إخطار آخر بالمسار الذي توجد به الملفات المضغوطة (Compressed or Zip Files) لتلك الحزم.

وإذا ما تم عرض الحِزم الموجودة في نظام R على الحاسوب، باستخدام الأمر `library()` من جديد، ستجد أن تلك الحزم الثلاثة الجديدة موجودة ضمنها. وفي جهازنا المُستخدَم، تظهر الحزم الجديدة ضمن المسار "C:/Users/PIXEL-PC/Documents/R/win-library/3.2"، أما المسار الأصلي لحزم R الافتراضية فهو "C:/Program Files/R/R-3.2.0/library"، كما يظهر أدناه في الشكل (14.1). وبنفس الطريقة، يمكن في أي وقت لاحق تحميل أي حزمة ترغب بإضافتها.



شكل 14.1: مسارات الحِزم الافتراضية والجديدة المضافة إلى نظام R

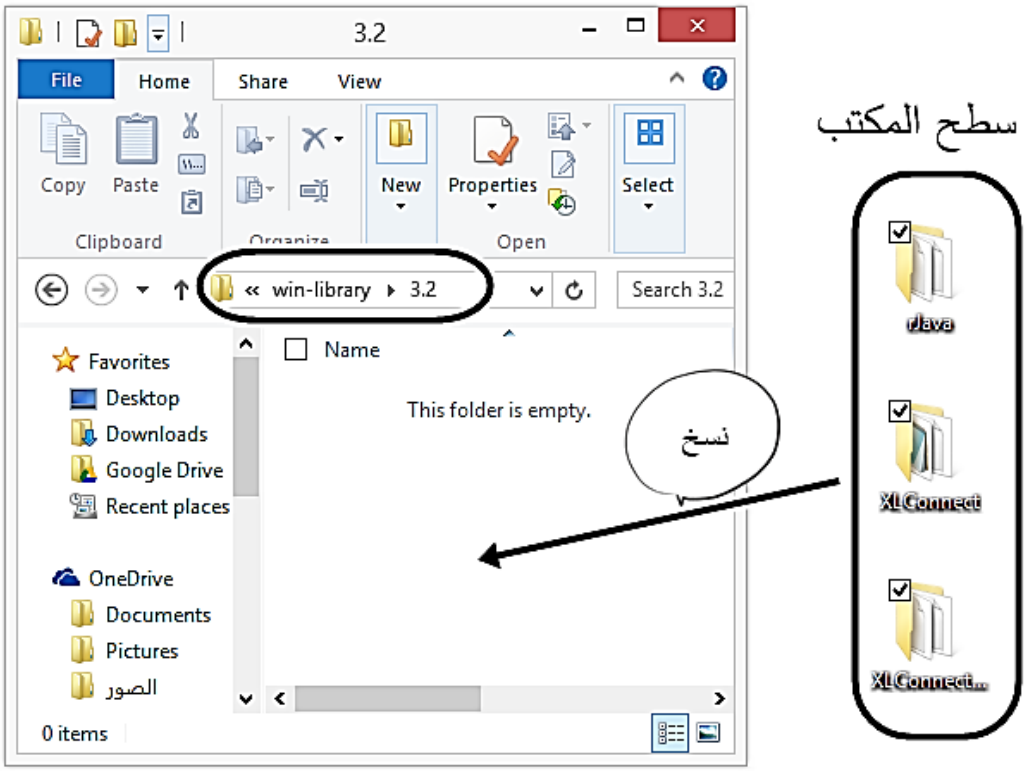
ملاحظة هامة:

من أجل استخدام الحزمة `rJava` واستدعاؤها لاحقا، لابد من وجود البرنامج المساعد الشهير **جافا** (Java™ Oracle) على حاسوبك، والذي يمكن تحميله أو تحديثه للإصدار 32-bit أو 64-bit لنظام ويندوز من الموقع الرسمي له "http://www.java.com" مباشرة وبصورة مجانية.

أما الطريقة الثانية لتثبيت الحِزم الإضافية؛ فهي أبسط بكثير من الأولى لأنها لا تتطلب الكثير من الخطوات ولا تتطلب أيضا الاتصال بالإنترنت، إلا أنه لابد أن تتوفر لديك الحزم الإضافية التي ترغب باستخدامها في صورة ملفات مخزنة[1] مسبقا.

علما بأن مسار الحافظة الذي سيتم فيها تنصيب حزم R الإضافية على جهازنا المُستخدم هو نفس المسار الذي تم إنشاؤه في الطريقة الأولى، (وذلك لغرض عدم تشتيت انتباه المُستخدم بين عدة مسارات مختلفة لحفظ الحِزم)؛ "C:/Users/PIXEL-PC/Documents/R/win-library/3.2"، وبافتراض أن الحافظات التي تحتوي على الحِزم الثلاثة المطلوبة قد تم نقلها إلى سطح المكتب[2]، عندئذ يتم نقلها من موقعها الحالي إلى مسار الحزم الإضافية كما يوضح الشكل (15.1)، وبهذا تكون هذه الحِزم متوفرة في نظام R لديك بصورة دائمة.

إلا أن هذا لا يغني عن ضرورة استدعاء هذه الحِزم في كل جلسة، شأنها كشأن كل الحزم الإضافية، عندما يتطلب الأمر استخدامها كما ذكرنا سابقا.



شكل 15.1: طريقة نسخ الحِزم المطلوبة إلى مسار الحِزم الإضافية

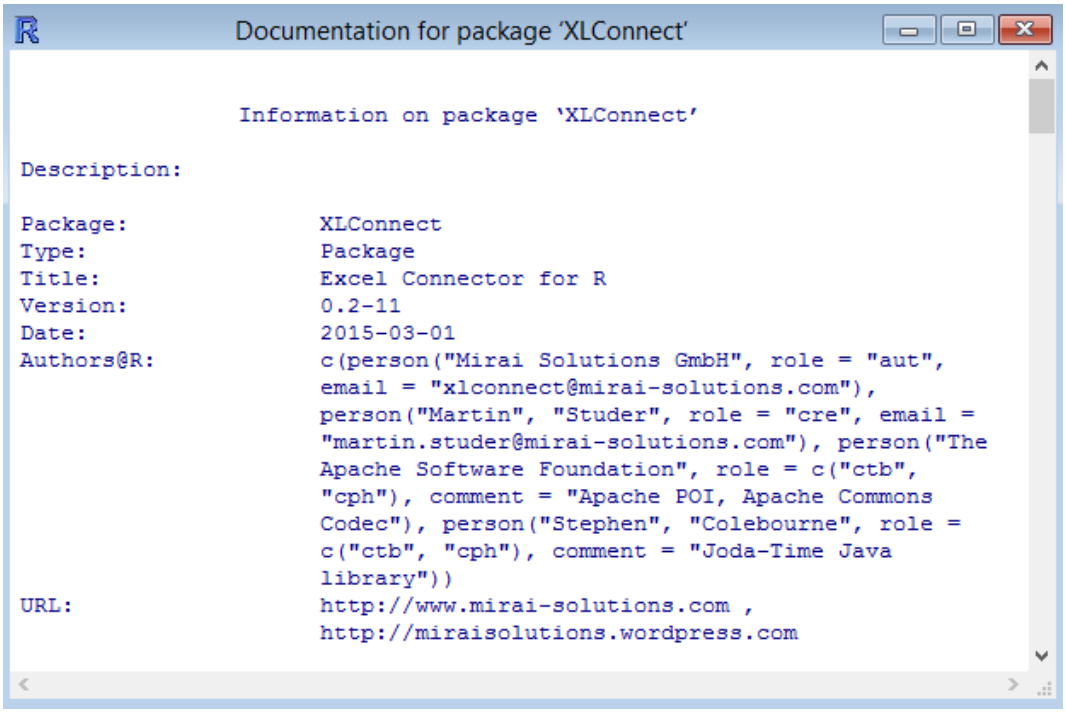
[1] لكي تتمكن من نقل أي حزمة إضافية إلى مسار نظام R مباشرة لابد أن تكون على هيئة حافظة لها نفس اسم الحزمة.

[2] إذا ما وُجدت الحافظات المطلوبة في قرص مضغوط، أو ذاكرة وميضية أو غيرها، فإنه يمكن نسخها مباشرة إلى المسار المطلوب، أو نسخها إلى سطح المكتب ومن ثم إلى المسار المطلوب.

وللتعرف على المكونات أو الدوال أو حتى ملفات البيانات المتوفرة في أي حزمة تم تحميلها في نظام R لديك، يمكنك استخدام الأمر `library(help="*")`، حيث تمثل * اسم الحزمة المطلوبة، فمثلا للتعرف على مكونات الحزمة `XLConnect` بعد تنصيبها نقوم بتنفيذ الأمر:

```
> library(help="XLConnect")
```

وستشاهد ظهور نافذة فرعية بعنوان (Documentation for package 'XLConnect')، كما في الشكل (16.1)، تضم كل المعلومات المتعلقة بالحزمة `XLConnect`. وقد تكون تلك المعلومات غير مفهومة بشكل واضح وتام للقارئ في هذه المرحلة، إلا أن هذا الأمر سيكون مفيدا في المراحل المتقدمة.



شكل 15.1: المعلومات المتعلقة بالحزمة XLConnect

في الفصلين القادمين سنوضح بعض المصطلحات والدوال الهامة في لغة R، ومنها ما يتعلق بتحديد طبيعة البيانات التي سيتم التعامل معها، وسنقوم أيضا بشرح كيفية إدخال تلك البيانات كمتجهات ومصفوفات وأُطر بيانات واستدعاؤها وتعديلها وغير ذلك من الأمور التفصيلية من أجل التمهيد لمرحلة التحليل الإحصائي للبيانات.

الفصل الثاني

المتجهات، المصفوفات، والقوائم في R

(Vectors, Matrices, and Lists in R)

1.2 التعيينات والأشياء (Assignments and Objects)

2.2 المتجهات في R (Vectors in R)

1.2.2 بعض الدوال الحسابية على المتجهات (Some Arithmetic Functions on Vectors)

2.2.2 توليد السلاسل العددية (Generating Sequences)

3.2 نُظم الصفوف والمصفوفات في R (Arrays and Matrices in R)

1.3.2 إنشاء المصفوفات (Constructing Matrices)

2.3.2 دمج المتجهات والمصفوفات (Merging Vectors and Matrices)

3.3.2 استدعاء القيم من المصفوفات (Calling values from Matrices)

4.3.2 استبدال القيم في المصفوفات (Replacing values in Matrices)

4.2 الدوال الحسابية الأساسية على المصفوفات في R
(Basic Arithmetic Functions on Matrices in R)

5.2 نظام القوائم (Lists)

بعد أن تم تحميل وتنصيب برنامج R على جهاز الحاسوب، وتوضيح أهم الإعدادات الأولية داخل نافذة البرنامج، وشرح كيفية بدء جلسة التعامل مع لغة R، ننتقل في هذا الفصل إلى المرحلة التالية والتي تبدء بتعريف أهم المصطلحات التي يعتمدها نظام R في تنظيم أولويات المُدخلات والأوامر، والتي سيتم من خلالها التعامل مباشرة مع البيانات.

في الفصل السابق قمنا بتخزين ملف العمل باسم "work1" في الحافظة "myR" على سطح المكتب، وسنفترض الآن أن برنامج R مغلق وأننا سنقوم بتشغيله باستخدام أيقونة البرنامج الموجودة على سطح المكتب، حيث أننا سنقوم بتعيين ملف عمل مستقل لكل فصل من فصول الكتاب بهدف تنظيم الأوامر التي يتم التعامل معها في كل فصل من جهة، ومن جهة أخرى فإن ذلك سيدرب المُستخدم على تنسيق الأوامر والدوال لكل بحث أو دراسة في ملفات عمل مستقلة.

بعد فتح برنامج R من أيقونة سطح المكتب سنقوم بتغيير مسار العمل الحالي، (نقصد للفصل الثاني)، إلى مسار الحافظة "myR" على سطح المكتب:

```
> getwd()
[1] "C:/Users/PIXEL-PC/Documents"

> setwd(dir="C:/Users/PIXEL-PC/Desktop/myR")
> getwd()
[1] "C:/Users/PIXEL-PC/Desktop/myR"
```

الآن لنعطي الاسم "work2" كاسم لملف العمل الحالي باستخدام أمر الحفظ save.image بالصورة التالية:

```
> save.image("C:/Users/PIXEL-PC/Desktop/myR/work2.RData")
```

وهذا سيؤدي بدوره لظهور ملف العمل "work2" داخل الحافظة "myR"، ويمكنك بعد ذلك فتح[1] برنامج R من داخل الحافظة "myR" باستخدام هذا الملف لتطبيق أوامر الفصل الثاني مباشرة. كذلك يمكنك إنشاء ملف لحفظ سطور الأوامر، كما هو الحال في الفصل الأول، باستخدام الأمر:

```
> savehistory("C:/Users/PIXEL-PC/Desktop/myR/his2.txt")
```

[1] يمكنك أيضا استدعاء الملف "work2" من داخل لوحة مراقبة R باستخدام أمر **الاستدعاء**:
load("C:/Users/PIXEL-PC/Desktop/myR/work2.RData").

ونُذكّر بضرورة استخدام أوامر الحفظ save.image وsavehistory، (أو استخدام زر التخزين ضمن رموز شريط الأوامر أو من داخل قائمة الملف (File > Save Workspace...) أو حتى استخدام الاختصار CTRL+S في لوحة المفاتيح)، في نهاية كل جلسة عمل إذا ما أردت حفظ ما تم تنفيذه في تلك الجلسة. الآن أنت مستعد لبدء المرحلة التالية.

1.2 التعيينات والأشياء (Assignments and Objects)

قمنا في الفصل الأول بإجراء بعض العمليات الرياضية البسيطة باستخدام لغة R وحصلنا على النتائج بصورة مباشرة، إلا أن تلك العمليات والنتائج لم يتم تعيينها إلى رموز، وبالتالي لا يمكن استدعاء أو تعديل تلك النتائج إذا ما رغبنا بذلك. لهذا نقوم عادة في نظام R بتنظيم التعامل عن طريق إعطاء **تعيينات** (Assignments) لبعض القيم أو الأوامر أو النتائج التي قد نرغب بالتعامل معها أو استدعاؤها من جديد.

ولتعيين أي قيمة أو أمر أو نتيجة يُستخدم **رمز التعيين** (Assignment Operator) الذي يُعرف[1] بالشكل "<-"، (إشارة "أصغر من" متبوعة بإشارة الطرح)، ويكون على يسار رمز التعيين الاسم الذي يضعه المستخدم وعلى يمين رمز التعيين القيمة أو الأمر أو النتيجة المطلوب تعيينها لذلك الاسم. وكمثال توضيحي سنقوم بتعيين القيمة 7 للاسم x كما يلي:

```
> x<-7
```

ستلاحظ بعد ضغط زر الإدخال عدم ظهور نتيجة مشاهدة لأن عملية التعيين هذه تم تخزينها في ذاكرة البرنامج فقط. قم الآن باستدعاء x عن طريق كتابة x والضغط على زر الإدخال فتحصل على التالي:

```
> x
[1] 7
```

إن ما قمنا به ببساطة هو تعريف **شيء** (Object) باسم x وتعيين القيمة 7 له. ومصطلح **أشياء** هو المصطلح اللغوي المعتمد في نظام R لتعريف كل الكيانات التي يتم إنشاؤها، أو بمعنى آخر هو المصطلح الذي تطلقه R على كل البيانات التي تأخذ أسماء ويتم تخزينها في ذاكرة البرنامج. وفيما يلي نسرد بعض الملاحظات التي تتعلق بإنشاء الأشياء بصورة عامة:

- يمكن استخدام الأحرف الإنجليزية الصغيرة أو الكبيرة كأسماء للأشياء مع التفريق بينها، بمعنى أنه يمكن مثلا استخدام age<-7 و AGE<-3 و Age<-14 كأسماء مختلفة لقيم مختلفة، حيث أن نظام R يتعامل مع الأحرف الصغيرة والكبيرة كأشياء مختلفة.

[1] يمكن استخدام إشارة "يساوي (=)" لنفس الغرض، إلا أن القائمين على مشروع R "يحبذون" استخدام رمز التعيين "<-" عند تعريف أسماء الأشياء واستخدام إشارة "يساوي" ضمن صيغ الدوال والأوامر وذلك من الناحية التنظيمية.

- يمكن استخدام الأرقام، النقاط "."، و إشارة "_" مع الأحرف للتسمية، فمثلا يمكن تعيين (y1، var_12، STUDENT.Num) وغيرها كأسماء للأشياء. مع الإخذ بالاعتبار أن الأرقام لا يمكن استخدامها في بداية الاسم.

- توجد بعض الأسماء "المحتجَزة" من قِبل نظام R والتي لا يمكن للمستخدم استخدامها لتسمية الأشياء. هذه الأسماء يمكن الاطلاع عليها عن طريق كتابة الأمر help(reserved) في نافذة R. نذكر من تلك الأسماء المحجوزة على سبيل المثال (TRUE، FALSE، if، for، ...).

- يجب الانتباه إلى أنه إذا ما تم استخدام اسم معين لتعيين قيمة أو دالة ما ثم تم استخدام نفس الاسم لتعيين قيمة جديدة فإن الاسم سيتم تعيينه لتلك القيمة الجديدة وتُلغى القيمة القديمة دون عرض تحذير من قِبل R، فمثلا إذا ما أدخلنا a<-1/2 ثم أدخلنا a<-5 فإن قيمة a التي ستبقى في ذاكرة R هي الرقم 5. لذلك يجب التأكد دائما عند استخدام أسماء جديدة للأشياء أنه لم يتم استخدامها سابقا في نفس ملف العمل، مع العلم بأنه يمكن استخدام نفس الأسماء في ملفات عمل مختلفة، ولذلك يُنصح دائما باستخدام ملف عمل مختلف لكل بيانات دراسة جديدة أو تحليل إحصائي تود القيام به.

وعموما، يمكن دائما التأكد من أسماء الأشياء التي تم إنشاؤها وحفظها في ذاكرة R في ملف العمل الحالي عن طريق استخدام أمر **استدعاء الأشياء** objects() أو ls() حيث سيتم عرض كل أسماء الأشياء التي قمت أنت بإنشائها في ملف العمل. والآن استخدام أحد الأمرين السابقين وسيظهر لك:

```
> ls()
[1] "x"
```

أما إذا ظهرت لك أشياء (أسماء) أخرى فهذا يعني أنك قمت بتعريفها خلال الجلسة الحالية. ويمكنك متى شئت حذف شيء أو مجموعة من الأشياء باستخدام الأمر rm(*)، حيث (*) يرمز لاسم الشيء أو الأشياء التي ترغب بحذفها (على أن تضع بينها فواصل إذا كانت أكثر من شيء واحد). وكمثال توضيحي لعملية حذف الأشياء، لنقم بإنشاء الأشياء التالية واستدعاؤها ثم حذفها:

```
> x1<-10
> x2<-20

> x1
[1] 10

> x2
[1] 20

> ls()
[1] "x"  "x1" "x2"
> rm(x1,x2)
```

```
> ls()
[1] "x"
```

وهذا يعني أن الأشياء المحفوظة في ملف العمل الحالي "work2" هو x فقط، إذا لم توجد أشياء أخرى قمت بتعيينها، علما بأنه يمكنك إلغاؤها هي الأخرى بنفس الطريقة. ويمكن أيضا استخدام الأمر؛ rm(list=ls()) لإلغاء كل الأشياء الموجودة في ملف العمل.

لاحظ في المثال التوضيحي السابق أنه قد تم تعيين قيم كلا من x1 وx2 في سطر أمر مستقل، ثم تم استدعاء كل شيء بمفرده، إلا إنه للاختصار يمكننا متى أردنا تعيين أشياء مختلفة، أو كتابة أوامر متعددة في سطر أمر واحد، وكذلك يمكننا استدعاء أو تنفيذ مجموعة من الأوامر باستخدام نفس سطر الأمر أيضا باستخدام رمز الفاصلة المنقوطة ";" للفصل بينها، كما سنوضح في المثال التوضيحي التالي:

```
> a<-6;b<-5;c<-a*b;a;b;c

[1] 6
[1] 5
[1] 30
```

ولاحظ أن عملية تعيين قيم للأشياء a و b ثم تعيين ناتج ضربهما إلى c، وكذلك كتابة أمر استدعاء الأشياء الثلاثة كلها قد تم في سطر واحد عن طريق الفصل بينها بالفاصلة المنقوطة دون استخدام مسافات بينها.

بعد هذا المثال، يمكنك التدرب على تعيين القيم العددية إلى الأشياء مستخدما أية أسماء تريدها، ثم استخدام تلك الأشياء لتنفيذ عمليات حسابية مختلفة في أسطر أوامر مختلفة أو في نفس سطر الأمر.

بعد ذلك قم بحذف كل الأشياء الموجودة في ملف العمل الحالي باستخدام الأمر rm لكي تصبح محتويات ملف العمل لديك متطابقة مع ما سنتناوله في بنود هذا الفصل، ثم قم بالتأكد من إتمام ذلك باستخدام الأمر ls() بالصورة:

```
> ls()
character(0)
```

وهذا يعني عدم وجود أية تعيينات أو أشياء في الذاكرة، ثم نقوم بحفظ ملف العمل "work2".

2.2 المتجهات في R (Vectors in R)

إن **المتجه** (Vector) يُعَد أبسط تركيبة للبيانات في R، فهو كما يُعرّف في نظام R كيان فردي (شيء) يضم عدة قيم منظمة بصورة محددة. وتوجد ثلاثة أنواع رئيسية من المتجهات الرياضية في لغة R؛ **المتجهات العددية** (Numeric vectors)، **المتجهات المُميّزة** (Character vectors)، و**المتجهات المنطقية** (Logical vectors)، وسوف نتناول المتجهات العددية أولا.

- **المتجهات العددية:**

هي تلك المتجهات التي تضم فئة الأعداد الصحيحة، ويُطلَق عليها (Integers)، أو التي تضم فئة الأعداد الحقيقية، ويُطلق عليها (Double). ولإنشاء متجه نستخدم دالة **إنشاء متجه** `c()` عادة لإدخال القيم (العددية وغير عددية) كما سنرى في المثال التالي، ونُذكّر هنا من جديد بضرورة فتح برنامج R باستخدام ملف العمل "work2":

```
> x1<-c(2,4,5,0,-7)
```

وهذا يعني أنه تم تعريف متجه عددي اسمه `x1` يضم القيم (2, 4, 5, 0, -7). ويمكن أيضا استخدام دالة **تعيين الأشياء** `assign()` للقيام بنفس العملية بالصورة التالية:

```
> assign("x2",c(1,3,8,10,3))
```

بمعنى أنه تم تعيين القيم (1, 3, 8, 10, 3) للمتجه `x2`. ولنقم الآن بتنفيذ بعض العمليات الحسابية مستخدمين المتجهين `x1` و`x2` كالتالي:

```
> # x1 , x2 مشاهدة ما يحتويه المتجهين
> x1;x2

[1]  2  4  5  0 -7
[1]  1  3  8 10  3

> x1^2
[1]  4 16 25  0 49

> 1/(x2-1)
[1]       Inf 0.5000000 0.1428571 0.1111111 0.5000000
```

ولاحظ أنه عند استخدام أي عملية رياضية على متجه عددي فإنه يتم تنفيذها على كل قيم المتجه. وبالنسبة للنتيجة الأخيرة فإن المصطلح `Inf` هو اختصار (Infinity) وتعني اللانهاية (∞) حيث تمت القسمة على الصفر في العملية الأخيرة. نتابع تنفيذ عمليات أخرى:

```
> x2*3
[1]  3  9 24 30  9

> x1+x2
[1]  3  7 13 10 -4

> x3<-x1+(5*x2)
> x3
[1]  7 19 45 50  8

> x4<-c(x1,x2) # دمج متجهين
> x4
```

```
[1]  2  4  5  0 -7  1  3  8 10  3
```

```
> sin(x2)
[1]  0.8414710 0.1411200  0.9893582 -0.5440211  0.1411200
```

في العملية الأخيرة تم استخدام دالة جيب الزاوية sin والتي توجد في لغة R إلى جانب أخواتها من الدوال المثلثية؛ cos و tan وغيرهما. نستمر بتنفيذ عملية أخرى وهي حساب اللوغاريتم الطبيعي ($\log_e$ أو Ln)؛

```
> log(x1)
```

```
[1] 0.6931472 1.3862944 1.6094379  -Inf   NaN
Warning message:
In log(x1) : NaNs produced
```

والنتيجة NaN هي اختصار (Not a Number) وتعني أن ناتج العملية الحسابية غير مُعرّف، وتم ظهور الرسالة التحذيرية (Warning message) التي تبين وجود قيمة أو قيم غير مُعرّفة عند تنفيذ العملية. وما حدث هنا قد يدفعنا أحيانا إلى تنفيذ بعض العمليات الحسابية على مجموعة يتم اختيارها من القيم الموجودة ضمن متجه ما وليس على كل قيم المتجه، ويتم تنفيذ ذلك باستخدام الأقواس [ ] كما يوضح المثال التالي؛

أولا نوضح كيفية استدعاء القيم الثلاثة الأولى في المتجه x1 بأكثر من طريقة:

```
> x1[1:3]
[1] 2 4 5
```

أو

```
> x1[c(1,2,3)]
[1] 2 4 5
```

أو

```
> x1[-c(4,5)]
[1] 2 4 5
```

أي أنه يمكن اختيار أي مجموعة من القيم المتسلسلة ضمن المتجه باستخدام الأمر الأول، ويمكن اختيار مجموعة من القيم المتفرقة ضمن المتجه باستخدام الأمر الثاني أو الثالث. ويتم تنفيذ العملية الحسابية المطلوبة (وهي أخذ اللوغاريتم الطبيعي) على القيم المُختارة (القيم الثلاثة الأولى) من المتجه x1 بالصورة:

```
> log(x1[1:3])
[1] 0.6931472 1.3862944 1.6094379
```

وكمثال آخر للتعامل مع بعض القيم دون الأخرى داخل المتجهات، لنفرض أننا نريد قسمة القيمة الأولى والثالثة والرابعة من المتجه x3 على القيم الثانية والرابعة والخامسة من المتجه x2، وتعيين النتيجة باسم المتجه x5، عندئذ نكتب:

```
> x5<-x3[c(1,3,4)]/x2[c(2,4,5)]
```

```
> x5
[1]  2.333333  4.500000 16.666667
```

مثال آخر يُعرّف متجه جديد هو x2.1 والذي سيحتوي على كل القيم الأكبر من 3 في المتجه x2؛

```
> x2
[1]  1  3  8 10  3

> x2.1<-x2[x2>3]
> x2.1
[1]  8 10
```

إضافة إلى ذلك، يمكننا استبدال أي قيمة أو قيم في المتجه العددي (أو غير العددي) بالطريقة التالية، حيث سنقوم باستبدال القيمة الثانية في المتجه x2 بالقيمة 5:

```
> x2[2]<-5
> x2
[1]  1  5  8 10  3
```

وسنقوم الآن بإرجاع المتجه x2 إلى وضعه الأصلي؛

```
> x2[2]<-3
> x2
[1]  1  3  8 10  3
```

من جديد يمكن تغيير أكثر من قيمة بالطريقة التالية:

```
> x2[c(2,4,5)]<-c(6,7,15)
> x2
[1]  1  6  8  7 15
```

ويمكن أيضا إرجاع المتجه x2 إلى قيمه الأصلية بنفس الطريقة:

```
> x2[c(2,4,5)]<-c(3,10,3)
> x2
[1]  1  3  8 10  3
```

قبل الاسترسال أكثر من ذلك في التعامل مع العمليات الحسابية التي يمكن تنفيذها على المتجهات العددية، سنقوم بتعريف النوعين الآخرين من المتجهات التي تم ذكرها سابقا؛

- **المتجه المُميَز:**

هو متجه يضم قيم نصية غير عددية، ويتم إدخال القيم النصية في R عموما باستخدام علامات الاقتباس المزدوجة (" ") أو المفردة (' '). وكمثال، لنقم بتعيين أسماء المدن التالية في ليبيا إلى المتجه Libya:

```
> Libya<-c("Benghazi","Tripoli","Tobruk")
> Libya
[1] "Benghazi" "Tripoli"  "Tobruk"
```

ويُعتبر المتجه **العاملي** (Factor)، والذي يتعامل مع المستويات أو الفئات (مثل المستوى الدراسي أو الفئة العمرية)، الحالة الخاصة الأهم من المتجهات المميزة، وسوف نتناوله في الفصل القادم.

- **المتجه المنطقي**:

وهو متجه يأخذ القيمتين[1] المنطقيتين **صحيح** أو **خاطئ** واللذين يتم تعريفهما بلغة R بالمصطلحات TRUE أو T، وFALSE أو F اختصارا على الترتيب، ويتم إدخالها بدون استخدام علامات الاقتباس. فمثلا يمكننا تعيين المتجه المنطقي بالصور التالية:

```
> c(TRUE,FALSE)
[1]  TRUE FALSE

> c(T,T,F,T,F,T)
[1]  TRUE  TRUE FALSE  TRUE FALSE  TRUE
```

إلا أنه عمليا لا يتم تعيين المتجهات المنطقية بالصورة السابقة، بل يتم استخدامها عادة مع دوال أو عمليات حسابية شرطية[2] للاستفادة من النتيجة المنطقية التي يوفرها هذا النوع من المتجهات. ولتوضيح هذه النقطة، لنفرض أنه لدينا متجه عددي يمثل درجات ستة طلبة في امتحان من 10 درجات، عندئذ يمكن تنفيذ العملية التالية مثلا لمعرفة عدد وترتيب الطلبة الذين تحصلوا على أقل من أو يساوي 5 درجات في الامتحان:

```
> grades<-c(6,7,4,10,8,5)
> grades
[1]  6  7  4 10  8  5

> bad.grades<-grades<=5

> bad.grades
[1] FALSE FALSE  TRUE FALSE FALSE  TRUE
```

ما تم أعلاه هو أننا قمنا بتعيين درجات الطلبة إلى المتجه grades وقمنا باستدعائه، ثم وضعنا الشرط (<=5) والذي يُقرأ "أقل من أو يساوي 5"، على جميع قيم المتجه grades وقمنا بتسمية هذه العملية (النتيجة) باسم bad.grades وهو متجه منطقي. ونلاحظ من المثال أن طالبين فقط (هما الثالث والسادس) قد تحصلوا على درجات أقل من أو يساوي 5 في الامتحان.

---

[1] قد يأخذ المتجه المنطقي أحيانا القيمة NA، وهي اختصار (Not Available) وترمز للقيم المفقودة.

[2] سنأتي لشرح الدوال الشرطية بتفصيل أكثر في الفصول القادمة.

ويمكن أيضا استخدام دالة which للحصول على متجه منطقي يعبر عن هذه النتيجة كما يلي:

```
> which(grades<=5)
[1] 3 6
```

ويمكن استخدام هذه الدالة مع أنواع المتجهات الأخرى، مثل المتجهات المُميَزة كالتالي:

```
> Libya
[1] "Benghazi" "Tripoli"  "Tobruk"
> which(Libya=="Benghazi")
[1] 1
```

وعموما فإن **المعامِلات المنطقية** (Logical Operators) الأكثر بساطة وتداولا في لغة R هي؛ أقل من (<)، أقل من أو يساوي (<=)، أكبر من (>)، أكبر من أو يساوي (>=)، يساوي تماما[1] (==)، ولا يساوي (=!). فمثلا يمكن كتابة:

```
> grades==10
[1] FALSE FALSE FALSE  TRUE FALSE FALSE

> grades!=10
[1] TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

وأما إذا كان لدينا متجهين منطقيين فإن الرمز (&) يستخدم للحصول على التقاطع (Intercept) بينهما، والرمز (|) يستخدم لإيجاد الاتحاد (Union) بينهما، والرمز (!) يستخدم للحصول على المتجه المكمل (Complement) للمتجه الأصلي. ولنأخذ المثال التالي لتوضيح طريقة استخدام عمليات الفئات السابقة:

لنفرض أن المتجه العددي age يمثل أعمار ستة أشخاص بالسنوات، والمتجه العددي BP يمثل ضغط الدم لهؤلاء الأشخاص، عندئذ نستطيع تعريف متجهات منطقية جديدة بالصورة التالية:

```
> age<-c(25,70,21,30,65,18)
> BP<-c(120,180,110,125,130,111)

> sen<-age>60 # أشخاص مسنون
> sen
[1] FALSE  TRUE FALSE FALSE  TRUE FALSE

> highBP<-BP>140 # أشخاص لديهم ضغط دم مرتفع
> highBP
[1] FALSE  TRUE FALSE FALSE FALSE FALSE
```

[1] لاحظ أنه إذا ما تم استخدام الإشارة (=) بدلا من (==) فإن هذا سيعني إعادة تعيين قيمة المتجه إلى القيمة الجديدة التي تلي إشارة يساوي.

لدينا الآن متجهان منطقيان هما sen وhighBP حيث يمثل الأول تبيان الأشخاص المسنين (نعم- لا)، (وهما الشخصين الثاني والخامس)، ويمثل المتجه الثاني تبيان الأشخاص الذين لديهم ضغط دم مرتفع (نعم- لا)، (وهو الشخص الثاني فقط). وسنقوم باستخدام عمليات الفئات على هذين المتجهين المنطقيين كما يلي:

```
> set1<-sen&highBP
> set2<-sen|highBP
> set3<-!highBP

> set1;set2;set3
[1] FALSE  TRUE FALSE FALSE FALSE FALSE
[1] FALSE  TRUE FALSE FALSE  TRUE FALSE
[1]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
```

حيث يُظهر المتجه المنطقي (أو الفئة) set1 ترتيب الأشخاص المسنين والذين لديهم ضغط دم مرتفع، (وهو الشخص الثاني)، والمتجه set2 ترتيب الأشخاص المسنين أو الذين لديهم ضغط دم مرتفع، (وهما الشخصين الثاني والخامس)، وأما المتجه set3 فيُظهر ترتيب الأشخاص الذين ليس لديهم ضغط دم مرتفع، (وهم جميع الأشخاص باستثناء الشخص الثاني).

استخدام آخر للمتجهات المنطقية، هو لإظهار ترتيب القيم المفقودة في البيانات، إن وُجدت، وذلك عن طريق الأمر is.na بالصورة التالية:

```
> x6<-c(5,1,4,NA,0,1)
> is.na(x6)
[1] FALSE FALSE FALSE  TRUE FALSE FALSE

> x7<-c("a","d","e",NA)
> is.na(x7)
[1] FALSE FALSE FALSE  TRUE
```

ولاحظ في نتيجة استخدام الدالة is.na مع المتجه العددي x6 ظهور TRUE في الترتيب الرابع والذي ظهرت فيه القيمة المفقودة في المتجه، وكذلك الأمر مع المتجه المُميز x7.

وعموما، يمكن استخدام الأوامر mode، class، أو str للتعرف[1] على نوع المتجه ما إذا كان متجه؛ عددي، مُميز، أو منطقي كما نرى في الأمثلة التالية:

```
> mode(x1)
[1] "numeric"

> class(Libya)
[1] "character"
```

[1] توجد دوال أخرى يمكن استخدامها لتغيير نوع المتجه، وسيتم تناولها لاحقا بحسب ما يتطلبه التحليل الإحصائي.

```
> str(highBP)
 logi [1:6] FALSE TRUE FALSE FALSE FALSE FALSE
```

و الاختصار logi في الأمر الأخير هو اختصار مصطلح (Logical) أي متجه منطقي، و[1:6] تفيد بوجود ستة قيم هي تلك المعروضة بعدها.

1.2.2 بعض الدوال الحسابية على المتجهات (Some Arithmetic Functions on Vectors)

بعد أن تم تصنيف أنواع المتجهات الرئيسية إلى عددية، مميزة، ومنطقية سنتناول الآن بعض الدوال الحسابية التي تستخدم في التعامل مع البيانات بصورة عامة والمتجهات العددية بصورة خاصة.

سنبدأ بدالتي **القيمة الصغرى** (Minimum)، والتي تعطي أقل قيمة ضمن مجموعة من القيم، و**القيمة الكبرى** (Maximum) التي تعطي أكبر قيمة ضمن مجموعة القيم، وهما min و max على الترتيب في لغة R، وكمثال، يمكن حساب هاتين القيمتين للمتجه الذي تم تعيينه سابقا باسم grades كما هو موضح أدناه، ولاحظ أنه يمكن تعيين قيم هاتين الدالتين إلى أشياء بحيث يمكن استخدامهما أو استدعاؤهما متى أردنا، وهذا بالطبع شأن كل الدوال المستخدمة في R:

```
> grades
[1]  6  7  4 10  8  5
> min.grades<-min(grades)
> min.grades
[1] 4

> max.grades<-max(grades)
> max.grades
[1] 10
```

ويمكن أيضا استخدام الدالة range للحصول على القيمتين الصغرى والكبرى مباشرة؛

```
> range(grades)
[1]  4 10
```

ولاحظ أن الدالة range لا تُعطي قيمة مقياس **المدى**[1] (Range)، وهو الفرق بين بين القيمتين الصغرى والكبرى، بل تقوم بعرض هاتين القيمتين فقط.

وتُستخدم دالة **المجموع** sum للحصول على مجموع القيم في المتجه، فمثلا:

```
> sum(x2)
[1] 25
```

[1] يمكن استخدام (max.grades-min.grades) لحساب المدى في البيانات grades في المثال.

أما الدالة length فتعطي عدد العناصر الموجودة في المتجه، (طول المتجه)؛

```
> length(x2)
[1] 5
```

والدالة prod تقوم بحساب ناتج ضرب قيم المتجه ببعضها البعض، على الصورة:

```
> prod(x2)
[1] 720
```

نأتي الآن لذكر بعض الدوال الرياضية التي تستخدم لترتيب القيم. فالدالة sort تقوم بترتيب قيم المتجه تصاعديا (أو تنازليا باستخدام الخيار[1] decreasing أو الدالة rev)، فمثلا:

```
> x1
[1]  2  4  5  0 -7

> sort(x1,decreasing=FALSE) # للترتيب التصاعدي
[1] -7  0  2  4  5
```

أو

```
> sort(x1) # للترتيب التصاعدي
[1] -7  0  2  4  5

> sort(x1,decreasing=TRUE) # للترتيب التنازلي
[1]  5  4  2  0 -7
```

أو

```
> rev(sort(x1)) # للترتيب التنازلي
[1]  5  4  2  0 -7
```

أما للحصول على **رُتب** (Rank) القيم في المتجه، فإنه يمكن استخدام الدالة order أو الدالة sort.list بالصورة التالية:

```
> order(x1,decreasing=FALSE) # لعرض الرتب تصاعديا
[1] 5 4 1 2 3
```

أو

```
> order(x1) # للترتيب التصاعدي
[1] 5 4 1 2 3
```

بمعنى أن القيمة الخامسة (وهي 7−) هي أقل قيمة، يليها القيمة الرابعة (وهي 0)، يليها القيمة الأولى (وهي 2)، ...، وهكذا. وأيضا:

```
> sort.list(x1,decreasing=TRUE) # لعرض الرتب تنازليا
[1] 3 2 1 4 5
```

[1] الخيارات هي إضافات تكون مدرجة في الكثير من دوال لغة R، وتُستخدم لإعطاء المُستخدِم مرونة في تطبيق هذه الدوال والحصول على النتيجة أو النتائج التي يرغب بها.

وفي سياق الحديث عن ترتيب المتجهات، يمكن استخدام الدالتين pmin أو pmax للحصول على متجه جديد يضم القيم الصغرى (أو الكبرى) المناظرة في المتجهين. كمثال على ذلك لنأخذ المتجهين x1 و x2 ونعيّن المتجه min.x1.x2 كمتجه يضم القيم الصغرى المناظرة في كلا المتجهين، والمتجه max.x1.x2 كمتجه يضم القيم الكبرى المناظرة في كلا المتجهين؛

```
> x1;x2
[1]  2  4  5  0 -7
[1]  1  3  8 10  3

> min.x1.x2<-pmin(x1,x2)
> min.x1.x2
[1]  1  3  5  0 -7

> max.x1.x2<-pmax(x1,x2)
> max.x1.x2
[1]  2  4  8 10  3
```

ونلاحظ في النتيجة الأخيرة مثلا أن المتجه max.x1.x2 يمثل الخمسة قيم الكبرى في كلا المتجهين x1 و x2 حيث تمت مقارنة كل عدد بنظيره في المتجه الآخر ثم تم اختيار القيمة الأكبر. ويمكن استخدام هذه الدوال مع أكثر من متجهين.

2.2.2 توليد السلاسل العددية (Generating Sequences)

يُعد استخدام السلاسل العددية أمرا مهما في لغة R ضمن كتابة الأوامر المُركبة، كما سنرى لاحقا، وكذلك لتوفير الوقت وعدد السطور في عملية كتابة الأوامر في العموم؛ فبدلا من كتابة أرقام متسلسلة كثيرة في متجه يمكن استخدام دالة السلسلة العددية لاختصار ذلك. وأبسط طريقة لتوليد سلاسل الأعداد بخطوة مساوية للواحد هي باستخدام رمز النقاط الشارحة (:) كما يلي:

```
> 1:7
[1] 1 2 3 4 5 6 7

> 3:9
[1] 3 4 5 6 7 8 9

> -5:5
[1] -5 -4 -3 -2 -1  0  1  2  3  4  5
> 1.5:7.5
[1] 1.5 2.5 3.5 4.5 5.5 6.5 7.5

> 20:15
[1] 20 19 18 17 16 15
```

لاحظ من العملية الأخيرة أنه يمكن توليد سلاسل عكسية أو تنازلية إذا ما تم وضع القيمة الأكبر أولا. ويمكن تعيين متجه للسلسلة العددية وإجراء عمليات حسابية مختلفة باستخدامها عند الحاجة كما توضح الأمثلة التالية:

```
> s1<-20:25
> s1
[1] 20 21 22 23 24 25

> s2<-(1:5)*2
> s2
[1]  2  4  6  8 10

> s3<-s2-(10:14)
> s3
[1] -8 -7 -6 -5 -4
```

أما إذا أردنا توليد السلسلة العددية بمزيد من الخيارات، مثل التحكم بالخطوات بين قيم السلسلة، فيمكن استخدام دالة **السلسلة** (Sequence) في R؛ seq بحيث يتم تعريف بداية السلسلة ونهايتها وعدد الخطوات المطلوب. والأمثلة التالية توضح الصورة:

```
> seq(1:6) # 1:6 تعطي نفس نتيجة
[1] 1 2 3 4 5 6

> seq(1,6,2) # 2 سلسلة بخطوة تساوي
[1] 1 3 5
```

ويُعرف الشكل العام لدالة seq بالصورة التالية، حيث أن الخياران from و to يحددان بداية ونهاية السلسلة على الترتيب والخيار by يحدد قيمة الخطوة:

```
> seq(from=5,to=8,by=.5)
[1] 5.0 5.5 6.0 6.5 7.0 7.5 8.0
```

ويتم توليد السلسلة بصورة عكسية أو تنازلية بوضع إشارة سالبة أمام قيمة الخطوة في الخيار by:

```
> seq(from=12,to=0,by=-4) # 4 سلسلة تنازلية بخطوة تساوي
[1] 12  8  4  0
```

أو ببساطة يمكن كتابة الأمر السابق بالصورة:

```
> seq(12,0,-4)
[1] 12  8  4  0
```

ويمكن استخدام خيار إضافي مع دالة seq هو length والذي يحدد طول السلسلة كما نشاهد في المثالين التاليين:

```
> seq(from=-1,length=10,by=0.2)
[1] -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8
```

```
> x2
[1]  1  3  8 10  3

> sum.x2<-sum(x2);sum.x2
[1] 25

> seq(from=sum.x2,length=6,by=-1.5)
[1] 25.0 23.5 22.0 20.5 19.0 17.5
```

ولاحظ في المثال الثاني أنه تم استخدام مجموع قيم المتجه x2 (والذي تم تعيين الاسم sum.x2 له)، كقيمة ابتدائية للسلسلة التنازلية.

ويمكن أيضا استخدام الخيار الإضافي along مع الدالة المولدة للسلسلة seq، والذي يتم تعيينه إلى قيم متجه آخر (سلسلة أولية) للحصول على عدد قيم في السلسلة الجديدة مساوي لعدد القيم في ذلك المتجه كما نشاهد في المثال التالي:

```
> s4<-1:9
> seq(1,5,along=s4)
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

ولاحظ أنه لم يتم تحديد حجم الخطوة في المثال السابق، بل تم الاعتماد على الخيار along لتحديد حجم الخطوة بصورة تلقائية تعتمد على عدد القيم في المتجه s4 وعلى نقطتي البداية والنهاية في دالة seq، ولاحظ أيضا أن عدد القيم في السلسلة الجديدة هو 9 وهو مساوي لعدد القيم في المتجه s4.

نعرّف الآن دالة **التكرار** (Replicate) التي تُعطى بالصورة rep، وتُستخدم لتوليد تكرارات للقيم، لذلك فهي تُعد مرتبطة بدالة السلسلة. هذه الدالة تضم خيارين هما times وeach واللذان يمكن توظيفهما بالصورة التالية:

```
> x1
[1]  2  4  5  0 -7

> rep(x1,times=4)
[1]  2  4  5  0 -7  2  4  5  0 -7  2  4  5  0 -7  2  4  5
0 -7

> rep(x1,each=4)
[1]  2  2  2  2  4  4  4  4  5  5  5  5  0  0  0  0 -7 -7
-7 -7
```

وتستطيع أن ترى الفرق بين الخيارين السابقين؛ فالخيار times تم استخدامه لتكرار قيم المتجه كمجموعة واحدة أربع مرات، وأما الخيار each فاستُخدِم لتكرار كل قيمة من قيم المتجه أربع مرات.

ونختم هذا البند بتذكير القارئ من جديد بضرورة تخزين كل ما تم تعريفه من أشياء حتى الآن في ملف العمل الحالي "work2" لمواكبة شرح الكتاب. وللتأكد من تلك المواكبة نقوم باستدعاء الأشياء المخزنة في ملف العمل، والتي من المفروض[1] أن تكون:

```
> ls()
[1] "age"    "bad.grades"    "BP"    "grades"      "highBP"
[6] "Libya"    "max.grades"    "max.x1.x2"    "min.grades"
"min.x1.x2"
[11] "s1"        "s2"         "s3"         "s4"         "sen"
[16] "set1"      "set2"       "set3"       "sum.x2"     "x1"
[21] "x2"        "x2.1"       "x3"         "x4"         "x5"
[26] "x6"        "x7"
```

ونلاحظ هنا ظهور أرقام أخرى بين الأقواس غير الرقم [1] الذي اعتدنا رؤيته في النتائج السابقة. هذه الأرقام في بداية كل سطر هي ببساطة ترقيم تسلسلي للنتائج أو الأشياء الموجودة، فالترقيم [6] مثلا في بداية السطر الثاني في النتائج يعني أن الشيء "Libya" هو سادس نتيجة، والترقيم [11] في بداية السطر ثالث يعني أن الشيء "s1" هو النتيجة الحادية عشر، وهكذا. فالفكرة العامة من هذا الترقيم هو تنظيم مظهر النتائج، وإذا ما قمتَ بتوسيع[2] نافذة لوحة مراقبة R ثم أعدت تنفيذ أمر استدعاء الأشياء ls() مرة أخرى فسوف تلاحظ تغيرا في ترقيم النتائج.

3.2 نُظم الصفوف والمصفوفات في R (Arrays and Matrices in R)

في البنود السابقة، تعرّفنا على بعض الدوال الحسابية والأوامر التي تم تطبيقها على قيم مفردة أو متجهات عددية وغير عددية. والآن سنوضح كيفية التعامل مع النُظم التي تحتوي على أكثر صف أو عامود. وتُعد **نُظم الصفوف** (Arrays) بمثابة الحالة العامة التي توجد عليها قواعد البيانات التي تحتوي على k من الأبعاد (k-Dimentions)، وتُعتبر **المصفوفات** حالة خاصة من نُظم الصفوف حيث تحتوي على بُعدين (2-Dimentions (2D))، ويمكننا رؤية الفرق بين تركيبة كلا من المتجه والمصفوفة ونظام الصفوف[3] من

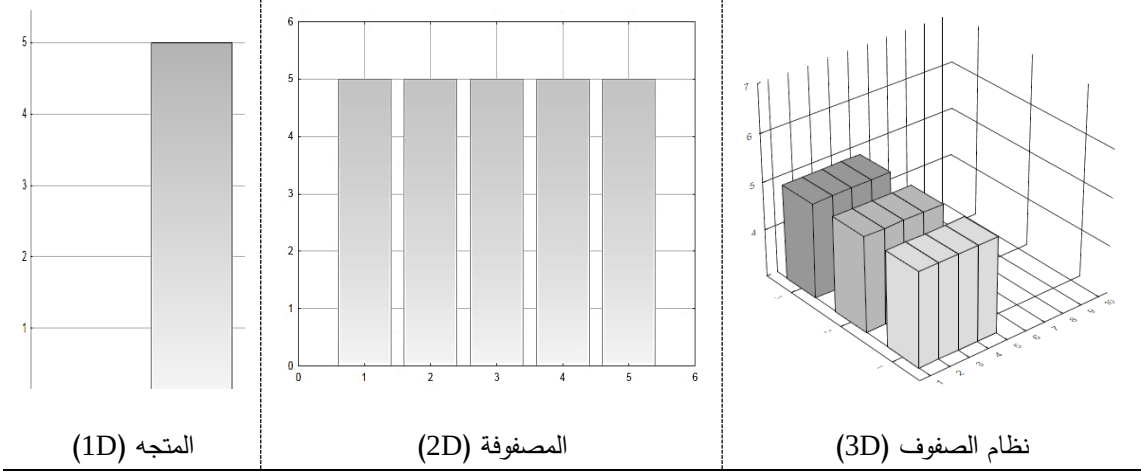
---

[1] إذا لم تكن الأشياء في ملف العمل الخاص بك مطابقة للأشياء الموجودة هنا يمكنك إعادة تعيين الناقص وحذف الأشياء الزائدة.

[2] يمكن القيام بذلك إما باستخدام الزر الأيسر للفأرة لتوسيع النافذة بالتدريج حسب المساحة المرغوبة، أو الضغط على مربع التكبير في زاوية النافذة.

[3] نعرض في الشكل (1.2) نظام صفوف ثلاثي الأبعاد كحالة خاصة من الـ k بُعد، علما بأنه لا يمكن تمثيل أكثر من ثلاثة محاور بيانيا.

خلال التمثيل البياني لبيانات افتراضية، كما هو موضح في الشكل (1.2)، علما بأن كل هذه النُظم يمكن أن تحوي قيماً عددية أو غير عددية.



شكل 1.2: الأبعاد العامة للمتجه، المصفوفة، ونظام الصفوف

وسنقوم بالتركيز على إنشاء المصفوفات على وجه الخصوص لأنها تُعد الأكثر استخداما بالنسبة لمستخدمي التحليل الإحصائي أو الرياضي، إلا أننا سوف نعرض مثال واحد أولا على نظام صفوف ذو ثلاثة أبعاد وذلك باستخدام الدالة `array` والتي تُستخدم لإنشاء نُظم الصفوف والمصفوفات بصورة عامة؛

```
> v1<-1:24
> ar1<-array(v1,dim=c(3,4,2))
> ar1

, , 1

     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12

, , 2

     [,1] [,2] [,3] [,4]
[1,]   13   16   19   22
[2,]   14   17   20   23
[3,]   15   18   21   24
```

ولاحظ أنه تم وضع قيم البيانات كمتجه عددي (سلسلة عددية في مثالنا) داخل خيارات الدالة `array` ثم تم استخدام الخيار `dim` لتخصيص عدد القيم للبُعد أو المحور الأول (عدد الصفوف يساوي 3)، وعدد القيم للمحور الثاني (عدد الأعمدة يساوي 4)، وعدد القيم للمحور الثالث أو الارتفاع بيانيا (تكرار الصفوف والأعمدة

يساوي 2)، ولهذا ظهرت قيمة نظام الصفوف ar1 على شكل مصفوفتين متتاليتين كل منهما ذات بُعدين ولها الترتيب 4×3.

1.3.2 إنشاء المصفوفات (Constructing Matrices)

إذا ما أردنا إنشاء مصفوفة (نظام صفوف ذو بُعدين) فيمكن استخدام الدالة array بالصورة التالية مثلا:

```
> mat1<-array(v1,dim=c(4,6))
> mat1
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    5    9   13   17   21
[2,]    2    6   10   14   18   22
[3,]    3    7   11   15   19   23
[4,]    4    8   12   16   20   24
```

ونلاحظ من المثال أعلاه أن الترتيب المُعطى في الخيار dim لابد أن يتناسب مع عدد القيم في المتجه v1، بمعنى أن حاصل ضرب عدد الصفوف (4) وعدد الأعمدة (6) لابد أن يساوي دائما عدد العناصر الكلي (وهو 24) في المتجه. ملاحظة أخرى خاصة بترتيب القيم وهي أنه تم ترتيبها افتراضيا في المصفوفة عن طريق ملء الأعمدة وليس الصفوف، إلا أن ذلك الترتيب يمكن تغيره كما سنوضح لاحقا. ويكون؛

```
> class(mat1)
[1] "matrix"
```

ولاحظ أيضا في الشكل العام أن الصفوف في المصفوفة تم إعطاؤها دلالات أو "أسماء" على الصورة [1,],...,[4,]، حيث يمثل الرقم في البداية ترتيب الصف والفاصلة (,) بعده للدلالة على وجود ترتيب آخر هو الأعمدة، وبالمثل نلاحظ أن الأعمدة تأخذ الأسماء [,1],...,[,6] حيث تكون الفاصلة هي في البداية للدلالة على وجود ترتيب للصفوف ويليها ترقيم الأعمدة.

طريقة أخرى لإنشاء المصفوفات هي باستخدام دالة **الأبعاد** dim ، (والتي استُخدمت كخيار في المثال السابق)، مباشرة لتحويل المتجه إلى مصفوفة كما يوضح المثال التالي:

```
> v2<-c(2,1,5,0,-3,8,12,6,4,-1,15,7)
> v2
 [1]  2  1  5  0 -3  8 12  6  4 -1 15  7

> dim(v2)<-c(4,3)
> v2
     [,1] [,2] [,3]
[1,]    2   -3    4
[2,]    1    8   -1
[3,]    5   12   15
[4,]    0    6    7
```

```
> class(v2)
[1] "matrix"
```

نأتي الآن لدالة **المصفوفة** matrix والتي تُستخدم أيضا، كما هو واضح من اسمها، لتكوين المصفوفات باستخدام الخيار nrow (أو ncol)[1] والخيار byrow كما سنرى من المثال التالي، حيث سنقوم باستدعاء v2 من جديد بصورته الأصلية، (وذلك عن طريق استخدام السهم العلوي "↑" في لوحة المفاتيح كما أوضحنا سابقا)، واستخدامه كمتجه لأنه حاليا يُعتبر مصفوفة:

```
> v2<-c(2,1,5,0,-3,8,12,6,4,-1,15,7)
> v2
 [1]  2  1  5  0 -3  8 12  6  4 -1 15  7

> mat2<-matrix(v2,nrow=6,byrow=T)
> mat2
     [,1] [,2]
[1,]    2    1
[2,]    5    0
[3,]   -3    8
[4,]   12    6
[5,]    4   -1
[6,]   15    7

> class(mat2)
[1] "matrix"
```

ولاحظ أنه عندما تم طلب إنشاء المصفوفة السابقة بعدد صفوف يساوي 6 فإن عدد الأعمدة أصبح 2 ليتوافق مع عدد القيم الكلية في المتجه v2 (وهو 12 قيمة). والملاحظة الأخرى هي أن الخيار byrow تم استخدامه وكتابته مساويا لـ T، (اختصارا لـ TRUE)، لاختيار إنشاء المصفوفة عن طريق ملء الصفوف وليس الأعمدة. وعموما، يمكننا إنشاء أي مصفوفة من متجه بهذه الطريقة عن طريق تحديد عدد الصفوف أو الأعمدة المرغوب فيه. وإذا لم يتم استخدام الخيار byrow فإنه سيتم ملء الأعمدة وليس الصفوف في المصفوفة كما نرى من المثال التالي:

```
> mat3<-matrix(v2,nrow=6)
> mat3
     [,1] [,2]
[1,]    2   12
[2,]    1    6
[3,]    5    4
[4,]    0   -1
[5,]   -3   15
[6,]    8    7
```

[1] يتم استخدام أحد الخيارين nrow أو ncol وليس كلاهما لتحديد عدد الصفوف أو عدد الأعمدة لأن برنامج R يقوم بحساب البُعد الآخر تلقائيا بحيث يتناسب مع عدد العناصر الكلي.

ويمكن من ناحية تنظيمية إعطاء الأسماء التي نريدها للأعمدة أو الصفوف أو كلاهما وذلك باستخدام الدالتين colnames وrownames كما يتضح من المثال التالي، والذي سنقوم فيه باستخدام المتجه v2 أيضا بصورته الأصلية من جديد وإنشاء مصفوفة جديدة لها الترتيب 4×3 ، ثم إعادة تسمية الأعمدة بالأسماء (c1, c2, c3, c4) والصفوف بالأسماء (r1, r2, r3):

```
> mat4<-matrix(v2,nrow=3)
> mat4

     [,1] [,2] [,3] [,4]
[1,]    2    0   12   -1
[2,]    1   -3    6   15
[3,]    5    8    4    7

> colnames(mat4)<-c("c1","c2","c3","c4")
> mat4

     c1 c2 c3 c4
[1,]  2  0 12 -1
[2,]  1 -3  6 15
[3,]  5  8  4  7

> rownames(mat4)<-c("r1","r2","r3")
> mat4

   c1 c2 c3 c4
r1  2  0 12 -1
r2  1 -3  6 15
r3  5  8  4  7
```

ولاحظ أنه يمكن إدراج أسماء الصفوف والأعمدة في سطر أوامر واحد، فمثلا؛

```
>colnames(mat3)<-c("a1","a2");rownames(mat3)<-
c("b1","b2","b3","b4","b5","b6")

>  mat3
   a1 a2
b1  2 12
b2  1  6
b3  5  4
b4  0 -1
b5 -3 15
b6  8  7
```

كما يمكن تسمية الأعمدة أو الصفوف بشكل سريع بأحرف الأبجدية الإنجليزية باستخدام الدالة LETTERS للأحرف الكبيرة والدالة letters للأحرف الصغيرة كما يوضح المثال التالي:

```
> mat5<-mat1
> mat5
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    5    9   13   17   21
[2,]    2    6   10   14   18   22
[3,]    3    7   11   15   19   23
[4,]    4    8   12   16   20   24

> colnames(mat5)<-LETTERS[1:6]
> mat5
     A B  C  D  E  F
[1,] 1 5  9 13 17 21
[2,] 2 6 10 14 18 22
[3,] 3 7 11 15 19 23
[4,] 4 8 12 16 20 24
```

وهنا لدينا نقطتين؛ الأولى تتعلق بتعريف mat5 مساوية للمصفوفة mat1 وذلك لعدم تغيير الحالة الأصلية التي عليها mat1، وإجراء تغيير الأسماء على مصفوفة جديدة، وأما النقطة الثانية فهي استخدام القوسين [1:6] بهذا التسلسل حيث أنه يوجد في المصفوفة ستة أعمدة يُراد تغيير أسمائها بالأحرف الصغيرة الستة الأولى من الأبجدية الإنجليزية، إلا أنه يمكن اختيار أي تسلسل آخر للأحرف. وإذا ما أردنا إعطاء أرقام متسلسلة لصفوف المصفوفة، يمكننا تنفيذ التالي:

```
> rownames(mat5)<-1:4

> mat5
  A B  C  D  E  F
1 1 5  9 13 17 21
2 2 6 10 14 18 22
3 3 7 11 15 19 23
4 4 8 12 16 20 24
```

2.3.2 دمج المتجهات والمصفوفات (Merging Vectors and Matrices)

ومن ضمن الطرق المستخدمة لإنشاء المصفوفات هي دمج المتجهات مع بعضها البعض أو حتى دمج المصفوفات مع بعضها لتكوين مصفوفة جديدة، هذا الدمج يمكن أن يتم عن طريق رص الصفوف تحت بعضها باستخدام دالة rbind، أو رص الأعمدة إلى جانب بعضها باستخدام دالة cbind. ويجب عند استخدام هاتين الدالتين أن يكون عدد العناصر في صفوف المتجهات متساوي عند دمج الصفوف، وأن يكون عدد العناصر في الأعمدة متساوي عند دمج الأعمدة، كما يتضح من المثالين التاليين:

```
> mat6<-rbind(x1,x2,x3)
> mat6

   [,1] [,2] [,3] [,4] [,5]
x1    2    4    5    0   -7
x2    1    3    8   10    3
x3    7   19   45   50    8

> mat7<-cbind(x1,x2,x3)
> mat7

     x1 x2 x3
[1,]  2  1  7
[2,]  4  3 19
[3,]  5  8 45
[4,]  0 10 50
[5,] -7  3  8
```

ولاحظ كيف أن أسماء المتجهات الأصلية انتقلت كأسماء للصفوف في المثال الأول، وكأسماء للأعمدة في المثال الثاني. مثال آخر على دمج المتجهات المكونة من سلاسل عددية كأعمدة، مع تسميتها بالأحرف الكبيرة للغة الإنجليزية، يمكن أن يكون على الصورة:

```
> mat8<-cbind(A=1:5,B=6:10,C=11:15,D=16:20)
> mat8

     A  B  C  D
[1,] 1  6 11 16
[2,] 2  7 12 17
[3,] 3  8 13 18
[4,] 4  9 14 19
[5,] 5 10 15 20
```

ويمكن دمج مصفوفتين (أو أكثر) للحصول[1] على مصفوفة جديدة بالصورة:

```
> cbind(mat7,mat8)

     x1 x2 x3 A  B  C  D
[1,]  2  1  7 1  6 11 16
[2,]  4  3 19 2  7 12 17
[3,]  5  8 45 3  8 13 18
[4,]  0 10 50 4  9 14 19
[5,] -7  3  8 5 10 15 20
```

[1] يجب التنويه هنا إلى أن دمج المصفوفات أو المتجهات بحد ذاته لا يعني تنفيذ أي عملية حسابية مثل الجمع أو الضرب أو غيرها.

لاحظ أنه لا يمكن دمج المصفوفتين mat7 وmat8 باستخدام الدالة rbind لأن عدد الأعمدة في كل منهما غير متساوي. ولاحظ أيضا أننا لم نقم بتعيين اسم للمصفوفة الجديدة ولذلك فإنه لن يتم تخزينها في الذاكرة حتى بعد تخزين ملف العمل الحالي (work2).

كما ذكرنا في بداية هذا البند، فإن نُظم الصفوف والمصفوفات يمكن أن تحتوي على قيم غير عددية أيضا، والمثال التالي يوضح استخدام أمر الدمج لتكوين مصفوفة من متجهين مُميزين:

```
> Libya
[1] "Benghazi" "Tripoli"  "Tobruk"

> Egypt<-c("Cairo","Alex","Suiz")

> Lib.Egy<-rbind(Libya,Egypt)
> Lib.Egy
      [,1]       [,2]      [,3]
Libya "Benghazi" "Tripoli" "Tobruk"
Egypt "Cairo"     "Alex"    "Suiz"
```

3.3.2 استدعاء القيم من المصفوفات (Calling values from Matrices)

الآن وقد شرحنا كيفية تكوين نُظم الصفوف بصورة عامة والمصفوفات بصورة خاصة، سنقدم بعض الأمثلة على كيفية استدعاء وعرض[1] بعض القيم من المصفوفة. والطريقة مشابهة لتلك المستخدمة مع المتجهات سابقا، إلا أننا نتعامل هنا مع مصفوفات ذات بُعدين، لذلك سنستخدم الأقواس المربعة [ ] بحيث تكون القيمة الأولى ضمن القوسين لترتيب الصف وتكون القيمة الثانية لترتيب العامود كما سنرى في الأمثلة التالية:

لعرض القيمة التي في الصف الخامس والعامود الأول في المصفوفة mat7 نكتب:

```
> mat7
     x1 x2 x3
[1,]  2  1  7
[2,]  4  3 19
[3,]  5  8 45
[4,]  0 10 50
[5,] -7  3  8

> mat7[5,1]
x1
-7
```

[1] يمكن كذلك استخدام تلك القيم المختارة من المصفوفات ضمن العمليات الحسابية المختلفة كما سنرى لاحقا.

ولعرض القيم في العامود الثالث وكل الصفوف ما عدا الصف الرابع نكتب:

```
> mat7[-4,3]

[1]  7 19 45  8
```

ولعرض القيم الموجودة في الصف الثالث إلى الخامس ضمن العامود الثاني نكتب:

```
> mat7[3:5,2]

[1]  8 10  3
```

أما لعرض أي صف أو عامود كامل فإننا نكتب ترتيبه فقط ونحذف ترتيب الاتجاه الآخر:

```
> mat7[2,]

x1 x2 x3
 4  3 19

> mat7[,3]

[1]  7 19 45 50  8
```

وإذا ما أردنا أن تظهر النتائج السابقة بصورة مصفوفة وليس بالصورة التقليدية، يمكننا استخدام الخيار drop لعمل ذلك كما نرى في الأمثلة:

```
> mat7[2,,drop=F]

     x1 x2 x3
[1,]  4  3 19

> mat7[,3,drop=F]

     x3
[1,]  7
[2,] 19
[3,] 45
[4,] 50
[5,]  8
```

ويمكن أيضا الحصول على مصفوفات فرعية من المصفوفة الأصلية بالصورة:

```
> mat7[1:4,1:2]

     x1 x2
[1,]  2  1
[2,]  4  3
[3,]  5  8
[4,]  0 10
```

```
> mat7[1:4,c(1,3)]

     x1 x3
[1,]  2  7
[2,]  4 19
[3,]  5 45
[4,]  0 50

> mat7[c(1,3,4),c(1,3)]

     x1 x3
[1,]  2  7
[2,]  5 45
[3,]  0 50

> mat7[2:4,]

     x1 x2 x3
[1,]  4  3 19
[2,]  5  8 45
[3,]  0 10 50

> mat7[,c(1,3)]

     x1 x3
[1,]  2  7
[2,]  4 19
[3,]  5 45
[4,]  0 50
[5,] -7  8

> mat7[-5,-1]

     x2 x3
[1,]  1  7
[2,]  3 19
[3,]  8 45
[4,] 10 50
```

4.3.2 استبدال القيم في المصفوفات (Replacing values in Matrices)

يمكن أيضا استبدال أي قيمة أو مجموعة من القيم أو حتى صف أو عامود كامل من المصفوفة بقيم أخرى كما نرى في الأمثلة التالية على المصفوفة mat8:

لاستبدال القيمة (12) الموجودة في الصف الثاني والعامود الثالث بالقيمة (21) نكتب:

```
> mat8
     A  B  C  D
[1,] 1  6 11 16
[2,] 2  7 12 17
[3,] 3  8 13 18
[4,] 4  9 14 19
[5,] 5 10 15 20
> mat8[2,3]<-21
> mat8
     A  B  C  D
[1,] 1  6 11 16
[2,] 2  7 21 17
[3,] 3  8 13 18
[4,] 4  9 14 19
[5,] 5 10 15 20
```

لاستبدال ستة قيم مثلا من الصفين الأولين والأعمدة الثلاثة الأولى نكتب:

```
> mat8[1:2,1:3]<-c(5,4,14,9,31,35)
> mat8
     A  B  C  D
[1,] 5 14 31 16
[2,] 4  9 35 17
[3,] 3  8 13 18
[4,] 4  9 14 19
[5,] 5 10 15 20
```

ولاحظ أن الاستبدال تم عن طريق ملء الأعمدة المُختارة وليس الصفوف. ولاستبدال الصف الثالث مثلا بأكمله بالقيم (1 ,23 ,16 ,2-) نكتب:

```
> mat8[3,]<-c(-2,16,23,1)
> mat8
      A  B  C  D
[1,]  5 14 31 16
[2,]  4  9 35 17
[3,] -2 16 23  1
[4,]  4  9 14 19
[5,]  5 10 15 20
```

4.2 الدوال الحسابية الأساسية على المصفوفات في R

(Basic Arithmetic Functions on Matrices in R)

في البند (1.2.2) تم عرض بعض الدوال التي تُستخدم لإجراء العمليات الحسابية على المتجهات، (والتي يسري معظمها على القيم المفردة على اعتبار أنها حالة خاصة من المتجهات)، وفي هذا البند سنتناول بعض أهم الدوال والعمليات الحسابية التي يمكن استخدامها للتعامل مع المصفوفات في R.

ومن أبسط تلك العمليات هي الجمع، الطرح، الضرب، والقسمة باستخدام قيمة مفردة، بمعنى أن يتم مثلا جمع قيمة واحدة مع كل قيمة من قيم المصفوفة، كما نرى في المثال التالي:

```
> mat8
```

```
      A  B  C  D
[1,]  5 14 31 16
[2,]  4  9 35 17
[3,] -2 16 23  1
[4,]  4  9 14 19
[5,]  5 10 15 20
```

```
> mat8+10
```

```
      A  B  C  D
[1,] 15 24 41 26
[2,] 14 19 45 27
[3,]  8 26 33 11
[4,] 14 19 24 29
[5,] 15 20 25 30
```

وكذلك يمكن أخذ الجذر التربيعي للقيم في المصفوفة مثلا كالتالي:

```
> sqrt(mat8)
```

```
             A        B        C        D
[1,] 2.236068 3.741657 5.567764 4.000000
[2,] 2.000000 3.000000 5.916080 4.123106
[3,]      NaN 4.000000 4.795832 1.000000
[4,] 2.000000 3.000000 3.741657 4.358899
[5,] 2.236068 3.162278 3.872983 4.472136
Warning message:
In sqrt(mat8) : NaNs produced
```

ولاحظ ظهور التعبير NaN عند حساب الجذر التربيعي للقيمة السالبة (2−) وظهور رسالة تحذيرية بذلك. وبالطبع يمكن تنفيذ العمليات الأخرى مباشرة بنفس الطريقة.

ويمكنك ملاحظة أن إجراء أي عملية حسابية لا يغير من قيم المصفوفة الأصلية بل يُنتِج مصفوفة تحتوي على القيم الجديدة، وتلك المصفوفة لا يتم تخزينها إلا إذا تم تعيين اسم لها، فمثلا:

```
> mat9<-mat8*2
> mat9
      A  B  C  D
[1,] 10 28 62 32
[2,]  8 18 70 34
[3,] -4 32 46  2
[4,]  8 18 28 38
[5,] 10 20 30 40
```

أما إذا أردنا جمع (أو طرح) مصفوفتين (أو أكثر) فيمكن تنفيذ ذلك إذا ما كان للمصفوفات نفس الترتيب، أي نفس عدد الصفوف والأعمدة. فمثلا يمكننا كتابة:

```
> mat8+mat9
      A  B   C  D
[1,] 15 42  93 48
[2,] 12 27 105 51
[3,] -6 48  69  3
[4,] 12 27  42 57
[5,] 15 30  45 60
```

كذلك يمكن جمع متجهين أو متجه ومصفوفة كما نرى في الأمثلة التالية:

```
> x1
[1]  2  4  5  0 -7
> x2
[1]  1  3  8 10  3

> x1+x2
[1]  3  7 13 10 -4

> x2+mat9
      A  B  C  D
[1,] 11 29 63 33
[2,] 11 21 73 37
[3,]  4 40 54 10
[4,] 18 28 38 48
[5,] 13 23 33 43
```

ولاحظ أنه في العملية الأخيرة تم جمع قيم المتجه x2 (ذو الترتيب 5×1) مع قيم كل عامود في المصفوفة mat9 (ذات الترتيب 5×4).

- **ضرب المصفوفات:**

نأتي الآن للعمليات الحسابية المتعلقة بضرب المصفوفات والمتجهات، ونبدأها **بالضرب الخارجي** للمصفوفات (Outer Product) الذي يُعرف في لغة R بالدالة `outer` أو `%o%` (حيث `%` هو رمز النسبة، و `o` هو أحد حروف اللغة الإنجليزية). وتقوم هذه الدالة بضرب قيم المصفوفة الأولى بكل قيمة من قيم المصفوفة الثانية، حيث يظهر ناتج الضرب الخارجي كمصفوفات متعددة عددها يساوي عدد العناصر في المصفوفة الثانية، والمثال التالي يوضح الفكرة:

لنعرّف (نستخرج) مصفوفة ذات ترتيب 2×2 ، وأخرى ذات ترتيب 2×3 من المصفوفة `mat9` بالصورة التالية:

```
> mat9.1<-mat9[1:2,1:2]
> mat9.1

      A  B
[1,] 10 28
[2,]  8 18
> mat9.2<-mat9[3:5,3:4]
> mat9.2

      C  D
[1,] 46  2
[2,] 28 38
[3,] 30 40
```

وبتنفيذ الضرب الخارجي على المصفوفتين `mat9.1` و `mat9.2` نحصل على:

```
> mat9.3<-mat9.1%o%mat9.2
> mat9.3

, , 1, C
       A    B
[1,] 460 1288
[2,] 368  828

, , 2, C
       A   B
[1,] 280 784
[2,] 224 504

, , 3, C
       A   B
[1,] 300 840
[2,] 240 540
```

```
, , 1, D
      A  B
[1,] 20 56
[2,] 16 36

, , 2, D
       A    B
[1,] 380 1064
[2,] 304  684

, , 3, D
       A    B
[1,] 400 1120
[2,] 320  720
```

ويمكن استخدام سطر الأوامر `mat9.3<-outer(mat9.1,mat9.2,"*")` للحصول على نفس النتيجة.

- **المحورة، المعكوس، ومحدد المصفوفة:**

ومن ضمن الدوال الهامة في عملية ضرب المصفوفات دالة **التحوير** (Transpose) والتي تأخذ الصيغة `t` ببساطة. وتحوير المصفوفة هو قلب ترتيبها بتغيير أعمدتها إلى صفوف وتغيير صفوفها إلى أعمدة كما يوضح المثال التالي:

```
> mat9

      A  B  C  D
[1,] 10 28 62 32
[2,]  8 18 70 34
[3,] -4 32 46  2
[4,]  8 18 28 38
[5,] 10 20 30 40

> mat9.t<-t(mat9)
> mat9.t

  [,1] [,2] [,3] [,4] [,5]
A   10    8   -4    8   10
B   28   18   32   18   20
C   62   70   46   28   30
D   32   34    2   38   40
```

دالة أخرى لا تقل أهمية عن دالة التحوير في عمليات المصفوفات هي دالة **معكوس** المصفوفة (Matrix Inverse) والتي تُعرّف بالصيغة[1] solve. وللحصول على معكوس المصفوفة[2] يتم استخدام دالة المعكوس مباشرة، إلا أننا سنقوم أولا هنا بإضافة عامود إضافي افتراضي للمصفوفة mat9 لتصبح مصفوفة مربعة:

```
> mat9.sq<-cbind(mat9,x1)
> mat9.sq
```

```
      A  B  C  D x1
[1,] 10 28 62 32  2
[2,]  8 18 70 34  4
[3,] -4 32 46  2  5
[4,]  8 18 28 38  0
[5,] 10 20 30 40 -7
```

من الناحية الشكلية نلاحظ أن العامود (المتجه) الذي تمت إضافته يحتفظ باسم المتجه الأصلي x1 لذلك سنقوم بتغيير اسمه إلى E ليتوافق مع نسق المصفوفة؛

```
> colnames(mat9.sq)<-LETTERS[1:5]
> mat9.sq
```

```
      A  B  C  D  E
[1,] 10 28 62 32  2
[2,]  8 18 70 34  4
[3,] -4 32 46  2  5
[4,]  8 18 28 38  0
[5,] 10 20 30 40 -7
```

(لاحظ أنه كان من الملائم أن يتم تغيير اسم العامود في المتجه أولا ثم إضافته إلى المصفوفة). على كل حال لنقم الآن بحساب المعكوس للمصفوفة:

```
> solve(mat9.sq)
```

```
         [,1]        [,2]         [,3]        [,4]         [,5]
A  0.24305104 -0.11659818 -0.091286673 -0.02987253 -0.062389142
B  0.04163867 -0.05381242  0.018458945  0.01807886 -0.005668235
C -0.01166698  0.02956406  0.002500854 -0.03291322  0.015346649
D -0.06229550  0.02825304  0.008631802  0.04829292  0.004511452
E  0.06020778 -0.03216957 -0.017627163  0.14388172 -0.156628364
```

أما لحساب **محدد** المصفوفة (Matrix Determinant) فنستخدم الدالة det، فمثلا يمكن حساب محدد المصفوفة السابقة mat9.sq بالصورة التالية؛

---

[1] توجد استخدامات إضافية أخرى للدالة solve والتي يمكنك التعرف عليها باستخدام دالة المساعدة help.

[2] من الناحية الرياضة، يتم حساب المعكوس للمصفوفات المربعة فقط.

```
> det(mat9.sq)
[1] 1452304
```

لنتناول الآن عملية ضرب مصفوفتين فقط[1]، ولتنفيذ ذلك رياضيا لابد أن يكون عدد الأعمدة في المصفوفة الأولى مساويا لعدد الصفوف في المصفوفة الثانية ويكون الناتج مصفوفة جديدة لها عدد صفوف المصفوفة الأولى وعدد أعمدة المصفوفة الثانية. فمثلا يمكننا ضرب المصفوفة mat9.t (ذات الترتيب 5×4) في المصفوفة mat9 (ذات الترتيب 4×5) لنحصل على مصفوفة جديدة لها الترتيب 4×4 باستخدام الصيغة[2] %*% كما يلي:

```
> mat9.t%*%mat9
```

```
     A    B     C    D
A  344  640  1520 1288
B  640 2856  5572 3056
C 1520 5572 12544 6720
D 1288 3056  6720 5228
```

الدالة diag (اختصار (Diagonal) أي قُطْري)، من ناحية أخرى هي دالة هامة أيضا واستخدامها يعتمد على ما نريد تنفيذه، فاستخدام هذه الدالة مع متجه يُعطي مصفوفة قُطرية عناصر القُطر فيها هي قيم المتجه، واستخدامها، أي الدالة، مع قيمة مفردة يُنتج مصفوفة **الوحدة** (Identity Matrix) بترتيب مساوي لتلك القيمة المستخدمة. أما استخدام الدالة القُطرية مع مصفوفة مربعة[3] فيقوم بعرض عناصر القطر. والأمثلة التالية توضح ذلك:

```
> x3
[1]  7 19 45 50  8

> diag(x3)
     [,1] [,2] [,3] [,4] [,5]
[1,]    7    0    0    0    0
[2,]    0   19    0    0    0
[3,]    0    0   45    0    0
[4,]    0    0    0   50    0
[5,]    0    0    0    0    8
```

[1] يمكن بالطبع ضرب أي عدد من المصفوفات ببعضها البعض، (إذا ما تم ترتيبها بصورة مناسبة)، إلا أننا سنترك ذلك بحسب حاجة المستخدم.

[2] يمكن استخدام معامل الضرب (*) بمفرده لضرب المصفوفات المربعة متساوية الترتيب.

[3] يمكن أيضا استخدام دالة diag مع المصفوفات الغير مربعة حيث أنها ستعرض عناصر القُطر لأصغر ترتيب مربع للمصفوفة، فمثلا لمصفوفة ترتيبها 4×6 سيتم عرض أول أربعة عناصر في القُطر.

```
> diag(7)

     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    1    0    0    0    0    0    0
[2,]    0    1    0    0    0    0    0
[3,]    0    0    1    0    0    0    0
[4,]    0    0    0    1    0    0    0
[5,]    0    0    0    0    1    0    0
[6,]    0    0    0    0    0    1    0
[7,]    0    0    0    0    0    0    1
```

```
> diag(mat9.t%*%mat9)

    A     B     C     D
  344  2856 12544  5228
```

وفي نهاية هذا البند، نذكر القارئ بالخيارين nrow و ncol اللذان تم استخدامهما سابقا لتحديد عدد الصفوف وعدد الأعمدة عند تكوين المصفوفة، حيث أنه يمكن استخدامهما أيضا للتحقق من عدد الصفوف والأعمدة للمصفوفات؛

```
> nrow(mat9)
[1] 5

> ncol(mat9)
[1] 4
```

5.2 نظام القوائم (Lists)

إن القوائم (وأُطر البيانات التي سنتناولها في الفصل التالي) هي دوال أساسية في R يتم من خلالها إدراج وتنظيم البيانات بشكل محدد تمهيدا لاستخدامها حسب الحاجة.

وتُعرّف **القائمة** في نظام R بأنها شيء يحتوي على مجموعة مرتبة من الأشياء تُسمى **مكونات** القائمة. وما يميز المكونات أو البيانات في القائمة أنه ليس من الضروري أن تكون كلها من نفس النوع (على غرار المتجهات)، بمعنى أنه يمكن للقائمة أن تضم قيم أو متجهات عددية وغير عددية في نفس الوقت، وأيضا يمكن أن تكون هذه المكونات ذات أحجام (أي أعداد قيم) مختلفة. باختصار، يمكن اعتبار القائمة بشكل عام بمثابة سجل يضم معلومات متعددة تتعلق بأي شيء.

ولتوضيح الصورة، لنفرض أننا نريد إدراج معلومات تتعلق باسم مريض وعمره بالسنوات ورقمه التسلسلي في المستشفى ورمز الجناح الذي يمكث فيه، عندئذ يمكن استخدام دالة **القائمة** وصيغتها list بالطريقة التالية:

```
> list1<-list(pait.name="Ahmed Omar",pait.age=39,
pait.no=155024,pait.ward="card74")
```

لاحظ أن البيانات غير العددية لابد أن يتم إدراجها بين علامات الاقتباس، كما هي الطريقة المتبعة دائما في إدخالها ضمن دوال R. وإذا ما تم استدعاء القائمة التي أعطيناها الاسم list1 فإننا سنحصل على المعلومات (وهي مكونات القائمة) بنفس ترتيب إدخالها كالتالي:

```
> list1

$pait.name
[1] "Ahmed Omar"

$pait.age
[1] 39

$pait.no
[1] 155024
$pait.ward
[1] "card74"
```

ويمكن عرض مكونات محددة من القائمة، إذا لم نرغب في عرض كل مكوناتها باستخدام الطرق التالية[1]:

```
> list1[1] #لعرض اسم المريض

$pait.name
[1] "Ahmed Omar"

> list1[1];list1[2] # لعرض اسم المريض وعمره

$pait.name
[1] "Ahmed Omar"

$pait.age
[1] 39

> list1[4] # لعرض جناح المريض

$pait.ward
[1] "card74"

> list1$pait.no #لعرض رقم المريض

[1] 155024
```

يمكن أيضا الحصول على أسماء مكونات القائمة باستخدام دالة names كالتالي:

[1] يمكن استخدام الأقواس المزدوجة لعرض مكونات القائمة بدون أسماء، مثلا؛ list1[[1]]. والرمز $ هو رمز الدولار الموجود في لوحة المفاتيح.

```
> names(list1)
[1] "pait.name" "pait.age"  "pait.no"   "pait.ward"
```

وبالإمكان تغيير هذه الأسماء باستخدام نفس الدالة إذا ما أردنا ذلك:

```
> names(list1)<-c("p.name","p.age","p.no","p.ward")
> names(list1)
[1] "p.name"  "p.age"  "p.no"  "p.ward"
```

ولإضافة مكونات جديدة للقائمة، مثلا إضافة جنس المريض، فإن ذلك يتم بالصورة التالية:

```
> list1$p.gender<-c("female")
> list1

$p.name
[1] "Ahmed Omar"

$p.age
[1] 39

$p.no
[1] 155024

$p.ward
[1] "card74"

$p.gender
[1] "female"
```

وبالطبع يمكن استخدام دالة length للتعرّف على عدد المكونات في القائمة؛

```
> length(list1)
[1] 5
```

ويمكن أيضا حذف مكونات معينة من القائمة، فمثلا يمكن حذف جنس المريض بالصورة:

```
> list1<-list1[-5]
> list1

$p.name
[1] "Ahmed Omar"

$p.age
[1] 39

$p.no
[1] 155024

$p.ward
[1] "card74"
```

وكما هو الحال مع المتجهات والمصفوفات، فإنه يمكن دمج أكثر من قائمة مع بعضها البعض بالطريقة التالية؛ أولا لنقم بإنشاء قائمة جديدة تضم اسم الطبيب وتخصصه[1]:

```
> list2<-c(dr.name="Ali Salem",dr.spec="cardiology")
> list2
```

```
     dr.name      dr.spec
 "Ali Salem" "cardiology"
```

وثانيا نقوم بدمج القائمتين list1 و list2 بالصورة التالية:

```
> list3<-c(list1,list2)
> list3
```

```
$p.name
[1] "Ahmed Omar"

$p.age
[1] 39

$p.no
[1] 155024

$p.ward
[1] "card74"


$dr.name
[1] "Ali Salem"

$dr.spec
[1] "cardiology"
```

إذا ما تم استخدام الدالة unlist مع أي قائمة، فإنها "ستتفكك" أو تتحول إلى متجه يضم كل مكونات تلك القائمة، إلا أن هذه الدالة قد لا تُعد ذات أهمية كبيرة من الناحية العملية لأن المكونات المفككة ستكون غالبا على هيئة قيم غير عددية، (حتى وإن كانت عددية الأصل). لنقم الآن بفك القائمة list1 على سبيل المثال:

```
> unlist(list1)
```

```
      p.name        p.age         p.no       p.ward
"Ahmed Omar"         "39"     "155024"     "card74"
```

[1] لاحظ التغير في طريقة عرض القائمة list2 عن القائمة list1 بسبب احتوائها على مكونات متجانسة، أي من نفس النوع.

ولاحظ أن عمر المريض، وهي قيمة عددية قد تم اعتبارها بعد فك القائمة قيمة غير عددية. كذلك لاحظ أن الدالة unlist لا تغير حالة القائمة list1 المُخزنة في ملف العمل الحالي، ويمكنك التأكد من ذلك عن طريق استدعائها من جديد.

ويمكن استخدام دالة mode هنا للتعرف على الفرق بين نوع القائمة list1 ونوع المتجه الناتج عن فك القائمة، (والذي سيظهر بصورة متجه مُميز لأن القيم السائدة هي قيم غير عددية)، بالصورة:

```
> mode(list1)
[1] "list"

> mode(unlist(list1))
[1] "character"
```

الفصل الثالث

أُطر البيانات واستيراد وتصدير الملفات في R

(Data Frames and Importing and Exporting Data Files in R)

1.3 طُرق إنشاء أُطر البيانات (Methods of Creating Data Frames)

1.1.3 تكوين أُطر البيانات من المتجهات (Constructing Data Frames from Vectors)

2.1.3 تحويل القوائم والمصفوفات إلى أُطر بيانات
(Transforming Lists and Matrices into Data Frames)

3.1.3 استخدام محرر بيانات R (Using R Data Editor)

2.3 التعامل مع مكونات أُطر البيانات (Manipulating Components of Data Frames)

1.2.3 استخدام الدوال الشرطية مع أُطر البيانات
(Using Conditional Functions with Data Frames)

2.2.3 تكوين أُطر البيانات الفرعية (Forming Sub-Data Frames)

3.2.3 بعض العمليات الإضافية على أُطر البيانات
(Some Additional Operations on Data Frames)

3.3 استيراد وتصدير ملفات البيانات (Importing and Exporting Data Files)

1.3.3 استيراد الملفات النصية (Importing Text Files)

2.3.3 استيراد ملفات بيانات اكسل (Importing Excel Data Files)

1.2.3.3 استيراد ملف اكسل كملف نصي (Importing Excel File as a Text File)

2.2.3.3 استيراد ملف اكسل بالامتداد الأصلي
(Importing Excel File with Original Extension)

3.3.3 تصدير ملفات البيانات من R (Exporting Data Files from R)

لتنظيم العمل على موضوعات هذا الفصل، سيتم استخدام ملف عمل جديد، باسم "work3"، في نفس الحافظة "myR" على سطح المكتب لتخزين البيانات والأشياء التي سنتعامل معها. لعمل ذلك سيتم أولا فتح برنامج R من الأيقونة الأصلية له على سطح المكتب، ثم تغيير مسار العمل إلى الحافظة "myR" عن طريق كتابة[1]:

```
> getwd()
[1] "C:/Users/PIXEL-PC/Documents"

> setwd(dir="C:/Users/PIXEL-PC/Desktop/myR")

> getwd()
[1] "C:/Users/PIXEL-PC/Desktop/myR"
```

بعد ذلك سنقوم بحفظ ملف العمل الحالي باسم "work3" بكتابة:

```
> save.image("C:/Users/PIXEL-PC/Desktop/myR/work3.RData")
```

حيث سيكون "work3" فارغا الآن من أي بيانات أو أشياء؛

```
> ls()
character(0)
```

سيتم إنشاء أيقونة جديدة بشعار برنامج R باسم "work3" داخل الحافظة "myR"، وهذه الأيقونة هي التي سوف يتم فتح نظام R منها للعمل على ملف العمل "work3" والخاص بالفصل الثالث من هذا الكتاب. إضافة إلى ذلك، يمكن تخزين سطور الأوامر، باسم "his3" مثلا، كما يلي:

```
> savehistory("C:/Users/PIXEL-PC/Desktop/myR/his3.txt")
```

يمكن اعتبار أُطر البيانات حالة خاصة من القوائم، علما بأنها الأكثر استخداما للتعامل مع البيانات في نظام R. كما أنه يمكن النظر إليها من زاوية إحصائية كمصفوفة بيانات تحتوي على أعمدة وصفوف، ويُشترط أن تكون فيها قيم كل عامود متجانسة مع بعضها البعض، مع إمكانية وجود أعمدة عددية وغير عددية فيها بشرط أن يكون لكل عامود اسم وأن تتساوى هذه الأعمدة في الطول (أي تتساوى في عدد القيم الموجودة بها).

[1] تم تناول كيفية تخزين ملف العمل وسطور الأوامر في السابق إلا أننا نذكّر القارئ دائما في بداية كل فصل بهذه الكيفية لكي يتمكن من مواكبة ترتيب شرح موضوعات الفصل الحالي.

1.3 طُرق إنشاء أُطر البيانات (Methods of Creating Data Frames)

توجد عدة طرق لإنشاء أو تكوين إطار بيانات في نظام R، وسنقوم فيما يلي باستعراض أهم ثلاثة من هذه الطرق وأكثرها استخداما.

1.1.3 تكوين أُطر البيانات من المتجهات (Constructing Data Frames from Vectors)

تعتمد هذه الطريقة على استخدام الدالة `data.frame` مباشرة إما مع متجهات تم تعيينها مسبقا وهي الحالة الأولى، (بشرط أن تستوفي الشروط المذكورة أعلاه)، أو عن طريق إدراج متجهات جديدة ضمن الدالة، وهي الحالة الثانية.

بالنسبة **للحالة الأولى**؛ وهي إنشاء إطار البيانات عن طريق ضم متجهات مُعرّفة مسبقا، لنفرض أنه تم تعريف ثلاثة متجهات جديدة هي `s.level`، `s.age`، و`s.openion` والتي تمثل المستوى الدراسي لعشرة طلبة في إحدى الكليات، وأعمارهم بالسنوات، وآراؤهم في طرق التدريس في الكلية، على الترتيب[1]؛

```
> s.level<-c(45,61,25,85,77,74,90,50,64,71)

> s.age<-c(20,21,20,24,23,20,19,26,23,22)

> s.openion<-c("average","good","bad","average","bad",
"bad","average","bad","good","bad")
```

عندئذ يمكن إنشاء إطار بيانات من هذه المتجهات، وليكن باسم `data.f1` مثلا، كالتالي:

```
> data.f1<-data.frame(s.level,s.age,s.openion)
> data.f1
   s.level s.age s.openion
1       45    20   average
2       61    21      good
3       25    20       bad
4       85    24   average
5       77    23       bad
6       74    20       bad
7       90    19   average
8       50    26       bad
9       64    23      good
10      71    22       bad
```

[1] على اعتبار أن المستوى الدراسي مُقاس على الدرجات من 0 إلى 100، والآراء هي جيدة، متوسطة، أو سيئة (good, average, bad).

ونستطيع أن نرى أن الملف data.f1 يبدو كمصفوفة بيانات بها ثلاثة أعمدة، (العامودان الأولان عدديان والثالث غير عددي)، وعشرة صفوف تمثل عدد المشاهدات.

ونلاحظ أيضا أنه إذا ما تم استخدام دالة mode للتحقق من طبيعة الملف أو الشيء data.f1 فإنه سيظهر كقائمة، أما إذا تم استخدام الدالة class فإنه سيظهر كإطار بيانات كما نرى؛

```
> mode(data.f1)
[1] "list"

> class(data.f1)
[1] "data.frame"
```

أما بالنسبة **للحالة الثانية**، وهي إنشاء إطار البيانات عن طريق تكوين متجهات جديدة بداخله، فإنها ستعطي نفس النتيجة تماما؛

```
> data.f1<-data.frame(s.level=c(45,61,25,85,77,74,90,50,
64,71),s.age=c(20,21,20,24,23,20,19,26,23,22),s.openion=
c("average","good","bad","average","bad","bad","average",
"bad","good","bad"))
```

ويمكن للقارئ التحقق من ذلك باستدعاء إطار البيانات data.f1 من جديد.

2.1.3 تحويل القوائم والمصفوفات إلى أُطر بيانات

(Transforming Lists and Matrices into Data Frames)

الطريقة الثانية لإنشاء أُطر البيانات هي عن طريق تحويل القوائم أو المصفوفات أو حتى المتجهات إلى أُطر بيانات باستخدام **دالة التحويل**[1] **إلى أطر البيانات** as.data.frame كما سنرى في الحالات التالية؛

- تحويل القائمة إلى إطار بيانات:

لنفرض أنه تم إنشاء القائمة list4 بالصورة:

```
> list4<-list(x1=c(2,5,6,7),x2=c("a","b","c","d"),
x3=c(10,20,30,40),x4=c(-2,0,7,3))

> list4

$x1
[1] 2 5 6 7
```

[1] سنتعرّض في هذا البند وفي البنود القادمة أيضا لاستخدام رمز التحويل as كجزء من دوال متنوعة.

```
$x2
[1] "a" "b" "c" "d"

$x3
[1] 10 20 30 40

$x4
[1] -2  0  7  3
```

عندها يمكن تحويل تلك القائمة إلى إطار بيانات، باسم data.f2 مثلا، كالتالي:

```
> data.f2<-as.data.frame(list4)
> data.f2

  x1 x2 x3 x4
1  2  a 10 -2
2  5  b 20  0
3  6  c 30  7
4  7  d 40  3
```

ولاحظ أن

```
> class(list4)
[1] "list"

> class(data.f2)
[1] "data.frame"
```

- تحويل المصفوفة والمتجه إلى إطار بيانات:

لنقم أولا بتعريف المتجهات x1، x2، x3، و x4 وتكوين المصفوفة mat10 عن طريق دمج هذه المتجهات بالصورة التالية؛

```
> x1<-c(7,1,-4,9)

> x2<-c(3,11,0,8)

> x3<-c(6,2,15,7)

> x4<-c(21,3,17,30)

> mat10<-cbind(x1,x2,x3,x4);rownames(mat10)<-
c("case1","case2","case3","case4")

> mat10

      x1 x2 x3 x4
case1  7  3  6 21
case2  1 11  2  3
case3 -4  0 15 17
case4  9  8  7 30
```

ثم نقوم بتحويل المصفوفة mat10 إلى إطار بيانات، باسم data.f3 مثلا؛

```
> data.f3<-as.data.frame(mat10)
> data.f3
```

```
      x1 x2 x3 x4
case1  7  3  6 21
case2  1 11  2  3
case3 -4  0 15 17
case4  9  8  7 30
```

ولاحظ أن المصفوفة mat10 وإطار البيانات data.f3 متطابقان عند عرضهما في المظهر العام، إلا أن طبيعتهما مختلفة كما نرى:

```
> class(mat10)
[1] "matrix"

> class(data.f3)
[1] "data.frame"
```

- تحويل المتجهات إلى إطار بيانات:

نسوق هنا مثالين آخرين يتم فيهما تحويل متجه واحد ومتجهين إلى أُطر بيانات لاستكمال طريقة التحويل إلى أُطر البيانات، (وننوه هنا إلى أن المتجهات التي يتم تحويلها إلى أُطر بيانات لابد أن تكون لها الطبيعة integer أو character)؛

```
> class(x1);class(x2)
```

```
[1] "numeric"
[1] "numeric"
```

ولتحويل المتجهين x1 و x2 إلى متجهات لها طبيعة integer، (أو character في حالة القيم غير العددية)، بدلا من numeric نستخدم دالة التحويل as، كما استخدمناها سابقا، بالصورة التالية[1]:

```
> x1<-as.integer(x1);x2<-as.integer(x2)

> class(x1);class(x2)
[1] "integer"
[1] "integer"
```

الآن يمكن تنفيذ عملية تحويل المتجهات إلى أُطر بيانات كالتالي:

```
> data.f4<-data.frame(x1)
```

[1] تم كتابة الدوال مع المتجهين x1 و x2 في سطر أوامر واحد للاختصار فقط.

```
> data.f4

  x1
1  7
2  1
3 -4
4  9

> class(data.f4)
[1] "data.frame"

> data.f5<-data.frame(x1,x2)
> data.f5

  x1 x2
1  7  3
2  1 11
3 -4  0
4  9  8
```

ويمكن، (إضافة للتأكد من طبيعة أطر البيانات باستخدام الدالة class)، التعرّف على عدد الأعمدة والصفوف فيها، باستخدام الدوال ncol و nrow، أو الحصول على ملخص لطبيعتها وما تحتويه من بيانات باستخدام الدالة str كما نرى مما يلي:

```
> ncol(data.f3);nrow(data.f3)

[1] 4
[1] 4

> str(data.f3)

'data.frame':   4 obs. of  4 variables:
 $ x1: num  7 1 -4 9
 $ x2: num  3 11 0 8
 $ x3: num  6 2 15 7
 $ x4: num  21 3 17 30
```

3.1.3 استخدام محرر بيانات R (Using R Data Editor)

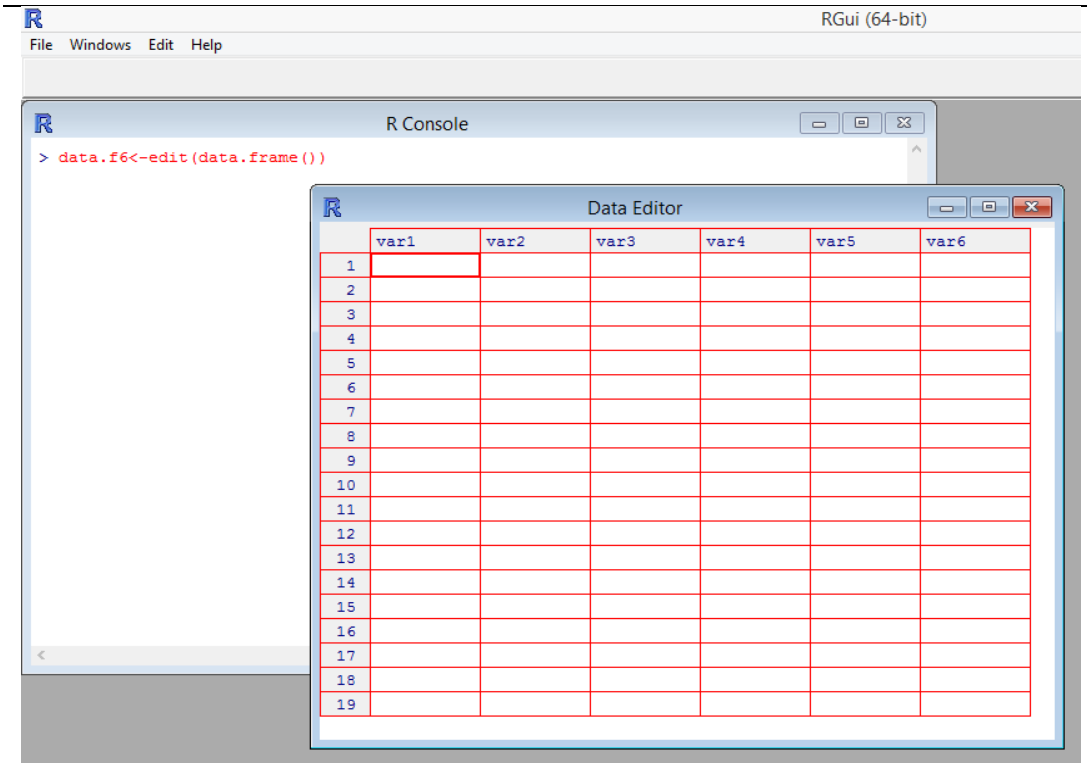
تُعد هذه الطريقة الأكثر مرونة واستخداما في إنشاء أُطر البيانات عندما يكون حجم البيانات كبيرا، وهي تُستخدم عموما في حال عدم وجود البيانات مسبقا في قوائم أو مصفوفات أو متجهات. هذه الطريقة تعتمد على استخدام **محرر بيانات R** (R Data Editor)، وهو عبارة عن نافذة منبثقة تظهر عند استدعائها وتشبه

نوافذ البرامج الإحصائية التقليدية التي تحتوي على جداول بها صفوف وأعمدة مقسّمة ومتسلسلة ويتم إدخال البيانات من خلالها.

لاستخدام محرر البيانات لإدخال بيانات جديدة، وحفظها كإطار بيانات (اسمه `data.f6` مثلا)، نستخدم دالة **التعديل** `edit` بالصورة التالية:

```
> data.f6<-edit(data.frame())
```

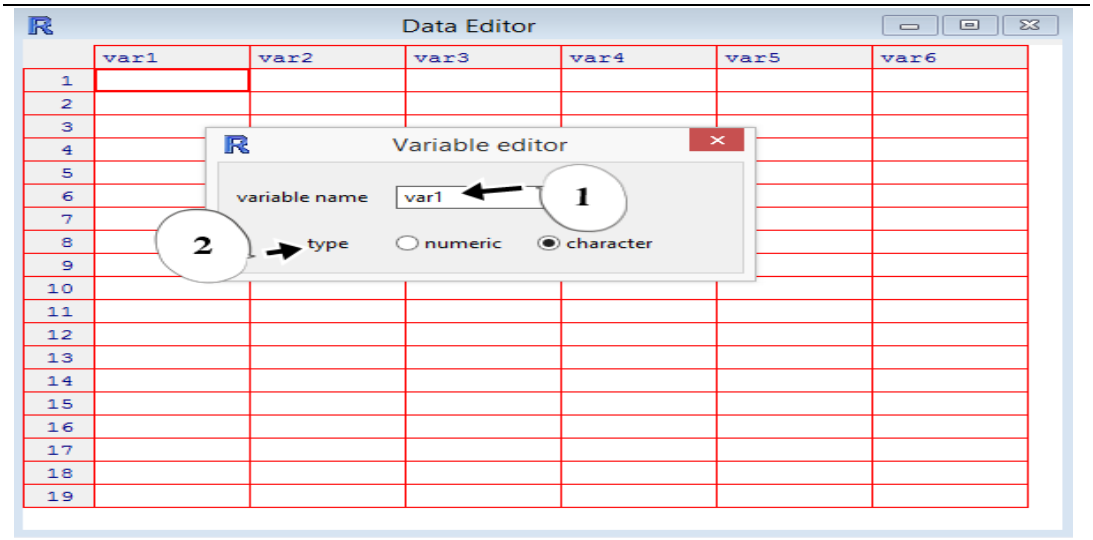
فتظهر نافذة محرر بيانات R كما يظهر في الشكل (1.3). ونرى في تلك النافذة جدول فارغ يضم مجموعة من الأعمدة لها الأسماء؛ (var1, var2, var3, …) وهي اختصار مصطلح **متغير**[1] (Variable)، وصفوف بأرقام متسلسلة.



شكل 1.3: نافذة محرر بيانات R

ويتم إدخال البيانات في خلايا الجدول مباشرة. وإذا ما أردنا تغيير أسماء الأعمدة للبيانات المُراد إدخالها فيمكن عمل ذلك عن طريق تغيير اسم كل عامود (ونوعه أيضا) على حِده بالنقر بالفأرة على الخلية الموجود بها اسم العامود، ولنقم بذلك مثلا مستخدمين العامود الأول إلى اليسار، حيث ستظهر نافذة صغيرة، (كما في الشكل (2.3))؛

[1] سنتطرق لمفهوم المتغيرات وأنواعها في علم الإحصاء في الفصل القادم ونكتفي هنا باستخدام تسمية عامود.



شكل 2.3: نافذة تغيير اسم العامود ونوع البيانات

نقوم بإلغاء الاسم الافتراضي للمتغير (var1) المشار إليه بالسهم (1) وإدخال الاسم الذي نرغب فيه، ثم نحدد نوع العامود أو المتغير المطلوب إدخاله وذلك بالنقر على إحدى الخيارين؛ متغير عددي (numeric) أو متغير مُميَز (character) في الموضع (type) المشار إليه بالسهم (2)، بعدها نقوم ببساطة بإغلاق هذه النافذة الصغيرة. ويفضل عادة القيام بهذه الخطوة قبل إدخال البيانات في الأعمدة، علما بأنه يمكن القيام بها أو العودة إليها في أي وقت لاحق.

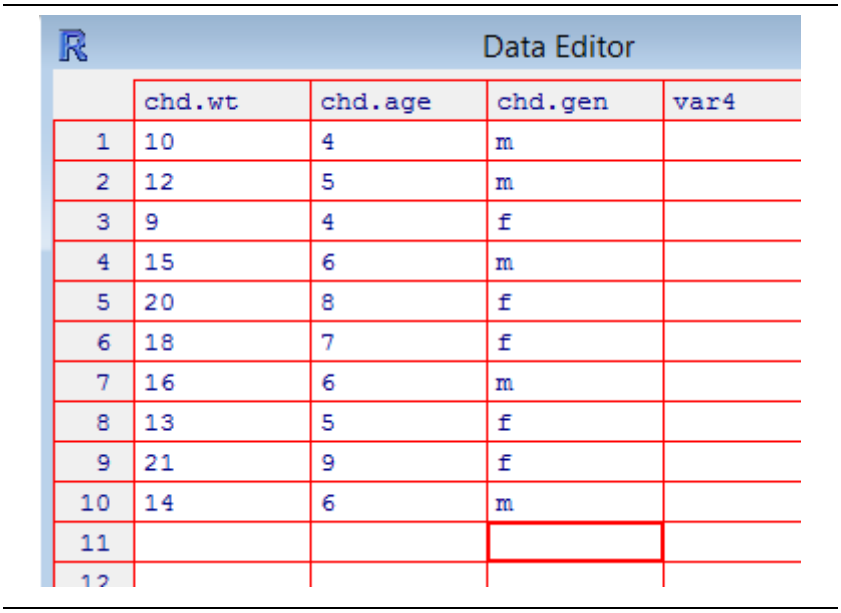
لنقم على سبيل المثال بإدخال البيانات الموجودة في الجدول التالي (جدول (1.3))، والتي تمثل أوزان عشرة مرضى من الأطفال في جناح الباطنية في مستشفى حكومي، وكذلك أعمارهم وجنسهم.

جدول 1.3: بيانات خاصة بمجموعة أطفال في جناح الباطنية

المشاهدات	1	2	3	4	5	6	7	8	9	10
وزن الطفل (كجم)	10	12	9	15	20	18	16	13	21	14
عمر الطفل (سنة)	4	5	4	6	8	7	6	5	9	6
جنس الطفل	ذكر	ذكر	أنثى	ذكر	أنثى	أنثى	ذكر	أنثى	أنثى	ذكر

يمكننا أولا، كإجراء تنظيمي، إبدال الأسماء الافتراضية للأعمدة الأولى الثلاثة من اليسار بالاختصارات chd.wt، chd.age، و chd.gen والتي ترمز لوزن، عمر، ونوع الطفل على الترتيب. والشكل (3.3) يبين شكل نافذة محرر البيانات بعد تغيير أسماء الأعمدة وإدخال البيانات، ولاحظ أن نوع الطفل تم إدخاله بالرموز (m) للذكر و(f) للأنثى، وأن هذا العامود يمثل قيم مُميزة لذلك لا ننس اختيار نوع المتغير المُميز (character) بعد

تغيير الاسم، أما وزن وعمر الطفل فهما عدديان وبالتالي نختار نوع المتغير العددي (numeric) بعد تغيير اسميهما.



	chd.wt	chd.age	chd.gen	var4
1	10	4	m	
2	12	5	m	
3	9	4	f	
4	15	6	m	
5	20	8	f	
6	18	7	f	
7	16	6	m	
8	13	5	f	
9	21	9	f	
10	14	6	m	
11				

شكل 3.3: نافذة محرر البيانات بعد إدخال بيانات الجدول (1.3)

بعد الانتهاء من عملية الإدخال، نقوم بإغلاق نافذة محرر البيانات ويتم بذلك حفظ إطار البيانات تلقائيا بالاسم الذي تمت كتابته سابقا وهو data.f6، مع ملاحظة أن اسم إطار البيانات لا يظهر على نافذة محرر بيانات R، لذلك يجب التأكد دائما من أننا نقوم بإدخال البيانات أو تحريرها لإطار البيانات المطلوب، وخاصة عند وجود تشابه في أسماء الأعمدة.

ويمكنك رؤية أن بياناتك قد تم تخزينها في إطار البيانات المحدد عن طريق استدعاؤه كالعادة بكتابة اسمه؛

```
> data.f6

   chd.wt chd.age chd.gen
1      10       4       m
2      12       5       m
3       9       4       f
4      15       6       m
5      20       8       f
6      18       7       f
7      16       6       m
8      13       5       f
9      21       9       f
10     14       6       m
```

ونوضح هنا أنه عند حدوث أي خطأ من قِبل المستخدم في عملية إدخال البيانات أو عندما يتم تغيير أسماء الأعمدة أو عند الحاجة لإجراء تعديل ما على البيانات، (ويشمل ذلك تغيير قيم أو إضافة صفوف أو أعمدة)،

في أي وقت لاحق، فإنه يمكن عمل ذلك باستخدام الدالة fix لتنفيذ ذلك. ولتوضيح الفكرة، لنفرض أننا نريد تغيير القيمة العاشرة في العامود الأول chd.wt وهي (14) بالقيمة (15) في إطار البيانات data.f6، فنقوم بكتابة:

```
> fix(data.f6)
```

وبعد ضغط زر الإدخال سيظهر محرر البيانات لإطار البيانات data.f6 من جديد، فنذهب للخلية المطلوبة ونضغط عليها ثم نكتب القيمة الجديدة (15) ونغلق النافذة، وبذلك يتم التعديل المطلوب كما نرى؛

```
> data.f6
```

	chd.wt	chd.age	chd.gen
1	10	4	m
2	12	5	m
3	9	4	f
4	15	6	m
5	20	8	f
6	18	7	f
7	16	6	m
8	13	5	f
9	21	9	f
10	15	6	m

ملاحظات هامة:

- لتعديل إطار البيانات السابق data.f6 يمكن أيضا كتابة:

```
> edit(data.frame(data.f6))
```

- انتبه! إلى أنه إذا ما تم كتابة الأمر data.f6<-edit(data.frame()) من جديد، فإن كل البيانات الموجودة في data.f6 سيتم إلغاؤها.
- يمكن بالطبع إنشاء إطار بيانات جديد بتعيين اسم جديد ضمن دالة edit واستخدامها كما وضحنا سابقا.
- إذا ما تم استخدام الأمر edit(data.frame()) بدون تعيين اسم، فإنه سيفتح إطار بيانات جديد، وإذا تم إدخال بيانات فيه وإغلاقه فإن تلك البيانات ستظهر مباشرة بعد الإغلاق، وبالطبع لا يمكن التعامل لاحقا مع تلك البيانات، لذلك لا نلجأ عادة لاستخدام هذا الأمر بهذه الصورة.

2.3 التعامل مع مكونات أُطر البيانات (Manipulating Components of Data Frames)

كما هو الحال مع نظام القوائم، يمكن استدعاء أي من مكونات أو أعمدة أُطر البيانات والتعامل معها كمتجهات مستقلة أو مصفوفات. ولعمل ذلك لابد من كتابة تلك المكونات من ثلاثة مقاطع؛ المقطع الأول هو اسم إطار البيانات، والمقطع الثاني هو رمز الدولار، وأما المقطع الثالث فهو اسم المكون نفسه داخل إطار البيانات. ولتوضيح الصورة نأخذ المثال التالي؛

في إطار البيانات المُستخدَم أخيرا وهو data.f6 لنفرض أننا نود التعامل مع العامود الذي يمثل عمر الطفل chd.age فقط، عندها يمكن استدعاؤه كالتالي:

```
> data.f6$chd.age
 [1] 4 5 4 6 8 7 6 5 9 6
```

ولاحظ تركيبة المقاطع الثلاثة؛ chd.age و $ و data.f6 والتي تُعد مجتمعة اسم المتجه الذي يمثل العامود الثاني في إطار البيانات data.f6. ومن الناحية العملية، يُلاحَظ أن اسم هذا المتجه يُعد طويلا بعض الشيء، لذلك يمكن دائما تعيين اسم مختصر لهذه المتجهات عوضا عن استدعائها بهذه الطريقة، ولنستخدم الاسم y2 مثلا في مثالنا؛

```
> y2<-data.f6$chd.age
> y2
 [1] 4 5 4 6 8 7 6 5 9 6

> class(y2)
[1] "numeric"
```

ولاحظ أن المتجه y2 هو من النوع العددي. كما أنه من الممكن استدعاء أية قيمة أو قيم مع إطار البيانات كما رأينا ذلك مع المصفوفات، فمثلا يمكن استدعاء القيمة الرابعة في العامود الثالث لإطار البيانات data.f6 بالصورة:

```
> data.f6[4,3]
[1] "m"
```

أو استدعاء صف أو عامود كامل كما نرى:

```
> data.f6[6,]
  chd.wt chd.age chd.gen
6     18       7       f

> data.f6[,2]
 [1] 4 5 4 6 8 7 6 5 9 6
```

1.2.3 استخدام الدوال الشرطية مع أُطر البيانات

(Using Conditional Functions with Data Frames)

يمكن استخدام الدوال والأوامر الشرطية لاستدعاء أو تعيين القيم التي تخضع لشرط معين، (كما هو الحال مع المتجهات والمصفوفات)، فمثلا إذا أردنا معرفة الأطفال الذين تزيد أعمارهم عن ستة سنوات في إطار البيانات data.f6 يمكننا كتابة:

```
> data.f6[data.f6$chd.age>6,]

  chd.wt chd.age chd.gen
5     20       8       f
6     18       7       f
9     21       9       f

> class(data.f6[data.f6$chd.age>6,])
[1] "data.frame"
```

ولاحظ أن النتيجة الأخيرة هي بحد ذاتها إطار بيانات فرعي من إطار البيانات الأصلي، ويمكن أيضا تعيين اسم جديد له والتعامل معه باستقلالية. وكمثال إضافي، لنقم بتعريف إطار بيانات آخر (باسم data.f6.f مثلا) بحيث يحتوي على الأطفال الإناث فقط في البيانات الأصلية:

```
> data.f6.f<-data.f6[data.f6$chd.gen=="f",]

> data.f6.f

  chd.wt chd.age chd.gen
3      9       4       f
5     20       8       f
6     18       7       f
8     13       5       f
9     21       9       f

> class(data.f6.f)
[1] "data.frame"
```

ولاحظ أنه إذا ما تم كتابة:

```
> data.f6$chd.gen=="f"
[1]  FALSE  FALSE  TRUE  FALSE  TRUE  TRUE  FALSE  TRUE  TRUE
FALSE
```

فإن النتيجة تكون متجه منطقي يعطي صحيح (TRUE) للإناث و خاطئ (FALSE) للذكور.

2.2.3 تكوين أُطر البيانات الفرعية (Forming Sub-Data Frames)

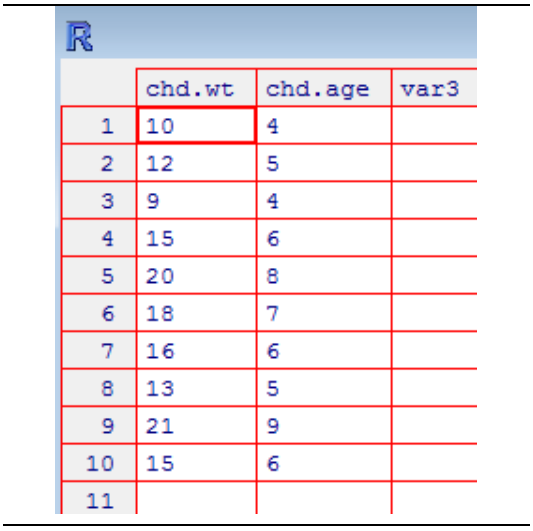
يمكن تكوين أو اختيار أي إطار بيانات فرعي من إطار البيانات الأصلي بصورة أكثر مرونة مما تعاملنا معه سابقا من خلال استخدام الدوال الشرطية وذلك باستخدام الدالة subset التي تمكننا من تكوين إطار بيانات جديد يضم فقط الأعمدة أو المتغيرات التي نرغب في التعامل معها. وكمثال على هذه الدالة، لنفرض أننا نرغب في تكوين إطار بيانات فرعي يضم فقط وزن وعمر الأطفال في مجموعة البيانات data.f6، وليكن اسمه data.f7، عندها نكتب:

```
> data.f7<-subset(data.f6,select=c(chd.wt,chd.age))
> data.f7
```

```
   chd.wt chd.age
1      10       4
2      12       5
3       9       4
4      15       6
5      20       8
6      18       7
7      16       6
8      13       5
9      21       9
10     15       6
```

ولاحظ استخدام الخيار select لتحديد الأعمدة المطلوبة. وبالطبع يمكن التعامل مع إطار البيانات الجديد data.f7، (بالتعديل أو الإضافة أو الحذف)، باستخدام الدالة fix أو edit لفتح محرر البيانات الخاص به كما يظهر في الشكل (4.3).

```
> fix(data.f7)
```



	chd.wt	chd.age	var3
1	10	4	
2	12	5	
3	9	4	
4	15	6	
5	20	8	
6	18	7	
7	16	6	
8	13	5	
9	21	9	
10	15	6	
11			

شكل 4.3: نافذة محرر البيانات لـ data.f7

كما يمكن استخدام الدالة subset مع أمر شرطي لتكوين إطار بيانات جديد يضم تلك القيم التي تحقق ذلك الشرط، فمثلا يمكن تنفيذ نفس الأمر السابق، والخاص بعرض بيانات الأطفال الذين تكون أعمارهم أكبر من 6 سنوات؛ data.f6[data.f6$chd.age>6] بصورة أخرى كالتالي:

```
> subset(data.f6,chd.age>6)

  chd.wt chd.age chd.gen

5     20       8       f
6     18       7       f
9     21       9       f
```

ويمكن أيضا استخدام أكثر من أمر شرطي في نفس الوقت، فمثلا إذا كان المطلوب هو عرض بيانات الأطفال الإناث اللواتي تقل أوزانهن عن 21 كجم في data.f6، فيمكن عندها استخدام الرمز "&"، لعمل ذلك:

```
> subset(data.f6,chd.wt<21&chd.gen=="f")

  chd.wt chd.age chd.gen
3      9       4       f
5     20       8       f
6     18       7       f
8     13       5       f
```

وللحصول على بيانات الأطفال الذين تزيد أوزانهم عن 20 كجم "أو" تكون أعمارهم أكبر من أو تساوي 7 سنوات مثلا، يمكن استخدام الرمز "|" بالصورة:

```
> subset(data.f6,chd.wt>20|chd.age>=7)

  chd.wt chd.age chd.gen

5     20       8       f
6     18       7       f
9     21       9       f
```

من جديد، يمكننا استدعاء أو تعيين أي إطار فرعي يضم أي مجموعة من الصفوف الأولى أو الأخيرة[1] باستخدام متسلسلة عددية بالصورة:

```
> data.f6[1:5,] # أول 5 صفوف

  chd.wt chd.age chd.gen
1     10       4       m
2     12       5       m
3      9       4       f
4     15       6       m
5     20       8       f
```

[1] يمكن استخدام الدوال head(data.f6) و tail(data.f6) أيضا للحصول على أول ستة صفوف وآخر ستة صفوف، على الترتيب، في إطار البيانات data.f6.

```
> data.f6[7:10,] # آخر 4 صفوف
```

```
   chd.wt chd.age chd.gen
7      16       6       m
8      13       5       f
9      21       9       f
10     15       6       m
```

وعند وجود عامود في إطار البيانات يأخذ قيما مُميزة، (كما هو الحال مع نوع الطفل في البيانات data.f6)، فإنه يمكن استخدام الدالة split لفصل مكونات الأعمدة العددية بناء على القيم المُميزة، وإظهار ذلك على صورة قائمة بها عدد من المتجهات بحسب عدد التقسيمات لتلك القيم المميزة، فمثلا يمكن عرض وزن الطفل بحسب النوع، وكذلك عرض عمر الطفل بحسب النوع في البيانات data.f6 كالتالي:

```
> split(data.f6$chd.wt,data.f6$chd.gen)
```

```
$f
[1]  9 20 18 13 21
$m
[1] 10 12 15 16 15
```

```
> split(data.f6$chd.age,data.f6$chd.gen)
```

```
$f
[1] 4 8 7 5 9
$m
[1] 4 5 6 6 6
```

3.2.3 بعض العمليات الإضافية على أُطر البيانات

(Some Additional Operations on Data Frames)

- **تحويل البيانات:**

نأتي الآن لنوع آخر من التعامل مع مكونات أُطر البيانات؛ فقد يلزمنا أحيانا تعريف عامود (متغير) جديد أو أكثر بحيث يمثل تحويل (أي عملية حسابية) من عامود آخر (أو أكثر) موجود ضمن البيانات، فعلى سبيل المثال لنفرض في إطار البيانات data.f6 أننا نرغب بإدراج وزن الأطفال بالرطل (Pound)، فهذا سيستدعي تحويل الوزن من وحدة كيلوجرام (في العامود chd.wt) إلى وحدة رطل وتعريف عامود جديد، وليكن اسمه chd.wt.p، بالصورة التالية[1]:

```
> data.f6$chd.wt.p<-data.f6$chd.wt*2.21
> data.f6
```

[1] نذكّر هنا أن كل 1 كيلوجرام يساوي 2.21 رطل تقريبا.

```
   chd.wt chd.age chd.gen chd.wt.p
1      10       4       m    22.10
2      12       5       m    26.52
3       9       4       f    19.89
4      15       6       m    33.15
5      20       8       f    44.20
6      18       7       f    39.78
7      16       6       m    35.36
8      13       5       f    28.73
9      21       9       f    46.41
10     15       6       m    33.15
```

- **التدوير:**

وإذا ما أردنا تدوير (تقريب) قيم المتغير الجديد إلى أقرب عدد صحيح يمكننا استخدام دالة **التدوير** round لعمل ذلك؛

```
> data.f6$chd.wt.p<-round(data.f6$chd.wt*2.21)
> data.f6
```

```
   chd.wt chd.age chd.gen chd.wt.p
1      10       4       m       22
2      12       5       m       27
3       9       4       f       20
4      15       6       m       33
5      20       8       f       44
6      18       7       f       40
7      16       6       m       35
8      13       5       f       29
9      21       9       f       46
10     15       6       m       33
```

- **حذف البيانات:**

أما لحذف أي عامود من إطار البيانات، (على افتراض أننا سنحذف العامود الرابع chd.wt.p)، فيمكن تنفيذ ذلك بعدة طرق، أبسطها هي التالية:

```
> data.f6<-data.f6[-4]
> data.f6
```

```
   chd.wt chd.age chd.gen
1      10       4       m
2      12       5       m
3       9       4       f
4      15       6       m
5      20       8       f
```

```
6       18        7        f
7       16        6        m
8       13        5        f
9       21        9        f
10      15        6        m
```

أو عن طريق كتابة الأمر التالي:

```
data.f6<-subset(data.f6,select=c(chd.wt,chd.age,chd.gen)).
```

- **ترتيب البيانات:**

يمكن أيضا استخدام دوال order وsort.list، والتي استُخدمت فيما سبق لغرض ترتيب المتجهات تصاعديا أو تنازليا، مع أُطر البيانات لنفس الغرض. فبافتراض أننا نريد ترتيب البيانات (الأعمدة) في data.f6 تصاعديا بناء على عمر الطفل، وتعيين البيانات المرتبة في إطار بيانات جديد هو data.f8، عندها نكتب:

```
> data.f8<-data.f6[order(data.f6[,2]),1:3]
> data.f8
   chd.wt chd.age chd.gen
1      10       4       m
3       9       4       f
2      12       5       m
8      13       5       f
4      15       6       m
7      16       6       m
10     15       6       m
6      18       7       f
5      20       8       f
9      21       9       f
```

حيث أن [،2] تعني أن المطلوب هو ترتيب الأعمدة بناء على العامود الثاني وهو عمر الطفل، والسلسلة 1:3 تعني أن الأعمدة المشمولة في الترتيب هي من العامود الأول إلى الثالث، أي كل البيانات في مثالنا.

ولتنفيذ المطلوب السابق بترتيب تنازلي يمكن كتابة[1]:

```
> data.f9<-data.f6[rev(order(data.f6[,2])),1:3]
> data.f9
```

[1] يمكن أيضا استخدام الخيار decreasing بالصورة:

```
data.f9<-data.f6[order(data.f6[,2],decreasing=T),1:3].
```

```
   chd.wt chd.age chd.gen
9      21       9       f
5      20       8       f
6      18       7       f
10     15       6       m
7      16       6       m
4      15       6       m
8      13       5       f
2      12       5       m
3       9       4       f
1      10       4       m
```

3.3 استيراد وتصدير ملفات البيانات (Importing and Exporting Data Files)

في كثير من الحالات، قد تكون البيانات التي يرغب الباحث أو المستخدم بالتعامل معها وتحليلها موجودة مسبقا في ملف أو قاعدة بيانات تم إنشاؤها باستخدام البرامج المتخصصة الأخرى التي قد تعتمد غالبا على أنظمة ملفات مختلفة عن نظام ملفات R، وبالتالي لا يمكن "فتح" أو استخدام مثل هذه الملفات مباشرة في نظام R. ولهذا توجد عدة طرق لاستيراد (Import) ملفات البيانات هذه إلى R اعتمادا على الصيغة التي أُنشأت بها. وبالمثل، يمكن تصدير (Export) أو إعادة نقل ملفات البيانات من نظام R إلى البرامج الأخرى باستخدام الدوال المناسبة.

وسيتم في هذا البند التنويه إلى بعض تلك الطرق لعرض أو استيراد ملفات البيانات المتوفرة بصيغ مختلفة، علما بأن الشرح سيتركز على الملفات النصية (Text Files) وملفات مايكروسوفت أُوفيس اكسل (MS Office Excel) باعتبارها الصيغ الأكثر استخداما والأكثر مرونة في التعامل ليس مع نظام R فقط بل مع معظم البرامج الإحصائية الأخرى.

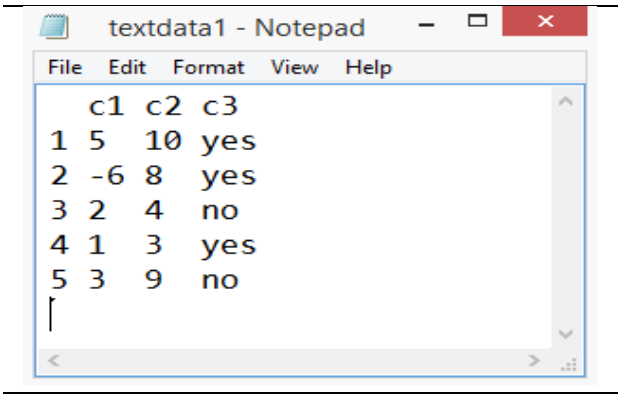
1.3.3 استيراد الملفات النصية (Importing Text Files)

الملفات النصية قد تأخذ عدة أشكال، (في نظام ويندوز)، إلا أن أشهرها هما **دفتر الملاحظات** (Notepad) و**دفتر الكلمات** (WordPad)، واللذان يُستخدَمان عموما في كتابة الملاحظات في ملفات لها الامتدادات[1] (TXT) و(RTF) على التوالي.

لنفترض أنه لدينا ملف نصي يحتوي على بيانات تمثل ثلاثة أعمدة وخمسة صفوف، (كما يوضح الشكل (5.3))، له الاسم (textdata1) وموقعه (مساره) هو سطح المكتب، عندها يمكن أن يتم استيراد هذا الملف من داخل نظام R بالطريقة التالية؛

[1] هنالك امتدادات أخرى للملفات النصية مثل (CSV) و(DIF) على سبيل المثال، ويمكن للقارئ استخدام دالة المساعدة للتعرف على طرق التعامل معها متى أراد ذلك.

أولا: يتم نقل أو نسخ الملف (textdata1) إلى الحافظة (myR) التي يوجد بها ملف العمل (work3) لأن مسار العمل الحالي مُعرّف على تلك الحافظة. (علما بأنه يمكن استيراد الملف المطلوب من أي مسار آخر في الحاسوب، إلا أن ذلك سيتطلب تعريف هذا المسار ضمن أمر الاستيراد).



شكل 5.3: الملف النصي (textdata1)

ثانيا: يتم استخدام الدالة `read.table` وتعيين اسم للبيانات التي سيتم استيرادها، وليكن `data.imp1` مثلا، كما يلي:

```
> data.imp1<-read.table("textdata1.txt")
> data.imp1

  c1 c2  c3
1  5 10 yes
2 -6  8 yes
3  2  4  no
4  1  3 yes
5  3  9  no

> class(data.imp1)
[1] "data.frame"
```

وبذلك يتم استيراد الملف النصي إلى إطار بيانات جديد هو `data.imp1`. (ولاحظ أيضا أن البيانات في الملف الأصلي `textdata1.txt` قد تم إدخالها صفا تلو الآخر باستخدام المسافة أو زر (TAB) في لوحة المفاتيح للفصل بين القيم ضمن الصف الواحد).

إلا أن هذا الأسلوب لا يُعد شائع الاستخدام ضمن الأساليب المتبعة لإدخال وتخزين البيانات لأن إدخال البيانات في جداول ذات صفوف وأعمدة يُعتبر الأفضل عمليا، لذلك سننتقل الآن للتعامل مع تلك النوعية من جداول البيانات.

2.3.3 استيراد ملفات بيانات اكسل (Importing Excel Data Files)

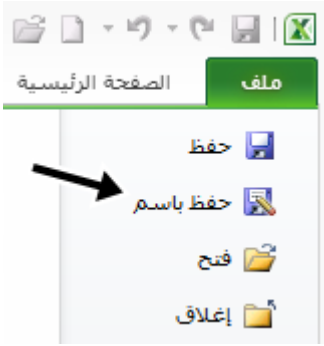
توجد في الواقع عدة أساليب لاستيراد ملفات اكسل وفتحها في نظام R، وتلك الأساليب تعتمد على كل من نسخة أو إصدار برنامج R وإصدار برنامج اكسل ونسخة الويندوز على الجهاز، وغيرها من الأمور التقنية التي قد تكون معقدة بعض الشيء لغير المتمرسين في التعامل مع الحاسوب، لذلك سيتم عرض أسلوبين يُعتبران مرنين إلى حد كبير مع معظم إصدارات كل من ويندوز واكسل؛

1.2.3.3 استيراد ملف اكسل كملف نصي (Importing Excel File as a Text File)

يعتمد هذا الأسلوب على حفظ ملف البيانات الموجود في برنامج اكسل بصيغة ملف نصي، (له الامتداد txt.)، ثم استخدام نفس الدالة `read.table` لاستيراده في R، كما توضح الخطوتان التاليتان، حيث تتناول الخطوة (1) ما سيتم في برنامج اكسل، وتتناول الخطوة (2) باقي الإجراءات في نظام R:

(1) في برنامج اكسل:

بفرض وجود ملف بيانات[1] باسم "excdata1" مُخزن في برنامج اكسل بالامتداد الافتراضي لملفات اكسل[2] (xls. أو xlsx.) في أي مسار في الحاسوب، عندها يتم نسخه إلى الحافظة "myR" على سطح المكتب والتي يوجد بها ملف العمل الحالي "work3". وبعد فتح الملف، (من موقعه الجديد داخل الحافظة "myR")، يتم اختيار "حفظ باسم"، كما يوضح الشكل (6.3 أ)؛



شكل 6.3 أ: موضع الخيار "حفظ باسم" في برنامج اكسل

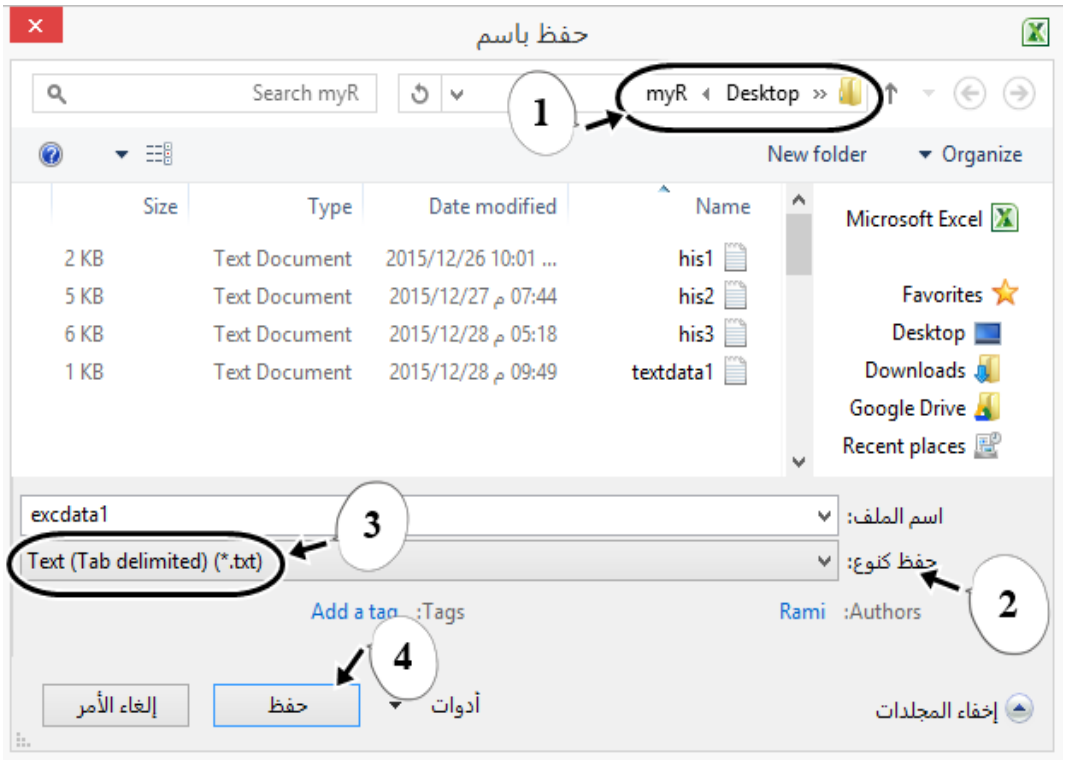
ستظهر بعد ذلك النافذة الفرعية، (الشكل (6.3 ب))، والتي يتم من خلالها تحديد كل من اسم الملف، الصيغة المطلوبة لحفظ الملف، ومسار حفظ الملف.

لنترك اسم الملف الأصلي كما هو، ونتأكد أولا أن مسار الحفظ هو في الحافظة "myR" على سطح المكتب كما يُوضح السهم الأول. ثم نقوم بعد ذلك بالضغط على سهم "حفظ كنوع" المشار إليه بالسهم الثاني، ونختار بعد ذلك الخيار "Text (Tab delimited)" كما يُوضح السهم الثالث، ثم نضغط "حفظ" حيث يشير السهم الرابع.

[1] البيانات الخاصة بالملف excdata1 موجودة في الجدول (م1.1) في الملحق (1)، ويمكن للقارئ إدخالها في ملف اكسل بلاسم المستخدَم.

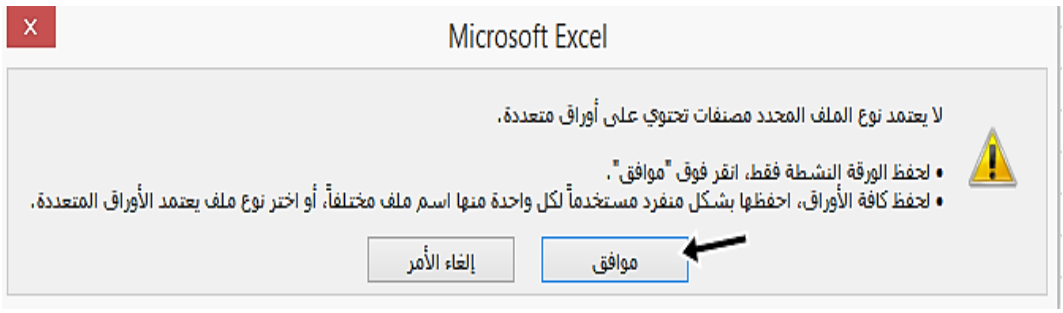
[2] الامتداد "xls." هو لإصدارات Excel 2003 فأقل، والامتداد "xlsx." هو لإصدارات Excel 2007 فأعلى، وقد تم استخدام الاصدار Excel 2010 في تنفيذ هذه الخطوات.

في الوضع الافتراضي، سيحتوي أي ملف اكسل على أكثر من ورقة عمل بداخله، (عادة ثلاث أوراق)، وسنفترض أن الورقة الأولى بها البيانات المطلوب نقلها لنظام R والأخريات إما فارغة أو بها بيانات أخرى.



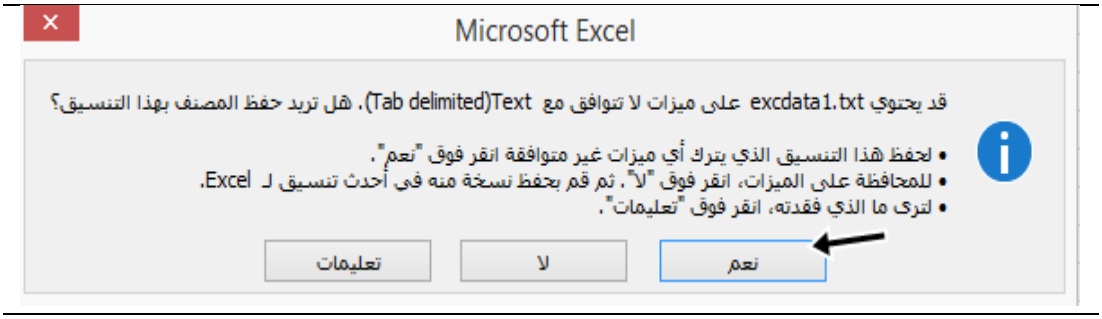
شكل 6.3 ب: النافذة الفرعية لحفظ باسم في اكسل والخيارات المطلوبة

وهذا الأسلوب الذي يتم شرحه الآن يسمح بتخزين الورقة المفتوحة أو النشطة فقط كملف نصي، لذلك ستظهر لنا رسالة في نافذة كتلك التي في الشكل (6.3 ج):



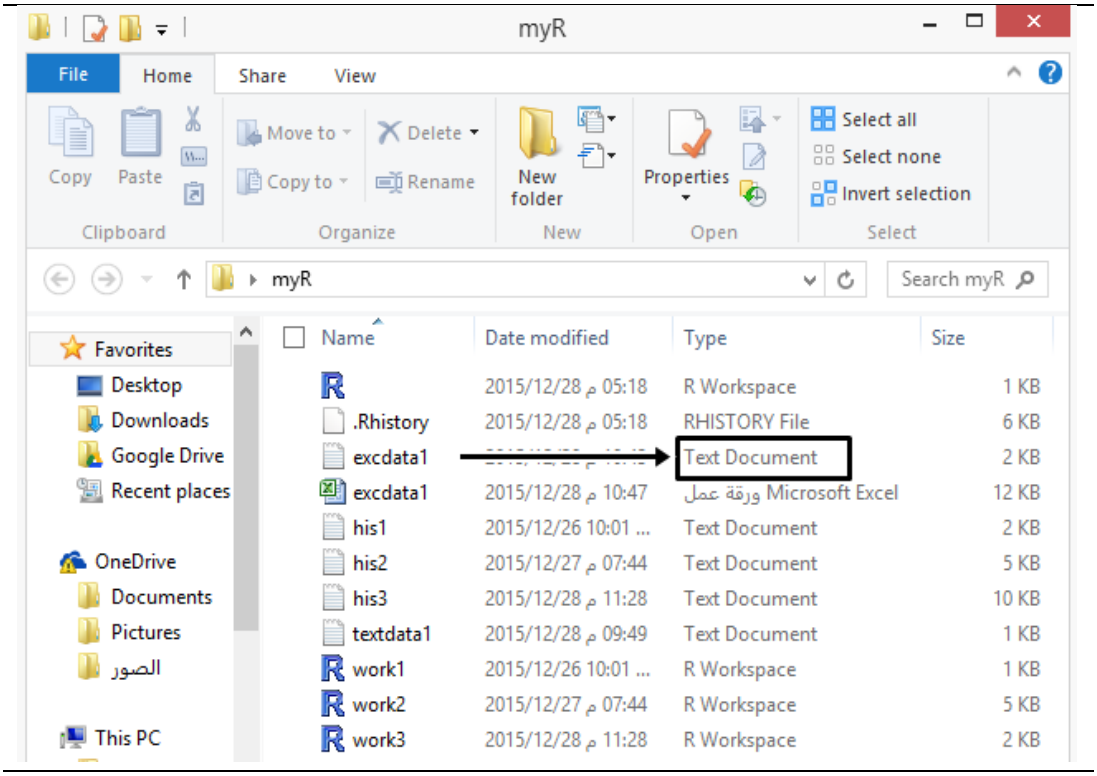
شكل 6.3 ج: رسالة نظام اكسل لحفظ الورقة النشطة

لحفظ الورقة النشطة التي تحتوي على البيانات المطلوبة يتم اختيار (موافق) فتظهر رسالة أخرى للتنبيه إلى أن الملف الجديد النصي سيفقد المميزات الخاصة بالملفات المخزنة بصيغة اكسل، وهذا ما يوضحه الشكل (6.3د)، علما بأن ذلك لن يؤثر على عملية استيراد الملف في برنامج R لاحقا؛



شكل 6.3 د: رسالة نظام اكسل الخاصة بفقد مميزات الامتداد xlsx.

نختار "نعم" فتتم عملية حفظ الملف "excdata1" بالامتداد ".txt"، ويمكنك رؤية الملف بعدها في الحافظة "myR" كما يوضح الشكل (6.3 هـ)؛



شكل 6.3 هـ: الملف "excdata1" كملف نصي في الحافظة "myR"

الآن نحن مستعدون للخطوة الثانية، والتي ستتضمن فتح وتخزين الملف النصي "excdata1.txt" في نظام R؛

(2) في برنامج R:

في لوحة مراقبة R، يتم استخدام الدالة `read.table` لاستيراد الملف النصي "excdata1.txt" بالصورة التالية:

```
> data.imp2<-read.table("excdata1.txt")
```

ولاحظ أنه تم تعيين الاسم data.imp2 كاسم لملف البيانات في R، (علما بأنه يمكن تعيين أي اسم آخر)، ويتم حفظه كإطار بيانات، (والذي له نفس مكونات الملف "excdata1.txt" والملف "excdata1.xlsx")، وهو يمثل المعدلات الدراسية لمجموعة من الطلبة[1] في ستة أقسام علمية في كلية العلوم بجامعة بنغازي في ليبيا خلال 41 فصل دراسي. ولنقوم الآن باستدعاء تلك البيانات؛

```
> data.imp2
```

	Statistics	Botany	Zoology	Chemistry	Physics	Mathematics
S1	1.51	1.76	1.68	1.67	1.40	1.47
S2	1.67	1.95	1.80	1.75	1.40	1.55
S3	1.88	1.70	1.96	1.67	1.70	1.56
S4	1.81	1.77	1.90	1.66	1.53	1.48
S5	1.79	1.89	1.83	1.58	1.55	1.55
S6	1.75	2.01	1.89	1.74	1.39	1.41
S7	1.69	1.90	1.97	1.88	1.51	1.66
S8	1.77	1.88	1.91	1.77	1.40	1.56
S9	1.72	1.66	1.70	1.68	1.62	1.66
S10	1.90	1.91	1.86	1.76	1.31	1.61
S11	1.96	2.01	1.78	1.71	1.27	1.51
S12	1.76	1.85	1.77	1.78	1.38	1.48
S13	1.90	1.83	1.73	1.83	1.51	1.55
S14	1.85	1.85	1.77	1.85	1.46	1.67
S15	1.91	1.93	1.79	1.84	1.65	1.72
S16	2.05	1.99	1.73	1.76	1.46	1.48
S17	1.81	2.00	1.80	1.78	1.48	1.54
S18	1.75	2.02	1.77	1.78	1.47	1.43
S19	1.81	1.86	1.84	1.84	1.57	1.28
S20	1.82	2.00	1.65	1.66	1.43	1.41
S21	1.84	2.04	1.96	1.67	1.63	1.45
S22	1.73	1.88	1.82	1.67	1.53	1.25
S23	1.69	2.00	1.93	1.73	1.41	1.43
…	…	…	…	…	…	…
S41	1.74	1.86	1.71	1.69	1.66	1.26

ولاحظ أن العامود الأول إلى اليسار يمثل أسماء الصفوف للبيانات، وهي الفصول الدراسية، ولا يمثل متغيرا كما هو الحال في الأعمدة الستة الباقية.

2.2.3.3 استيراد ملف اكسل بالامتداد الأصلي (Importing Excel File with Original Extension)

يعتمد هذا الأسلوب على استيراد ملف البيانات بصيغة اكسل الأصلية دون تحويلها مسبقا إلى صيغة أخرى كما في الأسلوب الأول، إلا أن ذلك لا يعني بالطبع أن البيانات سيتعامل معها نظام R بصيغة اكسل، بل سيتم تحويلها إلى الامتداد ".RData" ، ويتطلب استخدام هذا الأسلوب وجود حزم إضافية في R هي

[1] المعدلات مقاسة لمجموعات الطلبة في كل فصل دراسي على نظام الوحدات الفصلي من 0.00 إلى 4.00 وحدات.

rJava، XLConnect، وXLConnectJars . تلك الحِزم التي تم شرح تحميلها في البند (4.1) في الفصل الأول. لذلك سنقوم أولا باستدعاء[1] تلك الحِزم من مكتبة حِزم R بالصورة التالية:

```
> library(rJava)
> library(XLConnectJars)
> library(XLConnect)

XLConnect 0.2-11 by Mirai Solutions GmbH [aut],
  Martin Studer [cre],
  The Apache Software Foundation [ctb, cph] (Apache POI,
Apache Commons
    Codec),
  Stephen Colebourne [ctb, cph] (Joda-Time Java library)
http://www.mirai-solutions.com ,
http://miraisolutions.wordpress.com
```

ولابد من التنويه هنا إلى ضرورة استدعاء الحزم الثلاثة بالترتيب المعروض لأن تغيير هذا الترتيب قد يؤدي لحدوث أخطاء في عملية استيراد الملف. أما الرسالة التي ظهرت بعد استدعاء الحزمة الثالثة فهي مجرد معلومات خاصة بالحزمة XLConnect تشمل رقم الإصدار واسم مبرمج الحزمة والموقع الالكتروني وغيرها من المعلومات.

بعد ذلك يتم استخدام الدالة readWorksheetFromFile بعد تعيين اسم جديد للملف المستورد بالصورة التالية:

```
> data.imp3<-readWorksheetFromFile("excdata1.xlsx",
sheet=1,rownames=1)
```

حيث يتم كتابة اسم ملف البيانات في اكسل في البداية، يليه ترتيب ورقة العمل التي توجد بها البيانات المطلوبة، وهي الأولى في ملفنا، ثم تعريف العامود الذي توجد به أسماء الصفوف في ملف اكسل، والخيار الأخير لن يكون ضروريا إذا لم توجد في الملف الأصلي للبيانات أسماء للصفوف.

ولاحظ الآن أن إطاري البيانات data.imp2 و data.imp3 هما متماثلين تماما بالرغم من أن الصيغة الأصلية لملف كل منهما مختلفة عن الآخر. لذلك يمكن "الاستغناء" عن أحدهما، وليكن data.imp3؛

```
> rm(data.imp3)
```

[1] نذكّر من جديد هنا أنه إذا ما ظهرت رسالة تحذيرية بعد تنفيذ الأمر library(rJava) تفيد بوجود خطأ في الاستدعاء، فهذا يعني أن جهازك بحاجة لبرنامج (Java™ Oracle) كما أشرنا في الفصل الأول ضمن خطوات تحميل الحزم الإضافية. وعموما يمكنك تحميل هذا البرنامج مجانا من الموقع الالكتروني الرسمي له وهو (http://www.java.com).

وتوجد مع دالة استيراد ملفات اكسل بعض الخيارات الإضافية التي يمكن للقارئ الاطلاع عليها باستخدام دالة المساعدة؛ help(readWorksheetFromFile). ومن ضمن تلك الخيارات امكانية استيراد جزء محدد من البيانات الموجودة في الورقة النشطة في اكسل، عن طريق اختيار العامود أو الصف الذي يُراد البدء منه، (باستخدام الخيار startCol أو startRow، على الترتيب)، واختيار العامود أو الصف الذي يُراد الانتهاء عنده، (باستخدام الخيار endCol أو endRow، على الترتيب).

وكمثال على استخدام هذه الخيارات، لنفرض أننا نريد استيراد البيانات التي تخص الأقسام الثلاثة الأولى خلال السبعة فصول دراسية الأولى بدون استخدام أسماء الصفوف، عندها يتم كتابة:

```
> data.imp4<-
readWorksheetFromFile("excdata1.xlsx",sheet=1
,startCol=2,endCol=4,startRow=1,endRow=8)

> data.imp4

  Statistics   Botany  Zoology
1   1.511481 1.759762 1.680968
2   1.674444 1.950238 1.801935
3   1.876296 1.703333 1.961935
4   1.809259 1.772143 1.903226
5   1.786296 1.887143 1.825806
6   1.745556 2.009286 1.889677
7   1.692222 1.895476 1.969355
```

ولاحظ أن الاختيار بالنسبة للصفوف كان من الصف الأول startRow=1 لكي يتم أخذ أسماء الأعمدة في الاعتبار.

ملاحظة هامة:

نذكّر هنا أنه في معظم الحالات، عند إغلاق برنامج R وإعادة فتحه من جديد لابد من استدعاء الحزم الإضافية مرة أخرى، (باستخدام دالة library)، إذا ما أراد المستخدم العمل عليها.

3.3.3 تصدير ملفات البيانات من R (Exporting Data Files from R)

قد يرغب المُستخدِم أحيانا بتصدير بعض ملفات البيانات، (مثل أُطر البيانات، المصفوفات، أو المتجهات)، من R بغرض استخدامها في برامج أخرى، وتوجد بعض الطرق لعمل ذلك، حيث سنتناول طريقتين بسيطتين لتصدير ملفات البيانات من R إلى برنامج اكسل؛

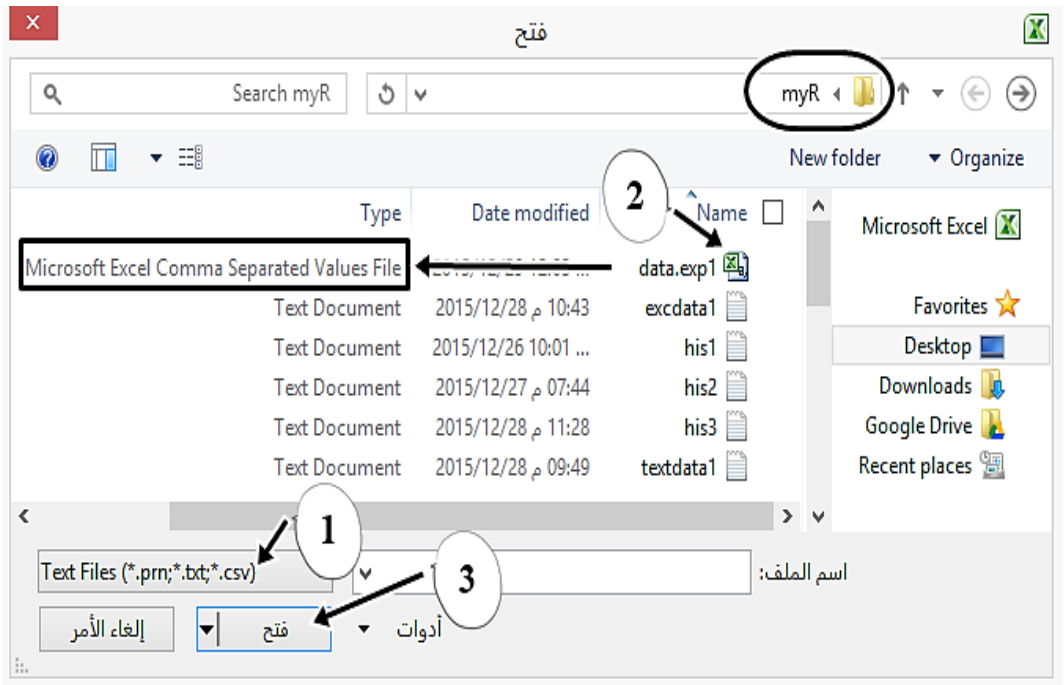
- **الطريقة الأولى:**

وتعتمد على استخدام الدالة[1] `write.csv` أو `write.csv2` لتصدير إطار البيانات، وهو عادة النوع الأكثر استخداما، بصيغة ملف نصي كما سيوضح المثال التالي:

بفرض أننا نريد تصدير إطار البيانات `data.f1` الموجود في ملف العمل "work3"، إلى ملف نصي له الامتداد "csv" باسم "data.exp1" مثلا، يتم عندها في لوحة مراقبة R كتابة:

```
> write.csv2(data.f1,file="data.exp1.csv")
```

فيتم على الفور إنشاء ذلك الملف في مسار الحافظة "myR". بعد ذلك يمكن فتح هذا الملف من داخل برنامج اكسل عن طريق اختيار فتح فتظهر نافذة فرعية كما في الشكل (7.3) فنقوم باختيار إظهار أنواع الملفات النصية حيثما يشير السهم الأول، بعدها يتم اختيار الملف المطلوب "data.exp1"، والمشار إليه في السهم الثاني، ثم نضغط على "فتح" حيث يشير السهم الثالث.



شكل 7.3: خطوات فتح الملف النصي "data.exp1" من الحافظة "myR"

ستظهر البيانات بعدها في ورقة عمل في برنامج اكسل، كما يظهر في الشكل (8.3)، نقوم بعد ذلك بحفظها بالامتداد الافتراضي لبرنامج اكسل وهو "xlsx" أو "xls". ويمكن بعد ذلك استخدام تلك البيانات في اكسل أو نقلها إلى أي برنامج تحليل بيانات آخر.

[1] استخدام إحدى هاتين الدالتين يعتمد على إصدار اكسل لدى المستخدِم، لذلك يمكنك تجربة كليهما للتحقق.

	s.level	s.age	s.openion
1	45	20	average
2	61	21	good
3	25	20	bad
4	85	24	average
5	77	23	bad
6	74	20	bad
7	90	19	average
8	50	26	bad
9	64	23	good
10	71	22	bad

شكل 8.3: بيانات ملف اكسل "data.exp1"

- **الطريقة الثانية:**

هي عبارة عن مجموعة من الأوامر ترتبط بالدالة `Workbook` ويتم استخدامها كما هو موضح في المثال التالي:

لنفرض من جديد أننا نود تصدير نفس إطار البيانات `data.f2` إلى صيغة اكسل مباشرة، حيث سيُعطى الاسم `data.exp2` مثلا، ويجب في هذه الطريقة تعيين اسم لورقة العمل المرغوب تصدير البيانات إليها، ولتكن "d1" مثلا، عندها يتم كتابة الأوامر التالية بنفس الترتيب:

```
> data.exp2<-loadWorkbook("data.exp2.xlsx",create=TRUE)
> createSheet(data.exp2,name="d1")
> writeWorksheet(data.exp2,data.f2,sheet="d1")
> saveWorkbook(data.exp2)
```

ستجد بعد ذلك ملف البيانات "data.exp2" موجودا بصيغة اكسل (بالامتداد "xlsx") في مسار الحافظة التي نعمل عليها "myR"، ويمكنك فتحه مباشرة في برنامج اكسل. ونذكر القارئ في نهاية الفصل الثالث بضرورة تخزين ملف العمل "work3" وسطور الأوامر "his3".

الفصل الرابع

التحليل الاستكشافي للبيانات باستخدام R

(Exploratory Data Analysis (EDA) using R)

1.4 أنواع البيانات (Data Types)

2.4 التحليل الاستكشافي للبيانات الأحادية (EDA for Univariate Data)

1.2.4 التمثيل البياني للبيانات الأحادية الكمية

(Graphical Display for Quantitative Univariate Data)

2.2.4 التمثيل البياني للبيانات الأحادية النوعية

(Graphical Display for Qualitative Univariate Data)

3.4 التحليل الاستكشافي للبيانات المتعددة (EDA for Multivariate Data)

1.3.4 التعامل مع متغيرات التقسيم (Dealing with Grouping Variables)

2.3.4 تكوين جداول البيانات في اتجاهين (Constructing two-way Data Tables)

3.3.4 التمثيل البياني للبيانات المتعددة (Graphical Display for Multivariate Data)

4.4 التحليل الاستكشافي للبيانات `stu.data1`: دراسة حالة

(EDA of `stu.data1`: Case Study)

1.4.4 استكشاف متغيرات الدراسة بصورة أحادية

(Exploring Data in Univariate Fashion)

2.4.4 الاستكشاف متعدد المتغيرات في الدراسة

(Exploring Data in Multivariate Fashion)

3.4.4 أهم استنتاجات التحليل الاستكشافي للبيانات (Important Conclusions of the EDA)

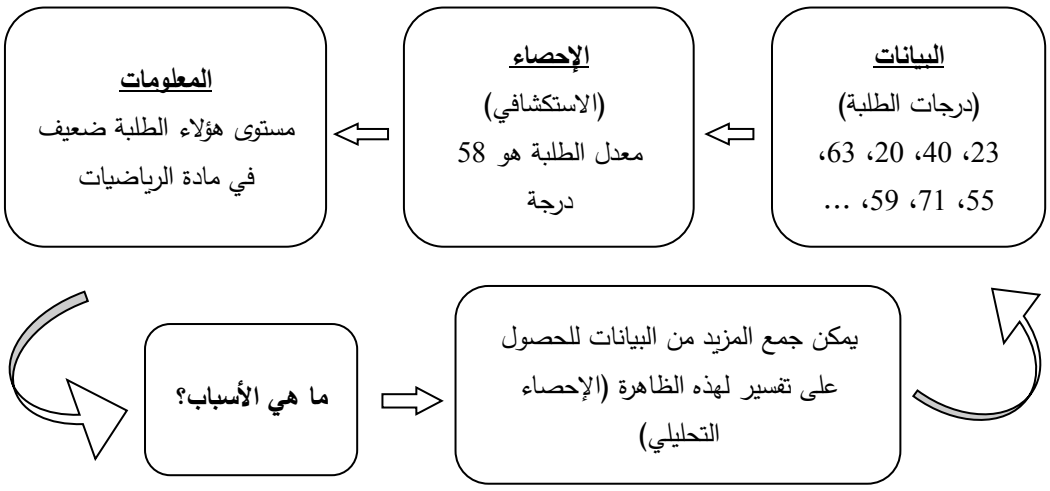
بداية من هذا الفصل، سنبدأ بتطبيق المقاييس والأساليب والنماذج الإحصائية على البيانات المختلفة باستخدام دوال R، على اعتبار أن القارئ مُلم بالحد الأدنى للمفاهيم والأساليب الأساسية لكل من الإحصاء الاستكشافي أو الوصفي (Exploratory or Descriptive Statistics) والإحصاء الاستدلالي (Inferential Statistics). وسيتم التركيز في هذا الفصل بالتحديد على تناول دوال R الخاصة **بالتحليل الاستكشافي للبيانات** ((Exploratory Data Analysis (EDA)، على أن يتم تناول تطبيقات حساب الاحتمال ودوال التقدير واختبارات الفروض الخاصة بأساليب الإحصاء الاستدلالي في الفصلين القادمين. وسيتم في البداية تناول أهم التعريفات الخاصة بأنواع البيانات باختصار لكي يسهل على القارئ توظيف المقياس أو الأسلوب الإحصائي المناسب حسب طبيعة البيانات المتوفرة بغية الوصول للهدف المطلوب.

1.4 أنواع البيانات (Data Types)

إن **البيانات** (Data) تمثل المادة الخام التي تمكننا من الحصول على المعلومات المفيدة من خلال استكشافها وتحليلها. ولفظ "بيانات" يظهر كثيرا في الكتب والمطبوعات ووسائل الإعلام والتعاملات اليومية في العموم، وهو مصطلح شمولي، فالبيانات الشخصية التي تضم الاسم، العمر، النوع، الحالة الاجتماعية، وغيرها هي نوع من البيانات، والتاريخ الطبي للمريض والذي يضم قراءات نسب السكر، ضغط الدم، درجات الحرارة المسجلة، وغيرها هي بيانات، وكذلك المشاهدات الناتجة عن تجربة كيميائية لقياس التفاعل الناتج عن دمج بعض المحاليل هي أيضا بيانات، وهكذا.

لذلك يمكننا القول بأن أي ظاهرة أو دراسة أو تجربة أو حتى مراقبة لعملية معينة يمكن أن ينتج عنها جميعا بيانات، فالبيانات هي المقياس الفعلي الذي نحصل عليه في النهاية.

وقد يخلط البعض أحيانا بين مفهوم "البيانات" ومفهوم "المعلومات"، وهما في الواقع مختلفان على الأقل من وجهة النظر الإحصائية، والشكل (1.4) يعطي مثلا بسيطا يمكن من خلاله توضيح كلا من المفهومين وعلاقة علم الإحصاء بكل منهما بصورة عامة.



شكل 1.4: هيكلية تعامل علم الإحصاء مع البيانات والمعلومات من خلال بيانات خاصة بدرجات مجموعة من الطلبة في مادة الرياضيات في إحدى الكليات العلمية، (الدرجة من 100)

أي أن الأساليب الإحصائية تقوم بالتعامل مع البيانات بصورة استكشافية أو تحليلية بغية الحصول على المعلومات المطلوبة بلغة الأرقام أو بالرسومات التوضيحية. وقواعد البيانات التي يتعامل معها علم الإحصاء عادة ما يكون لها بناء محدد مكون من مجموعة من **المفردات** (Individuals) أو **المشاهدات** (Observations) ناتجة عن مقياس **لمتغير** (Variable) أو أكثر.

والمتغير في علم الإحصاء، يشير إلى الصفة المميزة لشيء يمكن التعبير عنه بقيمة عددية أو غير عددية. فأي ظاهرة يمكن قياسها بوحدة قياس مناسبة يمكن أن يعبر عنها بمتغير، فالطول (بالسنتيمتر أو القدم) لمجموعة من الأشخاص هو متغير، والعمر الاستهلاكي (بالأشهر أو السنوات) لبطاريات السيارات هو متغير، وتقديرات طلبة الجامعة (ممتاز، جيد جدا، جيد، ...) تمثل أيضا متغير.

إن طرق التعامل مع البيانات واستخلاص المعلومات المباشرة وغير المباشرة منها يعتمد على طبيعة تلك البيانات، وعلى الباحث التأكد جيدا من هذه النقطة بغية تحديد الأسلوب الإحصائي المناسب للتعامل معها. ويمكن من حيث الطبيعة أن نقسم أنواع البيانات إلى نوعين رئيسيين هما؛ البيانات **الكمية** (Quantitative Data) والبيانات **النوعية** أو **الوصفية** (Qualitative Data)، والتي قد تسمى أيضا بالبيانات **القطاعية** أو **التصنيفية** (Categorical Data).

البيانات الكمية هي تلك البيانات التي تحتوي على مشاهدات تم قياسها ورصدها مباشرة بصورة أعداد لها قيمة كمية، مثل درجات الطلبة من 10 أو من 100، وأسعار النفط بالدولار. وهذا النوع من البيانات تمثله متغيرات يمكن التعامل معها بواسطة العمليات الحسابية مباشرة.

أما البيانات التي يتم التعبير عنها بصفات لا بأعداد، مثل النوع (ذكر، أنثى)، أو حالة الطقس المتوقعة (صحو، غائم، ممطر، ...)، فهذه البيانات تسمى بيانات نوعية أو قطاعية حيث أنها تأخذ قيما تحدد تصنيف المفردة إلى نوع أو فئة أو قطاع محدد ولا يمكن التعامل معها بالعمليات الحسابية المباشرة.

ويمكن تقسيم **البيانات الكمية** من حيث المقياس إلى نوعين أساسين هما:

1. بيانات **المقياس الفئوي** (Interval Scale): وهذا النوع من المقاييس يسمح بترتيب المشاهدات على مقياس محدد بحيث يمكن تحديد القيمة الفعلية لكل مشاهدة. إلا أن المقياس الفئوي لا يقيس الصفر كقيمة. ومن الأمثلة على هذا النوع من البيانات قياس درجة الحرارة عدديا.

2. بيانات **المقياس النسبي** (Ratio Scale): وهي مشابهة لبيانات المقياس الفئوي إلا أنها تعتبر الصفر من ضمن درجات القياس، لذلك فإنها تسمح بحساب النسبة بين قيمتين ضمن المفردات، مِثل أن نقول أن طفلا عمره 12 سنة هو أكبر بمرتين من طفل عمره 6 سنوات. وهذا النوع من المقاييس يكون شاملا للعديد من أنواع البيانات الكمية مثل العمر والوزن والطول والزمن والمبالغ النقدية وغيرها.

وهنالك تقسيم آخر للبيانات الكمية من حيث طبيعة الأرقام وهو التقسيم إلى بيانات **منفصلة** (Discrete)، وبيانات **متصلة** (Continuous). فالبيانات الكمية المنفصلة تأخذ قيما في مجموعات أعداد منتهية أو غير منتهية وقابلة للعد، مثل الأعداد الصحيحة (Integers) والتي تم ذكرها عندما تم تناول المتجهات في الفصل السابق.

أما البيانات الكمية المتصلة فهي تأخذ قيما ضمن فترة من الأرقام، وتشمل مجموعة الأعداد الحقيقية (Real Numbers)، والتي تُعرف في نظام R بالنوع "Double".

وتنقسم البيانات النوعية أو الوصفية هي الأخرى إلى نوعين هما:

1. بيانات **المقياس الاسمي** (Nominal Scale): وهي تمثل تصنيف المشاهدات إلى مستويات أو قطاعات مختلفة، إلا أنه لا يمكن تحديد القيمة الفعلية لكل مستوى، ولا حتى ترتيب هذه المستويات بشكل تصاعدي أو تنازلي. فمثلا، لا يمكن القول أن الإجابة "نعم" هي أفضل من الإجابة "لا" فيما يتعلق بسؤال معين. ومن أمثلة هذا النوع من البيانات نوع الشخص (ذكر، أنثى)، وتحديد أسماء الأشياء وأسماء الألوان وما شابه ذلك.

2. بيانات **المقياس الترتيبي أو الرُتبي** (Ordinal or Rank Scale): وهي على عكس النوع السابق، فطبيعتها تسمح بترتيب المشاهدات وفق تدرج محدد من الأقل قيمة إلى الأكثر قيمة أو العكس، مثل تقديرات الطلبة (ممتاز، جيد جدا، جيد، ...)، إلا أنه لا يمكن حساب الفرق في القيمة بين أي مستويين من المستويات المرتبة، ولا معنى للأفضلية في هذه التقسيمات أيضا.

من جديد، يمكن التعامل مع البيانات، كمية كانت أو وصفية، من زاوية إحصائية بشكل مفرد، أي التعامل مع متغير واحد، وهو ما يُعرف **بالبيانات الأحادية** (Univariate Data)، أو بشكل متعدد، أي التعامل مع عدة متغيرات في آن واحد، وهو ما يُعرف **بالبيانات المتعددة**[1] (Multivariate Data). والأساليب الإحصائية التي تُستخدم مع البيانات الأحادية تعني أنه سيتم حساب تلك المقاييس لكل متغير على حدة، حتى وإن كانت مجموعة أو قاعدة البيانات متكونة من متغيرات عديدة.

ومثال على ذلك؛ بافتراض وجود بيانات بها خمسة متغيرات تمثل درجات طلبة في خمسة كليات مختلفة، فإنه إذا ما تم التعامل، (أي استخدام الأساليب الإحصائية)، مع درجات الطلبة لأي متغير بمفرده فإن ذلك يكون ضمن إطار التعامل مع البيانات الأحادية أو يمكن أن يسمى تحليل أحادي حتى وإن تم تنفيذ ذلك لكل المتغيرات الخمسة. أما إذا ما تم استخدام أسلوب أو أساليب إحصائية تحتوي على صيغ تتعامل مع متغيرين أو أكثر في نفس الوقت فإن ذلك يُسمى تحليل ثنائي، (في حالة استخدام متغيرين)، أو تحليل متعدد إذا ما تم إدراج أكثر من متغيرين في التحليل الإحصائي. وسيتم الإشارة إلى نوع التحليل الإحصائي، إضافة إلى اسمه، عند تنفيذه في نظام R في البنود القادمة.

2.4 التحليل الاستكشافي للبيانات الأحادية (EDA for Univariate Data)

إن التحليل الاستكشافي للبيانات، سواء الأحادي أو المتعدد، هو ذلك التحليل الذي يشتمل على الطرق الرقمية والبيانية التي تهتم بجمع وتنظيم واستكشاف وعرض الصفات المميزة ضمن البيانات بغية الحصول على المعلومات الكامنة بداخلها. والهدف من استخدامه هو تمكين الباحث من الفهم الأولي لما تحتويه البيانات من معلومات.

ويمكن الوصول لهذا الفهم عن طريق تتبع السلوك أو النمط (Pattern) السائد في هذه البيانات، هذا النمط الذي يعبر عن التغير المنتظم، (الغير عشوائي)، في البيانات. ولهذا، يجب استخدام الأدوات المناسبة للتحليل الاستكشافي، والتي تتمثل في مجموعة من الإحصاءات الرقمية والرسومات البيانية، "لتلخيص" المعلومات المستخلصة من البيانات في صورة رقمية أو مرئية تكون واضحة وقابلة للتفسير من قِبل الباحث.

ولتنظيم العمل في هذا الفصل، كما هو النمط السائد في الكتاب، سيتم تخزين البيانات والأشياء في ملف عمل جديد، باسم "work4"، بعد تغيير مسار العمل إلى الحافظة "myR" على سطح المكتب كما هو الإجراء المتبع في كل فصل. وسوف يتم فتح نظام R من ملف العمل "work4" الخاص بالفصل الرابع. إضافة إلى تخزين سطور الأوامر، في ملف نصي باسم "his4".

[1] وتوجد حالة خاصة من البيانات المتعددة هي **البيانات الثنائية** (Bivariate Data).

الآن وقد تم تهيئة نظام R للعمل على محتويات هذا الفصل، سيتم استيراد ملف بيانات من برنامج اكسل لاستخدامه في تطبيق أساليب التحليل الاستكشافي، (كبيانات أحادية، ثنائية، ومتعددة). هذا الملف سيكون اسمه "studata1.xlsx"، وسيتم نقله إلى مسار الحافظة "myR" تمهيدا لاستيراده من داخل نظام R، كما تم توضيح هذه الخطوة في الفصل السابق.

ويمكن للقارئ متابعة هذه الخطوة مع الكتاب من خلال مراجعة بيانات الملف "studata1" في الجدول (م2.1) في الملحق (1) في نهاية الكتاب، ثم يقوم بإدخالها[1] في هذه المرحلة في برنامج اكسل في جهازه، تمهيدا بعد ذلك لاستيرادها من داخل برنامج R.

وسنستخدم الأسلوب الثاني الذي تم تناوله في الفصل السابق لاستيراد الملف المطلوب، والذي سيُعطَى الاسم "stu.data1" في R، والذي يعتمد على استدعاء الحزم الإضافية. لذلك، وبعد نقل الملف "studata1.xlsx" إلى الحافظة "myR"، يتم كتابة سطور الأوامر الخاصة باستيراد ملف البيانات المطلوب بالشكل التالي:

```
> library(rJava)
> library(XLConnectJars)
> library(XLConnect)

> stu.data1<-
readWorksheetFromFile("studata1.xlsx",sheet=1,rownames=1)

> stu.data1
      grd1 grd2 grd3 age gen sem fam hou
stu1    55   50   60  22   m   3  10   2
stu2    49   52   50  19   m   8  11   2
stu3    60   54   51  23   m   2  10   2
stu4    65   70   54  20   f   3   8   3
stu5    35   40   40  24   m   7  12   2
…       …    …    …   …    …   …  …    …
stu35   94   96   49  21   f   4   2   6
```

ولاحظ، لدواعي الاختصار، أننا لم نقم بعرض الملاحظة التي من المفروض أن تظهر بعد استدعاء `library(XLConnect)`، وكذلك قمنا باختصار عرض البيانات المتكونة من 35 مشاهدة، والتي ستظهر بشكل كامل لديك عند تنفيذ الأوامر في جهازك.

[1] يُراعى عند إدخال البيانات في برنامج اكسل ضرورة كتابة البيانات بنفس ترتيبها في الجدول، بمعنى كتابة أسماء الحالات stu1، stu2، ... في العامود الأول مع ترك الخانة الأولى فارغة، وكتابة أسماء المتغيرات grd1، grd2، ... في الصف الأول مع ترك الخانة الأولى فارغة.

قبل البدء بالتحليل الاستكشافي لإطار البيانات stu.data1، يجب التعرّف أولا على ما تمثله المتغيرات التي يحتويها، والتي تخص 35 طالبا جامعيا في إحدى الأقسام العلمية، (ويرمز لهم بالرمز stu*i* ، حيث $i = 1, 2, \ldots, 35$)؛ والمتغيرات الكمية grd1، grd2، وgrd3 تمثل درجات الطلبة، (من 100 درجة)، في ثلاثة مقررات دراسية مختلفة، المتغير الكمي age يمثل أعمار الطلبة، المتغير النوعي (الاسمي) gen يمثل نوع الطالب ذكر "m" أو أنثى "f"، المتغير النوعي (الرتبي) sem يمثل ترتيب الفصل الدراسي الذي تم رصد درجات المقررات فيه، المتغير الكمي fam يمثل عدد أفراد أسرة الطالب المقيمين في المنزل باستثناء الطالب، والمتغير الكمي hou يمثل عدد الغرف في منزل أسرة الطالب.

إن الأدوات الرئيسية في التحليل الاستكشافي هي عبارة عن مقاييس النزعة المركزية ومقاييس التشتت، إضافة للتمثيل البياني للبيانات والذي يُعتبر جزء هاما لا يتجزأ من منظومة التحليل الإحصائي في العموم. وسيتم فيما يلي عرض دوال R التي تُستخدم ضمن هذا الإطار؛

يمكن البدء بحساب ملخص سريع باستخدام الدالة summary والتي تحسب ستة مقاييس هي على الترتيب **القيمة الصغرى**، **الرُبيع الأول** (First Quartile)، **الوسيط** (Median)، **الوسط الحسابي** (Mean)، **الرُبيع الثالث** (Third Quartile)، و**القيمة الكبرى**؛

```
> summary(stu.data1)
      grd1            grd2            grd3            age       
 Min.   :35.00   Min.   :40.00   Min.   :27.00   Min.   :19.00  
 1st Qu.:63.50   1st Qu.:67.50   1st Qu.:49.50   1st Qu.:21.00  
 Median :75.00   Median :75.00   Median :53.00   Median :22.00  
 Mean   :71.31   Mean   :72.97   Mean   :55.34   Mean   :22.14  
 3rd Qu.:83.00   3rd Qu.:84.50   3rd Qu.:60.00   3rd Qu.:23.00  
 Max.   :94.00   Max.   :96.00   Max.   :95.00   Max.   :25.00  

    gen                 sem             fam          hou       
 Length:35          Min.   :2.000   Min.   : 2   Min.   :2.000  
 Class :character   1st Qu.:4.000   1st Qu.: 4   1st Qu.:3.000  
 Mode  :character   Median :5.000   Median : 5   Median :4.000  
                    Mean   :5.057   Mean   : 6   Mean   :4.114  
                    3rd Qu.:6.500   3rd Qu.: 9   3rd Qu.:5.000  
                    Max.   :8.000   Max.   :12   Max.   :6.000
```

لاحظ هنا أن دالة summary قد أعطت نتائج للبيانات بشكل فردي أو أحادي، بمعنى أن قيم المقاييس قد تم حسابها لكل متغير في إطار البيانات stu.data1 على حدة (في عامود مستقل)، ولا توجد نتيجة محسوبة باستخدام متغيرين أو أكثر في نفس الوقت، وهذا ما سيُلاحظ في كل المقاييس المُستخدمة في هذا البند. ولاحظ أيضا أنه للمتغير gen، الذي يأخذ قيما غير رقمية، قد تم عرض عدد المشاهدات (حجم العينة) فقط مع التنويه على طبيعة ذلك المتغير. وننوه هنا أنه لن يتم التعليق على هذه النتيجة أو النتائج التي ستظهر في البنود اللاحقة حيث سيتم تخصيص البند الأخير في هذا الفصل للتعليق والتحليل المنطقي لهذه النتائج كدراسة حالة.

من جديد يمكن حساب كل المقاييس الموجودة في دالة summary بدوال تأخذ صيغ منفصلة، إضافة لحساب مقاييس إحصائية أخرى. والجدول (1.4) يشتمل على أسماء بعض أهم المقاييس الإحصائية[1] التي تُستخدم في التحليل الاستكشافي للبيانات الأحادية في لغة R.

جدول 1.4: بعض مقاييس النزعة المركزية والتشتت في R

الدالة في R	تقوم بحساب
sum	مجموع البيانات
mean	الوسط الحسابي
median	الوسيط
quantile	التجزيئات (القيمة الصغرى، الرُبيع الأول، الرُبيع الثاني، الرُبيع الثالث، والقيمة الكبرى)
fivenum	ملخص الأرقام الخمسة[2] لتوكي (Tukey)، وهي (القيمة الصغرى، الرُبيع الأول، الرُبيع الثاني، الرُبيع الثالث، والقيمة الكبرى)
var	التباين[3]
sd	الانحراف المعياري
IQR	المدى الرُبيعي
skewness	معامل الالتواء[4]
kurtosis	معامل التفرطح

والمقاييس المذكورة في الجدول، تُستخدم على المتغير المفرد ولا يمكن استخدامها مع مصفوفة أو إطار البيانات ككل، أي أنه لا يتم كتابة أمر بالصورة؛ mean(stu.data1) مثلا، بل يمكن حساب الوسط الحسابي لأي متغير بمفرده، فمثلا:

```
> mean(stu.data1$grd1)
[1] 71.31429

> quantile(stu.data1$sem)

  0%  25%  50%  75% 100%
 2.0  4.0  5.0  6.5  8.0
```

[1] ننوه هنا بأن دالة المنوال (Mode) غير متوفرة بصورتها المعتادة في حزم R الأساسية حتى وقت إعداد هذا الكتاب، علما بأنه يمكن "تقدير" قيمة المنوال باستخدام دوال الحزمة الإضافية modeest. وسنقوم بكتابة دالة خاصة لحساب المنوال عند تناول موضوع إنشاء الدوال الخاصة.

[2] تظهر نتيجة الأرقام الخمسة في R بعناوين على شكل نسب مئوية تتناسب مع قيم التجزيئات (Quantiles).

[3] الدالة var تقوم بحساب تباين العينة، أي أن القسمة في صيغة التباين تكون على عدد المشاهدات ناقصا واحد. وكذلك الأمر بالنسبة لدالة الانحراف المعياري sd.

[4] يتطلب استخدام دالتي معامل الالتواء والتفرطح وجود الحزمة الإضافية (e1071).

```
> fivenum(stu.data1$grd3)
[1] 27.0 49.5 53.0 60.0 95.0

> sd(stu.data1$age)
[1] 1.647509

> IQR(stu.data1$fam)
[1] 5
```

ملاحظة:

عند وجود بعض القيم المفقودة في بياناتك، يجب استخدام الخيار na.rm=T ضمن الدوال (المقاييس) المستخدمة لكي يتم "إهمال" هذه القيم من بين المشاهدات المستخدمة في الحساب وإلا ستكون نتيجة المقياس هي NA، فعلى سبيل المثال لاحظ أن:

```
> mean(c(1,NA,3))
[1] NA
```

بينما

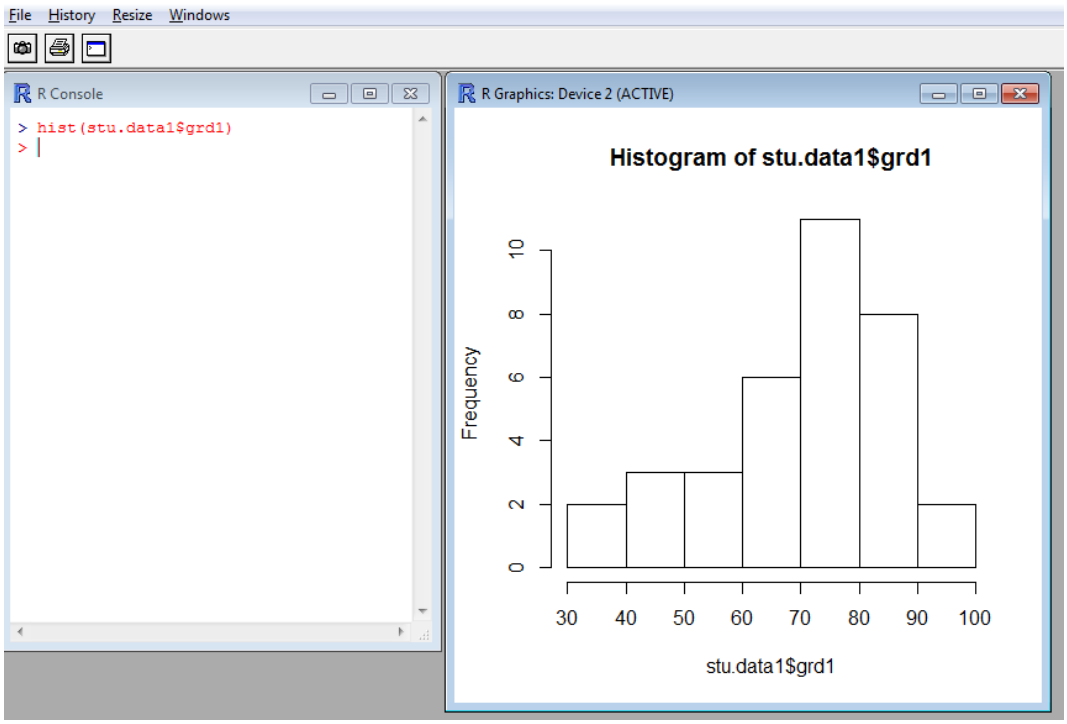
```
> mean(c(1,NA,3),na.rm=T)
[1] 2
```

1.2.4 التمثيل البياني للبيانات الأحادية الكمية

(Graphical Display for Quantitative Univariate Data)

نأتي الآن لاستخدام دوال الرسم أو التمثيل البياني في نظام R، وسيتم في هذا البند تعريف أهم دوال التمثيل البياني في لغة R والخاصة بتمثيل البيانات الأحادية الكمية، (وسيتم تناول تمثيل البيانات الأحادية النوعية في البند القادم)، ونود التنويه هنا إلى أنه عند استخدام أي أمر أو دالة للرسم في R فإن الشكل المطلوب تنفيذه سيظهر في نافذة مستقلة عن لوحة مراقبة R، كما يُلاحظ من المثال المستخدَم في الشكل (2.4).

ويكون لدى المستخدِم عندها عدة خيارات للتعامل مع ذلك الرسم، تلك الخيارات يتم استعراضها عن طريق وضع مؤشر الفأرة على نافذة الرسم ثم النقر على الزر الأيمن للفأرة. وأهم تلك الخيارات، (بحسب ترتيب عرضها)، هو الخيار الأول "Copy as metafile" والذي يمكّنك من أخذ نسخة من الرسم ولصقها في أي برنامج آخر (مثل برنامجي مايكروسوفت وورد واكسل)، وكذلك الخيار الثالث "Save as metafile" الذي يمكّنك من الاحتفاظ بنسخة من الرسم في المسار والاسم الذي تحدده وبالتالي يمكنك استدعاؤه أو إدراجه في برامج أخرى متى أردت، وهذا الخيار يُعتبر أفضل من الناحية العملية عند القيام بتنفيذ الكثير من الرسومات في الجلسة الواحدة.



شكل 2.4: الشكل العام لعرض الرسومات في R

ملاحظة:

سيكون من المفيد في كثير من الأحيان عرض أكثر من شكل بياني بشكل متجاور أو فوق بعضها البعض، أو حتى بشكل مصفوفة من الرسومات، ويتم هذا عادة باستخدام دالة مرافقة لدوال التمثيل البياني، هي دالة[1] `par(mfrow=c(*,**))`، حيث (* هو عدد صفوف مصفوفة الرسم، و** هو عدد أعمدة مصفوفة الرسم). وباستخدام هذه الدالة، ستتمكن من تنسيق مجموعة من الرسومات في شكل أو إطار واحد بحسب ترتيب عدد الصفوف والأعمدة المطلوب، وستلاحظ أن كل رسم ناتج عن دالة قمت بتنفيذها سيأخذ مكانه بالترتيب في الشكل العام إلى أن تنتهي من كل الرسومات.

- **تمثيل المدرج التكراري:**

نبدأ بدالة المدرج التكراري (Histogram) التي تكتب بالصورة `hist`، ويمكن استخدامها مع الخيار `prob=T` لعرض قيم التكرارات النسبية مقابل الأعمدة أو بدون ذلك الخيار لعرض التكرارات المعتادة. وكذلك يمكن رسم منحنى التوزيع الطبيعي "فوق" المدرج التكراري باستخدام دوال الرسم الإضافية `lines` و`density`.

[1] يُستخدم الخيار `mfrow` لترتيب الرسومات بحسب الصفوف، وإذا ما رغبنا بترتيب الرسومات بحسب الأعمدة فنستخدم الخيار `mfcol`.

وكمثال على الملاحظة السابقة؛ لنقم برسم ثلاثة مدرجات تكرارية للمتغير age في البيانات stu.data1 الأول باستخدام قيم التكرارات، والثاني باستخدام قيم التكرارات النسبية، والثالث بإدراج المنحنى الطبيعي مع قيم التكرارات النسبية، وذلك بشكل متجاور، مما يعني أننا سنستخدم عدد صفوف مساو للواحد وعدد أعمدة مساو لثلاثة كما يلي:

```
> par(mfrow=c(1,3))
```

سيتم بعد هذا الأمر فتح نافذة رسم فارغة، قم الآن بتنفيذ الرسم الأول؛

```
> hist(stu.data1$age)
```

سيظهر المدرج الأول في يسار نافذة الرسم، بعدها قم بتنفيذ الرسم الثاني؛

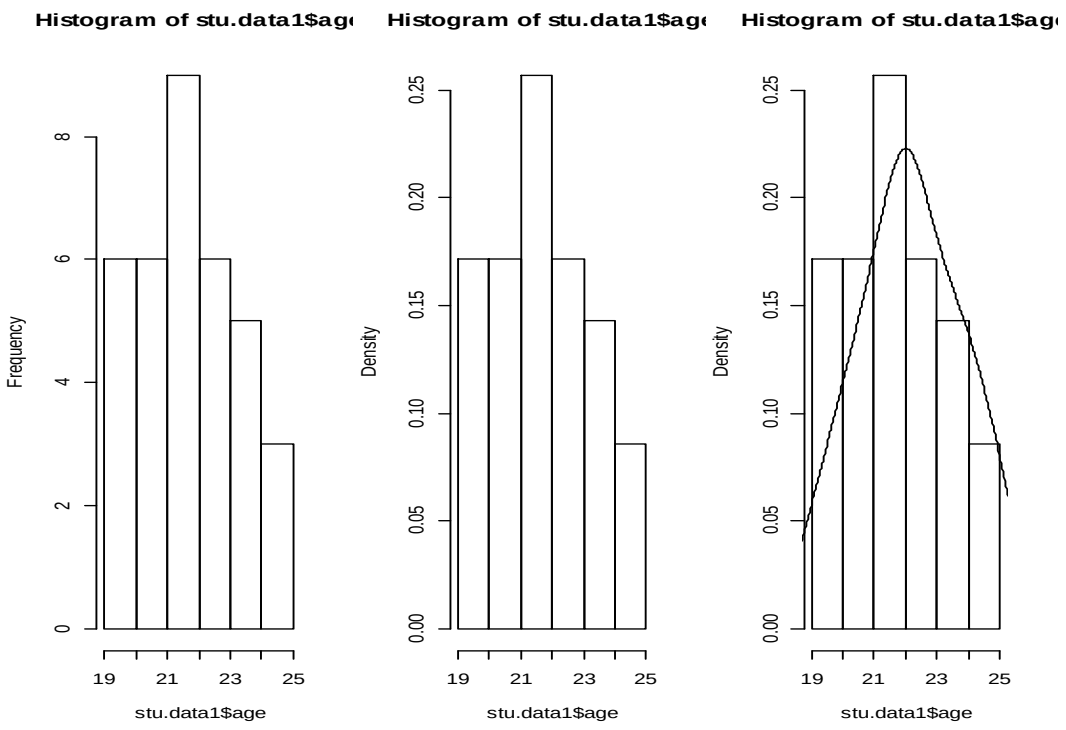
```
> hist(stu.data1$age,prob=T)
```

وسيظهر الرسم الثاني إلى جانب الأول، أخيرا قم بتنفيذ الرسم الثالث، والذي يتطلب كتابة السطرين التاليين؛

```
> hist(stu.data1$age,prob=T)
> lines(density(stu.data1$age))
```

وستجد أنه قد أصبح لديك شكل يضم المدرجات التكرارية الثلاثة، ويمكنك بعدها اختيار أي من الخيارات الخاصة بنسخ أو حفظ الرسم البياني كما وضحنا سابقا. والشكل (3.4) يوضح "مُحصلة" ما ستحصل عليه عند انتهاء الأوامر السابقة. أما إذا ما أردنا رسم مدرج تكراري واحد للمتغير، فيكفي استخدام أحد الأوامر أعلاه فقط.

وفيما يخص التعليق على التمثيل البياني في الشكل (3.4)، فسيتم تناوله بعد تعريف باقي الدوال المتعلقة بالتحليل الاستكشافي للبيانات كما ذكرنا سابقا.



شكل 3.4: المدرج التكراري للمتغير age في stu.data1 مع تكرارات (إلى اليسار)، ومع تكرارات نسبية (في الوسط)، ومع المنحنى الطبيعي (إلى اليمين)

ويمكنك إضافة إحدى الخيارين *=nclass أو *=breaks، (حيث ترمز * للعدد المطلوب)، لاختيار عدد الفترات أو الأعمدة في المدرج التكراري، إلا أن نظام R سيتجه "في الغالب" لاختيار عدد الفترات تلقائيا بالرغم من استخدامك لهذين الخيارين، فمثلا يمكنك تنفيذ دالة المدرج التكراري لنفس المتغير السابق age بالصور التالية، (بشكل مستقل[1])، وستحصل على مدرجات تكرارية لها أعداد فترات كما هو موضح في الملاحظات المرفقة مع الدوال:

```
> hist(stu.data1$age,nclass=4)# مدرج تكراري بأربع فترات
> hist(stu.data1$age,nclass=5)# مدرج تكراري بستة فترات
> hist(stu.data1$age,nclass=6)# مدرج تكراري بستة فترات
> hist(stu.data1$age,nclass=7)# مدرج تكراري بستة فترات
```

[1] يُفضل عادة إغلاق نافذة الرسم المفتوحة، بعد نسخها أو حفظها، لأن الرسم الجديد سيحل مكان القديم وقد يحدث أحيانا تداخل بين الرسومات.

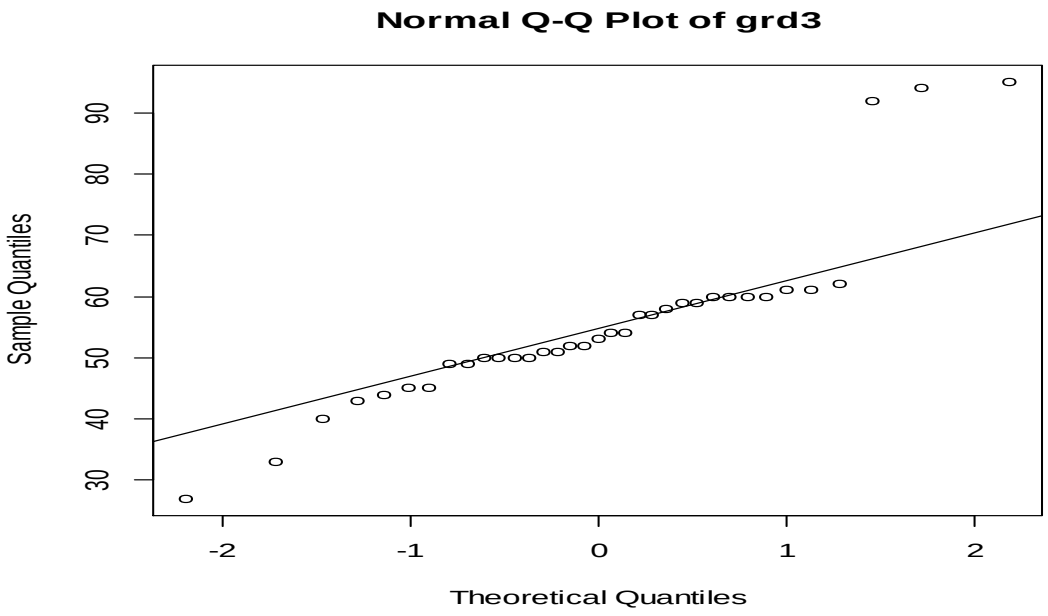
- **تمثيل شكل Q-Q الطبيعي:**

من ضمن الرسوم البيانية المساعِدة في مراقبة توزّع المتغيرات بتوزيع طبيعي، إضافة للمدرج التكراري، هو تمثيل **Q-Q الطبيعي**[1] (Normal Q-Q)، والذي تُعرّف دالته في R بالصيغة qqnorm، وعادة ما تُستخدم مع دالة رسم إضافية هي qqline والتي تُضيف خط مستقيم للرسم المرافق لها كما يوضح المثال التالي:

```
> qqnorm(stu.data1$grd3)# رسم Q-Q الطبيعي بدون خط مستقيم
> qqline(stu.data1$grd3)# إضافة خط مستقيم للرسم السابق
```

فنحصل على الرسم المطلوب، (والذي لن يتم عرضه هنا لأننا سنستخدم خيار إضافي معه تاليا)، إلا أنه في بعض الرسومات، ومن ضمنها هذا الرسم، لا يتم عرض اسم المتغير بشكل تلقائي، كما في تمثيل المدرج التكراري، لذلك يمكننا من ناحية تنظيمية إدراج اسم المتغير على الرسم، (في موضع عنوان الرسم العلوي كما سنرى هنا، أو في أسفل الرسم كما سنرى لاحقا)، حيث سنقوم بتغيير عنوان الرسم الافتراضي إلى العنوان الموجود في الشكل (4.4) والذي يشمل اسم المتغير grd3 باستخدام الخيار main ضمن دالة الرسم:

```
> qqnorm(stu.data1$grd3,main = "Normal Q-Q Plot of grd3")
> qqline(stu.data1$grd3)
```



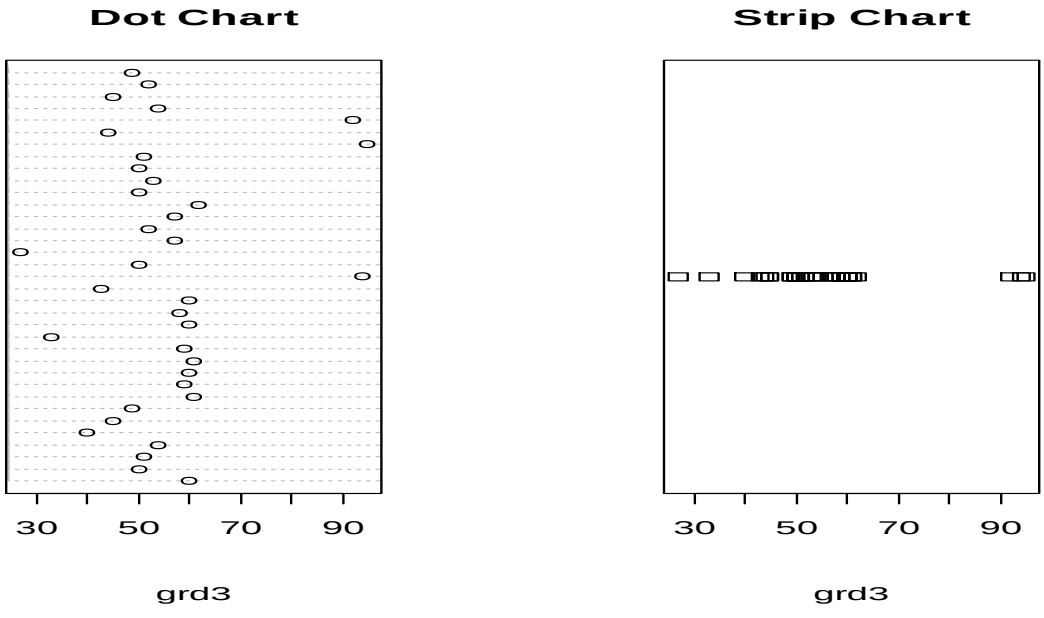
شكل 4.4: شكل Q-Q الطبيعي للمتغير grd3 في stu.data1 مع إدراج اسم المتغير في العنوان

[1] تستند فكرة هذا التمثيل البياني على رسم القيم الحقيقية للمتغير مقابل القيم المتوقعة للتوزيع الطبيعي الافتراضي لهذا المتغير.

- **التمثيل النقطي والتمثيل الشريطي:**

تعتبر دوال **الرسم النقطي** (Dot plot) dotchart و**الرسم الشريطي** (Strip plot) stripchart أيضا من دوال التمثيل البياني الخاصة بالبيانات الأحادية، ورغم بساطة التمثيل البياني الخاص بهما إلا أنهما تعكسان توزيع البيانات بشكل جيد. وسنقوم بتمثيل كلتا الدالتين في شكل واحد، (الشكل (5.4))، مع كتابة أسماء الدوال في عناوين الرسومات العلوية باستخدام الخيار main، وكتابة اسم المتغير، وهو grd3، في أسفل الرسم[1] باستخدام الخيار sub كما يلي:

```
> par(mfrow=c(1,2))
> dotchart(stu.data1$grd3,main="Dot Chart",sub="grd3")
>stripchart(stu.data1$grd3,main="Strip Chart",sub="grd3")
```



شكل 5.4: الرسم النقطي والشريطي للمتغير grd3 باستخدام دالة dotchart (إلى اليسار)، ودالة stripchart (إلى اليمين)

ولاحظ أن التمثيل النقطي يقوم بعرض انتشار البيانات على المحور الأفقي (محور X)، مع استخدام تدرج منتظم على المحور العامودي (محور Y) بحيث يعكس مواقع تركز نقاط المشاهدات في بُعدين (2D). أما الرسم الشريطي فيعرض انتشار البيانات في بُعد واحد هو المحور الأفقي. وفائدة هذه الرسومات تكمن في أنها تُظهر مواضع تركز أو انتشار المشاهدات وهذا يمنح مساعدة "مرئية" مع مقاييس التشتت لفهم كيفية توزع القيم أو النمط السائد في المتغير، وكذلك تحديد مواقع **القيم المتطرفة** (Outliers)، إن وُجدت.

[1] في الرسومات التي لا تتضمن مسميات في الجزء السفلي بشكل افتراضي، يمكن استخدام الخيار xlab لكتابة اسم المتغير أو أي مسمى آخر. فمثلا، يمكن ادراج الخيار "xlab="grd3 بدلا من الخيار "sub="grd3 في أي من الدوال الأخيرة.

ملاحظات:

1. توجد دالة رسم عامة[1] أخرى، تُعرّف في لغة R بالصيغة plot، يمكن للقارئ أن يستخدمها، وهي تُعطي رسم نقطي للمتغير بشكل مشابه لدالة dotchart، ويمكن استخدامها لتمثيل المتغير grd3 مثلا بالصورة plot(stu.data1$grd3). ولاحظ تغير طريقة العرض في الرسم عند استخدام الخيار "type=h" ضمن دالة plot؛ أي أن الأمر يُصبح plot(stu.data1$grd3,"type=h").

2. إذا لم ترغب في عرض مسميات معينة في أي موضع من المواضع الأربعة الرئيسية على الرسم، (أعلى، أسفل، يسار، أو يمين)، يمكنك استخدام أقواس الاقتباس الفارغة ("") بعد إشارة يساوي مع خيارات التسمية، فمثلا لرسم الدالة qqnorm بلا عنوان في الأعلى ولا مسمى في الأسفل يمكن كتابة:

```
> qqnorm(stu.data1$grd3,main = "",xlab="")
> qqline(stu.data1$grd3)
```

3. يمكن من الناحية الفنية اختيار ألوان معينة في معظم الرسومات البيانية باستخدام خيار اللون col ضمن دالة الرسم، فمثلا يمكن استخدام اللون الرصاصي في الرسم بكتابة col="grey". (استخدام دالة المساعدة للتعرّف على أسماء الألوان المتوفرة).

- **تمثيل الساق والورقة:**

تمثيل بياني آخر مناظر للمدرج التكراري، وهو **شكل الساق والورقة** (Stem-leaf plot)، والذي يَستخدم قيم البيانات نفسها داخل الرسم لعرض توزيع البيانات، وهو يُستخدم عادة مع البيانات ذات الأحجام المتوسطة والصغيرة. ودالة الرسم هي stem ويمكن استخدامها، على سبيل المثال مع المتغير grd1 بالصورة التالية:

```
> stem(stu.data1$grd1)

  The decimal point is 1 digit(s) to the right of the |

  3 | 5
  4 | 059
  5 | 015
  6 | 0345689
  7 | 1335557779
  8 | 0246889
  9 | 0034
```

[1] سيتم تناول دالة الرسم plot في بعض استخدامات الرسم الإضافية الأخرى لاحقا.

ولاحظ أن هذا التمثيل البياني لم يظهر في نافذة فرعية، بل ظهر كنتيجة اعتيادية في لوحة مراقبة برنامج R. ويتم النظر إلى القيم التي تظهر في الجانب الأيمن على أنها تَركّز قيم المشاهدات بشكل أعمدة "أفقية"، أو كأنه مدرج تكراري معروض بشكل أفقي. (ويمكن من الناحية الشكلية استخدام الخيار scale للتحكم بعدد الأعمدة الأفقية).

وتوجد صيغة أخرى لشكل الساق والورقة هي stem.leaf، إلا أنها تتطلب وجود الحزمة الإضافية[1] aplpack، والتي سنقوم بتحميلها من داخل لوحة مراقبة R باستخدام الأمر install.packages() واتباع الخطوات التي تم شرحها في الفصل الأول (البند (4.1))، وبعد اكتمال التحميل نقوم باستدعاء الحزمة المطلوبة؛

```
> library(aplpack)
Loading required package: tcltk
```

فنلاحظ ظهور رسالة تفيد بضرورة وجود الحزمة الإضافية tcltk، لذلك نقوم بتحميلها هي الأخرى، وبعد اكتمال التحميل يتم استدعاء الحزمة aplpack من جديد وتنفيذ الدالة stem.leaf؛

```
> library(aplpack)
> stem.leaf(stu.data1$grd1)
```

وسوف تلاحظ وجود اختلاف عن نتيجة استخدام الدالة stem، حيث أن الدالة stem.leaf تعطي تفصيلات إضافية.

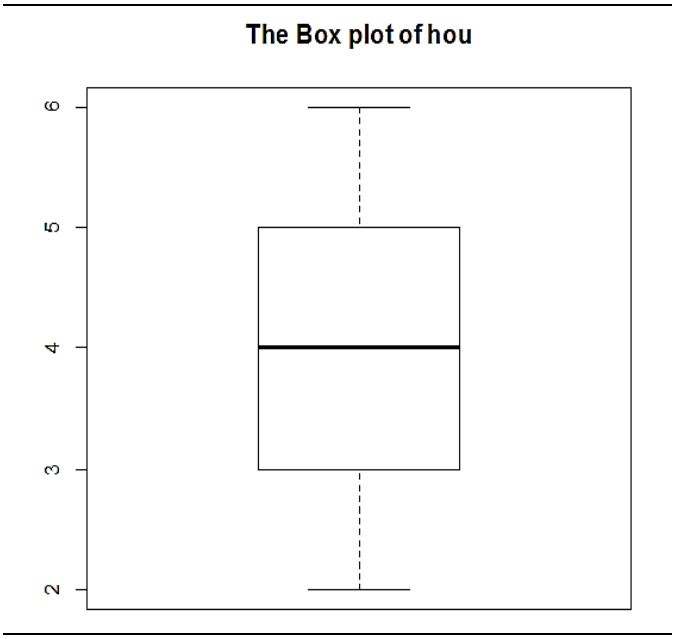
- **تمثيل الصندوق:**

نأتي الآن لتمثيل **شكل الصندوق** (Boxplot)، والذي يُعتبر من أهم الرسومات البيانية في التحليل الاستكشافي للبيانات الأحادية، وهو عبارة عن تمثيل بياني لمقياس الأرقام الخمسة. وتُعرّف دالة الرسم الخاصة به بالصيغة boxplot. وحيث أن الرسم الافتراضي لهذه الدالة لا يحتوي على عنوان أو أية مسميات، فسيتم إدراج عنوان للرسم واسم المتغير المطلوب، (وليكن المتغير hou مثلا في stu.data1)، وتنفيذ تلك الدالة بالصورة التالية:

```
> boxplot(stu.data1$hou,main="The Box plot of hou")
```

فيظهر لدينا رسم الصندوق، (الشكل (6.4))، حيث يضم "الصندوق" في الشكل معظم قيم المشاهدات التي تنحصر بين قيمتي الربيع الأول والثالث، والخطين المتصلين به من الأسفل والأعلى يمثلان القيمة الصغرى والكبرى على الترتيب.

[1] بعض الحِزم الإضافية، مثل الحزمة aplpack تتطلب تحميل (أو وجود) حِزم إضافية ثانوية أخرى لابد من تحميلها أيضا.



شكل 6.4: رسم الصندوق للمتغير hou في البيانات stu.data1

ويمكن من الناحية الشكلية عرض شكل الصندوق بشكل أفقي إذا تم إضافة خيار الرسم horizontal=T لدالة رسم الصندوق.

2.2.4 التمثيل البياني للبيانات الأحادية النوعية

(Graphical Display for Qualitative Univariate Data)

إن استخدام الدوال الخاصة بالتمثيل البياني للبيانات النوعية يتطلب بصورة عامة استخدام بعض الدوال الإضافية والخيارات المرتبطة بها، كما سنلاحظ لاحقا، بهدف الحصول على التمثيل الملائم الذي يصف تلك البيانات بشكل جيد.

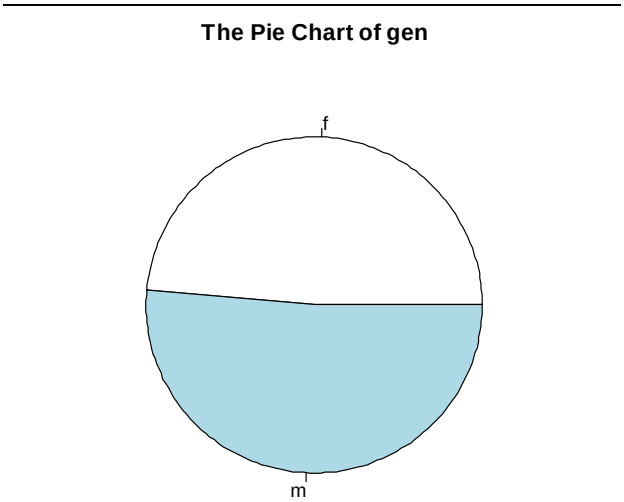
- **تمثيل القطاعات الدائرية:**

لنبدأ بدالة القطاعات الدائرية (Pie Chart) التي تصف نسب المستويات (التقسيمات) للمشاهدات. وإذا لم تتوفر البيانات بشكل جدول توزيع تكراري، كما هو الحال في معظم بيانات الدراسة عادة، فإن استخدام هذه الدالة، والتي تُعرّف بالصيغة Pie، يستدعي[1] استخدام دالة إضافية هي دالة **الجَدولة** table كما يوضح المثال التالي الذي سيُستخدم فيه المتغير النوعي gen في البيانات stu.data1:

```
> pie(table(stu.data1$gen),main="The Pie Chart of gen")
```

[1] تم استخدام الدالة table هنا هو لسببين؛ الأول هو أن المتغير gen مُعرّف بقيم غير رقمية، والسبب الثاني هو لتلخيص تكرارات القيم في مستويات، (لتوضيح الفكرة جرّب مثلا تنفيذ الأمر؛ pie(stu.data1$sem) ثم جرب تنفيذ؛ pie(table(stu.data1$sem)).

فنحصل على الشكل (7.4).



شكل 7.4: القطاعات الدائرية للمتغير gen في stu.data1

- **تمثيل الأعمدة البيانية:**

تُستخدم الأعمدة البيانية (Bar Charts) هي الأخرى لتمثيل البيانات النوعية، وتُشبه في الشكل العام المدرجات التكرارية رغم الاختلاف في نوع البيانات المُستخدم، وتُعرّف بالدالة barplot، ويتم إضافة دالة الجدولة table إلى سطر الأوامر للحصول على الرسم المطلوب.

ويمكن أيضا استخدام الدالة prop.table كدالة إضافية مع barplot للحصول على التكرارات النسبية في المحور العامودي، ويمكن تنفيذ الشكلين باستخدام متسلسلة الأوامر، (للمتغير sem مثلا في البيانات stu.data1)، بالصورة التالية:

```
> par(mfrow=c(1,2))
> barplot(table(stu.data1$sem),main="The Bar Chart of
sem",xlab="Semester",ylab="Frequency")
> barplot(prop.table(table(stu.data1$sem)),main="The
Bar Chart of sem",xlab="Semester",ylab="Relative
Frequency")
```

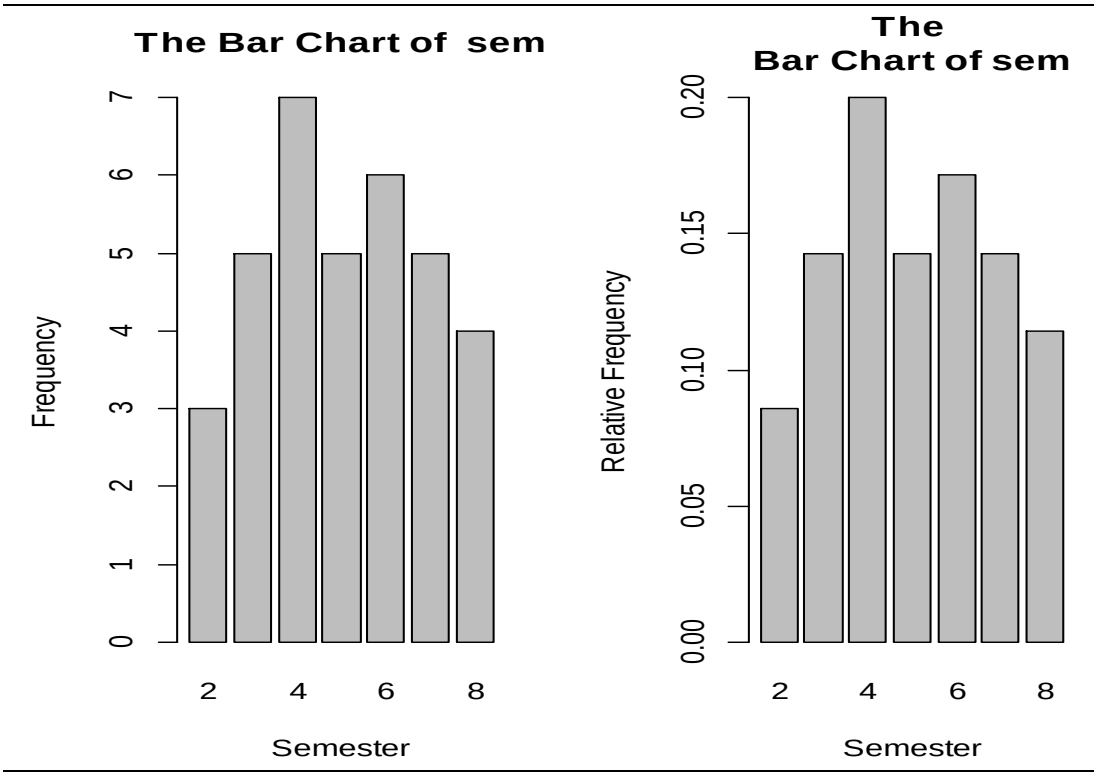
ويُشاهَد في الشكل (8.4) إلى اليسار الأعمدة البانية بالتكرارات، وإلى اليمين الأعمدة البيانية بالتكرارات النسبية. ولاحظ أن اللون الافتراضي للأعمدة هو اللون الرصاصي، (ويمكنك إضافة الخيار col="white" لأحد الأوامر السابقة لترى تغير لون الأعمدة إلى اللون الأبيض).

ويمكن للقارئ أن ينفّذ دالة الأعمدة البيانية، مع المتغير sem مثلا، بدون إضافة الدالة table لمراقبة الفرق في طريقة عرض الأعمدة؛

```
> barplot(stu.data1$sem)
```

أو يمكنه تنفيذها مع المتغير gen مثلا لرؤية رسالة الخطأ الخاصة باستخدام متغير يأخذ قيم غير رقمية؛

```
> barplot(stu.data1$gen)
Error in -0.01*height :non-numeric argument to binary operator
```



شكل 8.4: الأعمدة البيانية للمتغير sem في stu.data1

- **تمثيل الجداول التكرارية:**

لنفرض الآن أنه توفرت لدينا بيانات أحادية مُبوبة، (لمتغير كمي أو نوعي)، على هيئة جدول توزيع تكراري مثل تلك البيانات في الجدول (2.4)، فإنه يمكن أيضا استخدام دوال القطاعات الدائرية والأعمدة البيانية لتمثيل ذلك المتغير بصورة مباشرة وبدون استخدام دالة الجدولة.

جدول 2.4: توزيع أعداد مرضى بحسب حالة المريض في أحد الأقسام في مستشفى خاص

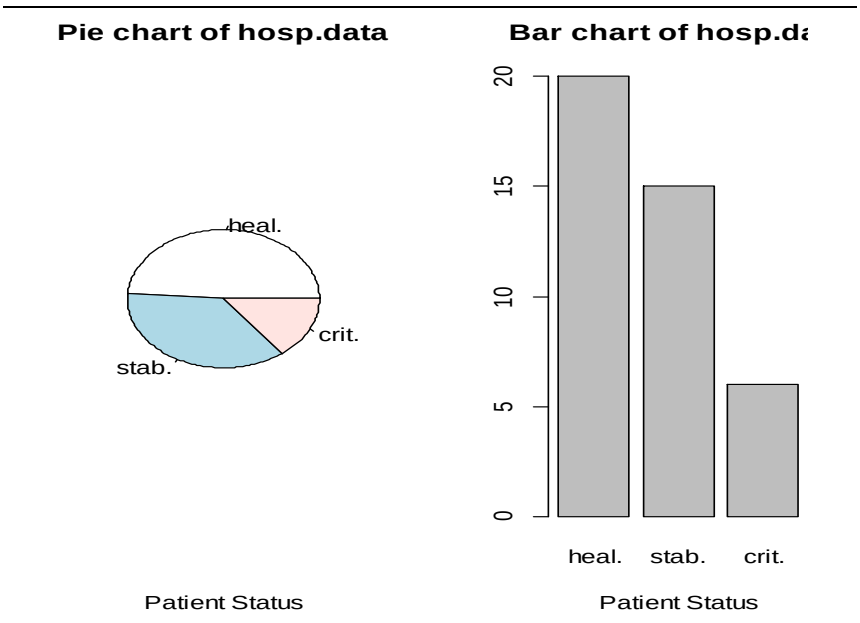
حالة المريض	يتماثل للشفاء (heal.)	مستقرة (stab.)	حرجة (crit.)
أعداد المرضى	20	15	6

ويمكن إدخال بيانات الجدول السابق، والتي تمثل أعداد (تكرارات) مرضى أحد الأقسام في مستشفى خاص موزعة بحسب حالة المريض، كمتجه عددي، (رغم أن المتغير "الأصلي" والذي يمثل مستويات حالة المريض هو متغير نوعي اسمي)، وإدخال مستويات المتغير كأسماء لذلك المتجه العددي بالصورة المُبسطة التالية:

```
> hosp.data<-c(20,15,6)
> names(hosp.data)<-c("heal.","stab.","crit.")
> hosp.data

heal. stab. crit.
   20    15     6

> class(hosp.data)
[1] "numeric"
```

ولاحظ كيف أن البيانات hosp.data تأخذ نفس هيئة الجدول (2.4) رغم أن طبيعتها في R هي متغير عددي. الآن يمكن استخدام دوال القطاعات والأعمدة البيانية لتمثيل هذه البيانات، (الشكل (9.4))، بصورة مباشرة كالتالي:

```
> par(mfrow=c(1,2))
> pie(hosp.data,main="Pie chart of hosp.data",xlab=
"Patient Status")
> barplot(hosp.data,main="Bar chart of hosp.data",xlab=
"Patient Status")
```



شكل 9.4: القطاعات والأعمدة البيانية للبيانات hosp.data

3.4 التحليل الاستكشافي للبيانات المتعددة (EDA for Multivariate Data)

سنستكمل في هذا البند تناول المقاييس والرسوم البيانية للتحليل الاستكشافي للبيانات، حيث سنتعامل هنا مع البيانات المتعددة باستخدام لغة R. ويشمل هذا التعامل؛ استكشاف التوزيعات الثنائية ضمن المتغيرات المتعددة مع إنشاء الرسومات البيانية الخاصة بها، وتصنيف المتغيرات الكمية اعتمادا على المتغيرات النوعية، وتكوين الجداول ذات الاتجاهين، وغير ذلك من الطرق الوصفية، دون التطرق للأساليب التي تستخدم طرق الاستدلال الإحصائي والتي ستتم مناقشتها ضمن موضوعات الفصول القادمة.

1.3.4 التعامل مع متغيرات التقسيم (Dealing with Grouping Variables)

تناولنا في ما سبق استخدام دالة summary مع البيانات الأحادية، والتي نحصل من خلالها على أهم المقاييس الإحصائية الوصفية لكل متغير على حده، وفي هذا البند سيتم توضيح كيفية حساب المقاييس الإحصائية للمتغيرات باستخدام متغير تقسيم أو تصنيف إلى مجموعات، والذي غالبا ما يكون متغيرا نوعيا. وهذا النوع من التحليل الاستكشافي يكون مفيدا عندما يرغب الباحث في الحصول على معلومات تفصيلية حول متغير أو أكثر بعد تقسيم مفرداته بحسب مستويات متغير نوعي آخر، فمثلا قد نرغب في حساب قيمة الوسيط لدرجات الطلبة الذين هم في الفصل الدراسي السابع، أو حساب مجموع إنتاج النفط في حقول محددة، وهكذا.

وتوجد أكثر من دالة في لغة R يمكن استخدامها لهذا الغرض؛ أي حساب المقاييس لمتغير كمي، (أو حتى حساب قيم دوال رياضية أخرى)، اعتمادا على قيم متغير تصنيف آخر في البيانات، وسنتناول أهمها في هذا البند.

- **استخدام دالة** by:

نبدأ بالدالة by التي تُستخدم للحصول على ملخص لمقاييس دالة summary اعتمادا على قيمة متغير آخر نوعي أو كمي، فعلى سبيل المثال، في البيانات stu.data1، يمكن الحصول على ملخص لكل متغير بحسب جنس الطالب، (مُمثلا بالمتغير gen)، كما يلي:

```
> by(stu.data1,stu.data1["gen"],summary)
```

```
gen: f
```

```
      grd1            grd2            grd3            age
 Min.   :51.00   Min.   :55.00   Min.   :33.00   Min.   :19.00
 1st Qu.:77.00   1st Qu.:80.00   1st Qu.:50.00   1st Qu.:20.00
 Median :84.00   Median :85.00   Median :57.00   Median :22.00
 Mean   :81.65   Mean   :83.29   Mean   :60.06   Mean   :21.53
 3rd Qu.:89.00   3rd Qu.:91.00   3rd Qu.:61.00   3rd Qu.:23.00
 Max.   :94.00   Max.   :96.00   Max.   :95.00   Max.   :24.00

    gen                 sem             fam         hou
 Length:17          Min.   :2.000   Min.   :2   Min.   :3
 Class :character   1st Qu.:4.000   1st Qu.:3   1st Qu.:5
```

```
 Mode  :character   Median :4.000   Median :4   Median :5
                    Mean   :4.765   Mean   :4   Mean   :5
                    3rd Qu.:6.000   3rd Qu.:4   3rd Qu.:6
                    Max.   :8.000   Max.   :9   Max.   :6
------------------------------------------------------------
gen: m
      grd1            grd2            grd3            age
 Min.   :35.00   Min.   :40.00   Min.   :27.00   Min.   :19.00
 1st Qu.:51.25   1st Qu.:51.25   1st Qu.:46.00   1st Qu.:22.00
 Median :65.00   Median :69.50   Median :51.00   Median :22.50
 Mean   :61.56   Mean   :63.22   Mean   :50.89   Mean   :22.72
 3rd Qu.:72.50   3rd Qu.:73.00   3rd Qu.:59.75   3rd Qu.:24.00
 Max.   :77.00   Max.   :79.00   Max.   :61.00   Max.   :25.00

     gen                 sem             fam         hou
Length:18          Min.   :2.000   Min.   : 4.000   Min.   :2.000
Class:character    1st Qu.:4.250   1st Qu.: 5.250   1st Qu.:2.250
Mode :character    Median :5.500   Median : 8.000   Median :3.500
                   Mean   :5.333   Mean   : 7.889   Mean   :3.278
                   3rd Qu.:7.000   3rd Qu.:10.000   3rd Qu.:4.000
                   Max.   :8.000   Max.   :12.000    Max.  :5.000
```

ويمكن للقارئ تنفيذ نفس المثال السابق وحساب المقاييس الإحصائية الاستكشافية لنفس المتغيرات الكمية اعتمادا على متغير نوعي آخر، وليكن الفصل الدراسي مثلا، (المتغير sem)، باستخدام الأمر؛

```
.by(stu.data1,stu.data1["sem"],summary)
```

ملاحظة:

يمكن اختصار أسماء المتغيرات في البيانات stu.data1 عن طريق تعيين أسماء مختصرة لها في R تمكننا من كتابة سطور الأوامر بشكل أسرع، ولتكن مثلا على الصورة التالية:

```
> s.grd1<-stu.data1$grd1
> s.grd2<-stu.data1$grd2
> s.grd3<-stu.data1$grd3
> s.age<-stu.data1$age
> s.gen<-stu.data1$gen
> s.sem<-stu.data1$sem
> s.fam<-stu.data1$fam
> s.hou<-stu.data1$hou
```

وهكذا، فإن سطر الأمر الأخير الذي يحتوي دالة by مثلا يمكن كتابته على الصورة التالية؛

```
.by(stu.data1,s.sem,summary)
```

- **استخدام دالة** apply:

نأتي الآن لدالة apply (وتركيباتها المختلفة) والتي تُعد من أهم الدوال التي تتعامل مع البيانات المتعددة، وأكثرها شيوعا في الاستخدام نظرا لما تتمتع به من خيارات أكثر من دالة by، فهي تمكننا من حساب أي مقياس إحصائي أو رياضي للأعمدة (التي قد تكون متغيرات) أو الصفوف (التي قد تكون مشاهدات) وذلك لأي بيانات على شكل إطار بيانات أو مصفوفة، وتوجد لها بعض التركيبات التي تُعطي المستخدِم المرونة اللازمة لإجراء العمليات المطلوبة، ومن ضمنها حساب المقاييس الإحصائية المختلفة بصورة عامة للمتغيرات الكمية اعتمادا على تقسيم معين لمتغير نوعي أو كمي، كما هو الحال مع دالة by.

ومن الأمثلة على ذلك، حساب الوسط الحسابي للمتغيرات في stu.data1، إلا أنه يجب استثناء أي متغير يأخذ قيم غير عددية[1] من البيانات عند تنفيذ هذه الدالة:

```
> apply(stu.data1[-5],2,mean)
```

```
     grd1      grd2      grd3       age       sem       fam       hou
71.314286 72.971429 55.342857 22.142857  5.057143  6.000000  4.114286
```

ولاحظ أن stu.data1[-5] تعني استدعاء كل المتغيرات في هذه البيانات ما عدا المتغير الخامس والذي يمثل النوع. أما القيمة 2 في سطر الأمر فتعني اختيار الأعمدة، أي المتغيرات، (والقيمة 1 تعني اختيار الصفوف، أي المشاهدات).

ويمكن من الناحية الرياضية الشكلية استخدام دالة التدوير round للحصول على عرض أفضل للنتيجة السابقة، فيمكن تدوير القيم لأقرب خانتين بعد الفاصلة بالصورة؛

```
> round(apply(stu.data1[-5],2,mean),2)
```

```
 grd1  grd2  grd3   age   sem   fam   hou
71.31 72.97 55.34 22.14  5.06  6.00  4.11
```

ولحساب الوسط الحسابي[2] للمشاهدات (الصفوف) يتم كتابة:

---

[1] يمكن تغيير القيم غير العددية في أي متغير له الطبيعة العاملية إلى رموز (Codes) باستخدام دالة as.numeric، فمثلا لتغيير القيم في متغير النوع إلى رموز يجب أولا تحويله إلى متغير عاملي بكتابة s.gen<-as.factor(s.gen)، ثم كتابة s.gen2<-as.numeric(s.gen)، بحيث لا يتم المساس بقيم المتغير الأصلي.

[2] لاحظ أنه لا يوجد معنى منطقي لحساب الوسط الحسابي أو أي مقياس آخر لصفوف (مشاهدات) البيانات stu.data1 نظرا لاختلاف وحدات قياس المتغيرات، وقد تم ذلك فقط لتوضيح التطبيق حسابيا.

```
> round(apply(stu.data1[-5],1,mean),2)

 stu1  stu2  stu3  stu4  stu5  stu6  stu7  stu8  stu9 stu10
28.86 27.29 28.86 31.86 22.86 32.00 33.57 39.57 38.86 35.29

stu11 stu12 stu13 stu14 stu15 stu16 stu17 stu18 stu19 stu20
29.43 35.86 32.71 33.14 36.00 27.14 26.29 34.57 35.86 30.57

stu21 stu22 stu23 stu24 stu25 stu26 stu27 stu28 stu29 stu30
36.86 37.71 35.71 39.14 31.71 38.29 38.43 31.29 41.14 31.57

stu31 stu32 stu33 stu34 stu35
41.00 32.71 35.57 34.14 38.86
```

وعلى هذا المنوال، يمكن استخدام أي دالة حسابية عوضا عن الوسط الحسابي ضمن دالة apply وتطبيقها على الأعمدة أو الصفوف، ويمكن للقارئ في هذا الموضع تطبيق أي من المقاييس الإحصائية المذكورة في الجدول (1.4) السابق مع متغيرات (أو مشاهدات) البيانات stu.data1 ومراقبة النتائج.

- **استخدام الدالتين** lapply **و**sapply:

تعد الدالتان lapply وsapply تركيبتان إضافيتان من الدالة apply ويعطيان نتائج مشابهة لها، فالدالة lapply تقوم بحساب المقياس لأعمدة البيانات وإعطاء النتيجة على هيئة قائمة وهذا سبب وجود حرف "l" من كلمة "list"، فمثلا يمكن حساب الانحراف المعياري لمتغيرات stu.data1 بالصورة:

```
> lapply(stu.data1[-5],sd)

$grd1
[1] 15.605
$grd2
[1] 15.21896
$grd3
[1] 14.28892
$age
[1] 1.647509
$sem
[1] 1.846186
$fam
[1] 3.058258
$hou
[1] 1.278129
```

والدالة sapply تُعطي نفس النتيجة ولكنها تقوم "بتبسيط" عرض ناتج عملية الحساب، وهذا سبب وجود حرف "s" من كلمة "simple"، فالنتيجة السابقة ستظهر كالتالي:

```
> sapply(stu.data1[-5],sd)
```

```
     grd1      grd2      grd3       age       sem       fam       hou
15.604998 15.218962 14.288916  1.647509  1.846186  3.058258  1.278129
```

- **استخدام دالة** tapply:

نتناول الآن التركيبة الأهم من تركيبات دالة apply وهي الدالة tapply. هذه الدالة تمكن المستخدم أيضا من حساب المقاييس المختلفة للمتغيرات الكمية اعتمادا على تصنيف معين لمتغير نوعي أو كمي، وحرف "t" فيها للدلالة على أن النتيجة ستأخذ هيئة الجدول "table". ولنأخذ الأمثلة التالية باستخدام البيانات stu.data1؛

- حساب الوسيط لمتغير العمر s.age بناء على متغير النوع s.gen:

```
> tapply(s.age,s.gen,median)
```

```
   f    m
22.0 22.5
```

وهذا يعني أن القيمة الوسيطية لأعمار الطلبة الذكور هي 22.5 سنة، وللطالبات هي 22 سنة.

- إيجاد توزيع أعداد الطلبة (التكرارات) بحسب تقسيم النوع والفصل الدراسي s.sem، ولاحظ كيفية إدراج متغيرات التصنيف[1] داخل دالة list:

```
> tapply(s.age,list(s.gen,s.sem),length)
```

```
  2 3 4 5 6 7 8
f 1 3 6 1 3 1 2
m 2 2 1 4 3 4 2
```

ونلفت الانتباه هنا أنك لو قمت باستبدال المتغير s.age بأي متغير آخر فإنك ستحصل على نفس النتيجة لأن الدالة الحسابية المستخدمة هي دالة العَد length والتي تقوم هنا بحساب أعداد الطلبة اعتمادا على تقسيمات ترتيب الفصل الدراسي (سبعة تقسيمات) وتقسيم جنس الطالب (تقسيمين)، فمثلا في العامود الأول إلى اليسار؛ يُلاحظ في الفصل الدراسي الثاني وجود طالبة واحدة وطالبان من الذكور، وهكذا تتوزع باقي التكرارات بحيث يكون مجموعها مساويا لعدد المشاهدات أو حجم العينة وهو 35 طالب.

- حساب متوسط أعمار الطلبة بناء على تصنيف النوع والفصل الدراسي:

```
> round(tapply(s.age,list(s.gen,s.sem),mean),2)
```

```
     2     3    4     5     6     7    8
f 20.0 21.33 21.5 23.00 22.33 19.00 22.0
m 23.5 22.00 22.0 22.75 22.33 23.75 21.5
```

[1] إذا ما تم استخدام متغير نوعي ثالث ضمن دالة list، فإننا سنحصل على عدد من جداول النتائج (ثنائية الأبعاد) مساو لعدد تقسيمات هذا المتغير الثالث.

والقيم في هذا الجدول لهذا المثال تمثل متوسطات أعمار بحسب التقسيمات المحددة، فمثلا القيمة 21.5 في العامود الثالث من اليسار تمثل متوسط أعمار الطلبة الستة (بالتوافق مع تكرارات الطلبة في جدول المثال السابق) اللذين هم في الفصل الدراسي الرابع وهن من الإناث.

- **استخدام الدالة** aggregate:

في بعض الحالات، قد نحتاج إلى التعامل مع أو جدولة أكثر من متغير كمي في نفس الوقت، أو قد نرغب في استخدام مقياس أو دالة حسابية تُعطي نواتج متعددة، وهذا لا يمكن تنفيذه بدالة tapply، لذلك يمكن عندئذ استخدام الدالة aggregate لتنفيذ ذلك.

لحساب مقياس ما لأكثر من متغير كمي مع استخدام متغير تصنيف يمكن استخدام الدالة aggregate مع دالة تجميع الأعمدة cbind بالصورة التي يوضحها المثال التالي، (للبيانات stu.data1):

```
> aggregate(cbind(s.fam,s.hou)~s.gen,stu.data1,sd)

  s.gen    s.fam     s.hou
1     f 1.936492 0.9354143
2     m 2.720054 0.9582800
```

في هذا المثال تم حساب الانحراف المعياري لكل من عدد أفراد أسرة الطالب s.fam وعدد غرف منزل الطالب s.hou بناء على النوع. ولاحظ أن تركيبة دالة aggregate تعتمد على إدراج المتغير أو المتغيرات الكمية في البداية، ثم إدراج متغير التصنيف بعد العلامة "~"، يلي ذلك إدراج اسم البيانات التي تنتمي لها هذه المتغيرات، وأخيرا تعريف المقياس أو دالة العملية الحسابية المطلوب تنفيذها.

وفي نتيجة هذا المثال، القيمة (2.720054) مثلا تعني أن الانحراف المعياري لعدد أفراد الأسرة للطلبة الذكور هو ثلاثة أفراد تقريبا، والقيمة (0.9354143) تعني أن الانحراف المعياري لعدد غرف المنزل للطالبات هو غرفة واحدة تقريبا، وهكذا.

ويمكن للقارئ في هذا المثال استبدال متغير التصنيف s.gen بالمتغير s.sem مثلا، وكذلك تغيير مقياس الانحراف المعياري sd بمقاييس أخرى وملاحظة التغير في النتائج.

ويمكن أيضا باستخدام الدالة aggregate لحساب مقياس ما لأكثر من متغير كمي بناء على أكثر من متغير تصنيف كما يتضح من المثال التالي لنفس البيانات stu.data1:

```
> aggregate(cbind(s.fam,s.hou)~s.gen+s.sem,stu.data1,mean)

   s.gen s.sem     s.fam    s.hou
1      f     2  5.000000 5.000000
2      m     2  8.000000 3.000000
3      f     3  4.666667 4.666667
4      m     3  8.000000 3.000000
5      f     4  3.166667 5.333333
6      m     4  9.000000 3.000000
7      f     5  2.000000 6.000000
8      m     5  6.000000 4.000000
9      f     6  5.666667 4.666667
10     m     6  5.000000 4.333333
11     f     7  3.000000 5.000000
12     m     7  9.750000 2.750000
13     f     8  4.000000 4.500000
14     m     8 11.500000 2.000000
```

ولاحظ أن التركيبة العامة الدالة لم تتغير باستثناء إضافة متغير التصنيف الثاني باستخدام إشارة الجمع "+".

ويمكنك أيضا استخدام متغير كمي واحد مع أكثر من متغير تصنيف في نفس الوقت، (وفي هذه الحالة يتم الاستغناء عن دالة cbind وكتابة اسم المتغير الكمي قبل علامة "~". (جرّب ذلك بكتابة s.fam بدلا من cbind(s.fam,s.hou) في سطر الأمر الأخير)).

ذكرنا في بداية الحديث عن الدالة aggregate أنه من ضمن مميزاتها إمكانية استخدامها لحساب مقياس يُعطي نتائج متعددة، فمثلا يمكن حساب القيمتين الصغرى والكبرى (بدالة المدى[1] (range)) لمتغيرين كميين بناء على متغير تصنيف بالصورة:

```
> aggregate(cbind(s.fam,s.hou)~s.gen,stu.data1,range)

  s.gen s.fam.1 s.fam.2 s.hou.1 s.hou.2
1     f       2       9       3       6
2     m       4      12       2       5
```

ويُلاحظ في هذه النتيجة ظهور ترقيم للمتغيرات الكمية s.fam وs.hou متوافق مع عدد قيم نواتج الدالة range، فمثلا القيمتين 2 و9 تحت s.fam.1 وs.fam.2 على الترتيب هما القيمتين الصغرى

[1] تُستخدم دالة المدى أيضا للحصول على القيمتين الصغرى والكبرى لمجموعة من القيم بالصورة التالية مثلا:

```
> range(2,-1,5,0)
[1] -1  5
```

والكبرى لعدد أفراد أسر الطالبات. (ويمكنك تجربة استخدام الدالة quantile بدلا من الدالة range في سطر الأمر الأخير).

نود الإشارة هنا إلى أن كل هذه الجداول الناتجة عن استخدام المقاييس الإحصائية أو العمليات الحسابية مع دالتي tapply وaggregate يمكن تعيينها إلى مسميات (أشياء) بحيث يمكن استخدامها في عمليات حسابية أخرى لاحقا، فمثلا، (للبيانات stu.data1)، يمكن إنشاء جدول يضم الأوساط الحسابية لدرجات الطلبة في المقررات الثلاثة بناء على تقسيم النوع، وتعيينه باسم grd.gen بالصورة التالية:

```
> grd.gen<-
aggregate(cbind(s.grd1,s.grd2,s.grd3)~s.gen,stu.data1,mean)

> grd.gen

  s.gen   s.grd1   s.grd2   s.grd3
1     f 81.64706 83.29412 60.05882
2     m 61.55556 63.22222 50.88889

> class(grd.gen)
[1] "data.frame"
```

وتأخذ الجداول المكوَنة باستخدام الدالة aggregate عادة طبيعة أُطر البيانات، كما هو ملاحظ أعلاه، أما جداول دالة tapply فتأخذ طبيعة المصفوفة.

2.3.4 تكوين جداول البيانات في اتجاهين (Constructing two-way Data Tables)

تناولنا في ما سبق التعامل مع جداول البيانات الأحادية النوعية، (الجدول (2.4))، وتم توضيح كيفية إدخالها في نظام R، وسنقوم في هذا البند بالتوسع قليلا في هذه الجزئية وشرح طريقة إدخال جداول البيانات ثنائية الاتجاه[1]، وذلك في حالتين؛ الحالة الأولى هي عند توفر البيانات من مصدرها على هيئة جدول في اتجاهين، (كما هو الحال في الجدول (3.4) أدناه أو الجداول الناتجة عن استخدام دالة tapply أو aggregate مثلا)، والحالة الثانية هي عند وجود البيانات في شكلها التقليدي أو "الخام".

- **الحالة الأولى:**

توجد عدة طرق لتكوين الجداول في اتجاهين في نظام R، سنختار منها الطريقة التالية، والتي سيتم تطبيقها على البيانات الموجودة في جدول (3.4) التي تمثل عينة من الأشخاص تتوزع أعدادهم بحسب فئاتهم العمرية، (الاتجاه الأول العامودي في الجدول)، وبحسب برامجهم المفضلة في التلفاز، (الاتجاه الثاني الأفقي في الجدول)؛

[1] تُعرف هذه الجداول أيضا **بجداول الاقتران** (Contingency Tables)، وسيتم التعرض لها عند تناول اختبارات الفروض في الفصول القادمة.

جدول 3.4: توزيع الفئات العمرية لعينة من الأشخاص بحسب البرامج المفضلة لهم

		الفئة العمرية (age.g)		
		أطفال (chd.)	بالغون (adt.)	مسنون (sen.)
البرامج المفضلة (fav.prog)	رياضة (sport)	12	20	28
	أخبار (news)	6	15	30
	ترفيه (ent.)	24	20	8
	ديني (rel.)	7	18	10

يتم أولا إدخال القيم العددية الموجودة في الجدول كمصفوفة، ولتكن باسم tv.data مثلا؛

```
> tv.data<-matrix(data=c(12,20,28,6,15,30,24,20,8,7,18
,10),nrow=4,ncol=3,byrow=T)

> tv.data

     [,1] [,2] [,3]
[1,]   12   20   28
[2,]    6   15   30
[3,]   24   20    8
[4,]    7   18   10
```

ولاحظ أن إدخال الأعداد تم بالصف وليس بالعامود، ولهذا تم استخدام الخيار byrow=T. بعد ذلك، يتم استخدام دالة **تسمية الأبعاد** dimnames لإدراج أسماء الأعمدة أولا ثم الصفوف ثانيا عن طريق استخدام دالة القائمة list كالتالي:

```
> dimnames(tv.data)<-
list(fav.prog=c("sport","news","ent.","rel."), age.g=
c("chd.","adt.","sen."))

> tv.data

        age.g
fav.prog chd. adt. sen.
   sport   12   20   28
   news     6   15   30
   ent.    24   20    8
   rel.     7   18   10
```

وهكذا يمكنك ملاحظة مدى تشابه شكل المصفوفة tv.data مع الجدول (3.4) أعلاه.

- **الحالة الثانية:**

لنفرض وجود بيانات تمثل تقديرات طلبة في مقررين هما الإحصاء (.stat) والرياضيات (.math)، كما هو موضح في الجدول (4.4)؛

جدول 4.4: تقديرات عشرون طالبا في مقرري الإحصاء والرياضيات

المشاهدات	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
الإحصاء (.stat)	B	C	A	C	F	D	B	D	D	C	A	C	B	B	F	B	C	A	C	B
الرياضيات (.math)	C	C	B	B	F	D	C	D	D	F	B	C	C	C	F	C	D	B	D	C

ونقوم بإدخال هذه البيانات، إما بصورة مباشرة في نظام R، (باستخدام دالة المصفوفة matrix، أو دالة المتجه c()، أو دالة تعديل أُطر البيانات (edit(data.frame())، أو استيراد تلك البيانات كملف اكسل إذا توفرت بتلك الصيغة.

- **استخدام دالة** table margin.table:

لنفرض أن البيانات قد تم إدخالها باسم exc.data3. يتم بعد ذلك استخدام دالة table لتكوين جدول ذو اتجاهين، (وليكن باسم stu.data2)، كالتالي:

```
> stu.data2<-table(exc.data3)

> stu.data2

     math.
stat. B C D F
    A 3 0 0 0
    B 0 6 0 0
    C 1 1 3 1
    D 0 0 3 0
    F 0 0 0 2
```

- **استخدام الدالتين** margin.table **و**addmargins:

كخطوة إضافية، يمكن تكوين التوزيعات الهامشية للجدول ذو اتجاهين باستخدام دالة **التوزيع الهامشي** margin.table، فيتم إنشاء التوزيع الهامشي للصفوف أو الأعمدة. ولنقم بتنفيذ التوزيعات الهامشية للجدول tv.data على سبيل المثال حيث نبدأ بالتوزيع الهامشي للصفوف:

```
> margin.table(tv.data,1)

fav.prog
sport  news  ent.  rel.
   60    51    52    35
```

ولاحظ أنه قد تم استخدام القيمة 1 كخيار لحساب مجموع الصفوف، وبالمثل يمكن لحساب التوزيع الهامشي للأعمدة باستخدام خيار القيمة 2 كالتالي:

```
> margin.table(tv.data,2)

age.g
chd. adt. sen.
  49   73   76
```

وكذلك يمكن حساب مجاميع الصفوف والأعمدة باستخدام الدالة addmargins مباشرة كالتالي:

```
> addmargins(tv.data)

        age.g
fav.prog chd. adt. sen. Sum
   sport   12   20   28  60
   news     6   15   30  51
   ent.    24   20    8  52
   rel.     7   18   10  35
   Sum     49   73   76 198
```

وكمثال آخر على تكوين الجداول الثنائية؛ لنفرض أنه تم اختيار عينة مكونة من 45 مراهق لدراسة طبيعة تعلّقهم بالألعاب الإلكترونية، حيث تم سؤالهم عن نوع جهاز اللعب المفضل لديهم، (يمثله المتغير "Devi" في الاتجاه الأول)، وهل يُفضل اللعب بمفرده أم مع الأصدقاء، (يمثله المتغير "Pref" في الاتجاه الثاني)، فكانت الإجابات كما هو موضح في الجدول (5.4)، حيث تم استخدام الاختصار "alo" للدلالة على تفضيل المراهق اللعب بمفرده، والاختصار "frd" للدلالة على تفضيل اللعب مع الأصدقاء.

بعد إدخال هذه البيانات في نظام R، (أو استيرادها من اكسل)، باسم teen.age مثلا، نقوم بتحويل المتغيرين "Devi" و"Pref" إلى متغيرات أو متجهات عاملية (Factors) قبل دمجها في جدول ذو اتجاهين بالصورة التالية:

```
> Devi<-factor(teen.age$Devi)
> Pref<-factor(teen.age$Pref)

> class(Devi);class(Pref)
[1] "factor"
[1] "factor"
```

بعد ذلك يتم تكوين جدول الاقتران باستخدام دالة table، وليكن باسم teen.aget مثلا، بالصورة التالية:

```
> teen.aget<-table(Pref,Devi)

> teen.aget

     Devi
Pref  PC PS4 XBOX
  alo  6   5    8
  frd  5  15    6
```

جدول 5.4: إجابات المراهقين حول نوع جهاز اللعب المفضل وطريقة اللعب المفضلة

	Devi.	Pref.		Devi.	Pref.		Devi.	Pref.
1	PS4	frd	16	PC	alo	31	PC	frd
2	XBOX	alo	17	PS4	frd	32	PS4	alo
3	PC	alo	18	PC	alo	33	PC	frd
4	PS4	frd	19	PS4	frd	34	PC	frd
5	XBOX	alo	20	PS4	frd	35	PS4	alo
6	PS4	frd	21	PS4	frd	36	PC	frd
7	PC	alo	22	XBOX	alo	37	XBOX	frd
8	PS4	frd	23	PS4	frd	38	PC	frd
9	PS4	alo	24	PC	alo	39	XBOX	frd
10	XBOX	alo	25	PS4	frd	40	PS4	alo
11	PS4	frd	26	XBOX	alo	41	XBOX	frd
12	XBOX	alo	27	PS4	frd	42	XBOX	frd
13	PC	alo	28	PS4	frd	43	PS4	frd
14	PS4	frd	29	XBOX	alo	44	XBOX	frd
15	XBOX	alo	30	PS4	alo	45	XBOX	frd

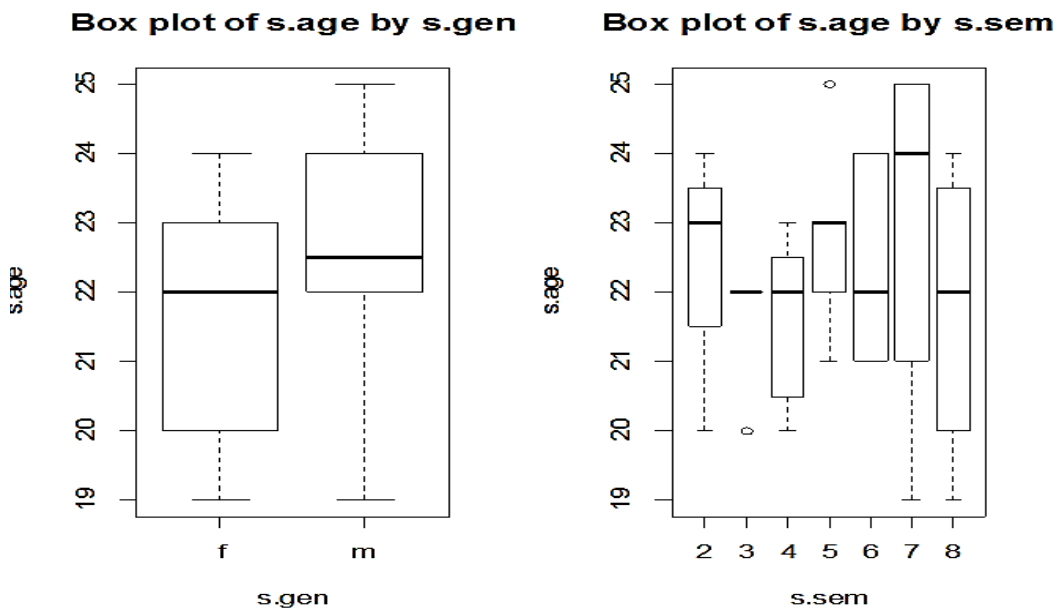
3.3.4 التمثيل البياني للبيانات المتعددة (Graphical Display for Multivariate Data)

- **تمثيل الصندوق لمتغير كمي اعتمادا على متغير تصنيف:**

في سياق التعامل مع البيانات المتعددة، يمكن استخدام رسم الصندوق لتمثيل أحد المتغيرات اعتمادا على قيم متغير آخر، (والذي غالبا ما يكون متغيرا وصفيا). ولنقم على سبيل المثال بتنفيذ رسم الصندوق لمتغير عمر الطالب s.age في البيانات stu.data1 اعتمادا على جنس الطالب s.gen، ومرة أخرى اعتمادا على الفصل الدراسي للطالب s.sem وذلك باستخدام الأوامر التالية:

```
> par(mfrow=c(1,2))
> boxplot(s.age~s.gen,main="Box plot of s.age by s.gen",
xlab="s.gen",ylab="s.age")
> boxplot(s.age~s.sem,main="Box plot of s.age by s.sem",
xlab="s.sem",ylab="s.age")
```

ولاحظ أن المتغير النوعي الذي يتم التقسيم بناء عليه يتم كتابته بعد علامة "~" ضمن الدالة. سنحصل بعد ذلك على رسمين لشكل الصندوق، (شكل (10.4))، يوضحان التغير في عمر الطالب بالنسبة إلى كونه ذكرا أو أنثى (الرسم إلى اليسار)، وكذلك بالنسبة إلى ترتيب الفصل الدراسي للطالب (الرسم إلى اليمين).



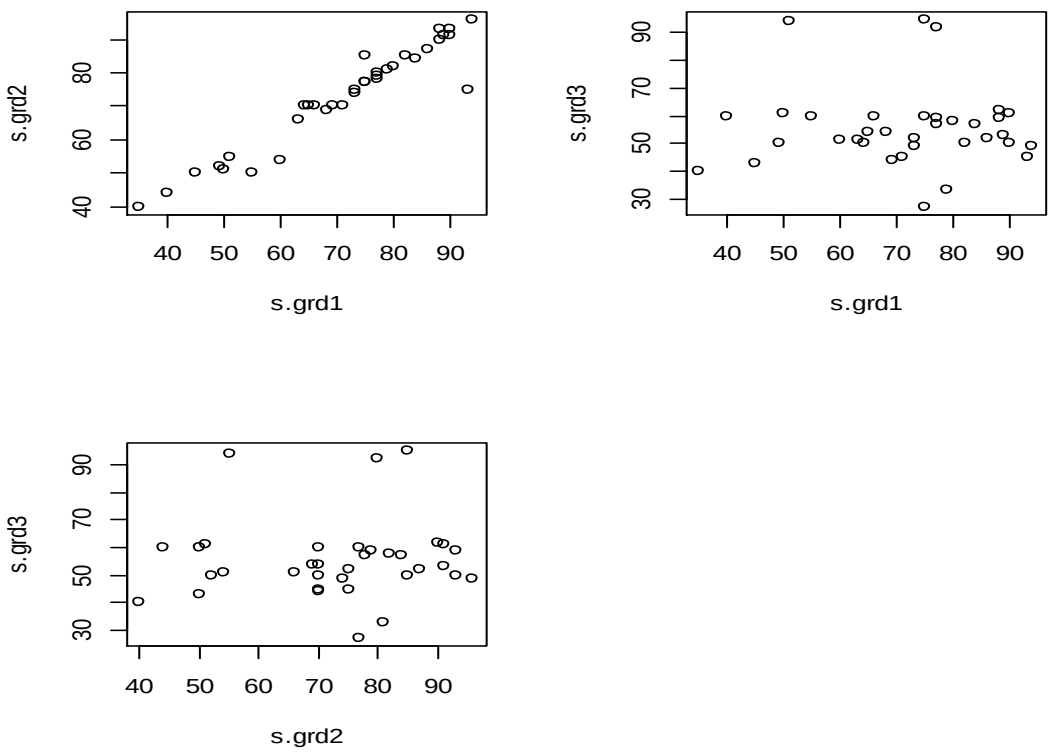
شكل 10.4: رسم الصندوق للمتغير `s.age` اعتمادا على المتغير `s.gen` (إلى اليسار)، واعتمادا على المتغير `s.sem` (إلى اليمين)

- **تمثيل شكل الانتشار:**

يُعد شكل الانتشار (Scatter Plot) من الرسومات الأساسية التي تُستخدم في فهم طبيعة العلاقة بين متغيرين، إذ أنه يُعتبر الخطوة الأولى في دراسة أو تحليل الارتباط والانحدار[1] بين المتغيرات. ولرسم شكل الانتشار بين متغيرين، نستخدم دالة `plot` التي تناولناها سابقا، وكأمثلة على هذا الرسم، لنراقب العلاقة الثنائية بين المتغيرات `s.grd1`، `s.grd2`، و`s.grd3` في ثلاثة رسومات منفصلة داخل إطار واحد، (شكل (11.4))، وذلك بتنفيذ التالي:

```
> par(mfrow=c(2,2))
> plot(s.grd1,s.grd2)
> plot(s.grd1,s.grd3)
> plot(s.grd2,s.grd3)
```

[1] سيتم التعرض لتطبيقات الارتباط والانحدار وحساب المعاملات الخاصة بهما لاحقا في الفصول القادمة.



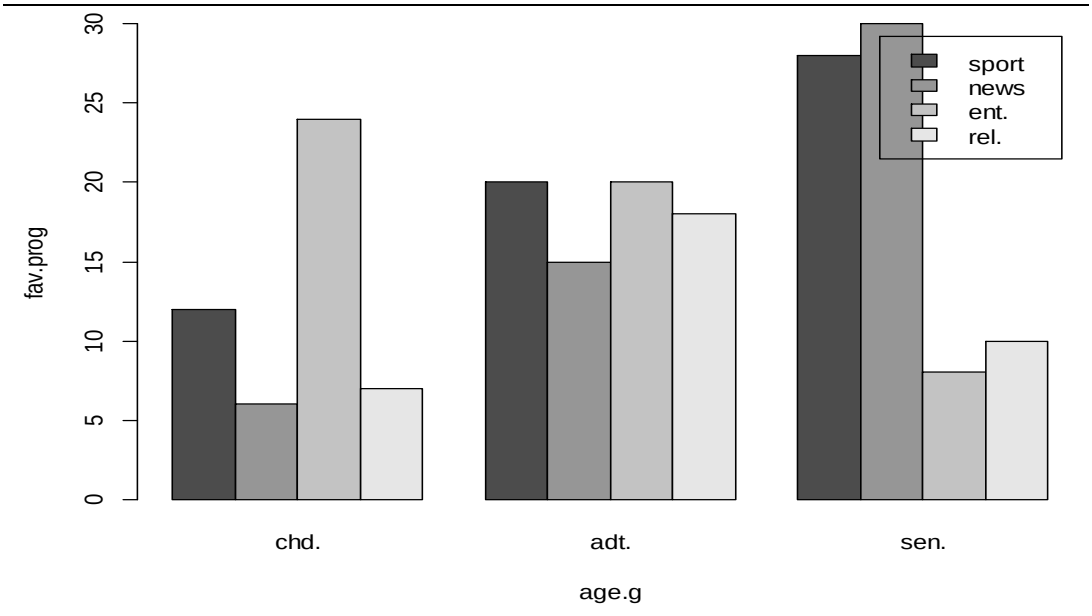
شكل 11.4: أشكال الانتشار للمتغيرات s.grd1، s.grd2، وs.grd3

- **تمثيل الأعمدة البيانية باستخدام متغير تصنيف (لجداول الاقتران):**

من جديد نعود لجداول الاقتران ذات الاتجاهين، حيث سيتم استخدام الأعمدة البيانية لتمثيل تلك الجداول، ولنأخذ البيانات tv.data في الجدول (3.4) كمثال، حيث سيتم رسم الأعمدة البيانية لمتغير العمر بحسب تصنيف البرنامج المفضل؛

```
> barplot(tv.data,xlab="age.g",ylab="fav.prog",beside=T,
legend.text=T)
```

وقد تم استخدام الخيار beside=T في سطر الأمر السابق لعرض مستويات (صفوف) جدول البيانات في أعمدة متجاورة، واستُخدم الخيار legend.text=T لعرض دليل تفسيري لمستويات المتغير الذي يمثل الصفوف fav.prog. كما نرى في الشكل (12.4).



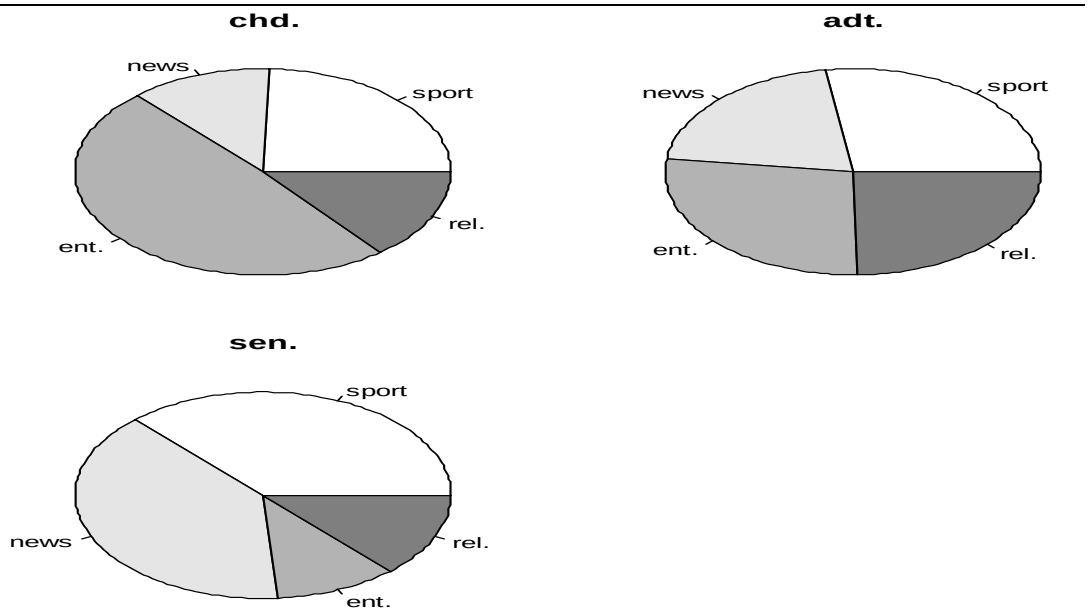
شكل 12.4: الأعمدة البيانية للمتغير age.g بحسب تصنيف المتغير fav.prog للبيانات tv.data

ويمكن للقارئ استخدام الأمر barplot(tv.data) بشكل مبسط للحصول على الأعمدة البيانية بتقسيمات داخل الأعمدة نفسها.

- **تمثيل القطاعات الدائرية باستخدام متغير تصنيف (لجداول الاقتران):**

يمكن استخدام القطاعات الدائرية هي الأخرى للحصول على "وصف" بياني، (الشكل (13.4))، مشابه لذلك الذي حصلنا عليه من الأعمدة البيانية في الشكل السابق. ولنستخدم نفس البيانات tv.data في الجدول (3.4) لتوضيح ذلك التشابه في النتيجة؛

```
> par(mfrow=c(2,2),mex=0.8, mar=c(1,1,2,1))
> slices1 <- c("white","grey90","grey70","grey50")
> pie(tv.data[,"chd."],main="chd.",col=slices1)
> pie(tv.data[,"adt."],main="adt.",col=slices1)
> pie(tv.data[,"sen."],main="sen.",col=slices1)
```



شكل 13.4: القطاعات الدائرية للمتغير age.g بحسب تصنيف المتغير fav.prog للبيانات tv.data

ونشير هنا إلى أنه تم في سطور الأوامر السابقة إجراء بعض التعديلات الإضافية والخاصة بتغييرات في حجم الرسومات (القطاعات) وألوانها بغرض الحصول على مظهر أفضل للتمثيل البياني، وذلك باستخدام خيارات الرسم؛ mex الخاص بتغيير حجم الرسم، وmar الخاص بتغيير تباعد هوامش الرسم، وتم أيضا تعيين متجه يمثل الألوان التي نرغب باستخدامها بالترتيب المذكور، وباستخدام اللون الأبيض وثلاث تدرجات للون الرصاصي، (بإعطاء أرقام للون). ويمكن الرجوع للملحق (2) الخاص بخيارات الرسم للحصول على المزيد من المعلومات.

4.4 التحليل الاستكشافي للبيانات stu.data1: دراسة حالة

(EDA of stu.data1: Case Study)

إن الخطوة الأولى في أي دراسة إحصائية بسيطة كانت أو متقدمة، عادة ما تكون استكشاف ما تمثله أو ما تصفه هذه البيانات، ومعرفة ما يمكن أن تقدمه من معلومات حول مجتمع الدراسة. وتتم هذه الخطوة عادة باستخدام أدوات التحليل الاستكشافي من مقاييس وجداول ورسومات بيانية كما شاهدنا في بنود هذا الفصل.

والآن سيتم التعامل مع البيانات stu.data1، (الجدول (م2.1) في الملحق (1))، كدراسة حالة للوقوف على ما يمكن تنفيذه من دوال برنامج R الخاصة باستكشاف البيانات للحصول على معلومات مفيدة حول مجتمع الطلبة الذي تمثله هذه العينة من البيانات. وننوه هنا أنه لن يتم استخدام كل الطرق والمقاييس الإحصائية الوصفية لهذا الغرض لأن الهدف من هذا الكتاب لا يشمل مناقشة الشرح النظري لهذه الطرق، بل سيكون الغرض هنا هو تقديم مثال فقط عن كيفية استخدام دوال لغة R في استكشاف ووصف البيانات بصورة مبسطة.

البيانات stu.data1 كما وضحنا سابقا، هي عينة من بيانات طلبة جامعيين تحتوي على متغيرات تمثل الدرجات في ثلاثة مقررات (الدرجة من 100)، والفصل الدراسي للطالب، إضافة لبعض المتغيرات الشخصية والاجتماعية الأخرى وهي العمر، النوع، عدد أفراد الأسرة، وعدد الغرف في منزل الطالب. والهدف سيكون استخدام هذه البيانات لوصف حالة ومستوى الطلبة والتي من المفروض أن تعكس حالة الطلبة في العموم، (تحت شروط العينة الجيدة).

ولتنظيم عرض النتائج والتعليقات المرتبطة بها بما يتوافق مع اسلوب السرد في هذا الفصل، سيتم البدء مع طرق الاستكشاف للبيانات الأحادية أولا.

1.4.4 استكشاف متغيرات الدراسة بصورة أحادية (Exploring Data in Univariate Fashion)

يمكن البدء بمقارنة درجات الطلبة في المقررات الثلاثة grd1، grd2، وgrd3 مثلا، وهذا يمكن تنفيذه أولا باستخدام دالة summary كما وضحنا سابقا، وسنقوم هنا بحساب هذه الدالة لدرجات الطلبة فقط، أي للمتغيرات الثلاثة الأولى:

- **مراقبة النزعة المركزية:**

```
> summary(stu.data1[1:3])
```

```
      grd1            grd2            grd3
 Min.   :35.00   Min.   :40.00   Min.   :27.00
 1st Qu.:63.50   1st Qu.:67.50   1st Qu.:49.50
 Median :75.00   Median :75.00   Median :53.00
 Mean   :71.31   Mean   :72.97   Mean   :55.34
 3rd Qu.:83.00   3rd Qu.:84.50   3rd Qu.:60.00
 Max.   :94.00   Max.   :96.00   Max.   :95.00
```

وإذا ما نظرنا إلى الوسط الحسابي للدرجات فإننا نلاحظ أن المقررين grd1 وgrd2 لهما أوساط متقاربة تقترب من تقدير جيد جدا، أما المقرر الثالث grd3 فيختلف وسطه عن المقررين الأولين ويتجه نحو التقدير مقبول. وضمن عائلة الأوساط أيضا، يُلاحظ تساوي الوسيط للمقررين الأولين واختلاف (ارتفاع) قيمته عن المقرر الثالث بصورة كبيرة. هذه النتيجة تعطي انطباعا "مبدئيا" بأن أداء الطلاب في المقررين grd1 و grd2 هو أفضل من أداؤهم في المقرر grd3.

من ناحية أخرى، هذا الارتفاع الطفيف في قيم الوسيط للمقررين الأولين يعكس وجود التواء بسيط إلى اليسار في توزيع درجات الطلاب في هذين المقررين، بمعنى أن الدرجات تتجه إلى اليمين، أي تتجه نحو الدرجات الأعلى، وهذه الملاحظة تتوافق مع قيم الأوساط للدرجات.

وضمن الحديث عن توزيع البيانات، إذا ما نظرنا إلى قيم الربيعات في النتيجة السابقة فإننا نلاحظ أن قيم الربيع الأول هي متقاربة في المقررين الأولين (63.5 و67.5) وهذه القيم تعني أن تقديرات 75% من الطلاب (أي الأكثرية) هي تقريبا جيد في المقررين. كما أن قيم الربيع الثالث للمقررين هي متقاربة جدا (83 و84.5) مما يدل على تقارب توزيع درجات الطلاب في هذين المقررين. أما بالنسبة للمقرر الثالث grd3، فمن الملاحظ أن قيم الربيع الأول والثالث له متدنية كثيرا عن المقررين الأولين، وحيث أن قيمة الربيع الثالث له هي 60، (مما يعني أن 75% من درجات الطلاب في هذا المقرر هي أقل من جيد)، فهذا كله يؤكد أن أداء الطلاب في هذا المقرر كان تحت المستوى المُرضي مقارنة بالمقررين grd1 وgrd2.

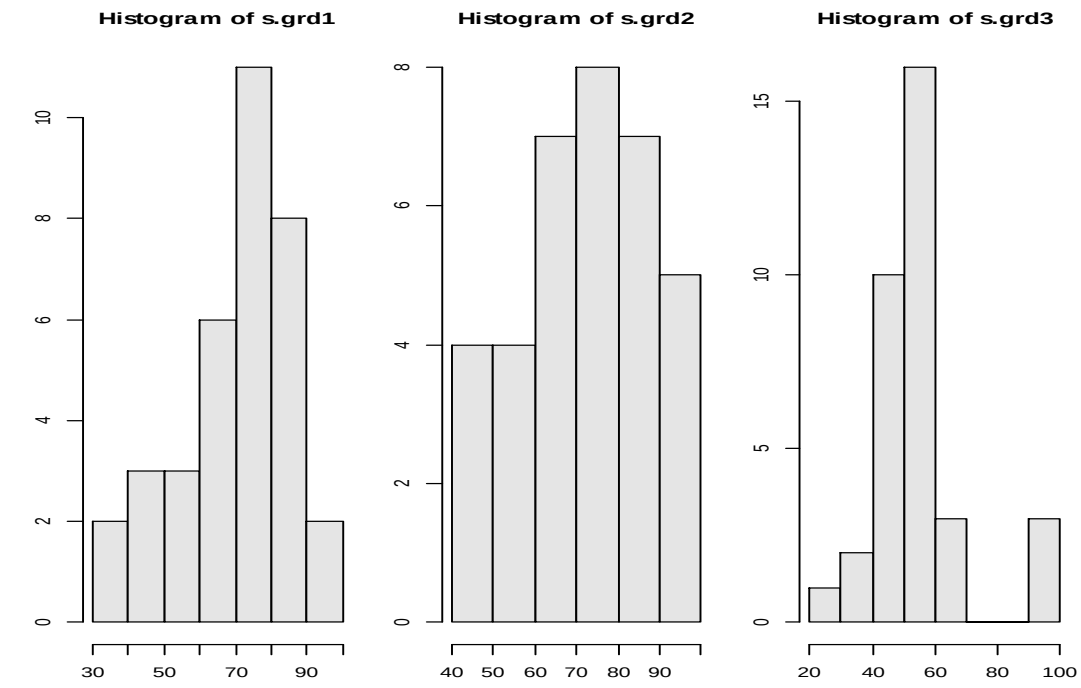
وإذا ما نظرنا إلى القيم الكبرى (العظمى) للمقررات الثلاثة والتي يُلاحظ تقاربها (94، 96، و95 على الترتيب)، فإننا، وبعد قراءة النتائج السابقة، نستنتج وجود قيم متطرفة عليا في المقرر الثالث، بمعنى أن طالب أو أكثر قد حققوا درجات عالية جدا في هذا المقرر رغم تدني مستوى الأداء لغالبية الطلاب مما يُعد من الناحية الإحصائية "تطرفا" في قيم ذلك المتغير.

- **المدرج التكراري:**

إن الكثير من تلك الاستنتاجات السابقة وغيرها يمكن قراءتها "بشكل مرئي" باستخدام التمثيل البياني، وعلى رأسها المدرج التكراري، ويمكن عرض المدرجات التكرارية الثلاثة في إطار واحد (شكل (14.4))، بتنفيذ الدوال التالية، ولاحظ تعديل بعض خيارات الرسم ،(من أحجام للمدرجات ومسافات للهوامش ولون الأعمدة)، بغرض عرض المدرجات الثلاثة بصورة أفضل:

```
> par(mfrow=c(1,3),mex=1.2, mar=c(2,2,3,1))
> hist(s.grd1,col="grey90")
> hist(s.grd2,col="grey90")
> hist(s.grd3,col="grey90")
```

ويُلاحظ من المدرجات في الشكل (14.4) وجود التواء بسيط إلى اليسار في توزيع المقررين الأولين grd1 وgrd2، كما ذكرنا سابقا، أما قيم المقرر الثالث grd3 فتُظهر التواء بسيط إلى اليمين والذي ساعد على ظهوره هو وجود القيم المتطرفة العليا، (العامود الأخير الذي يَظهر بشكل منفصل في يمين المدرج).



شكل 14.4: المدرجات التكرارية للمتغيرات grd1، grd2، وgrd3 في البيانات stu.data1

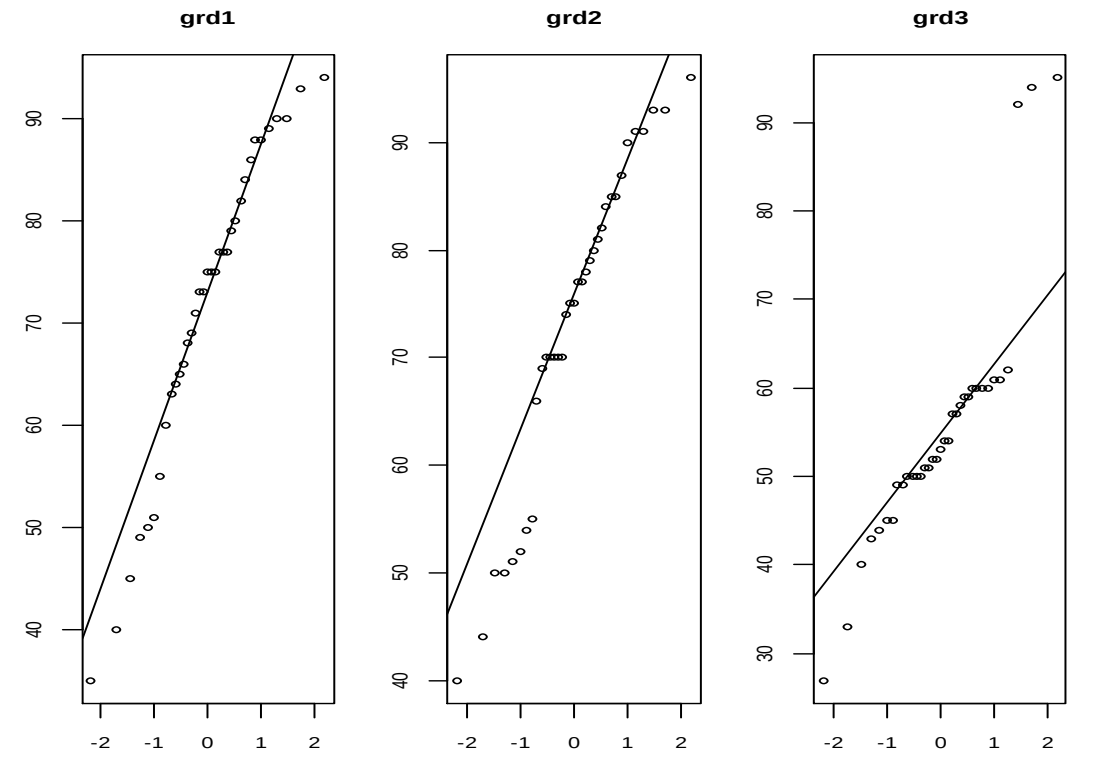
- **استخدام رسم Q-Q الطبيعي:**

في بعض الأحيان قد لا نتمكن من الحكم بصريا على توزع البيانات بتوزيع طبيعي[1] من خلال استخدام المدرج التكراري فقط، لذلك يمكن اللجوء لرسم Q-Q الطبيعي بالصورة:

```
> par(mfrow=c(1,3),mex=1.5, mar=c(2,2,3,1))
> qqnorm(s.grd1,main="grd1")
> qqline(s.grd1)
> qqnorm(s.grd2,main="grd2")
> qqline(s.grd2)
> qqnorm(s.grd3,main="grd3")
> qqline(s.grd3)
```

ويُلاحظ من الشكل (15.4) أن المقرر (المتغير) الأول grd1 يقترب إلى حد كبير من التوزيع الطبيعي (رغم وجود ذلك الالتواء)، حيث نشاهد اقتراب النقاط من الخط المستقيم بشكل منظم. يليه في ذلك المقرر الثاني grd2 إلى حد ما، أما المقرر الثالث grd3 فلا يبدو طبيعيا خاصة مع وجود القيم المتطرفة، (النقاط في أعلى الرسم إلى اليمين)، وعدم توزع النقاط الأخرى على الخط المستقيم بانتظام. ويمكن لمزيد من الإيضاح الرجوع للرسم النقطي أو الشريطي أيضا لمشاهدة هذه القيم المتطرفة، (في الشكل (5.4)).

[1] سيتم التطرق لاختبارات الفروض الخاصة بالتوزيع الطبيعي في الفصول القادمة.

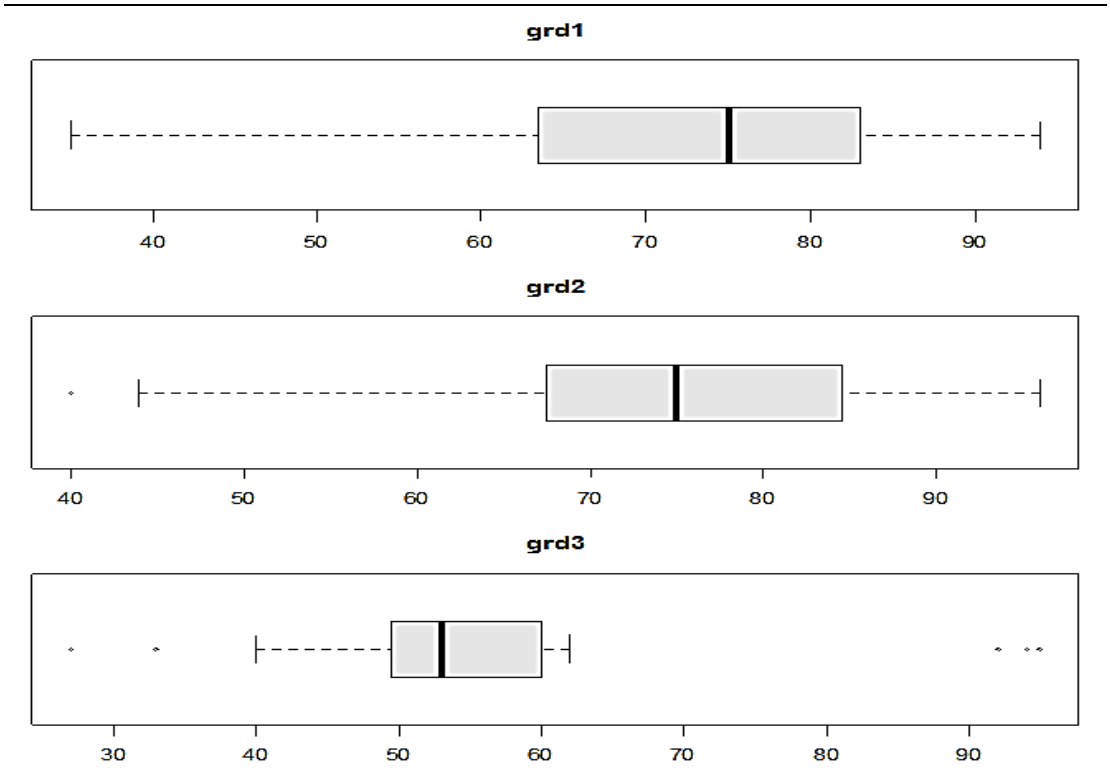


شكل 15.4: تمثيل Q-Q الطبيعي للمتغيرات grd1، grd2، وgrd3 في البيانات stu.data1

- **استخدام شكل الصندوق:**

يمكن أن يتم استكشاف توزيع البيانات أيضا بشكل بسيط وواضح باستخدام شكل الصندوق، كما هو موضح في شكل (16.4)، والذي ينتج عن سطور الأوامر التالية، التي تم فيها استخدام خيار العرض الأفقي horizontal=T واللون الرصاصي إضافة لخيارات الرسم الأخرى:

```
> par(mfrow=c(3,1),mex=1.2, mar=c(2,2,3,1))
> boxplot(s.grd1,main="grd1",col="grey90",horizontal=T)
> boxplot(s.grd2,main="grd2",col="grey90",horizontal=T)
> boxplot(s.grd3,main="grd3",col="grey90",horizontal=T)
```



شكل 16.4: شكل الصندوق للمتغيرات grd1، grd2، وgrd3 في البيانات stu.data1

ونلاحظ من الشكل (16.4) ما يلي:

1. اقتراب الصندوق، والذي يمثل كتلة البيانات، في المتغيرين أو المقررين الأولين grd1 وgrd2 إلى الجانب الأيمن في الرسم مما يُظهر التواء إلى اليسار في توزيع البيانات، وهذه دلالة على ارتفاع المستوى الدراسي للطلاب في هذين المقررين. إلا أن ما نراه في المتغير الثالث grd3 ليس التواء بالمعنى الصحيح، لأن كل المشاهدات في هذا المتغير، باستثناء النقاط (الدرجات) الثلاثة المتطرفة في يمين الرسم، تنحصر في الجانب الأيسر.
2. أغلبية درجات الطلاب في المقررين grd1 وgrd2 تنحصر في الفترة (65 إلى 85) تقريبا، أما أغلبية الدرجات في المقرر grd3 فتنحصر في الفترة (50 إلى 60)، مما يدل على انخفاض مستوى الطلاب في المقرر الثالث، كما ذكرنا سابقا.
3. عدم ظهور قيم متطرفة في المقرر الأول، وظهور قيمة واحدة متطرفة (هي الدرجة 40) في المتغير الثاني، وكذلك ظهور قيم متطرفة دنيا (هما الدرجتان 27 و33) اللتان لم تظهرا في الرسومات السابقة بوضوح، إضافة للقيم المتطرفة الثلاثة الكبرى الظاهرة.

4. انتشار البيانات (عرض الصندوق) والذي يمثل تشتت درجات الطلاب يبدو أكبر في المقررين grd1 وgrd2 مما هو عليه في المقرر grd3، إلا أن ذلك لم يكن جليا واضحا في النتائج والرسومات السابقة، ولا حتى بحساب الانحراف المعياري للمقررات الثلاثة؛

```
> sd(s.grd1);sd(s.grd2);sd(s.grd3)

[1] 15.605
[1] 15.21896
[1] 14.28892
```

والتي تبدو متقاربة[1]، وهذه في الواقع إحدى مزايا التمثيل الجيد لتوزيع البيانات في رسم الصندوق. من ناحية أخرى، قد يكون استخدام مقياس المدى الربيعي كمقياس للتشتت، بحسب سلوك بياناتنا الحالية أفضل، حيث تظهر الفروقات في انتشار الدرجات بين المقررات الثلاثة بشكل أوضح من الانحراف المعياري كما نرى؛

```
> IQR(s.grd1);IQR(s.grd2);IQR(s.grd3)

[1] 19.5
[1] 17
[1] 10.5
```

- **استخدام شكل الساق والورقة:**

سنقوم الآن بتنفيذ رسم الساق والورقة للمتغيرات grd1، grd2، وgrd3، (والذي سيُعرض هنا بشكل مُعدّل):

```
> stem(s.grd1);stem(s.grd2);stem(s.grd3)
```

المتغير grd1؛

```
  3 | 5
  4 | 059
  5 | 015
  6 | 0345689
  7 | 1335557779
  8 | 0246889
  9 | 0034
```

المتغير grd2؛

```
  4 | 04
  5 | 001245
  6 | 69
  7 | 000004557789
  8 | 0124557
  9 | 011336
```

[1] اقتراب قيمة الانحراف المعياري للمتغير grd3 من قيم المتغيرين الأولين سببه وجود القيم المتطرفة العليا التي أدت لرفع قيمته.

والمتغير grd3؛

```
  2 | 7
  3 | 3
  4 | 0345599
  5 | 0000112234477899
  6 | 0000112
  7 |
  8 |
  9 | 245
```

ولاحظ مدى اقتراب شكل الأعمدة الأفقية في رسم المتغير grd1 من التوزيع الطبيعي، (مع ظهور ذلك الالتواء البسيط إلى اليسار). كذلك يظهر لنا من رسم المتغير الثاني grd2 تلك "الفجوة" في تكرار الدرجات حول الدرجة 60 والتي أدت إلى ابتعاد توزيع المتغير قليلا عن التوزيع الطبيعي. أما بالنسبة للمتغير الثالث grd3 فيظهر تركز الدرجات في الفترة من 40 إلى 60، إضافة إلى ظهور القيم المتطرفة الثلاثة في النهاية 92، 94، و95 وابتعاد توزيع المتغير عن التوزيع الطبيعي في العموم.

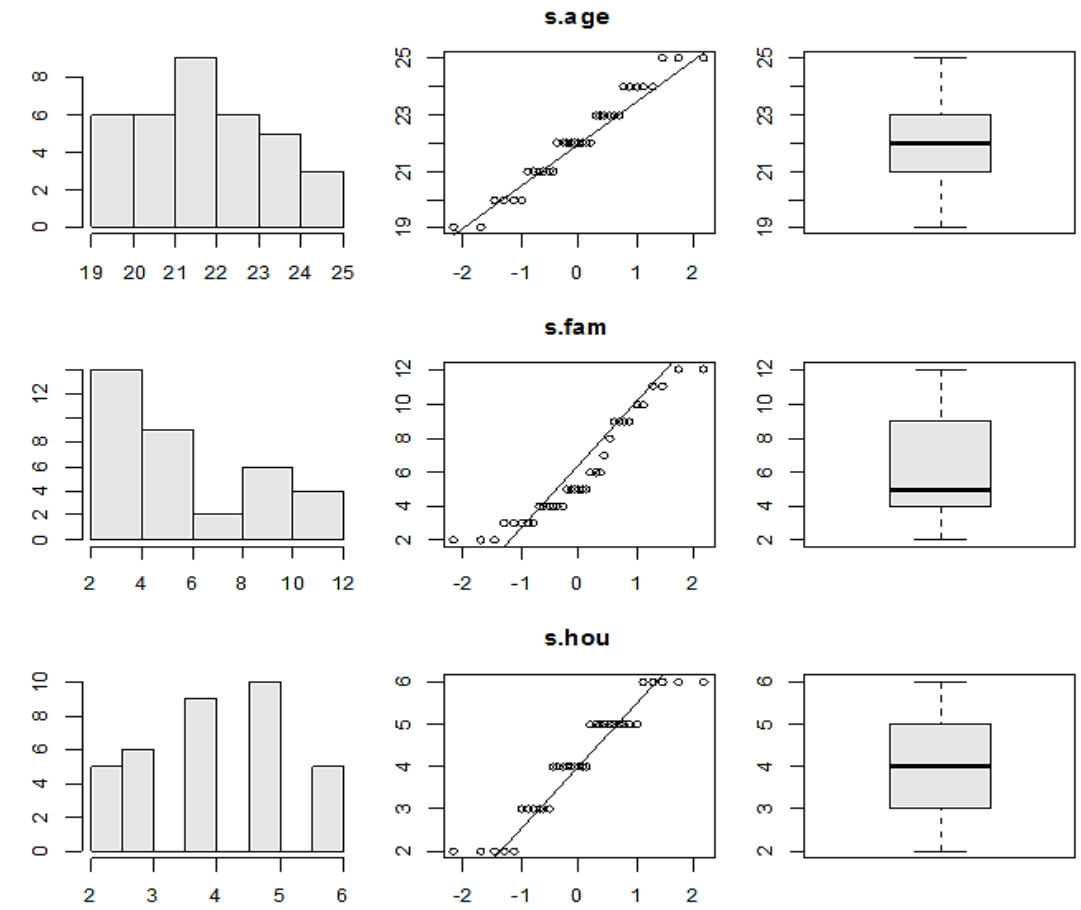
- **استكشاف باقي المتغيرات الكمية:**

فيما يخص باقي المتغيرات الكمية؛ عمر الطالب age، عدد أفراد أسرة الطالب fam وعدد غرف منزل الطالب hou، فيتم دراسة توزيعاتها بشكل منفصل أي بدون مقارنتها مع بعضها البعض، (كما كان الحال مع درجات الطلاب)، لأنها مقاسة بوحدات قياس مختلفة؛ فالعمر مُقاس بالسنة وأفراد الأسرة بعدد الأشخاص وغرف المنزل مُقاسة بعددها.

في الشكل (17.4)، قد تُعطي النظرة الأولى لشكل الصندوق للمتغيرات الثلاثة انطباعا أوليا بأنها تتبع التوزيع الطبيعي، إلا أن المدرجات التكرارية تُظهر صورة مختلفة عن ذلك الاستنتاج، فمن الواضح أن المتغيرين fam وhou يسلكان سلوكا غير طبيعي. أما متغير العمر age فيبدو أنه يسلك سلوكا طبيعيا. وبالنسبة لشكل Q-Q الطبيعي، فهي لم تكن ذات فائدة كبيرة مع هذه المتغيرات وذلك ربما بسبب وجود "فجوات" أو قيم غير موجودة ضمن قيم هذه المتغيرات أدت لتوزع النقاط على الخط المستقيم بصورة غير معبرة. وللحصول على الشكل (17.4) تم تنفيذ مجموعة السطور التالية:

```
> par(mfrow=c(3,3),mex=1.2,mar=c(2,2,3,1))
> hist(s.age,col="grey90",main="")
> qqnorm(s.age,main="s.age")
> qqline(s.age)
> boxplot(s.age,col="grey90")
> hist(s.fam,col="grey90",main="")
> qqnorm(s.fam,main="s.fam")
> qqline(s.fam)
> boxplot(s.fam,col="grey90")
```

```
> hist(s.hou,col="grey90",main="")
> qqnorm(s.hou,main="s.hou")
> qqline(s.hou)
> boxplot(s.hou,col="grey90")
```



شكل 17.4: رسوم المدرج التكراري، Q-Q الطبيعي، وشكل الصندوق للمتغيرات age، fam، وhou في البيانات stu.data1

وللمزيد من التوضيح، نستخدم شكل الساق والورقة، والذي يُلاحظ فيه أولا أن كل القيم على يمين الرسم هي أصفار وذلك لأن خانتي الآحاد والعشرات تم استخدامها على يسار الرسم (الساق) ولم يتبق خانات لليمين (الأوراق) فتم وضع هذه الأصفار. ويتضح من الأشكال الثلاثة توزع المتغير الأول (age) بالتوزيع الطبيعي، وابتعاد توزيع المتغيرين الآخرين (fam وhou) عن ذلك التوزيع؛

```
> stem(s.age)

  19 | 00
  20 | 0000
  21 | 000000
  22 | 000000000
  23 | 000000
  24 | 00000
  25 | 000
```

```
> stem(s.fam)

   2 | 00000000
   4 | 000000000000
   6 | 0000
   8 | 00000
  10 | 0000
  12 | 00

> stem(s.hou)

  2 | 00000
  2 |
  3 | 000000
  3 |
  4 | 000000000
  4 |
  5 | 0000000000
  5 |
  6 | 00000
```

كما توضح المدرجات التكرارية، وأيضا أشكال الساق والورقة، أن معظم أعمار الطلاب في العينة تتراوح ما بين 20 و23 سنة وهذا يبدو ملائما في هذه المرحلة التعليمية، وكذلك فإن معظم أسر الطلاب تتراوح أعداد أفرادها ما بين 2 و6 أفراد، وأما المنازل فهي ذات 4 أو 5 غرف لدى غالبية الطلاب.

- **مراقبة تشتت المتغيرات:**

وبملاحظة التشتت في هذه المتغيرات[1]، عن طريق حساب الانحراف المعياري والمدى الربيعي؛

```
> sd(s.age);sd(s.fam);sd(s.hou)

[1] 1.647509
[1] 3.058258
[1] 1.278129

> IQR(s.age);IQR(s.fam);IQR(s.hou)

[1] 2
[1] 5
[1] 2
```

يتضح بالنسبة لمتغير العمر أن متوسط الفروقات بين أعمار الطلاب هو أقل من سنتين، ويُلاحظ أن هنالك درجة تشتت كبيرة نوعا ما ضمن عدد أفراد أسر هؤلاء الطلاب، (وهذا أيضا واضح من ارتفاع الأعمدة في

[1] دون إجراء مقارنة بين هذه المتغيرات كما وضحنا، بل مقارنة المقاييس فقط.

أطراف المدرج التكراري للمتغير fam وانخفاضها في المنتصف)، وأما عدد غرف منازل تلك الأسر فهي غير مختلفة كثيرا فيما بينها، حيث أن مقاييس تشتتها لها قيم منخفضة.

- **التعليق على الأعمدة البيانية والقطاعات الدائرية:**

بالنسبة للمتغيرات النوعية في الدراسة، وهما نوع الطالب وترتيبه في الفصل الدراسي، فيمكن من الشكل (7.4) والشكل (8.4) السابقين، ملاحظة تقارب عدد الطلبة الذكور والإناث في عينة الدراسة، وكذلك ملاحظة أن الفصلين الرابع والسادس يضمان أعداد أكبر من الطلاب، وأن أقل عدد للطلاب في العينة موجود في الفصل الدراسي الثاني.

2.4.4 الاستكشاف متعدد المتغيرات في الدراسة (Exploring Data in Multivariate Fashion)

بعد أن تم "مراقبة" سلوك أو توزيع المتغيرات بشكل منفرد في البند السابق، سنقوم هنا بمتابعة استكشاف متغيرات الدراسة باستخدام أدوات التحليل الاستكشافي المتعدد، ثم التعليق على النتائج.

- **مراقبة النزعة المركزية:**

أولا، من النتيجة الخاصة باستخدام دالة summary بناء على تقسيم النوع للطلاب، (المُتحصَل عليها في بداية البند (1.3.4)) والتي تم تلخيصها في الجدول (6.4)، يمكن إجراء مقارنة لأنماط المتغيرات بين الطلبة الذكور والإناث وملاحظة التالي:

- معدلات (أوساط) درجات الطالبات في المقررات الثلاثة grd1، grd2، وgrd3 أعلى من تلك التي للطلبة الذكور بدرجة كبيرة نوعا ما. وكذلك فإن أعلى درجات تم الحصول عليها في هذه المقررات كانت للطالبات، إضافة إلى أن الحد الأدنى لدرجات الطالبات في كل المقررات كان أعلى من الطلبة الذكور. ومن قيم الربيعات يتضح أن معظم درجات الطالبات أفضل بشكل ملحوظ من درجات الطلبة الذكور.

- معدلات أعمار الطلبة الذكور هي أعلى بمقدار ضئيل من الطالبات.

- من قيم الربيعات لترتيب الفصل الدراسي (المتغير sem) يتضح أن الطلبة الذكور يدرسون في مراحل دراسية أعلى بقليل من الطالبات في عينة الدراسة.

- الطلبة الذكور في العينة لديهم أسر ذات أعداد أكبر من الطالبات، (من المتغير fam).

- من المتغير hou، يتضح أن الطالبات يُقمْن في منازل أوسع، (لها أعداد غرف أكثر)، من الطلبة الذكور.

جدول 6.4: ملخص مقاييس دالة summary للمتغيرات في البيانات stu.data1 بناء على التقسيم إلى ذكور وإناث

gen: f

	grd1	grd2	grd3	age	sem	fam	hou
Min.	:51.00	55.00	33.00	19.00	2.00	2	3
1st Qu.	:77.00	80.00	50.00	20.00	4.00	3	5
Median	:84.00	85.00	57.00	22.00	4.00	4	5
Mean	:81.65	83.29	60.06	21.53	4.77	4	5
3rd Qu.	:89.00	91.00	61.00	23.00	6.00	4	6
Max.	:94.00	96.00	95.00	24.00	8.00	9	6

gen: m

	grd1	grd2	grd3	age	sem	fam	hou
Min.	:35.00	40.00	27.00	19.00	2.00	4.00	2.00
1st Qu.	:51.25	51.25	46.00	22.00	4.25	5.25	2.25
Median	:65.00	69.50	51.00	22.50	5.50	8.00	3.50
Mean	:61.56	63.22	50.89	22.72	5.33	7.89	3.28
3rd Qu.	:72.50	73.00	59.75	24.00	7.00	10.0	4.00
Max.	:77.00	79.00	61.00	25.00	8.00	12.0	5.00

- **مراقبة تشتت البيانات:**

يمكن إجراء مقارنة إضافية بين الطلبة الذكور والطالبات من ناحية تشتت توزيع المتغيرات في العموم لكل من الفئتين باستخدام دالة aggregate بالصورة التالية:

```
> aggregate(cbind(s.grd1,s.grd2,s.grd3,s.sem,s.fam,s.hou)
~s.gen,stu.data1,sd)
```

```
  s.gen   s.grd1   s.grd2    s.grd3    s.sem    s.fam     s.hou
1     f 10.87969 10.07946 17.422940 1.786386 1.936492 0.9354143
2     m 12.96249 12.66563  8.910594 1.909727 2.720054 0.9582800
```

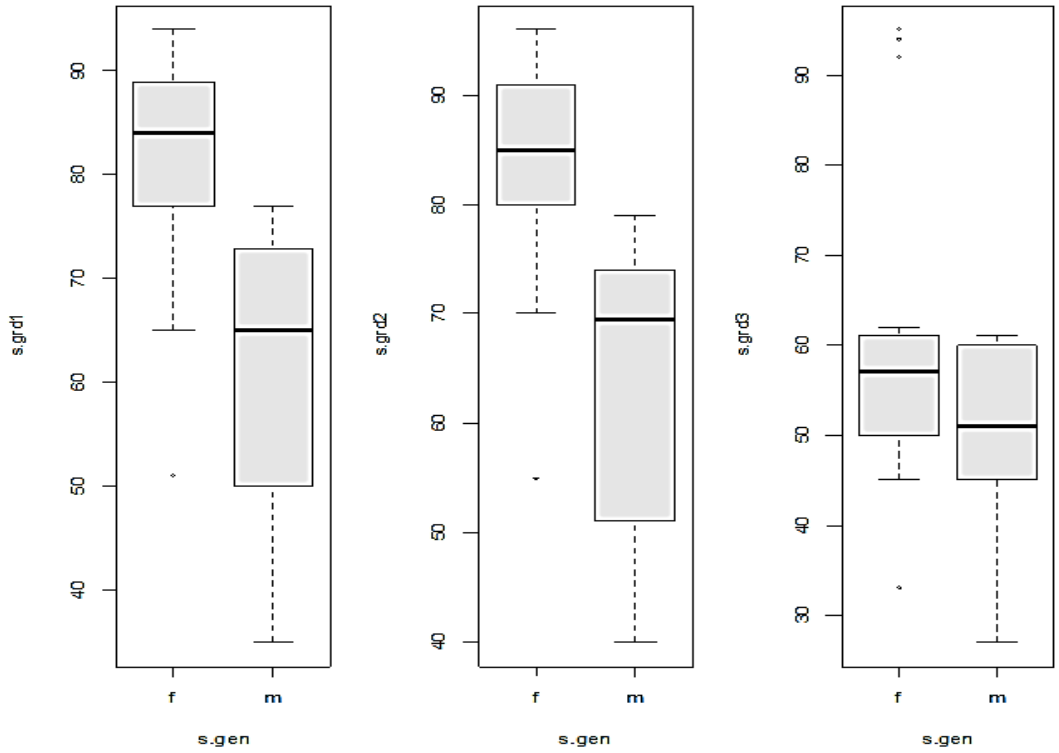
وأهم ما يُلاحظ في هذه النتيجة وجود اختلاف كبير في تشتت الدرجات في المقرر grd3 بين الذكور والإناث حيث كانت درجات الطالبات أكثر تشتتا، وهذا قد يشير إلى تفاوت أداء الطالبات في هذا المقرر. أما بالنسبة للمتغيرات الأخرى، فالفرق بين درجات التشتت ليس بالكبير في العموم.

- **التعليق على شكل الصندوق:**

ومن خلال استخدام شكل الصندوق، (الشكل (18.4))، يمكن على سبيل المثال، مقارنة التغير في درجات الطلبة والطالبات كما لاحظناه في النتيجة السابقة ولكن بشكل مرئي، من خلال تنفيذ الأوامر التالية؛

```
> par(mfrow=c(1,3),mex=1.2,mar=c(4,4,1,1))
> boxplot(s.grd1~s.gen,ylab="s.grd1",xlab="s.gen",col="grey90")
> boxplot(s.grd2~s.gen,ylab="s.grd2",xlab="s.gen",col="grey90")
> boxplot(s.grd3~s.gen,ylab="s.grd3",xlab="s.gen",col="grey90")
```

ويُلاحظ من الشكل تفوق الطالبات على الطلبة الذكور في المقررات الثلاثة بصورة عامة، وبالتحديد يمكن مشاهدة ارتفاع درجات الطالبات (كتلة البيانات مُمَثلة بالصندوق) في المقررين الأولين grd1 وgrd2 بشكل ملحوظ عما هو عنه في المقرر الثالث grd3.



شكل 18.4: شكل الصندوق للمتغيرات grd1، grd2، وgrd3 بحسب تقسيم الذكور والإناث في البيانات stu.data1

- **استخدام مصفوفة شكل الانتشار:**

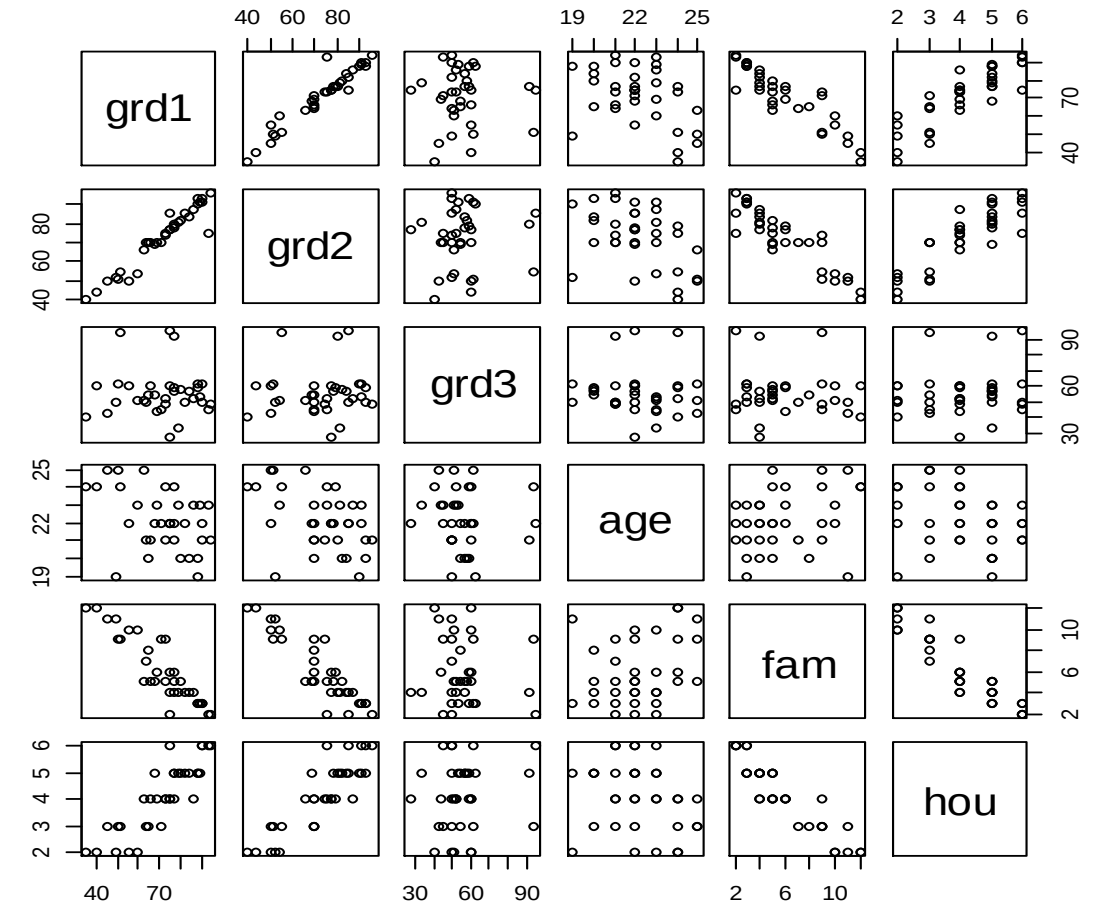
من جديد، يمكن استخدام شكل الانتشار لاستكشاف وجود علاقات خطية أو غير خطية بين المتغيرات بشكل مرئي قبل استخدام مقاييس الارتباط المعروفة. وقد سبق استخدام دالة شكل الانتشار بين متغيرين plot، (في الشكل (11.4))، إلا أننا سنقوم هنا بعرض شكل الانتشار بين كل متغيرين من المتغيرات الكمية في البيانات، وهو ما يُعرف **بمصفوفة شكل الانتشار** (Scatter Plot Matrix)، وهذا يتم في لغة R باستخدام الدالة pairs بالصورة التالية:

```
> pairs(stu.data1[c(-5,-6)])
```

فنحصل على الشكل (19.4)، والذي يساعدنا في استكشاف العلاقات الثنائية بين المتغيرات الكمية في البيانات. (ولاحظ أننا استثنينا المتغيران النوعيان gen وsem من دالة الرسم باستخدام خيار الأقواس المربعة والإشارة السالبة، حيث أن المتغيران النوعيان لهما الترتيب الخامس والسادس في البيانات).

وكما نرى، فإن الشكل (19.4) يبدو على هيئة مصفوفة متماثلة عناصرها هي أشكال الانتشار الثنائية بين المتغيرات وقطرها يحتوي على أسماء المتغيرات، وأشكال الانتشار في المثلث السفلي للمصفوفة هي انعكاس للأشكال في المثلث العلوي. وأهم ما يُلاحظ من أشكال الانتشار التالية هو ما يلي:

- وجود علاقة خطية طردية قوية بين كل من المتغيرين grd1 وgrd2، المتغيرين grd1 وhou، والمتغيرين grd2 وhou.
- وجود علاقة خطية عكسية قوية بين كل من المتغيرين grd1 وfam، المتغيرين grd2 وfam، والمتغيرين fam وhou.
- وجود علاقات خطية عكسية تقترب من القوية بين المتغير age وكل من grd1 وgrd2.



شكل 19.4: مصفوفة شكل الانتشار للمتغيرات الكمية في البيانات stu.data1

- **استخدام جداول الاقتران والأعمدة البيانية لها:**

لننتقل الآن إلى نوعية أخرى من التحليل الاستكشافي، والتي ستشمل تكوين جداول الاقتران وتمثيلها بيانيا. لنقم على سبيل المثال بتكوين جدول اقتران لمتغير العمر age مع متغير النوع gen أولا:

```
> table(s.gen,s.age)

     s.age
s.gen 19 20 21 22 23 24 25
    f  1  4  3  4  4  1  0
    m  1  0  3  5  2  4  3
```

ثانيا، سنقوم باختزال الجدول السابق عن طريق عرض متغير العمر على هيئة فترات عمرية مقترنة بالنوع، وذلك بإنشاء ثلاثة فترات للعمر هي (أقل من 21 سنة)، (من 21 إلى 23 سنة)، و(24 سنة فأكثر) ودمج التكرارات المناظرة لهذه الفترات، ثم إدخال التكرارات الجديدة كمصفوفة بيانات باسم age.gent؛

```
> age.gent<-
matrix(data=c(5,11,1,1,10,7),nrow=2,ncol=3,byrow=T)

> age.gent

     [,1] [,2] [,3]
[1,]    5   11    1
[2,]    1   10    7
```

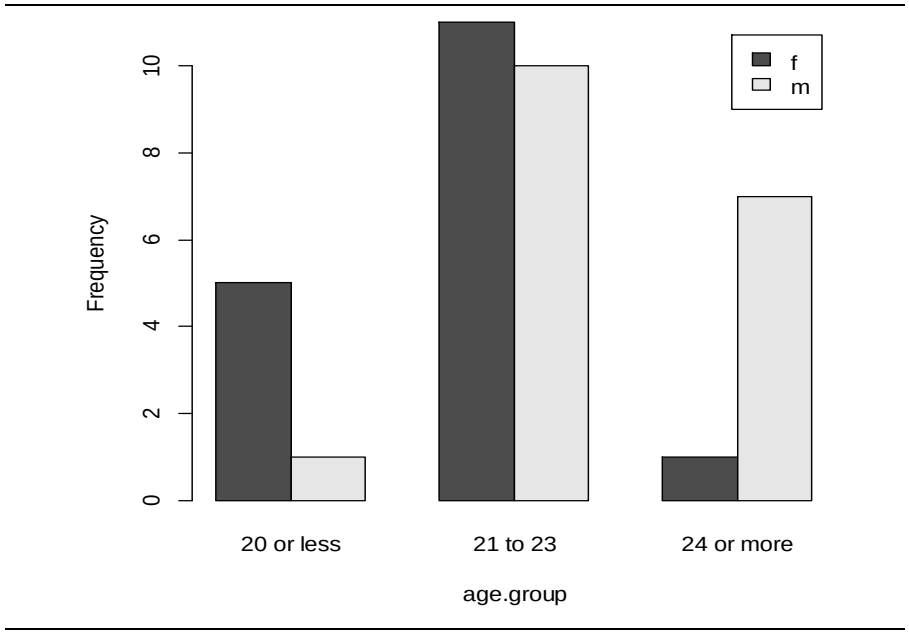
يلي ذلك تعريف الأسماء المناظرة للتكرارات؛

```
> dimnames(age.gent)<-
list(gender=c("f","m"),age.group=c("20 or less","21 to
23" ,"24 or more"))

> age.gent

      age.group
gender 20 or less 21 to 23 24 or more
     f          5       11          1
     m          1       10          7
```

الآن يمكن تمثيل جدول الاقتران age.gent باستخدام دالة الأعمدة البيانية:

```
> barplot(age.gent,xlab="age.group",ylab="Frequency",
beside=T,legend.text=T)
```

ويُلاحظ من الشكل (20.4) أن أكثرية أعمار الطلاب من الجنسين تقع في الفترة العمرية الوسطى، (من 21 إلى 23 سنة)، وأن الطالبات هن أكثر عددا من الطلبة الذكور في الفترة العمرية الدنيا بينما هن أقل عددا في الفترة العمرية الأكبر.



شكل 20.4: الأعمدة البيانية لجدول الاقتران age.gent

لنأخذ المتغيران؛ عدد أفراد الأسرة fam وعدد غرف المنزل hou كمثال آخر على تكوين جداول الاقتران وتمثيلها بيانيا، حيث سيتم اختزال قيم المتغير fam في ثلاثة فترات هي (4 أفراد فأقل)، (من 5 إلى 8 أفراد)، و(9 أفراد فأكثر)، واختزال قيم المتغير hou في فترتين هما (4 غرف فأقل) و(5 غرف فأكثر) بالصورة التالية، ونبدأ بالجدولة الأولية:

```
> table(s.hou,s.fam)

     s.fam
s.hou 2 3 4 5 6 7 8 9 10 11 12
    2 0 0 0 0 0 0 0 0  2  1  2
    3 0 0 0 0 0 1 1 3  0  1  0
    4 0 0 2 3 3 0 0 1  0  0  0
    5 0 3 4 3 0 0 0 0  0  0  0
    6 3 2 0 0 0 0 0 0  0  0  0
```

ثم دمج التكرارات؛

```
> fam.hou<-
matrix(data=c(2,8,10,12,3,0),nrow=2,ncol=3,byrow=T)

> fam.hou

     [,1] [,2] [,3]
[1,]    2    8   10
[2,]   12    3    0
```

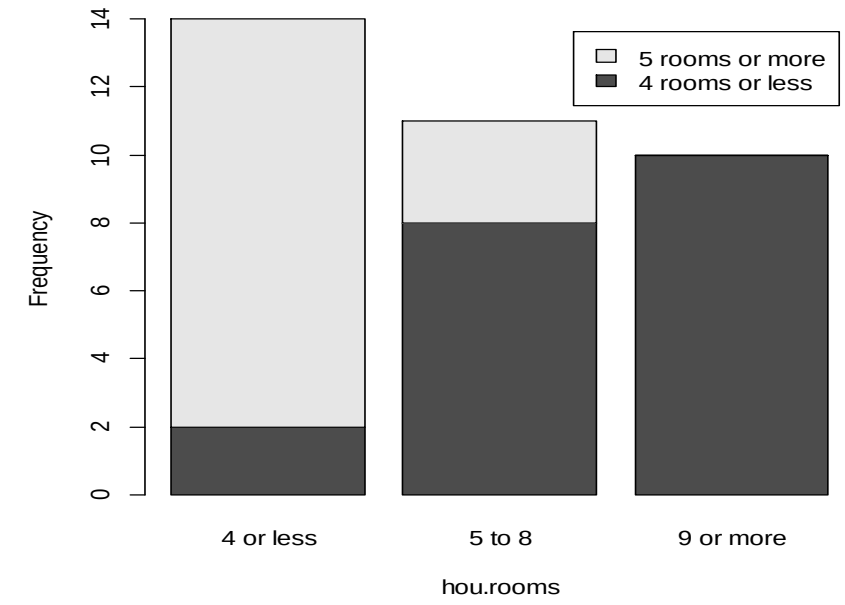
ثم تسمية الأعمدة والصفوف؛

```
> dimnames(fam.hou)<-list(hou.rooms=c("4 rooms or
less","5 rooms or more"),fam.members=c("4 or less","5 to
8","9 or more"))

> fam.hou

                fam.members
hou.rooms         4 or less 5 to 8 9 or more
  4 rooms or less         2      8        10
  5 rooms or more        12      3         0
```

وأخيرا رسم الأعمدة البيانية، والتي سيُستخدم فيها الخيار beside=F لتغيير طريقة العرض الأعمدة؛

```
> barplot(fam.hou,xlab="hou.rooms",ylab="Frequency",
beside=F,legend.text=T)
```

بعد ذلك نحصل على الشكل (21.4)، والذي يُلاحظ منه أن عدد الأفراد الأقل من أسر الطلاب (4 أفراد فأقل) يقطنون في منازل بها عدد غرف كبير، (5 غرف فأكثر)، وبالعكس، نجد أن عدد الأفراد الأكبر يقطنون في منازل بها عدد غرف أقل من 4، وهذا قد يبدو متناقضا بعض الشيء، إلا أن تفاوت الوضع الاقتصادي لتلك الأسر قد يكون أحد الأسباب الرئيسية لتلك النتيجة.



شكل 21.4: الأعمدة البيانية لجدول الاقتران fam.hou

3.4.4 أهم استنتاجات التحليل الاستكشافي للبيانات (Important Conclusions of the EDA)

إن أهم ما تم استنتاجه من المعلومات التي تم الحصول عليها من تطبيق أدوات التحليل الاستكشافي على البيانات stu.data1 يمكن تلخيصه في هذا البند، مع التذكير أن الفكرة الرئيسية كانت تدريب القارئ على التعامل مع أدوات التحليل الاستكشافي في نظام R من خلال تطبيق عملي على بيانات افتراضية، وأن هذه البيانات ليست بذات أهمية بحد ذاتها نظرا لأنها بيانات افتراضية من جهة، ومن جهة أخرى فإن عدد المشاهدات (حجم العينة) ليس كبيرا بشكل مُعتبّر من الناحية العملية.

وأهم ما يمكن ملاحظته من نتائج التحليل يمكن تلخيصه في النقاط التالية:

1. أداء الطلاب الدراسي في المقررين grd1 وgrd2 كان جيدا ويتبع أيضا توزيعا طبيعيا (رغم اتجاه الكثير من الدرجات للقيم الأعلى)، مقارنة بأدائهم في المقرر grd3، (والذي يبتعد عن التوزيع الطبيعي)، مما قد يعكس تعرض الطلاب لظروف دراسية مختلفة في هذا المقرر مثل صعوبة مفردات المقرر، أو عدم "توافق" الطلاب مع استاذ المقرر، أو غير ذلك من الأسباب. وحصول طالب أو أكثر على درجات عالية جدا في هذا المقرر الثالث (قيم متطرفة) يؤيد هذا الاستنتاج.

2. يُلاحظ وجود تفاوت في الأداء الدراسي للطلاب في كل من المقررين grd1 وgrd2، مع ذلك الأداء الجيد في العموم، وعدم وجود هذا التفاوت في درجات المقرر grd3.

3. أعمار الطلاب age تتبع توزيعا طبيعيا تقريبا وتتراوح ما بين 20 و23 سنة، بتشتت يُعادل السنتين تقريبا، وهذا يبدو طبيعيا في هذه المرحلة التعليمية.

4. تتراوح معظم أعداد أفراد أسر الطلاب fam ما بين 2 و6 أفراد، بتشتت كبير إلى حد ما، وأما منازل هذه الأسر hou فهي ذات 4 أو 5 غرف لدى الغالبية. ولا يتبع هذان المتغيران توزيعا طبيعيا.

5. يُلاحظ أن معظم الطلاب في العينة هم في الفصل الدراسي الرابع أو السادس، وأن عدد الطلبة الذكور والإناث متقارب جدا.

6. الطالبات بحسب هذه العينة هن الأفضل أداء في المقررات الدراسية، ويعشن ضمن أسر صغيرة في منازل أكثر اتساعا مقارنة بالطلبة الذكور.

7. عدد غرف المنزل مرتبط طرديا بأداء الطلاب في المقررين grd1 وgrd2، أي أنه بزيادة الأول يزيد الثاني، أما عدد أفراد الأسرة فيبدو أنه مرتبط عكسيا مع أداء الطلاب في هذين المقررين. من ناحية أخرى، يبدو من علاقة متغير العمر بالمقررين grd1 وgrd2 أنه بزيادة عمر الطالب يقل أداؤه الدراسي. وكذلك توجد علاقة عكسية بين عدد غرف المنزل وعدد أفراد الأسرة.

الفصل الخامس

الاحتمال والتوزيعات الاحتمالية في R
(Probability and Probability Distributions in R)

1.5 حساب الاحتمال (Calculating Probability)

1.1.5 فراغ العينة والأحداث (Sample Space and Events)

2.1.5 تكوين فئات جزئية من فراغ العينة (Making Subsets of Sample Space)

3.1.5 بعض العمليات الأساسية على الفئات (Some Basic Operations on Sets)

4.1.5 حساب الاحتمالات للأحداث (Calculating Probabilities for Events)

2.5 التوزيعات الاحتمالية المنفصلة (Discrete Probability Distributions)

3.5 أهم التوزيعات المنفصلة الخاصة (Most Important Special Discrete Distributions)

1.3.5 التوزيع المنتظم المنفصل (Discrete Uniform Distribution)

2.3.5 توزيع ذي الحدين (Binomial Distribution)

3.3.5 التوزيع متعدد الحدود (Multinomial Distribution)

4.3.5 التوزيع الهندسي (Geometric Distribution)

5.3.5 توزيع ذي الحدين السالب (Negative Binomial Distribution)

6.3.5 التوزيع فوق الهندسي (Hyper-geometric Distribution)

7.3.5 توزيع بواسون (Poisson Distribution)

4.5 التوزيعات الاحتمالية المتصلة (Continuous Probability Distributions)

5.5 أهم التوزيعات المتصلة الخاصة (Most Important Special Continuous Distributions)

1.5.5 التوزيع المنتظم المتصل (Continuous Uniform Distribution)

2.5.5 التوزيع الطبيعي (Normal Distribution)

3.5.5 توزيع جاما (Gamma Distribution)

4.5.5 توزيع بيتا (Beta Distribution)

5.5.5 التوزيع الأسي (Exponential Distribution)

6.5.5 توزيع استيودنت *t* (Student's *t* Distribution)

7.5.5 توزيع مربع كاي (Chi-Square Distribution)

8.5.5 توزيع فيشر *F* (Fisher's *F* Distribution)

6.5 حساب العزوم (Calculating Moments)

1.6.5 العزوم والدالة المولدة للعزوم للتوزيعات الخاصة

(Moments and MGF for Special Distributions)

كما هو الحال في بداية كل فصل، لنقم بإنشاء مسار عمل جديد خاص بمواضيع هذا الفصل، إضافة لملف جديد لحفظ سطور الأوامر. ليكن اسم مسار العمل كما هو النسق المتبع في فصول الكتاب؛ "work5" واسم ملف حفظ الأوامر هو "his5".

1.5 حساب الاحتمال (Calculating Probability)

يُعد مفهوم العشوائية وعلم الاحتمالات كما هو معلوم الأسُس التي تقوم عليها كل نظريات الإحصاء الاستدلالي، وهو يُمثل عمليا حلقة الوصل بين مفهومي الإحصاء الاستكشافي والإحصاء الاستدلالي. لذلك سنفرد هذا الفصل لمناقشة كيفية تنفيذ وحساب الاحتمالات باستخدام دوال R، والتي تُعد بحد ذاتها مهمة في سحب العينات وتكوين فراغ العينة وحل المسائل الاحتمالية المختلفة، إضافة لتكوين واستدعاء التوزيعات الاحتمالية المنفصلة والمتصلة وحساب الاحتمالات منها، وغير ذلك من المواضيع المتعلقة بنظرية الاحتمال.

ملاحظة:

للعمل على الدوال المتعلقة بالاحتمال سنحتاج لتحميل الحزمة الإضافية "prob"، (إن لم تكن متوفرة لديك مسبقا)، والتي تتبعها حزم إضافية أخرى[1] يتم تحميلها تلقائيا عند اختيارك للحزمة.

1.1.5 فراغ العينة والأحداث (Sample Space and Events)

الأحدث هي النتائج التي يتم الحصول عليها عند إجراء التجربة العشوائية، ويتم تنظيم هذه الأحداث عادة داخل فراغ العينة. وسنتناول في هذا البند بعض الأمثلة لتكوين فراغ العينة باستخدام دوال R.

ويمكن للمستخدم تكوين أي فراغ عينة عن طريق إدخال العناصر أو الأحداث الناتجة عن التجربة العشوائية باستخدام دالة إطار البيانات أو دالة c()، فمثلا في تجربة إنجاب الأطفال يتكون فراغ العينة من نتيجتين فقط، (إذا ما تم استثناء حالات التوائم)، هما ذكر (m) أو أنثى (f)، ويمكن باستخدام رمز فراغ العينة المعتاد "S" كتابة S={m, f}، ويمكن تكوين فراغ العينة في R، وليكن باسم S1، بالصورة:

```
> S1<-data.frame(Delivery=c("m","f"))
> S1
  Delivery
1        m
2        f
```

[1] الحِزم الإضافية المرتبطة بحزمة الاحتمال prob هي combinat، AsianOptions، timeSeries، timeDate، fBasics، fOptions، hypergeo، elliptic، contfrac، وVGAM.

ولاحظ أنه تم تعيين الاسم (Delivery) أي الإنجاب، للمتجه الذي يمثل عناصر فراغ العينة فقط لتوضيح ما تمثله هذه العناصر. وبنفس الكيفية يمكن كتابة فراغ العينة لتجربتي إنجاب، (أو تجربة إنجاب لمرتين)، بالصورة:

```
> S2<-data.frame(Delivery=c("mm","mf","fm","ff"))
> S2
  Delivery
1       mm
2       mf
3       fm
4       ff
```

مثال آخر على تكوين فراغ العينة؛ هو تحليل فصيلة الدم لشخص ما، والذي ينتج عنه أربع حالات (فصائل)، حيث سيتم إدخاله كمتجه نوعي:

```
> S3<-c("A","B","AB","O")
> S3
[1] "A"  "B"  "AB" "O"
```

وقبل تناول دوال الاحتمالات في R، نقوم أولا باستدعاء الحزمة الإضافية prob، والتي ستُظهر بعد استدعائها المعلومات الخاصة بالمطوِر وبعض الرسائل التي توضح تحميل الحزم الإضافية الضرورية لها، (وهذا هو الحال مع معظم الحزم الإضافية، وننوه هنا أننا لن نقوم بعرض ناتج استدعاء الحزم الإضافية بصورة عامة إلا عند وجود ضرورة لذلك)؛

```
> library(prob)
```

ومن ضمن مزايا الحزمة prob احتواؤها على بعض فراغات العينة الناتجة عن إجراء تجارب عشوائية معروفة، وبذلك لا داعي لكتابتها باستخدام دالة إطار البيانات. ومن أشهر الأمثلة على ذلك فراغ العينة لتجربة رمي عملة معدنية، والذي يُعرّف بالدالة tosscoin؛

```
> tosscoin(1)
  toss1
1     H
2     T
```

حيث يتم تحديد عدد الرميات أو عدد العملات بالرقم داخل القوسين، وبالتالي للحصول على فراغ العينة لتجربة إلقاء عملة معدنية ثلاث مرات نكتب:

```
> tosscoin(3)

  toss1 toss2 toss3
1     H     H     H
2     T     H     H
3     H     T     H
4     T     T     H
5     H     H     T
6     T     H     T
7     H     T     T
8     T     T     T
```

ولاحظ أنه يمكن دائما تعيين اسم لهذا الفراغ إذا ما رغبنا في ذلك.

ومن ضمن التجارب العشوائية الشهيرة أيضا رمي زهر النرد، والذي يُعرّف بالدالة rolldie، فمثلا لتنفيذ تجربة إلقاء زهر نرد مرة واحدة نكتب:

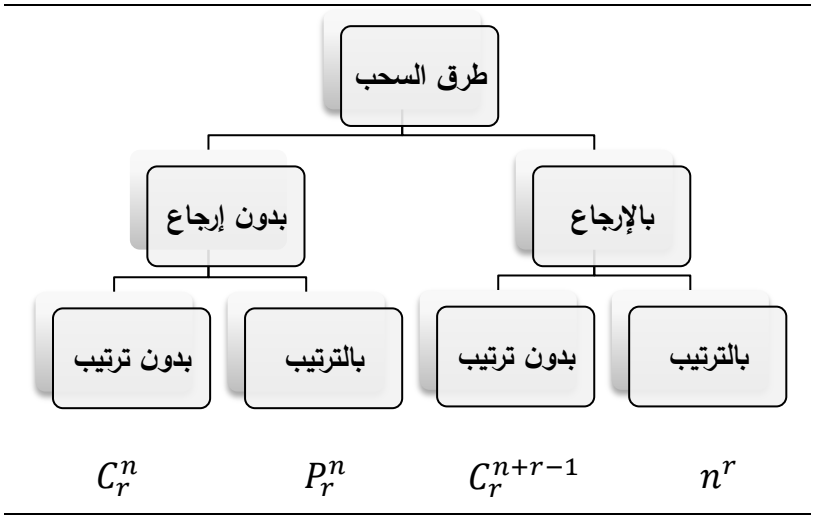
```
> rolldie(1)

  X1
1  1
2  2
3  3
4  4
5  5
6  6
```

وللحصول على فراغ العينة عند رمي زهر النرد مرتين نكتب rolldie(2)، وهكذا. ويمكن اعتبار الدالة rolldie دالة عامة يمكن من خلالها تنفيذ عدد معين من التجارب للحصول على عدد معين من الخيارات باستخدام الخيار nsides مع هذه الدالة، فمثلا إذا كان السؤال المطروح هو؛ ما هي الحالات الكلية لاختيار كرتين من صندوق به 5 كرات مرقمة من 1 إلى 5؟، فإن الإجابة يمكن الحصول عليها بتنفيذ الآتي:

```
> rolldie(2,nsides=5)

   X1 X2
1   1  1
2   2  1
3   3  1
…   …  …

25  5  5
```

وهذه الإجابة، التي تم عرضها بصورة مختصرة، تحتوي على 25 حالة أو عنصر. إلا أن موضوع اختيار العناصر، والذي يندرج تحت مفهوم إيجاد **طرق العد** (Counting Methods)، به تفصيلات يجب أخذها

بالاعتبار، حيث أنه يجب تحديد ما إذا اختيار أو سحب العناصر قد تم **بالإرجاع** (With Replacement) أو بدون إرجاع، و**بالترتيب** (Ordered) أو بدون ترتيب. وللتذكير نُدرج المخطط التفصيلي التالي (شكل (1.5)) الذي يوضح طرق السحب وقوانينها ضمن طرق العد؛



شكل 1.5: طرق سحب العينات المختلفة

حيث ترمز "n" لعدد العناصر الكلي و"r" للعدد المسحوب. وعلى هذا، فإن المثال الأخير كان الاختيار فيه بالإرجاع والترتيب وبالتالي كان عدد الحالات الكلية هو؛ $n^r = 5^2 = 25$ حالة.

وعموما، فإنه يمكن استخدام دالة **المعاينة** `urnsamples` لتنفيذ أي عملية سحب بحسب الطريقة المطلوبة. فمثلا يمكن إيجاد الطرق الكلية لاختيار عنصرين من ثلاثة عناصر بدون إرجاع وباعتبار الترتيب، أي حساب **التباديل** (Permutation)، بالصورة:

```
> urnsamples(1:3,size=2,replace=F,ordered=T)

  X1 X2
1  1  2
2  2  1
3  1  3
4  3  1
5  2  3
6  3  2
```

حيث أن `1:3` يعني أن الاختيار سيتم من ضمن العناصر الكلية (1، 2، 3)، والخيار `size=2` يوضح أن العدد المطلوب سحبه هو عنصرين، والخيار `replace=F` يعني أن السحب بدون إرجاع، والخيار `ordered=T` يعني أن السحب بالترتيب. وكذلك إذا كان المطلوب إيجاد الطرق الكلية لاختيار ثلاثة عناصر من خمسة عناصر بدون إرجاع وبدون ترتيب، أي حساب **التوافيق** (Combination)، فإننا نكتب:

```
> urnsamples(1:5,size=3,replace=F,ordered=F)

   X1 X2 X3
1   1  2  3
2   1  2  4
3   1  2  5
4   1  3  4
5   1  3  5
6   1  4  5
7   2  3  4
8   2  3  5
9   2  4  5
10  3  4  5
```

ولاختيار ثلاثة عناصر من ثلاثة عناصر كلية بالإرجاع وبالترتيب، (والنتيجة معروضة باختصار)، نكتب الآتي:

```
> urnsamples(1:3,size=3,replace=T,ordered=T)

   X1 X2 X3
1   1  1  1
2   2  1  1
3   3  1  1
…   …  …  …
27  3  3  3
```

وهكذا فإن الأمرين؛ `urnsamples(1:6,size=2,replace=T,ordered=T)` و `rolldie(2)` يكون لهما نفس الناتج.

ويمكن أيضا استخدام هذه الدالة للاختيار من ضمن عناصر تأخذ قيم نوعية غير رقمية، فمثلا إذا كان المطلوب هو حساب عدد الطرق الكلية لاختيار ثلاثة حالات من حالات الإنجاب في فراغ العينة S2={mm, mf, fm, ff} بدون إرجاع وبدون ترتيب نكتب:

```
> urnsamples(S2,size=3,replace=F,ordered=F)

[[1]]
[1] mm mf fm
Levels: ff fm mf mm

[[2]]
[1] mm mf ff
Levels: ff fm mf mm

[[3]]
[1] mm fm ff
Levels: ff fm mf mm
```

```
 [[4]]
[1] mf fm ff
Levels: ff fm mf mm
```

ويمكن اختيار فصيلتي دم من الفصائل الموجودة في فراغ العينة S3={A, B, AB, O}، بدون إرجاع وبالترتيب بعدد طرق هو؛

```
> urnsamples(S3,size=2,replace=F,ordered=T)
```

```
   X1 X2
1   A  B
2   B  A
3   A AB
4  AB  A
5   A  O
6   O  A
7   B AB
8  AB  B
9   B  O
10  O  B
11 AB  O
12  O AB
```

ولاحظ أن الاختلاف في طريقة عرض نتائج المعاينة بين S2 وS3، هو نتيجة اختلاف تصنيف كل منهما، فالأول هو إطار بيانات أما الثاني فهو متجه نوعي[1].

ويمكن استخدام الدالة nsamp لحساب عدد العناصر الناتج عن استخدام طرق العد المختلفة؛ بالإرجاع أو بدون إرجاع وبالترتيب أو بدون ترتيب، فمثلا:

```
> nsamp(n=5,k=2,replace=T,ordered=T)
[1] 25

> nsamp(n=4,k=2,replace=F,ordered=T)
[1] 12

> nsamp(n=3,k=3,replace=F,ordered=F)
[1] 1
```

حيث يمثل k عدد العناصر المطلوب سحبه من n عنصر كلي.

[1] المتجهات العاملية تأخذ أيضا نفس طريقة العرض الخاصة بالمتجهات النوعية.

2.1.5 تكوين فئات جزئية من فراغ العينة (Making Subsets of Sample Space)

عادة ما نرغب في التعامل مع جزء أو بعض الأحداث فقط من فراغ العينة وليس كل الكل، لهذا سنحتاج لتكوين فئات جزئية أحيانا من فراغ العينة، ويمكن عمل ذلك باستخدام الدالة subset مع دوال المعاينة. فمثلا، يمكن الحصول على الحالات التي يتساوى فيها وجهي زهري نرد بالصورة:

```
> subset(rolldie(2),X1==X2)
```

```
   X1 X2
1   1  1
8   2  2
15  3  3
22  4  4
29  5  5
36  6  6
```

وكذلك يمكن الحصول على كل الحالات التي يظهر فيها عدد أكبر من 4 في الرمية الثانية عند رمي زهر نرد مرتين بالصورة التالية:

```
> subset(rolldie(2),X1>4)
```

```
   X1 X2
5   5  1
6   6  1
11  5  2
12  6  2
17  5  3
18  6  3
23  5  4
24  6  4
29  5  5
30  6  5
35  5  6
36  6  6
```

وفي تجربة رمي عملة معدنية ثلاث مرات، يمكن تكوين فئة جزئية تضم كل الحالات التي يظهر فيها صورة في العملة الأولى والثانية (أو الرمية الأولى والثانية) كالتالي:

```
> subset(tosscoin(3),toss1=="H"&toss2=="H")
```

```
  toss1 toss2 toss3
1     H     H     H
5     H     H     T
```

وتوجد بعض الدوال التي قد تُستخدم للمساعدة في تكوين الفئات الجزئية من فراغ العينة مثل الدالة %in% والتي لها عدة استخدامات، منها تحديد جزء أو فئة من قيم متجه أو متغير، فمثلا لتكوين فئة تحتوي على الحالات التي يظهر فيها العددين 3 و4 في الرمية الأولى عند إلقاء زهري نرد نكتب:

```
> subset(rolldie(2),X1%in%3:4)
```

```
   X1 X2
3   3  1
4   4  1
9   3  2
10  4  2
15  3  3
16  4  3
21  3  4
22  4  4
27  3  5
28  4  5
33  3  6
34  4  6
```

ويمكن استخدام الدالة %in% كدالة منطقية للتعرف على إذا ما كانت العناصر في فئة معينة موجودة في فراغ العينة (أو في فئة أخرى)، فمثلا:

```
> y1<-c(2,4,5)
> y2<-c(1,3,4,5,7)

> y1%in%y2

[1] FALSE  TRUE  TRUE
```

ولاحظ أن النتيجة هي متجه منطقي يضم ثلاثة قيم (مناظرة لعدد القيم في y1) توضح وجود أو عدم وجود كل قيمة من قيم y1 في y2. ويمكن إضافة الدالة all للدالة %in% للتعرف على ما إذا كانت فئة جزئية ما هي جزء من فراغ العينة، حيث تكون الإجابة بنعم أو لا، كما يوضح المثال التالي:

```
> all(y1%in%y2)

[1] FALSE
```

والدالة isin يمكن أيضا أن تؤدي نفس عمل الدالة %in%، وسنتناول فيما يلي مثالا حول تطبيق هذه الدالة باستخدام نتيجة رمي عملتين وثلاث عملات، (والتي نعرضها في جدول (1.5) لتسهيل المقارنة):

جدول 1.5: فراغ العينة لتجربة رمي عملتين وثلاث عملات معدنية

```
> tosscoin(2)

  toss1 toss2
1     H     H
2     T     H
3     H     T
4     T     T
```

```
> tosscoin(3)

  toss1 toss2 toss3
1     H     H     H
2     T     H     H
3     H     T     H
4     T     T     H
5     H     H     T
6     T     H     T
7     H     T     T
8     T     T     T
```

في البداية إذا ما كتبنا:

```
> isin(tosscoin(2),tosscoin(3))

[1] FALSE FALSE FALSE FALSE
```

فهذا يعني أن الصفوف الأربعة الأولى في فراغ العينة الأول tosscoin(2) (والتي تتضمن عامودين) غير موجودة كما هي في فراغ العينة الثاني tosscoin(3) (والتي تتضمن ثلاثة أعمدة). أما إذا ما كتبنا:

```
> isin(tosscoin(2),tosscoin(3)[1,2])

[1]  TRUE  TRUE  TRUE FALSE
```

فهذا يعني أن البحث عن الصفوف الأربعة الأولى في tosscoin(2) سيكون في العامودين الأولين في tosscoin(3)، وبالتالي دلت النتيجة على وجود العناصر الثلاثة الأولى (HH، TH، وHT) من فراغ العينة الأول في نفس الموضع في فراغ العينة الثاني، أما العنصر الرابع (TT) فلم يكن في نفس الموضع.

3.1.5 بعض العمليات الأساسية على الفئات (Some Basic Operations on Sets)

من أهم العمليات التي يرتكز عليها حساب الاحتمالات هي الاتحاد، التقاطع، والفرق بين فئتين، وهذه العمليات يقابلها في لغة R الدوال union، intersect، وsetdiff على الترتيب.

وكمثال، لنعتبر أن فراغ العينة S4 يمثل الأحرف الأبجدية الإنجليزية، ولنقم بتعريف الأحداث A، B، وC على فراغ العينة S4 بالصورة التالية:

```
> S4<-letters
> S4

[1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m"
[14] "n" "o" "p" "q" "r" "s" "t" "u" "v" "w" "x" "y" "z"
```

```
> A<-letters[1:7]
> A

[1] "a" "b" "c" "d" "e" "f" "g"

> B<-letters[6:10]
> B

[1] "f" "g" "h" "i" "j"

> C<-letters[c(2,4,12,14)]
> C

[1] "b" "d" "l" "n"
```

والآن لنقم بتنفيذ العمليات التالية على تلك الفئات:

```
> union(A,B)

[1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"

> intersect(A,C)

[1] "b" "d"
> intersect(B,C)

character(0)

> setdiff(A,B)

[1] "a" "b" "c" "d" "e"
```

من جديد، في تجربة إلقاء عملة معدنية وزهر نرد معا يكون فراغ العينة، وليكن S5 معرف كالتالي:

```
> S5<-
data.frame(coin=c("H","H","H","H","H","H","T","T","T","T"
,"T","T"),die=c(1:6,1:6))

> S5

   coin die
1     H   1
2     H   2
3     H   3
4     H   4
5     H   5
6     H   6
7     T   1
8     T   2
```

```
9     T   3
10    T   4
11    T   5
12    T   6
```

وبتعريف الحدث A1 بأنه يمثل ظهور الصورة H في S5، والحدث A2 بأنه يمثل ظهور عدد فردي في S5 على النحو التالي؛

```
> A1<-subset(S5,coin=="H")
> A1
  coin die
1    H   1
2    H   2
3    H   3
4    H   4
5    H   5
6    H   6

> A2<-subset(S5,die==1|die==3|die==5)
> A2
   coin die
1     H   1
3     H   3
5     H   5
7     T   1
9     T   3
11    T   5
```

فإنه يمكن عندئذ حساب العمليات التالية:

```
> intersect(A1,A2)
  coin die
1    H   1
3    H   3
5    H   5

> setdiff(A2,A1)
   coin die
7     T   1
9     T   3
11    T   5
```

```
> union(A1,A2)

   coin die
1     H   1
3     H   3
5     H   5
7     T   1
8     H   2
9     T   3
10    H   4
11    T   5
12    H   6
```

4.1.5 حساب الاحتمالات للأحداث (Calculating Probabilities for Events)

بعد تكوين فراغ العينة وتعريف الأحداث المطلوبة عليه، عادة ما سنرغب في حساب الاحتمالات المناظرة لتلك الأحداث، ومن المعلوم أنه عند تعريف مجموعة من الأحداث المتنافية فيما بينها بحيث يشكل اتحادها فراغ العينة، ثم حساب الاحتمالات لها فإن ذلك يسمى تكوين توزيع احتمالي لتلك التجربة العشوائية. ومن أبسط الأمثلة على ذلك تجربة إلقاء عملة معدنية وتعريف الحدث A1 بأنه يمثل ظهور الصورة والحدث A2 بأنه يمثل ظهور الكتابة، فاتحاد الحدثان المتنافيان A1 و A2 يُشكل فراغ العينة للتجربة، ويكون P(A1) = 0.5 و P(A2) = 0.5.

ولحساب الاحتمالات المناظرة للتجارب المعروفة مثل رمي العملة المعدنية وزهر النرد يمكن استخدام الخيار makespace=T مع دوال تلك التجارب كالتالي:

```
> tosscoin(3,makespace=T)

  toss1 toss2 toss3 probs
1     H     H     H 0.125
2     T     H     H 0.125
3     H     T     H 0.125
4     T     T     H 0.125
5     H     H     T 0.125
6     T     H     T 0.125
7     H     T     T 0.125
8     T     T     T 0.125
```

وأيضا

```
> rolldie(1,makespace=T)

  X1     probs
1  1 0.1666667
2  2 0.1666667
3  3 0.1666667
```

```
4  4 0.1666667
5  5 0.1666667
6  6 0.1666667
```

وكذلك يمكن استخدام الدالة probspace بصورة عامة لتكوين التوزيع الاحتمالي بالشكل التالي:

```
> probspace(tosscoin(2))

  toss1 toss2 probs
1     H     H  0.25
2     T     H  0.25
3     H     T  0.25
4     T     T  0.25
```

وأيضا مع فراغ العينة الخاص بفصائل الدم؛

```
> S3

[1] "A"  "B"  "AB" "O"

> probspace(S3)

   x probs
1  A  0.25
2  B  0.25
3 AB  0.25
4  O  0.25
```

مع الأخذ بالاعتبار أن هذه الدوال السابقة تقوم فقط بتوزيع الاحتمالات بشكل متساوي "منتظم" على أحداث فراغ العينة. أما إذا رغبنا بتوزيع الاحتمالات بشكل غير متساوي على الأحداث فيمكن استخدام الخيار probs مع الدالة probspace كما يوضح المثال التالي لتجربة إنجاب لمرتين:

```
> S2
  Delivery
1       mm
2       mf
3       fm
4       ff
> probspace(S2,probs=c(1/8,3/8,3/8,1/8))

  Delivery probs
1       mm 0.125
2       mf 0.375
3       fm 0.375
4       ff 0.125
```

(والذي يعني أن احتمالات إنجاب توائم متطابقة هو أقل).

من جديد، يمكن استخدام الدالة[1] Prob لحساب الاحتمالات بالصورة التالية؛ لنفرض أننا نريد حساب احتمال الحدث A1 والذي يمثل ظهور الصورة في تجربة رمي عملة معدنية وزهر نرد (S5) السابقة، عندها يجب أولا كتابة فراغ العينة S5 كتوزيع احتمالي، وليكن باسم S6؛

```
> S6<-probspace(S5)
> S6
```

```
   coin die      probs
1     H   1 0.08333333
2     H   2 0.08333333
3     H   3 0.08333333
4     H   4 0.08333333
5     H   5 0.08333333
6     H   6 0.08333333
7     T   1 0.08333333
8     T   2 0.08333333
9     T   3 0.08333333
10    T   4 0.08333333
11    T   5 0.08333333
12    T   6 0.08333333
```

ثم نقوم بتكوين التوزيع الاحتمالي للحدث A1 من خلال تعريفه على S6 كالتالي:

```
> A1<-subset(S6,coin=="H")
```

```
> A1
```

```
  coin die      probs
1    H   1 0.08333333
2    H   2 0.08333333
3    H   3 0.08333333
4    H   4 0.08333333
5    H   5 0.08333333
6    H   6 0.08333333
```

الآن يمكن حساب احتمال الحصول على صورة في هذه التجربة؛

```
> Prob(A1)
```

```
[1] 0.5
```

ومن الممكن أيضا حساب احتمال الحدث A3 والذي يمثل ظهور عدد أكبر من 4 في زهر النرد في S6 كالتالي:

[1] لاحظ أن الحرف P في هذه الدالة مكتوب كحرف كبير بالإنجليزية.

```
> A3<-subset(S6,die>4)
> A3

   coin die      probs
5     H   5 0.08333333
6     H   6 0.08333333
11    T   5 0.08333333
12    T   6 0.08333333

> Prob(A3)

[1] 0.3333333
```

إضافة لحساب الاحتمال الاعتيادي، فإنه يمكن حساب الاحتمال الشرطي (Conditional Probability) أيضا باستخدام الخيار given مع الدالة Prob؛ لنفرض أنه في تجربة إلقاء زهري نرد تم تعريف الحدث B1 بأنه يمثل الحصول على المجموع 6 لوجهي الزهرين، والحدث B2 بأنه يمثل الحصول على عدد أكبر من أو يساوي 4 في الوجه الثاني (أو الرمية الثانية)، عندئذ يمكن حساب الاحتمال الشرطي P(B1\B2) كالتالي:

```
> S7<-rolldie(2,makespace=T)

> S7

   X1 X2      probs
1   1  1 0.02777778
2   2  1 0.02777778
3   3  1 0.02777778
…   …  … …

36  6  6 0.02777778

> B1<-subset(S7,X1+X2==6)
> B1

   X1 X2      probs
5   5  1 0.02777778
10  4  2 0.02777778
15  3  3 0.02777778
20  2  4 0.02777778
25  1  5 0.02777778

> B2<-subset(S7,X1>=4)
> B2

   X1 X2      probs
4   4  1 0.02777778
5   5  1 0.02777778
```

```
6   6  1 0.02777778
10  4  2 0.02777778
11  5  2 0.02777778
12  6  2 0.02777778
16  4  3 0.02777778
17  5  3 0.02777778
18  6  3 0.02777778
22  4  4 0.02777778
23  5  4 0.02777778
24  6  4 0.02777778
28  4  5 0.02777778
29  5  5 0.02777778
30  6  5 0.02777778
34  4  6 0.02777778
35  5  6 0.02777778
36  6  6 0.02777778

> Prob(B1,given=B2)

[1] 0.1111111
```

2.5 التوزيعات الاحتمالية المنفصلة (Discrete Probability Distributions)

سنتناول في هذا البند كيفية تعريف أو إدخال دوال التوزيعات الاحتمالية المنفصلة (Discrete Probability Distribution Functions, (PDF) في R وما يتعلق بها من طرق إضافة المتغيرات العشوائية للتوزيع، وحساب التوقع والتباين لها.

في البداية، لنقم بتعريف التوزيع الاحتمالي المنفصل بالصورة التالية؛

إذا كان X متغير عشوائي يأخذ القيم المنفصلة $x_1, x_2, \dots, x_n$ باحتمالات مناظرة $P(x_1), P(x_2), \dots, P(x_n)$ فإن الدالة $f(x) = P(X = x)$ تُعرّف بأنها دالة احتمالية منفصلة أو دالة كتلة احتمالية إذا كان:

$$\text{(1} \quad 0 \le P(X = x) \le 1 \ , \forall x \qquad \text{(2 و} \quad \sum_x P(X = x) = 1 \ .$$

والآن لنوضح كيفية تعريف متغير عشوائي أو أكثر على تجربة عشوائية في لغة R وإيجاد توزيعهم الاحتمالي عن طريق إضافة هذه المتغيرات إلى التوزيع الأصلي.

لنفرض أنه في تجربة إلقاء زهري نرد، (والذي تم تعيينه سابقا لفراغ العينة S7)، تم تعريف المتغير العشوائي X بأنه يمثل مجموع الوجهين، فإنه يمكن إضافة هذا المتغير باستخدام دالة إضافة المتغير العشوائي addrv كالتالي:

```
> S8<-addrv(S7,X=X1+X2)
> S8
```

```
   X1 X2  X      probs
1   1  1  2 0.02777778
2   2  1  3 0.02777778
3   3  1  4 0.02777778
…   …  …  … …
36  6  6 12 0.02777778
```

ويمكن أيضا إضافة متغير عشوائي آخر، وليكن Y الذي يمثل الفرق المطلق بين وجهي النرد بنفس الطريقة:

```
> S8<-addrv(S8,Y=abs(X1-X2))
> S8
```

```
   X1 X2  X Y      probs
1   1  1  2 0 0.02777778
2   2  1  3 1 0.02777778
3   3  1  4 2 0.02777778
…   …  …  … … …
36  6  6 12 0 0.02777778
```

الأن ننتقل لطرق تعريف التوزيعات الاحتمالية المنفصلة في R، وسنتناول الطريقتين الأسهل لعمل ذلك؛

- **الطريقة الأولى:**

وهي الطريقة المختصرة، وتتم باستخدام الدالة DiscreteDistribution التي تتطلب تحميل الحزمة الإضافية distrEx. وكمثال على تطبيق هذه الدالة، لنفرض أنه في تجربة إلقاء عملة معدنية ثلاث مرات تم تعريف المتغير العشوائي X1 بأنه يمثل عدد الصور الناتج، عندئذ سيكون التوزيع الاحتمالي للمتغير X1 كما هو موضح في الجدول (2.5). ويتم إدخال هذا الجدول باستخدام دالة التوزيع الاحتمالي المنفصل في R عن طريق إدخال كلا من قيم المتغير العشوائي والقيم الاحتمالية المناظرة؛

جدول 2.5: التوزيع الاحتمالي لعدد الصور في تجربة رمي ثلاث عملات معدنية

X1	0	1	2	3
P(x)	$1/8$	$3/8$	$3/8$	$1/8$

```
> library(distrEx)
```

```
> X1<-DiscreteDistribution(0:3,prob=c(1/8,3/8,3/8,1/8))
> X1
```

```
Distribution Object of Class: DiscreteDistribution
```

ولاحظ أن تعيين الاسم X1 لهذا التوزيع لا يعني عرض جدول التوزيع عند استدعاؤه، بل يعني أن X1 يمثل هذا التوزيع في ذاكرة البرنامج، ويمكن استخدامه لحساب بعض المقاييس الإحصائية الهامة للتوزيعات الاحتمالية بالشكل التالي:

```
> median(X1)
[1] 1

> mean(X1)

[1] NA
Warning message:
In mean.default(X1) : argument is not numeric or logical: returning NA

> E(X1)
[1] 1.5

> var(X1)
[1] 0.75

> sd(X1)
[1] 0.8660254

> IQR(X1)
[1] 1
```

ولاحظ أن بعض الدوال، (مثل الوسيط، التباين، الانحراف المعياري، والمدى الربيعي)، يمكن استخدامها مباشرة في الحساب، ودوال أخرى، (مثل الوسط الحسابي)، لا يمكن استخدامها بصيغتها المعروفة، ولذلك تم استخدام دالة التوقع E() كبديل.

ونود الإشارة هنا إلى أنه في حال كون الاحتمالات المناظرة للمتغير العشوائي متساوية، فإنه يمكن إدخال قيم المتغير العشوائي فقط دون إدخال الاحتمالات المناظرة، فمثلا في تجربة إلقاء عملة معدنية مرة واحدة، وتعريف المتغير العشوائي X2 بأنه يمثل عدد الصور،(X2: 0, 1)، فإنه يمكن عندها إدخال التوزيع الاحتمالي له بالشكل:

```
> X2<-DiscreteDistribution(0:1)
```

- **الطريقة الثانية:**

تتلخص هذه الطريقة بإدخال قيم المتغير العشوائي واحتمالاته المناظرة في متجهين منفصلين، ثم استخدام الدوال الحسابية الاعتيادية لحساب أي مقياس مطلوب. فعلى سبيل المثال، لنفرض أنه في تجربة إلقاء عملة معدنية ثلاث مرات تم تعريف المتغير العشوائي X3 بأنه يمثل الفرق بين عدد الصور والكتابة في كل حالة، كما هو موضح في الجدول (3.5) والذي يمثل التوزيع الاحتمالي للمتغير X3؛

جدول 3.5: التوزيع الاحتمالي للفرق بين عدد الصور والكتابة في تجربة رمي ثلاث عملات معدنية

X3	−3	−1	1	3
P(x)	$1/8$	$3/8$	$3/8$	$1/8$

عندها يمكن إدخال هذا التوزيع الاحتمالي بالشكل التالي مثلا:

```
> X3.value<-c(-3,-1,1,3)
> X3.prob<-c(1/8,3/8,3/8,1/8)
```

ولحساب التوقع والتباين مثلا يمكن كتابة:

```
> EX3<-sum(X3.value*X3.prob)

> EX3
[1] 0

> VarX3<-sum((X3.value-EX3)^2*X3.prob)
> VarX3

[1] 3
```

أما لحساب **دالة التوزيع الاحتمالي المتجمع** ((Cumulative Distribution Function, (CDF)) فيمكن استخدام الدالة cumsum مع قيم احتمالات التوزيع؛

```
> cumsum(X3.prob)

[1] 0.125 0.500 0.875 1.000
```

3.5 أهم التوزيعات المنفصلة الخاصة (Most Important Special Discrete Distributions)

سنتناول في هذا البند كيفية حساب الاحتمالات ضمن بعض التوزيعات الاحتمالية الشهيرة المعرّفة في لغة R، وكذلك كيفية توليد العينات التي تتبع هذه التوزيعات لغرض استخدامها في تطبيقات النماذج الإحصائية المختلفة. وسنعرض أهم هذه التوزيعات المنفصلة فيما يلي، ويمكن للقارئ استخدام الجدول (4.5) للتعرّف على صيغ دوال التوزيعات المنفصلة (والمتصلة أيضا) التي سنتناولها[1] والخيارات الإضافية التي تتضمنها.

1.3.5 التوزيع المنتظم المنفصل (Discrete Uniform Distribution)

يُعرّف التوزيع المنتظم المنفصل كالتالي:

إذا كان X متغير عشوائي منفصل يأخذ القيم $x_1, x_2, \ldots, x_k$ باحتمالات متساوية، فإن التوزيع المنتظم المنفصل يعرّف بالصورة:

[1] يمكن أيضا استخدام دالة المساعدة help للتعرّف على تفاصيل استخدام دوال التوزيعات الاحتمالية المنفصلة والمتصلة، وذلك عن طريق كتابة اسم الدالة ضمن أقواس دالة المساعدة، فمثلا لقراءة تفاصيل استخدام دالة التوزيع المنتظم يتم كتابة help(unif).

$$P(x;k) = \frac{1}{k} \quad , \quad x = x_1, x_2, \dots, x_k$$

ويقال أن المتغير X يتبع التوزيع المنتظم المنفصل بمعلمة[1] k، حيث k = max(X)-min(X).

ويكون $E(X) = \frac{\max(X)+\min(X)}{2}$, $Var(X) = \frac{(\max(X)-\min(X))^2}{12}$.

جدول 4.5: أهم دوال التوزيعات الاحتمالية المنفصلة والمتصلة في R

التوزيع الاحتمالي	الدالة في R	الخيارات الإضافية في الدالة
Uniform	`unif`	`min, max`
Binomial	`binom`	`size, prob`
Multinomial	`multinom`	`size, prob`
Geometric	`geom`	`prob`
Negative Binomial	`nbinom`	`size, prob`
Hyper-geometric	`hyper`	`m, n, k`
Poisson	`pois`	`lambda`
Normal	`norm`	`mean, sd`
Gamma	`gamma`	`shape, scale`
Beta	`beta`	`shape1, shape2`
Exponential	`exp`	`rate`
Student's t	`t`	`df`
Chi-squared	`chisq`	`df`
F	`f`	`df1, df2`

ويأخذ التوزيع المنتظم في R الصيغة `unif`، وبإضافة حروف (اختصارات) معينة في بداية الدالة يتم الحصول على نتائج متنوعة حول التوزيع الاحتمالي، وهذه القاعدة تسري في الواقع على معظم التوزيعات الاحتمالية كما سنرى.

فإضافة الحرف `d` (اختصار Density) إلى صيغة الدالة `unif` يتم للحصول على دالة الكتلة أو الكثافة الاحتمالية، فمثلا إذا كان المتغير العشوائي X يتبع توزيعا منتظما بمعلمة تساوي 3، أي $X \sim Uniform\ (x;\ 3)$، فهذا يعني أن احتمال الحصول على أي قيمة من قيم X هو $\frac{1}{3}$.

إلا أنه لحساب احتمال أي قيمة من قيم المتغير العشوائي المنتظم X في R يجب تعيين الخيارات الإضافية في الدالة والمتمثلة في القيمتين الصغرى `min` والكبرى `max` وإدخال قيمة X التي يُرغب بحساب الاحتمال عندها، وهكذا فإن الدالة `unif` تصلح للتعامل مع نوعي التوزيع المنتظم؛ المنفصل والمتصل.

[1] معالم التوزيع الاحتمالي هي القيم التي تحدد الشكل المميز لدالة التوزيع الاحتمالي.

ولنأخذ المثال التالي حول التوزيع المنتظم المنفصل؛ ليكن المتغير العشوائي X يمثل نتيجة الكشف عن فصيلة الدم، حيث يحوي فراغ العينة العناصر S = {A, B, AB, O}، والذي يمكن إعادة كتابته بالرموز على النحو S = {1, 2, 3, 4} ويمكن عندها حساب احتمال الحصول على فصيلة الدم B وهو P(X=2) مثلا بالصورة:

```
> dunif(2,min=1,max=4)

[1] 0.3333333
```

مع ملاحظة أنه يمكن الاستغناء عن كتابة min وmax ضمن الدالة وكتابة قيمهما مباشرة، أي كتابة dunif(2,1,4).

وإضافة الحرف p (اختصار Probability) إلى صيغة دالة unif فينتج عنه الحصول على دالة التوزيع أو الدالة التراكمية، فمثلا بحساب التوزيع التراكمي عند كل نقاط X للمثال السابق سنحصل على:

```
> punif(1,1,4)
[1] 0

> punif(2,1,4)
[1] 0.3333333

> punif(3,1,4)
[1] 0.6666667

> punif(4,1,4)
[1] 1
```

ومن البديهي أنه إذا ما تم حساب الاحتمال عند قيم خارجة عن النطاق الذي يأخذه المتغير العشوائي فإن النتيجة ستكون صفرا؛

```
> dunif(0,1,4)
[1] 0

> dunif(5,1,4)
[1] 0
```

أما إضافة الحرف r (اختصار Random) إلى صيغة دالة unif فيكون **لتوليد**[1] (Generate) متغير عشوائي يتبع التوزيع المنتظم، أي أن الدالة runif تمكّنك من توليد عينة عشوائية، (بالحجم الذي ترغب به)، تتبع توزيعا منتظما ضمن الفترة التي تحددها (عن طريق تحديد القيمتين الصغرى والكبرى). فمثلا إذا كنا نرغب بتوليد عينة عشوائية حجمها 10 مفردات ضمن الفترة من 0 إلى 6 فإننا نكتب:

[1] توليد البيانات أو المتغيرات العشوائية التي تتبع توزيعات احتمالية معروفة يدخل ضمن إطار ما يُعرف بعملية **المحاكاة** (Simulation) والتي سنتطرق لها في الفصل السابع.

```
> runif(10,0,6)

 [1] 5.7024384 0.6698918 1.3976509 2.9052041 3.1416223
 [6] 2.2855054 5.1962294 3.2520485 0.4040828 1.4049851
```

ولاحظ أنه عند توليد القيم العشوائية لأي توزيع احتمالي فإننا لا نتوقع أن تتكرر تلك القيم عند إعادة توليد العينة، لأن هذا ببساطة هو مفهوم العشوائية أو التوليد العشوائي للبيانات، فمثلا إذا ما تم إعادة تنفيذ الأمر السابق بصورة متكررة فإننا سنحصل في كل مرة على قيم مختلفة، لذلك سيحصل المستخدم على نتائج مختلفة عن تلك الموجودة هنا في الكتاب عند إجراء أي عدد من المحاولات، وسنعرض هنا بضعة محاولات لتوضيح هذه النقطة:

```
> runif(10,0,6)

 [1] 0.3755041 0.7968575 4.6104343 4.0867463 2.8280864
 [6] 1.6205483 4.9467057 2.9874406 0.1349558 2.1418043

> runif(10,0,6)

 [1] 1.6890823 5.0324993 3.5582828 3.7200646 4.2474008
 [6] 2.3610673 2.2151045 4.2443638 0.3337554 4.5190498

> runif(10,0,6)

 [1] 2.2892482 5.9807932 0.0558689 4.6572957 3.7397783
 [6] 2.8461401 2.6711614 1.7168447 5.7030258 5.0334469
```

- **استخدام الحِزمة** `distr` **لحساب المقاييس الإحصائية:**

يمكن حساب التوقع والتباين، (أو أي مقياس إحصائي آخر)، للتوزيع المنتظم، (أو أي توزيع احتمالي بصورة عامة)، بطريقة بسيطة وذلك عن طريق استدعاء الحزمة الإضافية `distr` وكتابة اسم التوزيع باستخدام الحرف الأول الكبير للاسم وكتابة معلمة أو معالم التوزيع بين الأقواس، فمثلا بالنسبة للمثال السابق الخاص بفصيلة الدم، والذي يأخذ فيه المتغير العشوائي القيمة الصغرى 1 والقيمة الكبرى 4، يمكننا التعرّف على دالة الكتلة الاحتمالية أولا ثم حساب بعض المقاييس كالتالي:

```
> library(distr)

> Unif(1,4)

Distribution Object of Class: Unif
 Min: 1
 Max: 4

> E(Unif(1,4)) #حساب التوقع
[1] 2.5
```

```
> var(Unif(1,4)) #حساب التباين
[1] 0.75

> median(Unif(1,4)) #حساب الوسيط
[1] 2.5
```

وتوجد طريقة أخرى للتعامل مع التوزيعات الاحتمالية، وذلك عن طريق تعيين اسم للدالة الاحتمالية ثم حساب الاحتمالات المطلوبة. فمثلا باستخدام المثال السابق يمكننا تعيين الاسم X.unif مثلا بالصورة:

```
> X.unif<-Unif(1,4)

> X.unif
Distribution Object of Class: Unif
 Min: 1
 Max: 4
```

ثم حساب الاحتمال المطلوب، فمثلا لحساب الاحتمال $P(X=2)$ نستخدم فقط الحرف d مع الاسم المعرّف يليه القيمة المطلوب حساب الاحتمال عندها بين قوسين كما يلي:

```
> d(X.unif)(2)
[1] 0.3333333
```

ولحساب الاحتمال التراكمي $P(X \leq 3)$ نستخدم الحرف p بالصورة:

```
> p(X.unif)(3)
[1] 0.6666667
```

وكذلك يمكن إيجاد التوقع والتباين، (وغيرها من المقاييس)، للتوزيع بالصورة:

```
> E(X.unif)
[1] 2.5

> var(X.unif)
[1] 0.75
```

أما لتوليد عينة بحجم معين من هذا التوزيع، (في الفترة من 1 إلى 4)، فنستخدم الحرف r، فمثلا لتوليد عينة بحجم 25 مشاهدة من التوزيع المنتظم السابق نكتب:

```
> r(X.unif)(25)

 [1] 2.956958 1.913858 2.556338 2.254612 2.177222
 [6] 3.817270 2.156815 1.264344 3.401928 2.259328
[11] 1.374336 2.438217 2.780839 3.968048 3.510004
[16] 2.521756 2.910577 1.201664 3.407034 2.475410
[21] 1.591480 3.345632 2.008336 1.349134 3.887102
```

ملاحظات:

1. يمكن توليد عينة منتظمة، (تتوزع بمسافات متساوية بين القيم)، بحيث يكون لها متوسط يساوي الصفر وتباين يقترب من الواحد الصحيح إذا ما تم استخدام الدالة unif مباشرة، فمثلا لتوليد عينة حجمها 7 مشاهدات بالخصائص المذكورة نكتب:

```
> unif(7)
[1] -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5
```

ولاحظ أن متوسط العينة يساوي الصفر تماما وتباينها يقترب من الواحد الصحيح:

```
> mean(unif(7));var(unif(7))

[1] 0
[1] 1.166667
```

2. الدالة sample يمكن استخدامها أيضا لتوليد عينات تتبع التوزيع المنتظم المنفصل، (أو بصيغة أخرى، لتوليد عينة عشوائية بسيطة). فمثلا لتوليد عينة عشوائية حجمها 20 مشاهدة من الفترة [15−5] نكتب:

```
> sample(5:15,size=20,replace=T)

 [1] 11  5 15 12 12  5  7 15  6 15  7 11 14 15 10 11
[17]  9 12 10  9
```

3. يمكن استخدام دالة sample بطرق أخرى لتوليد العينات العشوائية، فمثلا للحصول على عينة من المشاهدات العشوائية الناتجة عن إلقاء زهر نرد 200 مرة بالصورة:

```
> sample(6,size=200,replace=T)

  [1] 5 3 3 6 3 2 3 3 2 2 4 6 3 6 5 4 1 5 4 3 4 4 1
 [24] 1 3 6 6 2 5 1 3 6 4 1 1 3 3 5 4 1 4 1 4 1 5 5
 [47] 3 1 6 4 1 4 2 1 6 3 5 4 3 5 3 4 5 6 6 1 6 5 5
 [70] 4 5 1 6 6 6 1 5 6 2 1 2 3 3 4 1 2 2 2 6 6 4 5
 [93] 6 2 5 4 6 2 3 5 5 1 6 6 2 4 5 6 2 6 6 2 6 3 2
[116] 4 3 5 4 4 3 1 3 2 6 4 4 6 1 2 1 3 3 4 5 1 1 1
[139] 5 3 6 3 2 6 1 1 5 4 3 6 3 5 1 2 2 6 2 5 1 6 3
[162] 3 3 3 2 5 5 4 2 6 2 2 1 4 2 5 3 6 4 4 2 5 6 2
[185] 1 6 3 6 1 6 3 3 6 2 5 6 6 1 1 6
```

وكذلك يمكن الحصول على عينة من المشاهدات العشوائية الناتجة عن إلقاء عملة معدنية 50 مرة كالتالي:

```
> sample(c("H","T"),size=50,replace=T)

 [1] "H" "H" "T" "H" "H" "H" "H" "T" "H" "T" "T" "T"
[13] "H" "T" "T" "T" "T" "H" "T" "H" "H" "H" "T" "T"
[25] "H" "H" "H" "T" "T" "T" "T" "H" "T" "H" "H" "H"
[37] "T" "T" "H" "T" "T" "H" "H" "H" "H" "T" "H" "H"
[49] "H" "T"
```

ولاحظ أنه لابد من استخدام خيار الإرجاع replace=T في هذه الأمثلة لأن عدد العينات المطلوب سحبها (200 في المثال الأول و50 في المثال الثاني) هو أكبر من عدد العناصر المتاح (6 في المثال الأول و2 في المثال الثاني).

4. يمكن أيضا استخدام الدالة sample للمعاينة العشوائية من توزيعات لها احتمالات غير متساوية، (أي من توزيعات أخرى غير التوزيع المنتظم)، فمثلا باستخدام التوزيع الاحتمالي في جدول (3.5) يمكننا سحب عينات ذات حجم 1، 2 وحتى 4 بدون إرجاع، أما باستخدام الخيار replace=T فيمكن سحب أي عدد من العينات، (والتي ستكون عشوائية في كل مرة يتم فيها تنفيذ نفس الأمر)؛

```
> sample(X3.value,size=1,prob=X3.prob)

[1] -3

> sample(X3.value,size=2,prob=X3.prob)

[1] 3 1

> sample(X3.value,size=10,prob=X3.prob,replace=T)

 [1] -1  1  3  1  1  3  3  1  3  3
```

2.3.5 توزيع ذي الحدين (Binomial Distribution)

يُعرّف توزيع ذي الحدين كالتالي:

إذا تم تعريف المتغير العشوائي X بأنه يمثل عدد مرات النجاح في n محاولة، فإن توزيع المتغير X يعرّف بتوزيع ذي الحدين ويعطى بالصيغة:

$$P(x; n, p) = C_x^n \, p^x (1-p)^{n-x} \quad , \quad x = 0, 1, 2, \dots, n$$

حيث n, p هي معالم توزيع ذي الحدين؛ $0 \le p \le 1$، $n = 1, 2, 3, \dots$، و p يمثل احتمال النجاح في أي محاولة و $C_x^n = \frac{n!}{(n-x)!\, x!}$. ويكون $Var(X) = n\,p\,(1-p)$, $E(X) = n\,p$.

وكما رأينا في الجدول (4.5)، فإن دالة توزيع ذي الحدين لها الصيغة binom، وبإضافة الأحرف d، p، و r لصيغة الدالة الأصلية نحصل على دالة الكتلة، دالة التوزيع، ودالة توليد العينات العشوائية على الترتيب.

وكمثال على استخدام هذه الدالة، لنفرض أنه تم رمي زهر نرد 10 مرات وكان المتغير العشوائي X يمثل العدد الظاهر في أي رمية، عندها يكون $X \sim Binomial\ (x;\ 10,\ 1/6)$. وبالتالي إذا كان المطلوب هو ايجاد احتمال ظهور العدد 4 ثلاث مرات (أي حساب $P(X=3)$) مثلا فإن الحل يمكن إيجاده باستخدام الأمر dbinom(3,size=10,prob=1/6) أو باختصار:

```
> dbinom(3,10,1/6)
[1] 0.1550454
```

ولإيجاد الاحتمال $P(X \leq 5)$ ، أي احتمال ظهور الأعداد 5 فأقل، فإننا نستخدم الدالة التراكمية:

```
> pbinom(5,10,1/6)
[1] 0.9975618
```

ويمكننا الحصول على نفس الحل عن طريق جمع الاحتمالات المناظرة لقيم X من 0 إلى 5 كالتالي:

```
> sum(dbinom(0:5,10,1/6))
[1] 0.9975618
```

أما لإيجاد الاحتمال $P(2 \leq X \leq 5)$ مثلا، أي احتمال ظهور الأعداد من 2 إلى 5 فنستخدم الفرق بين الدالتين:

```
> pbinom(5,10,1/6)-pbinom(2,10,1/6)
[1] 0.222335
```

لأن $P(2 \leq X \leq 5) = P(X \leq 5) - P(X \leq 2)$. أو يمكن استخدام دالة الفرق diff لحساب الفرق بين الاحتمالين السابقين بالصورة:

```
> diff(pbinom(c(2,5),10,1/6))
[1] 0.222335
```

وكذلك لحساب $P(X > 4)$ مثلا يمكن أن نكتب:

```
> 1-pbinom(4,10,1/6)
[1] 0.01546197
```

أو نكتب:

```
> sum(dbinom(5:6,10,1/6))
[1] 0.01519445
```

ولاحظ وجود اختلاف طفيف جدا في النتيجتين الأخيرتين نظرا لاختلاف طريقة الحساب. وكما وضحنا سابقا، يمكن حساب أية مقاييس لهذا التوزيع، (باعتبار نفس المثال السابق)، بالصورة التالية:

```
> Binom(10,1/6)

Distribution Object of Class: Binom
```

```
 size: 10
 prob: 0.166666666666667

> E(Binom(10,1/6))
[1] 1.666667

> var(Binom(10,1/6))
[1] 1.388889
```

أو يمكن تعيين اسم للدالة Binom، وليكن X.binom مثلا، واستخدامه لحساب الاحتمالات والمقاييس المختلفة كما رأينا في أمثلة التوزيع المنتظم.

ملاحظة:

يمكن استخدام دالة توزيع ذي الحدين كدالة لتوزيع **بيرنوللي** (Bernoulli) وذلك بوضع الخيار size=1. وكمثال على ذلك؛ إذا كان المتغير العشوائي X يمثل ظهور الصورة في تجربة رمي عملة معدنية مرة واحدة، فإن احتمال عدم الحصول على صورة P(X=0) يمكن حسابه كالتالي:

```
> dbinom(0,1,1/2)
[1] 0.5
```

3.3.5 التوزيع متعدد الحدود (Multinomial Distribution)

التوزيع متعدد الحدود، (وهو الحالة العامة لتوزيع ذي الحدين)، يُعرّف كالتالي:

إذا كانت الأحداث $E_1, E_2, ..., E_k$ هي نتائج أي محاولة في التجربة العشوائية باحتمالات مناظرة $p_1, p_2, ..., p_k$ ، حيث $\sum_{i=1}^{k} p_i = 1$ ، لكل $p_i > 0$ ، $i = 1, 2, ..., k$ ، فإن المتغيرات العشوائية $X_1, X_2, ..., X_k$ والتي تمثل عدد مرات حدوث هذه النتائج في n محاولة مستقلة يكون لها توزيع احتمالي مشترك يعرف بالتوزيع متعدد الحدود بالمعالم n و $p_1, p_2, ..., p_k$ ، وتكون له الصيغة:

$$P(x_1, x_2, ..., x_k; p_1, p_2, ... p_k, n) = \binom{n}{x_1, x_2, ..., x_k} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

حيث $\binom{n}{x_1, x_2, ..., x_k} = \frac{n!}{x_1! x_2! ... x_k!}$ و $\sum_{i=1}^{k} x_i = n$.

ويكون $E(X_i) = n P_i$, $Var(X_i) = n P_i (1 - P_i)$, $i = 1, 2, ..., k$.

ويأخذ في لغة R الصيغة multinom، ولنأخذ المثال التالي كتطبيق للدالة؛

في تجربة إلقاء زهري نرد 6 مرات كان المطلوب هو حساب احتمال الحصول على مجموع للوجهين يساوي 7 أو 11 مرتين، ورقمين متساويين في الوجهين مرة واحدة، وعدم الحصول على الحدثين السابقين ثلاث مرات. لاحظ

أن هذا الاحتمال يحتوي على ثلاثة أحداث ينتج عنها ثلاثة متغيرات عشوائية تأخذ القيم X_1=2، X_2=1، وX_3=3، وهذه المتغيرات تناظرها الاحتمالات؛ $\frac{2}{9}$، $\frac{1}{6}$، و$\frac{11}{18}$ على الترتيب، أي أن الاحتمال المطلوب هو $P\left(2,1,3\,;\frac{2}{9},\frac{1}{6},\frac{11}{18},6\right)$، والذي يتم حسابه بالصورة:

```
> dmultinom(c(2,1,3),size=6,prob=c(2/9,1/6,11/18))
[1] 0.112703
```

ولتوليد عينة عشوائية مسحوبة من 10 عناصر مثلا من التوزيع متعدد الحدود السابق نكتب:

```
> rmultinom(c(2,1,3),size=10,prob=c(2/9,1/6,11/18))
```

```
     [,1] [,2]
[1,]    2    2
[2,]    1    2
[3,]    7    6
```

وننوه هنا بأن الدالة التراكمية للتوزيع المتعدد، (والتي من المفترض أن تُكتب بالصورة pmultinom)، هي غير متوفرة في الحزمة الافتراضية stats، وكذلك هو الحال في الحزمة الإضافية distr الذي لا تتوفر فيه الدالة Multinom، وذلك بحسب التحديث الحالي لهما.

4.3.5 التوزيع الهندسي (Geometric Distribution)

يُعرف التوزيع الهندسي كالتالي:

يقال أن المتغير العشوائي X يتبع توزيع هندسي بمعلمة p، إذا كانت دالة الكتلة الاحتمالية له على الصورة التالية:

$$P(x;p) = p\,(1-p)^{x-1} \quad ,x = 1,2,3,\ldots \quad ,0 \le p \le 1$$

ويكون $E(X) = \frac{1-p}{p}$, $Var(X) = \frac{1-p}{p^2}$.

ويُعرف التوزيع الهندسي في R بالصيغة geom، ولنأخذ السؤال التالي كتطبيق؛

إذا كان احتمال حدوث عطل بإحدى مضخات المياه في مشروع زراعي هو 0.02 خلال ساعة واحدة. فما هو احتمال عدم حدوث عطل في أي مضخة في هذا المشروع خلال الساعة أو الساعتين القادمتين؟. يمكن الإجابة على هذا السؤال كالتالي:

احتمال حدوث عطل خلال الساعة (X=1) أو الساعتين (X=2) القادمتين يعني؛

```
> dgeom(1,0.02)+dgeom(2,0.02)
[1] 0.038808
```

وبالتالي احتمال عدم حدوث عطل خلال الساعة أو الساعتين القادمتين يساوي:

```
> 1-(dgeom(1,0.02)+dgeom(2,0.02))
[1] 0.961192
```

ويمكن استخدام الدوال pgeom، وrgeom لحساب دالة التوزيع التراكمية وتوليد عينات من التوزيع الهندسي على الترتيب.

ويمكن حساب التوقع والتباين للتوزيع الاحتمالي في المثال السابق كالتالي:

```
> X.geom<-Geom(0.02)

> X.geom

Distribution Object of Class: Geom
 size: 1
 prob: 0.02

> E(X.geom)
[1] 49

> var(X.geom)
[1] 2450
```

5.3.5 توزيع ذي الحدين السالب (Negative Binomial Distribution)

يُعرف توزيع ذي الحدين السالب كالتالي:

يقال أن المتغير العشوائي X يتبع توزيع ذي الحدين السالب بالمعالم p وr إذا كانت له دالة الكتلة الاحتمالية التالية:

$$P(x;r,p) = C_{r-1}^{x-1}\, p^r\, (1-p)^{x-r} \quad ,x = r, r+1, r+2, \dots \quad ,r = 2,3,\dots \quad ,0 \le p \le 1$$

إذا ما تم تعريف المتغير العشوائي $Y = X - r$، والذي يمثل عدد مرات الفشل قبل الوصول للنجاح ذو الترتيب r، فإن التوزيع الاحتمالي للمتغير Y يكون:

$$P_Y(y) = C_{r-1}^{y+r-1}\, p^r\, (1-p)^y = C_y^{y+r-1}\, p^r\, (1-p)^y \quad , y = 0,1,2,\dots$$

ويكون $E(X) = \frac{r(1-p)}{p}$, $Var(X) = \frac{r(1-p)}{p^2}$.

وتأخذ دالة توزيع ذي الحدين السالب الصيغة nbinom، ويمكن استخدامها لحساب الاحتمالات من خلال دالة الكتلة dnbinom و دالة التوزيع التراكمية pnbinom، ويمكن أيضا توليد العينات العشوائية التي تتبع هذا التوزيع باستخدام الدالة rnbinom.

وكمثال على تطبيق هذه الدالة؛ لنفرض أن بعض الدراسات النفطية أثبتت أن احتمال الحصول على الغاز الطبيعي في إحدى المناطق الصحراوية من أي بئر يتم حفرها عشوائيا هو 0.2، فما هو احتمال الحصول على الغاز في المرة الثالثة وذلك عند حفر خمسة آبار؟.

من المثال نستطيع القول أن المتغير العشوائي X يمثل عدد المحاولات (حفر الآبار) حتى الوصول للنجاح r (والذي يمثل الحصول على الغاز) باحتمال 0.2، وبالتالي يمكن كتابة شكل الدالة بالصورة؛ P(x; 3, 0.2) حيث x = 3, 4, 5, ...، إلا أن الصيغة التي تعتمدها لغة R ترتكز على أن قيم المتغير العشوائي ستبدأ من الصفر وليس من r ، لذلك فإنه يتطلب إدخال عدد حالات الفشل قبل الوصول للنجاح r = 3 وهذا يعني $Y = X - r$ = 5 – 3 = 2 حالة فشل قبل الوصول إلى النجاح، حيث أن y = 0, 1, 2, ...، وبالتالي يكون:

```
> dnbinom(2,size=3,prob=0.2)
[1] 0.03072
```

وإذا كان المطلوب هو إيجاد احتمال الحصول على الغاز في المرة الثالثة في أقل من خمسة محاولات لحفر الآبار، فإن الحل يكون بحساب الاحتمال عند Y = 0 وعند Y = 1 ثم جمع هذين الاحتمالين:

```
> dnbinom(0,size=3,prob=0.2)+dnbinom(1,size=3,prob=0.2)
[1] 0.0272
```

ويمكن تعيين اسم للدالة Nbinom، واستخدامه لحساب الاحتمالات والمقاييس المختلفة. فالتوقع والتباين مثلا يمكن حسابهما لعدد الآبار التي يتم الحصول على الغاز الطبيعي منها في ثلاث محاولات كالتالي:

```
> X.nbinom<-Nbinom(3,0.2)

> X.nbinom

Distribution Object of Class: Nbinom
 size: 3
 prob: 0.2

> E(X.nbinom)
[1] 12

> var(X.nbinom)
[1] 60
```

6.3.5 التوزيع فوق الهندسي (Hyper-geometric Distribution)

يُعرّف التوزيع فوق الهندسي كالتالي:

يقال أن المتغير العشوائي X يتبع التوزيع فوق الهندسي بالمعالم k، n، وm إذا كانت دالة الكتلة الاحتمالية له معرّفة بالصورة التالية:

$$P(x;m,n,k)=\frac{C_x^m\,C_{k-x}^n}{C_k^{m+n}}$$

حيث $(k-x)\le n$, $x\le m$, $x=0,1,2,\dots,k$.

ويكون $Var(X)=k\,p\,(1-p)\left(\frac{m+n-k}{m+n-1}\right)$,$E(X)=k\,p$ حيث، $p=\frac{m}{m+n}$، و $0\le p\le 1$.

وتأخذ دالة التوزيع فوق الهندسي الصيغة hyper، بالمعالم m, n, k. وتُستخدم الدوال dhyper، phyper، وrhyper لإيجاد دالة الكتلة الاحتمالية، دالة التوزيع التراكمية، ولتوليد العينات العشوائية على الترتيب. ولنأخذ المثال التالي كتطبيق على هذه الدالة الاحتمالية؛

مخزن به 100 جهاز تكييف ياباني الصنع و 200 جهاز تكييف صيني الصنع. تم اختيار أربع أجهزة تكييف بصورة عشوائية بدون إرجاع من المخزن، أوجد احتمال:

1. أن تكون كل الأجهزة التي تم اختيارها يابانية الصنع.
2. أن يوجد جهازين على الأقل في العينة يابانية الصنع.

في هذا المثال يكون شكل دالة الكثافة P(x; 100, 200, 4)، وبالنسبة للمطلوب الأول P(x = 4) لدينا؛

```
> dhyper(4,m=100,n=200,k=4)
[1] 0.01185408
```

أما بالنسبة للمطلوب الثاني P($2\le x\le 4$) فيمكن أن يتم حسابه بالطريقة التالية؛

```
> phyper(4,m=100,n=200,k=4)-phyper(1,m=100,n=200,k=4)
[1] 0.4074057
```

أو بطريقة أخرى؛

```
> diff(phyper(c(1,4),m=100,n=200,k=4))
[1] 0.4074057
```

ولحساب التوقع والتباين للتوزيع السابق يمكن كتابة:

```
> X.hyper<-Hyper(100,200,4)
> X.hyper

Distribution Object of Class: Hyper
 m: 100
 n: 200
 k: 4
> E(X.hyper)
[1] 1.333333

> var(X.hyper)
[1] 0.8799703
```

7.3.5 توزيع بواسون (Poisson Distribution)

يُعرّف توزيع بواسون كالتالي:

إذا كان X متغير عشوائي يمثل عدد النتائج التي تحدث خلال فترة زمنية معينة أو ضمن نطاق محدد، فإنه يقال أنه يتبع توزيع بواسون بمعلمة λ إذا كانت دالة الكتلة الاحتمالية له معرّفة بالصورة:

$$P_X(x) = P(x;\ \lambda) = \frac{e^{-\lambda}\ \lambda^x}{x!}\ \ ,\ \ x = 0, 1, 2, \ldots\ \ ,\lambda > 0$$

ويكون $E(X) = \lambda\ \ ,\ \ Var(X) = \lambda$.

وتأخذ دالة توزيع بواسون الصيغة pois، بالمعلمة (الخيار) lambda. وتُستخدم الدالة dpois لإيجاد دالة الكتلة الاحتمالية، والدالة ppois، لإيجاد دالة التوزيع التراكمية، والدالة rpois لتوليد العينات العشوائية من توزيع بواسون. ولنأخذ المثال التالي كتطبيق على هذه الدالة الاحتمالية؛

في إحدى القرى، كان معدل الإصابة بمرض الأنفلونزا الموسمية من النوع A هو شخصين في الأسبوع الواحد. فأوجد احتمال أن يصاب في تلك القرية في أي أسبوع:

1. ثلاثة أشخاص. 2. أكثر من شخصين.

لدينا $X \sim \text{Poisson}(x;\ 2)$، وبالتالي فإن احتمال أن يصاب ثلاثة أشخاص $P(X = 3)$ يمكن حسابه بالصورة التالية:

```
> dpois(3,lambda=2)
[1] 0.180447
```

وأما احتمال أن يصاب أثر من شخصين $P(X > 2)$ فيمكن حسابه كالتالي:

```
> 1-ppois(2,lambda=2)
[1] 0.3233236
```

وعمليا لا داعي لحساب التوقع والتباين لتوزيع بواسون حيث أن قيمة كل منهما تساوي معلمة التوزيع λ، إلا أنه يمكن استخدام الدالة Pois لتنفيذ ذلك.

4.5 التوزيعات الاحتمالية المتصلة (Continuous Probability Distributions)

نأتي الآن للتعامل مع التوزيعات المتصلة والتي تعتمد حسابيا على إيجاد التكامل للدوال الاحتمالية، كما نرى من التعريف التالي:

إذا كان X متغير عشوائي متصل فإن دالة الكثافة الاحتمالية له $f(x)$ تكون دالة تتوفر فيها الشروط التالية:

1) $f(x) \leq 0$

2) $\int_{-\infty}^{\infty} f(x)dx = 1$ ، حيث $(\infty \, ، \, -\infty)$ هي الفترة الافتراضية التي ينتمي لها المتغير X.

3) المساحة تحت منحنى الدالة $f(x)$ من القيمة x_1 إلى القيمة x_2، والتي تساوي احتمال أن تقع قيمة المتغير X بين x_1 و x_2 ، يتم حسابها بالصورة:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx$$

ولنأخذ المثال التالي؛ إذا كان X متغير عشوائي متصل له الدالة التالية:

$$f(x) = \begin{cases} \frac{x}{2} \, ; 0 \leq x \leq 2 \\ 0 \, ; \quad \text{خلاف ذلك} \end{cases}$$

1. أثبت أن $f(x)$ هي دالة كثافة احتمالية.
2. أوجد $P(X > 1)$.
3. أوجد $P(0 \leq X \leq 1)$.
4. أوجد الدالة التراكمية $F(1)$.
5. أوجد التوقع والتباين للمتغير X.

(للوصول إلى الحل لا تعتمد على دوال جاهزة في R، سنقوم في هذا المثال أولا باستخدام دوال المُستخدِم الخاصة في R والتي يتم توضيح طريقة إنشاؤها بالتفصيل في البند (2.7) في الفصل السابع)، حيث يمكننا تعريف الدالة الاحتمالية للمتغير X في هذا المثال باستخدام الدالة `function` بالصورة التالية:

```
> fx1<-function(x)x/2
> fx1
function(x)x/2
```

1. إثبات أن $f(x)$ هي دالة كثافة احتمالية يعني إثبات أن تكامل الدالة في الفترة $0 \leq x \leq 2$ يساوي الواحد الصحيح، وهذا يمكن تنفيذه باستخدام دالة **التكامل** `integrate` كالتالي؛

```
> integrate(fx1,lower=0,upper=2)
1 with absolute error < 1.1e-14
```

ولاحظ ترتيب الإدخال في دالة التكامل، حيث يتم إدخال الدالة المطلوب تكاملها أولا يليها قيم الحد الأدنى والحد الأعلى للتكامل. ويتم عرض قيمة التكامل، والذي يساوي الواحد الصحيح في مثالنا، مصحوبا بقيمة الخطأ التقريبي للتكامل، والمحصلة إثبات أن $f(x)$ هي دالة كثافة احتمالية.

2. لإيجاد $P(X > 1)$ نكتب:

```
> integrate(fx1,lower=1,upper=2)

0.75 with absolute error < 8.3e-15
```

3. كذلك لإيجاد $P(0 \leq X \leq 1)$ نكتب:

```
> integrate(fx1,lower=0,upper=1)

0.25 with absolute error < 2.8e-15
```

4. لدينا $F(1) = P(X \leq 1)$ وبالتالي يتم تنفيذ نفس الأمر في المطلوب السابق، وتكون نتيجة الاحتمال هي 0.25 .

5. لحساب التوقع والتباين للمتغير X يمكن تعريف دوال جديدة خاصة بهما، كما كان الحال مع الدالة fx1، ثم استخدام دالة التكامل لحساب التوقع والتباين للدالة المطلوبة، إلا أننا سنستخدم هنا دالة أخرى لتنفيذ هذا المطلوب كما سنرى أدناه.

لاحظ في المثال السابق أن الدالة fx1 تم تعريفها في R على أنها مجرد دالة رياضية وليست دالة احتمالية، أما إذا ما أردنا تعريفها كدالة احتمالية متصلة فيمكن استخدام دالة التوزيع المتصل AbscontDistribution ضمن الحزمة الإضافية distr أو distrEx على النحو التالي، (بعد استدعاء تلك الحِزم إن لم يكن قد تم استدعاؤها مسبقا في الجلسة الحالية)؛

```
> fX1<-AbscontDistribution(d=fx1,low1=0,up1=2)
> fX1

Distribution Object of Class: AbscontDistribution
```

حيث تم تعيين الاسم fX1 للدالة الجديدة، ويتم ادخال اسم الدالة الرياضية الأصلية، (وهو fx1 في مثالنا)، متبوعا بالحد الأدنى والحد الأعلى للفترة المعرفة عليها الدالة الاحتمالية.

الآن يمكن استخدام الدالة الاحتمالية fX1 الآن لحساب المطلوب في المثال السابق كالتالي:

```
> p(fX1)(2) # المطلوب الأول
[1] 1
> p(fX1)(2)-p(fX1)(1) # المطلوب الثاني
[1] 0.75

> p(fX1)(1)-p(fX1)(0) # المطلوب الثالث
[1] 0.25

> p(fX1)(1) # المطلوب الرابع
[1] 0.25
```

```
> E(fX1);var(fX1) # المطلوب الخامس

[1] 1.332845
[1] 0.2225477
```

5.5 أهم التوزيعات المتصلة الخاصة (Most Important Special Continuous Distributions)

1.5.5 التوزيع المنتظم المتصل (Continuous Uniform Distribution)

ذكرنا في البند (1.3.5) كيف أن دالة التوزيع المنتظم في R يمكن استخدامها مع نوعي التوزيع المنتظم؛ المنفصل والمتصل. وبالتالي يكون الاختلاف في نوع القيم التي يأخذها المتغير العشوائي؛ فإما أن تكون قيم منفصلة أو قيم مُعرفة على فترة متصلة.

ويمكن تعريف المتغير العشوائي الذي يتبع التوزيع المنتظم المتصل بالصورة التالية:

يقال أن X متغير عشوائي يتبع التوزيع المنتظم المتصل في الفترة (a, b) والتي تمثل معالم التوزيع، إذا كانت دالة الكثافة الاحتمالية لـ X معرّفة بالصورة؛

$$f_X(x) = f(x; a, b) = \begin{cases} \frac{1}{b-a} : a \leq x \leq b \\ 0 \quad : \text{خلاف ذلك} \end{cases}$$

ولنأخذ المثال التالي حول تطبيق التوزيع المنتظم المتصل؛

في أحد المتاجر كان زمن قدوم الزبائن إلى الخزينة يتبع التوزيع المنتظم في الفترة (30، 0) دقيقة، حيث أنه خلال أي 30 دقيقة يصل زبون واحد إلى الخزينة. أوجد احتمال قدوم أي زبون في الخمسة دقائق الأخيرة خلال فترة الـ 30 دقيقة، كذلك أوجد التوقع والتباين لهذا التوزيع.

في هذه التجربة، يكون احتمال قدوم أي زبون في الفترة (30 ، 0) دقيقة، هو؛

$$f(x; 0, 30) = \frac{1}{30-0} = \frac{1}{30} \quad , \qquad 0 \leq x \leq 30$$

وبالتالي يكون احتمال قدوم أي زبون في الخمسة دقائق الأخيرة هو $P(25 \leq x \leq 30)$ والذي يتم إيجاده بحساب التكامل $\int_{25}^{30} \frac{1}{30}\, dx$، والذي يساوي بدوره التكامل $\int_{0}^{5} \frac{1}{30}\, dx$، لأن الاحتمال في التوزيع المنتظم هو ثابت لكل قيم المتغير العشوائي، ويتم حساب التكامل الأخير بالصورة:

```
> punif(5,0,30)
[1] 0.1666667
```

ويمكن أيضا حساب التوقع والتباين لهذا التوزيع كالتالي:

```
> X1.unif<-Unif(0,30)

> X1.unif

Distribution Object of Class: Unif
 Min: 0
 Max: 30

> E(X1.unif);var(X1.unif)

[1] 15
[1] 75
```

ولاحظ أنه إذا ما تم استخدام الدالة dunif(5,0,30) بدلا من تلك المستخدمة أعلاه فإن ذلك يعني حساب احتمال المتغير العشوائي عند نقطة محددة، وهذا يكون بمثابة استخدام التوزيع المنتظم المنفصل.

2.5.5 التوزيع الطبيعي (Normal Distribution)

يمكن تعريف التوزيع الاحتمالي الأكثر شهرة وأهمية في علم الإحصاء بالصورة التالية:

يقال أن المتغير العشوائي X يتبع التوزيع الطبيعي بالمعالم μ و σ^2، إذا كانت له دالة الكثافة الاحتمالية:

$$f_X(x) = f(x;\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

حيث $-\infty < x < \infty$ ، $-\infty < \mu < \infty$ ، $\sigma > 0$ ، و $\pi = 3.14$. ويكون $E(X) = \mu$, $Var(X) = \sigma^2$.

وإذا ما تم تعريف المتغير $Z = \frac{X-\mu}{\sigma}$، فإن Z سيتبع توزيع طبيعي معياري، وتكون دالة كثافته الاحتمالية معرّفة بالصورة:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad , \qquad -\infty < z < \infty$$

ويكون $E(Z) = 0$, $Var(Z) = 1$.

وتُعرف دالة التوزيع الاحتمالي للتوزيع الطبيعي المعياري وغير المعياري في لغة R بالصيغة norm وتمثل الدالة dnorm دالة كثافة التوزيع الطبيعي، والدالة pnorm دالة التوزيع التراكمي، أما الدالة rnorm فتستخدم لتوليد عينات من التوزيع الطبيعي. وتكون الخيارات أو المعالم المصاحبة لدالة التوزيع الطبيعي غير المعياري هي mean وsd.

وكما هو معلوم، فإن حساب أي احتمال للتوزيع الطبيعي يعني حساب التكامل أو حساب المساحة المحصورة بين قيمتين للمتغير العشوائي تحت منحنى التوزيع الطبيعي (أو المنحنى الناقوسي كما يطلق عليه)، وهذا يعني أننا عمليا سنستخدم الدالة pnorm لحساب ذلك.

الدالة norm بتركيباتها المختلفة تَعتبر افتراضيا أن mean=0 و sd=1 (أي أن التوزيع هو طبيعي معياري) ما لم يتم إدراج قيم مختلفة لها، وبالتالي عندما نكتب؛

```
> pnorm(0)
[1] 0.5
```

فهذا يعني حساب الاحتمال $P(Z \leq 0)$. أما استخدام دالة الكثافة dnorm لحساب الاحتمال عند النقطة $Z = 0$ فهذا يعني تنفيذ:

```
> dnorm(0)
[1] 0.3989423
```

أما إذا ما أردنا حساب الاحتمال لتوزيع طبيعي بقيم أخرى للمتوسط والانحراف المعياري، فيجب عندها إدخال هذه القيم كما يوضح المثال التالي:

إذا علمت أن درجات الطلبة في اختبارات أحد مقررات الدراسات العليا تتبع التوزيع الطبيعي بمتوسط 75 درجة وتباين 100 درجة، فأوجد احتمال أن يتحصل أحد الطلبة:

1. على درجة ما بين 80 و90 .
2. على درجة أعلى من 89.5 .

المطلوب الأول هو حساب الاحتمال $P(80 < X < 90)$ ، ويمكن تنفيذه كالتالي:

```
> pnorm(90,mean=75,sd=10)-pnorm(80,mean=75,sd=10)
[1] 0.2417303
```

أو تنفيذه باستخدام دالة الفرق:

```
> diff(pnorm(c(80,90),mean=75,sd=10))
[1] 0.2417303
```

والمطلوب الثاني هو حساب الاحتمال $P(X > 89.5)$ ، ويتم حسابه كالتالي:

```
> 1-pnorm(89.5,mean=75,sd=10)
[1] 0.07352926
```

مع ملاحظة أنه يمكن للقارئ كتمرين أن يقوم بتحويل المتغير الطبيعي X إلى Z وحساب الاحتمالات المطلوبة مباشرة بدون إدخال قيم المتوسط والانحراف المعياري.

وكمثال آخر، لنقم بحساب الاحتمالات الثلاثة الشهيرة في التوزيع $P(\mu-\sigma<X<\mu+\sigma)$، $P(\mu-2\sigma<X<\mu+2\sigma)$، و $P(\mu-3\sigma<X<\mu+3\sigma)$ والتي تضم 68%، 95%، و99.7% من قيم المتغير العشوائي الطبيعي على الترتيب، وهذه الاحتمالات السابقة تكافئ حساب الاحتمالات $P(-1<Z<1)$، $P(-2<Z<2)$، و $P(-3<Z<3)$ على الترتيب. ويمكن حسابها بالصورة التالية:

```
> diff(pnorm(c(-1,1)))
[1] 0.6826895

> diff(pnorm(c(-2,2)))
[1] 0.9544997

> diff(pnorm(c(-3,3)))
[1] 0.9973002
```

أو يمكن استخدام أمر واحد بالصورة:

```
> pnorm(1:3)-pnorm(-(1:3))

[1] 0.6826895 0.9544997 0.9973002
```

أما إذا ما أردنا حساب قيمة المتغير العشوائي الطبيعي المعياري Z عند قيمة احتمالية معينة، فيتم استخدام الدالة qnorm، (حيث q هو اختصار Quantile)، لتنفيذ ذلك، وكمثال على استخدام هذه التركيبة لنفرض أننا نريد إيجاد القيمة $Z_{0.05}$ فنقوم بكتابة:

```
> qnorm(0.05)
[1] -1.644854
```

وقد ظهرت الإشارة السالبة هنا لأن الدالة تعتبر كخيار افتراضي أن قيمة Z المطلوبة هي على يسار المنحنى الطبيعي (lower.tail=T)، أما إذا أردنا القيمة على يمين المنحنى فنقوم بتغيير الخيار الافتراضي كالتالي:

```
> qnorm(0.05,lower.tail=F)
[1] 1.644854
```

ويمكن حساب أكثر من قيمة للمتغير Z في أمر واحد عن طريق استخدام متجه لإدخال القيم، فمثلا يمكن حساب قيم $Z_{0.025}$، و$Z_{0.005}$ بالصورة:

```
> qnorm(c(0.025,0.005),lower.tail=F)

[1] 1.959964 2.575829
```

ويمكن أيضا تعيين اسم لدالة التوزيع الطبيعي واستخدامها لحساب الاحتمالات والتوقع والتباين والمقاييس المختلفة الأخرى، فمثلا يمكن كتابة:

```
> X.norm<-Norm(mean=0,sd=1)
> X.norm

Distribution Object of Class: Norm
 mean: 0
 sd: 1
```

وكذلك يمكن تعيين اسم لتوزيع طبيعي غير معياري، فمثلا للمثال السابق الخاص بدرجات الطلبة ($\mu = 75$ و$\sigma = 10$) يمكننا كتابة:

```
> X1.norm<-Norm(mean=75,sd=10)
> X1.norm

Distribution Object of Class: Norm
 mean: 75
 sd: 10
```

ويمكن حساب بعض المقاييس الإحصائية مثل:

```
> E(X1.norm)
[1] 75

> var(X1.norm)
[1] 100

> median(X1.norm)
[1] 75

> IQR(X1.norm)
[1] 13.4898
```

ويمكن أيضا تعريف أو تعيين دالة جديدة في دالة التوزيع الطبيعي وحساب المقاييس لها، فمثلا يمكننا تعريف Y = 3X بالصورة:

```
> Y1.norm<-3*X1.norm

> E(Y1.norm)
[1] 225

> var(Y1.norm)
[1] 900
```

وفي نفس السياق، يمكن تحويل التوزيع الطبيعي الاعتيادي لتوزيع طبيعي معياري (أي حساب **القيم المعيارية** (Z Scores)) بالصورة:

```
> Z1.norm<-(X1.norm-75)/10

> E(Z1.norm);var(Z1.norm)

[1] 0
[1] 1
```

أو يمكننا تنفيذ ذلك بصورة أكثر عمومية كالتالي:

```
> Z1.norm<-(X1.norm-E(X1.norm))/sd(X1.norm)

> E(Z1.norm);var(Z1.norm)

[1] 0
[1] 1
```

3.5.5 توزيع جاما (Gamma Distribution)

يقال أن المتغير العشوائي X يتبع توزيع جاما بالمعالم $\alpha > 0$ و $\beta > 0$ إذا كانت دالة الكثافة الاحتمالية له معرّفة بالصورة التالية:

$$f_X(x) = f(x;\alpha,\beta) = \frac{x^{\alpha-1}\, e^{-\frac{x}{\beta}}}{\beta^{\alpha}\,\Gamma(\alpha)} \quad , \quad 0 \le x < \infty$$

حيث $\Gamma(\alpha)$ تُعرف **بدالة جاما**، وتأخذ الصيغة $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\, e^{-x}\, dx$. ويكون التوقع والتباين للتوزيع معرف بالصورة؛ $E(X) = \alpha\,\beta \quad , \quad Var(X) = \alpha\,\beta^2$.

ويأخذ توزيع جاما في لغة R الصيغة gamma بالخيارات shape والتي تعادل المعلمة α، وscale التي تعادل المعلمة β. ولنأخذ المثال التالي كتطبيق؛

إذا كان معدل خروج الطلبة من لجنة الامتحانات النهائية في الساعة الواحدة هو طالبين. فما هو احتمال مرور من ساعتين إلى أربع ساعات قبل خروج ثلاثة طلبة؟، على افتراض أن الامتحانات تجري بفترات متلاحقة.

نُعرف المتغير العشوائي X بأنه يمثل زمن الخروج من الامتحان في هذا المثال والذي يتبع توزيع جاما بالمعالم α =3 و β = 2 ، ونستطيع حساب الاحتمال المطلوب $P(2 \le X \le 4)$ كالتالي:

```
> diff(pgamma(c(2,4),shape=3,scale=2))
[1] 0.2430222
```

ولحساب التوقع والتباين لهذا المثال نستخدم الدالة Gammad، (حيث أن الصيغة Gamma مُعرّفة لعملية أخرى في لغة R)، بالصورة التالية:

```
> X.gamma<-Gammad(3,2)
> X.gamma
```

```
Distribution Object of Class: Gammad
 shape: 3
 scale: 2

> E(X.gamma)
[1] 6

> var(X.gamma)
[1] 12
```

4.5.5 توزيع بيتا (Beta Distribution)

يقال أن المتغير العشوائي X يتبع توزيع بيتا بالمعالم $\alpha > 0$ و $\beta > 0$ إذا كانت دالة الكثافة الاحتمالية له مُعرّفة بالصورة التالية:

$$f_X(x) = f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad , \quad 0 \le x \le 1$$

حيث $B(\alpha,\beta)$ هي **دالة بيتا** وتعرّف بالصورة:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}\,(1-x)^{\beta-1}\,dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

وحيث $\Gamma(\cdot)$ هي دالة جاما. ويأخذ المتغير X التوقع والتباين؛

$$E(X) = \frac{\alpha}{\alpha+\beta} \quad , \quad Var(X) = \frac{\alpha.\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}$$

حيث $\alpha > 0$ و $\beta > 0$.

ويأخذ توزيع بيتا في R الصيغة beta بالخيارات shape1 والتي تعادل المعلمة α، وshape2 التي تعادل المعلمة β. ولنأخذ المثال التالي؛

إذا كان X متغير عشوائي يمثل نسبة كمية المياه التي يتم ضخها إلى مجمع سكني في إحدى الضواحي كل أسبوع، وكان يتبع توزيع بيتا بالمعالم $\alpha = 4$ و $\beta = 2$، فأوجد احتمال أن يتم ضخ 90% على الأقل من كمية المياه خلال أحد الأسابيع.

الاحتمال المطلوب هو $P(X \ge 0.90)$، والذي يمكننا حسابه بإحدى الطريقتين التاليتين:

```
> 1-pbeta(0.9,shape1=4,shape2=2)
[1] 0.08146
```

أو

```
> diff(pbeta(c(0.9,1),shape1=4,shape2=2))
[1] 0.08146
```

حيث أن الحد الأعلى للتكامل في دالة بيتا هو الواحد الصحيح. ولحساب التوقع والتباين نقوم بتنفيذ التالي:

```
> X.beta<-Beta(shape1=4,shape2=2)
> X.beta

Distribution Object of Class: Beta
 shape1: 4
 shape2: 2
 ncp: 0

> E(X.beta)
[1] 0.6666667

> var(X.beta)
[1] 0.03174603
```

والخيار ncp هو لتحديد ما إذا كان التوزيع مركزي أم غير مركزي[1].

5.5.5 التوزيع الأسي (Exponential Distribution)

يُعرّف التوزيع الأسي بالصورة التالية، حيث يقال أن المتغير العشوائي X يتبع توزيعا أسيا بالمعلمة β > 0 إذا كان له دالة الكثافة الاحتمالية المعرّفة بالصورة:

$$f_X(x) = f(x;\beta) = \frac{1}{\beta}\, e^{-\frac{x}{\beta}} \quad , \quad 0 \le x < \infty$$

ويكون $E(X) = \beta \quad , \quad Var(X) = \beta^2$.

وصيغة التوزيع الأسي في R هي exp بالخيار rate والذي يعادل المعلمة[2] $\frac{1}{\beta}$ في التوزيع. ويمكن أن نأخذ المثال التالي للتطبيق؛

في إحدى الشركات المزودة لخدمات الانترنت تم تعريف دخول المشتركين على موقع الشركة بأنه يعتبر عملية بواسون بمتوسط 25 دخول في الساعة الواحدة. أوجد احتمال أن لا يتم أي دخول إلى موقع الشركة خلال فترة زمنية طولها 6 دقائق.

إذا ما اعتبرنا أن X هو متغير عشوائي يمثل الزمن (بالساعات) من بداية الفترة الزمنية إلى أول دخول للموقع، في هذه الحالة يكون للمتغير X توزيعا أسيا بمعلمة $\beta = 25$.

---

[1] يمكن مراجعة دالة المساعدة الخاصة بدالة بيتا، help(Beta) للمزيد من المعلومات حول ذلك الخيار.

[2] دالة التوزيع الأسي في R تعتمد على الصيغة؛ $f_X(x) = f(x;\lambda) = \lambda\, e^{-\lambda x} \quad , \quad 0 \le x < \infty$ ، وهذا يعني أن معلمة التوزيع ستكون في هذه الحالة $\lambda = {}^{1}/_{\beta}$.

الآن نحن مهتمون بإيجاد احتمال أن يأخذ X قيمة تفوق الستة دقائق، وحيث أنه يتم قياس وحدات المتغير X بالساعات فتكون 6 دقائق = 0.1 ساعة، وهكذا فإن المطلوب هو $P(X > 0.1)$، وذلك يمكن تنفيذه كالتالي:

```
> 1-pexp(0.1,rate=1/25)
[1] 0.996008
```

أو باستخدام الخيار lower.tail=F؛

```
> pexp(0.1,rate=1/25,lower.tail=F)
[1] 0.996008
```

ويمكن حساب التوقع والتباين بالصورة:

```
> X.exp<-Exp(rate=1/25)
> X.exp

Distribution Object of Class: Exp
 rate: 0.04

> E(X.exp)
[1] 25
> var(X.exp)
[1] 625
```

6.5.5 توزيع استيودنت *t* (Student's *t* Distribution)

إذا كان X متغير عشوائي يتبع توزيع t بدرجات حرية $\gamma = n - 1$ فإن دالة الكثافة الاحتمالية له تُعرّف بالصورة:

$$f_X(x) = k\left(1 + \frac{x^2}{\gamma}\right)^{-(\gamma+1)/2}, \quad -\infty < x < \infty$$

حيث k هو ثابت تحدد قيمته بحيث يكون $\int_{-\infty}^{\infty} f_X(x)dx = 1$. ويمكن أن تكون قيمة k مساوية للمقدار $\frac{\Gamma((\gamma+1)/2)}{\sqrt{\gamma\pi}\,\Gamma(\gamma/2)}$ ، ويكون $E(X) = 0$, $Var(X) = \frac{\gamma}{\gamma-2}$ حيث $\gamma > 2$.

وعن علاقة توزيع t بالتوزيع الطبيعي، فتنص النظرية على أنه إذا كان $\bar{X}$ وS^2 هما، على الترتيب، الوسط الحسابي والتباين لعينة عشوائية حجمها n مسحوبة من مجتمع يتوزع بتوزيع طبيعي له المتوسط μ والتباين σ^2 فإن المتغير العشوائي

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

سيتبع توزيع t بدرجات حرية $\gamma = n - 1$ ، حيث γ هي معلمة التوزيع.

ويأخذ توزيع t في R الصيغة t بالخيار df والذي يمثل درجة الحرية أو معلمة التوزيع γ، ولنأخذ الأمثلة التالية؛

إذا كان المتغير العشوائي X يتبع توزيع t بمعلمة γ فأوجد القيم التالية:

1. $t_{0.05}$ إذا كانت $n = 10$. 2. $t_{0.025}$ إذا كانت $n = 5$.

لإيجاد قيم المتغير X عند احتمال ودرجة حرية محددين نستخدم الدالة qt، مع الخيار lower.tail=F للحصول على القيمة الموجبة؛

```
> qt(0.05,df=9,lower.tail=F) # المطلوب الأول
[1] 1.833113
> qt(0.025,df=4,lower.tail=F) # المطلوب الثاني
[1] 2.776445
```

أما للبحث عن القيمة الاحتمالية α باستخدام قيمة المتغير العشوائي ودرجة حرية التوزيع فنستخدم دالة التوزيع التراكمية pt بالخيار lower.tail=F لأن الاحتمال المحصور تحت منحنى توزيع t يكون إلى اليمين، وكمثال لنفرض أننا في جدول t نود حساب القيمة الاحتمالية المناظرة للقيمة 1.714 عند درجة حرية 23، عندها نكتب؛

```
> pt(1.714,df=23,lower.tail=F)
[1] 0.04998802
```

والتي تساوي بعد التقريب $\alpha = 0.05$.

ولحساب المقاييس المختلفة، ومن ضمنها التوقع والتباين، لتوزيع t نقوم بتعيين اسم للدالة Td عند درجة الحرية المطلوبة كما يوضح المثال التالي:

```
> X.t<-Td(df=9)
> X.t

Distribution Object of Class: Td
 df: 9
 ncp: 0

> E(X.t)
[1] 0

> var(X.t)
[1] 1.285714
```

وكما هو الحال مع معظم التوزيعات المنفصلة والمتصلة، فإنه يمكن استخدام الدالة dt لحساب الاحتمال عند نقطة معينة بدرجة حرية ما، والدالة rt لتوليد عينات تتبع توزيع t، وذلك بدون استخدام الخيار lower.tail=F.

7.5.5 توزيع مربع كاي (Chi-Square Distribution)

يمكن تعريف توزيع مربع كاي (χ^2) بالشكل التالي:

إذا كان X متغير عشوائي يتبع توزيع مربع كاي بدرجات حرية γ فإن دالة كثافته الاحتمالية تعرّف بالصورة التالية:

$$f_X(x) = k\,(x)^{\frac{\gamma}{2}-1}\,e^{-x/2} \quad , \quad 0 \le x \le \infty$$

و k ثابت تحدد قيمته بحيث يكون $\int_0^\infty f_X(x)dx = 1$. ويمكن أن تكون قيمة k مساوية للمقدار $\frac{1}{\Gamma\left(\frac{\gamma}{2}\right)2^{\gamma/2}}$ ، ويكون $E(X) = \gamma \ , Var(X) = 2\,\gamma$.

وتوزيع مربع كاي يأخذ الصيغة chisq في R بالخيار df والذي يمثل درجة الحرية أو معلمة التوزيع γ، ولنأخذ الأمثلة التالية؛

أوجد القيم التالية للمتغير العشوائي X، والذي يتبع توزيع مربع كاي χ^2 :

1) $\chi^2_{0.025}$ إذا كان $n = 12$. 2) $\chi^2_{0.99}$ إذا كان $n = 26$.

لإيجاد قيم المتغير X عند احتمال ودرجة حرية محددين نستخدم الدالة qchisq، مع الخيار lower.tail=F لأن توزيع مربع كاي غير متماثل ولا يأخذ قيم سالبة مثل توزيع t أو التوزيع الطبيعي، ويتم حساب القيم أو الاحتمالات إلى يمين منحنى التوزيع؛

```
> qchisq(0.025,df=11,lower.tail=F) # المطلوب الأول
[1] 21.92005
```

ولاحظ أنه إذا لم يتم استخدام الخيار lower.tail=F مع الدالة qchisq فإننا سنحصل على قيمة المتغير التي تعطي الاحتمال 0.025 إلى يسار المنحنى في هذا المثال كما نرى؛

```
> qchisq(0.025,df=11)
[1] 3.815748
```

وبالمثل للمطلوب الثاني، فإننا نكتب؛

```
> qchisq(0.99,df=25,lower.tail=F) # المطلوب الثاني
[1] 11.52398
```

وللبحث عن القيمة الاحتمالية α باستخدام قيمة المتغير العشوائي ودرجة حرية التوزيع نستخدم دالة التوزيع التراكمية pchisq بالخيار lower.tail=F، وكمثال لنفرض أننا في جدول مربع كاي نود حساب القيمة الاحتمالية المناظرة للقيمة 24.74 عند درجة حرية 13، عندها نكتب؛

```
> pchisq(24.74,df=13,lower.tail=F)
[1] 0.02496698
```

والتي تساوي بعد التقريب $\alpha = 0.025$.

ولحساب التوقع والتباين وغيرها من المقاييس لتوزيع مربع كاي، نقوم بتعيين اسم للدالة Chisq عند درجة الحرية المطلوبة، فمثلا عند درجة حرية 13 يمكننا تعيين:

```
> X.chisq<-Chisq(13)
> X.chisq

Distribution Object of Class: Chisq
 df: 13
 ncp: 0
```

ومن ثمة حساب:

```
> E(X.chisq)
[1] 13

> var(X.chisq)
[1] 26
```

وتُستخدم الدالتين dchisq وrchisq على الترتيب لحساب كلا من الاحتمال عند نقطة ما بدرجة حرية معينة، ولتوليد عينات تتبع توزيع مربع كاي، وذلك بدون استخدام الخيار lower.tail=F.

فمثلا لحساب الاحتمال $P(X = 7)$ حيث X يتبع توزيع مربع كاي بدرجات حرية 21 نكتب:

```
> dchisq(7,df=21)
[1] 0.001964487
```

ولتوليد عينة عشوائية حجمها 25 مشاهدة مثلا من توزيع مربع كاي بدرجات حرية 21 نكتب:

```
> rchisq(25,df=21)

 [1] 19.910920 13.758032 18.443947 15.225653 19.072101
 [6] 23.577535 14.309176 18.965896 13.241045 29.483919
[11] 19.295767 19.650266 18.658548 32.179373 31.231335
[16] 19.327331 12.739611 18.723772 19.045722  9.622548
[21] 28.141897 19.077087 24.604474 23.967909 21.844188
```

8.5.5 توزيع فيشر F (Fisher's F Distribution)

يمكن تعريف توزيع F بالشكل التالي:

إذا كان X متغير عشوائي يتبع توزيع فيشر F بالمعالم γ_1 و γ_2 فإن دالة الكثافة الاحتمالية له تعرف بالصورة التالية:

$$f_X(x) = \frac{\Gamma((\gamma_1+\gamma_2)/2)}{\Gamma(\gamma_1/2)\Gamma(\gamma_2/2)} \left(\frac{\gamma_1}{\gamma_2}\right)^{\frac{\gamma_1}{2}} (x)^{\frac{\gamma_1}{2}-1} \left(1+\frac{\gamma_1}{\gamma_2}x\right)^{-(\gamma_1+\gamma_2)/2} \quad , \quad 0 \le x < \infty$$

ويعرّف توقعه وتباينه بالصورة:

$$E(F) = \frac{\gamma_2}{\gamma_2 - 2} \quad , \quad \gamma_2 > 2$$

و

$$Var(F) = \frac{2\,{\gamma_2}^2\,(\gamma_1+\gamma_2-2)}{\gamma_1(\gamma_2-2)^2\,(\gamma_2-4)} \quad , \quad \gamma_2 > 4$$

ويأخذ توزيع F الصيغة `f` في لغة R بالخيارين `df1` و`df2` واللذان يمثلان درجتي الحرية الأولى والثانية γ_1 وγ_2 على الترتيب.

ولإيجاد قيم المتغير X الذي يتبع توزيع F عند احتمال ودرجات حرية محددة، نستخدم الدالة `qf`، مع الخيار `lower.tail=F` كما هو الحال مع توزيع مربع كاي، ويتم حساب القيم أو الاحتمالات إلى يمين منحنى التوزيع. وكمثال لنقم بحساب قيم F عند القيمة الاحتمالية 0.05 ودرجات حرية 2 و4:

```
> qf(0.05,df1=2,df2=4,lower.tail=F)
[1] 6.944272
```

ولحساب القيمة الاحتمالية α باستخدام قيمة المتغير العشوائي ودرجات الحرية نستخدم دالة التوزيع التراكمية `pf` مع الخيار `lower.tail=F`، وكمثال من جدول F، لنقم بحساب القيمة الاحتمالية المناظرة للقيمة 9.36 عند درجات حرية 5 و4؛

```
> pf(9.36,df1=5,df2=4,lower.tail=F)
[1] 0.0250211
```

وهي القيمة الاحتمالية 0.025 بعد التقريب.

أما لحساب التوقع والتباين وغيرها من المقاييس لتوزيع F، فنقوم بتعيين اسم للدالة `Fd` عند درجات الحرية المطلوبة، فمثلا عند درجات حرية 5 و4 يمكننا تعيين:

```
> X.f<-Fd(5,4)
> X.f

Distribution Object of Class: Fd
```

```
 df1: 5
 df2: 4
 ncp: 0
```

ثم حساب التوقع والتباين؛

```
> E(X.f)
[1] 2

> var(X.f)
[1] NA
```

ولاحظ عدم إمكانية حساب التباين في هذا المثال نظرا لأن درجة الحرية الثانية ليست أكبر من القيمة 4، وكمثال آخر لنقم بتعريف الدالة التالية، بنفس الاسم X.f ثم حساب التوقع والتباين؛

```
> X.f<-Fd(7,9)

> E(X.f)
[1] 1.285714

> var(X.f)
[1] 1.322449
```

ويمكن استخدام الدالة df لحساب الاحتمال عند نقطة معينة بدرجات حرية محددة، والدالة rf لتوليد عينات تتبع توزيع F، وذلك بدون استخدام الخيار lower.tail=F.

6.5 حساب العزوم (Calculating Moments)

يمكن تعريف العزوم لأي متغير بالشكل التالي:

إذا كانت $x_1, x_2, \dots, x_N$ هي مشاهدات المتغير X التي لها الوسط الحسابي $\bar{X}$، فإن العزوم يمكن أن تقسم (حسب النقطة التي يحسب مقدار الابتعاد عنها) إلى القسمين التاليين:

- **العزوم حول الصفر، (أو حول نقطة الأصل)** (Moments about zero)

حيث يعرف العزم الرائي (r^{th} moment) حول الصفر بالصيغة:

$$\acute{\mu}_r = \frac{\sum_{i=1}^{N} x_i^r}{N} \quad , \qquad r = 1,2,3, \dots$$

ويتم التعويض في قانون العزوم بقيم r حيث ($r = 1,2,3, \dots$) فنحصل على العزم الأول، الثاني، الثالث، ... وهكذا.

- **العزوم حول الوسط الحسابي** (Moments about mean)

حيث يعرف العزم الرائي حول الوسط الحسابي بالصيغة:

$$\mu_r = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^r}{N} \quad , \qquad r = 1,2,3, \dots$$

ويمكن استخدام دالة العزوم moment في R عن طريق تحميل الحزمة الإضافية e1071، والتي تمكننا من حساب العزوم حول الصفر أو حول الوسط الحسابي لأي مجموعة من القيم كما يوضح المثال التالي:

أوجد العزوم الأربعة الأولى حول الصفر أولا، وحول الوسط الحسابي ثانيا، وذلك للبيانات التالية المفردة التالية 10، 8، 7، 3، 2.

الخيار order في دالة العزوم moment هو لاختيار ترتيب العزم المطلوب، والخيار center هو لتحديد ما إذا كانت العزوم هي حول الصفر أو حول الوسط الحسابي، فلحساب العزوم الأولى حول الصفر نكتب:

```
> library(e1071)

> moment(c(10,8,7,3,2),1,center=F)
[1] 6

> moment(c(10,8,7,3,2),2,center=F)
[1] 45.2

> moment(c(10,8,7,3,2),3,center=F)
[1] 378

> moment(c(10,8,7,3,2),4,center=F)
[1] 3318.8
```

ملاحظات:

1. يمكننا عوضا عن تكرار نفس الأمر السابق للحصول على العزوم المتتالية، استخدام إحدى دوال الحلقات (Loop functions) في R والتي تساعدنا في تكرار نفس الأمر مع تغيير الترتيب بشكل متتالي. ومن أشهر تلك الدوال هي دالة for، كما سنرى:

لحساب العزوم المتتالية حول الوسط الحسابي للمثال السابق باستخدام تكرار الأمر يمكننا كتابة الأمر التالي، (مع ملاحظة استخدام الخيار center=T):

```
> for(i in 1:4)print(moment(c(10,8,7,3,2),i,center=T))

[1] 0
[1] 9.2
[1] -3.6
[1] 122
```

حيث أن الأمر for(i in 1:4) يعني تماما؛ من $i = 1$ إلى 4، والدالة print هي الدالة التي تقوم بطباعة كل ما يليها، وهكذا فإن الناتج هو العزوم الأربعة الأولى حول الوسط الحسابي كما هو مطلوب.

2. يمكن للمُستخدم أيضا تعريف دالة خاصة للعزوم بحيث تتضمن إدخال قيم المتغير وترتيب العزوم المطلوب حسابها واستخدام دالة for معها لحساب العزوم المتتالية، كما سنوضح:

لحساب العزوم الأربعة الأولى حول الوسط الحسابي، نقوم أولا بتعريف الدالة ولتكن باسم mom.mean:

```
> mom.mean<-function(x,r)moment(x,r,center=T)
```

ثم نستخدم دالتي for وprint مع إدخال قيمة x والتي تمثل هنا في الدالة قيم المتغير، ويتم إدخال ترتيب العزوم r تلقائيا ضمن سطر الأمر؛

```
> for(r in 1:4)print(mom.mean(x=c(10,8,7,3,2),r))
```

```
[1] 0
[1] 9.2
[1] -3.6
[1] 122
```

ويمكننا حساب العزوم للمتغيرات العشوائية المنفصلة عن طريق إدخال قيم المتغير بدون تكرارات كما يوضح المثال التالي:

من الجدول (2.5) السابق، والذي يتضمن التوزيع الاحتمالي للمتغير X1 الذي يمثل عدد الصور الناتج في تجربة إلقاء عملة معدنية ثلاث مرات، يمكننا حساب العزوم الخمسة الأولى حول الصفر للمتغير العشوائي بالصورة التالية:

```
> for(i in 1:5)print(moment(c(0,1,1,1,2,2,2,3),i,center=F))
```

```
[1] 1.5
[1] 3
[1] 6.75
[1] 16.5
[1] 42.75
```

وكذلك حساب العزوم الخمسة الأولى حول الوسط الحسابي بالصورة:

```
> for(i in 1:5)print(moment(c(0,1,1,1,2,2,2,3),i,center=T))
```

```
[1] 0
[1] 0.75
[1] 0
[1] 1.3125
[1] 0
```

1.6.5 العزوم والدالة المولدة للعزوم للتوزيعات الخاصة

(Moments and MGF for Special Distributions)

تعرّف الدالة المولدة للعزوم للمتغير العشوائي X (المنفصل أو المتصل) بالصورة التالية:

$$\mu_x(t) = E(e^{tX}) = 1 + \frac{t}{1!}E(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots$$

ويمكن حساب العزوم واستخدام الدوال المولدة للعزوم لكثير من الدوال الاحتمالية الخاصة من خلال دوال الحزمة الإضافية actuar، ولنأخذ الأمثلة التالية:

- **العزوم والدالة المولدة للعزوم للتوزيع المنتظم:**

من المثال الخاص بالتوزيع المنتظم المتصل (البند (1.5.5))، لدينا الدالة:

$$f(x; 0, 30) = \frac{1}{30 - 0} = \frac{1}{30} \quad , \qquad 0 \leq x \leq 30$$

ولحساب العزمين الأولين حول الصفر مثلا لهذا التوزيع نقوم باستخدام الدالة munif بالشكل التالي:

```
> library(actuar)

> munif(order=1:2,min=0,max=30)

[1]  15 300
```

أما الدالة المولدة للعزوم للتوزيع المنتظم، والتي تُعرف في الحزمة بالصيغة mgfunif فيمكن استخدامها لعدة أغراض منها حساب قيمة الدالة عند نقطة معينة t أو التمثيل البياني أو غير ذلك[1]. ونظريا، تُعرف الدالة المولدة للعزوم نظريا للمتغير X بالصورة التالية:

$$\mu_x(t) = E(e^{tx}) = \frac{e^{tb} - e^{ta}}{t(b - a)}$$

ولحساب قيمة الدالة للتوزيع السابق عند قيم معينة للمعلمة t يتم كتابة التالي:

```
> mgfunif(x=c(0.5,1,2,4),min=0,max=1)

[1]  1.297443  1.718282  3.194528 13.399538
```

مع ملاحظة أن x في الدالة mgfunif تُمثل قيم t. ويمكن للقارئ استخدام العمليات الأساسية في R لحساب قيمة $\mu_x(t)$ عن طريق التعويض بقيم t في المثال ومقارنة النواتج كتمرين.

[1] وهذا ينطبق بالطبع على كل دوال التوزيعات الاحتمالية المتوفرة في الحزمة الإضافية actuar.

- **العزوم والدالة المولدة للعزوم للتوزيع الطبيعي:**

للتوزيع الطبيعي المعياري، يمكننا مثلا حساب العزوم الستة الأولى حول الصفر باستخدام الدالة mnorm بالشكل التالي:

```
> mnorm(order=1:6,mean=0,sd=1)
[1]  0  1  0  3  0 15
```

وكذلك يمكن حساب العزوم الخمسة الأولى للتوزيع الطبيعي غير المعياري التالي:

```
> mnorm(order=1:5,mean=5,sd=2.5)

[1]  5.000    31.250   218.750  1679.688 13867.188
```

وتُعرف الدالة المولدة للعزوم للتوزيع الطبيعي تعرف بالصورة:

$$\mu_x(t) = E(e^{tx}) = e^{\left(\mu t + \frac{t^2\sigma^2}{2}\right)}$$

وللتوزيع الطبيعي المعياري تكون:

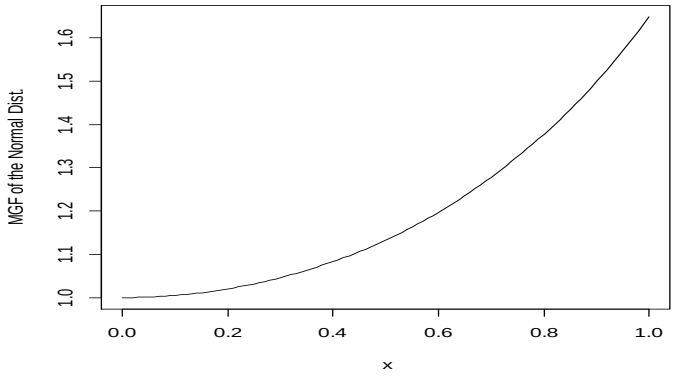
$$\mu_x(t) = E(e^{tx}) = e^{\left(\frac{t^2}{2}\right)}$$

ويمكن حساب قيمة الدالة المولدة للعزوم للتوزيع الطبيعي المعياري مثلا عند بعض نقاط t كالتالي:

```
> mgfnorm(x=c(0,-1,3.5),mean=0,sd=1)

[1]   1.000000   1.648721 457.144713
```

ويمكن أيضا تمثيل الدالة المولدة للعزوم للتوزيع الطبيعي المعياري بيانيا (الشكل (2.5)) عن طريق تعريف الدالة mgfnorm كدالة مُستخدِم خاصة ثم رسمها:

```
> plot(function(x){mgfnorm(x,mean=0,sd=1)},ylab="MGF of
the Normal Dist.")
```



شكل 2.5: التمثيل البياني للدالة المولدة للعزوم للتوزيع الطبيعي المعياري

ويمكن من خلال استخدام الأمر library(help=actuar) التعرف على دوال العزوم والدوال المولدة للعزوم للتوزيعات الاحتمالية المنفصلة والمتصلة الخاصة المُدرجة في الحزمة الإضافية actuar.

الفصل السادس

طرق الاستدلال الإحصائي في R

(Methods of Statistical Inference in R)

1.6 الاستدلالات حول المجتمع الواحد (Inferences about One Population)

1.1.6 الاستدلال حول الوسط الحسابي للمجتمع (Inference about the Population Mean)

2.1.6 الاستدلال حول تباين المجتمع (Inference about the Population Variance)

3.1.6 الاستدلال حول نسبة المجتمع (Inference about the Population Proportion)

4.1.6 اختبارات التوزيع الطبيعي (Tests of Normality)

5.1.6 تقدير معالم التوزيع الاحتمالي (Estimation of Distribution Parameters)

2.6 الاستدلالات حول مجتمعين (Inferences about Two Populations)

1.2.6 الاستدلال حول الفرق بين متوسطين

(Inference about Difference between Two Means)

2.2.6 اختبار تساوي تباينات عدة مجتمعات

(Testing the Equality of Several Populations Variances)

3.2.6 الاستدلال حول نسب مجتمعين (Inference about Two Population Proportion)

4.2.6 اختبار تبعية عينتين لنفس التوزيع

(Testing that Two Samples are Drawn from the Same Distribution)

3.6 اختبارات مربع كاي لبيرسون (Pearson's Chi-square Tests)

1.3.6 اختبار مربع كاي لجودة التوفيق (Chi-square Goodness of Fit Test)

2.3.6 اختبار مربع كاي للاستقلالية (Chi-square Test of Independence)

4.6 تحليل التباين (Analysis of Variance "ANOVA")

5.6 تحليل الارتباط والانحدار الخطي (Linear Correlation and Regression Analysis)

1.5.6 تحليل الارتباط الخطي (Linear Correlation Analysis)

1.1.5.6 معامل الارتباط الجزئي (Partial Correlation Coefficient)

2.5.6 تحليل الانحدار الخطي (Linear Regression Analysis)

1.2.5.6 التمثيل البياني للانحدار الخطي

(Graphical Display for Linear Regression)

6.6 توفيق النماذج الإحصائية بصورة عامة (Fitting Statistical Models in General)

في هذا الفصل، سنقوم بتناول دوال نظام R الخاصة بطرق التقدير واختبارات الفروض، والتي تشمل أيضا تقدير معالم النماذج الإحصائية تحت مفهوم علم الإحصاء الاستدلالي (أو الاستنتاجي)، ذلك العلم الذي يتكون من مجموعة من الطرق التي تستخدم في استنتاج تقديرات والوصول لدلالات واختبار الفرضيات الإحصائية المتعلقة بمجتمع ما أو أكثر.

وحيث أن معظم تلك الدوال في R تقوم بعرض نتائج التقديرات واختبارات الفروض معا، فلن نقوم بعرض مفهوم التقدير أولا ثم اختبار الفروض ثانيا كما هو تسلسل السرد المنطقي لهذين الموضوعين، بل سيتم عرض البنود تباعا بحسب المعلمة أو المعالم المطلوب إجراء الاستدلال حولها. وسنبدأ بالتعامل مع المجتمع الواحد (أي إجراء استدلال حول معلمة واحدة) ثم ننتقل للتعامل مع مجتمعين.

ونذكر القارئ هنا بضرورة إنشاء ملف عمل جديد، والذي سيكون باسم work6 وملف لحفظ سطور الأوامر باسم his6 كما هو متبع في تسلسل فصول الكتاب، بهدف تنظيم متابعة تطبيق الدوال والأوامر.

1.6 الاستدلالات حول المجتمع الواحد (Inferences about One Population)

يُقصد بالاستدلالات حول المجتمع هنا تقدير معلمة المجتمع بنقطة أو بفترة، واختبار مساواتها لقيمة مفترَضة، ويندرج هذا الموضوع في كثير من كتب الاستدلال الاحصائي تحت عنوان تقدير أو اختبار العينة الواحدة. وسنقوم بتقدير واختبار المعالم الأكثر استخداما في علم الإحصاء الاستدلالي وعلى رأسها معالم الوسط الحسابي والتباين والنسبة للمجتمع.

1.1.6 الاستدلال حول الوسط الحسابي للمجتمع (Inference about the Population Mean)

حيث أنه توجد عدة صيغ خاصة بتقدير فترة الثقة للوسط الحسابي للمجتمع μ وإجراء الاختبارات المتعلقة به، وهذه الصيغ تختلف باختلاف حجم العينة ومعلومية تباين المجتمع، لذلك سيتم عرض الصيغة النظرية لفترة الثقة بمعلومية تباين المجتمع σ^2؛

يتم تعريف $100(1-\alpha)\%$ فترة ثقة للوسط μ بالصيغة التالية:

$$\mu: \bar{X} \mp Z\alpha_{/2} \frac{\sigma}{\sqrt{n}}$$

حيث $\bar{X}$ هو الوسط الحسابي لعينة حجمها n مسحوبة من هذا المجتمع. ويتم اختبار الفرضية الصفرية $H_0: \mu = \mu_0$ عند مستوى معنوية α باستخدام الإحصاءة:

$$Z_c = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

ويتم استخدام دالة **اختبار الوسط الحسابي** t.test في R عادة لتقدير الوسط الحسابي للمجتمع، وكذلك لاختبار الفرضية الصفرية عند قيمة معينة. وتضم هذه الدالة الخيارات؛ mu لإدخال قيمة μ_0، وconf.level لإدخال نسبة الثقة، وalternative لتحديد الفرضية البديلة؛ ("two.sided"، "less" أو "greater" والتي تعني "لا يساوي"، "أقل من" أو "أكبر من" على الترتيب).

وتجدر الإشارة هنا إلى أنه في نظام R، (كما هو الحال في معظم البرامج الإحصائية الأخرى)، يتم إطلاق مسمى اختبار t.test على الاختبارات التي تشمل كلا من توزيع Z وتوزيع t حيث أن قيم الإحصاءة في التوزيع الثاني تقترب من التوزيع الطبيعي بزيادة حجم العينة.

ولتطبيق دالة الاختبار سنقوم باستيراد ملف البيانات "studata1.xlsx" الموجود في الحافظة myR بصيغة اكسل، (والذي سبق استيراده في مسار العمل work4 في الفصل الرابع)، إلى مسار العمل الحالي من جديد بنفس الاسم stu.data1؛

```
> library(rJava)
> library(XLConnectJars)
> library(XLConnect)

> stu.data1<-readWorksheetFromFile("studata1.xlsx",
sheet=1,rownames=1)
```

الآن إذا ما أردنا مثلا حساب 95% فترة الثقة للوسط الحسابي لمجتمع أعمار الطلبة المتمثل بالمتغير age، واختبار الفرضية القائلة بأن متوسط أعمار الطلبة هو 22 سنة، (أي اختبار الفرضية $H_0: \mu_{age} = 22$) عند مستوى معنوية 0.05، فإننا بعد تعريف متغير الدراسة s.age نكتب التالي[1]:

```
> s.age<-stu.data1$age

> age.mu.test<-t.test(s.age,mu=22,alternative=
c("two.sided"),conf.level=0.95)
```

[1] تم تعيين الاسم age.mu.test لنتيجة الاختبار حتى يتسنى لنا استدعاؤها متى أردنا، علما بأنه يمكن تنفيذ الاختبار بدون تعيين اسم له.

```
> age.mu.test

        One Sample t-test

data:  s.age
t = 0.513, df = 34, p-value = 0.6113
alternative hypothesis: true mean is not equal to 22
95 percent confidence interval:
 21.57692 22.70880
sample estimates:
mean of x
 22.14286
```

وأهم ما يُستخلص من هذه النتيجة أنه تم قبول الفرضية الصفرية (حيث أن p-value > 0.05) بنسبة ثقة 95%، مما يعني قبول الادعاء القائل بأن متوسط أعمار الطلبة هو 22 سنة، وأن تقدير القيمة الحقيقية (المعلمة) لهذا المتوسط سيتراوح ما بين (22≅21.576) و(23≅22.708) سنة.

ومن هذه النتيجة أيضا نرى أن القيمة الحسابية لإحصائي الاختبار، (والذي سيقترب من التوزيع الطبيعي لأن حجم العينة هو 35)، هي t=0.513 بدرجات الحرية df=34، مع توضيح أن الفرضية البديلة هي عدم التساوي، وكذلك عرض متوسط العينة، (أي التقدير بنقطة لمتوسط المجتمع)، وهو $\bar{X} = 22.142$.

ملاحظة:

عند استخدام دالة الاختبار t.test للاختبار من طرفين (أي اختيار عدم التساوي في الفرضية البديلة) وتحديد نسبة ثقة 95%، فإنه يمكن عندها الاستغناء عن كتابة الخيارين المتعلقين بهما. وبالتالي فإنه في المثال السابق يمكن استخدام سطر الأمر t.test(s.age,mu=22) للحصول على نفس النتيجة السابقة.

وإضافة للدالة t.test، فإنه يمكن استخدام الدالة z.test المتوفرة في الحزمة الإضافية TeachingDemos، إلا أنه يجب إدراج قيمة الانحراف المعياري المستخدم في الصيغة النظرية لقانون الإحصاءة، (إما تباين المجتمع أو العينة) كقيمة للخيار stdev في الدالة z.test، ويمكن استخدام الخيارات الإضافية مثل conf.level وalternative كما هو الحال مع الدالة t.test. وبتنفيذ المثال السابق باستخدام هذه الدالة، (واستخدام تباين العينة)، نحصل على التالي:

```
> library(TeachingDemos)
> z.test(s.age,mu=22,stdev=sd(s.age))

        One Sample z-test

data:  s.age
```

```
z = 0.513, n = 35.000, Std. Dev. = 1.648, Std. Dev. of
the sample mean = 0.278, p-value = 0.608
alternative hypothesis: true mean is not equal to 22
95 percent confidence interval:
 21.59705 22.68867
sample estimates:
mean of s.age
     22.14286
```

ولاحظ وجود اختلاف طفيف في بعض قيم نتائج كلا من الدالتين t.test وz.test نتيجة تغير طريقة الحساب، إلا أن الخلاصة هي نفسها.

2.1.6 الاستدلال حول تباين المجتمع (Inference about the Population Variance)

تُعرّف (1−α)100% فترة ثقة لتباين المجتمع σ^2 بالصيغة:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$$

حيث S^2 هو تباين العينة التي حجمها n، و $\chi^2_{\alpha/2}$ و $\chi^2_{1-\alpha/2}$ هي قيم المتغير العشوائي χ^2 الذي يتبع توزيع مربع كاي عند القيم الاحتمالية $\alpha/2$ و$(1-\alpha/2)$ على الترتيب. ويتم اختبار الفرضية الصفرية $H_0: \sigma^2 = \sigma_0^2$ عند مستوى معنوية α باستخدام الإحصاءة:

$$\chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

ويمكن استخدام دالة **اختبار التباين** sigma.test المتوفرة أيضا ضمن الحزمة الإضافية TeachingDemos لاختبار الفرضية الصفرية السابقة والحصول على فترة الثقة لتباين المجتمع.

وكمثال على تطبيق تلك الدالة، يمكننا اختبار الفرضية $H_0: \sigma^2_{age} = 7$ (باستخدام المتغير s.age) عند مستوى معنوية 0.01 بالصورة التالية:

```
> sigma.test(s.age,sigmasq=7,conf.level=0.99)

        One sample Chi-squared test for variance
data: s.age
X-squared = 13.1837, df = 34, p-value = 0.001004
alternative hypothesis: true variance is not equal to 7
99 percent confidence interval:
 1.565122 5.592642
sample estimates:
var of s.age
    2.714286
```

حيث يتم استخدام الخيار sigmasq لإدخال قيمة التباين المدعاة، ولاحظ أن قيمة إحصاءة الاختبار المحسوبة هي X-squared=13.1837.

3.1.6 الاستدلال حول نسبة المجتمع (Inference about the Population Proportion)

يمكن استخدام النظرية التالية لتحويل توزيع معلمة النسبة من توزيع ذي الحدين إلى التوزيع الطبيعي:

إذا كان X هو متغير عشوائي يتبع توزيع ذي الحدين بوسط $\mu = np$ وتباين $\sigma^2 = npq$ فإن المتغير العشوائي $z = \frac{X-np}{\sqrt{npq}}$ سيتبع التوزيع الطبيعي المعياري عندما $n \to \infty$. وبالتالي فإن توزيع النسبة $P = \frac{X}{n}$ سيؤول للتوزيع الطبيعي بوسط p وتباين $\frac{pq}{n}$.

وتُستخدم دالة **اختبار النسبة** prob.test في R لتقدير نسبة المجتمع، (بنقطة وفترة ثقة)، واختبار مساواتها لقيمة معينة. ولنأخذ المثال التالي:

في عينة عشوائية مكونة من 500 شخص من الذين يتناولون طعام الغذاء في أحد المطاعم خلال يوم الجمعة، وُجد أن 160 شخص منهم يفضلون تناول المأكولات البحرية. أوجد:

1. 95% فترة ثقة لنسبة مجتمع الأشخاص (P) الذين يفضلون تناول نفس النوع من المأكولات البحرية في ذلك المطعم يوم الجمعة.

2. اختبر الفرضية $H_1: P = 0.35$ عند مستوى معنوية 0.05.

يمكن تنفيذ المطلوب في هذا المثال عن طريق كتابة:

```
> prop.test(x=160,n=500,p=0.35,correct=F)

  1-sample proportions test without continuity correction

data:  160 out of 500, null probability 0.35
X-squared = 1.978, df = 1, p-value = 0.1596
alternative hypothesis: true p is not equal to 0.35
95 percent confidence interval:
 0.2806178 0.3621270
sample estimates:
   p
0.32
```

حيث أن x هو عدد النجاحات، n هو حجم العينة، p هي قيمة المعلمة الافتراضية، وcorrect=F هو خيار عدم استخدام **تصحيح "يتس"** (Yates' Continuity Correction).

ولاحظ أن $Z_c = \frac{160-(500\times0.35)}{\sqrt{500\times0.35\times0.65}} = -1.406$، وبإهمال الإشارة السالبة، (الناتجة عن كون عدد النجاحات في العينة أقل من عدد النجاحات المتوقع)، نرى بأن $\sqrt{1.978} = 1.406$، حيث أن القيمة المحسوبة X-squared=1.978 هي لإحصاءة مربع كاي والتي تساوي قيمة مربع التوزيع الطبيعي المعياري كما هو معلوم. ويمكن استخدام الخيارات الإضافية مثل conf.level وalternative مع دالة prop.test عند الضرورة.

4.1.6 اختبارات التوزيع الطبيعي (Tests of Normality)

في الفصل الرابع، تم التعرض لبعض الرسومات التي تعكس توزع المتغير ما إذا كان طبيعيا أم لا، وفي هذا البند سنعرض اختبارين هامين للاستدلال على توزع المتغير العشوائي بتوزيع طبيعي وهما اختبار شابيرو–ويلك واختبار كولموجوروف–سميرنوف؛

- **اختبار شابيرو- ويلك** (Shapiro-Wilk Normality Test)

يُعرّف اختبار شابيرو–ويلك في لغة R بالصيغة Shapiro.test ويستخدم لاختبار الفرضية الصفرية (العينة مسحوبة من مجتمع طبيعي :H_0).

وكمثال، يمكننا اختبار توزع المتغيرات grd1، grd2، وgrd3، (والتي تمثل درجات الطلبة في ثلاثة مقررات في البيانات stu.data1)، بتوزيع طبيعي بصورة منفردة، وسنقوم أولا بتعريف هذه المتغيرات في مسار العمل work6، ثم تنفيذ الاختبارات؛

```
> s.grd1<-stu.data1$grd1;s.grd2<-stu.data1$grd2;
s.grd3<-stu.data1$grd3

> shapiro.test(s.grd1)

        Shapiro-Wilk normality test

data:  s.grd1
W = 0.95, p-value = 0.1133


> shapiro.test(s.grd2)

        Shapiro-Wilk normality test

data:  s.grd2
W = 0.939, p-value = 0.05202
```

```
> shapiro.test(s.grd3)

        Shapiro-Wilk normality test

data:  s.grd3
W = 0.834, p-value = 0.000102
```

ومن نتائج القيم الاحتمالية (p-value) للاختبارات يمكن القول بأن المتغير s.grd1 يتوزع بتوزيع طبيعي وكذلك المتغير s.grd2 (بدرجة أقل)، إلا أن المتغير s.grd3 يبتعد عن التوزيع الطبيعي. ولاحظ أن هذه النتائج تتوافق مع ما تم استنتاجه من التحليل الاستكشافي لهذه المتغيرات في الفصل الرابع.

- **اختبار كولموجوروف - سميرنوف** (Kolmogorov-Smirnov Test)

يُعرّف هذا الاختبار في R بالصيغة ks.test، ويقوم باختبار الفرضية الصفرية التي تقضي بتبعية توزيع المتغير أو العينة بتوزيع طبيعي؛ (العينة مسحوبة من مجتمع طبيعي :H_0) ويجب في هذا الاختبار تحديد التوزيع المطلوب اختبار تبعية العينة له، (أي إجراء اختبار **جودة التوفيق** (Goodness of Fit))، بمعنى أنه إذا أردنا اختبار توزع المتغير بتوزيع طبيعي (كما هو الحال في مثالنا) فيجب إدراج الخيار[1] "pnorm" ضمن الدالة، وبالتالي يمكن اختبار توزع المتغيرات الثلاثة السابقة بالصورة؛

```
> ks.test(s.grd1,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  s.grd1
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(s.grd1, "pnorm") :
  ties should not be present for the Kolmogorov-Smirnov
test
> ks.test(s.grd2,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  s.grd2
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(s.grd2, "pnorm") :
  ties should not be present for the Kolmogorov-Smirnov test
```

[1] إذا ما كان المطلوب اختبار توزع المتغير العشوائي بتوزيع آخر، (المنتظم مثلا)، فيتم إدراج الخيار "punif" ضمن الدالة، وهكذا بالنسبة لباقي التوزيعات الاحتمالية.

```
> ks.test(s.grd3,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  s.grd3
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(s.grd3, "pnorm") :
  ties should not be present for the Kolmogorov-Smirnov test
```

ويُلاحظ أن النتائج متطابقة تماما للمتغيرات الثلاثة، وهي أنها لا تتبع التوزيع الطبيعي. وهذه النتيجة المعاكسة (بالنسبة للمتغيرين الأولين) لاختبار شابيرو- ويلك السابق قد تكون بسبب صغر حجم العينة وتكرر الكثير من قيم المشاهدات ضمن هذه المتغيرات مما تسبب في ظهور الرسالة التحذيرية.

ولتوضيح هذه النقطة، يمكننا إجراء هذا الاختبار من جديد باستخدام عينة كبيرة الحجم تم توليدها باستخدام دالة التوزيع الطبيعي rnorm وسنحصل في كل مرة[1] على عينة تجتاز الاختبار (أي تتبع التوزيع الطبيعي)، بنجاح؛

```
> ks.test(rnorm(1000),"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  rnorm(1000)
D = 0.0212, p-value = 0.7608
alternative hypothesis: two-sided
```

وعموما، فإن الكثير من الإحصائيين يعتبرون اختبار شابيرو- ويلك أكثر ثباتا من اختبار كولموجوروف - سميرنوف مع العينات صغيرة الحجم من خلال الممارسة والتطبيق.

5.1.6 تقدير معالم التوزيع الاحتمالي (Parameters Estimation of Distributions)

يمكن في لغة R تقدير معالم معظم التوزيعات الاحتمالية باستخدام طريقة التقدير المعروفة باسم طريقة **الامكان الأعظم** (Maximum Likelihood) من خلال العينة المتوفرة، وذلك عن طريق استخدام دالة **توفيق التوزيعات** fitdistr المتوفرة في الحزمة الإضافية MASS. فعلى سبيل المثال، يمكن باستخدام المتغير s.grd1 تقدير معالم التوزيع الطبيعي μ و σ^2 بالصورة التالية:

```
> library(MASS)
```

[1] حيث أنه سيتم توليد عينة عشوائية مختلفة في كل مرة يُجرى فيها الاختبار.

```
> fitdistr(s.grd1,"normal")

     mean          sd
  71.314286   15.380454
 ( 2.599771) ( 1.838316)
```

وهذا يعني أن $\hat{\mu} = 71.314$ و$\hat{\sigma} = 15.380$ بخطأ معياري يساوي 2.599 للتقدير الأول ويساوي 1.838 للتقدير الثاني على الترتيب.

وكمثال آخر، يمكننا تقدير معلمة التوزيع الأسي β لنفس المتغير بالصورة:

```
> fitdistr(s.grd1,"exponential")

      rate
  0.014022436
 (0.002370224)
```

مع ملاحظة أن $1/\hat{\beta} = 0.014$.

2.6 الاستدلالات حول مجتمعين (Inferences about Two Populations)

1.2.6 الاستدلال حول الفرق بين متوسطين (Inference about Difference between Two Means)

تُعرّف الصيغة العامة لإحصاءة اختبار الفرق بين متوسطين $(H_0: \mu_1 - \mu_2 = 0)$ بالصورة التالية:

$$Z_c = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

وذلك في الحالات التي تتضمن العينات الكبيرة أو عند توفر قيم التباين للمجتمعات، وتُستخدم إحصاءة t في الحالات الأخرى والتي يجب أن يكون من المعلوم فيها تساوي أو عدم تساوي تباينات المجتمعين.

وتُستخدم نفس الدالة `t.test` لتقدير فترة الثقة للفرق بين وسطي المجتمع وكذلك اختبار الفرضية الصفرية $H_0: \mu_1 - \mu_2 = 0$. إلا أنه للإحصاءة t، يجب التأكد أولا من تساوي أو عدم تساوي تباينات مجتمعي الدراسة، أي أنه يجب اختبار الفرضية الصفرية $H_0: \sigma_1^2 = \sigma_2^2$.

والفرضية الأخيرة يمكن اختبارها باستخدام دالة **اختبار التباين** `var.test` التي تُستخدم للاستدلال حول تبايني مجتمعين باستخدام الإحصاءة F، حيث $F_c = \frac{S_1^2}{S_2^2}$.

وكمثال على تطبيق كل من الدالة `var.test` والدالة `t.test`، للاستدلال حول معالم مجتمعين، لنتناول المشاهدات الـ 25 الأولى للمتغيرين `s.grd1` و`s.grd2`:

```
> var.test(s.grd1[1:25],s.grd2[1:25])

        F test to compare two variances

data:  s.grd1[1:25] and s.grd2[1:25]
F = 1.018, num df = 24, denom df = 24, p-value = 0.9655
alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval:
 0.4486027 2.3101330
sample estimates:
ratio of variances
          1.018004
```

ونلاحظ من نتيجة الاختبار قبول الفرضية الصفرية، أي قبول تساوي تباييني مجتمعي درجات الطلبة في المقررين s.grd1 و s.grd2.

كما يُلاحظ أن التقدير بنقطة لمعلمة النسبة بين تباييني المجتمعين $\frac{\sigma_1^2}{\sigma_2^2}$ هو $\frac{S_1^2}{S_2^2} = 1.018$ ، والتقدير بفترة لتلك المعلمة هو $0.448 < \frac{\sigma_1^2}{\sigma_2^2} < 2.310$ ، بنسبة ثقة 95%.

الآن يمكن تنفيذ الدالة t.test باستخدام الخيار var.equal=T لتحديد أن تباينات المجتمعين متساوية، والخيار paired=F لتحديد أن العينتين مستقلتين، (مع إمكانية الاستغناء عن الخيار الأخير لأنه الافتراضي ضمن الدالة):

```
> t.test(s.grd1[1:25],s.grd2[1:25],var.equal=T,paired=F)

        Two Sample t-test

data:  s.grd1[1:25] and s.grd2[1:25]
t = -0.4311, df = 48, p-value = 0.6683
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -11.100969   7.180969
sample estimates:
mean of x mean of y
    68.20     70.16
```

ونستخلص من النتيجة السابقة قبول تساوي متوسطي المجتمعين بمستوى معنوية 0.05، والذي يعني تقارب أداء الطلبة في المقررين s.grd1 وs.grd2 بشكل كبير، مع ملاحظة[1] أن أداء الطلبة في المقرر s.grd2 كان أفضل بقليل.

ولاختبار العينات غير المستقلة، نستخدم الخيار paired=T في الدالة t.test كما يوضح المثال التالي؛

البيانات التالية (المتغيرين A1 وA2) تمثل عينتين مرتبطتين (غير مستقلتين) تحتوي كل منهما على 10 مشاهدات:

```
> A1<-c(76,60,85,58,91,75,82,64,79,88)
> A2<-c(81,52,87,70,86,77,90,63,85,83)
```

ولتقدير 99% فترة ثقة للفرق بين وسطي مجتمعي الدراسة واختبار تساوي هذين الوسطين نقوم بتنفيذ الأمر التالي:

```
> t.test(A1,A2,conf.level=0.99,paired=T)

        Paired t-test

data:  A1 and A2
t = -0.793, df = 9, p-value = 0.4482
alternative hypothesis: true difference in means is not
equal to 0
99 percent confidence interval:
 -8.157191  4.957191
sample estimates:
mean of the differences
                  -1.6
```

فنستدل على تساوي متوسطات المجتمعين، ويكون تقدير الفرق بينهما بنقطة وبفترة كما هو واضح في النتيجة.

2.2.6 اختبار تساوي تباينات عدة مجتمعات

(Testing the Equality of Several Populations Variances)

اختبار تساوي التباين لعدة مجتمعات، والذي يُعرف باختبار **بارتليت** (Bartlett Test)، يأخذ الصيغة bartlett.test، ولتنفيذه يتم تنظيم بيانات أو عينات الاختبار ضمن قائمة أولا كما يوضح المثال التالي، والذي يختبر تساوي تباينات مجتمعات درجات الطلبة في المقررات الدراسية s.grd1، s.grd2، وs.grd3:

[1] من فترة الثقة يُلاحظ اقتراب تقدير قيمة الفرق بين متوسطي الدرجات في المجتمعين من الإشارة السالبة أكثر، مما يعني أفضلية الأداء في المقرر الثاني s.grd2 إلى حد ما.

```
> s.grd123<-list(s.grd1,s.grd2,s.grd3)

> bartlett.test(s.grd123)

        Bartlett test of homogeneity of variances

data:  s.grd123
Bartlett's K-squared = 0.2733, df = 2, p-value = 0.8723
```

ونلاحظ من نتيجة الاختبار قبول الفرضية $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ مما يدل على تجانس التباين بين المجتمعات الثلاثة.

3.2.6 الاستدلال حول نسب مجتمعين (Inference about Two Populations Proportions)

كما هو الحال مع الاستدلال حول نسبة واحدة، يتم استخدام الدالة `prop.test` للتعامل مع نسبتي مجتمعين، والمثال التالي يوضح طريقة تطبيق هذه الدالة؛

تم إجراء استفتاء في مدينتين متجاورتين لتحديد رغبة السكان في تغيير مسار الطريق السريع المار بالمدينتين. لتحديد ما إذا كان هنالك فرق معنوي في نسب المصوتين في كلا المدينتين تم اختيار عينة عشوائية من كل مدينة، فوجد أن 120 شخص من 200 شخص في المدينة الأولى موافقون على تغيير مسار الطريق، و240 من 500 شخص في المدينة الثانية أيضاً موافقون. هل توافق (عند مستوى معنوية 0.025) على أن نسبة الموافقين في المدينة الأولى هي أكبر منها في المدينة الثانية؟

نقوم بإدخال عدد النجاحات في العينة الأولى والثانية كقيم `x` وحجم العينتين كقيم `n` مع تحديد الخيارات الأخرى كالتالي:

```
> prop.test(x=c(120,240),n=c(200,500),conf.level=0.975,
alternative = c("greater"),correct=F)

    2-sample test for equality of proportions without
                continuity correction

data:  c(120, 240) out of c(200, 500)
X-squared = 8.2353, df = 1, p-value = 0.002054
alternative hypothesis: greater
97.5 percent confidence interval:
 0.03920763 1.00000000
sample estimates:
prop 1 prop 2
  0.60   0.48
```

وهذا يعني رفض الفرضية الصفرية $H_0: P_1 = P_2$ ، بمعنى أننا نعتقد أن نسب الموافقين في المدينة الأولى هي أعلى من نسب الموافقين في المدينة الثانية، ويُلاحظ أيضا أن قيمة إحصاءة الاختبار هي $Z_c = \sqrt{8.2353} = 2.869$.

4.2.6 اختبار تبعية عينتين لنفس التوزيع

(Testing that Two Samples follow the Same Distribution)

يمكن استخدام اختبار كولموجوروف–سميرنوف بنفس الصيغة ks.test لاختبار ما إذا كانت العينتين العشوائيتين المسحوبتين هما من نفس التوزيع الاحتمالي المتصل أم لا. ولنختبر ذلك بالنسبة للمتغيرين s.grd1 وs.grd2 أولا ثم للمتغيرين s.grd1 وs.grd3 ثانيا على سبيل المثال:

```
> ks.test(s.grd1,s.grd2)

        Two-sample Kolmogorov-Smirnov test

data:  s.grd1 and s.grd2
D = 0.1143, p-value = 0.9763
alternative hypothesis: two-sided

Warning message:
In ks.test(s.grd1, s.grd2) : cannot compute exact p-value
with ties


> ks.test(s.grd1,s.grd3)

        Two-sample Kolmogorov-Smirnov test

data:  s.grd1 and s.grd3
D = 0.6857, p-value = 1.425e-07
alternative hypothesis: two-sided

Warning message:
In ks.test(s.grd1, s.grd3) : cannot compute exact p-value
with ties
```

ونستنتج أن المتغيرين s.grd1 وs.grd2 مسحوبان من نفس التوزيع الاحتمالي، وهو التوزيع الطبيعي كما لاحظنا سابقا، وأن المتغيرين s.grd1 وs.grd3 لهما توزيعين مختلفين، وهذه أيضا نتيجة منطقية لأن توزيع المتغير s.grd3 يبتعد عن التوزيع الطبيعي كما شاهدنا سابقا.

3.6 اختبارات مربع كاي لبيرسون (Pearson's Chi-square Tests)

يُعرف اختبار مربع كاي لبيرسون في R بالصيغة chisq.test ويُستخدم عادة لإجراء اختبار جودة التوفيق للتوزيعات المنفصلة، واختبار الاستقلالية. ونبدأ بالاختبار الأول:

1.3.6 اختبار مربع كاي لجودة التوفيق (Chi-square Goodness of Fit Test)

تُعرف إحصاءة اختبار جودة التوفيق بالصيغة:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

حيث e_i و o_i هي قيم التكرارات المشاهدة والمتوقعة على الترتيب للحالة i، وk هو عدد الحالات الممكنة في التجربة. ولنأخذ المثال التالي؛

إذا علمت أنه حضر إلى أحد مكاتب البريد في أحد الأيام 90 شخصاً لإجراء مكالمات هاتفية وكان توزيعهم حسب المدة الزمنية للمكالمة هو:

الزمن بالدقائق (x)	1	2	3	4	5
التكرار (f)	16	30	25	11	8

هل ترى أن توزيع المدة الزمنية للمكالمة يتبع توزيع ذي الحدين؟

لاستخدام الدالة chisq.test يتم إدخال قيم المتغير X كمتجه أول وقيم الاحتمالات المناظرة له كمتجه ثاني في الدالة، ولاختبار توفيق هذه التجربة العشوائية لتوزيع ذي الحدين يجب حساب الاحتمالات المناظرة باستخدام الدالة الاحتمالية للتوزيع المطلوب اختباره.

لدينا $E(X) = \frac{\sum x.f}{\sum f} = n.p$، حيث n=5 هي عدد الحالات هنا، وبالتالي يمكن حساب احتمال النجاح p بالصورة؛ $\frac{235}{90} = 5 \times p$ ، وهكذا تكون p=0.522. عندئذ يمكننا كتابة:

```
> chisq.test(x=c(0,16,30,25,11,8),p=dbinom(0:5,size=5,
prob=0.522))

        Chi-squared test for given probabilities

data:  c(0, 16, 30, 25, 11, 8)
X-squared = 11.7712, df = 5, p-value = 0.03806

Warning message:
In chisq.test(x = c(0, 16, 30, 25, 11, 8), p=dbinom(0:5, size=
5,:  Chi-squared approximation may be incorrect
```

مما يعني توجهنا لرفض الفرضية الصفرية (المتغير يتبع توزيع ذي الحدين :H_0) عند مستوى معنوية 0.05، ولاحظ أننا قمنا بإضافة القيمة 0 للمتجه الأول x نظرا لأن توزيع ذي الحدين تبدأ قيم المتغير العشوائي فيه من تلك القيمة.

وكمثال آخر، يمكن اختبار ما إذا كان المتغير العشوائي الذي يمثل الرقم الظاهر على زهر النرد عند رميه 120 مرة، (كما هو موضح أدناه)، يتوزع بالتوزيع المنتظم كما يلي:

الرقم الظاهر (x)	1	2	3	4	5	6
التكرار (f)	20	22	17	18	19	24

```
> chisq.test(x=c(20,22,17,18,19,24),p=dunif(1:6,min=0,
max=6))

        Chi-squared test for given probabilities

data:  c(20, 22, 17, 18, 19, 24)
X-squared = 1.7, df = 5, p-value = 0.8889
```

وهذا يعني قبول الفرضية (المتغير يتبع التوزيع المنتظم :H_0) عند مستوى معنوية 0.05، وقد تم استخدام الخيارات min=0 وmax=6 لغرض تعريف الاحتمال المنتظم 1/6 لكل قيم المتغير العشوائي.

2.3.6 اختبار مربع كاي للاستقلالية (Chi-square Test of Independence)

يمكن استخدام نفس الدالة chisq.test لإجراء اختبار الاستقلالية لجداول الاقتران، والتي تناولنا طرق ادخالها وتعريفها في الفصل الرابع، وذلك بصورة مباشرة كما نرى في المثال التالي:

تم إجراء دراسة ميدانية لمعرفة ما إذا كان هنالك ارتباط بين فئة العمر وعدد ساعات مشاهدة التلفاز اليومية، فتم سحب عينة عشوائية مكونة من 1000 شخص من فئات عمرية مختلفة، وتم سؤالهم عن الأوقات التي يقضونها في مشاهدة التلفاز فكانت النتائج كما يلي:

		فئة العمر		
		أطفال	بالغون	مسنون
عدد ساعات المشاهدة اليومية	أقل من ساعتين	182	213	203
	ساعتين فأكثر	154	138	110

عند مستوى معنوية 0.05 هل ترى وجود علاقة بين عمر الشخص وعدد ساعات مشاهدته للتلفاز؟.

يمكننا إدخال جدول الاقتران السابق في R كمصفوفة بالصورة؛

```
> TV<-matrix(data=c(182,213,203,154,138,110),nrow=2,
ncol=3,byrow=T)

> TV

     [,1] [,2] [,3]
[1,]  182  213  203
[2,]  154  138  110
```

وبتنفيذ اختبار الاستقلالية نحصل على النتيجة التالية:

```
> chisq.test(TV)

        Pearson's Chi-squared test

data:  TV
X-squared = 7.8782, df = 2, p-value = 0.01947
```

وهذا يعني رفض الفرضية الصفرية (يوجد ارتباط بين فئة العمر ومشاهدة التلفاز :H_0)، أي يمكننا القول بأنه توجد استقلالية بين فئة العمر والرغبة في مشاهدة التلفاز من خلال تلك العينة.

4.6 تحليل التباين (Analysis of Variance "ANOVA")

إن تحليل التباين هو أحد المفاهيم الهامة والأساسية في علم الإحصاء الاستدلالي نظرا لكونه جزءا أساسيا من منظومة التحليل الإحصائي في معظم النماذج الإحصائية والأساليب المتعلقة بها. وسنقدم شرحا مختصرا **لتحليل التباين في اتجاه واحد** (One-Way ANOVA) فيما يلي؛ لنفرض أن x_{ij} هي القيمة في الصف رقم j والعامود رقم i وذلك في مصفوفة بيانات تضم k عامود (مجموعة أو متغير) وn صف (مشاهدة) بحيث يكون $N=n\times k$، كما يوضح جدول المشاهدات الافتراضية (جدول (1.6)) التالي؛

جدول 1.6: الشكل العام لجدول البيانات الافتراضية **X** التي تحوي $n\times k$ مشاهدة

x_{11}	x_{12}	⋯	x_{1k}
x_{21}	x_{22}	⋯	x_{2k}
⋯	⋯	x_{ij}	⋯
x_{n1}	x_{n2}	⋯	x_{nk}

المشاهدة x_{ij} يمكن "تفكيكها" أو تحليلها إحصائيا إلى المكونات التالية:

$$x_{ij} = \bar{x}_{.} + (\bar{x}_i - \bar{x}_{.}) + (x_{ij} - \bar{x}_i)$$

حيث $\bar{x}_.$ هو المتوسط العام لكل المشاهدات في المصفوفة و $\bar{x}_i$ هو متوسط العامود أو المجموعة، وبالتالي فإن المقدار $(\bar{x}_i - \bar{x}_.)$ يمثل انحراف متوسط المجموعة عن المتوسط العام، والمقدار $\left(x_{ij} - \bar{x}_i\right)$ يمثل انحراف المشاهدات عن متوسط المجموعة.

والمعادلة السابقة يمكن إعادة كتابتها بشكل نموذج إحصائي، (والذي يسمى عادة بنموذج تصميم التجارب)، بالصورة التالية:

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

حيث $\epsilon_{ij} \sim N(0, \sigma^2)$، ويتم من خلال هذا النموذج اختبار الفرضية الصفرية القائلة بأنه لا توجد انحرافات لمتوسطات المجموعات عن المتوسط العام، (أي $\alpha_i = 0$ لكل المجموعات أو الأعمدة).

وإذا ما تم كتابة المعادلة الأولى بالصورة:

$$(x_{ij} - \bar{x}_.) = (\bar{x}_i - \bar{x}_.) + \left(x_{ij} - \bar{x}_i\right)$$

فيمكن اعتبار أن المقدار $(x_{ij} - \bar{x}_.)$، الذي يمثل انحراف المشاهدات عن المتوسط العام، قد تم تحليله إلى مركبتين، وإذا ما تم التعامل مع مجاميع المربعات للمقادير الثلاثة فيمكن عندها القول بأن مجموع المربعات الكلي والذي يمثل التباين الكلي يمكن تحليله إلى مركبتين أو نوعين من الاختلاف أو التباين؛ الأول هو مجموع المربعات للمقدار $(\bar{x}_i - \bar{x}_.)$، والذي يُعرف **بالاختلاف بين المجموعات** (ويسمى **التباين المفسر** (Explained Variance))، والثاني هو مجموع المربعات للمقدار $\left(x_{ij} - \bar{x}_i\right)$، الذي يُعرف **بالاختلاف ضمن المجموعات** (ويسمى **التباين غير المفسر** (Unexplained Variance))، ويمكن كتابة الصيغ الخاصة بالمكونين بالصورة التالية:

- **الاختلاف بين المجموعات** (Variation Between Groups):

$$SSB = \sum_{i}^{k} n_i \, (\bar{x}_i - \bar{x}_.)^2$$

حيث n_i هو عدد المشاهدات في المجموعة i، ومنه يمكن حساب متوسط مجموع المربعات للاختلاف بين المجموعات بالصيغة:

$$MSB = SSB/(k-1)$$

- **الاختلاف ضمن المجموعات** (Variation Within Groups):

$$SSW = \sum_{i}^{k} \sum_{j}^{n} (\bar{x}_{ij} - \bar{x}_i)^2$$

ومنه يمكن حساب متوسط مجموع المربعات للاختلاف ضمن المجموعات بالصيغة:

$$MSW = SSW/(N-k)$$

وبالتالي يمكن كتابة التباين الكلي TSS بالصورة:

$$TSS = SSB + SSW$$

أو

$$\frac{1}{(N-1)}\sum_{i}^{k}\sum_{j}^{n}(x_{ij}-\bar{x}_{.})^2 = \frac{1}{(k-1)}\sum_{i}^{k} n_i\,(\bar{x}_i-\bar{x}_{.})^2 + \frac{1}{(N-k)}\sum_{i}^{k}\sum_{j}^{n}(\bar{x}_{ij}-\bar{x}_i)^2$$

ويتم اختبار الفرضية الصفرية (متوسطات مجتمعات المجموعات متساوية :H_0) باستخدام الإحصاءة F بالصيغة:

$$F_c = \frac{MSB}{MSW}$$

وهذا يعني أننا نختبر وجود فروق معنوية بين متوسطات المجموعات عن طريق مقارنة تباينين، (وهذا في الواقع السبب في تسمية الاختبار باختبار تحليل التباين رغم أن صيغة الفرضية تشمل اختبار متوسطات المجموعات).

وينطبق نفس المفهوم على **تحليل التباين في اتجاهين** (Two-Way ANOVA)، حيث تؤخذ الاختلافات بين متوسطات الصفوف في الاعتبار، فيُعتبر الاختلاف المفسر عندئذ ناتجا عن الاختلاف بين متوسطات مجموعات الأعمدة (الاتجاه الأول)، أو ناتجا عن الاختلاف بين متوسطات مجموعات الصفوف (الاتجاه الثاني)، أو ناتجا عن الاتجاهين معا.

ويتم كتابة التباين الكلي عندئذ بالصورة:

$$TSS = SSC + SSR + SSW$$

حيث SSC يمثل التباين الناتج عن الأعمدة، وSSR يمثل التباين الناتج عن الصفوف. ويتم اختبار الفرضيتين الصفريتين؛

(متوسطات الأعمدة متساوية :$H_0^{'}$) باستخدام الإحصاءة $F_c' = \frac{MSC}{MSW}$،

و(متوسطات الصفوف متساوية :$H_0^{''}$) باستخدام الإحصاءة $F_c'' = \frac{MSR}{MSW}$.

وعند مستوى معنوية معين يمكن إجراء اختبار تحليل التباين باستخدام دالتي **تحليل التباين** `aov` و`anova` كما يوضح المثال التالي:

البيانات في الجدول (2.6) تمثل أطوال سنابل قمح (بالسنتيمتر) تم زراعتها ضمن تجربة لاختبار وجود اختلاف بين ثلاثة أنواع من بذور القمح هي A، B، وC، وكذلك اختبار وجود اختلاف بين أربعة أنواع من السماد العضوي هي L1، L2، L3، وL4.

جدول 2.6: أطوال سنابل القمح تبعا لأنواع البذور وأنواع الأسمدة العضوية

	A	**B**	**C**
L1	74	72	64
L2	47	57	55
L3	58	66	59
L4	53	57	58

ويمكن تطبيق اختبار تحليل التباين للكشف عن وجود فروق معنوية بين أنواع القمح أو بين أنواع السماد، (حيث يُعتبر كلا من الاختبارين في اتجاه واحد)، أو اختبار وجود فروق بين أنواع القمح والسماد معا، (ويُعتبر ذلك اختبار في اتجاهين).

لنبدأ باختبار وجود فروق معنوية بين أنواع القمح، حيث سيتم أولا إدخال البيانات في الجدول بالصورة التالية:

```
> X.obs<-c(74,47,58,53,72,57,66,57,64,55,59,58)
```

```
> X.col<-c("a","a","a","a","b","b","b","b","c","c","c","c")
```

بمعنى أنه تم إدخال كل قيم المشاهدات في الجدول (2.6) كمتجه عددي، ثم إدخال الرمز المناظر لكل قيمة بحسب العامود كمتجه غير عددي يمثل أسماء أنواع بذور القمح.

بعد ذلك نقوم بتعيين اسم للدالة aov التي سيُدخل فيها المتجهان السابقان بالصورة:

```
> Xc.aov<-aov(X.obs~X.col)
```

ثم نقوم بعدها باستخدام الدالة anova بالصورة التالية فنحصل على جدول تحليل التباين في اتجاه واحد لاختبار أنواع القمح:

```
> anova(Xc.aov)
Analysis of Variance Table
Response: X.obs
          Df Sum Sq Mean Sq F value Pr(>F)
X.col      2     56  28.000  0.4158 0.6718
Residuals  9    606  67.333
```

ومن نتيجة الاختبار يتضح قبول الفرضية الصفرية، بمعنى أننا نعتقد أنه لا توجد فروق معنوية بين أنواع بذور القمح المستخدمة في الزراعة.

ملاحظة:

الاختبار السابق يفترض أن تباينات الأعمدة متساوية، أما إذا رغبنا في إجراء نفس الاختبار بافتراض أن تباينات المجموعات أو الأعمدة هي غير متساوية فيمكننا استخدام الدالة oneway.test، وباستخدام مثالنا الحالي لتطبيق هذه الدالة نحصل على:

```
> oneway.test(X.obs~X.col,var.equal=F)

 One-way analysis of means (not assuming equal variances)

data:  X.obs and X.col
F= 0.4563, num df= 2.000, denom df= 5.064, p-value=0.6573
```

ونلاحظ حدوث تغير طفيف في قيم هذه النتيجة، مع عدم تغير الاستنتاج الكلي، مقارنة بالنتيجة السابقة التي تفترض تساوي التباينات. ويمكن بالطبع الحصول على نفس النتيجة الأصلية عند استخدام الخيار var.equal=T ضمن الدالة.

يمكننا الآن استخدام تحليل التباين في اتجاه واحد أيضا، كمثال إضافي، لاختبار متوسطات الصفوف كما هو الحال مع الأعمدة، ولتنفيذ ذلك نقوم بإدخال الرمز المناظر لكل قيمة بحسب الصف كمتجه غير عددي؛

```
> X.row<-c("L1","L2","L3","L4","L1","L2","L3","L4","L1",
"L2","L3","L4")
```

الآن نقوم بنفس الخطوات السابقة مع الدالتين aov و anova ولكن باستخدام المتجه X.row كما يلي:

```
> Xr.aov<-aov(X.obs~X.row)

> anova(Xr.aov)

Analysis of Variance Table

Response: X.obs

          Df Sum Sq Mean Sq F value   Pr(>F)
X.row      3    498   166.0  8.0976 0.008297 **
Residuals  8    164    20.5
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ونتيجة الاختبار السابقة توضح رفض الفرضية الصفرية، بمعنى أننا نعتقد أنه توجد فروق معنوية بين نوعين على الأقل من أنواع الأسمدة العضوية المستخدمة. والسطر الأخير في النتيجة السابقة هو بمثابة مفتاح للرموز

المستخدمة لمستوى المعنوية الذي يتم من خلاله رفض الفرضية الصفرية في الاختبار، ففي هذا الاختبار نلاحظ رفض الفرضية الصفرية بمستوى معنوية عالي (ما بين القيمتين 0.001 و0.01).

نأتي الآن لتطبيق اختبار تحليل التباين في اتجاهين للمثال الحالي، حيث سيتم استخدام نفس الطريقة مع تعديل بسيط يتمثل في إضافة المتجه الغير عددي الممثل للصفوف إلى المتجه الغير عددي الممثل للأعمدة ضمن الدالة aov؛

```
> Xcr.aov<-aov(X.obs~X.col+X.row)

> anova(Xcr.aov)

Analysis of Variance Table

Response: X.obs

          Df Sum Sq Mean Sq F value  Pr(>F)
X.col      2     56      28  1.5556 0.28559
X.row      3    498     166  9.2222 0.01152 *
Residuals  6    108      18
---
Signif.codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

وهذه النتيجة تنص على قبول الفرضية الصفرية الأولى (متوسطات الأعمدة متساوية :H_0')، بمعنى الحكم بعدم وجود فروق معنوية بين أنواع بذور القمح، ورفض الفرضية الصفرية الثانية القائلة (متوسطات الصفوف متساوية :H_0'') عند مستوى معنوية 0.05، بمعنى اعتقادنا بوجود نوعين على الأقل من الأسمدة العضوية غير متساويي التأثير على أطوال القمح.

5.6 تحليل الارتباط والانحدار الخطي (Linear Correlation and Regression Analysis)

إن مفهوم الارتباط والانحدار يتعلق بصورة عامة بدراسة العلاقات الخطية أو غير الخطية بين المتغيرات، فالارتباط يسعى لقياس درجة العلاقة بين متغيرين أو أكثر، بينما تقوم نماذج الانحدار بتوفيق علاقة سببية بين هذه المتغيرات تعكس تأثيراتها المتبادلة. وسنقوم بتقديم شرح موجز فيما يلي لكل من المفهومين فيما يتعلق بالعلاقات الخطية فقط.

1.5.6 تحليل الارتباط الخطي (Linear Correlation Analysis)

يتم التعامل مع مفهوم الارتباط الخطي عادة من خلال حساب مقياس رقمي يعكس درجة ونوع العلاقة بين متغيرين، (ويسمى معامل **الارتباط البسيط** (Simple Correlation))، وكذلك باستخدام التمثيل البياني.

أو يمكن حساب الارتباط بين متغيرين بوجود تأثير متغير ثالث أو أكثر، (ويسمى معامل **الارتباط الجزئي** (Partial Correlation))، أو حساب الارتباط بين ثلاثة متغيرات فأكثر، (ويسمى معامل **الارتباط المتعدد** (Multiple Correlation)).

وأشهر هذه المعاملات وأكثرها استخداما من الناحية العملية هو معامل الارتباط الخطي البسيط، الذي تتراوح قيمته ما بين القيمتين 1− و1، حيث يدل اقتراب قيمته من إحداهما على وجود علاقة خطية قوية عكسية أو طردية بين المتغيرين، وعندما تقترب قيمة المعامل من الصفر فهذا يدل على ضعف العلاقة الخطية بين المتغيرين.

وتوجد عدة معاملات ارتباط مَعلمية (Parametric)، أي تعتمد على تقدير معلمة معينة، وغير مَعلمية (Non-Parametric)، لا تعتمد على تقدير المعالم، لحساب الارتباط بين متغيرين نذكر منها المعاملين التاليين:

- **معامل بيرسون للارتباط** (Pearson Correlation Coefficient):

معامل بيرسون هو مقياس معلمي يُعرف بالصيغة التالية لقياس الارتباط الخطي بين المتغيرين X و Y:

$$r_{XY} = \frac{Cov(X,Y)}{SD(X).SD(Y)} = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}$$

ولحساب معامل الارتباط البسيط في R، لبيرسون أو غيره من المعاملات، نستخدم الدالة `cor` كما سنرى من خلال استخدام البيانات `stu.data1` كمثال للتطبيق؛

في الفصل الرابع، تم استخدام التمثيل البياني بين متغيرين متمثلا بشكل الانتشار ومصفوفة شكل الانتشار للمتغيرات الكمية في البيانات `stu.data1`، وتم التعليق على طبيعة العلاقة بين هذه المتغيرات باستخدام الرسم البياني فقط. وهنا يمكن حساب قيمة معامل بيرسون للارتباط بين أي متغيرين من هذه المتغيرات باستخدام **دالة الارتباط البسيط** `cor` كالتالي:

```
> cor(s.grd1,s.grd2)
[1] 0.9593287
```

ملاحظة:

لاحظ أنه يمكن حساب قيمة الارتباط السابق من خلال علاقته بالتغاير بين المتغيرين والانحراف المعياري لكل منهما عن طريق استخدام **دالة التغاير** `cov` كالتالي:

```
> cov(s.grd1,s.grd2)/(sd(s.grd1)*sd(s.grd2))
[1] 0.9593287
```

إلا أنه سيكون من المناسب عمليا عندما نرغب باستكشاف العلاقات الخطية بين عدة متغيرات أن نقوم بإدخال كل المتغيرات الكمية[1] دفعة واحدة للحصول على مصفوفة الارتباط البسيط بالصورة التالية:

```
> cor(stu.data1[c(-5,-6)])

            grd1        grd2        grd3         age
grd1  1.00000000  0.95932871 -0.02226163 -0.41592912
grd2  0.95932871  1.00000000  0.05536356 -0.45379414
grd3 -0.02226163  0.05536356  1.00000000 -0.08210202
age  -0.41592912 -0.45379414 -0.08210202  1.00000000
fam  -0.90409356 -0.90743680 -0.10163064  0.29186983
hou   0.86522792  0.86959204  0.13629006 -0.27336391
            fam         hou
grd1 -0.9040936  0.8652279
grd2 -0.9074368  0.8695920
grd3 -0.1016306  0.1362901
age   0.2918698 -0.2733639
fam   1.0000000 -0.9255022
hou  -0.9255022  1.0000000
```

ونلاحظ بالطبع توافق درجة ونوع الارتباطات بين هذه المتغيرات مع ما تم ملاحظته من التمثيل البياني في الفصل الرابع في مصفوفة شكل الانتشار في الشكل (19.4).

من الناحية الإحصائية، تكون قيمة معامل الارتباط، (أو أي معلمة يتم تقديرها عن طريق عينة مسحوبة من المجتمع)، أكثر موثوقية عندما يتم اختبار معنويتها باستخدام الاختبار الإحصائي المناسب، ويتم اختبار معامل الارتباط بين متغيرين، (سواء معامل بيرسون أو غيره)، باستخدام دالة **اختبار معامل الارتباط** `cor.test`، وبتطبيق هذه الدالة مع المتغيرين `s.grd1` و`s.grd2` على سبيل المثال نحصل على النتيجة التالية:

```
> cor.test(s.grd1,s.grd2)

        Pearson's product-moment correlation

data:  s.grd1 and s.grd2
t = 19.5221, df = 33, p-value < 2.2e-16
alternative hypothesis:true correlation is not equal to 0
95 percent confidence interval:
 0.9202935 0.9794515
sample estimates:
      cor
0.9593287
```

[1] هذا هو السبب في كتابة `[c(-5,-6)]` لاستثناء المتغيرين النوعيين `gen` و`sem`، (واللذان يمثلان نوع الطالب وترتيب الفصل الدراسي على الترتيب)، من مصفوفة الارتباط.

وهذه النتيجة تعني رفض الفرضية الصفرية التي تنص على عدم وجود علاقة خطية معنوية بين المتغيرين؛ ($H_0: \rho_{\text{s.grd1,s.grd2}} = 0$)، مما يدل على وجود علاقة خطية قوية عند مستوى معنوية 0.05 بين المتغيرين. ولاحظ أن الدالة cor.test توفر قيمة معامل الارتباط، (أي التقدير بنقطة)، وكذلك فترة الثقة له، إضافة لقيمة إحصاءة الاختبار الحسابية.

وننوه إلى إمكانية استخدام الخيارات alternative وconf.level وغيرها لإجراء التغيير المطلوب في الاختبار السابق.

- **معامل سبيرمان لارتباط الرتب** (Spearman Rank Correlation Coefficient):

يُقدم معامل سبيرمان تقديرا لامعلميا للارتباط بين متغيرين، ويُستخدم عادة لحساب الارتباط بين المتغيرات النوعية عن طريق استخدام رتب المشاهدات وليس قيمها الفعلية، علما بأنه يمكن استخدامه أيضا عند وجود متغير كمي وآخر نوعي أو مع المتغيرات الكمية، ويُعرف هذا المعامل بالصيغة:

$$r_{XY} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

حيث d_i هو الفرق بين رتب كل من المتغيرين X وY، وn هو عدد المشاهدات.

ولحساب قيمة معامل سبيرمان في لغة R نستخدم نفس الدالة cor ولكن نقوم بتحديد طريقة الحساب عن طريق استخدام الخيار[1] method="spearman". ويجب عند التعامل مع متغيرات غير رقمية أن يتم إعطاؤها رموزا رقمية قبل إدخالها في الدالة، شأنها شأن معظم الدوال في R، وكمثال، يمكن حساب معامل سبيرمان بين المتغيرين النوعيين اللذان يمثلان النوع وترتيب الفصل الدراسي؛ s.gen وs.sem، إلا إنه يجب أولا إعطاء رموز لمتغير النوع بعد تحويله إلى متغير عاملي؛

```
> s.gen<-stu.data1$gen
> s.sem<-stu.data1$sem

> class(s.gen)
[1] "character"

> s.gen<-as.factor(s.gen)
> class(s.gen)
[1] "factor"

> s.gen2<-as.numeric(s.gen)
> class(s.gen2)
[1] "numeric"
```

[1] في الحالة الافتراضية للدالتين cor وcor.test يتم استخدام الخيار method="spearman" حتى وإن لم تتم كتابته.

الآن نكتب:

```
> cor(s.gen2,s.sem,method="spearman")
[1] 0.1689422
```

مما يدل على وجود علاقة طردية ضعيفة بين النوع وترتيب الفصل الدراسي، ويمكن اختبار هذه العلاقة بالصورة:

```
> cor.test(s.gen2,s.sem,method="spearman")

        Spearman's rank correlation rho

data:  s.gen2 and s.sem
S = 5933.752, p-value = 0.332
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1689422

Warning message:
In cor.test.default(s.gen2, s.sem, method = "spearman"):
  Cannot compute exact p-value with ties
```

وبذلك يتأكد عدم وجود علاقة خطية ذات معنوية بين المتغيرين.

وكمثال إضافي على استخدام معامل سبيرمان مع متغيرات كمية، يمكن حساب الارتباط بين درجات الطلبة في المقررين الأولين في البيانات stu.data1:

```
> cor.test(s.grd1,s.grd2,method="spearman")

        Spearman's rank correlation rho

data:  s.grd1 and s.grd2
S = 421.5886, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.940954

Warning message:
In cor.test.default(s.grd1, s.grd2, method = "spearman") :
  Cannot compute exact p-value with ties
```

ونلاحظ توافق هذه النتيجة مع نتيجة معامل بيرسون سابقا. ويمكن استخدام الدالة cor لحساب مصفوفة الارتباط بين المتغيرات في البيانات stu.data1 باستخدام طريقة معامل سبيرمان عن طريق كتابة الأمر؛

```
.cor(stu.data1[-5],method="spearman")
```

1.1.5.6 معامل الارتباط الجزئي (Partial Correlation Coefficient)

يتم حساب معامل الارتباط الجزئي بين المتغيرين X و Y بوجود تأثير المتغير Z باستخدام الصيغة التي تحتوي على معاملات الارتباط البسيط بين المتغيرات الثلاثة كالتالي:

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}\, r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

وكذلك يمكن حساب الارتباط الجزئي بين المتغيرين X و Y بوجود تأثير المتغيرين Z و W باستخدام الصيغة التي تحتوي على معاملات الارتباط الجزئي بين المتغيرات كالتالي:

$$r_{XY.ZW} = \frac{r_{XY.Z} - r_{XW.Z}\, r_{YW.Z}}{\sqrt{(1 - r_{XW.Z}^2)(1 - r_{YW.Z}^2)}}$$

ويمكن إيجاد معاملات الارتباط الجزئي عن طريق حساب الارتباطات البسيطة بين المتغيرات المطلوبة باستخدام دالة `cor` والتعويض بها في إحدى الصيغتين أعلاه، فمثلا، في البيانات `stu.data1` إذا كان المطلوب هو حساب قيمة الارتباط بين درجات الطلبة في المقرر `s.grd1` ودرجات الطلبة في المقرر `s.grd2` بوجود متغير العمر `s.age`، فيمكننا كتابة صيغة قانون الارتباط الجزئي بالصورة التالية:

```
> corXYZ<-(cor(s.grd1,s.grd2)-cor(s.grd1,s.age)*
cor(s.grd2,s.age))/sqrt((1-(cor(s.grd1,s.age))^2)*
(1-(cor(s.grd2,s.age))^2))

> corXYZ
[1] 0.9509025
```

وهذا قد يدل على وجود ارتباط طردي قوي بين درجات الطلبة في المقررين الأول والثاني بوجود تأثير متغير العمر.

إلا أن هذه الطريقة قد تستغرق وقتا طويلا في الإدخال عند التعامل مع عدة تباديل لمعاملات الارتباط الجزئي أو عند حساب الارتباط الجزئي لأربعة متغيرات أو أكثر، لذلك سيكون من الأنسب اللجوء لاستخدام الدوال التي توفرها الحزمة الإضافية `ppcor` المخصصة للتعامل مع معاملات الارتباط الجزئي والارتباط شبه الجزئي[1]. فالمثال السابق يمكن إعادة تنفيذه باستخدام الدالة `pcor` بالصورة التالية، علما بأن هذه الدالة توفر قيم معاملات الارتباط الجزئي بين كل متغيرين من المتغيرات المُدخلة بوجود باقي المتغيرات، وتوفر كذلك نتائج اختبار هذه الارتباطات المتمثلة في القيم الاحتمالية وقيم إحصاءات الاختبار، إضافة إلى عدد المشاهدات، عدد المتغيرات الهامشية، وطريقة الارتباط البسيط التي تم استخدامها، (وهي طريقة بيرسون في الحالة الافتراضية):

[1] الارتباط شبه الجزئي (Semi-partial Correlation) هو الارتباط بين متغيرين بحيث يتم أخذ تأثير متغير ثالث أو أكثر في الاعتبار بالنسبة للمتغير الثاني فقط.

```
> library(ppcor)

> pcor(stu.data1[c(1,2,4)])

$estimate
           grd1       grd2         age
grd1 1.00000000  0.9509025  0.07715562
grd2 0.95090248  1.0000000 -0.21339387
age  0.07715562 -0.2133939  1.00000000

$p.value
             grd1         grd2       age
grd1 0.000000e+00 1.158464e-67 0.6615581
grd2 1.158464e-67 0.000000e+00 0.2166078
age  6.615581e-01 2.166078e-01 0.0000000

$statistic
           grd1      grd2        age
grd1  0.0000000 17.380545  0.4377631
grd2 17.3805455  0.000000 -1.2355985
age   0.4377631 -1.235598  0.0000000

$n
[1] 35

$gp
[1] 1

$method
[1] "pearson"
```

ومن هذه النتائج نلاحظ على سبيل المثال وجود علاقة طردية قوية ذات معنوية بين المتغيرين s.grd1 وs.grd2 بوجود المتغير s.age، ووجود علاقات ضعيفة غير معنوية بين كلا من s.grd1 وs.age بوجود s.grd2، وأيضا s.grd2 وs.age بوجود s.grd1. ويمكن استخدام الخيار method="spearman" لحساب معاملات الارتباط البسيط بطريقة سبيرمان عوضا عن طريقة بيرسون.

ويمكن الحصول على نفس النتيجة ولكن بشكل أكثر اختصارا باستخدام الدالة pcor.test ضمن نفس الحزمة، إلا أن هذه الدالة تصلح للتعامل مع ثلاثة متغيرات فقط، فمثلا يمكن حساب واختبار معامل الارتباط الجزئي بين المتغيرين s.grd1 وs.grd2 بوجود متغير النوع المرموز s.gen2 باستخدام طريقة سبيرمان بالصورة التالية:

```
> pcor.test(s.grd1,s.grd2,s.gen2,method="spearman")
   estimate      p.value statistic  n gp   Method
1 0.8730431 4.165419e-24  10.12765 35  1 spearman
```

ولاحظ أن إدخال المتغيرات الثلاثة يجب أن يتم بشكل منفصل وليس كإطار بيانات أو مصفوفة كما هو الحال مع الدالة pcor، ومن هذه النتيجة نستنتج وجود علاقة طردية قوية ذات معنوية بين أداء الطلبة في المقررين الأولين بوجود تأثير متغير النوع.

وكمثال آخر على تطبيق الدالة pcor، يمكن حساب واختبار معاملات الارتباط الجزئي بين المتغيرات s.grd1، s.grd2، s.grd3، وs.age كالتالي:

```
> pcor(stu.data1[c(1,2,3,4)])

$estimate
            grd1       grd2        grd3         age
grd1  1.00000000  0.9542194 -0.26380135  0.06242748
grd2  0.95421942  1.0000000  0.25760643 -0.19434208
grd3 -0.26380135  0.2576064  1.00000000 -0.04518195
age   0.06242748 -0.1943421 -0.04518195  1.00000000

$p.value
             grd1          grd2      grd3       age
grd1 0.000000e+00 1.381836e-70 0.1278280 0.7276443
grd2 1.381836e-70 0.000000e+00 0.1377054 0.2699913
grd3 1.278280e-01 1.377054e-01 0.0000000 0.8011805
age  7.276443e-01 2.699913e-01 0.8011805 0.0000000

$statistic
           grd1      grd2       grd3        age
grd1  0.0000000 17.762405 -1.5227231  0.3482608
grd2 17.7624054  0.000000  1.4843902 -1.1030825
grd3 -1.5227231  1.484390  0.0000000 -0.2518196
age   0.3482608 -1.103083 -0.2518196  0.0000000

$n
[1] 35

$gp
[1] 2

$method
[1] "pearson"
```

ويُلاحَظ وجود علاقة واحدة ذات معنوية هي تلك التي بين المتغيرين s.grd1 وs.grd2 بوجود تأثير كلا من المتغيرين s.grd3 وs.age. أما بالنسبة للارتباط الخطي المتعدد فسنتطرق لطريقة حسابه في البند التالي عند تناول توفيق نموذج الانحدار الخطي.

2.5.6 تحليل الانحدار الخطي (Linear Regression Analysis)

كما رأينا أن معامل الارتباط يُستخدم لقياس درجة وطبيعة العلاقة بين متغيرين أو أكثر، يقوم نموذج الانحدار بتزويدنا بمعلومات إضافية حول مدى تأثير متغير أو أكثر على متغير ما ضمن منظومة من العلاقات السببية. وضمن إطار الانحدار الخطي، يمكننا كتابة نموذج الانحدار الخطي المتعدد لـ p **متغير توضيحي** (Explanatory variable) على المتغير التابع (Dependent variable) أو متغير **الاستجابة** (Response) Y بالصورة التالية:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

حيث $(x_{ij},\ i = 1, \dots, n,\ j = 1, \dots, p)$ هي قيمة المشاهدة رقم i للمتغير التوضيحي j، ومعالم النموذج $\beta_0,\ \beta_1, \dots,\ \beta_p$ هي معاملات الانحدار التي سيتم تقديرها، (عادة بطريقة **المربعات الصغرى الاعتيادية** (Ordinary Least Squares, (OLS))، و ϵ_i هو حد الخطأ العشوائي للنموذج. ويمكن كتابة النموذج السابق بصيغة المصفوفات للاختصار كالتالي:

$$Y = X\beta + \epsilon$$

حيث يمثل Y متجه قيم المتغير التابع، X مصفوفة المتغيرات التوضيحية، β متجه معالم الانحدار، و ϵ متجه قيم الخطأ العشوائي.

وتقوم طريقة المربعات الصغرى للتقدير على عدة فرضيات أهمها أن مصفوفة المتغيرات التوضيحية X تضم متغيرات غير عشوائية ومستقلة خطيا عن بعضها البعض، وأن يكون $\epsilon_i \sim NID(0, \sigma^2)$، وغير ذلك من الفرضيات. وعندما يكون عدد المتغيرات التوضيحية $p = 1$ يُسمى النموذج بنموذج الانحدار الخطي البسيط.

ولتوفيق نموذج الانحدار الخطي في R فإننا نستخدم دالة **النموذج الخطي** lm، وهي دالة عامة لكافة النماذج التي تأخذ الشكل الخطي وليست لنماذج الانحدار فقط. وكمثال لنفرض أننا مهتمون بدراسة تأثير كلا من عمر الطالب وترتيب الفصل الدراسي له وعدد أفراد أسرته وعدد حجرات منزله على درجاته في المقرر grd1، عندها نستطيع توفيق نموذج الانحدار الخطي؛

$$grd1 = \beta_0 + \beta_1 age + \beta_2 sem + \beta_3 fam + \beta_4 hou + \epsilon$$

بالصورة التالية:

```
> reg1<-lm(grd1~age+sem+fam+hou,data=stu.data1)
> reg1

Call:
lm(formula= grd1 ~age + sem + fam + hou, data= stu.data1)

Coefficients:
(Intercept)         age         sem         fam         hou
    121.250      -1.519      -1.182      -3.308       2.316
```

وهذه النتيجة تتضمن شكل العلاقة الخطية المُستخدمة، وتقديرات معالم النموذج فقط. لاحظ أن متغير الاستجابة أو المتغير التابع grd1 يتم إدخاله أولا قبل العلامة "~" ويلي ذلك المتغيرات التوضيحية، ثم استخدام الخيار data لتوضيح أن المتغيرات المُدخلة هي ضمن هذه البيانات[1]. وقد تم تعيين اسم لنموذج الانحدار لحفظ النتائج، ولأننا غالبا ما نرغب في الحصول على نتائج إضافية، فمثلا يمكن استخدام الدالة summary للحصول على الأخطاء المعيارية لتقديرات المعالم؛ $SE(\hat{\beta})$ وقيم إحصاءات الاختبار الخاصة بها $t_c = \frac{|\hat{\beta}|}{SE(\hat{\beta})}$ والقيم الاحتمالية لاختبار الفرضيات $(H_0: \beta_j = 0,\ j = 1, \dots, p)$ ونتيجة اختبار تحليل التباين للنموذج، وغير ذلك من النتائج الإضافية كما نلاحظ فيما يلي؛

```
> summary(reg1)

Call:
lm(formula= grd1 ~age + sem + fam + hou, data= stu.data1)
Residuals:
     Min       1Q   Median       3Q      Max
-16.5604  -2.4909  -0.5585   2.7540  13.1160

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 121.2501    20.0211   6.056 1.19e-06 ***
age          -1.5192     0.6610  -2.298 0.028697 *
sem          -1.1821     0.5812  -2.034 0.050888 .
fam          -3.3076     0.9064  -3.649 0.000992 ***
hou           2.3158     2.1507   1.077 0.290164
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.07 on 30 degrees of freedom
Multiple R-squared: 0.8665,   Adjusted R-squared:  0.8487
F-statistic: 48.68 on 4 and 30 DF,  p-value: 1.066e-12
```

[1] يمكن إدخال متغيرات النموذج في الدالة lm دون استخدام الخيار data عندما تكون هذه المتغيرات معرفة في R خارج إطار البيانات.

أهم ما يُلاحظ من النتيجة السابقة أن النموذج في العموم ذو معنوية أو أهمية إحصائية، (حيث أن القيمة الاحتمالية لاختبار الفرضية الصفرية (نموذج الانحدار غير معنوي :H_0) كانت أقل من القيمة 0.05)، إضافة إلى أنه يمكن القول من قيمة معامل التحديد[1] أن النموذج يفسر أكثر من 86% من الاختلاف الكلي أو العلاقة السببية بين المتغير التابع وباقي المتغيرات. ونرى أيضا أن قيمة الانحراف المعياري للبواقي هي $\hat{\sigma}_\varepsilon = \sqrt{MSE} = 6.07$.

ومن مراقبة تأثير كل متغير توضيحي على المتغير التابع يمكن أن نستنتج أن عمر الطالب وعدد أفراد أسرته لهما تأثير عكسي قوي على درجة الطالب في المقرر grd1 يليهما ترتيب الفصل الدراسي للطالب، بمعنى أنه بزيادة عمر الطالب وعدد أفراد الأسرة وترتيب الفصل الدراسي فإن درجات الطالب في المقرر grd1 تتناقص. أما عدد حجرات المنزل فكان له تأثير طردي ولكن ليس ذو أهمية على درجة الطالب.

ومن ضمن النتائج المعروضة ملخص للقيم الخمسة لقيم الخطأ المُقدَر أو **البواقي** (Residuals) للمساعدة في تقدير سلوك حد الخطأ العشوائي فيما إذا طبيعيا أم لا، إلا أن اللجوء لاختبار طبيعية التوزيع أو الرسم البياني قد يكون أفضل للتأكد كما سنرى لاحقا.

وإذا ما رغبنا في الحصول على جدول تحليل التباين للنموذج بشكل تفصيلي فيمكننا استخدام دالة تحليل التباين anova مع النموذج كالتالي:

```
> anova(reg1)

Analysis of Variance Table

Response: grd1
          Df Sum Sq Mean Sq  F value    Pr(>F)
age        1 1432.3  1432.3  38.8789 7.237e-07 ***
sem        1  813.7   813.7  22.0859 5.435e-05 ***
fam        1 4885.6  4885.6 132.6130 1.551e-12 ***
hou        1   42.7    42.7   1.1594    0.2902
Residuals 30 1105.2    36.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

والذي يقدم بدوره نتيجة مفصلة لمجاميع ومتوسطات مجاميع المربعات للمتغيرات والبواقي ودرجات الحرية وقيم إحصاءات الاختبار. وإذا ما رغبنا في عرض جدول تحليل التباين للنموذج بالصورة التقليدية المختصرة فيمكننا

[1] قيمة معامل التحديد لنموذج الانحدار هي في الواقع مربع قيمة معامل الارتباط المتعدد بين المتغير التابع وباقي المتغيرات التوضيحية. ففي مثالنا يكون $r_{grd1(age,sem,fam,hou)} = \sqrt{0.8665} = 0.9309$.

تعريف نموذج خطي جديد يتضمن المعلمة β_0 فقط، ثم يتم استخدام دالة anova مع النموذجين الأصلي والجديد كالتالي:

```
> reg1.1<-lm(grd1~1,data=stu.data1)

> anova(reg1.1,reg1)

Analysis of Variance Table

Model 1: grd1 ~ 1

Model 2: grd1 ~ age + sem + fam + hou

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     34 8279.5
2     30 1105.2  4    7174.3 48.684 1.066e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

حيث أن مجموع المربعات للانحدار يساوي 7174.3 ومجموع المربعات للبواقي يساوي 1105.2 والقيمة الحسابية لإحصاءة الاختبار F يساوي 48.684، وتكون القيمة الاحتمالية للاختبار هي نفس التي تم الحصول عليها سابقا.

إضافة إلى ما قد سبق، يمكن استخدام مجموعة من الدوال للحصول على نتائج محددة ضمن نموذج الانحدار الخطي الذي تم توفيقه، من هذه الدوال؛ دالة **معالم النموذج** coef للحصول على قيم معالم النموذج المقدرة فقط:

```
> coef(reg1)

(Intercept)         age         sem         fam         hou
 121.250128  -1.519219   -1.182092  -3.307620    2.315770
```

ويتم تقدير فترات الثقة لمعالم النموذج باستخدام دالة **فترة الثقة** confint كالتالي:

```
> confint(reg1)

                2.5 %       97.5 %
(Intercept) 80.361600 162.138655615
age         -2.869261  -0.169177953
sem         -2.369052   0.004868617
fam         -5.158779  -1.456460380
hou         -2.076451   6.707991062
```

ويمكن من هذه الفترات القول بأننا واثقون بنسبة 95% بأن تقدير معامل الانحدار لمتغير العمر مثلا لن يقل عن القيمة 2.869- ولن يزيد عن القيمة 0.169-.

وللحصول على قيم المتغير التابع المقدرة $\hat{Y}$ نستخدم[1] دالة **القيم المقدرة** fitted كالتالي، مع تعيين اسم لتلك القيم لغرض استخدامها لاحقا:

```
> fit1<-fitted(reg1)
> fit1
```

```
    stu1     stu2     stu3     stu4     stu5     stu6     stu7
55.83637 51.17595 55.49924 67.80582 41.45432 60.27767 62.93057
    stu8     stu9    stu10    stu11    stu12    stu13    stu14
87.07070 87.79337 73.69839 52.17373 71.84204 79.92809 74.97895
   stu15    stu16    stu17    stu18    stu19    stu20    stu21
83.54231 40.27223 45.55850 54.87505 82.62940 77.94945 84.48575
   stu22    stu23    stu24    stu25    stu26    stu27    stu28
72.88395 75.77550 85.76631 64.86585 83.23571 86.22573 70.08417
   stu29    stu30    stu31    stu32    stu33    stu34    stu35
91.56041 69.81499 78.23816 75.77550 87.67701 70.42130 91.89754
```

أما للحصول على قيم البواقي، فنستخدم دالة **البواقي** resid:

```
> resid1<-resid(reg1)
> resid1
```

```
      stu1       stu2       stu3       stu4       stu5
-0.8363695 -2.1759497  4.5007582 -2.8058179 -6.4543248
      stu6       stu7       stu8       stu9      stu10
10.7223322 10.0694344  2.9293031  0.2066345  1.3016111
     stu11      stu12      stu13      stu14      stu15
-2.1737350  5.1579582 -0.9280877 -8.9789533 -3.5423091
     stu16      stu17      stu18      stu19      stu20
-0.2722332 -0.5584953 -3.8750459 -0.6293986 -2.9494453
     stu21      stu22      stu23      stu24      stu25
-0.4857457 13.1160488  1.2244960  2.2336900 -0.8658520
     stu26      stu27      stu28      stu29      stu30
 5.7642925  3.7742670 -7.0841675 -16.5604083 -0.8149864
     stu31      stu32      stu33      stu34      stu35
-1.2381599 -7.7755040  5.3229943  2.5787048  2.1024640
```

من جديد، يمكن استخدام[2] الدالة summary بحيث يتم تعيين اسم لها، وليكن summ.reg1 مثلا:

```
> summ.reg1<-summary(reg1, correlation=T)
```

---

[1] يمكن استخدام الأمر predict(reg1) للحصول على نفس النتيجة، وسنتطرق للدالة predict لاحقا في هذا البند.

[2] تم إدراج الخيار correlation=T لكي نتمكن من استدعاء مصفوفة الارتباط لاحقا.

بحيث يتم استدعاء نتائج محددة منها للنموذج reg1 بصورة سريعة، فمثلا لاستدعاء مصفوفة التغاير بين معالم النموذج المقدرة نكتب:

```
> summ.reg1$cov
```

```
            (Intercept)           age           sem           fam
(Intercept) 10.88038789 -0.2519647387 -0.0335820550 -0.290341272
age         -0.25196474  0.0118613904 -0.0003767671 -0.001684993
sem         -0.03358206 -0.0003767671  0.0091688266 -0.001077184
fam         -0.29034127 -0.0016849927 -0.0010771835  0.022301244
hou         -0.81684176  0.0003244760  0.0004909213  0.048438987
                      hou
(Intercept) -0.8168417551
age          0.0003244760
sem          0.0004909213
fam          0.0484389872
hou          0.1255480009
```

ولاستدعاء مصفوفة الارتباط بين معالم النموذج المقدرة نكتب:

```
> summ.reg1$correlation
```

```
             (Intercept)           age         sem         fam
(Intercept)   1.0000000 -0.701374187 -0.10632324 -0.58941617
age          -0.7013742  1.000000000 -0.03612837 -0.10360141
sem          -0.1063232 -0.036128365  1.00000000 -0.07533005
fam          -0.5894162 -0.103601414 -0.07533005  1.00000000
hou          -0.6988932  0.008408333  0.01446938  0.91543099

                     hou
(Intercept) -0.698893240
age          0.008408333
sem          0.014469385
fam          0.915430991
hou          1.000000000
```

ولاستدعاء قيمة معامل التحديد نكتب:

```
> summ.reg1$r.squared
[1] 0.8665108
```

ويمكن للقارئ استخدام دالة المساعدة help(summary.lm) للحصول على المزيد من الاوامر المتعلقة بنتائج تحليل الانحدار.

الآن سنقوم بإجراء بعض الاختبارات الخاصة بتحقق فرضيات النموذج، والتي من ضمنها اختبار توزع البواقي بالتوزيع الطبيعي؛

```
> shapiro.test(resid1)

        Shapiro-Wilk normality test
data:  resid1
W = 0.9626, p-value = 0.2741
```

وهذا يعني توزع بواقي النموذج reg1 بالتوزيع الطبيعي مما يدل على ثبات الفرضية. وكذلك يمكن اختبار وجود علاقة بين قيم المتغير التابع المقدرة والبواقي بالصورة:

```
> cor.test(fit1,resid1)

        Pearson's product-moment correlation

data:  fit1 and resid1
t = 0, df = 33, p-value = 1
alternative hypothesis:true correlation is not equal to 0
95 percent confidence interval:
 -0.3332465  0.3332465

sample estimates:
          cor
-2.926716e-17
```

ونستطيع ملاحظة عدم وجود علاقة معنوية بين القيم المقدرة والبواقي وهذا يدل على استقرار أو صحة النموذج. وسيتم ملاحظة هذه النتائج بيانيا من خلال استخدام التمثيل البياني لنتائج تحليل البواقي في البند القادم.

- **التنبؤ** (Prediction)

تُعد دالة **التنبؤ** predict هي الأخرى من الدوال الهامة والتي كثيرا ما تُستخدم مع النماذج الإحصائية لغرض التقدير، وفي نماذج الانحدار الخطي يمكن استخدامها للحصول على قيم المتغير التابع المقدرة (كبديل للدالة fitted)، ويمكن استخدامها أيضا للحصول على القيم المقدرة مع خيار الحصول على فترات الثقة له، كما نرى من المثال التالي:

```
> predict(reg1,interval="confidence")

           fit      lwr      upr
stu1  55.83637 50.42960 61.24314
stu2  51.17595 43.88248 58.46942
stu3  55.49924 49.38051 61.61797
…       …          …          …

stu35 91.89754 88.01600 95.77907
```

إلا أن الاستخدام الأهم عمليا لهذه الدالة هو لغرض التنبؤ بقيم جديدة للمتغير التابع، فمثلا إذا أردنا التنبؤ بدرجات ثلاثة طلبة في المقرر grd1، (خارج العينة المستخدمة في توفيق النموذج)، لهم البيانات الموجودة في الجدول (3.6)؛

جدول 3.6: بيانات الطلبة الجدد في النموذج reg1

	age	sem	fam	hou
1	26	9	14	6
2	25	10	13	7
3	27	10	9	9

فإننا نقوم أولا بإدخال هذه البيانات الجديدة في إطار بيانات على الصورة:

```
> new.reg1<-data.frame(age=c(26,25,27),sem=c(9,10,10),
fam=c(14,13,9),hou=c(6,7,9))

> new.reg1

  age sem fam hou
1  26   9  14   6
2  25  10  13   7
3  27  10   9   9
```

ثم نستخدم الخيار newdata لإدخال البيانات الجديدة في دالة التنبؤ predict فنحصل على تقديرات درجات الطلبة الثلاث كالتالي:

```
> predict(reg1,newdata=new.reg1)

       1        2        3
38.69954 44.66006 59.48364
```

ويُلاحظ أنها درجات منخفضة مقارنة بدرجات الطلبة في عينة الدراسة.

- **تحديث النموذج** (Updating the Model)

بعد توفيق نموذج الانحدار الخطي، قد نرغب في إضافة أو حذف متغير توضيحي أو أكثر بغرض المقارنة بين نتائج هذه النماذج، ودالة التحديث[1] update تمكننا من عمل هذه التغييرات في النموذج الأصلي دون الحاجة لكتابة نموذج جديد في كل مرة نرغب فيها بعمل تغيير أو تحديث.

فمثلا في النموذج reg1، إذا أردنا حذف المتغير hou من ضمن المتغيرات التوضيحية ثم توفيق النموذج الجديد، نقوم بكتابة:

```
> reg1.up1<-update(reg1,~.-hou)
```

فنحصل على النموذج الجديد أو "المُحَدّث" reg1.up1 والذي يمكن التعامل معه بنفس الآلية السابقة؛

[1] يمكن استخدام الدالة update مع معظم النماذج الإحصائية وليس فقط نماذج الانحدار الخطي.

```
> reg1.up1

Call:
lm(formula = grd1 ~ age + sem + fam, data = stu.data1)

Coefficients:
(Intercept)          age          sem          fam
    136.317       -1.525       -1.191       -4.201
```

وكذلك يمكن إضافة المتغيرين grd2 وgrd3 للنموذج reg1 بنفس الكيفية:

```
> reg1.up2<-update(reg1,~.+grd2+grd3)
> reg1.up2

Call:
lm(formula = grd1 ~ age + sem + fam + hou + grd2 + grd3,
data = stu.data1)

Coefficients:
(Intercept)          age          sem          fam
   35.58588     -0.24934     -0.58376     -0.97394

        hou         grd2         grd3
    0.96578      0.70618     -0.09864
```

ملاحظات:

1. لتطبيق طريقة **بناء النموذج بالتدرّج** (Stepwise Model Selection) لاختيار النموذج الأمثل بالمتغيرات المتاحة، يتم استخدام دالة step مع النموذج الموفق بالخيار direction الذي يأخذ الطريقة التقدمية "forward" أو يأخذ الطريقة التراجعية "backward". فمثلا يمكن تطبيق طريقة الاختيار التراجعية للنموذج reg1 بكتابة الأمر:
   .step(reg1,direction="backward")

2. تُستخدم الدالة drop1 هي الأخرى مع النموذج الموفق لبناء النموذج الأفضل، فهي تقوم بحذف متغير توضيحي واحد من المتغيرات المدرجة في النموذج الأصلي الموفق وإعادة التوفيق في كل مرة وحساب تأثير ذلك الحذف على النتيجة الكلية للنموذج عن طريق تقديم جدول تحليل للتباين يوضح التأثير الناتج عن ذلك الحذف.

فمثلا، يمكن كتابة الأمر drop1(reg1) لمعرفة أي من المتغيرات التوضيحية الأربعة سيتم ترشيحه للحذف من النموذج reg1.

3. على عكس طريقة عمل الدالة drop1، يمكن استخدام الدالة add1 لبناء النماذج الخطية عن طريق إضافة متغير توضيحي واحد في كل مرة وحساب النتيجة المترتبة عن تلك الإضافة. فعلى سبيل المثال، يمكن كتابة الأمر add1(reg1.1,~age+sem+fam) لتحديد ما إذا كانت إضافة أي من المتغيرين sem أو fam سيكون ذو أهمية للنموذج reg1.1.

4. للمزيد من التحليل المتقدم لنتائج نموذج الانحدار الموفق يمكن استخدام المساعدة مع الدالة influence.measures، والتي تتضمن عدد من الدوال الفرعية المرافقة التي تقدم نتائج تفصيلية عن أداء كل مشاهدة ضمن كل متغير توضيحي في النموذج مثل توضيح المشاهدات التي لها أكبر تأثير ضمن نموذج الانحدار، وبالطبع لن يكون ذلك عمليا عند التعامل مع البيانات ذات الأحجام الضخمة.

1.2.5.6 التمثيل البياني للانحدار الخطي (Graphical Display for Linear Regression)

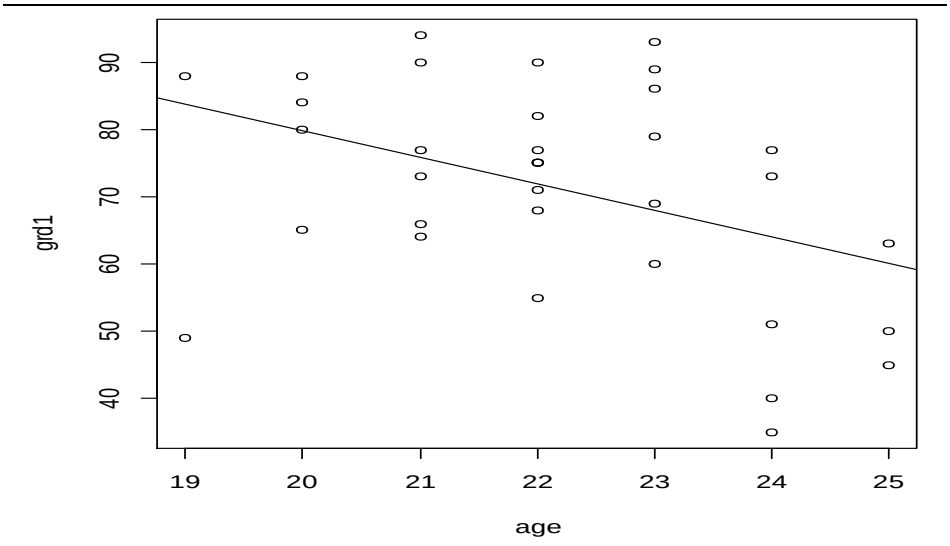
يُعد التمثيل البياني من الأدوات الهامة، إلى جانب اختبارات الفروض التقليدية، التي تساعد الباحث في تحليل طبيعة العلاقات المتشابكة بين المتغيرات من جهة، وفي مراقبة تحقق أو مخالفة فرضيات طريقة المربعات الصغرى في النموذج المُوفق من جهة أخرى.

ومن أهم تلك الرسومات البيانية رسم خط الانحدار المُوفق على شكل انتشار المتغير التابع مع كل متغير توضيحي على حدة، (حيث أن الرسم ذو بعدين فلا يمكن أن يتضمن أكثر من متغيرين من متغيرات النموذج). وعلى سبيل المثال لنقم بتوفيق نموذج الانحدار البسيط التالي:

```
> reg2<-lm(grd1~age,data=stu.data1)
```

بعدها يمكن رسم شكل الانتشار للمتغيرين ورسم خط انحدار grd1 على age، (شكل (1.6))، باستخدام الأوامر التالية:

```
> plot(grd1~age,data=stu.data1)
> abline(reg2)
```



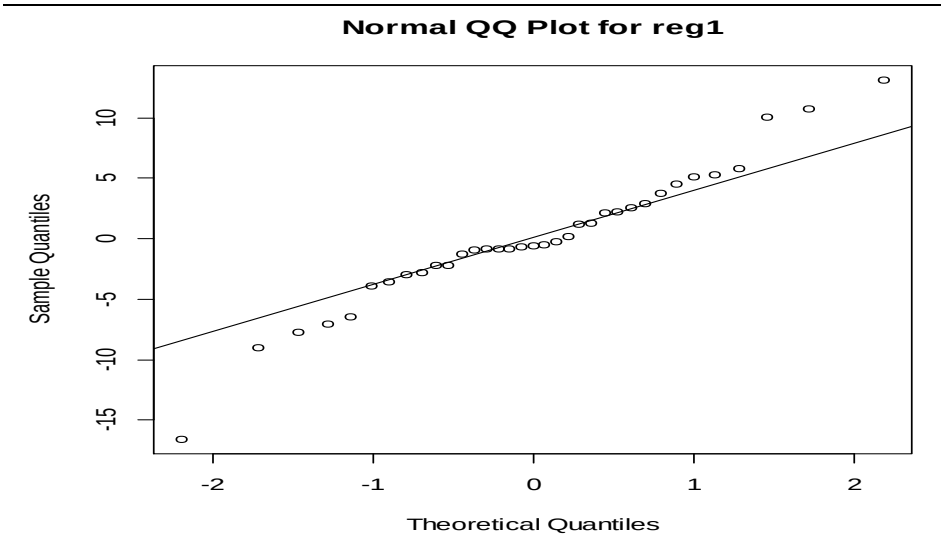
شكل 1.6: شكل انتشار مع خط انحدار المتغير grd1 على المتغير age

ويمكن ملاحظة العلاقة العكسية بين المتغيرين بوضوح. وكإضافة على الشكل السابق، يمكن للمستخدم تنفيذ الدالة segments مباشرة بعد سطري الأوامر السابقين للحصول على خطوط مستقيمة توضح المسافات بين النقاط المنتشرة في الرسم وخط الانحدار، أي البواقي. ويتم إدخال المتغيرات داخل هذه الدالة بالترتيب التالي؛ $(x_1, \hat{y}_1, x_2, y_2)$، أي أنه بالنسبة لمثالنا نكتب التالي:

```
> plot(grd1~age,data=stu.data1)
> abline(reg2)
> segments(s.age,fitted(reg2),s.age,s.grd1)
```

ومن ضمن الرسومات البيانية الهامة في تحليل البواقي، والتي تأتي كإضافة بيانية للاختبارات التي تم تنفيذها في البند السابق، هي رسم QQ الطبيعي للبواقي ورسم شكل انتشار قيم المتغير التابع المقدرة مع قيم البواقي، وكمثال لنقم بتنفيذ الرسمين للنموذج الأول الموفق reg1، حيث سيمثل الشكل (2.6) رسم QQ الطبيعي للبواقي؛

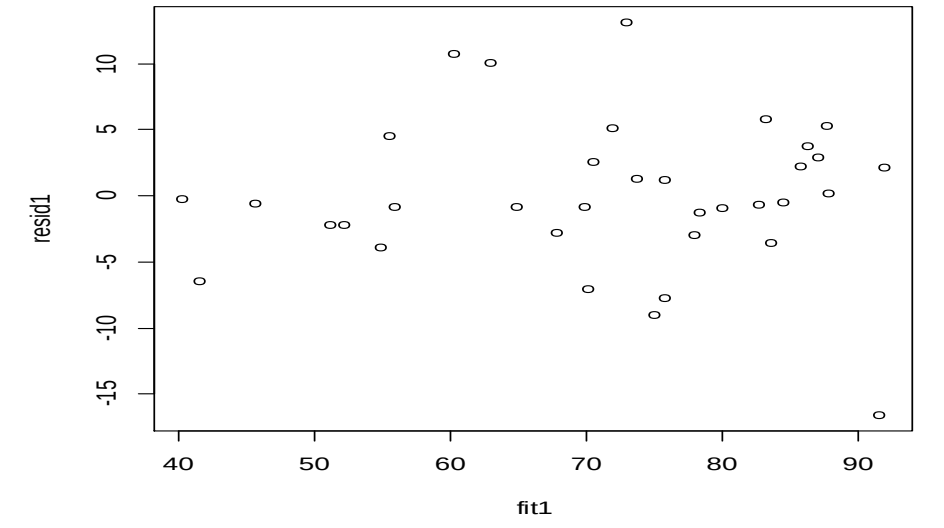
```
> qqnorm(resid1,main="Normal QQ Plot for reg1")
> qqline(resid1)
```



شكل 2.6: رسم QQ الطبيعي للبواقي في النموذج reg1

ويُلاحظ اقتراب توزع البواقي من التوزيع الطبيعي مما يدل على عدم مخالفة الفرضية الخاصة بتوزع البواقي طبيعيا. (ويمكن أيضا استخدام رسم المدرج التكراري للبواقي لمراقبة توزيعها عن طريق كتابة؛ `hist(residuals(reg1))`)، ويمثل الشكل (3.6) رسم شكل انتشار قيم المتغير grd1 المقدرة ضد قيم البواقي:

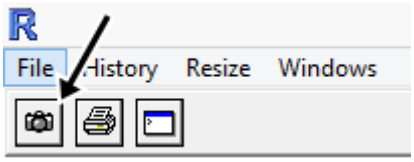
```
> plot(fit1,resid1)
```



شكل 3.6: شكل انتشار قيم المتغير grd1 المقدرة ضد قيم البواقي

ومن الشكل يمكن استنتاج عدم وجود علاقة خطية قوية بين القيم المقدرة والبواقي مما يدل على صحة النموذج. ويمكن بالمثل رسم شكل الانتشار للبواقي ضد قيم أي من المتغيرات التوضيحية بنفس الطريقة.

وإضافة إلى ذلك، يمكن ببساطة استخدام دالة الرسم plot مع النموذج الموفق للحصول على أربع رسومات مباشرة هي على الترتيب شكل انتشار قيم المتغير التابع المقدرة ضد قيم البواقي، رسم QQ الطبيعي للبواقي، شكل انتشار قيم المتغير التابع المقدرة ضد الجذر التربيعي لقيم البواقي المعيارية، وقيم المشاهدات **المُؤثِرة** (Leverage) ضد قيم البواقي المعيارية. مع ملاحظة أن ظهور هذه الرسومات الأربع سيكون تفاعليا (Interactive)، بمعنى أنه بمجرد تنفيذ الدالة، (ولتكن على سبيل المثال plot(reg1))، فستفتح نافذة فرعية خالية للرسم وبمجرد النقر بالفأرة على هذه النافذة، (أو استخدام زر الإدخال)، ستبدأ هذه الرسومات بالظهور تباعا عند كل نقرة، ولتخزين هذه الرسومات الواحدة تلو الأخرى يتم الضغط على الأيقونة المُشار إليها في الشكل (4.6) التي تأخذ شكل الكاميرا، والتي تقوم بنسخ الرسم الظاهر في النافذة الفرعية، ثم يتم لصق الرسم في المكان المطلوب، (صفحة وورد أو اكسل على سبيل المثال)، وهكذا بالمثل لباقي الرسومات.



شكل 4.6: أيقونة نسخ الرسم في النافذة الفرعية في R

6.6 توفيق النماذج الإحصائية بصورة عامة (Fitting Statistical Models in General)

تناولنا في البنود السابقة كيفية توفيق أو تقدير معالم نموذج الانحدار الخطي والنماذج الخطية الخاصة بتحليل التباين، وسنستعرض في هذا البند كيفية تكوين وتوفيق صيغ النماذج الإحصائية الخطية وغير الخطية في العموم بلغة R، وسنقوم بعرض كيفية التنفيذ العملي لهذه النماذج دون الخوض في تفاصيل تفسير النتائج لأن الهدف هنا هو تعلم تنفيذ هذه النماذج فقط.

الدالة الأساسية التي سنتعامل معها هنا هي دالة النماذج الخطية lm والتي تُستخدم في R لتوفيق النماذج غير الخطية أيضا كما ذكرنا سابقا، وسنقوم باستخدام المتغيرات التي تم تعريفها في هذا الفصل لكي يسهل على المُستخدم المقارنة بين النتائج كلما أمكن ذلك. وننوه هنا أيضا أن الدوال الفرعية مثل summary، anova، coef وغيرها يمكن استخدامها مع النماذج التي سيتم عرضها تِباعا فيما يلي؛

- **نموذج الانحدار الخطي:**

ذلك النموذج الذي تم تناوله سابقا باستخدام المتغيرات s.grd1، s.age، s.fam وغيرها ضمن إطار البيانات stu.data1، (ولاحظ ضرورة تعريف المتغير s.fam هنا باستخدام الأمر s.fam<-stu.data1$fam حيث أنه غير مُعرف حتى الآن في work6)، وتكون الصيغة العامة له هي:

```
> lm(s.grd1~1+s.age+s.fam)

Call:
lm(formula = s.grd1 ~ s.age + s.fam)

Coefficients:
(Intercept)        s.age        s.fam
    132.368       -1.574       -4.366
```

ويمكن استخدام الصيغة `lm(s.grd1~s.age+s.fam)` أيضا للحصول على نفس النتيجة.

- **نموذج الانحدار الخطي بدون الجزء المقطوع β_0:**

```
> lm(s.grd1~-1+s.age+s.fam)

Call:
lm(formula = s.grd1 ~ -1 + s.age + s.fam)

Coefficients:
 s.age   s.fam
 4.508  -4.865
```

كما يمكن استخدام الصيغة؛ `lm(s.grd1~0+s.age+s.fam)` أو الصيغة؛

`lm(s.grd1~s.age+s.fam-1)` لنفس الغرض.

- **نموذج الانحدار غير الخطي في المتغير التابع:**

توجد صيغ كثيرة لكتابة نموذج الانحدار غير الخطي في المتغير التابع تختلف باختلاف العلاقة بين متغيرات الدراسة، لذا يمكن استخدام أي صيغة غير خطية للمتغير التابع مع المحافظة على شكل النموذج كما سبق، وسنقدم هنا مثالا يحتوي على صيغة لوغاريتم المتغير التابع؛

```
> lm(log10(s.grd1)~s.age+s.fam)

Call:
lm(formula = log10(s.grd1) ~ s.age + s.fam)

Coefficients:
(Intercept)        s.age        s.fam
    2.27651     -0.01162     -0.02966
```

- **نموذج الانحدار غير الخطي في المتغير التوضيحي:**

كما هو الحال مع المثال السابق، فإنه يمكن استخدام صيغ غير خطية مع متغير توضيحي أو أكثر في النموذج، ولنأخذ مثالا يحتوي على الجذر التربيعي لأحد المتغيرات التوضيحية؛

```
> lm(s.grd1~s.age+sqrt(s.fam))

Call:
lm(formula = s.grd1 ~ s.age + sqrt(s.fam))

Coefficients:
(Intercept)        s.age  sqrt(s.fam)
    157.892       -1.626      -21.318
```

- **نموذج الانحدار متعدد الحدود:**

وهو أحد التركيبات المعروفة لنماذج الانحدار غير الخطية، ويمكن تنفيذه لنموذج متعدد الحدود من الدرجة الثالثة مثلا بالصورة:

```
> lm(s.grd1~poly(s.age,3))

Call:
lm(formula = s.grd1 ~ poly(s.age, 3))

Coefficients:
 (Intercept) poly(s.age,3)1 poly(s.age,3)2 poly(s.age,3)3
      71.314        -37.846        -32.989          1.432
```

حيث تمثل poly(s.age,3)1 تقدير معلمة المتغير s.age ذو الأس واحد، وpoly(s.age,3)2 وpoly(s.age,3)3 تقدير معلمة نفس المتغير للأسس اثنان وثلاثة.

- **نماذج الانحدار المختلطة:**

يمكن بناء نموذج يحتوي على متغيرات توضيحية خطية مع متغيرات متعددة الحدود مثلا كالتالي؛

```
> lm(s.grd1~s.age+s.sem+poly(s.fam,2))

Call:
lm(formula = s.grd1 ~ s.age + s.sem + poly(s.fam, 2))

Coefficients:
    (Intercept)            s.age            s.sem
        110.959           -1.524           -1.168
poly(s.fam, 2)1  poly(s.fam, 2)2
        -74.979           -1.145
```

- **نموذج تحليل التباين:**

يمكن استخدام الدالة الخطية lm لتنفيذ تحليل التباين باتجاه واحد وباتجاهين كما تم استخدام الدالة anova سابقا، ولنأخذ بالاعتبار المتغيرات X.obs، X.col، وX.row التي تم تعريفها من الجدول (2.6) سابقا. ولتوفيق نموذج تحليل تباين باتجاه واحد وتقدير المعالم يمكن كتابة؛

```
> lm(X.obs~X.col)

Call:
lm(formula = X.obs ~ X.col)


Coefficients:
(Intercept)        X.colb        X.colc
         58             5             1
```

وللحصول على جدول تحليل التباين نكتب الأمر؛

```
> anova(lm(X.obs~X.col))

Analysis of Variance Table

Response: X.obs
          Df Sum Sq Mean Sq F value Pr(>F)
X.col      2     56  28.000  0.4158 0.6718
Residuals  9    606  67.333
```

ولتوفيق نموذج تحليل تباين باتجاهين وتقدير المعالم نكتب؛

```
> lm(X.obs~X.col+X.row)

Call:
lm(formula = X.obs ~ X.col + X.row)

Coefficients:
(Intercept)        X.colb        X.colc       X.rowL2
         68             5             1           -17

    X.rowL3       X.rowL4
         -9           -14
```

ويتم الحصول على جدول تحليل التباين باستخدام دالة anova كما هو الحال مع تحليل التباين في اتجاه واحد. ويمكن أيضا توفيق نماذج تصميم التجارب ذات الصيغ المختلفة بنفس الطريقة، أي أنه يتم تعريف الصيغة في R كما هي موجودة في النموذج النظري، فمثلا يمكن تنفيذ نموذج تصميم تجارب ذو أثر مشترك بالصورة؛

```
> lm(X.obs~X.col*X.row)

Call:
lm(formula = X.obs ~ X.col * X.row)
```

```
Coefficients:
   (Intercept)          X.colb          X.colc
            74              -2             -10

       X.rowL2         X.rowL3         X.rowL4
           -27             -16             -21

X.colb:X.rowL2  X.colc:X.rowL2  X.colb:X.rowL3
            12              18              10

X.colc:X.rowL3  X.colb:X.rowL4  X.colc:X.rowL4
            11               6              15


> anova(lm(X.obs~X.col*X.row))

Analysis of Variance Table

Response: X.obs
            Df Sum Sq Mean Sq F value Pr(>F)
X.col        2     56      28
X.row        3    498     166
X.col:X.row  6    108      18
Residuals    0      0

Warning message:
In anova.lm(lm(X.obs ~ X.col * X.row)) :
  ANOVA F-tests on an essentially perfect fit are unreliable
```

وللمزيد من المعلومات حول استخدام الدالة lm يمكن للقارئ الرجوع دائما للمساعدة باستخدام دالة المساعدة help(lm).

الفصل السابع

بعض الدوال المتقدمة في R

(Some Advanced Functions in R)

1.7 الدوال الشرطية (Conditional Functions)

2.7 كتابة دوال المستَخدِم (Writing User-defined Functions)

1.2.7 تعريف دالة المستخدم لمتغير واحد (User-defined Function for One Variable)

2.2.7 تعريف دالة المستخدم لأكثر من متغير

(User-defined Function for more than One Variable)

3.7 الحلقات والمحاكاة (Loops and Simulation)

1.3.7 دوال الحلقات `for` و`while` (`for` and `while` Loops)

2.3.7 المحاكاة (Simulation)

4.7 أسلوب إعادة المعاينة (البوتستراب) (Bootstrap Sampling)

5.7 بعض دوال R الإضافية (Some Additional Functions of R)

في هذا الفصل، سنتناول بعض دوال لغة R المتقدمة مثل الدوال الشرطية ودوال الحلقات وغيرها من الدوال المتعلقة ببعض الأساليب الإحصائية الهامة، (مثل أسلوب إعادة المعاينة المعروف باسم البوتستراب)، إضافة إلى أنه سيتم شرح كيفية تكوين الدوال الخاصة التي قد يرغب المُستخدِم في تكوينها بنفسه وذلك في حال عدم توفرها ضمن الحزم الافتراضية أو الإضافية أو لرغبة المُستخدِم في إجراء تعديلات خاصة على بعض الصيغ الرياضية أو الإحصائية، وكذلك سيتم عرض طرق استخدام المحاكاة وتوليد البيانات العشوائية وغير ذلك من الدوال الإضافية.

ولا ننس هنا التذكير بضرورة إنشاء ملف عمل جديد كالمعتاد، والذي سيكون باسم work7 وملف لحف سطور الأوامر باسم his7، لتنظيم متابعة تطبيق الدوال والأوامر.

1.7 الدوال الشرطية (Conditional Functions)

تُعد **التعبيرات الشرطية** (Conditional Expressions) **والدوال الشرطية** من الأدوات الهامة المساعدة في نظام R التي تُستخدَم بشكلها المباشر أو ضمن دوال المُستخدِم التي سنتطرق لها في البند القادم. وعموما فإن التعبير أو الدالة الشرطية هي تساؤل يحتمل إحدى النتيجتين؛ صحيح (True) أو خطأ (False)، ومن أبسط استخدامات التعبيرات الشرطية الأمثلة التالية:

```
> 7>5
[1] TRUE

> 3*7<3*5
[1] FALSE

> "abc"=="abc"
[1] TRUE
```

ويمكن أيضا استخدام التعبيرات الشرطية للمقارنة بين قيم متجهين، فمثلا يمكن مقارنة القيم في المتجهين A1 و A2، كالتالي:

```
> A1<-c(76, 60, 85, 58, 91, 75, 82, 64, 79, 88)
> A2<-c(81, 52, 87, 70, 86, 77, 90, 63, 85, 83)
> A1>A2
 [1] FALSE TRUE FALSE FALSE TRUE FALSE FALSE TRUE FALSE TRUE
```

أو مقارنة قيم محددة ضمن متجهين بالصورة التالية:

```
> A1[3]==A2[3]
[1] FALSE
```

أما الدالة الشرطية %in% ، فتستخدم لمعرفة ما إذا كانت قيمة أو مجموعة من القيم هي ضمن متجه أو مصفوفة من القيم، فمثلا يمكن التحقق من التالي:

```
> 5 %in% c(1,2,3,4,5,6,7)
[1] TRUE

> c(2,5) %in% c(1,2,3,4,5,6,7)
[1] TRUE TRUE

> c("b","c","d") %in% c("b","c","r","h","n")
[1]  TRUE  TRUE FALSE
```

إضافة لما سبق، يمكن استخدام التعبير "و" (ورمزه &) والتعبير "أو" (ورمزه |) لتكوين أوامر شرطية مركبة، فمثلا يمكن كتابة:

```
> 2*5<2*6 & "a"=="b"
[1] FALSE

> 2*5<2*6 | "a"=="b"
[1] TRUE
```

حيث يُلاحظ في الأمر الأول (مع &) ضرورة تحقق جزئَي الشرط معا للحصول على "صحيح" TRUE، أما في الأمر الثاني (مع |)، فيكفي تحقق أحد جزئَي الشرط للحصول على "صحيح". وبالطبع يمكن أن يُستخدم أكثر من شرطين جزئيين لتكوين الشرط العام، فمثلا:

```
> (2*5<2*6 | "a"=="b") & 9>7
[1] TRUE
```

ونذكّر أنه يمكن دائما استخدام أمر النفي (!) ضمن التعبيرات الشرطية عند الحاجة؛

```
> (2*5<2*6 | "a"=="b") & !9>7
[1] FALSE
```

- **دالة if الشرطية:**

يتم استخدام الدالة الشرطية if بالصورة التقليدية التالية:

```
If (الشرط) الأمر
```

بمعنى أنه عند تحقق "الشرط" المطلوب فإنه يتم تنفيذ "الأمر". ويمكن التوسع في استخدام هذه الدالة بحيث يتم، بعد تحقق الشرط، تنفيذ سلسلة من الأوامر كما توضح القاعدة التالية؛

```
If (الشرط) { الأمر1 ، الأمر2 , ... }
```

ولنأخذ المثال الآتي:

```
> if(A1[1]<A2[1])A2[1]/A1[1]

[1] 1.065789
```

أما إذا لم يتم تحقق الشرط فإن سطر الأوامر لن ينتج عنه أي نتيجة، فمثلا يمكنك تجربة تنفيذ `if(A1[1]==A2[1])A2[1]/A1[1]` وستلاحظ عدم ظهور نتيجة.

من جديد يمكن إدراج أكثر من أمر مع الدالة الشرطية كالتالي:

```
> a1<-5;a2<-7;a3<-10

> if(a1>1|a3<8){print(a3/a1);print(a2^2)}

[1] 2
[1] 49
```

ولاحظ أن الدالة `print` يتم استخدامها للحصول على ناتج تنفيذ كلا من الأمرين `a3/a1` و`a2^2` حيث أنه عند عدم استخدامها فإننا نحصل على نتيجة الأمر الأخير فقط.

مثال آخر على استخدام دالة `if` الشرطية هو بإدراج دالة `warning` **الرسالة التحذيرية** ، (والتي تكون مفيدة أحيانا في كتابة دوال المستخدم)، بالصورة التالية:

```
> a4<-(-12)

> if(a4<0)warning("This value is negative")

Warning message:
This value is negative
```

- **دالة `if/else` الشرطية:**

يمكن اعتبار الدالة الشرطية `if/else` امتدادا للدالة `if`، فحيث أن الأخيرة تقوم بتنفيذ أمر أو مجموعة من الأوامر عند تحقق الشرط، فإن الدالة الأولى تقوم بعمل ذلك أيضا إضافة إلى أنها تقوم بتنفيذ أمر "بديل" في حال عدم تحقق الشرط، وآلية التنفيذ هي على النحو التالي:

```
If (الشرط) الأمر الأول else الأمر الثاني
```

أو بشكل موسع؛

```
If (الشرط) { مجموعة الأوامر الأولى } else { مجموعة الأوامر الثانية }
```

أو حتى بالصورة التالية الأكثر تعقيدا:

```
If (الشرط الأول) { مجموعة الأوامر الأولى } else if (الشرط الثاني) {مجموعة الأوامر الثانية} else
{ مجموعة الأوامر الثالثة }
```

والأمثلة التالية توضح استخدام الحالات السابقة لهذه الدالة؛

```
> if (a1==a2)a1/a2 else a1*a2
[1] 35

> if (7 %in% A1)print("yes") else print("no")
[1] "no"

> if (a1==a2){print(a1/a2);print(a2/a1)} else
{print(a1*a2);print(a1+1)}

[1] 35
[1] 6

> if (a1==a2){print(a1/a2);print(a2/a1)} else if
(a1>=a3){print(a1*a3);print(a1+1)} else {a3*a4}

[1] -120
```

- **دالة switch:**

دالة switch هي في الأصل دالة متعددة الاستخدامات، ومن ضمن تلك الاستخدامات استعمالها كدالة شرطية مثل دالة if/else، إلا أنه من الناحية العملية عادة ما يتم إدراجها ضمن دوال أخرى وعلى رأسها دوال المستخدم التي سنناقشها في البند التالي. وتكون القاعدة العامة لسطر الأمر الشرطي لها بالصورة التالية:

switch(المُدخَل , الأمر1, الأمر2, الأمر3, ...)

و"المُدخَل" هنا هي القيمة، (رقمية كانت أو نصية)، والتي بناء عليها سيتم تنفيذ أحد الأوامر التي تم تعريفها ضمن الدالة، أي عن طريق اختيار ترتيب الأمر المطلوب تنفيذه. فمثلا الأمر التالي؛

```
> switch("c",a=10*13,b=5+12,c=7/2)
[1] 3.5
```

يعني اختيار تنفيذ العملية التي اسمها c وهي 7/2، أو يمكن استخدام الأرقام لاختيار ترتيب العملية المطلوبة، فمثلا إذا ما كان المُدخَل هو الترتيب الثاني:

```
> switch(2,10*13,5+12,7/2,4*2,log(10))
[1] 17
```

فهذا يعني اختيار تنفيذ الأمر الثاني وهي عملية الجمع 5+12، ولاختيار تنفيذ الأمر الخامس مثلا نكتب:

```
> switch(5,10*13,5+12,7/2,4*2,log(10))
[1] 2.302585
```

2.7 كتابة دوال المُستَخدِم (Writing User-defined Functions)

تناولنا في الفصول السابقة الكثير من دوال لغة R التي تم توظيفها لتنفيذ العمليات الرياضية وحساب المقاييس الإحصائية ورسم الأشكال البيانية المختلفة وغير ذلك من الأوامر. وإضافة إلى الدوال التي يوفرها نظام R من خلال الحزم الافتراضية أو الإضافية، فإنه يمكن للمستخدم كتابة أو تعريف الدوال الخاصة التي يرغب بتنفيذها من خلال استخدام دالة **تكوين الدوال** function، سواء كانت هذه الدوال متوفرة أصلا ضمن حزم R أو غير متوفرة. فمثلا يمكن حساب الوسط الحسابي لمجموعة من القيم عن طريق استخدام الدالة mean أو يمكن استخدام دالة تكوين الدوال لكتابة صيغة حساب الوسط الحسابي ضمنها وتخزينها باسم محدد بحيث يمكن استخدام هذه الدالة الجديدة لحساب الوسط الحسابي لأي مجموعة جديدة من البيانات كما سنرى.

1.2.7 تعريف دالة المستخدم لمتغير واحد (User-defined Function for One Variable)

يتم كتابة الصيغة العامة التقليدية لدالة المستخدم أو دالة تكوين الدوال كما هو موضح في السطور التالية:

```
اسم الدالة <- function (أسماء المتغيرات داخل الدالة) {
مجموعة الأوامر
Return( النتائج )
}
```

حيث يتم عادة كتابة أجزاء أو مكونات الدالة في أسطر متتالية لتسهيل تتبع مكوناتها بالنسبة للمستخدم، إلا أنه يمكن كتابة الشكل العام باختصار كالتالي:

```
اسم الدالة <- function (أسماء المتغيرات داخل الدالة) { صيغ الحساب ضمن الدالة }
```

فمثلا؛ لتعريف دالة خاصة لحساب الوسط الحسابي لمتغير واحد كما ذكرنا أعلاه يمكننا كتابة الأمر التالي:

```
> fun1<-function(x){sum(x)/length(x)}
> fun1

function(x){sum(x)/length(x)}
```

وهذا يعني أنه تم تعيين الاسم fun1 لهذه الدالة الخاصة، والمتغير x لإدخال القيم في الدالة، والصيغة sum(x)/length(x) التي تمثل مجموع القيم مقسوما على عددها وهو قانون حساب الوسط الحسابي للمتغير x.

ملاحظة:

يمكن تعريف الدالة السابقة باستخدام الصيغة العامة التقليدية المطولة كالتالي:

```
fun1<-function(x){
sum(x)/length(x)
}
```

وننوه هنا إلى أن استخدام المتغير x في كتابة الدالة fun1 لا يعني أنه قد تم اعتباره خاص بالدالة fun1 فقط أو أنه لا يمكن استخدامه من جديد ضمن نفس مسار العمل، بمعنى أن استدعاء الرمز x، (إذا لم يكن قد تم تعيينه مسبقا إلى قيمة أو قيم معينة)، سيُعطي النتيجة التالية:

```
> x
Error: object 'x' not found
```

ويمكن أيضا عند استخدام الدالة function استخدام أي اسم أو رمز[1] للمتغير الذي يمثل القيم المطلوب التعامل معها، فمثلا يمكننا حساب الوسط الحسابي باستخدام "دالتنا الخاصة" fun1 للقيم (1، 2، 3) بأكثر من طريقة، منها الطريقتين التاليتين:

```
> fun1(c(1,2,3))
[1] 2

> obs1<-c(1,2,3)
> fun1(obs1)
[1] 2
```

ولاحظ أنه لا توجد ضرورة لاستخدام الرمز x بحد ذاته لتنفيذ الدالة، وأن استخدام الدالة fun1 يتم عن طريق كتابتها متبوعة بالأقواس الاعتيادية " ( ) " كما هو الحال عند استخدام دوال حِزم R المعتادة.

ويمكن الاستغناء عن كتابة أسم أو أسماء للمتغيرات ضمن الدالة، (أي كتابة function())، عندما لا يكون هنالك استخدام لهذه المتغيرات ضمن أوامر الدالة كما سنوضح في مثال لاحق، كما أنه يمكن تعيين قيم ثابتة ضمن مجموعة المتغيرات التي يتم تعريفها.

الآن بعد كتابة أي دالة جديدة من قِبل المستخدم وحفظها في مسار العمل الحالي فإن هذه الدالة يمكن استخدامها في أي جلسة أخرى، إلا أننا قد نرغب أحيانا بإجراء إضافات أو تعديلات على تلك الدالة، وفي هذه الحالة يمكن عمل ذلك عن طريق استخدام أمر التعديل edit عوضا عن إعادة كتابة دالة المستخدم من جديد، فمثلا في الدالة fun1 يُلاحظ أن الناتج يظهر بصورة مباشرة دون تسمية، وإذا ما أردنا أن يتم طباعة اسم المقياس المحسوب، وهو الوسط الحسابي، يمكننا تنفيذ الأمر التالي:

```
> fun1<-edit(fun1)
```

فتظهر مكونات الدالة في نافذة أوامر فرعية تحمل اسم الدالة، عندها نقوم بإجراء التعديل المطلوب بالصورة التالية:

```
function(x){
f.1<-list()
f.1$mean=sum(x)/length(x)
unlist(f.1)
}
```

[1] أو يمكن حتى استخدام القيم بعينها كمتجه بدون اسم.

حيث تم استخدام دالتي list وunlist لتعريف اسم المتوسط وكتابة النتيجة على شكل قائمة. ولاحظ أنه تم كتابة كل أمر في سطر مستقل كأسلوب تنظيمي لتسهيل قراءة الأوامر كل على حده، وهو الأسلوب الذي سيتم اتباعه في كتابة دوال المستخدم من الآن فصاعدا. يتم بعد ذلك الضغط_على أيقونة الحفظ في شريط الأوامر العلوي في برنامج R لحفظ التعديلات التي تم إجراؤها ثم تُغلق النافذة الفرعية. الآن إذا ما تم استخدام الدالة fun1 من جديد مع نفس القيم نحصل على النتيجة:

```
> fun1(obs1)

mean
   2
```

ملاحظة:

إن استخدام القائمة f.1 في الدالة السابقة fun1 لا يعني أنه تم تعريفها كشيء أو كبيانات ضمن مسار العمل الحالي، بمعنى أن f.1 ليس مُعرفا أو موجودا ضمن قائمة الأشياء في مسار العمل، وبالتالي إذا ما تم استدعاؤه بعد استخدام الدالة fun1 فإننا سنحصل على:

```
> f.1
Error: object 'f.1' not found
```

وهذا ينطبق بالطبع على كل الرموز والأسماء التي يتم استخدامها ضمن دوال المُستخدِم، (والتي تُعبِر في الغالب على بيانات أو نتائج)، بمعنى أن هذه الأسماء تكون مُعرفة ضمن دالة المستخدم فقط. أما إذا ما أردنا استدعاء أو التعامل مع هذه النتائج فيتم استخدام المعامل "<<-" كما سنوضح لاحقا.

لنقم الآن بإجراء تعديل آخر على نفس الدالة fun1 والذي يتضمن حساب الوسيط والانحراف المعياري لقيم المتغير؛

```
> fun1<-edit(fun1)
```

ثم نكتب في النافذة الفرعية:

```
function(x){
f.1<-list()
f.1$mean=sum(x)/length(x)
f.1$median=median(x)
f.1$SD=sd(x)
unlist(f.1)
}
```

وهكذا نحصل على نتيجة تطبيق الدالة fun1 على المتغير obs1 كالتالي:

```
> fun1(obs1)

  mean median      SD
     2      2       1
```

لقد رأينا في كثير من دوال R الأساسية وجود خيارات إضافية ضمن الدالة والتي يمكن للمستخدم اختيار تفعيلها متى أراد ذلك، أو يمكنه الاختيار من ضمن عدة خيارات في نفس الدالة، ويمكن تعريف هذه الخيارات ضمن دوال المستخدم الخاصة أيضا، فمثلا ضمن الدالة fun1 يمكن إدراج خيار حساب القيمتين الصغرى والكبرى كالتالي:

```
> fun1<-edit(fun1)
```

ونكتب في النافذة الفرعية:

```
function(x,extra=FALSE){
f.1<-list()
f.1$mean=sum(x)/length(x)
f.1$median=median(x)
f.1$SD=sd(x)
if(extra){
f.1$minimum<-min(x)
f.1$maximum<-max(x)
}
unlist(f.1)
}
```

وقد تم إضافة الخيار المُعرّف من قِبَلنا extra إلى الدالة مساويا للقيمة FALSE كخيار افتراضي (أي أنه لن يتم تنفيذه إذا لم يُكتب)، وإضافة ما يقوم هذا الخيار بحسابه (وهما القيمتين الصغرى والكبرى)، وذلك باستخدام الدالة الشرطية if والأقواس "{ }" كما رأينا.

الآن إذا ما كتبنا:

```
> fun1(obs1,extra=TRUE)

   mean  median      SD minimum maximum
      2       2       1       1       3
```

فهذا يعني تفعيل الخيار extra وطلب الحصول على النتائج الإضافية، أما إذا ما تم كتابة الدالة بالصورة fun1(obs1) أو الصورة fun1(obs1,extra=F)، فهذا يعني عدم طلب الحصول على النتائج الإضافية minimum وmaximum.

لنقم مرة أخرى بعمل إضافة جديدة على نفس الدالة fun1، حيث سنقوم هذه المرة بإضافة خيار الحصول على تمثيل بياني للمتغير المُدخل، حيث سيتمكن المُستخدِم للدالة fun1 من الاختيار من بين ثلاثة أنواع من التمثيل البياني هي المدرج التكراري، شكل الصندوق، ورسم QQ الطبيعي.

بعد تنفيذ أمر التعديل fun1<-edit(fun1) ، نكتب ما يلي في النافذة الفرعية ثم نقوم بتخزين التعديل الجديد؛

```
function(x,extra=F,Graph=F){
f.1<-list()
f.1$mean=sum(x)/length(x)
f.1$median=median(x)
f.1$SD=sd(x)
if(extra){
f.1$minimum<-min(x)
f.1$maximum<-max(x)
}
if(Graph=="Hist"){
hist(x)
}
if(Graph=="Box"){
boxplot(x)
}
if(Graph=="QQ"){
qqnorm(x)
qqline(x)
}
unlist(f.1)
}
```

ولنقم الآن بتحليل مكونات الدالة fun1 بعد إدخال التعديلات الجديدة عليها؛

- في السطر رقم 1: تم إضافة الخيار Graph الذي سيكون خاص بتنفيذ التمثيل البياني المطلوب، وقد تم تعيين الوضع الافتراضي بعدم استدعاء هذا التمثيل، (Graph=F).
- في السطور من رقم 2 إلى 9: لا يوجد أي تغيير عن السابق.
- في السطور (10 إلى 12)، (13 إلى 15)، و(16 إلى 19): تم إضافة الدالة الشرطية if لكل مجموعة من هذه السطور بحيث يتم تنفيذ المدرج التكراري عند استخدام الخيار "Graph="Hist، وشكل الصندوق عند استخدام الخيار "Graph="Box، ورسم QQ الطبيعي عند استخدام الخيار "Graph="QQ.

أما السطرين الأخيرين فيبقيان بدون تغير. ولاحظ أنه في الدالة fun1 لا يمكن الحصول على الرسومات الثلاثة في آن واحد، بل يجب على المُستخدِم الاختيار من بينها أو تجاهل الرسم عن طريق عدم استخدام الخيار Graph كما هو الحال مع الخيار extra. ونذكّر القارئ بضرورة التقيد بنفس الأسماء التي تم تعيينها للتمثيل البياني في دالة المُستخدِم؛ "Hist"، "Box"، و"QQ" وإلا ستظهر رسالة خطأ في نتيجة تنفيذ الدالة.

ولنستخدم الدالة fun1 الآن مع عينة حجمها 500 مفردة مثلا يتم توليدها من التوزيع الطبيعي المعياري[1]، وسنقوم في الحالة الأولى بإهمال الخيار extra واختيار رسم المدرج التكراري، وفي الحالة الثانية سيتم تفعيل الخيار extra واختيار شكل الصندوق، أما في الحالة الثالثة فسيتم إهمال خياري الدالة معا، ولن يتم عرض الرسم البياني هنا، (والذي سيظهر للمستخدم في نافذة فرعية كما هو المعتاد)، لعدم وجود ضرورة:

.1

```
> fun1(rnorm(500),Graph="Hist")

      mean     median         SD
0.04606487 0.06338109 1.01442118
```

.2

```
> fun1(rnorm(500),extra=T,Graph="Box")

        mean       median           SD      minimum
-0.027639211 -0.002994224  0.973570255 -2.741752663

     maximum
 3.038749441
```

.3

```
> fun1(rnorm(500))

       mean      median          SD
0.006984471 0.022136858 1.016244001
```

تكلمنا في البند السابق عن استخدام الدالة switch كدالة شرطية، ونسوق هنا المثال التالي كتطبيق على استخدامها ضمن دالة المستخدم. هذا المثال يتضمن تكوين دالة يتم من خلالها توليد عينة عشوائية من التوزيع الطبيعي أو المنتظم أو الأسي، حيث يقوم المُستخدِم بإدخال حجم العينة المرغوب ونوع التوزيع الاحتمالي فتقوم دالة switch بتنفيذ التوليد المطلوب:

```
f.switch<-function(n,d){
gen<-switch(d,
"normal"=rnorm(n),
"uniform"=runif(n),
"exponential"=rexp(n))
return(gen)
}
```

[1] نذكّر هنا بأن النتائج في هذا المثال ستكون مختلفة عما سيحصل عليه القارئ، كما هي مختلفة في الحالات الثلاث، لأن العينة ناتجة عن توليد عشوائي للبيانات.

وبالتالي إذا رغبنا مثلا بتوليد عينة عشوائية حجمها 20 مفردة من التوزيع المنتظم فإننا إما أن نكتب ترتيب التوزيع كما هو موجود ضمن الدالة أو أن نكتب اسم التوزيع المطلوب، (كما هو مُعرّف داخل الدالة)، كما نرى:

```
> f.switch(20,2)

 [1] 0.51427339 0.60156787 0.42551768 0.13643359 0.26592344
 [6] 0.33040552 0.69046760 0.13374269 0.33194107 0.97284936
[11] 0.10748860 0.09489916 0.51695334 0.93254113 0.40330952
[16] 0.83665483 0.31314746 0.34482694 0.77176410 0.17978409


> f.switch(20,"uniform")

 [1] 0.12513485 0.40155423 0.89029834 0.24529109 0.92518024
 [6] 0.66397400 0.41991732 0.01691235 0.38017183 0.80451872
[11] 0.71282659 0.61903813 0.43112541 0.11176716 0.68383298
[16] 0.01689613 0.57060349 0.68538793 0.36526210 0.31052811
```

- **دالة المنوال:**

إن مقياس **المنوال** (Mode) يُعد من ضمن المقاييس الإحصائية التي لم يتم إفراد دالة خاصة لحسابه في لغة R حتى وقت إعداد الكتاب، ويمكن للقارئ أن يجد عدة دوال مكتوبة من قِبل المستخدمين لحساب المنوال، وإحدى أبسط هذه الدوال المتوفرة على الانترنت يمكن أن يتم إعادة كتابتها بالصورة التالية:

```
fun.mode<-function(x){
f.m<-list()
x.m<-table(as.vector(x))
f.m$Mode<-names(x.m)[x.m==max(x.m)]
unlist(f.m)
}
```

ولاحظ أنه في السطر الثالث من الدالة fun.mode، تم استخدام الدالة table لوضع قيم المتغير في جدول يحتوي فيه السطر الأول على مفردات المتغير والسطر الثاني على تكرارات هذه المفردات. ثم يتم باستخدام الدالة names استدعاء القيمة أو القيم التي لها أكبر تكرار، والتي تمثل المنوال. ويمكن للدالة fun.mode حساب المنوال للمتغيرات الكمية والنوعية أيضا، ولنأخذ الأمثلة التالية التي تكون أحادية المنوال وثنائية المنوال وعديدة المنوال:

```
> fun.mode(c(2,2,2,3,5,5,6,7,7,7,7,8,11,15,21))

Mode
 "7"

> fun.mode(c(2,2,2,3,5,5,6,7,7,7,8,11,15,21))

Mode1 Mode2
  "2"   "7"
```

```
> fun.mode(c("a","b","a","b","b"))

Mode
 "b"

>  fun.mode(c(1,1,2,2,3,3))

Mode1 Mode2 Mode3
  "1"   "2"   "3"
```

2.2.7 تعريف دالة المستخدم لأكثر من متغير

(User-defined Function for more than One Variable)

لننتقل الآن لتكوين دالة مُستخدِم تتضمن التعامل مع أكثر من متغير أو مجموعة من البيانات، ولنتناول المثال التالي الذي يتعامل مع متغيرين؛

لنفرض أننا نرغب بتكوين دالة، (وليكن اسمها `fun2`)، تقوم بحساب فترة الثقة للفرق بين متوسطي مجتمعين، فتكون إحدى الطرق[1] هي مجموعة الأوامر التالية، (والتي يمكن كتابتها في ملف نصي[2] ثم تنفيذها في لوحة مراقبة R، وليكن الاسم "myfun2" هو الذي تم اختياره لتخزين الملف النصي):

```
fun2<-function(x1,x2,alpha){
f.2<-list()
n1=length(x1); n2=length(x2)
x1bar=sum(x1)/n1; x2bar=sum(x2)/n2
s1sq=var(x1); s2sq=var(x2)
f.2$diff.mean=x1bar-x2bar
f.2$SE=sqrt((s1sq/n1)+(s2sq/n2))
a=1-(alpha/2)
t.df=n1+n2-2
f.2$LCL=f.2$diff.mean-(qt(a,t.df)*f.2$SE)
f.2$UCL=f.2$diff.mean+(qt(a,t.df)*f.2$SE)
unlist(f.2)
}
```

ولاحظ في السطر الأول للدالة `fun2` أنه قد تم تعيين المتغيرين (`x1` و`x2`) لإدخال قيم العينتين، والخيار `alpha` لإدخال القيمة المرغوبة لمستوى المعنوية. بعد ذلك تم تعريف مكونات قانون حساب فترة الثقة للفرق

---

[1] توجد دائما عدة طرق لكتابة دوال المستخدم للحصول على نفس النتيجة المطلوبة.

[2] يتم تنفيذ ذلك عادة بإنشاء ملف نصي جديد (New Script) من قائمة الملف (File) في نظام R وكتابة سطور الأوامر فيه ثم تخزينها باسم معين، بعد ذلك يتم تنفيذ هذه الأوامر عن طريق تظليلها بالفأرة والضغط بالزر الأيمن على موضع التظليل ثم اختيار الخيار الأول في القائمة الفرعية (Run line or selection) فيتم تنفيذ الأوامر مباشرة في لوحة مراقبة R.

بين متوسطين[1]، والدالة fun2 تتعامل مع العينات المستقلة الكبيرة والصغيرة على حد السواء، (باعتبار أننا استخدمنا الدالة qt لحساب القيم الجدولية للتوزيع)، ويقوم المستخدِم بإدخال قيمة مستوى المعنوية المطلوب، حيث يتم قسمته على 2 وطرحه من الواحد الصحيح وإدخال تلك القيمة في قانون الحساب العام ضمن صيغة الدالة fun2.

وكنتيجة ظاهرة، سيظهر لمستخدم الدالة fun2 أربعة قيم هي الفرق بين متوسطي العينتين (diff.mean)، الخطأ المعياري للتقدير (SE)، والحدين الأدنى والأعلى لفترة الثقة للفرق بين وسطي المجتمعين (LCL وUCL) على الترتيب. وكمثال تطبيقي على هذه الدالة، لنقم بتقدير 95% فترة ثقة للفرق بين متوسطي مجتمعين من خلال عينتين عشوائيتين، (ذات الأحجام 150 و120 على الترتيب)، يتم توليدهما من التوزيع الطبيعي كالتالي:

```
> fun2(rnorm(150),rnorm(120),0.05)

diff.mean          SE         LCL         UCL
 0.0485197  0.2430325 -0.1945128  0.2915522
```

الآن لنقم بإدخال إضافة، (باستخدام الأمر fun2<-edit(fun2)، على الدالة بحيث تقوم بإجراء اختبار لفرضية تساوي متوسطي مجتمعين، تحت نفس الفرضيات التي تم ذكرها أعلاه. وهذه الإضافة ستقوم بعرض القيمة الحسابية لإحصاءة الاختبار والقرار بقبول أو رفض الفرضية الصفرية القائلة بتساوي متوسطي المجتمعين ($H_0: \mu_1 = \mu_2$)؛

```
fun2<-function(x1,x2,alpha){
f.2<-list()
n1=length(x1); n2=length(x2)
x1bar=sum(x1)/n1; x2bar=sum(x2)/n2
s1sq=var(x1); s2sq=var(x2)
f.2$diff.mean=x1bar-x2bar
f.2$SE=sqrt((s1sq/n1)+(s2sq/n2))
a=1-(alpha/2)
t.df=n1+n2-2
f.2$LCL=f.2$diff.mean-(qt(a,t.df)*f.2$SE)
f.2$UCL=f.2$diff.mean+(qt(a,t.df)*f.2$SE)
f.2$t.cal=f.2$diff.mean/f.2$SE
if(f.2$t.cal>=pt(a,t.df))f.2$Dec="Reject H0" else
f.2$Dec="Accept H0"
unlist(f.2)
}
```

[1] بافتراض تساوي تباينات المجتمعين المجهولة.

وهذه الإضافة قد تمت في السطور من 12 إلى 14، حيث تم إدراج صيغة قانون حساب قيمة إحصاءة الاختبار ثم الدالة الشرطية if/else والتي ستقوم بالتحقق ما إذا كانت القيمة الحسابية للإحصاءة هي أكبر من أو تساوي القيمة الاحتمالية الجدولية أم لا، فإذا تحقق هذا المضمون فإنه يتم كتابة رفض الفرضية الصفرية ("Reject H0")، وأما عند حدوث العكس فيتم كتابة قبول الفرضية الصفرية ("Accept H0").

ولنأخذ البيانات التي تمثل الخمس وعشرون مفردة الأولى للمتغيرين[1] s.grd1 وs.grd2 (والتي تم استخدامها في الفصل السابق عند تناول الاستدلال حول مجتمعين)، كتطبيق للدالة fun2 حتى يتمكن القارئ من مقارنة النتيجتين؛

```
> library(rJava)
> library(XLConnectJars)
> library(XLConnect)

> stu.data1<-readWorksheetFromFile("studata1.xlsx",
sheet=1,rownames=1)

> s.grd1<-stu.data1$grd1;s.grd2<-stu.data1$grd2

> fun2(s.grd1[1:25],s.grd2[1:25],0.05)

diff.mean           SE                  LCL
"-1.95999999999999" "4.54630985892222" "-11.1009686212787"

UCL                t.cal                   Dec
"7.18096862127873" "-0.431118876808068"   "Accept H0"
```

ويمكن ببساطة ملاحظة تشابه النتيجتين، إلا أنه من الناحية الشكلية فإن خانات الأعداد الظاهرة في النتيجة هي كثيرة العدد ويُفضل تقريبها، لذلك سنقوم بإضافة دالة التدوير (round) للدالة fun2 للحصول على مظهر أفضل؛

```
fun2<-function(x1,x2,alpha){
f.2<-list()
n1=length(x1); n2=length(x2)
x1bar=sum(x1)/n1; x2bar=sum(x2)/n2
s1sq=var(x1); s2sq=var(x2)
f.2$diff.mean=round(x1bar-x2bar,digits=3)
f.2$SE=round(sqrt((s1sq/n1)+(s2sq/n2)),digits=3)
a=1-(alpha/2)
t.df=n1+n2-2
f.2$LCL=round(f.2$diff.mean-(qt(a,t.df)*f.2$SE),digits=3)
```

[1] هذين المتغيرين هما ضمن ملف البيانات stu.data1 والذي لابد من إعادة استدعاؤه في ملف العمل الحالي work7.

```
f.2$UCL=round(f.2$diff.mean+(qt(a,t.df)*f.2$SE),digits=3)
f.2$t.cal=round(f.2$diff.mean/f.2$SE,digits=3)
if(f.2$t.cal>=pt(a,t.df))f.2$Decision="Reject H0" else
f.2$Decision="Accept H0"
unlist(f.2)
}
```

وبإعادة تنفيذ الدالة نحصل على نفس النتائج السابقة مُقربة إلى ثلاثة خانات:

```
> fun2(s.grd1[1:25],s.grd2[1:25],0.05)

  diff.mean          SE         LCL         UCL
    "-1.96"     "4.546"     "-11.1"      "7.18"

      t.cal    Decision
   "-0.431" "Accept H0"
```

ملاحظة:

يمكن استخدام الدالة source في استدعاء وتنفيذ دوال المُستخدِم، (أو أي سطور أوامر)، التي تم تخزينها مسبقا في ملف نصي (Script) في نظام R، فمثلا، لنفرض أن دالة المستخدم السابقة fun2 تم كتابتها في الملف النصي "myfun2" الذي تم تخزينه مسبقا ولم يتم تنفيذها في لوحة مراقبة R، عندئذ يمكن تنفيذ محتويات الملف النصي عن طريق كتابة اسم الملف متبوعا بالامتداد ".R" كالتالي:

```
> source("myfun2.R")
```

عندها سيتم تنفيذ سطور الأوامر الموجودة في الملف النصي بالكامل، فإذا كانت دالة مستخدم فسيتم تعريف اسمها ضمن قائمة الدوال في مسار العمل الحالي، وإن كانت محتويات الملف مجرد مجموعة من الأوامر أو الدوال التنفيذية فسيتم ظهور نتائجها مباشرة.

3.7 الحلقات والمحاكاة (Loops and Simulation)

لاحظنا في بعض المواضع في الفصول السابقة احتياجنا إلى تنفيذ عملية حسابية أو أكثر عدة مرات متتالية، وهذا بطبيعة الحال يستدعي كتابة نفس العملية في كل مرة، وكلما ازداد عدد مرات التكرار كلما استغرق وقتا أطول في الكتابة، وهنا تبرز الحاجة لدوال الحلقات التي تجعل العملية تكرر نفسها بحسب عدد مرات التكرار المطلوب، ومن أشهر دوال الحلقات هما دالتي for وwhile. وعمليا فإن أغلب استخدامات دوال الحلقات الأساسية تكون عادة ضمن دوال المستخدِم وضمن دوال المحاكاة عند توليد البيانات بصورة عشوائية إلى جانب بعض الاستخدامات الأخرى.

1.3.7 دوال الحلقات **for** و**while** (for and while Loops)

- **دالة for:**

يتم استخدام دالة for بالشكل التالي:

```
for (i in قيمة البداية:قيمة النهاية ( الأمر
```

أو بصورة أكثر عمومية؛

```
for (i in القيم من متجه ( { مجموعة من الأوامر }
```

حيث قيمة البداية والنهاية هما القيمتان اللتان تحددان بداية الحلقة ونهايتها، ويمكن بالطبع استخدام أي رمز آخر بدل من الرمز i في الصيغة السابقة. وكأمثلة بسيطة على استخدام هذه الدالة بالصورة السابقة يمكننا كتابة:

```
> a1
[1] 5

> for (i in 1:5) print(a1*i)
[1] 5
[1] 10
[1] 15
[1] 20
[1] 25
```

بمعنى أنه تم تنفيذ عملية ضرب قيمة a1 بسلسلة القيم من 1 إلى 5 وكتابة النتيجة في كل مرة، ولاحظ أنه إذا تم كتابة نفس السطر السابق بدون استخدام print أو أي من دوال إظهار النتائج فسيتم تنفيذ العملية بدون أن تظهر أية نتيجة.

ويمكن تنفيذ العملية السابقة على متجه من القيم وليس قيمة واحدة فقط كالتالي:

```
> A1

 [1] 76 60 85 58 91 75 82 64 79 88

> for (i in 1:5)print(A1*i)

 [1] 76 60 85 58 91 75 82 64 79 88
 [1] 152 120 170 116 182 150 164 128 158 176
 [1] 228 180 255 174 273 225 246 192 237 264
 [1] 304 240 340 232 364 300 328 256 316 352
 [1] 380 300 425 290 455 375 410 320 395 440
```

أو "بتوسيع" الأمر الذي سيتم إعادة حسابه ضمن الحلقة، بالصورة التالية:

```
> for (i in c(-2,5,7,10))print(round(((A1-i)/i^2),digits=2))
```

```
 [1] 19.50 15.5 21.75 15.0 23.25 19.25 21.00 16.50 20.25 22.50
 [1] 2.84 2.20 3.20 2.12 3.44 2.80 3.08 2.36 2.96 3.32
 [1] 1.41 1.08 1.59 1.04 1.71 1.39 1.53 1.16 1.47 1.65
 [1] 0.66 0.50 0.75 0.48 0.81 0.65 0.72 0.54 0.69 0.78
```

وأيضا يمكن تنفيذ مجموعة من الأوامر ضمن دوال الحلقات كالتالي؛

```
> for (i in A2[1:6]){m1=(A1[1:6]-mean(A1[1:6]));m2=
(A2[1:6]-mean(A2[1:6]));m=(m1-m2)*i;print(round
(m,digits=1))}
```

```
[1] -297    756    -54   -864    513    -54
[1] -190.7  485.3  -34.7 -554.7  329.3  -34.7
[1] -319    812    -58   -928    551    -58
[1] -256.7  653.3  -46.7 -746.7  443.3  -46.7
[1] -315.3  802.7  -57.3 -917.3  544.7  -57.3
[1] -282.3  718.7  -51.3 -821.3  487.7  -51.3
```

يمكننا أيضا كمثال آخر على استخدام دالة for ضمن دوال المُستخدِم، تكوين دالة تقوم بحساب وكتابة ناتج جدول الضرب (حتى العدد 12 مثلا) لأي قيمة يُدخلها المستخدِم بالشكل التالي؛

```
> mult.tab<-function(y) {
for (n in 1:12) {
mult<-paste(y,"x",n,"=",y*n)
print(mult) }
}
```

الآن يمكن استخدام هذه الدالة للحصول على جدول الضرب لأي قيمة عددية، فمثلا؛

```
> mult.tab(7)
```

```
[1] "7 x 1 = 7"
[1] "7 x 2 = 14"
[1] "7 x 3 = 21"
[1] "7 x 4 = 28"
[1] "7 x 5 = 35"
[1] "7 x 6 = 42"
[1] "7 x 7 = 49"
[1] "7 x 8 = 56"
[1] "7 x 9 = 63"
[1] "7 x 10 = 70"
[1] "7 x 11 = 77"
[1] "7 x 12 = 84"
```

ومن الناحية الشكلية، يمكن إدخال تعديل على دالة المستخدم السابقة بحيث تظهر نتيجة جدول الضرب بدون رقم السطر [1]، وذلك باستخدام دالة[1] cat عوضا عن print بالصورة التالية:

```
> mult.tab<-function(y) {
for (n in 1:12) {
mult<-paste(y,"x",n,"=",y*n)
cat(mult,"\n") }
}
```

وقد تم إضافة الخيار "\n" للحصول على ناتج سطر الأمر paste كل مرة في سطر جديد، فمثلا يمكن الآن الحصول على الناتج بالصورة التالية:

```
> mult.tab(9)

9 x 1 = 9
9 x 2 = 18
9 x 3 = 27
9 x 4 = 36
9 x 5 = 45
9 x 6 = 54
9 x 7 = 63
9 x 8 = 72
9 x 9 = 81
9 x 10 = 90
9 x 11 = 99
9 x 12 = 108
```

- **دالة while:**

تُعد دالة while من الدوال المناسبة للاستخدام عندما نرغب في تكرار أمر أو مجموعة من الأوامر حتى يتم تحقق شرط محدد[2]، وفي هذه الحالات غالبا ما يكون عدد التكرارات التي ستنفذ حتى تحقق الشرط مجهول وليس محدد مسبقا كما هو الحال مع دالة for، وتُعتبر دالة While هي الأخرى من الأدوات المساعدة ضمن دوال المُستخدِم. والصيغة العامة لهذه الدالة هي بالشكل التالي:

While (الشرط) الأمر

أو بصورة أكثر عمومية؛

While (الشرط) { مجموعة من الأوامر }

---

[1] تُعرف دالة cat بأنها دالة **الربط والطباعة** (Concatenate and Print)، حيث أنها تقوم بربط المُخرجات المطلوب طباعتها أو إظهارها في لوحة مراقبة R بشكل متجاور لبعضها البعض، فعلى سبيل المثال يمكن كتابة الأمر cat(2*6,5+4) ومراقبة النتيجة.

[2] في حال عدم تحقق هذا الشرط، فإن سطر الأمر سيقوم بتكرار نفسه بدون توقف. وفي هذه الحالة سيضطر المستخدم لإيقاف التنفيذ يدويا، (يمكن عمل ذلك بالضغط على ESC في لوحة المفاتيح أو أيقونة Stop في لوحة مراقبة R).

ومن الأمثلة على استخدام هذه الدالة؛

```
> b1<-2

> while(b1<5) {print(b1);b1=b1+1}

[1] 2
[1] 3
[1] 4
```

وكذلك يمكن كتابة دالة مُستخدِم تقوم مثلا بتوليد أرقام عشوائية من 1 إلى 10 وإدراج الدالة while بحيث تقوم بإيقاف توليد الأرقام عند ظهور أحد العددين 1 أو 10:

```
> choose1<-function() {
s=0
while(s!=1 & s!=10) {
s=sample(1:10,1)
print(s) }
}
```

ولاحظ في الدالة السابقة أنه لم يتم تعيين اسم للمتغير ضمن الدالة function نظرا لعدم الحاجة للتعويض بقيمة له، حيث أنه سيتم توليد الأرقام العشوائية من 1 إلى 10 تلقائيا والتوقف عند ظهور أحد الرقمين 1 أو 10، ولذلك سيكون ناتج تنفيذ هذه الدالة مختلفا في كل مرة كما يُلاحظ من النتائج العشوائية التالية؛

```
> choose1()

[1] 7
[1] 10

> choose1()

[1] 8
[1] 4
[1] 5
[1] 1

> choose1()

[1] 9
[1] 6
[1] 1
```

2.3.7 المحاكاة (Simulation)

إن دوال الحلقات يمكن أن تكون مفيدة أيضا ضمن مفهوم الإحصاء الاستدلالي، وخاصة ضمن ما يُعرف بعملية **المحاكاة**[1] والتي تتضمن توليد عينات عشوائية من التوزيعات الاحتمالية المنفصلة أو المتصلة بأحجام ومعالم محددة مسبقا، بمعنى أننا "نحاكي" هذا التوزيع الاحتمالي أو نستخدم بيانات نكون متأكدين من تبعيتها للتوزيع المطلوب، بهدف استخدامها في الدراسات المختلفة التي تشمل توفيق النماذج الإحصائية أو إثبات النظريات أو غير ذلك من الاستخدامات التطبيقية.

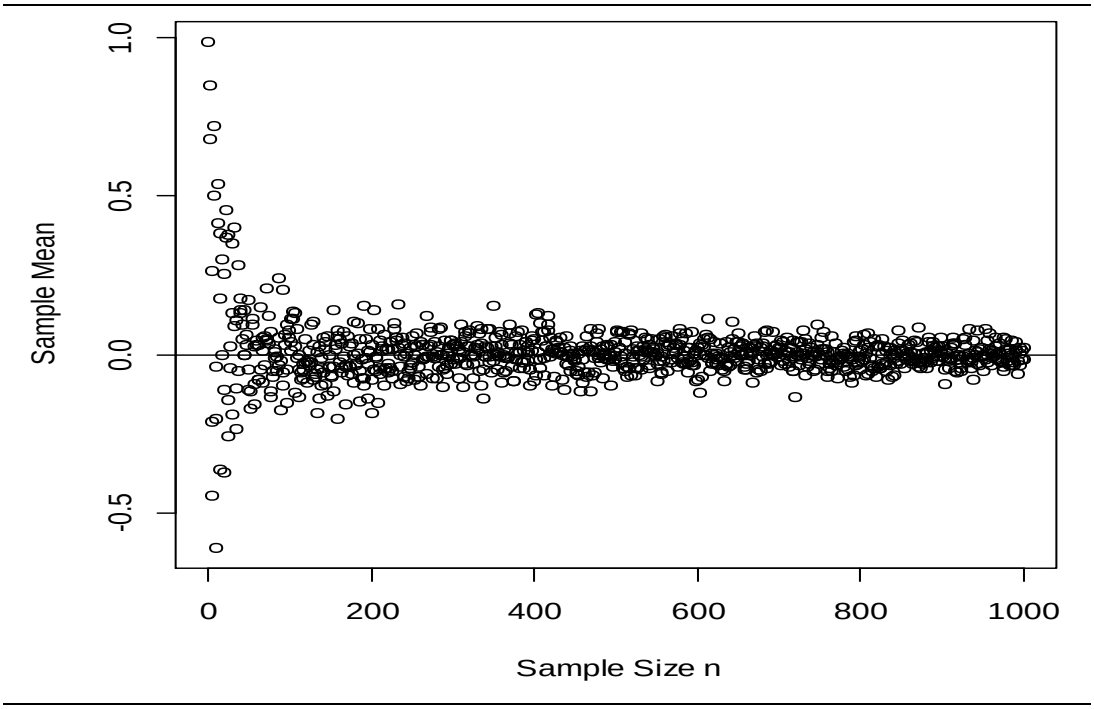
فعلى سبيل المثال، يمكننا استخدام دالة for ضمن دالة مُستخدِم تقوم بتطبيق عملية محاكاة عن طريق توليد عينات عشوائية ذات أحجام تصاعدية، (من حجم العينة الذي يساوي 1 حتى الحجم الذي يختاره المستخدم)، من التوزيع الطبيعي المعياري وحساب الوسط الحسابي لكل عينة تم توليدها ثم تمثيل قيم هذه المتوسطات ضد أحجام العينات بيانيا، وذلك بغرض توضيح كيفية تقارب قيمة متوسط العينة من متوسط المجتمع عند زيادة حجم العينة؛

```
> Xnorm<-function(n) {
Xn=numeric(n)
for (i in 1:n) {
Xn[i]=mean(rnorm(i,0,1))
plot(Xn,xlab="Sample Size n",ylab="Sample Mean")
abline(0,0) }
}
```

الآن يمكن اختيار عدد العينات المرغوب لتنفيذ الرسم البياني، وليكن 1000 عينة مثلا، عندها نحصل على انتشار للنقاط كما هو موضح في الشكل (1.7)، والذي نلاحظ منه اقتراب متوسط العينة من الصفر، (والذي هو متوسط مجتمع التوزيع الطبيعي المعياري)، كلما ازداد حجم العينة، ولاحظ أن الرسم قد يستغرق وقتا أطول من المعتاد ليكتمل نظرا لارتفاع عدد مرات تكرار رسم انتشار النقاط على الشكل:

```
> Xnorm(1000)
```

[1] تعرضنا في الفصل الخامس لمفهوم المحاكاة عندما تناولنا توليد العينات العشوائية من دوال التوزيعات الاحتمالية المنفصلة والمتصلة باستخدام الحرف r قبل اسم الدالة الاحتمالية.



شكل 1.7: شكل انتشار قيم متوسطات العينات العشوائية ضد أحجام العينات

لنفرض الآن أننا نود الحصول على قيم متوسطات العينات وحفظها لغرض استخدامها لاحقا في دراسات أخرى، عندها يمكن، (كما أشرنا في ملاحظة سابقة)، استخدام المعامل "<<-" لعمل ذلك، وهكذا يمكن تعديل الدالة Xnorm بحيث يتم استخدام التسمية Xmeans لحفظ قيم متوسطات العينات كالتالي:

```
> Xnorm<-function(n) {
Xn=numeric(n)
for (i in 1:n) {
Xn[i]=mean(rnorm(i,0,1))
plot(Xn,xlab="Sample Size n",ylab="Sample Mean")
abline(0,0)
}
Xmeans<<-Xn
return(Xmeans)
}
```

وهكذا فإنه باستخدام عينة صغيرة كمثال يمكن الحصول على قيم هذه المتوسطات، (ولن نقوم بعرض الرسم البياني هنا لعدم وجود ضرورة له)، بالصورة التالية:

```
> Xnorm(10)

 [1]  0.32449098  0.82495651  0.07446571  0.07110530
 [5] -0.27660252  0.16006970 -0.16074468 -0.04366095
 [9] -0.40114177 -0.12020905
```

كما أنه يمكننا استدعاء هذه المتوسطات واستخدامها متى أردنا عن طريق كتابة اسم المتجه الذي يمثلها وهو Xmeans.

لنتناول تطبيق آخر على المحاكاة يتضمن توليد عينة عشوائية من توزيع ذي الحدين ومراقبة اقتراب توزيعها من التوزيع الطبيعي عندما تكون قيمة كلا من $n.p$ أو $n.(1-p)$ أكبر من 5 حسب النظرية. في هذا التطبيق سنقوم بتعريف دالة مُستخدم تقوم بتنفيذ التالي:

- توليد عينة عشوائية من توزيع ذي الحدين بحسب تعريف المُستخدِم للمعالم x, n, p.
- تعيين الاسم Xbinom لتلك العينة لاستدعائها عند الحاجة.
- تمثيل تلك العينة العشوائية بيانيا باستخدام المدرج التكراري مصحوبا بمنحنى التوزيع الطبيعي الموفَق، وكذلك تنفيذ رسم QQ الطبيعي للعينة.
- عرض رسالة تحذيرية عندما تكون قيمة كلا من $n.p$ أو $n.(1-p)$ أقل من أو تساوي 5.

وهذه الدالة تم كتابتها كالتالي:

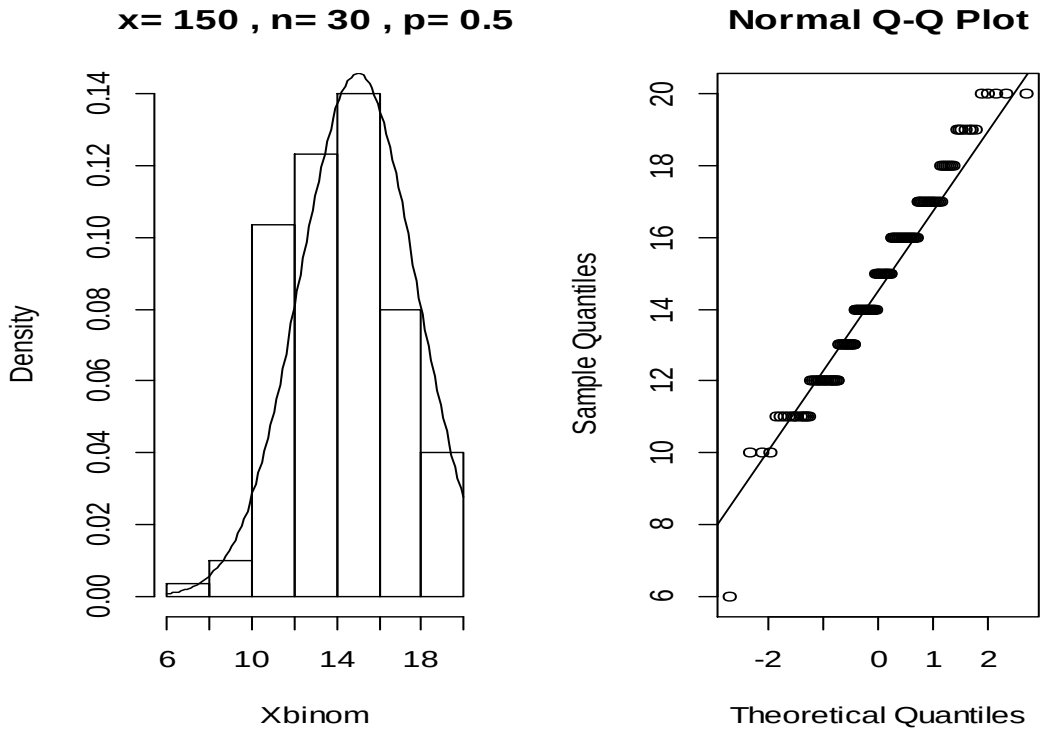
```
bi.to.norm<-function(x,n,p) {
simu=numeric(x)
simu=rbinom(x,n,p)
Xbinom<<-simu
par(mfrow=c(1,2))
hist(Xbinom,prob=TRUE,main=paste("x=",x,",","n=",n,",","p
=",p))
curve(dnorm(x,n*p,sqrt(n*p*(1-p))),add=TRUE)
qqnorm(Xbinom)
qqline(Xbinom)
if(n*p<=5|n*(1-p)<=5)warning("np or n(1-p)is less than or
equal to 5")
}
```

وعند تنفيذ[1] هذه الدالة عند x=150, n=30, p=0.5 مثلا؛

```
bi.to.norm(150,30,0.5)
```

نحصل على الشكل (2.7) والذي يمكن أن نلاحظ من خلاله اقتراب توزيع هذه العينة العشوائية (التي تم توليدها من توزيع ذي الحدين) من التوزيع الطبيعي عند زيادة قيمة حجم العينة n، مع ملاحظة ظهور النقاط بشكل متقطع في رسم QQ الطبيعي نظرا لطبيعة التوزيع المنفصل لذي الحدين.

[1] اختيار عدد محاولات كبير x=150 والذي يمثل في نفس الوقت حجم العينة التي سيتم توليدها، هو بغرض توضيح نمط توزيع قيم العينة بصورة أفضل.



شكل 2.7: المدرج التكراري (إلى اليسار) ورسم QQ الطبيعي (إلى اليمين) لعينة عشوائية مسحوبة من توزيع ذي الحدين

ويمكن للقارئ ملاحظة ظهور الرسالة التحذيرية عند اختيار قيم للمعالم n وp بحيث يكون حاصل ضربهما أقل من أو يساوي 5، (يمكنك تنفيذ الأمر[1] `bi.to.norm(2,2,0.5)` مثلا)، كما أنه سيلاحظ عدم اقتراب توزيع العينة الناتجة عن هذه المعالم من التوزيع الطبيعي.

ولاحظ أنه يمكن استخدام دوال الإحصاء الاستكشافي والنماذج والأساليب الإحصائية المختلفة في لغة R مع العينات التي تم توليدها كما هو الحال مع أي بيانات أخرى، فمثلا يمكن الحصول على المقاييس الملخصة للبيانات التي تم توليدها عن طريق دالة المستخدم السابقة `bi.to.norm` وهي `Xbinom` كالتالي:

```
> summary(Xbinom)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    3.0     4.0     5.0     4.8     6.0     6.0
```

[1] انتبه إلى أن تنفيذك لهذا الأمر باستخدام هذه المعالم سيُعطي تعريفا مختلفا للدالة `bi.to.norm` بحيث يلغي قيم المثال السابق `bi.to.norm(150,30,0.5)`، ولهذا يجب إعادة تنفيذ الأمر الأخير بالمعالم 150 و30 لمواكبة المثال، رغم أن النتائج العشوائية ستختلف بالتأكيد بعد كل إعادة تنفيذ للدالة.

من جديد، وكتطبيق على نظرية النهاية المركزية، يمكننا مراقبة اقتراب توزيع المتغير العشوائي $\bar{X}$ من التوزيع الطبيعي عندما يتبع المتغير العشوائي X توزيع احتمالي مختلف. فمثلا، دالة المُستخدم التي سنعرّفها تاليا ستقوم بما يلي؛

- توليد عينات عشوائية من التوزيع المنتظم بمعالم (min و max)، وأربعة أحجام مختلفة للعينات ($n1, n2, n3, n4$)، وأعداد عينات (حجمها m) كلها يتم تحديدها من قِبل المُستخدم.
- حساب المتوسطات $\bar{X}$ لهذه العينات عند كل حجم n_i ، حيث $i = 1, \ldots, 4$.
- عرض التمثيل البياني لدالة الكثافة الاحتمالية للمتغير $\bar{X}$ عند كل حجم عينة.

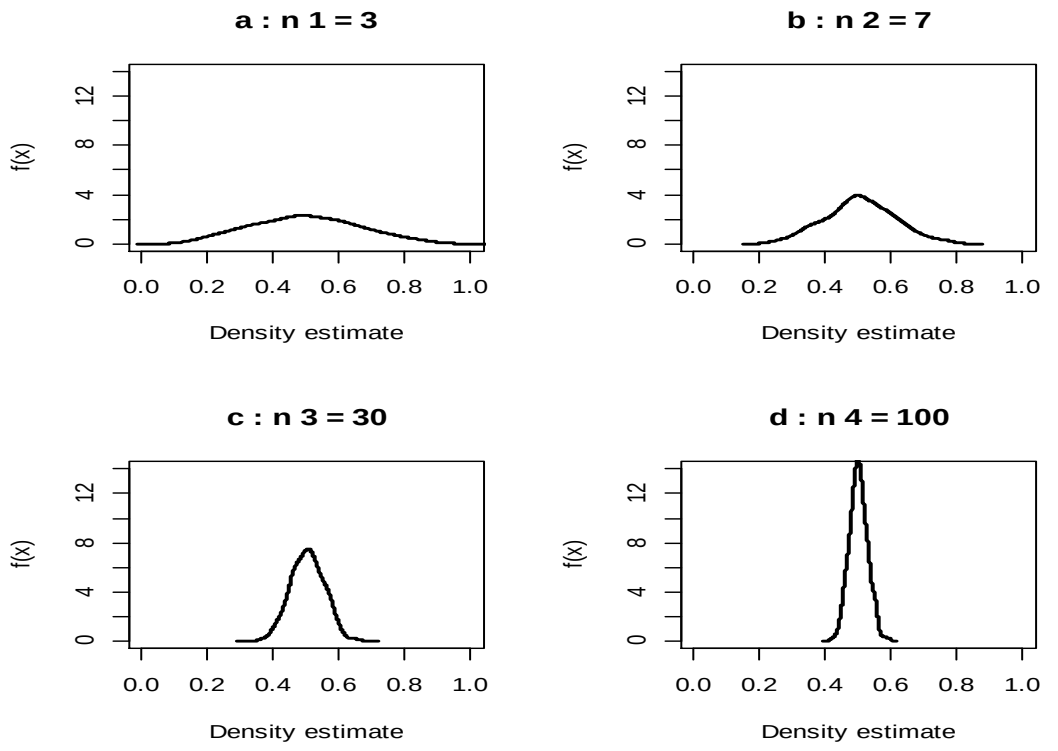
هذه الدالة تمت كتابتها بالصورة التالية:

```
> genf.unif<-function(m,min,max,n=c(n1,n2,n3,n4)) {
simu.n=c()
graph=c()
par(mfrow=c(2,2))
for (i in 1:4) {
graph[i]=plot(0,0,type="n",xlim=c(0,1),ylim=c(0,14),main=
paste(letters[i],":","n",i,"=",n[i]),xlab="Density
estimate",ylab="f(x)")
paste(graph[i])
for (j in 1:m) simu.n[j]=mean(runif(n[i],min,max))
lines(density(simu.n),lwd=2) }
}
```

وننوه إلى أن الخيارات الموجودة ضمن دالة `plot` ودالة `density` هي لإظهار الرسومات بشكل أكثر ملائمة. وإذا ما قمنا بتطبيق الدالة عند القيم التالية:

```
> genf.unif(1000,0,1,c(3,7,30,100))
```

فإننا نحصل على الشكل (3.7) التالي؛



شكل 3.7: التمثيل البياني لدوال الكثافة الاحتمالية للمتغير $\bar{X}$ عند أحجام العينات الأربعة المختلفة

ويُلاحظ من الشكل اقتراب توزيع المتغير $\bar{X}$ من التوزيع الطبيعي بزيادة حجم العينة. ويمكن للقارئ استخدام مقياس الوسيط `median` مثلا بدلا من دالة الوسط الحسابي في الدالة `genf.unif` ومقارنة النتائج.

4.7 أسلوب إعادة المعاينة (البوتستراب) (Bootstrap Sampling)

لنفرض أن لدينا مجتمعا حجمه N ونود تقدير وسطه الحسابي μ، عندها سنقوم أولا بسحب عينة عشوائية حجمها n من هذا المجتمع، ثم نحسب الوسط الحسابي لهذه العينة $\bar{X}$ والذي يُعد تقديرا لمتوسط المجتمع. والسؤال الذي يطرح نفسه دائما في نظرية التقدير؛ كيف لنا أن نعرف مدى معنوية أو صحة تقديرنا لمتوسط المجتمع؟، في الحقيقة سيبقى هذا التساؤل قائما طالما أن معلمة المجتمع مجهولة ولن نتمكن من مقارنتها بقيمة التقدير. إلا أنه يتم عادة استخدام التوزيع العيني للإحصاءة (وهي $\bar{X}$ في مثالنا) لعمل استدلالات، (مثل استخدام فترات الثقة)، حول قيمة معلمة المجتمع.

إلا أنه من الناحية العملية، فإن التوزيع العيني للإحصاءة يكون في الغالب مجهولا أيضا، وهذا ما أدى بدوره إلى وضع فرضيات حول توزيع المجتمع والتي تنعكس على توزيع العينة، فمثلا يتم افتراض أن المجتمع يتوزع بتوزيع طبيعي $N(\mu, \sigma^2)$ وبناء على نظرية النهاية المركزية يقال أن الإحصاءة $\bar{X}$ تتبع التوزيع الطبيعي تقريبا بمتوسط μ وتباين σ^2/n.

وكتوجه آخر، غير وضع الفرضيات حول معالم المجتمع، يمكن تقدير توزيع المجتمع نفسه، وهذا يتضمن الخوض في نظريات معقدة لا مجال للخوض فيها ضمن مواضيع هذا الكتاب، إلا أنه بعد تقدير توزيع المجتمع يتم سحب عينة عشوائية منه حجمها n، ويتم بعد ذلك سحب عينات عشوائية بالإرجاع من تلك العينة الأصلية، وهذا ما يُعرف **بإعادة المعاينة** (Resampling)، ويُعد أسلوب **البوتستراب** (Bootstrap Sampling) أحد أشهر طرق إعادة المعاينة المعروفة، وسنقوم هنا بشرح طريقة تنفيذه من خلال الخطوات التالية بإيجاز؛

- يتم سحب عينة عشوائية بسيطة حجمها n من المجتمع، (ولتكن $(x_1, x_2, \dots, x_n$)، ويتم حساب الإحصاءة لها، (فإذا كانت الإحصاءة هي الوسط الحسابي مثلا فهذا يعني ببساطة حساب $\bar{X}$).
- يتم سحب عينات عشوائية بالإرجاع من العينة الأصلية، (ولتكن $(n_1, \dots, n_m$)، ويكون عددها كبير جدا عادة وتسمى العينات المُعاد سحبها (Resamples).
- يتم بعد ذلك حساب قيمة الإحصاءة لكل عينة من العينات المُعاد سحبها، (أي أنه سيكون لدينا في مثال الوسط الحسابي؛ $(\bar{X}_1, \dots, \bar{X}_m$)، ويسمى توزيع الإحصاءة المحسوبة في هذه الحالة **بتوزيع البوتستراب** (Bootstrap Distribution).
- يتم استخدام توزيع البوتستراب للحصول على معلومات[1] حول توزيع المعاينة للإحصاءة الخاصة بالعينة الأصلية، (بمعنى أن يتم استخدام توزيع $\bar{X}_1, \dots, \bar{X}_m$ عوضا عن توزيع $\bar{X}$ في حالة الوسط الحسابي).

والآن ننتقل لعرض مثال حول كيفية استخدام دوال R لتطبيق أسلوب البوتستراب عمليا؛

لنقم أولا بسحب (توليد) عينة عشوائية بسيطة حجمها 100 من مجتمع له توزيع طبيعي بمتوسط هو $\mu = 4$ وانحراف معياري $\sigma = 1$، (علما بأن القيم الناتجة ستكون عشوائية في كل مرة يتم فيها تنفيذ هذا المثال، أي أنك لن تحصل على نفس القيم في النتائج المعروضة أدناه)؛

```
> samp1<-rnorm(100,mean=4,sd=1)
```

وهذا يعني أنه لدينا العينة العشوائية $samp1 = (x_1, x_2, \dots, x_{100})$. والوسط الحسابي لهذه العينة هو؛

```
> mean(samp1)
[1] 4.019315
```

أي أن $\bar{X} = 4.019315$، وهو تقدير قريب من الوسط الحقيقي للمجتمع الذي يساوي 4، ولاحظ أننا تمكنا هنا من مقارنة قيمة الإحصاءة المقدرة $\bar{X}$ بقيمة المعلمة الحقيقية μ لأن توزيع المجتمع وقيم معالمه μ و σ^2 معلومة في هذا المثال، أما في التجارب والدراسات الفعلية فإن ذلك لا يتحقق عادة.

[1] من أهم تلك المعلومات توزيع المشاهدات الذي يُعبر عن مركزية وانتشار البيانات والقيم المتطرفة وغير ذلك.

من جديد، نقوم بسحب عينات عشوائية من العينة الأصلية samp1، وليكن عددها $m = 2000$ عينة. ويمكن استخدام دالة الحلقة for لعمل ذلك إلا أننا سنستخدم دالة حلقة أخرى هي replicate لأنها أكثر ملائمة هنا؛

```
> resamp1<-replicate(2000,sample(samp1,100, replace=TRUE),
simplify=FALSE)
```

والأمر السابق يتضمن "تكرار" بمعنى توليد 2000 عينة (مُعاد سحبها) من العينة الأصلية samp1 حجم كل منها هو 100 مفردة بالإرجاع، (وإن أردت معاينة جزء من هذه العينات يمكنك كتابة الأمر head(resamp1) مثلا فتحصل على قيم المائة مفردة للعينات الستة الأولى)، والخيار simplify=FALSE يتم استخدامه لعدم تبسيط النتيجة في صورة أخرى مثل متجه أو مصفوفة أو غيرها، حيث أن simplify=TRUE هو الخيار الافتراضي في هذه الدالة، (راجع help(replicate)).

الآن لدينا $n_1 = n_1 = \cdots = n_{2000} = 100$ عينة مُعاد سحبها، وسنقوم بحساب الوسط الحسابي لكل عينة من هذه العينات، ويمكن استخدام الدالة sapply لتنفيذ ذلك بالصورة التالية:

```
> mean.resamp1<-sapply(resamp1,mean)
```

وبالتالي أصبح لدينا الأوساط الحسابية $\bar{X}_1, \bar{X}_2 \ldots, \bar{X}_{2000}$، لها توزيع البوتستراب، (ويمكن معاينة أي جزء من هذه الأوساط باستخدام الاستدعاء mean.resamp1[ ] مثلا)، وهذه الأوساط أو البيانات لها الوسط الحسابي:

```
> mean(mean.resamp1)
[1] 4.017494
```

والذي يُلاحَظ أنه قريب جدا من الوسط الحسابي للعينة الأصلية (4.019315)، أي أن قيمة التحيز (ابتعاد قيمة المتوسط المُقدَّر بأسلوب البوتستراب عن الوسط المُقدَّر بالطريقة التقليدية) هي:

```
> mean(samp1)-mean(mean.resamp1)
[1] 0.001821877
```

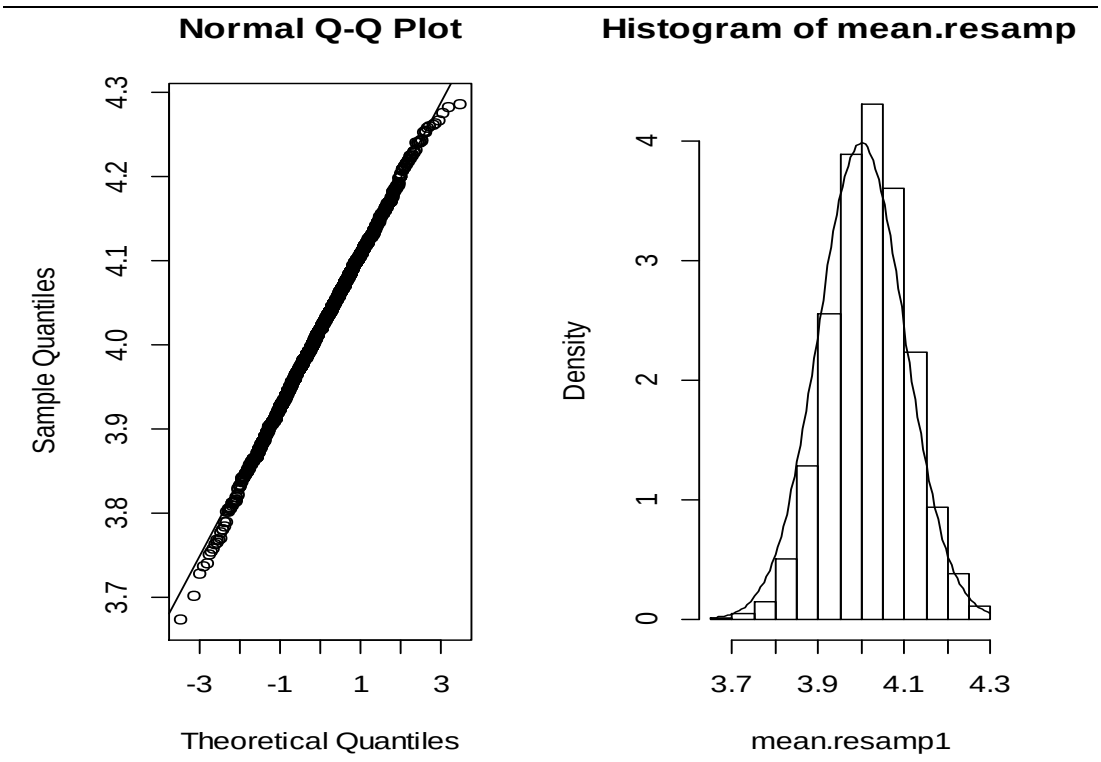
وهي قيمة يمكن اعتبارها منخفضة. ولاستكشاف طبيعة توزيع البوتستراب واختباره يمكننا استخدام أي من الاختبارات والرسومات البيانية المعتادة، فمثلا يمكن استخدام اختبار شابيرو–ويلك كالتالي:

```
> shapiro.test(mean.resamp1)

        Shapiro-Wilk normality test

data:  mean.resamp1
W = 0.999, p-value = 0.3201
```

وهذا يعني قبول الفرضية (البيانات تتبع التوزيع طبيعي :H_0)، أي أن توزيع الأوساط $\bar{X}_1, \bar{X}_2 \ldots, \bar{X}_{2000}$ يتبع التوزيع الطبيعي، ويمكن ملاحظة ذلك بيانيا باستخدام رسم QQ الطبيعي والمدرج التكراري لتوزيع البوتستراب، وسنقوم بتنفيذ كلى الرسمين مع إضافة منحنى توزيع طبيعي (بمتوسط يساوي 4 وحدات وانحراف معياري يساوي $\frac{\sigma}{\sqrt{n}} = 0.1$) إلى شكل المدرج التكراري بغرض مقارنته مع توزيع العينة الأصلية:

```
> par(mfrow=c(1,2))
> qqnorm(mean.resamp1)
> qqline(mean.resamp1)
> hist(mean.resamp1,prob=TRUE)
> curve(dnorm(x,4,(1/sqrt(100))),add=TRUE)
```

(ولاحظ أنه في السطر الأخير تم استخدام $\mu = 4$ و `1/sqrt(100)` $= \sigma/\sqrt{n}$). وبتنفيذ الأوامر السابقة نحصل على الشكل (4.7) الذي نلاحظ من خلاله اقتراب توزيع البيانات $\bar{X}_1, \bar{X}_2 \ldots, \bar{X}_{2000}$ من التوزيع الطبيعي إلى حد كبير، وكذلك لاحظ أن الخيار `prob=TRUE` ضمن الدالة `hist` هو لجعل شكل المنحنى الطبيعي ملائما للمدرج التكراري؛



شكل 4.7: رسم QQ الطبيعي (إلى اليسار) والمدرج التكراري (إلى اليمين) مضافا إليه منحنى التوزيع الطبيعي بمتوسط 4 وانحراف معياري 1 للبيانات `mean.resamp1`

ونشير من جديد هنا إلى أن النتائج السابقة ستكون كلها مختلفة عند إعادة تنفيذها من جديد باستخدام نفس حجم العينة (100) ونفس عدد مرات إعادة المعاينة (2000) لنفس التوزيع الطبيعي $N(4, 1)$، حيث أنه لدينا مصدرين للعشوائية؛ الأول ناتج عن توليد العينة الأصلية والثاني ناتج عن إعادة المعاينة (البوتستراب)، وفي

بعض المحاولات قد نجد اختلافا ملحوظا عن النتيجة السابقة، ويمكن للقارئ إجراء عدة محاولات لتنفيذ الأوامر السابقة لملاحظة ذلك.

ولتسهيل إعادة تنفيذ المثال السابق، سنضع سطور الأوامر السابقة ضمن دالة المستخدم التالية:

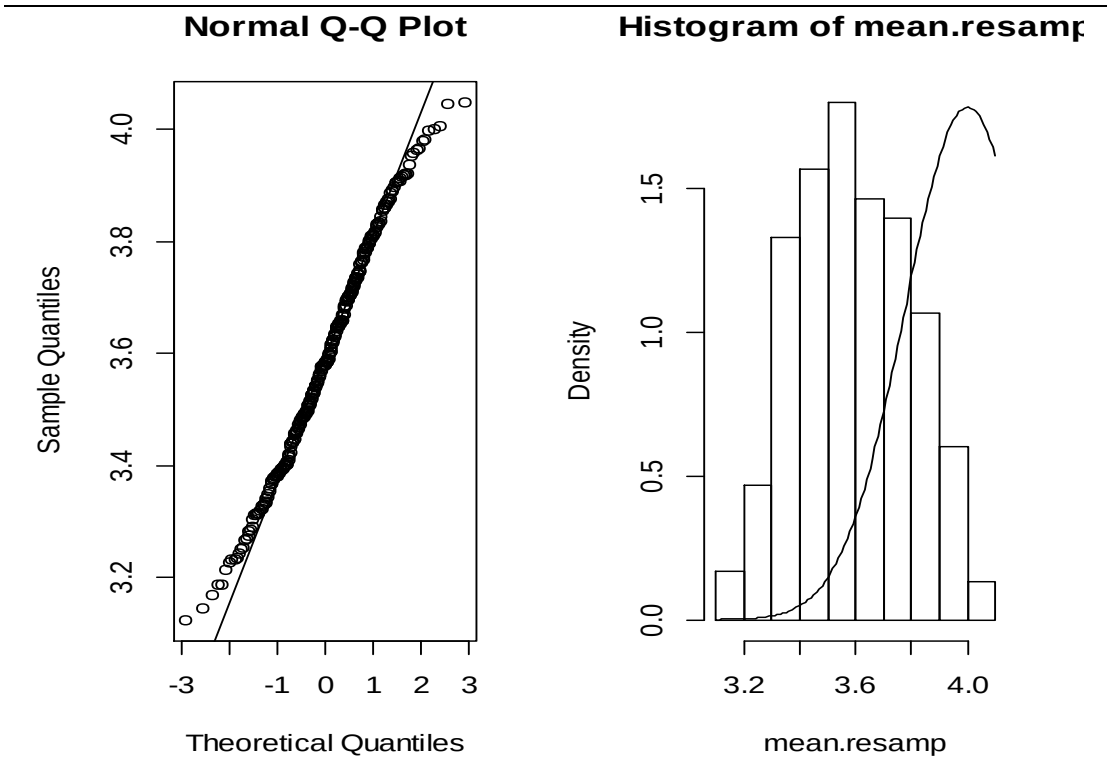
```
> bootstrap.fun1<-function(n,mean.norm,sd.norm,m) {
samp<-rnorm(n,mean=mean.norm,sd=sd.norm)
resamp<-
replicate(m,sample(samp,n,replace=TRUE),simplify=FALSE)
mean.resamp<-sapply(resamp,mean)

par(mfrow=c(1,2))
qqnorm(mean.resamp)
qqline(mean.resamp)
hist(mean.resamp,prob=TRUE)
curve(dnorm(x,mean.norm,(1/sqrt(n))),add=TRUE)

cat("mean.samp =",mean(samp),"\n")
cat("mean.resamp =",mean(mean.resamp),"\n")
cat("bias.resamp =",(mean(samp)-mean(mean.resamp)),"\n")

shapiro.test(mean.resamp)
}
```

الآن يمكن تنفيذ هذه الدالة، (باستخدام نفس القيم والمعالم السابقة أو قيم مختلفة)، عدة مرات ومراقبة النتائج بشكل أسهل، فمثلا يمكن اختيار حجم عينة أصغر وعدد مرات إعادة معاينة أقل، فتكون "إحدى" النتائج العشوائية بالصورة التالية:

```
> bootstrap.fun1(20,4,1,300)

mean.samp = 3.606401
mean.resamp = 3.59401
bias.resamp = 0.0123913

        Shapiro-Wilk normality test

data:  mean.resamp
W = 0.9896, p-value = 0.03158
```

لاحظ أن تقدير متوسط المجتمع قد ابتعد عن القيمة الحقيقية 4 بشكل أكبر، وأن مقدار التحيز قد ازداد بعض الشيء عن المثال السابق، وأيضا طبقا لاختبار شابيرو–ويلك فإن توزيع البوتستراب لم يعد طبيعيا، وهذا كله بسبب انخفاض حجم العينة الأصلية وتقليل عدد مرات إعادة المعاينة، ومن الشكل (5.7) نستطيع ملاحظة أن

متوسط البيانات $\bar{X}_1, \bar{X}_2 \ldots, \bar{X}_{300}$، (المدرج التكراري)، يبتعد عن الوسط الحسابي الفعلي 4، (المنحنى الطبيعي)، بشكل ملحوظ رغم أن المدرج التكراري يبدو بأنه يتوزع طبيعيا.



شكل 5.7: رسم QQ الطبيعي (إلى اليسار) والمدرج التكراري (إلى اليمين) مضافا إليه منحنى التوزيع الطبيعي بمتوسط 4 وانحراف معياري 0.1 باستخدام $n = 20, m = 300$

وتجدر الإشارة إلى أن أسلوب البوتستراب يتوفر أيضا ضمن الحزمة الإضافية boot، وسنقوم بعرض التطبيق التالي لاستخدامها بشكل موجز هنا؛

نقوم أولا باستدعاء الحزمة boot بعد تحميلها؛

```
> library(boot)
```

بعد ذلك يجب تعريف الإحصاءة التي سيتم استخدام أسلوب البوتستراب لتقديرها عن طريق دالة المستخدِم حتى وإن كانت متوفرة ضمن لغة R، فمثلا لتقدير الوسط الحسابي نقوم أولا بتوليد العينة الأصلية، ولنستخدم نفس القيم السابقة:

```
> samp2<-rnorm(100,mean=4,sd=1)
```

ثم نقوم بكتابة دالة المستخدِم للإحصاءة، وهي الوسط الحسابي، بالصورة التقليدية التالية:

```
> mean.fun<-function(x,indices)mean(x[indices])
```

واستخدام دالة مستخدِم لمتغيرين هو أساسي في دالة boot لأن الأول سيكون "محجوزا" للعينة الأصلية والثاني سيكون للعينات المُعاد سحبها، الآن نقوم باستخدام دالة boot بالصورة التالية:

```
> res.boot.mean<-
boot(data=samp2,statistic=mean.fun,R=2000)
```

ومكونات الدالة الأساسية كما نلاحظ هي العينة الأصلية data، الإحصاءة المستخدَمة statistic، والرمز R الذي يرمز هنا لعدد مرات إعادة المعاينة. وباستدعاء النتائج السابقة نحصل على:

```
> res.boot.mean

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = samp2, statistic = mean.fun, R = 2000)

Bootstrap Statistics :
    original      bias    std. error
t1* 3.961797 0.001241869   0.1020347
```

حيث توفر النتائج تقدير الإحصاءة من العينة الأصلية original ومقدار التحيز bias والخطأ المعياري للتقدير std.error.

وكمثال آخر على استخدام دالة boot، لنقم بتقدير إحصاءة الانحراف المعياري للعينة باستخدام نفس القيم السابقة؛

```
> sd.fun<-function(x,indices)sd(x[indices])
res.boot.sd<-boot(data=samp2,statistic=sd.fun,R=2000)
```

فنحصل على:

```
> res.boot.sd

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = samp2, statistic = sd.fun, R = 2000)
Bootstrap Statistics :

    original       bias    std. error
t1* 1.053829 -0.006807307  0.06181336
```

5.7 بعض دوال R الإضافية (Some Additional Functions of R)

في هذا البند سنتناول بعض دوال نظام R الإضافية الهامة والتي لم يتم التعرض لها في الفصول السابقة، وهذه الدوال هي بمثابة أدوات مساعدة يمكن استخدامها في حالات معينة مثل تعديل أُطر البيانات داخليا وإعادة ترميز المتغيرات وغيرها من الحالات.

- **دالة** transform:

تُستخدم دالة **التحويل** (transform) لإجراء التعديلات الثانوية أو الإضافية على الأشياء بصورة عامة، وبشكل عملي تُعتبر من الدوال الهامة لتعديل المتغيرات رياضيا ضمن المصفوفات وأُطر البيانات، والشكل العام لتنفيذ هذه الدالة هو؛

transform (التعديل المطلوب , اسم البيانات)

ويمكن، كما ذكرنا، استخدام هذه الدالة للتعديل أو التغيير لأي شيء أو متغير، إلا أنها تستخدم عادة لإجراء التعديلات ضمن أُطر البيانات كما يوضح المثال التالي:

لنقم بإدخال القيم الافتراضية في الجدول التالي، (جدول (1.7))، كإطار بيانات باسم data.man وذلك باستخدام محرر بيانات R؛

جدول 1.7: بيانات إطار البيانات data.man

x_1	2	4	5	6	8	1	5	5	9	7
x_2	17	8	15	12	10	-4	0	11	19	16
x_3	yes	yes	no	no	no	yes	no	no	no	yes
x_4	L1	L3	L1	L2	L2	L4	L1	L1	L3	L4

(لعمل ذلك يمكن مثلا تعريف إطار البيانات؛ data.man<-data.frame()، ثم كتابة؛ data.man<-edit(data.man)، ومن ثمة إدخال القيم في محرر البيانات).

بعد ذلك يصبح لدينا؛

```
> data.man

   x1 x2  x3 x4
1   2 17 yes L1
2   4  8 yes L3
3   5 15  no L1
4   6 12  no L2
5   8 10  no L2
6   1 -4 yes L4
7   5  0  no L1
8   5 11  no L1
9   9 19  no L3
10  7 16 yes L4
```

لنقم الآن باستخدام دالة transform لإضافة متغير جديد، وليكن x5، والذي يساوي سالب قيم المتغير x2، عندئذ نكتب؛

```
> data.man1<-transform(data.man,x5=-x2)

> data.man1

   x1 x2  x3 x4  x5
1   2 17 yes L1 -17
2   4  8 yes L3  -8
3   5 15  no L1 -15
4   6 12  no L2 -12
5   8 10  no L2 -10
6   1 -4 yes L4   4
7   5  0  no L1   0
8   5 11  no L1 -11
9   9 19  no L3 -19
10  7 16 yes L4 -16
```

(لاحظ أنه من الممكن تغيير قيم المتغير x2 نفسه إلى القيم السالبة بكتابة x2=-x2 بدلا من x5=-x2 في الأمر السابق). الآن لنقم بإضافة القيمة 3 إلى المتغير x1 وتربيع المتغير x2 وذلك للبيانات الأصلية data.man؛

```
> data.man2<-transform(data.man,x1=x1+3,x2=x2^2)

> data.man2

   x1  x2  x3 x4
1   5 289 yes L1
2   7  64 yes L3
3   8 225  no L1
4   9 144  no L2
5  11 100  no L2
6   4  16 yes L4
7   8   0  no L1
8   8 121  no L1
9  12 361  no L3
10 10 256 yes L4
```

وإذا ما أردنا إعادة ترميز متغير وصفي أو أكثر فيمكن عمل ذلك بسهولة باستخدام دالة التحويل، إلا أنه يجب الانتباه دائما إلى ضرورة تغيير طبيعة المتغيرات الوصفية من character إلى factor عند إعادة ترميزها، لذلك سنقوم أولا بذلك في إطار البيانات الأصلي، (للمتغيرين x3 و x4 في عملية واحدة)، تمهيدا للاستخدامات الأخرى:

```
> data.man<-
transform(data.man,x3=as.factor(x3),x4=as.factor(x4))
```

الآن نقوم بتغيير x3 إلى متغير كمي كالتالي:

```
> data.man3<-transform(data.man,x3=as.numeric(x3))

> data.man3

   x1 x2 x3 x4
1   2 17  2 L1
2   4  8  2 L3
3   5 15  1 L1
4   6 12  1 L2
5   8 10  1 L2
6   1 -4  2 L4
7   5  0  1 L1
8   5 11  1 L1
9   9 19  1 L3
10  7 16  2 L4
```

من جديد لنفرض أننا نود تعديل البيانات الأصلية بصورة أكثر تعقيدا، فعلى سبيل المثال لنقم بتعريف متغير جديد يأخذ القيمة 1 عندما يكون (x3= yes) أو (L4 أو x4= L1) ويأخذ القيمة 0 لغير ذلك. عندها نكتب:

```
> data.man4<-transform(data.man,x6=ifelse((x3=="yes"|
(x4=="L1"|x4=="L4")),1,0))

> data.man4

   x1 x2  x3 x4 x6
1   2 17 yes L1  1
2   4  8 yes L3  1
3   5 15  no L1  1
4   6 12  no L2  0
5   8 10  no L2  0
6   1 -4 yes L4  1
7   5  0  no L1  1
8   5 11  no L1  1
9   9 19  no L3  0
10  7 16 yes L4  1
```

وهكذا يمكن إجراء أي نوع من التغييرات أو التعديلات على المتغيرات ضمن أُطر البيانات بمرونة كبيرة كما رأينا باستخدام الدالة transform، وذلك إن لم يتم مسبقا إجراء هذه التغييرات داخل نظام R أو في الملف الأصلي للبيانات، إذا ما كان بصيغة اكسل أو غيرها.

- **دالة** cut:

تقوم دالة cut بتحويل المتغير الكمي إلى عامل أو متغير تصنيف (Grouping Factor) بفترات محددة، فقد تتطلب بعض الدراسات أو الأساليب الإحصائية الاستكشافية أو التحليلية المتقدمة أحيانا التعامل مع هذا النوع من المتغيرات، فمثلا قد نقوم بإدراج قيم متغيرات كمية مثل ضغط الدم أو العمر أو الدخل الشهري ضمن فترات لها تكرارات مجموعها هو عدد القيم بالطبع، وهذا يُشبه ببساطة تكوين **الجدول التكراري** (Frequency Table) من حيث المبدأ.

لنفرض مثلا أننا نرغب بتحويل المتغير الكمي s.grd1، والذي يمثل درجات الطلبة في المقرر الأول في البيانات stu.data1، إلى متغير تصنيف ذو 4 فترات، عندئذ نكتب ضمن دالة cut اسم المتغير متبوعا بعدد الفترات المرغوب كالتالي:

```
> grd1.int<-cut(s.grd1,4)
```

لكن لاحظ أنه عند استدعاء المتغير الجديد grd1.int فإننا لا نحصل على الشكل المطلوب، (وهي فترات مناظرة لتكرارات قيم المتغير)، بل نحصل على النتيجة التالية:

```
> grd1.int
 [1] (49.8,64.5] (34.9,49.8] (49.8,64.5] (64.5,79.2]
 [5] (34.9,49.8] (64.5,79.2] (64.5,79.2] (79.2,94.1]
 [9] (79.2,94.1] (64.5,79.2] (49.8,64.5] (64.5,79.2]
[13] (64.5,79.2] (64.5,79.2] (79.2,94.1] (34.9,49.8]
[17] (34.9,49.8] (49.8,64.5] (79.2,94.1] (64.5,79.2]
[21] (79.2,94.1] (79.2,94.1] (64.5,79.2] (79.2,94.1]
[25] (49.8,64.5] (79.2,94.1] (79.2,94.1] (49.8,64.5]
[29] (64.5,79.2] (64.5,79.2] (64.5,79.2] (64.5,79.2]
[33] (79.2,94.1] (64.5,79.2] (79.2,94.1]
Levels: (34.9,49.8] (49.8,64.5] (64.5,79.2] (79.2,94.1]
```

بمعنى أنه قد تم استبدال كل قيمة من قيم المتغير s.grd1 بالفترة التي يقع ضمنها، (مع وجود توضيح للفترات التي تم إنشاؤها في السطر الأخير من النتيجة)، وهذا بحد ذاته لا يُعد مفيدا كثيرا من الناحية التطبيقية، لذلك عادة ما نقوم "بجدولة" المتغير الجديد باستخدام الدالة table:

```
> table(grd1.int)

grd1.int
(34.9,49.8] (49.8,64.5] (64.5,79.2] (79.2,94.1]
          4           6          14          11
```

```
> length(grd1.int)
[1] 35

> range(s.grd1)
[1] 35 94
```

ولاحظ أن الأقواس "( )" هي للكناية عن الفترة المفتوحة أما الأقواس "[ ]" فهي للفترات المغلقة، ونستطيع رؤية أن الفترات هي مفتوحة من الحد الأدنى ومغلقة من الحد الأعلى، فمثلا الفترة الثانية [49.8,64.5) تضم كل الدرجات الأقل من أو تساوي 64.5 درجة، علما بأن القيم الأصلية تمثل أعداد أو درجات صحيحة لذلك نقول أن الفترة الأولى تضم درجات الطلبة من 35 إلى 49 درجة، والفترة الثانية تضم الدرجات من 50 إلى 64 درجة، وهكذا.

ويمكن اتباع طريقة أخرى لتكوين الفترات تعتمد على تحديد طول الفترة وليس عدد الفترات كما في المثال السابق، وذلك باستخدام الدالة seq وإدراج الحد الأدنى للقيم ناقصا الواحد الصحيح، الحد الأعلى زائدا الواحد الصحيح، وطول الفترة المرغوب، فمثلا لتكوين فترات طولها 10 درجات نكتب التالي:

```
> grd1.int<-cut(s.grd1,seq(34,95,10))
> table(grd1.int)

grd1.int
(34,44] (44,54] (54,64] (64,74] (74,84] (84,94]
      2       4       4       7      10       8
```

ويمكن استخدام الخيارين right وinclude.lowest اللذان يأخذان القيمتين المنطقيتين T أو F للتحكم بجعل حدود الفترات مفتوحة أو مغلقة، وفيما يلي نعرض تطبيق للحالات الأربع الممكنة لهذه الخيارات، علما بأن الحالة التي تتضمن right=T,include.lowest=F هي الحالة الافتراضية كما سنلاحظ مما يلي:

1. الحد الأدنى للفترة الأولى والحد الأعلى للفترة الأخيرة كلاهما مغلقان:

```
> grd1.int<-
cut(s.grd1,seq(34,95,10),right=T,include.lowest=T)
> table(grd1.int)

grd1.int
[34,44] (44,54] (54,64] (64,74] (74,84] (84,94]
      2       4       4       7      10       8
```

2. الحد الأدنى للفترة الأولى مفتوح والحد الأعلى للفترة الأخيرة مغلق:

```
> grd1.int<-
cut(s.grd1,seq(34,95,10),right=T,include.lowest=F)
> table(grd1.int)

grd1.int
(34,44] (44,54] (54,64] (64,74] (74,84] (84,94]
      2       4       4       7      10       8
```

3. الحدود الدنيا لكل الفترات مغلقة وكذلك الحد الأعلى للفترة الأخيرة:

```
> grd1.int<-
cut(s.grd1,seq(34,95,10),right=F,include.lowest=T)
> table(grd1.int)

grd1.int
[34,44) [44,54) [54,64) [64,74) [74,84) [84,94]
      2       4       3       8       9       9
```

4. الحدود الدنيا لكل الفترات مغلقة والحدود العليا لكل الفترات مفتوحة:

```
> grd1.int<-
cut(s.grd1,seq(34,95,10),right=F,include.lowest=F)
> table(grd1.int)

grd1.int
[34,44) [44,54) [54,64) [64,74) [74,84) [84,94)
      2       4       3       8       9       8
```

ويمكن عرض تنسيق آخر لإنشاء الفترات يتضمن تقسيم القيم الأصلية إلى فترات تضم التقسيمات الجزئية لهذه القيم، بمعنى إنشاء الفترات التي تضم النسب المتوالية التالية:

(0% - 25%)، (25% - 50%)، (50% - 75%)، و(75% - 100%) كما يلي؛

```
> grd1.q<-quantile(s.grd1)

> grd1.q

  0%  25%  50%  75% 100%
35.0 63.5 75.0 83.0 94.0

> grd1.int<-cut(s.grd1,grd1.q)
> table(grd1.int)

grd1.int
(35,63.5] (63.5,75]   (75,83]   (83,94]
        8        11         6         9
```

وأما إذا أردنا الحصول على نتيجة أكثر ملائمة من الناحية الشكلية، فيمكننا ببساطة تغيير أسماء الفترات إلى أعداد صحيحة تتناسب مع قيم المتغير، (أو إلى أية أسماء أخرى قد نرغب بإدراجها)، كالتالي:

```
> levels(grd1.int)<-c("35-63","64-75","76-83","84-94")
> table(grd1.int)

grd1.int
35-63 64-75 76-83 84-94
    8    11     6     9
```

- **دالة** factor:

قمنا فيما سبق بتحويل المتغيرات إلى متغيرات عاملية باستخدام الدالة as.factor، إضافة إلى أنه تم تناول تطبيق دالة factor في الفصل الرابع لتعيين المتجهات العاملية، وسوف نعرض أهم الخيارات الخاصة بهذه الدالة فيما يلي؛

عند تعريف المتغير أو المتجه العاملي للمرة الأولى فإن الخيار levels يتم استخدامه لترميز المستويات في المتغير، أما الخيار labels فيُستخدم لإعطاء أسماء لهذه المستويات، فمثلا لنفرض أنه لدينا المتغير التالي (marks)، والذي يمثل المستوى الدراسي لمجموعة مكونة من 20 طالبا على تقسيم يتدرج من 1 إلى 5؛

```
> marks<-c(2,1,1,5,4,4,2,4,3,1,3,5,1,5,5,1,4,4,5,1)

> class(marks)
[1] "numeric"
```

عندها يمكن تعريف المتغير marksf مثلا كمتغير عاملي من المتغير marks بالصورة التالية:

```
> marksf<-
factor(marks,levels=1:5,labels=c("F","D","C","B","A"))

> marksf

 [1] D F F A B B D B C F C A F A A F B B A F
Levels: F D C B A

> class(marksf)
[1] "factor"
```

ولاحظ كيفية استخدام الخيار levels لإدخال مستويات التدرج من المستوى 1 إلى المستوى 5، واستخدام الخيار labels لتعيين الأسماء لهذه المستويات الخمسة. لنقم الآن باختيار أسماء أخرى للمستويات السابقة كالتالي؛

```
> marksf2<-factor(marks,levels=1:5,labels=c("v.bad","bad"
,"average","good","excellent"))
```

```
> marksf2

 [1] bad       v.bad     v.bad     excellent good
 [6] good      bad       good      average   v.bad
[11] average   excellent v.bad     excellent excellent
[16] v.bad     good      good      excellent v.bad
Levels: v.bad bad average good excellent
```

ويمكن أيضا الحصول على الجدول التكراري للمتغير marksf2 بالصورة التالية؛

```
> table(marksf2)

marksf2
    v.bad       bad   average      good excellent
        6         2         2         5         5
```

وفي بعض الأحيان، قد نرغب بدمج بعض المستويات مع بعضها البعض مما ينتج عنه تغيير في عدد المستويات، هذا الدمج يمكن تنفيذه بعدة طرق أبسطها الطريقة التالية؛ لنفرض أننا نرغب بدمج المستويين الأول "v.bad" والثاني "bad" في مستوى جديد باسم "bad" في المتغير السابق marksf2، عندها يمكن استخدام الدالة list لتنفيذ هذا الدمج كالتالي:

```
> marksf3<-marksf2 #إجراء التغيير على متغير جديد

> levels(marksf3)<-list(bad=c("v.bad","bad"),average=
"average",good="good",excellent="excellent")

> marksf3

 [1] bad       bad       bad       excellent good
 [6] good      bad       good      average   bad
[11] average   excellent bad       excellent excellent
[16] bad       good      good      excellent bad
Levels: bad average good excellent

> table(marksf3)

marksf3
      bad   average      good excellent
        8         2         5         5
```

وننوه هنا إلى أنه في حالة التعامل مع متغير يأخذ قيما رقمية، كما هو الحال مع المتغير marks، فإنه يمكن الاستغناء عن كتابة الخيار levels لأن المستويات مرتبة مسبقا بحسب القيم الموجودة في المتغير الأصلي، أما عندما تكون قيم المتغير عبارة عن قيم نصية فإن الوضع يتغير عندها، ولنأخذ المثال التالي:

```
> marks2<-c("v.bad","bad","average","bad","good",
"excellent","good","good")

> class(marks2)
[1] "character"

> factor(marks2)

[1] v.bad     bad       average   bad       good
[6] excellent good      good
Levels: average bad excellent good v.bad
```

لاحظ كيف أن المستويات في السطر الأخير مرتبة ترتيبا أبجديا وليس منطقيا (من المستوى الأدنى إلى المستوى الأعلى)، لذلك إذا ما أردنا عرض المستويات مرتبة منطقيا فيجب علينا كتابة الخيار levels كمتجه بالصورة التالية:

```
> factor(marks2,levels=c("v.bad","bad","average","good"
,"excellent"))

[1] v.bad     bad       average   bad       good
[6] excellent good      good
Levels: v.bad bad average good excellent
```

- **دوال التدوير** (Rounding Functions)

توجد بعض الدوال في نظام R التي تُستخدم في تقريب أو تدوير الأرقام إلى جانب الدالة round التي تم استخدامها في الفصول السابقة، وأهم هذه الدوال هي ceiling، floor، trunc، وsignif، وسنقوم فيما يلي بتعريفها وعرض بعض الأمثلة عليها؛

- الدالة ceiling:

تقوم الدالة ceiling بالتقريب إلى العدد الصحيح الأكبر، فمثلا:

```
> ceiling(c(1.236,3.624,-0.059,0.075,-25.95))

[1]   2   4   0   1 -25
```

- الدالة floor:

الدالة floor تقوم بالتقريب إلى العدد الصحيح الأصغر، فمثلا:

```
> floor(c(1.236,3.624,-0.059,0.075,-25.95))

[1]   1   3  -1   0 -26
```

- الدالة trunc:

أما الدالة trunc فتقوم بالتقريب لأقرب عدد صحيح ما بين الرقم الأصلي والصفر، بمعنى أنها تعمل مثل الدالة ceiling مع القيم الأقل من الصفر، وتعمل مثل الدالة floor مع القيم الأكبر من الصفر، وكمثال على ذلك:

```
> trunc(c(1.236,3.624,-0.059,0.075,-25.95))
[1]   1   3   0   0 -25
```

- الدالة signif:

تقوم الدالة signif بالتدوير إلى عدد من الخانات يحدده المُستخدِم، وهي تشبه الدالة round من حيث التطبيق، ولنأخذ المثال التالي لتوضيح الفرق بينهما:

```
> round(c(5.263,14.628,-5.263,-14.628),digits=2)
[1]   5.26  14.63  -5.26 -14.63
> signif(c(5.263,14.628,-5.263,-14.628),digits=2)
[1]   5.3  15.0  -5.3 -15.0
```

ولاحظ أن الخيار digits=2 مع الدالة round يعني تقريب القيمة إلى خانتين بعد الفاصلة العشرية، أما مع الدالة signif فيعني استخدام الخانتين الأولى والثانية بعد الفاصلة العشرية للتقريب إلى أقرب قيمة ممكنة.

في نهاية الفصل السابع، نذكّر القارئ بإمكانية استخدام دالة المساعدة help دائما للتعرف على المزيد من الخيارات الإضافية الخاصة بدوال R وأيضا لاستكشاف الدوال الأخرى التي لم يتم التطرق إليها فيهذا الكتاب.

ملحق 1: الجداول الخاصة بملفات البيانات "excdata1" و"studata1"

جدول م1.1: البيانات الخاصة بالملف "excdata1"

	Statistics	Botany	Zoology	Chemistry	Physics	Mathematics
S1	1.51	1.76	1.68	1.67	1.40	1.47
S2	1.67	1.95	1.80	1.75	1.40	1.55
S3	1.88	1.70	1.96	1.67	1.70	1.56
S4	1.81	1.77	1.90	1.66	1.53	1.48
S5	1.79	1.89	1.83	1.58	1.55	1.55
S6	1.75	2.01	1.89	1.74	1.39	1.41
S7	1.69	1.90	1.97	1.88	1.51	1.66
S8	1.77	1.88	1.91	1.77	1.40	1.56
S9	1.72	1.66	1.70	1.68	1.62	1.66
S10	1.90	1.91	1.86	1.76	1.31	1.61
S11	1.96	2.01	1.78	1.71	1.27	1.51
S12	1.76	1.85	1.77	1.78	1.38	1.48
S13	1.90	1.83	1.73	1.83	1.51	1.55
S14	1.85	1.85	1.77	1.85	1.46	1.67
S15	1.91	1.93	1.79	1.84	1.65	1.72
S16	2.05	1.99	1.73	1.76	1.46	1.48
S17	1.81	2.00	1.80	1.78	1.48	1.54
S18	1.75	2.02	1.77	1.78	1.47	1.43
S19	1.81	1.86	1.84	1.84	1.57	1.28
S20	1.82	2.00	1.65	1.66	1.43	1.41
S21	1.84	2.04	1.96	1.67	1.63	1.45
S22	1.73	1.88	1.82	1.67	1.53	1.25
S23	1.69	2.00	1.93	1.73	1.41	1.43
S24	1.51	1.95	1.90	1.67	1.45	1.54
S25	1.60	1.84	1.84	1.72	1.34	1.61
S26	1.42	1.94	1.73	1.60	1.54	1.39
S27	1.77	1.96	1.86	1.71	1.56	1.49
S28	1.64	1.89	1.67	1.75	1.50	1.42
S29	1.79	1.84	1.71	1.78	1.46	1.42
S30	1.90	1.79	1.62	1.71	1.46	1.47
S31	1.95	1.85	1.71	1.74	1.57	1.50
S32	1.65	1.83	1.61	1.70	1.52	1.31
S33	1.90	1.94	1.77	1.71	1.56	1.40
S34	1.72	1.86	1.56	1.69	1.35	1.00
S35	1.65	1.93	1.74	1.71	1.61	1.29
S36	1.65	2.00	1.54	1.71	1.43	0.97
S37	1.71	1.89	1.61	1.69	1.58	1.11
S38	1.69	2.06	1.70	1.64	1.44	1.24
S39	1.83	2.03	1.67	1.69	1.51	1.44
S40	1.85	1.85	1.65	1.67	1.43	1.22
S41	1.74	1.86	1.71	1.69	1.66	1.26

جدول م2.1: البيانات الخاصة بالملف "studata1"

	grd1	grd2	grd3	age	gen	sem	fam	hou
stu1	55	50	60	22	m	3	10	2
stu2	49	52	50	19	m	8	11	2
stu3	60	54	51	23	m	2	10	2
stu4	65	70	54	20	f	3	8	3
stu5	35	40	40	24	m	7	12	2
stu6	71	70	45	22	m	4	9	3
stu7	73	74	49	21	m	5	9	4
stu8	90	91	61	22	f	4	3	6
stu9	88	93	59	20	f	4	3	5
stu10	75	77	60	22	m	3	6	4
stu11	50	51	61	25	m	7	9	3
stu12	77	79	59	24	m	2	6	4
stu13	79	81	33	23	f	4	4	5
stu14	66	70	60	21	m	6	5	4
stu15	80	82	58	20	f	2	5	5
stu16	40	44	60	24	m	8	12	2
stu17	45	50	43	25	m	7	11	3
stu18	51	55	94	24	f	6	9	3
stu19	82	85	50	22	f	3	4	5
stu20	75	77	27	22	m	5	4	4
stu21	84	84	57	20	f	4	4	5
stu22	86	87	52	23	f	8	4	4
stu23	77	78	57	22	f	6	5	5
stu24	88	90	62	19	f	7	3	5
stu25	64	70	50	21	m	7	7	3
stu26	89	91	53	23	f	4	3	5
stu27	90	93	50	21	f	6	3	6
stu28	63	66	51	25	m	5	5	4
stu29	75	85	95	22	f	3	2	6
stu30	69	70	44	23	m	5	6	4
stu31	77	80	92	21	f	8	4	5
stu32	68	69	54	22	m	6	5	5
stu33	93	75	45	23	f	5	2	6
stu34	73	75	52	24	m	6	5	4
stu35	94	96	49	21	f	4	2	6

ملحق 2: خيارات التمثيل البياني في نظام R

(Graphical Options in R)

عند استخدام الرسم أو التمثيل البياني في نظام R، كثيرا ما يرغب المُستخدِم بإجراء تغييرات على ذلك الرسم مثل إضافة عنوان رئيسي أو فرعي، تغيير أسماء المحاور، إضافة منحنى أو خط مستقيم على الرسم الأصلي، تصغير أو تكبير الخطوط على الرسم، إضافة وتعديل الألوان، وغير ذلك من التعديلات. لذلك سنتناول في هذا الملحق بعض أهم الخيارات التي يمكن استخدامها لإجراء التعديلات المختلفة في التمثيل البياني.

وحيث أنه يوجد العديد من التعديلات أحيانا للخيار الواحد في الرسم إضافة إلى وجود العشرات من دوال الخيارات، فإن المحصلة قد تكون عدد ضخم جدا من التغييرات التي يمكن تنفيذها في الرسم الواحد إذا ما تم استخدام كل أو حتى معظم الخيارات المتاحة، لذلك سنقوم بعرض أمثلة متنوعة على تعديل خيارات رسومات سبق أن تناولناها في الفصول السابقة، ويمكن للقارئ دائما تجربة باقي الخيارات الأخرى.

ننوه هنا إلى أن سيتم في هذا الملحق استخدام متغيرات لبيانات سبق التعامل معها في فصول الكتاب، وللمحافظة على النسق العام السائد في الكتاب من حيث التعامل مع نظام R عمليا، لنقم بتخزين الأوامر في ملف عمل جديد باسم "workA2" وحفظ سطور الأوامر باسم "hisA2". ثم بعد ذلك لنقم باستيراد البيانات التي سنتعامل معها هنا من جديد، وهي البيانات stu.data1؛

```
> library(rJava)
> library(XLConnectJars)
> library(XLConnect)
> stu.data1<-readWorksheetFromFile("studata1.xlsx",
sheet=1,rownames=1)
```

والخطوة التالية ستكون إعادة ترميز المتغيرات في البيانات stu.data1؛

```
> s.grd1<-stu.data1$grd1
> s.grd2<-stu.data1$grd2
> s.grd3<-stu.data1$grd3
> s.age<-stu.data1$age
> s.gen<-stu.data1$gen
> s.sem<-stu.data1$sem
> s.fam<-stu.data1$fam
> s.hou<-stu.data1$hou
```

م1.2 الخيارات العامة للرسم البياني (General Options for Graphical Display)

يمكن تصنيف خيارات الرسم في لغة R إلى خيارات عامة أو أساسية يمكن استخدامها حسب الحاجة ضمن أي رسم، وتُعرف في R بدوال الرسم ذات **المستوى العالي** (High-level)، وخيارات أخرى مخصصة للرسومات بحسب طبيعتها، وتُعرف بأوامر الرسم ذات **المستوى المنخفض** (Low-level)، وسنسلط الضوء هنا على أهم خيارات الرسم العامة، ونتناول النوع الآخر من الخيارات في البند القادم.

من أجل رؤية أوضح لهذه الخيارات العامة، (والتي قمنا بتنفيذ الكثير منها سابقا)، لنقم بتنفيذ رسم افتراضي من "العدم"، بمعنى أن نبدأ برسم فارغ ثم نقوم بإضافة أو "إظهار" مكوناته العامة أو الرئيسية خطوة بخطوة، ولنبدأ بالأمر التالي للدالة `plot`، (باستخدام المتغيرين `s.grd1` و`s.age` ضمن البيانات `stu.data1`):

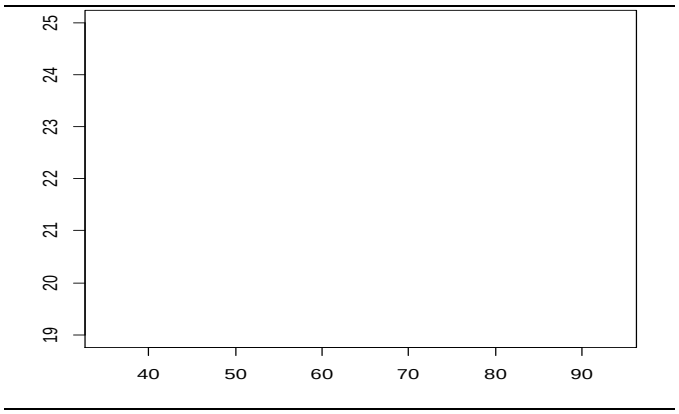
```
> plot(s.grd1,s.age,type="n",xlab="",ylab="",axes=F)
```

ستلاحظ ظهور نافذة الرسم فارغة نظرا لأنه تم اختيار عدم إظهار أي نقاط أو خطوط أو أي شيء آخر داخل الشكل باختيار (`type="n"`)[1]، وتم اختيار عناوين محاور فارغة (`xlab=""` و`ylab=""`)، وأيضا اخترنا عدم رسم صندوق الرسم والمحورين السيني والصادي (`axes=F`).

وبالتالي لإظهار صندوق الرسم مع محوريه الرئيسيين يجب اختيار `axes=T`، وإعادة تنفيذ سطر الأمر السابق بالصورة التالية:

```
> plot(s.grd1,s.age,type="n",xlab="",ylab="",axes=T)
```

وتكون النتيجة هي ظهور الشكل (م1.2)؛



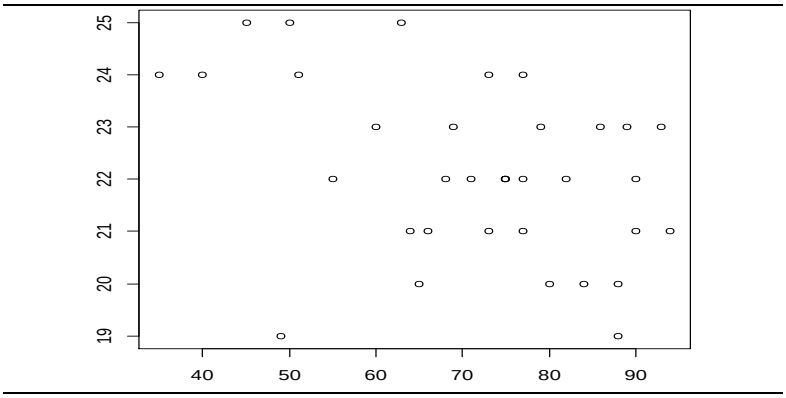
شكل م1.2: نتيجة تنفيذ الأمر:
`plot(s.grd1,s.age,type="n",xlab="", ylab="",axes=T)`

[1] أهم الاختيارات الأخرى للخيار `type` هي `"p"` لرسم النقط، `"l"` لرسم الخطوط المستقيمة، `"b"` لرسم نقاط متصلة بخطوط، `"o"` لرسم نقاط تمر بها الخطوط، `"h"` لرسم خطوط من النقاط إلى المحور السيني، و`"s"` و`"S"` لرسم دوال الشكل السُلّمي.

من جديد لنقم بعرض نقاط شكل الانتشار، (وهو ما تقوم بتنفيذه دالة plot في هذا الأمر)، وذلك عن طريق حذف الخيار type="n" من الأمر السابق؛

```
> plot(s.grd1,s.age,xlab="",ylab="",axes=T)
```

فتظهر النقاط كما يتضح من الشكل (م2.2) التالي:



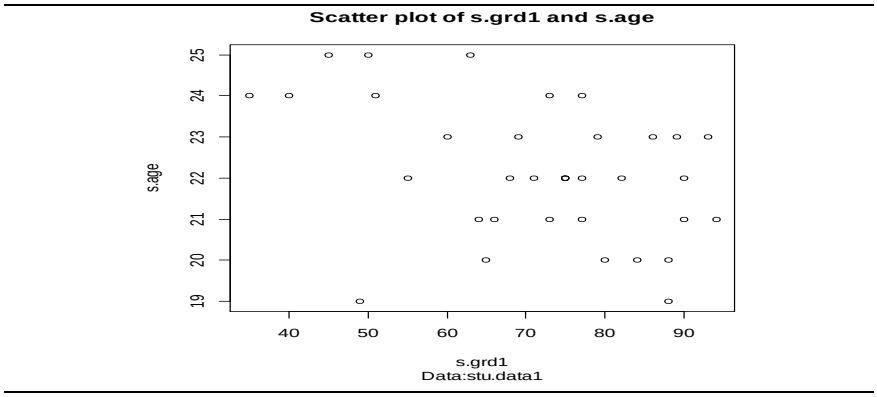
شكل م2.2: نتيجة تنفيذ الأمر:

```
plot(s.grd1,s.age,xlab="",ylab="",axes=T)
```

ويمكننا أيضا كتابة عنوان أساسي وآخر فرعي للرسم وكتابة عناوين للمحورين السيني والصادي بالصورة التالية:

```
> plot(s.grd1,s.age,main="Scatter plot of s.grd1 and
s.age",sub="Data:stu.data1",xlab="s.grd1",ylab="s.age",ax
es=T)
```

وبتنفيذ الأمر السابق نحصل على الشكل (م3.2) التالي. كما يمكن استخدام الخيار add=T مع بعض دوال الرسم لإضافة شكل ما إلى الرسم الأصلي، فمثلا يمكن باستخدام هذا الخيار إضافة منحنى التوزيع الطبيعي إلى رسم المدرج التكراري، (كما هو الحال في سطر الأمر للدالة curve والخاص برسم الشكل (5.7) في الفصل السابع).



شكل م3.2: نتيجة تنفيذ الأمر:

```
plot(s.grd1,s.age, main="Scatter plot of s.grd1
and s.age",xlab="s.grd1",sub="Data:stu.data1",
ylab="s.age",axes=T)
```

م2.2 الخيارات المُخصصة في الرسم البياني (Particular Options for Graphical Display)

ننتقل هنا إلى الحديث عن خيارات الرسم ذات **المستوى المنخفض** ونعني بها تلك الخيارات التي من الممكن استخدامها في بعض وليس كل الرسومات نظرا لأن كل تمثيل بياني له طبيعة خاصة به، فمثلا لا معنى لاستخدام خيار رسم النقاط مع التمثيل البياني الخاص بالمدرج التكراري، فعند عمل ذلك في بعض الحالات قد لا يحدث شيء وفي حالات أخرى قد تظهر رسالة تحذيرية تدل على وقوع مثل هذا الخطأ، فعلى سبيل المثال يمكنك تنفيذ سطري الأوامر:

```
> hist(s.grd1)
> points(s.grd1)
```

فتلاحظ أن النتيجة هي ظهور المدرج التكراري فقط، أما عند تنفيذ الأمر:

```
> hist(s.grd1,points=T)
```

فإنك ستلاحظ ظهور الرسالة التحذيرية.

سنقوم الآن بإعادة تنفيذ نفس أمر الرسم "الفارغ" الذي تناولناه في بداية البند السابق؛

```
> plot(s.grd1,s.age,type="n",xlab="",ylab="",axes=F)
```

إلا أننا سنستخدم هنا وبشكل متتالي الخيارات المخصصة "لبناء" الشكل الكلي للرسم خطوة بخطوة. الآن بعد تنفيذ الأمر السابق، وظهور النافذة الفارغة سنقوم بعرض المحور الأفقي السيني مثلا على الشكل الفارغ باستخدام دالة `axis` كالتالي:

```
> axis(1)
```

ثم عرض المحور العامودي الصادي بكتابة[1]:

```
> axis(2)
```

ثم عرض مربع أو صندوق الرسم باستخدام دالة `box()` كالتالي:

```
> box()
```

فيظهر لنا شكل مطابق للشكل (م1.2) أعلاه. الآن نقوم بعرض نقاط شكل الانتشار باستخدام الدالة المخصصة `points` كالتالي:

```
> points(s.grd1,s.age)
```

فنحصل على شكل انتشار مطابق للشكل (م2.2) أعلاه. وأخيرا، يمكننا استخدام دالة `title` لإضافة العنوان الرئيسي والفرعي وعناوين المحاور؛

[1] يمكن كتابة `axis(3)` و `axis(4)` لعرض المحور السيني في أعلى الرسم وعرض المحور الصادي على يمين الرسم على الترتيب.

```
> title(main="Scatter plot of s.grd1 and s.age",sub=
"Data: stu.data1",xlab="s.grd1",ylab="s.age")
```

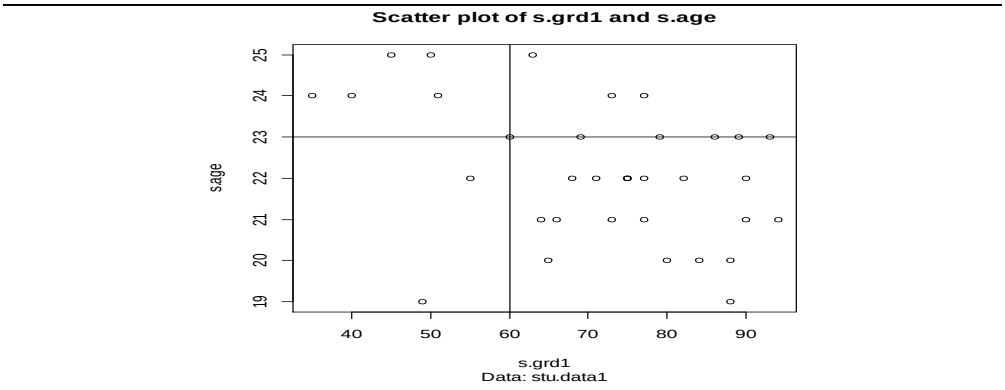
فتكون النتيجة تمثيل بياني مطابق للشكل (م3.2) السابق. وكملاحظة على ما تم ذكره في بداية هذا البند، فإنه إذا ما تم استخدام دالة الرسم المخصصة lines مثلا بدلا من أو بالإضافة إلى دالة رسم النقاط points فإن النتيجة ستكون ظهور خطوط مستقيمة تربط بين "مواقع" نقاط شكل الانتشار، وهذا بالطبع لن يكون المطلوب عرضه ضمن هذه النوعية من الرسومات، (يمكنك تجربة تنفيذ lines(s.grd1,s.age) بدلا من points(s.grd1,s.age) وملاحظة الفرق في الشكل الناتج).

ويمكن استخدام دالة رسم مخصصة أخرى هي دالة رسم أو توفيق الخطوط المستقيمة abline مع رسومات النقاط بصورة عامة، فمثلا يمكننا، بعد تنفيذ سلسلة الأوامر السابقة والحصول على شكل مطابق للشكل (م3.2)، كتابة[1] الأمر:

```
> abline(h=23,v=60)
```

لإضافة خطين مستقيمين الأفقي منهما يمر بالقيمة 23 على المحور الصادي والعامودي يمر بالقيمة 60 على المحور السيني كما يوضح الشكل (م4.2).

ومن أكثر استخدامات هذه الدالة هو التمثيل البياني لتوفيق خط الانحدار على شكل الانتشار، حيث يمكن تنفيذ ذلك بتوفيق نموذج الانحدار الخطي أولا ثم استخدام دالة abline عن طريق كتابة اسم نموذج الانحدار الموفق بين القوسين، (راجع البند (1.2.5.6) والشكل (1.6)).



شكل م4.2: إضافة خطين مستقيمين متعامدين عند النقطة (60،23) باستخدام دالة الرسم abline

[1] ننبه أن استخدام الأمر abline(h=23,v=60) هنا لابد أن يتم تنفيذه بالتسلسل بعد تنفيذ الأمر السابق له مباشرة وهو؛

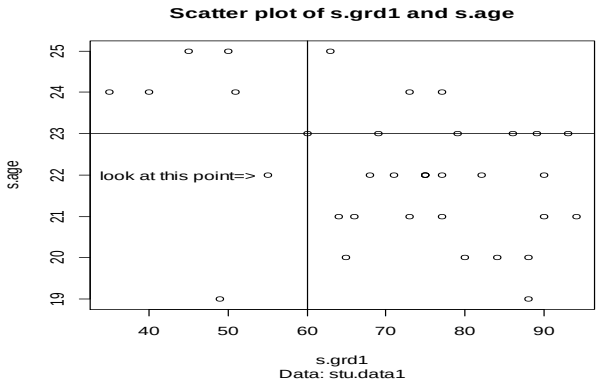
```
title(main="Scatter plot of s.grd1 and s.age",sub="Data:
stu.data1",xlab="s.grd1",ylab="s.age").
```

كما يمكن كتابة abline(a,b) لإضافة خط مستقيم يمر بالجزء المقطوع من المحور الصادي عند النقطة a و له الميل b، إلا أن الطريقة الأولى تُعتبر أفضل من الناحية الإحصائية المنهجية لرسم خط الانحدار.

الدالة text يمكن اعتبارها أيضا من دوال الرسم المخصصة الثانوية التي يمكن استخدامها لإضافة ملاحظات نصية ضمن مساحة الرسم الداخلية، فمثلا يمكن توجيه الانتباه إلى النقطة (55،22) على الشكل السابق (م4.2) عن طريق كتابة:

```
> text(55,22,"look at this point=>",pos=2)
```

فتظهر هذه الملاحظة النصية "look at this point =>"، والتي تعني "انظر إلى هذه النقطة"، كما هو موضح في الشكل (م5.2)؛



شكل م5.2: إضافة ملاحظة نصية عند النقطة (55،22) باستخدام دالة text

وقد تم استخدام الخيار pos=2 في الأمر السابق لتغيير موضع النص إلى يسار النقطة المحددة، (حيث يمكن استخدام الأرقام 1، أو 3، أو 4 لتغيير موضع النص إلى أسفل، أو أعلى، أو يمين النقطة المحددة على الترتيب)، علما بأن عدم استخدام هذا الخيار سيجعل النص يظهر على النقطة مباشرة بحيث يغطيها.

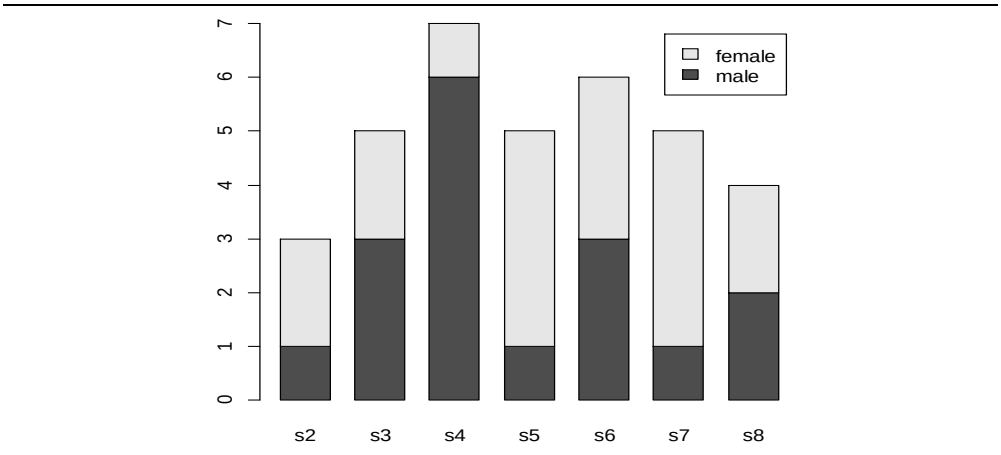
وسنستكمل الآن استعراض بعض خيارات الرسم المخصصة الإضافية من خلال تناول الأمثلة التطبيقية التالية؛

- **مثال على خيارات رسم الأعمدة البيانية:**

سنقوم برسم الأعمدة البيانية باستخدام المتغيرين gen وsem اللذان يمثلان نوع الطالب والفصل الدراسي له في البيانات stu.data1، حيث سنستخدم أولا الخيارات name، beside، space، وlegend بالصورة التالية:

```
> barplot(table(stu.data1[5:6]),names=c("s2","s3","s4",
"s5","s6","s7","s8"),beside=F, space=0.5,legend.text=
c("male","female"))
```

ولاحظ أن الخيار names يتضمن الأسماء التي سيتم عرضها أسفل كل عامود، والخيار beside=F هو لعرض تصنيف متغير النوع gen لكل مستوى (عامود) من مستويات المتغير sem، (واستخدام beside=T يعني عرض الذكور والإناث في عامودين متجاورين[1])، وفي الخيار space يتم تحديد المسافة المرغوبة بين الأعمدة، أما الخيار legend.text فهو أحد حالات الخيار legend الخاص بعرض دليل تفسيري للرسم، والشكل (م6.2) يمثل تمثيل الأعمدة البيانية المطلوب:



شكل م6.2: الأعمدة البيانية لتكرارات الفصل الدراسي للطلبة بحسب النوع في البيانات stu.data1

ونقدم طريقة أخرى لعرض الشكل السابق تتضمن استخدام دالة legend بصيغة مختلفة في سطر أوامر مستقل، وذلك عندما نرغب بتغيير تدرّج الألوان للأعمدة، (باستخدام دالة اللون col)؛

```
> barplot(table(stu.data1[5:6]),space=0.5,beside=F,names=
c("s2","s3","s4","s5","s6","s7","s8"),col=c("grey90","gre
y60"))

> legend(x=8.5,y=7,legend=c("female","male"),fill=
c("grey90","grey60"))
```

ونود الإشارة هنا إلى إمكانية استخدام الخيار horiz=T لعرض الأعمدة البيانية بشكل أفقي.

- **مثال على خيارات رسم شكل الانتشار:**

تناولنا في بداية الحديث عن الخيارات العامة للرسم مثالا توضيحيا عن شكل الانتشار، وسنقوم هنا بعرض مثال آخر يتضمن المزيد من خيارات الرسم الإضافية. لنأخذ المتغيرين الافتراضيين X1 وX2 المعطاة قيمهما كالتالي:

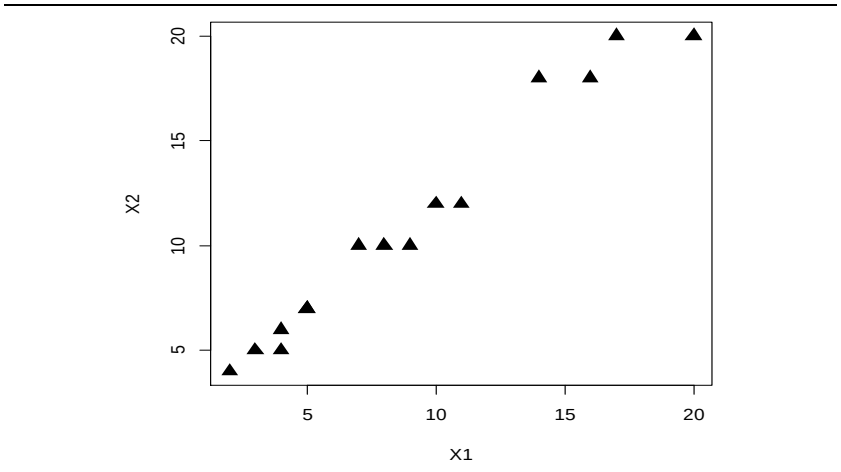
```
> X1<-c(2,4,5,7,9,14,4,16,10,11,8,17,20,5,3)
> X2<-c(4,6,7,10,10,18,5,18,12,12,10,20,20,7,5)
```

[1] لاحظ أن تنفيذ ذلك يستدعي حذف الخيار space من الأمر السابق.

سنقوم الآن برسم شكل الانتشار لهذين المتغيرين مع استخدام الخيار pch لتغيير شكل النقاط داخل الرسم والخيار cex لتغيير حجم هذه النقاط وذلك باستخدام دالة الرسم plot؛

```
> plot(X1,X2,pch=17,cex=1.5)
```

والنتيجة ستكون ظهور الشكل (م7.2) التالي، علما بأن القيم الافتراضية للخيارين السابقين هي pch=1 و cex=1، وسنتعرف على المزيد من طرق استخدام هذين الخيارين لاحقا في هذا البند.



شكل م7.2: شكل الانتشار للمتغيرين X1 وX2 باستخدام pch=17 وcex=1.5

الآن لنفرض أننا نرغب بإضافة نقاط جديدة على الرسم عن طريق إضافة بيانات جديدة، (هي newX1 و newX2 مثلا)، إلى البيانات الأصلية وإعادة رسم شكل الانتشار الجديد للمقارنة؛

```
> newX1<-c(30,24,30,27,35)
> newX2<-c(20,24,30,30,21)
```

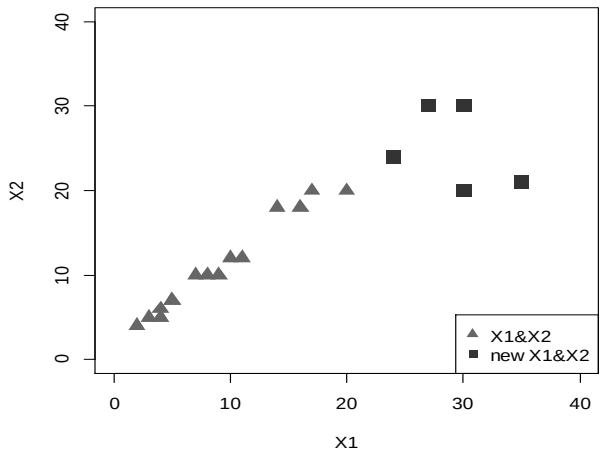
عندئذ نستخدم دالة إدراج النقاط points لتنفيذ ذلك على الشكل السابق، إلا أنه من المهم دائما التأكد من مدى قيم البيانات الجديدة فإذا كان مداها مختلف عن مدى البيانات الأصلية فيجب عندها إعادة تغيير مدى المحور السيني والصادي على الرسم بما يتناسب مع مدى البيانات الجديدة، وإلا فإن بعض أو أحيانا كل النقاط الجديدة قد لا تظهر على الرسم. وسنقوم أيضا بتغيير شكل النقاط الجديدة ولونها. وحيث أن مدى البيانات الجديدة newX1 و newX2 يتجاوز مدى البيانات الأصلية فيجب تغيير مدى محاور الرسم باستخدام الخيارين xlim وylim كما يلي:

```
> plot(X1,X2,pch=17,cex=1.5,col="grey40",xlim=c(0,40)
,ylim=c(0,40))

> points(x=newX1,y=newX2,pch=15,col="grey20",cex=1.7)
```

ولنقم أيضا بإضافة دليل تفسيري على الرسم بالخيارات التالية:

```
> legend(x="bottomright",legend=c("X1&X2","new X1&X2")
,pch=c(17,15),col=c("grey40","grey20"))
```

فنحصل على الشكل (م8.2) التالي الذي تظهر فيه نقاط الانتشار للبيانات الأصلية والمضافة[1]:



شكل م8.2: شكل الانتشار للمتغيرين X1 وX2 بعد إضافة البيانات الجديدة newX1 وnewX2

من جديد، لنقم بتوفيق خط انحدار بسيط لكل من البيانات الأصلية والجديدة وتمثيل هذين الخطين بيانيا، ولنبدأ بتوفيق نموذجي الانحدار الخطي؛

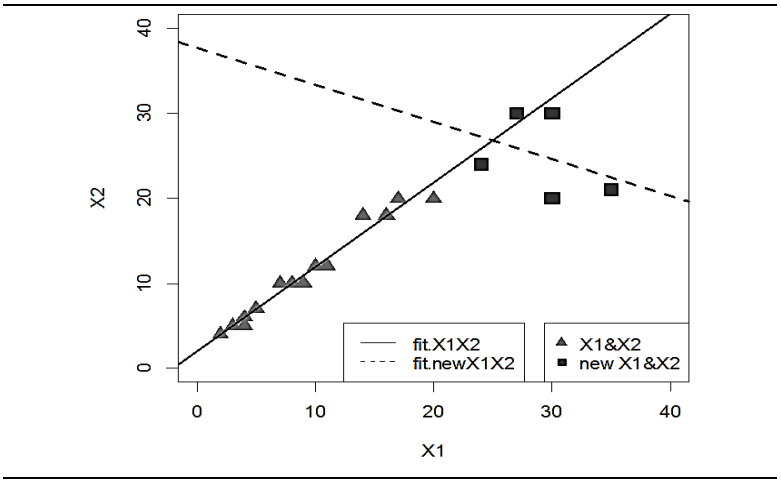
```
> reg.X1X2<-lm(X2~X1)
> reg.newX1X2<-lm(newX2~newX1)
```

بعدها يتم تمثيل معدلات الانحدار على الرسم باستخدام الخيار abline، مع إضافة دليل تفسيري جديد لتمييز الخطوط المستقيمة تلك المعادلات، كالتالي:

```
> abline(reg=reg.X1X2,lty=1,lwd=2)
> abline(reg=reg.newX1X2,lty=2,lwd=2)
> legend(x="bottom",legend=c("fit.X1X2","fit.newX1X2"),
lty=c(1,2))
```

فنحصل على الشكل (م9.2) التالي الذي يمثل شكل الانتشار لمجموعتي البيانات مع خطي الانحدار لها:

[1] في الشكل (م8.2) كان يكفي تغيير شكل النقاط أو لونها للبيانات الأصلية والجديدة للتمييز بينها، إلا أننا قمنا بتغير الاثنين معا لغرض توضيح استخدام خيارات الرسم الإضافية.



شكل م9.2: شكل الانتشار لمجموعتي البيانات (X1 و X2) و (newX1 و newX2) مع خطوط الانحدار الموفقة لهما

ولاحظ أنه قد تم استخدام كل من الخيار lty لتعديل نوع أو طبيعة الخط المستقيم، و الخيار lwd لتعديل سُمك تلك الخطوط، ويمكنك استخدام help(lines) للمزيد من المعلومات حول هذين الخيارين.

م3.2 معالم التمثيل البياني ودالة par (Graphical Parameters and **par** Function)

سنقوم هنا بتسليط الضوء على بعض أدوات وخواص الرسم التي قد تُستخدَم ضمن دوال التمثيل البياني الرياضي والإحصائي أو ضمن دالة معالم الرسم par التي قمنا باستخدامها في أكثر من موضع في بعض فصول الكتاب.

وتشمل دالة par في الواقع العديد من خيارات الرسم البياني المتنوعة والتي تهتم بأدق تفاصيل الرسم، فخيارات تلك الدالة تساعد في التحكم في نوع وسمك محاور الشكل البياني والخطوط المستقيمة داخله، والتحكم في حجم ونوع ولون الخط للنصوص والرموز المستخدَمة، وكذلك تحديد طرق حساب وعرض التقسيمات على محاور الرسم، وتحديد حجم الرسم العام و والمساحة الداخلية، وغير ذلك من التفاصيل، إضافة إلى إمكانية التعرف على القيم الافتراضية للخيارات المختلفة عن طريق تنفيذ الأمر par أو كتابة اسم الخيار أو الخيارات المطلوبة كمتجه بين الأقواس، فمثلا يمكن كتابة par(c("cex","col")) للتعرف على القيم الافتراضية للخيارين cex و col.

وتجدر الإشارة هنا إلى أن معظم هذه الخيارات التي سيتم تناولها في هذا البند يمكن تنفيذها أو تحديد قيمها من خلال تنفيذ الأمر على النحو التالي:

par(الخيار = قيمة الخيار)

أو يمكن ببساطة إدراجها ضمن دوال الرسم المستخدمة، وهذا ما سيتم التركيز عليه من خلال الأمثلة على هذه الخيارات لأن ذلك من الناحية العملية يُعتبر أفضل.

وتُقسم معالم الرسم عموما إلى ثلاثة أقسام هي أدوات التمثيل البياني، المحاور وما يتعلق بها، وهوامش الرسم البياني، وسنقوم بتقديم شرح موجز [1] لهذه المعالم فيما يلي مع عرض بعض الأمثلة عليها؛

1. أدوات التمثيل البياني:

- `pch`: وهي الأداة أو الخيار الخاص بنوع أو طبيعة نقاط الرسم (Points Character) حيث تتوفر عدة أشكال يمكن استخدامها لعرض النقاط مثل الدوائر والمربعات والنجوم وغيرها. ويمكن للمستخدم الاختيار من بين 26 نوع أساسي من هذه الأشكال إما بكتابة رقم الشكل المطلوب أو كتابة اسمه، والشكل الافتراضي للنقاط هو `pch=1`، وللتعرف على كل هذه الأشكال يمكن مثلا استخدام الأمر؛

```
> plot(0:25,0:25,pch=0:25,lab=c(25,25,25))
```

أو يمكن للمستخدم كتابة أي حرف أو رقم أو رمز مُعرف بين القوسين "" ليتم اعتباره شكل للنقاط، فمثلا يمكن تنفيذ شكل انتشار افتراضي تكون فيه النقاط هي الحرف "R" باستخدام الأمر:

```
> plot(runif(50),pch="R")
```

- `lty`: وهو الخيار الخاص بنوع الخطوط المستقيمة داخل الرسم (Line Type)، وتوجد 6 أشكال للخطوط المستقيمة يمكن التعرف عليها من خلال تنفيذ سطري الأوامر التاليين مثلا؛

```
> plot(0:6,0:6,pch="",lab=c(3,7,7))
> for(i in 0:6)abline(h=i,lty=i)
```

علما بأن الخيار `lty=0` يعني رسم خط مستقيم غير منظور، والنوع الافتراضي هو `lty=1`.

- `lwd`: ويُستخدم لتحديد سمك الخط المستقيم (Line Width)، ويمكن لهذا الخيار أن يأخذ أي قيمة موجبة، (وليس بالضرورة أن تكون عدد صحيح)، إلا أن القيم العشرية ما بين الأعداد الصحيحة قد لا تُظهر اختلافا واضحا على الرسم، ويمكن تنفيذ الأوامر التالية للتوضيح؛

```
> plot(1:7,1:7,pch="",lab=c(3,7,7))
> for(i in c(1,1.5,2,3,4,4.5,5,6,7))abline(h=i,lwd=i)
```

علما بأنه نظريا لا توجد قيمة عليا لسُمك الخط.

- `col`: وهي أداة عامة تُستخدم لتحديد الألوان (Colors) الخاصة بنقاط الرسم، الخطوط المستقيمة، العناوين والنصوص، وغيرها من ملامح الرسم. ويمكنك كتابة `colors()` للتّعرف على أسماء هذه الألوان وعددها 657 لون، وكتابة `demo("colors")` لرؤية هذه الألوان.

[1] يمكن للقارئ دائما الرجوع للمساعدة `help(par)` للمزيد من المعلومات حول معالم التمثيل البياني.

ويمكن استخدام الأرقام أو الأسماء مع الخيار col، مع ملاحظة أن الألوان الأساسية المتاحة مع الأرقام هي 8 ألوان[1]، أما لاستخدام المزيد من الألوان فيمكن كتابة اسم اللون المطلوب.

إن استخدام خيار اللون ضمن دالة الرسم مباشرة يؤدي في معظم الحالات إلى تعديل لون النقاط أو الخطوط المستقيمة أو الشكل الداخلي في الرسم دون تغيير لون القيم على المحاور أو العناوين أو غيرها من مكونات الرسم، ويمكن تنفيذ الأمثلة التالية للمزيد من التوضيح؛

```
> plot(rnorm(100),rnorm(100),pch=16,col=2)
```

```
> hist(rnorm(100),col=2)
```

```
> boxplot(rnorm(100),col=2)
```

أما لتغيير لون قيم المحاور، لون عناوين المحاور، ولون العنوان الرئيسي والفرعي فيتم استخدام الخيارات col.axis، col.lab، col.main، وcol.sub على الترتيب. ويمكنك تجربة المثال التالي:

```
> plot(rnorm(100),rnorm(100),main="Scatter plot",sub="A
colorful example",pch=16,col=2,col.axis="navyblue",
col.lab="magenta3",col.main="orange2",col.sub=8)
```

- font: وهو الخيار الخاص بنوع خط النصوص المستخدَمة في الشكل، وتوجد 4 أنواع أساسية من الخطوط هي الخط العادي (plain)، الخط الغامق (**Bold**)، الخط المائل (*Italic*)، والخط الغامق المائل (***Bold Italic***)، وهذه الأنواع يُعبر عنها بالأرقام من 1 إلى 4 على الترتيب.

ويمكن استخدام المزيد من أنواع الخطوط اعتمادا على عدد الخطوط المعرفة في جهاز الحاسوب، لذلك يمكنك تجربة اختيار الأرقام بشكل متسلسل من الرقم 5 فصاعدا ومراقبة التغير في نوع الخط بحسب أنواع الخطوط المُعرفة في جهازك، مع العلم بأن نظام R قد لا يتقبل أو يعرض بعض الخطوط المتوفرة في جهازك وخاصة بعض خطوط اللغة العربية، وعموما لمعرفة الخطوط المتوفرة يمكنك استدعاؤها بالأمر windowsFonts().

واستخدام خيار نوع الخط font بشكل مباشر مع دالة الرسم يؤدي في معظم الحالات إلى تغيير نوع الخط الخاص بقيم المحاور، وأحيانا قد لا يتم تغيير الخط على الإطلاق اعتمادا على نوع الرسم، ويمكن للقارئ تنفيذ الأمثلة التالية:

```
> hist(rnorm(50),font=3)
```

[1] للتعرف على هذه الألوان الثمانية يمكن استخدام الدالة:

```
> palette()
[1]"black" "red" "green3" "blue" "cyan" "magenta" "yellow" "gray"
```

```
> plot(runif(50),font=4)

> boxplot(runif(50),font=3)
```

ويمكن، كما هو الحال مع دالة اللون، استخدام الخيارات font.axis، font.lab، font.main وfont.sub لتغيير نوع الخط في قيم المحاور، عناوين المحاور، وخط العنوان الرئيسي والفرعي على الترتيب. ويمكن تنفيذ المثال التالي للمقارنة:

```
> plot(rnorm(100),rnorm(100),main="Scatter plot",sub=
"Different fonts",font.axis=3,font.lab=1,font.main=15,
font.sub=4)
```

- adj: ويُستخدم لتعديل موضع النصوص على الرسم، حيث يمكن تغيير موضع النص من أقصى اليسار، (adj=0)، إلى أقصى اليمين، (adj=1)، أي أنه بإعطاء قيم لهذا الخيار ما بين 0 و1 يتغير موضع النصوص، بما فيها العنوان الرئيسي والفرعي وعناوين المحاور، علما بأن القيمة الافتراضية هي adj=0.5 في المنتصف. ويمكن تنفيذ الأمر التالي مرارا مع تغيير قيم adj في كل مرة ومراقبة التغير:

```
> plot(rnorm(100),adj=0.3,rnorm(100),main="Scatter plot")
```

- cex: وهو الخيار الخاص بحجم الخط المُستخدم في الرسم، وعند استخدامه بشكل مباشر ضمن دالة الرسم فإن التغير في الحجم سيشمل ما في داخل الرسم من نقاط، أما غير ذلك من خطوط مستقيمة أو أعمدة أو غيرها فلا يتأثر في الغالب. والقيمة الافتراضية لخيار الحجم هو cex=1، ويمكن تنفيذ المثال التالي لمراقبة التغير من خلال استخدام أكثر من قيمة للخيار:

```
> plot(s.grd1,s.grd2,cex=2,main="Scatter plot")
```

لكن بالنسبة لبعض الرسومات الأخرى فإن استخدام الخيار cex لا يكون له تأثير، ومن ضمن تلك الرسومات تلك التالي:

```
> hist(s.grd1,cex=0.5,main="Histogram")

> boxplot(s.grd1,cex=2.5,main="Boxplot")
```

أما لتغيير حجم الخط لقيم المحاور، عناوين المحاور، والعنوان الرئيسي والفرعي فيتم استخدام الخيارات cex.axis، cex.lab، وcex.main وcex.subعلى الترتيب. ويمكنك ملاحظة التغير في أحجام الخطوط، (إضافة للتغير في حجم النقاط داخل الرسم)، من خلال المثال التالي:

```
> plot(s.grd1,s.grd2,main="Scatter plot",sub="change in
font
sizes",cex.axis=1.4,cex.lab=1.5,cex.main=2.5,cex.sub=1.3
,cex=0.7)
```

2. المحاور وتقسيماتها:

● lab: ويتعلق بتغيير عدد تقسيمات المحاور (Tick Marks)، ويأخذ شكل متجه به ثلاثة قيم، القيمة الأولى والثانية هي لتعديل عدد تقسيمات المحور السيني والصادي على الترتيب، إلا أن هذا التعديل غالبا ما يكون تقريبيا، بمعنى أن المستخدم قد يحدد عدد 6 تقسيمات مثلا للمحور الأفقي أو العامودي ثم يجد أقل أو أكثر من ذلك العدد على الرسم لأن نظام R سيقوم بعرض أفضل تقسيم ملائم للمظهر العام.

أما القيمة الثالثة في متجه قيم الخيار lab فهي لتعديل طول المساحة المحجوزة لعنوان المحور، وهذه القيمة لا تتغير عادة إلا من خلال دالة par وقد تكون معقدة بعض الشيء، لذلك لن نتطرق إليها هنا ويمكن للقارئ العودة للمساعدة help(par) للمزيد من المعلومات حول هذه النقطة.

أما القيمة الافتراضية للخيار lab فهي lab=c(5,5,7)، ويمكن ملاحظة التغير في عدد تقسيمات المحاور من خلال المثال التالي:

```
> plot(X1,X2,lab=c(5,5,7))
> plot(X1,X2,lab=c(7,8,7))
```

حيث ستلاحظ أن عدد تقسيمات المحاور قد تغير ولكن ليس بنفس العدد الذي تم إدراجه.

● las: وهو الخيار الخاص بتغيير اتجاه قيم المحاور أفقيا أو عاموديا. وتوجد أربعة اختيارات هي 0 (القيمة الافتراضية)، 1، 2، و3. والمثال التالي يشمل كل تلك الحالات:

```
> plot(X1)
> plot(X1,las=1)
> plot(X1,las=2)
> plot(X1,las=3)
```

● mgp: ويُستخدم لتعديل المسافات المتعلقة بقيم وعناوين المحاور إضافة لإمكانية "إزاحة" المحاور من مكانها الأصلي. وهذه التعديلات تشمل زيادة أو إنقاص المسافة من صندوق الرسم إلى الخارج، (وذلك باستخدام القيم الموجبة)، أو إلى الداخل، (وذلك باستخدام القيم السالبة). علما بأن المسافات الافتراضية هي mgp=c(3,1,0). والأمثلة التالية هي لتوضيح مدى التغير في تلك المسافات خارج صندوق الرسم:

```
> plot(X1,X2)
> plot(X1,X2,mgp=c(2.5,1,0))
> plot(X1,X2,mgp=c(3,2,0))
> plot(X1,X2,mgp=c(3,1,1))
```

وأما الأمثلة التالية، فهي لتوضيح تغير المسافات في الاتجاه العكسي، أي بالاتجاه داخل صندوق الرسم، (مع ملاحظة ظهور رسائل تحذيرية تتبع هذا التغيير نظرا لخروجه عن النمط الطبيعي للرسم):

```
> plot(X1,X2,mgp=c(-1.5,1,0))
> plot(X1,X2,mgp=c(3,-2,0))
> plot(X1,X2,mgp=c(3,1,-1))
```

● tck: وهذا الخيار هو لتعديل طول مسافة تقسيمات المحاور. وباستخدام القيم السالبة يتم زيادة أو إنقاص طول التقسيمات خارج صندوق الرسم، أما القيم الموجبة فتفعل نفس الشيء ولكن داخل صندوق الرسم. علما بأن القيمة الافتراضية لطول مسافة التقسيم يتم حسابها كنسبة من حجم الرسم ككل، ويمكنك تجربة الحالات التالية على سبيل المثال:

```
> plot(X1,X2)
> plot(X1,X2,tck=-0.025)
> plot(X1,X2,tck=-0.05)
> plot(X1,X2,tck=0)
> plot(X1,X2,tck=0.03)
> plot(X1,X2,tck=1)
```

ولاحظ أن الحالة الأخيرة، (tck=1)، يتم فيها استخدام الطول الكامل للتقسيمات مما ينتج عنه شبكة التقسيمات الداخلية للرسم (Plot Grid).

● xaxs وyaxs: وهي خيارات عرض القيم على المحورين السيني والصادي على الترتيب، وتوجد طريقتين للعرض؛ الأولى هي الاعتيادية (Regular)، وهي الافتراضية، التي تقوم بعرض قيم المحاور الأكثر ملائمة للبيانات، وتأخذ هذه الخيارات الرموز "xaxs="r و"yaxs="r، والطريقة الثانية هي طريقة الفترات (Intervals) وتقوم بعرض قيم المحاور على هيئة فترات متساوية، وفيها تأخذ الخيارات الرموز "xaxs="i و"yaxs="i، إلا أن هذه الطريقة قد لا تناسب كل أنواع الرسومات أو قيم البيانات المعروضة.

ويمكن ملاحظة التغير في عرض قيم المحاور من خلال الأمثلة التالية:

```
> plot(X1,X2)
> plot(X1,X2,xaxs="r",yaxs="r")
> plot(X1,X2,xaxs="r",yaxs="i")
> plot(X1,X2,xaxs="i",yaxs="i")
```

وستلاحظ عدم ظهور بعض نقاط الرسم أو اختفاؤها عند استخدام طريقة الفترات للعرض.

3. هوامش الرسم البياني:

تختص خيارات الهوامش (Margins) بالمسافات الأربع، (يمين، يسار، أعلى، وأسفل)، والتي تفصل بين مساحة الرسم الداخلي وإطار الشكل العام، وتضم الخيارات الأساسية التالية؛

- `mai`: وهو الخيار الخاص بتحديد الهوامش الأربع بين الرسم الداخلي وإطار الشكل العام عن طريق تحديد قيم هذه الهوامش، (والمُعرفة وحداتها بالبوصة (Inch))، بالترتيب التالي؛

`mai=c(bottom,left,top,right).`

وللتعرف على القيم الافتراضية لهذا الخيار في جهازك يمكنك تنفيذ الأمر `par("mai")` فتظهر القيم الأربع لمسافات الهوامش بالبوصة. ولتغيير هذه الهوامش في رسم معين[1] يتم تعيين قيم الهوامش المطلوبة ثم تنفيذ الرسم كما يوضح المثال التالي:

```
> par(mai=c(2,1,1.3,0.5))
> plot(X1,X2,main="Scatter plot",sub="sub")
```

ولاحظ أنه بعد ظهور الرسم إذا ما قمت بتنفيذ أمر استدعاء قيم الهوامش `par("mai")` فإن قيم الهوامش التي ستحصل عليها هي تلك المستخدمة في الرسم. أما إذا قمت بإغلاق نافذة الرسم وتنفيذ أمر استدعاء الهوامش من جديد فستظهر القيم الافتراضية.

- `mar`: وهو خيار مشابه للخيار `mai` السابق إلا أن وحدة قياس الهامش فيه هي سطر النص (Text Line) وليس مقياس البوصة. ويمكن بالمثل تنفيذ الأمر `par("mar")` للتعرف على الهوامش الافتراضية. وكمثال على هذا استخدام هذا الخيار يمكن كتابة الأوامر:

```
> par(mar=c(8,7,7.5,4.2))
> plot(X1,X2,main="Scatter plot",sub="sub")
```

م4.2 إدراج عدة رسومات في شكل واحد (Multiple Figures Handling)

يمكنك في نظام R تكوين شكل أو إطار يحتوي على عدة رسومات بيانية مختلفة ويُعد ذلك عمليا في مجال المقارنة أو التسلسل في الإيضاح. ويمكن التعامل أو تعديل كل رسم داخل الشكل العام بصورة مستقلة بحيث يشمل ذلك هوامشه الخاصة.

ولتنفيذ الرسومات المتعددة (Multiple Figures)، الواحد تلو الآخر في ما يُشبه مصفوفة الرسم، يتم استخدام إحدى الخيارين `mfrow` أو `mfcol` ضمن دالة `par`، حيث أنه في الخيار الأول يتم إدراج الرسومات في الصف الأول، (ابتداء من اليسار)، فالثاني وهكذا، ويتم في الخيار الثاني إدراج الرسومات في العامود الأول،

[1] تغيير الهوامش بهذه الطريقة يتم للرسم الذي يتم تنفيذه بعد الأمر `par("mai")` فقط ولا تصبح هذه القيم هي قيم الهوامش الافتراضية.

(إلى اليسار)، فالعامود الثاني وهكذا. وهذا بالطبع عندما تحتوي مصفوفة الشكل العام على أكثر من صف أو عامود.

وعلى سبيل المثال، يمكن تكوين شكل بياني يضم ستة رسومات على هيئة مصفوفة رسم لها الترتيب 2×3 بالصورة التالية، (علما بأنه يمكن تنفيذ سطور الأوامر الواحد تلو الآخر لغرض مراقبة ظهور الرسومات بالترتيب داخل الشكل العام):

```
> par(mfrow=c(3,2))
> plot(X1,main="Scatter plot of X1")
> plot(X2,main="Scatter plot of X2")
> boxplot(X1,main="Boxplot of X1",xlab="X1")
> boxplot(X2,main="Boxplot of X2",xlab="X2")
> hist(X1)
> hist(X2)
```

ولاحظ أن ترتيب الرسومات الستة سوف يتغير داخل الشكل العام إذا ما تم استخدام الخيار `par(mfcol=c(3,2))` مع مجموعة الرسومات السابقة.

ولتغيير[1] مسافات الهوامش الخارجية (Outer Margins) للرسومات يتم استخدام الخيارين `oma` أو `omi`، حيث يستخدم الأول سطر النص كوحدة قياس للهامش، ويستخدم الثاني البوصة. ولمراقبة الفرق بين استخدام خياري الهوامش `mar` و`mai` وخياري الهوامش الخارجية `oma` و`omi` يمكنك تنفيذ الأمثلة الثلاثة التالية؛ (الأول يتضمن استخدام الهوامش والهوامش الخارجية الافتراضية، الثاني يتضمن استخدام هوامش معدلة، أما المثال الثالث فيتضمن استخدام هوامش خارجية معدلة):

.1

```
par(mfcol=c(3,2))
plot(X1,main="Scatter plot of X1")
> plot(X2,main="Scatter plot of X2")
> boxplot(X1,main="Boxplot of X1",xlab="X1")
> boxplot(X2,main="Boxplot of X2",xlab="X2")
> hist(X1)
> hist(X2)
> par("mar")
> par("mai")
> par("oma")
> par("omi")
```

[1] يُقصد بتغيير الهوامش الخارجية زيادة قيمتها لأن الأشكال البيانية بصورتها الافتراضية لا توجد بها هوامش خارجية، بمعنى أن مسافات الهوامش الخارجية تساوي الصفر.

.2

```
> par(mfcol=c(3,2),mar=c(4.5,7.5,5.3,4.2))
> plot(X1,main="Scatter plot of X1")
> plot(X2,main="Scatter plot of X2")
> boxplot(X1,main="Boxplot of X1",xlab="X1")
> boxplot(X2,main="Boxplot of X2",xlab="X2")
> hist(X1)
> hist(X2)
> par("mar")
> par("mai")
> par("oma")
> par("omi")
```

.3

```
> par(mfcol=c(3,2),oma=c(4.5,7.5,5.3,4.2))
> plot(X1,main="Scatter plot of X1")
> plot(X2,main="Scatter plot of X2")
> boxplot(X1,main="Boxplot of X1",xlab="X1")
> boxplot(X2,main="Boxplot of X2",xlab="X2")
> hist(X1)
> hist(X2)
> par("mar")
> par("mai")
> par("oma")
> par("omi")
```

تناولنا فيما سبق استخدام الدالة text لإضافة النصوص داخل المساحة الداخلية للرسم البياني، ونضيف هنا إلى أنه يمكن أيضا استخدام الدالة[1] mtext، بعد تنفيذ الرسم أو الرسومات البيانية، لإضافة نص أو أكثر على هامش أو أكثر من هوامش الرسم، (أو حتى على الهوامش الخارجية للشكل بحسب المسافة التي تم تحديدها باستخدام إحدى الخيارين oma أو omi)، حيث يتم تحديد الجهة أو الجهات المطلوب عرض النص ضمنها من خلال استخدام الخيار side كما يوضح المثال التالي:

```
> par(mfcol=c(3,2),oma=c(2,2,2,2))
> plot(X1,main="Scatter plot of X1")
> plot(X2,main="Scatter plot of X2")
> boxplot(X1,main="Boxplot of X1",xlab="X1")
> boxplot(X2,main="Boxplot of X2",xlab="X2")
> hist(X1)
> hist(X2)
> mtext("These are six plots",side=1,outer=T,font=4)
> mtext("These are six plots",side=2,outer=T,font=4)
> mtext("These are six plots",side=3,outer=T,font=4)
> mtext("These are six plots",side=4,outer=T,font=4)
```

[1] يمكن استخدام الدالة mtext مع الأشكال البيانية التي تضم رسم واحد أو عدة رسومات على حد السواء.

م5.2 التعامل التفاعلي مع التمثيل البياني (Interactive Handling with Graphics)

يُقصد بالتعامل التفاعلي مع التمثيل البياني إمكانية استخلاص المعلومات أو إضافة بيانات أو نصوص إلى تلك الرسومات، فمثلا يمكن معرفة إحداثيات نقاط أو أشكال معينة تقع على الرسم، وإمكانية تمييز بعض النقاط أو المواقع على الرسم، وكذلك إضافة رموز أو نصوص في أي مكان على الرسم اعتمادا على الاختيار البصري للمستخدم، وغير ذلك من التعاملات التي تسمى تفاعلية لأنها تعتمد على تعامل المُستخدِم في الوقت الفعلي مع الرسم.

وأهم وأبسط دالة تفاعلية مع الرسم البياني في R هي الدالة التفاعلية `locator`، وسنتناول فيما يلي أهم استخداماتها؛

- `locator()`: يُستخدم هذا الأمر بصيغته المباشرة بعد أي دالة تمثيل بياني للحصول على موقع أو إحداثيات لأي نقطة على الرسم، وهذا قد يكون مفيدا في كثير من الأحيان للتعرف على إحداثيات بعض النقاط على الرسم مثل القيم المتطرفة، أو لتحديد إحداثيات موضع إدراج نص ما مسبقا، وهكذا.

ولاستخدام هذا الأمر، قم أولا بتنفيذ الرسم البياني الذي ترغب بالتعامل معه، ثم نفذ الدالة `locator()` وقم بتمرير مؤشر الفأرة على الرسم فستلاحظ ظهور إشارة زائد "+" بدلا من المؤشر الاعتيادي للفأرة، يمكنك الآن استخدام الزر الإيسر للفأرة لاختيار أي موضع (أو نقطة) أو أكثر على الرسم، وعندما ترغب في إنهاء هذه العملية فإنك تقوم بالضغط على الزر الأيمن للفأرة، (بدون الابتعاد عن منطقة الرسم)، فيظهر لك خيارين الأول هو (Stop) لإيقاف العملية، والثاني هو (Continue) للاستمرار بعملية الاختيار. بعد الإيقاف ستلاحظ ظهور نتيجة اختيار المواضع على الرسم ممثلة بمتجهين على هيئة قائمة يمثلان احداثيات النقاط التي قمت باختيارها في لوحة مراقبة R. والمثال التالي يوضح الفكرة؛

قم بتنفيذ السطرين:

```
> plot(X1,X2)
> locator()
```

فسيظهر شكل الانتشار للمتغيرين `X1` و`X2`، قم الآن باختيار أي عدد ترغب به من النقاط على الرسم، ثم اختر إيقاف كما وضحنا أعلاه وستلاحظ ظهور إحداثيات تلك النقاط في النتيجة.

- `locator(n,type)`: يُعد هذا الأمر امتدادا للأمر السابق، حيث يتم من خلاله تحديد عدد المواضع، (المطلوب معرفة إحداثياتها أو إضافة نقاط عليها)، مسبقا باستخدام القيمة `n`، وتحديد كيفية الإضافة باستخدام الخيار `type`. والخيارات المتوفرة لـ `type` هي:

`"n"`؛ والتي تتيح للمستخدم تحديد إحداثيات المواضع بدون رسم على الشكل الأصلي.

"p"؛ لإضافة نقاط فعلية إلى الشكل.

"l"؛ لتحديد مواضع على الشكل وتوصيلها على الترتيب بخطوط مستقيمة.

و"o"؛ لإضافة نقاط فعلية على الشكل وتوصيلها على الترتيب بخطوط مستقيمة.

ويمكن عموما استخدام خيارات إضافية خاصة بالرسم، (مثل تغيير شكل النقاط المضافة أو نوع وسماكة الخطوط المستقيمة المضافة وغير ذلك من التعديلات)، داخل دالة locator كما ستلاحظ عند تنفيذ الأمثلة التالية[1]:

```
> plot(X1,X2)
> locator(n=5,type="n")
> locator(n=5,type="p")
> locator(n=4,type="l")
> locator(n=3,type="o")
> locator(n=6,type="p",pch=8)
> locator(n=7,type="o",lwd=2.5)
```

● إضافة النصوص على الرسم باستخدام دالة locator: يمكن إضافة النصوص على الرسم باستخدام دالة locator ضمن دالة كتابة النصوص text، والمثال التالي يوضح إضافة نصين مختلفين على الشكل البياني:

قم بتنفيذ الآتي:

```
> plot(X1)
> text(locator(n=1),"extreme value 1",pos=4,font=4)
```

الآن قم بالنقر على النقطة الظاهرة في أسفل الشكل إلى اليسار، وستلاحظ على الشكل إضافة النص "extreme value 1" إلى يمين النقطة. من جديد قم بتنفيذ:

```
> text(locator(n=1),"extreme value 2",pos=2,font=4)
```

وقم بالضغط على النقطة الظاهرة في أعلى الشكل إلى اليمين فيتم إضافة النص الثاني وهو " extreme value 2" إلى يسار النقطة.

[1] لاحظ أنه لابد من تنفيذ كل دالة تفاعلية locator على حده حتى انتهاء عملها ثم البدء بتنفيذ الأمر التالي، كذلك يمكنك تنفيذ كل الدوال التفاعلية على نفس الشكل أو تنفيذ الرسم من جديد ثم تنفيذ الدالة التفاعلية.

● الدالة identify: وهي من الدوال التفاعلية مع الرسومات البيانية أيضا، ولها عدة استخدامات أهمها ما يلي؛

تحديد الترتيب التسلسلي للنقاط على الرسم عن طريق النقر عليها، فمثلا إذا ما تم تنفيذ التالي:

```
> X1;X2
> plot(X1,X2)
> identify(X1,X2)
```

عندها يمكنك الضغط على أي من النقاط على الرسم فيظهر ترتيبها التسلسلي بجانبها، ويمكنك إنهاء العملية بالنقر على الزر الأيمن للفأرة واختيار إيقاف.

استخدام آخر لدالة identify هو إظهار قيمة متغير ما غير مُمَثل بيانيا باستخدام قيم متغير آخر مُمَثل على الرسم، فمثلا عند تنفيذ الأوامر التالية:

```
> plot(s.age)
> identify(s.age,labels=s.gen)
```

فإنه بالنقر على أي نقطة على الرسم، (والتي تمثل انتشار قيم المتغير s.age)، ستظهر القيم المناظرة له في المتغير s.gen، وهي نوع الطالب، إلى أن يتم إيقاف العملية. وكمثال آخر لنفس النمط من العمليات يمكنك تنفيذ التالي:

```
> plot(s.age)
> identify(s.age,labels=s.sem)
```

حيث سيمكنك ذلك من الحصول على ترتيب الفصل الدراسي المناظر لعمر الطالب.

ويمكن توسيع استخدام هذه العملية لتشمل إظهار قيم متغير ما باستخدام شكل الانتشار لمتغيرين، والذي يمكن تنفيذه كالتالي:

```
> plot(s.grd1,s.age)
> identify(s.grd1,s.age,labels=s.gen)
```

ملحق 3: دوال وأوامر لغة R المُستخدَمة في الكتاب

(R Commands used in the Book)

%*%	Chisq (ضمن الحزمة distr)	drop1
%in%	chisq.test	dt
%o%	class	dunif
AbscontDistribution (ضمن الحزمة distr)	coef	edit(data.frame())
add1	colnames	exp
addmargins	confint	Exp (ضمن الحزمة distr)
addrv	cor	factor
aggregate	cor.test	f
all	cos	Fd (ضمن الحزمة distr)
anova	cov	Fitdistr (ضمن الحزمة MASS)
aov	createSheet	fitted
apply	cumsum	fivenum
array	cut	fix
as.data.frame	data.frame	floor
as.integer	dbinom	for
assign	dchisq	Fd (ضمن الحزمة distr)
Barplot	density	Fitdistr (ضمن الحزمة MASS)
bartlett.test	det	fitted
beta	df	fivenum
Beta (ضمن الحزمة distr)	dgeom	fix
binom	dhyper	floor
Binom (ضمن الحزمة distr)	diag	for
Boot (ضمن الحزمة boot)	diff	function
boxplot	dim	gamma
by	DiscreteDistribution	Gammad (ضمن الحزمة distr)
c( )	dmultinom	geom
cat	dnbinom	Geom (ضمن الحزمة distr)
cbind	dnorm	getwd()
ceiling	dotchart	head
chisq	dpois	help(reserved)

hist	Mgfunif (ضمن الحزمة actuar)	pgamma
history()	min	pgeom
hyper	Mnorm (ضمن الحزمة actuar)	phyper
Hyper (ضمن الحزمة distr)	mode	Pie
if	Mgfunif (ضمن الحزمة actuar)	plot
if/else	min	pmax
install.packages()	Mnorm (ضمن الحزمة actuar)	pmin
integrate	mode	pnbinom
intersect	Moment (ضمن الحزمة e1071)	pnorm
IQR	multinom	pois
is.na	Munif (ضمن الحزمة actuar)	Pois (ضمن الحزمة distr)
isin	names	ppois
ks.test	nbinom	predict
Kurtosis	Nbinom (ضمن الحزمة distr)	print
lapply	ncol	Prob
length	norm	prob.test
LETTERS	Norm (ضمن الحزمة distr)	probspace
Letters	nrow	prod
library()	nsamp	prop.table
library(help=" ")	objects()	pt
lines	oneway.test	punif
list	order	qchisq
lm	outer	qf
loadhistory	pairs	qnorm
loadWorkbook	par	qqline
log	paste	qqnorm
log10	pbeta	qt
ls()	pbinom	quantile
matrix	pchisq	range
max	Pcor (ضمن الحزمة ppcor)	rbind
mean	pcor.test (ضمن الحزمة ppcor)	rchisq
median	pexp	read.table
Mgfnorm (ضمن الحزمة actuar)	pf	readWorksheetFromFile

rep	sort	which
replicate	sort.list	while
resid	source	write.csv
rev	split	write.csv2
rf	sqrt	writeWorksheet
rgeom	stem	z.test (ضمن الحزمة TeachingDemos)
rhyper	stem.leaf	
rm()	step	
rmultinom	str	
rnbinom	stripchart	
rnorm	subset	
rolldie	subset	
round	sum	
round	sum	
rownames	Summary	
rpois	switch	
rt	t	
runif	t.test	
sample	tail	
sapply	tan	
save.image	tapply	
savehistory	Td	
saveWorkbook	tosscoin	
sd	transform	
segments	trunk	
seq	unif	
setdiff	Unif (ضمن الحزمة distr)	
setwd	union	
Shapiro.test	unlist	
sigma.test (ضمن الحزمة TeachingDemos)	update	
signif	urnsamples	
sin	var	
Skewness	var.test	
solve	warning	

ملحق 4: حِزم R الإضافية المُستخدَمة في الكتاب

(Additional R Packages used in the Book)

الفصل الأول

```
rJava
XLConnect
XLConnectJars
```

الفصل الرابع

```
modeest
e1071
aplpack
```

الفصل الخامس

```
Prob
distrEx
distr
actuar
```

الفصل السادس

```
TeachingDemos
MASS
ppcor
```

الفصل السابع

```
boot
```

المراجع
(References)

1. Bluman, A., (2005), *Probability Demystified,* The McGraw-Hill Companies, Inc. U.S.A.
2. Brink, D., (2008), *Statistics*, David Brink and Ventus Publishing ApS. U.S.A.
3. Crawley, M., (2005): *Statistics: An Introduction using R*. John Wiley and Sons, Inc. U.K.
4. Dalgaard, P., (2008): *Introductory Statistics with R*. Springer Science and Business Media, LLC. U.S.A.
5. Douglas, M. and George, R., (2003), *Applied Statistics and Probability for Engineers,* John Wiley and Sons, Inc. U.K.
6. Fernandes, M., (2009), *Statistics for Business and Economics,* Marcelo Fernandes and Ventus Publishing. U.S.A.
7. Frank, H., and Althoen, S., (1994), *Statistics: Concepts and Applications,* Cambridge University Press. U.K.
8. Han, J. and Kamber, M., (2000), *Data Mining: Concepts and Techniques,* Morgan Kaufmann Publishers. U.S.A.
9. Kerns, G., (2010): *Introduction to Probability and Statistics Using R*. Free Software Foundation, Inc. GNU Free Documentation License.
10. Knell, R., (2014): *Introductory R*. Robert Knell, University of London. U.K.
11. Moore, D., (2003), *The Basic Practice of Statistics,* W. H. Freeman Publishers. U.S.A.
12. Quick, J., (2010): *Statistical Analysis with R*. Packt Publishing Ltd. U.K.
13. Spiegel, M., Schiller, J. and Srinivasan, R., (2001), *Schaum's Easy Outline of Probability and Statistics,* The McGraw-Hill Companies, Inc. U.S.A.
14. Stephens, L., (2006), *Schaum's Outline Beginning Statistics,* The McGraw-Hill Companies, Inc. U.S.A.
15. Stowell, S., (2013): *Using R for Statistics*. Friendsof-Apress. U.S.A.
16. Torgo, L., (2011): *Data Mining with R*. Taylor and Francis Group, LLC. U.S.A.
17. Venables, W., Smith, D., and R Core Team, (2015): *An Introduction to R*. The R Foundation for Statistical Computing. U.S.A.
18. Verzani, J., (2005): *Using R for Introductory Statistics*. Chapman and Hall/CRC Press. U.S.A.
19. Vries, A., and Meys, J., (2012): *R for Dummies*. John Wiley and Sons, Inc. U.K.
20. Wackerly, D., Mendenhall, W., and Scheaffer, R., (2002), *Mathematical Statistics with Applications*. Duxbury Thomson Learning, Inc. U.S.A.

في هذا الكتاب

بعد استعراض وتطبيق أساليب التحليل الإحصائي الحديث باستخدام لغة البرمجة الرياضية الشهيرة R والتي تتميز أوامرها بالقوة والمرونة إلى جانب البساطة في صياغتها . حيث تكون البداية بتعريف نظام R وتوضيح طرق إدخال البيانات باستخدام المقاييس الإحصائية المناسبة والتمثيل البياني لها ، إضافة لتطبيق نظرية الاحتمال وسحب العينات والتوزيعات الاحتمالية المتصلة والمنفصلة ، ثم تناول استخدام الأساليب والنماذج الإحصائية المخصصة لتوفيق البيانات حسب طبيعة الدراسة . وإنتهاء بتطبيق بعض الأساليب الإحصائية المتقدمة وتوضيح طرق كتابة دوال المستخدم التي تطلق العنان للقارئ لتصميم دواله الخاصة التي يرغب بها

طباعة وتصميم
مركز القيصوم