

مقدمه‌ای بر نرم‌افزار آماری R

مؤلفان:

نادر نجاری - رحیم محمودوند

سرشناسه : نجاری، نادر، ۱۳۵۸.
 عنوان و نام پدیدآور : مقدمه‌ای بر نرم‌افزار آماری R / مولفان نادر نجاری، رحیم محمودوند.
 مشخصات نشر : تهران: صانعی شه میرزادی، ۱۳۹۰.
 مشخصات ظاهری : ۱۲۴ ص.: مصور، جدول، نمودار.
 شابک : ۹۷۸-۹۶۴-۷۳۲۰-۸۲-۲
 وضعیت فهرست‌نویسی: فیپا
 موضوع : آمار ریاضی -- داده‌پردازی
 شناسه افزوده : محمودوند، رحیم، ۱۳۸۵.
 رده‌بندی کنگره : QA۲۷۶/۲/ن۳م۷ ۱۳۹۰
 رده‌بندی دیویی : ۵۱۹/۵۰۲۸۵
 شماره کتابشناسی ملی : ۲۲۷۰۱۲۷



انتشارات صانعی شه میرزادی

مقدمه‌ای بر نرم‌افزار آماری R

مؤلفین: نادر نجاری، رحیم محمودوند

ناظر فنی و مدیر اجرایی: محسن صانعی

نوبت و تاریخ چاپ: اول / ۱۳۹۰

تیراژ: ۱۰۰۰

چاپخانه: ستاره سبز

صحافی: کیمیا

قیمت: ۳۰.۰۰۰ ریال

کلیه حقوق برای ناشر محفوظ است.

آدرس فروشگاه: تهران، خ انقلاب، بین ۱۲ فروردین و اردیبهشت، فروشگاه انتشارات صانعی، پلاک ۱۳۱۲

تلفن: ۶۶۴۰۹۹۲۴-۶۶۴۰۵۳۸۵ فاکس: ۶۶۴۶۲۱۷۷

آدرس دفتر نشر و پخش: تهران، خ انقلاب، خ ۱۲ فروردین، خ وحید نظری، پلاک ۱۰۵، واحد ۱

تلفن: ۶۶۴۱۲۳۹۵ فاکس: ۶۶۴۰۸۱۶۵

فهرست

پیش گفتار.....	۹
مقدمه.....	۱۱
آشنایی اولیه با R	۱۲
فصل اول: ساختار داده‌ها و محاسبات روی آنها.....	۱۷
۱-۱ مقدمه.....	۱۷
۲-۱ انواع داده‌ها در R	۱۸
۱-۲-۱ بردار.....	۱۸
۲-۲-۱ ماتریس.....	۲۱
۳-۲-۱ آرایه.....	۲۵
۴-۲-۱ فهرست.....	۲۵
۵-۲-۱ سری زمانی.....	۲۶
۳-۱ چارچوب اطلاعاتی.....	۲۷
۴-۱ ذخیره و بازخوانی داده‌ها.....	۲۷
فصل دوم: محاسبات ریاضی.....	۲۹
۱-۲ عملگرهای حسابی و منطقی.....	۲۹
۲-۲ مشتق.....	۳۲
۳-۲ ریشه‌های چندجمله‌ای.....	۳۲
۴-۲ دستگاه معادلات خطی.....	۳۳

فصل سوم: رسم نمودار	۳۵
۱-۳ رسم نمودارهای ساده و نحوه بهبودنمایش نمودار	۳۵
۲-۳ نمودارهای آماری	۴۲
فصل چهارم: توزیع‌های احتمال و تولید داده‌های تصادفی از آنها	۵۱
۱-۴ احتمال و رسم نمودار توزیع داده‌ها	۵۱
۲-۴ نمودارهای چندک-چندک	۵۵
۳-۴ تولید نمونه تصادفی	۵۸
فصل پنجم: آزمون‌های آماری	۵۹
۱-۵ آزمون‌های نیکویی برازش	۵۹
۲-۵ آزمون برابری میانگین با یک مقدار مفروض و آزمون برابری میانگین‌های دو جامعه	۶۱
۳-۵ آزمون برابری واریانس ها	۶۴
۴-۵ آزمون ضریب همبستگی	۶۵
۵-۵ آزمون برابری نسبت جامعه با یک مقدار مفروض (آزمون دقیق دو جمله‌ای)	۶۶
۶-۵ آزمون نسبت تک نمونه‌ای و برابری نسبت‌های دو یا چند جامعه	۶۷
۷-۵ جداول پیش‌بینی و آزمونهای استقلال	۶۸
۱-۷-۵ آزمون استقلال خی‌دو	۶۸
۲-۷-۵ آزمون استقلال فیشر	۷۰
۳-۷-۵ آزمون استقلال مانتل-هنزل	۷۰
فصل ششم: رگرسیون	۷۳
۱-۶ برازش مدل	۷۳
۲-۶ بررسی مناسبیت مدل	۷۵

فصل هفتم: طرح آزمایش‌ها و تحلیل واریانس	۷۹
۱-۷ وارد کردن داده‌ها	۷۹
۲-۷ بررسی داده‌ها	۸۰
۳-۷ تحلیل آماری طرح	۸۲
فصل هشتم: سری زمانی	۸۵
۱-۸ ساخت سری زمانی	۸۵
۲-۸ بررسی نموداری سری زمانی	۸۷
۳-۸ کار با مدل ARIMA	۹۰
فصل نهم: تعریف توابع جدید	۹۳
۱-۹ ساختار کلی تابع در R	۹۴
۲-۹ ساختار حلقه‌ها در R	۹۶
۳-۹ ساختار دستورهای شرطی در R	۹۸
۴-۹ برنامه‌نویسی بهینه	۹۸
پیوست: فهرست توابع و مجموعه داده‌های R	۱۰۳
واژه‌نامه انگلیسی به فارسی	۱۰۵
فهرست منابع	۱۲۳

پیش‌گفتار

امروزه آمار به عنوان یکی از کاربردی‌ترین شاخه‌های علوم، مورد استفاده سایر علوم است و به جرات، می‌توان از این علم به عنوان ابزاری نام برد که پیشرفت سایر علوم به آن وابسته است. در یک نگاه اجمالی می‌توان آمار را به دو بخش توصیفی و استنباطی تقسیم نمود. در بخش اول روش‌های مربوط به جمع‌آوری، دسته‌بندی و توصیف داده‌ها مورد بحث قرار می‌گیرد؛ اما در آمار استنباطی، تحلیل داده‌ها و اخذ تصمیم بر مبنای نتایج به دست آمده، هدف اصلی می‌باشد. گستردگی استفاده از آمار در سایر علوم باعث شده است تا کاربرانی با نگرش‌های متفاوت در زمینه‌های تحقیقاتی گوناگون و با سطوح علمی متفاوت، از آن استفاده کنند. عدم تسلط محققان غیر آماری به مبانی نظری از یک سو و نیاز آنها به استفاده از روش‌های آمار کاربردی از سوی دیگر، باعث ظهور نرم‌افزارهای متعددی با قابلیت‌های متفاوت، برای انجام محاسبه‌های آماری، شده است. اغلب این نرم‌افزارها تجاری و غیر رایگان هستند و در نتیجه، تهیه‌ی آنها برای کاربران مستلزم پرداخت هزینه‌های بالایی خواهد بود. در این میان R یک نرم‌افزار رایگان با قابلیت‌های محاسباتی برجسته است که اخیراً مورد توجه بسیاری از محققین حوزه‌ی علم آمار قرار گرفته است. در کتاب حاضر نویسندگان کوشیده‌اند تا روش استفاده از این نرم‌افزار را به صورت اجمالی ارائه کنند. با وجود اینکه نویسندگان تمام تلاش خود را در جهت ارائه یک اثر مناسب و کاربردی به کار گماشته‌اند، اما مطمئناً، کتاب حاضر خالی از ایراد و ابهام نیست. بر همین اساس موجب امتنان خواهد بود چنانچه خوانندگان محترم نظرها و پیشنهادهای خود و همچنین ایرادهای نوشته حاضر را به آدرس پست الکترونیکی:

nader_najari85@yahoo.com و با

r_mahmodvand@yahoo.com

ارسال نمایند.

در اینجا بر خود لازم می‌دانیم از زحمات کلیه کسانی که ما را در تهیه‌ی این کتاب یاری نمودند تقدیر و تشکر کنیم. در خاتمه اما نه کمتر بلکه همواره و خالصانه‌تر، نویسندگان از خانواده‌های خویش سپاس‌گزارند که همواره پشتیبان و مشوق ایشان در گردآوری مطالب بوده‌اند.

مولفان: بهار ۱۳۹۰

مقدمه

نویسندگان این نوشته همواره سعی داشته‌اند تا با ساده‌ترین شکل و با استفاده از ابزارهای متفاوت، مفاهیم را آموزش دهند. شاید بیان نوع برخورد نویسندگان با مسایل پیش‌آمده در حین گردآوری مطالب، خواننده را بیشتر در راستای نوع آموزش و اهداف این نوشته قرار دهد. در نگارش این اثر، دو نکته همواره مورد توجه بوده است که ذکر آنها ضروری است. اول اینکه مفاهیم ارائه شده کاربردی بوده و از بیان مباحث نظری خودداری شود و فرض شده است که یا خواننده از مطالب نظری آگاه است و یا در صورت نیاز، به منابع مناسب در آن زمینه رجوع خواهد کرد. دوم اینکه در این کتاب، هدف اصلی آموزش استفاده از امکانات موجود در نرم‌افزار R است و نه آموزش R. امید است خواننده همواره این دو نکته را در سرتاسر این نوشته در ذهن داشته باشد.

مطالب کتاب حاضر به این صورت است که فصل اول، ساختار انواع داده‌ها در R را مشخص نموده و به نحوه‌ی ورود اطلاعات و اجرای پردازش‌ها و ویرایش‌های لازم پرداخته است. لذا خواننده همراه با آموزش نرم‌افزار با روش‌های مربوط به مدیریت داده‌ها در این فصل آشنا می‌شود. سعی شده است روش‌های مورد استفاده تا حد ممکن ساده و قابل فهم باشند. در فصل دوم به بیان برخی از محاسبات ریاضی که در حین برنامه‌نویسی ممکن است به استفاده از آنها نیاز باشد، پرداخته شده است. فصل سوم انواع نمودارهای توصیف کننده داده‌ها را توضیح داده است. بررسی توزیع‌ها و نحوه‌ی تولید داده‌های تصادفی از توزیع‌های مختلف که مبنای محاسبات شبیه‌سازی است، در فصل چهارم ارائه شده است. در فصول پنجم تا هشتم کتاب نیز برخی از دستورهای مربوط به آزمون فرضیه‌های آماری، تحلیل رگرسیونی، طرح آزمایش‌ها و تحلیل واریانس و تحلیل سری‌های زمانی به همراه مثال‌های کافی برای آشنایی خواننده، شرح داده شده است. توانایی این نرم‌افزار در تهیه توابع مناسب برای اهداف خاص نیز در فصل پایانی این کتاب مورد توجه قرار گرفته است.

آشنایی اولیه با R

R نرم‌افزار آماری با قابلیت برنامه‌نویسی است که شباهت زیادی به نرم‌افزار S-PLUS دارد و برای انواع محاسبات آماری طراحی شده است. نرم‌افزار R توسط دو فرد به نام‌های رابرت جنتلمن^۱ و راس ایحاکا^۲ (این نرم‌افزار به خاطر اینکه نام این دو پژوهشگر با حرف R شروع شده به این نام، نام‌گذاری شده است!) در گروه آمار دانشگاه اوکلند^۳ نیوزلند در سال ۱۹۹۵ طراحی شده است.

نشانی صفحه وب این نرم‌افزار <http://www.r-project.org> بوده و آدرس اینترنتی <http://cran.um.ac.ir> جهت دانلود فایل اجرایی نرم‌افزار و بسته‌های آن است.

مزایای مهم نرم‌افزار R

- رایگان بودن و کد باز (Open Source) بودن؛
- قابلیت نصب روی اکثر سیستم‌های عامل از جمله توزیع‌های مختلف لینوکس، ویندوز و مک؛
- سرعت در دست‌یابی به تکنیک‌های جدید در قالب کتابخانه‌ها و توابع آماده؛
- کمک به یادگیری ادبیات جدید آمار با توجه به استفاده‌ی گسترده متخصصان برجسته‌ی آمار از R؛
- قابلیت اضافه کردن و نوشتن برنامه‌های جدید در R.

1 Robert Gentleman

2 Ross Ihaka

3 Auckland

طریقه نصب بر روی سیستم عامل ویندوز و به روز کردن R

اگر برای اولین بار است که می‌خواهید R را بر روی سیستم عامل ویندوز خود نصب نمایید، ابتدا فایل اجرایی R را از مسیر <http://cran.um.ac.ir> دانلود کرده و مراحل نصب آن را همانند دیگر نرم‌افزارها دنبال می‌کنیم. با اتمام مراحل نصب آیکونی به شکل R در صفحه میزکار قرار می‌گیرد. با دوبار کلیک بر روی این آیکون، پنجره‌ای باز می‌شود که شامل قسمت‌های زیر است:

- پنجره R Console که مربوط به محیط برنامه‌نویسی و خروجی‌های حاصل از اجرای فرامین است؛
- نوار منو که شامل منوهای ذخیره و بازیابی، ویرایش، بارگذاری بسته‌ها، راهنما و ... است؛
- نوار ابزار که شامل برخی آیکون‌های میان‌بر برای فرمان‌های پرکاربرد R است.

در صورتی که قبلاً R را بر روی سیستم عامل ویندوز خود نصب کرده‌اید و اکنون با ورود نسخه جدیدی از R مایل به به‌روز کردن نرم‌افزار R هستید، کافی است به مسیر `C:\Program Files\R\R-2.xx.x` رفته و محتویات پوشه‌ی `library` را در محل دیگری خارج از این مسیر کپی نمایید. سپس نسخه قبلی را `uninstall` نموده و نسخه جدید را به روشی که در بالا عنوان شد، نصب کنید. حال محتویات فایل `library` که قبلاً کپی نموده‌اید را در مسیر نصب R جدید در پوشه `library` کپی می‌کنیم و دستور زیر را برای به‌روز کردن کتابخانه‌های نصب شده در نسخه قبلی اجرا می‌کنیم:

```
> update.packages(ask=F, repos="http://cran.um.ac.ir")
```

با این‌کار تمامی کتابخانه‌هایی که قبلاً نصب کرده‌اید به‌صورت به‌روز شده قابل دسترسی و استفاده خواهند بود.

از گزینه‌های موجود در منوی Packages، که در قسمت نوار ابزار قرار دارد، نیز می‌توان برای این کار استفاده کرد.

نکات قابل توجه در R

— خط فرمان در R به صورت `>` و به رنگ قرمز است. در صورتی که دستور یا برنامه‌ای ناقص باشد این علامت به علامت `+` تبدیل می‌شود؛

— پاک کردن صفحه با کلیدهای میان‌بر `Ctrl+L` انجام می‌شود؛

— فراخوانی دستورات نوشته شده با استفاده از کلیدهای  و  امکان‌پذیر است؛

— پاسخ عبارت یا دستور در R در مقابل اعلان جدیدی (`[ ]`) به رنگ آبی نمایش داده می‌شود؛

— R نسبت به کوچک و بزرگ بودن حروف حساس است و از فاصله‌های اضافی چشم‌پوشی می‌کند؛

— برای انتساب نام به یک دستور یا شی، از `=` (یا گاهی از `<-`) استفاده می‌کنیم. بنابراین برای فراخوانی آن شی یا دستور، از نام انتساب داده شده به آن استفاده می‌کنیم. به مثال زیر توجه کنید:

```
د. d<-30 >
```

با این دستور برای عدد ۳۰، نام `d` انتساب داده می‌شود و با دستور:

```
د. d >
```

`d` را فراخوانی می‌کنیم. با اجرای این دستور عدد ۳۰ در خروجی به شکل زیر نمایش داده می‌شود:

```
[1] 30
```

— برای توقف اجرای یک دستور یا برنامه، از کلید `ESC` استفاده می‌کنیم؛

— R مقادیر عددی بی‌نهایت ($-\infty, +\infty$) را با Inf (یا به ترتیب با $-\text{Inf}$ و Inf) نشان می‌دهد؛

— هر شی در R حداقل دارای دو خصوصیت زیر است:

- ویژگی mode: نوع عناصر شی را مشخص می‌کند که ۴ نوع عمده‌ی آن عبارت است از: عددی، کاراکتری، منطقی و مختلط؛
- ویژگی length: طول (تعداد) عناصر شی را مشخص می‌کند.

— مقادیر گم‌شده در R، با NA^1 نشان داده می‌شود و نماد NaN^2 به مفهوم آن است که مقدار

محاسبه شده، عدد مبهمی مثل $\frac{0}{0}$ و یا لگاریتم یک عدد منفی و یا ... است؛

— پس از اجرای دستورات مربوط به رسم نمودار در R، پنجره‌ی جدیدی در R باز می‌شود و نمودار در آن نمایش داده می‌شود. برای کپی یا ذخیره کردن نمودار، روی آن کلیک راست نموده و گزینه‌ی Copy as metafile یا Copy as bitmap را برای کپی کردن فایل و گزینه‌ی Save as metafile... و گزینه‌ی Save as postscript... را به ترتیب برای ذخیره‌ی فایل با پسوند emf و پسوند eps استفاده می‌کنیم. با ذخیره کردن فایل با پسوند eps، فایل با حجم خیلی کم و کیفیت بالا ذخیره می‌شود؛

— برای استفاده از کدهای برنامه‌ی یک تابع موجود در R، اغلب کافی است نام تابع را فراخوانی کنیم. به عنوان مثال دستور زیر کدهای برنامه تابع arima را فراخوانی می‌کند:

```
> arima
```

— راهنمای R

- منوی Help از چندین قسمت تشکیل شده است.

1 Not Available
2 Not a Number

قسمت Manuals (in PDF) شامل فایل‌های راهنمایی در قالب pdf است که هم

برخی از دستورات R و هم مبانی نظری و روش‌های آماری را در بر می‌گیرد:

- استفاده از دستور (نام تابع یا "کاراکتر") help یا به‌جای آن از استفاده از دستور (نام تابع یا "کاراکتر")؟. برای مثال، دستور help(plot)، راهنمایی‌های لازم را برای تابع plot فراهم می‌کند. البته متذکر می‌شویم که این دستور در صورتی قابل استفاده است که نام تابع را به‌طور دقیق بدانیم. در غیر این صورت از روشی که در ادامه گفته شده است، استفاده می‌شود:
- استفاده از دستور (نام تابع یا "کاراکتر") help.search یا به‌جای آن استفاده از (نام تابع یا "کاراکتر")؟، برای جستجو و یافتن صفحات در موضوعات مبهم استفاده می‌شود. با اجرای این دستور، جستجو در تمام مستندات، عناوین، کلمات کلیدی و ... که به نحوی با نام تابع مورد جستجو در ارتباط هستند نمایش داده می‌شود. برای مثال، با دستور garch؟؟؟، کلیمی مواردی که به نحوی با تابع garch در ارتباط است نمایش داده می‌شود.