LUXEMBOURG INSTITUTE OF HEALTH

Multiomics Data Science Group (MODAS)
Department of Cancer Research, LIH

Bioinformatics Platform (BIOINFO)
Department of Medical Informatics, LIH

BIOSTATISTICS for PhDs

Lecture 1

Descriptive Statistics, Distributions, Sampling

Peter Nazarov

Email: petr.nazarov@lih.lu
Skype: pvn.public

05-02-2024

http://edu.modas.lu

COURSE OVERVIEW

Outline (*to be updated during the course*)

- **Lecture 1, 2024-02-05**
  - numerical measures (location/variability/association), parametric/nonparametric
  - basic summary and visualization in R: barplot, boxplot, scatter plot
  - z-score, detection of outliers
  - continuous distributions (normal, Student, χ^2, F), linkage to probability
  - sampling distribution, methods for sampling

https://cran.r-project.org/ https://posit.co/downloads/

- **Lecture 2, 2024-02-19**
  - interval estimations for mean and proportion
  - hypotheses testing for mean(s), p-value, tails
  - number of samples
  - power of a test
  - multiple comparisons

- **Lecture 3, 2024-03-04**
  - interval estimations and hypotheses for variance
  - model fitting and test for independence
  - linear models, ANOVA, posthoc analysis
  - simple and multiple linear regression
  - factors in linear regression
  - logistic regression

- **Lecture 4, 2024-03-18** *(please, propose!)*
  - omics data analysis?
  - survival analysis?
  - clustering?
  - more practical exercise?

Let's work at a comfortable speed!

Materials and other courses:

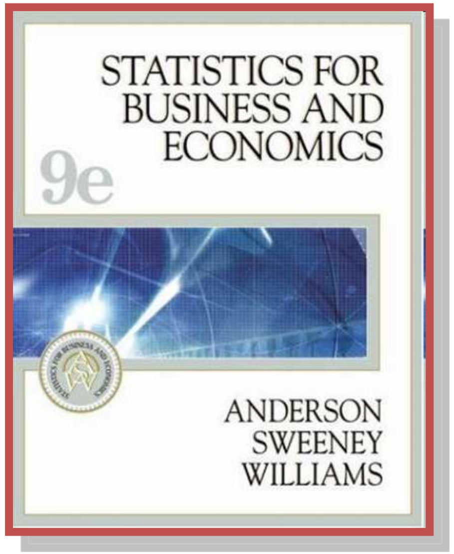
http://edu.modas.lu

LUXEMBOURG INSTITUTE OF HEALTH

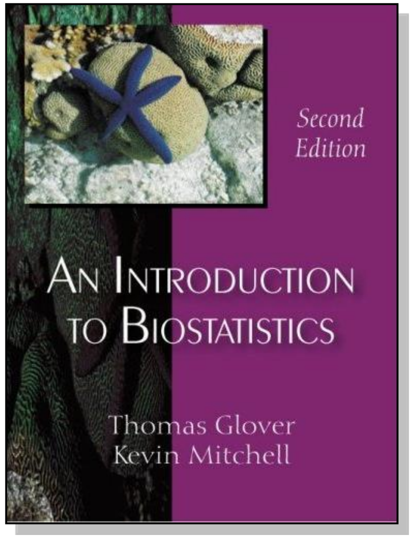
COURSE OVERVIEW

Recommended Literature

presentation methodology

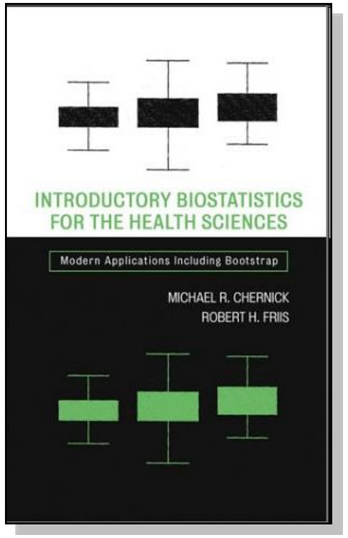


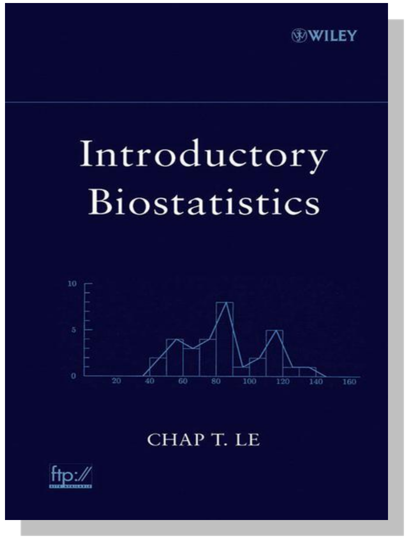
https://nibmehub.com/opac-service/pdf/read/Statistics%20for%20business%20and%20economics-%20%20Anderson-%20D.R..pdf



https://cran.r-project.org/

WIKIPEDIA
The Free Encyclopedia





ChatGPT

LUXEMBOURG INSTITUTE OF HEALTH

NUMERICAL MEASURES

Population and sample

Measures of location and variability

Parametric and non-parametric measures

Quantiles, quartiles and percentiles

Covariation, correlation

Exploratory data analysis

z-score, detection of outliers

NUMERICAL MEASURES

Population and Sample

Population parameter
A numerical value used as a summary measure for a population of size ***N*** (e.g., the population mean μ, variance σ^2, standard deviation σ)

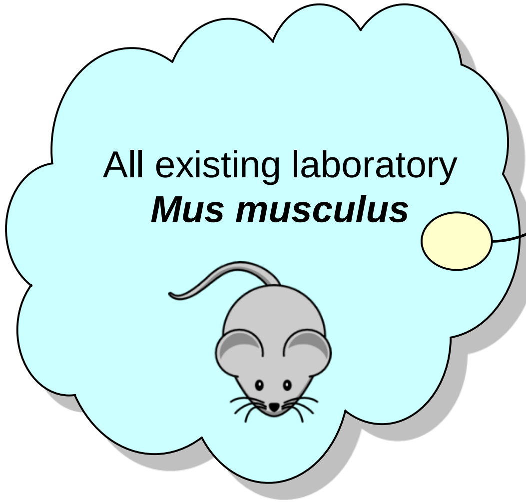
POPULATION

μ – mean
σ^2 – variance
N – number of elements (usually $N \to \infty$)

SAMPLE

m, $\bar{x}$ – mean
s^2 – variance
n – number of elements

Sample statistic (parameter)
A numerical value used as a summary measure for a sample of size ***n*** (e.g., the sample mean ***m***, the sample variance s^2, and the sample standard deviation ***s***)



`mice`

790 mice from different strains
http://phenome.jax.org

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvImJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvImJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvImJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvImJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvImJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvImJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvImJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvImJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvImJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvImJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvImJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvImJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvImJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvImJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvImJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvImJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvImJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvImJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvImJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

Load the data:

```
Mice = read.table("http://edu.modas.lu/data/txt/mice.txt", sep="\t", header=TRUE, stringsAsFactors = TRUE)
```

NUMERICAL MEASURES

Measures of Location

Mean
A measure of central location computed by summing the data values and dividing by the number of observations.

Median
A stable measure of central location provided by the value in the middle when the data are arranged in ascending order.

Mode
A measure of location, defined as the value that occurs with greatest frequency.

sample mean

$$\bar{x} = m = \frac{\sum x_i}{n}$$

population mean

$$\mu = \frac{\sum x_i}{N}$$

proportion

$$p = \frac{\sum (x_i = true)}{n}$$

Weight
12
16
19
22
23
23
24
32
36
42
63
68

Mode = 23

Median = 23.5

Mean = 31.7

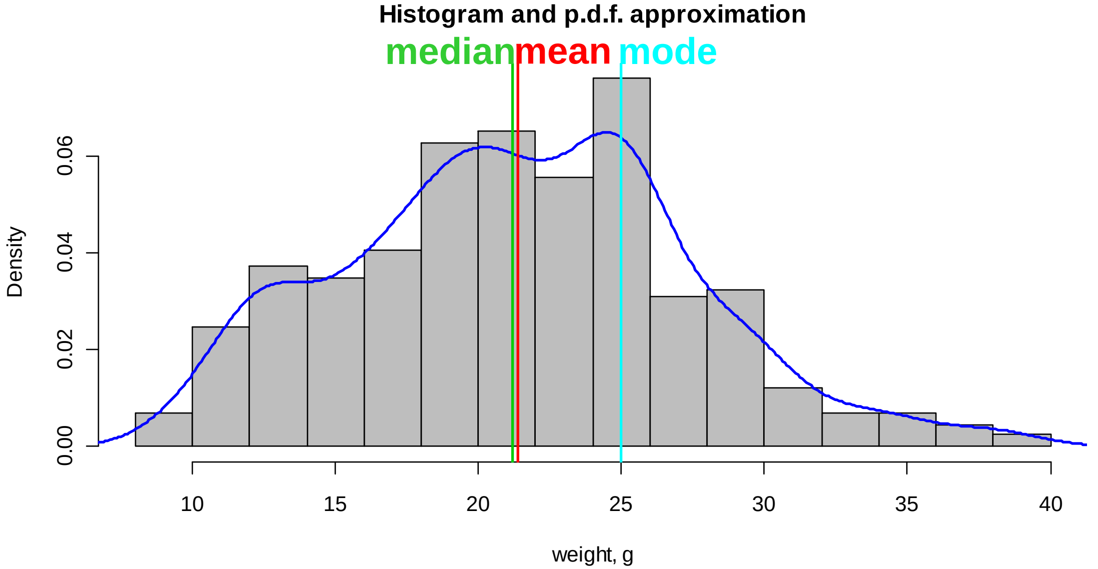
NUMERICAL MEASURES

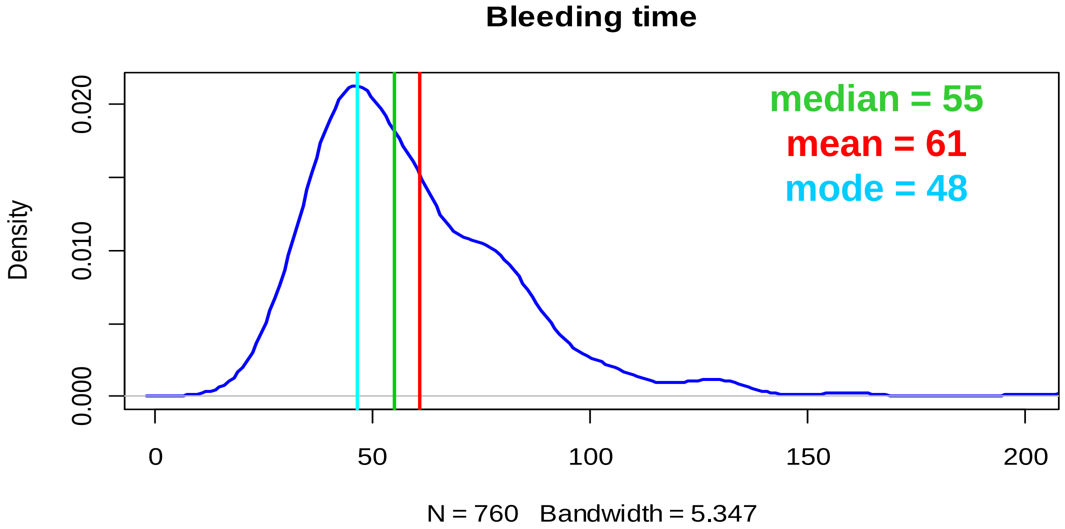
Measures of Location

mice

Female proportion

$p_f = 0.501$





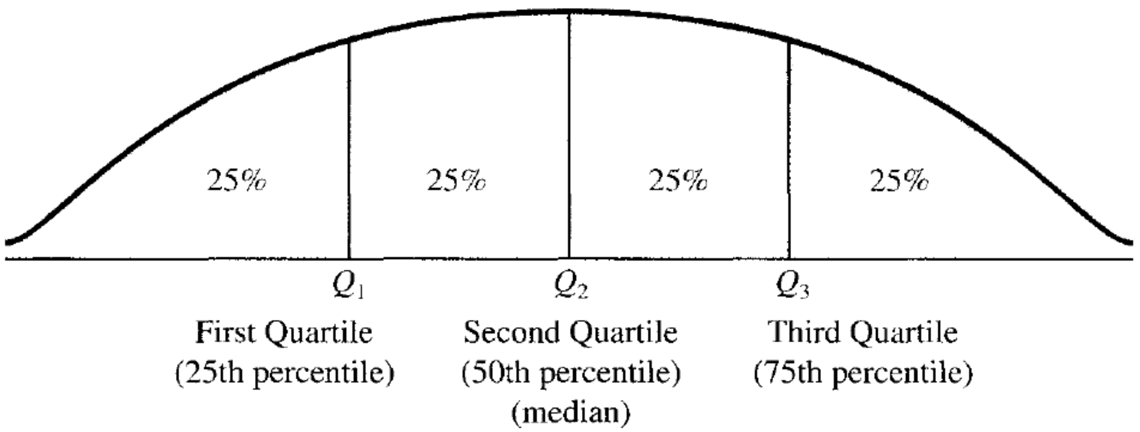

```
x = Mice$Bleeding.time
# measures of location
mean(x, na.rm=TRUE)
median(x , na.rm=TRUE)

# we need a package `modeest`
library(modeest)
mlv(x, na.rm=TRUE)

# show distribution
plot(density(x,na.rm=TRUE))
```

NUMERICAL MEASURES

Quantiles, Quartiles and Percentiles

Percentile
A value such that at least p% of the observations are less than or equal to this value, and at least (100-p)% of the observations are greater than or equal to this value. The 50-th percentile is the ***median***.

Quartiles
The 25th, 50th, and 75th percentiles, referred to as the first quartile, the second quartile (median), and third quartile, respectively.




```
# define your data
x = c(12, 16, 19, 22, 23, 23, 24, 32, 36, 42, 63, 68)
# overview Q1, Q2, Q3
quantile(x)
# calculate 1st quartile
quantile(x, 0.25)
```

Weight	12	16	19	22	23	23	24	32	36	42	63	68

$Q_1 = 21$ $Q_2 = 23.5$ $Q_3 = 39$

NUMERICAL MEASURES

Measures of Variability

Interquartile range (IQR)
A robust non-parametric measure of variability, defined to be the difference between the third and first quartiles.

$$IQR = Q_3 - Q_1$$

Variance
A measure of variability based on the squared deviations of the data values about the mean.

population $$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

sample $$s^2 = \frac{\sum (x_i - m)^2}{n-1}$$

Standard deviation
A measure of variability computed by taking the square root of the variance.

$$Sample\ standard\ deviation = s = \sqrt{s^2}$$

$$Population\ standard\ deviation = \sigma = \sqrt{\sigma^2}$$

Weight	12	16	19	22	23	23	24	32	36	42	63	68

IQR = 18 *Variance* = 320.2 *St. dev.* = 17.9

```
# Measures of variability
var(x)
sd(x)
IQR(x)
```

NUMERICAL MEASURES

Measures of Variability

Coefficient of variation
A measure of relative variability computed by dividing the standard deviation by the mean.

Weight	12	16	19	22	23	23	24	32	36	42	63	68

$$\left(\frac{Standard\ deviation}{Mean}\times 100\right)\%$$

$CV = 57\%$

Median absolute deviation (MAD)
MAD is a robust non-parametric measure of the variability of a univariate sample of quantitative data.

$$MAD = median\left(\left|x_i - median(x)\right|\right)$$

Set 1	Set 2
23	23
12	12
22	22
12	12
21	21
18	81
22	22
20	20
12	12
19	19
14	14
13	13
17	17

	Set 1	Set 2
Mean	17.3	22.2
Median	18	19
St.dev.	4.23	18.18
MAD	5.93	5.93

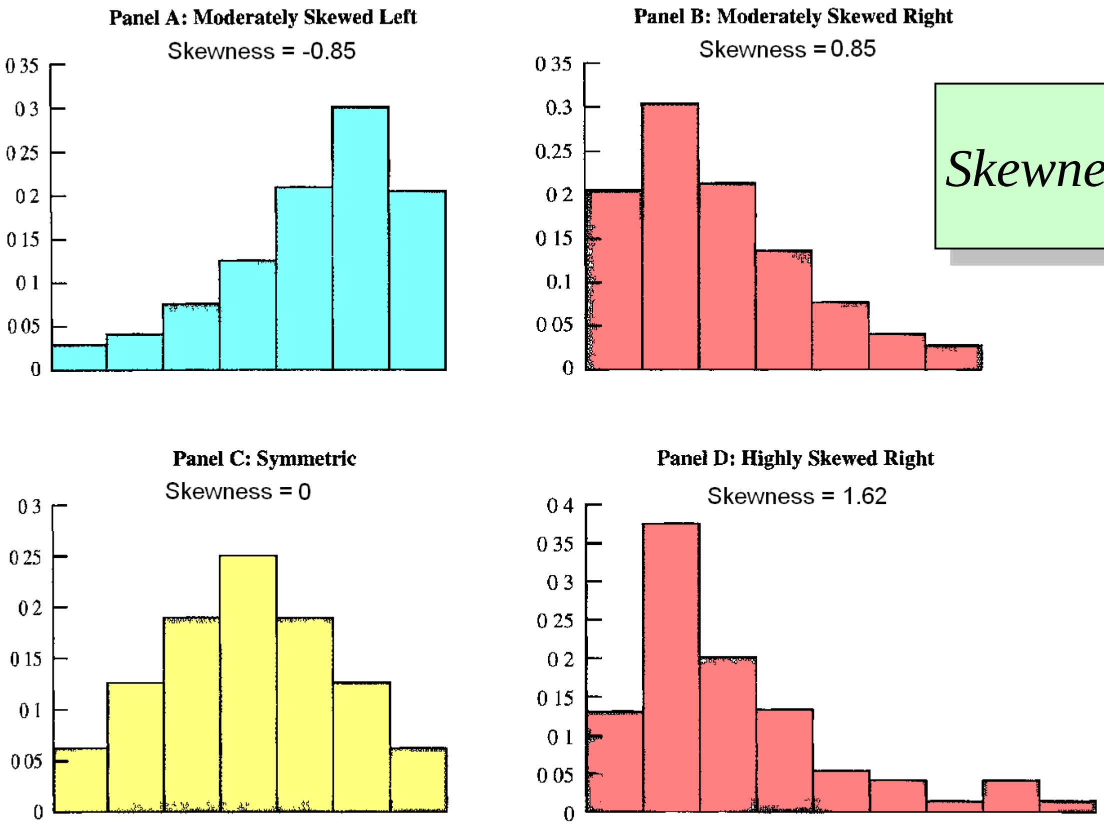
```
# define your data
x1 = c(23,12,22,12,21,18,22,20,12,19,
14,13,17)
x2 = c(23,12,22,12,21,81,22,20,12,19,
14,13,17)
# Parametric measures
mean(x1); mean(x2)
var(x1); var(x2)
sd(x1); sd(x2)
# Non-parametric measures
median(x1); median(x2)
mad(x1); mad(x2)
IQR(x1); IQR(x2)
```

LUXEMBOURG INSTITUTE OF HEALTH

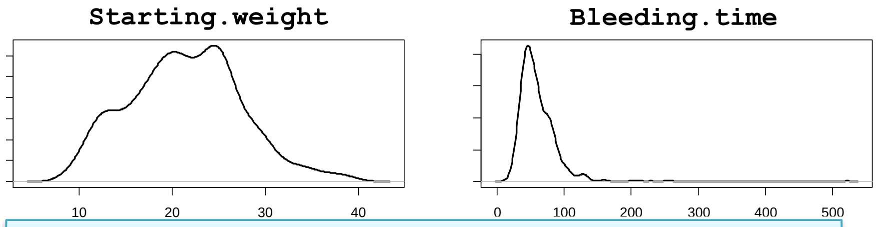
NUMERICAL MEASURES

Skewness (3rd central moment)

Skewness
A measure of the shape of a data distribution. Data skewed to the left result in negative skewness; a symmetric data distribution results in zero skewness; and data skewed to the right result in positive skewness.



$$Skewness = \frac{n}{(n-1)(n-2)} \sum_i \left(\frac{x_i - m}{s} \right)^3$$




```
# we need package `e1071`
library(e1071)
# skewness
skewness(Mice$Starting.weight)
skewness(Mice$Bleeding.time,na.rm=TRUE)
```

adapted from Anderson et al Statistics for Business and Economics

NUMERICAL MEASURES

Measure of Association between 2 Variables

Covariance
A measure of linear association between two variables. Positive values indicate a positive relationship; negative values indicate a negative relationship.

population

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

sample

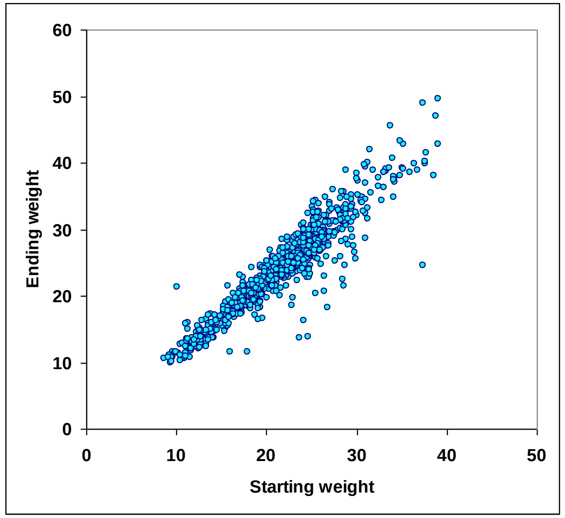
$$s_{xy} = \frac{\sum (x_i - m_x)(y_i - m_y)}{n-1}$$

```
# let's use variables (less typing)
x = Mice$Starting.weight
y = Mice$Ending.weight
# plot
plot(x, y, pch=19, col=4)
# covariance
cov(x,y)
```

For missing data add the parameter:
`use = "pairwise.complete.obs"`

mice.xls

Ending weight vs. Starting weight



s_{xy} = 39.8

hard to interpret

NUMERICAL MEASURES

Measure of Association between 2 Variables

Correlation (Pearson product moment correlation coefficient)

A measure of linear association between two variables that takes on values between -1 and +1. Values near +1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship; and values near zero indicate the lack of a linear relationship.

```
# correlation (Pearson)
cor(x, y)
# correlation (Spearman)
cor(x, y, method="spearman")
```

population

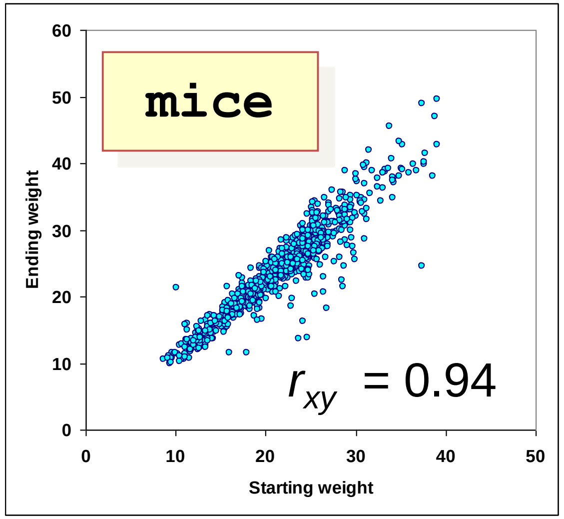
$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y N}$$

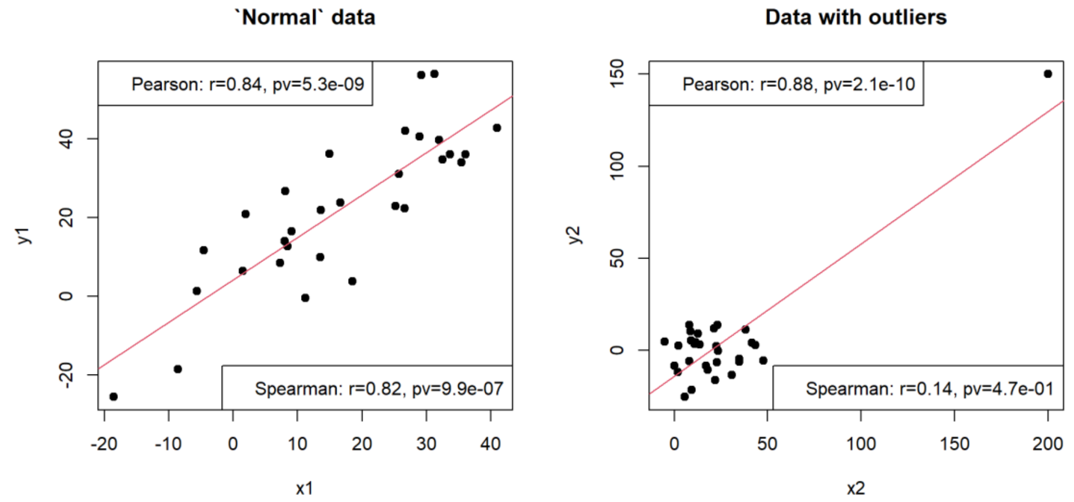
sample

$$\rho = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - m_x)(y_i - m_y)}{s_x s_y (n-1)}$$

Spearman Correlation

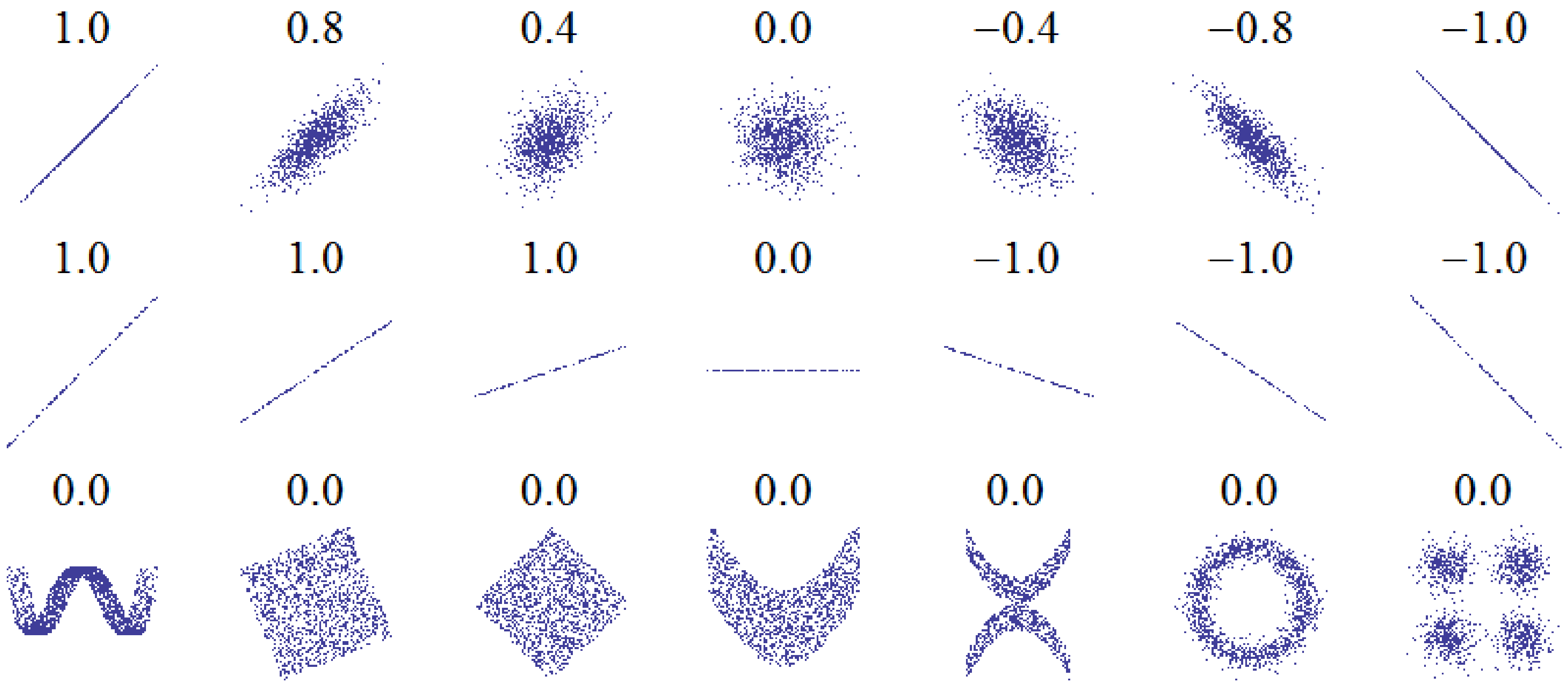
Non-parametric stable measure of association, equal to Pearson correlation between ranks





NUMERICAL MEASURES

Pearson Coefficient

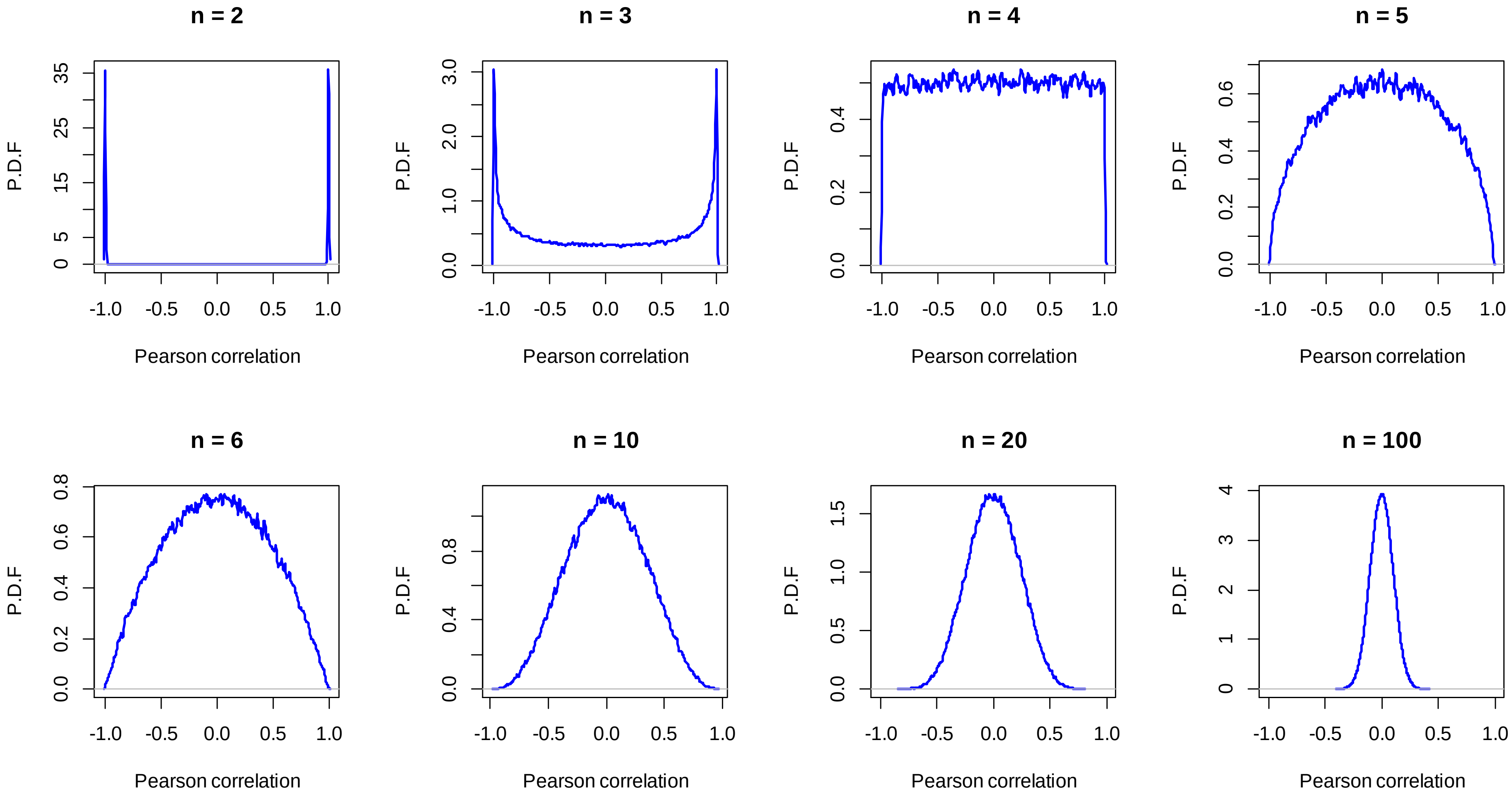


Wikipedia

? If we have only 2 data points in *x* and y datasets, what values would you expect for correlation b/w *x* and y ?

NUMERICAL MEASURES

Pearson Correlation: Effect of Sample Size



EXPLORATORY DATA ANALYSIS

Summarizing Data with Relative Frequency Distribution

pancreatitis

```
# load dataset

Panc = read.table(
"http://edu.modas.lu/data
/txt/pancreatitis.txt",
sep="\t", header=TRUE,
as.is = FALSE)

str(Panc)
```

Frequency distribution
A tabular summary of data showing the number (frequency) of items in each of several nonoverlapping classes.

Frequency distribution:

Smoking	Cases	Controls
Never	2	56
Ex-smokers	13	80
Smokers	38	81
Total	53	217

Relative frequency distribution
A tabular summary of data showing the fraction or proportion of data items in each of several nonoverlapping classes. Sum of all values should give 1

Relative frequency distribution:

Smoking	Cases	Controls
Never	0.038	0.258
Ex-smokers	0.245	0.369
Smokers	0.717	0.373
Total	1	1

Estimation of probability distribution
When number of experiments $n \to \infty$,
R.F.D. $\to$ P.D.

```
# frequency distribution (crosstabulation)
FD = table(Panc[,-1])
FD

# relative frequency distribution
RFD = prop.table(table(Panc[,-1]),2) # 2 - sum by columns
RFD
```

```
            Disease
Smoking      other pancreatitis
  Ex-smoker    80            13
  Never        56             2
  Smoker       81            38
```

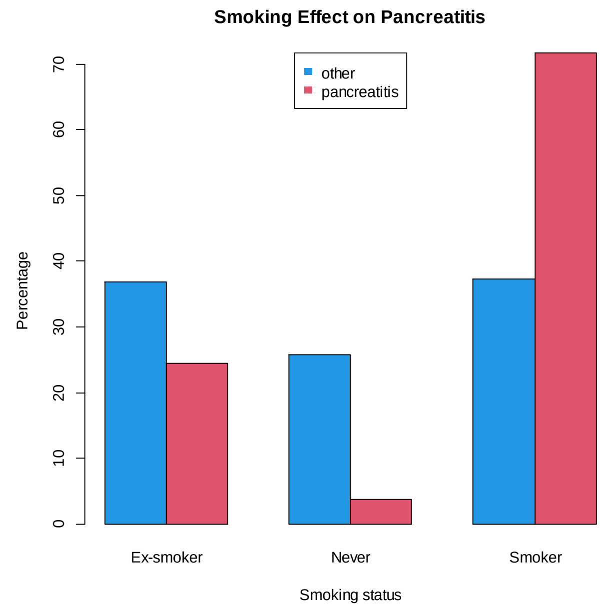
```
             Disease
Smoking            other pancreatitis
  Ex-smoker 0.36866359    0.24528302
  Never     0.25806452    0.03773585
  Smoker    0.37327189    0.71698113
```

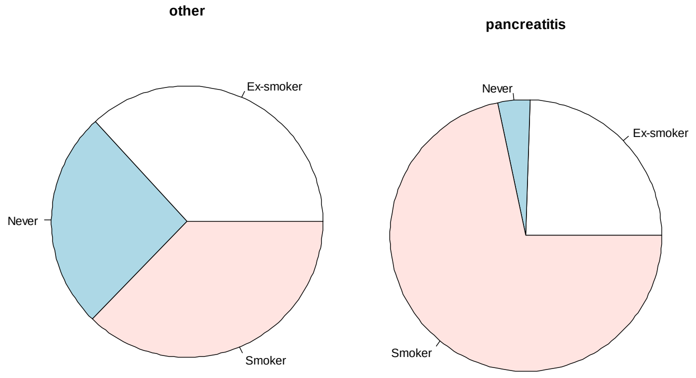
EXPLORATORY DATA ANALYSIS

Bar and Pie Charts

pancreatitis

```
# load dataset
Panc = read.table(
"http://edu.modas.lu/data/txt/pancreatitis.txt", sep="\t",
header=TRUE, as.is = FALSE)
str(Panc)
# crosstabulation
RFD = prop.table(table(Panc[,-1]),2)
# barplot
barplot(t(RFD), beside=TRUE)
# let's add some beauty :)
barplot(t(RFD)*100, beside=TRUE, col=c(4,2), main="Smoking Effect
on Pancreatitis", xlab="Smoking status", ylab="Percentage")
legend("top",colnames(RFD),col=c(4,2),pch=15)
# pies
par(mfcol=c(1,2)) # define 1x2 windows
pie(RFD[,1], main = colnames(RFD)[1])
pie(RFD[,2], main = colnames(RFD)[2])
```





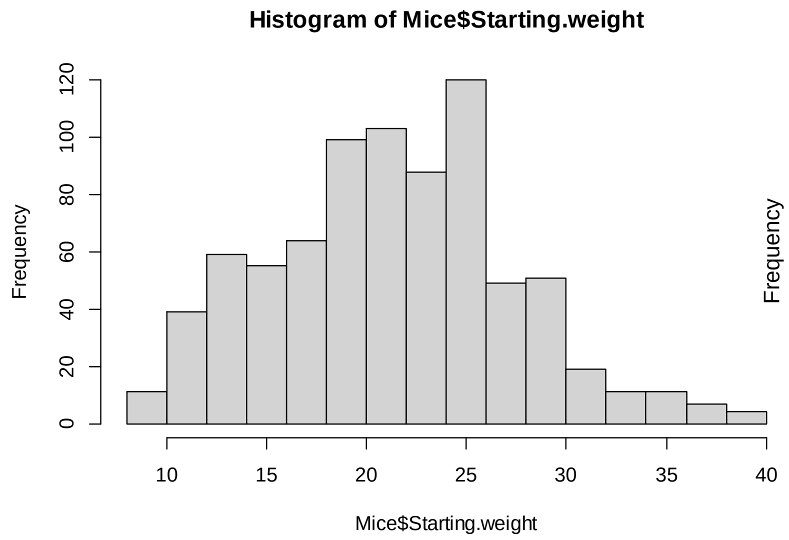
Try to avoid using pie-charts in scientific reports.
For public/business presentations only!

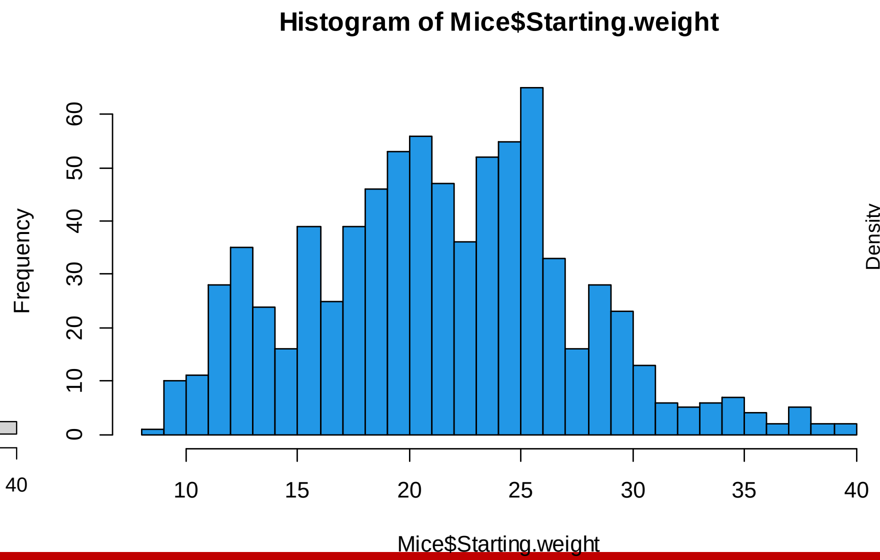
EXPLORATORY DATA ANALYSIS

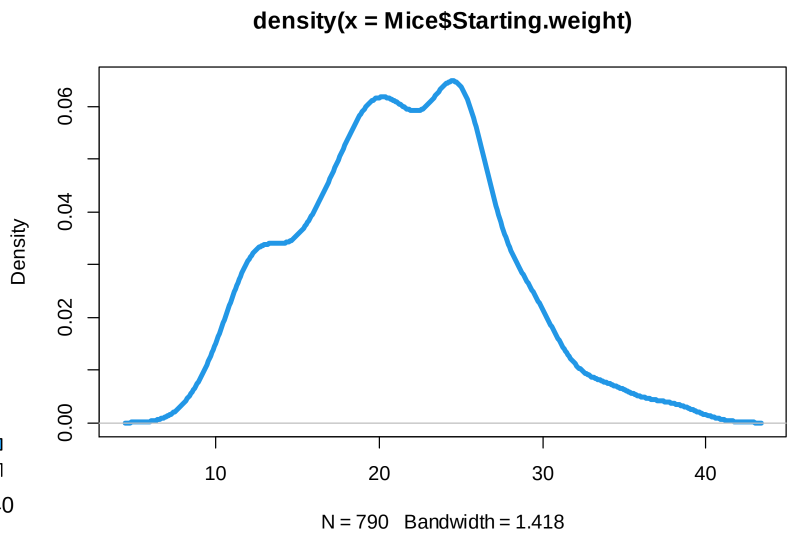
Histogram and Probability Density Function

```
# load dataset
Mice = read.table( "http://edu.modas.lu/data/txt/mice.txt", sep="\t", header=TRUE, as.is = FALSE)
str(Mice)
# histogram
hist(Mice$Starting.weight)
hist(Mice$Starting.weight, breaks = seq(8,40),col=4)
# Probability density function (convolution with a smooth kernel)
plot(density(Mice$Starting.weight), lwd=2, col=4)
```

- if data contain NA, use **`na.rm=TRUE`** parameter in `density()`
- play with kernel **`width`**
- **`cut=0`** to ensure value limits







LUXEMBOURG INSTITUTE OF HEALTH

EXPLORATORY DATA ANALYSIS

Box Plot

Five-number summary

An exploratory data analysis technique that uses five numbers to summarize the data: smallest value, first quartile, median, third quartile, and largest value

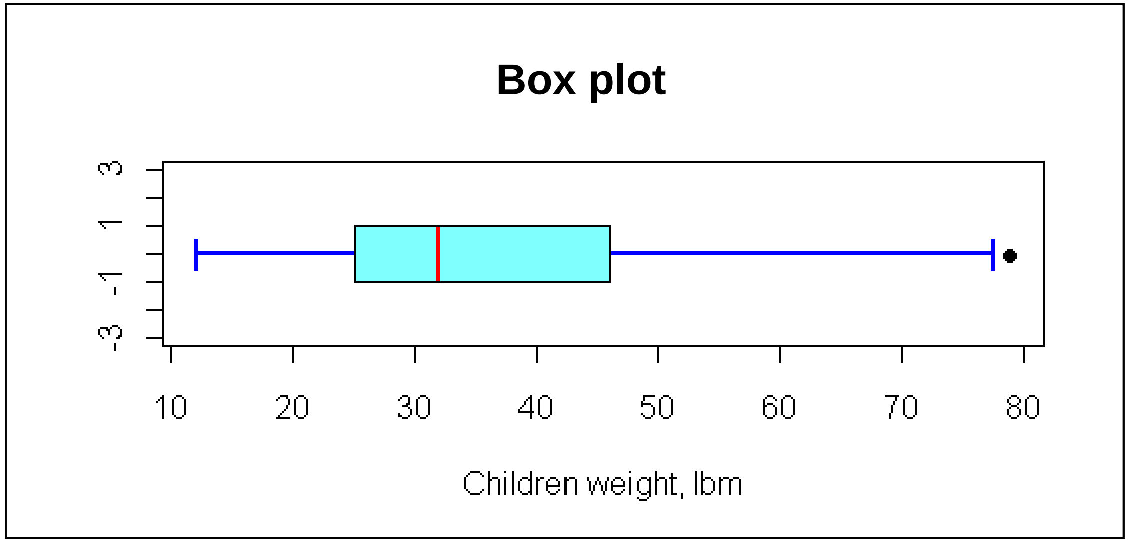
`children`

Min. :	12
Q_1 :	25
Median:	32
Q_3 :	46
Max. :	79

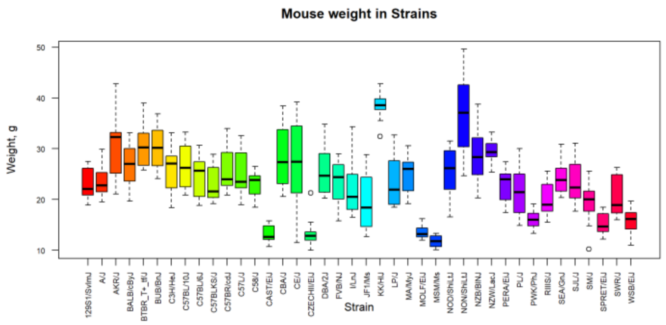
`boxplot(x)`

Box plot

A graphical summary of data based on a five-number summary




```r
## define colors for strains
col_strain =
  rainbow(nlevels(Mice$Strain))


## build boxplots
boxplot(Ending.weight ~ Strain,
  data = Mice,
  las = 2,
  col = col_strain,
  cex.axis = 0.7,
  ylab="Weight, g",
  main="Mouse weight ~ Strains")
```



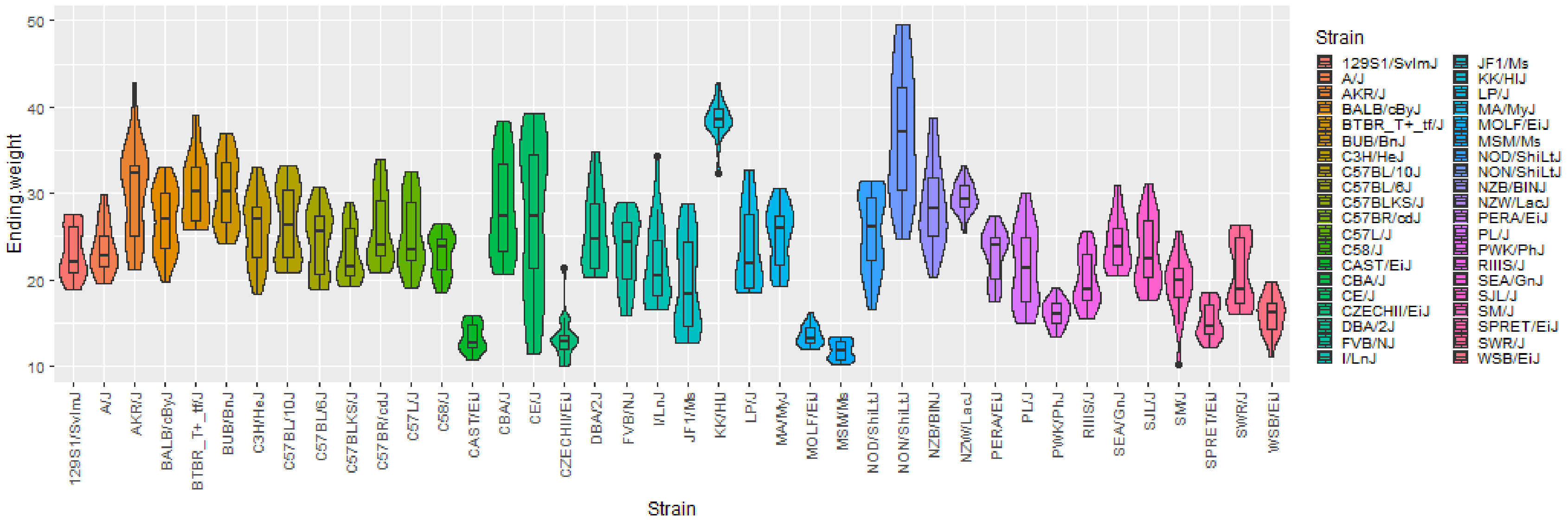
EXPLORATORY DATA ANALYSIS

Violin Plot

Violin plot

Violin plot is a more advanced visualization tool that shows the distribution of the data in categories

```
library(ggplot2)
p = ggplot(Mice, aes(x=Strain, y=Ending.weight, fill=Strain))
p = p + geom_violin(scale="width") + geom_boxplot(width=0.3)
p = p + theme_grey(base_size = 10)
p = p + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
p = p + theme(legend.key.size = unit(0.3, 'cm'))
print(p)
```



DETECTION OF OUTLIERS

z-score

z-score

This value is computed by dividing the deviation from the mean by the standard deviation *s*. A ***z-score*** is referred to as a standardized value and denotes the number of standard deviations x_i is from the mean.

$$z_i = \frac{x_i - m}{s}$$

```
# z-score

z = scale(x)
```

Weight	z-score
12	-1.10
16	-0.88
19	-0.71
22	-0.54
23	-0.48
23	-0.48
24	-0.43
32	0.02
36	0.24
42	0.58
63	1.75
68	2.03

Chebyshev's theorem

For **any data set**, at least **$(1 - 1/z^2)$** of the data values must be within z standard deviations from the mean, where z – any value > 1.

For ANY distribution:

- At least **75 %** of the values are within **z = 2** standard deviations from the mean
- At least **89 %** of the values are within **z = 3** standard deviations from the mean
- At least **94 %** of the values are within **z = 4** standard deviations from the mean
- At least **96%** of the values are within **z = 5** standard deviations from the mean

DETECTION OF OUTLIERS

Normal and other bell-shaped

For bell-shaped distributions:

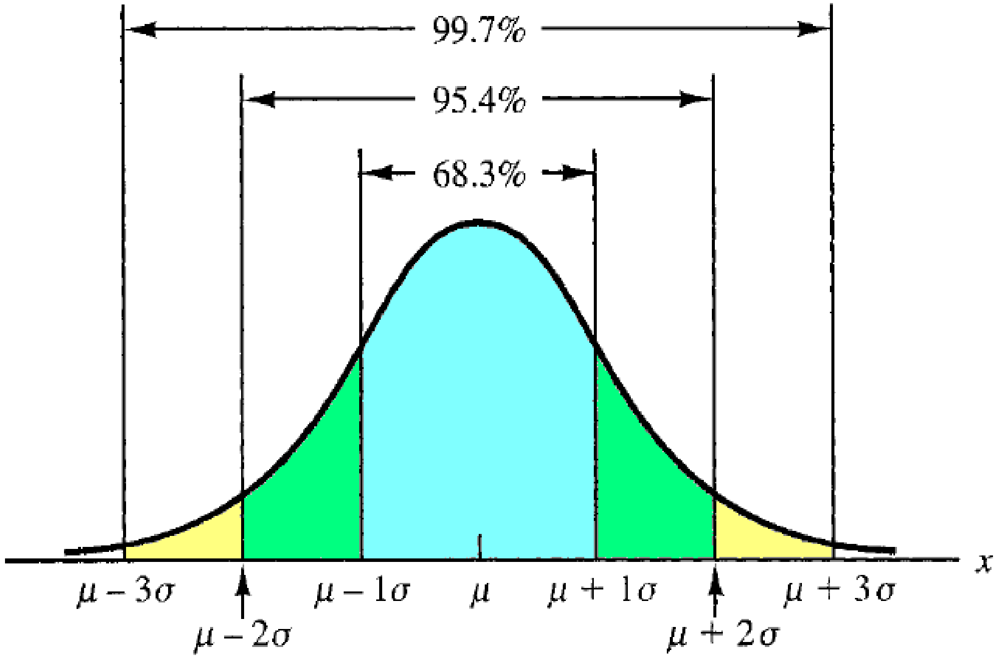
- Approximately 68 % of the values are within 1 st.dev. from mean
- Approximately 95 % of the values are within 2 st.dev. from mean
- Almost all data points are inside 3 st.dev. from mean

Outlier
An unusually small or unusually large data value.

For bell-shaped distributions data points with |z|>3 can be considered as outliers.

Weight	z-score
23	0.04
12	-0.53
22	-0.01
12	-0.53
21	-0.06
81	**3.10**
22	-0.01
20	-0.11
12	-0.53
19	-0.17
14	-0.43
13	-0.48
17	-0.27

Example: Gaussian / normal distribution



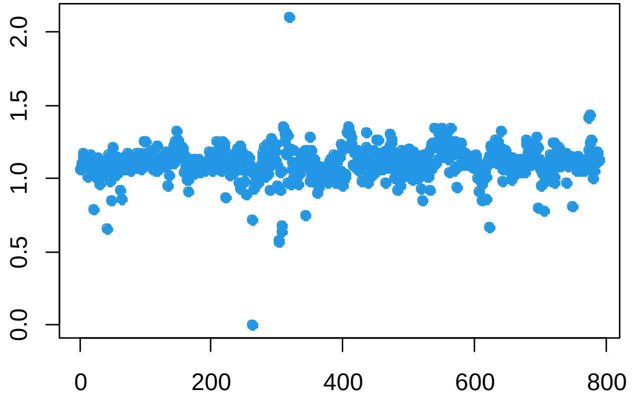
DETECTION OF OUTLIERS

Task: Detection of Outliers

mice

Using R, try to identify outlier mice on the basis of ***Weight change*** variable

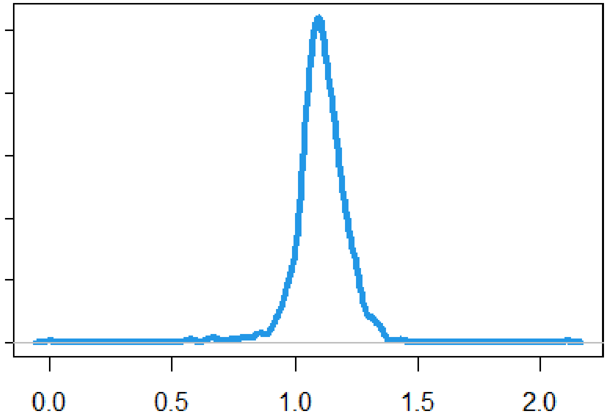
```
# load Mice dataset
Mice = read.table(
"http://edu.modas.lu/data/txt/mice.txt", sep="\t", header=TRUE, as.is = FALSE)

str(Mice)
```

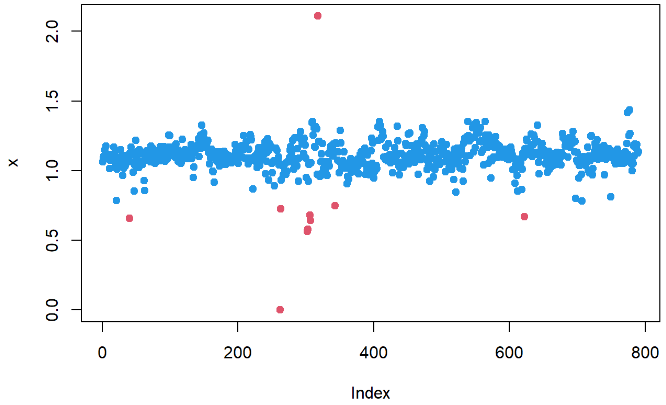


$$z_i = \frac{x_i - m}{s}$$

For bell-shaped distributions data points with |z|>3 can be considered as outliers.

```
# take and show the data

x = Mice$Weight.change

plot(x,pch=19,col=4)

plot(density(x,na.rm=TRUE))

# z-score

z = scale(x)

# show outlier values

x[abs(z)>3]
# show outlier mice

Mice[abs(z)>3,]
```





DETECTION OF OUTLIERS

Iglewicz-Hoaglin Method

Iglewicz-Hoaglin method: modified Z-score

$$z_i = 0.6745 \frac{x_i - median(x)}{MAD(x)}$$

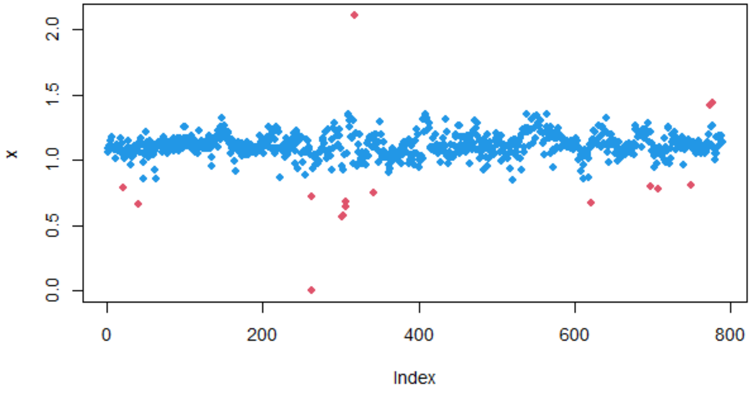
$$MAD = median\left(\left|x_i - median(x)\right|\right)$$

$$|z|>3.5 \Rightarrow \text{outlier}$$

These authors recommend that modified Z-scores with an absolute value of greater than 3.5 be labeled as potential outliers.

```
x = Mice$Weight.change
z = (x-median(x))/mad(x)
# index of outlier mice
iout = abs(z)>3.5
## plot
plot(x, pch=19, col=
c(4,2)[as.integer(iout)+1] )
```

Boris Iglewicz and David Hoaglin (1993), "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor



More methods are at:

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

DETECTION OF OUTLIERS

Grubbs' Method

Grubbs' test is an iterative method to detect outliers in a data set assumed to come from a normally distributed population.

Grubbs' statistics at step k+1:

$$G_{(k+1)} = \frac{\max|x_i - m_{(k)}|}{s_{(k)}} = \max|z_i|_{(k)}$$

(k) – iteration k
m – mean of the rest data
s – st.dev. of the rest data

The hypothesis of no outliers is rejected at significance level α if

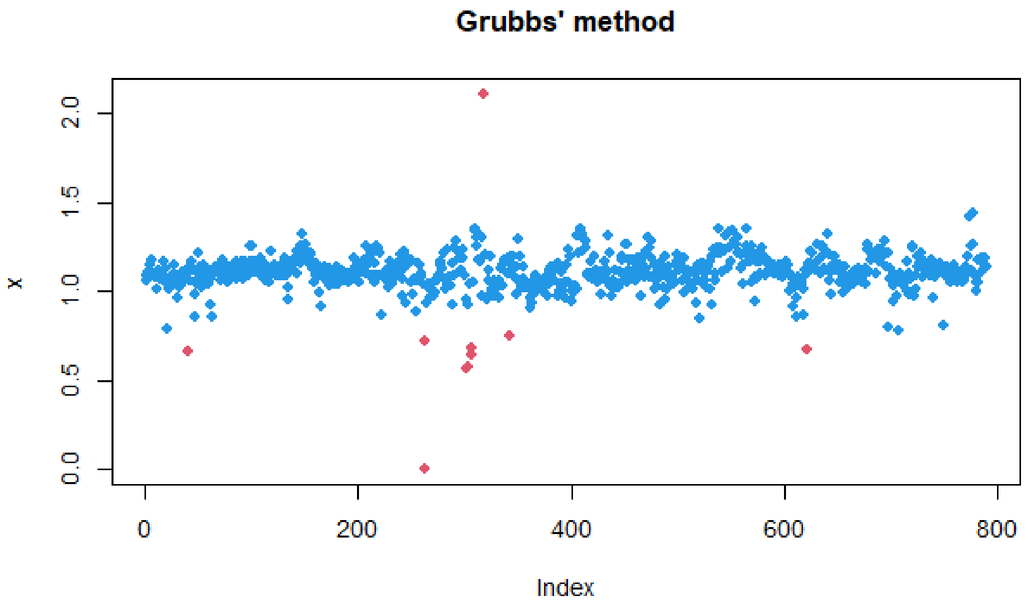
$$G > \frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2}{n-2+t^2}}$$

where

$$t^2 = t^2_{\alpha/(2n),\ d.f.=n-2}$$

t – Student statistics

```
library(outliers)
x1 = x
while(grubbs.test(x1)$p.value<0.05)
  x1[ x1==outlier(x1) ] = NA

plot(x, pch=19, col=2,
     main="Grubbs' method")points(x1, pch=19, col=4)
```

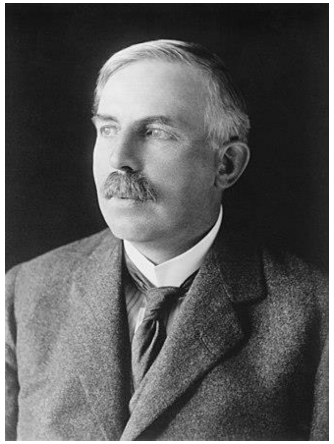


DETECTION OF OUTLIERS

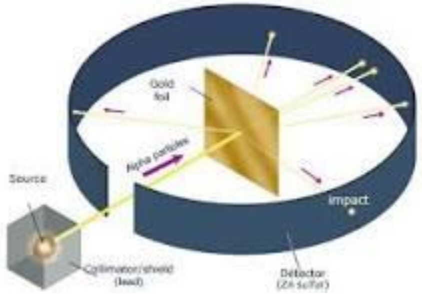
Remember!

Generally speaking, removing of outliers is a dangerous procedure and cannot be recommended!

Instead, potential outliers should be investigated and only (!) if there is other evidence that data come from experimental error – removed.

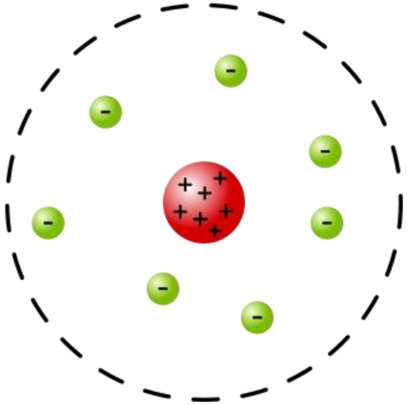


Ernest Rutherford



Outliers

~ 0.01%



Rutherford's atom model

LUXEMBOURG INSTITUTE OF HEALTH

DISTRIBUTIONS

Probability density function

Normal distribution

Other: *t*, χ^2, F distributions

Sampling distribution

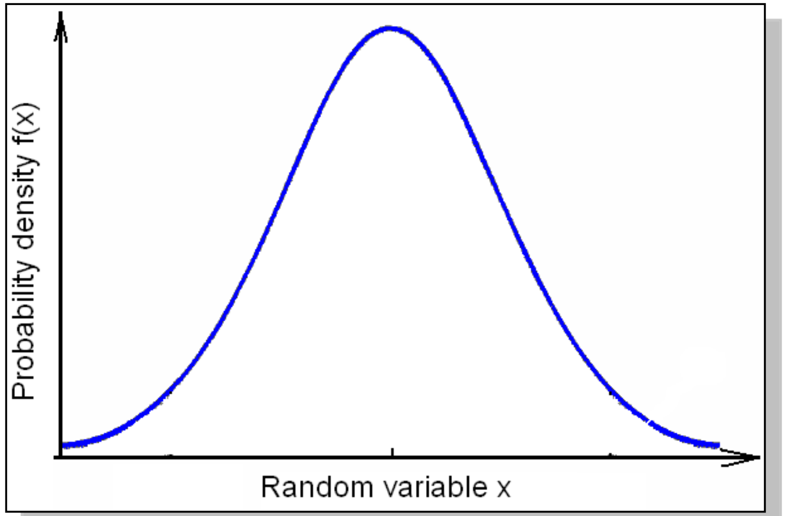
Point estimation

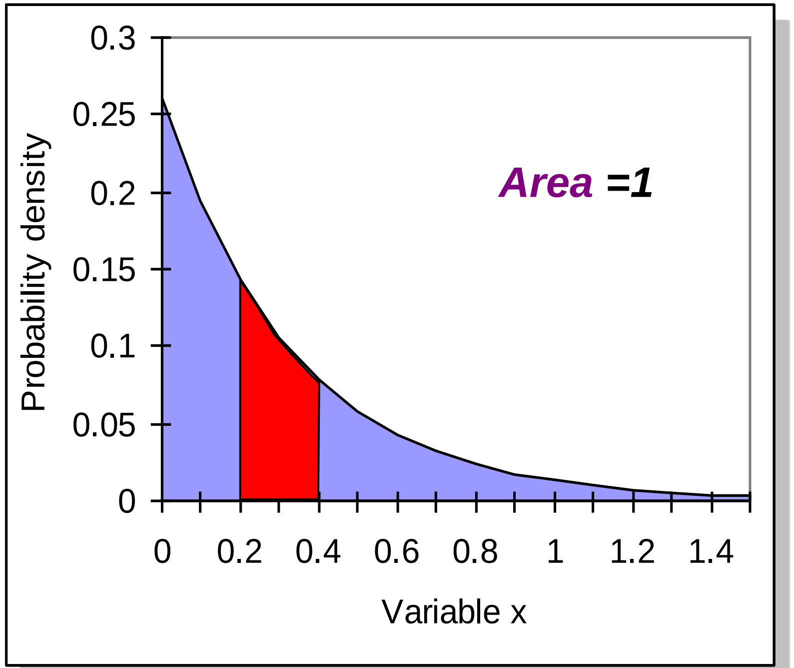
DISTRIBUTIONS

Probability Density

Probability density function

A function used to compute probabilities for a continuous random variable. The area under the graph of a probability density function over an interval represents probability.





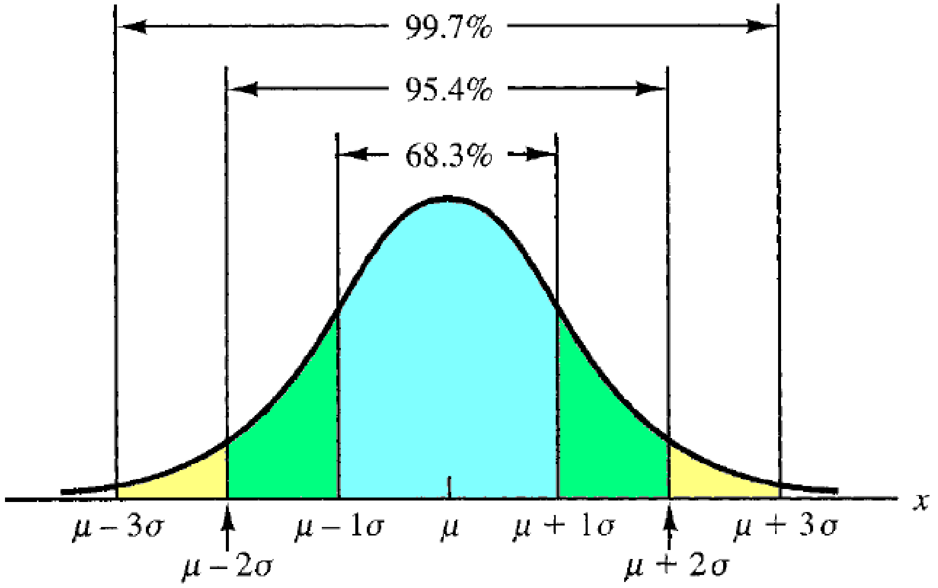
$$\int_x f(x) = 1$$

NORMAL DISTRIBUTION

Normal Probability Density Function

Normal (Gaussian) probability distribution
A continuous probability distribution. Its probability density function is bell shaped and determined by its mean μ and standard deviation σ.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

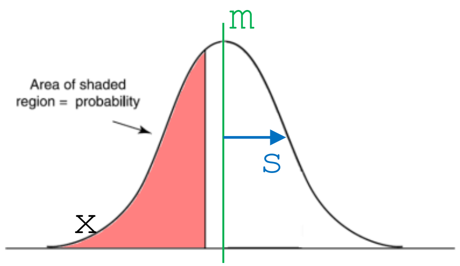


(cumulative) **Probability** function:

```
pnorm(x,m,s)
```

Probability **density** function:

```
dnorm(x,m,s)
```



probability density (x->y):	dnorm()
cumulative probability (x->p):	pnorm()
quantile (p->x):	qnorm()
generate random variables (x):	rnorm()

NORMAL DISTRIBUTION

Standard Normal Probability Distribution

Standard normal probability distribution
A normal distribution with a mean of zero and a standard deviation of one. We will call it "normal statistics" later :)

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

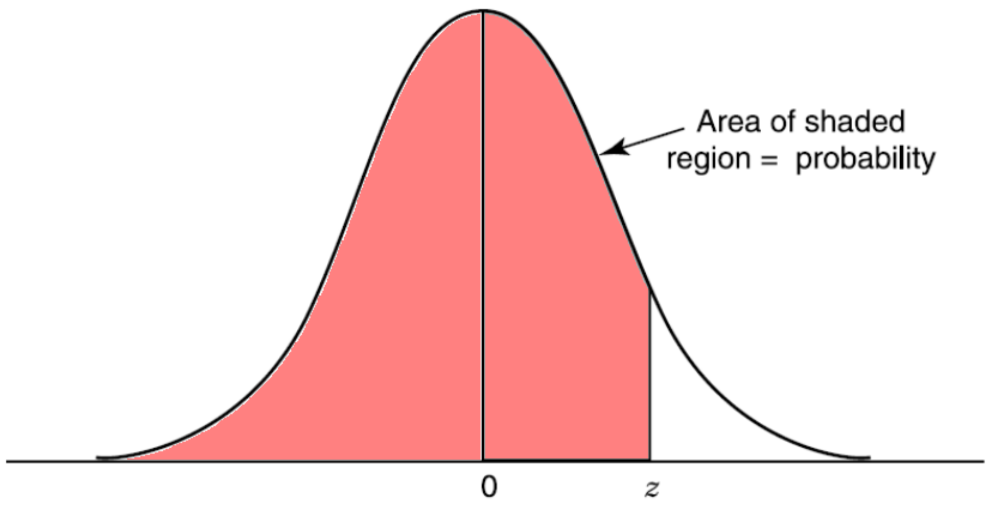
remember z-score?

x → z

$$z = \frac{x - \mu}{\sigma}$$

z → x

$$x = \sigma z + \mu$$



`pnorm(z)`

LUXEMBOURG INSTITUTE OF HEALTH

NORMAL DISTRIBUTION

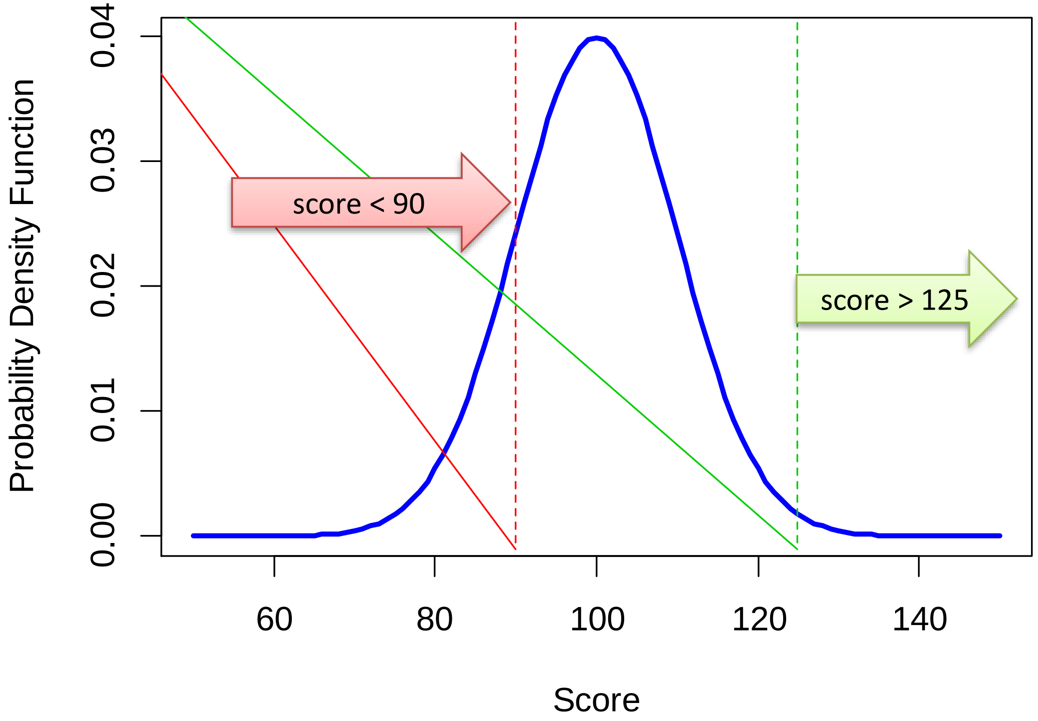
Example: Aptitude Test

Example

Suppose that the score on an aptitude test are normally distributed with a mean of 100 and a standard deviation of 10. (Some original IQ tests were purported to have these parameters)

What is the probability that a randomly selected score is below 90?

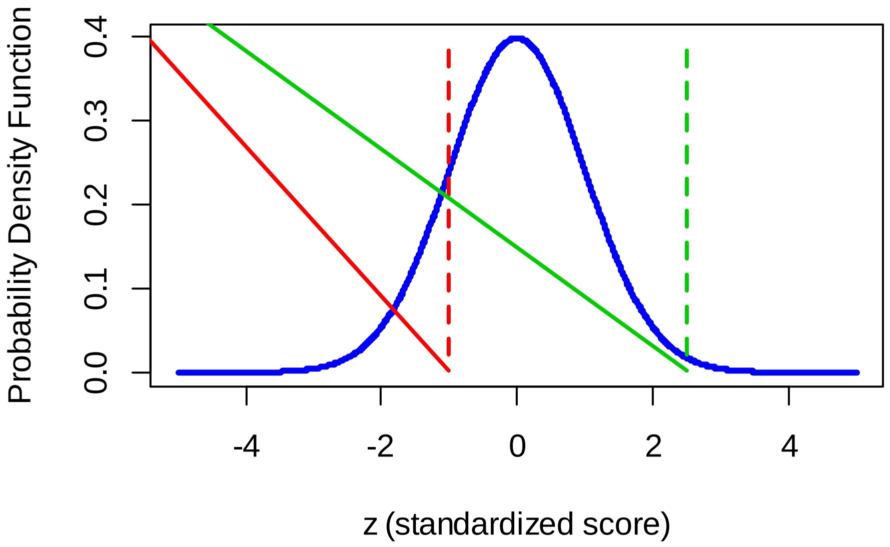
What is the probability that a randomly selected score is above 125?



Glover & Mitchel. An introduction to biostatistics

NORMAL DISTRIBUTION

Example: Aptitude Test



Example

Suppose that the score on an aptitude test are normally distributed with a mean of 100 and a standard deviation of 10. (Some original IQ tests were purported to have these parameters.) What is the probability that a randomly selected score is below 90?
What is the probability that a randomly selected score is above 125?

Classical way:

Let's transform Normal distribution **x** to Standard Normal **z**

$$z_{x=90} = \frac{90-100}{10} = -1 \qquad z_{x=125} = \frac{125-100}{10} = 2.5$$

Calculate the area under the curve before these z-values:

P(x<90) = P(z< -1) = NORM.S.DIST(-1;TRUE) = **0.159**
P(x>125) = P(z>2.5) = 1- P(z<2.5) =1- NORM.S.DIST(2.5,TRUE) = **0.006**

Easier way:

We can directly work with Normal distribution if we know its *mean* and *standard deviation*.

```
pnorm(90,100,10)
1-pnorm(125,100,10)
```

NORMAL DISTRIBUTION

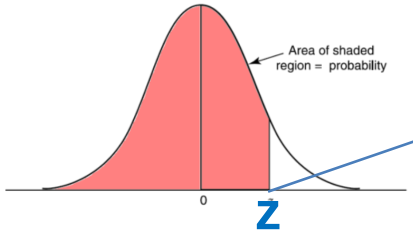
Example: Inverted situation

Example

Suppose that the score on an aptitude test are normally distributed with a mean of 100 and a standard deviation of 10.

Find the score cutting top 5% respondent?

Glover & Mitchel. An introduction to biostatistics



Assume that we know red area (probability *p*).
Then limiting z can obtained using:

```
qnorm(p)

qnorm(p,m,s)
```

```
qnorm(1-0.05, 100, 10)
> 116
```

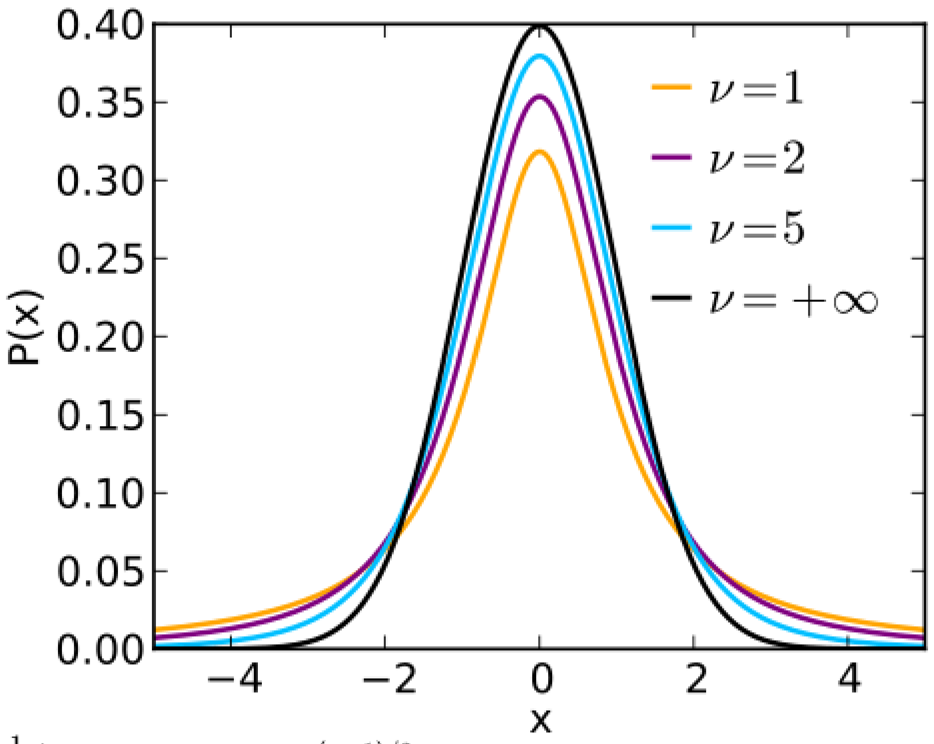
OTHER CONTINUOUS DISTRIBITIONS

Student's (t) Distribution

Student's *t*-distribution
is a continuous probability distribution that generalizes the standard normal distribution. It has very similar properties but heavier tails.

Degrees of freedom
A parameter of many distributions that is usually linked to the **number of independent observations**. E.g. when t distribution is used for the computation of an interval estimate of a population mean, the appropriate t distribution has $\nu = n - 1$ degrees of freedom, where n is the size of the simple random sample.

Student ***t*** distribution with d.f. $\nu \to \infty$
becomes normal ***z*** distribution



probability density (x->y):	`dnorm()`	`dt()`
cumulative probability (x->p):	`pnorm()`	`pt()`
quantile (p->x):	`qnorm()`	`qt()`
generate random variables (x):	`rnorm()`	`rt()`

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi \nu}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

ν - degree of freedom

Wikipedia

OTHER CONTINUOUS DISTRIBITIONS

Chi-squared (χ^2) Distribution

χ^2-distribution

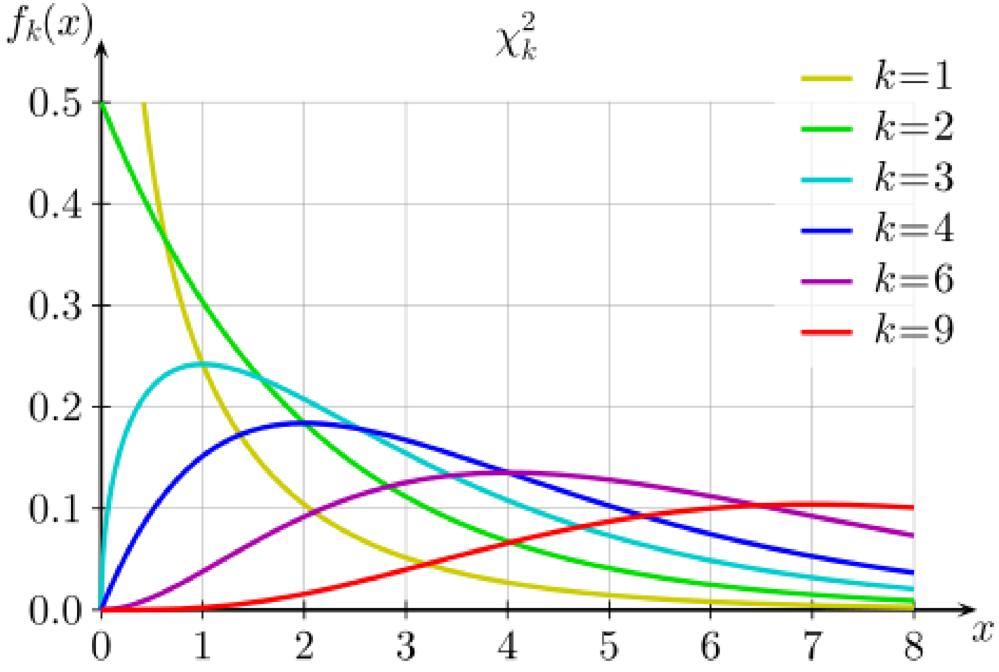
the chi-squared distribution (also chi-square or χ^2-distribution) with *k* degrees of freedom is the distribution of a sum of the squares of *k* independent standard normal random variables *z*. It describes the behavior of sampling variance.

$$\chi^2_{df=k} = \sum_{i=1}^{k} x_i^2 \quad where\ x_i - normal$$

Some applications of χ^2 distribution:

- interval estimations for variance
- goodness of fit of statistical model to observations

probability density (x->y):	**dchisq()**
cumulative probability (x->p):	**pchisq()**
quantile (p->x):	**qchisq()**
generate random variables (x):	**rchisq()**



$$f(x;\,k) = \begin{cases} \dfrac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise} \end{cases}$$

Wikipedia

OTHER CONTINUOUS DISTRIBITIONS

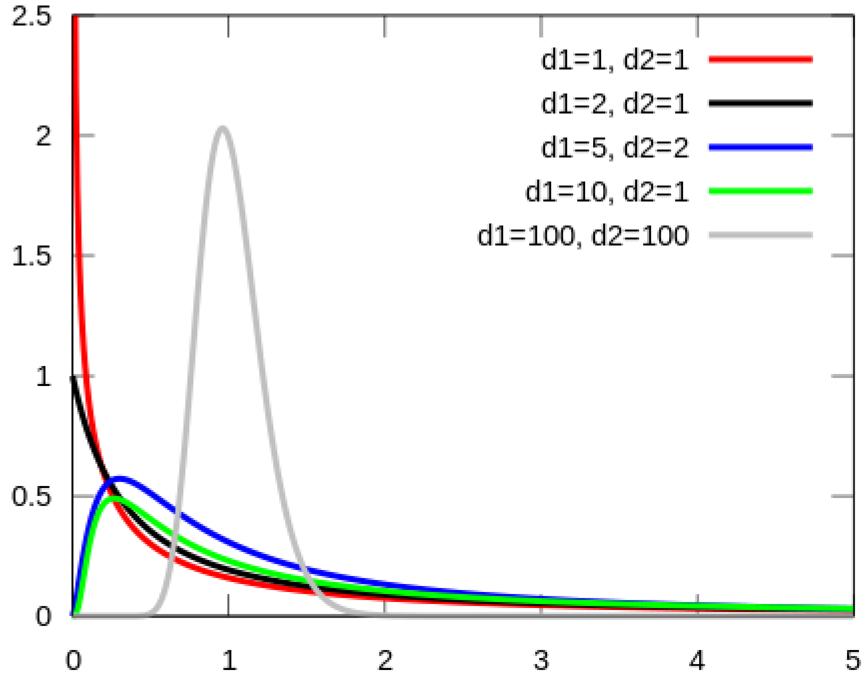
F-Distribution (Fisher–Snedecor)

***F*-distribution**

The F-distribution was introduced as a distribution of a ratio of two χ^2 random variables. It has **2 degrees of freedom** (numerator and denominator) and is used frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA) and F-test

$$X = \frac{S_1/d_1}{S_2/d_2}$$

The function is "invariant" to the function 1/x ☺.
So usually only values F > 1 are considered



Wikipedia

probability density (x->y):	**df()**
cumulative probability (x->p):	**pf()**
quantile (p->x):	**qf()**
generate random variables (x):	**rf()**

SAMPLING DISTRIBUTION

Population and Sample

Population parameter

A numerical value used as a summary measure for a population of size ***N*** (e.g., the population mean μ, variance σ^2, standard deviation σ)

POPULATION

μ – mean
σ^2 – variance
N – number of elements
(usually $N \to \infty$)

SAMPLE

m, $\bar{x}$ – mean
s^2 – variance
n – number of elements

Sample statistic

A numerical value used as a summary measure for a sample of size ***n*** (e.g., the sample mean ***m***, the sample variance s^2, and the sample standard deviation ***s***)

All existing laboratory
Mus musculus

`mice`

790 mice from different strains

http://phenome.jax.org

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvImJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvImJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvImJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvImJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvImJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvImJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvImJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvImJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvImJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvImJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvImJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvImJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvImJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvImJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvImJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvImJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvImJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvImJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvImJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

Load the data:

```
Mice = read.table("http://edu.modas.lu/data/txt/mice.txt", sep="\t", header=TRUE, stringsAsFactors = TRUE)
```

SAMPLING DISTRIBUTION

Example: Making a Random Sampling

```
mice
```

790 mice from different strains

http://phenome.jax.org

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvImJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvImJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvImJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvImJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvImJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvImJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvImJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvImJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvImJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvImJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvImJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvImJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvImJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvImJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvImJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvImJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvImJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvImJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvImJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

```
sample(x,size)
```

Assume that these mice is a population with size N=790. Build 5 samples with *n*=20

Calculate *m*, *s* for *Ending weight* and *p* – proportion of *males* for each sample

```
m = double(0)
s = double(0)
p = double(0)

for (i in 1:5){
  ix = sample(1:nrow(Mice),20)
  m[i] = mean(Mice$Ending.weight[ix])
  s[i] = sd(Mice$Ending.weight[ix])
  p[i] = mean(Mice$Sex[ix] == "m")
}

summary(m)
summary(s)
summary(p)
```

Point estimator

The sample statistics, such as ***m***, ***s***, or ***p*** (proportion) that provide the point estimations to the population parameters μ, σ, π. are called point estimators

Now, replace 5 with 1000
and check the distributions:

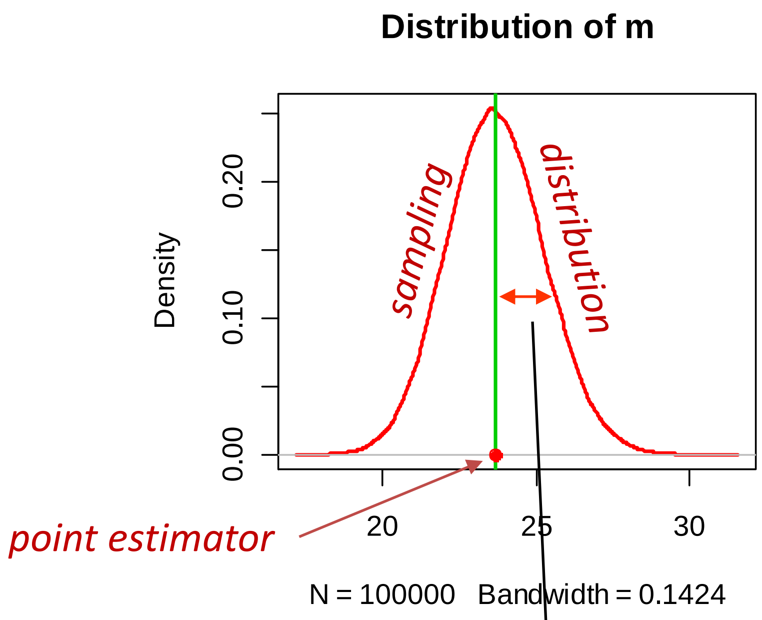
```
plot(density(m))
plot(density(s))
plot(density(p))
```

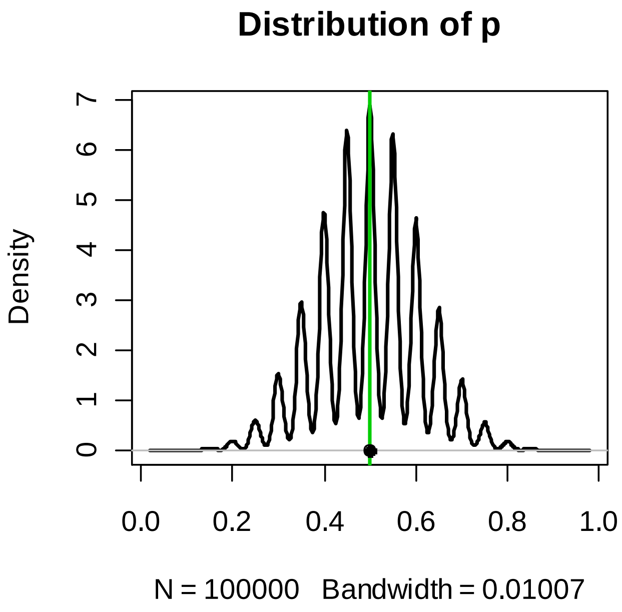
SAMPLING DISTRIBUTION

Sampling Distribution

Sampling distribution
A probability distribution consisting of all possible values of a sample statistic.

Point estimator
The sample statistic, such as *m*, *s*, or *p*, that provides the point estimation the population parameters μ, σ, π.





$$E(m) = \mu$$

$$E(p) = \pi$$

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

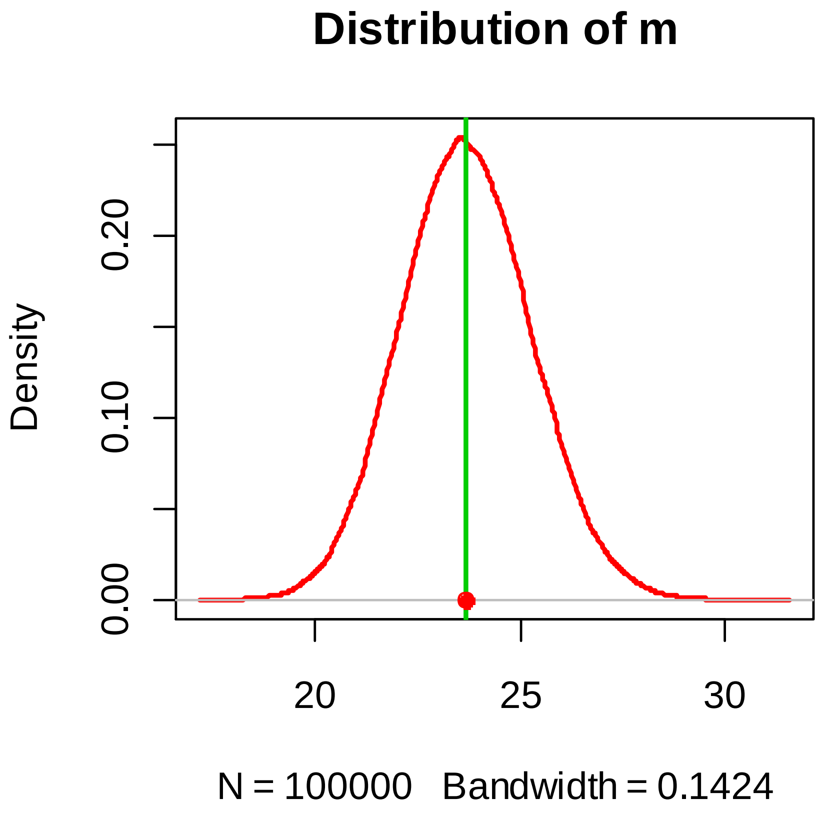
$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

The standard deviation of the point estimator - "Standard error"

SAMPLING DISTRIBUTION

Unbiased Point Estimator: Mean

Unbiased
A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.



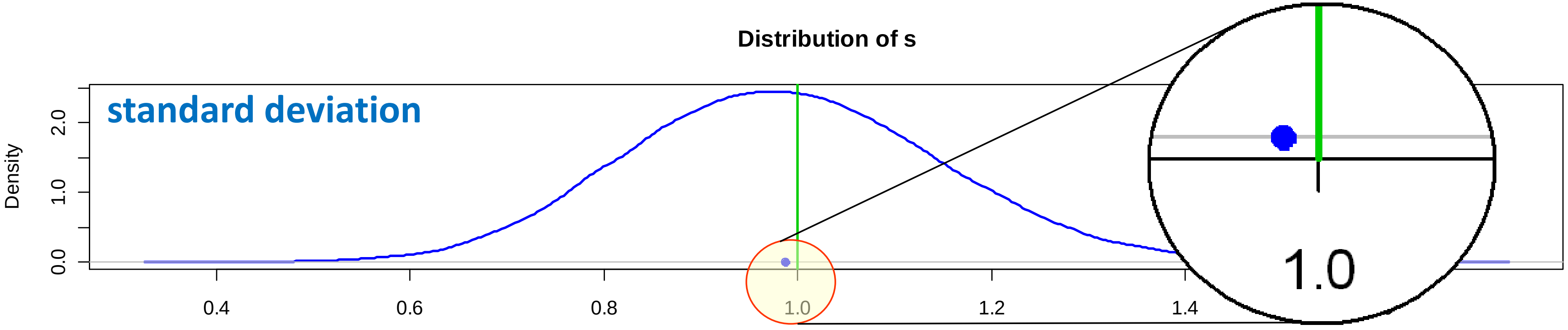
$$E(m) = \mu$$

SAMPLING DISTRIBUTION

Unbiased Point Estimator: Variance (but not St.Dev!)

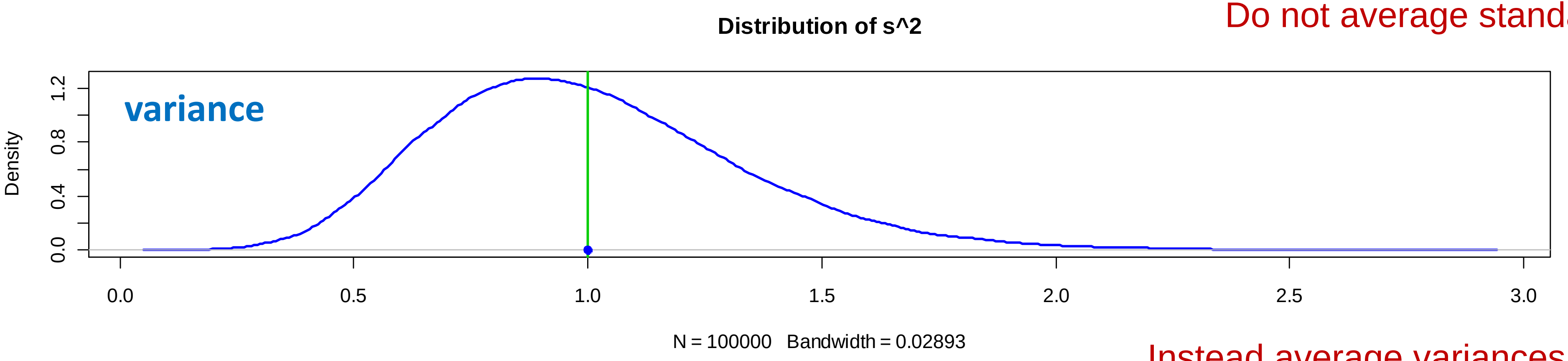
Unbiased
A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.

Biased



Do not average standard deviations!!!

Unbiased



Instead average variances and take sq.root

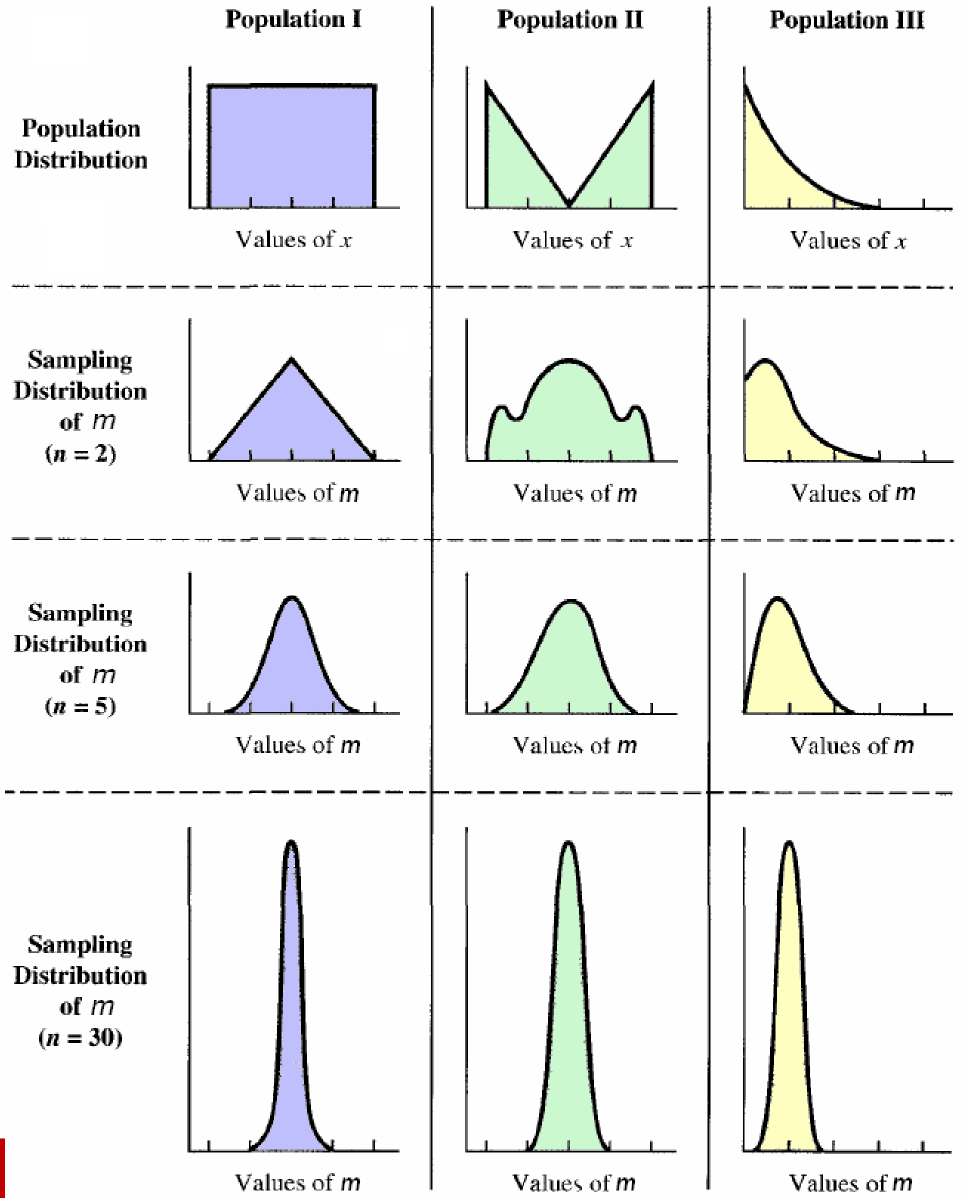
SAMPLING DISTRIBUTION

Central Limit Theorem

Central limit theorem

In selecting simple random sample of size *n* from a population, the ***sampling distribution of the sample mean m can be approximated by a normal distribution*** as the sample size becomes large

In practice, if the sample size *n*>30, the **normal distribution** is a good approximation for the sample mean for any initial distribution.

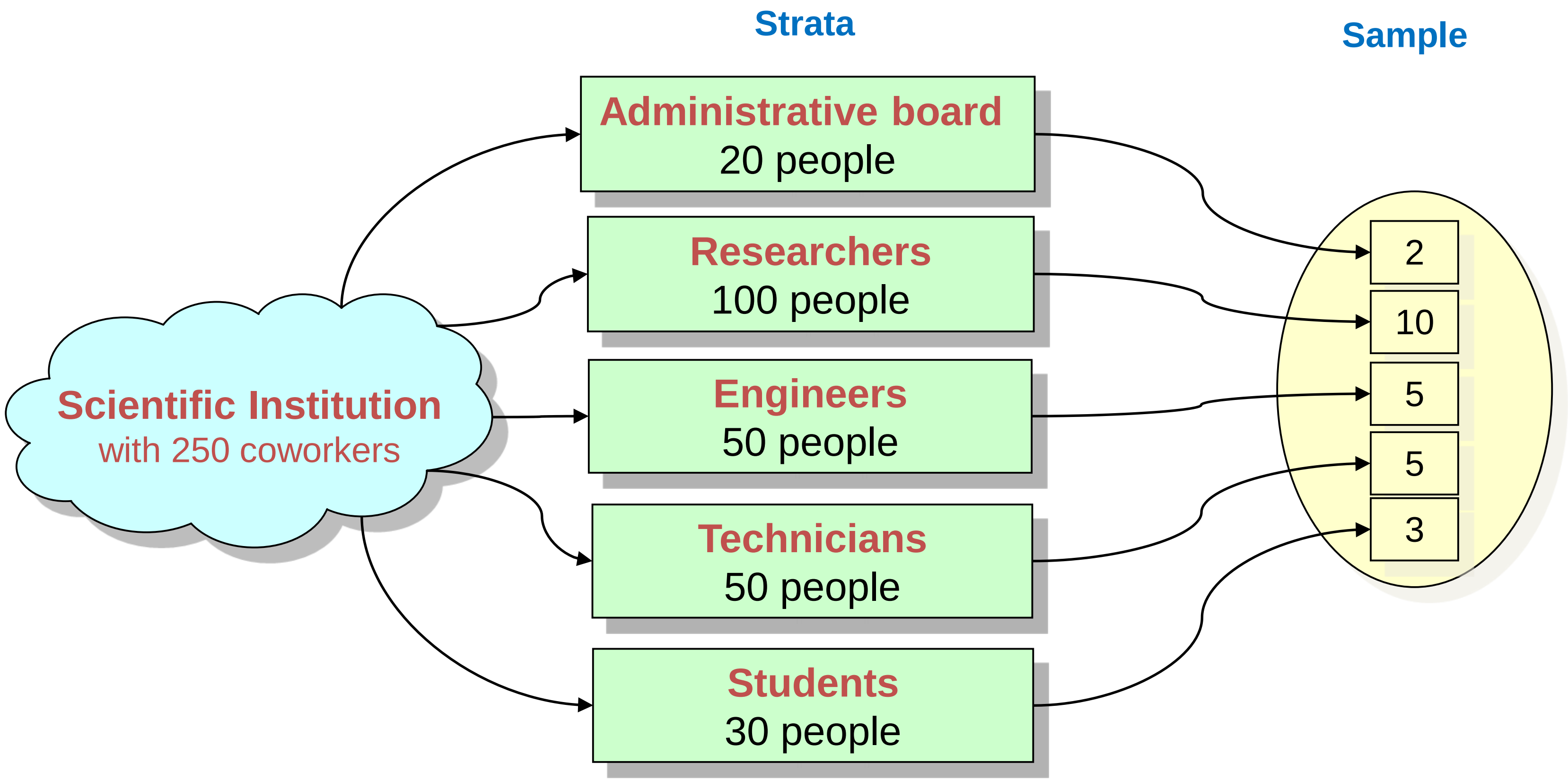


adapted from Anderson et al Statistics for Business and Economics

SAMPLING METHODS

Stratified Sampling

Stratified random sampling
A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.



SAMPLING METHODS

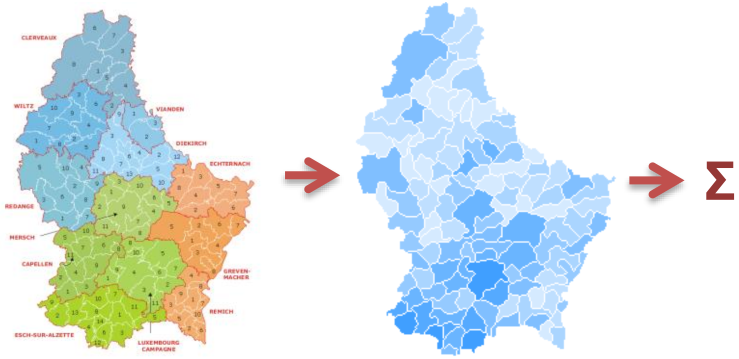
Stratified Sampling Strategies

Stratified random sampling
A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.

Strategies of Stratified Sampling

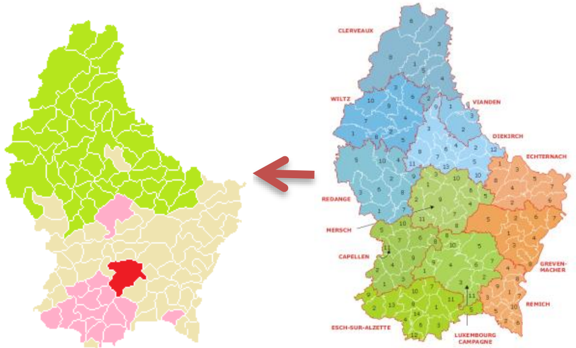
Proportionate allocation

Selected size $\mathbf{n_i}$ of a sub-sample depends on **population size** $\mathbf{N_i}$ of a strata



Optimum (disproportionate) allocation

Selected size $\mathbf{n_i}$ of a sub-sample depends on **variance** $\boldsymbol{\sigma_i^2}$ of a strata



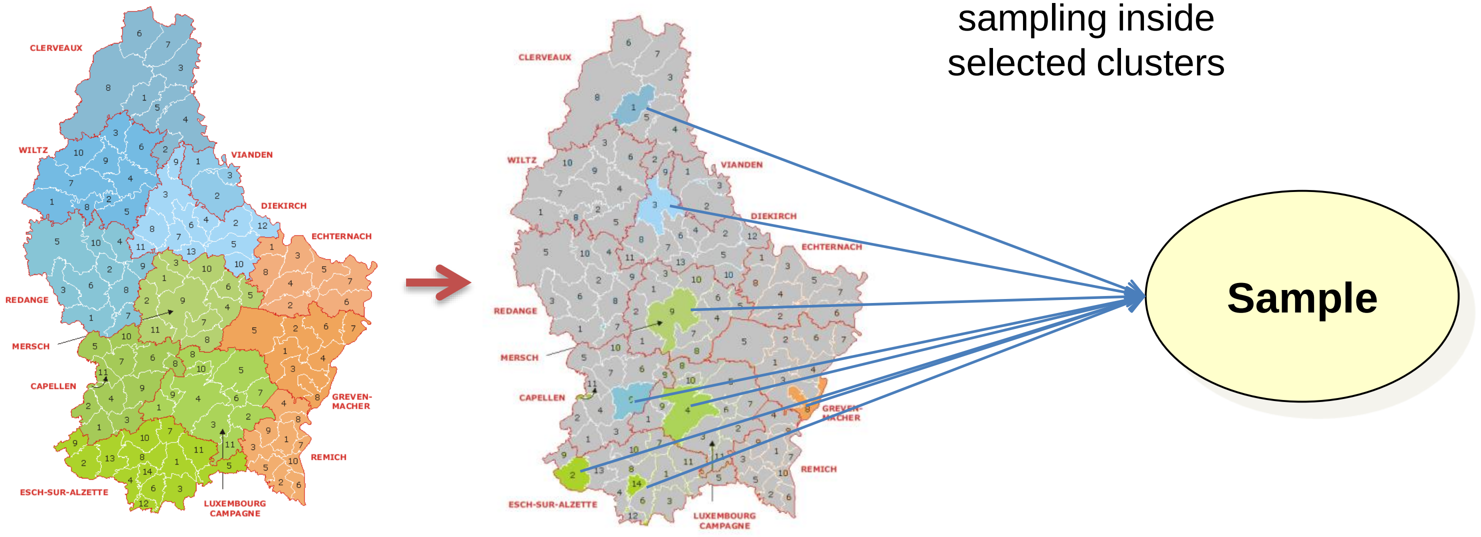
SAMPLING METHODS

Cluster Sampling

Cluster sampling
A probability sampling method in which the population is first divided into clusters and then a simple random sample of the clusters is taken.

1. Random sampling of sampling based on cost optimization

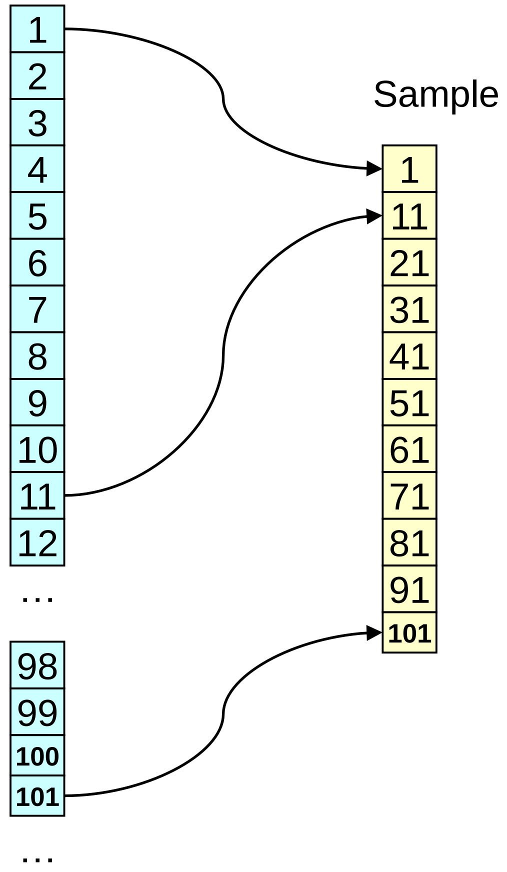
2. Simple random sampling inside selected clusters



SAMPLING METHODS

Systematic Sampling

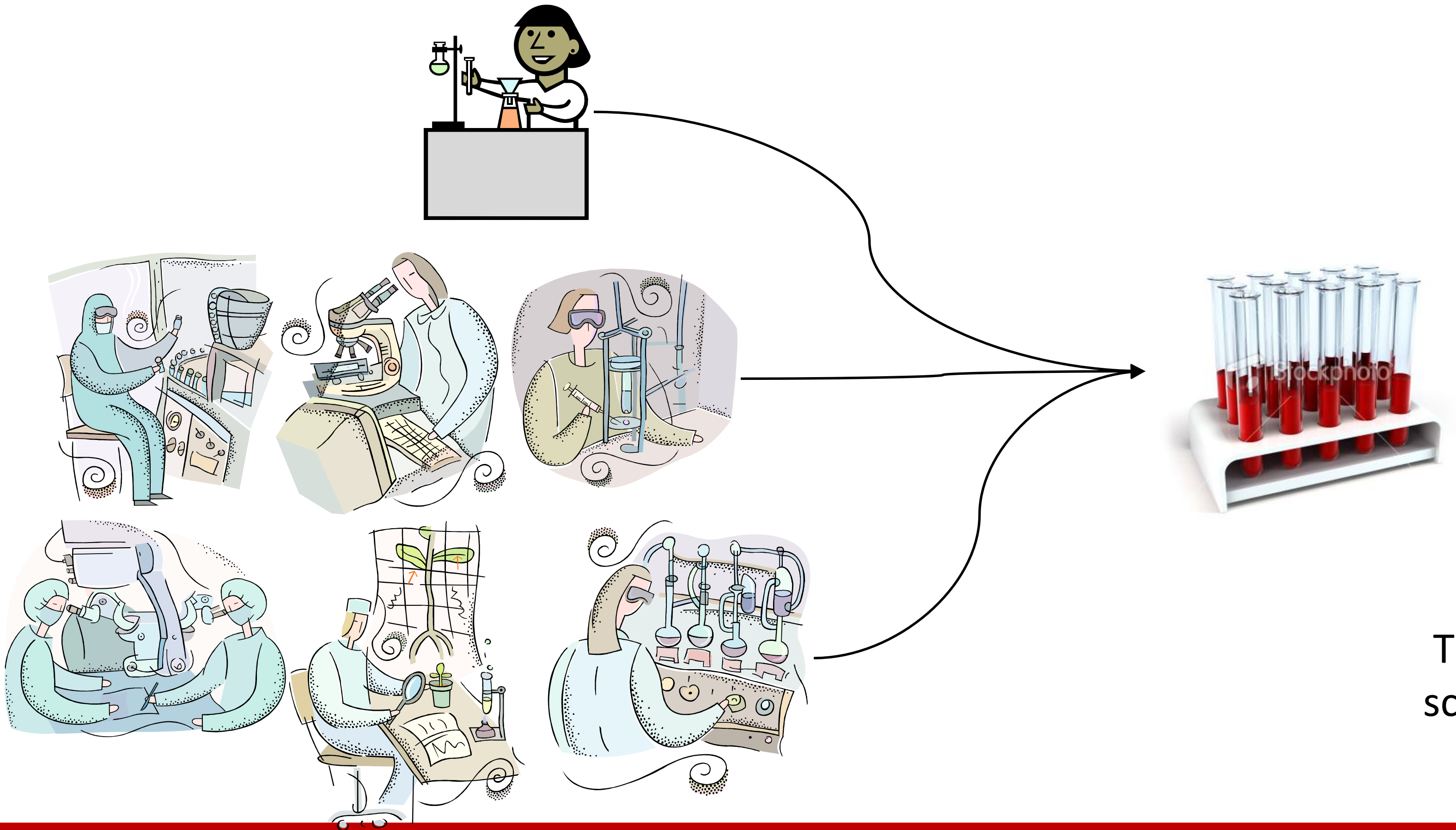
Systematic sampling
A probability sampling method in which we randomly select one of the first *k* elements and then select every *k*-th element thereafter.



SAMPLING METHODS

Convenience Sampling

Convenience sampling
A nonprobability method of sampling whereby elements are selected for the sample on the basis of convenience.



This is what we use often in science… though it is not too scientific ☺

SAMPLING METHODS

Judgment Sampling

Judgment sampling
A nonprobability method of sampling whereby elements are selected for the sample based on the judgment of the person doing the study.

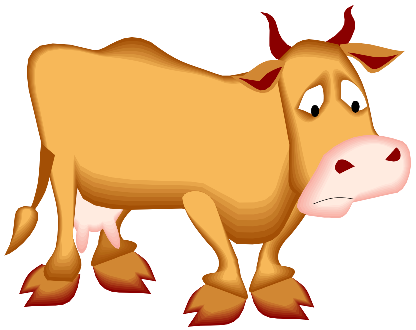
Perform of a selection of most confident or most experienced experts.

SAMPLING METHODS

The Wisdom of the Crowd

The wisdom of the crowd
is the process of taking into account the collective opinion of a group of individuals rather than a single expert to answer a question. A large group's aggregated answers to questions involving quantity estimation has generally been found to be as good as, and often better than, the answer given by any of the individuals within the group.

The classic wisdom-of-the-crowds finding involves point estimation of a continuous quantity. At a 1906 country fair in Plymouth, eight hundred people participated in a contest to estimate the weight of a slaughtered and dressed ox. Statistician Francis Galton observed that the median guess, 1207 pounds, was accurate within 1% of the true weight of 1198 pounds.

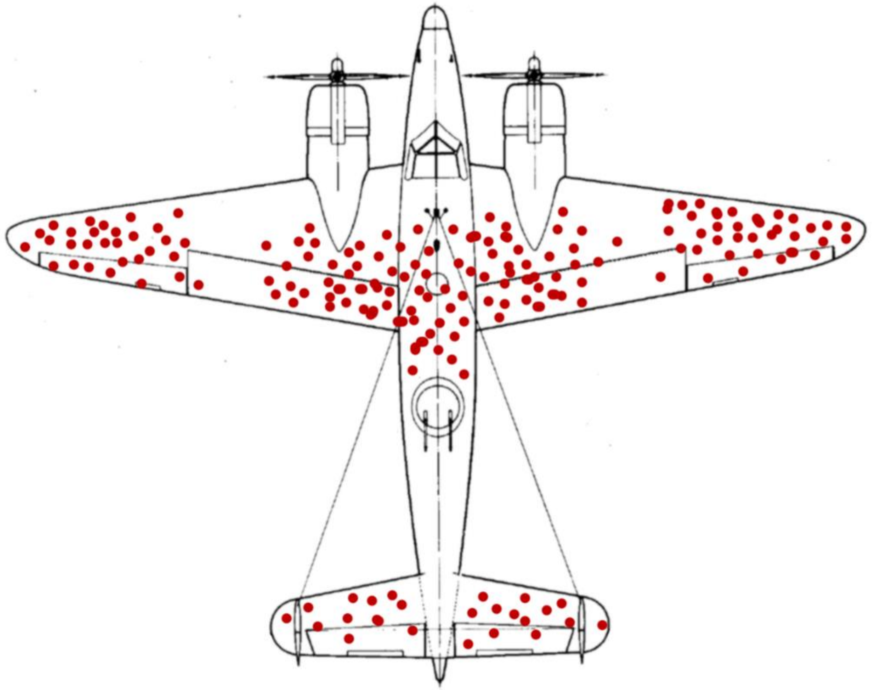


http://www.youtube.com/watch?v=r-FonWBEb0o

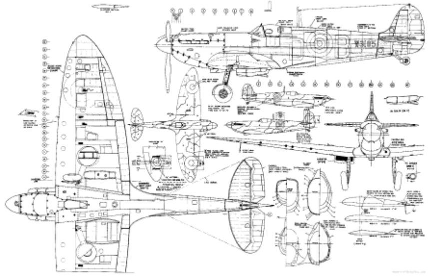
SAMPLING BIAS

Be Careful with Sampling

‘Spitfire’: damage analysis



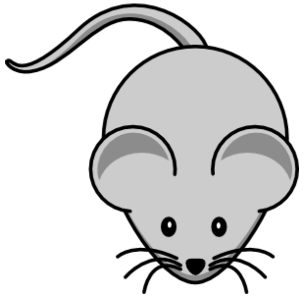
Were to put an additional protection?



Other examples: Paleolithic remains & lifestyle, kind dolphins, …

QUESTIONS ?

Thank you for your attention



to be continued…