محاسبات آماری پیشرفته

ترم اول سال تحصیلی ۹۳

جلسه دهم: روش‌های خودگردان (*Bootstrap Methods*)

حسین باغیشنی

دانشگاه شاهرود

۲ آذر ۱۳۹۳

استنباط آماری با مشاهدات ناقص

آمار، مطالعه روش‌های استخراج استنباط‌هایی از داده‌های ناقص و محدود شده است. مثلا مایلیم بدانیم:

- چطور یک نرون در مغز یک موش به کشیده شدن یکی از موهای صورتش، پاسخ می‌دهد

استنباط آماری با مشاهدات ناقص

آمار، مطالعه روش‌های استخراج استنباط‌هایی از داده‌های ناقص و محدود شده است. مثلا مایلیم بدانیم:

- چطور یک نرون در مغز یک موش به کشیده شدن یکی از موهای صورتش، پاسخ می‌دهد
- چند تا موش در جنگل ابر زندگی می‌کنند

استنباط آماری با مشاهدات ناقص

آمار، مطالعه روش‌های استخراج استنباط‌هایی از داده‌های ناقص و محدود شده است. مثلا مایلیم بدانیم:

- چطور یک نرون در مغز یک موش به کشیده شدن یکی از موهای صورتش، پاسخ می‌دهد
- چند تا موش در جنگل ابر زندگی می‌کنند
- در اردیبهشت، تا چه ارتفاعی آب رودخانه زاینده‌رود زیر پل خواجو بالا خواهد آمد

استنباط آماری با مشاهدات ناقص

آمار، مطالعه روش‌های استخراج استنباط‌هایی از داده‌های ناقص و محدود شده است. مثلا مایلیم بدانیم:

- چطور یک نرون در مغز یک موش به کشیده شدن یکی از موهای صورتش، پاسخ می‌دهد
- چند تا موش در جنگل ابر زندگی می‌کنند
- در اردیبهشت، تا چه ارتفاعی آب رودخانه زاینده‌رود زیر پل خواجو بالا خواهد آمد
- متوسط دمای هوا در شاهرود در طول سال، چقدر است
- ...

عدم قطعیت

برای همه این چیزها، مجموعه‌ای از مشاهدات داریم. اما می‌دانیم

عدم قطعیت

برای همه این چیزها، مجموعه‌ای از مشاهدات داریم. اما می‌دانیم

- داده‌ها ناقص هستند
- تکرار آزمایش یا مشاهدات، حتی اگر نهایت سعی خود را برای ثابت نگه داشتن شرایط به کار ببریم، همیشه کم و بیش نتایج متفاوتی خواهد داشت

عدم قطعیت

برای همه این چیزها، مجموعه‌ای از مشاهدات داریم. اما می‌دانیم

- داده‌ها ناقص هستند
- تکرار آزمایش یا مشاهدات، حتی اگر نهایت سعی خود را برای ثابت نگه داشتن شرایط به کار ببریم، همیشه کم و بیش نتایج متفاوتی خواهد داشت

بنابراین مضحک است اگر استنباط استخراج‌شده از داده‌ها را قطعی بدانیم.

مدل‌های تصادفی و تابعی‌ها

اگرچه تکرار یک آزمایش نتایج متفاوتی به دست می‌دهد، بعضی از نتایج نسبت به سایرین بیشتر رخ می‌دهند.

و فراونی‌های نسبی این پیشامدها، پایدار هستند.

بنابراین، مکانیسم تولید داده‌ها را می‌توان به وسیله توزیع‌های احتمالی و فرآیندهای تصادفی، مدل‌بندی کرد.

مدل‌های تصادفی و تابعی‌ها

اگرچه تکرار یک آزمایش نتایج متفاوتی به دست می‌دهد، بعضی از نتایج نسبت به سایرین بیشتر رخ می‌دهند.

و فراونی‌های نسبی این پیشامدها، پایدار هستند.

بنابراین، مکانیسم تولید داده‌ها را می‌توان به وسیله توزیع‌های احتمالی و فرآیندهای تصادفی، مدل‌بندی کرد.

کمیت‌هایی مانند مثال‌هایی که ذکر کردیم، به صورت توابعی از مدل تصادفی، یعنی توزیع احتمال زیربنایی، نمایش داده می‌شوند.

چون یک تابع از یک تابع را تابعی می‌نامند، و کمیت‌های مورد نظر توابعی از تابع توزیع احتمال واقعی هستند، آن‌ها را تابعی‌های آماری (پارامتر)، $\theta(F)$، نیز می‌نامند.

تابعی‌ها می‌توانند عدد حقیقی (مانند تعداد کل موش‌ها)، بردار یا کل یک منحنی (منحنی رگرسیون قد افراد بر روی وزن)، باشند.

استنباط آماری، برآورد کردن این تابعی‌ها یا آزمودن فرضیه‌هایی در مورد آن‌هاست.

مدل‌های تصادفی، عدم قطعیت، استنباط آماری

برآورد پارامترها و سایر استنباط‌ها، توابعی از مقادیر داده هستند.

این به آن معنی است که: آن‌ها عدم‌قطعیت همراه با فرآیند تصادفی زیربنایی را به ارث می‌برند.

اگر آزمایش تکرار شود، ما داده‌های متفاوتی خواهیم داشت اما با یک توزیع قطعی مشخص و اجرای یک روش استنباطی ثابت، نتایج استنباطی متفاوتی در بر خواهد داشت اما دوباره با یک توزیع قطعی مشخص.

مدل‌های تصادفی، عدم قطعیت، استنباط آماری

برآورد پارامترها و سایر استنباط‌ها، توابعی از مقادیر داده هستند.

این به آن معنی است که: آن‌ها عدم‌قطعیت همراه با فرآیند تصادفی زیربنایی را به ارث می‌برند.

اگر آزمایش تکرار شود، ما داده‌های متفاوتی خواهیم داشت اما با یک توزیع قطعی مشخص و اجرای یک روش استنباطی ثابت، نتایج استنباطی متفاوتی در بر خواهد داشت اما دوباره با یک توزیع قطعی مشخص.

آماردانان مایل به استفاده از این توزیع برای کمی‌سازی عدم قطعیت همراه با استنباط‌ها هستند:

مدل‌های تصادفی، عدم قطعیت، استنباط آماری

برآورد پارامترها و سایر استنباط‌ها، توابعی از مقادیر داده هستند.

این به آن معنی است که: آن‌ها عدم‌قطعیت همراه با فرآیند تصادفی زیربنایی را به ارث می‌برند.

اگر آزمایش تکرار شود، ما داده‌های متفاوتی خواهیم داشت اما با یک توزیع قطعی مشخص و اجرای یک روش استنباطی ثابت، نتایج استنباطی متفاوتی در بر خواهد داشت اما دوباره با یک توزیع قطعی مشخص.

آماردانان مایل به استفاده از این توزیع برای کمی‌سازی عدم قطعیت همراه با استنباط‌ها هستند:

- خطای معیار، پاسخی است به این سوال که تا چه حد برآورد پارامتر مورد نظر از یک تکرار آزمایش به دیگری، تغییر می‌کند؟
- یک ناحیه اطمینان برای پارامتر، پاسخ به این سوال است که همه مقادیری از پارامتر که این داده‌ها را، با حداقل یک احتمال مشخص، می‌توانند تولید کنند، کدامند؟

مدل‌های تصادفی، عدم قطعیت، توزیع‌های نمونه‌ای

برای داشتن چیزهایی مانند خطای استاندارد و فاصله اطمینان، نیازمند دانستن توزیع برآوردگرهای به دست آمده هستیم.

به توزیع برآوردگرها، توزیع‌های نمونه‌ای گویند.

توزیع‌های نمونه‌ای از توزیع داده‌ها پیروی می‌کنند، زیرا برآوردگرها توابعی از داده‌ها هستند.

مدل‌های تصادفی، عدم قطعیت، توزیع‌های نمونه‌ای

برای داشتن چیزهایی مانند خطای استاندارد و فاصله اطمینان، نیازمند دانستن توزیع برآوردگرهای به دست آمده هستیم.

به توزیع برآوردگرها، توزیع‌های نمونه‌ای گویند.

توزیع‌های نمونه‌ای از توزیع داده‌ها پیروی می‌کنند، زیرا برآوردگرها توابعی از داده‌ها هستند.

از دیدگاه ریاضی، با یک مساله خوش تعریف روبه‌رو هستیم. اما محاسبه داستان دیگری است!!!

مدل‌های تصادفی، عدم قطعیت، توزیع‌های نمونه‌ای

برای داشتن چیزهایی مانند خطای استاندارد و فاصله اطمینان، نیازمند دانستن توزیع برآوردگرهای به دست آمده هستیم.

به توزیع برآوردگرها، توزیع‌های نمونه‌ای گویند.

توزیع‌های نمونه‌ای از توزیع داده‌ها پیروی می‌کنند، زیرا برآوردگرها توابعی از داده‌ها هستند.

از دیدگاه ریاضی، با یک مساله خوش تعریف روبه‌رو هستیم. اما محاسبه داستان دیگری است!!!

معمولا، برآوردگرها توابع پیچیده‌ای از داده‌ها هستند و صحبت در مورد محاسبه شکل‌های بسته برای توزیع آن‌ها، ناامیدکننده است.

در این موارد، دو راهکار کلاسیک آماردانان عبارتند از:

- تمرکز بر روی موارد خاص و ساده
- استفاده از نظریه مجانبی توزیع‌ها

مدل‌های تصادفی، عدم قطعیت، توزیع‌های نمونه‌ای، محاسبات

تا حدود دهه ۶۰ میلادی، علم آمار در توسعه دو راهکار بالا متمرکز شده بود.

اما با انقلابی که دنیای انفورماتیک و کامپیوترها ایجاد کردند، باعث به کار آمدن مدل‌های آماری پیچیده‌تر (و البته واقعی‌تر) و خارج شدن از دنیای کوچک مدل‌های ساده و سرراست شد.

از طرف دیگر، ممکن است نظریه بزرگ‌نمونه برای مدل‌های پیچیده نیز، درمانی داشته باشد، اما همگرایی به توزیع‌های حدی ممکن است به طور غیرقابل قبولی کند باشد.

تا قبل از دهه ۷۰ میلادی، آمار با مساله کمی‌سازی عدم قطعیت استنباط‌ها بدون در نظر گرفتن پذیره‌های غیرمفید و آمار مجانبی، مواجه شد.

همه راه حل‌ها تبدیل شد به محاسبات آماری بیشتر و بیشتر.

یکی از موفق‌ترین راه حل‌ها که توسط بردلی افرون پیشنهاد شد، خودگردان‌سازی، *Bootstrapping*، است.

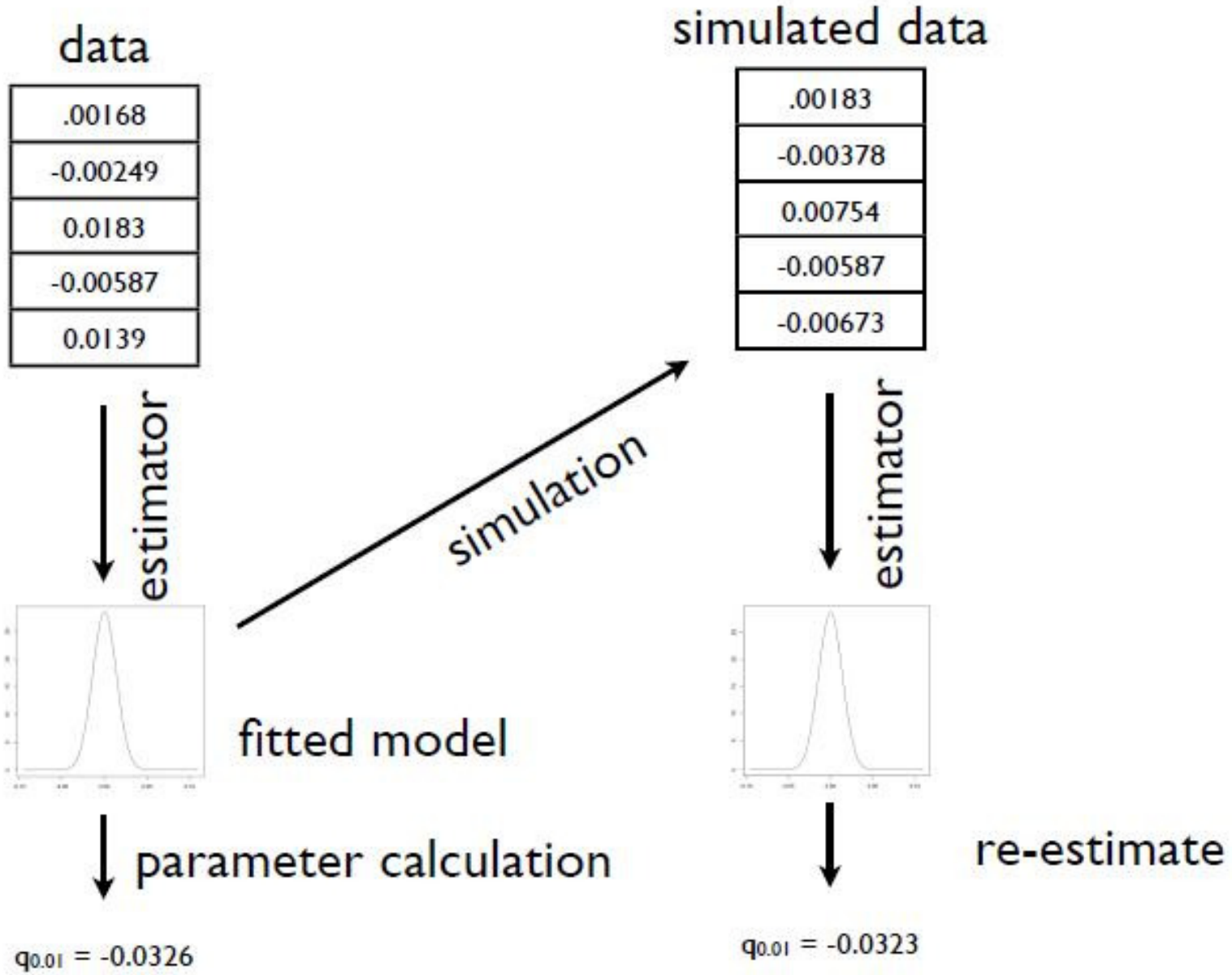
اصل خودگردان‌سازی

کلید بررسی عدم قطعیت در برآوردگرها، توزیع نمونه‌ای آن‌ها می‌باشد.

زیرکی افرون در این بود که گفت می‌توانیم تکرار آزمایش را شبیه‌سازی کنیم:

- مدلی را که حدس می‌زنیم مکانیسم واقعی تولید داده‌ها باشد را به داده‌ها برازش می‌دهیم
- اجرای آن مکانیسم، داده‌های شبیه‌سازی‌شده را تولید می‌کند که دارای توزیع مشابه داده‌های واقعی است
- محاسبه مقدار برآوردگر بر روی داده‌های شبیه‌سازی‌شده، یک تحقق از توزیع نمونه‌ای برآوردگر را نتیجه می‌دهد
- تکرار این فرآیند به تعداد فراوان، توزیع نمونه‌ای برآوردگر را تقریب می‌زند

چون از مدل برای محاسبه عدم قطعیت خودش استفاده می‌شود، افرون آن را خودگردان نامید.


data
.00168
-0.00249
0.0183
-0.00587
0.0139
estimator
simulation
simulated data
.00183
-0.00378
0.00754
-0.00587
-0.00673
estimator
fitted model
parameter calculation
re-estimate
$q_{0.01}$ = -0.0326
$q_{0.01}$ = -0.0323

نمادگذاری

داده‌های اصلی را با X نشان می‌دهیم. دقت کنید در اینجا منظور از X چارچوب داده است نه یک عدد تنها.

پارامتر برآوردشده به وسیله داده‌ها: $\hat{\theta}$

مجموعه داده‌های شبیه‌سازی‌شده از مدل برازش‌شده: $\tilde{X}_1, \ldots, \tilde{X}_B$

برآوردهای خودگردان پارامتر، حاصل از داده‌های شبیه‌سازی‌شده: $\tilde{\theta}_1, \ldots, \tilde{\theta}_B$

علاوه بر این فرض کنید، تابعی مورد نظر، $\theta(F)$، توسط آماره T به صورت $\hat{t} = T(X)$ برآورد شود و به طور مشابه برای داده‌های شبیه‌سازی‌شده:

$$\tilde{t}_1 = T(\tilde{X}_1), \ldots, \tilde{t}_B = T(\tilde{X}_B),$$

$$\tilde{t} = (\tilde{t}_1, \ldots, \tilde{t}_B).$$

آماره T ممکن است تابعی مستقیم از پارامترهای برآوردشده باشد، یا تابعی غیرمستقیم.

در این جا فرض می‌کنیم مدل برای مقداری از θ یک مدل درست است. این مقدار را با θ_0 نشان می‌دهیم. به طور مشابه فرض کنید مقدار درست تابعی، t_0 است.

واریانس و خطای استاندارد

ساده‌ترین معیار برای کمی‌سازی عدم قطعیت، واریانس یا خطای استاندارد است:

$$\begin{aligned} \hat{Var}(\hat{t}) &= Var(\tilde{t}) \\ \hat{se}(\hat{t}) &= sd(\tilde{t}) \end{aligned}$$

منطق این روابط اینست که $\tilde{X}$ شبیه‌سازی‌شده توزیع یکسانی با متغیر واقعی X دارد که داده‌های اصلی، x، از آن استخراج شده‌اند. بنابراین اجرای مشابه روش برآورد بر روی داده‌های شبیه‌سازی‌شده، توزیع نمونه‌ای را نتیجه خواهد داد.

دقت کنید در این جا فرض بر آن است که مدل در نظر گرفته شده درست است و $\hat{\theta}$ خیلی از مقدار واقعی θ. دور نیست.

```
rboot <- function(B, statistic, simulator, ...) {
         tboots <- replicate(B, statistic(simulator(...)))
         return(tboots)
}
bootstrap.se <- function(simulator, statistic, B, ...) {
                tboots <- rboot(B, statistic, simulator, ...)
                se <- sd(tboots)
                return(se)
}
```

تصحیح اریبی

می‌توان از روش خودگردان برای تصحیح اریبی یک برآوردگر اریب، استفاده کرد.

از آن‌جا که توزیع نمونه‌ای $\tilde{t}$ نزدیک به توزیع $\hat{t}$ است، و $\hat{t}$ نیز نزدیک به t_0، پس

$$\mathbb{E}(\hat{t}) - t_0 \approx \mathbb{E}(\tilde{t}) - \hat{t}.$$

سمت چپ تساوی، اریبی هست که به دنبال محاسبه آن هستیم و سمت راست مقداری است که از روی نمونه‌های خودگردان قابل دستیابی است.

دقت کنید که تساوی بالا مادامی معتبر است که توزیع نمونه‌ای $\hat{t} - t_0$ نزدیک به توزیع نمونه‌ای $\tilde{t} - \hat{t}$ باشد. این لزوم، شرط ضعیفتری نسبت به نزدیک بودن توزیع‌های نمونه‌ای $\tilde{t}$ به $\hat{t}$ و $\hat{t}$ به t_0 است.

یک شرط کافی (اما نه لازم) برای برقراری تساوی بالا، آن است که کمیت $\hat{t} - t_0$ محوری یا تقریبا محوری باشد.

```
bootstrap.bias <- function(simulator, statistic, B, t.hat, ...) {
                  tboots <- rboot(B, statistic, simulator, ...)
                  bias <- mean(tboots) - t.hat
                  return(bias)
}
```

فاصله اطمینان

$$Pr(t_{\cdot} \in C) = ۱ - \alpha,$$

در زمان محاسبه فاصله اطمینان، عدم شناخت دقیق توزیع نمونه‌ای به آن معنی است که سطح اطمینان واقعی (نرخ پوشش فاصله) مقدار دلخواه $۱ - \alpha$ نخواهد بود.

هر چه نرخ پوشش تجربی نزدیک‌تر به مقدار اسمی باشد، به معنی تقریب بهتر و در نتیجه فاصله اطمینان دقیق‌تر است.

فرض کنید C_l و C_u به ترتیب کران‌های پایین و بالای فاصله اطمینان باشند. برای فواصل اطمینان با دم‌های برابر داریم:

$$\begin{aligned} \frac{\alpha}{2} &= Pr(C_l \geq t_0) = Pr(C_l - \hat{t} \geq t_0 - \hat{t}) \\ &= Pr(\hat{t} - C_l \leq \hat{t} - t_0) \\ \frac{\alpha}{2} &= Pr(\hat{t} - C_u \geq \hat{t} - t_0) \end{aligned}$$

فاصله اطمینان خودگردانی پایه

روش خودگردانی، توزیع $\tilde{t} - \hat{t}$ را نتیجه می‌دهد که تقریبا با توزیع $\hat{t} - t$، یکی است. با محاسبه این توزیع و $\hat{t}$، می‌توان کران‌های C_l و C_u را به دست آورد:

$$\begin{aligned} C_l &= \hat{t} - \left(Q_{\tilde{t}}(1 - \frac{\alpha}{2}) - \hat{t} \right) \\ C_u &= \hat{t} - \left(Q_{\tilde{t}}(\frac{\alpha}{2}) - \hat{t} \right) \end{aligned}$$

که در آن $Q_{\tilde{t}}$ تابع چندک $\tilde{t}$ است. یعنی چندک‌های توزیع نمونه‌ای نمونه‌های شبیه‌سازی‌شده را به دست می‌دهد.

تمرین: دو رابطه بالا را اثبات کنید.

این فاصله، یک فاصله اطمینان خودگردانی پایه یا فاصله اطمینان مبتنی بر کمیت محوری است. به سادگی قابل محاسبه است و دقت آن هم قابل قبول است.

```
bootstrap.ci.basic <- function(simulator, statistic, B, t.hat,
                               alpha, ...) {
   tboots <- rboot(B,statistic, simulator, ...)
   ci.lower <- 2*t.hat - quantile(tboots,1-alpha/2)
   ci.upper <- 2*t.hat - quantile(tboots,alpha/2)
   return(list(ci.lower=ci.lower,ci.upper=ci.upper))
}
```

فواصل اطمینان خودگردانی استودنت‌شده

فاصله اطمینان خودگردانی پایه، مبتنی بر پذیره تقریبا یکسان بودن توزیع $\tilde{t} - \hat{t}$ با توزیع $\hat{t} - t_0$ است.

اما حتی زمانی که این پذیره نادرست باشد، توزیع

$$\tau = \frac{\hat{t} - t_0}{\hat{se}(\hat{t})}$$

نزدیک به توزیع

$$\tilde{\tau} = \frac{\tilde{t} - \hat{t}}{se(\tilde{t})}$$

خواهد بود. این کمیت‌ها شبیه آماره t در آزمون tـ استودنت هستند و چون آزمون t توسط استودنت کشف شد، به آن‌ها کمیت‌های استودنت‌شده می‌گویند.

فواصل اطمینان خودگردانی استودنت‌شده: ادامه

اگر τ و $\tilde{\tau}$ توزیع یکسانی داشته باشند، می‌توان فاصله اطمینان زیر را به دست آورد:

$$(\hat{t} - \hat{se}(\hat{t})\, Q_{\tilde{\tau}}(۱ - \frac{\alpha}{۲}), \hat{t} - \hat{se}(\hat{t})\, Q_{\tilde{\tau}}(\frac{\alpha}{۲}))$$

این فاصله با فاصله پایه یکی خواهد بود هرگاه $\hat{se}(\hat{t}) = se(\tilde{t})$ و در غیر این صورت متفاوت خواهند بود.

برای محاسبه $se(\tilde{t})$ نیاز به سطح دومی از خودگردان‌سازی داریم. به الگوریتم زیر دقت کنید:

فواصل اطمینان خودگردانی استودنت‌شده: الگوریتم

۱. مدل را با $\hat{\theta}$ برازش بده و $\hat{t}$ را محاسبه کن

۲. برای $i = ۱, \ldots, B_۱$:

الف) $\tilde{X}_i$ را از $\hat{\theta}$ تولید کن

ب) مقادیر $\tilde{\theta}_i$ و $\tilde{t}_i$ را محاسبه کن

ج) برای $j = ۱, \ldots, B_۲$:

۱. $X^{\dagger}_{ij}$ از $\tilde{\theta}_i$ تولید کن

۲. $t^{\dagger}_{ij}$ را محاسبه کن

د) $\tilde{\sigma}_i$ را برابر انحراف استاندارد $t^{\dagger}_{ij}$ها قرار بده

ه) برای تمام jها قرار بده $\tilde{\tau}_{ij} = \frac{t^{\dagger}_{ij} - \tilde{t}_i}{\tilde{\sigma}_i}$

۳. $\hat{se}(\hat{t})$ را برابر انحراف استاندارد $\tilde{t}_i$ها قرار بده

۴. چندک‌های $\alpha/۲$ و $۱ - \alpha/۲$ توزیع $\tilde{\tau}$ را محاسبه کن

۵. همه کمیت‌های مورد نظر را در فاصله اطمینان جایگذاری کن

فواصل اطمینان خودگردانی چندکی

مزیت فواصل استودنت‌شده نسبت به فواصل پایه، دقت بالاترشان است.

عیب آن‌ها نیز زمان‌بر بودن محاسبات است.

نوع دیگری از فواصل اطمینان خودگردانی، فواصل چندکی است که به راحتی به صورت زیر تشکیل می‌شود:

$$(Q_{\tilde{t}}(\alpha/۲), Q_{\tilde{t}}(۱ - \alpha/۲)).$$

این نوع فواصل اطمینان به سادگی قابل محاسبه‌اند، اما دقیق نیستند.

البته همه این فواصل اطمینان مطرح‌شده، صورت‌های مختلفی دارند که توسط افراد مختلف پیشنهاد شده‌اند.

آزمون فرضیه خودگردانی

برای آزمون فرضیه‌ها، دو توزیع نمونه‌ای متفاوت را محاسبه می‌کنیم:

- توزیع آماره آزمون تحت فرضیه صفر که برای محاسبه اندازه آزمون و سطح معنی‌داری است
- توزیع آماره آزمون تحت فرضیه جانشین که برای محاسبه توان آزمون است

هر دو توزیع با روش خودگردان‌سازی قابل محاسبه هستند.

در آزمون فرضیه، آماره مورد نظر، t، آماره آزمون می‌باشد.

```
boot.pvalue <- function(test,simulator,B,testhat, ...) {
   testboot <- rboot(B=B, statistic=test, simulator=simulator, ...)
   p <- (sum(test >= testhat)+1)/(B+1)
   return(p)
}
```

این کد، pـ مقدار خودگردانی را برای یک آزمون محاسبه می‌کند. دقت کنید که $testhat$ مقدار آماره آزمون برای داده‌های اصلی است و $test$ تابعی است که معرف آماره آزمون است.

برای محاسبه توان آزمون به روش مشابه می‌توان عمل کرد. فقط باید نمونه‌های خودگردانی تحت فرضیه جانشین تولید شوند و $testhat$ مقدار بحرانی آزمون باشد نه مقدار مشاهده‌شده آماره آزمون.

مثال: قانون نابرابری ثروت پارتو

توزیع پارتو (توزیع قانون توانی) یک مدل معمول برای داده‌های با دم‌های سنگین است. به این معنی که چگالی احتمال $f(x)$ زمانی که $x \longrightarrow \infty$ خیلی کند به صفر میل می‌کند. یا به عبارت دیگر، این توزیع به شدت به راست چوله است و میانگین آن خیلی بزرگتر از میانه است.

$$f(x) = \frac{\theta - ۱}{x_۰} \left(\frac{x}{x_۰} \right)^{-\theta}$$

که در آن $x_۰$ مقیاس می‌نیمم توزیع است.

تمرین. نشان دهید که $x_۰$ مد توزیع پارتو است.

اگر $x_۰$ معلوم باشد، آن گاه:

$$\hat{\theta} = ۱ + \frac{n}{\sum_{i=۱}^{n} \log \frac{x_i}{x_۰}},$$

که برآوردگری سازگار و کاراست.

مثال پارتو: ادامه

فایل $pareto.R$ شامل تعدادی تابع مرتبط با توزیع پارتو است. یکی از توابع آن $pareto.fit$ است که مدل بالا را به داده‌ها برازش می‌دهد.

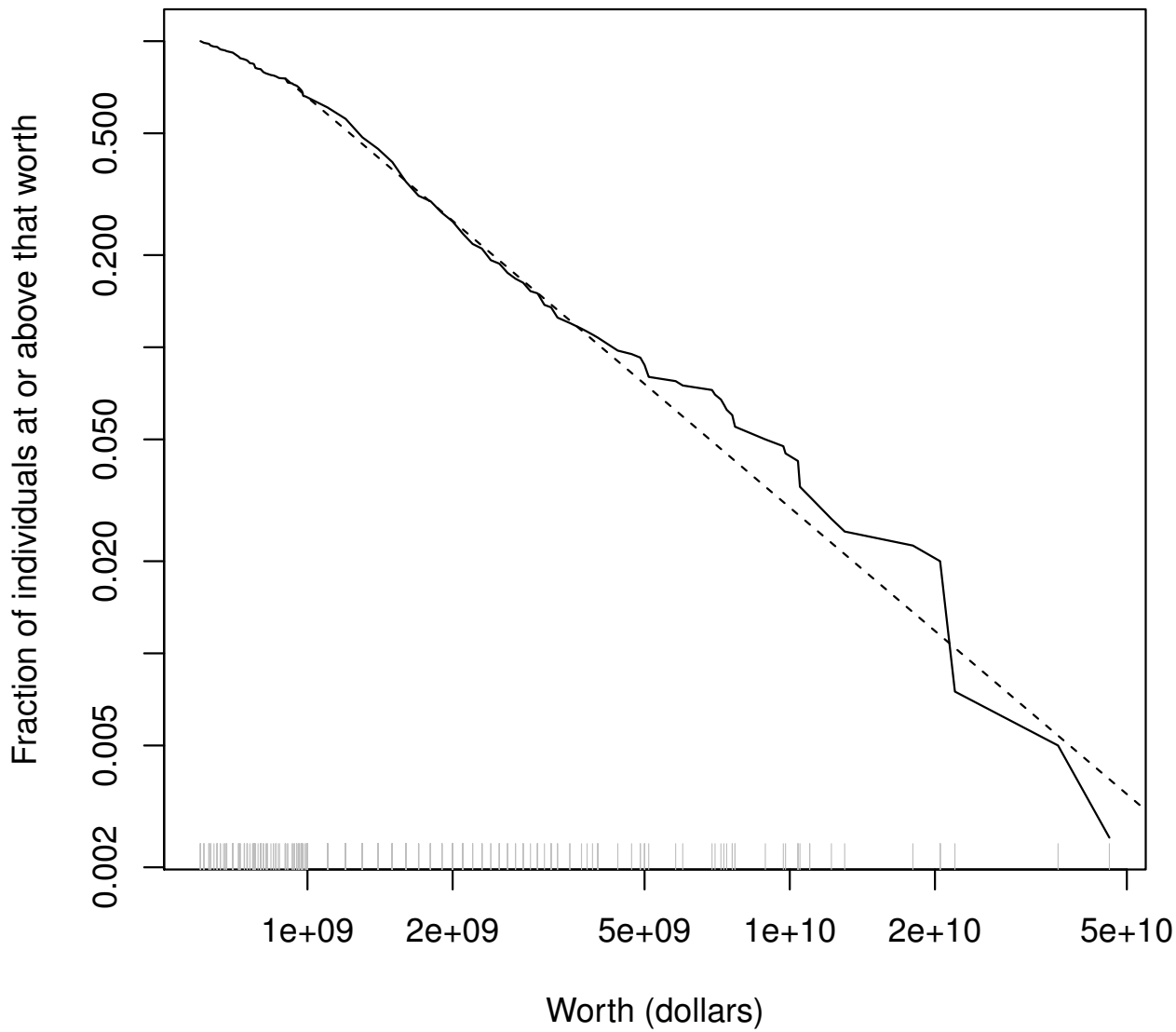
پارتو، وقتی که می‌خواست توزیع داده‌های ثروت را مدل‌بندی کند، به چگالی پارتو رسید.

تقریبا در کشورهای مختلف و زمان‌های متفاوت، دم بالایی توزیع درآمد و ثروت از یک قانون توانی پیروی می‌کند.

پارامتر θ در این توزیع با مقدار پولی که کم و بیش در دست ثروتمندترین مردم کشور هست، تغییر می‌کند.

شکل صفحه بعد، توزیع ثروت را برای ۴۰۰ پولدار اول آمریکا در سال ۲۰۰۳ نشان می‌دهد. با در نظر گرفتن $x_. = ۹ \times ۱۰^۸$، تعداد افرادی که در دم قرار می‌گیرند، ۳۰۲ نفر است و $\hat{\theta} = ۲/۳۴$:

```
source("pareto.R")
wealth <- scan("wealth.dat")
wealth.pareto <- pareto.fit(wealth,threshold=9e8)
> signif(wealth.pareto$exponent,3)
[1] 2.34
```


Fraction of individuals at or above that worth
0.002
0.005
0.020
0.050
0.200
0.500
1e+09
2e+09
5e+09
1e+10
2e+10
5e+10
Worth (dollars)

مثال پارتو: ادامه

چه میزان عدم قطعیت در این برآورد θ وجود دارد؟

با روش خودگردانی می‌توان به این سوال پاسخ داد:

```
rboot.pareto <- function(B,exponent,x0,n) {
   replicate(B,pareto.fit(rpareto(n,x0,exponent),x0)$exponent)
}
pareto.se <- function(B,exponent,x0,n) {
   return(sd(rboot.pareto(B,exponent,x0,n)))
}
pareto.bias <- function(B,exponent,x0,n) {
   return(mean(rboot.pareto(B,exponent,x0,n)) - exponent)
}
```

با $\hat{\theta} = ۲/۳۴$، $x_۰ = ۹ \times ۱۰^۸$، $n = ۳۰۲$ و $B = ۱۰^۴$، خطای معیار (عدم قطعیت برآورد) برابر ۰/۰۷۷ به دست می‌آید که با نتایج مجانبی هم هم‌خوانی دارد.

به طور مجانبی نیز (با توجه به سازگاری برآوردگر) اریبی به سمت صفر میل می‌کند. میزان اریبی خودگردان نیز برابر ۴×۱۰^{-۳} به دست آمده است که قابل صرفنظر است.

مثال پارتو: ادامه

می‌توان فاصله اطمینان خودگردانی پایه را نیز محاسبه کرد:

```
pareto.ci <- function(B,exponent,x0,n,alpha) {
   tboot <- rboot.pareto(B,exponent,x0,n)
   ci.lower <- 2*exponent - quantile(tboot,1-alpha/2)
   ci.upper <- 2*exponent - quantile(tboot,alpha/2)
   return(list(ci.lower=ci.lower, ci.upper=ci.upper))
}
```

با استفاده از همان مشخصات قبلی، فاصله ٪۹۵ پایه به صورت $(۲/۱۸, ۲/۴۸)$ به دست می‌آید.

روش خودگردان‌سازی ناپارامتری

خودگردان‌سازی، توزیع نمونه‌ای را همراه با سه منبع خطای برآورد، تقریب می‌زند:

۱. خطای شبیه‌سازی: استفاده از تعداد تکرار فراوان اما متناهی برای دستیابی به توزیع نمونه‌ای کامل. تعداد کافی تکرار و طراحی مناسب شبیه‌سازی، می‌تواند این خطا را به اندازه دلخواه کوچک کند.

۲. خطای آماری: توزیع نمونه‌ای پارامترهای برآوردشده خودگردانی تحت مدل برازش‌شده، دقیقا با توزیع نمونه‌ای پارامترهای برآوردشده تحت مدل واقعی مکانیسم تولید داده‌ها یکی نیست. توزیع نمونه‌ای با تغییر پارامترها، تغییر می‌کند و برآورد اولیه ما کاملا دقیق نیست. اما اغلب مشخص شده است که توزیع برآوردها حول مقدار واقعی پایاتر از توزیع خود برآوردهاست. بنابراین کم کردن برآورد اولیه از مقادیر خودگردان، باعث کاهش خطای آماری می‌شود.

۳. خطای مشخص‌سازی: داده‌ها از مدل‌هایی که ما در نظر می‌گیریم دقیقا پیروی نمی‌کنند. بنابراین شبیه‌سازی مدل هرگز با توزیع نمونه‌ای واقعی یکی نمی‌شود.

خطاها

خطای آماری:

$$F_n \longrightarrow \tilde{X} \longrightarrow F_n^*,$$

$$F_n \longrightarrow \tilde{t} \longrightarrow F_n^*$$

خطای مشخص‌سازی:

$$F \longrightarrow X \longrightarrow F_n$$

$$F \longrightarrow t \longrightarrow F_n$$

برای درک خطای آماری، به مثال ۱ ۰ ۷ کتاب توجه کنید.

روش خودگردان‌سازی ناپارامتری: ادامه

افرون، یک ایده معرکه دومی هم داشت، که مربوط به خطای مشخص‌سازی می‌شد.

وی پیشنهاد کرد برای فرار از این خطا، به جای شبیه‌سازی از مدل، از خود داده‌ها (یا به عبارتی توزیع تجربی داده‌ها) بازنمونه‌گیری کنیم.

در واقع خود داده‌ها (یا توزیع تجربی داده‌ها) آماره بسنده برای پارامتر است.

از دیدگاه دیگر، توزیع تجربی داده‌ها کم‌تعصب‌ترین برآورد ممکن از توزیع واقعی زیربنایی است. هر چیز دیگری اریبی و پیش‌داوری وارد می‌کند. ممکن است مدل انتخاب‌شده دقیق باشد، اما احتمال نامناسب بودن آن نیز کم نیست.

بسیاری از کمیت‌ها را می‌توان به طور مستقیم، و بدون وساطت یک مدل پارامتری، از روی توزیع تجربی برآورد کرد.

خودگردان‌سازی ناپارامتری افرون، مجموعه داده اصلی را به عنوان یک جامعه متناهی در نظر می‌گیرد و یک نمونه شبیه‌سازی‌شده جدید، که در آن هر مشاهده با احتمال برابر انتخاب می‌شود، از آن استخراج می‌کند. سپس برآورد را با نمونه جدید تکرار می‌کند.

در واقع، معمولا وقتی آماردانان اسم خودگردان را می‌آورند، منظورشان این نوع خودگردان است.

الگوریتم خودگردان‌سازی ناپارامتری

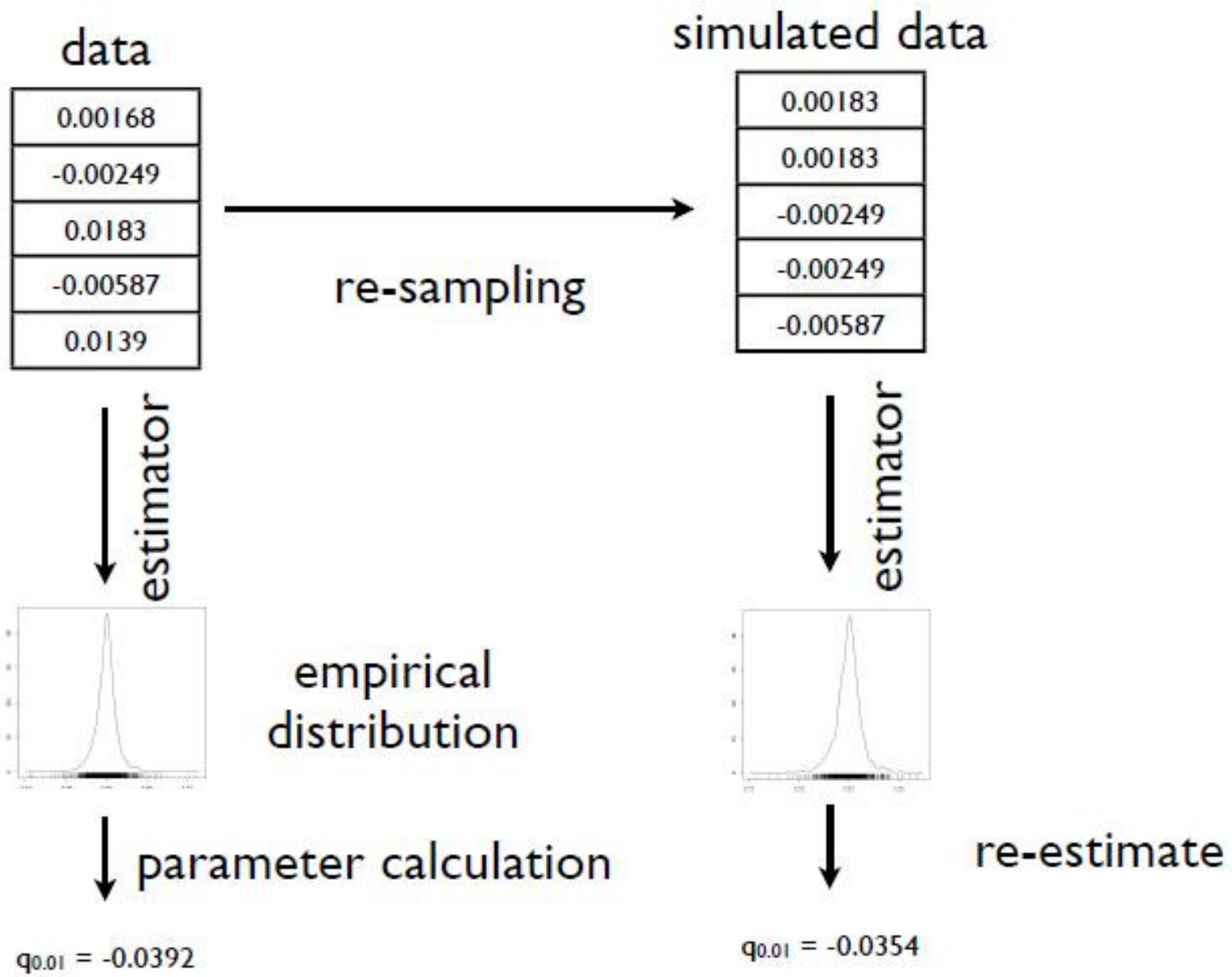
توزیع نمونه‌ای $\hat{t}$ را می‌توان به کمک روش خودگردان‌سازی به صورت زیر تقریب زد:

۱. برای هر تکرار خودگردان $b = ۱, \ldots, B$:

الف) نمونه $\tilde{x}^{(b)} = (\tilde{x}_۱, \ldots, \tilde{x}_n)$ را با نمونه‌گیری با جایگذاری از نمونه مشاهده شده $x_۱, \ldots, x_n$ تولید کن

ب) مقادیر $\tilde{\theta}^{(b)}$ و $\tilde{t}^{(b)}$ را محاسبه کن

۲. برآورد توزیع نمونه‌ای خودگردان، F_n^*، توزیع تجربی $\tilde{t}^{(B)}, \ldots, \tilde{t}^{(۱)}$ می‌باشد


data
0.00168
-0.00249
0.0183
-0.00587
0.0139
re-sampling
simulated data
0.00183
0.00183
-0.00249
-0.00249
-0.00587
estimator
estimator
empirical
distribution
parameter calculation
re-estimate
$q_{0.01}$ = -0.0392
$q_{0.01}$ = -0.0354

روش خودگردان‌سازی ناپارامتری: ادامه

هر آنچه که با خودگردان‌سازی پارامتری مطرح شد، با خودگردان‌سازی ناپارامتری نیز قابل طرح و انجام است.

تنها تفاوت در توزیع داده‌های شبیه‌سازی‌شده خودگردانی است که همان توزیع تجربی داده‌های اصلی است.

```
resample <- function(x) { sample(x,size=length(x),replace=TRUE) }
resamp.pareto <- function(B,data,x0) {
   replicate(B,pareto.fit(resample(data),threshold=x0)$exponent)
}
> sd(resamp.pareto(B,wealth,x0))
[1] 0.07762305
> mean(resamp.pareto(B,wealth,x0))-exponent
[1] 0.003534574
resamp.pareto.CI <- function(B,data,alpha,x0) {
   thetahat <- pareto.fit(data,threshold=x0)$exponent
   thetaboot <- resamp.pareto(B,data,x0)
   ci.lower <- thetahat - (quantile(thetaboot,1-alpha/2) - thetahat)
   ci.upper <- thetahat - (quantile(thetaboot,alpha/2) - thetahat)
   return(list(ci.lower=ci.lower,ci.upper=ci.upper))
}
> resamp.pareto.CI(B,wealth,0.05,x0)
$ci.lower
   97.5%
2.173528

$ci.upper
    2.5%
2.476391
```

خودگردان‌سازی پارامتری در مقابل ناپارامتری

اگر یک مدل مشخص (پارامتری) مناسب داشته باشیم، شبیه‌سازی از آن مدل نتایج دقیق‌تری را، با حجم نمونه ثابت n، نسبت به بازنمونه‌گیری از توزیع تجربی داده‌ها (ناپارامتری) به دست خواهد داد.

در واقع، در این حالت، برآورد پارامتری توزیع سریع‌تر از توزیع تجربی، به توزیع واقعی همگرا می‌شود.

در مقابل اگر مدل پارامتری به اشتباه مشخص شده باشد، به سرعت به یک توزیع نادرست همگرا می‌شود. در چنین مواردی استفاده از روش ناپارامتری، نتایج خیلی بهتری به دست خواهد داد.

معمولا، چون در اغلب کاربردها نسبت به پذیره‌های مدل‌های پارامتری تردید زیادی وجود دارد، ترجیح داده می‌شود از روش‌های بازنمونه‌گیری استفاده شود. مگر آن که بتوان قانع شد که یک مدل پارامتری، تقریب خیلی خوبی از واقعیت پدیده تصادفی خواهد بود.

مثال: مجموعه داده آزمون ورودی دانشکده حقوق

در بسته $bootstrap$ در R، که مربوط به مباحث و مثال‌های کتاب افرون و تیبشیرانی است، مجموعه داده‌ای به نام داده‌های دانشکده حقوق وجود دارد که شامل معدل نمره آزمون‌های ورودی ۸۲ دانشکده حقوق، $LSAT$، و معدل نمرات دبیرستان شرکت‌کنندگان در آزمون، GPA، است.

این داده‌ها در این بسته با نام $law۸۲$ قابل دسترسی است. یک نمونه ۱۵ تایی از این داده‌ها در صفحه ۱۸۵ کتاب آورده شده است.

هدف برآورد ضریب وابستگی بین این دو متغیر و محاسبه برآورد خودگردانی خطای استاندارد این برآورد می‌باشد.

در این مثال تابعی t، ضریب همبستگی بین دو متغیر $LSAT$ و GPA می‌باشد.

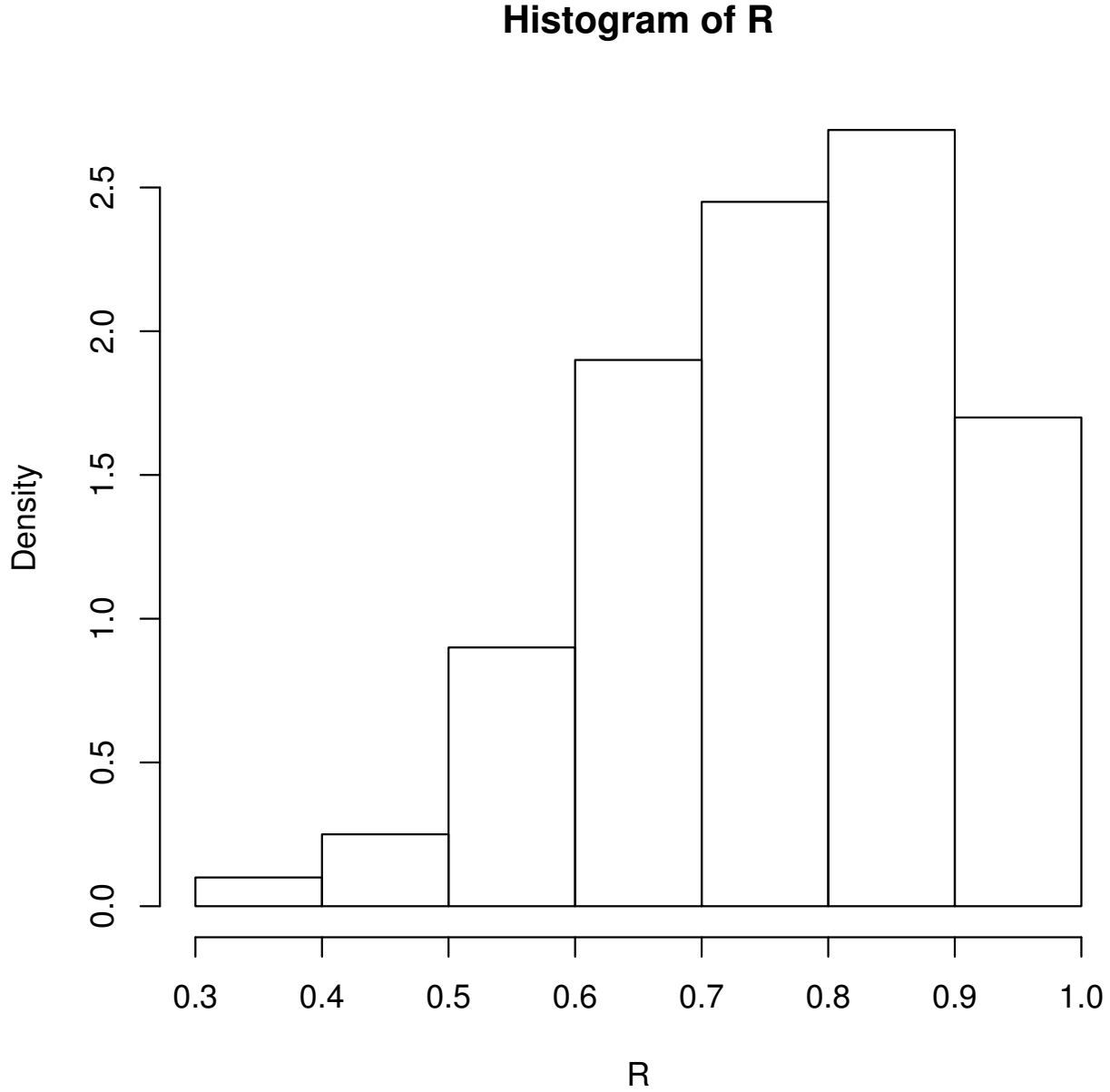
```
library(bootstrap)  # for the law data
> print(cor(law$LSAT, law$GPA)) # for a sample with size 15
[1] 0.7763745
> print(cor(law82$LSAT, law82$GPA))
[1] 0.7599979
```

ادامه مثال: مجموعه داده آزمون ورودی دانشکده حقوق

```
B <- 200  # number of replicates
n <- nrow(law) # sample size
R <- numeric(B) # storage for replicates
# bootstrap estimate of standard error of R
for (b in 1:B) {
    # randomly select the indices
    i <- sample(1:n, size = n, replace = TRUE)
    LSAT <- law$LSAT[i]       #i is a vector of indices
    GPA <- law$GPA[i]
    R[b] <- cor(LSAT, GPA)
}
# output
> print(se.R <- sd(R))
[1] 0.1348793
```

برآورد خطای استاندارد بر اساس نظریه نرمال برابر ۰/۱۱۵ است. (چرا؟)

تمرین: با تابع *boot* در بسته *boot* می‌توان محاسبات بالا را انجام داد. برای مثال آزمون ورودی دانشکده حقوق، محاسبات را تکرار کنید و با نتایج بالا مقایسه کنید.


Histogram of R
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
R

ادامه مثال: مجموعه داده آزمون ورودی دانشکده حقوق

برای برآورد اریبی نیز می‌توان مشابه مثال‌های قبلی عمل کرد:

```
# sample estimate for n=15
theta.hat <- cor(law$LSAT, law$GPA)
# bootstrap estimate of bias
B <- 2000 # larger for estimating bias
n <- nrow(law)
theta.b <- numeric(B)
for (b in 1:B) {
     i <- sample(1:n, size = n, replace = TRUE)
     LSAT <- law$LSAT[i]
     GPA <- law$GPA[i]
     theta.b[b] <- cor(LSAT, GPA)
}
bias <- mean(theta.b - theta.hat)
> bias
[1] -0.002769211
```

برآورد اریبی برای یک برآوردگر نسبت

داده‌های $patch$ نیز در بسته $bootstrap$ قرار دارد که توضیح کامل آن در کتاب افرون و تیبشیرانی (۱۹۹۳) هست.

این داده‌ها شامل اندازه‌های یک هرمون خاص در گردش خون ۸ آزمودنی بعد از پانسمان پزشکی است. پارامتر مورد نظر عبارتست از:

$$\theta = \frac{E(new) - E(old)}{E(old) - E(placebo)}$$

که در آن $E(old)$ و $E(new)$ پانسمان‌های قدیمی و جدید هستند. برای این داده‌ها، اگر $|t| \leq ۰/۲۰$، نشان‌دهنده برابری عملکرد پزشکی پانسمان‌های قدیم و جدید است.

تابعی مورد نظر آماره $t = \frac{\bar{Y}}{\bar{Z}}$ است. هدف برآورد خودگردانی اریبی آماره نسبت برابری عملکرد است.

```
data(patch, package = "bootstrap")
> patch
  subject placebo oldpatch newpatch     z     y
1       1    9243    17649    16449  8406 -1200
2       2    9671    12013    14614  2342  2601
3       3   11792    19979    17274  8187 -2705
4       4   13357    21816    23798  8459  1982
5       5    9055    13850    12560  4795 -1290
6       6    6290     9806    10157  3516   351
7       7   12412    17208    16570  4796  -638
8       8   18806    29044    26325 10238 -2719
```

برآورد اریبی برای یک برآوردگر نسبت: ادامه

```
n <- nrow(patch) # in bootstrap package
B <- 2000
theta.b <- numeric(B)
theta.hat <- mean(patch$y) / mean(patch$z)
# bootstrap
for (b in 1:B) {
     i <- sample(1:n, size = n, replace = TRUE)
     y <- patch$y[i]
     z <- patch$z[i]
     theta.b[b] <- mean(y) / mean(z)
}
bias <- mean(theta.b) - theta.hat
se <- sd(theta.b)
> print(list(est=theta.hat, bias = bias,se = se, cv = bias/se))
$est
[1] -0.0713061

$bias
[1] 0.004647646

$se
[1] 0.09888232

$cv
[1] 0.04700179
```

معمولا اگر $\frac{|bias|}{se} \leq$ ۰/۲۵، نیازی به تصحیح اریبی نیست. در این مثال این نسبت کمتر از ۰/۰۵ است.