مجله مدل‌سازی پیشرفته ریاضی
سال ۱۴۰۲، دوره ۱۳، شماره ۱، ص ۱۷-۳۰

شاپا چاپی: ۲۲۵۱-۸۰۸۸
DOI: 10.22055/JAMM.2023.41837.2093

برازش توزیع‌های احتمالی به کمک نرم‌افزار R و کاربرد آن در پزشکی

مهدی مهدی‌زاده [۱] ×، الهه مهدی‌زاده [۲]

[۱] گروه آمار، دانشگاه حکیم سبزواری، سبزوار، ایران
[۲] بیمارستان مهر مادر، دانشکده علوم پزشکی تربت‌جام، تربت‌جام، ایران

دبیر مسئول: سعید قهرمانی

تاریخ دریافت: ۱۴۰۱/۶/۱۹ تاریخ پذیرش: ۱۴۰۱/۱۲/۱

چکیده: محققان علوم مختلف اغلب با پدیده‌هایی رو به رو هستند که ماهیت تصادفی دارند. گاهی می‌توان از توزیع‌های احتمالی برای توصیف و پیش‌بینی این‌گونه پدیده‌ها استفاده کرد. هر توزیع دارای تعدادی پارامتر مجهول است که مقدار آنها بر اساس داده‌ها برآورد می‌شوند. در برخی مسائل، چند توزیع رقیب برای برازش به یک مجموعه داده وجود دارد. در این صورت، لازم است مدل مناسب را بر اساس معیارهایی انتخاب کرد. این مقاله به معرفی امکانات نرم‌افزار آماری R برای اجرای مراحل فوق می‌پردازد. کاربرد روش‌های مطرح شده را به کمک یک مجموعه داده پزشکی نشان می‌دهیم.

واژه‌های کلیدی: انتخاب مدل، برآورد پارامتر، توزیع گاما، توزیع لگ‌نرمال.

رده‌بندی ریاضی: 62F10; 62P10

۱ مقدمه

نظریه توزیع‌ها ابزاری قوی برای مدل‌سازی آماری پدیده‌ها در علوم مختلف بوده و تحقیقات زیادی در این زمینه انجام شده است. تحت شرایطی، روش‌های آماری توسعه یافته بر مبنای فرض‌های توزیعی، نسبت به معادل ناپارامتری خود کاراتر هستند. به عنوان نمونه می‌توان به توزیع‌های نرمال و نمایی اشاره کرد که شکل ریاضی انعطاف‌پذیری دارند و رویه‌های آماری بسیاری بر اساس آنها پیشنهاد شده‌اند. در زمینه پزشکی، توزیع گاما در ارزیابی سرطان سینه بر اساس MRI پخش وزنی (بورلینهاس و همکاران [۱])، توزیع لگ‌نرمال در مدل‌سازی جذب سلولی رادیواکتیویته (نتی و هوول [۴])، و توزیع گومپرتز در مدل‌سازی سن در زمان مرگ (اسکریور و همکاران [۷]) کاربرد دارد. همچنین، توزیع‌های بور، لومکس، پارتو، لوژستیک و برخی توزیع‌های دیگر در مدل‌سازی نویز نقطه‌ای استفاده شده‌اند (پارکر و پول [۵] ؛ ونگ و همکاران [۱۰]).

در سال‌های اخیر، معرفی خانواده جدید از توزیع‌ها مورد توجه مؤلفان زیادی قرار گرفته است. این توزیع‌ها معمولاً با اعمال برخی تبدیلات روی توزیع‌های موجود حاصل می‌شوند و علاوه بر پارامترهای مکان و مقیاس، دارای یک یا چند پارامتر شکل نیز هستند. با افزایش تعداد

×نویسنده مسئول مقاله
رایانامه: mahdizadeh.m@hsu.ac.ir;mahdizadeh.m@live.com (M. Mahdizadeh)

پارامترها، توزیع حاصل امکان برازش بهتر به داده را دارد. به عنوان مثال، شارما و همکاران [۶] توزیع لیندلی معکوس را معرفی و ویژگی‌های آن را بررسی کردند. سپس این توزیع را برای مدل‌سازی طول عمر در بیماران مبتلا به سرطان سر و گردن به کار بردند.

برای مدل‌سازی به کمک یک توزیع، ابتدا باید مقدار پارامترهای آن را بر اساس داده‌ها برآورد کرد. برخی روش‌های رایج برآورد عبارتند از ماکسیمم درستنمایی، گشتاوری و چندکی. اگر چند توزیع برای برازش به یک مجموعه داده وجود داشته باشد، آنگاه باید بهترین توزیع را بر اساس معیارهایی انتخاب کرد. در انجام این مراحل، غالباً باید از نرم‌افزارهای ریاضی کمک گرفت. به عنوان مثال، R یک نرم‌افزار رایگان و متن‌باز برای برنامه‌نویسی و محاسبات آماری است. این نرم‌افزار در سال ۱۹۹۳ توسط راس ایهاکا و رابرت جنتلمن در دانشگاه اوکلند (کشور نیوزلند) طراحی شد. در سال‌های اخیر، استفاده از R با استقبال خوبی از سوی محققان علوم مختلف رو به رو شده است. برای بارگیری آخرین نسخه این نرم‌افزار می‌توان به شبکه آرشیو جامع R [†] مراجعه کرد.

در بخش ۲، روش‌های ماکسیمم درستنمایی، گشتاوری و چندکی مرور می‌شوند. چند معیار اصلی انتخاب مدل در بخش ۳ آمده‌اند. در بخش ۴، نحوه برازش توزیع‌ها در نرم‌افزار R بررسی می‌شود. بخش ۵ کاربرد رویه‌های معرفی شده را در زمینه پزشکی نشان می‌دهد. خلاصه‌ای از مقاله در بخش ۶ بیان می‌شود. نمودارهای خروجی نرم‌افزار در یک ضمیمه آمده‌اند.

۲ روش‌های برآورد پارامتر

فرض کنید $x_1, \ldots, x_n$ مشاهدات یک نمونه تصادفی به حجم n از توزیعی با تابع جرم (چگالی) احتمال $f(x;\boldsymbol{\theta})$ و تابع توزیع تجمعی $F(x;\boldsymbol{\theta})$ باشد که در آن $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ بردار پارامترهای توزیع است. همانطور که در بخش قبل ذکر شد، روش‌های مختلفی برای برآورد پارامترهای مجهول θ_j $(j = 1, \ldots, k)$ وجود دارند. روش ماکسیمم درستنمایی به برآوردگرهایی منجر می‌شود که تحت شرایطی، دارای ویژگی‌های بسیار مطلوب مانند پایایی، سازگاری و توزیع مجانبی نرمال هستند.

اساس کار روش ماکسیمم درستنمایی، تابع درستنمایی است که به صورت حاصلضرب تابع جرم (چگالی) احتمال مورد نظر به ازای مشاهدات نمونه تعریف می‌شود. تابع درستنمایی متناظر با $f(x;\boldsymbol{\theta})$ برابر است با

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}).$$

برآوردگر ماکسیمم درستنمایی پارامترها مقادیری هستند که تابع فوق را ماکسیمم می‌کنند. به عبارت دیگر، برای محاسبه این برآوردگرها باید یک مسئله بهینه‌سازی را حل کرد. معمولاً به جای خود تابع درستنمایی، از لگاریتم آن به صورت زیر استفاده می‌شود:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i;\boldsymbol{\theta}).$$

با مشتق‌گیری از تابع فوق نسبت به پارامترها و حل دستگاه حاصل از برابر صفر قرار دادن مشتق‌های مذکور، برآوردگرهای ماکسیمم درستنمایی به دست می‌آیند. به عبارت دیگر، برآوردگرهای ماکسیمم درستنمایی جواب دستگاه معادلات زیر هستند (با فرض اینکه در برخی شرط‌های لازم صدق کنند):

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} = 0 \quad (j = 1, \ldots, k).$$

حل دستگاه معادلات فوق برای برخی توزیع‌ها مثل نرمال و نمایی، آسان بوده و به جواب‌های صریح منجر می‌شود. اما در مورد بسیاری از توزیع‌ها مثل گاما و بتا، برآوردگرهای ماکسیمم درستنمایی شکل بسته ندارند و برای به دست آوردن آنها باید از روش‌های عددی استفاده کرد.

روش گشتاوری، یکی از قدیمی‌ترین روش‌های برآورد است (ووز [۹]). ایده روش گشتاوری این است که اگر نمونه‌ای معرف و به قدر کافی بزرگ در اختیار داشته باشیم، باید گشتاورهای نمونه با گشتاورهای جامعه تقریباً برابر باشند. فرض کنید m'_r و μ'_r به ترتیب گشتاور مرتبه r ام نمونه و جامعه حول مبدأ باشند، یعنی

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r,$$

$$\mu'_r = E(X^r),$$

[†] https://cran.r-project.org/

که در آن $E(.)$ نماد امید ریاضی است. برآوردگرهای گشتاوری با حل دستگاه معادلات زیر به دست می‌آیند:

$$\mu'_r = m'_r \quad (r = ۱, \dots, k),$$

که در آن k تعداد پارامترهای مجهول است. در این دستگاه، m'_r ها مقادیر معلوم و μ'_r ها مقادیر مجهول معادله‌ها هستند. اگر دستگاه فوق دارای جواب صریح نباشد، باید آن را با روش‌های عددی حل کرد. می‌دانیم برای برخی از توزیع‌های دم‌بلند مانند کوشی، گشتاورها از هیچ مرتبه‌ای وجود ندارند. بنابراین، نمی‌توان روش گشتاوری را به کار برد.

روش چندکی، یک رویکرد دیگر برای برآورد پارامتر است (تسه [۸]). این روش، ایده‌ای مشابه روش گشتاوری دارد اما از چندک‌های نمونه و جامعه برای برآورد پارامتر استفاده می‌کند. فرض کنید $Q_{n,p}$ و $F^{-۱}(p;\boldsymbol{\theta})$ به ترتیب چندک مرتبه p نمونه و جامعه باشند. برآوردگرهای چندکی با حل دستگاه معادلات زیر به دست می‌آیند:

$$F^{-۱}(p_r;\boldsymbol{\theta}) = Q_{n,p_r} \quad (r = ۱, \dots, k),$$

که در آن k همانند قبل است، و p_r ها احتمالاتی هستند که برای تطابق چندک‌ها انتخاب می‌شوند. در این روش نیز، گاهی باید دستگاه فوق را با روش‌های عددی حل کرد.

۳ معیارهای انتخاب مدل

فرض کنید چند توزیع رقیب برای برازش به یک مجموعه داده وجود دارند. در این صورت، می‌توان بر اساس شاخص‌هایی آنها را با هم مقایسه کرد. یک معیار ساده، $\ell(\hat{\boldsymbol{\theta}})$ است که در آن $\hat{\boldsymbol{\theta}}$ برآورد پارامترها به یکی از روش‌های بخش قبل است. هر چقدر این کمیت بزرگتر باشد، توزیع متناظر برازش بهتر دارد. استفاده از این معیار زمانی مجاز است که توزیع‌های رقیب دارای تعداد پارامتر یکسان باشند. در غیر این صورت، معمولاً توزیع با پارامترهای بیشتر انتخاب می‌شود. این مشکل، بیش‌برازش [‡] نام دارد و برای حل آن می‌توان از معیار اطلاع آکائیکه (AIC[§]) و معیار اطلاع بیزی (BIC[¶]) استفاده کرد. این دو معیار عبارتند از

$$\text{AIC} = -۲\ell(\hat{\boldsymbol{\theta}}) + ۲k,$$

و

$$\text{BIC} = -۲\ell(\hat{\boldsymbol{\theta}}) + k\log(n),$$

که در آن k تعداد پارامترهای توزیع و n حجم نمونه است. هر چقدر این کمیت‌ها کوچکتر باشد، توزیع متناظر برازش بهتر دارد. واضح است که افزایش تعداد پارامترها در شاخص BIC بیشتر جریمه می‌شود.

در آمار ناپارامتری، تابع توزیع تجربی یک برآوردگر ناریب برای $F(x;\boldsymbol{\theta})$ است. این تابع به صورت زیر تعریف می‌شود

$$F_n(x) = \frac{۱}{n}\sum_{i=۱}^{n} I(x_i \le x),$$

که در آن $I(.)$ تابع نشانگر است. آماره‌های نیکویی برازش که بر اساس فاصله‌های مختلف بین $F_n(x)$ و $F(x;\hat{\boldsymbol{\theta}})$ ساخته می‌شوند، راه دیگری برای مقایسه برازش چند توزیع به یک مجموعه داده هستند. سه آماره معروف کولموگوروف–اسمیرنوف (KS[ǁ])، کرامر–فون میزس (CvM[**])، و اندرسون–دارلینگ (AD[††]) عبارتند از

$$\text{KS} = \sup_x \left| F_n(x) - F(x;\hat{\boldsymbol{\theta}}) \right|,$$

$$\text{CvM} = n\int_{-\infty}^{\infty} \left(F_n(x) - F(x;\hat{\boldsymbol{\theta}}) \right)^۲ \mathrm{d}F(x;\hat{\boldsymbol{\theta}}),$$

[‡] Overfitting
[§] Akaike Information Criterion
[¶] Bayesian Information Criterion
[ǁ] Kolmogorov-Smirnov
[**] Cramer-von Mises
[††] Anderson-Darling

و

$$\mathrm{AD} = n \int_{-\infty}^{\infty} \frac{\left(F_n(x) - F(x;\hat{\boldsymbol{\theta}})\right)^{۲}}{F(x;\hat{\boldsymbol{\theta}})\left(۱ - F(x;\hat{\boldsymbol{\theta}})\right)} \,\mathrm{d}F(x;\hat{\boldsymbol{\theta}})\,.$$

هر چقدر مقدار آماره‌های فوق کوچکتر باشد، توزیع متناظر برازش بهتر دارد. لازم به ذکر است که این آماره‌ها پیچیدگی مدل (تعداد پارامترها) را در نظر نمی‌گیرند.

۴ بسته fitdistrplus

در بسته اصلی نرم‌افزار R ، بسیاری از تحلیل‌های آماری استاندارد انجام می‌شود. برای برخی روش‌های پیچیده‌تر یا دارای کاربرد خاص، بسته‌های جانبی وجود دارد که به طور رایگان در شبکه آرشیو جامع R ارائه می‌شوند. این بسته‌ها به سادگی در نرم‌افزار اصلی نصب شده و امکانات آنها قابل دسترسی است. fitdistrplus یک بسته مفید برای برازش توزیع‌های احتمالی به انواع داده‌ها در نرم‌افزار R می‌باشد (دلیگنت-مولر و داتنگ [۳]). در این بخش، نحوه نصب و استفاده از این بسته بررسی می‌شود.

پنجره اصلی نرم‌افزار R با نام R Console شناخته می‌شود. در صورتی که رایانه شما به اینترنت متصل باشد، برای نصب بسته فوق ابتدا دستور زیر را در پنجره R Console تایپ کرده و با فشار دادن کلید Enter آن را اجرا کنید:

```
install.packages("fitdistrplus")
```

سپس پنجره جدیدی باز شده که شامل لیستی از سرورهای موجود در کشورهای مختلف است. با انتخاب یکی از سرورها، فرایند نصب به طور خودکار ادامه و پایان می‌یابد. برای استفاده از امکانات این بسته، ابتدا باید آن را فراخوانی کرد. این کار با اجرای دستور زیر انجام می‌شود:

```
library(fitdistrplus)
```

هر چند می‌توان سایر دستورات لازم را نیز در پنجره R Console تایپ و اجرا کرد اما بهتر است آنها را در یک فایل متنی نوشت تا راحت‌تر قابل ویرایش باشد. به این منظور، از منوی File در نرم افزار R ، گزینه New script را انتخاب کنید. اکنون، پنجره‌ای جدید با نام Untitled – R Editor باز می‌شود که یک محیط متنی برای نوشتن دستورات است. با استفاده از کلیدهای ترکیبی Ctrl+S ، می‌توان دستورات را در یک فایل (با پسوند R) با نام دلخواه ذخیره کرد. بعد از آخرین تغییرات فایل، استفاده مجدد از کلیدهای ترکیبی مذکور ضروری است. برای اجرای محتوای این فایل، ابتدا بخش مورد نظر را انتخاب کرده و سپس با کلیک راست روی آن، گزینه Run line or selection را انتخاب کنید تا خروجی در پنجره R Console نمایش داده شود. برای انتخاب همه محتوای یک فایل باید از کلیدهای ترکیبی Ctrl+A استفاده کرد. انتخاب بخشی از محتوای یک فایل، با نگه داشتن کلید سمت چپ ماوس و کشیدن اشاره‌گر آن روی بخش مورد نظر انجام می‌شود.

اگر داده‌ها در بردار data ذخیره شده باشند، آنگاه توزیع مناسب برای برازش را می‌توان به کمک دستور زیر انتخاب کرد:

```
descdist(data)
```

این دستور، آماره‌های توصیفی مینیمم، ماکسیمم، میانه، میانگین، انحراف معیار، و ضرایب چولگی و کشیدگی گشتاوری را گزارش می‌کند. برای سه آماره آخر، برآورد ناریب محاسبه می‌شود. دستور فوق، نمودار چولگی-کشیدگی معرفی شده توسط کولن و فری [۲] را نیز رسم می‌کند. نقطه مرجع در این نمودار، چولگی و کشیدگی داده‌ها را نشان می‌دهد. ناحیه متناظر با چولگی و کشیدگی توزیع‌های رایج نیز روی این نمودار تعیین می‌شود. مقدار چولگی و کشیدگی برخی توزیع‌ها (مثل نرمال، یکنواخت، لوژستیک و نمایی)، مستقل از پارامترها بوده و یک عدد ثابت است. بنابراین، توزیع با یک نقطه روی نمودار مشخص می‌شود. مجموعه مقادیر ممکن برای چولگی و کشیدگی در برخی توزیع‌ها (مثل گاما و لگ‌نرمال) به صورت یک خط، و در برخی توزیع‌های دیگر (مثل بتا) شامل نواحی بزرگتر است. هر چقدر نقطه مرجع به ناحیه متناظر یک توزیع نزدیکتر باشد، آن توزیع برای برازش به داده‌ها مناسب‌تر است.

شکل ساده دستور برازش یک توزیع به یک مجموعه داده به صورت زیر است:

```
out=fitdist(data,distr)
```

که در آن data بردار شامل داده‌ها، و distr توزیعی است که می‌خواهیم برازش دهیم. اکثر توزیع‌های پرکاربرد در نرم‌افزار R یا برخی بسته‌های تولید شده برای آن تعریف شده‌اند. برای هر توزیع، یک نام استاندارد وجود دارد. تابع جرم (چگالی) احتمال، تابع توزیع تجمعی و تابع چندک متناظر، به ترتیب با اضافه شدن حروف p ، d و q به ابتدای این نام ساخته می‌شوند. مثلاً norm نام استاندارد برای توزیع نرمال است و تابع چگالی احتمال، تابع توزیع تجمعی و تابع چندک متناظر عبارتند از dnorm ، pnorm و qnorm . با اجرای دستور ?Distributions در نرم‌افزار R می‌توانید فهرستی از نام‌های استاندارد برای توزیع‌های مختلف را ببینید. در دستور فوق برای برازش توزیع، به جای distr باید نام استاندارد توزیع مورد نظر را داخل گیومه گذاشت. مثلاً برای توزیع نرمال می‌نویسیم "norm" . نتایج حاصل از برازش به صورت زیر قابل مشاهده است:

```
summary(out)
```

موارد گزارش شده در خروجی این دستور عبارتند از: $\hat{\boldsymbol{\theta}}$ و برآورد خطای معیار مولفه‌های آن، $\ell(\hat{\boldsymbol{\theta}})$ و مقدار شاخص‌های AIC و BIC ، و برآورد ماتریس همبستگی $\hat{\boldsymbol{\theta}}$.

به طور پیش فرض، تابع fitdist() روش ماکسیمم درستنمایی را اجرا می‌کند. آماره‌های خروجی نیز بر همین اساس محاسبه می‌شوند. روش گشتاوری با شناسه‌های method و order در این تابع مشخص می‌شود. مقدار شناسه اول را mme ، و مقدار شناسه دوم را نام بردار شامل مرتبه گشتاورهای به کار رفته در برآورد قرار می‌دهیم. مثلاً در اجرای این روش برای توزیع گاما می‌نویسیم:

```
fitdist(data,"gamma",method="mme",order=1:2)
```

به طور مشابه، روش چندکی با شناسه‌های method و probs مشخص می‌شود. مقدار شناسه اول را qme ، و مقدار شناسه دوم را نام بردار شامل احتمالات p_r در روش چندکی قرار می‌دهیم. مثلاً در اجرای این روش برای توزیع لگ‌نرمال می‌نویسیم:

```
fitdist(data,"lnorm",method="qme",probs=c(0.25,0.75))
```

همانطور که در بخش ۲ ذکر شد، گاهی برآورد پارامتر به کمک روش‌های عددی انجام می‌شود. اغلب الگوریتم‌های موجود به یک نقطه اولیه برای شروع جستجو نیاز دارند. در بیشتر توزیع‌ها، این نقطه به طور خودکار توسط تابع fitdist() تعیین می‌شود. اگر این تابع نتواند نقطه مناسب را انتخاب کند، باید آن را از طریق شناسه start معرفی کرد. برای جزئیات بیشتر، دستور ?fitdist را اجرا کنید. برخی شناسه‌های دیگر نیز در تابع fitdist() وجود دارند که از بیان آنها صرفنظر می‌کنیم.

در بسته fitdistrplus ، توابعی برای رسم چهار نمودار استاندارد نیکویی برازش وجود دارد (کولن و فری [۲]). نمودار اول، بافت‌نگار ‡‡ داده‌ها به همراه تابع چگالی برازش شده است. نمودار دوم، تابع توزیع تجربی داده‌ها به همراه تابع توزیع تجمعی برازش شده است. نمودار سوم، نمودار چندک–چندک §§ است که در آن محور افقی (عمودی) چندک‌های نظری (تجربی) را نشان می‌دهد. نمودار چهارم، نمودار احتمال–احتمال ¶¶ است که در آن محور افقی (عمودی) احتمالات تجمعی نظری (تجربی) را نشان می‌دهد. این نمودارها به ترتیب با توابع denscomp() ، cdfcomp() ، qqcomp() و ppcomp() تولید می‌شوند. همچنین، آماره‌های نیکویی برازش با تابع gofstat() محاسبه می‌شوند. جزئیات استفاده از این توابع به کمک داده واقعی در بخش بعد بیان می‌شود.

۵ کاربرد

بیش از ۲ میلیون نفر در جهان که غالباً زنان وکودکان هستند دچار فقر آهن می‌باشند. حدود ۹۰٪ از افراد مبتلا در کشورهای درحال توسعه زندگی می‌کنند. فقر آهن، شایع‌ترین کم‌خونی در جهان بوده و پیامدهای آن در بیماران مبتلا از جمله در افراد نوجوان و بزرگسال، باعث کاهش توانایی کاری، و نقص در عملکرد سیستم ایمنی می‌شود. همچنین، تحقیقات نشان می‌دهد که فقر آهن با کاهش توانایی باروری و تولید مثل بیماران در ارتباط است.

مطالعه وضعیت فقر آهن در جوامع مختلف از اهمیت ویژه‌ای برخوردار است. اندازه‌گیری فری‌تین ××× ، شاخص خوبی برای ذخایر آهن در بدن است. فری‌تین پروتئین اصلی ذخیره‌کننده آهن است که غلظت آن ارتباط مستقیم با ذخایر آهن دارد. سطح فری‌تین با افزایش سن به طور دائم در مردان و زنان (پس از یائسگی) افزایش می‌یابد. کاهش سطح فری‌تین نشانه کم شدن ذخایر آهن و کم‌خونی فقر آهن است. هنگامی که ذخایر پروتئینی بدن به شدت کاهش یابد، سطح فری‌تین نیز کاهش می‌یابد. در بارداری نیز کاهش سطح فری‌تین مشاهده می‌شود.

‡‡Histogram
§§Q-Q plot
¶¶P-P plot
***Ferritin

داده‌های موسسه ورزش استرالیا [†††] به طور گسترده در مطالعه توزیع‌های چوله استفاده شده‌اند. این داده‌ها شامل اندازه‌گیری ۱۳ متغیر در مورد ۲۰۲ ورزشکار (۱۰۰ زن و ۱۰۲ مرد) استرالیایی هستند. یکی از این متغیرها غلظت فری‌تین است که در این بخش از آن استفاده می‌شود. برای دسترسی به داده‌های این متغیر، ابتدا بسته sn را در نرم‌افزار R نصب می‌کنیم (جزئیات این کار در بخش قبلی بیان شد). با اجرای دستورات زیر، داده‌ها در بردار Fer ذخیره شده، سپس بافت‌نگار و نمودار جعبه‌ای آنها رسم می‌شود:

```
library(sn)
data(ais)
Fer=ais[,7]
par(mfrow=c(2,1))
hist(Fer,xlab="Ferritin",main="")
boxplot(Fer,ylab="Ferritin")
```

با توجه به نمودارهای شکل ۱ می‌توان دید که توزیع متغیر مورد نظر، چولگی مثبت دارد. برای تشخیص توزیع مناسب، دستور زیر را بعد از فراخوانی بسته fitdistrplus به کار می‌بریم:

```
descdist(Fer)
```

مقدار آماره‌های توصیفی به صورت زیر در خروجی گزارش می‌شود:

```
> descdist(Fer)
summary statistics
------
min:  8   max:  234
median:  65.5
mean:  76.87624
estimated sd:  47.50124
estimated skewness:  1.290184
estimated kurtosis:  4.486265
```

نمودار چولگی-کشیدگی حاصل نیز در شکل ۲ رسم شده است. با توجه به موقعیت نقطه مرجع (دایره آبی رنگ توپر)، توزیع‌های گاما و لگ‌نرمال را برای مدل‌سازی داده‌های غلظت فری‌تین به کار می‌بریم. توزیع بتا به دلیل محدود بودن تکیه‌گاه آن، انتخاب نمی‌شود.

تابع چگالی احتمال توزیع گاما در نرم‌افزار R عبارتست از

$$f_G(x) = \frac{s^a}{\Gamma(a)} x^{a-1} \exp\{-sx\}, \quad x > 0; a, s > 0,$$

که در آن $\Gamma(.)$ تابع گاما، a پارامتر شکل و s پارامتر نرخ است. تابع چگالی احتمال توزیع لگ‌نرمال در این نرم‌افزار عبارتست از

$$f_{LN}(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-(\log x - \mu)^2/(2\sigma^2)\right\}, \quad x > 0; \mu \in \mathbb{R}, \sigma > 0,$$

که در آن μ و σ به ترتیب میانگین و انحراف معیار متغیر تصادفی $\log X$ هستند. همچنین، نام استاندارد توزیع گاما و لگ‌نرمال عبارتند از gamma و lnorm.

برای برازش این دو توزیع با روش ماکسیمم درستنمایی به داده‌ها (که قبلاً در بردار Fer ذخیره شده‌اند) کافی است دستورات زیر را اجرا کنیم:

```
mle.g=fitdist(Fer,"gamma")
mle.ln=fitdist(Fer,"lnorm")

summary(mle.g)
summary(mle.ln)
```

[†††] Australian Institute of Sport

خروجی متناظر در پنجره R Console به صورت زیر است:

```
> mle.g=fitdist(Fer,"gamma")
> mle.ln=fitdist(Fer,"lnorm")
>
> summary(mle.g)

Fitting of the distribution ' gamma ' by maximum likelihood
Parameters :
        estimate  Std. Error
shape 2.90803574 0.273720869
rate  0.03783072 0.003883684
Loglikelihood:  -1030.719   AIC:  2065.438   BIC:  2072.055
Correlation matrix:
          shape      rate
shape 1.0000000 0.9158039
rate  0.9158039 1.0000000

> summary(mle.ln)

Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate Std. Error
meanlog 4.1605298 0.04347532
sdlog   0.6179004 0.03074133
Loglikelihood:  -1029.804   AIC:  2063.608   BIC:  2070.225
Correlation matrix:
        meanlog sdlog
meanlog       1     0
sdlog         0     1
```

با توجه به اینکه تعداد پارامترهای هر دو توزیع یکسان است، می‌توان از ملاک $\ell(\hat{\boldsymbol{\theta}})$ برای مقایسه کیفیت برازش استفاده کرد. مقدار این کمیت برای توزیع گاما و لگ‌نرمال به ترتیب برابر است با ۷۱۹.۱۰۳۹- و ۸۰۴.۱۰۲۹-. بنابراین، توزیع لگ‌نرمال برای مدل‌سازی داده‌ها اندکی برتری دارد. شاخص‌های AIC و BIC نیز از این توزیع حمایت می‌کنند.

چهار نمودار استاندارد نیکویی برازش که در بخش قبل معرفی شد، با اجرای دستورات زیر تولید می‌شوند:

```
par(mfrow=c(2,2))
plot.legend=c("Gamma","Lognormal")
denscomp(list(mle.g,mle.ln),legendtext=plot.legend)
qqcomp(list(mle.g,mle.ln),legendtext=plot.legend)
cdfcomp(list(mle.g,mle.ln),legendtext=plot.legend)
ppcomp(list(mle.g,mle.ln),legendtext=plot.legend)
```

با توجه به نتایج حاصل در شکل ۳ به نظر می‌رسد که توزیع لگ‌نرمال برازش بهتری دارد (قاب بالا و سمت چپ را ببینید). لازم به ذکر است که این نمودارها فقط برای مقایسه کیفیت برازش به صورت بصری مفیدند، و نتیجه‌گیری بر اساس آنها قطعاً با قضاوت شخصی همراه است.

آماره‌های نیکویی برازش، ابزار دقیق‌تری نسبت به نمودارهای فوق برای مقایسه هستند. این آماره‌ها به کمک دستور زیر محاسبه می‌شوند:

```
gofstat(list(mle.g,mle.ln),fitnames=c("Gamma","LN"))
```

خروجی حاصل به صورت زیر است:

```
> gofstat(list(mle.g,mle.ln),fitnames=c("Gamma","LN"))
```

```
Goodness-of-fit statistics
                                  Gamma         LN
Kolmogorov-Smirnov statistic 0.05425914 0.03568929
Cramer-von Mises statistic   0.10449303 0.03019391
Anderson-Darling statistic   0.72780643 0.23834663

Goodness-of-fit criteria
                                  Gamma       LN
Akaike's Information Criterion 2065.438 2063.608
Bayesian Information Criterion 2072.055 2070.225
```

ملاحظه می‌شود بر اساس هر سه آماره KS ، CvM و AD ، باید توزیع لگ‌نرمال را انتخاب کرد.

برای برازش توزیع‌های گاما و لگ‌نرمال با روش گشتاوری به داده‌های غلظت فری‌تین کافی است دستورات زیر را اجرا کنیم:

```
mme.g=fitdist(Fer,"gamma",method="mme",order=1:2)
mme.ln=fitdist(Fer,"lnorm",method="mme",order=1:2)

summary(mme.g)
summary(mme.ln)
```

خروجی حاصل به صورت زیر است:

```
> mme.g=fitdist(Fer,"gamma",method="mme",order=1:2)
> mme.ln=fitdist(Fer,"lnorm",method="mme",order=1:2)
>
> summary(mme.g)
Fitting of the distribution ' gamma ' by matching moments
Parameters :
       estimate
shape 2.6322654
rate  0.0342403
Loglikelihood:  -1031.261   AIC:  2066.522   BIC:  2073.139
> summary(mme.ln)
Fitting of the distribution ' lnorm ' by matching moments
Parameters :
         estimate
meanlog 4.1811910
sdlog   0.5674607
Loglikelihood:  -1031.49   AIC:  2066.979   BIC:  2073.596
```

برای محاسبه آماره‌های نیکویی برازش داریم:

```
gofstat(list(mme.g,mme.ln),fitnames=c("Gamma","LN"))
```

خروجی این دستور به صورت زیر است:

```
> gofstat(list(mme.g,mme.ln),fitnames=c("Gamma","LN"))
Goodness-of-fit statistics
                                  Gamma         LN
Kolmogorov-Smirnov statistic 0.04753879 0.06484533
Cramer-von Mises statistic   0.08773110 0.08816265
Anderson-Darling statistic   0.73600043 0.71344878

Goodness-of-fit criteria
                                  Gamma       LN
Akaike's Information Criterion 2066.522 2066.979
Bayesian Information Criterion 2073.139 2073.596
```

در هر دو توزیع، مقادیر برآورد گشتاوری پارامترها به مقادیر مشابه در روش ماکسیمم درستنمایی نزدیک است. برخلاف روش ماکسیمم درستنمایی، در اینجا تقریباً همه معیارها (به جز آماره AD) از توزیع گاما حمایت می‌کنند.

۶ نتیجه‌گیری

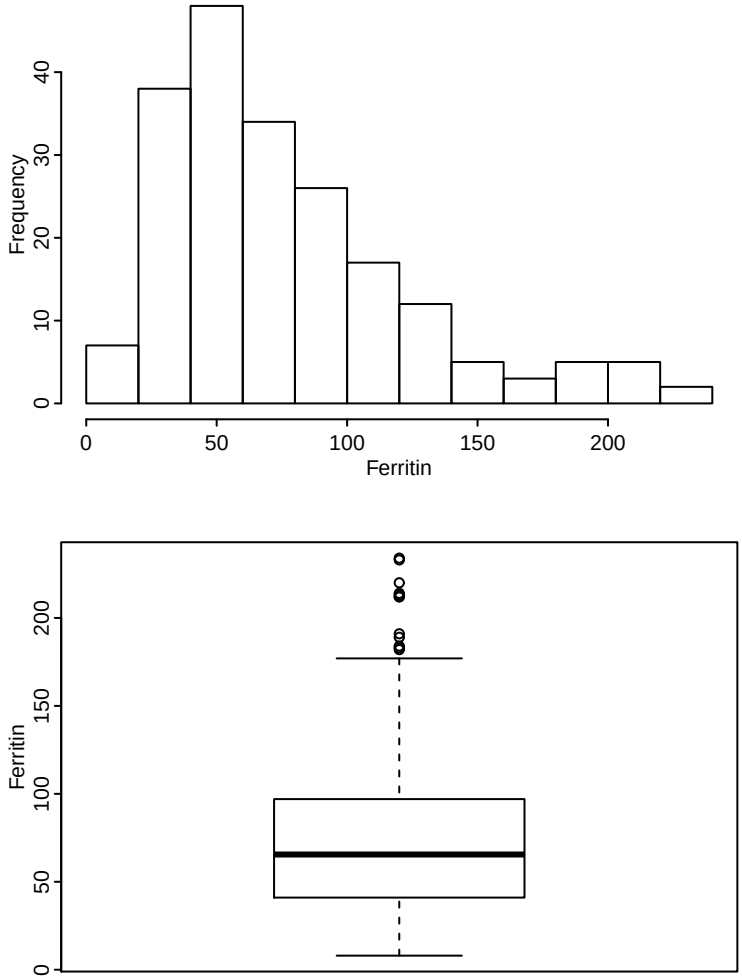
در علوم مختلف، بسیاری از متغیرهای تصادفی را می‌توان به کمک توزیع‌های احتمالی مدل‌سازی کرد. این کار نقش مهمی در توصیف و پیش‌بینی پدیده‌هایی دارد که رفتار آنها از مدل‌های قطعی پیروی نمی‌کنند. نرم‌افزار R قابلیت‌های خوبی در زمینه محاسبات آماری دارد و کاربرد آن در بسیاری از شاخه‌های علمی رایج است. این مقاله به موضوع برازش توزیع‌های احتمالی به کمک این نرم‌افزار می‌پردازد. ابتدا روش‌های ماکسیمم درستنمایی، گشتاوری و چندکی برای برآورد پارامتر مرور می‌شوند. سپس، معیارهای انتخاب مدل بر اساس تابع درستنمایی، و آماره‌های نیکویی برازش معرفی می‌شوند. در ادامه، نحوه استفاده از بسته fitdistrplus در برازش توزیع‌ها بررسی می‌شود. در پایان، روش‌های مطرح شده روی یک مجموعه داده پزشکی اعمال می‌شوند.

لازم به ذکر است که هر چند داده‌های غلظت فری‌تین در بخش قبل پیوسته هستند، اما محدودیتی در استفاده از بسته fitdistrplus برای داده‌های گسسته وجود ندارد. همچنین، نسخه فعلی این بسته امکان برازش توزیع به داده‌های سانسور شده را نیز دارد (البته برخی محدودیت‌ها در مقایسه با تحلیل داده‌های کامل وجود دارد). برای جزئیات بیشتر در این زمینه، علاقه‌مندان می‌توانند به بخش ۳.۳ در دلیگنت–مولر و داتنگ [۳] مراجعه کنند. با توجه به امکانات خوب این بسته، استفاده از آن برای محققان علوم مختلف سودمند است.

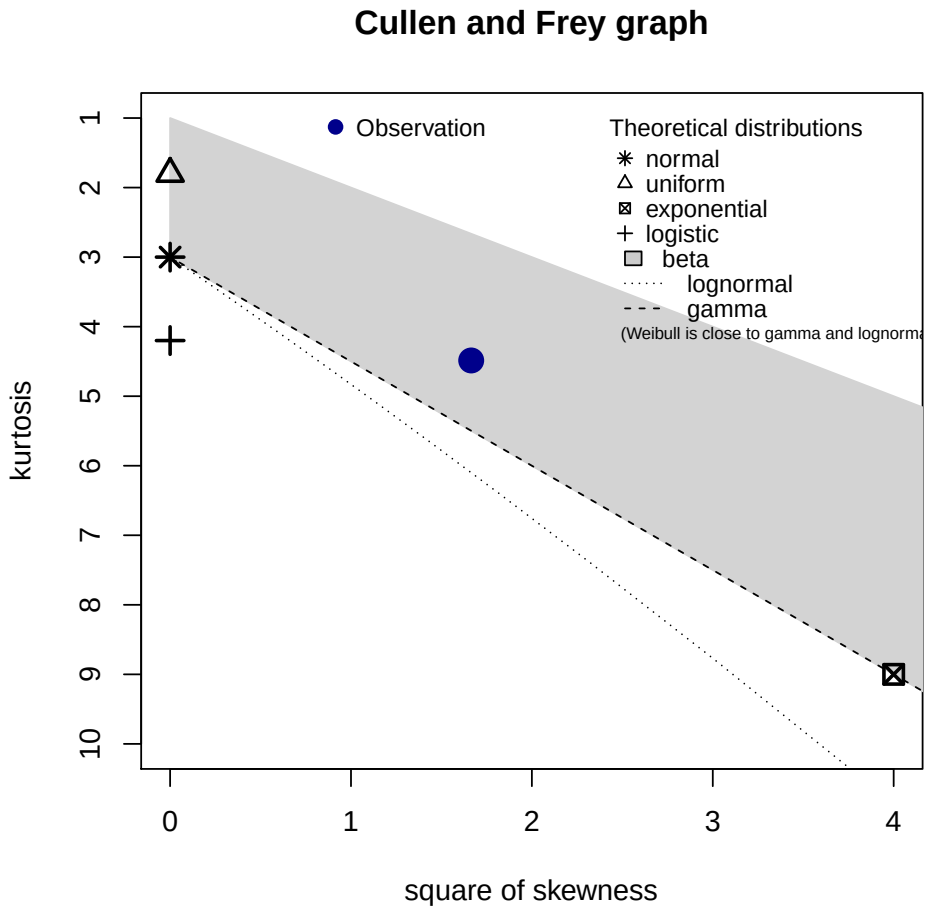
تشکر و قدردانی

نویسندگان از دبیر مسئول و داور محترم که با صرف وقت فراوان ارزیابی علمی این مقاله را به عهده داشتند، صمیمانه قدردانی می‌کنند.

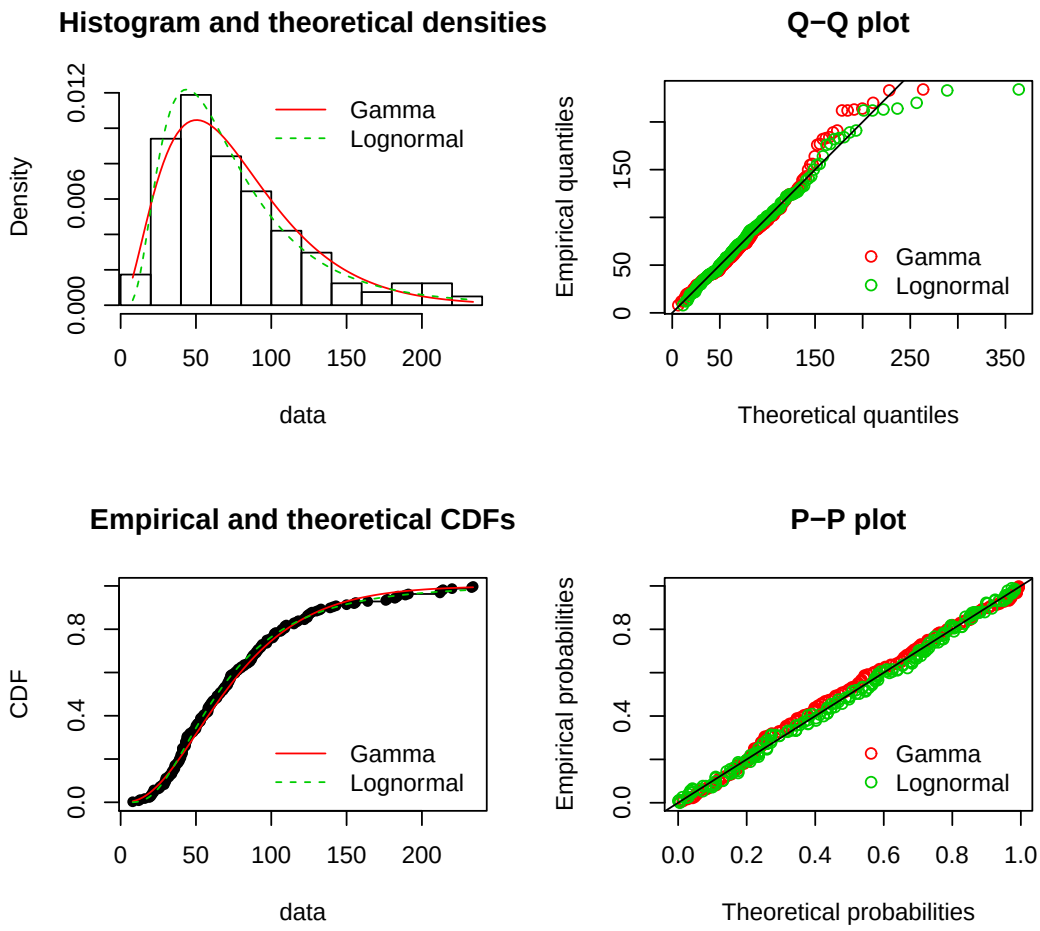
ضمیمه



شکل ۱: بافت‌نگار و نمودار جعبه‌ای برای داده‌های غلظت فری‌تین.



شکل ۲: نمودار چولگی–کشیدگی برای داده‌های غلظت فری‌تین.



شکل ۳: نمودارهای نیکویی برازش توزیع‌های گاما و لگ‌نرمال برای داده‌های غلظت فری‌تین.

فهرست منابع

[1] Borlinhas, F., Loução, R., Conceição, R.C., and Ferreira, H.A. (2019). Gamma Distribution Model in the Evaluation of Breast Cancer Through Diffusion-Weighted MRI: A Preliminary Study. Journal of Magnetic Resonance Imaging 50, 230-238.

[2] Cullen, A.C., and Frey, H.C. (1999). Probabilistic Techniques in Exposure Assessment (1st Edition). Plenum Publishing Co.

[3] Delignette-Muller, M.L., and Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software 64, 1–34.

[4] Neti, P.V.S.V., and Howell, R.W. (2006). Log Normal Distribution of Cellular Uptake of Radioactivity: Implications for Biologic Responses to Radiopharmaceuticals. Journal of Nuclear Medicine 47, 1049-1058

[5] Parker, K.J., and Poul, S.S. (2020). Burr, Lomax, Pareto, and Logistic Distributions from Ultrasound Speckle. Ultrasonic Imaging 42, 203-212.

[6] Sharma, V.K., Singh, S.K., Singh, U., and Agiwal, V. (2015) The inverse Lindley distribution: a stress-strength reliability model with application to head and neck cancer data. Journal of Industrial and Production Engineering 32, 162-173.

[7] Skriver, M.V., Væth, M., and Støvring, H. (2018). Loss of life expectancy derived from a standardized mortality ratio in Denmark, Finland, Norway and Sweden. Scandinavian Journal of Public Health 46, 767-773.

[8] Tse, Y.K. (2009). Nonlife Actuarial Models: Theory, Methods and Evaluation. International Series on Actuarial Science (1st Edition). Cambridge University Press.

[9] Vose, D. (2010). Quantitative Risk Analysis: A Guide to Monte Carlo Simulation Modelling (1st Edition). John Wiley & Sons.

[10] Wang, R., He, N., Wang, Y., and Lu, K. (2020). Adaptively weighted nonlocal means and TV minimization for speckle reduction in SAR images. Multimedia Tools and Applications 79, 7633–7647.

Fitting probability distributions using R software and its application in medicine

M. Mahdizadeh[1, ‡‡‡] **, E. Mahdizadeh**[2]

[(1)] Department of Statistics, Hakim Sabzevari University, Sabzevar, Iran
[(2)] Mehr-e-Madar Hospital, Torbat Jam Faculty of Medical Sciences, Torbat Jam, Iran

Communicated by: Saeed Ghahramani

Received: 2022/9/10 Accepted: 2023/2/20

Abstract: Researchers in different disciplines often face phenomena of random nature. Sometimes it is possible to use probability distributions to describe and predict such phenomena. Each distribution has a number of unknown parameters, whose values are estimated from data. In some problems, there are a few competing distributions for fitting to a data set. In this setup, selecting a suitable model based on some criteria is necessary. This article introduces facilities of R statistical software in performing the above steps. Application of the discussed methods is illustrated using a medical data set.

Keywords: Gamma distribution, Lognormal distribution, Model selection, Parameter estimation.

‡‡‡Corresponding author.
E-mail addresses: mahdizadeh.m@hsu.ac.ir;mahdizadeh.m@live.com (M. Mahdizadeh).