Applied Statistics

(Applied Biostatistics, Statistics and R)

Erlis Ruli
Department of Statistical Sciences

based in part on course material
by A.R Brazzale, G. Masarotto and D. Risso

Second cycle degree courses:
Molecular Biology
Quantitative and Computational Biosciences

Current version: October 11, 2024

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Part 1

Introduction

Table of Contents

1. Background information
2. Statistical inference: an overview

Background information

About us

Who are we?

Erlis Ruli, Francesco Lollato
Department of Statistical Sciences
via Cesare Battisti, 241
e-mail: erlis.ruli@unipd.it

Class schedule

- Monday 08:30 – 09:30 (**1D @ Botta**)
- Tuesday 08:30 – 09:30 (**1D @ Botta**)
- Friday 11:30 – 13:30 (**1D @ Botta**)
- mid-Oct. – Jan.: computer labs (3hrs, Thur. 14:30 - 17:30)

Office hours

to be agreed over e-mail

Software used

an open source software environment for statistical computing and graphics

Homepage	http://www.r-project.org
Download	https://cloud.r-project.org/
	an integrated development environment (IDE) for R https://www.rstudio.com/products/rstudio/

Installation instructions on Moodle.

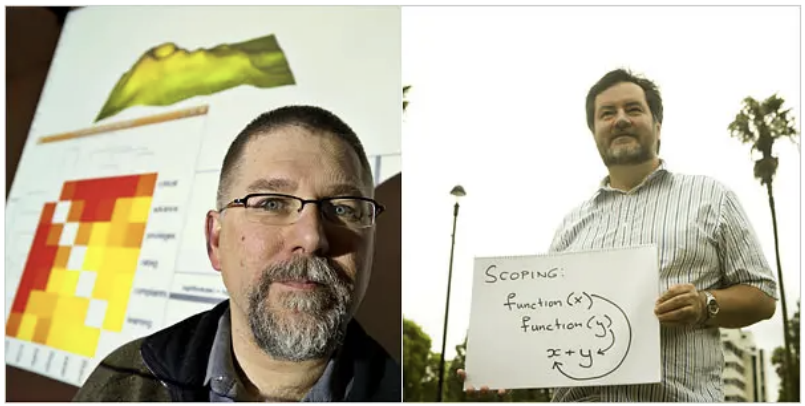
Why R?

- *open-source*, available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form
- Runs on a wide variety of UNIX platforms, Windows and MacOS
- Highly extensible (currently more than 18,600 packages available; many of these are of interest to biology/ bioinformatics)
- Freely available, and ...

The New York Times

Data Analysts Captivated by R's Power

Give this article



R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

Left, Stuart Isett for The New York Times; right, Kieran Scott for The New York Times

By Ashlee Vance

Jan. 6, 2009

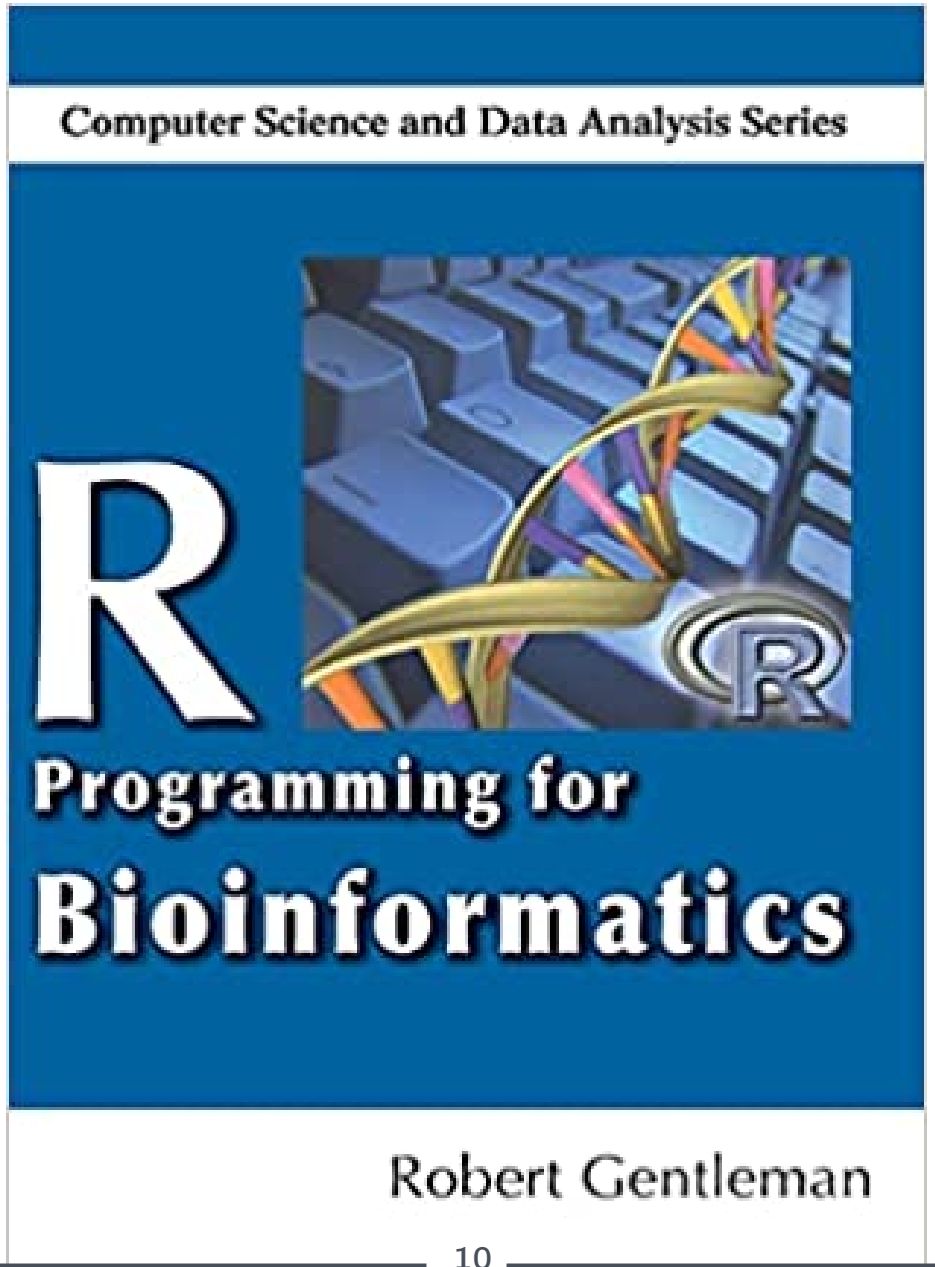
Why R?

To some people, R is just the 18th letter of the alphabet. To others, . . . R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia.

Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it. But R has also quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use.

"The great beauty of R is that you can modify it to do all sorts of things," said Hal Varian, chief economist at Google. "And you have a lot of prepackaged stuff that's already available, so you are standing on the shoulders of giants".

Why R?


Computer Science and Data Analysis Series
R
Programming for
Bioinformatics
Robert Gentleman

Why R?

There are many reasons to prefer R to other languages for scientific computation. The existence of a substantial collection of good statistical algorithms, access to high-quality numerical routines, and integrated data visualisation tools are perhaps the most obvious ones.

Reproducibility is an essential part of any scientific investigation. The ability to integrate text and software into a single document greatly facilitates the writing of scientific papers. It helps to ensure that all figures, tables, and facts are based on the same data and are reproducible by the reader.

The exam

- Written and closed-book.
- A few "theoretical" questions to verify that you have mastered a minimum of terminology.
- Exercises focus on interpreting and communicating the results of statistical analyses.
- Some templates will be handed out during the course.
- There will be a **mock exam** session.

Contents

1. Statistical Inference: background and extensions

hypothesis testing, confidence intervals, non-parametric tests, power function, width of a confidence interval, sample size calculations, ...

2. Multivariate techniques

correlation, multiple regression, basics of logistic regression, principal components, and cluster analysis.

Style guide:

The style with be informal, using intuition, examples, and only a minimum of maths. Study the textbook.

In particular, we will see . . .

1. Why Mendel shouldn't have been perplexed by his first results on round and wrinkled peas. And, hence, how he may have avoided falsifying his final results given that the preliminary results were in agreement with his theories!
2. Why MAO (monoamine oxidase) activity levels differ among different groups of schizophrenia patients.
3. which are the morphological differences among female and male crabs, or blue and orange ones.
4. and much more . . .

What is statistics?

Statistics is the *art of making numerical conjectures* about puzzling questions. (Freedman et al., 1978).

The objective of statistics is to make *inferences* (predictions, decisions) *about a population* based on information contained in a sample. (Mendenhall, 1987).

What is statistics?

Far better an *approximate answer to the right question*, which is often vague, than an *exact answer to the wrong question*, which can always be made precise. (John W. Tukey, 1962)

All models are wrong, but some are useful. (George E. P. Box, 1987)

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of. (R. A. Fisher, 1938)

Why statistics?

There is a long history of joint development of biology and statistics.
For instance, Mendel's laws of heredity were entirely based on statistical inference.
Biostatistics has become an integral part of modern biomedical research.
Perhaps the most critical example is the rigorous practice of *clinical trials*, in which statistics plays a crucial role in deciding:

- how many subjects to enrol
- when to stop the trial
- if the new treatment is indeed an advancement over the standard of care

Biostatistics remains at the forefront of biomedical research today.
Think of *genomics* and *precision medicine*.

To wind up

- Doing statistics means answering questions about our world; this is what we will be doing.
- Given the nature of this course, we will tackle basic questions and obtain answers using standard and widely applicable techniques.
- I'll try and guide you in this journey; but ***you*** must help me/each other; so, for instance, try and seek a couple of "true" questions of practical use for every "answer" I'll give you.

What I expect from you

1. Spirit of enquiry and willingness to use common sense.
2. Averages, variance/standard deviation (to quantify *location and dispersion*).
3. Quantiles?
4. Random variable? Binomial and normal distributions?
5. But, most of all, spirit of enquiry and willingness to use common sense.

What do you expect from me (this course)?

Please post your answer after this lecture and as soon as possible on the course's **social platform** (login via SSO): https://unipd.padlet.org/erlisruli/applied-statistics-2024-8pr229e9snsv7zbx

Textbooks and lecture notes

1. Study material on Moodle
2. The textbook

   M. C. Whitlock & D. Schluter (2015+)
   The Analysis of Biological Data (2nd ed. is fine),
   Macmillan learning.

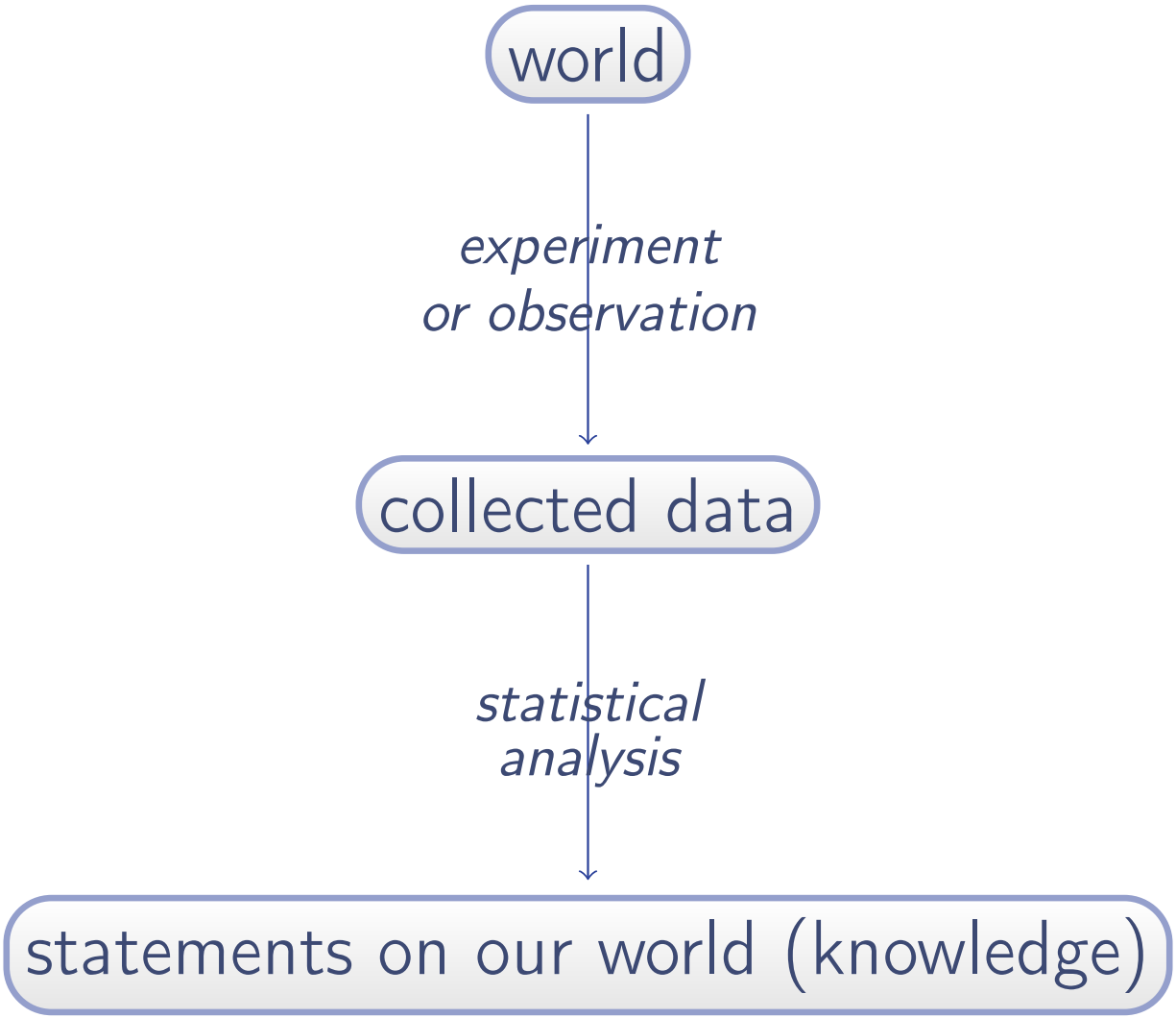
3. Another modern book (available online for free!) is

   S. Holmes & W. Huber (2019)
   Modern Statistics for Modern Biology,
   Cambridge.

Statistical inference: an overview

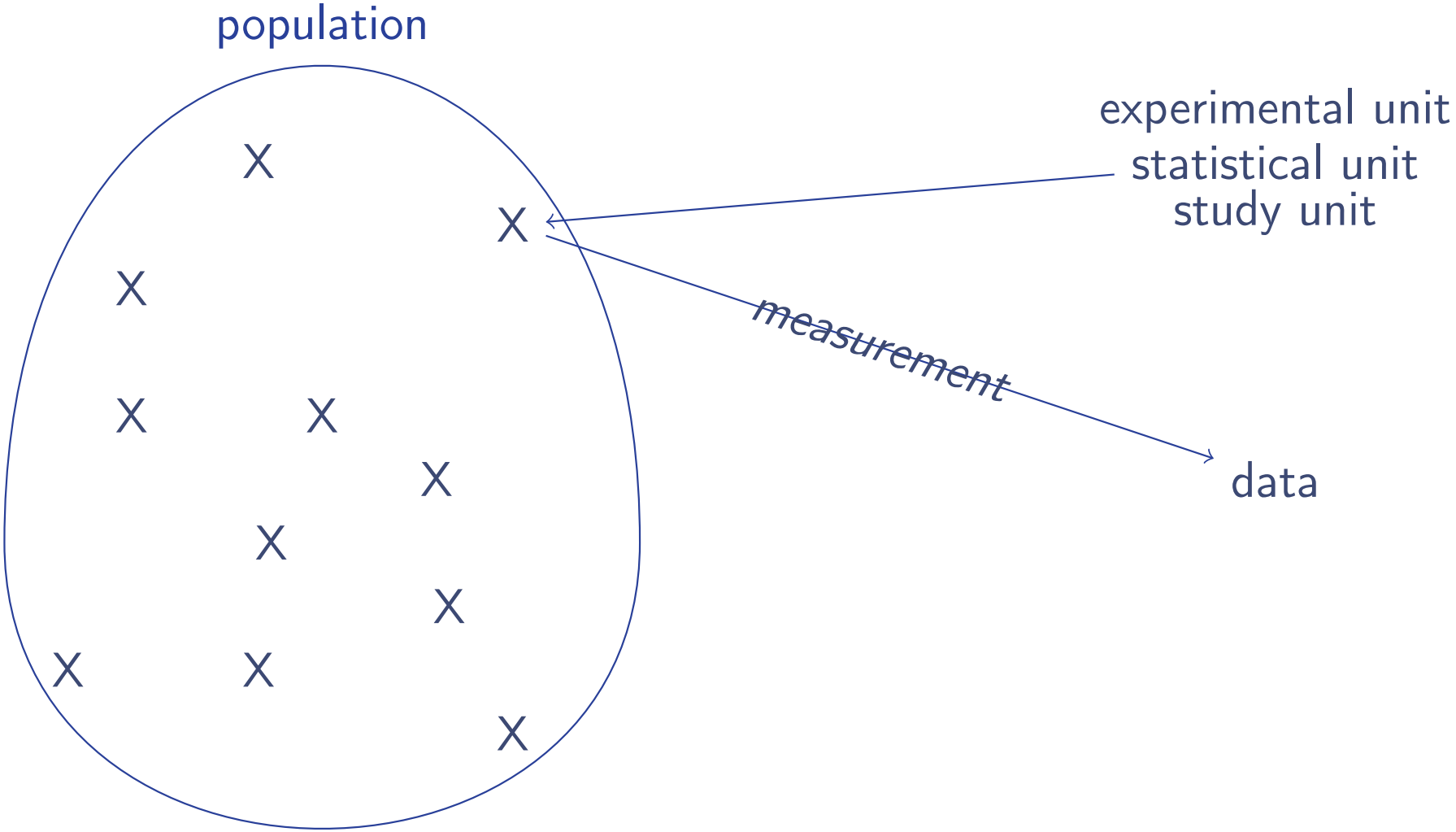
Statistical inference: an overview

Basic problem



Statistical inference

An overview: population, sample, data



Statistical inference

An overview: population, sample, data

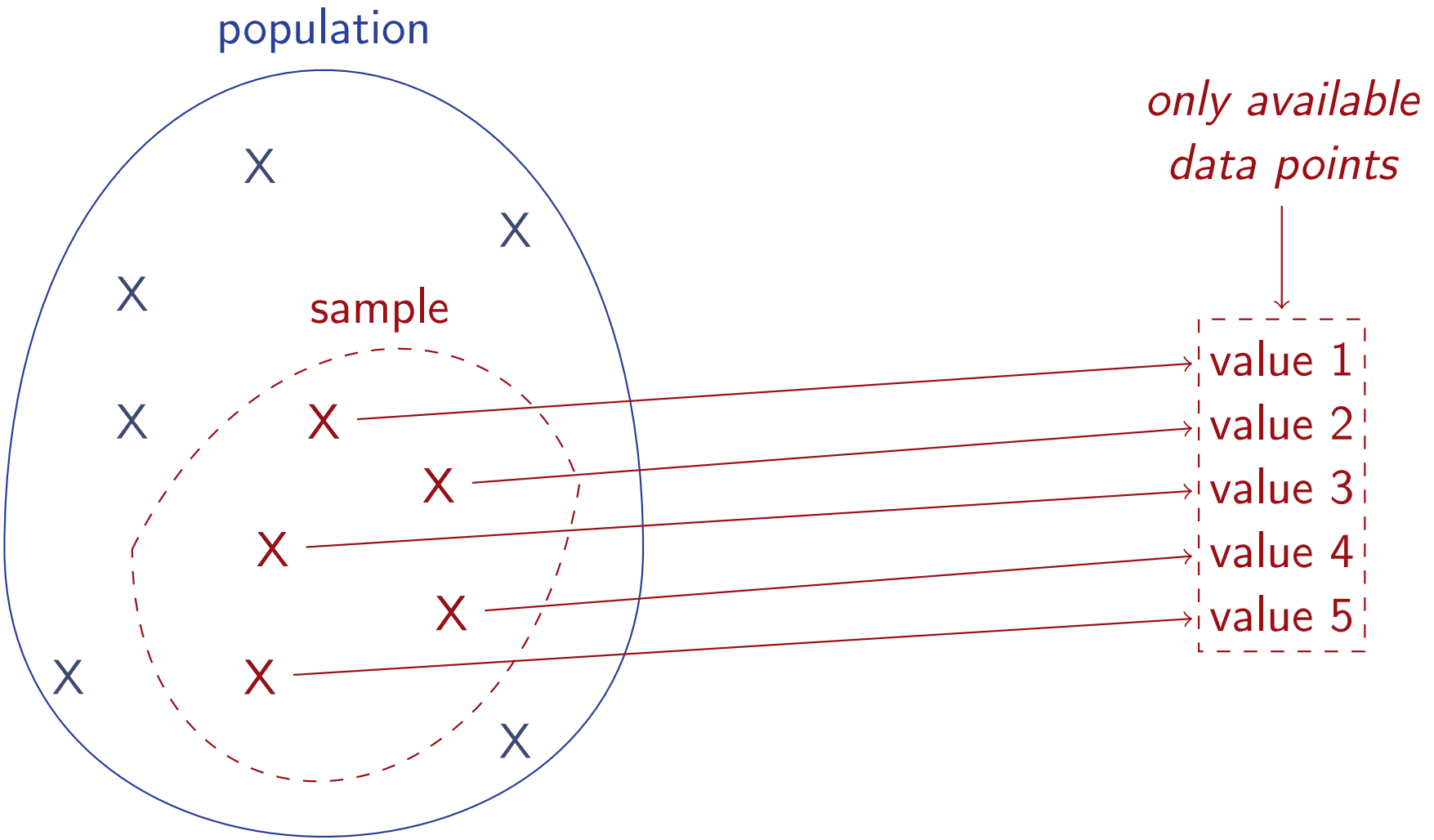
- We associate the part of the world *"we want to know more about"* with a set/ensemble called target (study) population.
- The elements of the target population are called statistical units.
- These can be of any type: humans, humans with particular kinds of allergies, giant octopuses of the South China Sea, galaxies, birches, cells, neutrinos, ...
- They all differ because of variability (or statistical variability).
- Data are obtained by measuring the variable of interest on the statistical units.

- The population can be real and finite.
- For instance, we may be interested in studying how a certain virus spreads among the capuchin monkeys who live on the Sugar Loaf in Rio de Janeiro in October 2018. The target population is the set of all capuchin monkeys who, in October 2018, live on the Sugar Loaf in Rio de Janeiro.

- The population can also be virtual and infinite.
- So, if we talk about schizophrenia, this means considering all human beings who have been suffering from schizophrenia since Ask and Embla were given Midgard and will be suffering until Ragnarok announces the final battle between the Aesir and Giants (at least according to Norse mythology).

Statistical inference

An overview: population, sample, data



- Often, we do not measure what we are interested in on all population elements, but only on a sample.
- We are, however, ambicious: we want to say something about the entire world (the population), not only for the sample.
- 1^{st} side effect: if statistical units differ (because of variability – statistical? biological?), there is only one thing we can be certain about, that is, that we will make mistakes (errors).
- 2^{nd} side effect: in some way, we have to know the relationship between the sample and the population.

Further terminology

Anything measured on a statistical unit is called variable.

In some fields, variables are called *parameters*, but to statisticians, these terms have different meanings.

Variables can be classified in several ways; most importantly, we have

(i) continuous variables (height, weight, blood pressure, etc.)

(ii) count variables (number of bacteria per leaf, etc)

(iii) categorical variables (sex, colour, species, etc.)

A rule of thumb: continuous v. take on real numbers, count variables take on integers (0,1,2,...), categorical variables take on labels.

Sometimes, labels of a categorical variable can be ordered (e.g., tumour grade); in this case, they are called ordinal variables.

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Part 2

Analysing proportions

Table of Contents

3. A gardener monk meets a binomial distribution which tells him "you are right".

4. What would Mendel write in a (scientific) paper was he alive today?

5. While we are on the topic, let Mendel (and us!) meet the normal distribution.

A gardener monk meets a binomial distribution which tells him “you are right”.

Mendel's "pilot" experiment (I)

Materials Two types of pea plants; the first type produces green pods, the second yellow pods. Both groups of plants belong to a "pure line" (= maintains its traits constant across generations).

1st generation Obtained by cross-pollinating "yellow" plants with pollen obtained from "green" plants. This leads to hybrid offspring. *All plants of this generation have green pods.*

Mendel's "pilot" experiment (II)

2nd generation Obtained by self-pollinating 1st generation plants.

The data 56 2nd generation plants of which:

- 39 are "green"
- 17 "yellow"

Question: Do you see the variable? What type is it?

Why did Mendel run this experiment?

- Mendel was interested in

$$\vartheta = \text{Prob}\left(\begin{array}{c}\text{a } 2^{\text{nd}} \text{ generation plant} \\ \text{produces a green pod}\end{array}\right)$$

- Indeed, according to his theory

$$\vartheta = \frac{3}{4}.$$

All initial plants are either "yy" (yellow group) or "GG" (green group); hence, we obtain "Gy" in the first generation. "G" is the dominant trait, and all first-generation plants "look" green. The second-generation plants are "GG", "Gy", "yG" or "yy". All four combinations have the same probability of occurring. Hence, ...

What does ϑ represent?

Target population

the set of 2^{nd} generation plants we would obtain if we repeated Mendel's experiment an infinite number of times.
The population is, hence, virtual and infinite.

Possible interpretation of ϑ

the fraction of plants in the population which produce a green pod

Note that ...

ϑ is a feature of our world!

Let's guess the value of ϑ

Notation

- $y = 39$ number of plants with a green pod and
- $n = 56$ plants in total

were observed in the 2nd generation.

"Intuitive" estimate (= guess) of ϑ

$$\hat{\vartheta} = \frac{y}{n} = \frac{\text{number of "green" plants}}{\text{"total" number of plants}}.$$

With Mendel's data

$$\hat{\vartheta} = \frac{39}{56} \approx 0.696.$$

So, was Mendel right?

Beware that . . .

If there is a guess (= estimate), there is an error!

Unless we are incredibly lucky, we expect

$$\hat{\vartheta} \text{ to differ from } \vartheta$$

or, stated differently,

$$(\text{estimate of } \vartheta) \neq (\text{true value } \vartheta)$$

This is because

- ϑ is a "constant" feature of our world (the infinite number of 2^{nd} generation pea plants we may obtain);
- $\hat{\vartheta}$ is a feature of the sample (the 56 pea plants grown by Mendel).
  If we repeat the experiment, we will likely obtain a different result.

Question:

Does this experiment support Mendel's theory?
In other words, if the "proportion of green plants in the world" is 75%, in a sample, can we observe less than 70% of "green" pea plants?

Formally,

does the observed outcome "support" the hypothesis

$$H_0 : \vartheta = 0.75$$

or does it suggest that

$$H_1 : \vartheta \neq 0.75$$

A possible solution

1. Pretend H_0 is true and identify all the possible experimental outcomes.
2. Check if the observed data point (actual experiment's outcome) belongs to the possible experimental outcomes.
   If so, we accept H_0.
   If not, we reject H_0.

Where's the issue?

How do we identify the outcomes we "expect" to observe under the hypothesis that Mendel was right?

How Mendel was eventually saved by the binomial

Working assumption

Mendel did not "intentionally select" the plants!

If the hypothesis is true

the 56 pea plants that Mendel grew can be seen as 56 "randomly selected" plants from our target population with infinite plants.

Hence

$y = 39$ is the observed value of a binomial random variable with $n = 56$ trials and, success probability ϑ; i.e. the probability that a pea plant will be "green".

What do we mean by random variable?

Crash course (with Kolmogorov's permission)

What is a dice?

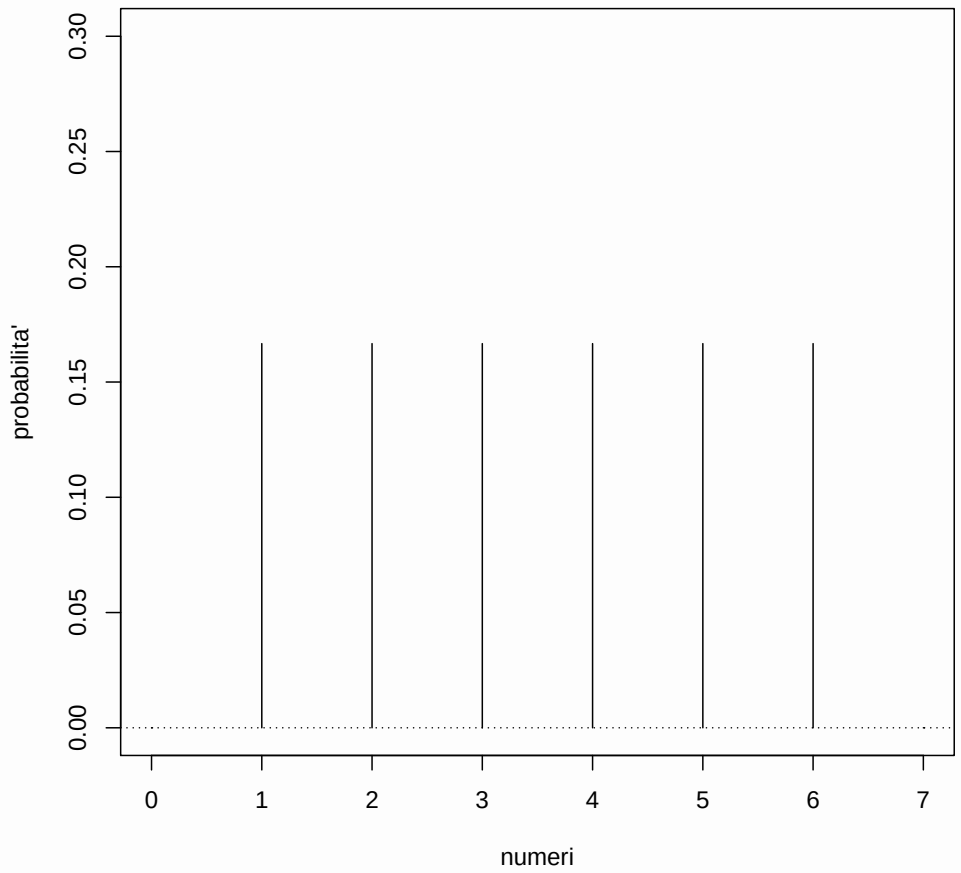
a random generator of numbers

We roll the dice a first time and obtain "3"; we roll it a second time and obtain "6"; a third time ...

How can we describe how a dice works?

1. by the numbers it generates, that is, $\{1, 2, 3, 4, 5, 6\}$.
2. by the probability with which the different numbers "occur". So, for instance, if the die is fair, the probability of obtaining any of the six faces is $1/6$.

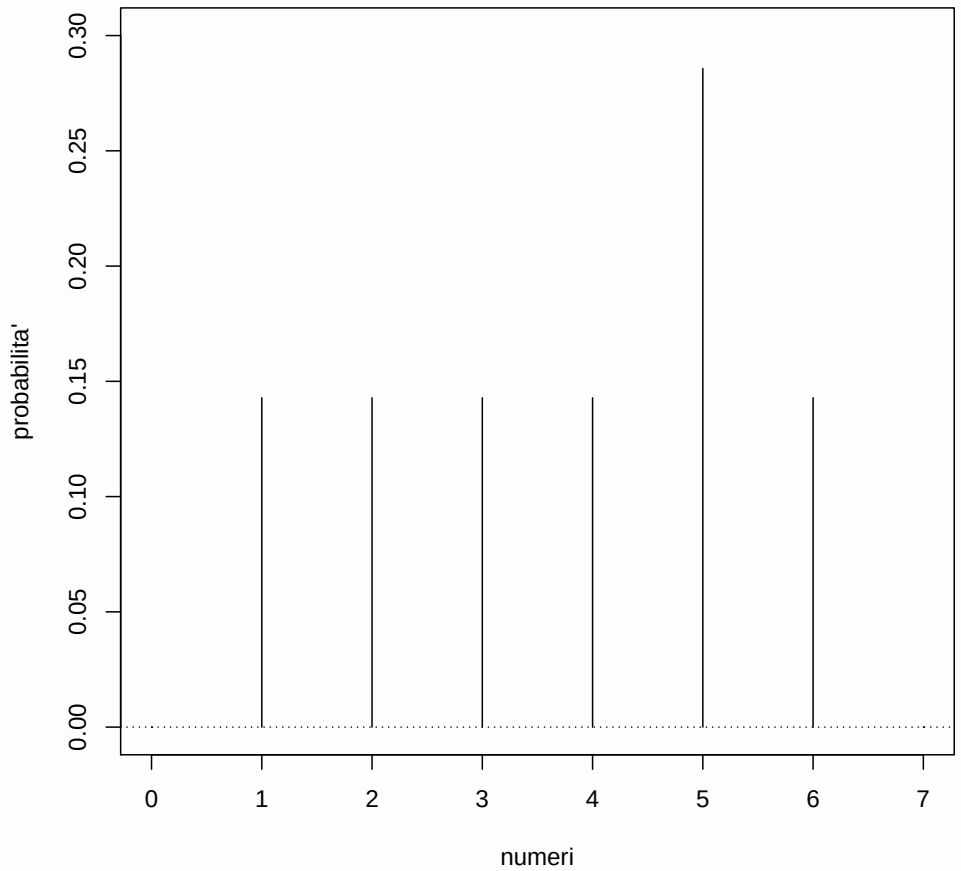
Probability distribution of a fair die



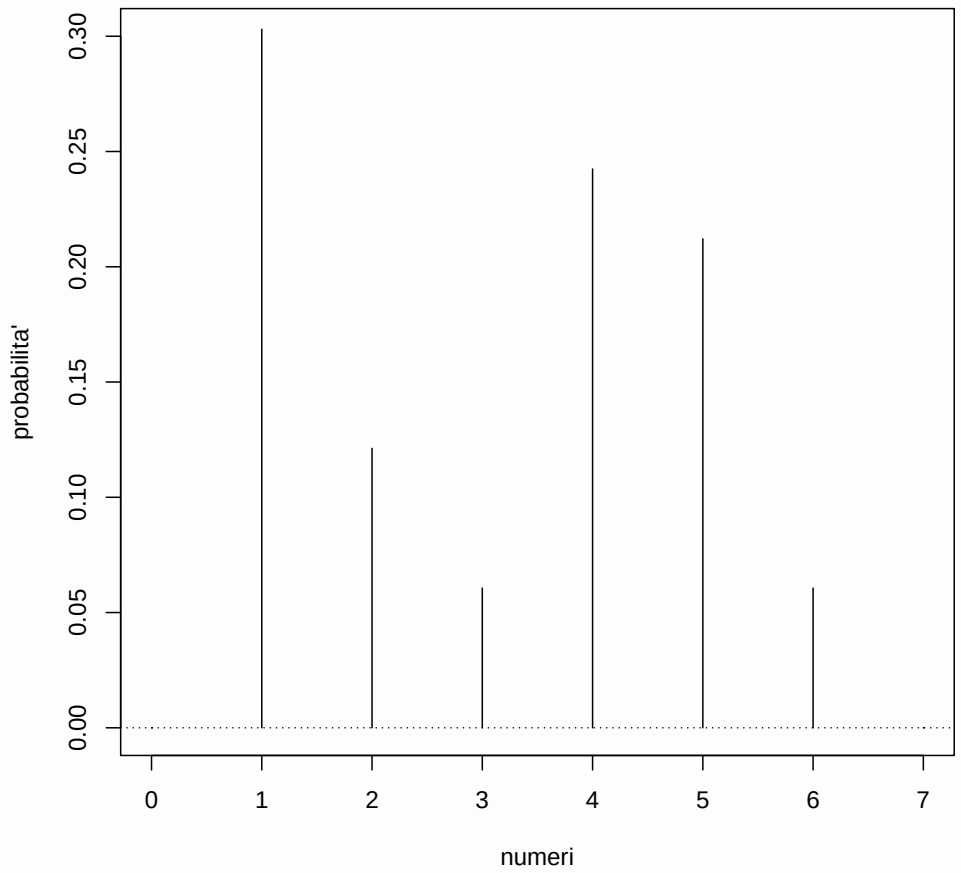
What do we mean by random variable?

Crash course (with Kolmogorov's permission)

Probability distribution of a "loaded" die



Probability distribution of another loaded die



What do we mean by random variable?

Crash course (with Kolmogorov's permission)

A random variable

is a machinery capable of generating random numbers.
I run an experiment and obtain a result; you run the same experiment and obtain a different result; and so on ...

So, what should a random variable do?

It tells us which values we may obtain and with what frequency.

For instance, from the previous figures, we deduce that "possible values" are $1, \ldots, 6$. They also reveal the frequency with which we would observe these values if we rolled the die repeatedly (= repeated the experiment).

The binomial distribution

Underlying experiment

- We conduct n sub-experiments (or trials).
- The result of a single trial can be a "success" with probability ϑ or a "failure" with probability $1 - \vartheta$.
- The trials are independent of each other (the outcome of a trial doesn't affect the result of a second one).

Definition

The number of "successes" follows a binomial random variable with number of trials equal to n and success probability ϑ.
The possible outcomes are $\{0, 1, \ldots, n\}$. Given n and ϑ, we can calculate the probability of observing these values.

So, how did the binomial save Mendel?

The outcome

$y = 39$ is the observed value of a binomial random variable with $n = 56$ trials and success probability ϑ, the true probability that a pea plant is "green".

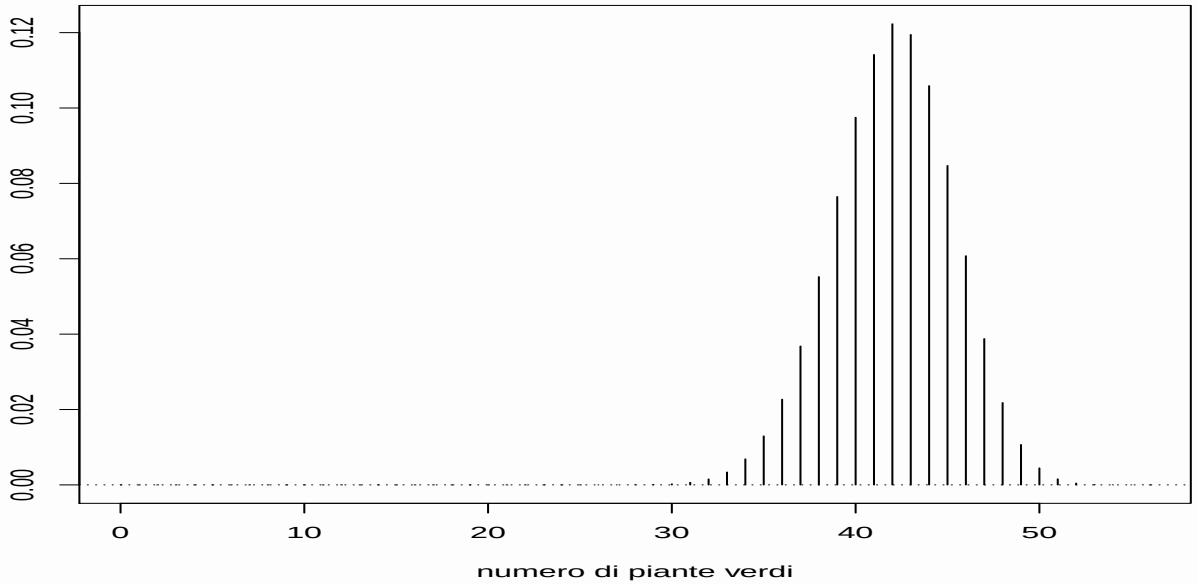
Why is this statement so important?

It describes the relationship between what we know (our data, "39 green" peas out of a "total of 56") and what we want to know (ϑ).

Binomial tells to Mendel

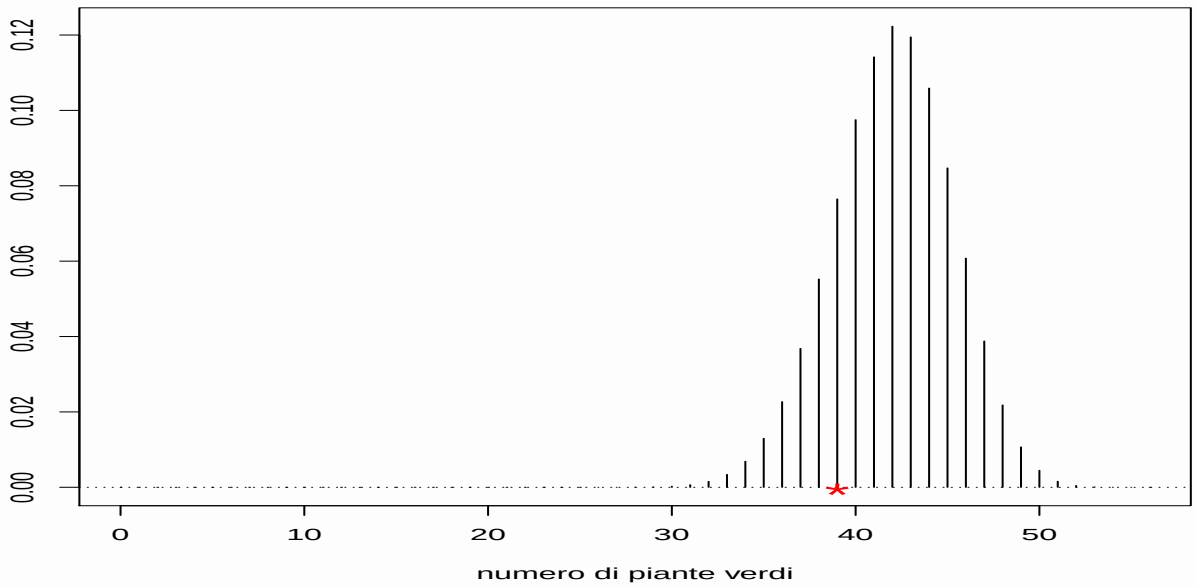
"Dear Gregor, these would be the possible outcomes if you are right!"

Probability distribution of a binomial with 56 trials and success probability 0,75



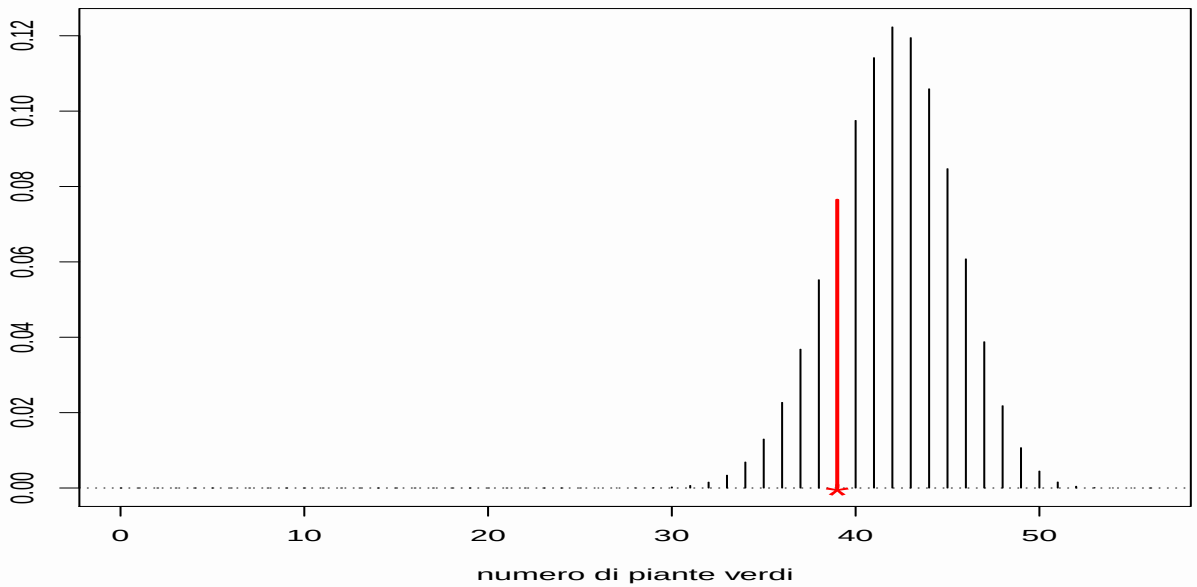
This distribution tells us how many "green plants" we can expect to observe if (i) the world satisfies Mendel's hypothesis ($\vartheta = 0{,}75$) and (ii) we repeat Mendel's experiment over and over again.

Probability distribution of a binomial with 56 trials and success probability 0,75



The "red star" pinpoints the experimental results (39).

Probability distribution of a binomial with 56 trials and success probability 0,75



The "red bar" highlights the probability of observing 39 if Mendel's hypothesis is true.

Final remarks

- We observe a result which is likely to occur under the hypothesis that

$$H_0 : \vartheta = 0{,}75.$$

  Indeed, under this hypothesis, observing 39 (that is, around 70% of) "green" plants out of a total of 56 second-generation plants doesn't surprise us; on the contrary, it is an event that occurs with an acceptable probability.

- Hence, the data advise us to "accept H_0"; that is, we produced the following statement about our world:

  "Given the result of our experiment, there is not enough evidence to question Medel's theory".

What would Mendel write in a (scientific) paper was he alive today?

Synonyms

p-value, p, (observed significance level)

What is it?

a measure of how much evidence the data provide in support of H_0
Assumes values in 0 and 1.
The larger it is, the more do the data "reproduce" what we expect to see under H_0.

We use it to communicate the findings of a test.

Excerpt of the "paper" Mendel may have written a couple of months ago

2° generation plants	"green" ones	p
56	39 ($\approx 70\%$)	0.36

p-value

as implemented in statistical software (an example from R)

```
> binom.test(x=39,n=56,p=0.75)

Exact binomial test

data:  39 and 56
number of successes = 39, number of trials = 56,
p-value = 0.3559
alternative hypothesis:  true probability of success
is not equal to 0.75
95 percent confidence interval:
0.5590326 0.8122013
sample estimates:
probability of success
0.6964286
```

Definition

$$\text{Prob}\left(\begin{array}{c}\text{of obtaining an experimental result}\\ \text{which is as far or farther}\\ \text{from } H_0 \text{ than the observed outcome}\end{array}\right)$$

calculated assuming the hypothesis H_0 is true.

Translation

- Suppose the world satisfies H_0;
- think of an infinite number of replications of the experiment;
- hence derive the frequency with which we would observe experimental results "as much (or more) inconsistent" with the null hypothesis H_0 then the observed outcome.

$p = 0$: interpretation

- If H_0 is true
- and we repeat the experiment an infinite number of times,
- we will NEVER observe a result as inconsistent or more "inconsistent with H_0" than the observed one.
- The observed outcome is hence "very, very far" from H_0; it's so much of an "astonishing" result that we wouldn't expect it in a world in which H_0 is true.
- *Conclusion:* the experiment conveys evidence against H_0.

$p = 1$: interpretation

- If H_0 is true
- and we repeat the experiment an infinite number of times,
- we will ALWAYS observe a result as inconsistent or more "inconsistent with H_0" than the observed one.
- The observed outcome is hence "very, very close" to H_0; indeed, it couldn't be closer given that all other possible experimental outcomes are either as "extreme" or more extreme.
- *Conclusions:* the experimental outcome doesn't allow us to question H_0.

The experimental outcome which is "closest" to H_0 is

$$56 \times 0{,}75 = 42 \text{ green plants}$$

If $y = 42$, then 75% of the sampled plants is green; under this scenario, it would be hard to question Mendel's theory.

Hence, "far from H_0" means "far from 42".

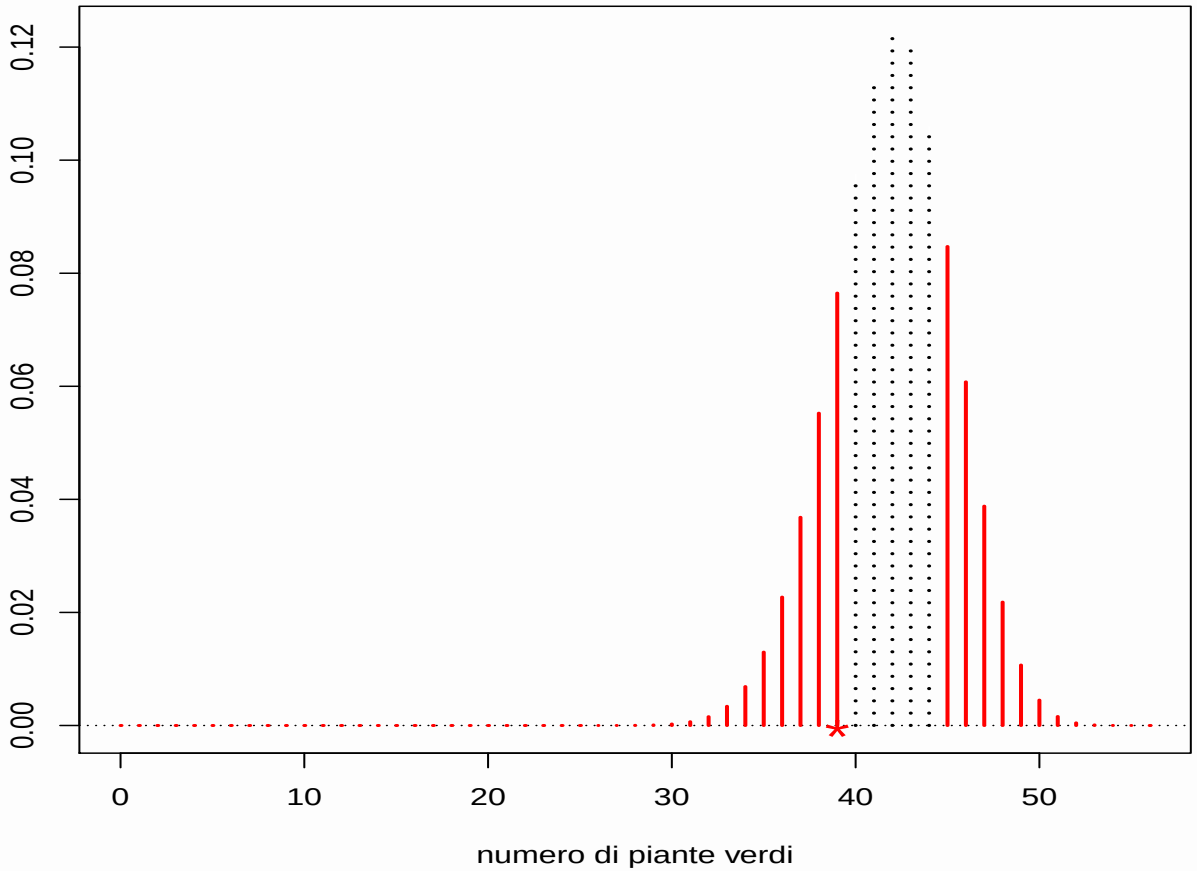
The experimental outcomes which are as far from H_0 or farther off are

to the left: $39, 38, 37, \ldots, 0$ green plants;

to the right: $45, 46, \ldots, 56$ green plants.

Note that: 39 e 45 are "as far" from the observed value; the remaining values are "farther off' the observed value.

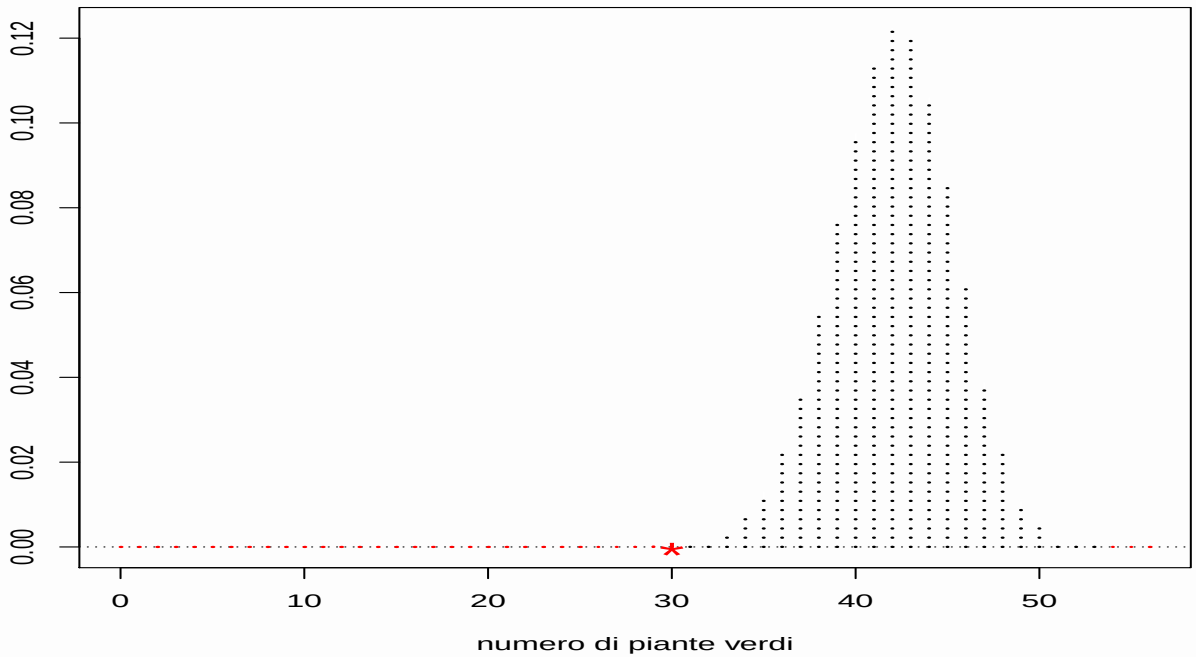
If $y = 39$, the p-value (= sum of red bars) is $\approx 0{,}36$



Interpretation

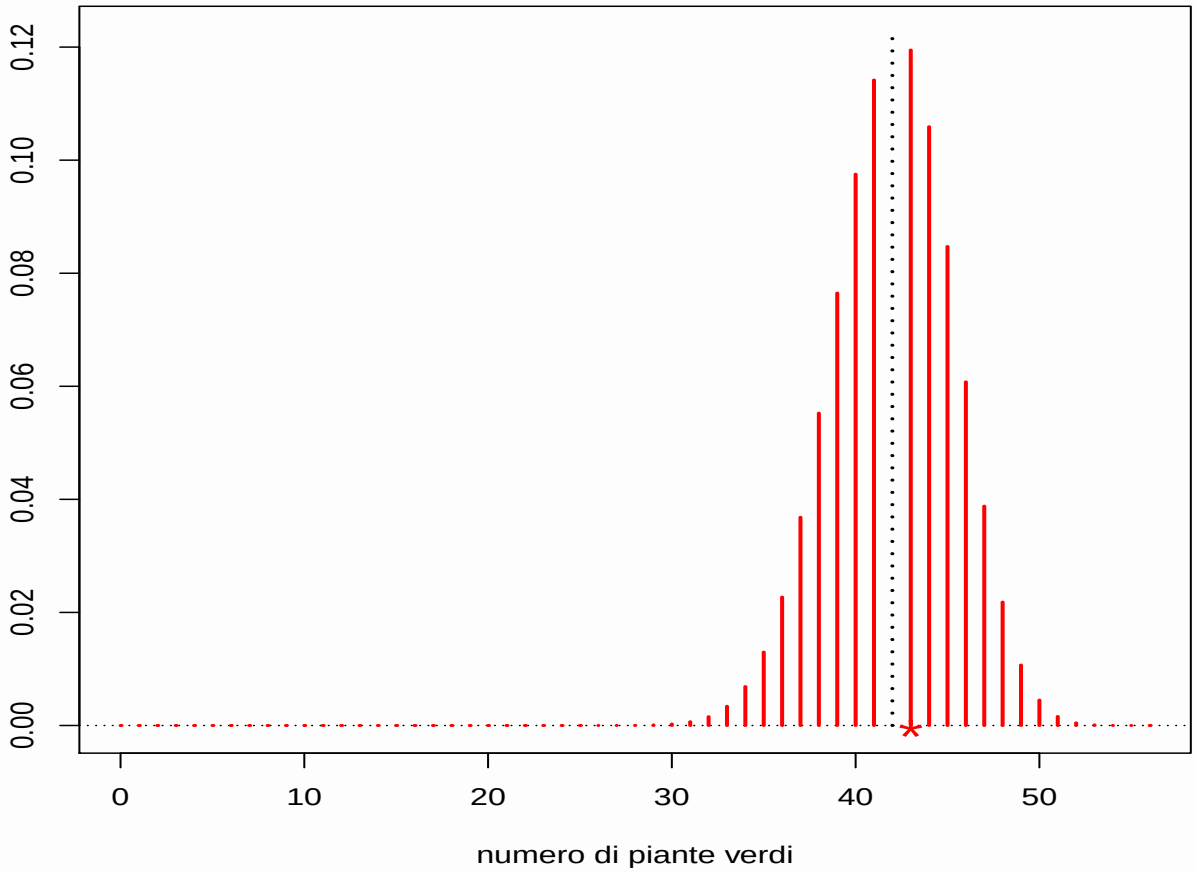
- (i) If Mendel is right ($\vartheta = 0{,}75$)
  (ii) and if we repeat the experiment again and again,
  around 36% of them will produce results which are as extreme or more extreme than that observed by Mendel.
- Hence, the experiment's outcome is consistent with what we would expect to observe if Mendel was right.
- *Conclusion.* The data do not provide enough evidence to doubt the theory developed by Mendel.

If $y = 30$, the p-value (= sum of red bars) is smaller than 0,001.



Interpretation: If Mendel's hypothesis were true, we would expect a result as "unusual" or more unusual than the observed one in less than once every 1000 repetitions of the experiment. This frequency is so small that we have to doubt the theory developed by Mendel!

If $y = 43$, the p-value (= sum of red bars) is 0,88.



threshold	meaning	"decision"
$p \leq \frac{1}{100}$	the results are highly significant (highly inconsistent with H_0)	reject H_0
$\frac{1}{100} < p \leq \frac{1}{20}$	the results are significant (inconsistent with H_0)	reject H_0 (though with less emphasis!)
$\frac{1}{20} < p \leq \frac{1}{10}$	the results are "borderline" (on the edge of significance)	why not repeating the experiment and collecting more data?
$\frac{1}{10} < p$	the results are not significant	accept H_0 (the larger p is, the less should we question our decision)

p-value

Caution! Do not over-interpret!

- There are no automatic rules. Every statistical conclusion needs to be accompanied by a biological one. Hence, *significant does not necessarily mean relevant*.
- The p-value IS NOT the probability that the null hypothesis is true. It measures how consistent the results are with what is expected under the null hypothesis.
- The more hypotheses you test, the more likely you will obtain a small p-value (more on this later!)

While we are on the topic, let Mendel (and us!) meet the normal distribution.

Introduction

- What is the purpose?
  To repeat the previous analysis of Mendel's data using the normal approximation for the binomial distribution.
- Why do we do it?
  To show, using a simple case, that we can carry out a statistical test even in situations where we know the sampling distribution only approximately.
- Don't worry! (ͨ☹) We must review the normal distribution and introduce a theorem.
- Indeed (ͨ☺): we will see how to derive confidence intervals for the binomial distribution (and not only).

Normal distribution

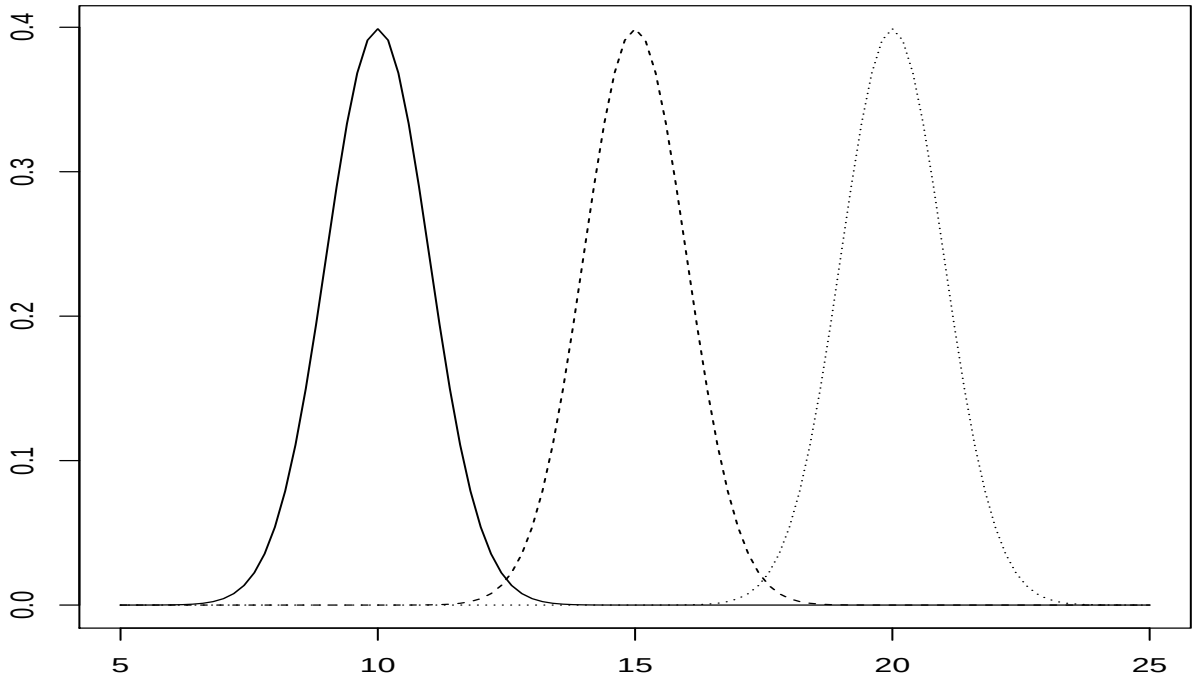
Remember!

- Also known as Gaussian, Gauss or Laplace-Gauss distribution.
- Depends on two parameters:
  1. the mean μ, determines the center of the distribution;
  2. the variance σ^2, governs the amount of variability.

  The standard notation for this distribution is $N(\mu, \sigma^2)$.
- If $\mu = 0$ and $\sigma^2 = 1$, we call it the *standard normal* distribution.

Normal distribution

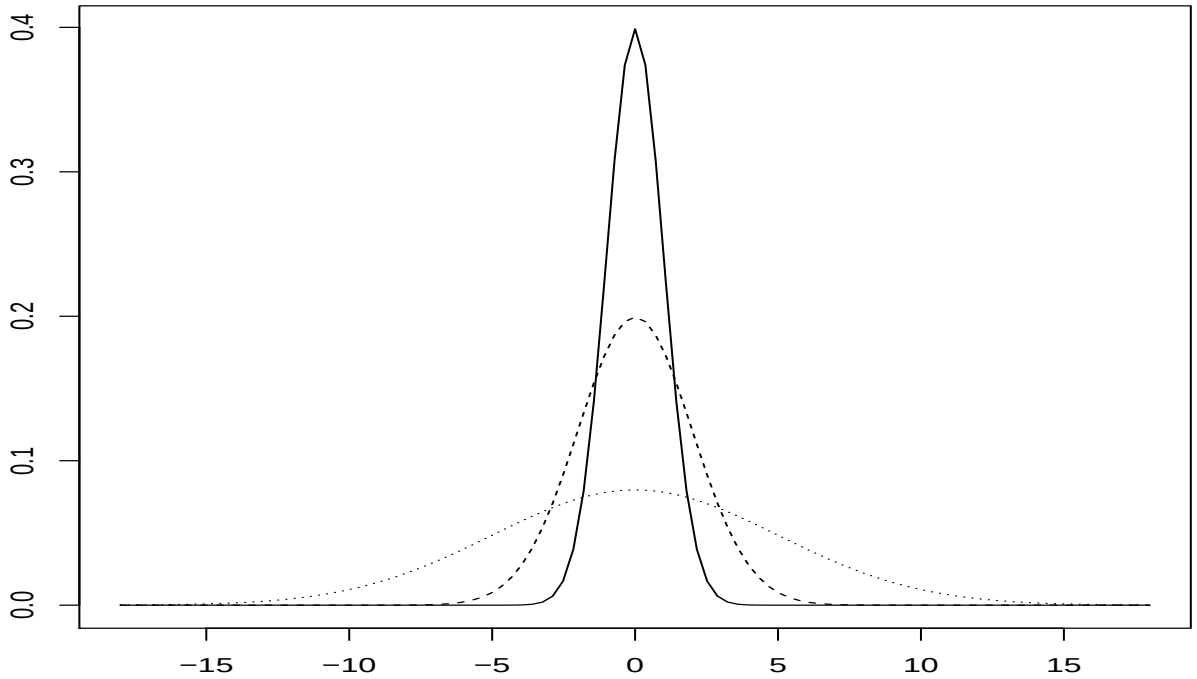
Three different values for μ



The solid line represents a normal distribution with mean 10, the dashed line with mean 15, the dotted line with mean 20; the variance is always 1.

Normal distribution

Three different values for σ^2

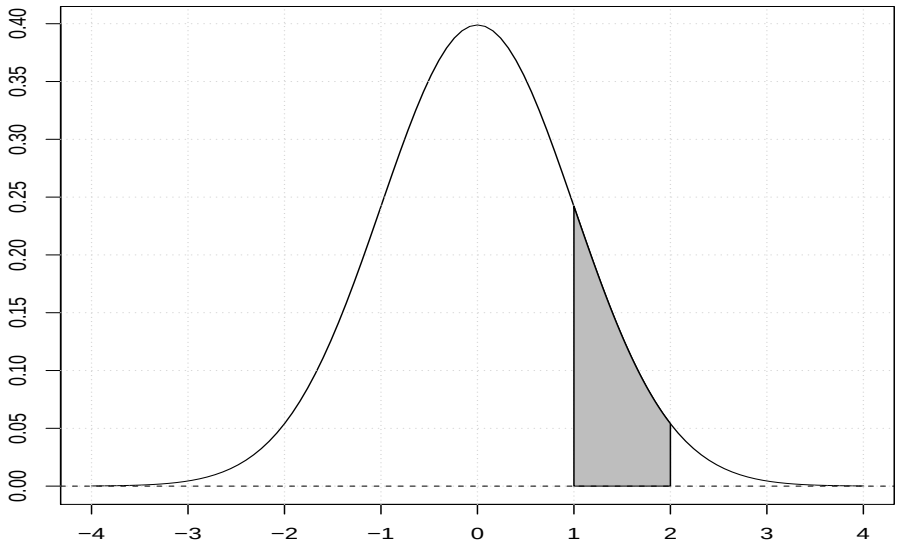


The solid line represents a normal distribution with variance 1, the dashed with variance 4, the dotted with variance 25; the mean is always 0.

Normal distribution

area=probability

The area under the curve of a normal distribution gives us the probability of an interval.



The curve describes the probability distribution of the *standard normal*.
The grey area is the probability of observing a value between 1 and 2.

Normal distribution

Remember!

- The normal distribution is symmetric around μ.
- Given two values $a \leq b$ we can calculate

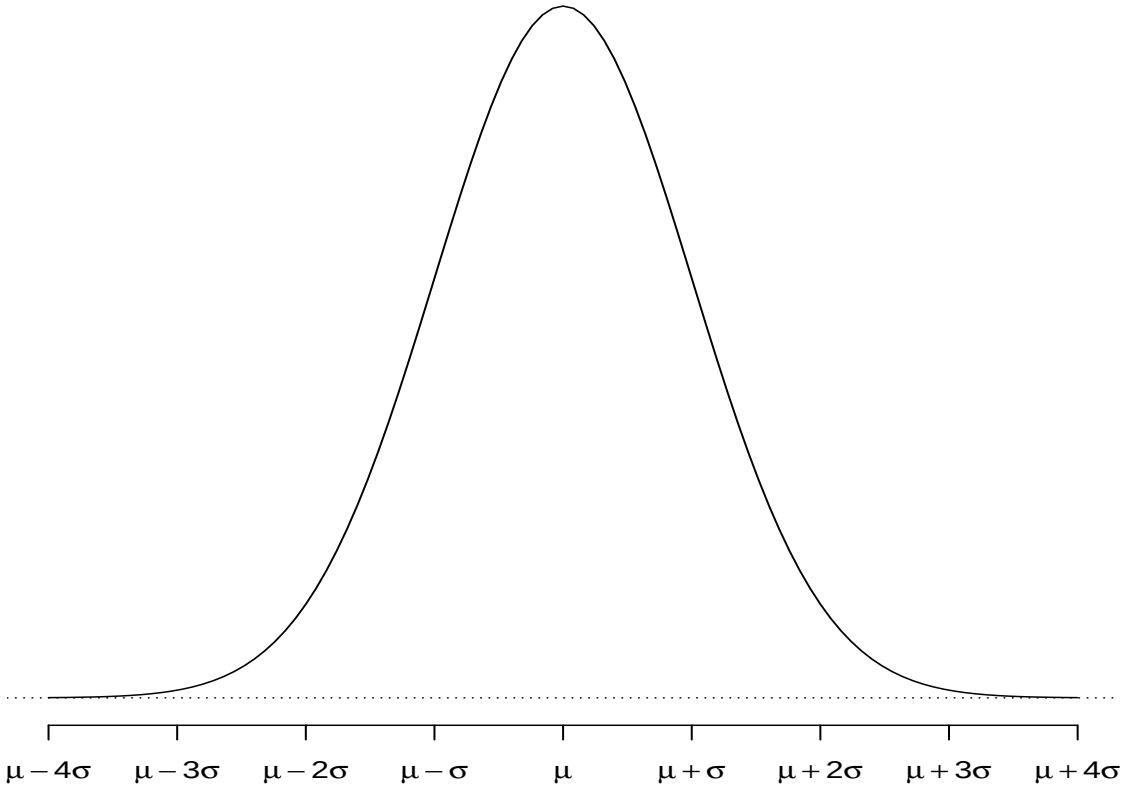
$$\text{Prob}(a \leq N(\mu, \sigma^2) \leq b).$$

- An experiment with a normal distribution yields any numerical value.

  So, the normal distribution isn't, for instance, bound to integer values between 0 and n as for the binomial distribution with several trials equal to n.

- Some values are almost impossible.

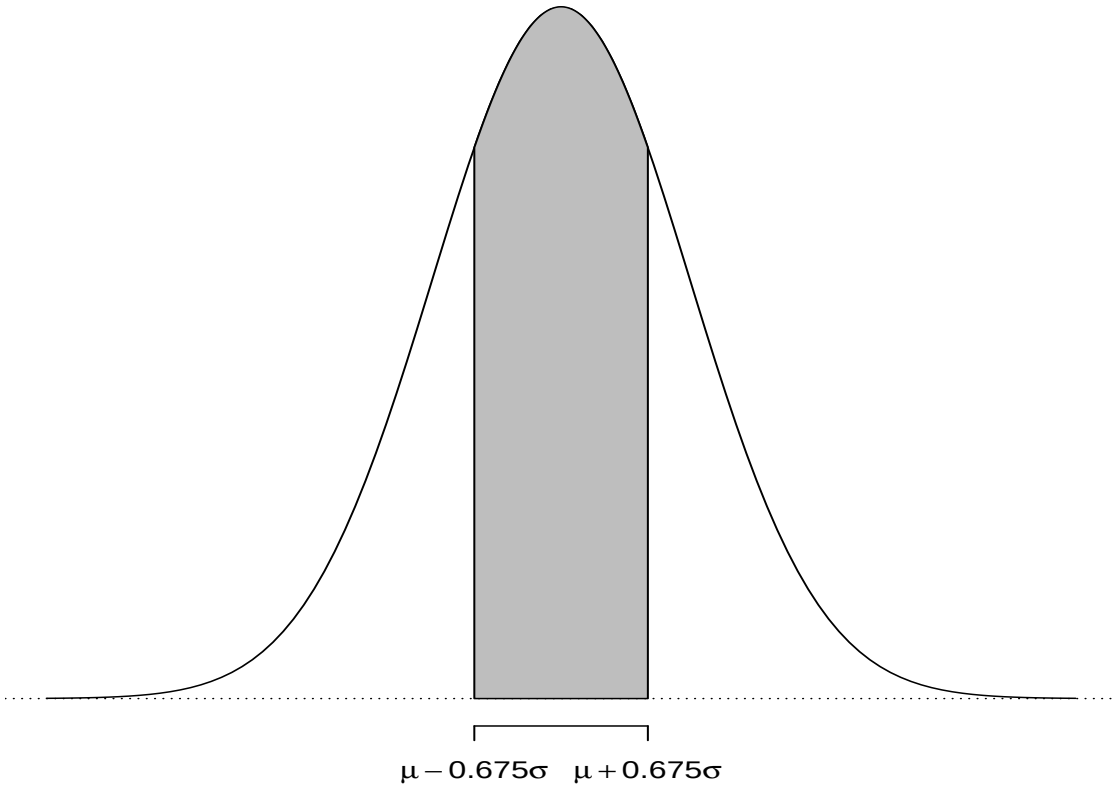
Normal distribution



almost the entire probability lies in the interval

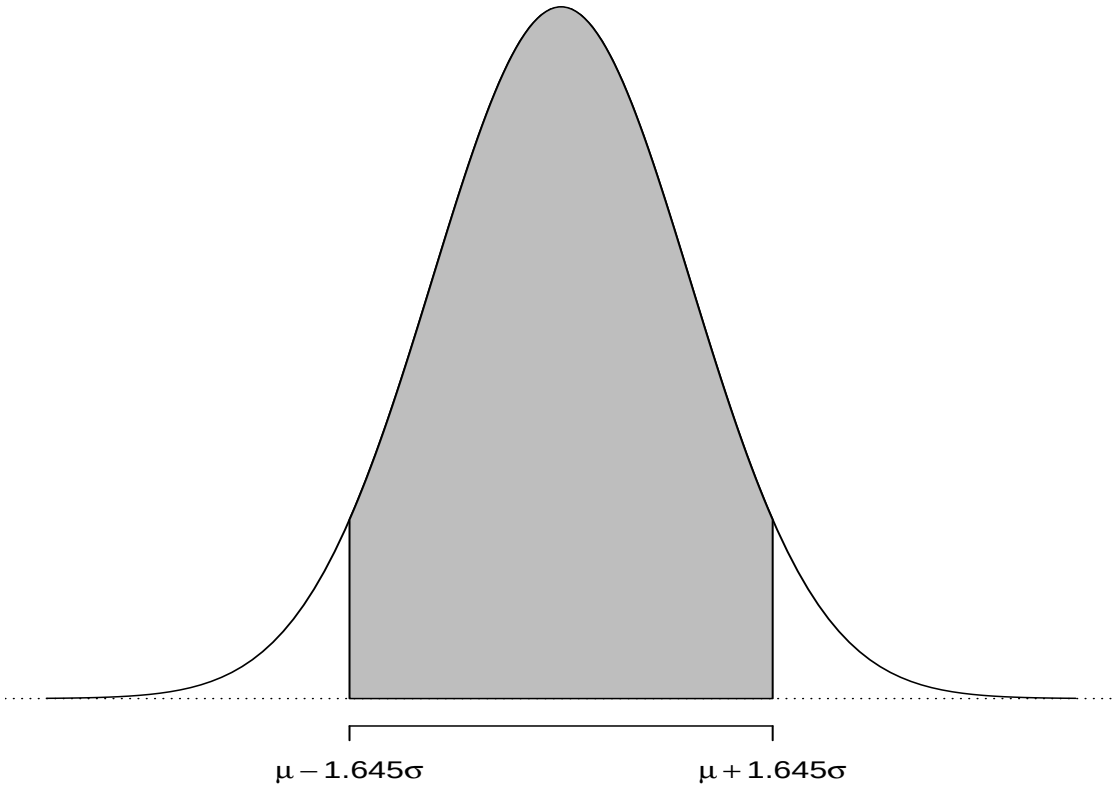
$$[\mu - 3\sigma, \mu + 3\sigma].$$

Normal distribution



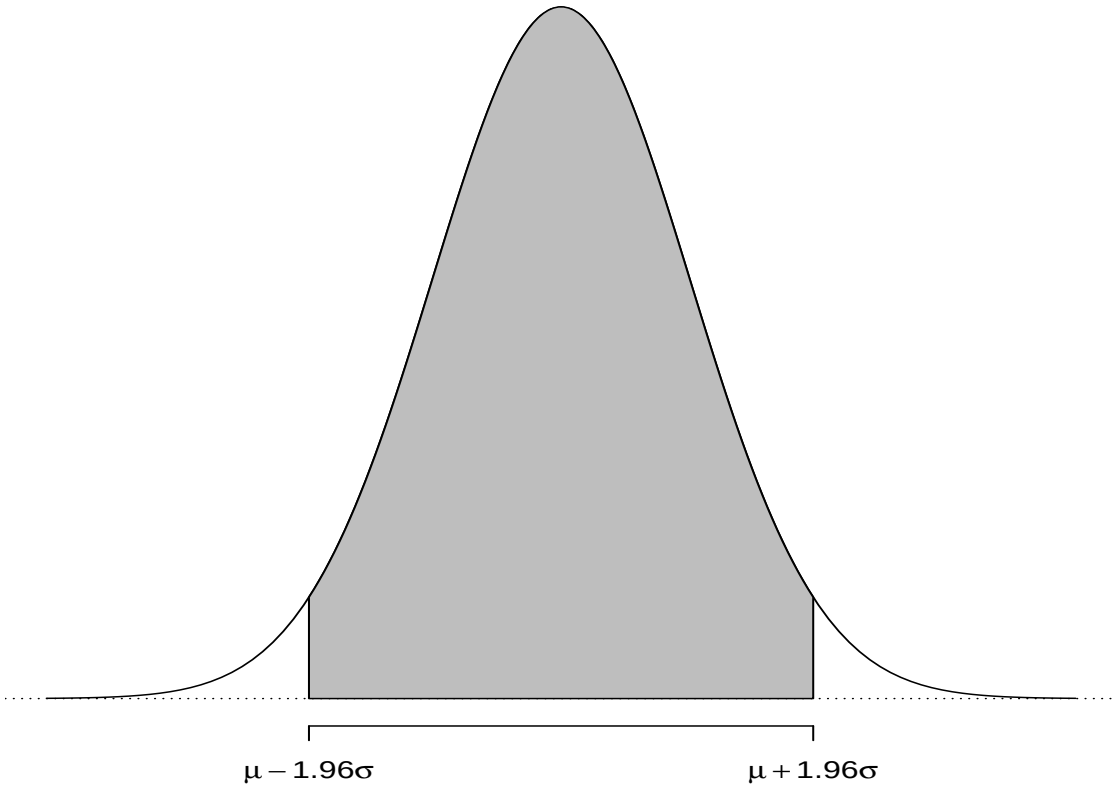
area under curve (= the probability) is 0.5.

Normal distribution



area under curve (= the probability) is 0.9.

Normal distribution

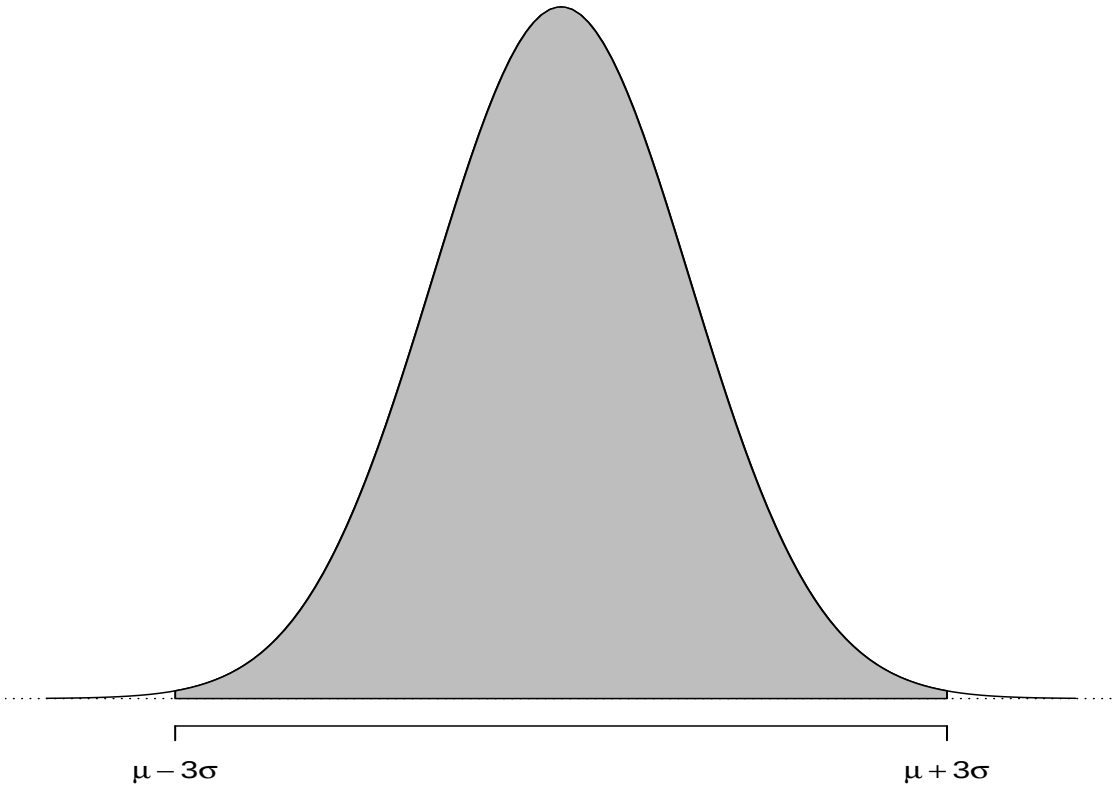


area under curve (= the probability) is 0.95.

Normal distribution

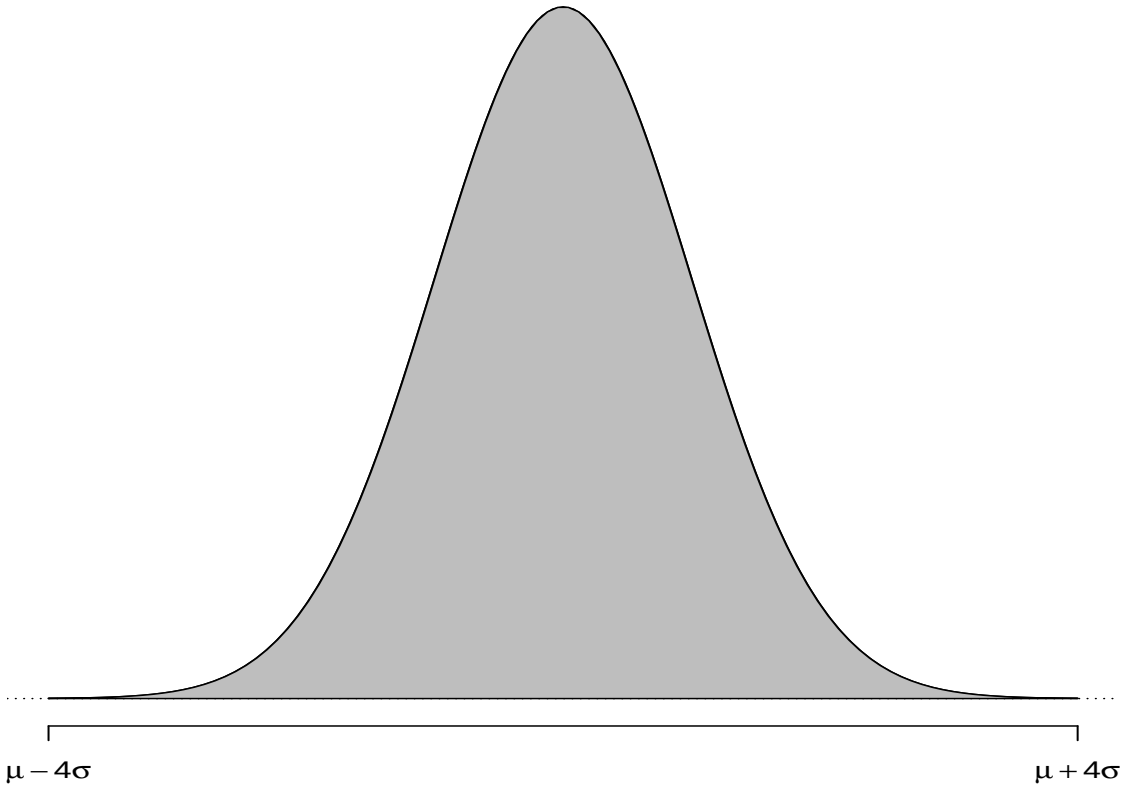
area under curve (= the probability) is 0.99.

Normal distribution



area under curve (= the probability) is 0.997.

Normal distribution



area under curve (= the probability) is 0.99994.

Normal approximation for the binomial

The De Moivre-Laplace theorem

Let Y be a binomial random variable with n independent trials and success probability ϑ. Write

$$w = \frac{\hat{\vartheta} - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} \quad \text{and} \quad z = \frac{\hat{\vartheta} - \vartheta}{\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}}$$

where $\hat{\vartheta} = Y/n$.
Then, if n is sufficiently large, the distribution of w and z can be approximated by the standard normal distribution ($\mu = 0, \sigma^2 = 1$).

Normal approximation for the binomial

Does it work?

- The theorem states that, if n is sufficiently large,

$$\text{Prob}(a \leq w \leq b) \approx \text{Prob}(a \leq N(0,1) \leq b)$$

where $a \leq b$ are two arbitrary numbers and $N(0,1)$ represents the standard normal distribution (with zero mean and unit variance). The same holds for z.

- The quality of the approximation improves with increasing n and becomes "acceptable" when

both, $n \cdot \vartheta$ and $n \cdot (1 - \vartheta)$, are larger than 5.

So, for instance, in Mendel's case, the condition is satisfied given that

$$n \cdot \vartheta = 56 \cdot 0.75 = 42 \quad \text{and} \quad n \cdot (1 - \vartheta) = 56 \cdot 0.25 = 14.$$

Is this of any use to Mendel?

- Compute

$$w = \frac{\hat{\vartheta} - 0.75}{\sqrt{\frac{0.75(1-0.75)}{56}}} = \frac{\frac{39}{56} - 0.75}{\sqrt{\frac{0.75(1-0.75)}{56}}} = -0.926$$

- If our world works according to Mendel's theory, we expect that $\hat{\vartheta}$ is close to 0.75.

  That is, H_0 true $\Rightarrow$ w close to zero.

- On the other hand, if Mendel's hypothesis is false, we expect that $\hat{\vartheta}$ is far from 0.75.

  That is, H_0 false $\Rightarrow$ w away from zero.

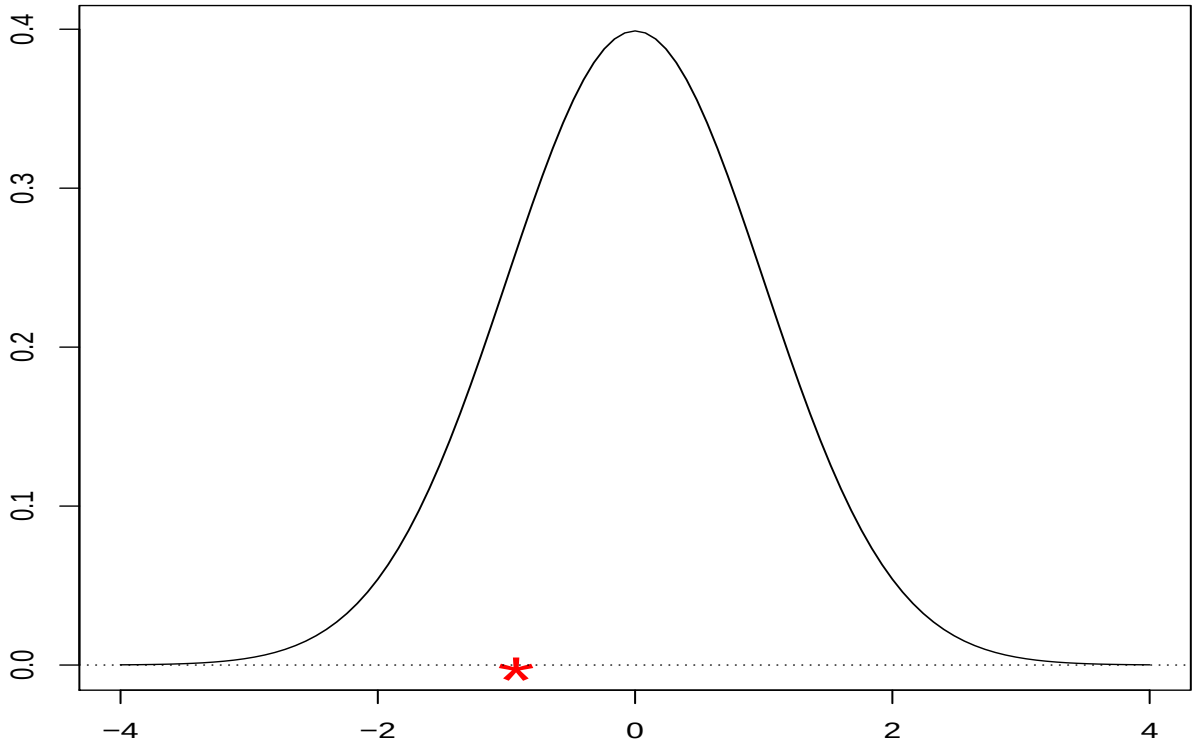
Is this of any use to Mendel?

- It seems sensible to
  - accept Mendel's hypothesis if $|w|$ is sufficiently small;
  - reject it if $|w|$ is too large.
- But, how large must $|w|$ be to have us doubting Mendel's hypothesis?

  Or, which are the values of w we expect to observe if Mendel was right?
- If Mendel is right, w follows the standard normal distribution.

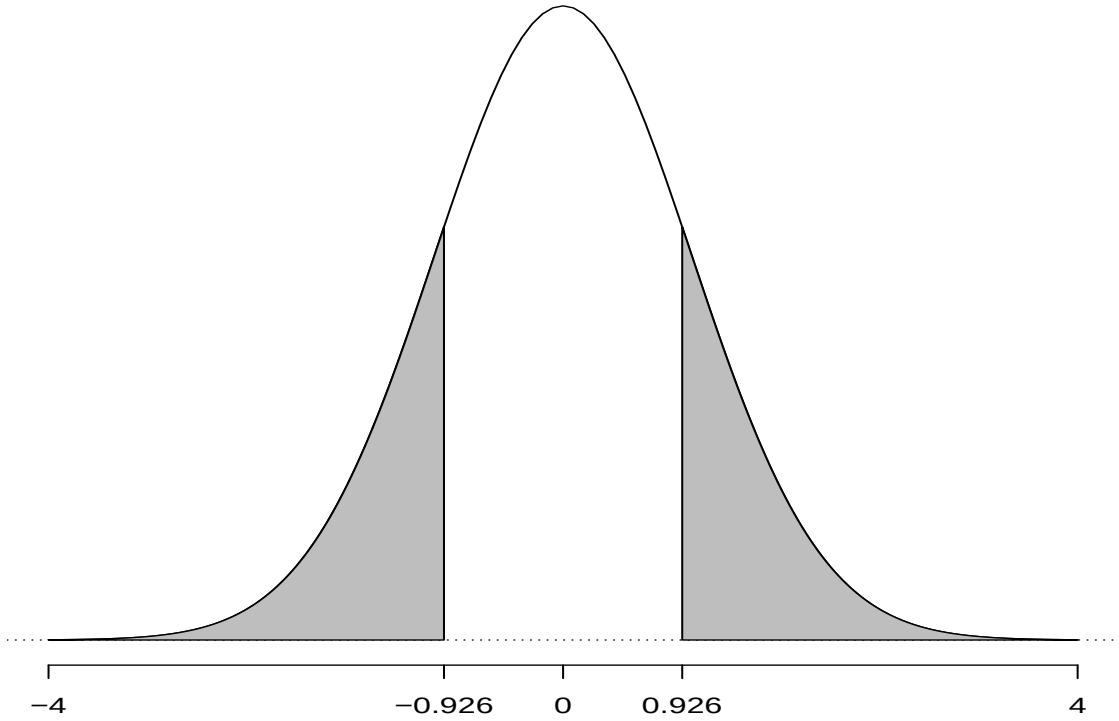
  Hence ... we know the answer!

Is this of any use to Mendel?



The red star pinpoints the value of w obtained from the data (-0.926). The solid curve represents the standard normal $N(0, 1)$. The observed outcome may have been generated from this distribution. There is not enough evidence to question Mendel's hypothesis.

Is this of any use to Mendel?



If we think in terms of w

"far from H_0" $\Leftrightarrow$ "far from 0".

The shaded grey area hence approximates the *p-value*.

Is this of any use to Mendel?

(approximate) computation of the p-value using R

```
> prop.test(x=39,n=56,p=0.75,correct=FALSE)

  1-sample proportions test without continuity correction

data:  39 out of 56, null probability 0.75
X-squared = 0.8571, df = 1, p-value = 0.3545
alternative hypothesis:  true p is not equal to 0.75
95 percent confidence interval:
0.5666413 0.8009967
sample estimates:
p
0.6964286
```

Note that. `X-squared` represents the squared value of w. To question H_0 for large $|w|$ amounts to question H_0 for large w^2.

How to carry out a statistical test

1. We want to know whether the world satisfies a conjecture of us.
   Example: Does 0.75 represent the probability that a second generation plant is "green"?
2. We hence collect data on the part of the world we are interested in.
   Example: 39 out of 56 second generation plants are "green".
3. Using the data we calculate a test statistic, say T, which tends to assume different values depending on whether our hypothesis is true or is false.
   Example: $T = w = (\hat{\vartheta} - 0.75)/\sqrt{0.75(1-0.75)/56}$
4. We match the observed value of T with the probability distribution of the outcomes we expect to observe if our hypothesis is coherent with the world. In particular, we use the p-value to measure how "close" the observed value of T is to the values we expect to observe if the hypothesis is true.

But, ... what about z?

- Good question! To understand the answer, we need some (little!) algebra.
- As z follows approximately the standard normal, we can write

$$\text{Prob}(-1.96 \leq z \leq 1.96) \approx 0.95.$$

But, ... what about z?

- Now, let's replace z by its expression, to give

$$\text{Prob}\left(-1.96 \leq \frac{\hat{\vartheta} - \vartheta}{\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}} \leq 1.96\right) \approx 0.95.$$

- Then rewrite the inequality by isolating ϑ, so that

$$\text{Prob}\left(\hat{\vartheta} - 1.96\frac{s}{\sqrt{n}} \leq \vartheta \leq \hat{\vartheta} + 1.96\frac{s}{\sqrt{n}}\right) \approx 0.95.$$

where

$$s = \sqrt{\hat{\vartheta}(1-\hat{\vartheta})}.$$

Wow! Nice result!

- We want to learn about ϑ, the "true" probability of growing a 2° generation plant with "green" pods;
- According to the observed data, ϑ could be 0.75; On the other hand, we cannot exclude 0.70 as another "plausible" value since the data support it.
- The previous "sentence" tells us, indeed, that the interval

$$[\hat{\vartheta} - 1.96\frac{s}{\sqrt{n}}; \hat{\vartheta} + 1.96\frac{s}{\sqrt{n}}]$$

which we obtain from our data, includes the true value of ϑ with a level of confidence approx. 95%.

Applied to Mendel's data

- We "produced" the following statement about our world
  "we don't know the probability of growing a "green" plant; but, with high confidence, i.e. approximately 95%, this probability is between 0.58 *and* 0.82*."*
- Indeed

$$\hat{\vartheta} = \frac{39}{56} = 0.70, \; s = \sqrt{\hat{\vartheta}(1-\hat{\vartheta})} = \sqrt{0.7 \times 0.3} \approx 0.46$$

  Hence, our interval is

$$\hat{\vartheta} \pm 1.96 \frac{s}{\sqrt{56}} \approx 0.7 \pm 1.96 \frac{0.46}{\sqrt{56}} \approx [0.58; 0.82].$$

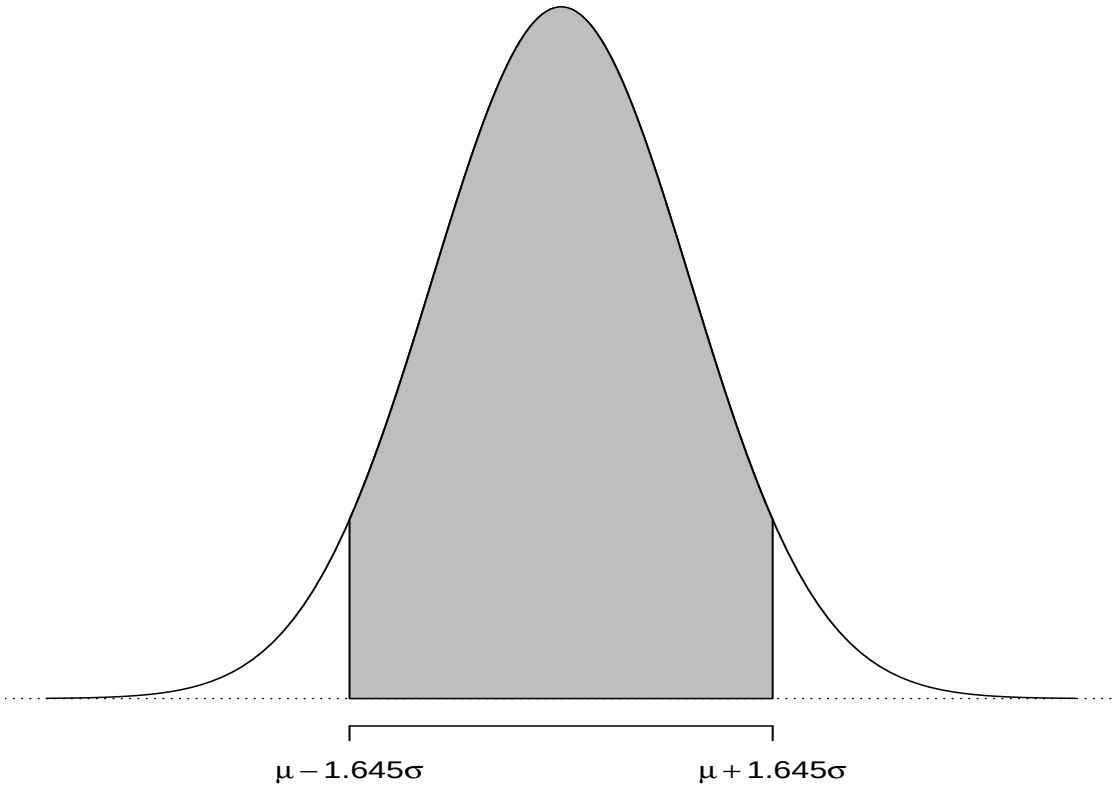
Confidence intervals

- We've just constructed a confidence interval with 95% confidence level.
- We may replace 1.96 with "other numbers" to obtain confidence intervals with varying probabilities of including the true value of the parameter.
- Generally speaking, the starting point is $z_{1-\alpha/2}$, which satisfies

$$\text{Prob}(-z_{1-\alpha/2} \leq N(0,1) \leq z_{1-\alpha/2}) = 1 - \alpha.$$

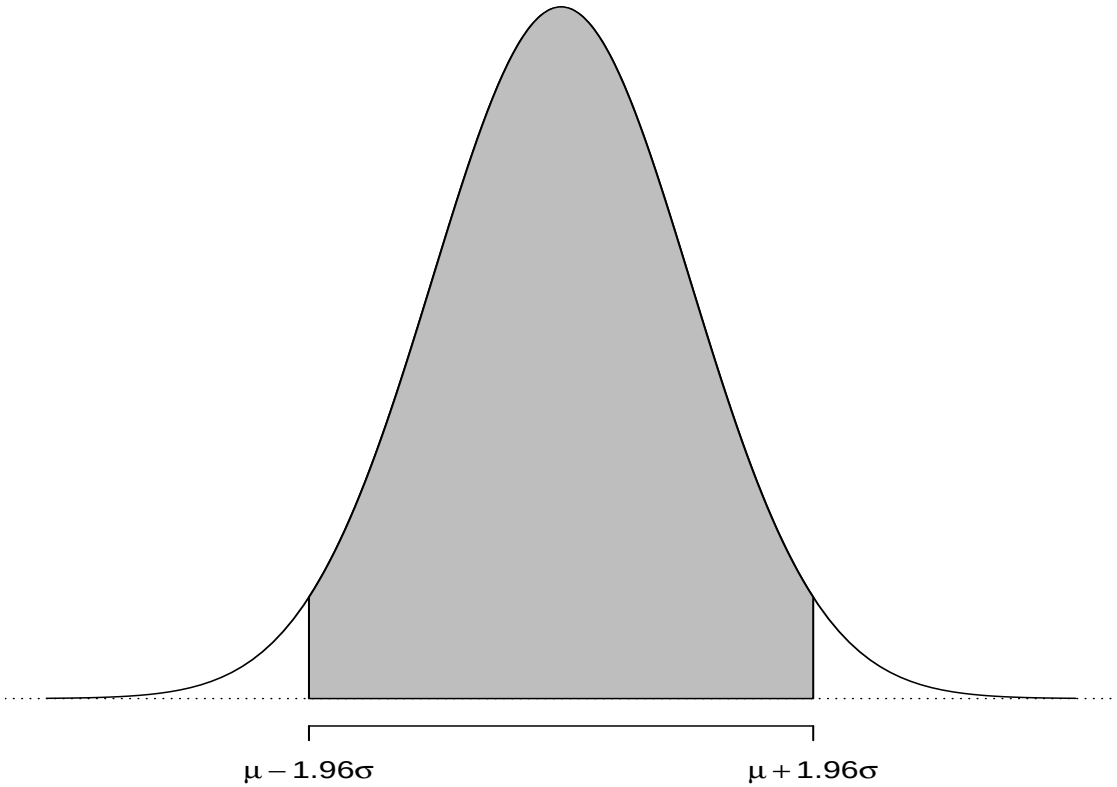
α	$1-\alpha$	$z_{1-\alpha/2}$
0.10	0.90	1.645
0.05	0.95	1.960
0.01	0.99	2.576

Normal distribution



area under curve (= probability) is 0.9.

Normal distribution



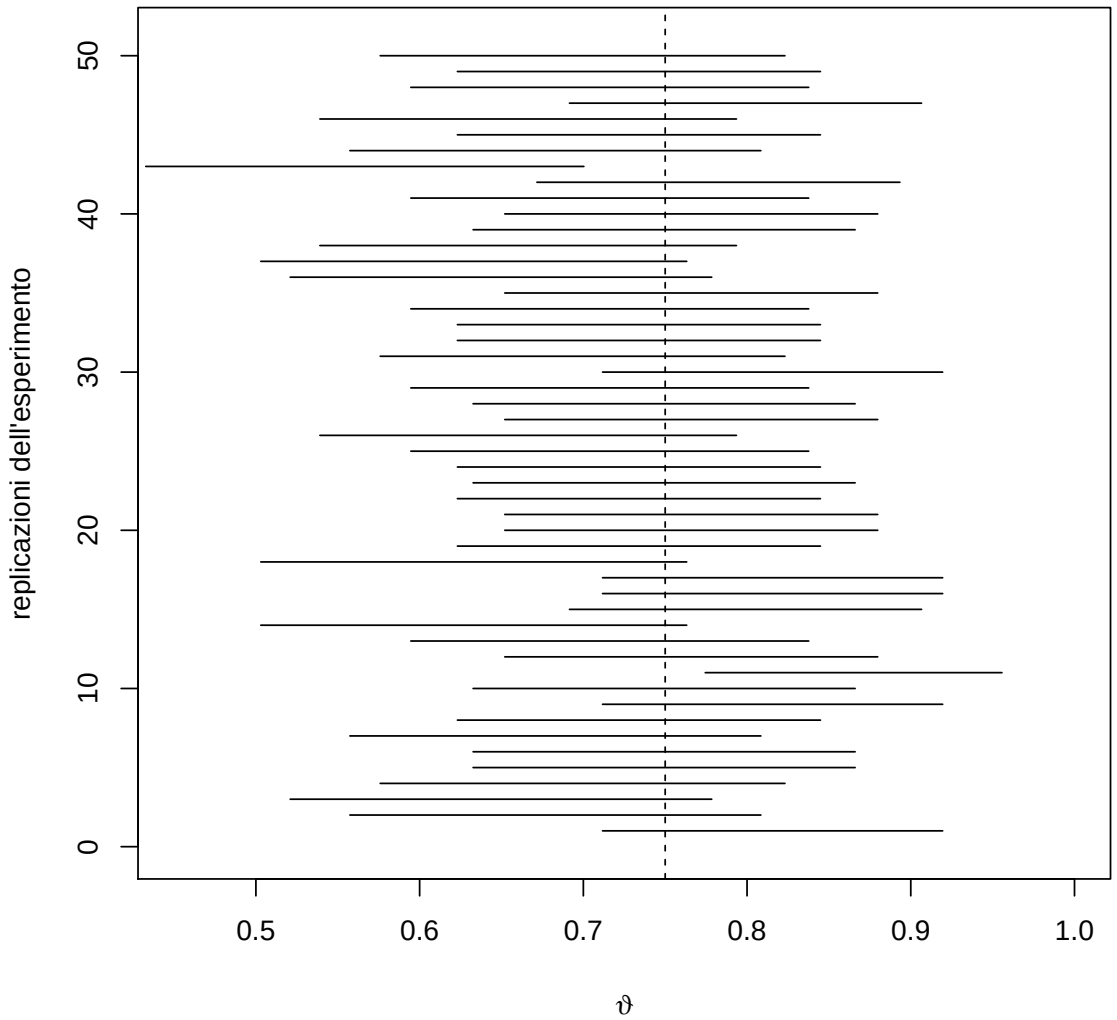
area under curve (= probability) is 0.95.

Normal distribution

area under curve (= probability) is 0.99.

Beware!

If we repeat the experiment, the interval changes (as it depends on the data), not ϑ (which represents a feature of our world).



The curse of $\sqrt{n}$

- The width of a confidence interval decreases with increasing sample size n, that is, with increasing number of trials.
- However, the speed at which this occurs is only $1/\sqrt{n}$. Hence, if we want
  - to halve the width, we must quadruple the sample size;
  - to reduce the width by a factor of 10, we must multiply n by a factor of 100;
  - ...

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Part 3

Analysing one or several means

Table of Contents

6 Checking the calibration of a laboratory instrument

7 Why 8 measurements?

8 More on measuring the INR

9 Further tests based on the normal distribution

10 A brewer statistician meets butterflies, cuckoos, wrens and redbreasts

Checking the calibration of a laboratory instrument

Description of the problem

The International Normalised Ratio INR

is a standardised measure (i.e., it doesn't depend on the instrument and method of measurement used) for blood clotting speed.

Instrument

receives a "drop of plasma" and provides an INR measurement.

Measurement error

The measurement is "dirtied" by

- random error: unavoidable and always present;
- a possible systematic error: due to imperfect calibration of the instrument
  This latter error can be eliminated but then "reappears" because of "breakages" or "wrong settings" of the instrument.

Description of the problem

Distribution of measurements

well-calibrated instrument	$N(\text{INR}, \sigma_0^2)$
wrongly calibrated instrument	$N(\text{INR} + \delta, \sigma_0^2)$

where

σ_0^2: random measurement error
We assume it to be constant and equal to 0.07^2.

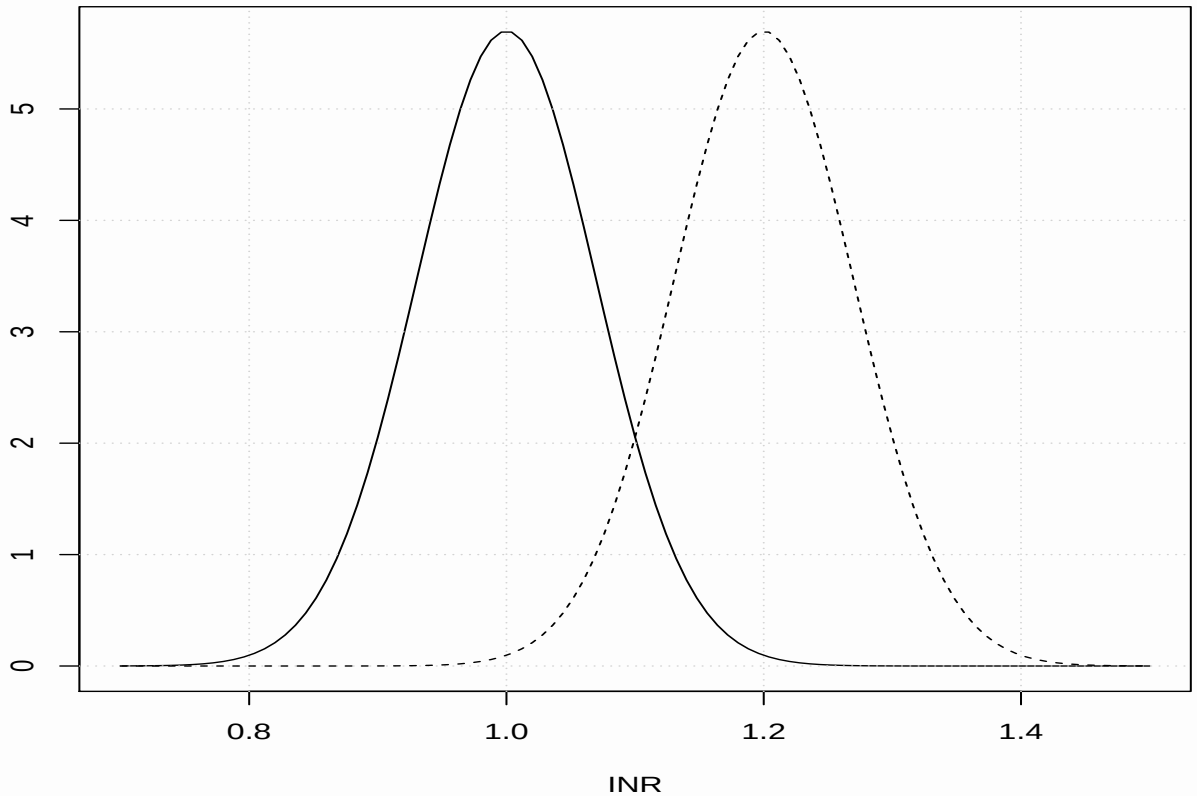
δ: systematic error

In addition, we assume that the different measurements are independent, even when obtained from the same blood sample.

Note. The above settings (normality, independence, value of σ_0) were obtained by analysing thousands of measurements.

Description of the problem

Distribution of measurements (INR = 1)



The solid line shows the distribution of the measurements when the instrument is well calibrated. The dashed line shows the same distribution but assumes that there is a systematic error ($\delta = 0.2$).

Description of the problem

Data

To verify whether the instrument is well-calibrated,

- 8 measurements are taken at the beginning of each daily shift
- using a (possibly artificial) blood sample with a known INR value equal to μ_0.

D&R. How is it possible to obtain blood samples with "known" INR? Well, this is a question you will be able to answer one day . . .

Excerpt

day	μ_0	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
A	1.41	1.346	1.401	1.422	1.41	1.291	1.433	1.376	1.518
B	0.94	0.904	1.044	0.979	1.070	1.019	1.070	1.048	1.052

Description of the problem

Why do we collect these data?

We want to "determine" whether

$$H_0 : \text{the instrument is well calibrated}$$

or

$$H_1 : \text{the instrument isn't well calibrated.}$$

If we decide "in favour of H_1", the instrument will be recalibrated. This is a rather long-lasting and costly operation that we want to avoid unless it is needed (that is, as long as the instrument looks well-calibrated).

Testing hypotheses on the mean of a normal distribution (with known variance)

The problem

data: n independent observations

$$y_1, \ldots, y_n$$

from a normal distribution with unknown mean μ and known variance σ_0^2

hypothesis: We want to verify whether

$$H_0 : \mu = \mu_0 \quad \text{or} \quad H_1 : \mu \neq \mu_0$$

where μ_0 is a preassigned value.

A cornerstone result

Working assumption

$y_1, \ldots, y_n$ are independent draws from a normal distribution with mean μ and variance σ^2.

Result

The sample mean of the n measurements

$$\overline{y} = \frac{y_1 + \cdots + y_n}{n}$$

distributes like a normal distribution with mean μ and variance σ^2/n. Hence,

$$z = \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

distributes like the standard normal.

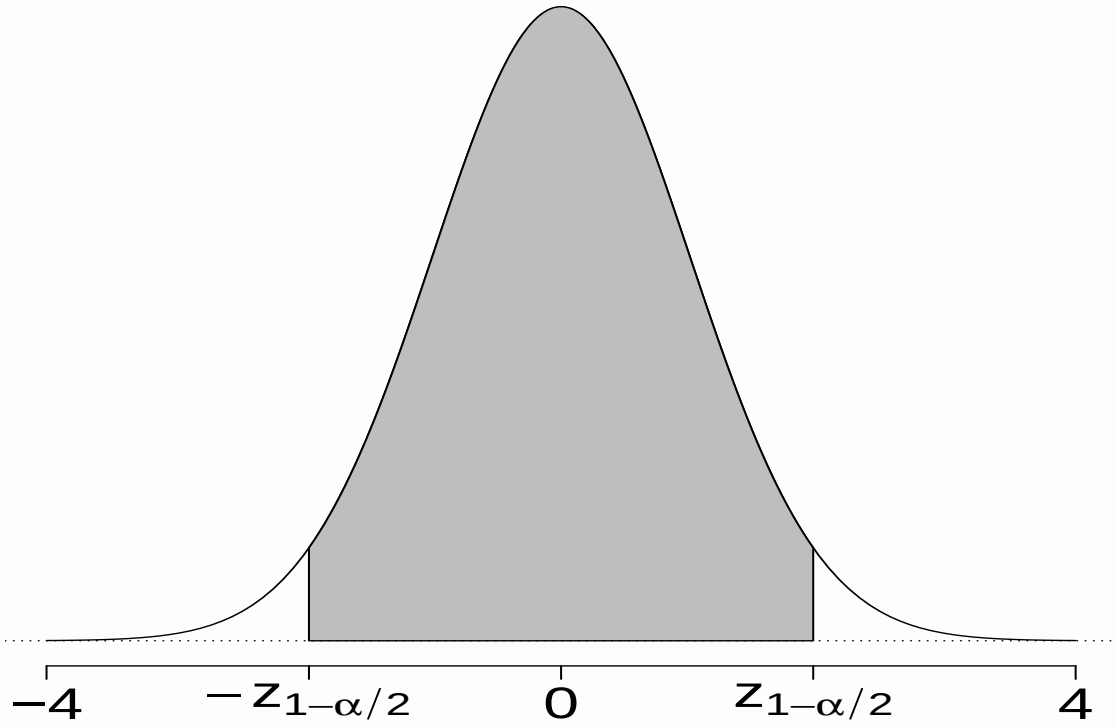
Testing hypotheses on the mean of a normal distribution (with known variance)

Methodology (without using the p-value)

1. Fix α and determine $z_{1-\alpha/2}$, that is, the $1-\alpha/2$ quantile of the standard normal.
2. Compute the sample mean (using the data)
$$\overline{y} = \frac{y_1 + \cdots + y_n}{n}.$$
3. Compute
$$z = \frac{\overline{y} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}.$$
4. Check whether z lies between $-z_{1-\alpha/2}$ and $z_{1-\alpha/2}$.
If the answer is yes, accept H_0.
If the answer is no, reject H_0.

Testing hypotheses on the mean of a normal distribution (with known variance)

Graphical interpretation of $z_{1-\alpha/2}$



The "grey" area is $1-\alpha$. Both "white" areas, on the left and the right, are $\alpha/2$.

Testing hypotheses on the mean of a normal distribution (with known variance)

What does α represent?

α is the probability of rejecting H_0 when it is in effect true.
For example, in our case, it is the probability of

- saying that the instrument lost the calibration
- when it is, in fact, well-calibrated.

Application to calibration control

How to chose α

In our case, by fixing $\alpha = 1/c$ for a fixed integer value c, we let the procedure wrongly cry

> *"Beware a wolf!" (Or rather, "The instrument lost its calibration!")*

one time every c day when, and vice versa, the instrument is well-tuned.
Since tolerating a false alarm (that is, a "useless recalibration") every 20th one was considered to be the "maximum bearable" amount, α was set to

$$\alpha = \frac{1}{20} = 0{,}05.$$

The corresponding $z_{1-\alpha/2}$ value is hence 1,96.

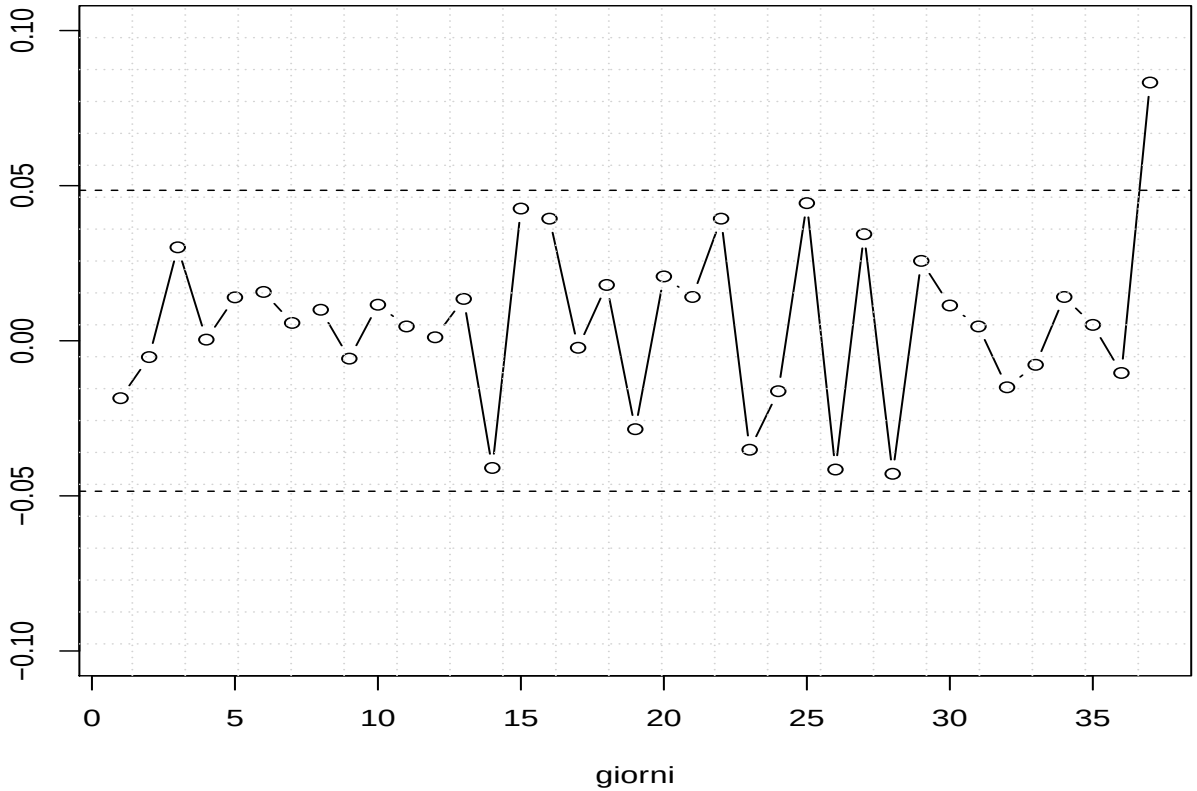
Application to calibration control

Numerical results for two days

day	$\overline{y}$	$\hat{\delta} = \overline{y} - \mu_0$	z	decision
A	1,40	-0,01	-0,42	The instrument seems well calibrated. We can start immediately with the analysis of today's blood samples.
B	1,02	0,08	3,36	The instrument isn't well calibrated. Before starting the analysis of today's blood samples, we must recalibrate it.

Application to calibration control

Trace plot of $\hat{\delta} = \overline{y} - \mu_0$



The figure considers 37 succeeding "working" days. The two previously analysed days are the last two (days 36 and 37). The "dashed bands" are at $\pm 1{,}96 \cdot 0{,}07/\sqrt{8}$. Every time we leave these bands, the instrument is recalibrated!

Testing hypotheses on the mean of a normal distribution (with known variance)

Using the p-value

We may also carry out the test using the p-value.
In our case,

$$\text{“far from } H_0\text{”} \iff \text{“}z \text{ far from } 0\text{”}.$$

Hence,

$$p = \text{Prob}(N(0,1) < -|z|) + \text{Prob}(N(0,1) > |z|).$$

Of course, the conclusions will be the same.

day	$\overline{y}$	z	p-value
A	1,40	-0,42	0,6745
B	1,02	3,36	0,0008

Why 8 measurements?

Motivation

- I want to mention how it is possible to determine the sample size when the main interest focuses on hypothesis testing.
- Obviously, the larger the sample is, the more the data will be able to identify the correct hypothesis.

- 8 INR measurements
- These are independent draws from a $N(\mu_0, \sigma_0^2)$ if the instrument is well calibrated.
- These are independent draws from a $N(\mu_0 + \delta, \sigma_0^2)$ if the instrument is not well calibrated.
- We want to use the data to discriminate between the following two hypotheses:

$$H_0 : \delta = 0 \quad \text{or} \quad H_1 : \delta \neq 0.$$

- If we "reject" H_0, the instrument will be recalibrated.

Two days measurements

day	μ_0	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
A	1.41	1.346	1.401	1.422	1.41	1.291	1.433	1.376	1.518
B	0.94	0.904	1.044	0.979	1.070	1.019	1.070	1.048	1.052

Result

day	decision
A	The instrument looks well calibrated. Let's start immediately with the analysis of today's blood samples.
B	The instrument is not well calibrated. Before starting with the analysis, we must recalibrate it.

The procedure in short

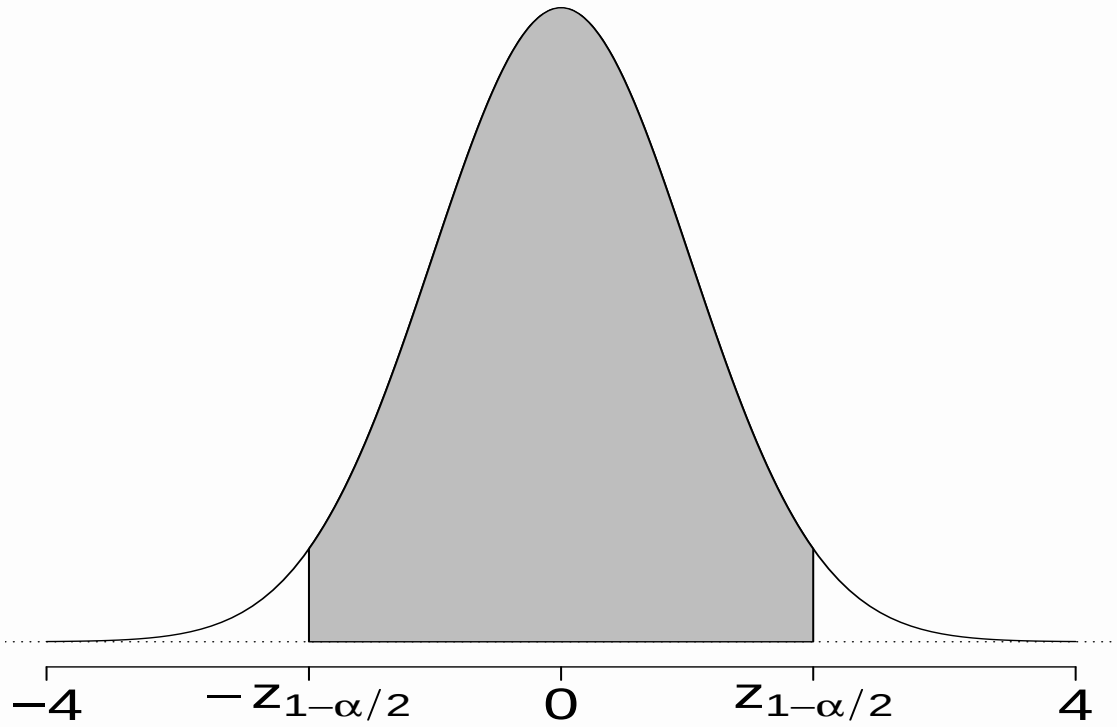
without using the p-value

- Compute
$$z = \frac{\overline{y} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}.$$
- If
$$-z_{1-\alpha/2} \leq z \leq z_{1-\alpha/2},$$
accept H_0 otherwise H_1.
- The "threshold" $z_{1-\alpha/2}$ is such that
$$P(-z_{1-\alpha/2} \leq z \leq z_{1-\alpha/2}) = 1 - \alpha$$
when $\delta = 0$, that is, when H_0 is true.

The procedure in short

without using the p-value

The distribution, a $N(0, 1)$, "tells us" which values of z we expect to observe if the instrument is well calibrated.



The "grey" area is $1 - \alpha$. Both "white" areas, on the left and on the right, are $\alpha/2$.

The procedure in short

with the p-value – same results!

- Compute

$$z = \frac{\overline{y} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}$$

and the p-value as

$$p = \text{Prob}(N(0,1) < -|z|) + \text{Prob}(N(0,1) > |z|).$$

- If $p > \alpha$ accept H_0 otherwise H_1.

Two possible errors

For our instrument

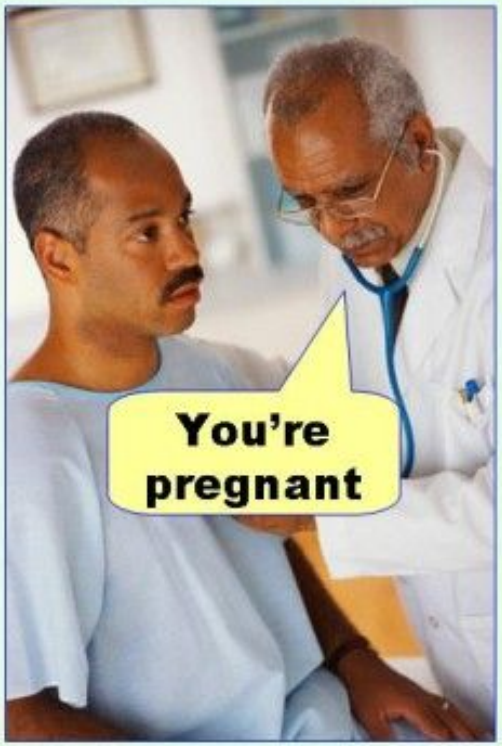
		we decide not to recalibrate	we decide to recalibrate
the instrument is	calibrated	(ˆ‿ˆ)	(ˆ⌢ˆ)
	not calibrated	(ˆ⌢ˆ)	(ˆ‿ˆ)

In general

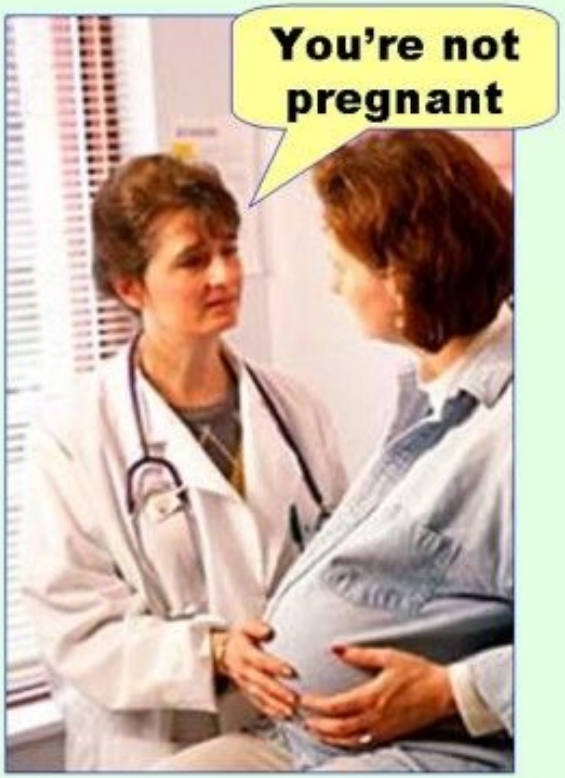
		the test goes for H_0	the test goes for H_1
the real situation is	H_0	OK	type I error
	H_1	type II error	OK

Type I and Type II errors

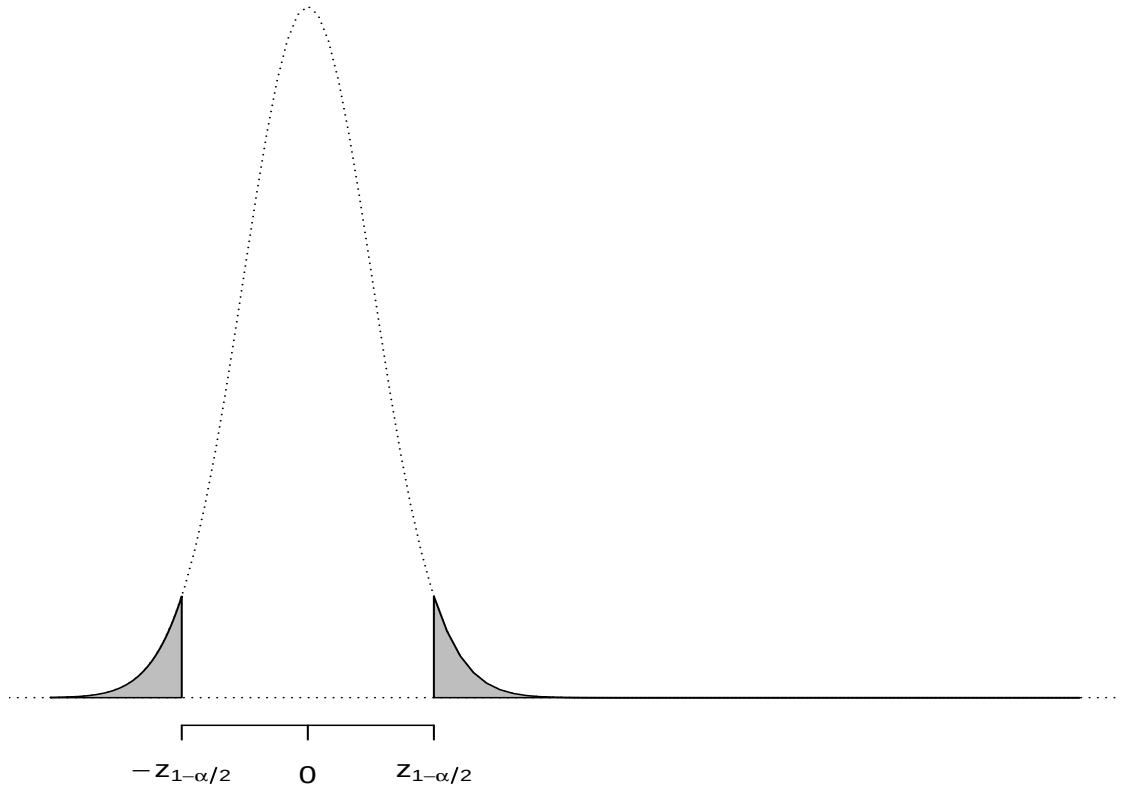
Type I error
(false positive)



Type II error
(false negative)

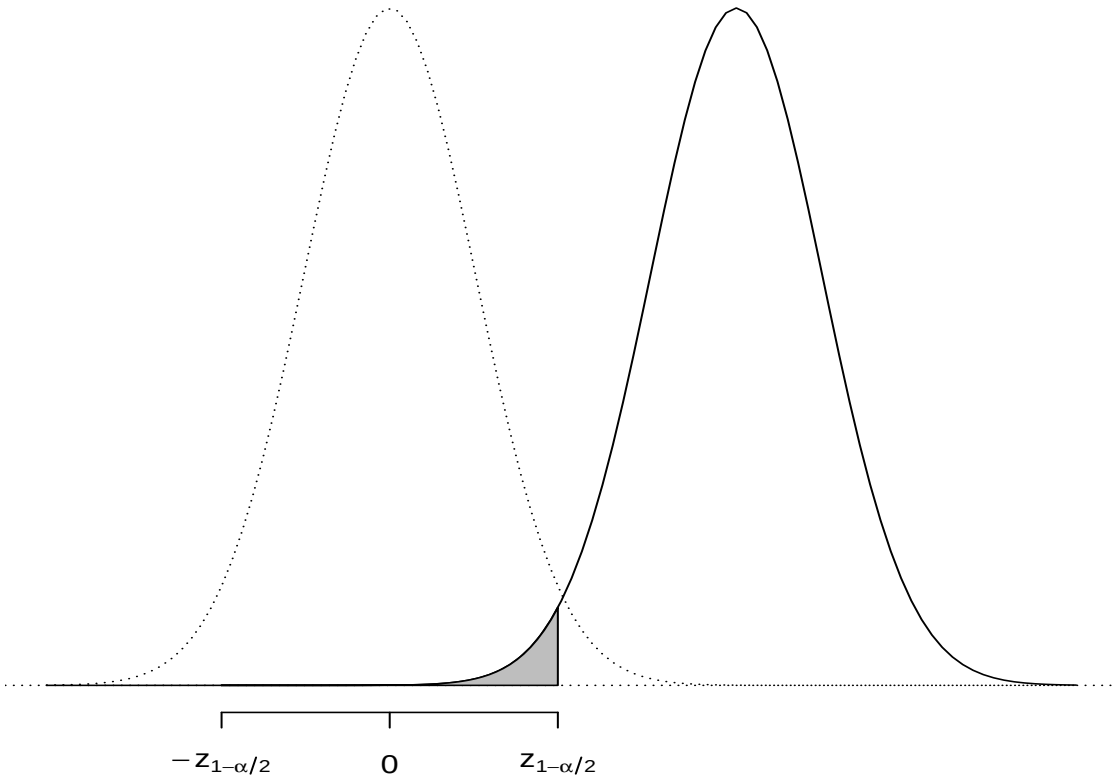


We already talked about type I errors



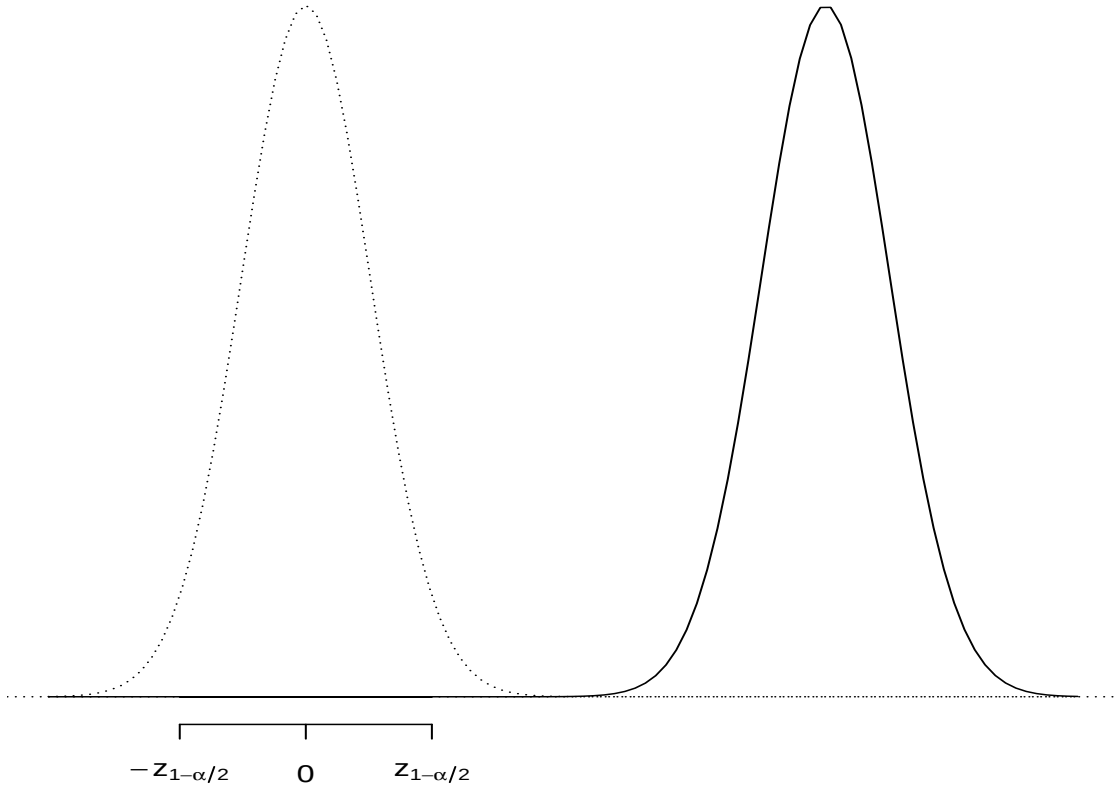
The solid curve is the standard normal distribution. The grey area represents the probability of the type I error. This error is controlled by α.

Type II error ($n = 8$, $\alpha = 0.05$ $\delta = 0{,}1$)



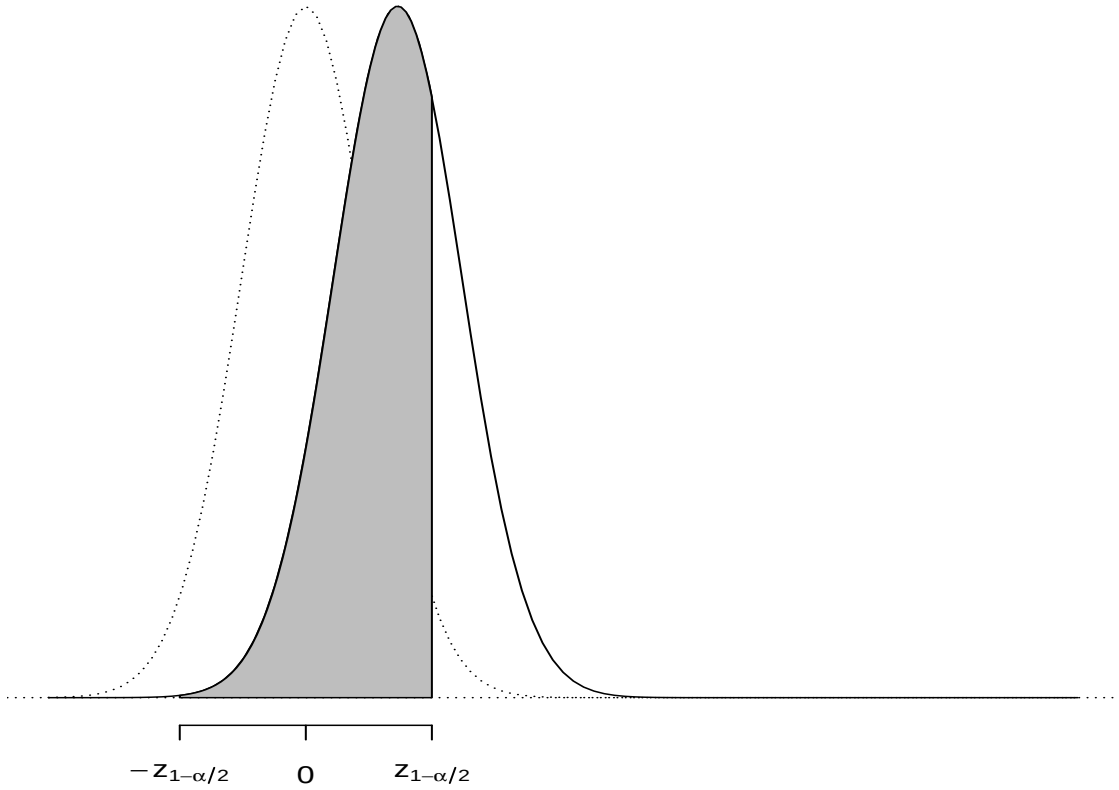
The dotted curve represents the distribution of z when H_0 is true. The solid curve represents the distribution of z when $\delta = 0{,}1$. The shaded area is the probability of committing a type II error.

Type II error ($n = 8$, $\alpha = 0.05$ $\delta = 0{,}2$)



The dotted curve represents the distribution of z when H_0 is true. The solid curve represents the distribution of z when $\delta = 0{,}2$. The probability of committing a type II error is very very small.

Type II error ($n = 8$, $\alpha = 0.05$ $\delta = 0{,}01$)



The dotted curve represents the distribution of z when H_0 is true. The solid curve represents the distribution of z when $\delta = 0{,}01$.

How to choose n

Our aim

We want to fix n such that the probability of committing a type II error is acceptable.

Note that the probability of committing a type I error is controlled by α.

Probability of the type II error

n	$\delta = 0.1$	$\delta = 0.2$
1	0.702	0.185
3	0.304	0.001
5	0.109	4×10^{-6}
6	0.062	2×10^{-7}
7	0.034	1×10^{-8}
8	0.019	4×10^{-10}
10	0.005	7×10^{-13}

The power of a statistical test

How to choose n

What we have seen together can also be described in terms of

> *"the sample size was chosen in such a way that the power of the test is accetable (above* 98% *if* $\delta = 0.1$ *and, actually, equal to* 100% *if* $\delta \geq 0.2$*)"*

Definition

The power of a test is the probability that the test rejects H_0, that is, decides in favour of H_1. If H_1 is true

$$\text{power} = 1 - (\text{prob. of type II error}).$$

More on measuring the INR

An additional problem

- Suppose our instrument is well calibrated.
- In this case, a single measurement produced by the instrument follows a normal distribution with mean equal to the true INR value and variance $0{,}07^2$.
- The error of a single measurement may be too large for some specific purposes.
- How can we reduce it?

We may think of ...

$$\left(\begin{array}{c}\text{error of the}\\ \text{average}\\ \text{taken for } n\\ \text{measurements}\end{array}\right) < \left(\begin{array}{c}\text{error of a}\\ \text{single measurement}\end{array}\right)$$

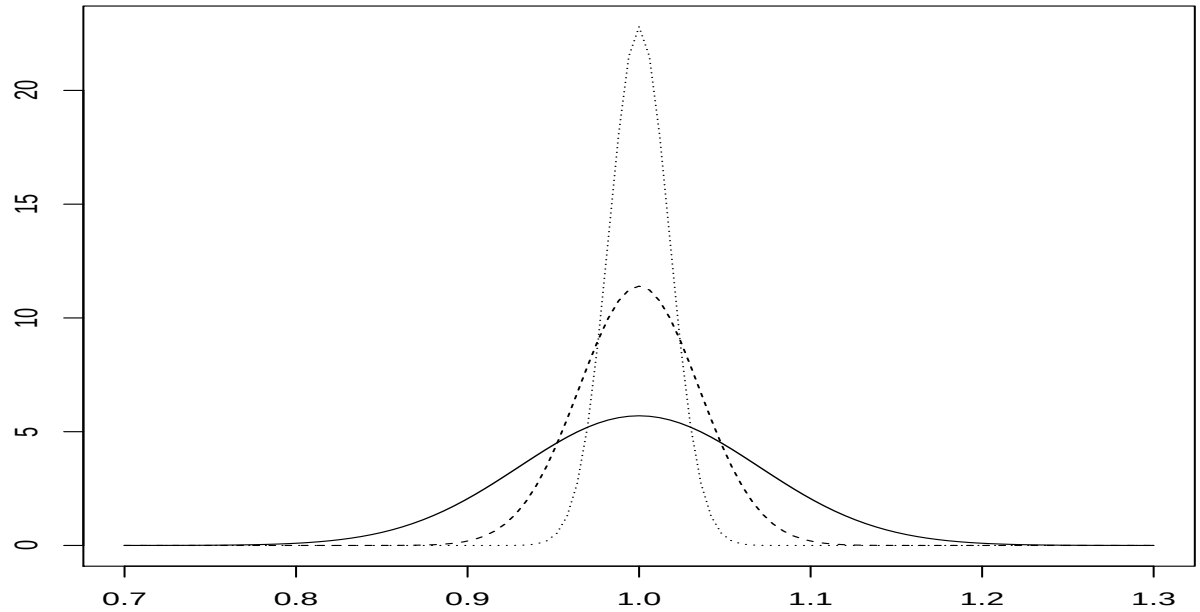
Indeed, ...

If $y_1, \ldots, y_n$ are independent draws from a $N(\mu, \sigma^2)$, then

$$\overline{y} = \frac{y_1 + \cdots + y_n}{n}$$

will follow a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

The distribution of the sample mean of n independent measurements is a normal with mean INR and variance σ_0^2/n



The solid curve represents the distribution of a single measurement – $N(INR, 0.07^2)$. The dashed curve represents a mean of 4 measurements – $N(INR, 0.07^2/4)$; the dotted curve a sample mean of 16 measurements – $N(INR, 0.07^2/16)$. All three cases refer to $INR = 1$.

A possible solution

Plan of action

1. Take a blood sample from the patient.
2. Split it up into n sub-samples.
3. Measure the INR separately for each sub-sample.
4. Report, that is, "comunicate" to the patient and to his medical doctor, the average of the n measurements.

But, how do we choose n?

First of all, by remembering that we aren't Dracula. Of course, we may obtain thousands of different measurements (or even more) by bleeding the patient ... though you can imagine its side effect.

Confidence interval for the mean of a normal distribution with known variance

dati	n independent observations $$y_1, \ldots, y_n$$ from a normal distribution with unknown mean μ and known variance σ_0^2
interval	The true mean μ belongs with probability $1 - \alpha$ to the interval $$\overline{y} \pm z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}$$ where $$\overline{y} = \frac{y_1 + \cdots + y_n}{n}.$$

A possible choice for n

The idea

Let's fix n such that the width of the confidence interval for the true *INR* value is smaller than a preassigned value d.

Solution

$$n \geq \left(\frac{2 z_{1-\alpha/2} \sigma_0}{d} \right)^2 .$$

Usually, and also because of cost considerations, we take the smallest integer value that satisfies the inequality.

Property

$$\text{Prob}(|\overline{y} - INR| < d/2) \geq 1 - \alpha.$$

Numerical example

1. Set $\alpha = 0{,}1$, and hence $z_{1-\alpha/2} = 1{,}645$.
2. Set $d = 0{,}1$.
3. Compute

$$\left(\frac{2z_{1-\alpha/2}\sigma_0}{d}\right)^2 = \left(\frac{2 \cdot 1{,}645 \cdot 0{,}07}{0{,}1}\right)^2 = 5{,}3.$$

4. Choose $n = 6$.

This choice guarantees that, in more than nine times out of ten, the distance between the estimate, $\overline{y}$, reported to the patient and the true INR value is less than 0,05.

A second numerical example (Dracula would have liked)

1. Set $\alpha = 0{,}01$, and hence $z_{1-\alpha/2} = 2{,}576$.
2. Set $d = 0{,}002$.
3. Compute

$$\left(\frac{2z_{1-\alpha/2}\sigma_0}{d}\right)^2 = \left(\frac{2 \cdot 2{,}576 \cdot 0{,}07}{0{,}002}\right)^2 = 3215{,}3.$$

4. Choose $n = 3216$.

This choice guarantees that, in more than 99 times out of 100, the distance between the estimate, $\overline{y}$, reported to the patientand the true INR value is less than 0,001.
. . . and Dracula for sure would have been happy to help out!

And Mendel? How might he have fixed n?

- In case of a binomial distribution, the confidence interval becomes

$$\hat{\vartheta} \pm z_{1-\alpha/2} \frac{\sqrt{\hat{\vartheta}(1-\hat{\vartheta})}}{\sqrt{n}}.$$

- In this case, the width depends on quantities which we don't know before carrying out the experiment.
- Possibility:
  1. Fix the sample size assuming $\hat{\vartheta} = \vartheta_0$, where ϑ_0 is an a priori chosen value. Mendel may have chosen 0,75.
  2. Compute n while considering the least favourable situation given that it is possible to prove that

$$\sqrt{\hat{\vartheta}(1-\hat{\vartheta})} \leq \frac{1}{2} \text{ whatever the value of } \hat{\vartheta}.$$

Further tests based on the normal distribution

The cornerstone result

$$\frac{\left(\begin{array}{c}\text{estimate}\\ \text{computed}\\ \text{from the}\\ \text{data}\end{array}\right) - \left(\begin{array}{c}\text{true value of}\\ \text{parameter of}\\ \text{interest}\end{array}\right)}{s.e.}$$

distributes, exactly or approximateley, like the standard normal distribution.

Standard error

s.e., acronym for standard error, represents the square root of the variance of the estimation error.
We use it, in the previous expression, to "standardise" (= obtain a unit variance for) the estimation error at the numerator.

reference distribution	parameter of interest	estimate	standard error
binomial	$\vartheta =$ population proportion	$\hat{\vartheta} =$ sample proportion	$\sqrt{\frac{\vartheta \cdot (1-\vartheta)}{n}}$ $\sqrt{\frac{\hat{\vartheta} \cdot (1-\hat{\vartheta})}{n}}$
normal	$\mu =$ population mean	$\overline{y} =$ sample mean	$\frac{\sigma}{\sqrt{n}}$

Our hypotheses

$$H_0 : \left(\begin{array}{c} \text{true value of} \\ \text{parameter of} \\ \text{interest} \end{array} \right) \begin{array}{c} \leq \\ = \\ \geq \end{array} \left(\begin{array}{c} \text{preassigned} \\ \text{value} \end{array} \right)$$

$$H_1 : \left(\begin{array}{c} \text{true value of} \\ \text{parameter of} \\ \text{interest} \end{array} \right) \begin{array}{c} > \\ \neq \\ < \end{array} \left(\begin{array}{c} \text{preassigned} \\ \text{value} \end{array} \right)$$

The procedure in short

Compare $$\frac{\left(\begin{array}{c} \text{estimate computed} \\ \text{from the data} \end{array} \right) - \left(\begin{array}{c} \text{preassigned} \\ \text{value} \end{array} \right)}{s.e.}$$

with the values "we expect to observe from the standard normal".

Confidence intervals: bilateral case

$$\left(\begin{array}{c}\text{estimate computed}\\ \text{from the data}\end{array}\right) \pm z_{1-\alpha/2} \times s.e.$$

Comment. We need a "computable" version of the standard error.

Generalisation

We may use procedures similar to the above ones also in other situations.
This possibility descends directly from the central limit theorem (which generalises the theorem seen for the binomial distribution).

Considered cases

- Inference on the mean of a population for "large" samples.
- Inference on the difference between two proportions.
- Inference on the difference between the means of two populations (as, for instance, healthy/diseased or treated/not treated subjects, . . .),

If $y_1, \ldots, y_n$ are independent draws from a random variable with arbitrary mean μ and finite variance σ^2, then, for a sufficiently large n, having written

$$\overline{y} = \frac{y_1 + \cdots + y_n}{n}$$

and

$$s^2 = \frac{(y_1 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2}{n-1},$$

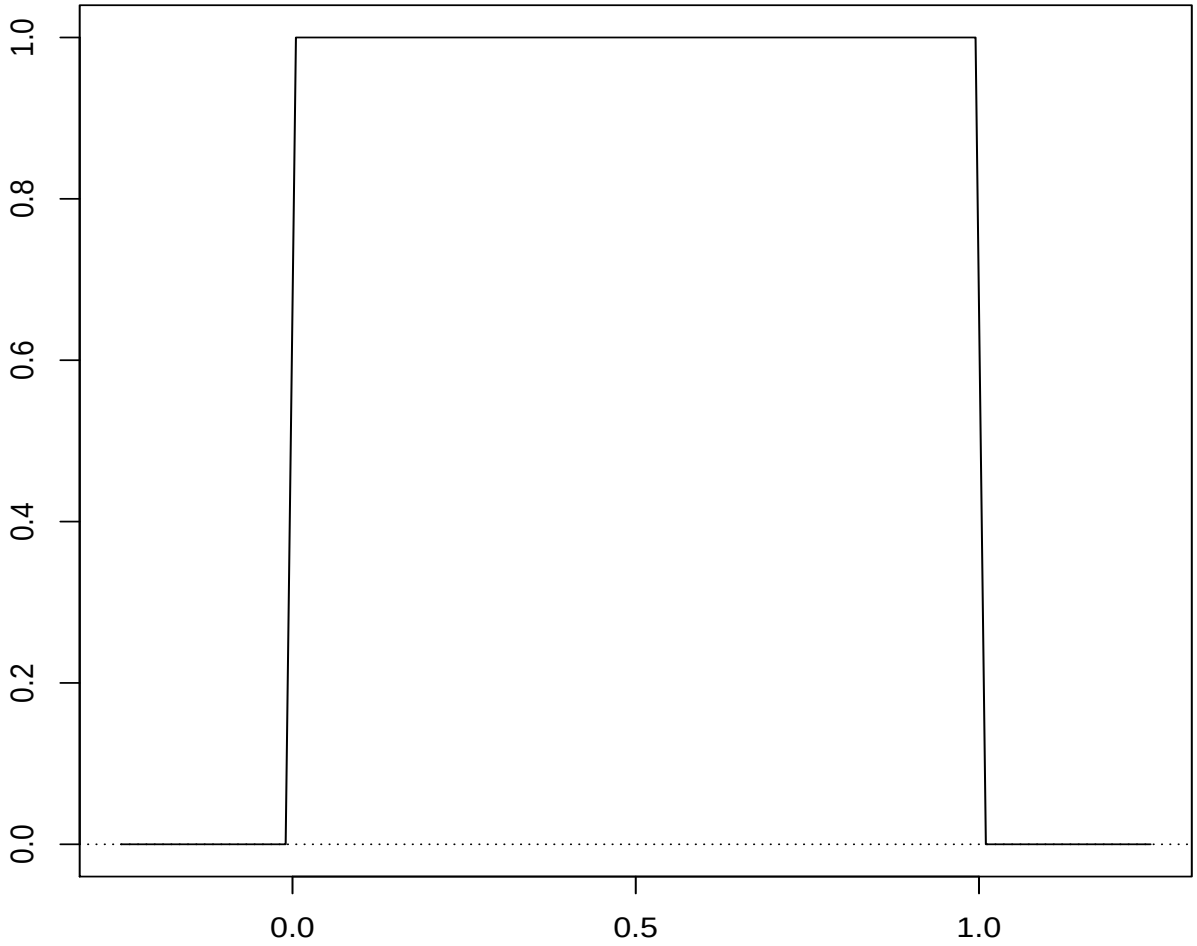
the distribution of

$$z = \frac{\overline{y} - \mu}{\frac{s}{\sqrt{n}}}$$

can be approximated by the $N(0,1)$.

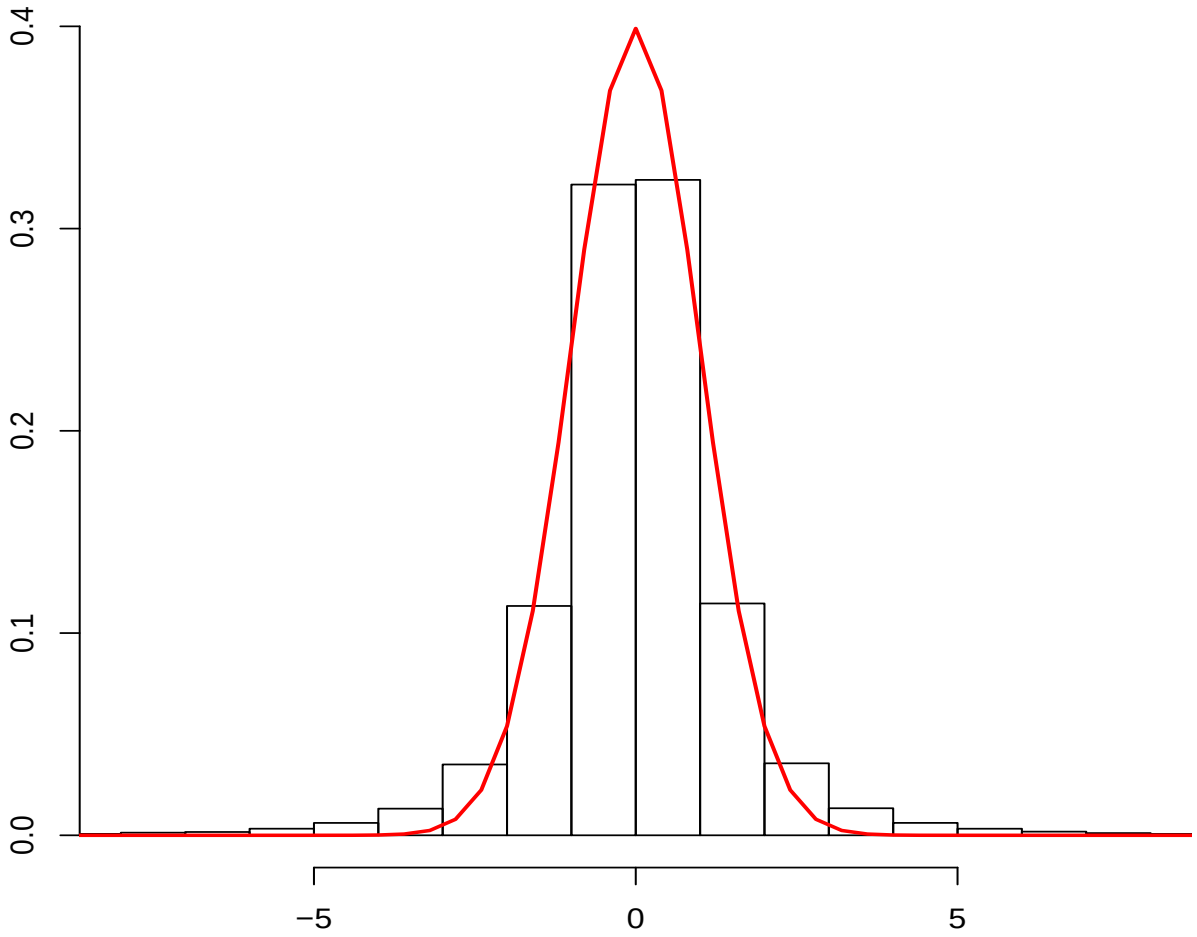
The central limit theorem

Example – A first possible distribution of the data



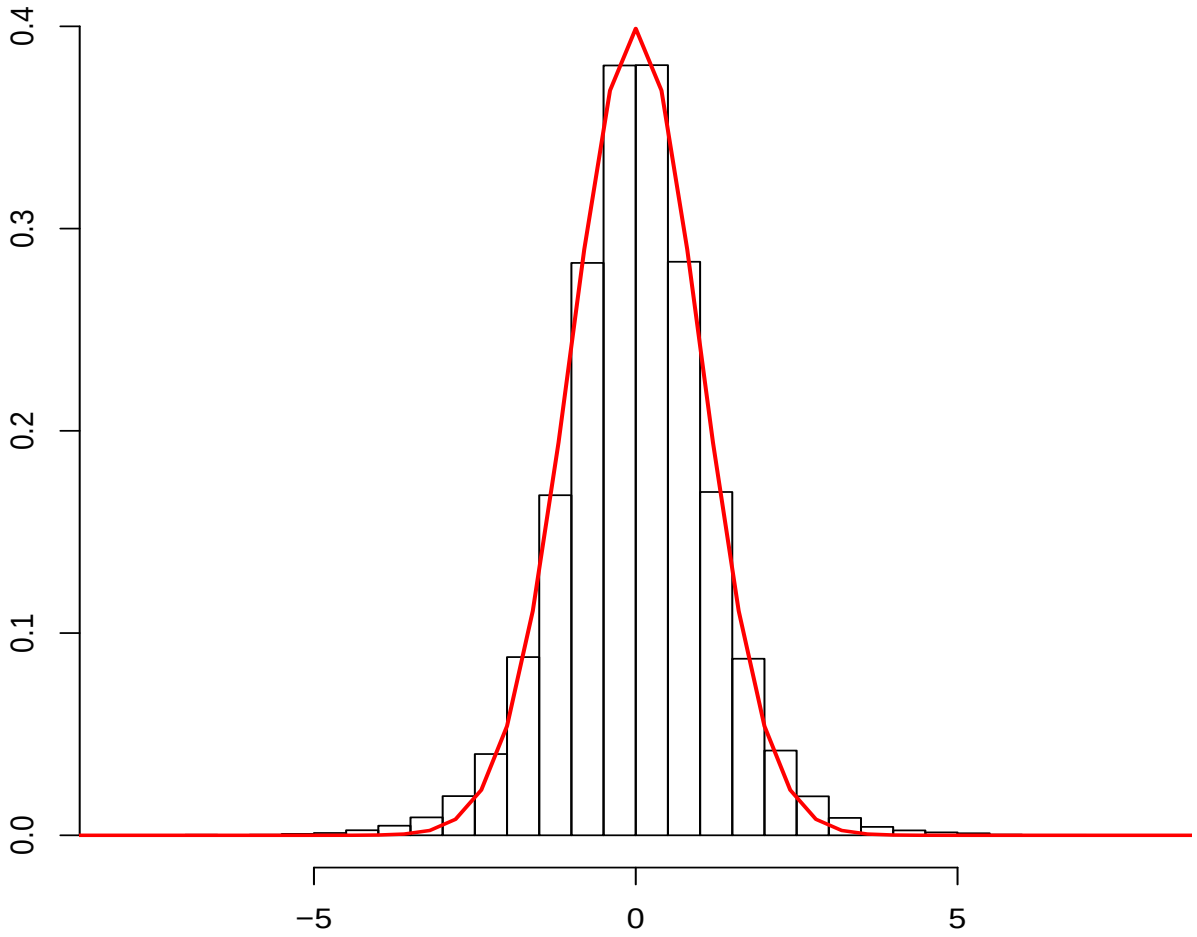
The central limit theorem

Example – Distribution of z when $n = 5$



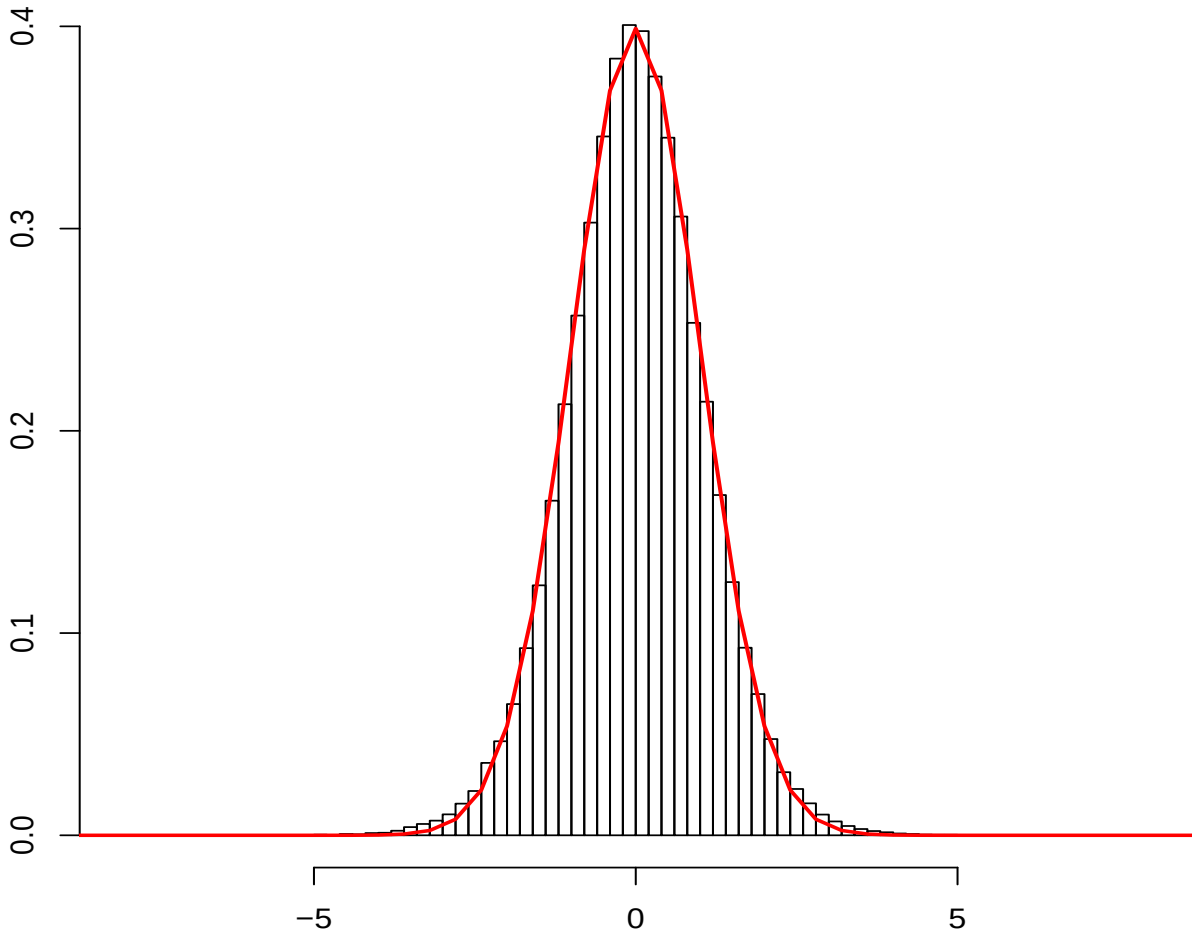
The central limit theorem

Example – Distribution of z $n = 10$



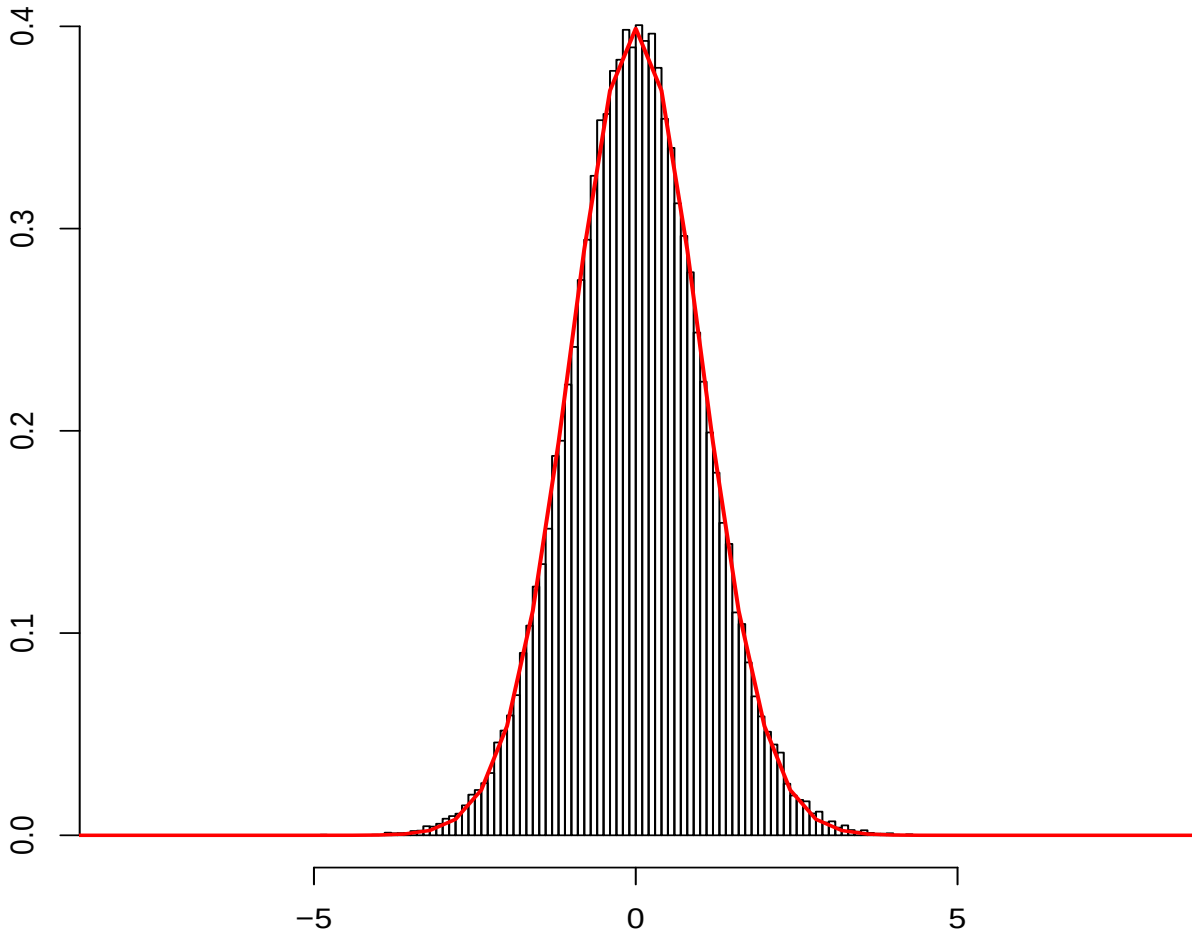
The central limit theorem

Example – Distribution of z when $n = 20$



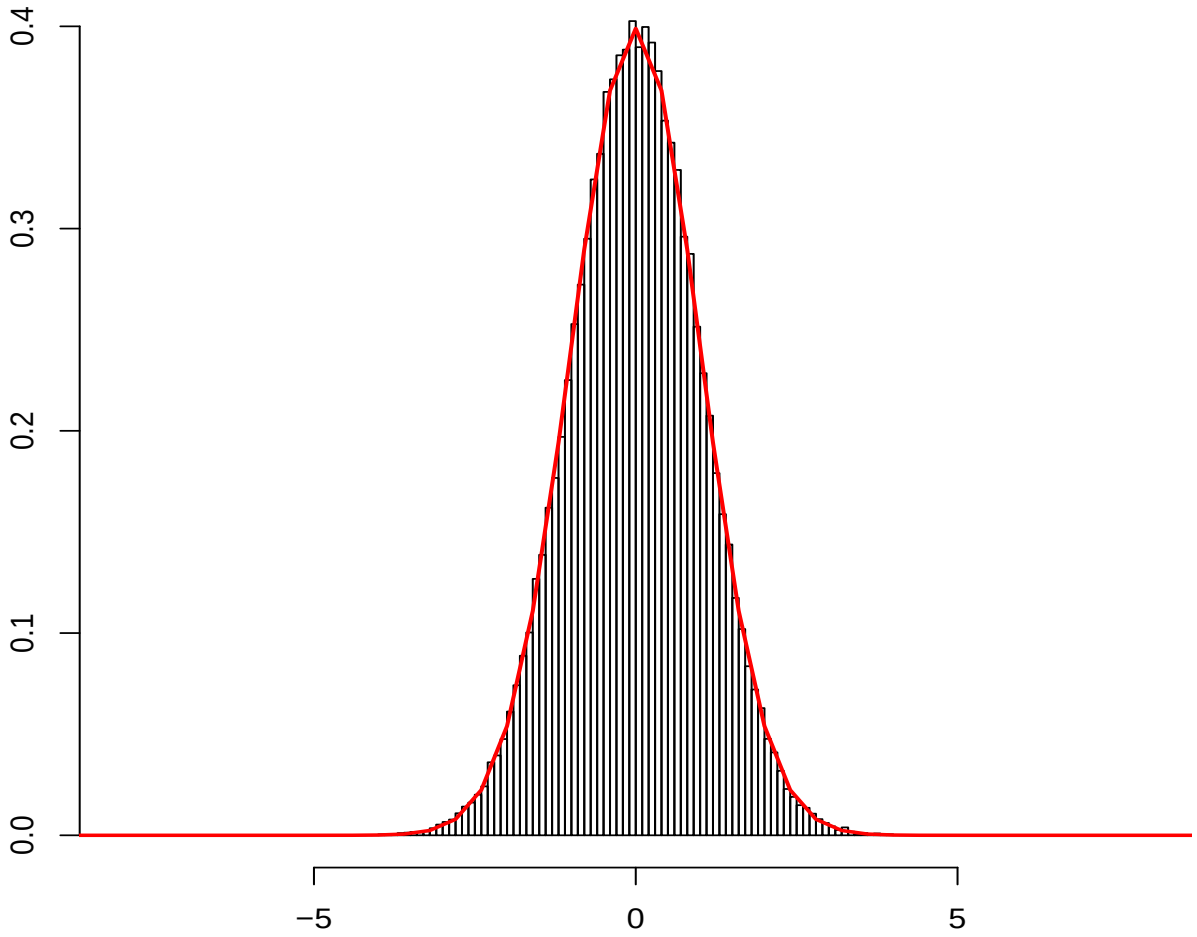
The central limit theorem

Example – Distribution of z when $n = 40$



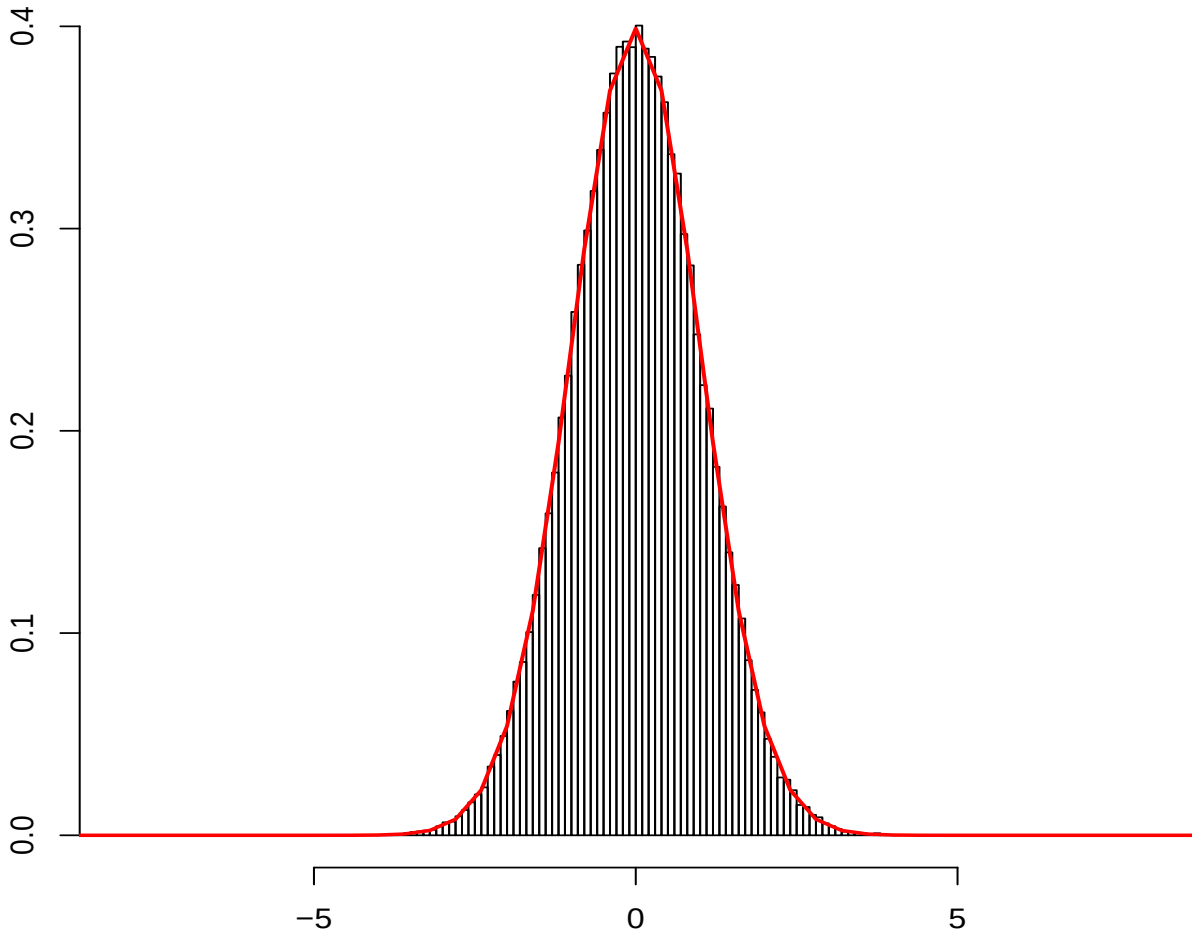
The central limit theorem

Example – Distribution of z when $n = 80$



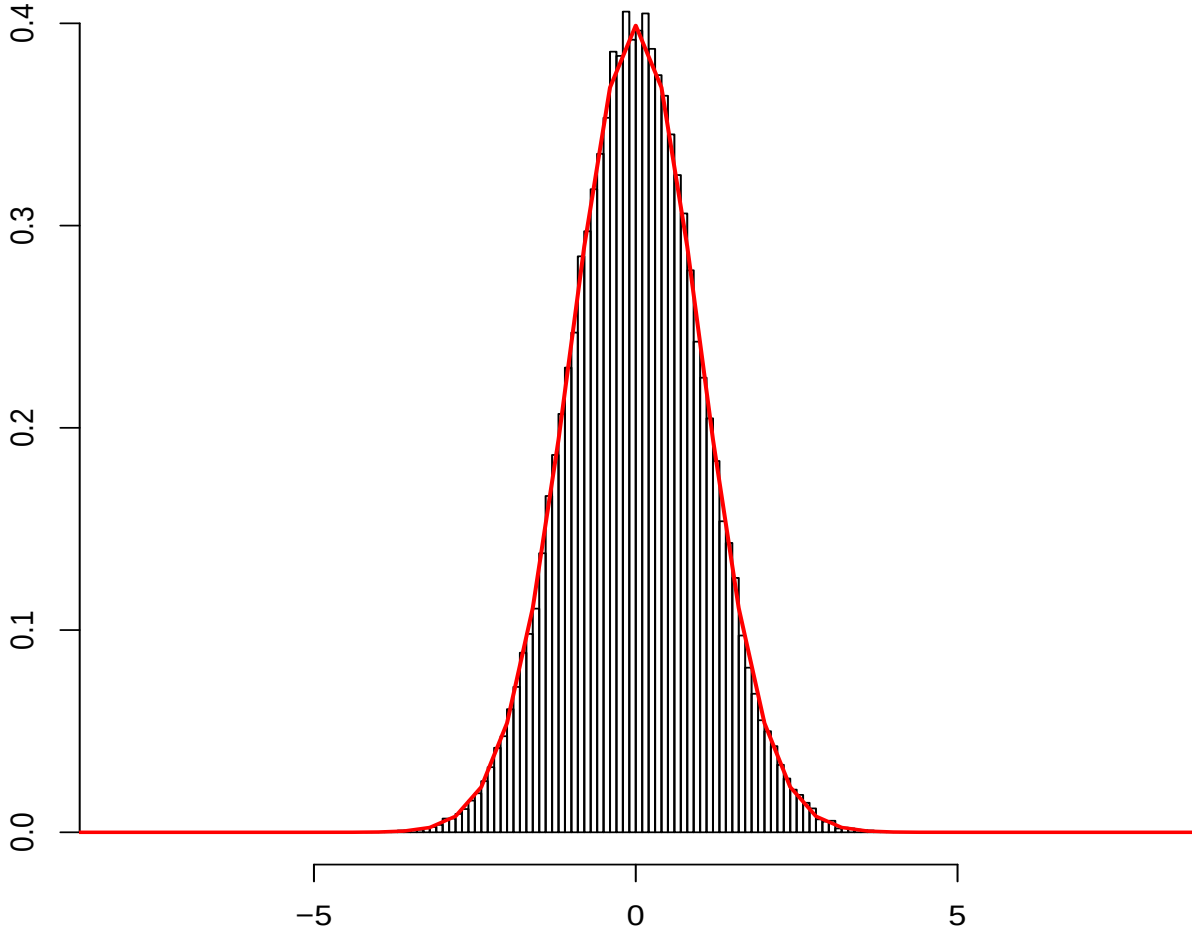
The central limit theorem

Example – Distribution of z when $n = 160$



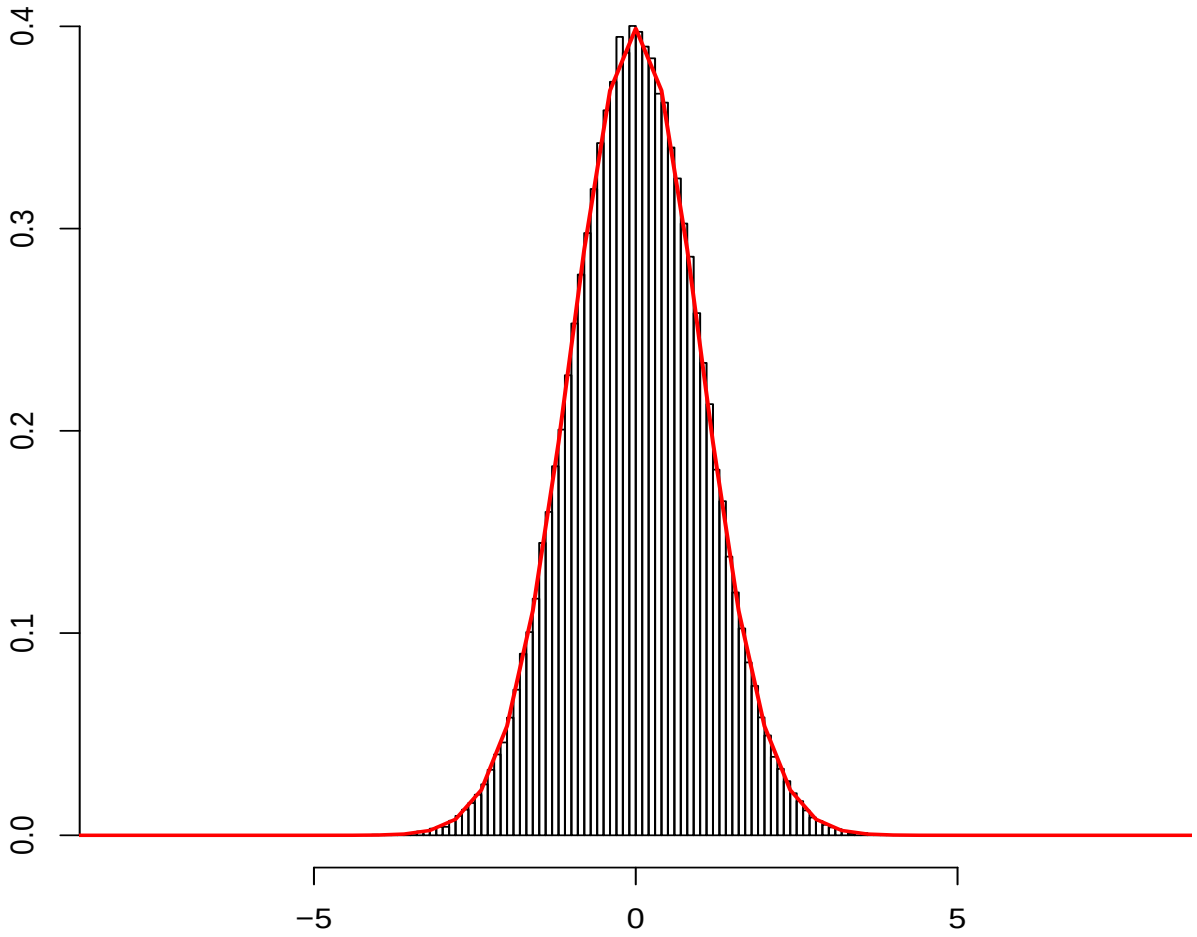
The central limit theorem

Example – Distribution of z when $n = 320$



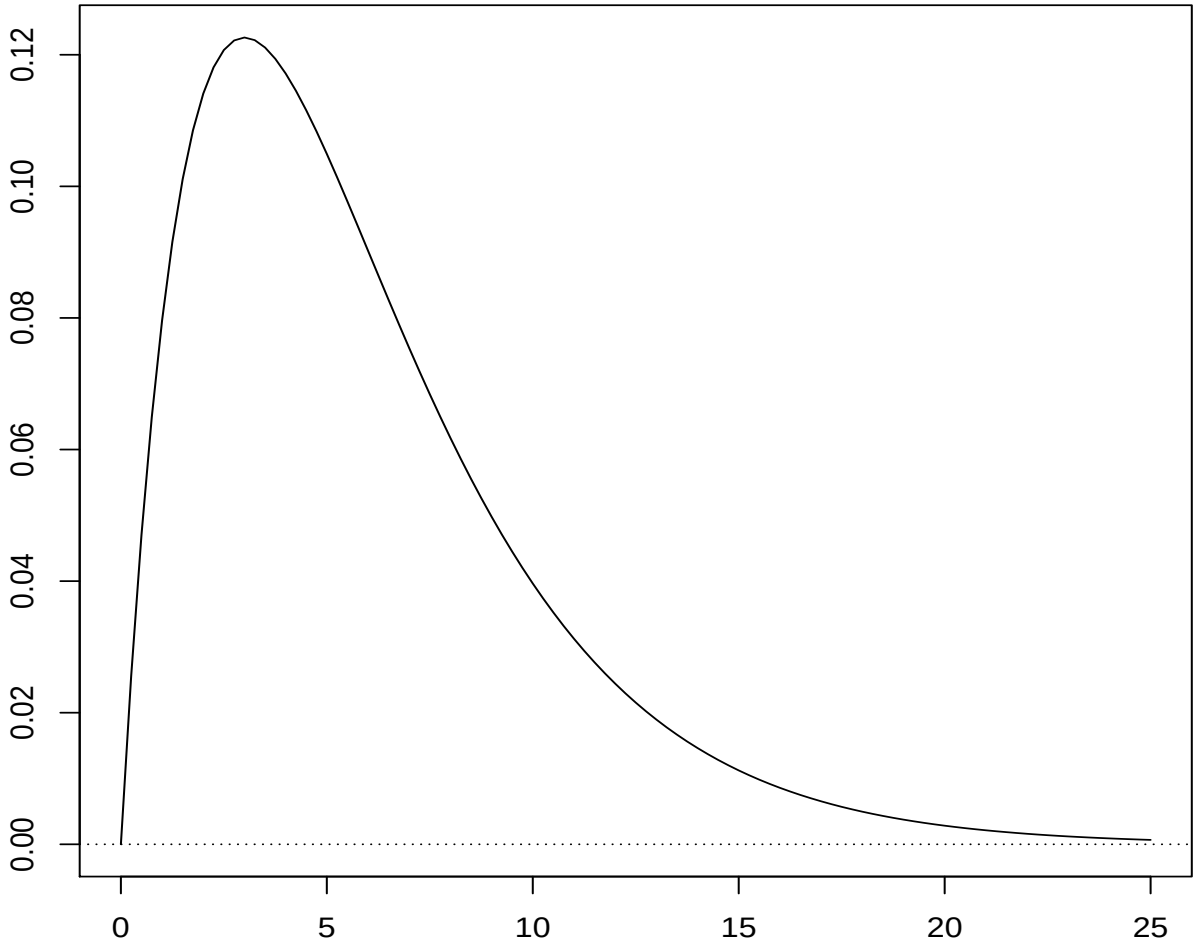
The central limit theorem

Example – Distribution of z when $n = 640$



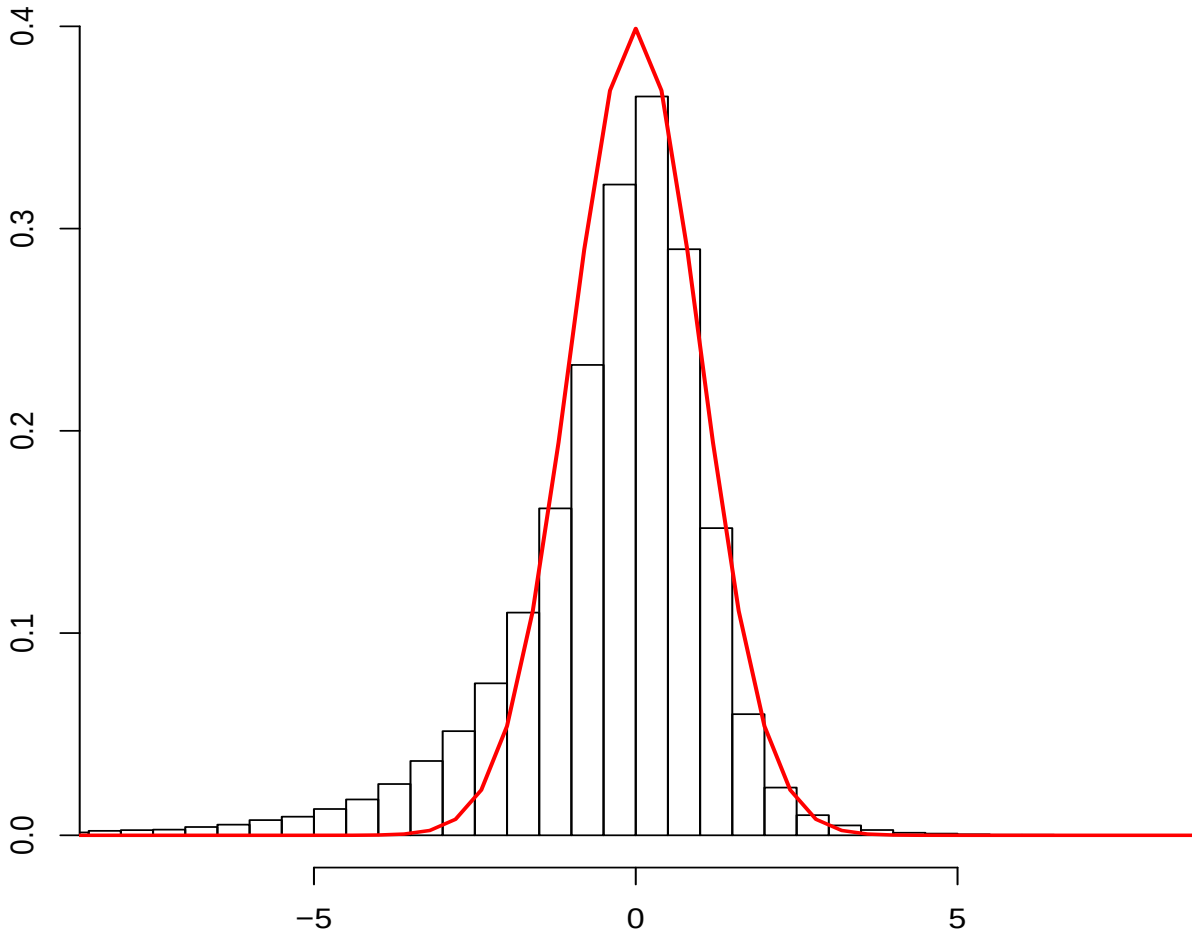
The central limit theorem

Example – A second possible distribution of the data



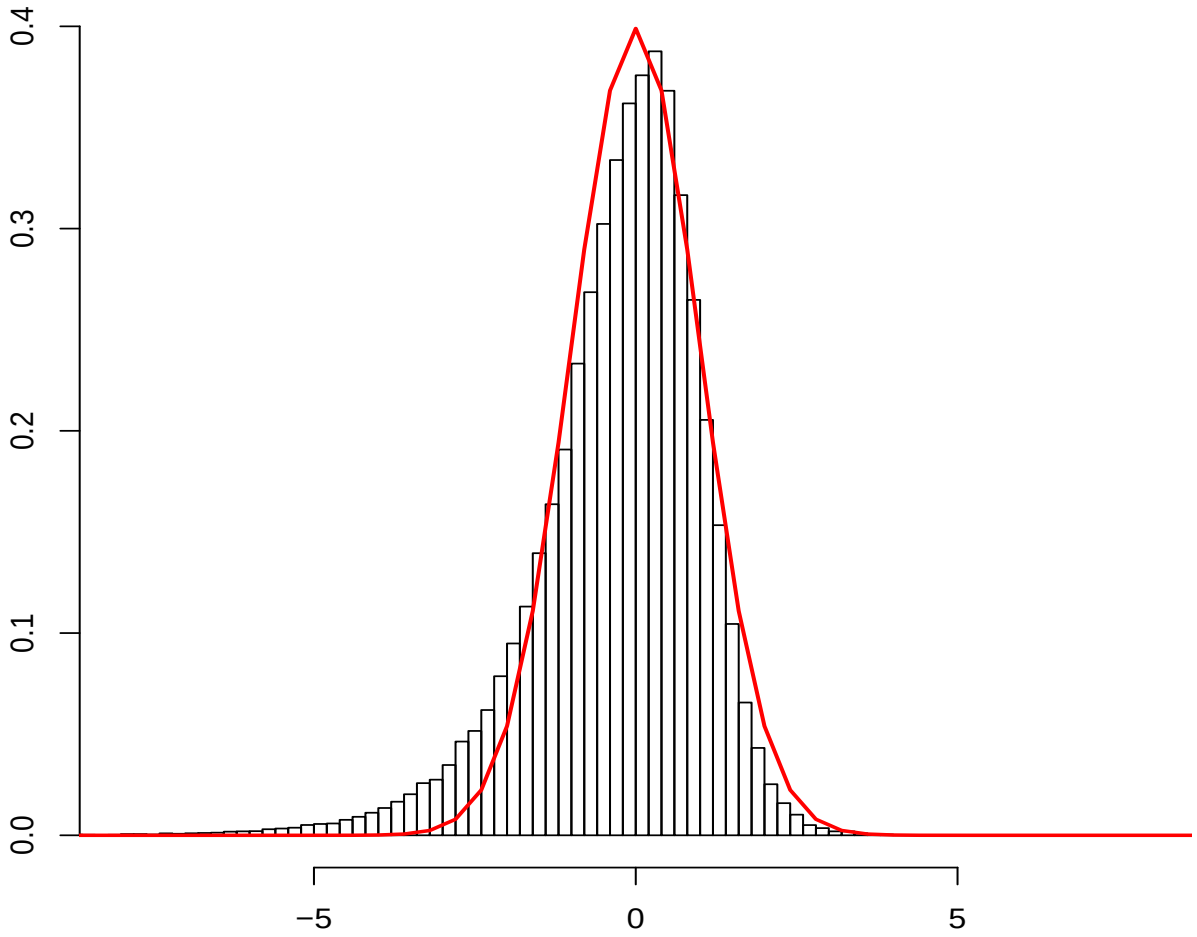
The central limit theorem

Example – Distribution of z when $n = 5$



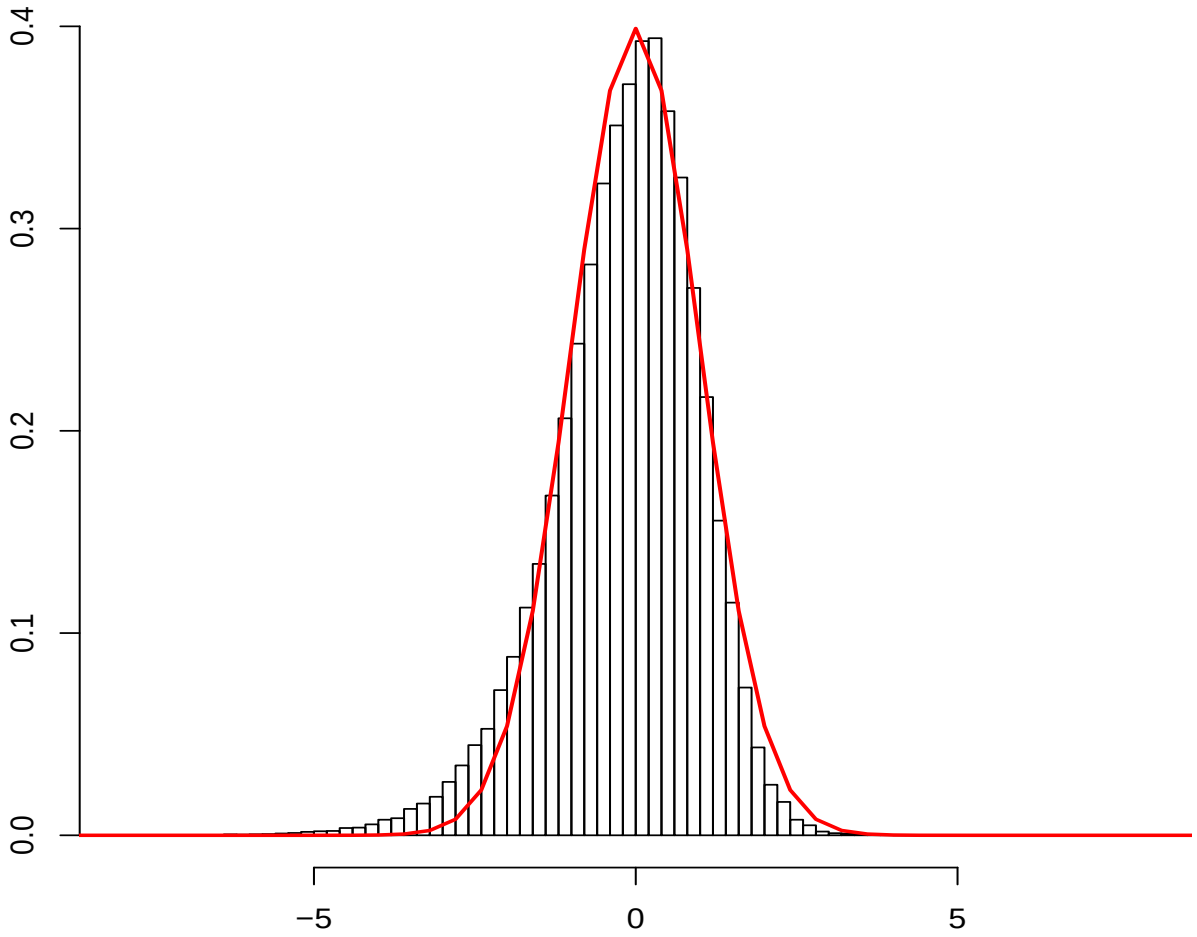
The central limit theorem

Example – Distribution of z when $n = 10$



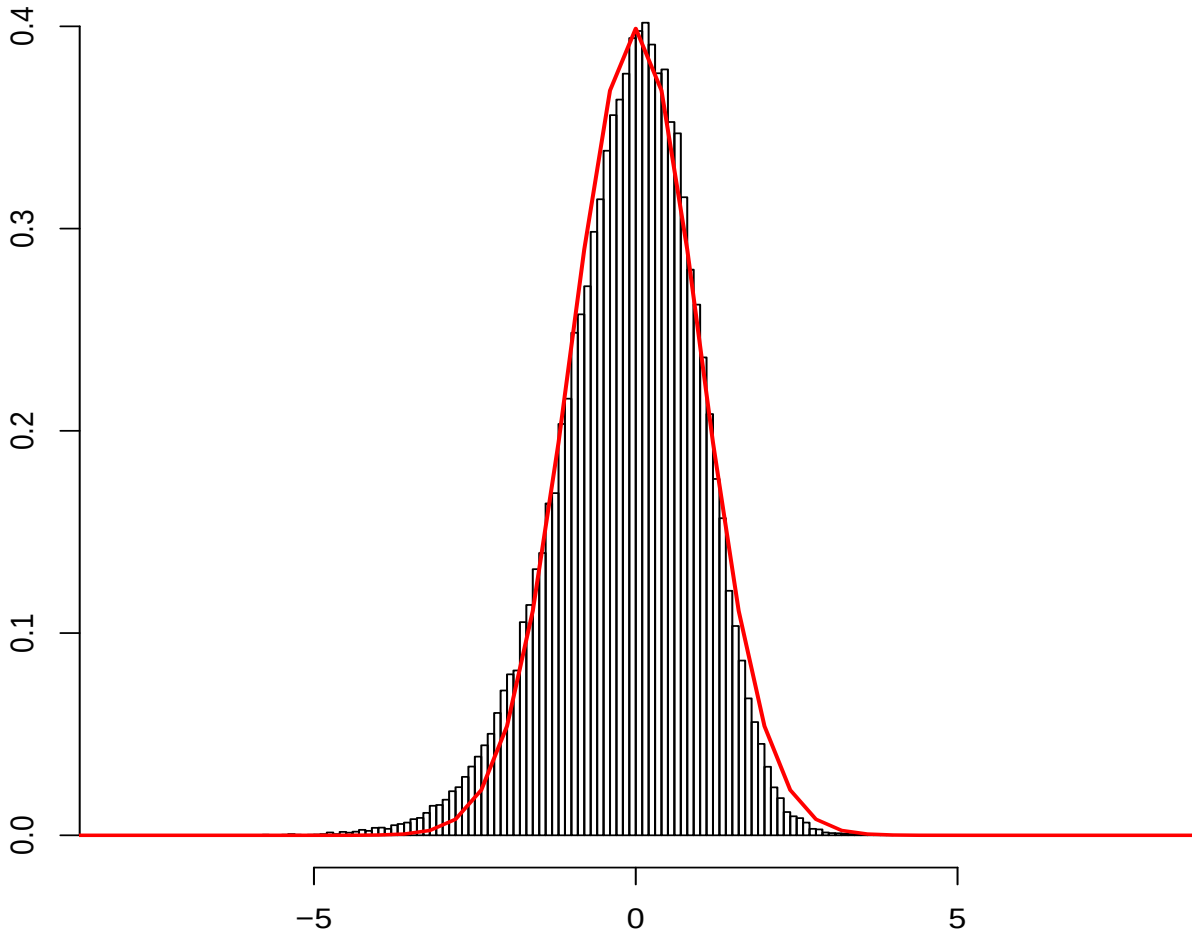
The central limit theorem

Example – Distribution of z when $n = 20$



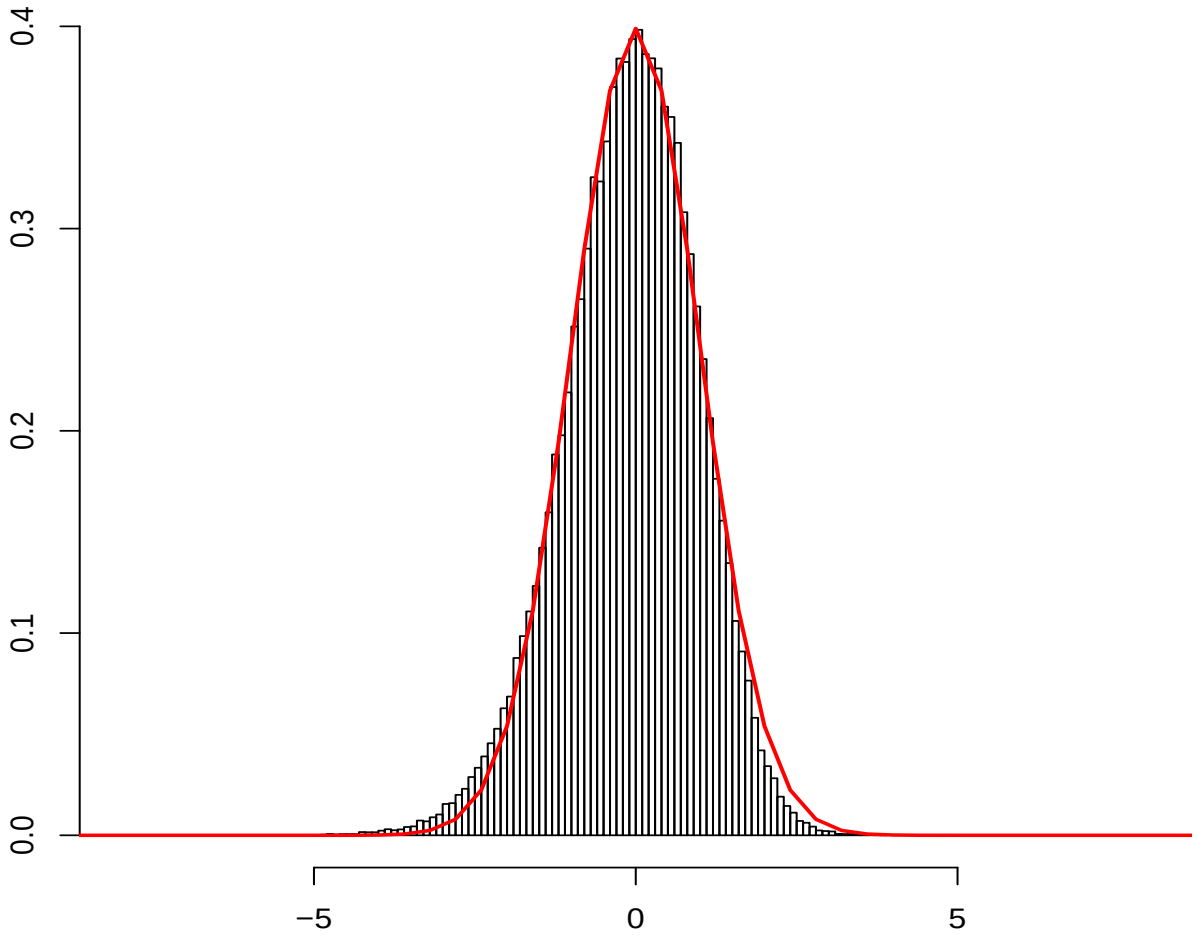
The central limit theorem

Example – Distribution of z when $n = 40$



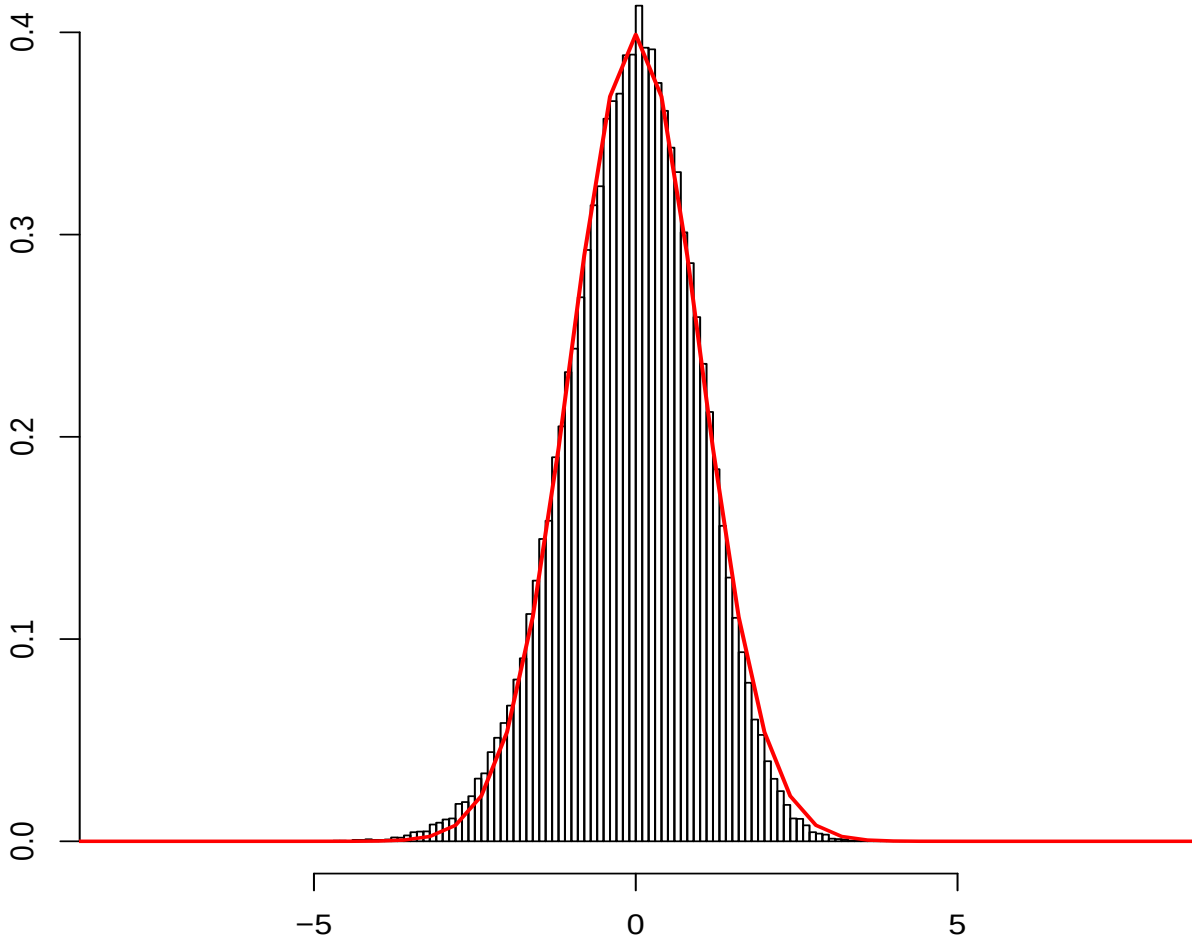
The central limit theorem

Example – Distribution of z when $n = 80$



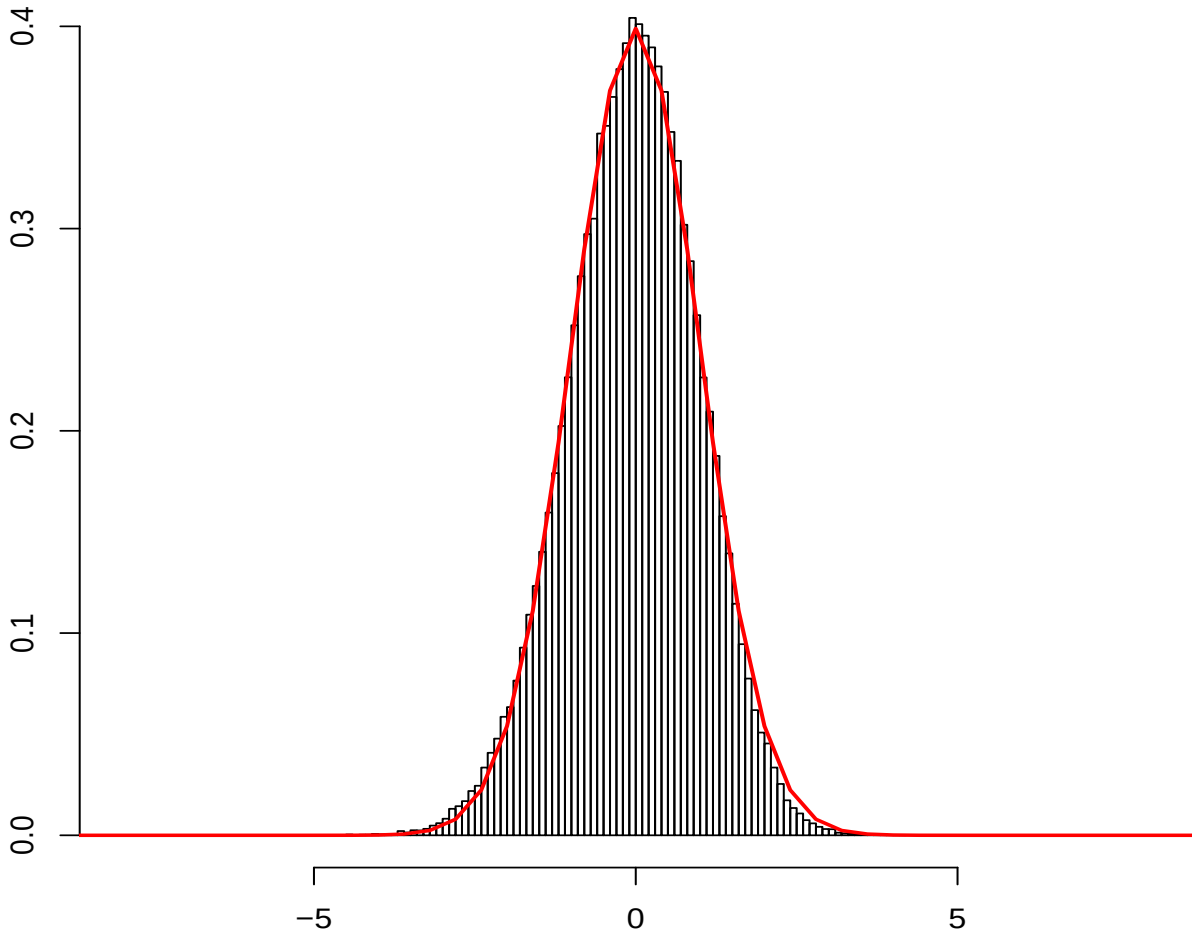
The central limit theorem

Example – Distribution of z when $n = 160$



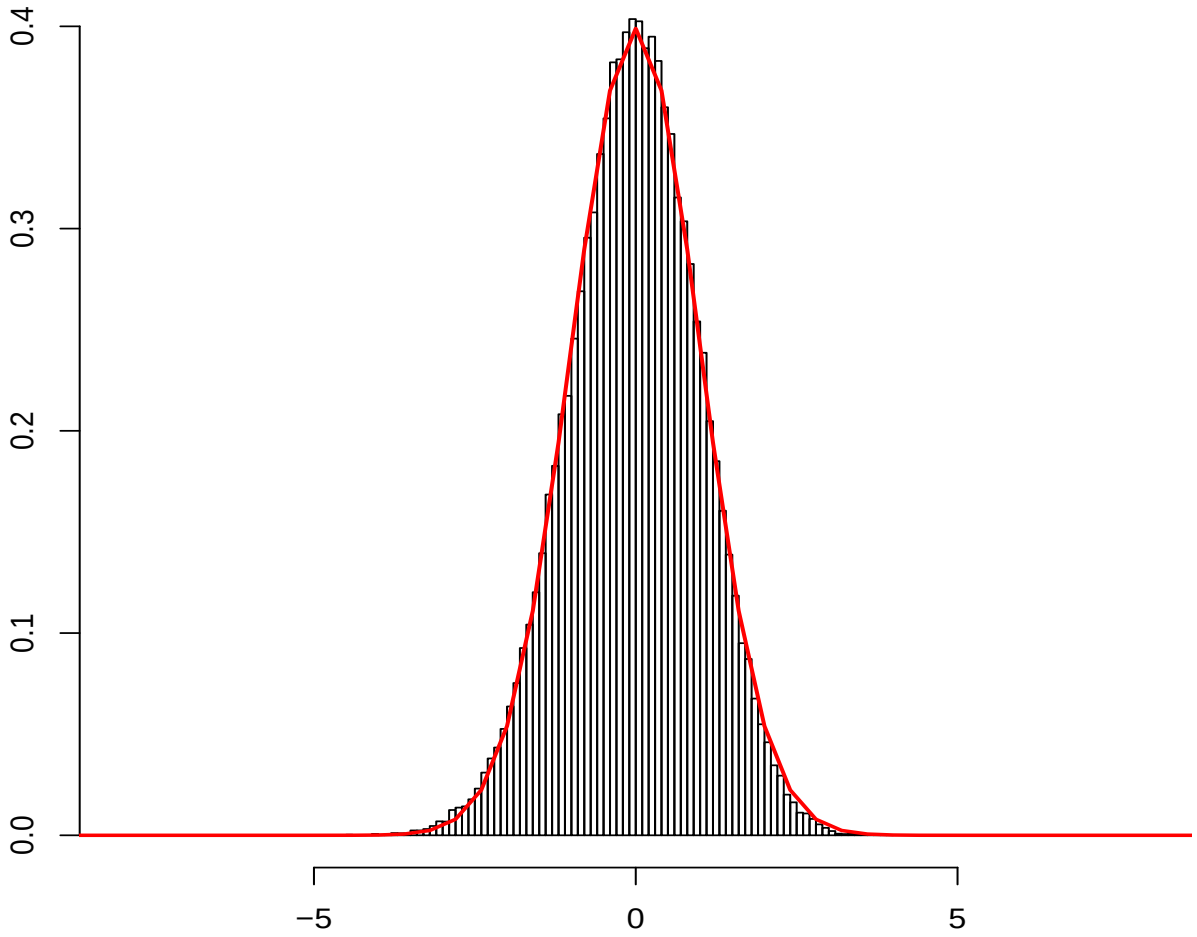
The central limit theorem

Example – Distribution of z when $n = 320$



The central limit theorem

Example – Distribution of z when $n = 640$



The central limit theorem

Comments

When can we use it?

The approximation is considered to be acceptable if $n \geq 50$.
If the distribution is symmetric, it may already work for $n \geq 30$.

Generalisation

A similar result also holds for analysing the difference between two means.

Bone mineral density (BMD)

A first example

Description of the study

Low BMD values may lead to an increased risk of fractures, particularly of the hip.
A study was set up to evaluate the effectiveness of hormone replacement therapy. In all, 94 women between 45 and 64 years of age took EEC (an estrogen) for 36 months. At the end of the study, BMD was measured in all 94 women.
We want to calculate a confidence interval for the mean value of BMD of the population consisting of all women aged between 45 and 64 who may "apply" for the same therapy.
Given that the sample size is large, we can use the previous result without having to bother about the true distribution of the data.

Bone mineral density (BMD)

A first example

Some results

$$n = 94 \quad \overline{y} = 0{,}878\, g/cm^2 \quad s = 0{,}126\, g/cm^2$$

Confidence interval ($\alpha = 0.1$)

$$\overline{y} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}} = 0{,}878 \pm 1{,}645\frac{0{,}126}{\sqrt{94}} = [0{,}857; 0{,}899].$$

That is, ...

we can state, with a 90% confidence level, that the true mean of BMD in the target population lies between $0{,}857\, g/cm^2$ and $0{,}899\, g/cm^2$.

Genes and alcohol

Description of the experiment

- Some researchers conjectured that alcoholism may be associated with a particular gene.
- To assess this hypothesis, they collected the preferences for a "holy" and a "alcoholic" version of a certain beverage.
- The preferences were gathered for 27 "normal" mice and 25 "knockout" ones.
- Below the results

```
> mices
              alcohol not alcohol
not knockout       18           9
knockout           12          13
```

Pure chance or effect?

In the "sample" $\frac{2}{3}$ of the normal mice prefer the "spirited" version of the beverage. This proportion falls to less than 50% in case of "knockout" pigs.

The difference corresponds to a true effect for the considered gene or is it simply due to chance?

That is, the observed difference in behaviour is the consequence of a "feature of the world" or is it simply a "non replicable property" of the 52 mice used for the experiment?

Comparing two proportions

- y is a draw from a binomial distribution with n_y trials and "success" probability ϑ_y.
  In our case, $y = 18$ "alcoholic" mice out of $n_y = 27$ normal ones. ϑ_y is the probability that a normal mice prefers the alcoholic version of the beverage.
- x is a draw from a binomial distribution with n_x trials and "success" probability ϑ_x.
  In our case, $x = 12$ "alcoholic" mice out of $n_x = 25$ "knockout" ones. ϑ_x is the probability that a "knockout" mice prefers the alcoholic version of the beverage.

Comparing two proportions

- We want to "make inference" (test, confidence interval,. . .) on $\delta = \vartheta_y - \vartheta_x$.
  In our case, we want to test the following set of hypotheses

$$H_0 : \vartheta_y = \vartheta_x \quad \textit{vs.} \quad H_1 : \vartheta_y > \vartheta_x$$

which we may also rewrite as

$$H_0 : \delta = 0 \quad \textit{vs.} \quad H_1 : \delta > 0.$$

Comparing two proportions

- We may again find the solution using the fact that

$$z = \frac{\hat{\delta} - \delta}{s.e.}$$

distributes approximately like the $N(0,1)$ with a suitably chosen "$s.e.$".
In the previous formulation

$$\hat{\delta} = \hat{\vartheta}_y - \hat{\vartheta}_x = \frac{y}{n_y} - \frac{x}{n_x} = \frac{18}{27} - \frac{12}{25} = 0{,}19.$$

Let's skip further formulae and have a look to the corresponding R code.

```
> prop.test(mices, alternative="greater")

2-sample test for equality of proportions
with continuity correction

data:  mices
X-squared = 1.1672, df = 1, p-value = 0.14
alternative hypothesis: greater
95 percent confidence interval:
 -0.07384283  1.00000000
sample estimates:
   prop 1    prop 2
0.6666667 0.4800000
```

The difference is not significant. The bilateral confidence intervals "tells" us that, indeed, we know rather little about δ.

```
> prop.test(mices)

2-sample test for equality of proportions
with continuity correction

data:  mices
X-squared = 1.1672, df = 1, p-value = 0.28
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1163704  0.4897037
sample estimates:
   prop 1    prop 2
0.6666667 0.4800000
```

The power of the test based upon $n_y = n_x = 26$ mice is about 30% for true proportions which are close to the observed ones..

```
> power.prop.test(n=26, p1=0.66, p2=0.50,
                  alternative="one.sided")

Two-sample comparison of proportions power calculation

              n = 26
             p1 = 0.66
             p2 = 0.5
      sig.level = 0.05
          power = 0.3147615
    alternative = one.sided

 NOTE: n is number in *each* group
```

To reach a power of about 95% (probability of committing a type II error of approximately 5%) we would need 2×204 mice.

```
> power.prop.test(p1=0.66, p2=0.50, power=0.95,
                  alternative="one.sided")

     Two-sample comparison of proportions power calculation

              n = 203.2449
             p1 = 0.66
             p2 = 0.5
      sig.level = 0.05
          power = 0.95
    alternative = one.sided

 NOTE: n is number in *each* group
```

Comparing two proportions: applicability

- The two samples must be chosen randomly from the target population.

  For instance (trivial case), we must avoid to select the most alcoholic subjects among the "normal" ones.

- The preferences accorded by mice must be independent.

  For instance, a "normal" subject which drinks the alcoholic version of the beverage must not prevent a "knockout" subject to do the same.

- All of y, $n_y - y$, x, $n_x - x$ must be larger than 5.

  Otherwise, the normal approximation will not hold.
  This holds in our case given that $y = 18$, $n_y - y = 27 - 18 = 9$, $x = 12$ *e* $n_x - x = 25 - 12 = 13$.

Does calcium uptake reduce blood pressure?

The conjecture

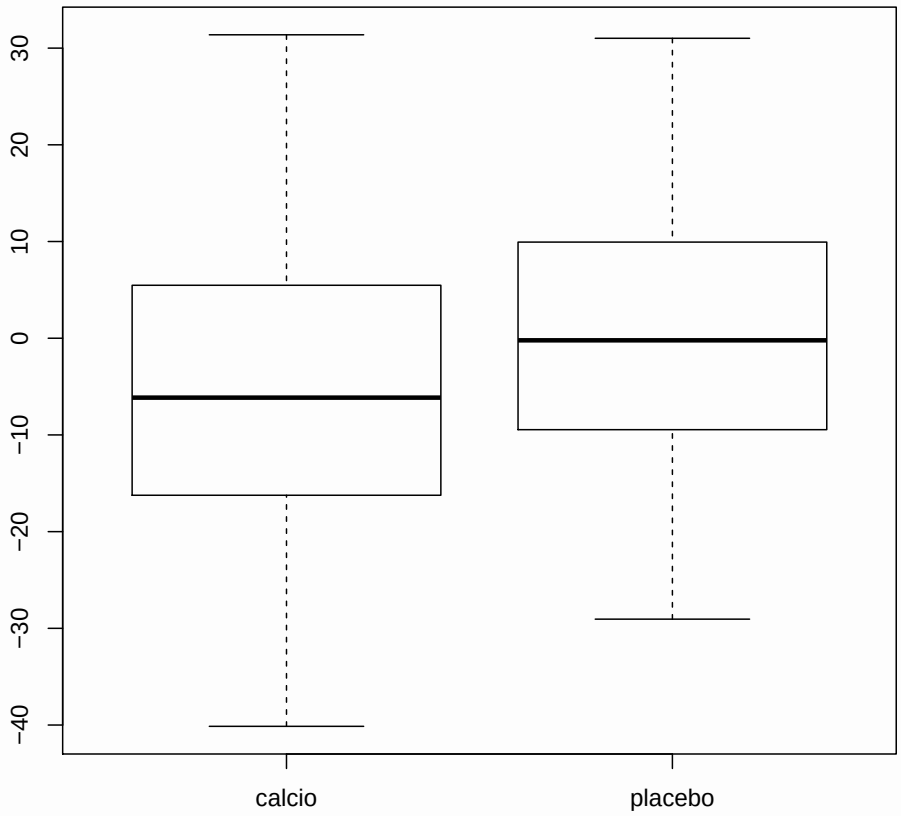
Calcium uptake reduces blood pressure, especially in the afro-american ethnic group.

The experiment

- 200 healthy individuals are randomly split into two groups of 100 individuals each.
- During a 12 weeks period, one group takes a placebo, the other calcium supplement.
- We want to analyse the difference between the final and the initial blood pressure.
- This value is available for all 100 individuals of the "treatment" group and for 94 individuals of the placebo group.

Does calcium uptake reduce blood pressure?

The data (graphical representation)



Does calcium uptake reduce blood pressure?

Questions

- Can we conclude that the reduction observed for the sample will also apply to the target population?
- If so, how large is the effect?

Comparing means of two populations: fundamental result

Proposition

- $y_1, \ldots, y_n$ are draws from a random variable with mean μ_y and variance σ_y^2.
- $x_1, \ldots, x_n$ are draws from a random variable with mean μ_x and variance σ_x^2.
- All observations are independent of each other.
- Then, if n and m are sufficiently large (say, both ≥ 30),

$$\frac{(\overline{y} - \overline{x}) - (\mu_y - \mu_x)}{\sqrt{\frac{s_y^2}{n} + \frac{s_x^2}{m}}}$$

distributes, at least approximately, as the $N(0, 1)$.

Comparing means of two populations: fundamental result

Comments

Notation

$\overline{y}$, $\overline{x}$, s_y^2 e s_x^2 represent the sample means and variances of the "y" and "x".

Why is this useful?

It allows us to make inference on

$$\delta = \mu_y - \mu_x = \left(\begin{array}{c} \text{mean difference for the} \\ \text{phenomenon of interest in} \\ \text{the two populations} \end{array} \right).$$

Other versions

If both variances are equal, that is, if we can assume that $\sigma_y^2 = \sigma_x^2$, there exists an equivalent version of the same result.

Does calcium uptake reduce blood pressure?

Some answers

```
> t.test(placebo, calcio, alternative="greater")

Welch Two Sample t-test

data:  placebo and calcio
t = 2.5424, df = 191.393, p-value = 0.0059
alternative hypothesis:
true difference in means is greater than 0
95 percent confidence interval:
 1.839415      Inf
sample estimates:
 mean of x  mean of y
-0.1671057 -5.4242036
```

A brewer statistician meets butterflies, cuckoos, wrens and redbreasts

Winged monarchs

- The pictures shows a male monarch butterfly.
- This species originates from North America to the extent that it became the national insect of various US states as, for instance, of Texas.

More about their wings

- In the framework of a study aimed at investigating the body structure of different species of insects, a group of researchers captures 14 male monarch butterflies in one of California's national parks (*Ocean Dunes State Park)*.
- They retrieve the following wing measurements (in cm^2).

```
> monarche
 [1] 33.9 33.0 30.6 36.6 36.5 34.0 36.1
     32.0 28.0 32.0 32.2 32.2 32.3 30.0
```

- Suppose we want to calculate a confidence interval for the mean width of the wings of "all" male monarch butterflies of the *Ocean Dunes State Park*.

Butterflies and dark beers

- We just saw how to solve the problem for large n.
- The solution uses the fact that

$$t = \frac{\overline{y} - \mu}{\frac{s}{\sqrt{n}}}$$

  follows approximately the standard normal distribution whatever the underlying distribution for the phenomenon in the target population is.
- But here, $n = 14$ isn't large enough to justify this result.
- A possible solution was suggested by W.S. Gossett under the assumption of normality.

Butterflies and dark beers

W.S. Gossett was a statistician who worked at Guinness as "Head Experimental Brewer" (which sounds like a cool job title!).

Student's t: the origins

Hypothesis

Let $y_1, \ldots, y_n$ be n independent observations from a random variable $N(\mu, \sigma^2)$.

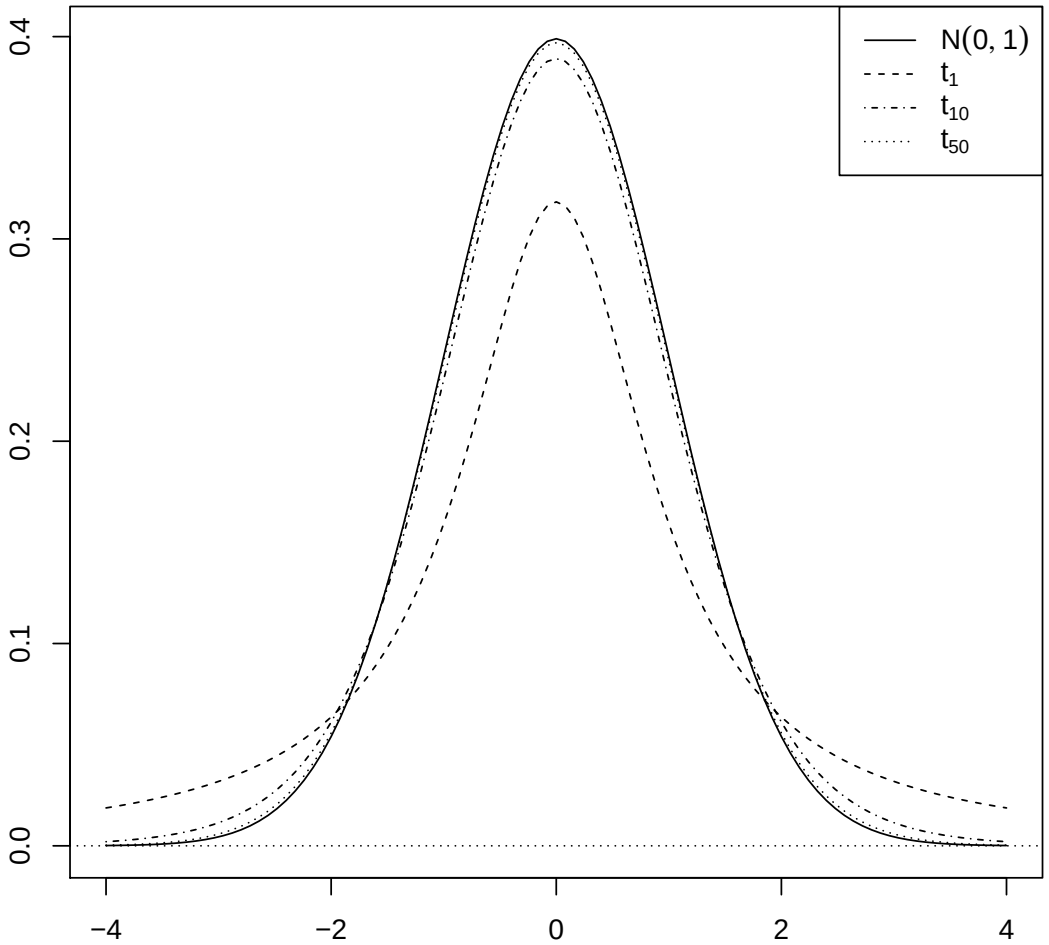
Result

Write $\overline{y}$ and s^2 for the sample mean and sample variance (= the estimates for μ and σ^2 obtained from the data). Then

$$t = \frac{\overline{y} - \mu}{\frac{s}{\sqrt{n}}}$$

distributes like a Student t random variable with $n - 1$ degrees of freedom.

Student's t: graphical representation



The standard normal and three Student t distributions with differing degrees of freedom. Note how the t distribution approaches the normal distribution with increasing degrees of freedom.

Student's t: comments

Name

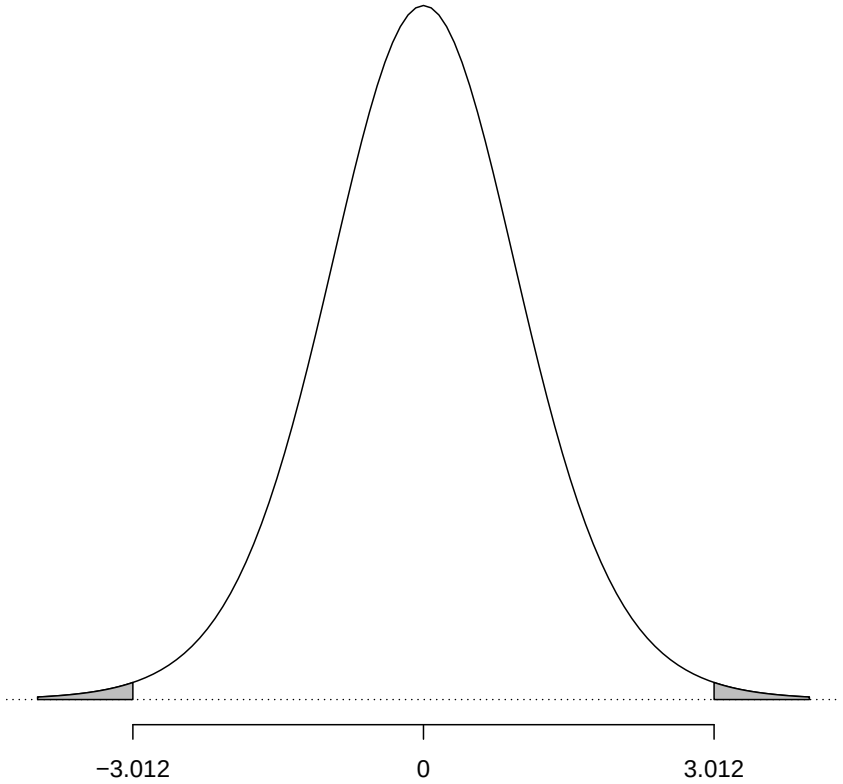
Gosset published his results using the pseudonym of Student. And Gosset used the letter t to indicate his distribution. Therefore, we name it Student's t.

Why is this result important?

It allows us to construct tests and confidence intervals for the mean of a normal distribution for an arbitrary value of n when we don't known the variance.
Indeed, we can proceed as done so far by replacing the normal distribution with a t distribution.

Let's get back to the butterflies

. . . assuming their wings are "normal"



The curve shows a t distribution with 13 degrees of freedom. The grey areas, on the left and on the right, amount to 0,005. Hence, the probability of the interval $[-3,.012; 3,012]$ is 0,99.

Let's get back to the butterflies

. . . assuming their wings are "normal"

Therefore, if $n = 14$,

$$P\left(-3{,}012 \leq \frac{\overline{y} - \mu}{\frac{s}{\sqrt{n}}} \leq 3{,}012\right) = 0{,}99.$$

By isolating μ we find that

$$P\left(\overline{y} - 3{,}012\frac{s}{\sqrt{n}} \leq \mu \leq \overline{y} + 3{,}012\frac{s}{\sqrt{n}}\right) = 0{,}99.$$

That is, the interval

$$[\overline{y} - 3{,}012\frac{s}{\sqrt{n}}; \overline{y} + 3{,}012\frac{s}{\sqrt{n}}]$$

includes the true mean with a confidence level of 99%.

Let's get back to the butterflies

... assuming their wings are "normal"

From the data,

```
> monarche
 [1] 33.9 33.0 30.6 36.6 36.5 34.0 36.1
     32.0 28.0 32.0 32.2 32.2 32.3 30.0
```

we find that

$$\overline{y} = \frac{33.9 + \cdots + 30.0}{14} = 32.8$$

and

$$s^2 = \frac{(33.9 - 32.8)^2 + \cdots + (30.0 - 32.8)^2}{14 - 1} = 6.1.$$

The interval is

$$32.8 \pm 3.012 \cdot \sqrt{\frac{6.1}{14}} = [30.8; 34.8].$$

Let's get back to the butterflies

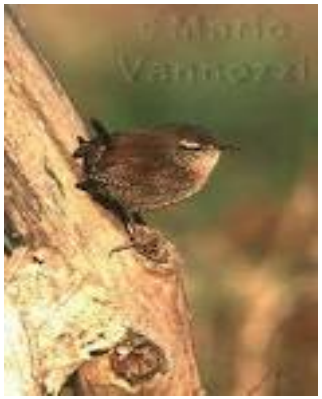
. . . assuming their wings are "normal"

Yet, better let it do by R.

```
> t.test(monarche, conf.level=0.99)

One Sample t-test

data:  monarche
t = 49.5943, df = 13, p-value = 3.332e-16
alternative hypothesis: true mean is not equal to 0

99 percent confidence interval:
 30.82120 34.80737

sample estimates:
mean of x
 32.81429
```

Darwin's cuckoos

. . . not forgetting wrens and redbreasts



Darwin's cuckoos

. . . not forgetting wrens and redbreasts

- It is well known that cuckoos lay their eggs in the nests of other birds, who are then left with the task of hatching.
- It is possible to observe an association between territory and bird chosen as "hosts". That is, in some territories the cuckoos seem to prefer one species of bird as "host", in others another.
- According to the theory of natural selection, we therefore expect some form of adaptation of the cuckoo egg to that of the "host" bird.
- In fact, the probability of an egg being hatched (which, given the habits of the cuckoo, influences the survival of its genetic heritage to a great extent) should be all the higher the more the "abusive" eggs are similar to those of the "host" bird.

Darwin's cuckoos

. . . not forgetting wrens and redbreasts

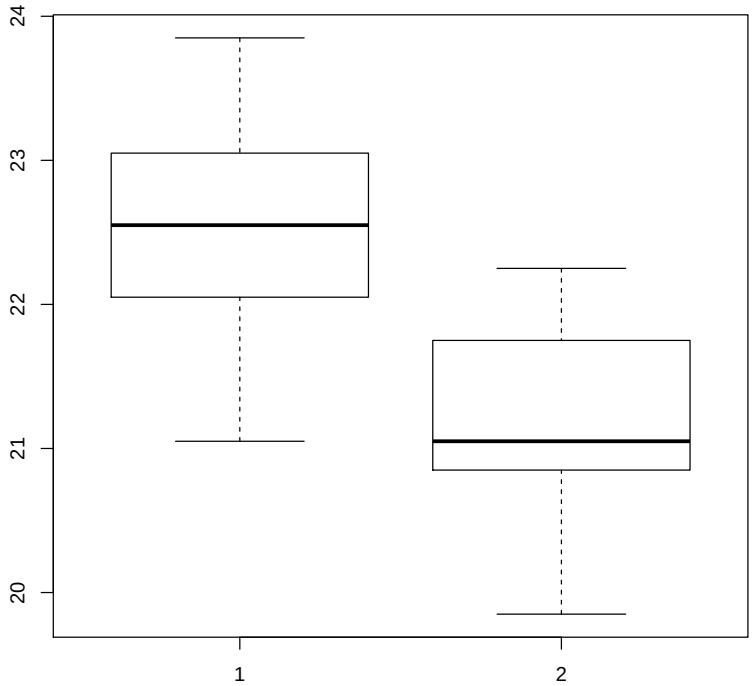
To verify this idea, a study considered the length (in *mm*) of some cuckoo eggs found in nests of redbreasts and wrens of two territories, one in which the cuckoos "prefer" the redbreasts, the other where they "prefer" the wrens.

```
> redbreast
  21.05 21.85 22.05 22.05 22.05 22.25 22.45 22.45
  22.65 23.05 23.05 23.05 23.05 23.05 23.25 23.85
> wren
  19.85 20.05 20.25 20.85 20.85 20.85 21.05 21.05
  21.05 21.25 21.45 22.05 22.05 22.05 22.25
```

Darwin's cuckoos

. . . not forgetting wrens and redbreasts

host	n	$\overline{y}$	s
redbreast	16	22,57	0,68
wren	15	21,13	0,74



Darwin's cuckoos

. . . not forgetting wrens and redbreasts

The question . . . is always the same.

- Wrens are among the smallest birds. Their eggs are therefore smaller than redbreast eggs!
- The observed data (the mean, the position of the boxplots) suggest that the cuckoos adapted to their "hosts".
- Can the difference between the two mean lengths observed in our data be due to chance? That is, can it be due to the fact that we considered a small sample of laid eggs? Do we expect it to happen *in general*?

A bit of math

- The target population is split into two groups.
- The first (second) group includes all eggs which the cuckoos in the considered territories lay in redbreast (wren) nests.
  All eggs means also those we don't know, not only those of our sample.
- Let $\mu_{\text{redbreast}}$ and μ_{wren} be the average length of eggs in the two groups.
  Using the available data we are interested in verifying the two hypotheses

$$H_0 : \mu_{\text{redbreast}} = \mu_{\text{wren}}$$
$$H_1 : \mu_{\text{redbreast}} > \mu_{\text{wren}}$$

Darwin's cuckoos

. . . not forgetting wrens and redbreasts

The cuckoos, wrens, redbreasts and Darwin eventually meet Student.

- The problem here is that the sample sizes (= the number of observations in the two groups) do not allow us to use the results based on the central limit theorem, and variants of it.
- If we assume, however, that the distribution within both groups is normal, a solution exists based on Student's t distribution.
- Two versions are available:
  (i) both groups share the same variance;
  (ii) the two groups have different variances.

Darwin's cuckoos

. . . not forgetting wrens and redbreasts

Let's forget about the details, but rather have a look at the results. This is the version with equal variances.

```
> t.test(redbreasts, wrens, alternative="greater",
         var.equal=TRUE)

Two Sample t-test

data:  redbreasts and wrens
t = 5.633, df = 29, p-value = 2.189e-06
alternative hypothesis:
true difference in means is greater than 0
95 percent confidence interval:
 1.009136      Inf
sample estimates:
mean of x mean of y
   22.575    21.130
```

Darwin's cuckoos

. . . not forgetting wrens and redbreasts

This is the version which doesn't necessarily assume equal variances.

```
> t.test(redbreasts, wrens, alternative="greater")

Welch Two Sample t-test

data:  redbreasts and wrens
t = 5.6175, df = 28.369, p-value = 2.462e-06
alternative hypothesis:
true difference in means is greater than 0
95 percent confidence interval:
 1.00761     Inf
sample estimates:
mean of x mean of y
   22.575    21.130
```

Interpretation and comments

Conclusion The data strongly suggest that there was an adaptation, or, if you prefer, that

"the cuckoos vote for Darwin".

Two-sample t test Allows us to make inference on the mean difference of two groups under the assumption that the distribution of the phenomenon of interest within the groups is normal.

Welch yes or no? Use Welch unless you are sure that the variances are equal.

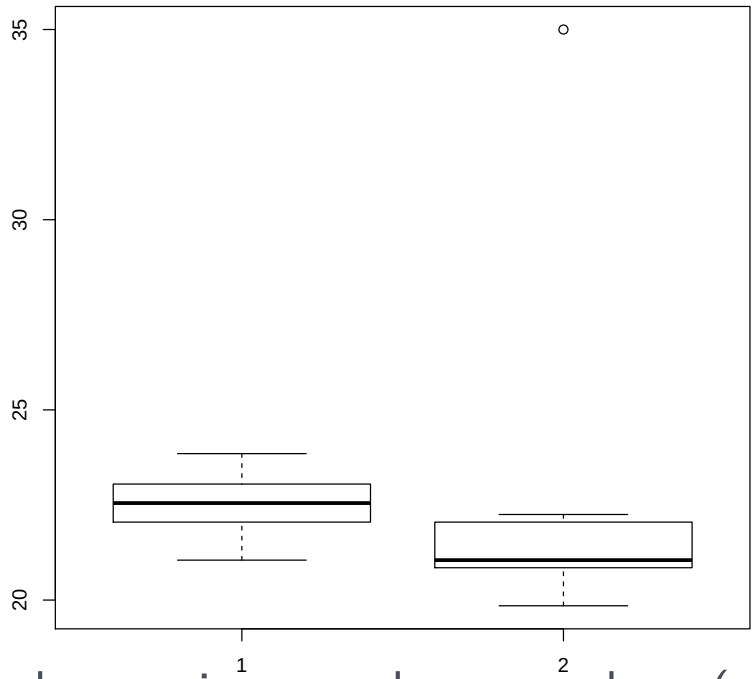
Yet, are butterflies, cuckoos, readbreasts and wrens normal?

Better pay attention

- Watch out for outlying observations.
- Use a test for normality.
  They aren't very powerful for small n, but at least they highlight severe problems.
- Cross-check the results with those obtained from a non-parametric test.

Darwin's cuckoos

. . . not forgetting wrens, redbreasts and why not ostriches



I added an outlying observations to the wren data (a so-called *outlier*). Now, the p-value increases to 0,26. That is, a single observation changes our conclusion.

Testing for normality

- There are several tests for normality.
- The hypotheses are

$$H_0 : \left(\begin{array}{c} \text{the distribution of the} \\ \text{phenomenon considered in the} \\ \text{target population is normal} \end{array} \right)$$

against

$$H_1 : \text{the distribution isn't normal.}$$

- Shapiro-Wilk's test is among the most suitable ones.

Testing for normality

Butterfly wings

```
> shapiro.test(monarche)

Shapiro-Wilk normality test

data:  monarche
W = 0.9458, p-value = 0.4978
```

Cuckoo eggs in redbreast nests

```
> shapiro.test(redbreasts)

Shapiro-Wilk normality test

data:  redbreasts
W = 0.9521, p-value = 0.5239
```

Testing for normality

Cuckoo eggs in wren nests

```
> shapiro.test(wrens)

Shapiro-Wilk normality test

data:  wrens
W = 0.9329, p-value = 0.3019
```

Cuckoo eggs in wren nests (plus one ostrich)

```
> shapiro.test(wrens_ostrich)

Shapiro-Wilk normality test

data:  wrens_ostrich
W = 0.4614, p-value = 1.032e-06
```