Parameter estimation – Maximum Likelihood Estimation (MLE)

Let $\{X_1, \ldots, X_n\}$ be a random sample for the considered stochastic model and $(x_1, \ldots, x_n)' \in S^n$ the statistical data available.

The likelihood function is

$$L(x_1, \ldots, x_n; \theta) = \begin{cases} p(x_1, \ldots, x_n; \theta) & \text{for a discreet model} \\ f(x_1, \ldots, x_n; \theta) & \text{for a continuous model} \end{cases}, \theta \in \Theta$$

MLE

Obs if $x_1, \dots, x_n$ are i.i.d then

$$L(x_1, \dots, x_n; \theta) = \begin{cases} \prod_{i=1}^{n} p(x_i; \theta) & \text{for a discreet model} \\ \prod_{i=1}^{n} f(x_i; \theta) & \text{for a continuous model} \end{cases}$$

Def MLE is an estimator $\hat{\theta}_{MLE}(x_1, \dots, x_n)$ such that, for any $(x_1, \dots, x_n) \in S^n$, it has the value $\hat{\theta}_{MLE}(x_1, \dots, x_n)$ as solution of the optimization problem $\sup_{\theta \in \Theta} L(x_1, \dots x_n; \theta)$ or of the equivalent problem $\sup_{\theta \in \Theta} \ln L(x_1, \dots, x_n; \theta)$.

Examples

1. MLE for the parameters of $N(\mu, \sigma^2)$

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}$$

$$\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$$

Examples

Let $X_1, \dots, X_n$ be iid $N(\mu, \tau^2)$

$(x_1, \dots, x_n)$ the statistical data

$$L(x_1, \dots x_n; \mu, \tau^2) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2} (x_i - \mu)^2 \right) \right\} =$$

$$= (2\pi)^{-\frac{n}{2}} (\tau^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\tau^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

Examples

$$\ln L(x_1,\dots,x_n;\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\begin{cases} \dfrac{\partial \ln L}{\partial \mu} = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = 0 \end{cases} \quad \begin{cases} \displaystyle\sum_{i=1}^{n}(x_i-\mu) = 0 \\ -\dfrac{n}{2}\dfrac{1}{\sigma^2} + \dfrac{1}{2(\sigma^2)^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \end{cases} \quad \begin{cases} \hat{\mu}(x_1,\dots,x_n) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i = \bar{x} \\ \hat{\sigma}^2(x_1,\dots,x_n) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{cases}$$

Examples

$$\hat{\mu}_{MLE}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$$

$$\hat{\sigma}^2_{MLE}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$E(\hat{\mu}_{MLE}) = \frac{1}{n} \cdot n \cdot \mu = \mu$ - is unbiased for μ

$$Var(\hat{\mu}_{MLE}) = Var\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

Examples

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i-\mu)-(\bar{x}-\mu)\right]^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right]$$

$$E\left(\hat{\sigma}^2_{MLE}\right) = \frac{1}{n}\left[\sum_{i=1}^{n} \underbrace{E\left[(x_i-\mu)^2\right]}_{=\mathrm{Var}(x_i)} - n\,\underbrace{E\left[(\bar{x}-\mu)^2\right]}_{=\mathrm{Var}(\hat{\mu}_{MLE})}\right]$$

$$= \frac{1}{n}\left[n\sigma^2 - n\cdot\frac{\sigma^2}{n}\right] = \frac{n-1}{n}\sigma^2 \text{ - is biased for } \sigma^2$$

An unbiased estimator for σ^2 is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{n}{n-1}\hat{\sigma}^2_{MLE}$$

$$E(s^2) = \sigma^2$$

Examples

2. MLE for the parameters of the negative binomial distribution

$X \sim NBinom(k, p)$, $k \in \mathbb{N}$, $0 < p < 1$ p = prob. of "success"

X = the number of failures until we get k successes in a sequence of independent Bernoulli trials.

$$p(x) = P(X = x) = C_{x+k-1}^{k-1} \cdot p^k (1-p)^x \qquad x = 0, 1, \dots$$

if $(x_1, \dots, x_n)$ are the statistical data, then

$$L(x_1, \dots, x_n; k, p) = \prod_{i=1}^{n} p(x_i)$$

$$\ln L(x_1, \dots, x_n; k, p) = \sum_{i=1}^{n} \ln p(x_i)$$

Examples

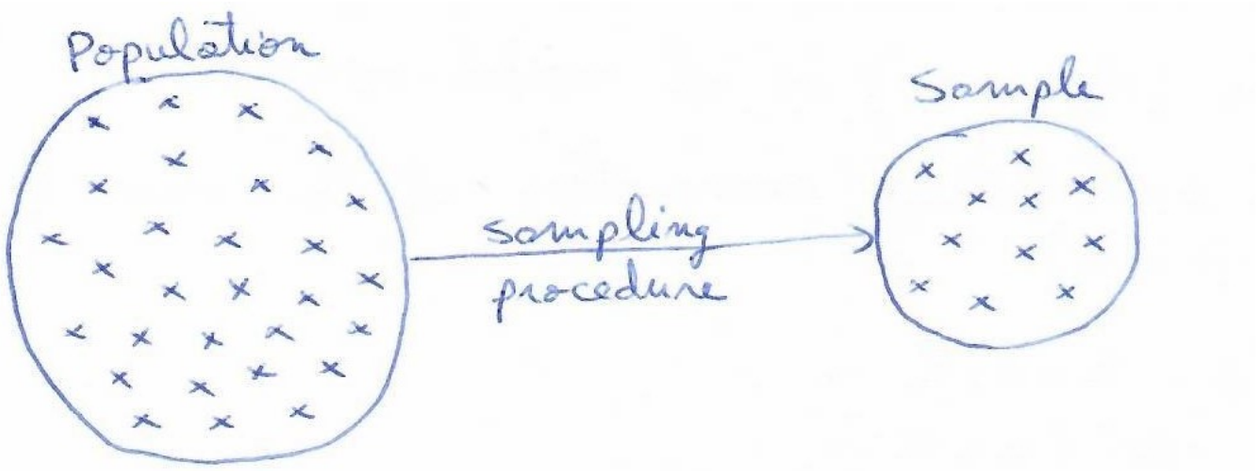
There is no closed-form solution to maximize $\ln L$, it must be solved numerically.

As statistical data, consider the following "game":

- someone tosses a coin until K heads are accumulated. We are not told the value of K, but the number of "failures" (i.e. tails) in each experiment $x_1, \ldots, x_n$ (also, the number of all tosses in each experiment would be good, instead of the number of failures).
- We are asked to estimate K (in this example $p = \frac{1}{2}$)

Sampling, resampling*

A sample is a subset from a larger set of data, called population.



It is assumed that the population follows an underlying (but unknown) distribution.
What is available for us: the sample data and its empirical distribution.

*Peter Bruce, Andrew Bruce. Practical Statistics for Data Scientists, O'Reilly Media, 2017

Sampling, resampling

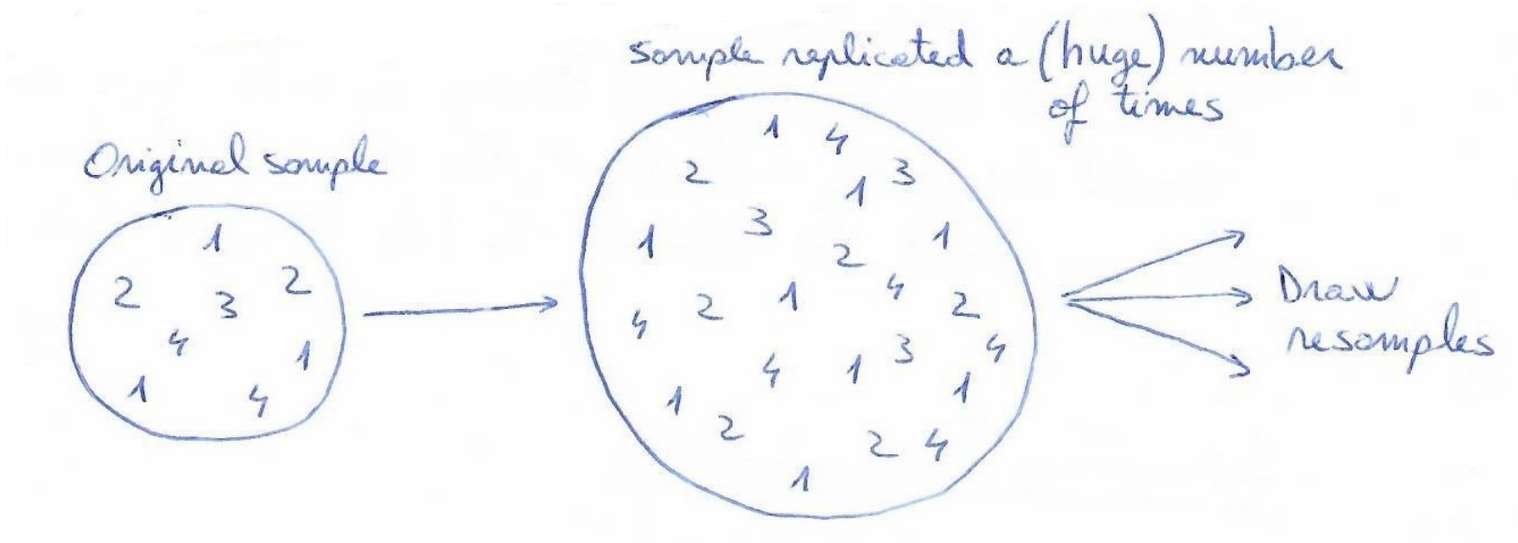
In a random sampling procedure, each member of the population has an equal chance of being chosen for the sample.

Sampling
- with replacement – observations are returned to the population after each draw and they are available for future selections.
- without replacement – a selected observation is no longer available for future selections.

Sampling, resampling

Bootstrapping

A bootstrap sample is a sample taken with replacement from an observed data set, with the general goal of evaluating random variability in a statistic.



Sampling, resampling

Bootstrapping allows us to estimate the properties of an estimator. We can estimate the sampling distribution of a statistic (e.g. mean) or of model parameters.

Example: a bootstrap resampling of the mean (for a sample size = n)

for $r = 1, R$
  for $i = 1, n$
    draw a sample value, record it and replace it
  Compute the mean $\bar{X}_r$ and record it

Sampling, resampling

Use $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_R$

- to compute the standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{R} (\bar{x}_i - \bar{\bar{x}})^2}{R-1}}, \quad \bar{\bar{x}} = \frac{1}{R} \sum_{i=1}^{R} \bar{x}_i$$

s is the standard error for the sample mean

- plot a histogram
- find a confidence interval.

Sampling, resampling

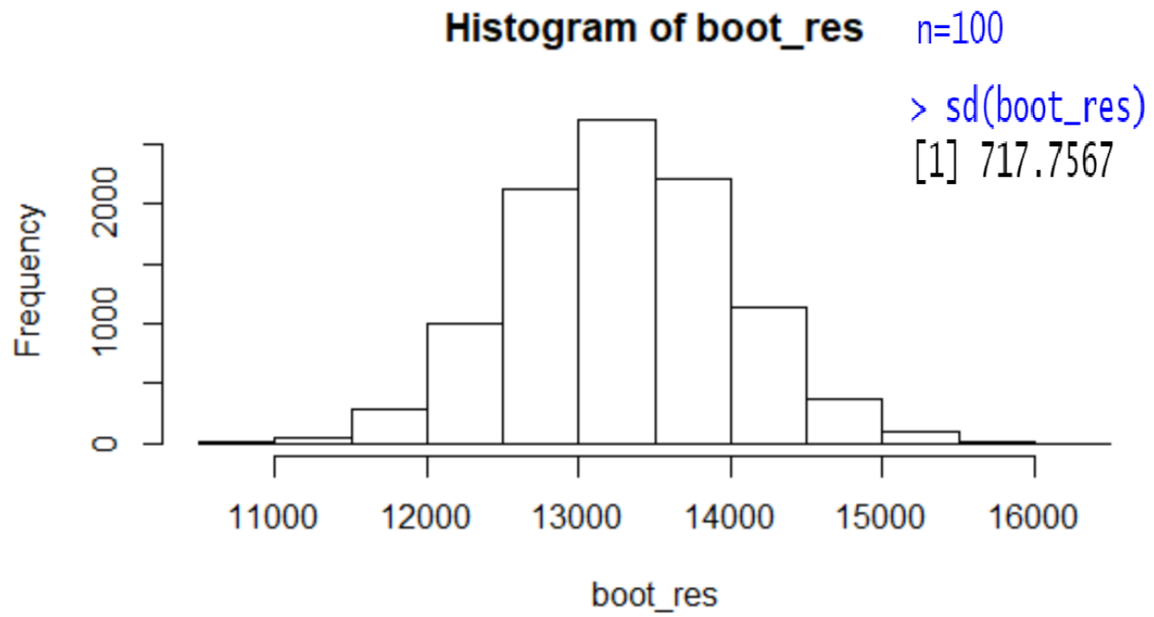
The Central Limit Theorem says that the means from multiple samples resemble the gaussian curve, even if the initial population is not normal (the samples size should be large enough).

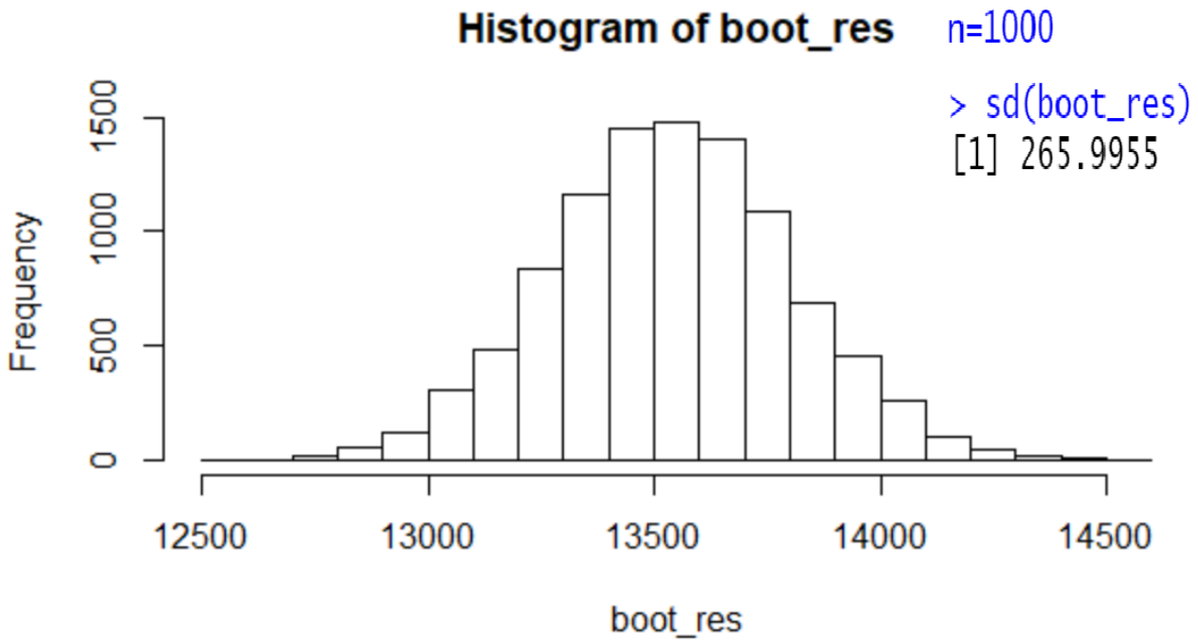
The bootstrap can be used to determine the sample size: try different values for n to see how the sampling distribution is affected.

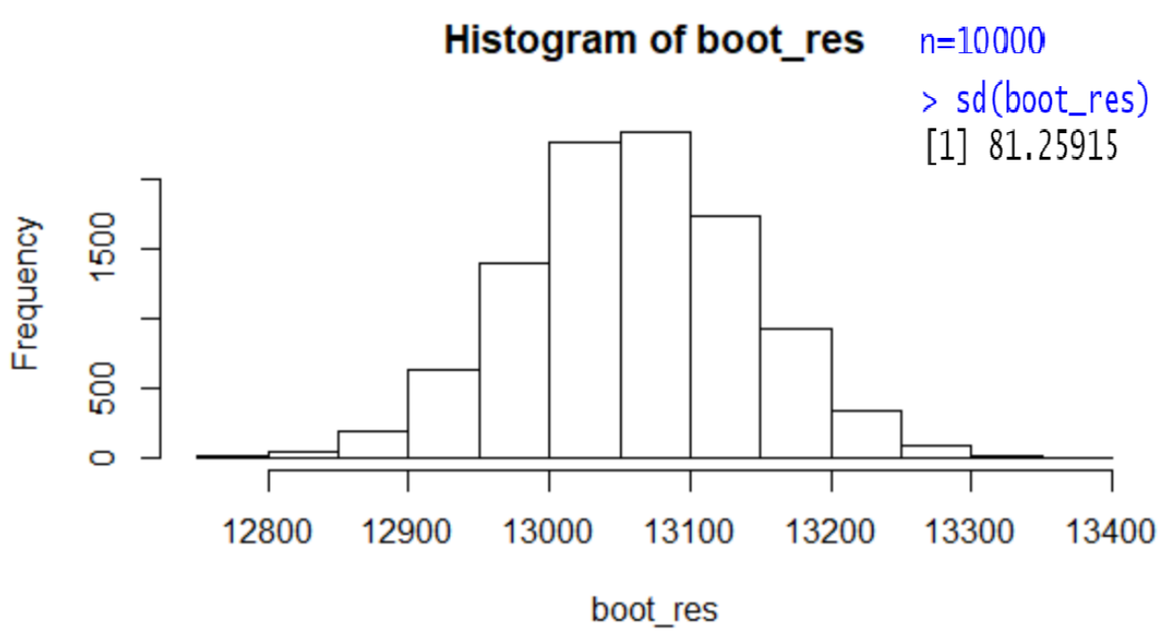
The bootstrap does not compensate for a small sample size as it does not create new data. It just tells us about how many samples would behave when drawn from the original sample.

Sampling, resampling

R=10000







Confidence intervals*

It is a tool to understand how variable a sample result might be (what is the potential error in a sample estimate).

Confidence intervals have a high percentage associated (90%, 95%), called level of confidence.

For example, a 90% confidence interval contains the central 90% of the bootstrap sampling distribution of a sample statistic.

*Peter Bruce, Andrew Bruce. Practical Statistics for Data Scientists, O'Reilly Media, 2017

Confidence intervals

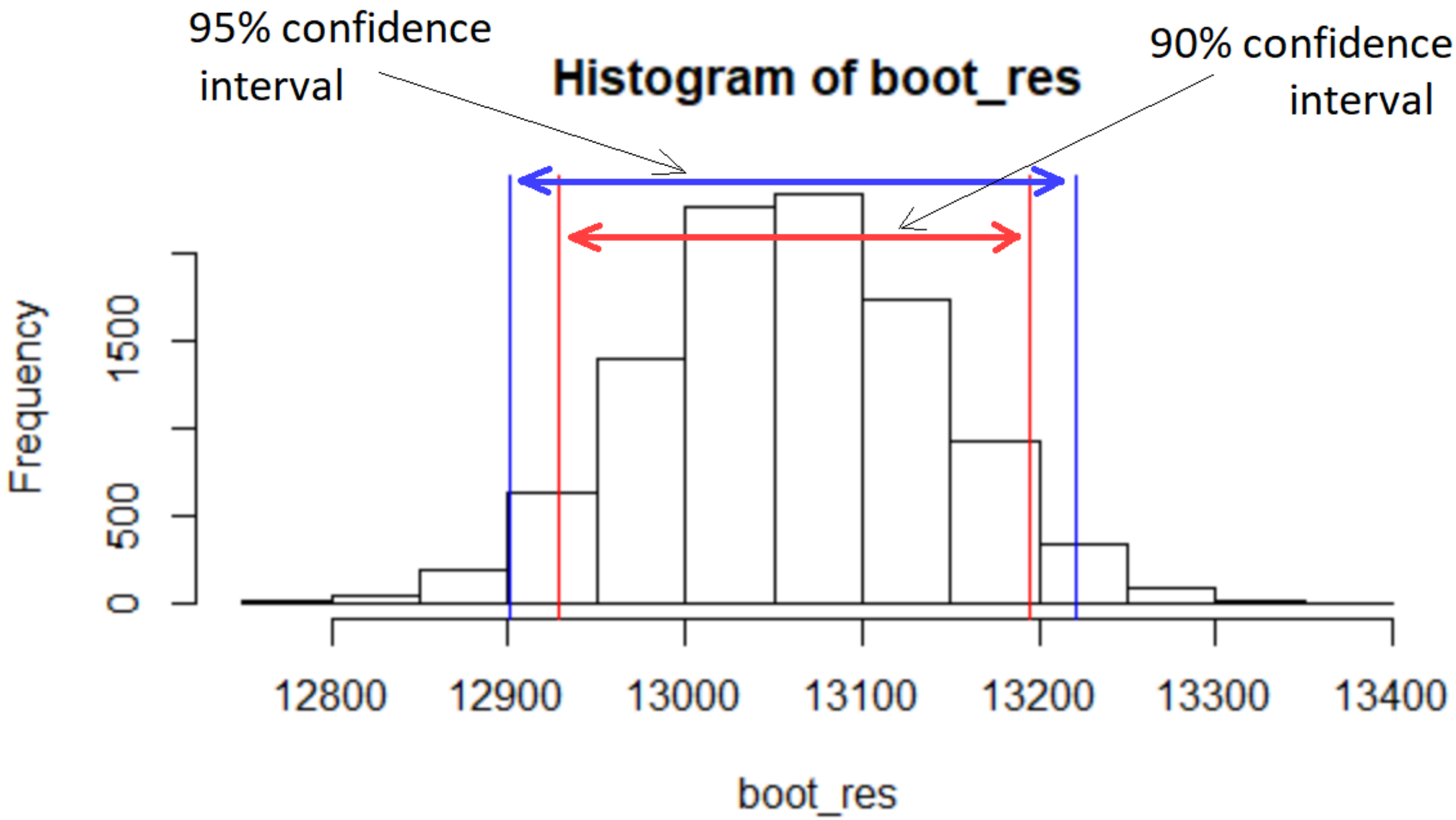
Input: the sample size n
a sample statistic of interest (e.g. mean)

for $r = 1, R$
  draw a random sample of size n with replacement
  record the statistic of interest for that sample

for an x% confidence interval, trim $\left[\frac{100-x}{2}\right]$% of the R sample results from both ends of the distribution.

Output: the trim points are the endpoints of an x% bootstrap confidence interval

Confidence intervals



Permutation tests*

Humans have the tendency to misinterpret randomness, to underestimate it.

Ask a person to imagine a series of 50 coin flips and then put the person to actually flip it 50 times and write the results.

In the real sequence, it is not unusual to see 5-6 of heads in a row (or tails).

In the "imagined" sequence after 3 consecutive heads, the person feels that it would be better to switch to a tail.

We are inclined to attribute a sequence of 5 heads in a row to a different cause, not just chance.

*Peter Bruce, Andrew Bruce. Practical Statistics for Data Scientists, O'Reilly Media, 2017

Permutation tests

Hypothesis tests are used to find (significant) differences between groups. There is a baseline assumption, called null hypothesis, that assumes that the groups are "equivalent" (that is, any difference between the groups is due to chance).

In experiments, we ask for proof that the difference between groups is greater than what chance might produce.

Permutation tests are used to test hypotheses, involving two (or more) groups.

Permutation tests

Suppose we have two groups A and B.

The permutation procedure is as following:

for $r = 1, R$

- combine the two groups in a single data set
- shuffle the combined data, then randomly draw without replacing a resample of the same size as group A; the remaining data go to group B
- the statistic calculated for the original samples is calculated now for the resamples and is recorded

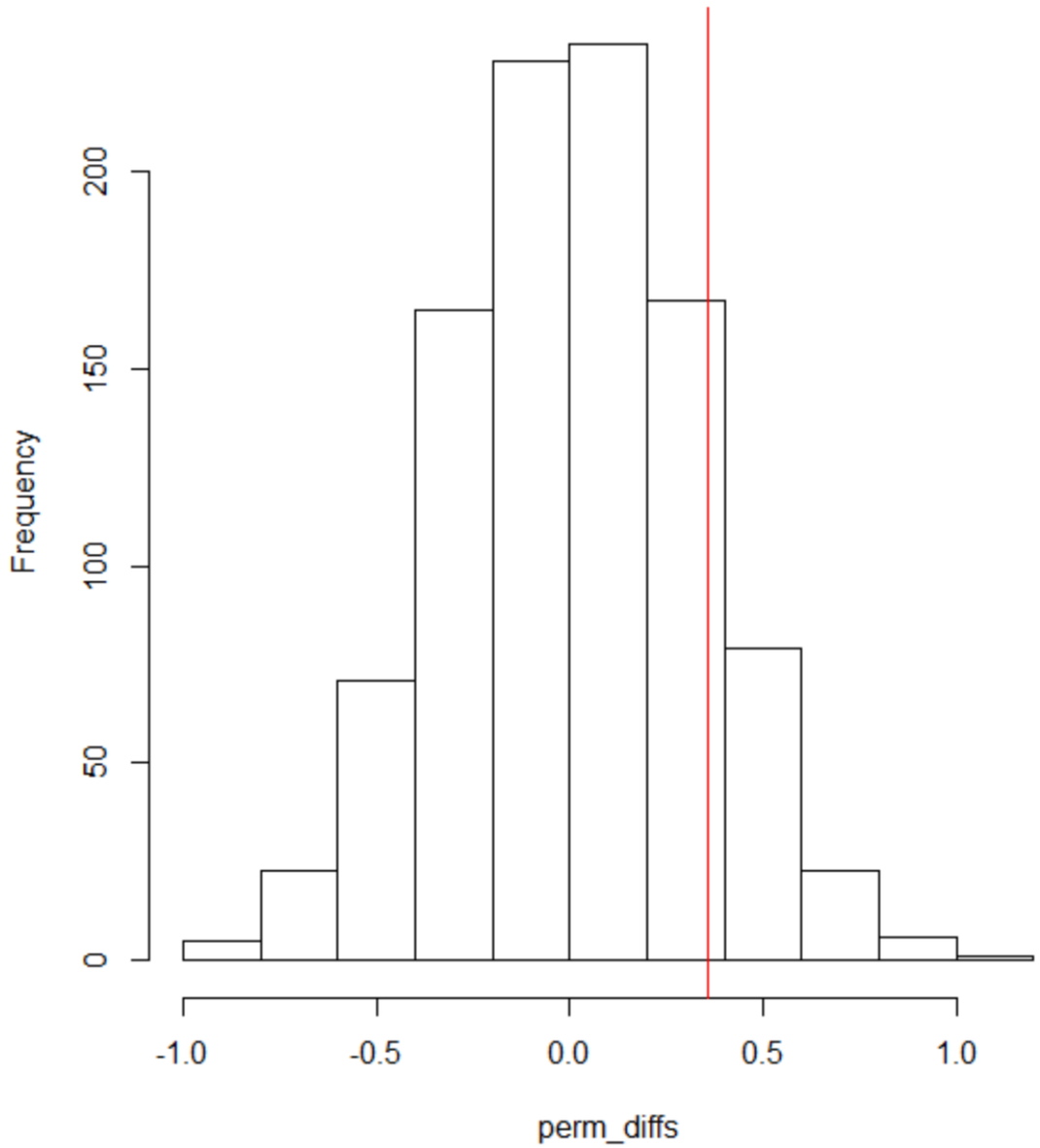
Permutation tests

We compare the statistic of the original samples to the set of permuted statistics. If the initial statistic lies well within the set of permuted statistics, then the observed difference between groups is due to chance. Otherwise, we conclude that there is a significant statistical difference.

Example The stetistic calculated for the two groups A and B may be $mean_A - mean_B$.

Permutation tests

Difference between groups due to chance



Significant statistical difference