PhD Course

Advanced Biostatistics

Lecture 2
Basic Statistics in R

Peter Nazarov

petr.nazarov@lih.lu

29-05-2017

Descriptive Statistics

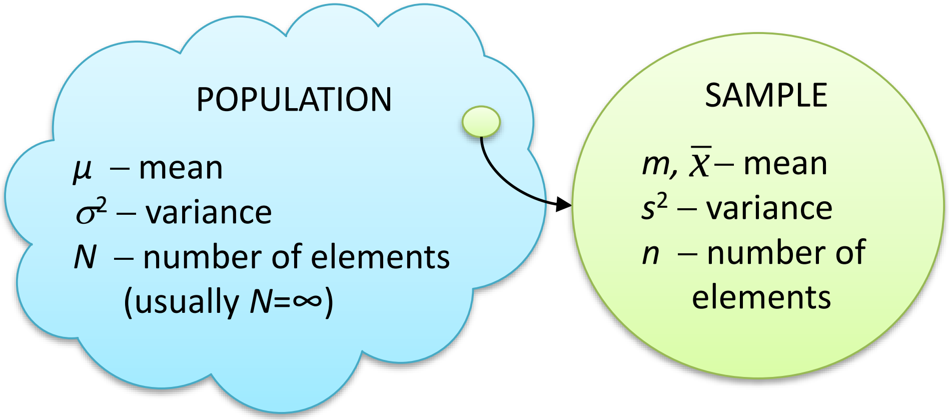
Outline

- **Descriptive statistics in R (L2.1)**
  - sum, mean, median, sd, var, cor, *etc*.
- **Statistical tests (L2.2)**
- **Detection of outliers (L2.3)**
  - z-score, Iglewicz-Hoaglin, Grubb's test

Population and Sample

Population parameter

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



POPULATION

μ – mean

σ^2 – variance

N – number of elements (usually $N=\infty$)

SAMPLE

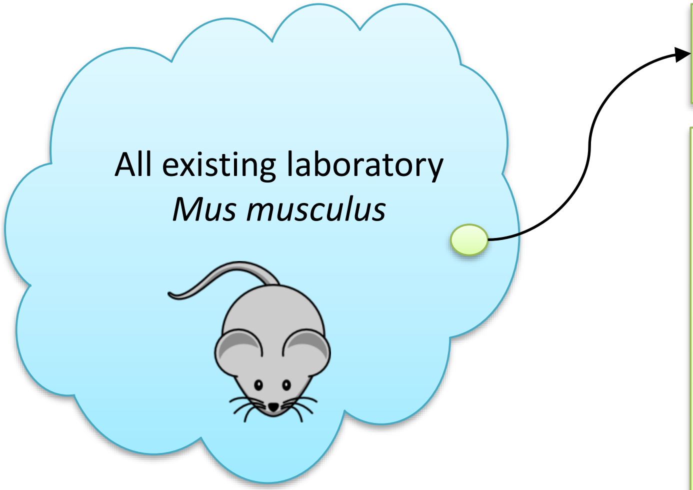
m, $\bar{x}$ – mean

s^2 – variance

n – number of elements

Sample statistic

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



All existing laboratory *Mus musculus*

`mice.txt`

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

L2.1. Descriptive Statistics in R

Measures of Location

Mean

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

Median

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

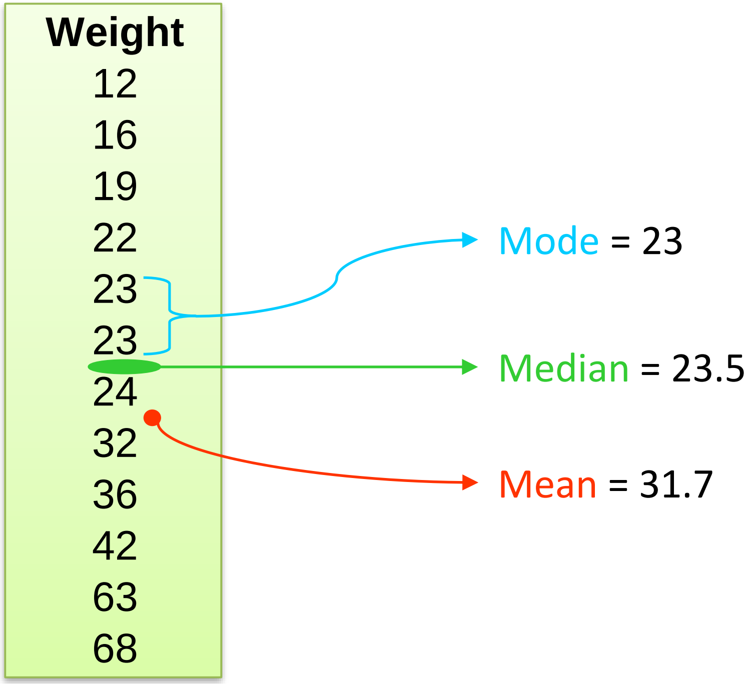
Mode

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

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

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

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



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

Mode = 23

Median = 23.5

Mean = 31.7

L2.1. Descriptive Statistics in R

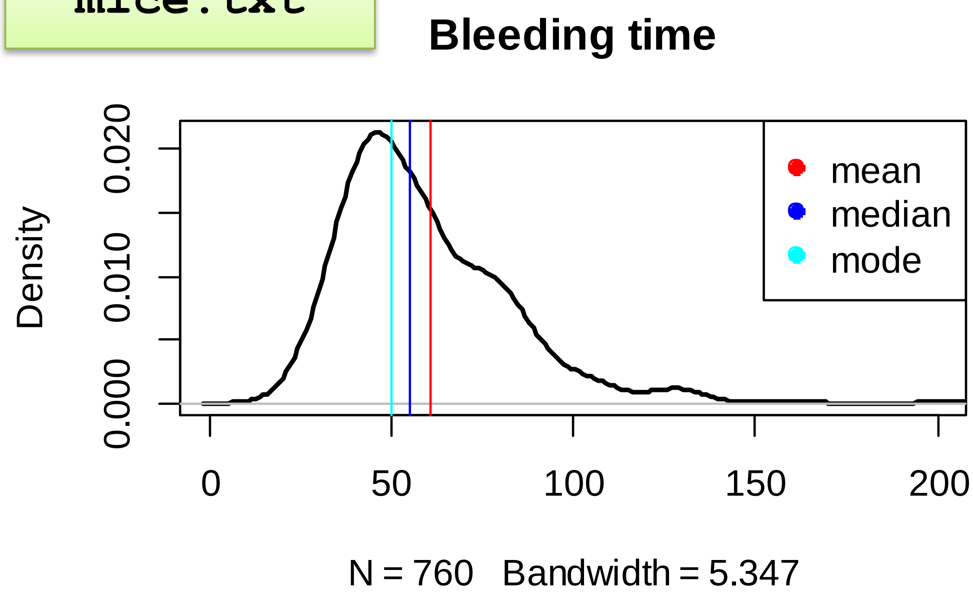
Measures of Location

In R use the following functions:

- `mean(x, na.rm=T)`
- `median(...)`
- `library(modeest)`
  `mlv(...)$M`

To by applied to data with missing elements (NA), use parameter: `..., na.rm = T`

`mice.txt`



In Excel use the following functions:

- `=AVERAGE(data)`
- `=MEDIAN(data)`
- `=MODE(data)`

To calculate proportion – count occurrence and divide by total number of elements:

```
prop.f = sum(Mice$Sex=="f")/nrow(Mice)
> 0.501
```

L2.1. Descriptive Statistics in R

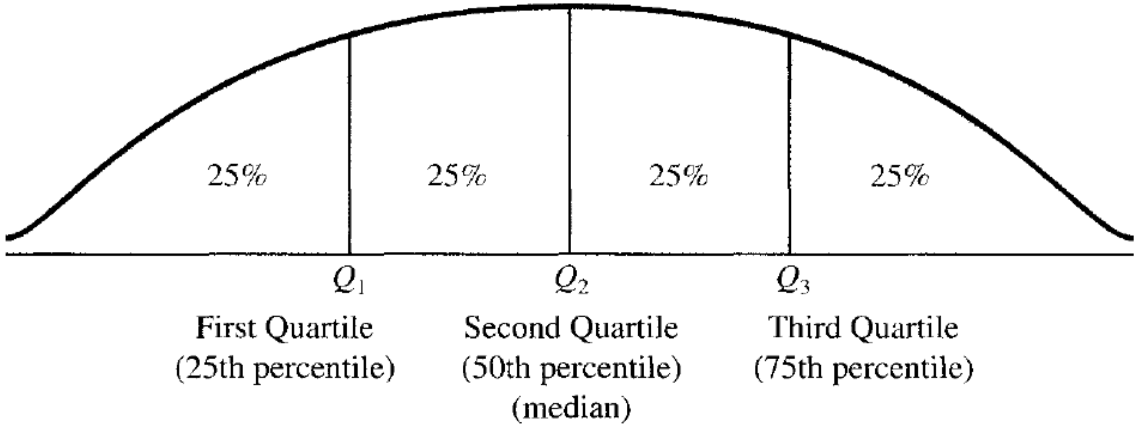
Quantiles, Percentiles and Quartiles

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.



In R use:

- `quantile(x,...)`

R provides up to 7 methods to estimate quantiles. Use parameter:

`type =` put a number 1-7

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

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

figure is adapted from Anderson et al Statistics for Business and Economics

In Excel use the following functions:

- `=PERCENTILE(data,p)`

L2.1. Descriptive Statistics in R

Measures of Variation

Interquartile range (IQR)

A 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 positive square root of the variance.

In R use:

- `IQR (x,...)`
- `sd (x,...)`
- `var (x,...)`

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

IQR = 18

Variance = 320.2

St. dev. = 17.9

In Excel use the following functions:

- `= STDEV(data)`
- `= VAR(data)`

L2.1. Descriptive Statistics in R

Measures of Variation

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

$$C_V = \frac{\sigma}{\mu}$$

C_V = 57%

Median absolute deviation (MAD)

MAD is a robust measure of the variability of a univariate sample of quantitative data.

$$MAD = 1.4826 \cdot med(|x_i - med(x)|)$$

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

Constant 1.4826 is introduced to ensure that MAD → σ for normal distribution. Can be modified by `constant = ...`

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

In R use:

◆ `mad (x,...)`

L2.1. Descriptive Statistics in R

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.txt`

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

In R use:

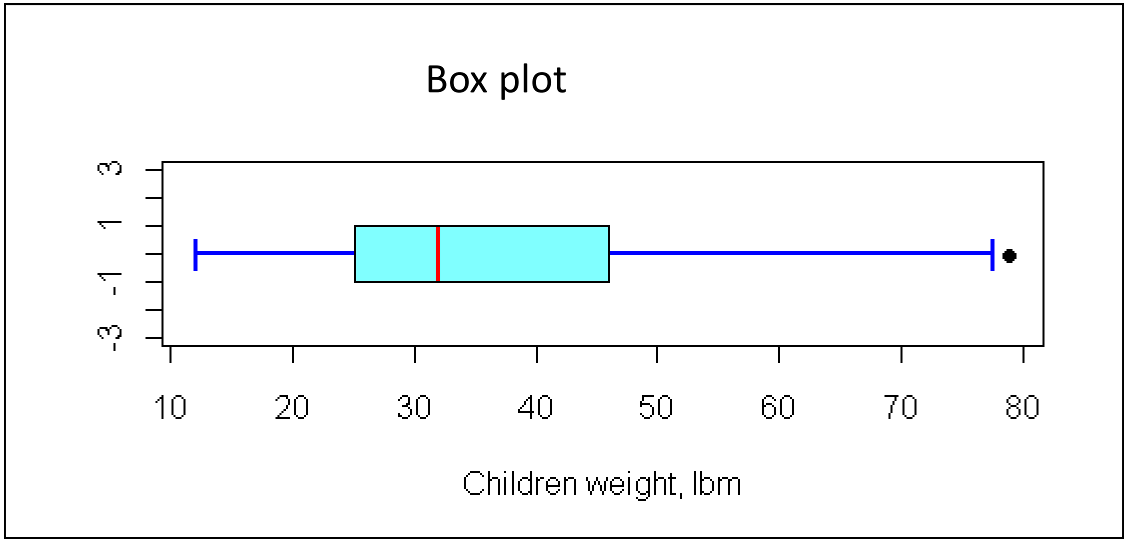
◆ `summary(x)`

Box plot

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

In R use:

◆ `boxplot(...)`



L2.1. Descriptive Statistics in R

Other Parameters

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.

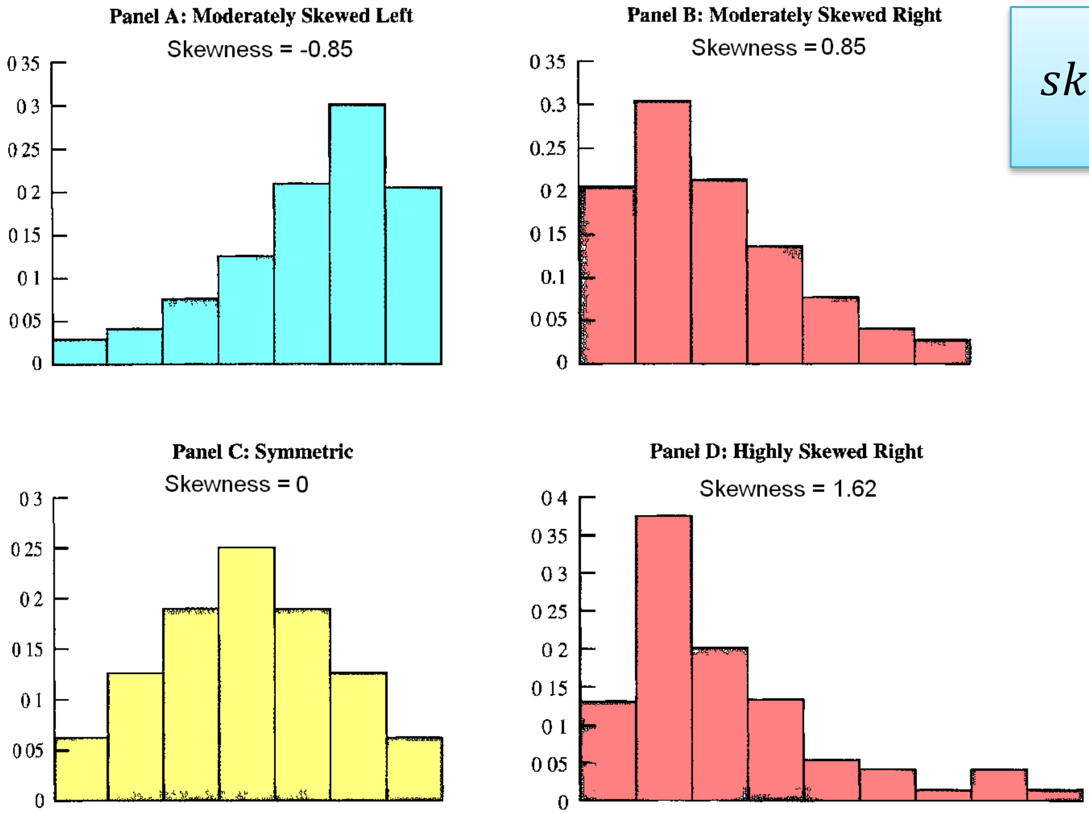


figure is adapted from Anderson et al Statistics for Business and Economics

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

In R use:

```
library(e1071)
 skewness(x,...)

library(modeest)
 skewness(x,...)
```

L2.1. Descriptive Statistics in R

Measure of Association between 2 Variables

Pearson 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.

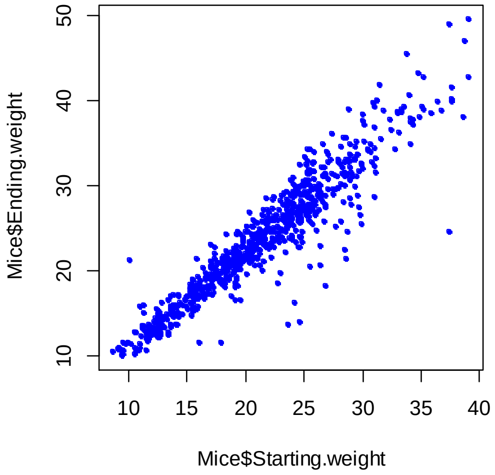
population

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

sample

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

`mice.xls`



r_{xy} = 0.94

In R use:

◆ `cor (x,...)`

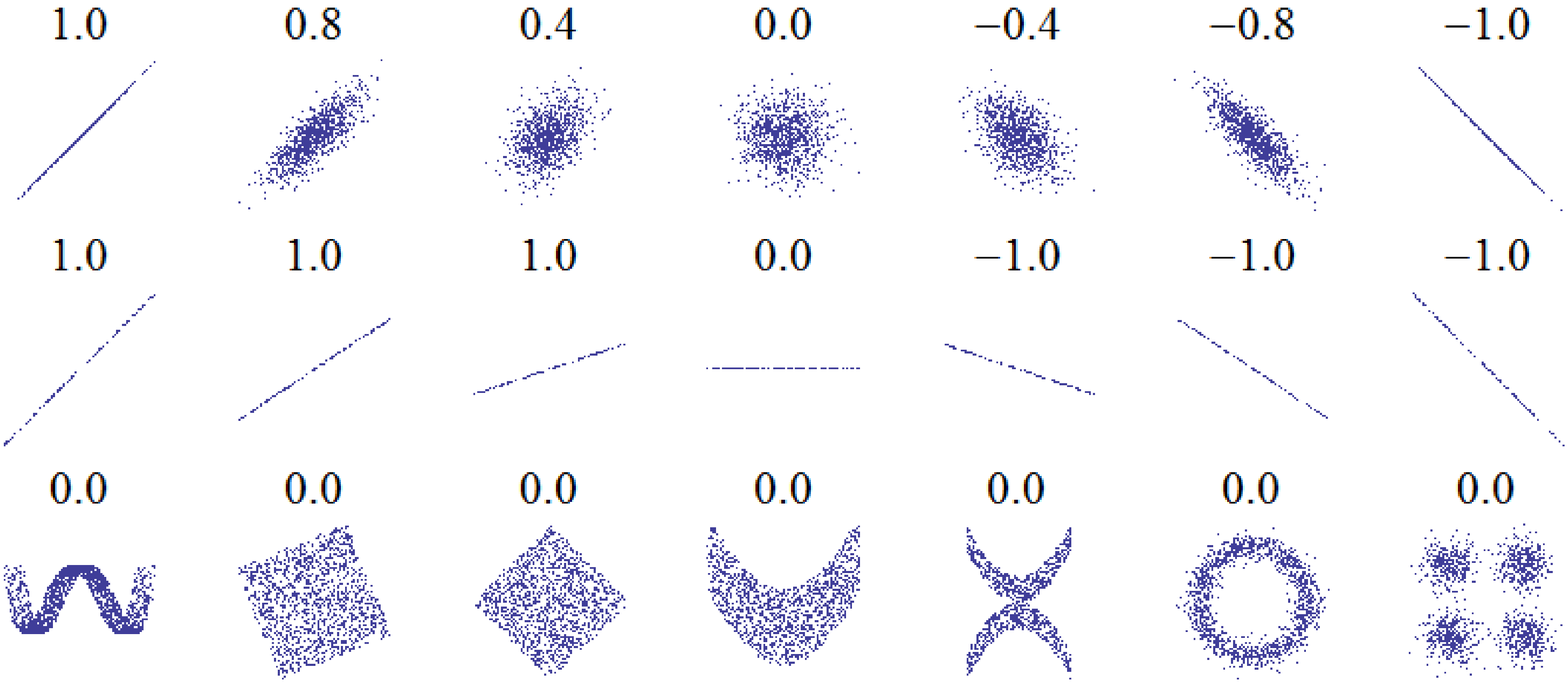
For missing data add parameter
`use = "pairwise.complete.obs"`

In Excel use function:

◆ `=CORREL(data)`

L2.1. Descriptive Statistics in R

Pearson Correlation



Wikipedia

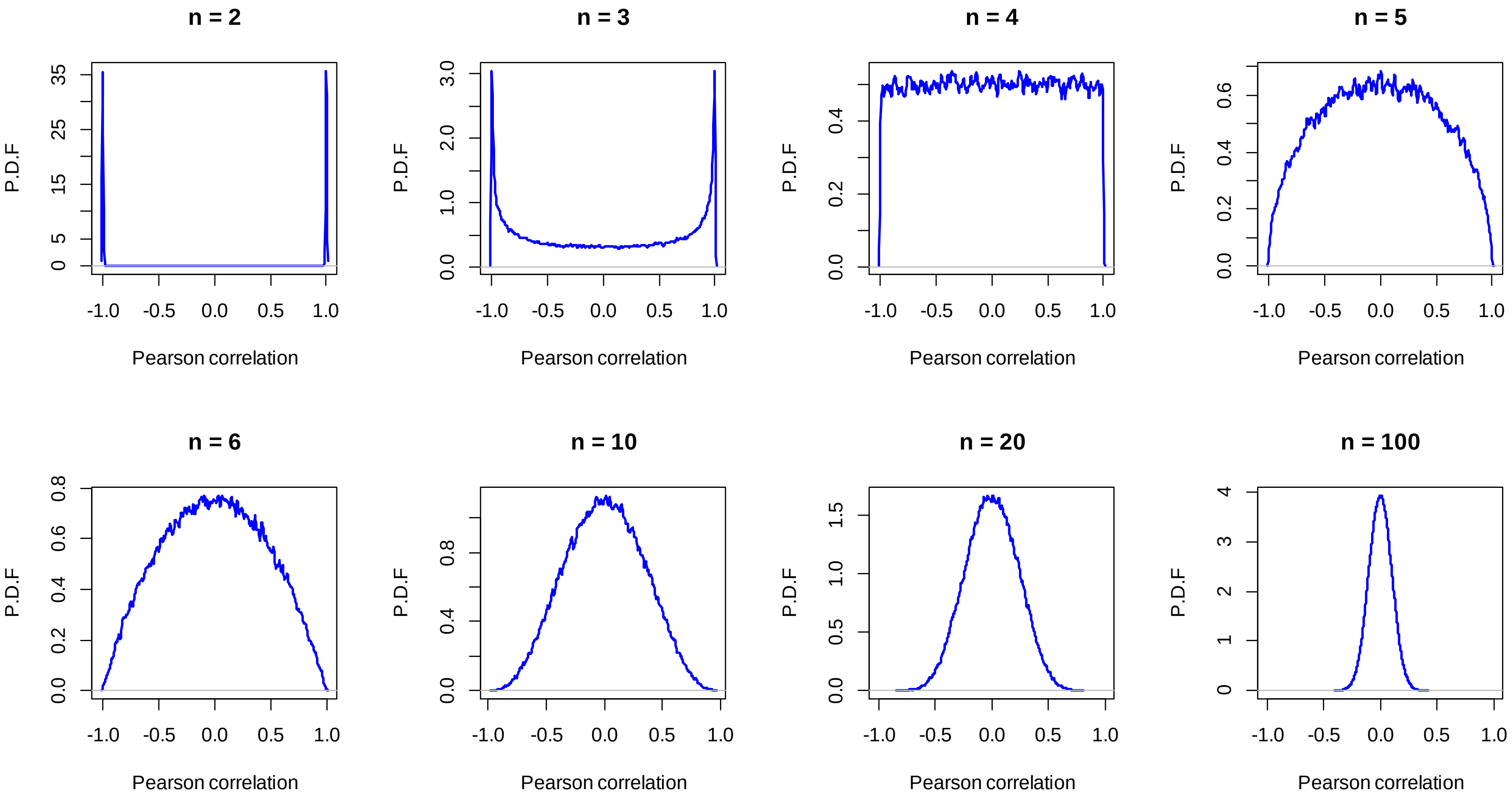
? If we have only 2 data points in *x* and y datasets, what values would you expect for correlation b/w *x* and y ?

UNIVERSITÉ DU LUXEMBOURG

L2.1. Descriptive Statistics in R

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Pearson Correlation: Effect of Sample Size



L2.1. Descriptive Statistics in R

Nonparametric Measures of Association

Kendal Correlation, τ (Kendall tau rank correlation)

a non-parametric measure of rank correlation: that is, the similarity of the orderings of the data when ranked by each of the quantities.

All combination of data pairs (x_i, y_i), (x_j, y_j) are checked.

2 pairs are concordant if:

$$(x_i - x_j)(y_i - y_i) > 0$$

2 pairs are discordant if:

$$(x_i - x_j)(y_i - y_i) < 0$$

In case of = 0 pair is not considered.

Let number of corresponding pairs be $n_{concordant}$ and $n_{discordant}$

$$\tau = 2\frac{n_{concordant} - n_{discordant}}{n(n-1)}$$

Spearman's Correlation, ρ (Spearman's rank correlation)

a non-parametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function.

Data (x_i, y_i) are replaced by their ranks, let's denote them (X_i, Y_i). Then Person correlation is measured b/w ranks:

$$\rho_{xy} = \frac{\sum(X_i - m_X)(Y_i - m_Y)}{\sum(X_i - m_Y)\sum(Y_i - m_Y)}$$

In R use:

- `cor(x,method="kendal",...)`
- `cor(x,method="spearman",...)`

For missing data add parameter

`use = "pairwise.complete.obs"`

L2.1. Descriptive Statistics in R

Please go through the code at:

http://edu.sablab.net/abs2017/scripts2.html

Section 2.1

Do Exercises 2.1

L2.2. Statistical Tests in R

- **Hypotheses testing for means and proportions**
  - Hypotheses about means
  - Hypotheses about proportions
  - 1-tail vs. 2-tail
- **Hypotheses testing for means of 2 populations**
  - Independent and matched samples
  - Unpaired t-test
  - Paired t-test
  - Hypotheses about proportions of 2 populations
- **Testing hypothesis about variances of 2 populations**
- **Testing hypothesis about correlations**
- **Power of a test**

L2.2. Hypotheses about Mean of a Population

Hypotheses

Here we continue the discussion of statistical inference by showing how hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected.

In hypothesis testing we begin by making a tentative assumption about a population parameter, i.e. by formulation of a null hypothesis.

Null Hypothesis

The hypothesis tentatively assumed true in the hypothesis testing procedure, H_0

Alternative hypothesis

The hypothesis concluded to be true if the null hypothesis is rejected, H_a

H_0: $\mu \leq$ const

H_a: $\mu >$ const

H_0: $\mu \geq$ const

H_a: $\mu <$ const

H_0: $\mu =$ const

H_a: $\mu \neq$ const

H_0: $\mu_1 \leq \mu_2$

H_a: $\mu_1 > \mu_2$

H_0: $\mu_1 \geq \mu_2$

H_a: $\mu_1 < \mu_2$

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

L2.2. Hypotheses about Mean of a Population

Errors

Type I error

The error of rejecting H_0 when it is true.

Type II error

The error of accepting H_0 when it is false.

Level of significance

The probability of making a Type I error when the null hypothesis is true as an equality, α

poor sensitivity

False Negative, β error

		Population Condition	
		H_0 **True**	H_a **True**
Conclusion	**Accept** H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

False Positive, α error

poor specificity

L2.2. Hypotheses about Mean of a Population

One-tailed Test for Mean

One-tailed test

A hypothesis test in which rejection of the null hypothesis occurs for values of the test statistic in one tail of its sampling distribution

$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$

A Trade Commission (TC) periodically conducts statistical studies designed to test the claims that manufacturers make about their products. For example, the label on a large can of Hilltop Coffee states that the can contains 3 pounds of coffee. The TC knows that Hilltop's production process cannot place exactly 3 pounds of coffee in each can, even if the mean filling weight for the population of all cans filled is 3 pounds per can. However, as long as the population mean filling weight is at least 3 pounds per can, the rights of consumers will be protected. Thus, the TC interprets the label information on a large can of coffee as a claim by Hilltop that the population mean filling weight is at least 3 pounds per can. We will show how the TC can check Hilltop's claim by conducting a lower tail hypothesis test.

μ_0 = 3 lbm

Suppose sample of n=36 coffee cans is selected. From the previous studies it’s known that σ = 0.18 lbm

L2.2. Hypotheses about Mean of a Population

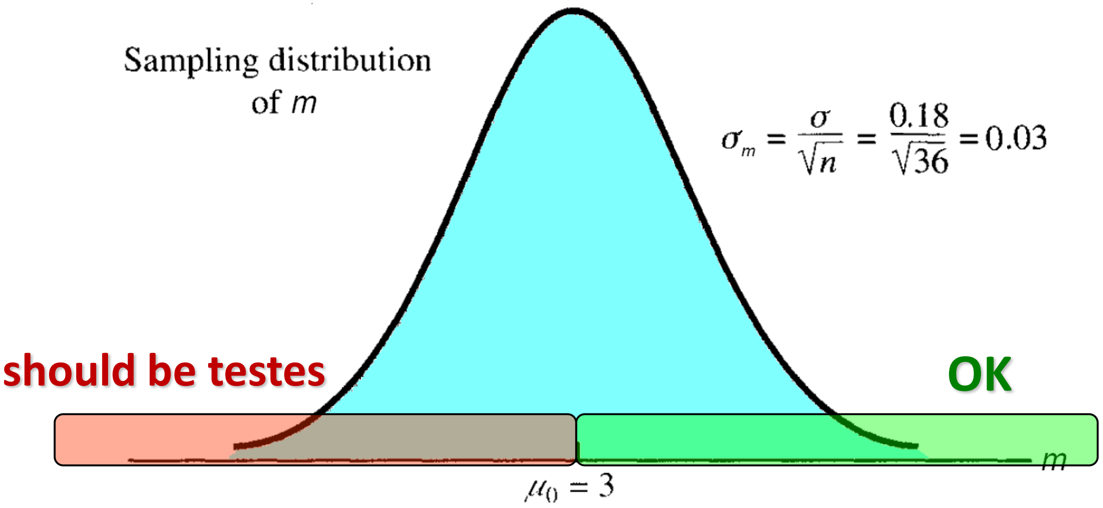
One-tailed Test: Example

μ_0 = 3 lbm

H_0: $\mu \geq 3$ **no action**

H_a: $\mu < 3$ **legal action**

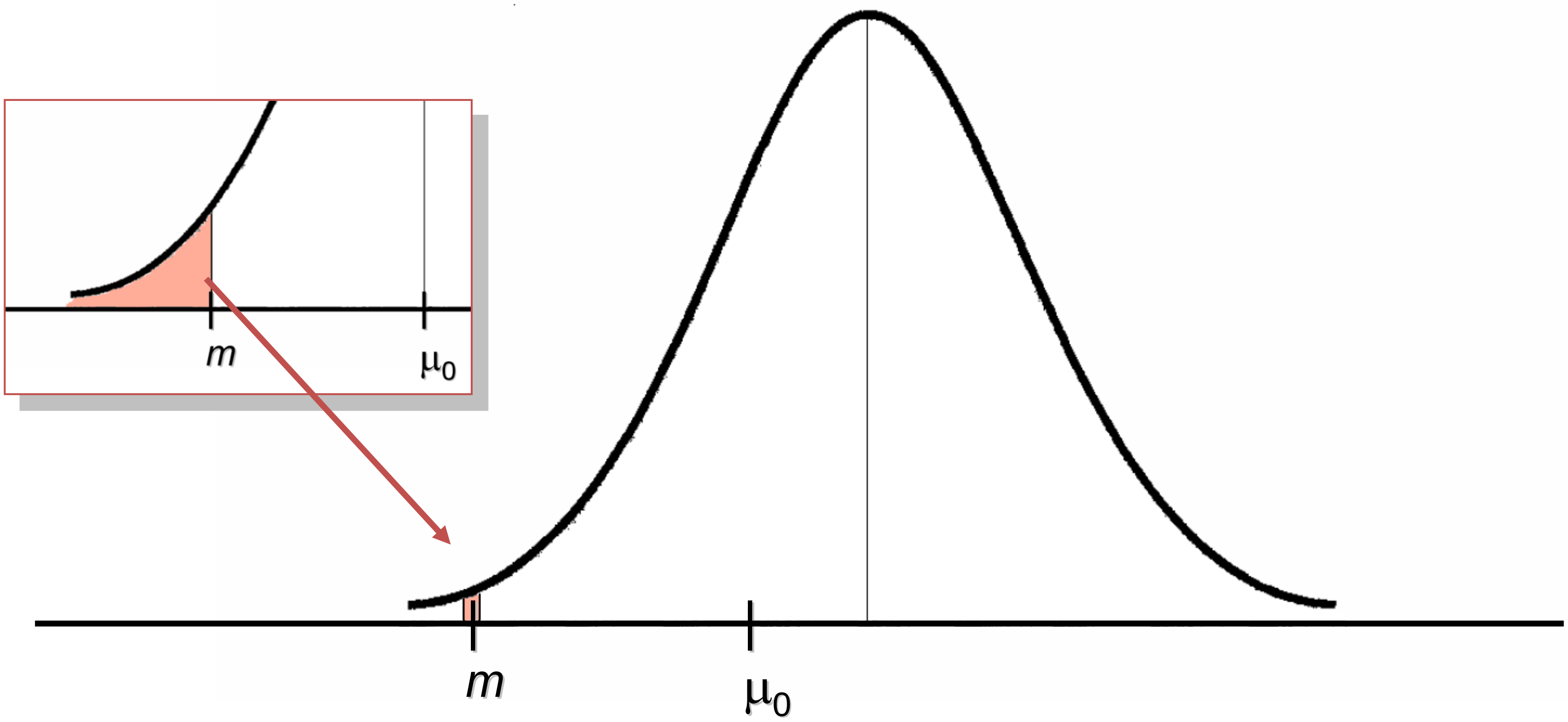
Let's say: in the extreme case, when μ=3, we would like to be 99% sure that we make no mistake, when starting legal actions against Hilltop Coffee. It means that selected significance level is **α = 0.01**


Sampling distribution
of m
$\sigma_m = \frac{\sigma}{\sqrt{n}} = \frac{0.18}{\sqrt{36}} = 0.03$
should be testes
OK
m
$\mu_0 = 3$

L2.2. Hypotheses about Mean of a Population

One-tailed Test: Example

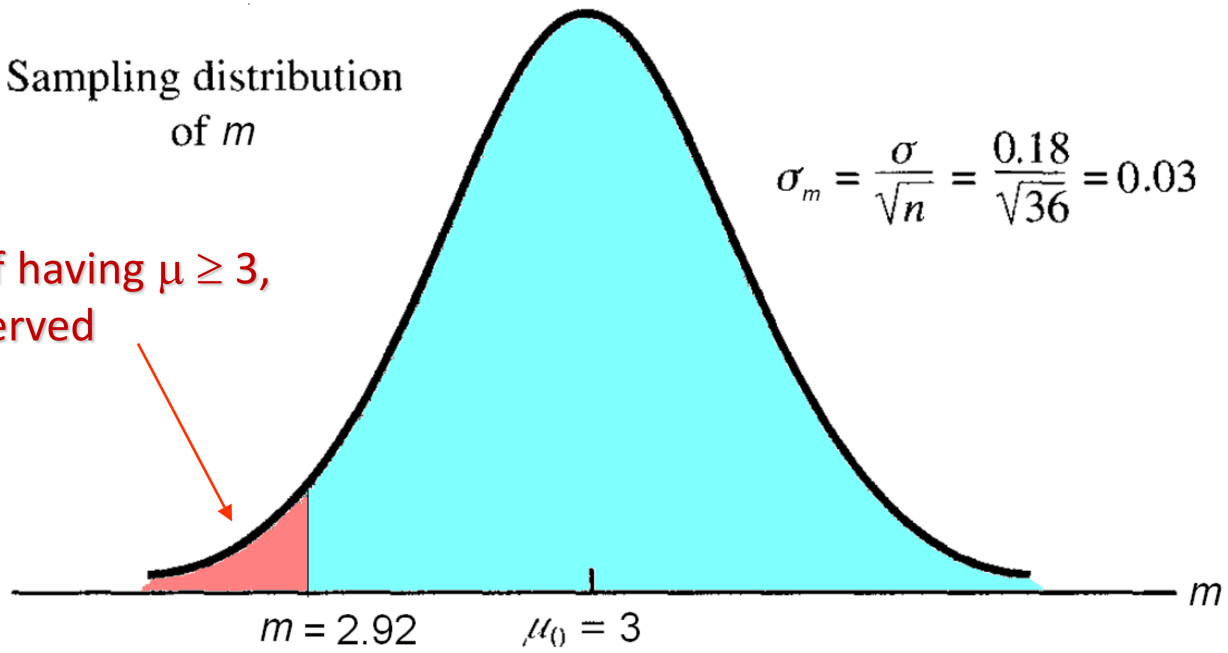
Let's find the probability of observation *m* for all possible $\mu \geq 3$. We start from an extreme case ($\mu=3$) and then probe all possible $\mu>3$. See the behavior of the small probability area around measured *m*. What you will get if you summarize its area for all possible $\mu \geq 3$?



P(*m*) for all possible $\mu \geq \mu_0$ is equal to *P*(*x*<*m*) for an extreme case of $\mu=\mu_0$

L2.2. Hypotheses about Mean of a Population

p-value



The probability of having μ ≥ 3, if *m* = 2.92 is observed

In other words, red area characterizes the probability of the null hypothesis.

To be completely correct, the **red area** gives us a **probability of making an error** when rejecting the null hypothesis, or the **p-value.**

L2.2. Hypotheses about Mean of a Population

Pipeline to Test Hypothesis about Population Mean (manual)

if σ in unknown:
σ → *s*
z → *t*

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0 : \mu \geq \mu_0$ $H_a : \mu < \mu_0$	$H_0 : \mu \leq \mu_0$ $H_a : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$
Test Statistic	$t = \frac{m - \mu_0}{s/\sqrt{n}}$	$t = \frac{m - \mu_0}{s/\sqrt{n}}$	$t = \frac{m - \mu_0}{s/\sqrt{n}}$
Rejection Rule: **p-Value Approach**	Reject H_0 if p-value ≤ α	Reject H_0 if p-value ≤ α	Reject H_0 if p-value ≤ α
Rejection Rule: **Critical Value Approach**	Reject H_0 if $t \leq -t_\alpha$	Reject H_0 if $t \geq t_\alpha$	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

L2.2. Hypotheses about Mean of a Population

Pipeline to Test Hypothesis about Population Mean (R)

In fact in R it is much simpler:

In R use (parametric):

- **`t.test(x, mu =`** μ_0, alternative =…)

In R use (non parametric):

- **`wilcox.test(x, mu =`** μ_0, alternative =…)

alternative = c("**two.sided**", "less", "greater")

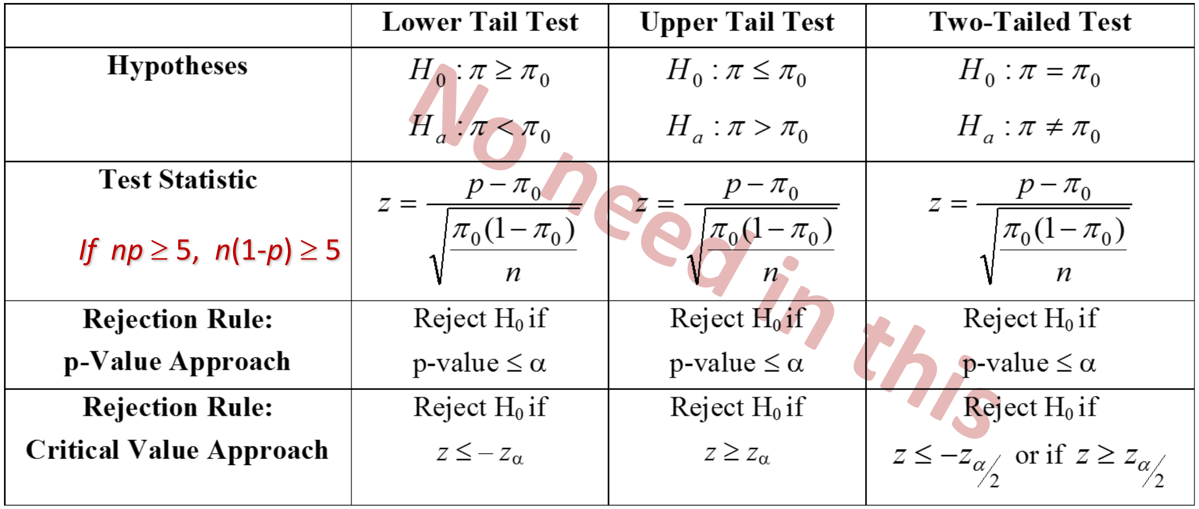
Pipeline to Test Hypothesis about Population Proportion (manual)

Proportions

π – population proportion

p – sample proportion

π_0 – testing value



	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0 : \pi \ge \pi_0$ $H_a : \pi < \pi_0$	$H_0 : \pi \le \pi_0$ $H_a : \pi > \pi_0$	$H_0 : \pi = \pi_0$ $H_a : \pi \ne \pi_0$
Test Statistic *If* $np \ge 5$, $n(1\text{-}p) \ge 5$	$z = \dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$	$z = \dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$	$z = \dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\le \alpha$	Reject H_0 if p-value $\le \alpha$	Reject H_0 if p-value $\le \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $z \le -z_\alpha$	Reject H_0 if $z \ge z_\alpha$	Reject H_0 if $z \le -z_{\alpha/2}$ or if $z \ge z_{\alpha/2}$

No need in this

Is used with big **n**

In R use:

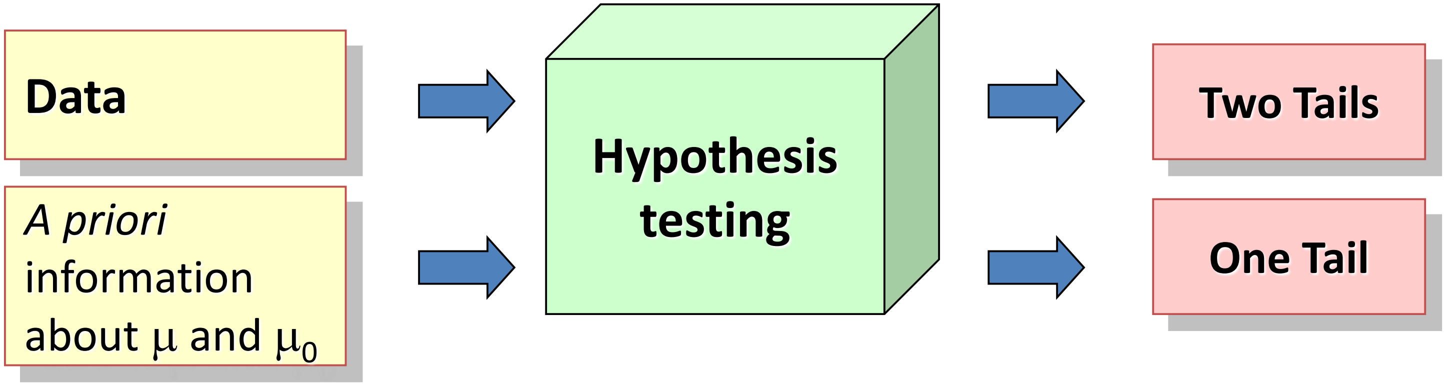
- `prop.test(x, n, p =` π_0 `)`
- `binom.test(x, n, p =` π_0 `)`

Exact test, works always (but is 10 times slower than prop.test)

L2.2. Hypotheses about Mean of a Population

One-tail Test vs. Two-tail Test

There is a raging controversy (for about the last hundred years) on whether or not it is ever appropriate to use a one-tailed test. The rationale is that if you already know the direction of the difference, why bother doing any statistical tests. While it is **generally safest to use a two-tailed tests**, there are situations where a one-tailed test seems more appropriate. The bottom line is that **it is the choice of the researcher** whether to use one-tailed or two-tailed research questions.



$$2\times\text{p-value}_{(1\ \text{tail})} = \text{p-value}_{(2\ \text{tails})}$$

Reminder: discussion around NDAs submitted to FDA, USA

L2.2. Hypotheses about Mean of a Population

Example: Hypothesis about Mean

Number of living cells in **5 wells** under some conditions are given in the table, with average value of **4705**. In a reference literature source authors clamed a mean quantity of **5000** living cells under the same conditions. Is our result significantly different?

Well	Cells
1	5128
2	4806
3	5037
4	4231
5	4322

Two Tails

H_0: $\mu = 5000$

H_a: $\mu \neq 5000$

Let’s use α=0.05

$$t = \frac{m - \mu_0}{s/\sqrt{n}}$$

n	5
mean	4704.8
stdev	409.49
mu	5000
t	-1.612
p-value 2 t	0.1823
p-value 1 t	0.0911

```
x =c(5128,4806,5037,4231,4222)
n=length(x)
m=mean(x)
s=sd(x)
mu=5000
t=(m-mu)/s*sqrt(n)
p.val.1 = pt(t,df=n-1)
p.val.2 = 2*pt(t,df=n-1)
```

In R use:

```
t.test(x,mu=5000)
```

L2.2. Hypotheses about Mean of a Population

Example: Hypothesis about Proportion

During a study of a new drug against viral infection, you have found that **70 out of 100** mice survived, whereas the survival after the standard therapy is **60%** of the infected population. Is this enhancement statistically significant? Use error level α=0.05

One Tail

H_0: $\pi \le 0.6$

H_a: $\pi > 0.6$

Let's use α=0.05

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

In R use:

- `prop.test(…)`
- `binom.test(…)`

```
p   = 0.7
p0  = 0.6
sp  = sqrt(p0*(1-p0)/n)
z   = (p-p0)/sp
p.val.1 = 1-pnorm(z)
```

```
> p.val.1
[1] 0.02061342
```

```
prop.test(x=70,n=100,p=0.6,
                    alternative="greater")
```

```
data:  70 out of 100, null probability 0.6
X-squared = 3.7604, df = 1, p-value = 0.02624
alternative hypothesis: true p is greater
than 0.6
```

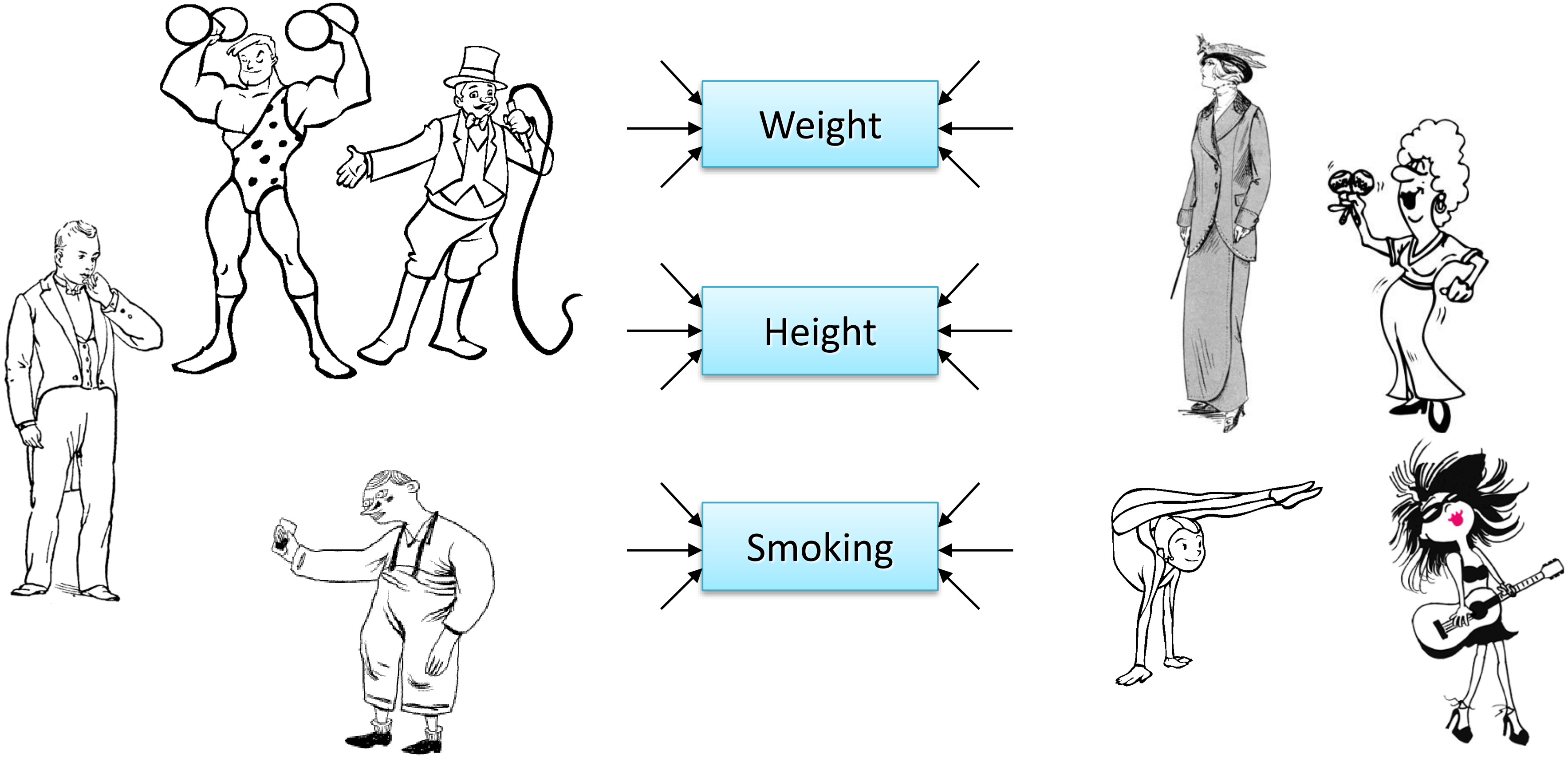
Discrepancy for prop.test() comes from continuity correction.

L2.2. Hypotheses about Means of 2 Populations

Independent Samples

Independent samples

Samples selected from two populations in such a way that the elements making up one sample are chosen independently of the elements making up the other sample.



Weight

Height

Smoking

L2.2. Hypotheses about Means of 2 Populations

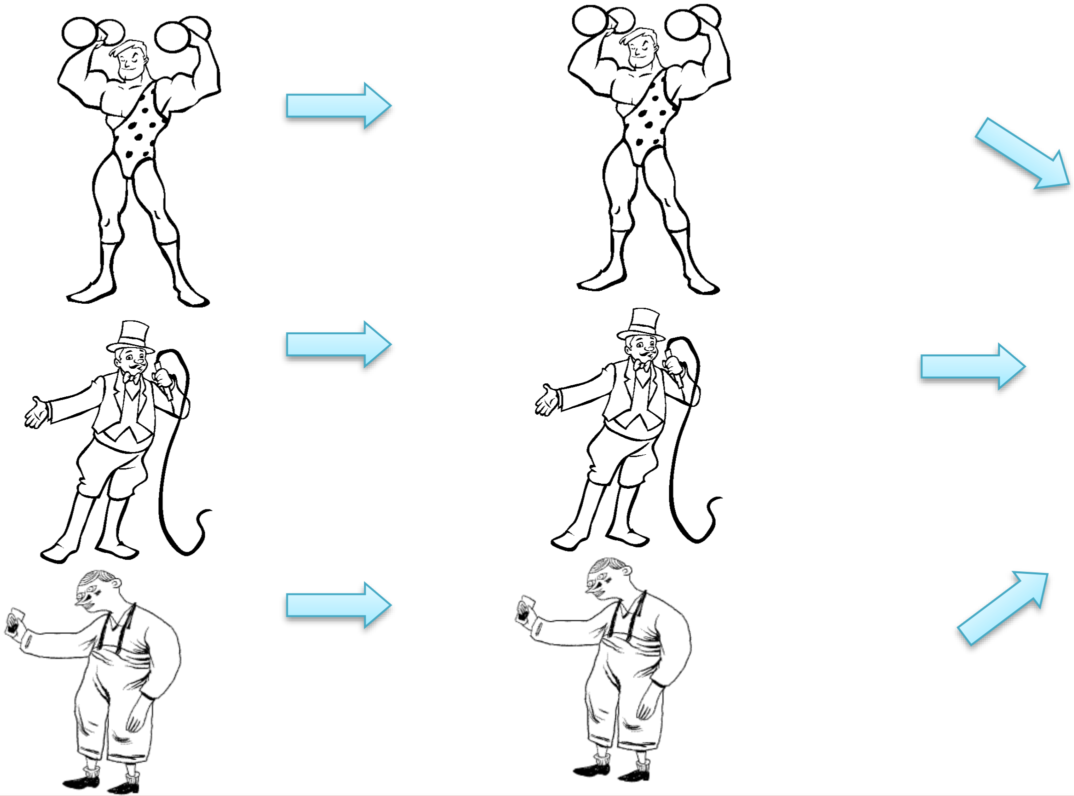
Matched Samples

Matched samples

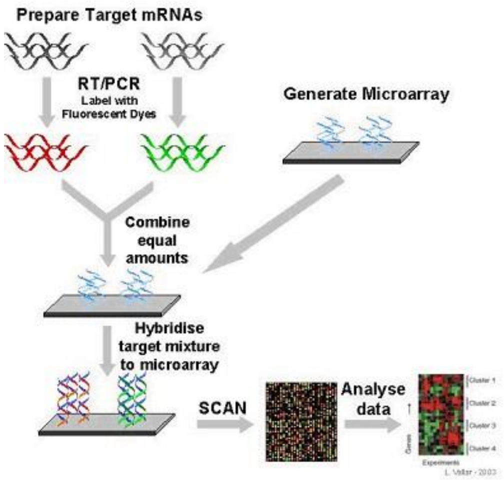
Samples in which each data value of one sample is matched with a corresponding data value of the other sample.

Before treatment

After treatment



Analysis



L2.2. Hypotheses about Means of 2 Populations

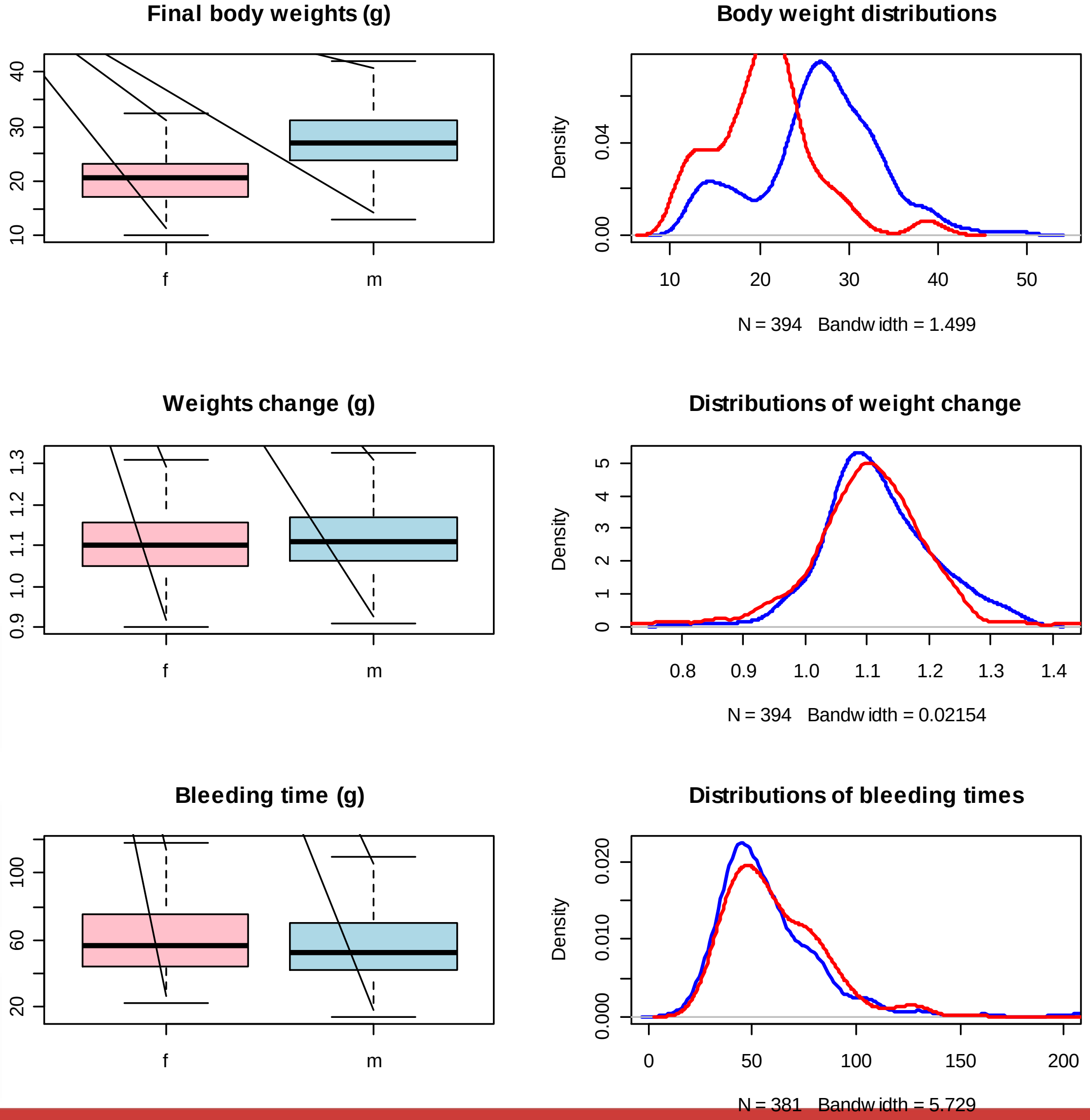
Example: Independent Samples

mice.txt

Q1: Is **body weight** for male and female significantly different?

Q2: Is **weight change** for male and female significantly different?

Q3: Is **bleeding time** for male and female significantly different?



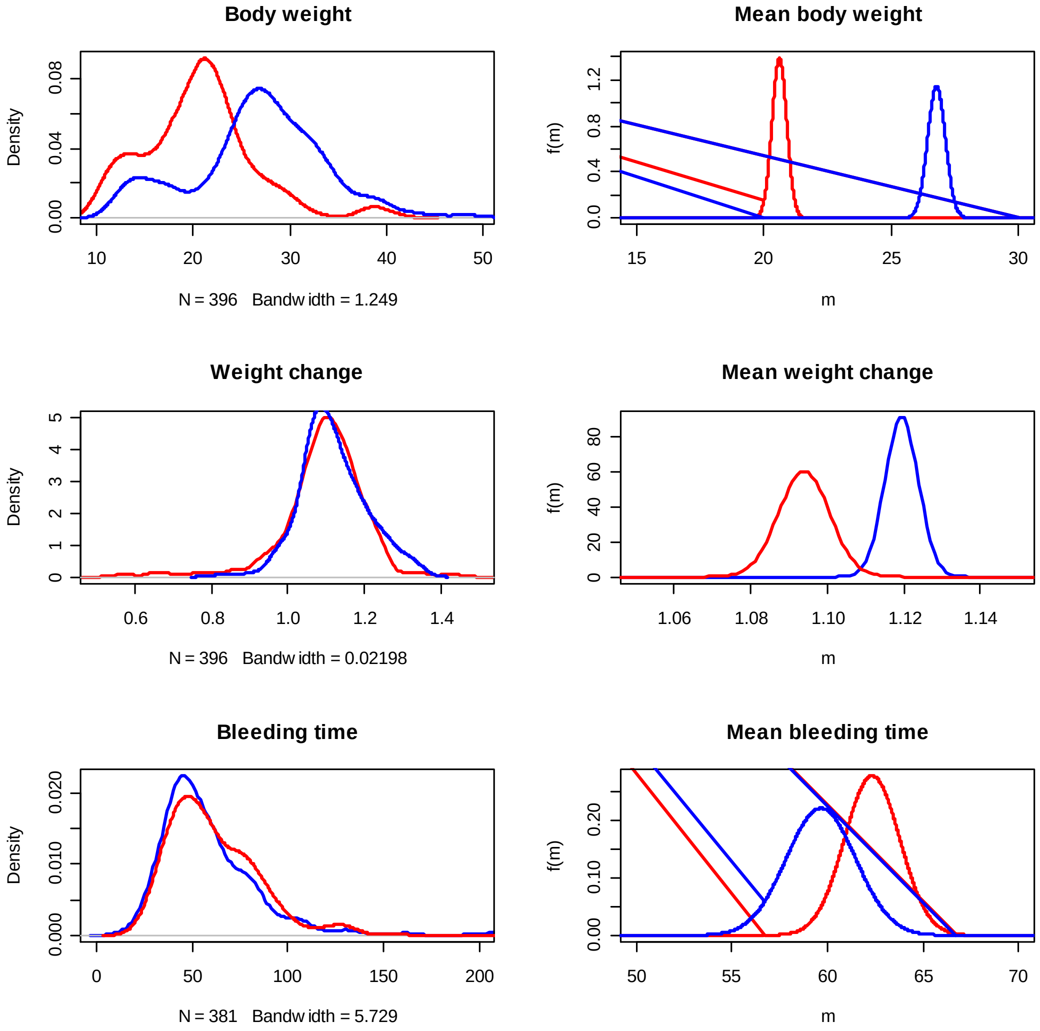
L2.2. Hypotheses about Means of 2 Populations

Example: Independent Samples

Q1: Is **body weight** for male and female significantly different?

Q2: Is **weight change** for male and female significantly different?

Q3: Is **bleeding time** for male and female significantly different?



L2.2. Hypotheses about Means of 2 Populations

Algorithm (manual… only for stat-geeks ☺)

H_0: $\mu_2 - \mu_1 = D_0$

H_a: $\mu_2 - \mu_1 \neq D_0$

Usually $D_0 = 0$

$$s_{m_2 - m_1} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

1. Build the statistics to be used for hypothesis testing:

$$t = \frac{m_2 - m_1 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-distribution has following degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$df = (n-1)\frac{\left(s_1^2 + s_2^2\right)^2}{\left(s_1^4 + s_2^4\right)}$$

$(n_1+n_2)/2 < \boldsymbol{df} < n_1+n_2$

2. Calculate p-value

```
= 2*pt(-abs(t),df)
```

In Excel use:

```
=T.TEST(x,y,2,3)
```

L2.2. Hypotheses about Means of 2 Populations

In R

In fact we do not need these calculations:

In R use (parametric):

```
t.test( x, y, alternative=…)
```

In R use (non-parametric):

```
wilcox.test ( x, y, alternative=…)
```

L2.2. Hypotheses about Means of 2 Populations

Paired Samples

`bloodpressure.txt`

Systolic blood pressure (mmHg)

Subject	BP before	BP after
1	122	127
2	126	128
3	132	140
4	120	119
5	142	145
6	130	130
7	142	148
8	137	135
9	128	129
10	132	137
11	128	128
12	129	133

Test	p-value
unpaired	0.414662
paired	0.014506

The systolic blood pressures of n=12 women between the ages of 20 and 35 were measured before and after usage of a newly developed oral contraceptive.

Q: Does the treatment affect the systolic blood pressure?

Unpaired test

```
= t.test (x, y)
```

Paired test

```
= t.test (x, y, paired=T)
```

In R use (parametric):

- `t.test(x, y, paired=T)`

In R use (non-parametric):

- `wilcox.test(x, y, paired=T)`

L2.2. Hypotheses about Means of 2 Populations

Hypothesis about Proportions of 2 Populations

$H_0: \pi_1 = \pi_2$

$H_a: \pi_1 \neq \pi_2$

$H_0: \pi_1 - \pi_2 = 0$

$H_a: \pi_1 - \pi_2 \neq 0$

$$\sigma_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Pooled estimator of π

An estimator of a population proportion obtained by computing a weighted average of the point estimators obtained from two independent samples.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\sigma_{p_1-p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

◆ `= 2*pnorm(-abs(z))`

◆ `= 2*pnorm(-abs(z))`

L2.2. Hypotheses about Means of 2 Populations

Example: Hypothesis about Proportions of 2 Populations

SWR/J	MA/MyJ
f	f
f	f
f	f
f	f
f	f
f	f
f	f
f	f
f	m
f	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
	m
	m
	m
	m

mice.txt

Q: Is the male proportion significantly different in these mouse strains (0.47 and 0.65)?

	SWR/J	MA/MyJ	*pooled*
count male	9	15	*24*
n	19	23	*42*
p	0.474	0.652	*0.571*
z	-1.16		
p-val	**0.244658997**		

In R use:

- `prop.test(...)`

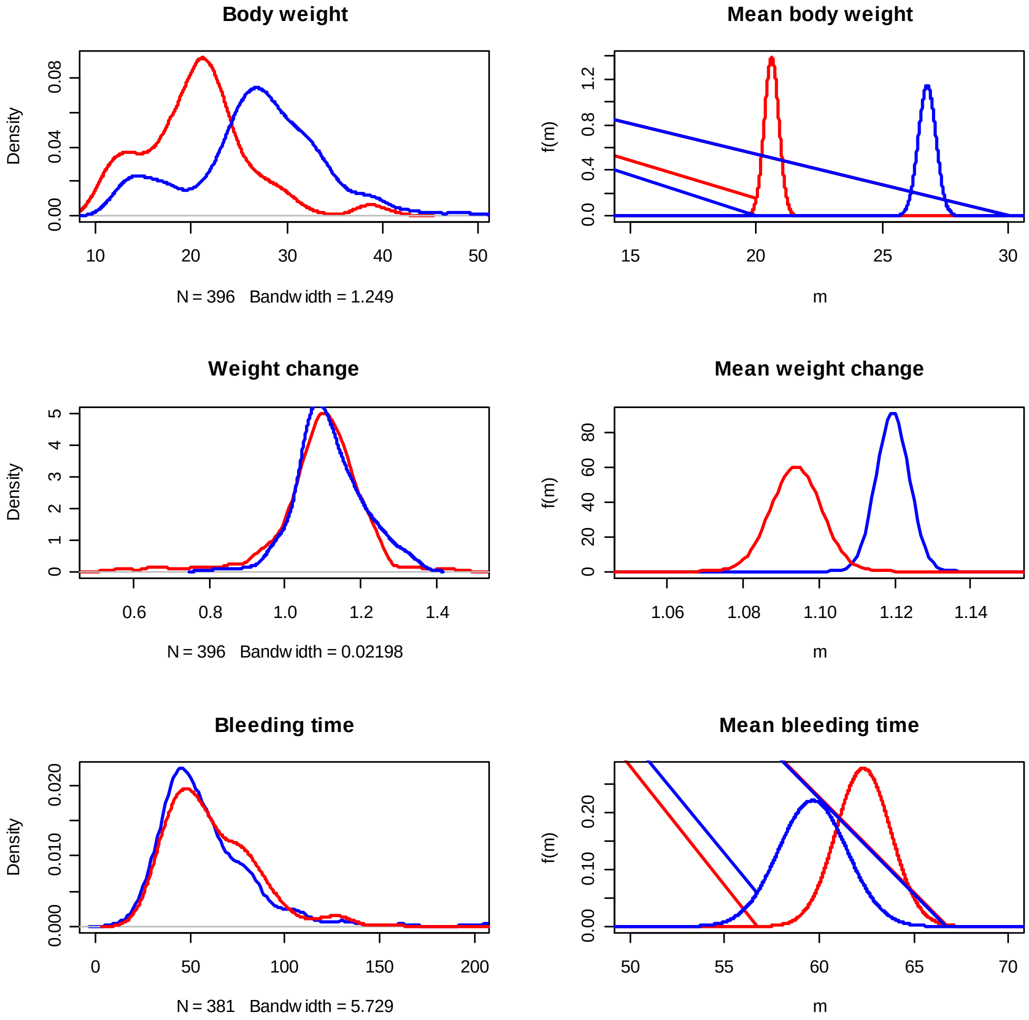
```
prop.test(c(9,15),n=c(19,23),correct=F)
prop.test(c(9,15),n=c(19,23))
```

p-value = 0.3952

Discrepancy comes from continuity correction.

L2.2. Hypotheses about Means of 2 Populations

Non-parametric Tests



- **T-test**, p-val < 2.2e-16
- **Wilcox**, p-val < 2.2e-16

- **T-test**, p-val = 0.0014
- **Wilcox**, p-val = 0.0299

Explanations?

- **T-test**, p-val = 0.2487
- **Wilcox**, p-val = 0.0178

In R use:

- `wilcox.test(x,y)`

L2.2. Testing Hypothesis about Variances

Hypotheses about Population Variance

H_0: $\sigma^2 \leq$ const

H_a: $\sigma^2 >$ const

H_0: $\sigma^2 \geq$ const

H_a: $\sigma^2 <$ const

H_0: $\sigma^2 =$ const

H_a: $\sigma^2 \neq$ const

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0 : \sigma^2 \geq \sigma_0^2$ $H_a : \sigma^2 < \sigma_0^2$	$H_0 : \sigma^2 \leq \sigma_0^2$ $H_a : \sigma^2 > \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_a : \sigma^2 \neq \sigma_0^2$
Test Statistic	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Rejection Rule: **p-Value Approach**	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: **Critical Value Approach**	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha)}$	Reject H_0 if $\chi^2 \geq \chi^2_{\alpha}$	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha/2)}$ or if $\chi^2 \geq \chi^2_{\alpha/2}$

L2.2. Testing Hypothesis about Variances

Sampling Distribution

In many statistical applications we need a comparison between variances of two populations. In fact well-known ANOVA-method is base on this comparison.

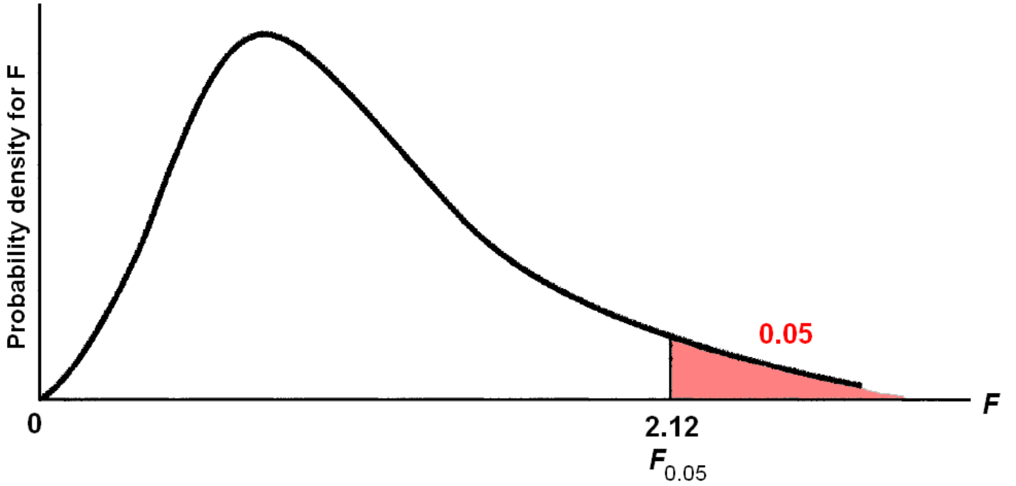
$$F = \frac{s_1^2}{s_2^2}$$

The statistics is build for the following measure:

Sampling distribution of s_1^2/s_2^2 when $\sigma_1^2 = \sigma_2^2$

Whenever a independent simple random samples of size n_1 and n_2 are selected from two normal populations with equal variances, the sampling of s_1^2/s_2^2 has F-distribution with n_1-1 degree of freedom for numerator and n_2-1 for denominator.

F-distribution for 20 d.f. in numerator and 20 d.f. in denominator



In Excel use functions:

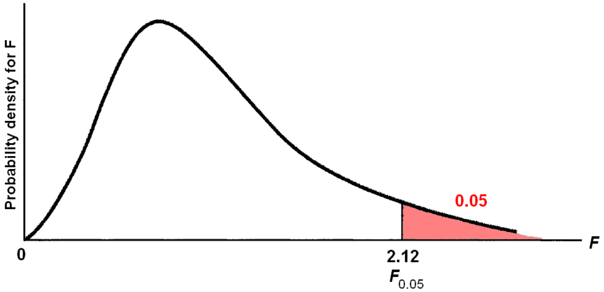
- = FTEST(x,y)

In R use:

- `var.test(x,y)`

L2.3. Testing Hypothesis about Variances

Hypotheses about Variances of Two Populations



H_0: $\sigma_1^2 \leq \sigma_2^2$

H_a: $\sigma_1^2 > \sigma_2^2$

H_0: $\sigma_1^2 = \sigma_2^2$

H_a: $\sigma_1^2 \neq \sigma_2^2$

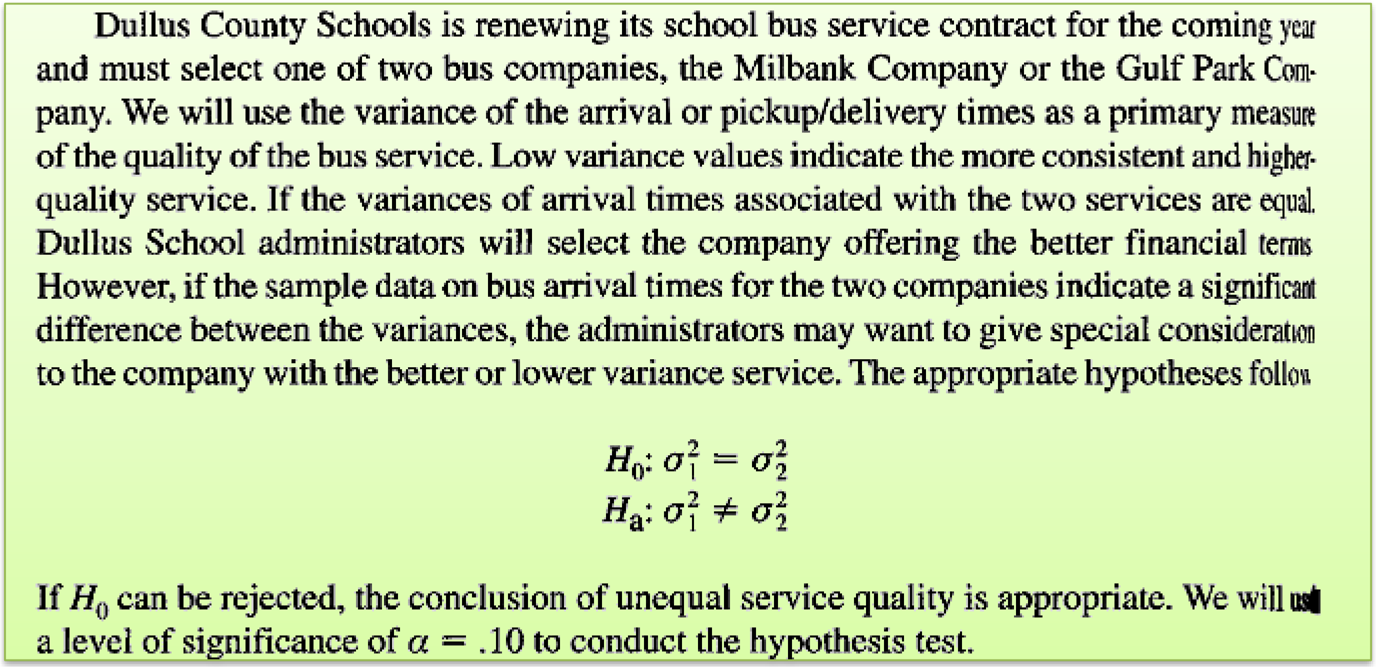
	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \sigma_1^2 \leq \sigma_2^2$ $H_a: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$ *Note: Population 1 has the lager sample variance*
Test Statistic	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value ≤ α	Reject H_0 if p-value ≤ α
Rejection Rule: Critical Value Approach	Reject H_0 if $F \geq F_\alpha$	Reject H_0 if $F \geq F_\alpha$

L2.2. Testing Hypothesis about Variances

Example

schoolbus.txt

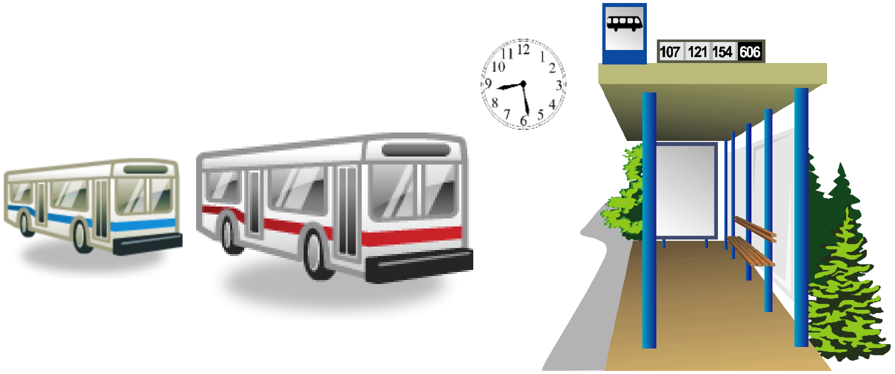
#	Milbank	Gulf Park
1	35.9	21.6
2	29.9	20.5
3	31.2	23.3
4	16.2	18.8
5	19.0	17.2
6	15.9	7.7
7	18.8	18.6
8	22.2	18.7
9	19.9	20.4
10	16.4	22.4
11	5.0	23.1
12	25.4	19.8
13	14.7	26.0
14	22.7	17.1
15	18.0	27.9
16	28.1	20.8
17	12.1	
18	21.4	
19	13.4	
20	22.9	
21	21.0	
22	10.1	
23	23.0	
24	19.4	
25	15.2	
26	28.2	



Dullus County Schools is renewing its school bus service contract for the coming year and must select one of two bus companies, the Milbank Company or the Gulf Park Company. We will use the variance of the arrival or pickup/delivery times as a primary measure of the quality of the bus service. Low variance values indicate the more consistent and higher-quality service. If the variances of arrival times associated with the two services are equal, Dullus School administrators will select the company offering the better financial terms. However, if the sample data on bus arrival times for the two companies indicate a significant difference between the variances, the administrators may want to give special consideration to the company with the better or lower variance service. The appropriate hypotheses follow.

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

If H_0 can be rejected, the conclusion of unequal service quality is appropriate. We will use a level of significance of $\alpha = .10$ to conduct the hypothesis test.



L2.2. Testing Hypothesis about Variances

Example

schoolbus.txt

#	Milbank	Gulf Park
1	35.9	21.6
2	29.9	20.5
3	31.2	23.3
4	16.2	18.8
5	19.0	17.2
6	15.9	7.7
7	18.8	18.6
8	22.2	18.7
9	19.9	20.4
10	16.4	22.4
11	5.0	23.1
12	25.4	19.8
13	14.7	26.0
14	22.7	17.1
15	18.0	27.9
16	28.1	20.8
17	12.1	
18	21.4	
19	13.4	
20	22.9	
21	21.0	
22	10.1	
23	23.0	
24	19.4	
25	15.2	
26	28.2	

1. Let us start from estimation of the variances for 2 data sets

Milbank: $s_1^2 = 48$

Gulf Park: $s_2^2 = 20$

Milbank: $\sigma_1^2 \approx 48$ $(29.5 \div 91.5)$

Gulf Park: $\sigma_2^2 \approx 20$ $(10.9 \div 47.9)$

2. Let us calculate the *F*-statistics

$$F = \frac{s_1^2}{s_2^2} = \frac{48}{20} = 2.40$$

3. … and p-value = 0.08

```
var.test(x,y)

 F test to compare two variances

data:  Bus[, 1] and Bus[, 2]
F = 2.401, num df = 25, denom df = 15, p-value = 0.08105
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8927789 5.7887880
sample estimates:
ratio of variances 
          2.401036
```

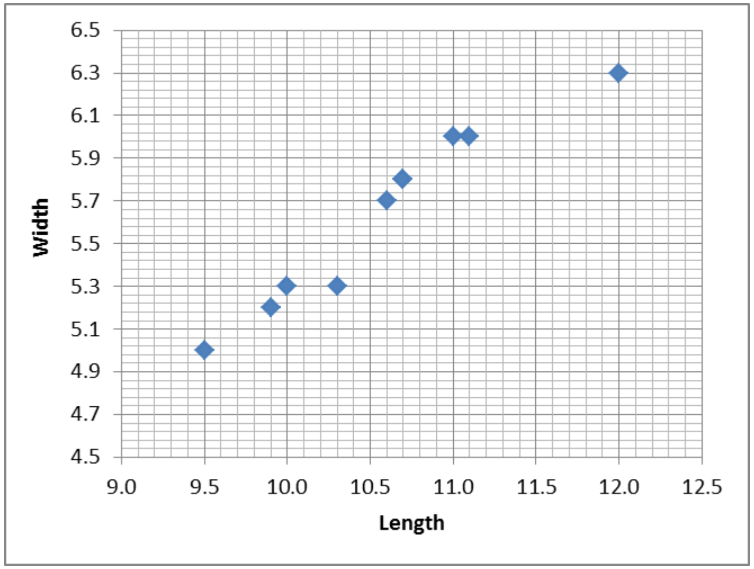
L2.2. Testing Hypothesis about Correlation

Significance of Correlation

A malacologist interested in the morphology of West Indian chitons, *Chiton olivaceous*, measured the length and width of the eight overlapping plates composing the shell of 10 of these animals.

chiton.txt

Length	Width
10.7	5.8
11.0	6.0
9.5	5.0
11.1	6.0
10.3	5.3
10.7	5.8
9.9	5.2
10.6	5.7
10.0	5.3
12.0	6.3



r = 0.9692, is it significant?

Test hypotheses:

H_0: ρ = 0

H_a: ρ ≠ 0

Assume x,y has normal distributions, ρ = 0, then perform a one sample t-test with following parameters:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Degree of freedom $df = n - 2$

L2.4. Testing Hypothesis about Correlation

Significance of Correlation

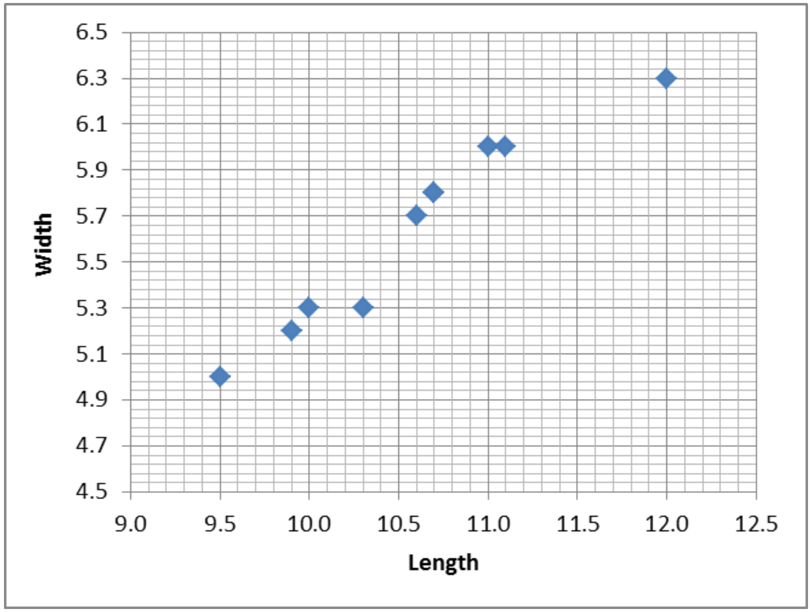
r = 0.9692

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Degree of freedom *df = n - 2*

$$t = \frac{r-0}{s_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

t = 11.14,
p-value = 4e-6

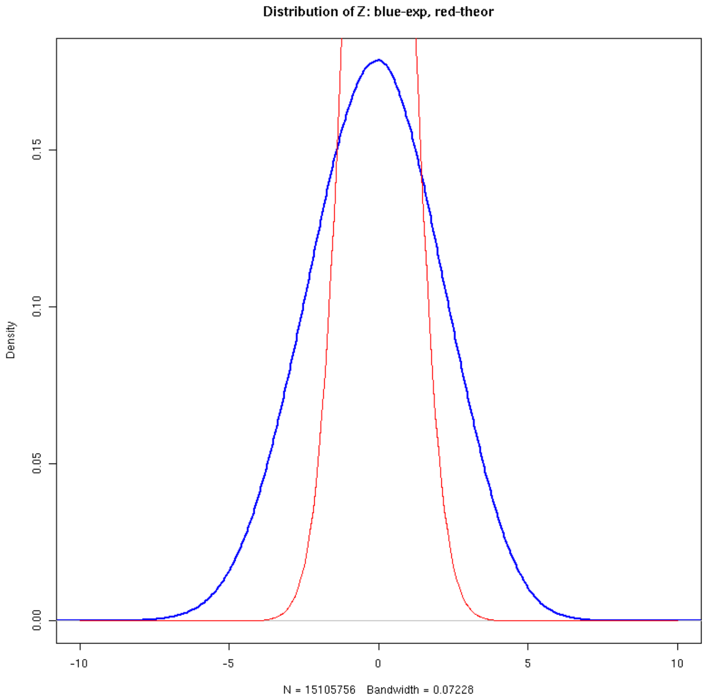


In R use:

```
cor.test(x,y)
```

L2.2. Testing Hypothesis about Correlation

Comparison of 2 Correlations



Fisher's transformation

$$z = 0.5\ln\left(\frac{1+r}{1-r}\right)$$

z-statistics for the difference in correlation

$$Z = \frac{z_1 - z_2}{\sqrt{(n_1-3)^{-1} + (n_2-3)^{-1}}}$$

Use standard normal distribution to assign p-value of identified Z

L2.2. Power of a Test

Type II Error

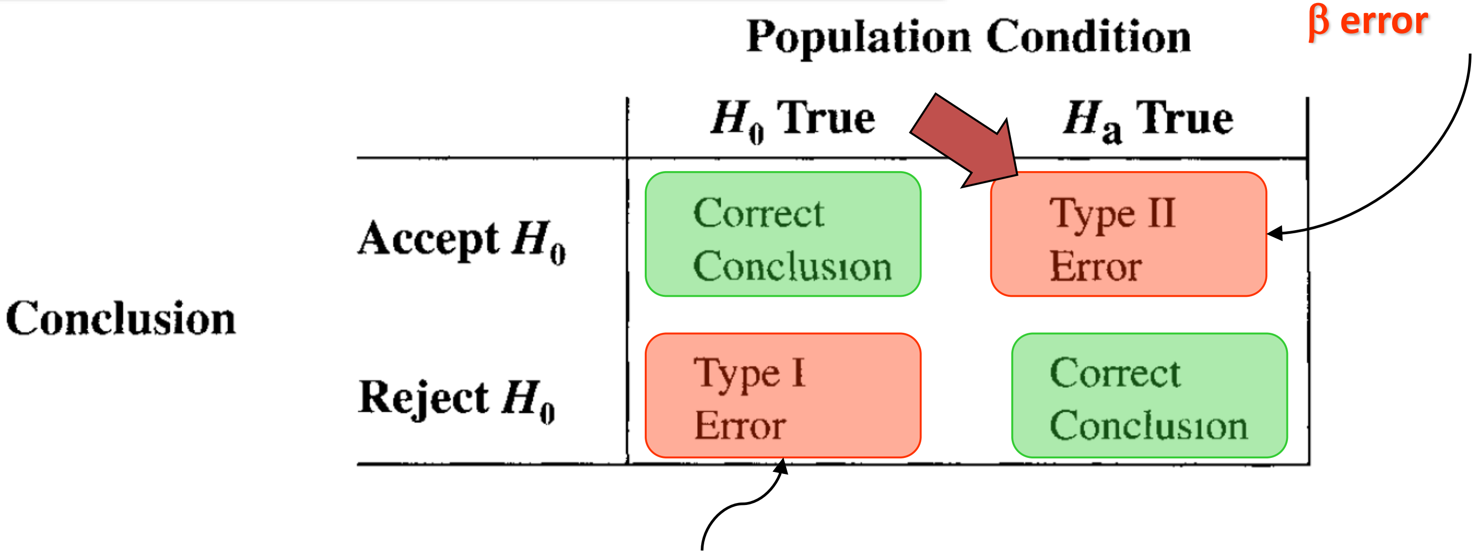
Type I error
The error of rejecting H_0 when it is true.

Type II error
The error of accepting H_0 when it is false.

Level of significance
The probability of making a Type I error when the null hypothesis is true as an equality

poor sensitivity

False Negative,
β error



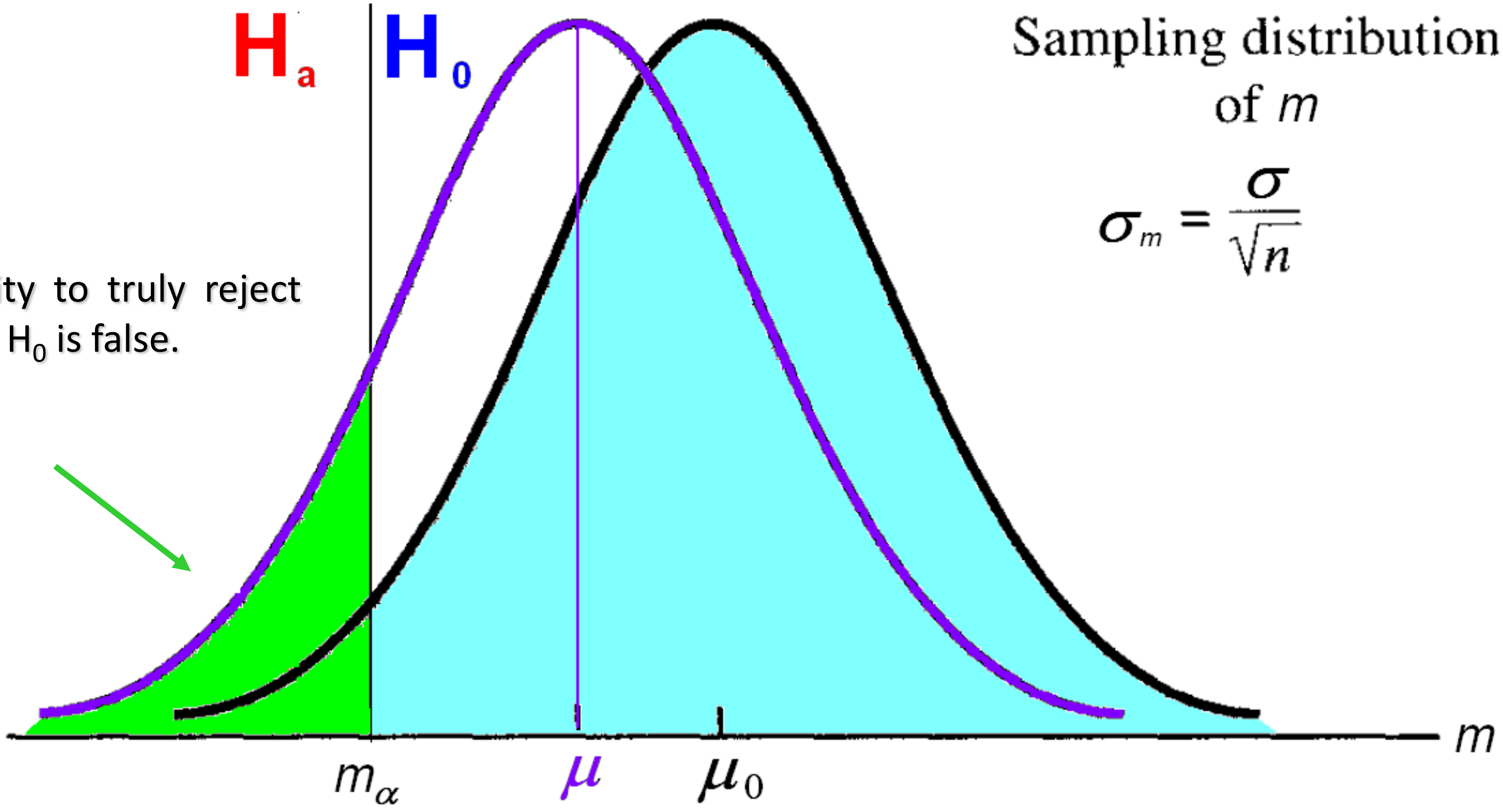
		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

False Positive,
α error

poor specificity

L2.2. Power of a Test

Power Curve


H_a
H_0
Sampling distribution of m
$\sigma_m = \frac{\sigma}{\sqrt{n}}$
m_α
μ
μ_0
m

Probability to truly reject H_0 when H_0 is false.

L2.2. Power of a Test

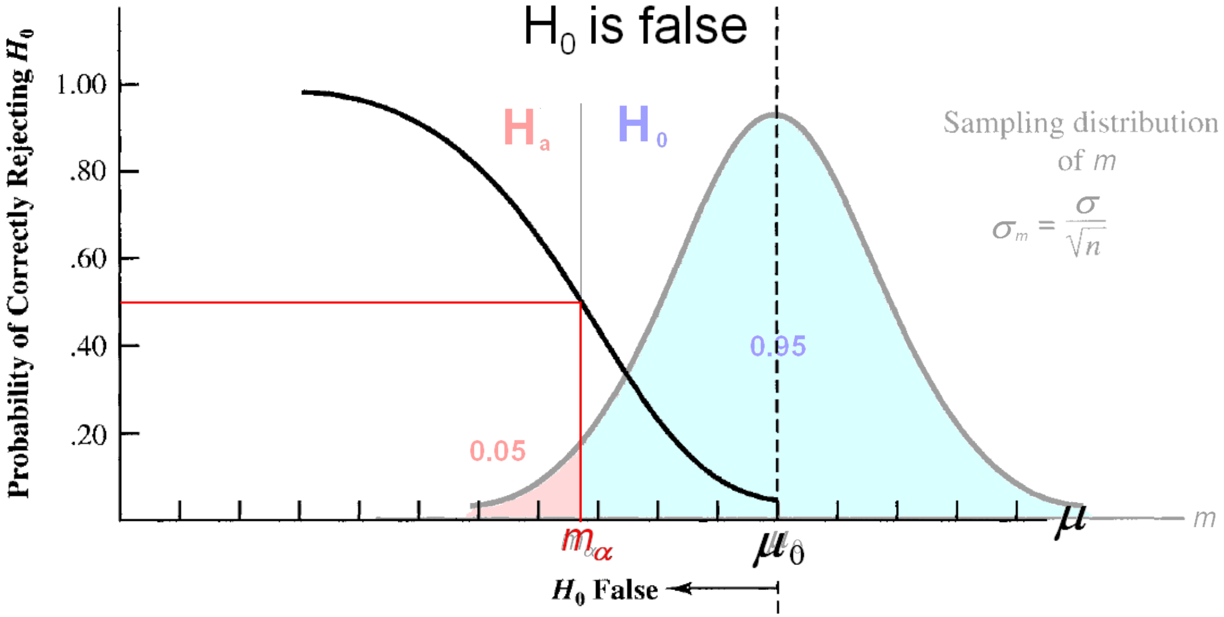
Power Curve

Power

The probability of correctly rejecting H_0 when it is false

Power curve

A graph of the probability of rejecting H_0 for all possible values of the population parameter not satisfying the null hypothesis. The power curve provides the probability of correctly rejecting the null hypothesis



L2.2. Power of a Test

Power Analysis in R

In R use:

- `power.t.test (…)`
- `power.prop.test (…)`

Or pwr package

- `pwr.2p.test(…)`
- `pwr.t.test(…)`
- …

function	power calculations for
pwr.2p.test	two proportions (equal n)
pwr.2p2n.test	two proportions (unequal n)
pwr.anova.test	balanced one way ANOVA
pwr.chisq.test	chi-square test
pwr.f2.test	general linear model
pwr.p.test	proportion (one sample)
pwr.r.test	correlation
pwr.t.test	t-tests (one sample, 2 sample, paired)
pwr.t2n.test	t-test (two samples with unequal n)

http://www.statmethods.net/stats/power.html

L2.2. Statistical Tests in R

Please go through the code at:

http://edu.sablab.net/abs2017/scripts2.html

Section 2.2

Do Exercises 2.2

L2.3. Detection of Outliers

z-score and Chebyshev's Theorem

z-score

A value computed by dividing the deviation about the mean (x_i x) 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.

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

In R use:

- **`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

L2.3. Detection of Outliers

Outliers

For bell-shaped distributions:

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

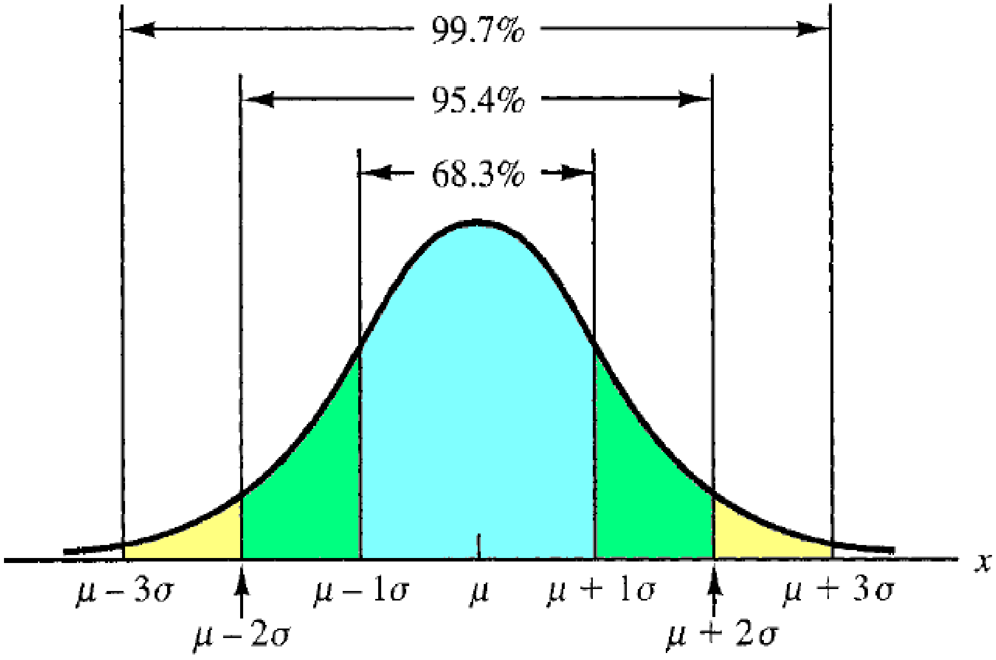
Outlier

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 distribution



L2.3. Detection of Outliers

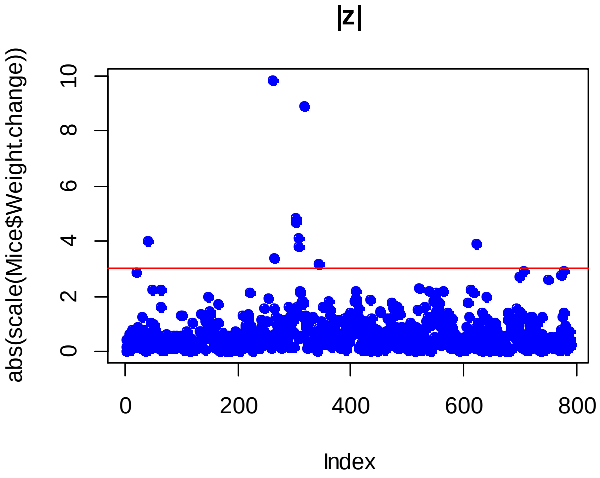
Simplest Method to Detect Outliers

`mice.xls`

Try to identify outlier mice on the basis of *Weight change* variable

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

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



- Calculate z-score by **`scale(...)`**
- Measurements with z-score > 3 are potential outliers

L2.3. Detection of Outliers

Iglewicz-Hoaglin Method

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

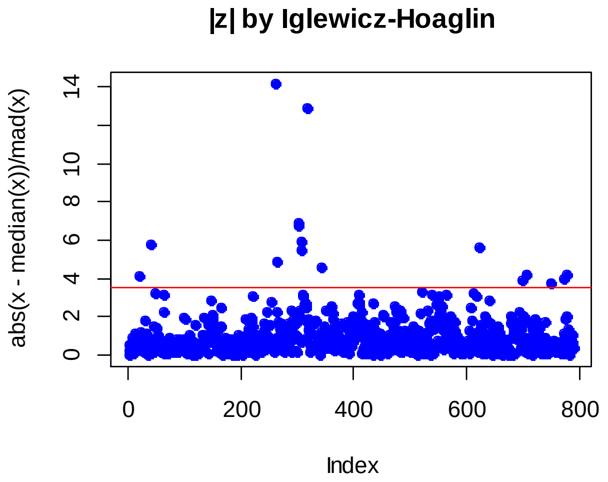
***med*(x)** – median
***MAD*(x)** – median absolute deviation with constant = 1.4826

$$if \;\; |z_i| > 3.5 \Rightarrow x_i \;\; - outlier$$

`mice.xls`

In R use:

- `abs(x-median(x))/mad(x)`



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

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

L2.3. Detection of Outliers

Grubb's 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:

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

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

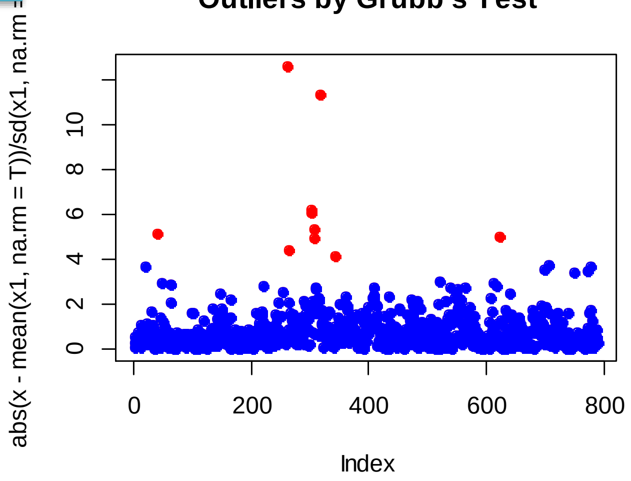
The hypothesis of no outliers is rejected at significance level α if

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

where

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

t – Student statistics


Outliers by Grubb's Test

In R use:

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

L2.3. 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 are other evidences that data come from experimental error – removed.

L2.3. Detection of Outliers

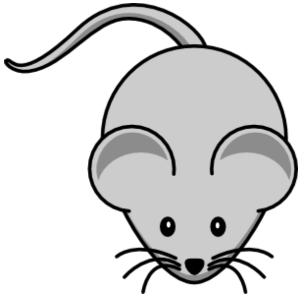
Please go through the code at:

http://edu.sablab.net/abs2017/scripts2.html

Section 2.3

Questions ?

Thank you for your attention



to be continued…