

# Experimental Design and Data Analysis, Lecture 3

Eduard Belitser

VU Amsterdam

# Lecture overview

- ① one sample, normal
  - $t$ -test
- ② one sample, not normal
  - sign test
  - Wilcoxon signed rank test
- ③ two paired samples
  - paired  $t$ -test
  - Pearson's correlation test
  - Spearman's rank correlation test

one sample from a normal distribution

# Setting

An experiment with:

- one **numerical outcome** per experimental unit

Interest is in the **location** of the population distribution.

**EXAMPLE** Measurement of the **height** of 4 years old children.

**EXAMPLE** Measurement of the **time** it takes to find a certain document in a web design for different users.

**EXAMPLE** Measurement of the **yearly amount of sun hours** in different countries.

# Design

- Take a random sample of experimental units from the relevant population
- Measure the outcome on each unit

# Analysis

Data  $(X_1, \dots, X_N)$ .

The **t-test** assumes that the **data**  $X_1, \dots, X_N$  are a random sample from a **normal** population.

We **test** the null hypothesis  $H_0 : \mu = \mu_0$  that the mean of this population is  $\mu_0$ , e.g.  $\mu_0 = 0$ .

The **test statistic** is

$$T = \frac{\bar{X}_N - \mu_0}{S_N},$$

which has the  $t_{N-1}$ -distribution under  $H_0$ .

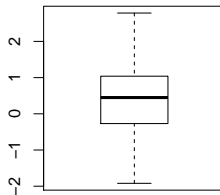
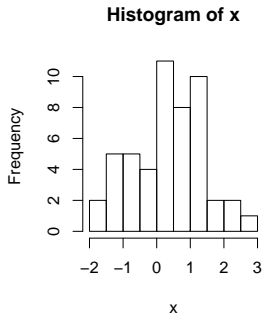
We **estimate**  $\mu$ .

## Analysis in R — data input

```
> mu=0.2  
> x=rnorm(50,mu,1); # creating artificial data
```

# Analysis in R — graphics

```
> par(mfrow=c(1,2))  
> hist(x)  
> boxplot(x)
```





# Analysis in R — estimation and testing

```
> t.test(x)
```

## One Sample t-test

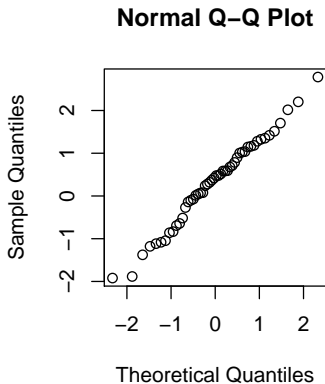
```
data:  x
t = 2.2701, df = 49, p-value = 0.02764
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.03804252 0.62504011
sample estimates:
mean of x
0.3315413
```

## Conclusion?

(By default `t.test` tests  $H_0 : \mu = 0$ .)

# Analysis in R — diagnostics

```
> qqnorm(x)
```

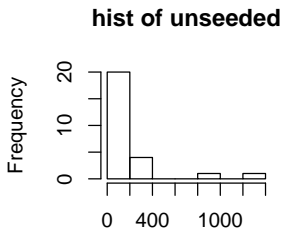
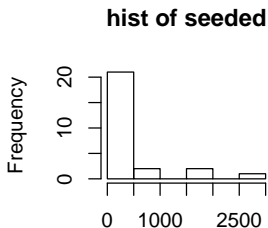


# Discussion (1)

The assumption of normality is crucial. If the data do not follow a normal distribution, the  $p$ -value from the  $t$ -test **cannot be trusted**.

Not all data can be assumed to come from a normal distribution. Histograms and QQ-plots can be used to **check the normality assumption**.

**EXAMPLE** Cloud seeding is a technique used to change the amount and type of precipitation, by dispersing substances into clouds. Precipitation values of seeded and unseeded clouds were measured.

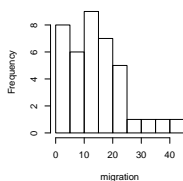


Assuming normality is clearly **wrong**.

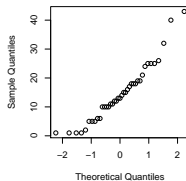
## Discussion (2)

**EXAMPLE** From a sample of 39 Peruvian men that had moved from a native culture to a modern society, the following variables were measured (amongst others): years since migration, systolic and diastolic blood pressure, heart rate, weight, length.

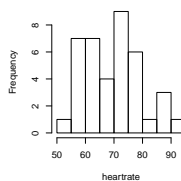
Histogram of migration



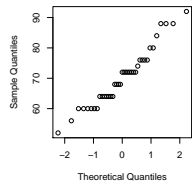
Normal Q-Q Plot



Histogram of heartrate



Normal Q-Q Plot



Normality is doubtful for both migration (seems not symmetric) and heartrate (rounded data).

one sample from a nonnormal distribution

# Setting

An experiment with:

- one **numerical outcome** per experimental unit

Interest is in the **location** of the population distribution.

**EXAMPLE** The **number of infected people** by a certain disease in different countries.

**EXAMPLE** The **times between eruptions** of a volcano.

**EXAMPLE** The **exam grades** for a certain course.

# Design

- Take a random sample of experimental units from the relevant population
- Measure the outcome on each unit

# Analysis A

Data  $(X_1, \dots, X_N)$ .

The **sign test** assumes that the **data**  $X_1, \dots, X_N$  are a random sample from a population with a certain **median**  $m$ .

We **test** the null hypothesis  $H_0 : m = m_0$  that the median of this population is  $m_0$ , e.g.  $m_0 = 0$ .

The **test statistic** is  $T = \#(X_i > m_0)$ , which has the  $\text{bin}(N, 0.5)$ -distribution under  $H_0$ .

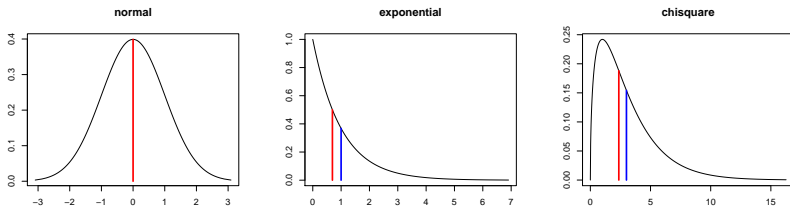


# The median — recap

The **median** of a population is the middle value in the sorted population values.

For a given population median  $m$ , we have that  $P(X < m) = P(X > m) = \frac{1}{2}$  for a random value  $X$  from the population. Being bigger or smaller than the median is like **tossing a fair coin**.

For skewed distributions (e.g. clouds) the mean is highly influenced by the high/low values. In such cases it is better to test location in terms of the **median**, in stead of the **mean**.



The more skewed, the bigger the distance between **median** and **mean**.

# Analysis in R — data input

We want to test whether an exam is of adequate level, that is whether the median is equal to 6. Because of the small sample size, we are not sure about normality. (Grades are never really normally distributed!)

```
> examresults = c(3.7, 5.2, 6.9, 7.2, 6.4, 9.3, 4.3, 8.4,  
+ 6.5, 8.1, 7.3, 6.1, 5.8)
```

(Data are exam grades of 13 randomly selected students.)

# Analysis A in R — testing

The **sign test** computes the number of values bigger than  $m_0$ . If  $m = m_0$  then we expect about  $N/2$  values bigger/smaller than  $m_0$ .

```
> examresults  
[1] 3.7 5.2 6.9 7.2 6.4 9.3 4.3 8.4 6.5 8.1 7.3 6.1 5.8  
> sum(examresults>6)  
[1] 9  
> binom.test(9,13,p=0.5)
```

Exact binomial test

data: 9 and 13

number of successes = 9, number of trials = 13, p-value = 0.2668

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.3857383 0.9090796

sample estimates:

probability of success

0.6923077

**Conclusion:**  $H_0$  is not rejected.

# Analysis B

Data  $(X_1, \dots, X_N)$ .

The Wilcoxon **signed rank test** assumes that the **data**  $X_1, \dots, X_N$  are a random sample from a **symmetric** population with a certain **median**  $m$ . This is a stronger assumption than the one for the sign test!

We **test** the null hypothesis  $H_0 : m = m_0$  that the median of this population is  $m_0$ , e.g.  $m_0 = 0$ .

The **test statistic**  $T$  is based on the ranks  $R_i$  of the absolute differences  $|X_i - m_0|$ .

$$T = \sum_{i: X_i > m_0} R_i.$$

The distribution of  $T$  under  $H_0$  is known, and can be approximated by a normal distribution if  $N$  is large.

Large values of  $T$  indicate that  $m > m_0$ , whereas small values of  $T$  indicate that  $m < m_0$ .

# Analysis B in R — testing

The **signed rank test** takes into account the **ranks** of the deviations from the proposed median  $m_0$ . If the data are symmetric around  $m_0$ , the ranks at both sides should be approximately equal.

```
> examresults-6
[1] -2.3 -0.8  0.9  1.2  0.4  3.3 -1.7  2.4  0.5  2.1  1.3  0.1 -0.2
> rank(abs(examresults-6))
[1] 11  5  6  7  3 13  9 12  4 10  8  1  2
> rank(abs(examresults-6))[examresults-6>0]
[1]  6  7  3 13 12  4 10  8  1
> sum(rank(abs(examresults-6))[examresults-6>0])
[1] 64
> wilcox.test(examresults,mu=6)
```

Wilcoxon signed rank test

data: examresults

V = 64, p-value = 0.2163

alternative hypothesis: true location is not equal to 6

**Conclusion:**  $H_0$  is not rejected.

## two paired samples

# Setting A

An experiment with:

- two **numerical outcomes** per experimental unit.

Interest is in a possible **difference** between the two outcomes.

**EXAMPLE** Measurement of **blood pressure** of a person before and after a drug treatment.

**EXAMPLE** Comparing **pain relief** by a dedicated drug or by a placebo. Both treatments are applied to every individual (with recovery time in between, order assumed to have no effect).

**EXAMPLE** Comparing two **car tire brands** by putting both brands of tire on the same car and measuring the tires' wear.

# Design A

- Take a random sample of experimental units from the relevant population.
- Measure the two outcomes on each unit.

The two outcomes are related, because measured on the same experimental unit. The experiment should be set up so that any other type of “dependence” is eliminated and a difference in outcomes is due to the “treatment” only.

**EXAMPLE** If subjects must perform two tasks, then they should be allowed sufficient time between the tasks to recover and forget.

If a **learning effect** (the first measurement influences the second) is suspected, then, if possible:

- Randomize the order of the two treatments within the units.

The analysis must then follow the **cross over design** (lecture 7), not the paired samples design as discussed here.



# Analysis A

Data  $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ .

The **paired t-test** assumes that the **differences**  $X_1 - Y_1, X_2 - Y_2, \dots, X_N - Y_N$  are a random sample from a **normal** population.

We **test** the null hypothesis  $H_0 : \mu = 0$  that the mean of this population is 0.

The **test statistic** is

$$T = \frac{\bar{Z}_N}{S_N}$$

where  $\bar{Z}_N$  is the average of the differences  $Z_i = X_i - Y_i$ . Under  $H_0$   $T$  has the  $t_{N-1}$ -distribution.

We **estimate**  $\mu$ .

The analysis is simply a **one sample analysis** on the differences, and  $\mu$  is the *difference of the means of the X-population and the Y-population*.

# Analysis A in R — data input

Create the two samples as parallel vectors, e.g. as two columns of a data.frame.

```
> ashina=read.table("ashina.txt",header=TRUE)
```

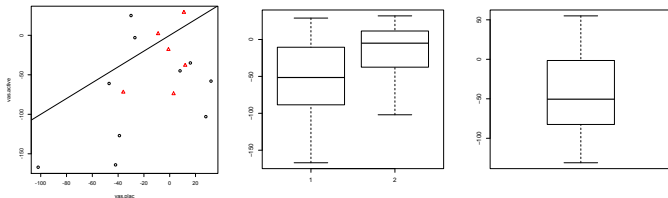
```
> ashina[,1:2]
```

	vas.active	vas.plac
1	-167	-102
2	-127	-39
3	-58	32
4	-103	28
5	-35	16
6	-164	-42
7	-3	-27
8	25	-30
9	-61	-47
10	-45	8
11	-38	12
12	29	11
13	2	-9
14	-18	-1
15	-74	3
16	-72	-36

(The rows correspond to 16 subjects and give measures of pain (for chronic tension-type headache) when treated with an active drug or a placebo.)

# Analysis A in R — graphics

```
> ashina
  vas.active vas.plac grp
1      -167   -102   1
2      -127   -39   1
[ some output deleted ]
16     -72    -36   2
> plot(vas.active~vas.plac,pch=grp,col=grp,data=ashina); abline(0,1)
> boxplot(ashina[,1],ashina[,2])
> boxplot(ashina[,1]-ashina[,2])
```



(The third column of the data.frame `ashina` indicates the order of measurement (1=placebo first, 2=active first). This is used in the first plot (only) to determine the plotting character; this plot does not suggest that the order is important.)

# Analysis A in R — estimation and testing (1)

The paired  $t$ -test:

```
> t.test(ashina[,1],ashina[,2],paired=TRUE)
      Paired t-test
data:  ashina[, 1] and ashina[, 2]
t = -3.2269, df = 15, p-value = 0.005644
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -71.1946 -14.5554
sample estimates:
mean of the differences
      -42.875
```

**Conclusion:**  $H_0$  is rejected, mean of the differences is different from 0.

(A possible effect of the ordering of the measurements is ignored.

Without the option `paired=TRUE` the function `t.test` with 2 arguments assumes that the 2 samples are independent.)

## Analysis A in R — estimation and testing (2)

The one sample  $t$ -test applied to the differences:

```
> t.test(ashina[,1]-ashina[,2])
      One Sample t-test
data:  ashina[, 1] - ashina[, 2]
t = -3.2269, df = 15, p-value = 0.005644
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -71.1946 -14.5554
sample estimates:
mean of x
 -42.875
```

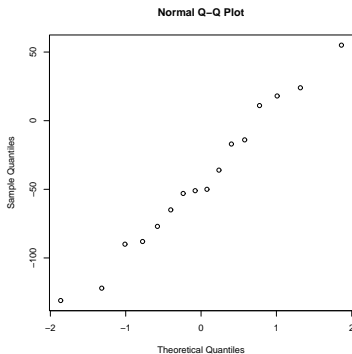
**Conclusion:**  $H_0$  is rejected, mean of the differences is different from 0.

(With 1 argument the function `t.test` performs a one sample  $t$ -test. Applied to the differences this is equivalent to a paired two sample  $t$ -test – the shown  $p$ -values are identical.)

# Analysis A in R — diagnostics

Check the normality assumption on the differences:

```
> qqnorm(ashina[,1]-ashina[,2])
```



(No reason to suspect that the differences are not taken from a normal population.)

# Setting B

An experiment with:

- two **numerical outcomes** per experimental unit.

Interest is in a possible **dependence** between the two outcomes per unit.

**EXAMPLE** Relation between **shoe size** and **body mass index** of a person.

**EXAMPLE** Relation between **average course grade** and **number of students taking the course** for courses at the VU.

**EXAMPLE** Relation between amount of **precipitation** and **sun hours** for different cities in Europe.

## Design B

- Take a random sample of experimental units from the relevant population.
- Measure the two outcomes on each unit.

The two outcomes are in principal related, because measured on the same experimental unit. However, we possibly have measured unrelated quantities of the units.



# Analysis B1

Data  $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ .

Pearson's **correlation test** assumes **normality** of both  $X$ 's and  $Y$ 's. It is based on the linear correlation.

The **test statistic** is based on the sample correlation coefficient:

$$\rho_{X,Y} = \frac{\sum_{i=1}^N (X_i - \bar{X}_N)(Y_i - \bar{Y}_N)}{\sqrt{\sum_{i=1}^N (X_i - \bar{X}_N)^2 \sum_{i=1}^N (Y_i - \bar{Y}_N)^2}}$$

We **test** the null hypothesis  $H_0 : \rho = 0$  that the correlation between the two populations is 0.

We **estimate** the correlation  $\rho$ .

# Analysis B2

Data  $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ .

Spearman's **rank correlation test** does **not assume normality**. The test considers the ranks in the two samples, and compares the ordering of the ranks in the  $X_i$  and the  $Y_i$ . If the data are rank correlated, these sequences of ranks will run (approximately) in parallel or in opposite order.

The test statistic is based on the correlation coefficient  $\tilde{\rho}$  of the rank vectors.

We **test** the null hypothesis  $H_0 : \tilde{\rho} = 0$  that the rank correlation between the two populations is 0.

We **estimate** the rank correlation  $\tilde{\rho}$ .

# Analysis B in R — data input

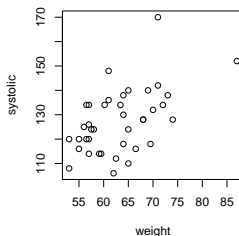
Create the two samples as parallel vectors, e.g. as two columns of a data.frame.

```
> peruvians=read.table("peruvians.txt",header=TRUE)
> peruvians[,c(3,9)]
      weight systolic
1      71.0      170
2      56.5      120
3      56.0      125
4      61.0      148
[ some output deleted ]
38     70.0      132
39     87.0      152
```

(The rows correspond to 39 men that moved from a native culture to a modern society. Amongst others, weight and systolic blood pressure were measured.)

# Analysis B in R — graphics

```
> peruvians
  age migration weight length chin  arm calf wrist systolic diastolic
1  21          1  71.0  1629  8.0  7.0 12.7   88      170      76
2  22          6  56.5  1569  3.3  5.0  8.0   64      120      60
3  24          5  56.0  1561  3.3  1.3  4.3   68      125      75
[ some output deleted ]
39 54         40  87.0  1542 11.3 11.7 11.3   92      152      88
> attach(peruvians)
> plot(systolic~weight)
```



(Based on this picture we expect dependence.)

# Analysis B1 in R — estimation and testing

```
> cor.test(systolic,weight)
```

Pearson's product-moment correlation

```
data:  systolic and weight
t = 3.7164, df = 37, p-value = 0.0006654
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2463759 0.7186619
sample estimates:
      cor
0.5213643
```

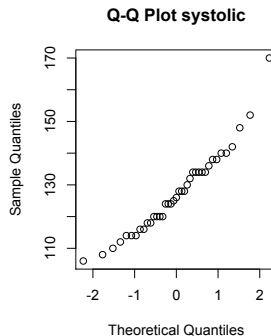
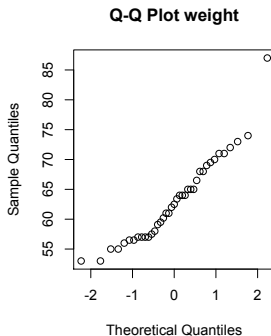
**Conclusion:** there is significant correlation, if normality is assumed.

(The default for `cor.test` is Pearson's correlation test, based on normality.)

# Analysis B1 in R — diagnostics

Check the normality assumption on the two samples:

```
> qqnorm(weight,main="Q-Q Plot weight")  
> qqnorm(systolic,main="Q-Q Plot systolic")
```



Normality is not plausible for the weight sample. Hence, use the rank correlation test of Spearman.

# Analysis B2 in R — estimation and testing

```
> cor.test(systolic,weight,method="spearman")
```

Spearman's rank correlation rho

data: systolic and weight

S = 5322.352, p-value = 0.003119

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.4613004

Warning message:

In cor.test.default(systolic, weight, method = "spearman") :

Cannot compute exact p-values with ties

**Conclusion:** there is indeed significant rank correlation.

(There is a warning about ties, which means that some values occur multiple times in weight and/or systolic. Therefore *R* uses an approximation for the *p*-value)

to finish



# To wrap up

Today we saw:

- 1 one sample, normal
- 2 one sample, not normal
- 3 two paired samples

Next time: two independent samples,  $k$  samples