ADVANCED BIOSTATISTICS
ABSTAT18

Monte Carlo and Bootstrap methods

Carina Silva
(*carina.silva@estesl.ipl.pt*)
Higher School of Technologies and Health of Lisbon &
Center of Statistics and Applications, University of Lisbon| CEAUL

ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA
INSTITUTO POLITÉCNICO DE LISBOA

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

IGC, April 3rd - 6th, 2018

Motivation

- Small samples that are NOT from a normal distribution
- In the old days: non-parametric tests
- More common now: Simulation based statistics:
  - Confidence intervals are much easier to achieve
  - They are much easier to apply in more complicated situations

What is a simulation study

Simulation: A numerical technique for conducting experiments on the computer

Monte Carlo simulation: Computer experiment involving random sampling from probability distributions

- Invaluable in statistics
- Usually, when statisticians talk about *simulations*, they mean *Monte Carlo simulations*

What is a simulation?

- (Pseudo) random numbers generated from a computer.
- A random number generator is an algorithm that can generate x_{i+1} from x_i.
- Require a “start”called “seed”if you want to get reproducible data.
- Basically the uniform distribution is simulated in this way.

- **Task**: Do exercises 1 and 2 from `simulation.md` file.

Monte Carlo (MC) simulation to the rescue

- An estimator or test statistic has a true sampling distribution under a particular set of conditions (finite sample size, true distribution of the data, etc.).

Monte Carlo (MC) simulation to the rescue

- An estimator or test statistic has a true sampling distribution under a particular set of conditions (finite sample size, true distribution of the data, etc.).
- Ideally, we would want to know this true sampling distribution.

Monte Carlo (MC) simulation to the rescue

- An estimator or test statistic has a true sampling distribution under a particular set of conditions (finite sample size, true distribution of the data, etc.).
- Ideally, we would want to know this true sampling distribution.
- But sometimes derivation of the true sampling distribution is not tractable or impossible.

Monte Carlo (MC) simulation to the rescue

- An estimator or test statistic has a true sampling distribution under a particular set of conditions (finite sample size, true distribution of the data, etc.).
- Ideally, we would want to know this true sampling distribution.
- But sometimes derivation of the true sampling distribution is not tractable or impossible.
- MC simulation allows approximate the sampling distribution of an estimator or test statistic **under a particular set of conditions**.

Monte Carlo simulation

A typical Monte Carlo simulation involves the following:

- Generate S independent data sets under the conditions of interest.

Monte Carlo simulation

A typical Monte Carlo simulation involves the following:

- Generate S independent data sets under the conditions of interest.
- Compute numerical values of the estimator/test statstic, for each data set $\Longrightarrow t_1^*, ...; t_S^*$

Monte Carlo simulation

A typical Monte Carlo simulation involves the following:

- Generate S independent data sets under the conditions of interest.
- Compute numerical values of the estimator/test statstic, for each data set $\Longrightarrow t_1^*, ...; t_S^*$
- If S is large enough, summary statistics across $t_1^*, ..., t_S^*$ should be good approximations to the true sampling properties of the estimator/test statistic under the conditions of interest.

Exercise 1

1. Generate 1000 MC replicates from samples of size 15 of a normal distribution with μ=1 and σ=1.7.

Exercise 1 (cont.)

1. Get the following MC estimates: MC mean, MC standard deviation, MC bias and MC MSE, of the true parameter μ, considering the sample mean, 20% trimmed mean and median estimators.

Simulation procedure

For a particular choice of $\mu = 1$, $n = 1000$ and true underlying distribution (normal):

- Generate independent draws $Y_1, ..., Y_n$ from the distribution
- Compute $T^{(1)}$, $T^{(2)}$ and $T^{(3)}$
- Repeat *S=1000* times:
  $T_1^{(1)}, ..., T_S^{(1)}$; $T_1^{(2)}, ..., T_S^{(2)}$; $T_1^{(3)}, ..., T_S^{(3)}$

Simulation procedure

- Compute for $k = 1, 2, 3$:

$$\widehat{mean}_{MC}^{(k)} = \frac{\sum_{s=1}^{S} T_s^{(k)}}{S} = \bar{T}(k);$$

$$\widehat{bias}_{MC}^{(k)} = \bar{T}^{(k)} - \theta;$$

$$\widehat{SD}_{MC}^{(k)} = \sqrt{\frac{\sum_{i=1}^{S}(T_s^{(k)} - \bar{T}^{(k)})^2}{S-1}};$$

$$\widehat{MSE}_{MC}^{(k)} = \frac{\sum_{s=1}^{S}(T_s^{(k)} - \mu)^2}{S} \approx (\widehat{SD}^{(k)})^2 + (\widehat{bias}^{(k)})^2$$

Mean Squared Error

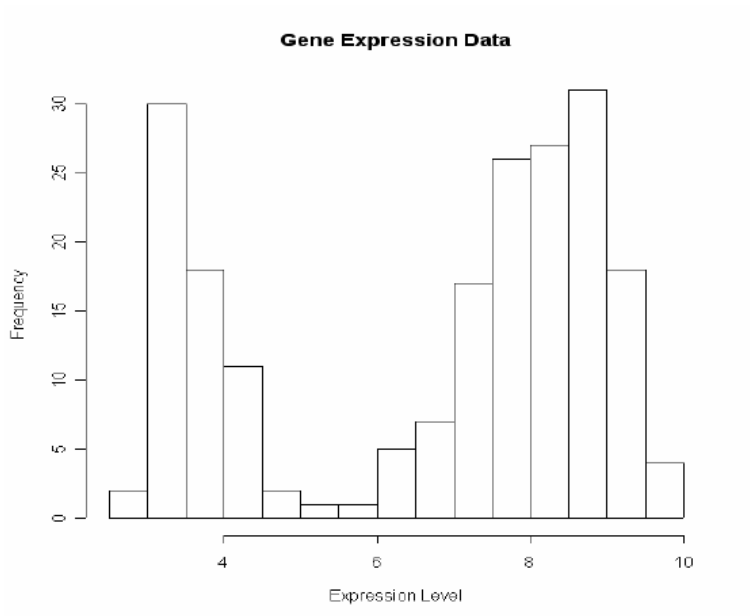
MSE: the mean squared error (MSE) of an estimator measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated.
An MSE of zero, meaning that the estimator $\hat{\theta}$ predicts observations of the parameter θ with perfect accuracy, is the ideal, but is practically never possible.

Exercise 1 (cont.)

1. Compare the results and choose the best estimator.

Bootstrap - Motivation

- You collect gene expression data for a particular gene from 200 samples of cells.



Bootstrap - Motivation

- Clearly this is not a nice distribution pattern in terms of being able to fit a standard probability model.
- Computing a mean expression level alone given this distribution may not have meaning and does not describe the situation well.
- Hence, a more important objective of study is primarily to understand the variability of the expression level for this particular gene.

Motivation

- This is a case where the bootstrap can be a useful technique.
- bootstrap is a technique developed by Brad Efron to find standard errors (measures of variability in the data) or confidence intervals in complicated situations where analytical computation is impossible.
- In bioinformatics literature bootstrapping is used extensively in phylogeny analysis and in microarray data analysis.

Introduction

- Bootstrapping is a method which uses random sampling techniques to estimate properties (such as bias, variance, confidence intervals, etc.) of an estimator, $\hat{\theta}$, when we don't know the true distribution, F_θ, of our data (and we cannot feasibly draw new samples from our population).

Introduction

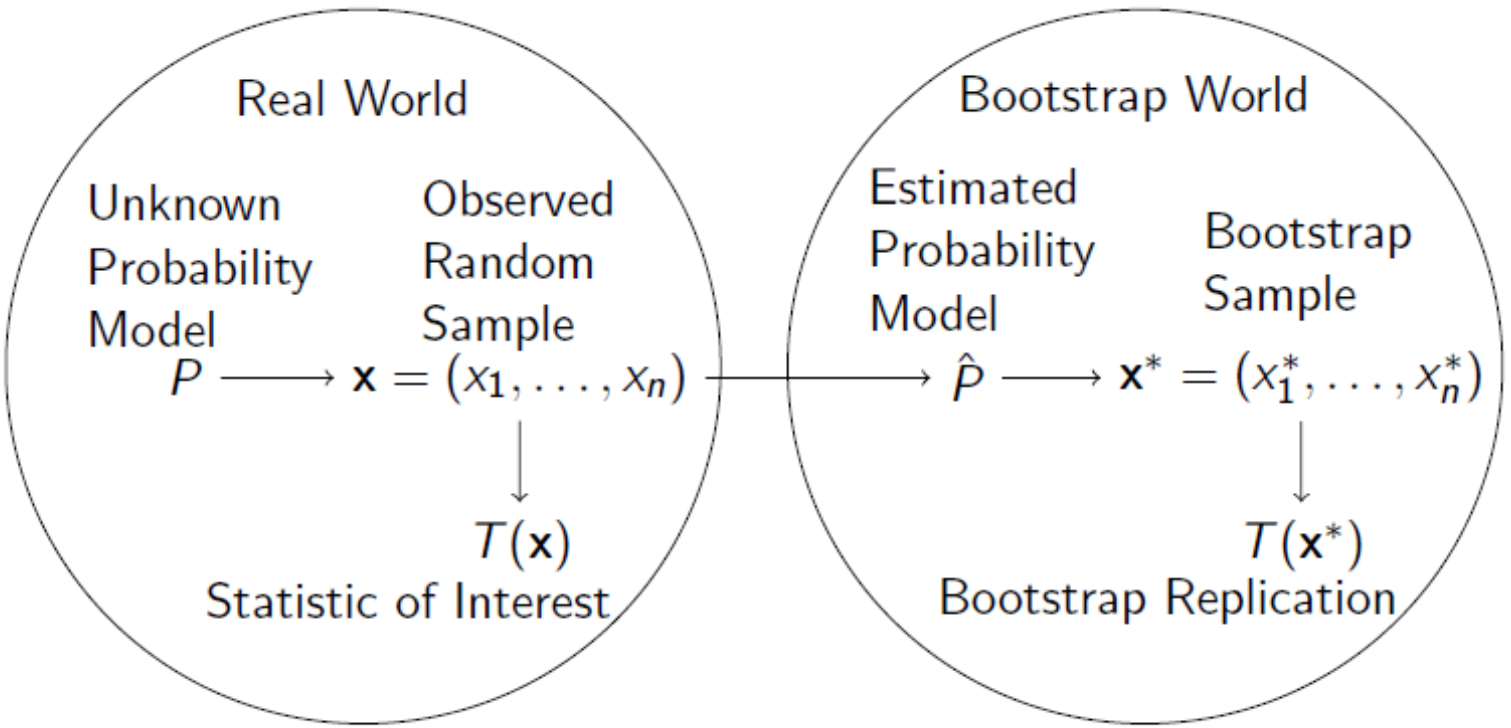
The term "bootstrap" comes from literature. In "The Adventures of Baron Munchausen", by Rudolph Erich Raspe, the Baron had fallen to the bottom of a deep lake, and he thought to get out by pulling himself up by his own bootstraps.

How does it work?

- *Step 1: Resampling*. A sampling distribution is based on many random samples from the population. In place of many samples from the population, create many resamples by repeatedly sampling with replacement from this one random sample. Each resample is the same size as the original random sample.

How does it work?

- *Step 2: Bootstrap distribution*. The sampling distribution of a statistic collects the values of the statistic from many samples. The bootstrap distribution of a statistic collects its values from many resamples.


Real World
Unknown Probability Model
Observed Random Sample
$P \longrightarrow \mathbf{x} = (x_1, \ldots, x_n)$
$T(\mathbf{x})$
Statistic of Interest
Bootstrap World
Estimated Probability Model
Bootstrap Sample
$\hat{P} \longrightarrow \mathbf{x}^* = (x_1^*, \ldots, x_n^*)$
$T(\mathbf{x}^*)$
Bootstrap Replication

Bootstrap

We will focus on two approaches to generating bootstrap samples:

- **Parametric bootstrap**: Estimate $\hat{\theta}$ from our original sample, $X_1, ..., X_n$ and generate samples $X_1^*, ..., X_n^*$ from $F_{\hat{\theta}}$, which approximates F_θ.
- **Non-parametric bootstrap**: Take samples $X_1^*, ..., X_n^*$ with replacement from our original sample $X_1, ..., X_n$.

Once we have B bootstrap samples, we can generate B estimates of $\hat{\theta}$:

$$X_1^{*(1)}, ..., X_n^{*(1)} \quad \rightarrow \quad \hat{\theta}^{*(1)}$$

$$\vdots$$

$$X_1^{*(B)}, ..., X_n^{*(B)} \quad \rightarrow \quad \hat{\theta}^{*(B)}$$

We can use these bootstrapped estimators, $\hat{\theta}^{*(1)}, ..., \hat{\theta}^{*(B)}$, to estimate properties of $\hat{\theta}$

Introduction

The number of bootstrap replications B

- A small number of replications B=25, is usually informative according to Efron.
- Fifty replications is often enough to give good estimate of standard error.
- Much bigger values of B are required for bootstrap confidence intervals.
- 1000 replications is recommended by Harrel and others for stable confidence intervals.

Parametric Bootstrap

Recall that we are using bootstrapping because we don't know the true parameter θ, but we want to identify the distribution of our estimator $\hat{\theta}$ (e.g. the MLE).

Suppose that we have observed a sample $X_1, ..., X_n$, $X_i \sim F_\theta$. Suppose further that we know F, but θ is unknown (For example, we know that our sample is $N(\theta, 2)$ distributed, but we don't know θ).

We can calculate $\hat{\theta} = T(X_1, ..., X_n)$ for our sample.

- To identify the distribution of our estimator, $\hat{\theta}$, we want to obtain more observations of $\hat{\theta}$, but we only have one sample!

How can we get more (approximated) values of $\hat{\theta}$?

Parametric Bootstrap

Why not approximate the distribution, F_θ, using our estimate, $\hat{\theta}$?
We can then draw samples from the approximated distribution $F_{\hat{\theta}}$:

$$F_{\hat{\theta}} \overset{simulate}{\rightarrow} \begin{cases} X_1^{*(1)}, ..., X_n^{*(1)} & \rightarrow \hat{\theta}^{*(1)} \\ \vdots & \\ X_1^{*(N)}, ..., X_n^{*(N)} & \rightarrow \hat{\theta}^{*(N)} \end{cases}$$

We can now estimate properties of $\hat{\theta}$ using these bootstrapped estimates, for example:

$$Var_{boot}(\hat{\theta}) = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\theta}_i^* - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j^* \right)^2$$

Exercise 2

Suppose it was drawn a sample of 300 observations with sample mean $\bar{x} = 2$ from an $exp(\lambda)$ distribution. Obtain the boostrap estimate of λ using B=1000 replicates and the bootstrap estimate of the standard error of $\hat{\lambda}$.

Classic Confidence Interval

A symmetric $100(1-\alpha)\%$ confidence interval (CI) has the form:

$$\hat{\theta} \pm t_{\alpha/2}\sigma_{\hat{\theta}}$$

where $\hat{\theta}$ is our estimate of θ, $\sigma_{\hat{\theta}}$ is the standard error of $\hat{\theta}$, $t_{\alpha/2}$ is the critical value of the test statistic, i.e., $P(T \leq t_{\alpha/2}) = \alpha/2$.

- Assumes that distribution of test statistic is symmetric around zero.
- As $n \to \infty$ we often have $\hat{\theta} \dot{\sim} N(0, \sigma_{\hat{\theta}})$, so that $t_{\alpha/2} = z_{\alpha/2}$

More generally we can write a $100(1-\alpha)\%$ CI as:

$$[\hat{\theta} - t_{1-\alpha/2}\sigma_{\hat{\theta}}; \hat{\theta} - t_{\alpha/2}\sigma_{\hat{\theta}}]$$

Proper interpretation of CIs

Unfortunately, (frequentist) confidence intervals don't have the interpretation that one might expect (or hope) for...

- Incorrect interpretations of CIs are prevalent in scientific papers.

Interpreting a 99% Confidence Interval:

- **Wrong:** through one sample; e.g., there is a 99% chance the confidence interval around my $\hat{\theta}$ contains the true θ (with $\alpha = .01$)
- **Correct:** through repeated samples, e.g., 99 out of 100 confidence intervals would be expected to contain true θ with $\alpha = .01$

Some properties of CIs

Two properties we can use to describe a conï¬¬*denceinterval* :

- length $= \hat{\theta}_{up} - \hat{\theta}_{lo}$
- shape $= \frac{\hat{\theta}_{up} - \hat{\theta}}{\hat{\theta} - \hat{\theta}_{lo}}$

Note that...

- Length: describes the overall size of the CI
- Shape: describes the asymmetry of the CI.
  shape > 1 indicates a greater distance between $\hat{\theta}_{up}$ to $\hat{\theta}$ than between $\hat{\theta}_{lo}$ to $\hat{\theta}$

Defining a Good CI

What is a "good" bootstrap confidence interval?

- If an exact CI can be formed (e.g., sample mean), bootstrap CI should closely match exact CI.
- If an exact CI cannot be formed (e.g., sample median), bootstrap CI should give accurate coverage probabilities.

Note that a narrower CI is not necessarily a better CI. Length and shape are only important if the coverage probabilities are accurate.

Different bootstrap CI methods have different coverage accuracies.

Parametric bootstrap CI

Ingredients to use parametric bootstrap approach to estimate a confidence interval for a parameter:

- Data $x_1, x_2, ..., x_n$ drawn from a distribution $F(\theta)$ with unknown parameter θ.
- A statistic $\hat{\theta}$ that estimates θ.
- B Bootstrap samples drawn from $F(\hat{\theta})$.

Bootstrap CI

Percentile Method to a (1-α)*100% CI for θ.

- For each bootstrap sample, $x_1^*, x_2^*, ..., x_n^*$, we compute $\hat{\theta}^*$.
- Get the empirical percentiles $\hat{\theta}^*_{(\alpha/2)}$ and $\hat{\theta}^*_{(1-\alpha/2)}$.
- The bootstrap CI for θ will be given by $(\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)})$.

Bootstrap CI

- **Normal interval**:

$$\hat{\theta} \pm q_{(1-\alpha/2)} \sqrt{Var_{boot}(\hat{\theta})}$$

where q is the $1 - \alpha/2$ quantile of the standard normal distribution.

- **Pivotal interval**:

$$(2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)})$$

Exercise 2 (cont.)

Suppose it was drawn a sample of 300 observations with sample mean $\bar{x} = 2$ from an $exp(\lambda)$ distribution. Estimate λ using a 95% parametric bootstrap confidence interval for λ.

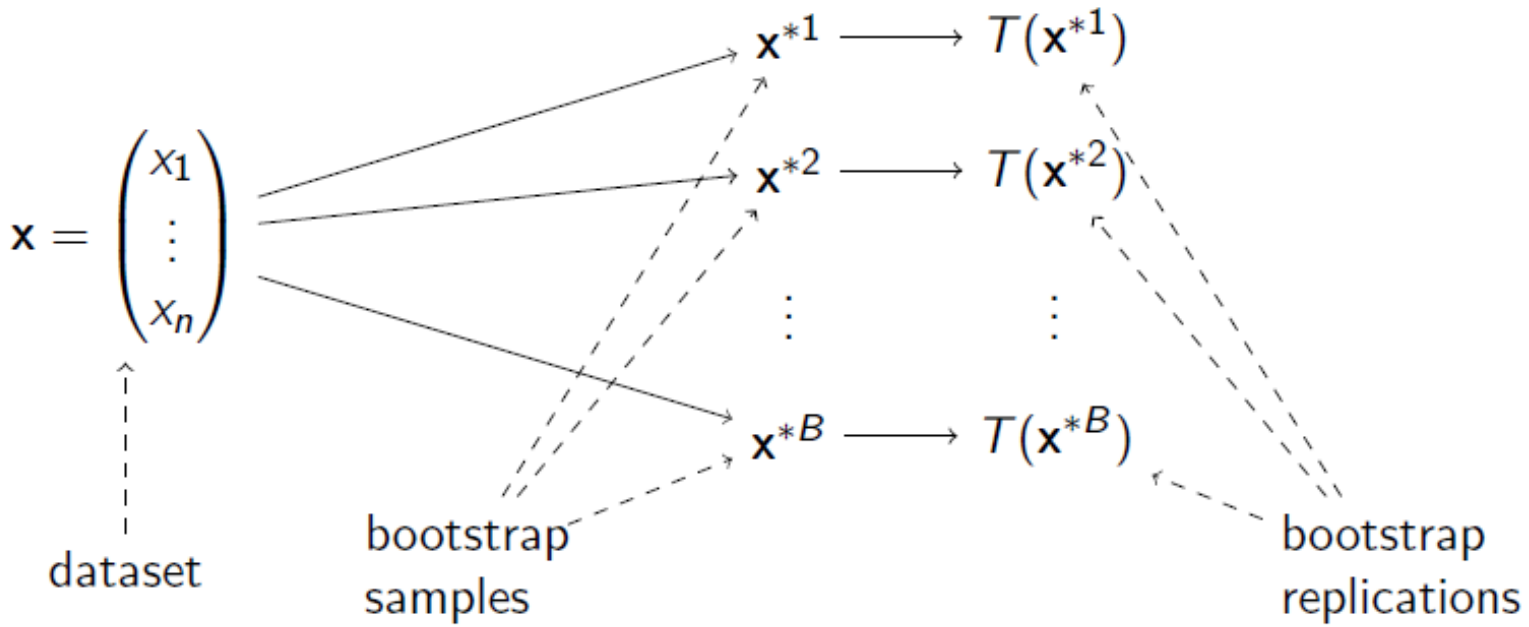
Exercise 3

The following measurements were given for weights (Kg) of 11 children with ages between 8 and 10 years old with renal disfunction: 38.43, 38.43, 38.39, 38.83, 38.45, 38.35, 38.43, 38.31, 38.32, 38.48, 38.50.

a) Find the 95% parametric bootstrap confidence interval for μ assuming the normal distribution for the observations and $\sigma = 0.57$. Compare with the classical analytic approach based on the t-distribution.
N.B:Use B=1000 bootstrap samples (each sample hence consisting of 11 measurements).

Non-parametric Bootstrap

- Consider the situation:
  - We have a sample $\mathbf{x} = (x_1, ..., x_n)$ from an unknown distribution function F.
  - We wish to make inferences about a parameter $\theta = t(F)$ based on $\mathbf{x}$.
- Let $\hat{F}$ be the empirical distribution function, which assigns the probability $1/n$ to each observed value $x_i, i = 1, ..., n$.
- A bootstrap sample, $\mathbf{x}^* = (x_1^*, ..., x_n^*)$ is a random sample of size n (the size of the original sample) from $\hat{F}$, ie, is a sample of size n obtained, with replacement, from a population of n objects $x_1, ..., x_n$.


$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$
$\mathbf{x}^{*1} \longrightarrow T(\mathbf{x}^{*1})$
$\mathbf{x}^{*2} \longrightarrow T(\mathbf{x}^{*2})$
$\mathbf{x}^{*B} \longrightarrow T(\mathbf{x}^{*B})$
dataset
bootstrap samples
bootstrap replications

Non-parametric Bootstrap: bootstrap estimation

- Based on data $\mathbf{x}$ we obtain an estimate of θ, say $\hat{\theta} = t(\mathbf{x})$.
- Based on the bootstrap sample $\mathbf{x}^*$ we obtain a new estimate for θ, say $\hat{\theta}^* = t(\mathbf{x}^*)$
- Repeating the procedure m times, the bootstrap estimate of θ is

$$\hat{\theta}_B = \frac{1}{m} \sum_{i=1}^{m} t(\mathbf{x}_i^*) \ .$$

Non-parametric Bootstrap: estimation of the standard error

- Bootstrap can be used to estimate the standard error of $\hat{\theta}$, $se(\hat{\theta})$. It is given by the standard error of the bootstrap estimate of $\hat{\theta}$, ie,

$$\hat{se}_B(\hat{\theta}) = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}[\hat{\theta}_i^* - \hat{\theta}_B]^2}.$$

Exercise 4

To illustrate the bootstrap procedure, let's bootstrap a small random sample:

3.12 0.00 1.57 19.67 0.22 2.20

1. Create 1000 replicates of size 6 with replacement.
2. Calculate the sample mean for each of the replicates.
3. Make a histogram and a normal quantile plot of the 1000 means. Make the density plot of the 1000 replicates. This is the bootstrap distribution.
4. Calculate the bootstrap estimates of the mean and the standard error.

Confidence interval for a correlation coefficient **Exercise 5**

(Adapted from Applied Statistics for Bioinformatics using R, Wim P. Krijnen)

Consider two sets of expression values of the MCM3 gene of the Golub *et al.* (1999) data. This data set is a gene expression data (3051 genes and 38 tumor mRNA samples) from the leukemia microarray study. This gene encodes for highly conserved mini-chromosome maintenance proteins (MCM) which are involved in the initiation of eukaryotic genome replication.

```
source("https://bioconductor.org/biocLite.R")
biocLite("multtest")
library(multtest)
data(golub)
 x <- golub[2289,]; y <- golub[2430,]
 cor(x,y)
[1] 0.6376217
```

1. Obtain a bootstrap sample from (x, y), and compute the correlation coefficient for the bootstrap sample.
2. Repeat the procedure [1] B=1000 times.
3. From the sample of size $n = 38$ of the bootstraped correlation coefficients obtain the 0.025 and 0.975 percentiles.
4. This pair is a bootstrap 95% confidence interval for the correlation coefficient ρ.

Compare with the interval obtained using normality assumption for the sampling distribution of the empirical correlation coefficient.

Non-parametric Bootstrap: test of hypothesis

- Suppose X follows an unknown distribution F, with expected value μ. We want to perform the test

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

Non-parametric Bootstrap: test of hypothesis

- Suppose X follows an unknown distribution F, with expected value μ. We want to perform the test

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

- Let $X_1, ..., X_n$ be a random sample from X. Let $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$, be the test statistic where $\overline{X}$, is the sample mean an S the sample standard deviation.

Non-parametric Bootstrap: test of hypothesis

- Suppose X follows an unknown distribution F, with expected value μ. We want to perform the test

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

- Let $X_1, ..., X_n$ be a random sample from X. Let $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$, be the test statistic where $\overline{X}$, is the sample mean an S the sample standard deviation.
- Let $\mathbf{x} = (x_1, ..., x_n)$ an observed sample and t_{obs} the observed value of the test statistic.

Non-parametric Bootstrap: test of hypothesis

- Suppose X follows an unknown distribution F, with expected value μ. We want to perform the test

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

- Let $X_1, ..., X_n$ be a random sample from X. Let $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$, be the test statistic where $\overline{X}$, is the sample mean an S the sample standard deviation.
- Let $\mathbf{x} = (x_1, ..., x_n)$ an observed sample and t_{obs} the observed value of the test statistic.
- Since we need an estimate of the sampling distribution of T under the null hypothesis, we cannot use the empirical distribution function $\hat{F}$ since it does not obey H_0.

Non-parametric Bootstrap: test of hypothesis

- One way of overcoming this problem is to apply a transformation to F such that the expected value is μ_0. This can be done if we consider the empirical distribution function $\mathbf{z} = (z_1, ..., z_n)$, where $z_i = x_i - \overline{x} + \mu_0, i = 1, ..., n$.

Non-parametric Bootstrap: test of hypothesis

- One way of overcoming this problem is to apply a transformation to F such that the expected value is μ_0. This can be done if we consider the empirical distribution function $\mathbf{z} = (z_1, ..., z_n)$, where $z_i = x_i - \overline{x} + \mu_0, i = 1, ..., n$.
- We obtain then m bootstrap samples from $\mathbf{z}$, $\mathbf{z}_1^*, ..., \mathbf{z}_m^*$, their averages and standard deviations and with them we build a sequence of statistics $t_i^*, i = 1, ..., m$ with

$$t_i^* = \frac{\overline{z}_i^* - \mu_0}{s_i^*/\sqrt{n}},$$

Non-parametric Bootstrap: test of hypothesis

- One way of overcoming this problem is to apply a transformation to F such that the expected value is μ_0. This can be done if we consider the empirical distribution function $\mathbf{z} = (z_1, ..., z_n)$, where $z_i = x_i - \overline{x} + \mu_0, i = 1, ..., n$.
- We obtain then m bootstrap samples from $\mathbf{z}$, $\mathbf{z}_1^*, ..., \mathbf{z}_m^*$, their averages and standard deviations and with them we build a sequence of statistics $t_i^*, i = 1, ..., m$ with

$$t_i^* = \frac{\overline{z}_i^* - \mu_0}{s_i^*/\sqrt{n}},$$

- The $p-$value, $p = 2P_{H_0}(T > |t_{obs}|)$ is then approximated by

$$p \approx \frac{\#\{i : |t_i^*| > |t_{obs}|\}}{m}.$$

Exercise 6: Gdf5 gene from the Golub et al. (1999) data.

(Adapted from Applied Statistics for Bioinformatics using R, Wim P. Krijnen)

- The corresponding expression values are contained in row 2058.
- A quick search through the NCBI site makes it likely that this gene is not directly related to leukemia.
- Hence, we may hypothesize that the population mean of the ALL expression values equals zero.
- Accordingly, we test $H_0 : \mu = 0$ vs. $H_1 : \mu > 0$.
- a t test gives a $p - value = 0.499$ and clearly H_0 is not rejected.
- How can we use bootstrap to test the present hypothesis?

Non-parametric Bootstrap: test of hypothesis - two samples

- Suppose we have $X_1 \sim F_1$, with expected value μ_1 and $X_2 \sim F_2$, with expected value μ_2, where X_1 is independent of X_2 and F_1 e F_2 are unknown.
- The objective is to test

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_1 : \mu_1 \neq \mu_2.$$

based on the test statistic

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}},$$

where $(\overline{X}_k, S_k^2, n_k), k = 1, 2$ are, respectively, the sample means, sample variances and sample sizes of the samples $\mathbf{X}_k, k = 1, 2$, from the two populations.

Non-parametric Bootstrap: test of hypothesis - two samples

- Since we need the distribution of T under H_0 and this hypothesis assumes only the equality of the mean values we have to resample from samples which obey this condition. For that we do as follows:
  1. Compute the combined mean of the two samples $\overline{x} = \frac{\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j}}{n_1 + n_2}$,
  2. Consider $z_{ki} = x_{ki} - \overline{x}_k + \overline{x}, i = 1, ..., n_k, k = 1, 2$.
  3. Obtain m bootstrap samples from $(z_{1i}, i = 1, ..., n_1)$ and m bootstrap samples from $(z_{2i}, i = 1, ..., n_2)$, and compute the respective sample means and sample variances. $(\overline{z}^*_{ki}, s^{2*}_{ki}), i = 1, ..., m, k = 1, 2$
  4. Obtain the test statistic for each bootstrap samples

$$t_i^* = \frac{\overline{z}_1^* - \overline{z}_2^*}{\sqrt{s_1^{2*}/n_1 + s_2^{2*}/n_2}},$$

Non-parametric Bootstrap: test of hypothesis - two samples

- An approximation for the $p-$value $p = 2P_{H_0}(T > |t_{obs}|)$ is

$$p \approx \frac{\#\{i : |t_i^*| > |t_{obs}|\}}{m}.$$

Exercise 7: gene CCND3 Cyclin D3

- Golub et al. (1999) argue that gene CCND3 Cyclin D3 plays an important role with respect to discriminating ALL from AML patients.
- We are then interested in testing the null hypothesis of equal means, ie, we want to test

$$H_0 : \mu_{ALL} = \mu_{AML} \qquad vs. H_1 : \mu_{ALL} \neq \mu_{AML}.$$

- A t test for the equality of means would give strong support for the rejection of H_0.
- How would you implement a bootstrap test for this problem?

Parametric vs non-parametric

- Simulation from empirical distribution is computationally cheap and easier to implement;
- simulation from the fitted model may be computationally expensive and/or difficult to implement;

Anyway, one of the main motivations of the bootstrap was a desire to shake off the restrictions of conventional parametric modelling.

- Acknowledgements:
  Antónia Turkman (DEIO-FCUL), for allowing the use of some material produced by us in previous courses.
- A good book for introducing bootstrap methods: Efron, B and Tibshirani (1993) An Introduction to the Bootstrap, New York: Chapman and Hall.