An Introduction to Bootstrap Methods and their Application

Prof. Dr. Diego Kuonen, CStat PStat CSci

Statoo Consulting, Berne, Switzerland

@DiegoKuonen + kuonen@statoo.com + www.statoo.info

'**WBL in Angewandter Statistik ETHZ 2017/19**' — January 22 & 29, 2018

Presentation code: 'WBL.Statistik.ETHZ.2018'.

Typesetting: LaTeX, version 2ε. PDF producer: pdfTeX, version 3.141592-1.40.3-2.2 (Web2C 7.5.6).

Compilation date: 12.01.2018.

About myself (about.me/DiegoKuonen)

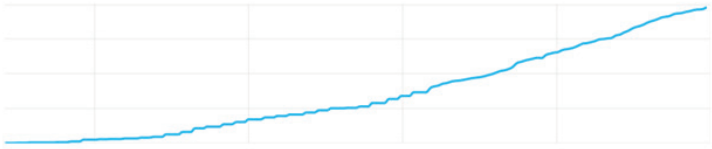
- ◇ PhD in Statistics, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
- ◇ MSc in Mathematics, EPFL, Lausanne, Switzerland.

- CStat ('Chartered Statistician'), Royal Statistical Society, UK.
- PStat ('Accredited Professional Statistician'), American Statistical Association, USA.
- CSci ('Chartered Scientist'), Science Council, UK.

- Elected Member, International Statistical Institute, NL.
- Senior Member, American Society for Quality, USA.
- President of the Swiss Statistical Society (2009-2015).

- ▷ Founder, CEO & CAO, Statoo Consulting, Switzerland (since 2001).
- ▷ Professor of Data Science, Research Center for Statistics (RCS), Geneva School of Economics and Management (GSEM), University of Geneva, Switzerland (since 2016).
- ▷ Founding Director of GSEM's new MSc in Business Analytics program (started fall 2017).
- ▷ Principal Scientific and Strategic Big Data Analytics Advisor for the Directorate and Board of Management, Swiss Federal Statistical Office (FSO), Neuchâtel, Switzerland (since 2016).

@DiegoKuonen



- ➢ 30.11.2013: 3 followers
- ➢ 18.11.2014: 1'404
- ➢ 12.01.2018: 15'133

Onalytica,
New York & London

- ➢ #26 Big Data (May 2017)
- ➢ #29 Internet of Things, IoT (February 2016)
- ➢ #45 Artificial Intelligence & Machine Learning (March 2016)

About Statoo Consulting (www.statoo.info)

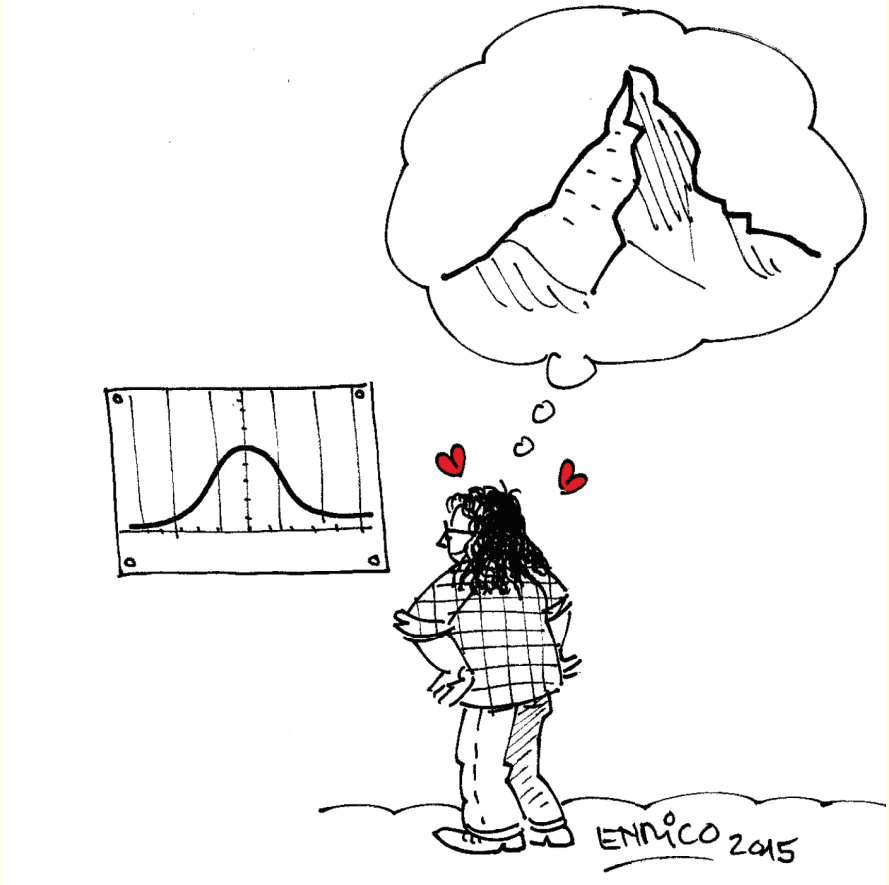
- Founded Statoo Consulting in **2001**.

$$\rightsquigarrow 2018 - 2001 = 17 + \epsilon.$$

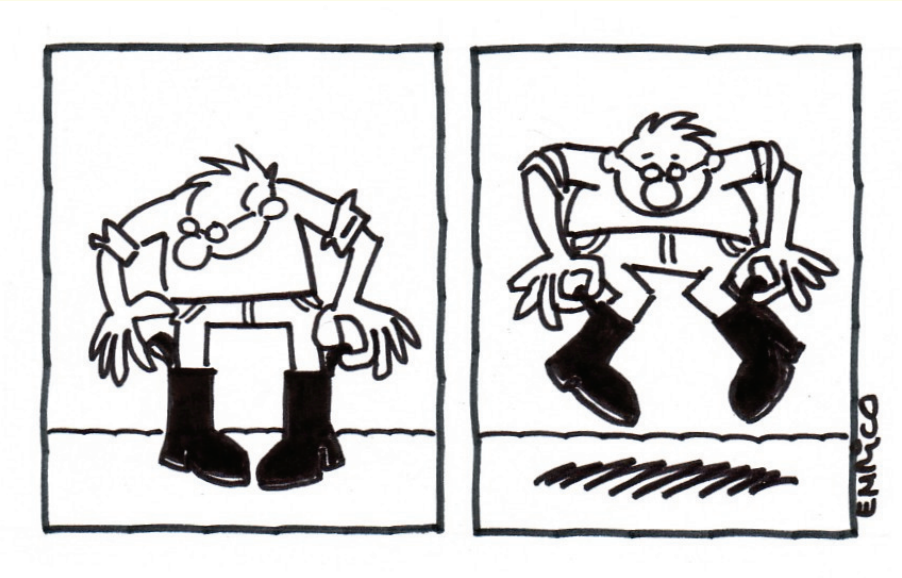
- Statoo Consulting is a software-vendor independent Swiss consulting firm specialised in statistical consulting and training, data analysis, data mining (data science) and big data analytics services.
- Statoo Consulting offers consulting and training in statistical thinking, statistics, data mining and big data analytics in English, French and German.

$\rightsquigarrow$ **Are you drowning in uncertainty and starving for knowledge?**

$\rightsquigarrow$ **Have you ever been Statooed?**



Contents



1. Introduction

• 'Monte Carlo' methods of inference for difficult statistical problems, where standard methods can fail or are unreliable.

$\rightsquigarrow$ Use computer in essential way.

◇ Randomisation methods based on symmetry arguments, mostly used for tests (as seen in block 'Nichtparametrische Methoden').

◇ Jackknife used for bias and variance estimation.

◇ Bootstrap is general tool for confidence intervals, assessment of uncertainty.

2. Introduction to the jackknife

• Used to obtain improved estimates and confidence intervals for complicated statistics.

$\rightsquigarrow$ Mainly used for bias and variance estimation.

$\rightsquigarrow$ The jackknife was introduced by Quenouille (1949, 1956) as a general method to remove bias from estimators.

$\rightsquigarrow$ And Tukey (1958) suggested its use for variance and interval estimation.

Example. For motivation, consider average of $Y_1, \ldots, Y_n$, *i.e.* $\bar{Y} = n^{-1} \sum_{j=1}^{n} Y_j$.

$\rightsquigarrow$ Average of $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$ is $\bar{Y}_{-i} = (n-1)^{-1} \left(\sum_{j=1}^{n} Y_j - Y_i \right)$.

$\rightsquigarrow$ Hence $Y_i = n\bar{Y} - (n-1)\bar{Y}_{-i}$, which we can average to get estimate $\bar{Y}$ and its variance

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 .$$

- Aim to estimate the parameter of interest θ using complicated statistic

$$T = t(Y_1, \ldots, Y_n),$$

for which no variance is available.

- Define the n 'leave one out' estimates

$$T_{-i} = t(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n),$$

computed with Y_i left out, and so-called 'pseudo values'

$$T_i = nT - (n-1)T_{-i}.$$

$\rightsquigarrow$ By analogy with the average $\bar{Y}$, an improved estimate of θ and its variance are

$$T_{\text{jack}} = n^{-1} \sum_{i=1}^{n} T_i, \quad \text{var}(T_{\text{jack}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (T_i - T_{\text{jack}})^2 .$$

$\rightsquigarrow$ The jacknife estimate of bias is defined by

$$b_{\text{jack}}(T) = T_{\text{jack}} - T = (n-1) \left(T - n^{-1} \sum_{i=1}^{n} T_{-i} \right) .$$

Example. For illustration, consider the following artificial data:

5	7	2	10	3

- The classical biased estimate of $\theta = \sigma^2$, the variance, is

$$\hat{\sigma}^2 = n^{-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 = n^{-1} \sum_{j=1}^{n} y_j^2 - \bar{y}^2.$$

$\rightsquigarrow$ We want to improve $T = \hat{\sigma}^2$ using $T_{\text{jack}} = \hat{\sigma}^2_{\text{jack}}$, the jackknife estimate of σ^2.

- We have $\hat{\sigma}^2 = 8.24$ and we can calculate for $i = 1, \ldots, 5$ the pseudo values

$$T_i = nT - (n-1)T_{-i} = n\hat{\sigma}^2 - (n-1)\hat{\sigma}^2_{-i}.$$

$\rightsquigarrow$ Yields

$$T_{\text{jack}} = \hat{\sigma}^2_{\text{jack}} = n^{-1} \sum_{i=1}^{n} T_i = 10.3,$$

which in this case corresponds to $s^2 = 10.3$, where

$$s^2 = (n-1)^{-1} \sum_{j=1}^{n} (y_j - \bar{y})^2$$

is the classical unbiased estimate of $\theta = \sigma^2$.

$\rightsquigarrow$ The jacknife estimate of bias is

$$b_{\text{jack}}(\hat{\sigma}^2) = \hat{\sigma}^2_{\text{jack}} - \hat{\sigma}^2 = 10.3 - 8.24 = 2.06.$$

- It can be shown in general that if $T = \hat{\sigma}^2 = n^{-1} \sum_{j=1}^{n} (y_j - \bar{y})^2$ the jacknife estimate $T_{\text{jack}} = \hat{\sigma}^2_{\text{jack}}$ equals $s^2 = (n-1)^{-1} \sum_{j=1}^{n} (y_j - \bar{y})^2$.

Example. Generate $n = 20$ values from the $N(0, \theta^2)$ distribution, and estimate θ using the biased estimate

$$T = \left\{ n^{-1} \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \right\}^{1/2}.$$

- Original value is $t = 1.03$.

$\rightsquigarrow$ Jackknife estimate $t_{\text{jack}} = 1.10$, with estimated variance $\text{var}(T_{\text{jack}}) = 0.27^2$.

$\rightsquigarrow$ A 95% jackknife normal confidence interval is

$$t_{\text{jack}} \pm 1.96 \cdot \text{var}(T_{\text{jack}})^{1/2} = 1.10 \pm 1.96 \cdot 0.27 = (0.57, 1.63).$$

Comments

- Jackknife 'works' for smooth statistics, *e.g.* averages, variances, most estimators based on moments, but not for rough ones, *e.g.* medians, maxima.

Example. Jackknifing the sample median gives a numerical variance estimate, but in fact this number does not approach the true variance of the median, even as $n \to \infty$.

$\rightsquigarrow$ Consider data:

10	27	31	40	46	50	52	104	146

$\rightsquigarrow$ We have $T = \text{med}(y) = 46$.

$\rightsquigarrow$ The values $T_{-i} = \mathrm{med}(y)_{-i}$ are

48	48	48	48	45	43	43	43	43

$\rightsquigarrow$ Only three different values and the standard variance of the jacknife estimate is 6.68^2 which is too small!

$\rightsquigarrow$ Note that, for example, $s^2 = 42.41^2$.

- Jackknife confidence intervals can be poor if n small.

$\rightsquigarrow$ Depends on normality of T_i.

$\rightsquigarrow$ Check this by normal Q–Q plot of T_i.

$\rightsquigarrow$ If T_i non-normal, a confidence interval for a transformation of θ may be better.

$\rightsquigarrow$ For example, might apply jackknife to $U = \log T$, giving variance estimate and confidence interval for $\phi = \log \theta$.

'The bootstrap has shown us how to use the power of the computer and iterated calculations to go where theoretical calculations can not, which introduces a different way of thinking about all of statistics.'

George Casella, 2003

3. Introduction to the bootstrap

- The bootstrap was introduced by Efron (1979) as a general method for assessing the statistical accuracy of an estimator.

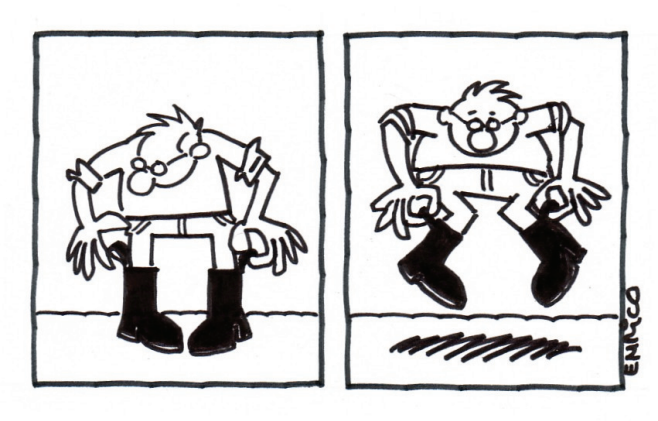
$\rightsquigarrow$ Found an immediate place in statistical theory and, more slowly, in practise (although well suited to the computer age).

- Bootstrap: a marriage of computers and statistics.

- Bootstrap: simulation methods for frequentist inference.

• Efron (1979) coined the term 'bootstrap', named after Baron Münchhausen who, it is said,

> 'finding himself at the bottom of a deep lake, thought to pull himself up by his bootstraps'.



• The basic idea is quite simple:

- simulate data from one or more plausible 'models',
- apply the same procedure to the simulated data sets as was applied to the original data, and
- then analyse the results.

- For example, the bootstrap is used to find
  - standard errors for estimators;
  - confidence intervals for unknown parameters; or
  - p-values for test statistics under a null hypothesis.

$\rightsquigarrow$ The bootstrap is typically used to estimate quantities associated with the sampling distribution of estimators and test statistics.

- Useful when
  - standard assumptions invalid, *e.g.* n small, data not normal;
  - standard problem has non-standard twist;
  - standard methods can fail;
  - complex problem has no (reliable) theory;
  - or (almost) anywhere else.

$\rightsquigarrow$ Aim to describe
  - basic ideas;
  - confidence intervals;
  - significance tests;
  - application to simple linear regression.

'The bootstrap is the most important new idea in statistics introduced in the last 20 years, and probably in the last 50 years.'

Jerome H. Friedman, 1998

3.1 Scene-setting

- Single homogeneous random sample of data $y_1, \ldots, y_n$, outcomes of 'independent, identically distributed' (iid) variables $Y_1, \ldots, Y_n$, which have unknown 'Cumulative Distribution Function' (CDF) F and 'Probability Density Function' (PDF) f.

- Possibilities for F:
  - parametric: $F \equiv F_\psi$ determined by unknown ψ;
  - nonparametric: 'mild' conditions on F, *e.g.* moments, symmetry.

- In parametric model, CDF and PDF known up to values of unknown parameters.

$\rightsquigarrow$ For example,

$$f(y;\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

is $N(\mu,\sigma^2)$ density.

$\rightsquigarrow$ We set $\psi = (\mu,\sigma^2)$ and write $F_\psi(y)$ to emphasise dependence on ψ.

- In nonparametric model, just use the fact that the sample is homogeneous.

- Need to estimate F:
  - parametric: $\hat{F} = F_{\hat{\psi}}(y)$, where $\hat{\psi}$ is a suitable estimate of ψ; often 'Maximum Likelihood Estimates' (MLE);
  - nonparametric: $\hat{F} = F_n(y) = n^{-1}\#\{y_j \le y\}$ is the 'Empirical Distribution Function' (EDF) of $y_1,\ldots,y_n$.

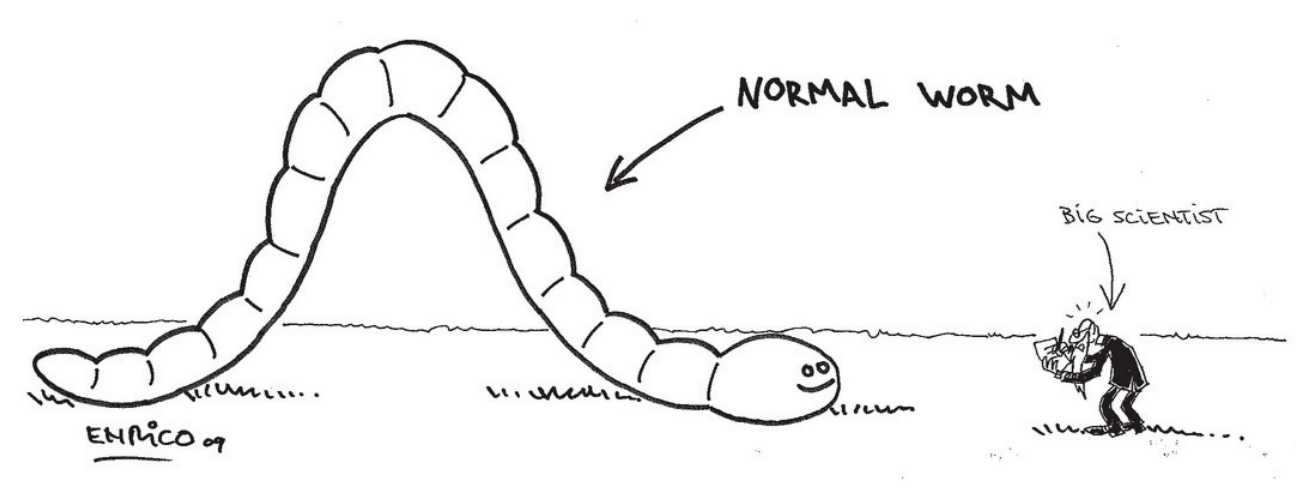
    $\rightsquigarrow$ More formally $F_n(y) = n^{-1}\sum_{j=1}^n H(y-y_j)$, where $H(u)$ is the unit step function which jumps from 0 to 1 at $u=0$.

• Want inferences for parameter $\theta = t(F)$, population characteristic, estimated by statistic T with observed value $t = t(\hat{F})$, where $\hat{F}$ is estimate of F.

$\rightsquigarrow$ For example, might want bias and variance of T, confidence intervals for θ, or test of whether θ_0 is a plausible value for θ.

• Think of 'statistical function' $t(\cdot)$ as an algorithm for calculating θ from F.

• Heroic assumption: $\hat{F} \doteq F$, so properties of t calculated from $\hat{F}$ similar to those calculated from F.



Normal approximation

• Statistic T is an estimator of θ.

$\rightsquigarrow$ Initially focus on bias and variance of T (for use with normal approximation), or confidence limits for θ.

• Suppose T normal, then $T \sim N(\theta + \beta, \nu)$ as $T - \theta \sim N(\beta, \nu)$, where bias and variance of T are

$$\begin{aligned} \beta &= b(F) = \mathrm{E}(T \mid Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F) - \theta, \\ \nu &= v(F) = \mathrm{var}(T \mid Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F), \end{aligned}$$

so

$$\Pr(T \leq t \mid F) \doteq \Phi\left\{\frac{t - (\theta + \beta)}{\nu^{1/2}}\right\},$$

where $\Phi(\cdot)$ is the standard normal CDF.

• If α-quantile of $N(0,1)$ distribution is $z_\alpha = \Phi^{-1}(\alpha)$, and if β, ν known, then an approximate $(1 - 2\alpha)$ confidence interval for θ is

$$t - \beta \pm z_\alpha \nu^{1/2},$$

because $z_{1-\alpha} = -z_\alpha$ in

$$\Pr(\beta + z_\alpha \nu^{1/2} \leq T - \theta \leq \beta + z_{1-\alpha} \nu^{1/2}) \doteq 1 - 2\alpha.$$

• In practise β and ν replaced by estimates.

$\rightsquigarrow$ Replace F by estimate $\hat{F}$, to get

$$B = b(\hat{F}) = \mathrm{E}(T \mid \hat{F}) - t, \quad V = v(\hat{F}) = \mathrm{var}(T \mid \hat{F}).$$

$\rightsquigarrow$ Use B and V in place of β and ν in confidence limits above.

Example (Air-conditioning data). The data are $n = 12$ times between failures of air-conditioning equipment in a Boeing 720 jet aircraft.

$\rightsquigarrow$ We wish to estimate the underlying mean (or its reciprocal, the failure rate).

$\rightsquigarrow$ The data are:

3	5	7	18	43	85	91	98	100	130	230	487

$\rightsquigarrow$ A simple model for this problem is that the times are sampled from an exponential distribution, $\mathcal{E}(1/\mu)$, having CDF $F_\mu(y) = 1 - \exp(-y/\mu)$ for $y > 0$ and 0 for $y \leq 0$.

$\rightsquigarrow$ To fit the exponential distribution set mean μ equal to the sample average, $\bar{y} = 108.083$.

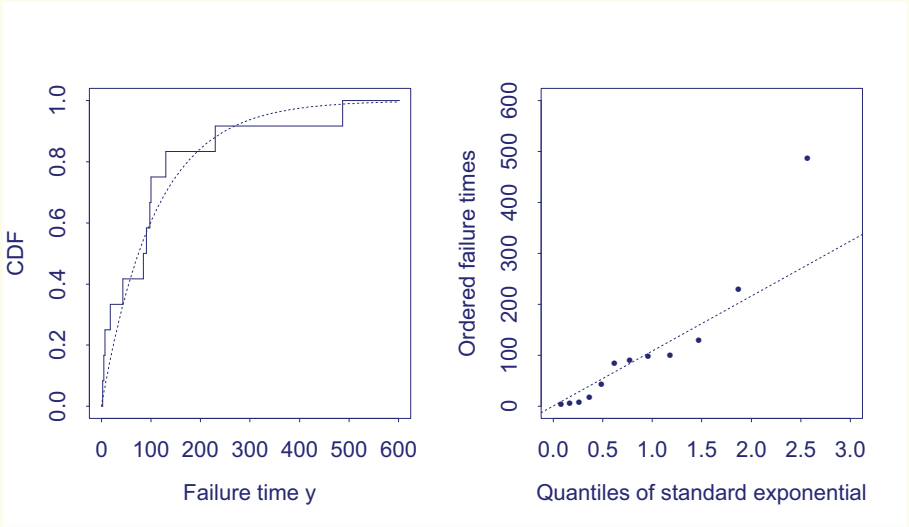


Figure 1: Summary displays for the air-conditioning data. The left panel shows the EDF for the data, F_n (solid), which places equal probabilities $n^{-1} = 1/12 = 0.083$ at each sample value, and the CDF of a fitted exponential distribution, $F_{\hat{\mu}} = F_{\bar{y}}$ (dotted). The right panel shows a plot of the ordered failure times against exponential quantiles, with the fitted exponential model shown as the dotted line.

$\rightsquigarrow$ Although these plots suggest reasonable agreement with the exponential model, the sample is rather too small to have much confidence in this.

• Might be better to fit the more general gamma model with mean μ and index κ; its density is

$$f_{\mu,\kappa}(y) = \frac{1}{\Gamma(\kappa)} \left(\frac{\kappa}{\mu}\right)^{\kappa} y^{\kappa-1} \exp(-\kappa y/\mu), \qquad y > 0, \quad \mu, \kappa > 0.$$

$\rightsquigarrow$ For the sample the estimated index is $\hat{\kappa} = \bar{y}^2/s^2 = 0.71$.

• Basic properties of the estimator $T = \bar{Y}$ for μ and corresponding confidence intervals are easy to obtain theoretically under the exponential model.

$\rightsquigarrow$ But, things are more complicated under the more general gamma model, because the index κ is only estimated, and so in a traditional approach we would use approximations, *e.g.* a normal approximation, for the distribution of T.

• Under the exponential model for the data, the mean failure time μ is estimated by the average $T = \bar{Y}$.

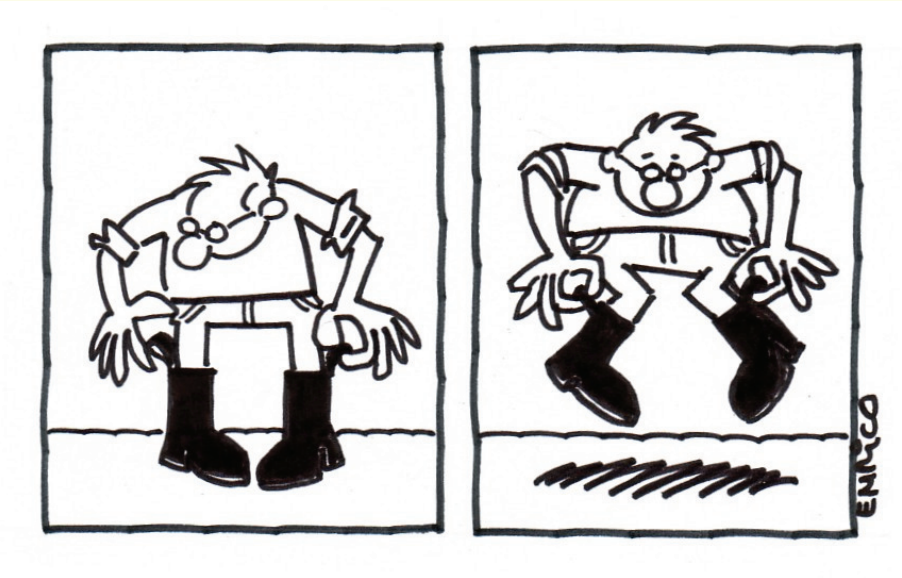
$\rightsquigarrow$ Since $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{E}(1/\mu)$, *i.e.* $\mathrm{E}(Y_j) = \mu$ and $\mathrm{var}(Y_j) = \mu^2$, $j = 1, \ldots, n$, we have

$$\mathrm{E}(T) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}(Y_j) = \mu \text{ and } \mathrm{var}(T) = \frac{1}{n^2}\sum_{j=1}^{n} \mathrm{var}(Y_j) = \frac{\mu^2}{n}.$$

$\rightsquigarrow$ The bias and variance of T are $b(F) = 0$ and $v(F) = \mu^2/n$, and these are estimated by 0 and $\bar{y}^2/n$.

$\rightsquigarrow$ Since $n = 12$, $\bar{y} = 108.083$ and $z_{0.025} = -1.96$, a 95% normal confidence interval for μ, *i.e.* based on the normal approximation, is

$$\bar{y} \pm 1.96 \cdot n^{-1/2}\bar{y} = (46.93, 169.24).$$



3.2 Parametric bootstrap simulation

- Random sample $Y_1, \ldots, Y_n$ from F.

- Parametric model for data, with CDF $F = F_\psi$ and PDF $f = f_\psi$.

$\rightsquigarrow$ Parametric estimate is $\hat{\psi}$, giving fitted model $\hat{F} = F_{\hat{\psi}}$.

$\rightsquigarrow$ We focus on properties of statistics calculated from $\hat{F}$.

- Estimator T is algorithm given by $T = t(Y_1, \ldots, Y_n)$:
  - applied to F gives parameter $\theta = t(F)$;
  - applied to $\hat{F}$ gives estimate $t = t(\hat{F})$.

$\rightsquigarrow$ Usually hard to find theoretical properties of T.

$\rightsquigarrow$ Use simulation from fitted model $\hat{F} = F_{\hat{\psi}}$ instead: parametric bootstrap.

• Simulated data set (resample): $Y_1^*, \ldots, Y_n^* \overset{\text{iid}}{\sim} \hat{F} = F_{\hat{\psi}}$, where the Y_j^* are independently sampled from $\hat{F}$, giving T^*.

$\rightsquigarrow$ Repeat R times to get $t_1^*, \ldots, t_R^*$.

$\rightsquigarrow$ Estimate properties of $T - \theta$ empirically.

◇ Bias $b(F) = \mathrm{E}(T \mid F) - \theta$ of T is estimated by

$$B = b(\hat{F}) = \mathrm{E}(T \mid \hat{F}) - t = \mathrm{E}^*(T^*) - t,$$

and this in turn is estimated by

$$b^* = R^{-1} \sum_{r=1}^{R} t_r^* - t = \bar{t}^* - t.$$

$\rightsquigarrow$ In the simulation t is the parameter value for the model, so that $T^* - t$ is the simulation analogue of $T - \theta$.

◇ Variance $\mathrm{var}(T)$ estimated by empirical variance of $t_1^*, \ldots, t_R^*$, *i.e.*

$$v^* = (R-1)^{-1} \sum_{r=1}^{R} (t_r^* - \bar{t}^*)^2,$$

with similar estimators for other moments.

• Empirical approximations based on law of large numbers; if B^* and V^* are random variables corresponding to b^* and v^*, then $B^* \to B$ and $V^* \to V$ as $R \to \infty$.

◇ Two sources of error: statistical error because $\hat{F} \neq F$ (reduce by thought), and simulation error because R finite (reduce by taking R 'large enough').

$\rightsquigarrow$ R in range $50-200$ usually OK for bias and variance estimation; need $1'000-2'000$ for confidence interval estimation.

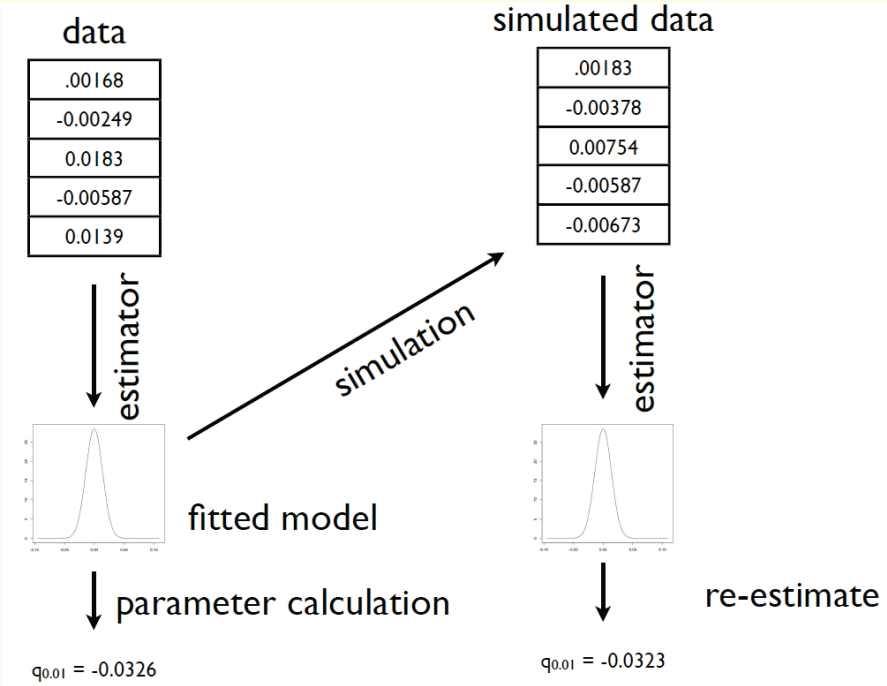


Figure 2: Schematic for parametric bootstrap simulation: simulated values are generated from the fitted model, then treated like the original data, yielding a new estimate of the 'functional' of interest, here called $q_{0.01}$; see Shalizi (2013, Lecture

Example (Air-conditioning data, continued). Fitted model is an exponential distribution for the failure times, with mean estimated by the sample average $\hat{\psi} = \hat{\mu} = \bar{y} = 108.083$.

$\rightsquigarrow$ Simulate bootstrap samples from this distribution.

- As $T^* = \bar{Y}^* = n^{-1}(Y_1^* + \cdots + Y_n^*)$ theoretical calculation yields

$$\mathrm{E}^*(\bar{Y}^*) = \bar{y}, \quad \mathrm{var}^*(\bar{Y}^*) = \bar{y}^2/n.$$

$\rightsquigarrow$ The estimated bias of $T = \bar{Y}$ is 0, being the difference between $\mathrm{E}^*(\bar{Y}^*) = \bar{y}$ and the value $t = \bar{y}$ for the mean of the fitted distribution.

$\rightsquigarrow$ The estimated variance is $\bar{y}^2/n = (108.083)^2/12 = 973.5$.

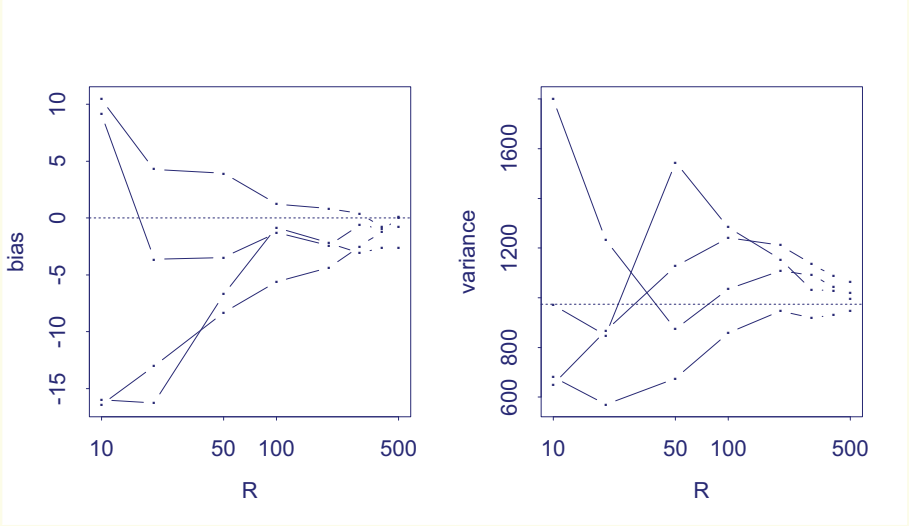


Figure 3: Empirical biases and variances of $T^* = \bar{Y}^*$ for the air-conditioning data from four repetitions of parametric simulation. Each line shows how the estimated bias and variance for $R = 10$ initial simulations change when further simulations are successively added. $\rightsquigarrow$ Note how the variability decreases as the simulation size increases, and how the simulated values converge to the exact values under the fitted exponential model, given by the horizontal dotted lines.

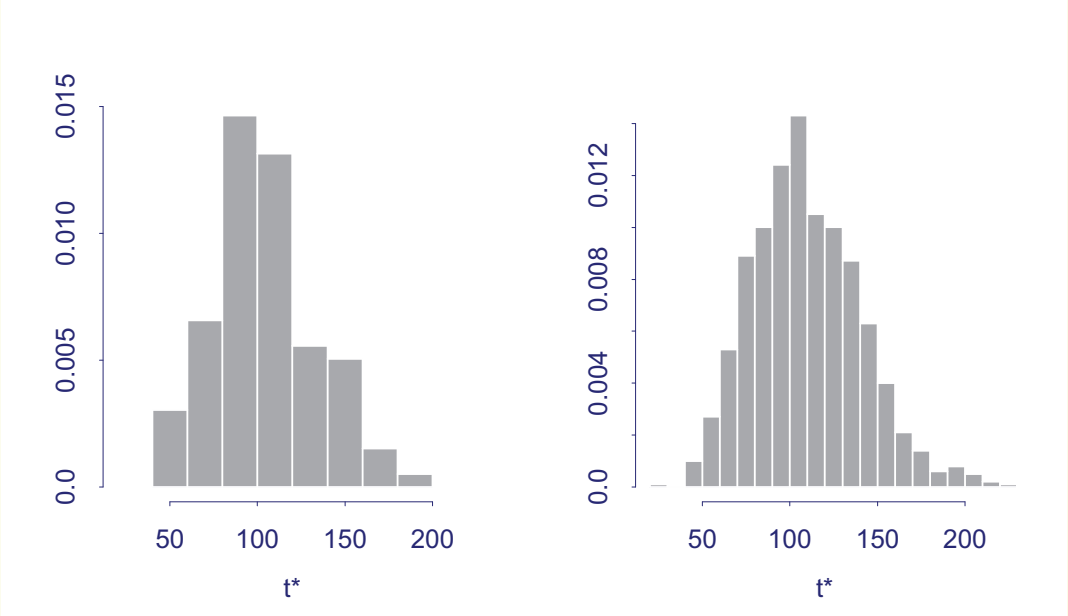


Figure 4: Histograms of t^* values based on $R = 99$ (left panel) and $R = 999$ (right panel) simulations from the fitted exponential model for the air-conditioning data. $\rightsquigarrow$ Clearly a normal approximation would not be accurate in the tails, and this is already fairly clear with $R = 99$.

◇ Estimate quantiles of T.

$\rightsquigarrow$ For example, use $t_1^*, \ldots, t_R^*$ to estimate quantiles of T.

$\rightsquigarrow$ More demanding than bias and variance estimation, so must take R at least $1'000$ or so.

$\rightsquigarrow$ Estimate α-quantile by $(R+1)\alpha$th ordered value of

$$t_{(1)}^* \leq t_{(2)}^* \leq \cdots \leq t_{(R)}^*,$$

so $R = 999$, $\alpha = 0.025$ gives $t_{(25)}^*$.

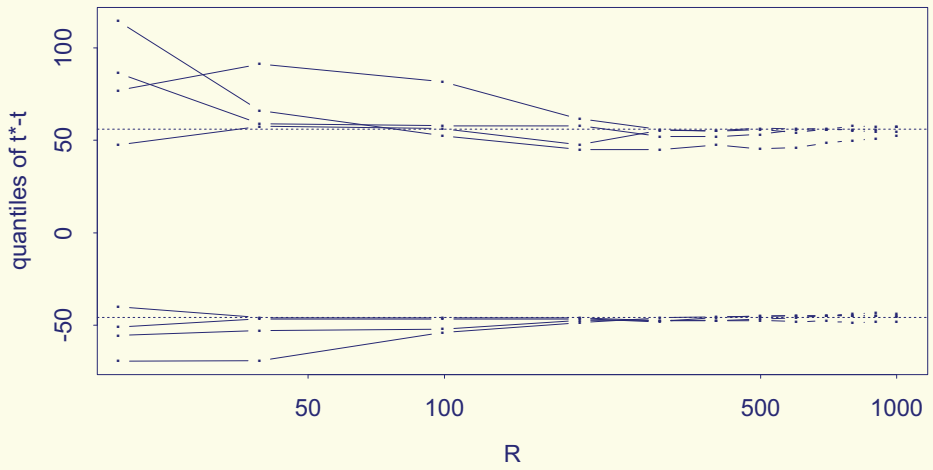


Figure 5: Empirical quantiles ($\alpha = 0.05, 0.95$) of $T^* - t$ (approximated by $t^*_{((R+1)\alpha)} - t$) under resampling from the fitted exponential model for the air-conditioning data. The horizontal dotted lines are the exact quantiles under the model. $\rightsquigarrow$ There is large variability in the approximate quantiles for R less than 100 and it appears that 500 or more simulations are required to get accurate results.

'Although we often hear that data speak for themselves, their voices can be soft and sly.'

(Charles) Frederick Mosteller, 1983

3.3 Nonparametric bootstrap simulation

• No parametric model, but assume data are iid according to an unknown CDF F.

$\rightsquigarrow$ Use the EDF $F_n = n^{-1}\#\{y_j \leq y\}$ to estimate the unknown CDF F.

$\rightsquigarrow$ Use F_n just as we would a parametric model: theoretical calculation if possible (unusual), otherwise empirical calculation from simulated data sets.

• Simulation: EDF puts equal probabilities n^{-1} on the original data values $y_1, \ldots, y_n$, so each Y_j^* is independently sampled at random from these.

$\rightsquigarrow$ Simulated sample $Y_1^*, \ldots, Y_n^*$ is a random sample taken with replacement from the data.

$\rightsquigarrow$ This is nonparametric bootstrap.

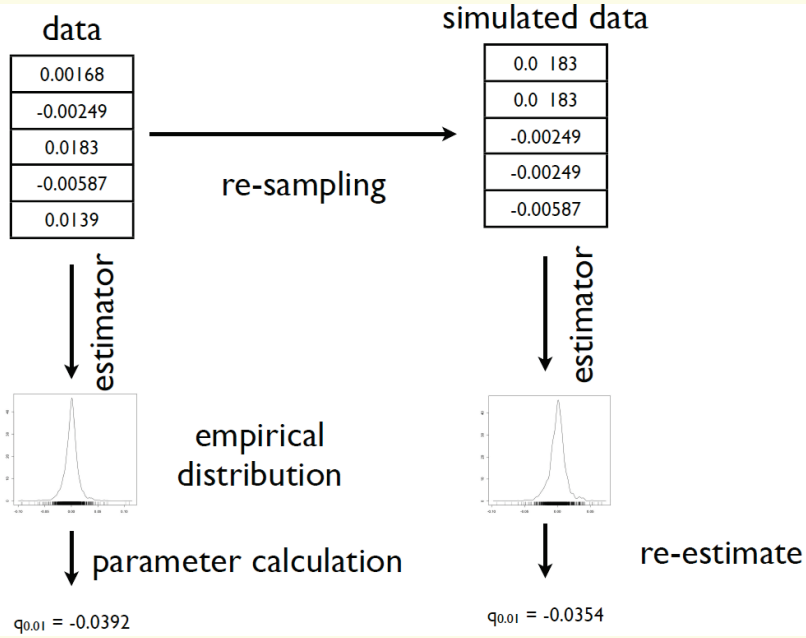


Figure 6: Schematic for nonparametric bootstrap simulation: new data are simulated by sampling with replacement from the original data, and parameters are calculated either directly from the empirical distribution, or by applying a model to this surrogate data; see Shalizi (2013, Lecture 6).

Example (City population data). Populations in thousands of $n = 49$ large United States cities in 1920 (u) and in 1930 (x):

u	x	u	x	u	x
138	143	76	80	67	67
93	104	381	464	120	115
61	69	387	459	172	183
179	260	78	106	66	86
48	75	60	57	46	65
37	63	507	634	121	113
29	50	50	64	44	58
23	48	77	89	64	63
30	111	64	77	56	142
2	50	40	60	40	64
38	52	136	139	116	130
46	53	243	291	87	105
71	79	256	288	43	61
25	57	94	85	43	50
298	317	36	46	161	232
74	93	45	53	36	54
50	58				

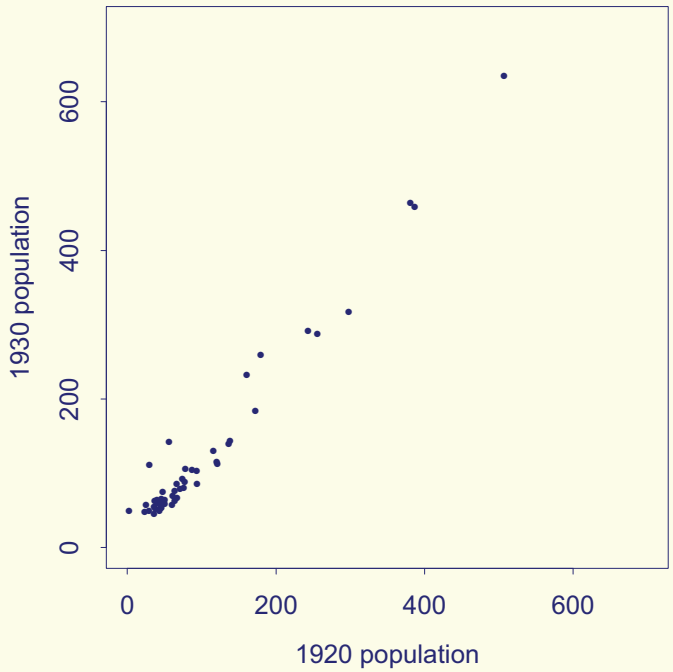


Figure 7: Populations of 49 large United States cities (in $1'000$s) in 1920 and 1930.

• Interest here is in the ratio of means, because this would enable us to estimate the total population of the United States in 1930 from the 1920 figure.

$\rightsquigarrow$ If the cities form a random sample with (U, X) denoting the pair of population values for a randomly selected city, then the total 1930 population is the product of the total 1920 population and the ratio of expectations

$$\theta = \frac{\mathrm{E}(X)}{\mathrm{E}(U)},$$

which is the parameter of interest.

$\rightsquigarrow$ There is no obvious parametric model for the joint distribution of (U, X), so it is natural to estimate θ by its empirical analogue,

$$T = \frac{\bar{X}}{\bar{U}},$$

the ratio of sample averages.

$\rightsquigarrow$ We are interested in the uncertainty in T.

• In this case F is the bivariate CDF of $Y = (U, X)$, and the bivariate EDF F_n puts probability n^{-1} at each of the data pairs (u_i, x_i).

$\rightsquigarrow$ The corresponding estimate of θ is

$$t = t(F_n) = \frac{\bar{x}}{\bar{u}}.$$

• This is an application with no obvious parametric model for the joint distribution of (U, X), so nonparametric simulation makes good sense.

i	1	2	3	4	5	6	7	8	9	10
u	138	93	61	179	48	37	29	23	30	2
x	143	104	69	260	75	63	50	48	111	50
i^*	6	7	2	2	3	3	10	7	2	9
u^*	37	29	93	93	61	61	2	29	93	30
x^*	63	50	104	104	69	69	50	50	104	111

Table 1: A subset of the data, comprising the first 10 pairs. The values i^* are chosen randomly with equal probability from $\{1, \ldots, n = 10\}$ with replacement; the simulated pairs are (u_{i^*}, x_{i^*}). $\rightsquigarrow$ Here $i^* = 1$ never occurs and $i^* = 2$ occurs three times, so that the first data pair is never selected, the second is selected three times, and so forth.

i	1	2	3	4	5	6	7	8	9	10
u	138	93	61	179	48	37	29	23	30	2
x	143	104	69	260	75	63	50	48	111	50

	Numbers of times each pair sampled										Statistic
Data	1	1	1	1	1	1	1	1	1	1	$t = 1.520$
Replicate r											
1		3	2			1	2		1	1	$t_1^* = 1.466$
2	1		1		2	2	1		2	1	$t_2^* = 1.761$
3	1	1		1		1			4	2	$t_3^* = 1.951$
4		1	2		1	1	2	2		1	$t_4^* = 1.542$
5	3			1	3		1	1	1		$t_5^* = 1.371$
6	1	1	2			1		1	1	3	$t_6^* = 1.686$
7	1	1	2	2	2		1			1	$t_7^* = 1.378$
8	2		1		3	1	1	1	1		$t_8^* = 1.420$
9		1	1	1	2	1		2	1	1	$t_9^* = 1.660$

Table 2: Frequencies with which each original data pair (out of the first 10 pairs) appears in each of $R = 9$ nonparametric bootstrap samples (previous sample plus eight more) for the city population data. The ratio $t^* = \bar{x}^*/\bar{u}^*$ of the averages $\bar{x}^*$ and $\bar{u}^*$ for each simulated sample is recorded in the last column.

$\rightsquigarrow$ For the $R = 9$ replicates shown, the results are

$$b^* = \bar{t^*} - t = 1.582 - 1.520 = 0.062, \quad v^* = (R-1)^{-1} \sum_{r=1}^{R} (t_r^* - \bar{t^*})^2 = 0.0391.$$

$\rightsquigarrow$ A simple approximate distribution for $T - \theta$ is $N(b^*, v^*)$. With the results so far, this is $N(0.062, 0.0391)$, but this is unlikely to be accurate enough and a larger value of R should be used.

$\rightsquigarrow$ In a simulation with $R = 999$ we obtained $b^* = 1.5755 - 1.5203 = 0.0552$ and $v^* = 0.0601$.

• For the entire data set of size $n = 49$, the empirical bias and variance of T are $b^* = 0.00118$ and $v^* = 0.001290$.

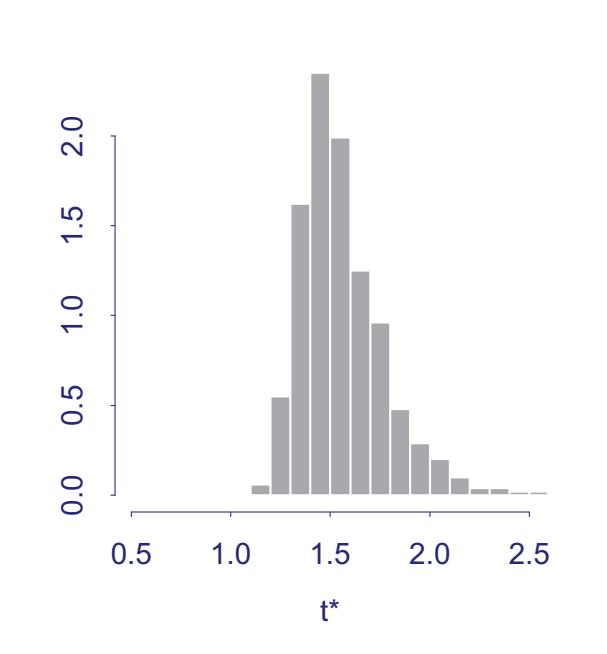


Figure 8: City population data. Histogram of t^* under nonparametric resampling for the entire data set of size $n = 49$, $R = 999$ simulations. $\rightsquigarrow$ The theoretical normal approximation for T would be inaccurate in view of the strong skewness of t^*.

'The bootstrap is very simple but remarkably powerful.'

Persi Diaconis, 1998

3.4 Computing: R-ng

- These ideas are only useful if they can be easily implemented.
- If a suite of reliable random number generators is available, the basic programming of bootstrap resampling is not difficult, but it is much more work to implement carefully some of the more esoteric methods in the literature.
- The key requirement is a statistical environment in which a wide range of functions and methods are already available.

$\rightsquigarrow$ We will use R.

• The main package for bootstrapping in R is `boot`, which is available on the 'Comprehensive R Archive Network' (CRAN) at

`CRAN.R-project.org/package=boot`

and accompanies Davison and Hinkley (1997).

• As `boot` is a 'recommended' R package it is already included in all binary distributions of R and one simply has to type

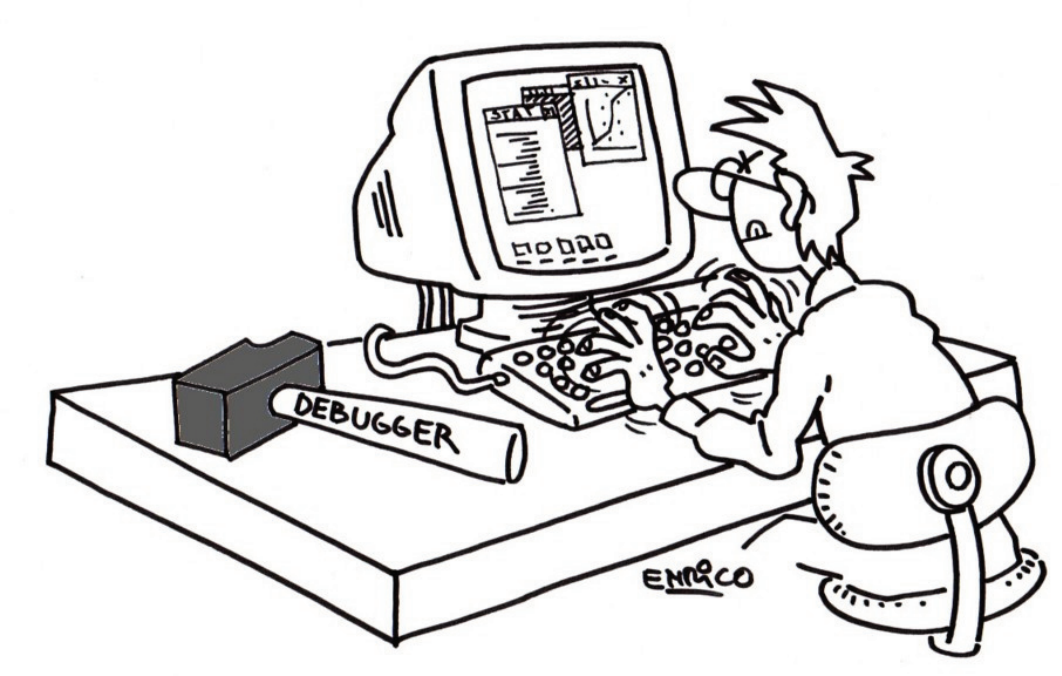
```
> require(boot)
```

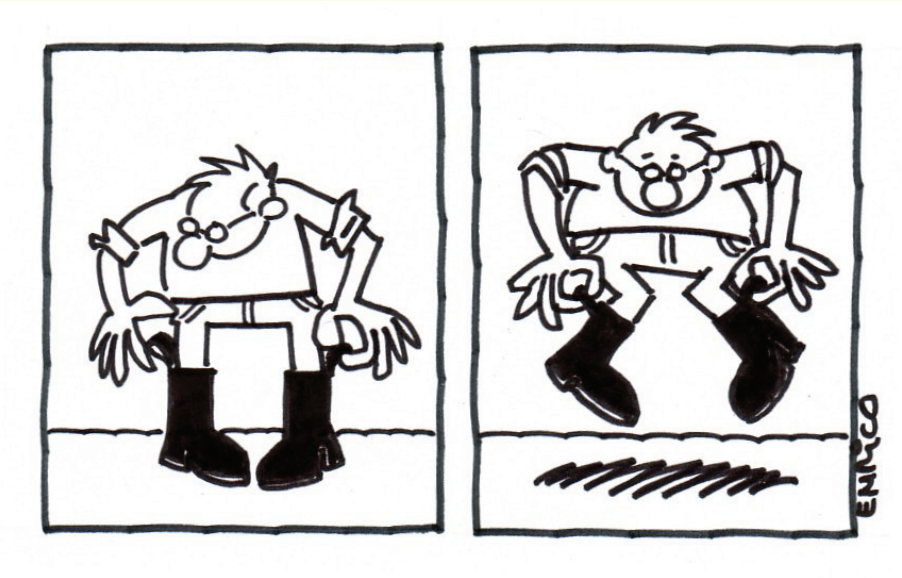
at the R prompt.

⇝ A good starting point is to carefully read the documentations of the R functions `boot` and `boot.ci` (more later)

```
> help(boot)
> help(boot.ci)
```

and to try out one of the examples given in the 'Examples' section of the corresponding help file.





3.5 Bootstrap confidence intervals

3.5.1 Bootstrap normal confidence intervals

- Assume $T \sim N(\theta + \beta, \nu)$, where β is the bias and ν is the variance of T.

$\rightsquigarrow$ If β, ν known, then $(1 - 2\alpha)$ normal confidence interval for θ is

$$t - \beta \pm z_\alpha \nu^{1/2},$$

where z_α is the α-quantile of a $N(0, 1)$ distribution.

$\rightsquigarrow$ Replace unknown β and ν by proxies b^* and v^*, estimated by simulation.

$\rightsquigarrow$ Interval used in practise is the **bootstrap normal confidence interval**

$$t - b^* \pm z_\alpha v^{*1/2},$$

where the simulation estimates of β and ν, based on $t_1^*, \ldots, t_R^*$, are

$$b^* = R^{-1} \sum_{r=1}^{R} t_r^* - t, \quad v^* = (R-1)^{-1} \sum_{r=1}^{R} (t_r^* - \bar{t}^*)^2.$$

- Check normality of T^* by Q–Q plot of $t_1^*, \ldots, t_R^*$.

3.5.2 Bootstrap confidence intervals

- Quantile estimation.

$\rightsquigarrow$ Estimate quantiles of T and derived quantities.

$\rightsquigarrow$ For example, use $t_1^*, \ldots, t_R^*$ to estimate quantiles of T.

$\rightsquigarrow$ More demanding than bias and variance estimation, so must take R at least $1'000$ or so.

$\rightsquigarrow$ Estimate α-quantile by $(R+1)\alpha$th ordered value of

$$t_{(1)}^* \leq t_{(2)}^* \leq \cdots \leq t_{(R)}^*,$$

so $R = 999$, $\alpha = 0.025$ gives $t_{(25)}^*$.

Simple bootstrap confidence intervals

- Basic bootstrap confidence interval (also known as 'reverse bootstrap percentile interval'): treat $T - \theta$ as pivot, *i.e.* combination of data and parameter whose distribution is independent of underlying model, get

$$t - (t_{((R+1)(1-\alpha))}^* - t), \quad t - (t_{((R+1)\alpha)}^* - t).$$

- Percentile interval: use empirical quantiles of $t_1^*, \ldots, t_R^*$:

$$t_{((R+1)\alpha)}^*, \quad t_{((R+1)(1-\alpha))}^*.$$

Better bootstrap confidence intervals

• Hope properties of $t_1^*, \ldots, t_R^*$ mimic effect of sampling from original model.

$\rightsquigarrow$ False in general, but often (not always) more nearly true for a pivot.

Example. Suppose $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then Student-$t$ statistic

$$\frac{\bar{Y} - \mu}{(S^2/n)^{1/2}} \quad \sim \quad t_{n-1},$$

independent of μ, σ^2, can be used to construct exact confidence intervals for μ, even with σ^2 unknown.

$\rightsquigarrow$ Distribution depends on normality.

• Exact pivot generally unavailable in nonparametric case.

$\rightsquigarrow$ What is a good (approximate) pivot in nonparametric estimation problems?

• A good (approximate) pivot in nonparametric estimation problems is the t-like studentized statistic,

$$Z = \frac{T - \theta}{V^{1/2}},$$

where T is an estimator of location of θ and V is an estimator of its variance.

$\rightsquigarrow$ Simple approximation is to take Z to be $N(0, 1)$. Often valid as $n \to \infty$ and is only an approximation for finite samples.

$\rightsquigarrow$ Slightly better approximation is Student's t distribution on $n - 1$ degrees of freedom. One can expect that Z will be behave like a t-statistic, but there is no guarantee of having $n - 1$ degrees of freedom.

• The studentized bootstrap or bootstrap-t method is known to be accurate to estimate the distribution of Z.

$\rightsquigarrow$ Uses replicates of the studentized bootstrap statistic,

$$Z^* = \frac{T^* - t}{V^{*1/2}},$$

where t denotes the observed value of the statistic T, $T^\star$ and $V^\star$ are based on a simulated bootstrap sample, $Y_1^\star, \ldots, Y_n^\star$.

$\rightsquigarrow$ Get R bootstrap copies of T, V:

$$\begin{array}{cccc} t_1^* & t_2^* & \cdots & t_R^* \\ v_1^* & v_2^* & \cdots & v_R^* \end{array}$$

and the corresponding studentized quantities

$$z_1^* = \frac{t_1^* - t}{v_1^{*1/2}}, \; z_2^* = \frac{t_2^* - t}{v_2^{*1/2}}, \; \ldots, \; z_R^* = \frac{t_R^* - t}{v_R^{*1/2}}.$$

$\rightsquigarrow$ EDF of z_r^* estimates distribution of Z^*.

- **Studentized bootstrap confidence interval**: get $(1-2\alpha)$ confidence interval

$$t - v^{1/2} z^*_{((R+1)(1-\alpha))}, \quad t - v^{1/2} z^*_{((R+1)\alpha)},$$

where $z^*_r = (t^*_r - t)/v^{*1/2}_r$ is the value of Z^* for rth bootstrap sample.

- **Improved percentile intervals** (BC_a, ABC, ABC_d, ABC^e_d, ...).
  - Replace percentile interval with

$$t^*_{((R+1)\alpha')}, \quad t^*_{((R+1)(1-\alpha''))},$$

    where α', α'' chosen to improve properties.
  - Scale-invariant.

Example (Visual acuity data). Data from an experiment in which two laser treatments ('blue argon' and 'red krypton') were randomised to eyes on patients.

$\rightsquigarrow$ The response is visual acuity, measured by the number of letters correctly identified in a standard eye test.

$\rightsquigarrow$ Some patients had only one suitable eye, and they received one treatment allocated at random.

$\rightsquigarrow$ There are 20 patients with paired data and 20 patients for whom just one observation is available.

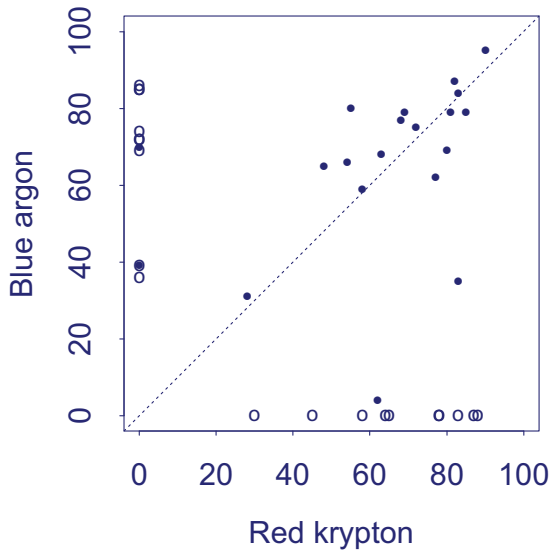
$\rightsquigarrow$ We have a mixture of paired comparison and two-sample data.

$\rightsquigarrow$ A standard analysis is available for each, but we should like to combine them.

$\rightsquigarrow$ Aim to compare treatments.

- In R:

```
> red = c(62, 80, 82, 83, 0, 81, 28, 69, 48, 90, 63, 77, 0, 55, 83, 85,
          54, 72, 58, 68, 88, 83, 78, 30, 58, 45, 78, 64, 87, 65)
> blue = c(4, 69, 87, 35, 39, 79, 31, 79, 65, 95, 68, 62, 70, 80, 84,
           79, 66, 75, 59, 77, 36, 86, 39, 85, 74, 72, 69, 85, 85, 72)
> acuity = data.frame(str=c(rep(0, 20), rep(1, 10)), red, blue)
> acuity
   str red blue
1    0  62    4
2    0  80   69
...
20   0  68   77
21   1  88   36
...
29   1  87   85
30   1  65   72
```



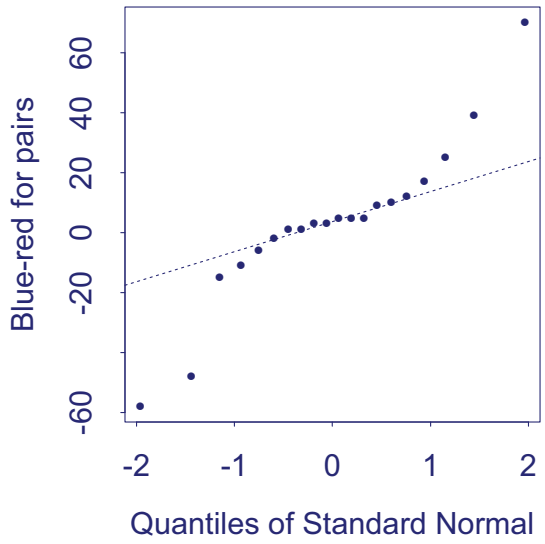


Figure 9: Paired comparison analysis. Left panel: paired (blobs) and unpaired data (circles). Right panel: normal Q–Q plot of differences of 'blue argon' and 'red krypton' treatments, based on paired data.

⇝ The usual normal model might be thought unreliable here, because the responses are quite skewed.

⇝ In summary:

- data non-normal.
- Student-t 95% confidence interval reliable?
- effect of high kurtosis?

• We denote the fully observed pairs $y_j = (r_j, b_j)$, the responses for the eyes treated with 'red' and 'blue' treatments, and for these n_d patients we let $d_j = b_j - r_j$.

• Individuals with just one observation give data $y_j = (?, b_j)$ or $y_j = (r_j, ?)$; there are n_b and n_r of these.

• The unknown variances of the d's, r's and b's are σ_d^2, σ_r^2 and σ_b^2.

$\rightsquigarrow$ For illustration purposes, we will perform a standard analysis for each.

◇ First, we could only consider the paired data and construct the classical Student-t 95% confidence interval for the mean of the differences, of form

$$\bar{d} \pm t_{n_d-1}(0.025)\, s_d / n_d^{1/2},$$

where $\bar{d} = 3.25$, s_d is the standard deviation of the d's and $t_{n_d-1}(0.025)$ is the quantile of the appropriate t distribution.

⇝ This can be done in R by means of

```
> acu.pd = acuity[acuity$str==0, ]
> acu.pd
   str red blue
1    0  62    4
2    0  80   69
...
19   0  58   59
20   0  68   77

> dif = acu.pd$blue - acu.pd$red
> tmp = qt(0.025, nrow(acu.pd) - 1) * sd(dif)/sqrt(nrow(acu.pd))
> c(mean(dif) + tmp, mean(dif) - tmp)
[1] -9.270335 15.770335
```


• But the normal Q–Q plot of the differences shown in the right panel of Figure 9 looks more Cauchy than normal, so the usual model might be thought unreliable.

⇝ The bootstrap can help to check this.

• To perform a nonparametric bootstrap in this case we first need to define the bootstrap function, corresponding to the algorithm $t(\cdot)$:

```
acu.pd.fun = function(data, i)
 {
 d = data[i, ]
 dif = d$blue - d$red
 c(mean(dif), var(dif)/nrow(d))
 }
```

$\rightsquigarrow$ A set of $R = 999$ bootstrap replicates can then be easily obtained with

```
> require(boot)
> acu.pd.boot = boot(acu.pd, acu.pd.fun, R=999)
```

$\rightsquigarrow$ The resulting nonparametric 95% bootstrap confidence intervals can be calculated as shown previously or using directly

```
> boot.ci(acu.pd.boot, type=c("norm", "basic", "stud"))
...
Normal             Basic              Studentized
(-8.200, 14.948)   (-8.100, 15.050)   (-8.662, 15.768)
```

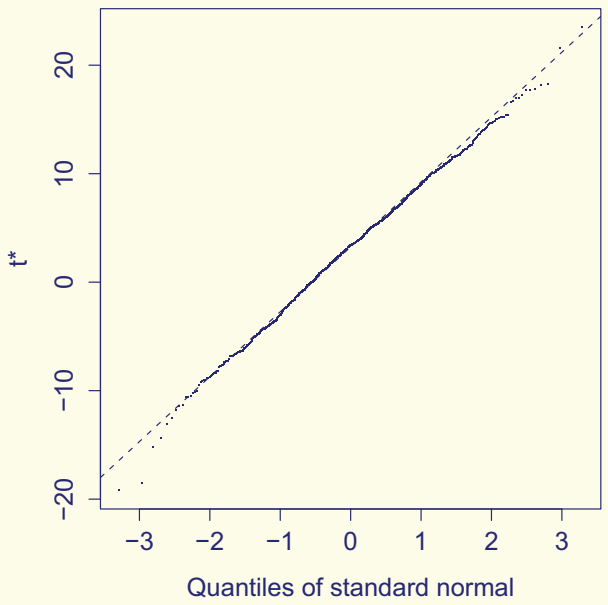



Figure 10: Normal Q–Q plot of bootstrap estimate t^* ($R = 999$) for the paired analysis. $\rightsquigarrow$ The normal Q–Q plot underlines the fact that the studentized and the other two bootstrap intervals are essentially equal.

◇ An alternative is to consider only the two-sample data and compare the means of the two populations issuing from the patients for whom just one observation is available.

⇝ The classical normal 95% confidence interval for the difference of the means is

$$(\bar{b} - \bar{r}) \pm z_{0.025} \, (s_b^2/n_b + s_r^2/n_r)^{1/2},$$

where s_b and s_r are the standard deviations of the b's and r's, and $z_{0.025}$ is the 2.5% quantile of the standard normal distribution.

• In R:

```
> acu.ts = acuity[acuity$str==1, ]
> acu.ts
   str red blue
21   1  88   36
22   1  83   86
...
29   1  87   85
30   1  65   72

> dif = mean(acu.ts$blue) - mean(acu.ts$red)
> tmp = qnorm(0.025) *
        sqrt(var(acu.ts$blue)/nrow(acu.ts) +
             var(acu.ts$red)/nrow(acu.ts))
> c(dif + tmp, dif - tmp)
[1] -13.76901  19.16901
```

- The obvious estimator and its estimated variance are

$$t = \bar{b} - \bar{r}, \quad v = s_b^2/n_b + s_r^2/n_r,$$

whose values for these data are 2.7 and 70.6.

$\rightsquigarrow$ To construct bootstrap confidence intervals we generate $R = 999$ replicates of t and v, with each simulated data set containing n_b values sampled with replacement from the bs and n_r values sampled with replacement from the rs.

$\rightsquigarrow$ In R:

```
> y = c(acuity$blue[21:30], acuity$red[21:30])
> acu = data.frame(col=rep(c(1, 2), c(10, 10)), y)
> acu
   col  y
1    1 36
2    1 86
..
10   1 72
11   2 88
..
19   2 87
20   2 65
```

```
acu.ts.fun = function(data, i)
 {
 d = data[i, ]
 m = mean(d$y[1:10]) - mean(d$y[11:20])
 v = var(d$y[1:10])/10 + var(d$y[11:20])/10
 c(m, v)
 }
```

```
> acu.ts.boot = boot(acu, acu.ts.fun, R=999, strata=acu$col)
```

⇝ Here `strata=acu$col` ensures stratified simulation.

```
> boot.ci(acu.ts.boot, type=c("norm", "basic", "perc", "stud"))
...
Normal              Basic               Percentile          Studentized
(-13.446, 19.006)  (-13.200, 19.100)  (-13.700, 18.600)  (-14.883, 21.341)
```

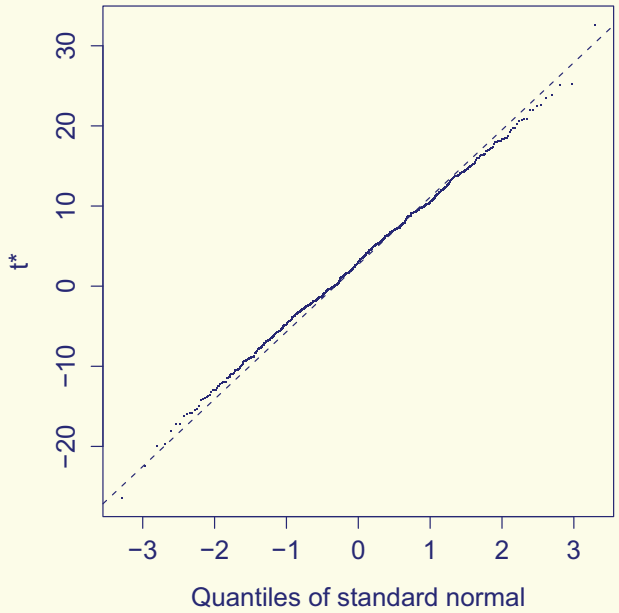



Figure 11: Normal Q–Q plot of bootstrap estimate t^* ($R = 999$) for the two-sample analysis. ⇝ The Q–Q plot is close to normal, and the bootstrap intervals differ little from the classical normal interval.

◇ We now combine the analyses, hoping that the resulting confidence interval will be shorter.

⇝ If the variances σ_d^2, σ_r^2 and σ_b^2 of the ds, rs and bs were known, a 'minimum variance unbiased' estimate of the difference between responses for blue and red treatments would be

$$\frac{n_d\bar{d}/\sigma_d^2 + (\bar{b}-\bar{r})/(\sigma_b^2/n_b + \sigma_r^2/n_r)}{n_d/\sigma_d^2 + 1/(\sigma_b^2/n_b + \sigma_r^2/n_r)}.$$

⇝ As σ_d^2, σ_r^2 and σ_b^2 are unknown, we replace them by estimates, giving estimated treatment difference and its variance

$$\begin{aligned} t &= \frac{n_d\bar{d}/\hat{\sigma}_d^2 + (\bar{b}-\bar{r})/(\hat{\sigma}_b^2/n_b + \hat{\sigma}_r^2/n_r)}{n_d/\hat{\sigma}_d^2 + 1/(\hat{\sigma}_b^2/n_b + \hat{\sigma}_r^2/n_r)}, \\ v &= \left\{n_d/\hat{\sigma}_d^2 + 1/(\hat{\sigma}_b^2/n_b + \hat{\sigma}_r^2/n_r)\right\}^{-1}. \end{aligned}$$

⇝ Here $t = 3.07$ and $v = 4.873^2$, so a naive 95% confidence interval for the treatment difference is $(-6.48, 12.62)$.

• One way to apply the bootstrap here is to generate a bootstrap data set by taking n_d pairs randomly with replacement from $\hat{F}_y$, n_b values with replacement from $\hat{F}_b$ and n_r values with replacement from $\hat{F}_r$, each resample being taken with equal probability.

```
acu.fun = function(data, i)
 {
 d = data[i, ]
 m = sum(data$str)
 if(length(unique((i)==(1:nrow(data))))!=1){
   d$blue[d$str==1] = sample(d$blue, size=m, replace=TRUE)
   d$red[d$str==1] = sample(d$red, size=m, replace=TRUE)
 }
 dif = d$blue[d$str==0] - d$red[d$str==0]
 d2 = d$blue[d$str==1]
 d3 = d$red[d$str==1]
 v1 = var(dif)/length(dif)
 v2 = var(d2)/length(d2) + var(d3)/length(d3)
 v = 1/(1/v1 + 1/v2)
 c((mean(dif)/v1 + (mean(d2) - mean(d3))/v2) * v, v)
 }
```


⇝ And

```
> acu.boot = boot(acuity, acu.fun, R=999, strata=acuity$str)
> boot.ci(acu.boot,
          type=c("norm", "basic", "perc", "stud", "bca"))
```

yields all five sets of confidence limits.

⇝ Using $R = 999$ we get $b^* = 0.22$ and $v^* = 6.086^2$.

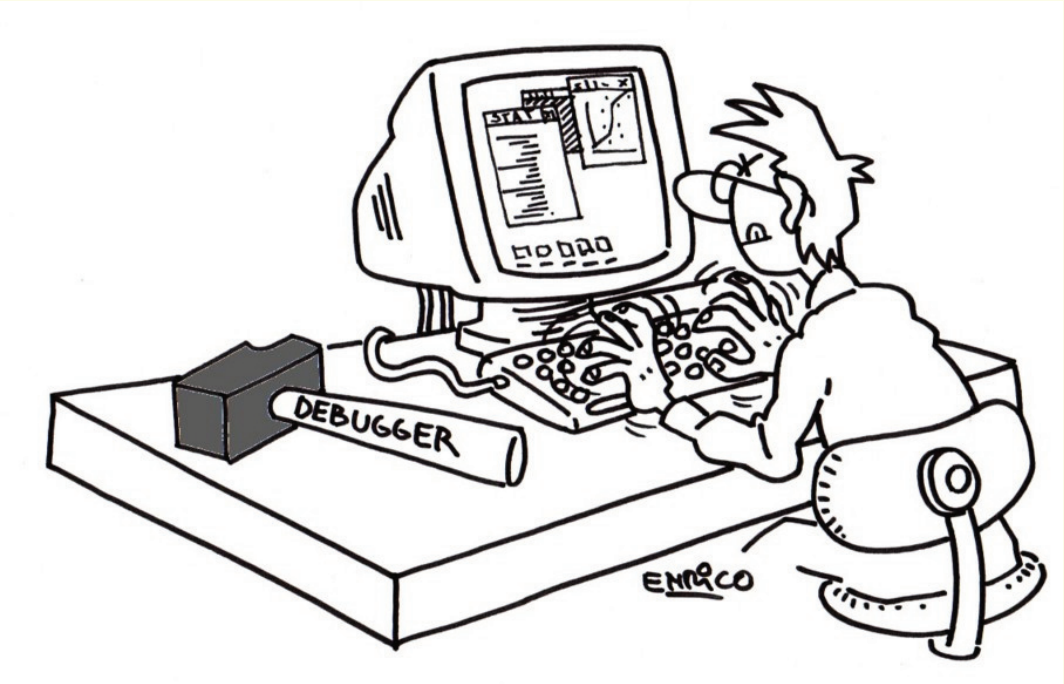
⇝ Hence 95% bootstrap confidence intervals are:

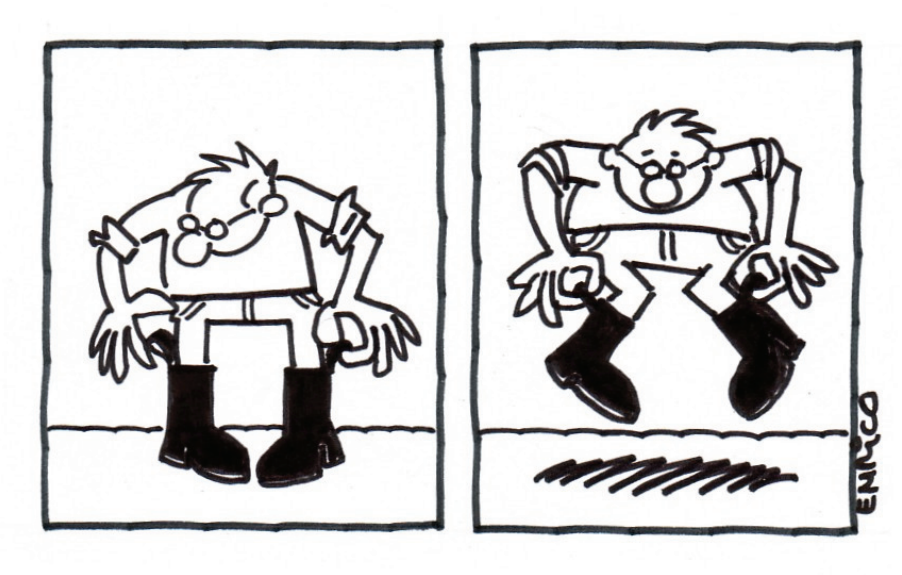
Normal	-9.08	14.78
Basic	-9.17	15.54
Percentile	-8.33	15.26
Studentized	-8.98	14.83
BC_a	-7.75	15.67

⇝ All bootstrap intervals are similar.

3.5.3 Comparison of bootstrap confidence intervals

- All tend to under-cover.
- Normal, basic and studentized intervals depend on scale.
- Percentile interval often too short (over-under-covers) but is transformation-invariant.
- Studentized intervals give best coverage overall, but
  - depend on scale, can be sensitive to V;
  - length can be very variable;
  - best on transformed scale, where V is approximately constant.
- Improved percentile intervals have same error in principle as studentized intervals, but often shorter — so lower coverage.





3.6 Significance tests

• Familiar tests include both parametric, *e.g.* Student t-tests for means, and nonparametric, *e.g.* rank tests.

$\rightsquigarrow$ The aim here is to provide more general methods for computing significance once the test statistic is chosen.

• The ingredients of a significance test problem are:

– data $y_1, \ldots, y_n$;

– the hypothetical model M_0 to be tested, *e.g.* $N(0, \sigma^2)$;

– a test statistic T chosen to be sensitive to departures from M_0, with large positive values giving evidence against M_0.

• The significance, or p-value, $p = \Pr(T \geq t \mid M_0)$ is the standard measure of evidence against M_0 — the smaller p the stronger the evidence against M_0.

• Reject M_0 at the 0.05 level, say, if $p \leq 0.05$, and 0.05 is the error rate of this rejection rule.

• General difficulties include:

– p depends upon 'nuisance' parameters, those of M_0, *e.g.* σ when M_0 is $N(0, \sigma^2)$;

– p is hard to calculate.

$\rightsquigarrow$ The first difficulty can sometimes be dealt with by appropriate choice of t, *e.g.* studentizing an estimate.

$\rightsquigarrow$ The second difficulty can be dealt with by use of approximations, or simulation — we are concerned with the latter.

3.6.1 Simulation (Monte Carlo) calculation of p

• Estimate $p = \Pr(T \geq t \mid M_0)$ by simulation from completely known null hypothesis model M_0, *i.e.* M_0 does not involve any nuisance parameters.

$\rightsquigarrow$ If M_0 completely determines a sampling distribution for the data, with no unknowns, then the algorithm is:

– For $r = 1, \ldots, R$:

⋆ simulate data set $y_1^*, \ldots, y_n^*$ from M_0;

⋆ calculate test statistic t_r^* from $y_1^*, \ldots, y_n^*$.

– Calculate simulation (Monte Carlo) estimate

$$\hat{p} = \frac{1 + \#\{t_r^* \geq t\}}{R + 1}.$$

$\rightsquigarrow$ The exact significance p is approximated by the empirical proportion of times that the test statistic equals or exceeds the observed value t in repeated sampling under the null hypothesis model M_0.

- Take R big enough to get small standard error for $\hat{p}$, typically one would need ≥ 100 to get a good approximation, if this is necessary — we usually use about $1'000$.

- Taking R too small blurs power function of test.

Example (Fir seedlings data). Data are $n = 50$ counts of balsam-fir seedlings in five feet square quadrants:

0	1	2	3	4	3	4	2	2	1
0	2	0	2	4	2	3	3	4	2
1	1	1	1	4	1	5	2	2	3
4	1	2	5	2	0	3	2	1	1
3	1	4	3	1	0	0	2	7	0

$\rightsquigarrow$ We wish to test the null hypothesis that these data are an independent random sample from a Poisson distribution with unknown mean (M_0).

$\rightsquigarrow$ The concern is that the data are over-dispersed relative to the Poisson distribution, which strongly suggests that we take as test statistic the dispersion index

$$T = \frac{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}{\bar{Y}}.$$

$\rightsquigarrow$ Under M_0, $(n-1)^{-1}T$ should be close to 1, since then the population mean and variance are equal.

$\rightsquigarrow$ Under the null hypothesis Poisson model (M_0), the conditional distribution of $Y_1, \ldots, Y_n$ given $n\bar{Y}$ is multinomial with denominator $n\bar{Y}$ and n categories each having probability n^{-1} — which is too complicated for exact calculation of p.

• But, it is easy to simulate from this multinomial distribution.

$\rightsquigarrow$ For the data, we have $t = 55.15$.

$\rightsquigarrow$ In the first $R = 99$ simulated values t^*, 24 are larger than $t = 55.15$.

$\rightsquigarrow$ So the simulation p-value is equal to $\hat{p} = 0.25$.

$\rightsquigarrow$ We conclude that the data dispersion is consistent with Poisson dispersion.

$\rightsquigarrow$ Increasing R to 999 makes little difference, giving $\hat{p} = 0.235$.

$\rightsquigarrow$ For this simple problem the theoretical null distribution of T given $n\bar{Y}$ is approximately χ^2_{n-1}.

$\rightsquigarrow$ The p-value obtained with this approximation is 0.253, close to the simulated value.

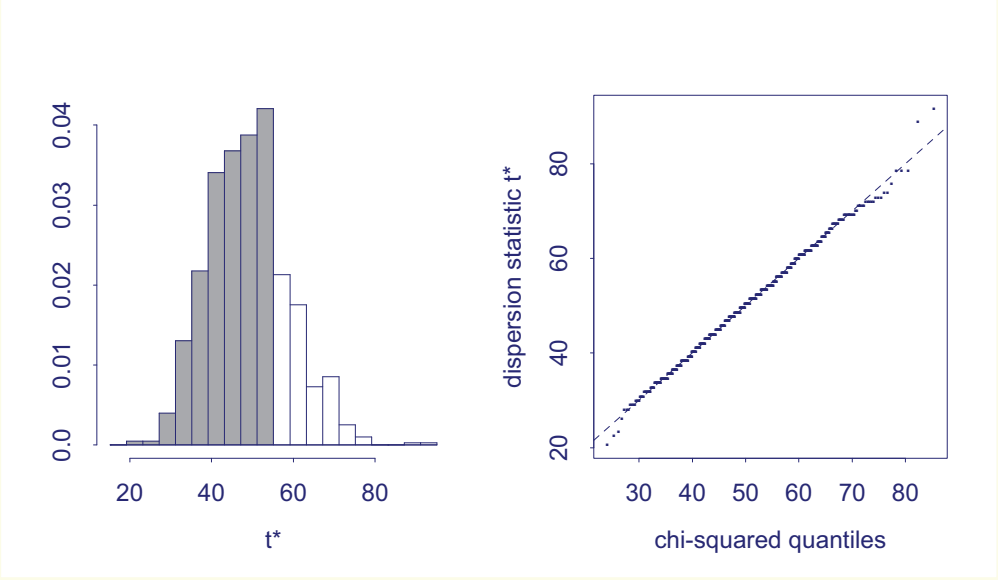


Figure 12: Simulation results for dispersion test. Left panel: histogram of $R = 999$ values of the dispersion statistic t^* obtained under multinomial sampling: the data value is $t = 55.15$ and $\hat{p} = 0.235$. Right panel: chi-squared plot of ordered values of t^*, dotted line corresponding to χ^2_{49} approximation to null conditional distribution.

Comments

• First, the simulation results enable us to check on the accuracy of the theoretical approximation: if the approximation is good, then we can use it; but if it is not, then we have the simulation p-value.

• Second, the simulation method does not require knowledge of a theoretical approximation, which may not even exist in more complicated problems, such as spatial analysis of these data.

$\rightsquigarrow$ The simulation method applies very generally.

3.6.2 Parametric bootstrap test

• Suppose that we have a parametric model for the data, with M_0 a special (parametric) case to be tested. The latter depends on nuisance parameters ψ.

$\rightsquigarrow$ If the fitted null model is denoted by $\hat{M}_0$, obtained by estimating the parameters ψ of M_0 (often MLE), the adapted algorithm is:

- Fit the model $\hat{M}_0$ to the data $y_1, \ldots, y_n$;
- For $r = 1, \ldots, R$:
  - ⋆ simulate data set $y_1^*, \ldots, y_n^*$ from $\hat{M}_0$;
  - ⋆ calculate test statistic t_r^* from $y_1^*, \ldots, y_n^*$.
- Calculate (parametric bootstrap) simulation estimate

$$\hat{p} = \frac{1 + \#\{t_r^* \geq t\}}{R + 1}.$$

Example (Two-way table). Suppose we have a two-way table of counts:

1	2	2	1	1	0	1
2	0	0	2	3	0	0
0	1	1	1	2	7	3
1	1	2	0	0	0	1
0	1	1	1	1	0	0

$\rightsquigarrow$ Table contains a set of observed multinomial counts, for which we wish to test the null hypothesis of row-column independence (M_0).

$\rightsquigarrow$ If the count in row i and column j is y_{ij}, then the null fitted values are $\hat{\mu}_{ij,0} = y_{i+}y_{+j}/y_{++}$, where $y_{i+} = \sum_j y_{ij}$ and so forth.

$\rightsquigarrow$ The log likelihood ratio or deviance test statistic is

$$t = 2 \sum_{i,j} y_{ij} \log(y_{ij}/\hat{\mu}_{ij,0}).$$

$\rightsquigarrow$ According to standard theory, T is approximately distributed as χ^2_d under the null hypothesis with $d = (5-1) \times (7-1) = 24$.

$\rightsquigarrow$ Since $t = 38.52$, the approximate p-value is $\Pr(\chi^2_{24} \geq 38.52) = 0.031$.

$\rightsquigarrow$ However, the chi-squared approximation is known to be quite poor for such a sparse table, so we apply the parametric bootstrap.

$\rightsquigarrow$ The model $\hat{M}_0$ is the fitted multinomial model, sample size $n = y_{++}$ and (i,j)th cell probability $\hat{\mu}_{ij,0}/n$.

$\rightsquigarrow$ We generate R tables from this model $\hat{M}_0$ and calculate the corresponding log likelihood ratio statistics $t^*_1, \ldots, t^*_R$.

$\rightsquigarrow$ With $R = 999$ we obtain 47 statistics larger than the observed value $t = 38.52$.

$\rightsquigarrow$ The parametric bootstrap p-value is $(1+47)/(999+1) = 0.048$.

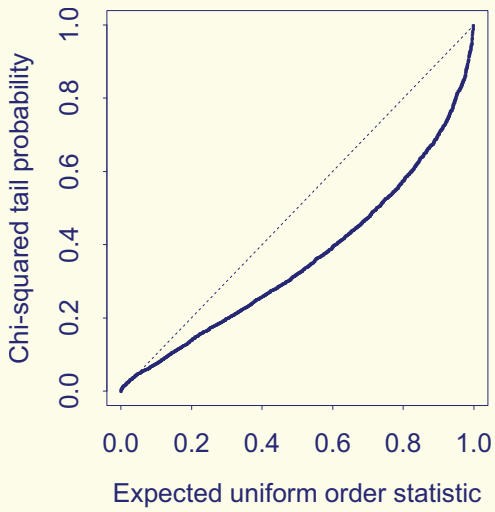


Figure 13: Ordered values of $\Pr(\chi^2_{24} \geq t^*)$ *versus* expected uniform order statistics from $R = 999$ bootstrap simulations under the null fitted model for two-way table. The dotted line corresponds to the theoretical chi-squared approximation. $\rightsquigarrow$ This illustrates the inaccuracy of the theoretical approximation. Indeed, ideally the p-values should be uniformly distributed on $(0, 1)$ if the usual error rate interpretation is to be valid. However, here the bootstrap p-value turns out to be quite non-uniform.

3.6.3 Nonparametric bootstrap test

- For more general application we copy the idea of the parametric bootstrap test.

$\rightsquigarrow$ This requires fitting the model $\hat{M}_0$ without full parametric assumptions.

$\rightsquigarrow$ There is no unique way to do this.

$\rightsquigarrow$ One general principle is to ensure that $\hat{M}_0$ makes practical sense.

$\rightsquigarrow$ Adapted algorithm is:

- Fit the model $\hat{M}_0$ to the data $y_1, \ldots, y_n$;
- For $r = 1, \ldots, R$:
  - $\star$ simulate data set $y_1^*, \ldots, y_n^*$ from $\hat{M}_0$;
  - $\star$ calculate test statistic t_r^* from $y_1^*, \ldots, y_n^*$.
- Calculate (nonparametric bootstrap) simulation estimate

$$\hat{p} = \frac{1 + \#\{t_r^* \geq t\}}{R + 1}.$$

Example (Correlation test). Suppose that $Y = (U, X)$ is a random pair and that n such pairs are observed.

$\rightsquigarrow$ The objective is to see if U and X are independent, this being the null hypothesis H_0.

• An illustrative data set is given by the Claridge data, where $u = \texttt{dnan}$ is a genetic measure and $x = \texttt{hand}$ is an integer measure of left-handedness.

$\rightsquigarrow$ The alternative hypothesis is that x tends to be larger when u is larger.

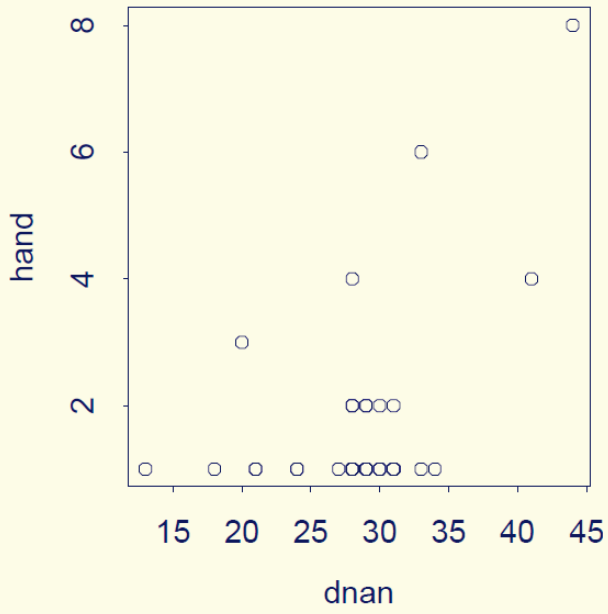


Figure 14: Scatter plot of $n = 37$ pairs of measurements in a study of handedness (provided by Gordon Claridge, University of Oxford, United Kingdom).

- One simple test statistic is the sample correlation, $T = \rho(F_n)$ say.

$\rightsquigarrow$ The correlation coefficient for the Claridge data is $t = 0.509$.

- The correlation is zero for any distribution that satisfies H_0.

- Note that here the EDF F_n puts probabilities n^{-1} on each of the n data pairs (u_i, x_i).

• If the two variables are independent, the distribution of $Y = (U, X)$ factorises, to give

$$F(y) = F(u, x) = F_1(u)F_2(x),$$

and under the null hypothesis the closest equivalent to this is

$$F_{0,n}(y) = F_{1,n}(u)F_{2,n}(x),$$

where $F_{1,n}$ is the EDF of the u_i, and $F_{2,n}$ is the EDF of the x_i.

$\rightsquigarrow$ Under the null hypothesis we therefore generate bootstrap samples as

$$(u_1^*, x_1^*), \ldots, (u_n^*, x_n^*),$$

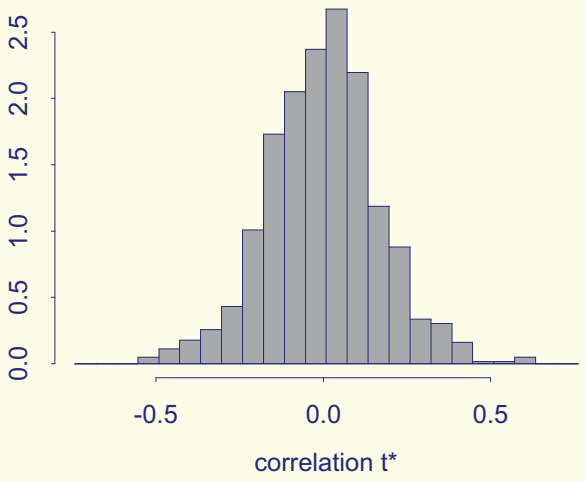
where the u_i^* generated from $F_{1,n}^*$, and the x_i^* are generated from $F_{2,n}^*$.

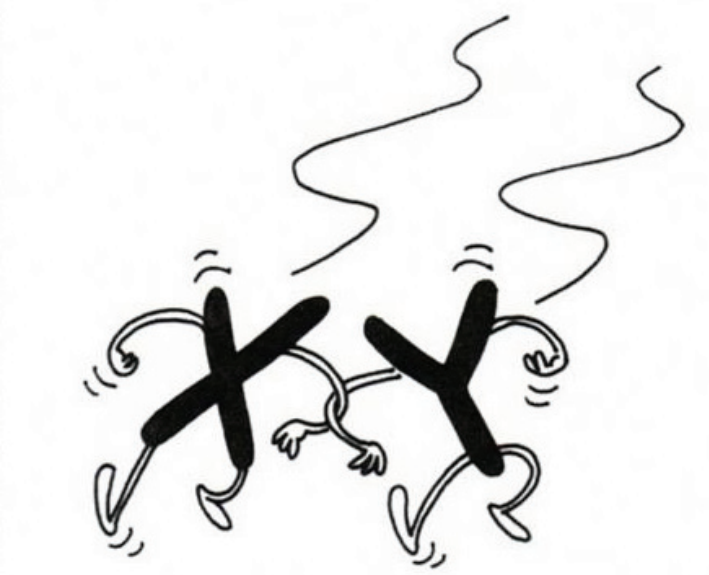
$\rightsquigarrow$ This means that the u_i and x_i are essentially detached from one another under the null hypothesis sampling distribution.

• We then repeat this sampling R times, and use the values t_r^* generated to calculate the nonparametric bootstrap p-value $\hat{p}$.

$\rightsquigarrow$ With $R = 999$ the bootstrap p-value is $\hat{p} = (1 + 6)/(999 + 1) = 0.007$.

$\rightsquigarrow$ Histogram of correlation t^* values for the $R = 999$ bootstrap samples:





3.7 Simple linear regression

- Independent data $(x_1, y_1), \ldots, (x_n, y_n)$.
- The simple linear regression model is

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, \ldots, n,$$

where the ε_j's are uncorrelated with 0 means and equal variances σ^2.

$\rightsquigarrow$ Ordinary least squares estimates for $\beta = (\beta_0, \beta_1)^T$ are

$$\hat{\beta}_1 = \frac{\sum_{j=1}^n (x_j - \bar{x}) y_j}{SS_x}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{x} = n^{-1} \sum_{j=1}^n x_j$ and $SS_x = \sum_{j=1}^n (x_j - \bar{x})^2$. Conventional estimate of the error variance σ^2 is $\hat{\sigma}^2 = (n-2)^{-1} \sum_{j=1}^n r_j^2$, where $r_j = y_j - \hat{y}_j$ are raw residuals with $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j$ the fitted values, or estimated mean values, for the response at the observed x values.

• For linear regression with normal random errors ε_j having constant variance, the least squares theory of regression estimation and inference provides clean, exact and optimal methods for analysis.

• But for generalisations to non-normal errors and non-constant variance, exact methods rarely exist, and we are faced with approximate methods based on linear approximations to estimators and central limit theorems.

$\rightsquigarrow$ So, resampling methods have the potential to provide more accurate analysis.

• With simple least squares linear regression, where in ideal conditions resampling essentially reproduces the exact theoretical analysis, but also offers the potential to deal with non-ideal circumstances such as non-constant variance.

Case resampling in linear regression

• Case resampling, *i.e.* resampling cases (x_1, y_1), ..., (x_n, y_n):

For $r = 1, \ldots, R$,

⋆ sample $i_1^*, \ldots, i_n^*$ randomly with replacement from $\{1, 2, \ldots, n\}$;

⋆ for $j = 1, \ldots, n$, set $x_j^* = x_{i_j^*}$, $y_j^* = y_{i_j^*}$; then

⋆ fit least squares regression to $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$, giving estimates $\hat{\beta}_{0,r}^*$, $\hat{\beta}_{1,r}^*$, $\hat{\sigma}_r^{*2}$.

• Varying design but robust to model failure; assumes (x_j, y_j) sampled from population.

$\rightsquigarrow$ Usually design variation no problem; can prove awkward when n is small, or if a few observations have a strong influence on some aspect of the design.

Model-based resampling in linear regression

- Change raw residual r_j to have constant variance, that is using the modified residuals $e_j = r_j/(1-h_j)^{1/2}$, where $h_j = n^{-1} + (x_j - \bar{x})^2/SS_x$ are the leverages.
- Model-based resampling, *i.e.* residuals resampling:

  For $r = 1, \ldots, R$,
  - for $j = 1, \ldots, n$,
    - ⋆ set $x_j^* = x_j$;
    - ⋆ randomly sample ε_j^* from centred modified residuals $e_1 - \bar{e}, \ldots, e_n - \bar{e}$;
    - ⋆ set $y_j^* = \hat{\beta}_0 + \hat{\beta}_1 x_j + \varepsilon_j^*$.
  - Fit least squares regression to $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$, giving estimates $\hat{\beta}_{0,r}^*$, $\hat{\beta}_{1,r}^*$, $\hat{\sigma}_r^{*2}$.
- Fixes design but not robust to model failure; assumes ε_j sampled from population.

$\rightsquigarrow$ Careful model-checking is needed before it can be used!

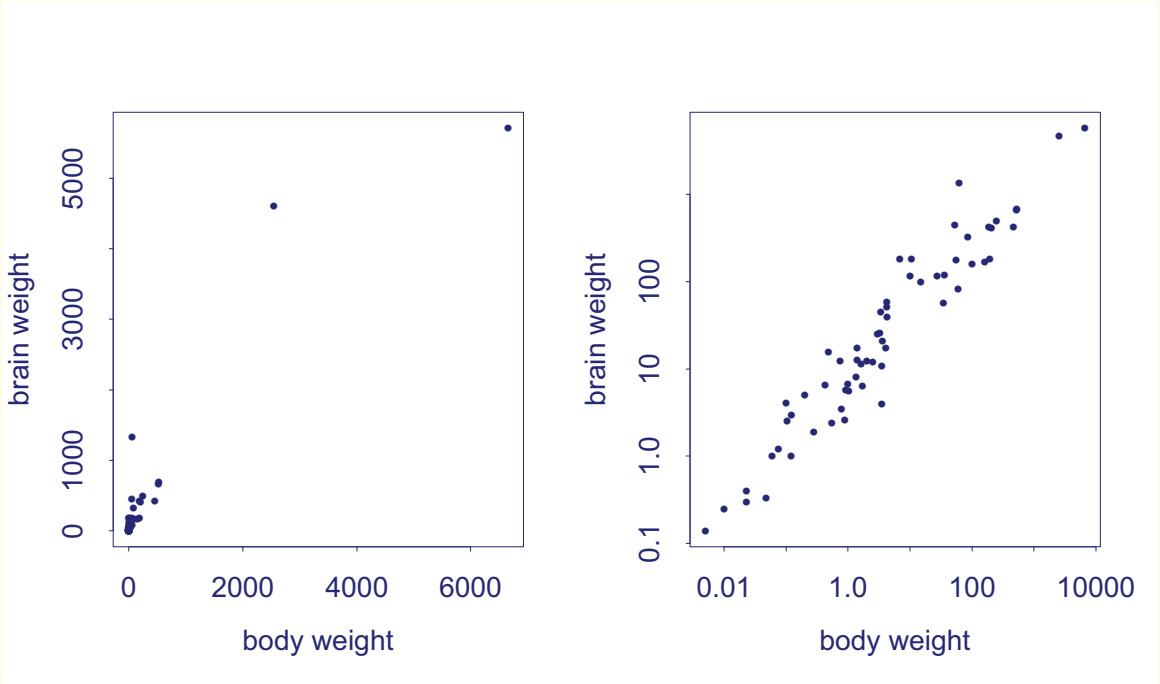


Figure 15: Mammals data. Average body weight (kg) and brain weight (g) for 62 species of mammals, plotted on original scales and logarithmic scales.

Example (Mammals data, continued). Data well-described by a simple linear regression after the two variables are transformed logarithmically, so that

$$y = \log(\text{brain weight}), \quad x = \log(\text{body weight}).$$

- The simple linear regression model is

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, \ldots, n,$$

where the ε_j's are uncorrelated with zero means and equal variances σ^2.

$\rightsquigarrow$ This constancy of variance, or homoscedasticity, seems roughly right.

$\rightsquigarrow$ Standard analysis suggests that errors are approximately normal.

$\rightsquigarrow$ The standard parameter estimates are $\hat{\beta}_0 = 2.135$ and $\hat{\beta}_1 = 0.752$.

		Theoretical	Model-based resampling	Resampling cases
$\hat{\beta}_0^*$	bias	0	0.0004	0.0006
	standard error	0.0960	0.0958	0.091
$\hat{\beta}_1^*$	bias	0	0.0001	0.0002
	standard error	0.0285	0.0285	0.0223

Table 3: Mammals data. Comparison of bootstrap biases and standard errors of intercept and slope with theoretical standard results. Model-based resampling with $R = 499$ and resampling cases with $R = 999$. $\rightsquigarrow$ So, as expected for a moderately large 'clean' data set, the resampling results agree closely with those obtained from standard methods.

Example (Survival data). The `survival` data are survival percentages for rats at a succession of doses of radiation, with two or three replicates at each dose.

$\rightsquigarrow$ The data come with the package `boot` and can be loaded using

```
> require(boot)
> survival
     dose   surv
1   117.5 44.000
...
14 1410.0  0.006
```

• The theoretical relationship between survival rate (`surv`) and dose (`dose`) is exponential, so linear regression applies to

$$x = \texttt{dose}, \quad y = \log(\texttt{surv}).$$

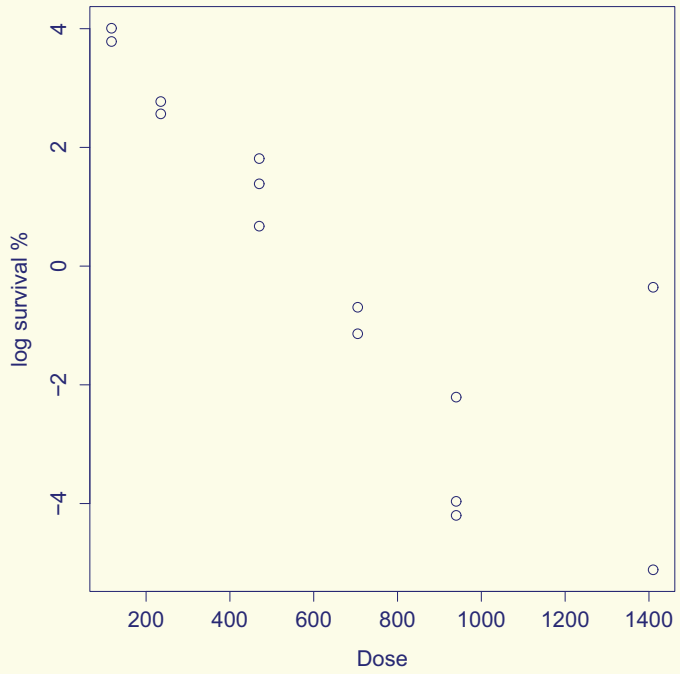


Figure 16: Scatter plot of `survival` data.

$\rightsquigarrow$ There is a clear outlier, case 13, at $x = 1'410$.

• The least squares estimate of slope is -59×10^{-4} using all the data, changing to -78×10^{-4} with standard error 5.4×10^{-4} when case 13 is omitted.

$\rightsquigarrow$ To illustrate the potential effect of an outlier in regression we resample cases, using

```
surv.fun = function(data, i)
 {
 d = data[i, ]
 d.reg = lm(log(d$surv) ~ d$dose)
 c(coef(d.reg))
 }
```

```
> surv.boot = boot(survival, surv.fun, R=999)
```

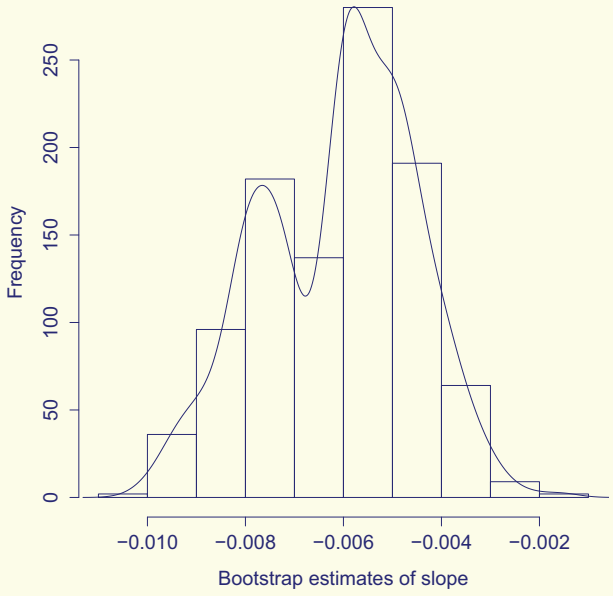



Figure 17: Histogram of $R = 999$ bootstrap estimates of least squares slope $\hat{\beta}_1^*$ with superposed kernel density estimate, showing the effect of the outlier on the resampled estimates.

$\rightsquigarrow$ The two groups of bootstrapped slopes correspond to resamples in which case 13 does not occur and to samples where it occurs once or more.

$\rightsquigarrow$ The resampling standard error of $\hat{\beta}_1^*$ is 15.6×10^{-4}, but only 7.8×10^{-4} for samples without case 13.

• A 'jackknife-after-bootstrap' plot shows the effect on $T^* - t$ of resampling from data sets from which each of the observations has been removed.

$\rightsquigarrow$ In R:

```
> jack.after.boot(surv.boot, index=2)
```

$\rightsquigarrow$ We expect deletion of case 13 to have a strong effect.

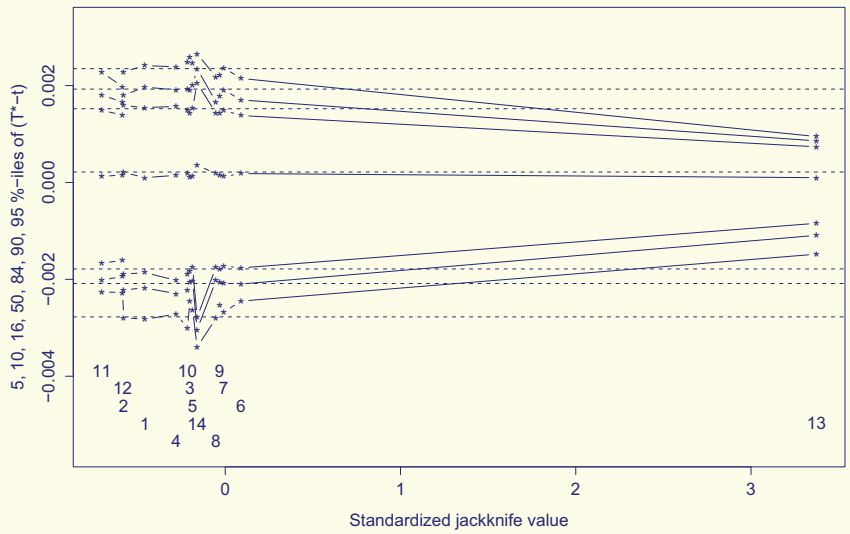
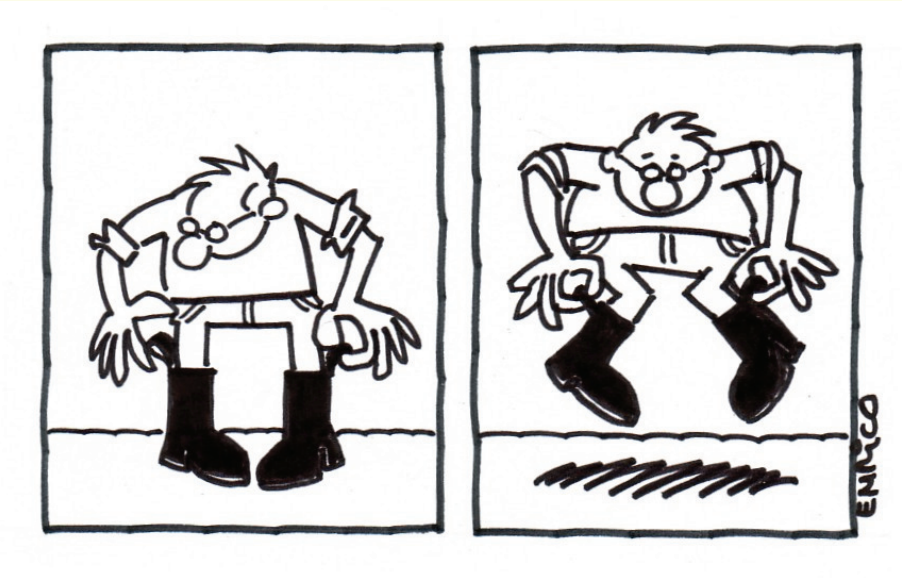


Figure 18: 'Jackknife-after-bootstrap' plot for the slope. The vertical axis shows quantiles of $T^* - t$ for the full sample (horizontal dotted lines) and without each observation in turn, plotted against the influence value for that observation. $\rightsquigarrow$ This shows clearly that this case has an appreciable effect on the resampling distribution, and that its omission would give much tighter confidence limits on the slope.



Conclusion

• Bootstrap resampling allows fast (the computer time involved is often close to negligible) empirical assessment of standard approximations, and may indicate ways to fix them when they fail.

• Bootstrap methods offer considerable potential for modelling in complex problems.

• In principle the estimator chosen should be appropriate to the model used, or there is a loss of efficiency.

• In practise, however, there is often some doubt about the exact error structure, and a well-chosen resampling scheme can give inferences robust to the precise error structure of the data.

• Although the bootstrap is sometimes touted as a replacement for 'traditional statistics', this belief is misguided.

• It is unwise to use a powerful tool without understanding why it works, and the bootstrap rests on 'traditional' ideas, even if their implementation via simulation is not 'traditional'.

$\rightsquigarrow$ Populations, parameters, samples, sampling variation, pivots and confidence limits are fundamental statistical notions, and it does no one a service to brush them under the carpet.

• Indeed, it is harmful to pretend that mere computation can replace thought about central issues such as the structure of a problem, the type of answer required, the sampling design and data quality.

• Moreover, as with any simulation experiment, it is essential to monitor the output to ensure that no unanticipated complications have arisen and to check that the results make sense, and this entails understanding how the output will be used.

'Despite its scope and usefulness, resampling must be carefully applied. Unless certain basic ideas are understood, it is all too easy to produce a solution to the wrong problem, or a bad solution to the right one. Bootstrap methods are intended to help avoid tedious calculations based on questionable assumptions, and this they do. But they can not replace clear critical thought about the problem, appropriate design of the investigation and data analysis, and incisive presentation of conclusions.'

Anthony C. Davison and David V. Hinkley, 1997

References and resources

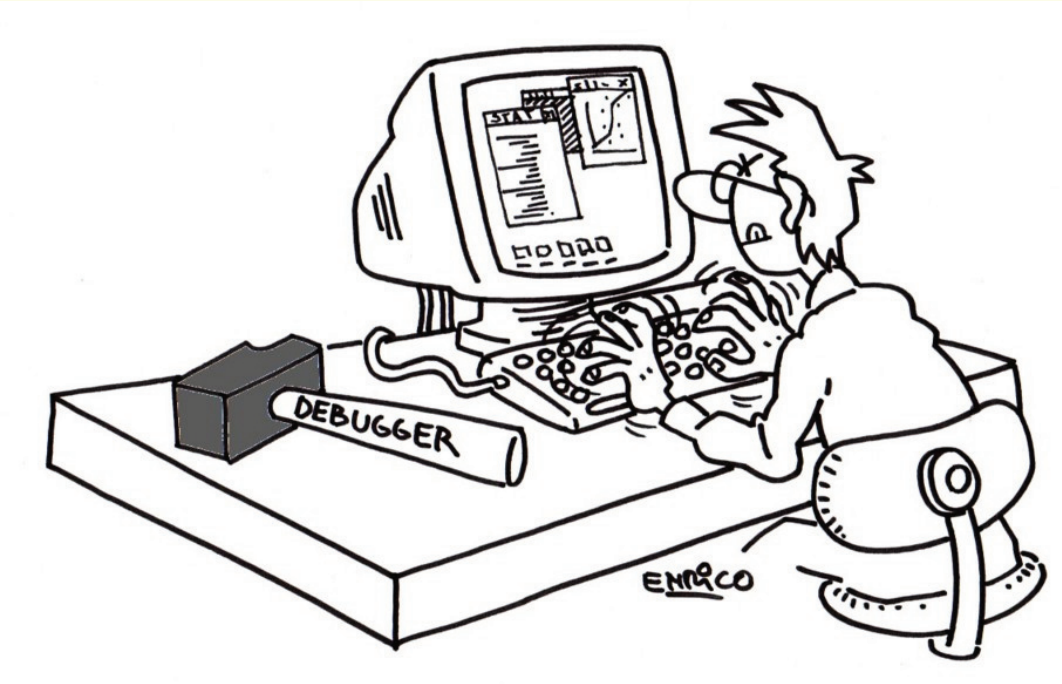
- Boos, D. D. & Osborne, J. A. (2015). Assessing variability of complex descriptive statistics in Monte Carlo studies using resampling methods. *International Statistical Review*, **25**, 775–792.
- Carpenter, J. & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, **19**, 1141–1164.
- Chernick, M. R. & LaBudde, R. A. (2011). *An Introduction to Bootstrap Methods with Applications to R*. Hoboken, NJ: Wiley.
- Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University.
- Davison, A. C. & Kuonen, D. (2002). An introduction to the bootstrap with applications in R. *Statistical Computing and Statistical Graphics Newsletter*, **13**, 6–11.
- Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, **248**, 96–108.
- DiCiccio, T. J. & Efron, B. (1996). Bootstrap confidence intervals (with Discussion). *Statistical Science*, **11**, 189–228.

- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 1–26.
- Efron, B. & Tibshirani, R. J. (1991). Statistical analysis in the computer age. *Science*, **253**, 390–395.
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Friedl, H. & Stampfer, E. (2006). Jackknife resampling. In *Encyclopedia of Environmetrics*. New York: Wiley.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer.
- Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, **3**, 497–526.
- Hesterberg, T. (2015). What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. *The American Statistician*, **69**, 371–386.
- Hesterberg, T. (2018). *Bootstrap and Resampling Resources*.
  ⇝ Online at `www.timhesterberg.net/bootstrap`
- Kuonen, D. (2005a). Studentized bootstrap confidence intervals based on M-estimates. *Journal of Applied Statistics*, **32**, 443–460.

- Kuonen, D. (2005b). Saddlepoint approximations to studentized bootstrap distributions based on M-estimates. *Computational Statistics*, **20**, 231–244.
- Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B*, **11**, 68–84.
- Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353-360.
- Shalizi, C. (2013). *36-402, Undergraduate Advanced Data Analysis Course, Spring 2013*. Pittsburgh, PA: Department of Statistics, Carnegie Mellon University.
  ⇝ Online at `www.stat.cmu.edu/~cshalizi/uADA/13/`
- Shao, J. & Tu, D. (1995). *The Jackknife and Bootstrap.* New York: Springer.
- Special issue ('Silver Anniversary of the Bootstrap') of *Statistical Science*, **18** (May 2003; `www.imstat.org/sts/`).
- Tukey, J. W. (1958). Bias and confidence in not quite large samples (Abstract). *The Annals of Mathematical Statistics*, **29**, 614.
- Young, G. A. (1994). Bootstrap: more than a stab in the dark (with Discussion)? *Statistical Science*, **9**, 382–415.

‘What we have to learn to do, we learn by doing.’

Aristotle



Have you been Statooed?

Prof. Dr. Diego Kuonen, CStat PStat CSci

Statoo Consulting

Morgenstrasse 129

3018 Berne

Switzerland

email kuonen@statoo.com

@DiegoKuonen

web www.statoo.info

Presentation code: 'WBL.Statistik.ETHZ.2018'.

Typesetting: LaTeX, version 2ε. PDF producer: pdfTeX, version 3.141592-1.40.3-2.2 (Web2C 7.5.6).

Compilation date: 12.01.2018.