Bootstrap

Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch

December 18, 2025

- Parametric models are determined by a finite vector $\theta \in \Theta$.
- 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** (or equivalently the indicator function).

- 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.

Example: non-parametrically estimating the mean with $\widehat{\theta} = \int x d\widehat{G}(x) = \frac{1}{n}\sum X_i$
If G is Gaussian, we know the sampling distribution $\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2/n)$.
Without Gaussianity, there is still a sampling distribution, we just don't know what it is.

Inference about θ is based on the sampling distribution, which is given by the sampling process

- If we control the sampling process, we can approximate the sampling distribution by Monte Carlo
- G unknown but $\widehat{G}$ is known. Then, the (re)sampling distribution can be studied/approximated by Monte Carlo

The Bootstrap Idea: The (re)sampling process from $\widehat{G}$ can mimic the sampling process from G itself

$$\begin{aligned} \text{Sampling (real world):} \quad & G \Longrightarrow X_1, \ldots, X_N \Longrightarrow \widehat{\theta} = \theta(\widehat{G}) \\ \text{Resampling (bootstrap world):} \quad & \widehat{G} \Longrightarrow X_1^\star, \ldots, X_N^\star \Longrightarrow \widehat{\theta}^\star = \theta\left(\widehat{G}^\star\right) \end{aligned}$$

$\rightarrow$ removes need for mathematical skills but still perform well in practice (usually!)

- 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 $y_1^*, \ldots, y_n^*$,

  output 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) *i.e. 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 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.

Example

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?

- 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, $p = 1/2$.

- 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 (nonparametric) 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 $u = 0$,

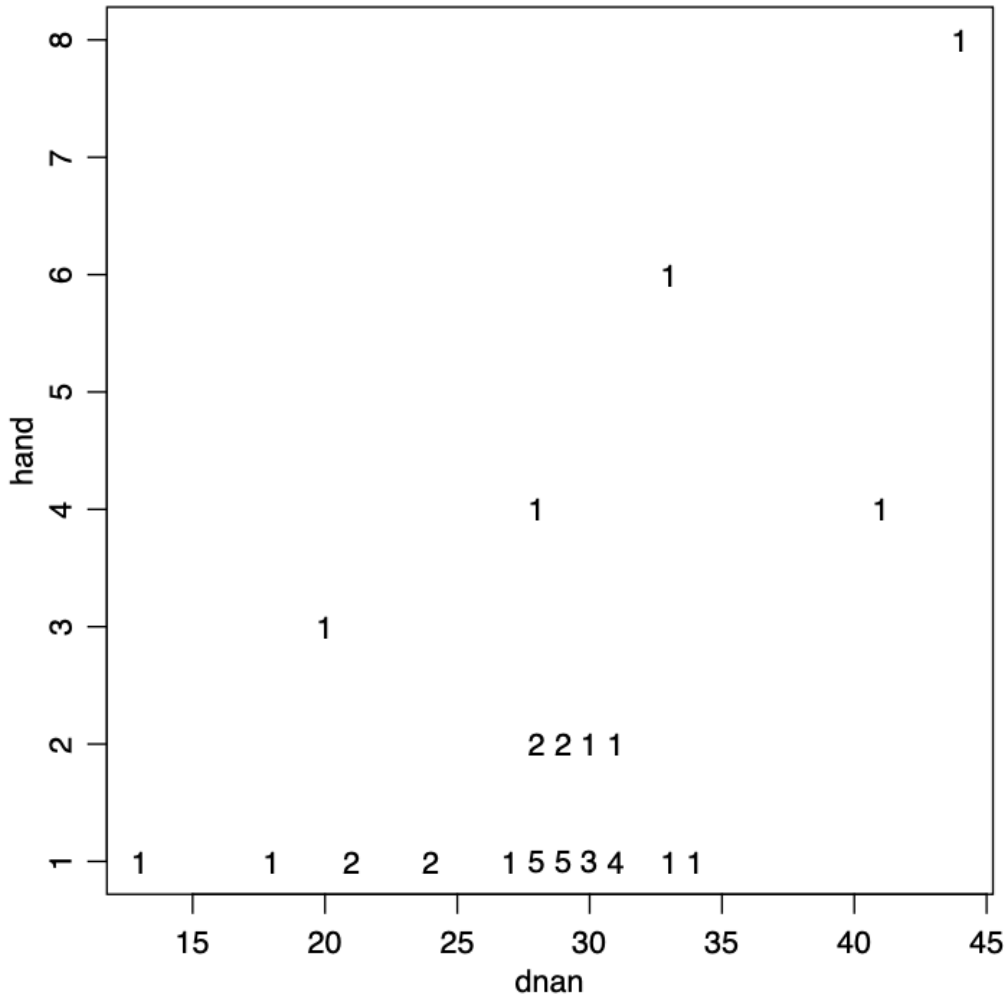
$$\begin{aligned}
t_2(\widehat{G}) &= \operatorname{argmax}_\theta \int \log f(y;\theta)\, \mathrm{d}\widehat{G}(y) \\
&= \operatorname{argmax}_\theta \int \log f(y;\theta)\, \mathrm{d}\left\{ n^{-1} \sum_{j=1}^{n} H(y - Y_j) \right\} \\
&= \operatorname{argmax}_\theta n^{-1} \sum_{j=1}^{n} \int \log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

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

	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			

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

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



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

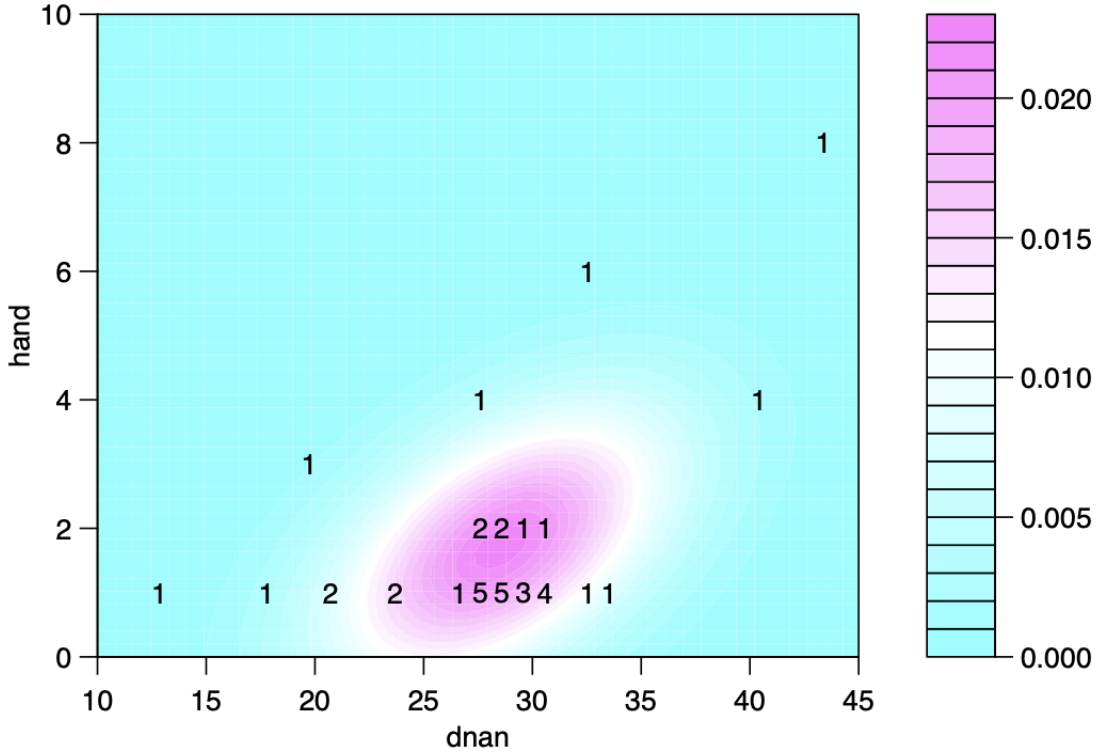
$$\begin{aligned} \widehat{\theta} = t(\widehat{G}) &= \frac{\sum_{j=1}^n (\mathtt{dnan}_j - \overline{\mathtt{dnan}})(\mathtt{hand}_j - \overline{\mathtt{hand}})}{\left\{\sum_{j=1}^n (\mathtt{dnan}_j - \overline{\mathtt{dnan}})^2 \sum_{j=1}^n (\mathtt{hand}_j - \overline{\mathtt{hand}})^2\right\}^{1/2}}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

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

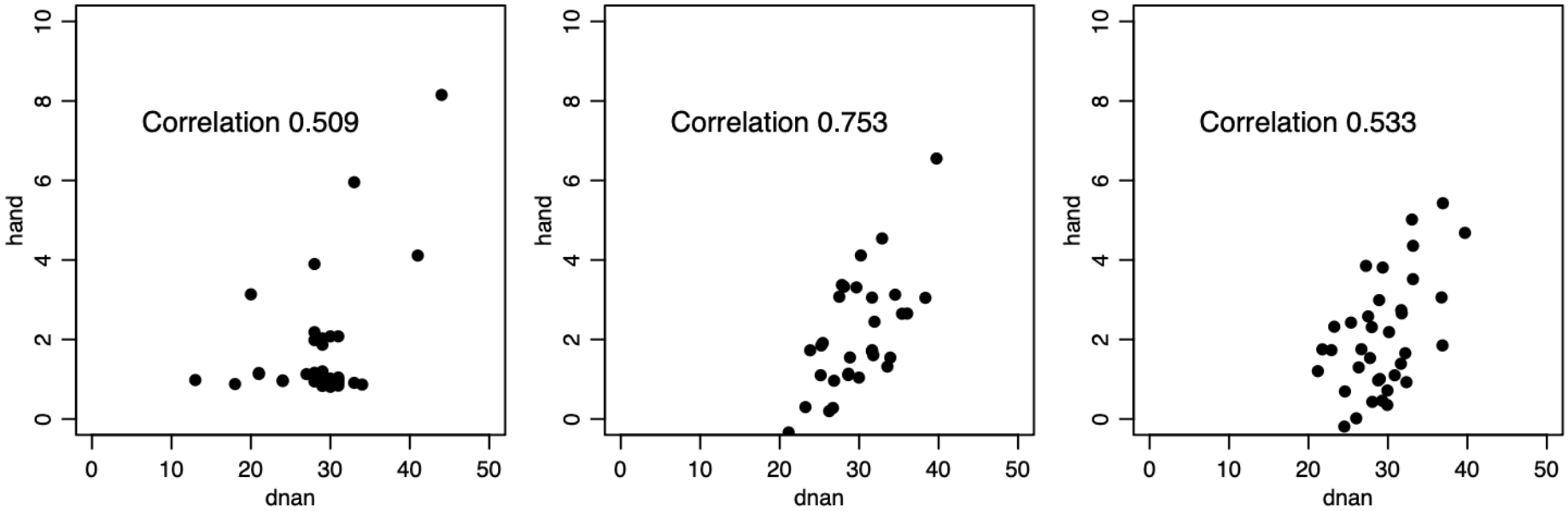
Try simulation approach ...

- Data $y_1, \ldots, y_n \overset{\text{iid}}{\sim} G$ are used to estimate the unknown parameter $\theta = t(G)$
- The estimator is $\widehat{\theta} = t(\widehat{G})$, where $\widehat{G}$ estimates G.
- For the handedness data,
  - $y = (u, v) = (\texttt{dnan}, \texttt{hand})$,
  - F puts probability mass on a subset of $\mathbb{R}^2$, and
  - $\widehat{\theta}$ is the sample correlation coefficient.
- Key questions
  - How does $\widehat{\theta}$ behave when samples are repeatedly taken from F?
  - How can we use knowledge of this to learn about θ?
- If G was known, we could answer these questions by
  - analytical calculation, or
  - simulation from G.
- This motivates the idea of simulating from a (parametric or nonparametric) estimate $\widehat{G}$ of G.

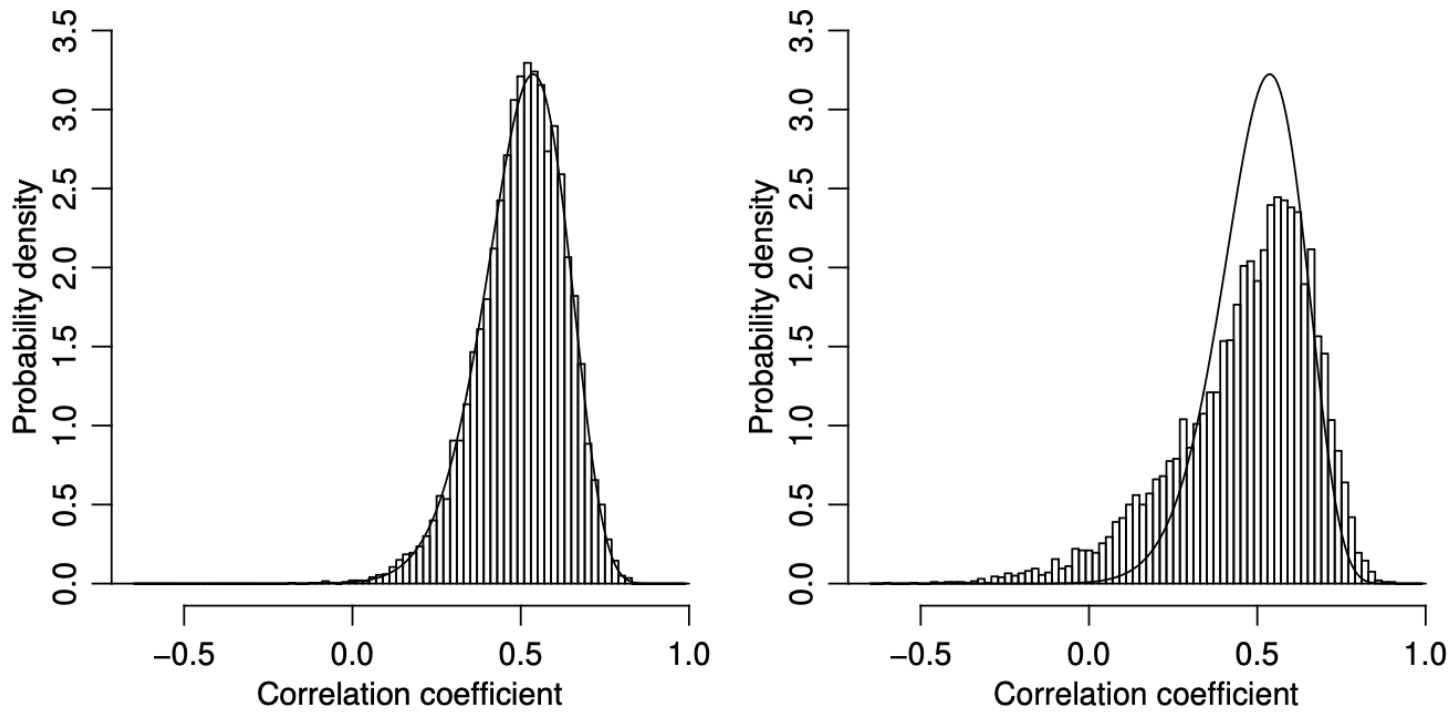
Contours of bivariate normal distribution fitted to handedness data; parameter estimates are $\widehat{\mu}_1 = 28.5$, $\widehat{\mu}_2 = 1.7$, $\widehat{\sigma}_1 = 5.4$, $\widehat{\sigma}_2 = 1.5$, $\widehat{\rho} = 0.509$. The data are also shown.



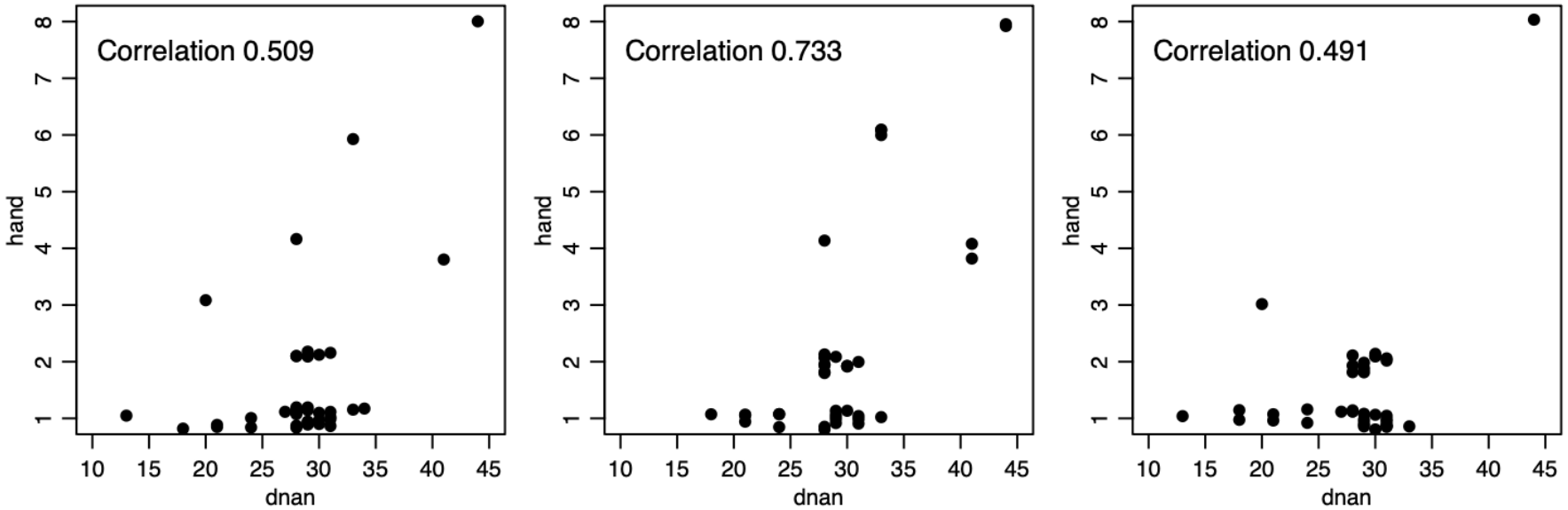
Left: original data, with jittered vertical values. Centre and right: two samples generated from the fitted bivariate normal distribution.



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



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



- 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 $\widehat{G}$:

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

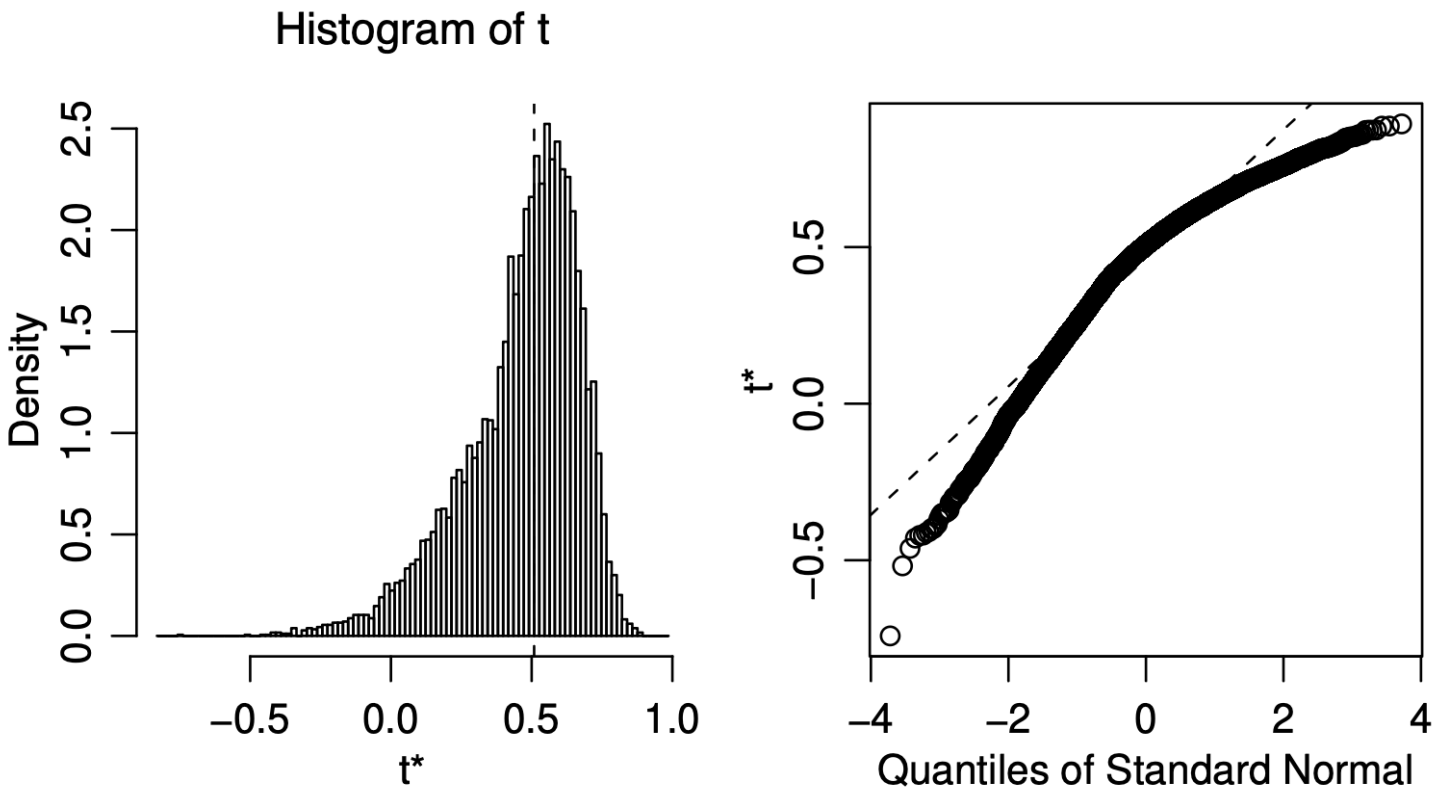
- The Monte Carlo approximations to $\beta(\widehat{G})$ and $\nu(\widehat{G})$ are

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

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

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



- **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.

- 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

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

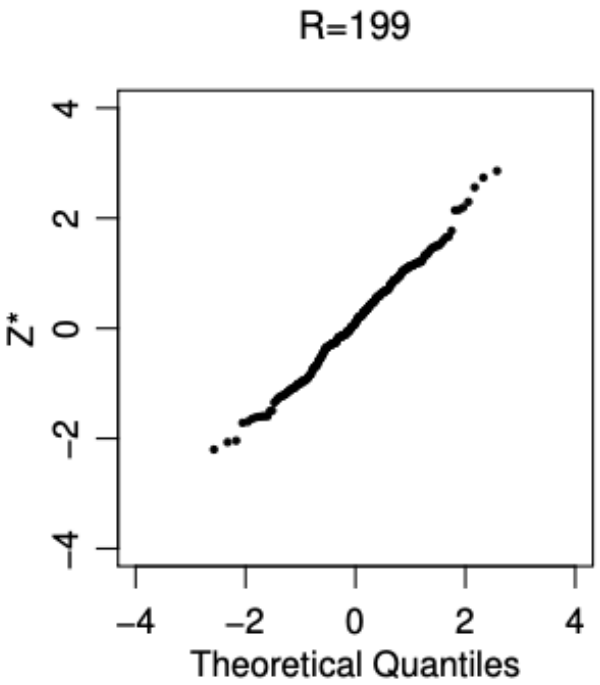
so exact calculations of the variance etc. are possible.

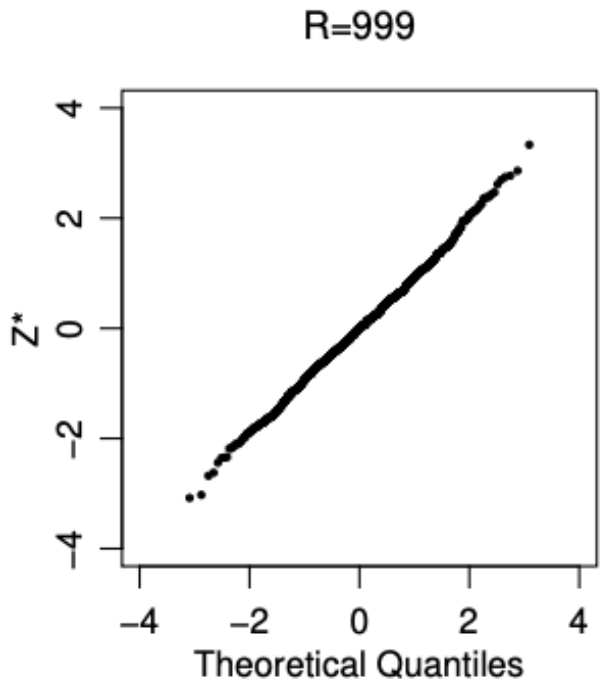
- However the median is very vulnerable to bad sample values, so for the median (and other 'non-smooth' statistics) much larger n is needed for reliable inference.

What about outliers?

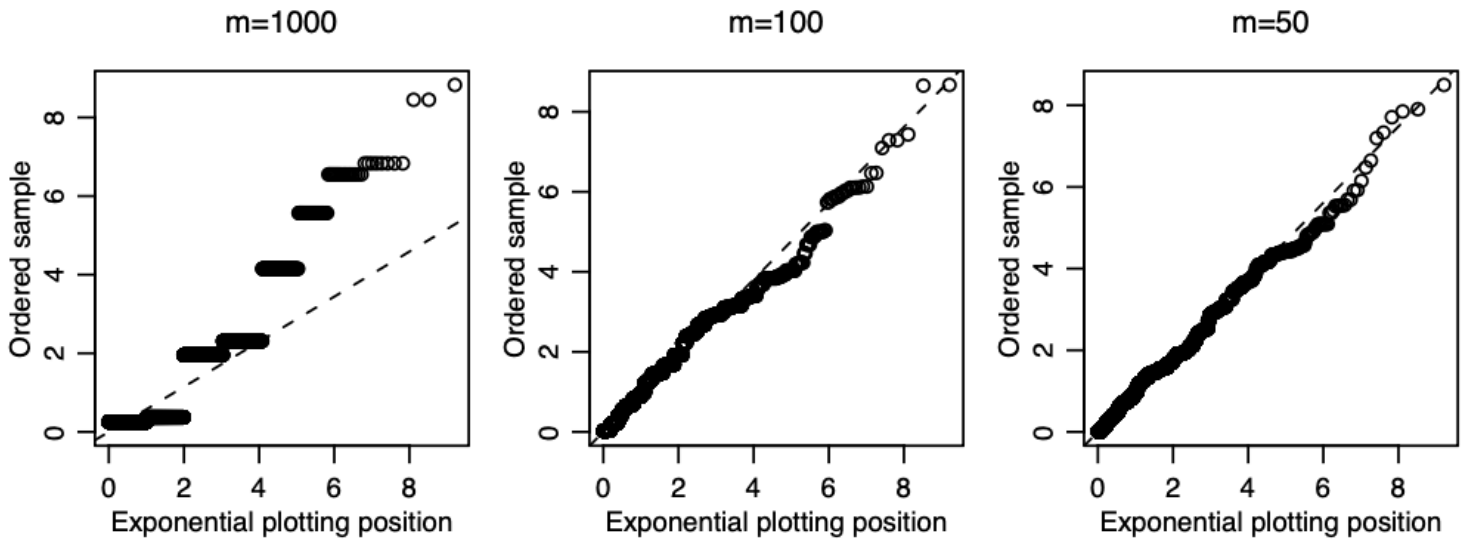
- Bad sample values not specific to bootstrap methods—need to consider them in any analysis
- Can reduce their effects by using robust estimators, but beware, as outliers may affect bootstrapped estimators even if they don't affect original value

- 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





- 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?



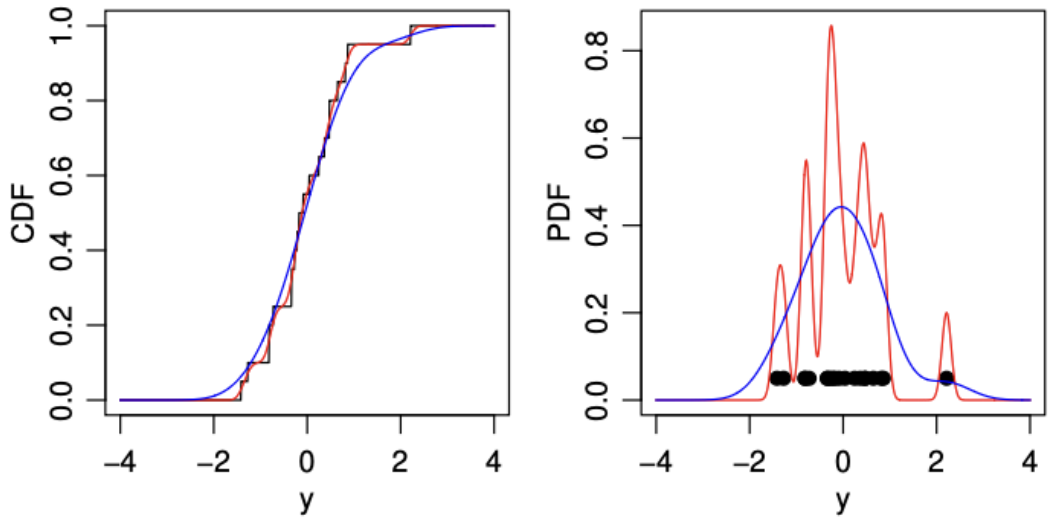
- Can be useful to simulate from a smoothed EDF, given by

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

  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

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



- '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
$$H_{G,n}(q) = \Pr_G \{Q(Y_1, \ldots, Y_n; G) \leq q\},$$
where subscript G indicates that $Y_1, \ldots, Y_n$ is a random sample from G.
- Bootstrap estimate of this is
$$H_{\widehat{G},n}(q) = \Pr_{\widehat{G}} \left\{ Q(Y_1^*, \ldots, Y_n^*; \widehat{G}) \leq q \right\}$$
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 $n \to \infty$.

- 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 $b(\cdot)$,

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

- 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 example, for the minimum the convergence is not uniform on suitable neighbourhoods of G.

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

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

Theorem (Consistency of Bootstrap)

Consider the Bootstrap estimator

$$\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu_n}) \leq t | X_1, \ldots, X_n)$$

for

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t).$$

If $\mathrm{E}[|X_i|^3] < \infty$, *then*

$$\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P(n^{-1/2}).$$

Theorem (Berry-Esseen Theorem)

Let X_i be i.i.d. with mean μ and variance σ^2. Let $\mathrm{E}[|X_i - \mu|^3] < \infty$, and

$$Z_n = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}.$$

Then,

$$\sup_z |\mathbb{P}(Z_n \leq z) - \Phi(z)| \leq \frac{33}{4}\frac{\mathrm{E}[|X_i - \mu|^3]}{\sqrt{n}}$$

Proof.

Let $\Phi_\sigma(t)$ be the CDF of a Normal with mean 0 and variance σ^2. Let $\widehat{\sigma}^2 = \frac{1}{n}\sum(X_i - \widehat{\mu})^2$. Then, $\widehat{\sigma}^2 = \text{Var}(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu})|X_1, \ldots, X_n)$. By the triangle inequality,

$$\sup_t |\widehat{F}_n(t) - F_n(t)| \leq \underbrace{\sup_t |F_n(t) - \Phi_\sigma(t)|}_{I} + \underbrace{\sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)|}_{II} + \underbrace{\sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)|}_{III}.$$

Then, by the Berry-Esseen theorem,

$$I = \sup_t |F_n(t) - \Phi_\sigma(t)| = \sup_t |\mathbb{P}(\sqrt{n}(\widehat{\mu} - \mu) \leq t) - \mathbb{P}(\sigma Z \leq t)|$$

$$= \sup_t |\mathbb{P}(\sqrt{n}\tfrac{\widehat{\mu}-\mu}{\sigma} \leq t) - \mathbb{P}(Z \leq t/\sigma)| \leq \tfrac{33}{4}\tfrac{\mathbb{E}[|X_i-\mu|^3]}{\sqrt{n}}.$$

The same can be done for (III),

$$III = \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \leq \tfrac{33}{4}\tfrac{\frac{1}{n}\sum|X_i-\widehat{\mu}|^3}{\sqrt{n}}.$$

□

contd.

By the SLLN, $\frac{1}{n}\sum |X_i - \widehat{\mu}|^3 \xrightarrow{a.s} \mathrm{E}[|X_i - \mu|^3]$, so for n large enough, $\frac{1}{n}\sum |X_i - \widehat{\mu}|^3 \leq 2\mathrm{E}[|X_i - \mu|^3]$ and so,

$$III \leq \frac{33}{4}\frac{2\mathrm{E}[|X_i-\mu|^3]}{\sqrt{n}}.$$

Then, using the fact that $\widehat{\sigma} - \sigma = O_P(n^{-1/2})$, we can see that $II = O_P(n^{-1/2})$ (check: taylor expand $\phi_{\widehat{\sigma}}$ around σ). Thus, all terms converge at the same rate and the proof is completed. This can also be used to prove asymptotically correct coverage of the bootstrap confidence interval. □

- **Estimator is algorithmic**:
  - 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:
  - $\widehat{G} \approx G$ — strong 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).

- 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

$$\Pr(\theta \leq \theta_\alpha) = \alpha, \quad 0 < \alpha < 1, \theta \in \Theta. \tag{1}$$

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

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

- In practice exact intervals are rarely available, and we seek intervals such that (1) is satisfied as closely as possible. If $Y \equiv Y_1, \ldots, Y_n$, then we typically have

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

  and the corresponding two-sided interval satisfies

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

- 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

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

  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.

- For the handedness data, try Fisher's z transformation:

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

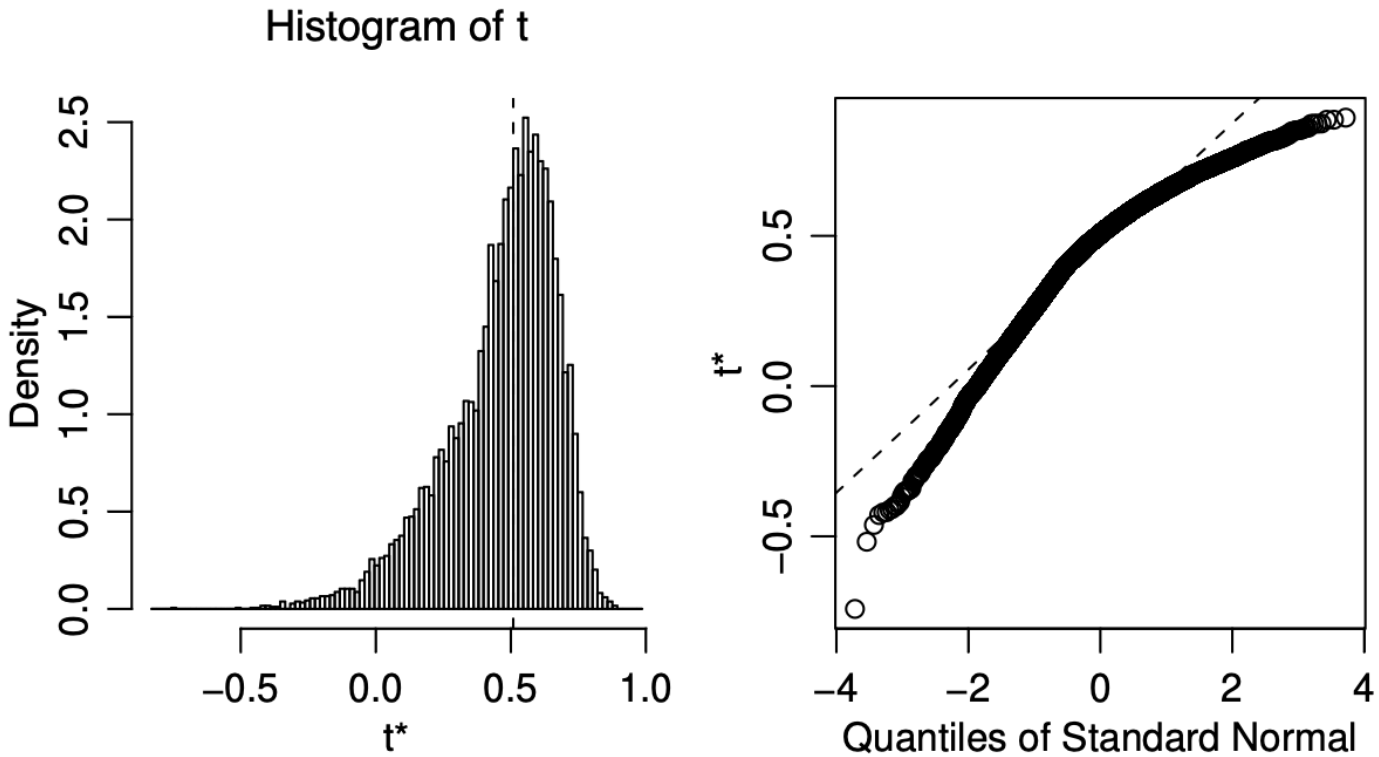
  with bias and variance estimates

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

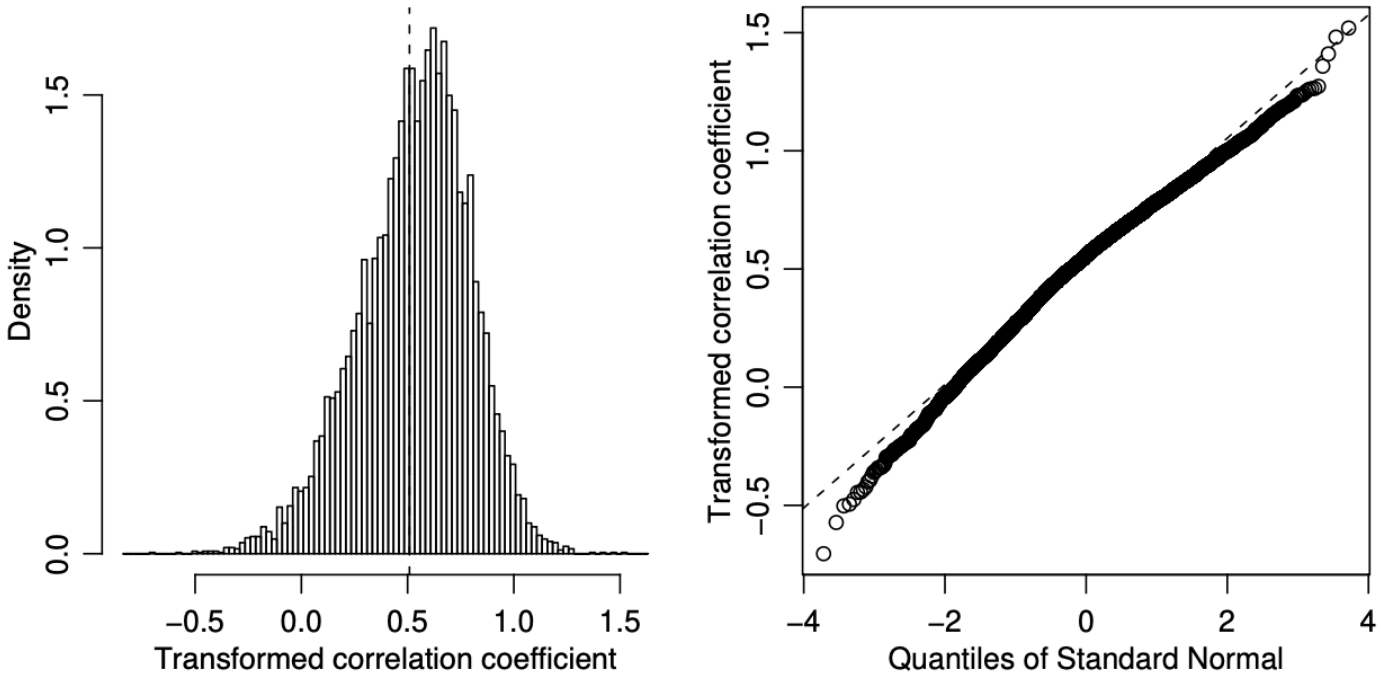
- Then the $(1-2\alpha)$ confidence interval for θ is

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

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



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



But how do we choose a transformation in general?

- 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.
- Recall that a **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

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

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

- 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

$$z_1^* = \frac{\widehat{\theta}_1^* - \widehat{\theta}}{V_1^{*1/2}}, \quad z_2^* = \frac{\widehat{\theta}_2^* - \widehat{\theta}}{V_2^{*1/2}}, \quad \ldots, \quad z_R^* = \frac{\widehat{\theta}_R^* - \widehat{\theta}}{V_R^{*1/2}},$$

  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 $\widehat{\theta}_r^*$.

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

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

  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

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

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

$$\Pr_{\widehat{G}}(Z^* \le z) - \Pr_G(Z \le z) = O_p(n^{-1}),$$

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

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

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

  and

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

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

$$\Pr_{\widehat{G}}(Z^* \leq z) - \Pr_G(Z \leq z) = O_p(n^{-1/2}),$$

  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 $O_p(n^{-1/2})$.

- 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.

- 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.