Experimental Design and Data Analysis
Lecture 1

Eduard Belitser

VU Amsterdam

Lecture Overview

1. course parameters (lecturers, literature, assignments, etc.)
2. experimental design
3. recap of statistical concepts
4. small R demonstration (?)

Course parameters

Organisation

Lecturer Eduard Belitser

Teaching assistant(s)

Lectures 10-11 lectures

Assignments 6 assignments, made by groups of two students

Final project to be submitted in the last week of May

Grade based on assignments (67%) and final project (33%)

Prerequisites basic statistics course

Literature

Schedule of lectures and assignments available on blackboard

Lecture slides available on blackboard (attend lectures!)

Assignments available on blackboard

R manual(s) available on blackboard

Other literature suggestions on blackboard

Assignments

Sign up in groups of two students on bb.vu.nl.

R is an open software package, widely adopted in the academic community. It is

- a programming language
- a statistical package
- a graphics environment
- free

R is object oriented. This means that the input and output of functions are structured objects that can be manipulated with general purpose functions.

RStudio IDE is a powerful user interface for R.

R can be downloaded from `http://www.r-project.org/`.

Blackboard

All relevant information is on blackboard:

- up-to-date schedule of lecture topics
- assignments + due dates
- lecture slides (some may be updated after the lecture)
- ...

experimental design

Statistical design

Statistics allows to generalise from a sample of data to a true state of nature.

To make this work the data must be obtained by a carefully designed (chance) experiment (or at least it must be possible to think about the data in this way).

EXAMPLE To compare two fertilisers we prepare 20 plots of land, apply the first fertilisers to 10 randomly chosen plots and the second one to the remaining plots. We plant a crop and measure the total yield from each plot.

EXAMPLE To compare two web designs we randomly select 50 subjects and measure the time needed to find some information. All 50 subjects perform this task with both designs, but for each subject the order of the two designs is based on tossing a coin.

Randomisation

Any good design involves a chance element: "experimental units" are assigned to "treatments" by chance.

The purpose is to exclude other possible explanations of an observed difference.

EXAMPLE If an experiment involves subjects, then it is wrong to assign "task A" to the first 10 subjects who arrive and "task B" to the last 10. (There may be a reason for arriving early that correlates with the outcome.) Instead assign the tasks at random. Then an observed difference is due to the task or chance.

Price to pay: we need probability to quantify the chance element in the data.

Pseudo randomisation

In practice randomisation is implemented with a random number generator.

In R we create a sequence of 5 A's and 5 B's in random order, as

```
> x=rep(c("A","B"),each=5)
> x
 [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B"
> sample(x)
 [1] "A" "B" "A" "B" "B" "A" "B" "A" "A" "B"
```

In R we can toss a fair coin 10 times, by

```
> rbinom(10,1,0.5)
[1] 1 0 1 1 1 0 1 0 0 0
> rbinom(10,1,0.5)
[1] 1 0 0 0 0 1 0 1 1 0
```

and a biased coin (succes probability=0.8) 5 times by

```
> rbinom(5,1,0.8)
[1] 1 1 0 1 1
```

Observational studies

Data obtained by registering an ongoing phenomenon, without randomisation or applying other controls, is called observational.

Statistical inference from such data is often impossible. If possible at all, it requires assumptions and mathematical modelling.

EXAMPLE The incidence of lung cancer among 500 smokers is observed to be higher than among 500 non-smokers. Does this finding generalise to the full population? Does this show that smoking causes lung cancer?

Overview EDDA

An overview of the topics in this course can be found in the schedule under Course Information on `bb.vu.nl`.

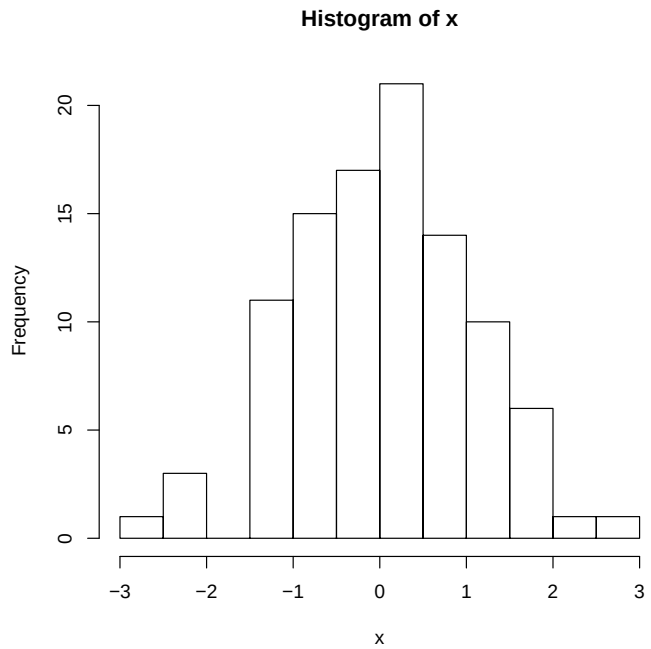
recap statistics

Histogram

The histogram corresponding to numerical measurements $x_1, x_2, ..., x_N$ is a barplot, where the area of the bar over an interval (a, b) corresponds to the fraction

$$\frac{1}{N}\#\big(1 \le n \le N : a \le x_n \le b\big).$$

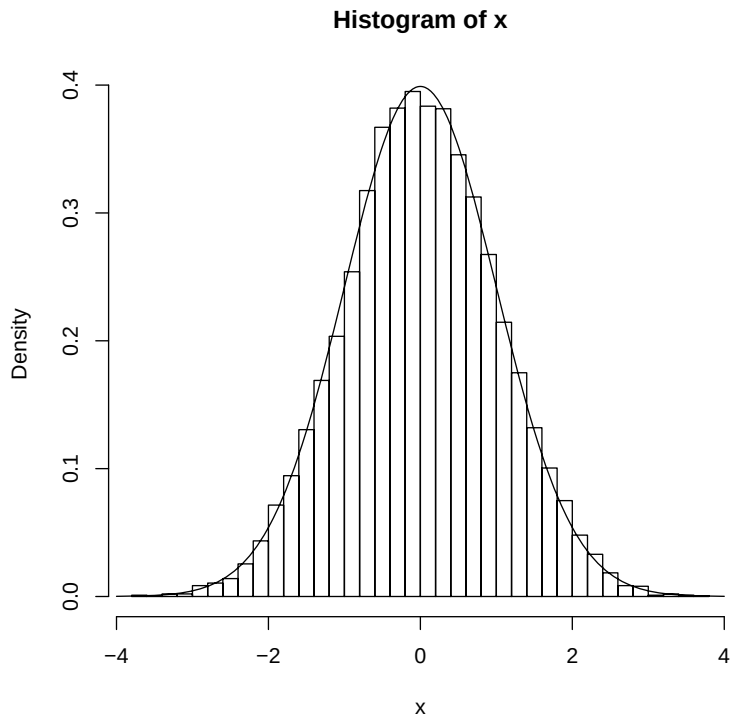
```
> x=rnorm(100)
> hist(x)
```



Population distribution

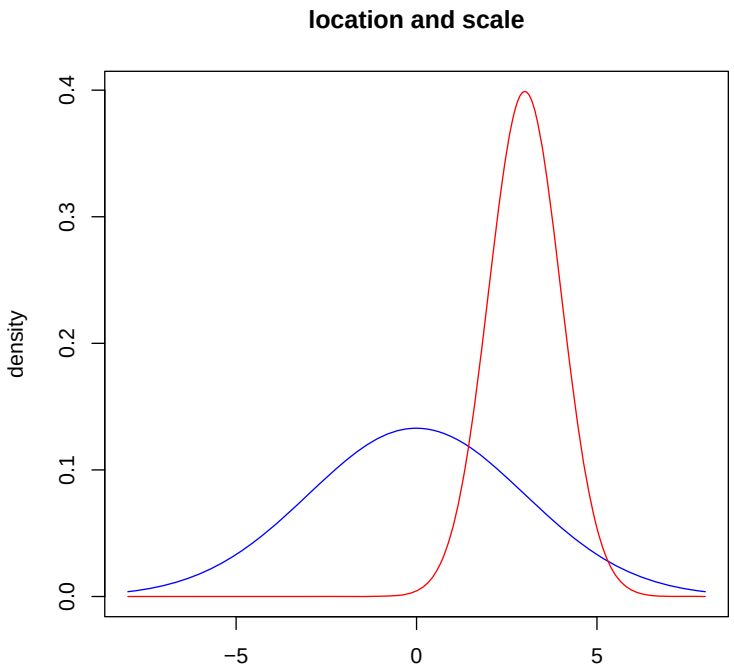
A population curve or population density is a (smoothed) histogram of a population of values.

A population can be an actual population, e.g. the heights of all men in the Netherlands. It can also be the (imaginary) infinite number of outcomes obtained by repeating an experiment over and over, e.g. throwing a die many times.



Location and scale

Two important characteristics of a population are location and scale (or mean and standard deviation).



The normal density curve is given by the function

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

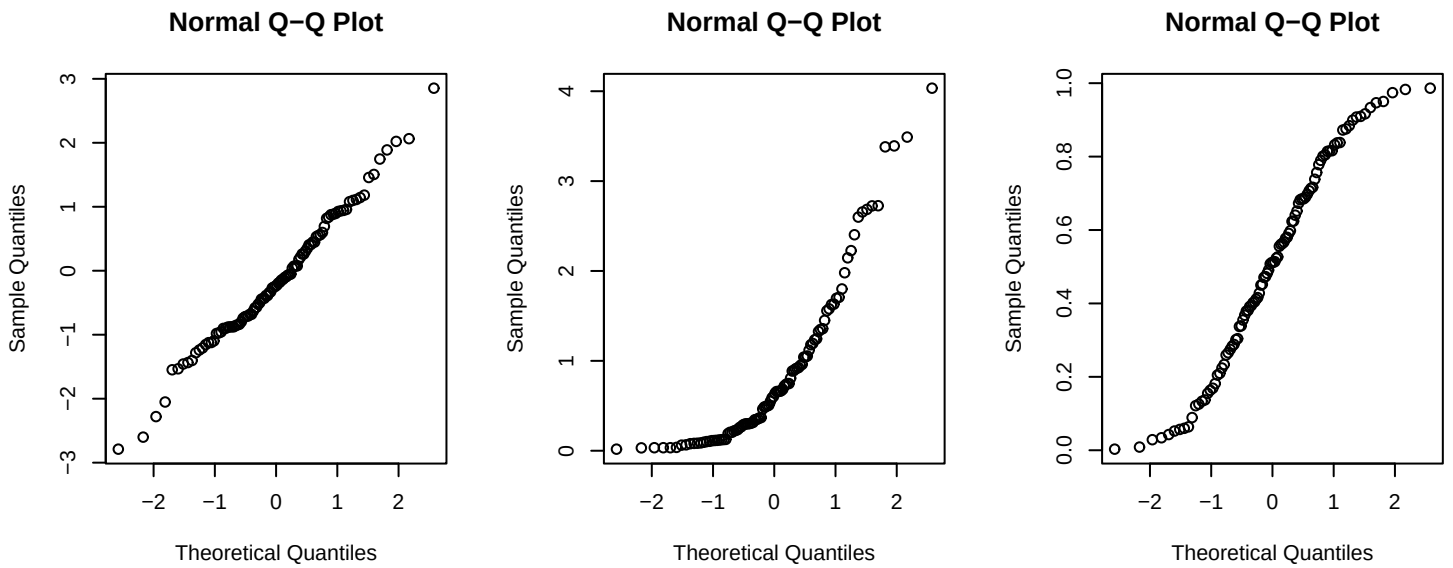
The parameters μ and σ are the location and scale.

Remark The normal curve is very particular! There are many "bell shaped" curves that are far from normal.

QQ-plots

A QQ-plot can reveal whether data (approximately) follows a certain population curve, e.g. the normal curve.
It plots the ordered data $x_{(1)}, x_{(2)}, \ldots, x_{(N)}$ versus the values $\alpha_{N,1}, \alpha_{N,2}, \ldots, \alpha_{N,N}$ that are typical for ordered values from the population. A fraction of i/N of the population is smaller than the i/N-quantile $\alpha_{N,i}$.

If the points are approximately on a straight line, then the data can be assumed to be sampled from the population, possibly with different location and scale.



Statistical test (1)

A statistical test chooses between two possibilities: the null hypothesis H_0 and the alternative hypothesis H_1.

EXAMPLE H_0 says that genetic algorithm 1 performs better than genetic algorithm 2, H_1 the opposite.
We think of values $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$ obtained using the two algorithms during a fixed CPU time as random samples from the two (imaginary) populations of all values that would be obtained if the experiments were repeated infinitely often. If the algorithms are supposed to find a minimal value, then H_0 says that the mean of the first population is smaller than the mean of the second one.

Statistical test (2)

In a statistical test we either reject H_0 (and accept H_1) or do not reject H_0 (and treat the analysis as inconclusive).

Statistical tests are typically not perfect, but make two types of errors:

- Error of the first kind rejecting H_0 while it is true.
- Error of the second kind not rejecting H_0 while it is false.

Tests are constructed to have small probability of an error of the first kind ($<$ 5%). This is called the level of the test.

The probability of an error of the second kind depends (among others) on the amount of data. 1 minus the probability of an error of the second kind is called the power of the test. The power of a test is specified for each possibility under H_1.

You may apply different tests for the same H_0, each having its own power.

One sample t-test

Given a population with mean μ, we wish to test $H_0 : \mu = \mu_0$ against $H_1; \mu \neq \mu_0$ for some given number μ_0 e.g. $\mu_0 = 0$. We take a random sample $X_1, \ldots, X_N$ from the population.

We can test H_0 by the t-test, which compares $\overline{X}_N$ to μ_0, taking into account the variation amongst $X_1, \ldots, X_N$.

```
> mu=0.2
> x=rnorm(50,mu,1)  # creating artificial data
> t.test(x,mu=0)

        One Sample t-test

data:  x
t = 2.4211, df = 49, p-value = 0.01922
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.05219746 0.56202370
sample estimates:
mean of x
0.3071106
```

Two sample t-test (1)

Given two populations with means μ and ν, we wish to test $H_0 : \mu = \nu$ against $H_1; \mu \neq \nu$. We take a random sample $X_1, \ldots, X_N$ from the first population and, independently, $Y_1, \ldots, Y_M$ from the second population.

We can test H_0 by the two sample t-test.

```
> mu=0;nu=0.5
> x=rnorm(50,mu,1); y=rnorm(50,nu,1) # creating artificial data
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = -2.4339, df = 96.574, p-value = 0.01677
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.85202520 -0.08659066
sample estimates:
 mean of x  mean of y
0.06552453 0.53483246
```

Two sample t-test (2)

The outcome of the test is based on the difference $\overline{X}_M - \overline{Y}_N$ which is a reasonable estimate for $\mu - \nu$. If it is very different from 0 we reject H_0.

How different?

$\overline{X}_M - \overline{Y}_N$ will not exactly be $\mu - \nu$. The (random) size of the estimation error depends on M and N and the standard deviations of the populations. The t-statistic therefore divides $\overline{X}_M - \overline{Y}_N$ by an estimate $S_{M,N}$ of its standard error:

$$T = \frac{\overline{X}_M - \overline{Y}_N}{S_{M,N}}, \qquad S_{M,N} = S_{X,Y}\sqrt{\frac{1}{M} + \frac{1}{N}},$$

where $S^2_{X,Y} = \frac{1}{M+N-2}\left(\sum_{i=1}^{M}(X_i - \overline{X}_M)^2 + \sum_{j=1}^{N}(Y_j - \overline{Y}_N)^2\right)$. If H_0 holds, the test statistic T follows the t_{N+M-2}-distribution. Therefore, the observed value of T is compared to the critical value, which is a quantile from this distribution.

Remark The standard t-test assumes that the two populations are normal. If the sample sizes M and N are large, then the test performs well even without this assumption, but the test is unreliable for M,N less than 20 and skewed or otherwise nonnormal populations.

p-value

The p-value of a test is the probability that an experiment in the situation that the null hypothesis is true will deliver data that is more extreme than the data actually observed.

So: a small p-value indicates that the observed data is unlikely if H_0 is true.

Typically H_0 is rejected if the p-value is less than 5%. The data are then said to be statistically significant at level 5%.

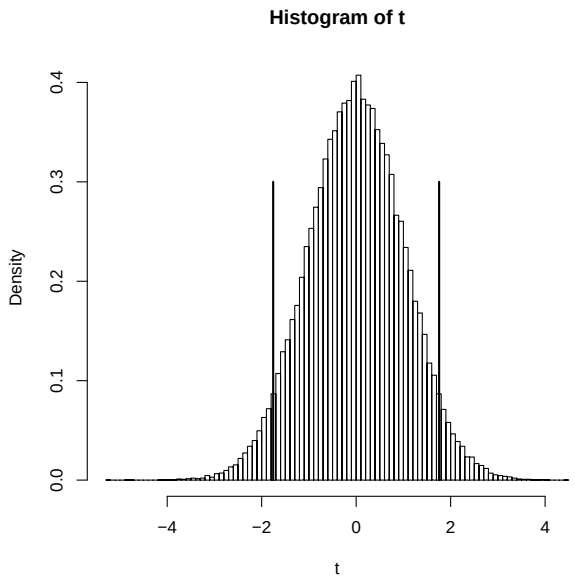
By construction, the p-value is like a uniform draw from the numbers $[0, 1]$ if H_0 is true (regardless of the level).

p-value for two sample t-test (1)

For the two sample t-test we can construct a p-value as follows.
Let $X_1, \ldots, X_M$ and $Y_1, \ldots, Y_N$ be independent random samples from two populations with the same means, and

$$T = \frac{\overline{X}_M - \overline{Y}_N}{S_{M,N}}.$$

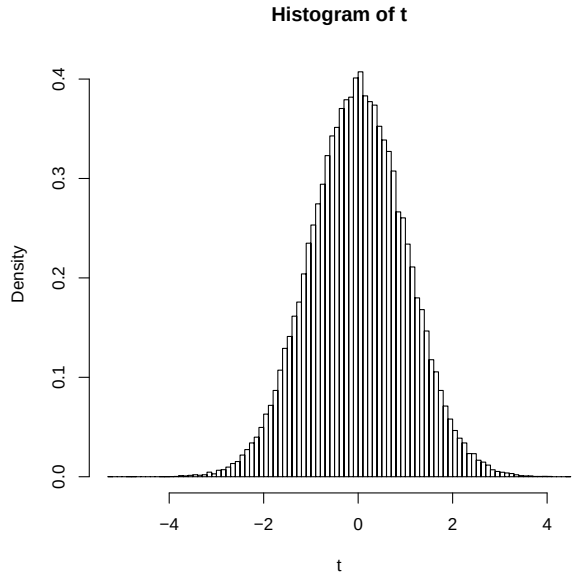
We can form a population of T-values under H_0 by repeating the sampling.



The p-value of the observed value t is the fraction of this population that is bigger than $|t|$ or smaller than $-|t|$.

p-value for two sample t-test (2)

```
> mu=nu=0; t=numeric(100000)
> for (i in 1:100000){x=rnorm(50,mu,1); y=rnorm(50,nu,1)
+ t[i]=t.test(x,y)[[1]]}
> hist(t,breaks=seq(-5,5,length=100),prob=TRUE)
> sum(abs(t)>= 1.7563)/length(t)
[1] 0.08171    ## cf. 2*(1-pt(1.7563,98))=0.08217
```



Practical significance

Data can be statistically significant even though the deviation from H_0 is very small!

Statistical significance is about generalisation: an observed effect is not due to chance, but would be observed again if a new experiment were performed.

Practical significance is about the size of the effect.

EXAMPLE Suppose that a coin has probabilities $1/2-10^{-10}$ and $1/2+10^{-10}$ to land HEAD or TAIL.
If we use the coin to decide who will kick-off in a soccer game, then TAIL has a slight advantage, but the difference is negligible. A statistical test based on observing 100 tosses of the coin will not reject H_0, but a test based on observing 10^{21} coin tosses almost certainly will.

Point and interval estimation

If a null hypothesis is rejected the practical significance can be determined by estimating the effect size.

A point estimate is a "best guess" while an interval estimate or confidence interval gives a set of "plausible values".

For the t-test the effect size is the difference $\mu - \nu$ of the population means.

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

        Welch Two Sample t-test

data:  x and y
t = -1.7563, df = 97.913, p-value = 0.08217
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.80258158  0.04896641
sample estimates:
 mean of x  mean of y
0.04370384 0.42051143
```

Standard error

A standard error of an estimator is a measure of its precision. If the estimator is normally distributed, then `Estimate` $\pm$ `1.96 Std. Error` gives a 95 % confidence interval.

The bigger the sample size, the smaller the standard errors and the confidence intervals. That means, the estimates get more precise, because more information is available.

Estimate population mean (true value = 0) from standard normal sample

sample size	`Estimate`	`Std. Error`
10	0.3564	0.3604
50	0.2198	0.1510
100	0.1098	0.1067
1000	-0.007433	0.031466

In all cases the true value 0 is in the 95 % confidence interval `Estimate` $\pm$ `1.96 Std. Error`.

to finish

To wrap up

Today we saw

1. course parameters (lecturers, literature, assignments, etc.)
2. experimental design
3. recap of statistical concepts

Assignment in groups of 2 students. Enroll in groups!

Questions ask teaching assistant(s)

Next time bootstrap methods