Course 02402 Introduction to Statistics Lecture 7:

Statistics - Simulation based

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

1. Introduction to simulation - what is it really?
   - Example, Area of plates
2. Propagation of error
3. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals

Oversigt

1. Introduction to simulation - what is it really?
   - Example, Area of plates
2. Propagation of error
3. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals

Motivation

- Many (most?) relevant statistics("computed features") have complicated sampling distributions:
  - A trimmed mean
  - The median
  - Quantiles in general, i.e. IQR$= Q_3 - Q_1$
  - The coefficient of variation
  - ANY non-linear function of one or more input variables
  - (The standard deviation)
- The data distribution itself may be non-normal, complicating the statistical theory for even the simple mean
- We may HOPE for the magic of CLT (Central Limit Theorem)
- BUT but: We NEVER really no whether CLT is good enough - simulation can tell us!!
- Require : Use of computer - R is a super tool for this!

What is simulation really?

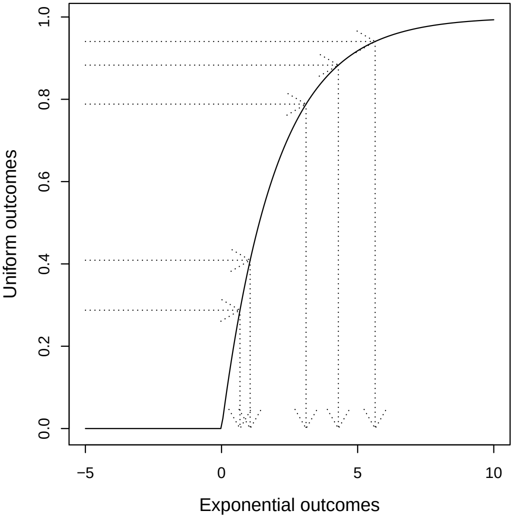
- (Pseudo) random numbers generated from a computer
- A random number generator is an algorithm that can generate x_{i+1} from x_i
- A sequence of numbers appears random
- Require a "start"called a "seed"(Using e.g. the computer clock)
- Basically the uniform distribution is simulated in this way, and then:

Theorem 2.51: All distributions can be extracted from the uniform

If $U \sim \text{Uniform}(0,1)$ and F is a distribution function for any probability distribution, then $F^{-1}(U)$ follow the distribution given by F

Example: the exponential distribution, $\lambda = 0.5$:

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}$$



In practice in R

Most distributions are ready for simulation, for instance:

`rbinom`	Binomial distribution
`rpois`	Poisson distribution
`rhyper`	The hypergeometric distribution
`rnorm`	normal distribution
`rlnorm`	log-normal distributions
`rexp`	exponential
`runif`	The uniform distribution
`rt`	t-distribution
`rchisq`	χ^2-distribution
`rf`	F distribution

Example: Area of plates

A company produces rectangular plates. The length of plates (in meters), X is assumed to follow a normal distribution $N(2, 0.01^2)$ and the width of the plates (in meters), Y are assumed to follow a normal distribution $N(3, 0.02^2)$. We are interested in the area of the plates which of course is given by $A = XY$.

- What is the mean area?
- What is the standard deviation in the areas from plate to plate?
- how often such plates have an area that differ by more than $0.1m^2$ from the targeted $6m^2$?
- The probability of other events?
- Generally: what is the probability distribution of the random variable A

Example: Area of plates, Solution by simulation

```
set.seed(345)
k = 10000 # Number of simulations
X = rnorm(k, 2, 0.01)
Y = rnorm(k, 3, 0.02)
A = X*Y
```

```
mean(A)

## [1] 6

sd(A)

## [1] 0.04957

mean(abs(A-6)>0.1)

## [1] 0.0439
```

Oversigt

1. Introduction to simulation - what is it really?
   - Example, Area of plates
2. Propagation of error
3. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals

Propagation of error

Must be able to find:

$$\sigma^2_{f(X_1,\ldots,X_n)} = \mathsf{Var}(f(X_1,\ldots,X_n))$$

We already know:

$$\sigma^2_{f(X_1,\ldots,X_n)} = \sum_{i=1}^{n} a_i^2 \sigma_i^2, \text{ if } f(X_1,\ldots,X_n) = \sum_{i=1}^{n} a_i X_i$$

Method 4.3: for non-linear functions:

$$\sigma^2_{f(X_1,\ldots,X_n)} \approx \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2$$

Example, cont.

We already used the simulation method in the first part of the example. Given two specific measurements of X and Y, $X = 2.00m$ and $y = 3.00m$. What is the variance of $A = 2.00 \times 3.00 =$ 6.00 using the error propagation law?

Example, cont.

The varianses are:

$$\sigma_1^2 = Var(X) = 0.01^2 \text{ og } \sigma_2^2 = Var(Y) = 0.02^2$$

The function andn the derivarives are:

$$f(x,y) = xy,\ \frac{\partial f}{\partial x} = y,\ \frac{\partial f}{\partial y} = x$$

So the result becomes:

$$\begin{aligned} Var(A) &\approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_2^2 \\ &= y^2\sigma_1^2 + x^2\sigma_2^2 \\ &= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2 \\ &= 0.0025 \end{aligned}$$

Propagation of error - by simulation

Method 4.4: Error propagation by simulation

Assume we have actual measurements $x_1, \ldots, x_n$ with known/assumed error variances $\sigma_1^2, \ldots, \sigma_n^2$.

1. Simulate k outcomes of all n measurements from assumed error distributions, e.g. $N(x_i, \sigma_i^2)$: $X_i^{(j)}, j = 1 \ldots, k$
2. Calculate the standard deviation directly as the observed standard deviation of the k simulated values of f:

$$s_{f(X_1,\ldots,X_n)}^{\text{sim}} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(f_j - \bar{f})^2}$$

where

$$f_j = f(X_1^{(j)}, \ldots, X_n^{(j)})$$

Example, Area, cont.

Actually one can deduce the variance of A theoretically,

$$\begin{aligned}
\mathsf{Var}(XY) &= \mathsf{E}\left[(XY)^2\right] - [\mathsf{E}(XY)]^2 \\
&= \mathsf{E}(X^2)\mathsf{E}(Y^2) - \mathsf{E}(X)^2\mathsf{E}(Y)^2 \\
&= \left[\mathsf{Var}(X) + \mathsf{E}(X)^2\right]\left[\mathsf{Var}(Y) + \mathsf{E}(Y)^2\right] - \mathsf{E}(X)^2\mathsf{E}(Y)^2 \\
&= \mathsf{Var}(X)\mathsf{Var}(Y) + \mathsf{Var}(X)\mathsf{E}(Y)^2 + \mathsf{Var}(Y)\mathsf{E}(X)^2 \\
&= 0.01^2 \times 0.02^2 + 0.01^2 \times 3^2 + 0.02^2 \times 2^2 \\
&= 0.00000004 + 0.0009 + 0.0016 \\
&= 0.00250004
\end{aligned}$$

Example, Area, cont. - in summary

Three different approaches:

1. The simulation based approach
2. A theoretical derivation
3. The analytical, but approximate, error propagation method

The simulation approach has a number of crucial advantages:

1. It offers a simple tool to compute many other quantities than just the standard deviation (the theoretical derivations of such other quantities could be much more complicated than what was shown for the variance here)
2. It offers a simple tool to use any other distribution than the normal, if we believe such better reflect reality.
3. It does not rely on any linear approximations of the true non-linear

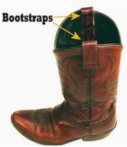
Oversigt

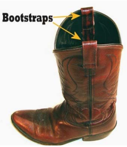
1. Introduction to simulation - what is it really?
   - Example, Area of plates
2. Propagation of error
3. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals

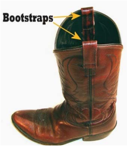
Bootstrapping

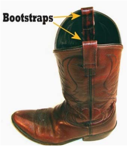
Bootstrapping exists in two versions:

1. Parametric bootstrap: Simulate multiple samples from the assumed (and estimated) distribution.
2. Non-parametric bootstrap: Simulate multiple samples directly from the data.









Example: Confidence interval for the exponential rate or mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

$$32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0$$

From the data we estimate

$$\hat{\mu} = \bar{x} = 26.08 \text{ and hence: } \hat{\lambda} = 1/26.08 = 0.03834356$$

Our distributional assumption:

The waiting times come from an exponential distribution

What is the confidence interval for μ?

Based on previous knowledge in this course: We don't know!

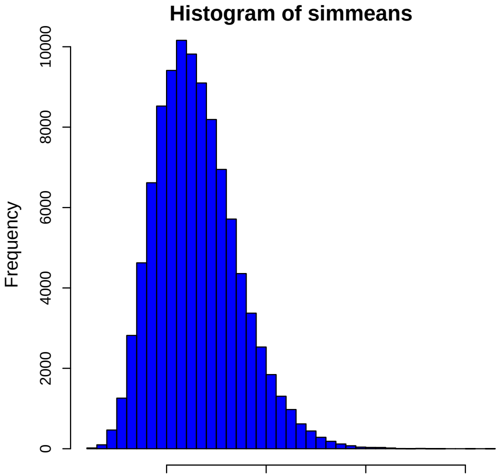
Example: Confidence interval for the exponential rate or mean

```
## Set the number of simulations:
k <- 100000
## 1. Simulate 10 exponentials with the right mean k times:
set.seed(9876.543)
simsamples <- replicate(k, rexp(10, 1/26.08))
## 2. Compute the mean of the 10 simulated observations k time
simmeans <- apply(simsamples, 2, mean)
## 3. Find the two relevant quantiles of the k simulated means
quantile(simmeans, c(0.025, 0.975))

##  2.5% 97.5%
## 12.59 44.63
```

Example: Confidence interval for the exponential rate or mean

```
hist(simmeans, col="blue", nclass=30)
```



Example: Confidence interval for the median of an exponential distribution

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

$$32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0$$

From the data we estimate

$$\text{Median } = 21.4 \text{ and } \hat{\mu} = \bar{x} = 26.08$$

Our distributional assumption:

The waiting times come from a an exponential distribution

What is the confidence interval for the median?

Based on previous knowledge in this course: We don't know!

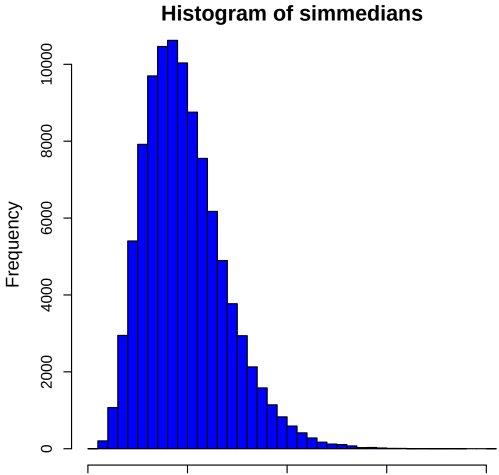
Example: Confidence interval for the median of an exponential

```
## Set the number of simulations:
k <- 100000
## 1. Simulate 10 exponentials with the right mean k times:
set.seed(9876.543)
simsamples <- replicate(k, rexp(10, 1/26.08))
## 2. Compute the median of the n=1010 simulated observations
simmedians <- apply(simsamples, 2, median)
## 3. Find the two relevant quantiles of the k simulated media
quantile(simmedians, c(0.025, 0.975))

##   2.5%  97.5%
##  7.038 38.465
```

Example: Confidence interval for the median of an exponential

```
hist(simmedians, col="blue", nclass=30)
```



Confidence interval for any feature (including μ)

Method 4.7: Confidence interval for any feature θ by parametric bootstrap

Assume we have actual observations $x_1, \ldots, x_n$ and assume that they stem from some probability distribution with density f.

1. Simulate k samples of n observations from the assumed distribution f where the mean [a] is set to $\bar{x}$.
2. Calculate the statistic $\hat{\theta}$ in each of the k samples $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%},\, q^*_{100(1-\alpha/2)\%}\right]$

[a] And otherwise chosen to match the data as good as possible: Some distributions have more than just a single mean related parameter, e.g. the normal or the log-normal. For these one should use a distribution with a variance that matches the sample variance of the data. Even more generally the approach would be to match the chosen distribution to the data by the so-called *maximum likelihood* approach

Another example: 99% confidence interval for Q_3 assuming a normal distribution

```
## Read in the heights data:
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
n <- length(x)
## Define a Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
 ## Set the number of simulations:
k <- 100000
## 1. Simulate k samples of n=10 normals with the right mean and variance:
set.seed(9876.543)
simsamples <- replicate(k, rnorm(n, mean(x), sd(x)))
## 2. Compute the Q3 of the n=10 simulated observations k times:
simQ3s <- apply(simsamples, 2, Q3)
## 3. Find the two relevant quantiles of the k simulated medians:
quantile(simQ3s, c(0.005, 0.995))

##  0.5% 99.5%
## 172.8 198.0
```

Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ (including $\mu_1 - \mu_2$)

Method 4.10: Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ by parametric bootstrap

Assume we have actual observations $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ and assume that they stem from some probability distributions with density f_1 and f_2.

1. Simulate k sets of 2 samples of n_1 and n_2 observations from the assumed distributions setting the means [a] to $\hat{\mu}_1 = \bar{x}$ and $\hat{\mu}_2 = \bar{y}$, respectively.
2. Calculate the difference between the features in each of the k samples $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%},\, q^*_{100(1-\alpha/2)\%}\right]$

Example: Confidence interval for the difference of exponential means

```
## Day 1 data:
x <- c(32.6, 1.6, 42.1, 29.2, 53.4, 79.3,
       2.3 , 4.7, 13.6, 2.0)
## Day 2 data:
y <- c(9.6, 22.2, 52.5, 12.6, 33.0, 15.2,
       76.6, 36.3, 110.2, 18.0, 62.4, 10.3)
n1 <- length(x)
n2 <- length(y)
```

Example: Confidence interval for the difference of exponential means

```
## Set the number of simulations:
k <- 100000
## 1. Simulate k samples of each n1=10 and n2=12
## exponentials with the right means:
set.seed(9876.543)
simXsamples <- replicate(k, rexp(n1, 1/mean(x)))
simYsamples <- replicate(k, rexp(n2, 1/mean(y)))
## 2. Compute the difference between the simulated
## means k times:
simDifmeans <- apply(simXsamples, 2, mean) -
                    apply(simYsamples, 2, mean)
## 3. Find the two relevant quantiles of the
## k simulated differences of means:
quantile(simDifmeans, c(0.025, 0.975))

##   2.5%  97.5%
## -40.74  14.12
```

Parametric bootstrap - an overview

We assume SOME distribution!

Two confidence interval method boxes were given:

	One-sample	Two-sample
For any feature	Method 4.7	Method 4.10

Oversigt

1. Introduction to simulation - what is it really?
   - Example, Area of plates
2. Propagation of error
3. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals

Non-parametric bootstrap - an overview

We do NOT assume ANY distribution!

Two confidence interval method boxes will be given:

	One-sample	Two-sample
For any feature	Method 4.15	Method 4.17

Example: Women's cigarette consumption

In a study women's cigarette consumption before and after giving birth is explored. The following observations of the number of smoked cigarettes per day were the results:

before	after	before	after
8	5	13	15
24	11	15	19
7	0	11	12
20	15	22	0
6	0	15	6
20	20		

Compare the before and after means! (Are they different?)

Example: Women's cigarette consumption

A paired t-test setting, BUT with clearly non-normal data!

```
x1 <-  c(8, 24, 7, 20, 6, 20, 13, 15, 11, 22, 15)
x2 <-  c(5, 11, 0, 15, 0, 20, 15, 19, 12, 0, 6)
dif <- x1-x2
dif

##  [1]  3 13  7  5  6  0 -2 -4 -1 22  9

mean(dif)

## [1] 5.273
```

Example: Women's cigarette consumption - bootstrapping

```
t(replicate(5, sample(dif, replace = TRUE)))

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## [1,]    3    6    0    9    3    9   -4    0    0    -1     6
## [2,]   -1    9    5    5    6    9    3   13    3    22    22
## [3,]   -4   -2    3   -1    3   -1    7    3    9     6     0
## [4,]    6    3   -4    9    3   22    3   -1   -1    -4     7
## [5,]   13    0    5   22    0    9    9    5    0    22    -1
```

Example: Women's cigarette consumption - the non-parametric results:

```
k = 100000

simsamples = replicate(k, sample(dif, replace = TRUE))
simmeans = apply(simsamples, 2, mean)
quantile(simmeans, c(0.025,0.975))

##  2.5% 97.5%
## 1.364 9.818
```

One-sample confidence interval for any feature θ (including μ)

Method 4.15: Confidence interval for any feature θ by non-parametric bootstrap

Assume we have actual observations $x_1, \ldots, x_n$.

1. Simulate k samples of size n by randomly sampling among the available data (with replacement)
2. Calculate the statistic $\hat{\theta}$ in each of the k samples $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%},\, q^*_{100(1-\alpha/2)\%}\right]$

Example: Women's cigarette consumption

Let us find the 95% confidence interval for the median cigarette consumption change in the example from above:

```
k = 100000
simsamples = replicate(k, sample(dif, replace = TRUE))
simmedians = apply(simsamples, 2, median)
quantile(simmedians, c(0.025,0.975))

##  2.5% 97.5%
##    -1     9
```

Example: Tooth health and infant bottle use

In a study it was explored whether children who received milk from bottle as a child had worse or better teeth health conditions than those who had not received milk from the bottle. For 19 randomly selected children is was recorded when they had their first incident of caries:

bottle	age	bottle	age	bottle	Age
no	9	no	10	yes	16
yes	14	no	8	yes	14
yes	15	no	6	yes	9
no	10	yes	12	no	12
no	12	yes	13	yes	12
no	6	no	20		
yes	19	yes	13		

Example: Tooth health and infant bottle use - a 95% confidence interval for $\mu_1 - \mu_2$

```
## Reading in no group:
 x <- c(9, 10, 12, 6, 10, 8, 6, 20, 12)
## Reading in yes group:
 y <- c(14,15,19,12,13,13,16,14,9,12)

k <- 100000
simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
simmeandifs <- apply(simxsamples, 2, mean)-
                        apply(simysamples, 2, mean)
quantile(simmeandifs, c(0.025,0.975))

##    2.5%   97.5%
## -6.2333 -0.1444
```

Two-sample confidence interval for $\theta_1 - \theta_2$ (including $\mu_1 - \mu_2$) by non-parametric bootstrap

Method 4.17: Two-sample confidence interval for $\theta_1 - \theta_2$ by non-parametric bootstrap

Assume we have actual observations $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$.

1. Simulate k sets of 2 samples of n_1 and n_2 observations from the respective groups (with replacement)
2. Calculate the difference between the features in each of the k samples $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%}, q^*_{100(1-\alpha/2)\%}\right]$

Example: Tooth health and infant bottle use - a 99% confidence interval for the difference of medians

```
k <- 100000
simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
simmediandifs <- apply(simxsamples, 2, median)-
                        apply(simysamples, 2, median)
quantile(simmediandifs, c(0.005,0.995))

##  0.5% 99.5%
##    -8     0
```

Bootstrapping - an overview

We were given 4 similar method boxes

1. With distribution or not (parametric or non-parametric)
2. For one- or two-sample analysis

Note:

Means also included in *other features*. Or: These methods can be used not only for means!!

Hypothesis testing also possible

We can do hypothesis testing by looking at the confidence intervals!

Agenda

1. Introduction to simulation - what is it really?
   - Example, Area of plates
2. Propagation of error
3. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals