Bootstrap

Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch

December 9, 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 $u = 0$,

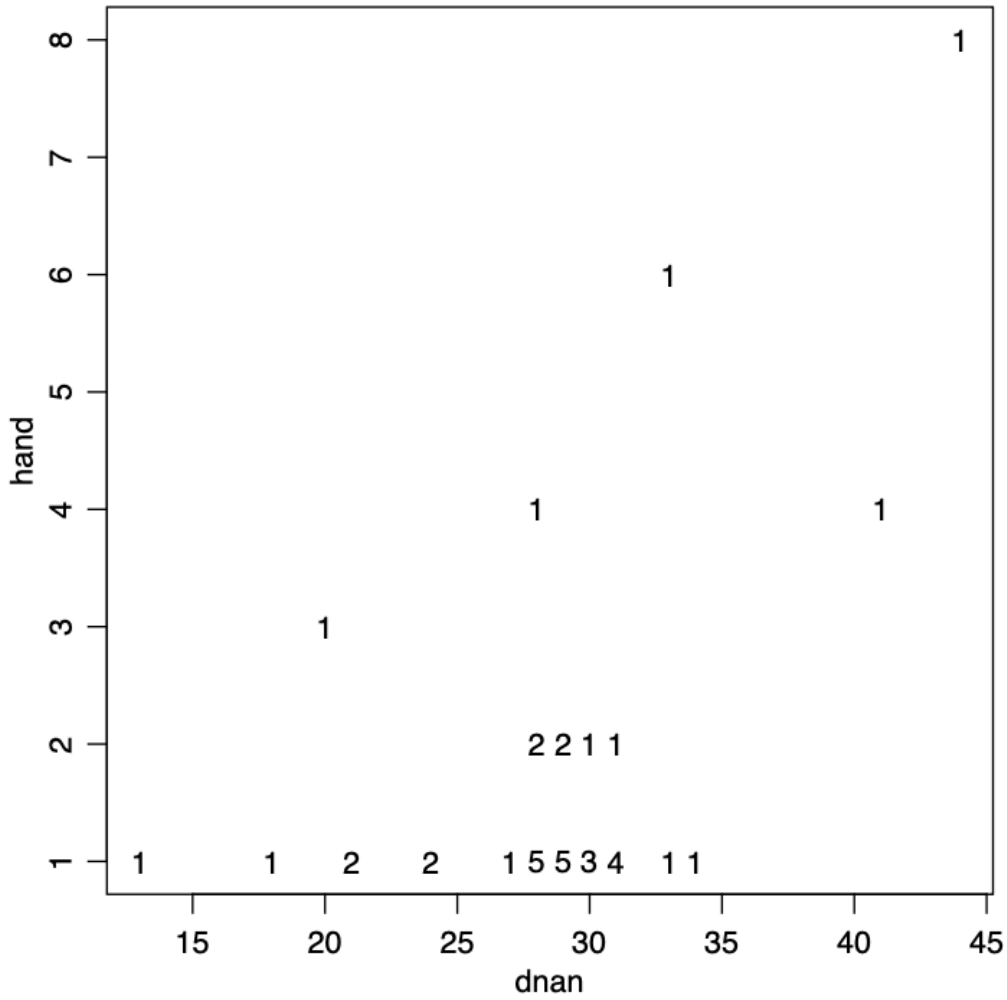
$$\begin{aligned}
t_2(\widehat{G}) &= \operatorname{argmax}_\theta \int \log f(y;\theta)\, \mathrm{d}\widehat{G}(y) \\
&= \operatorname{argmax}_\theta \int \log f(y;\theta)\, \mathrm{d}\left\{ n^{-1} \sum_{j=1}^{n} H(y - Y_j) \right\} \\
&= \operatorname{argmax}_\theta n^{-1} \sum_{j=1}^{n} \int \log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

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

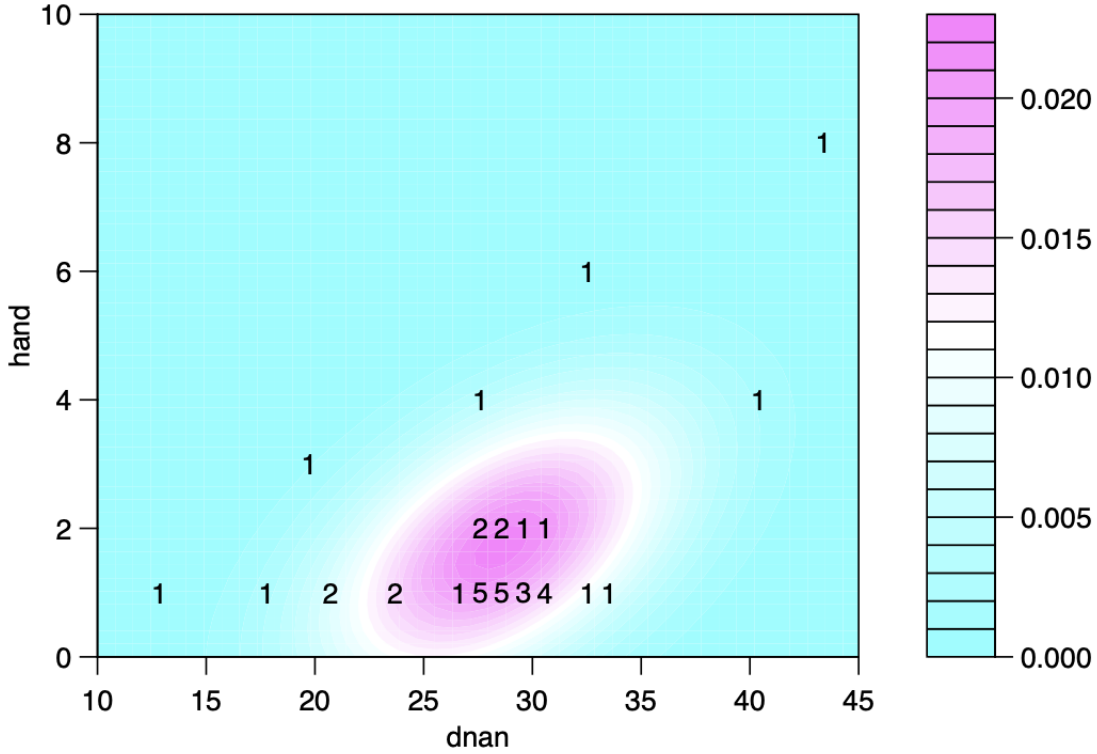
$$\begin{aligned} \widehat{\theta} = t(\widehat{G}) &= \frac{\sum_{j=1}^n (\texttt{dnan}_j - \overline{\texttt{dnan}})(\texttt{hand}_j - \overline{\texttt{hand}})}{\left\{\sum_{j=1}^n (\texttt{dnan}_j - \overline{\texttt{dnan}})^2 \sum_{j=1}^n (\texttt{hand}_j - \overline{\texttt{hand}})^2\right\}^{1/2}}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

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

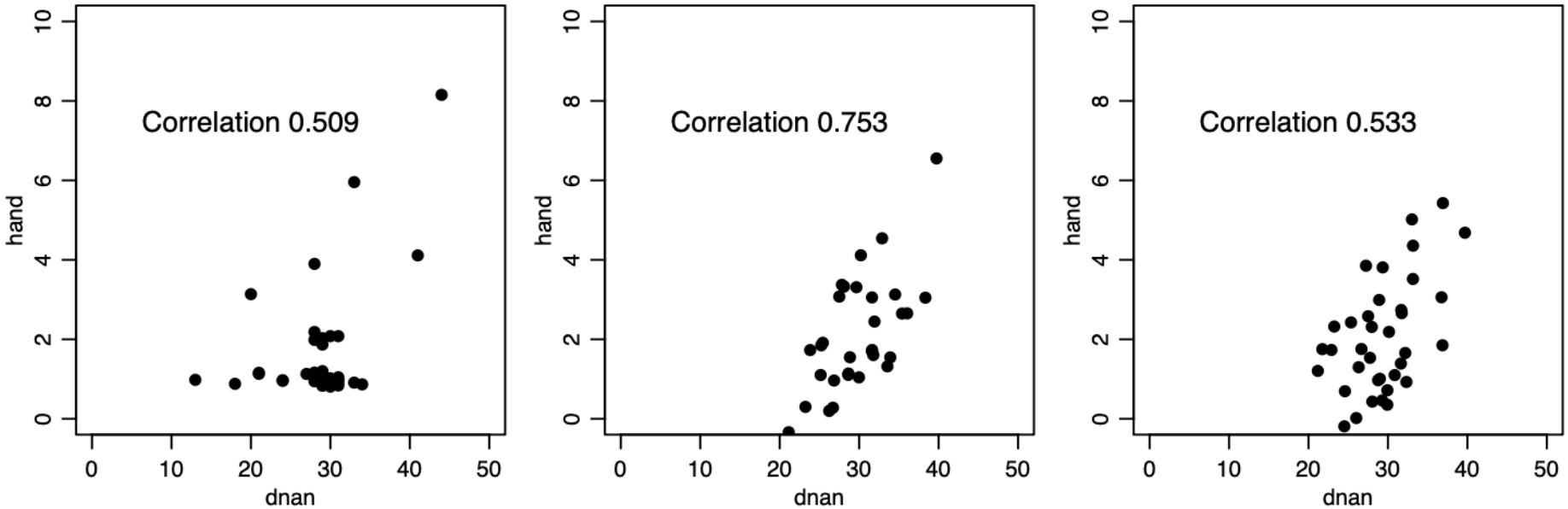
Try simulation approach ...

- 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.

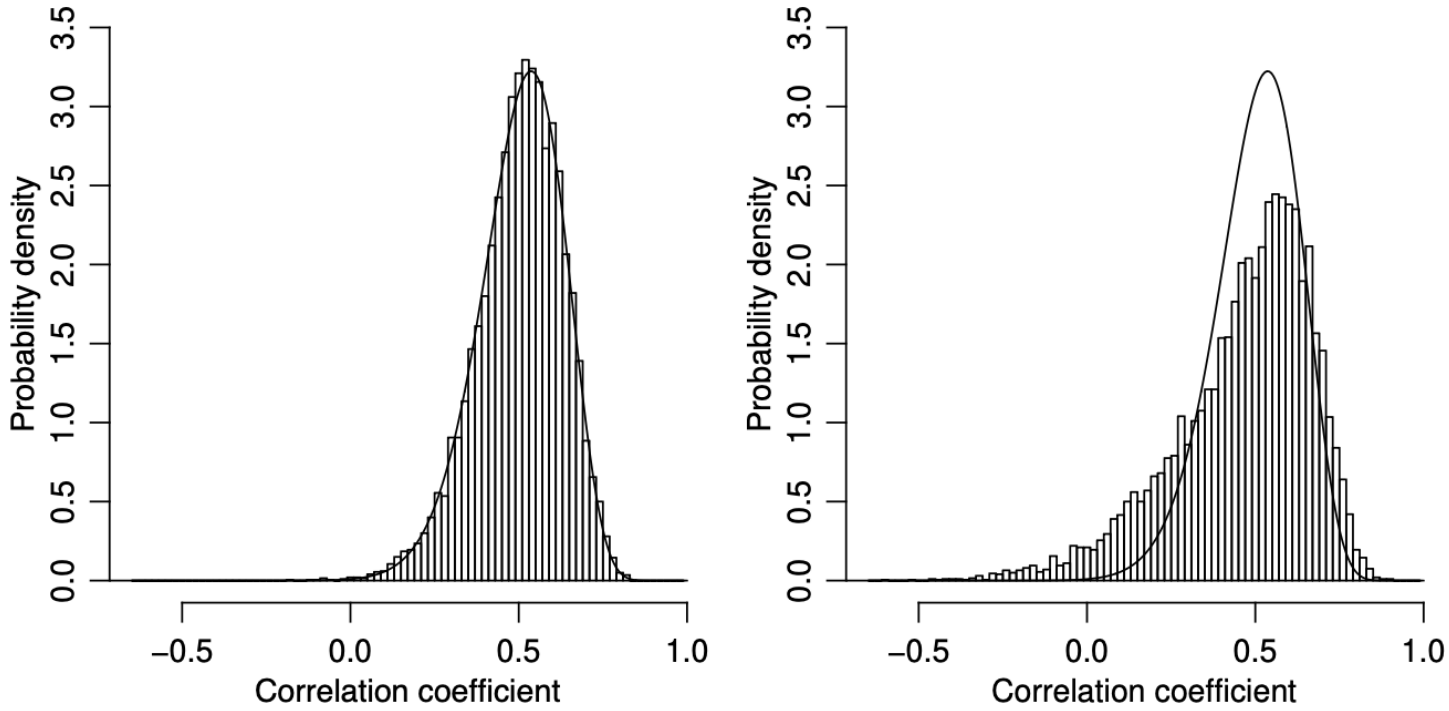
Contours of bivariate normal distribution fitted to handedness data; parameter estimates are $\widehat{\mu}_1 = 28.5$, $\widehat{\mu}_2 = 1.7$, $\widehat{\sigma}_1 = 5.4$, $\widehat{\sigma}_2 = 1.5$, $\widehat{\rho} = 0.509$. The data are also shown.



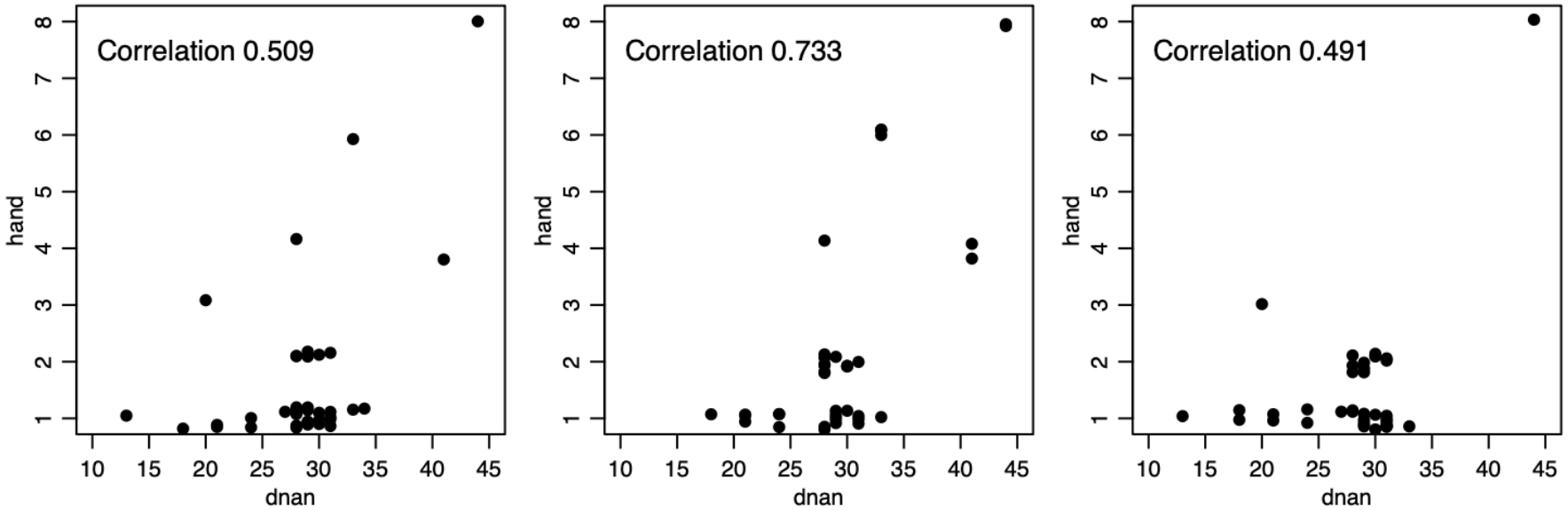
Left: original data, with jittered vertical values. Centre and right: two samples generated from the fitted bivariate normal distribution.



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 $\widehat{G}$:

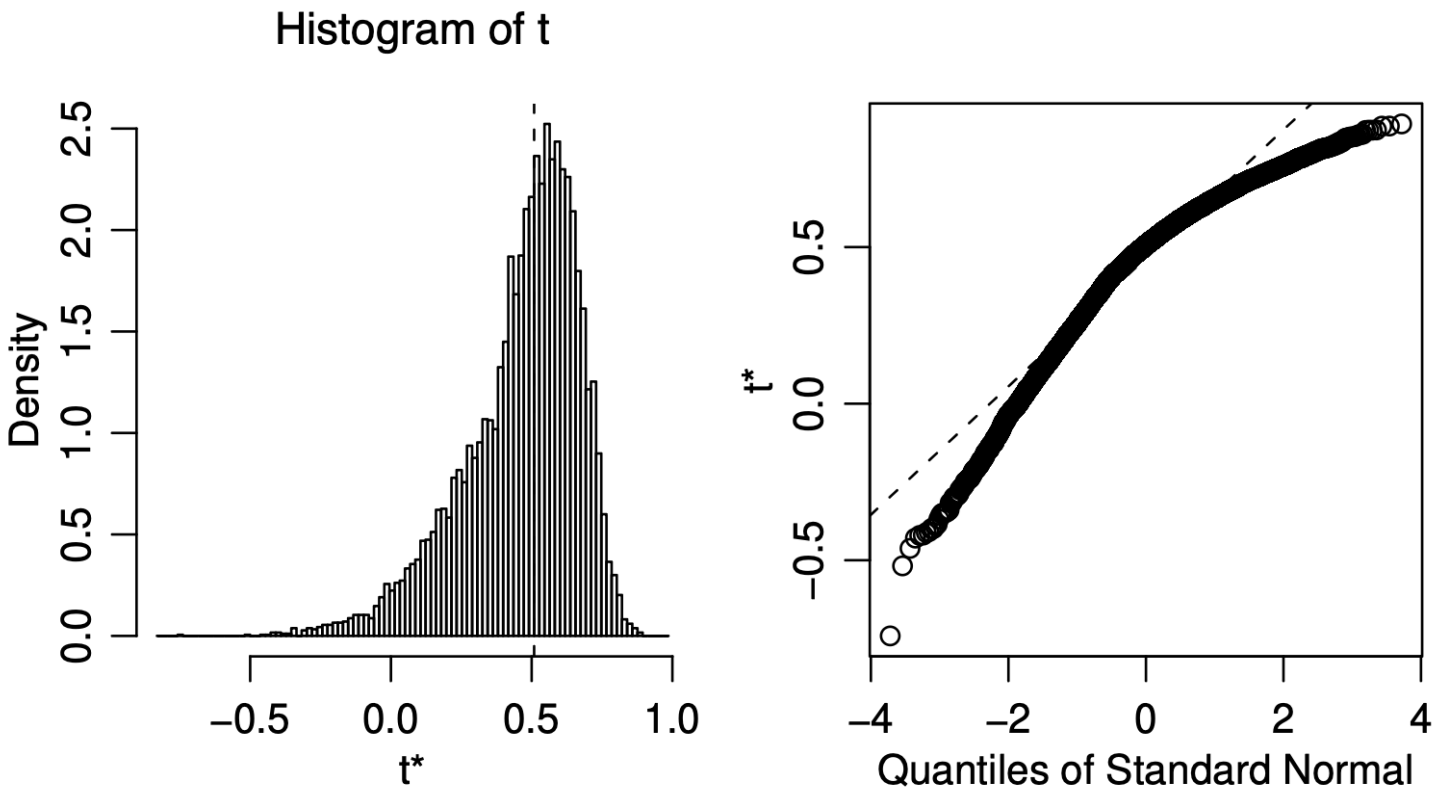
$$\beta(\widehat{G}) = \mathrm{E}(\widehat{\theta} \mid y_1, \ldots, y_n \overset{\text{iid}}{\sim} \widehat{G}) - t(\widehat{G}), \quad \nu(\widehat{G}) = \mathrm{var}(\widehat{\theta} \mid y_1, \ldots, y_n \overset{\text{iid}}{\sim} \widehat{G}).$$

- 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

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

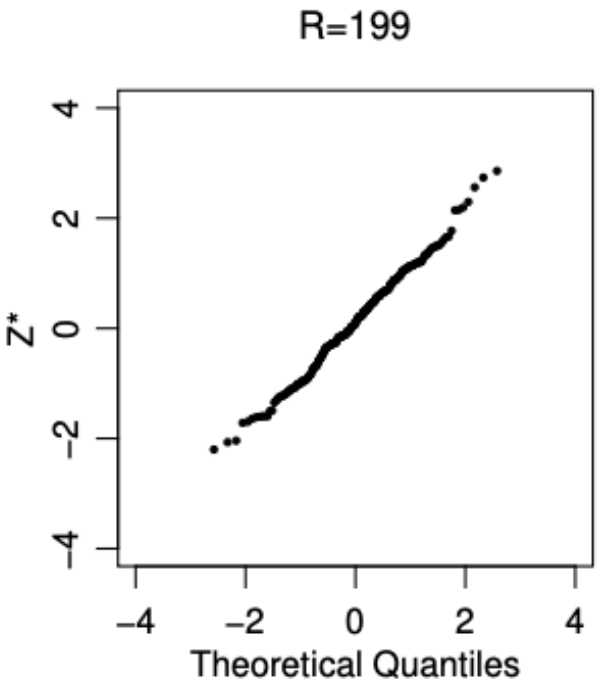
so exact calculations of the variance etc. are possible.

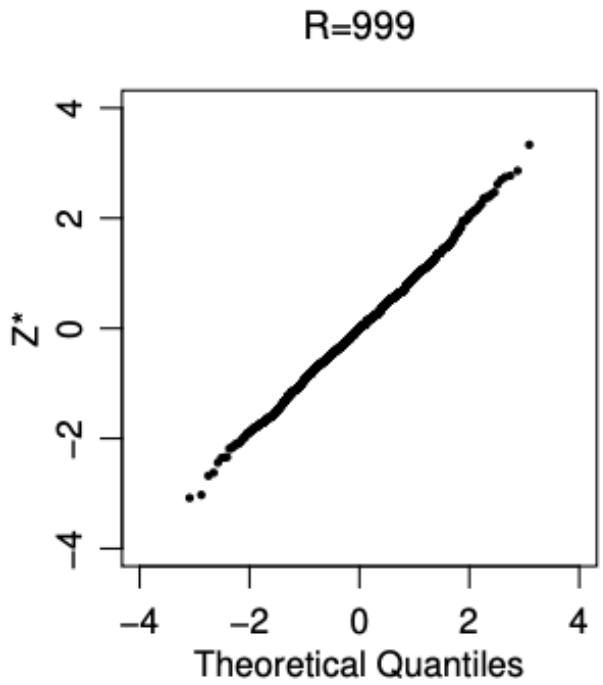
- However the median is very vulnerable to bad sample values, so for the median (and other 'non-smooth' statistics) much larger n is needed for reliable inference.

What about outliers?

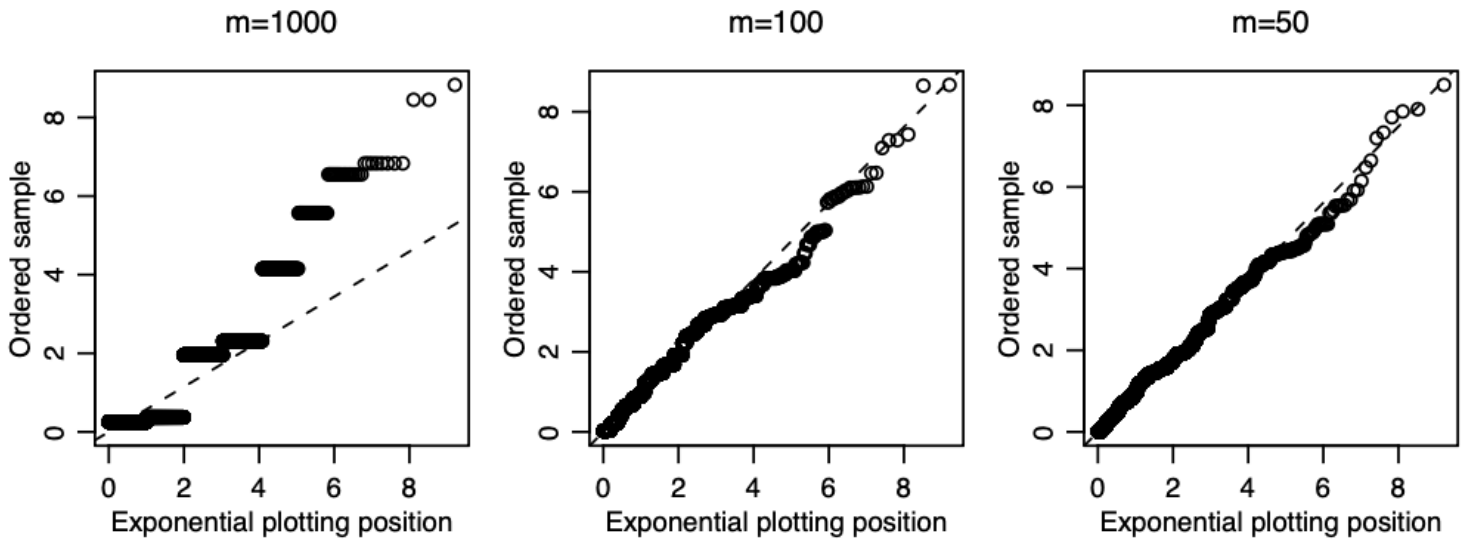
- Bad sample values not specific to bootstrap methods—need to consider them in any analysis
- Can reduce their effects by using robust estimators, but beware, as outliers may affect bootstrapped estimators even if they don't affect original value

- 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

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