Selection effects

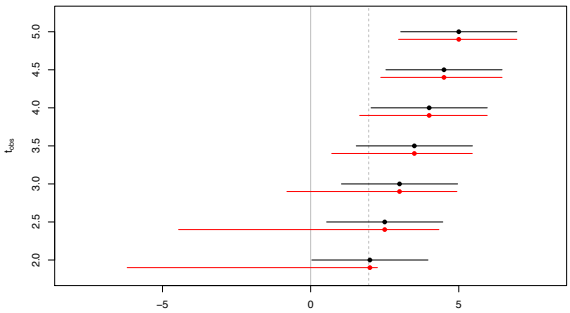
- Contrast
  - **exploratory analysis**, where we study data with no strong prior hypotheses, aiming to find something 'interesting' for future study, and
  - **confirmatory analysis**, where we specify an analysis protocol (hypotheses/tests/...) in advance and stick to it.
- Most statistical procedures assume we are doing the second, but there can be a strong temptation to cheat and treat an exploratory analysis as confirmatory.
- In 'the garden of forking paths' we make a series of choices (which response? transformation? which explanatory variables? ...) but do not then allow for them.
- This leads to non-reproducible results, 'false discoveries', bad science ...
- If we compute a confidence interval $\mathcal{I}$ for θ following a sequence of choices summarised in a selection event $\mathcal{S}$ that is *based on the same data*, and compute

$$\mathrm{P}(\theta \in \mathcal{I}) \quad \text{when we should compute} \quad \mathrm{P}(\theta \in \mathcal{I} \mid \mathcal{S}),$$

  we are effectively pretending that $\mathcal{S}$ did not exist.

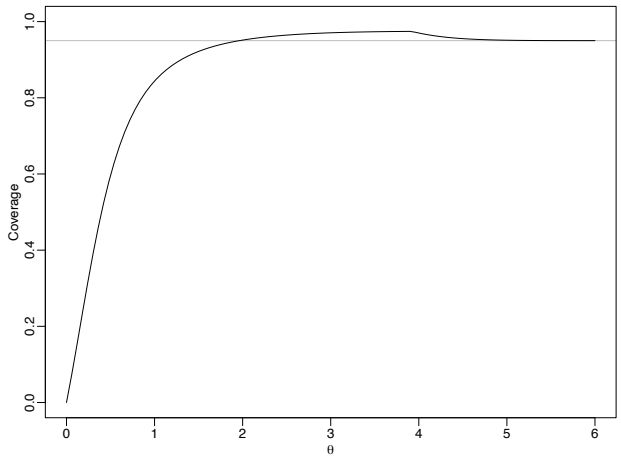
Simple example

Example 68 *Suppose $T \sim \mathcal{N}(\theta, 1)$ and we perform a two-sided test of $H_0 : \theta = 0$ at level $\alpha = 5\%$ and then construct a 95% confidence interval $\mathcal{I}_{95}$ around the observed t_{obs} if we reject H_0. Compare the resulting confidence intervals when we do and do not allow for selection. What is the coverage of $\mathcal{I}_{95}$ conditional on $\mathcal{S}$?*



95% confidence intervals for θ without (black) and with (red) allowance for selection on event $\mathcal{S} = \{T > z_{0.975}\}$.

Simple example II



Conditional coverage $\mathrm{P}(\theta \in \mathcal{I}_{95} \mid \mathcal{S})$ of $\mathcal{I}_{95}$ as a function of θ.

Note to Example 68

- Recall the basis of confidence intervals for θ based on an estimator T satisfying $T \sim \mathcal{N}(\theta, 1)$. We use the fact that $T - \theta \sim \mathcal{N}(0, 1)$ to argue that

$$\mathrm{P}(T \leq t_{\mathrm{obs}}) = \mathrm{P}(T - \theta \leq t_{\mathrm{obs}} - \theta) = \Phi(t_{\mathrm{obs}} - \theta)$$

  and then set this equal to α, $1 - \alpha$ to obtain the $(1 - 2\alpha)$ confidence interval $(t_{\mathrm{obs}} - z_{1-\alpha}, t_{\mathrm{obs}} - z_{\alpha})$, which reduces to the 95% confidence interval $\mathcal{I}_{95}$ with limits $t_{\mathrm{obs}} \pm 1.96$ when $\alpha = 0.025$.

- If we condition on the selection event $\mathcal{S}_R = \{T > z_{1-\beta}\}$ and, compute the 95% confidence interval for θ if this event occurs, we are effectively using the conditional distribution

$$\begin{aligned} \mathrm{P}(T \leq t_{\mathrm{obs}} \mid T > z_{1-\beta}) &= \mathrm{P}(T - \theta \leq t_{\mathrm{obs}} - \theta \mid T - \theta > z_{1-\beta} - \theta) \\ &= \frac{\Phi(t_{\mathrm{obs}} - \theta) - \Phi(z_{1-\beta} - \theta)}{1 - \Phi(z_{1-\beta} - \theta)} \end{aligned}$$

  and the $(1 - 2\alpha)$ interval for θ has as endpoints the solutions to

$$\frac{\Phi(t_{\mathrm{obs}} - \theta) - \Phi(z_{1-\beta} - \theta)}{1 - \Phi(z_{1-\beta} - \theta)} = \alpha, 1 - \alpha.$$

- If we set $\beta = \alpha = 0.025$, then we get the limits shown in the graph, which shows that even having $t_{\mathrm{obs}} = 3$ still leads to a 95% CI that contains 0 when we allow for selection. Hence making allowance for selection can radically change inferences, especially when H_0 is only just rejected.

- The second graph shows that if we ignore the selection and just use the interval $\mathcal{I}_{95}$ after observing the event $\mathcal{S} = \{|T| > z_{0.975}\}$, then the true coverage varies from 0 when $\theta = 0$ to 0.95 when $\theta \to \infty$, but does not pass its nominal value until $\theta > 2$.

Allowing for selection

- Lots of work in last decade, in two main categories:
- methods for specific algorithms (e.g., the lasso) with a selection event $\mathcal{S}$ of a specified form and for which $f(\mathcal{Y} \mid \mathcal{S})$ is tractable;
- more general approaches, including
  - methods that allow for all possible selection procedures, and hence are hyper-conservative (e.g., so-called universal inference, e-values, ...);
  - splitting the data into two or more groups (below);
  - adding noise (less general, since strictly applicable only to certain setttings).
- Garcia Rasines and Young (2023, *Biometrika*) have a good discussion and more references.

Sample splitting

- **Sample splitting** is a standard approach to dealing with selection.
- Partition (independent) original data $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ at random into subsets $\mathcal{Y}_0$ and $\mathcal{Y}_1$, of respective sizes $n_0 = pn$ and $n_1 = (1-p)n$, use $\mathcal{Y}_0$ for selection, and then perform inferences using $\mathcal{Y}_1$.
- As $\mathcal{Y}_1 \perp\!\!\!\perp \mathcal{Y}_0$ and $\mathcal{S}$ depends only on $\mathcal{Y}_0$, we have

$$f(\mathcal{Y}) = f(\mathcal{Y} \mid \mathcal{S}) f(\mathcal{S}) = f(\mathcal{Y}_0, \mathcal{Y}_1 \mid \mathcal{S}) f(\mathcal{S}) = f(\mathcal{Y}_1) f(\mathcal{Y}_0 \mid \mathcal{S}) f(\mathcal{S}),$$

  so any inference based on $\mathcal{Y}_1$ is unaffected by the selection.
- This approach is simple and widely applicable (at least for random samples), but
  - if the split is random, selections and inferences may be different for different splits;
  - there is a loss of power, both for finding any effects (using $\mathcal{Y}_0$) and for inference for them (using $\mathcal{Y}_1$);
  - if the data are not a random sample (e.g., in a regression setup, (y, x), with x treated as constant), then we should aim for similar information contents in $\mathcal{Y}_0$ and $\mathcal{Y}_1$ (more formally, ancillary statistics should be similar for both parts), and it may be hard to achieve this, particularly in high dimensions.

Randomisation

- ☐ Data (Y, X), with X (if present) treated as constant
- ☐ Have random variable W, maybe dependent on X, and base selection on $U = u(Y, W)$, e.g., setting selection variable $S = s(U)$ equal to s.
- ☐ Then base inference on $Y \mid U$, which is conditionally independent of S.
- ☐ If $Y \mapsto (U, V) = (u(Y, W), v(Y, W))$, where (U, V) are jointly sufficient for model and $U \perp\!\!\!\perp V$, then inference from $Y \mid U$ is equivalent to inference from V.
- ☐ Simple example: $Y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$ with σ^2 known, $U = Y + \sigma p W$ and $V = Y - \sigma p^{-1} W$, where $W \sim \mathcal{N}_n(0, I_n)$ for some $p \in (0, 1)$:
  - if $p \approx 0$, then $U \approx Y$ and the selection will be nearly the same as with the original data, but the inference will be poor because $V \not\approx Y$;
  - if $p \approx 1$, then $V \approx Y$ and the inference will be good but $U \not\approx Y$ so the selection may be very different from that based on Y.
  - Implies context-based trade-off between selection and inference.
- ☐ It can be shown that this beats sample splitting, at least in some special cases.

Example 69 *Discuss randomisation in Example 68.*

Note to Example 69

- ☐ Here $T \sim \mathcal{N}(\theta, 1)$, so we would take $U = T + pW$, where $W \sim \mathcal{N}(0, 1)$ independent of T. Note that if we set $V = T - W/p$, then

$$U \sim \mathcal{N}(\theta, 1 + p^2), \quad V \sim \mathcal{N}(\theta, 1 + 1/p^2), \quad \operatorname{cov}(U, V) = 0,$$

so $U \perp\!\!\!\perp V$, and we can write

$$T = \frac{U + p^2 V}{1 + p^2}.$$

Hence

$$T \mid U = u \stackrel{\mathrm{D}}{=} \frac{u + p^2 V}{1 + p^2},$$

which is equivalent to using the normal distribution of V for inference, as p and u are known.

Implications

- ☐ Need to be aware of possibility of selection effects and to read the literature critically.
- ☐ Must be clear if a study is exploratory or confirmatory:
  - if confirmatory, need to clarify protocol for inference **beforehand**;
  - if exploratory, need to avoid (any?) conclusions that might be due to 'forking paths'.
- ☐ Very active area of research, likely to keep on changing in next few years.
- ☐ At present it looks like randomisation is a good approach in cases with simple sufficient statistics . . . and asymptotically when σ^2 can be estimated reasonably well.

5 Bootstrap Inference

slide 166

5.1 Basic Notions

slide 167

Parameters and functionals

- ☐ Parametric models are determined by a finite vector $\theta \in \Theta$. Does this generalise?
- ☐ If $Y \sim G$, then we can define a parameter in terms of a **statistical functional**, e.g.,

$$\mu = t_1(G) = \int y \, \mathrm{d}G(y), \quad \sigma^2 = t_2(G) = \int y^2 \, \mathrm{d}G(y) - \left\{ \int y \, \mathrm{d}G(y) \right\}^2.$$

- ☐ Below we always assume that such functionals are well-defined.
- ☐ We apply the '**plug-in principle**' and replace G by an estimator $\widehat{G}$, giving

$$\widehat{\mu} = t_1(\widehat{G}) = \int y \, \mathrm{d}\widehat{G}(y), \quad \widehat{\sigma}^2 = t_2(\widehat{G}) = \int y^2 \, \mathrm{d}\widehat{G}(y) - \left\{ \int y \, \mathrm{d}\widehat{G}(y) \right\}^2.$$

- ☐ With a parametric model we can write $G \equiv G_\theta$ and $\widehat{G} \equiv G_{\widehat{\theta}}$, but a general estimator of G based on $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} G$ is the **empirical distribution function (EDF)**

$$\widehat{G}(y) = \frac{1}{n} \sum_{j=1}^{n} H(y - Y_j), \quad H(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0, \end{cases}$$

  where $H(\cdot)$ is the **Heaviside function**.

Algorithmic approach

Example 70 *Give general definitions of the median and the parameter obtained from a maximum likelihood fit of a density $f(y;\theta)$. What are the corresponding estimators (a) under a fitted exponential model, and (b) a nonparametric model?*

- ☐ This approach is essentially algorithmic: $t(\cdot)$ is an algorithm that
  - when applied to the distribution G gives the parameter $t(G)$;
  - when applied to an estimator $\widehat{G}$ based on data $Y_1, \ldots, Y_n$ gives the estimator $t(\widehat{G})$.
- ☐ The algorithm $t(\cdot)$ can be (almost) arbitrarily complex.
- ☐ This point of view suggests a sampling approach to frequentist inference:
  - if we knew G, we could assess the properties of $t(\widehat{G})$ by generating many samples $\widehat{G} \equiv \{Y_1, \ldots, Y_n\}$ from G and looking at the corresponding values of $t(\widehat{G})$;
  - since G is unknown, we replace it by $\widehat{G}$, generate samples $\widehat{G}^* \equiv \{Y_1^*, \ldots, Y_n^*\}$ from $\widehat{G}$, and use the corresponding values of $t(\widehat{G}^*)$ to estimate the distribution of $t(\widehat{G})$.
- ☐ The samples $\widehat{G}^* \equiv \{Y_1^*, \ldots, Y_n^*\}$ are known as **bootstrap samples**, and the overall procedure is known as a **bootstrap**, one of many possible **resampling** procedures.

Note to Example 70

☐ The usual definition of the p quantile is

$$t_1(G) = \inf\{x : G(x) \geq p\},$$

for $p \in (0, 1)$. For the median we set $p = 1/2$.

☐ The maximum likelihood estimator is defined as

$$t_2(G) = \text{argmax}_\theta \mathrm{E}_G\{\log f(Y; \theta)\} = \text{argmax}_\theta \int \log f(y; \theta)\, \mathrm{d}\widehat{G}(y),$$

which we earlier called θ_g.

☐ Under an exponential model

$$t_1(G) = \inf\{x : 1 - \exp(-\lambda x) \geq p\} = -\lambda^1 \log(1 - p) = \lambda^{-1} \log 2,$$

so if the fitted model has parameter $\widehat{\lambda}$, then $t_1(\widehat{G}) = \widehat{\lambda}^{-1} \log 2$.
Likewise θ_g is estimated by

$$\text{argmax}_\theta \int \log f(y; \theta)\, \widehat{\lambda} e^{-\widehat{\lambda} y}\, \mathrm{d}y;$$

note that f is not necessarily exponential.

☐ Under the general model and with order statistics $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$,

$$t_1(\widehat{G}) = \inf\{x : \widehat{G}(x) \geq p\} = Y_{(m)},$$

where $m = \lfloor (n + 1)/2 \rfloor$, and as $\mathrm{d}H(u)$ puts a unit mass at $u = 0$,

$$\begin{aligned}
t_2(\widehat{G}) &= \text{argmax}_\theta \int \log f(y; \theta)\, \mathrm{d}\widehat{G}(y) \\
&= \text{argmax}_\theta \int \log f(y; \theta)\, \mathrm{d}\left\{ n^{-1} \sum_{j=1}^{n} H(y - Y_j) \right\} \\
&= \text{argmax}_\theta n^{-1} \sum_{j=1}^{n} \int \log f(y; \theta)\, \mathrm{d}H(y - Y_j) \\
&= \text{argmax}_\theta n^{-1} \sum_{j=1}^{n} \log f(Y_j; \theta),
\end{aligned}$$

i.e., the maximum likelihood estimator of θ based on the sample.

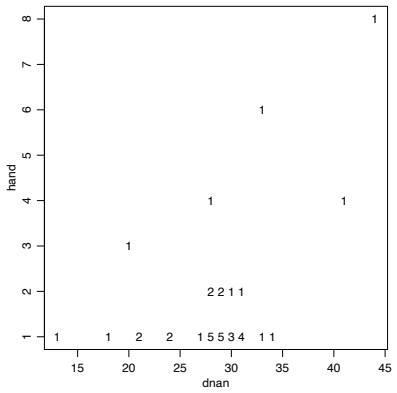
Example: Handedness data

Table 1: Data from a study of handedness; `hand` is an integer measure of handedness, and `dnan` a genetic measure. Data due to Dr Gordon Claridge, University of Oxford.

	dnan	hand		dnan	hand		dnan	hand		dnan	hand
1	13	1	11	28	1	21	29	2	31	31	1
2	18	1	12	28	2	22	29	1	32	31	2
3	20	3	13	28	1	23	29	1	33	33	6
4	21	1	14	28	4	24	30	1	34	33	1
5	21	1	15	28	1	25	30	1	35	34	1
6	24	1	16	28	1	26	30	2	36	41	4
7	24	1	17	29	1	27	30	1	37	44	8
8	27	1	18	29	1	28	31	1			
9	28	1	19	29	1	29	31	1			
10	28	2	20	29	2	30	31	1			

Example: Handedness data

Scatter plot of handedness data. The numbers show the multiplicities of the observations.



Example: Handedness data

- How do we quantify dependence between `dnan` and `hand` for these $n = 37$ individuals?
- A standard measure is the **product-moment (Pearson) correlation** for $G(u, v)$, i.e.,

$$\theta = t(G) = \frac{\int \{u - \int u \, \mathrm{d}G(u,v)\} \{v - \int v \, \mathrm{d}G(u,v)\} \, \mathrm{d}G(u,v)}{\left[\int \{u - \int u \, \mathrm{d}G(u,v)\}^2 \, \mathrm{d}G(u,v) \int \{v - \int v \, \mathrm{d}G(u,v)\}^2 \, \mathrm{d}G(u,v)\right]^{1/2}}.$$

- With $(u, v) = (\texttt{dnan}, \texttt{hand})$, the sample version is

$$\begin{aligned} \widehat{\theta} = t(\widehat{G}) &= \frac{\sum_{j=1}^n (\texttt{dnan}_j - \overline{\texttt{dnan}})(\texttt{hand}_j - \overline{\texttt{hand}})}{\left\{\sum_{j=1}^n (\texttt{dnan}_j - \overline{\texttt{dnan}})^2 \sum_{j=1}^n (\texttt{hand}_j - \overline{\texttt{hand}})^2\right\}^{1/2}} \\ &= 0.509. \end{aligned}$$

- Standard (bivariate normal) 95% confidence interval is $(0.221, \ 0.715)$, but this is obviously inappropriate (the data look highly non-normal).
- Try simulation approach ...

Bootstrap simulation

- Whether $\widehat{G}$ is parametric or non-parametric, we simulate as follows:
  - For $r = 1, \ldots, R$:
    - ▷ generate a bootstrap sample $y_1^*, \ldots, y_n^* \overset{\text{iid}}{\sim} \widehat{G}$,
    - ▷ compute $\widehat{\theta}_r^*$ using $y_1^*, \ldots, y_n^*$,

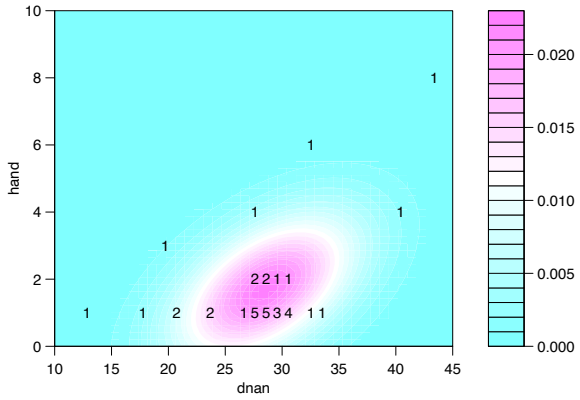
    so the output is a set of **bootstrap replicates**,

$$\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*.$$

- We then use $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ to estimate properties of $\widehat{\theta}$ (histogram, ...).
- If $R \to \infty$, then get perfect match to theoretical calculation based on $\widehat{G}$ (if this is available): Monte Carlo error disappears completely.
- In practice R is finite, so some Monte Carlo error remains.
- If $\widehat{G}$ is the EDF, then $y_1^*, \ldots, y_n^* \overset{\text{iid}}{\sim} \widehat{G}$ are sampled with replacement and equal probabilities from $y_1, \ldots, y_n$, so if $f_i^* = \#\{y_j^* = y_i\}$, then $(f_1^*, \ldots, f_n^*)$ has the multinomial distribution with denominator n and probability vector $(n^{-1}, \ldots, n^{-1})$.
- Although $\mathrm{E}^*(f_j^*) = 1$, y_j can appear $0, 1, \ldots, n$ times in the bootstrap sample.

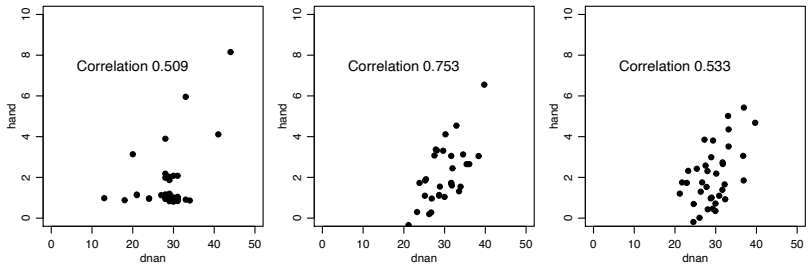
Handedness data: Fitted bivariate normal model

Contours of bivariate normal distribution fitted to handedness data; parameter estimates are $\widehat{\mu}_1 = 28.5$, $\widehat{\mu}_2 = 1.7$, $\widehat{\sigma}_1 = 5.4$, $\widehat{\sigma}_2 = 1.5$, $\widehat{\rho} = 0.509$. The data are also shown.



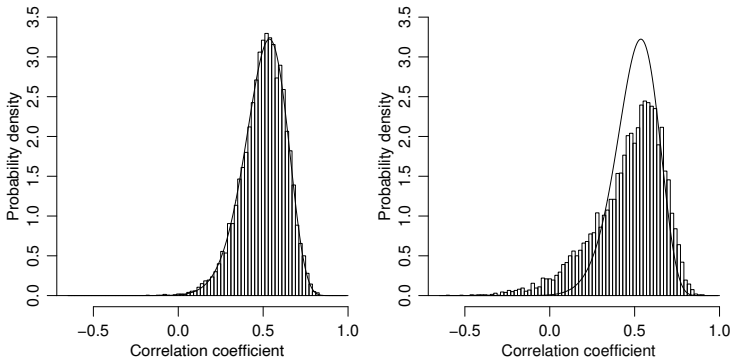
Handedness data: Parametric bootstrap samples

Left: original data, with jittered vertical values. Centre and right: two samples generated from the fitted bivariate normal distribution.



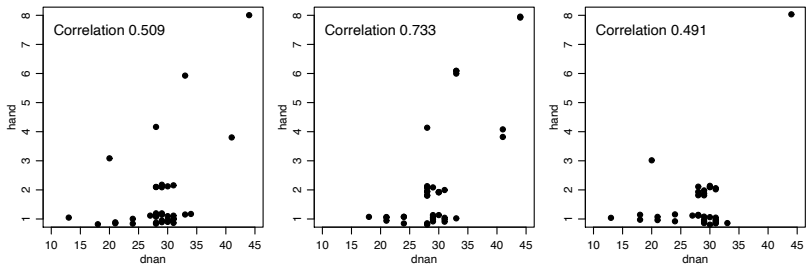
Handedness data: Correlation coefficient

Bootstrap distributions with $R = 10000$. Left: simulation from fitted bivariate normal distribution. Right: nonparametric sampling from the EDF. The lines show the theoretical probability density function of the correlation coefficient under sampling from a fitted bivariate normal distribution.



Handedness data: Bootstrap samples

Left: original data, with jittered vertical values. Centre and right: two bootstrap samples, with jittered vertical values.



Using the $\widehat{\theta}^*$

- The **bias** and **variance** of $\widehat{\theta}$ as an estimator of $\theta = t(G)$,

$$\beta(G) = \mathrm{E}(\widehat{\theta} \mid y_1, \ldots, y_n \overset{\text{iid}}{\sim} G) - t(G), \quad \nu(G) = \mathrm{var}(\widehat{\theta} \mid G),$$

  are estimated by replacing the unknown G by its known estimate $\widehat{G}$:

$$\beta(\widehat{G}) = \mathrm{E}(\widehat{\theta} \mid y_1, \ldots, y_n \overset{\text{iid}}{\sim} \widehat{G}) - t(\widehat{G}), \quad \nu(\widehat{G}) = \mathrm{var}(\widehat{\theta} \mid y_1, \ldots, y_n \overset{\text{iid}}{\sim} \widehat{G}).$$

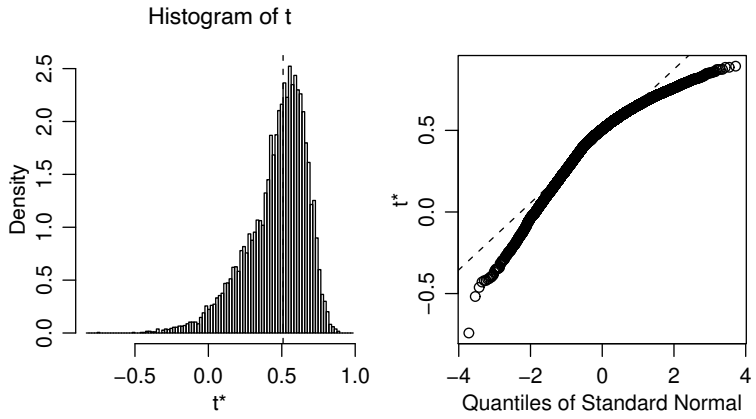
- The Monte Carlo approximations to $\beta(\widehat{G})$ and $\nu(\widehat{G})$ are

$$b = \overline{\widehat{\theta}^*} - \widehat{\theta} = R^{-1} \sum_{r=1}^{R} \widehat{\theta}^*_r - \widehat{\theta}, \quad v = \frac{1}{R-1} \sum_{r=1}^{R} \left(\widehat{\theta}^*_r - \overline{\widehat{\theta}^*} \right)^2.$$

  For the handedness data, $R = 10^4$ and $b = -0.046$, $v = 0.043 = 0.205^2$.
- We estimate the p **quantile** of $\widehat{\theta}$ using the p quantile of $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$, i.e., $\widehat{\theta}^*_{((R+1)p)}$.

Handedness data

Summaries of the $\widehat{\theta}^*$. Left: histogram, with vertical line showing $\widehat{\theta}$. Right: normal Q–Q plot of $\widehat{\theta}^*$.



Common questions

- **How big should n be?** — depends on the context
- **What if the sample is unrepresentative?** — this is always a potential problem in statistics, not specific to resampling methods.
- **How big should R be?** — at least 1000 for most purposes
- **Why take resamples of size n?**
  - We usually want to mimic the sampling properties of samples like the original one, so take resamples of size n,
  - but sometimes we take resamples of size $m \ll n$ in order to achieve validity of the bootstrap—e.g., for extreme quantiles.
- **Why resample from the EDF?**
  - The EDF is the nonparametric MLE of G, so is a natural choice, but
  - sometimes (e.g., testing) we resample from a constrained version of $\widehat{G}$,
  - sometimes it may be useful to smooth $\widehat{G}$;
  - sometimes it may be useful to simulate from (several) parametric fits.

How big should n be?

- For the **average** $\widehat{\theta} = \overline{y}$, the number of distinct samples is

$$m_n = \binom{2n-1}{n},$$

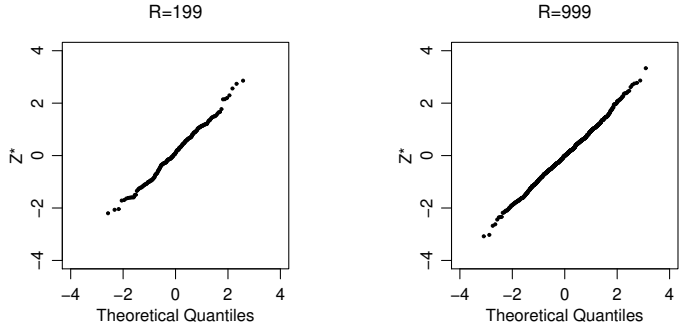
  the most probable of which has probability $p_n = n!/n^n$.
  For $n > 12$, we have $m_n > 10^6$ and $p_n < 6 \times 10^{-5}$.
- Bootstrapping of smooth statistics like the average will often work OK provided $n > 20$.
- For the **median** of a sample of size $n = 2m + 1$, the possible distinct values of $\widehat{\theta}^*$ are $y_{(1)} < \cdots < y_{(n)}$, and

$$\mathrm{P}^*(\widehat{\theta}^* > y_{(l)}) = \sum_{r=0}^{m} \binom{n}{r} \left(\frac{l}{n}\right)^r \left(1 - \frac{l}{n}\right)^{n-r},$$

  so exact calculations of the variance etc. are possible.
- However the median is very vulnerable to bad sample values, so for the median (and other 'non-smooth' statistics) much larger n is needed for reliable inference.

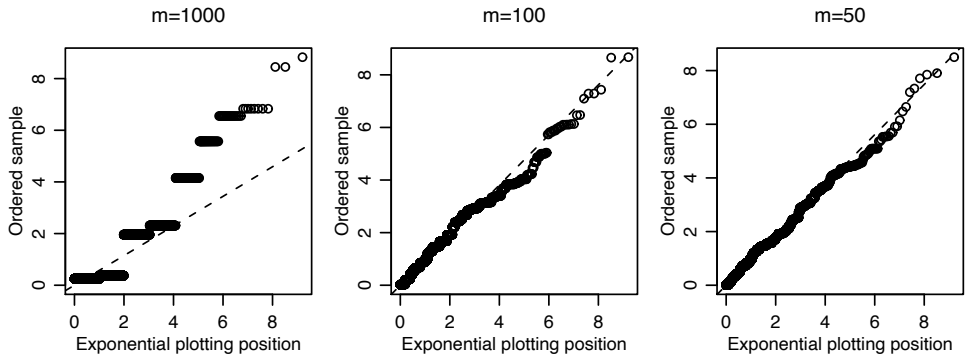
How many bootstraps?

- ☐ Must estimate moments and quantiles of $\widehat{\theta}$ and derived quantities. Often feasible to take $R \gg 1000$
- ☐ Need $R \geq 200$ to estimate bias, variance, etc.
- ☐ Need $R \gg 100$, preferably $R \geq 2500$ to estimate quantiles needed for 95% confidence intervals



Resamples of size n?

- ☐ Exponential sample of size $n = 1000$
- ☐ Distribution of $n \min(Y_1, \ldots, Y_n)$ is $\exp(1)$
- ☐ Resampling distribution $m \min(Y_1^*, \ldots, Y_m^*)$ using resamples of size $m = 1000$, 100, 50
- ☐ To avoid discreteness must choose $m \ll n$, but how?



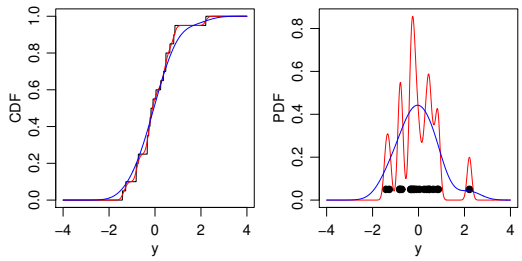
Variants of $\widehat{G}$?

- Can be useful to simulate from a smoothed EDF, given by

$$Y^* = y_{j^*} + h\varepsilon^*, \quad \varepsilon^* \sim \mathcal{N}(0,1) \perp\!\!\!\perp j^* \sim U\{1,\ldots,n\},$$

  equivalent to simulating from a kernel density estimate. Below, with $h = 0.1$ (red) and $h = 0.5$ (blue).
- Since $\mathrm{var}^*(Y^*) = \widehat{\sigma}^2 + h^2$, may prefer a shrunk smoothed estimate, given by

$$Y^* = \overline{y} + \frac{(y_{j^*} - \overline{y}) + h\varepsilon^*}{(1 + h^2/\widehat{\sigma}^2)^{1/2}}.$$



When does the bootstrap work?

- 'Work' might mean the bootstrap gives
  - **reliable** answers when used in practice, or
  - **mathematically correct** answers under 'suitable' regularity conditions.
- For the second of these, suppose we seek to estimate properties of a standardized quantity $Q = q(Y_1,\ldots,Y_n;G)$, maybe $Q = n^{1/2}(\overline{Y} - \theta)$. Let $n \to \infty$ to get limiting results for the distribution function

$$H_{G,n}(q) = \mathrm{P}_G\{Q(Y_1,\ldots,Y_n;G) \le q\},$$

  where subscript G indicates that $Y_1,\ldots,Y_n$ is a random sample from G.
- Bootstrap estimate of this is

$$H_{\widehat{G},n}(q) = \mathrm{P}_{\widehat{G}}\left\{Q(Y_1^*,\ldots,Y_n^*;\widehat{G}) \le q\right\}$$

  where $Q(Y_1^*,\ldots,Y_n^*;\widehat{G}) = n^{1/2}(\overline{Y}^* - \overline{y})$.
- We need conditions under which $H_{\widehat{G},n} \xrightarrow{D} H_{G,n}$ as $n \to \infty$.

Regularity conditions

- ☐ The true distribution G is surrounded by a neighbourhood $\mathcal{N}$ in a suitable space of distributions, and as $n \to \infty$, $\widehat{G}$ eventually falls into $\mathcal{N}$ with probability one. Also:
  1. for any $F \in \mathcal{N}$, $H_{F,n}$ converges weakly to a limit $H_{F,\infty}$;
  2. this convergence must be uniform on $\mathcal{N}$; and
  3. the function mapping F to $H_{F,\infty}$ must be continuous.
- ☐ Weak convergence of $H_{F,n}$ to $H_{F,\infty}$ means that for all integrable $b(\cdot)$,

$$\int b(u)\,\mathrm{d}H_{F,n}(u) \quad \to \quad \int b(u)\,\mathrm{d}H_{F,\infty}(u), \qquad n \to \infty.$$

- ☐ Under these conditions the bootstrap is **consistent**: for any q and $\varepsilon > 0$,

$$\mathrm{P}\{|H_{\widehat{G},n}(q) - H_{G,\infty}(q)| > \varepsilon\} \to 0, \quad n \to \infty.$$

- ☐ The first condition ensures that there is a limit for $H_{G,n}$ to converge to.
- ☐ As n increases, $\widehat{G}$ changes, so the second and third conditions are needed to ensure that $H_{\widehat{G},n}$ approaches $H_{G,\infty}$ along every possible sequence of $\widehat{G}$s.
- ☐ If any one of these conditions fails, the bootstrap can fail. For the minimum (for example) the convergence is not uniform on suitable neighbourhoods of G.

Summary

- ☐ **Estimator is algorithm**:
  - applied to original data $y_1, \ldots, y_n$ gives original $\widehat{\theta}$;
  - applied to simulated data $y_1^*, \ldots, y_n^*$ gives $\widehat{\theta}^*$;
  - $\widehat{\theta}$ can be of (almost) any complexity; but
  - for more sophisticated ideas to work, $\widehat{\theta}$ must often be smooth function of data.
- ☐ **Sample is used to estimate G**:
  - $\widehat{G} \approx G$ — heroic assumption
- ☐ **Simulation replaces theoretical calculation:**
  - removes need for mathematical skill;
  - does not remove need for thought; and in particular,
  - check code **very** carefully — garbage in, garbage out!
- ☐ **Two sources of error**:
  - statistical ($\widehat{G} \neq G$) — reduce by thought; and
  - simulation ($R \neq \infty$) — reduce by taking R large (enough).

Bootstrap confidence Intervals: Desiderata

- A $(1-\alpha)$ **upper confidence limit** for a scalar parameter θ based on data Y is a random variable $\theta_\alpha = \theta_\alpha(Y)$ for which

$$\mathrm{P}\left(\theta \leq \theta_\alpha\right) = \alpha, \quad 0 < \alpha < 1, \theta \in \Theta. \tag{7}$$

- We may seek invariance to monotone transformations $\psi = \psi(\theta)$, that is

$$\mathrm{P}\left\{\psi(\theta) \leq \psi_\alpha\right\} = \alpha, \quad 0 < \alpha < 1, \theta \in \Theta.$$

- In practice exact intervals are rarely available, and we seek intervals such that (7) is satisfied as closely as possible. If $Y \equiv Y_1, \ldots, Y_n$, then we typically have

$$\mathrm{P}\left(\theta \leq \theta_\alpha\right) = \alpha + \mathcal{O}(n^{-1/2}), \quad 0 < \alpha < 1, \theta \in \Theta,$$

and the corresponding two-sided interval satisfies

$$\mathrm{P}\left(\theta_\alpha < \theta \leq \theta_{1-\alpha}\right) = (1-2\alpha) + \mathcal{O}(n^{-1}), \quad 0 < \alpha < 1/2, \theta \in \Theta.$$

Normal confidence intervals

- If $\widehat{\theta} \dot{\sim} \mathcal{N}(\theta + \beta, \nu)$ with known bias $\beta = \beta(G)$ and variance $\nu = \nu(G)$, then a $(1-2\alpha)$ confidence interval is based on the equation

$$\mathrm{P}\left(z_\alpha < \frac{\widehat{\theta} - \theta - \beta}{\nu^{1/2}} \leq z_{1-\alpha}\right) = 1 - 2\alpha,$$

and has limits $\widehat{\theta} - \beta \pm z_\alpha \nu^{1/2}$, where $\Phi(z_\alpha) = \alpha$.

- We replace β, ν by the bootstrap estimates

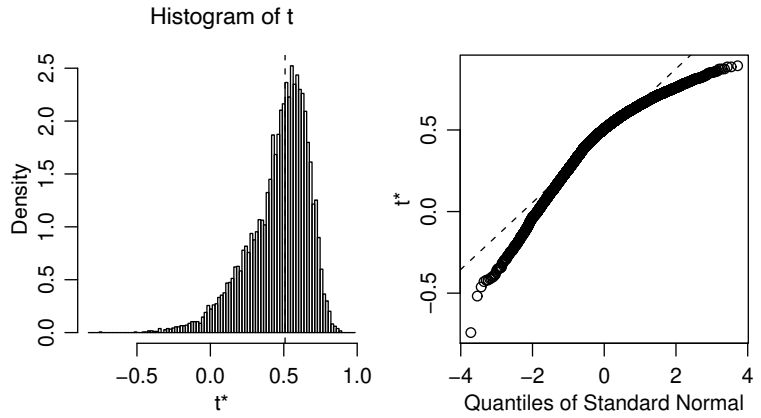
$$\begin{aligned} \beta(G) &\doteq \beta(\widehat{G}) \doteq b = \overline{\widehat{\theta}^*} - \widehat{\theta}, \\ \nu(G) &\doteq \nu(\widehat{G}) \doteq v = (R-1)^{-1} \sum_r (\widehat{\theta}_r^* - \overline{\widehat{\theta}^*})^2, \end{aligned}$$

to get the $(1-2\alpha)$ interval with limits $\widehat{\theta} - b \pm z_\alpha v^{1/2}$.

- For the handedness data we have $R = 10,000$, $b = -0.046$, $v = 0.205^2$, $\alpha = 0.025$, $z_\alpha = -1.96$, so 95% CI is $(0.147, 0.963)$
- We can use the $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ to check the quality of the normal approximation, and perhaps to suggest transformations.

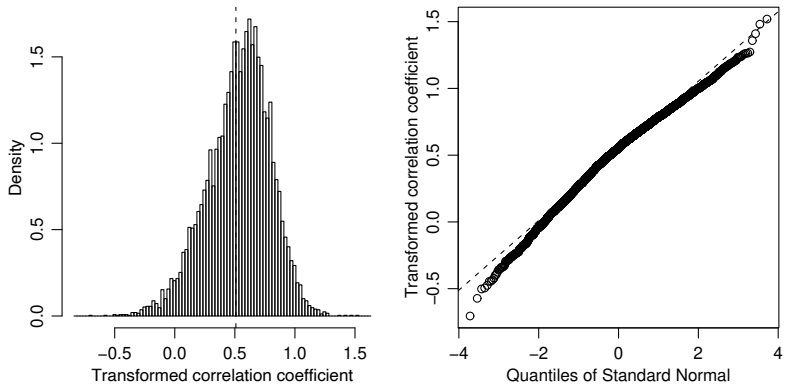
Handedness data

Summaries of the $\widehat{\theta}^*$. Left: histogram, with vertical line showing $\widehat{\theta}$. Right: normal Q–Q plot of $\widehat{\theta}^*$.



Handedness data: Transformed scale?

Plots for $\widehat{\psi}^* = \frac{1}{2}\log\{(1+\widehat{\theta}^*)/(1-\widehat{\theta}^*)\}$:



Normal confidence intervals

- Correlation coefficient: try Fisher's z transformation:

$$\widehat{\psi}^* = \psi(\widehat{\theta}^*) = \tfrac{1}{2}\log\{(1+\widehat{\theta}^*)/(1-\widehat{\theta}^*)\}$$

with bias and variance estimates

$$b_\psi = R^{-1}\sum_{r=1}^{R}\widehat{\psi}_r^* - \widehat{\psi}, \quad v_\psi = \frac{1}{R-1}\sum_{r=1}^{R}\left(\widehat{\psi}_r^* - \overline{\widehat{\psi}^*}\right)^2,$$

- Then the $(1-2\alpha)$ confidence interval for θ is

$$\psi^{-1}\left\{\widehat{\psi} - b_\psi - z_{1-\alpha}v_\psi^{1/2}\right\}, \quad \psi^{-1}\left\{\widehat{\psi} - b_\psi - z_\alpha v_\psi^{1/2}\right\}$$

- For handedness data, get $(0.074, 0.804)$ ... but how do we choose a transformation in general?

Pivots

- Assume properties of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ mimic effect of sampling from original model (plug-in principle) — false in general, but more nearly true for pivots.
- **Pivot** is combination of data and parameter whose distribution is independent of underlying model, such as t statistic

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

when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

- Exact pivot generally unavailable in nonparametric case, but if we can estimate the variance of $\widehat{\theta}^*$ using V, we use

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

- If the quantiles z_α of Z known, then

$$\mathrm{P}\left(z_\alpha \leq Z \leq z_{1-\alpha}\right) = \mathrm{P}\left(z_\alpha \leq \frac{\widehat{\theta} - \theta}{V^{1/2}} \leq z_{1-\alpha}\right) = 1 - 2\alpha$$

(z_α no longer denotes a normal quantile!) gives $(1-2\alpha)$ CI $(\widehat{\theta} - V^{1/2}z_{1-\alpha}, \widehat{\theta} - V^{1/2}z_\alpha)$

Studentized statistic

- Bootstrap sample gives $(\widehat{\theta}^*, V^*)$ and hence

$$Z^* = \frac{\widehat{\theta}^* - \widehat{\theta}}{V^{*1/2}}.$$

- We bootstrap to get R copies of $(\widehat{\theta}, V)$, i.e.,

$$(\widehat{\theta}^*_1, V^*_1), \quad (\widehat{\theta}^*_2, V^*_2), \quad \ldots, \quad (\widehat{\theta}^*_R, V^*_R),$$

  and the corresponding

$$z^*_1 = \frac{\widehat{\theta}^*_1 - \widehat{\theta}}{V^{*1/2}_1}, \quad z^*_2 = \frac{\widehat{\theta}^*_2 - \widehat{\theta}}{V^{*1/2}_2}, \quad \ldots, \quad z^*_R = \frac{\widehat{\theta}^*_R - \widehat{\theta}}{V^{*1/2}_R},$$

  then order these to estimate quantiles of Z, with z_p estimated by $z^*_{(p(R+1))}$.
- Get $(1-2\alpha)$ **Studentized bootstrap confidence interval**

$$\widehat{\theta} - V^{1/2} z^*_{((1-\alpha)(R+1))}, \quad \widehat{\theta} - V^{1/2} z^*_{(\alpha(R+1))}.$$

- This is not invariant to transformation and needs an estimated variance V^*_r for each $\widehat{\theta}^*_r$.

Why Studentize?

- If we Studentize, then $Z \xrightarrow{D} N(0,1)$ as $n \to \infty$, and we can use Edgeworth series to write

$$\mathrm{P}_G(Z \le z) = \Phi(z) + n^{-1/2} a(z)\phi(z) + O(n^{-1}),$$

  where $a(\cdot)$ is an even quadratic polynomial.
- For example, if we use $\widehat{\theta} = \overline{Y}$ and $V = n^{-1}S^2$ to compute Z for data with skewness γ, then $a(x) = \gamma(2x^2+1)/6$ and (next slide) $a'(x) = -\gamma(x^2-1)/6$.
- The corresponding expansion for Z^* is

$$\mathrm{P}_{\widehat{G}}(Z^* \le z) = \Phi(z) + n^{-1/2} \widehat{a}(z)\phi(z) + O_p(n^{-1}).$$

- Typically $\widehat{a}(z) = a(z) + O_p(n^{-1/2})$, so

$$\mathrm{P}_{\widehat{G}}(Z^* \le z) - \mathrm{P}_G(Z \le z) = O_p(n^{-1}),$$

  so the order of error is n^{-1}.

Why Studentize? II

- Without Studentization, $Z = n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} N(0, \nu')$, and then

$$\mathrm{P}_G(Z \leq z) = \Phi\left(\frac{z}{\nu'^{1/2}}\right) + n^{-1/2} a'\left(\frac{z}{\nu'^{1/2}}\right)\phi\left(\frac{z}{\nu'^{1/2}}\right) + O(n^{-1})$$

  and

$$\mathrm{P}_{\widehat{G}}(Z^* \leq z) = \Phi\left(\frac{z}{\widehat{\nu}'^{1/2}}\right) + n^{-1/2} \widehat{a}'\left(\frac{z}{\widehat{\nu}'^{1/2}}\right)\phi\left(\frac{z}{\widehat{\nu}'^{1/2}}\right) + O_p(n^{-1}).$$

- Typically $\widehat{\nu}' = \nu' + O_p(n^{-1/2})$, giving

$$\mathrm{P}_{\widehat{G}}(Z^* \leq z) - \mathrm{P}_G(Z \leq z) = O_p(n^{-1/2}),$$

  and the difference in the leading terms means that the overall error is of order $n^{-1/2}$.
- Thus Studentizing reduces error from $O_p(n^{-1/2})$ to $O_p(n^{-1})$: better than using large-sample asymptotics, for which error is usually $O_p(n^{-1/2})$.

Other confidence intervals

- Simpler approaches:
  - **Basic bootstrap** interval: treat $\widehat{\theta} - \theta$ as pivot, get

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

  - **Percentile interval**: use empirical quantiles of $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$:

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

- The percentile interval is transformation-invariant, not the basic bootstrap interval.
- **Bias-corrected and accelerated (BC_a)** intervals replace percentile interval with $(\widehat{\theta}^*_{((R+1)\alpha')}, \widehat{\theta}^*_{((R+1)(1-\alpha''))})$, where

$$\alpha' = \Phi\left\{w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right\}, \quad w = \Phi^{-1}\left\{\widehat{G}^*(\widehat{\theta})\right\}, \quad a = \frac{1}{6}\frac{\sum_{j=1}^n l_j^3}{\left(\sum_{j=1}^n l_j^2\right)^{3/2}},$$

  with $\widehat{G}^*$ the EDF of the $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$, and $l_1, \ldots, l_n$ the empirical influence values (soon).
- If the **bias** $w = 0$, then $\widehat{G}^*(\widehat{\theta}) = \frac{1}{2}$, so $\widehat{\theta}$ is at the median of the EDF of $\widehat{\theta}^*$
- If the **acceleration** $a = 0$, then the effect of the data $y_1, \ldots, y_n$ on $\widehat{\theta}$ is symmetric.

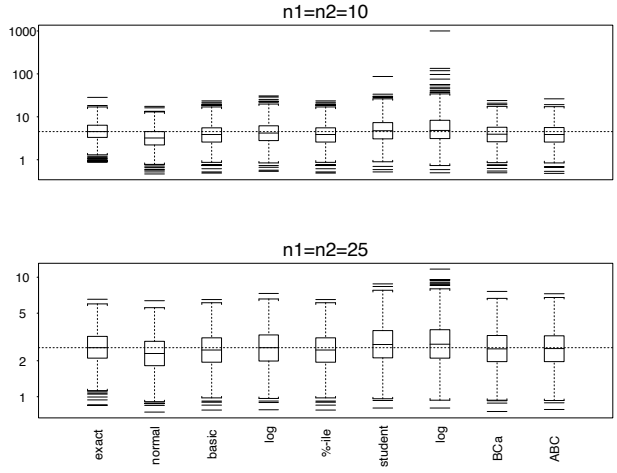
Comparisons

Table 2: Empirical error rates (%) for nonparametric bootstrap confidence limits in ratio estimation: rates for sample sizes $n_1 = n_2 = 10$ are given above those for sample sizes $n_1 = n_2 = 25$. $R = 999$ for all bootstrap methods. 10,000 data sets generated from Gamma distributions.

Method	Nominal error rate							
	Lower limit				Upper limit			
	1	2.5	5	10	10	5	2.5	1
Exact	1.0	2.8	5.5	10.5	9.8	4.8	2.6	1.0
	1.0	2.3	4.8	9.9	10.2	4.9	2.5	1.1
Normal approximation	0.1	0.5	1.7	6.3	20.6	15.7	12.5	9.6
	0.1	0.5	2.1	6.4	16.3	11.5	8.2	5.5
Basic bootstrap	0.0	0.0	0.2	1.8	24.4	21.0	18.6	16.4
	0.0	0.1	0.4	3.0	19.2	15.0	12.5	10.3
Basic bootstrap, log scale	2.6	4.9	8.1	12.9	13.1	7.5	4.8	2.5
	1.6	3.2	6.0	11.4	11.5	6.3	3.3	1.7
Studentized bootstrap	0.6	2.1	4.6	9.9	11.9	6.7	4.0	2.0
	0.8	2.3	4.6	9.9	10.9	5.9	3.0	1.4
Studentized bootstrap, log scale	1.1	2.8	5.6	10.7	11.6	6.3	3.5	1.7
	1.1	2.5	5.0	10.1	10.8	5.7	2.9	1.3
Bootstrap percentile	1.8	3.6	6.5	11.6	14.6	8.9	5.9	3.3
	1.2	2.6	5.1	10.1	12.6	7.1	4.2	2.1
BC_a	1.9	4.0	6.9	12.3	14.0	8.3	5.3	3.0
	1.4	3.0	5.6	10.9	11.8	6.8	3.8	1.9
ABC	1.9	4.2	7.4	12.7	14.6	8.7	5.5	3.1
	1.3	3.0	5.7	11.0	12.1	6.8	3.7	1.9

Confidence interval lengths

Lengths of 95% confidence intervals for the first 1000 simulated samples in the numerical experiment with Gamma data.



Discussion

- ☐ Bootstrap confidence intervals usually under-cover (i.e., are too short).
- ☐ Normal, basic, and studentized intervals depend on scale.
- ☐ Percentile interval often too short but is transformation-invariant.
- ☐ Studentized intervals give best coverage overall, but
  - they depend on scale, can be sensitive to V;
  - their lengths can be very variable;
  - they are best when V is approximately constant.
- ☐ Improved percentile intervals have same asymptotic error as Studentized intervals, but often are shorter, so give lower coverage probabilities.
- ☐ Caution: Edgeworth theory OK for smooth statistics, but beware rough statistics: must check output.
- ☐ Typically need $R > 1000$ for reliable estimation of quantiles.

5.3 Nonparametric Delta Method

Nonparametric delta method

- ☐ The **delta method** (Theorem 11) gives variance formulae for functions of averages.
- ☐ More generally we use the **nonparametric delta method**, which is based on the linear functional expansion

$$t(F) \doteq t(G) + \int L_t(x; G)\, \mathrm{d}F(x),$$

where L_t, the first derivative of $t(\cdot)$ at G, is defined by

$$L_t(y; G) = \lim_{\varepsilon \to 0} \frac{t\{(1-\varepsilon)G + \varepsilon H_y\} - t(G)}{\varepsilon} = \left. \frac{\partial t\{(1-\varepsilon)G + \varepsilon H_y\}}{\partial \varepsilon} \right|_{\varepsilon=0},$$

with $H_y(u) \equiv H(u - y)$ the Heaviside function jumping from 0 to 1 at $u = y$.

- ☐ The **influence function value** $L_t(y; G)$ for the statistical functional t for an observation at y when the background distribution is G, satisfies $\mathrm{E}_G\{L_t(Y; G)\} = 0$.
- ☐ If $\widehat{G}$ is based on a random sample $y_1, \ldots, y_n$, then the jth **empirical influence value** is

$$l_j = L_t(y_j; \widehat{G}),$$

and $\mathrm{E}_{\widehat{G}}\{L_t(Y; \widehat{G})\} = n^{-1} \sum_j l_j = 0$.

- ☐ The influence function also plays an important role in robust statistics.

Nonparametric delta method II

☐ If we replace F by the EDF $\widehat{G}$ for a random sample $Y_1, \ldots, Y_n$, then

$$t(\widehat{G}) \doteq t(G) + \int L_t(x; G)\, \mathrm{d}\widehat{G}(x) = t(G) + \frac{1}{n}\sum_{j=1}^{n} L_t(Y_j; G),$$

has variance

$$\mathrm{var}\{t(\widehat{G})\} \doteq \frac{1}{n^2}\sum_{j=1}^{n} L_t^2(Y_j; G) = V_L,$$

say, which we estimate based on a sample $y_1, \ldots, y_n$ by $v_L = n^{-2}\sum l_j^2$.

Example 71 *Apply the nonparametric delta method to the average $\overline{Y}$.*

Example 72 *Apply the nonparametric delta method to a statistic defined by an estimating equation, and hence find the variance of the ratio $\overline{V}/\overline{U}$ for data pairs $Y = (U, V)$.*

Note to Example 71

☐ The population mean and its empirical version are

$$\theta = t(G) = \int x\, dG(x), \quad \widehat{\theta} = t(\widehat{G}) = \int x\, d\widehat{G}(x) = n^{-1}\sum_{j=1}^{n} Y_j = \overline{Y}.$$

☐ If H_y puts unit mass at y, its 'density' is a Dirac delta function $\delta_y(x)$, and

$$\begin{aligned} \theta\{(1-\varepsilon)G + \varepsilon H_y\} &= \int x\, \mathrm{d}\{(1-\varepsilon)G + \varepsilon H_y\}(x) \\ &= (1-\varepsilon)\int x\, \mathrm{d}G(x) + \varepsilon \int x\, \mathrm{d}H_y(x) = (1-\varepsilon)\theta(G) + \varepsilon y \end{aligned}$$

and therefore

$$L(y; G) = \lim_{\varepsilon \to 0} \frac{\theta\{(1-\varepsilon)G + \varepsilon H_y\} - \theta(G)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{(1-\varepsilon)\theta(G) + \varepsilon y - \theta(G)}{\varepsilon} = y - \theta(G),$$

☐ Hence the empirical influence values and variance estimate are

$$l_j = L(y_j; \widehat{G}) = y_j - \overline{y}, \qquad v_L = \frac{1}{n^2}\sum (y_j - \overline{y})^2 = \frac{n-1}{n} n^{-1} s^2.$$

Note to Example 72

☐ The scalar parameter $\theta = t(G)$ is determined implicitly through the estimating equation

$$\int a(x;\theta)\,\mathrm{d}G(x) = \int a\{x;t(G)\}\,\mathrm{d}G(x) = 0.$$

We replace G by $G_\varepsilon = (1-\varepsilon)G + \varepsilon H_y$ and see that

$$\begin{aligned} 0 &= \int a\{x;t(G_\varepsilon)\}\,\mathrm{d}G_\varepsilon(x) \\ &= (1-\varepsilon)\int a\{x;t(G_\varepsilon)\}\,\mathrm{d}G(x) + \varepsilon\int a\{x;t(G_\varepsilon)\}\,\mathrm{d}H_y(x) \\ &= (1-\varepsilon)\int a\{x;t(G_\varepsilon)\}\,\mathrm{d}G(x) + \varepsilon a\{y;t(G_\varepsilon)\}, \end{aligned}$$

and differentiation using the chain rule gives

$$0 = a\{y;t(G_\varepsilon)\} - \int a\{x;t(G_\varepsilon)\}\,\mathrm{d}G(x) + \varepsilon a_\theta\{y;t(G_\varepsilon)\}\frac{\partial t(G_\varepsilon)}{\partial\varepsilon} + (1-\varepsilon)\int a_\theta\{x;t(G_\varepsilon)\}\frac{\partial t(G_\varepsilon)}{\partial\varepsilon}\,\mathrm{d}G(x),$$

which reduces to

$$0 = a\{y;t(G)\} + \int a_\theta\{x;t(G)\}\,\mathrm{d}G(x)\,\frac{\partial t(G)}{\partial\varepsilon}\bigg|_{\varepsilon=0}$$

on setting $\varepsilon = 0$. Hence

$$L_t(y;G) = \frac{\partial t(G_\varepsilon)}{\partial\varepsilon}\bigg|_{\varepsilon=0} = \frac{a(y;\theta)}{-\int a_\theta(x;\theta)\,\mathrm{d}G(x)}, \quad \text{where} \quad a_\theta(x;\theta) = \frac{\partial a(x;\theta)}{\partial\theta}.$$

☐ In the case of the ratio and with $y = (u,v)$, we take $a(y;\theta) = v - \theta u$, so

$$\theta = \theta(G) = \int v\,\mathrm{d}G(u,v) \Big/ \int u\,\mathrm{d}G(u,v), \quad \widehat{\theta} = \overline{v}/\overline{u},$$

and $a_\theta = -u$, so $l_j = (x_j - \widehat{\theta}u_j)/\overline{u}$, giving

$$v_L = \frac{1}{n^2}\sum\left(\frac{x_j - \widehat{\theta}u_j}{\overline{u}}\right)^2.$$

Comments

- For statistics involving only averages (ratio, correlation coefficient, ...), the nonparametric delta method retrieves the delta method.
- For example, the correlation coefficient may be written as a function of $\overline{xu} = n^{-1} \sum x_j u_j$, etc.:

$$\widehat{\theta} = \frac{\overline{xu} - \overline{x}\,\overline{u}}{\left\{(\overline{x^2} - \overline{x}^2)(\overline{u^2} - \overline{u}^2)\right\}^{1/2}},$$

  from which empirical influence values l_j can be derived, giving $v_L = 0.029$ for the handedness data, to be compared with $v = 0.043$ obtained by bootstrapping.
- v_L typically underestimates $\mathrm{var}(\widehat{\theta})$!
- The l_j can also be obtained by numerical differentiation if $t(\widehat{G})$ is coded appropriately, or approximated using a jackknife method.