UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Statistics

Inference

Sara Geremia

April 11th, 2024

Population, parameter, sample, estimate

- Assuming a given population,
  - e.g.: the subjects eligible to vote in the U.S.

Population, parameter, sample, estimate

- Assuming a given population,
  - e.g.: the subjects eligible to vote in the U.S.
- We are interested in a specific characteristic of the population, that we call parameter,
  - e.g.: the percentage of voters for the Democratic party

Population, parameter, sample, estimate

- Assuming a given population,
  - e.g.: the subjects eligible to vote in the U.S.
- We are interested in a specific characteristic of the population, that we call parameter,
  - e.g.: the percentage of voters for the Democratic party
- We cannot (or we do not want to) observe the entire population (too large, no time), but we can randomly select some subjects: the sample

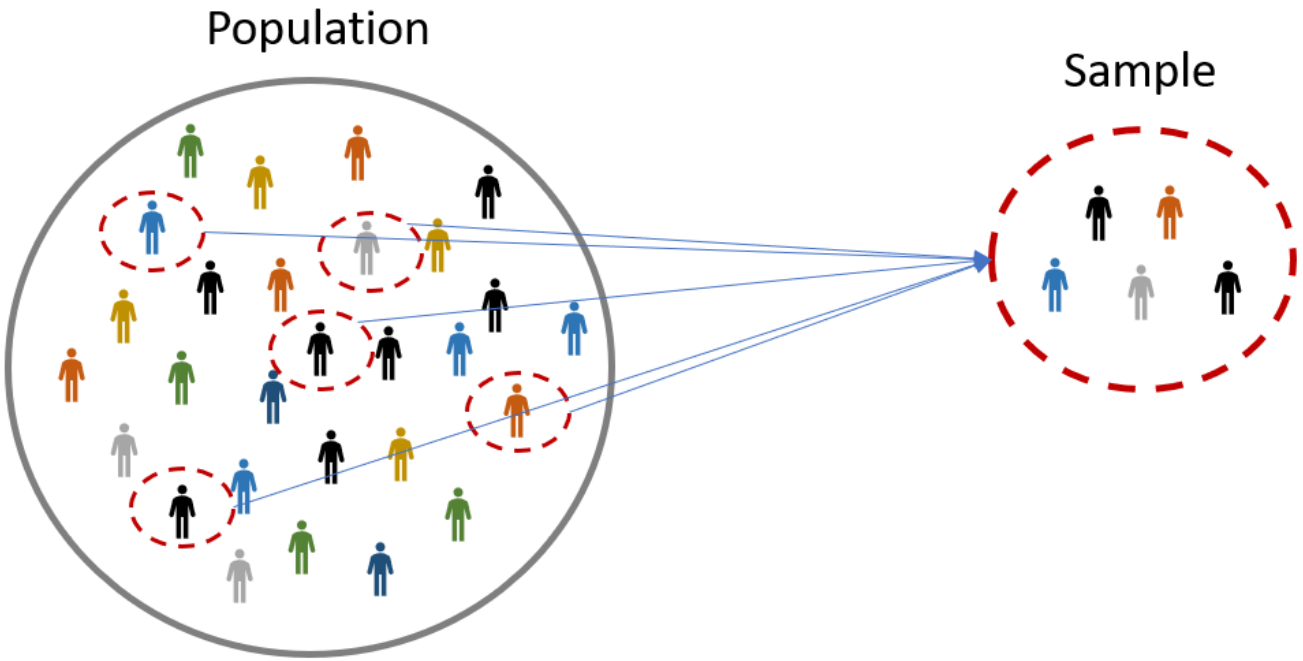
Population, parameter, sample, estimate

- Assuming a given population,
  - e.g.: the subjects eligible to vote in the U.S.
- We are interested in a specific characteristic of the population, that we call parameter,
  - e.g.: the percentage of voters for the Democratic party
- We cannot (or we do not want to) observe the entire population (too large, no time), but we can randomly select some subjects: the sample
- We observe the characteristics of interest of those subjects
  - e.g.: we may ask about their vote preference

Population, parameter, sample, estimate

- Assuming a given population,
  - e.g.: the subjects eligible to vote in the U.S.
- We are interested in a specific characteristic of the population, that we call parameter,
  - e.g.: the percentage of voters for the Democratic party
- We cannot (or we do not want to) observe the entire population (too large, no time), but we can randomly select some subjects: the sample
- We observe the characteristics of interest of those subjects
  - e.g.: we may ask about their vote preference
- Using this information, we obtain an estimate of the parameter of interest
  - e.g.: the percentage of the voters for the given party in the sample

Briefly: population-parameter, sample-estimate



Parameter

% subjects voting for Democrats in the population

Estimate

% subjects voting for Democrats in the sample

The estimate is a quantitative characteristic of the sample that we assume is 'similar' to the parameter

Example: electoral polling

How many people will vote for the Democratic party during the next elections?

Parameter: % of voters for the Democratic party among the 240 million eligible voters

Estimate: % of voters for the Democratic party among the 1000 randomly selected

Sample characteristics

- When we say that the sample is n individuals taken at random, this does not mean that any group of n individuals is fine
- There are various ways, even complex ones, of selecting a valid sample, called representative
- The simplest way to get a representative sample is to choose n individuals so that each individual in the population has an equal chance of being drawn
- Examples of NOT representative samples:
  - selecting the people in a class
  - selecting friends/relatives/acquaintances
  - by asking a question on an Instagram story and inviting the audience to answer

Example: expenditure

Sometimes the population might not be made by subjects

How much does an Italian family spend on average each month?

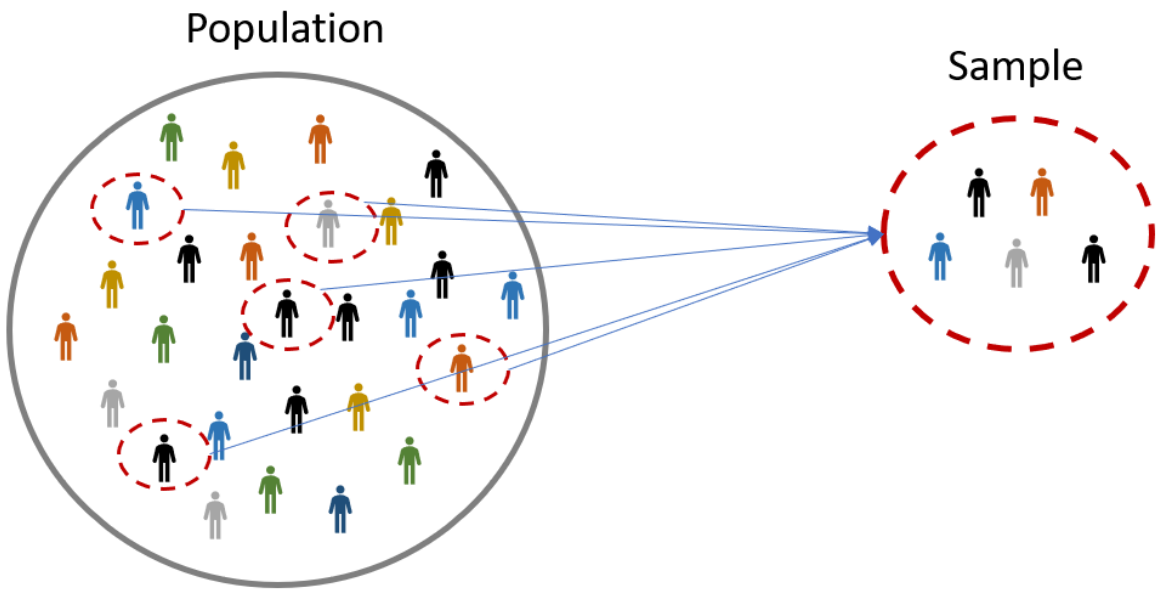
Parameter: Average monthly expenditure of families resident in Italy (23 881 224)

Estimate: Average monthly expenditure of 24400 families resident in Italy

According to ISTAT, the average expenditure is 2480 € (Current survey on household consumption) and the confidence interval is 2450-2510 €

Estimate and parameter

We have identified the characters and the relationship between some of them, but why do we expect the estimate to be close to the parameter value?



Parameter

% subjects voting for Democrats in the population

Average monthly expenditure of families in the population

Estimate

% subjects voting for Democrats in the sample

Average monthly expenditure of families in the sample

Samples and estimates

Let us consider again the U.S. presidential elections example, we want to find the percentage of voters for the Democratic party

This problem has been addressed by various sources conducting polls:

Poll source	sample size	% Democratic
Redfield & Wilton Strategies	8765	53%
SurveyMonkey/Axios	24930	52%
Leger	827	50%

Samples and estimates

Let us consider again the U.S. presidential elections example, we want to find the percentage of voters for the Democratic party

This problem has been addressed by various sources conducting polls:

Poll source	sample size	% Democratic
Redfield & Wilton Strategies	8765	53%
SurveyMonkey/Axios	24930	52%
Leger	827	50%

The three polls give different estimates, but they have been done in the same period on the same population and for the same parameter

Anyone is wrong?

Samples and estimates

Let us consider again the U.S. presidential elections example, we want to find the percentage of voters for the Democratic party

This problem has been addressed by various sources conducting polls:

Poll source	sample size	% Democratic
Redfield & Wilton Strategies	8765	53%
SurveyMonkey/Axios	24930	52%
Leger	827	50%

The three polls give different estimates, but they have been done in the same period on the same population and for the same parameter

Anyone is wrong?

In a sense, they are all wrong

The fact is that different people are interviewed, and therefore the result is random

Estimator, estimate, sampling distribution

An estimate is the result of a random experiment:

i. A sample is randomly extracted,
ii. The estimator (e.g. the sampling proportion) is applied,
iii. The estimate (the value of the estimator for that sample) is obtained.

If I were to extract a new sample, I would obtain a new estimate.

Therefore, is the result random?

Estimator, estimate, sampling distribution

An estimate is the result of a random experiment:

i. A sample is randomly extracted,
ii. The estimator (e.g. the sampling proportion) is applied,
iii. The estimate (the value of the estimator for that sample) is obtained.

If I were to extract a new sample, I would obtain a new estimate.

Therefore, is the result random?

- Yes, the result is random, but not useless

Estimator, estimate, sampling distribution

An estimate is the result of a random experiment:

i. A sample is randomly extracted,
ii. The estimator (e.g. the sampling proportion) is applied,
iii. The estimate (the value of the estimator for that sample) is obtained.

If I were to extract a new sample, I would obtain a new estimate.

Therefore, is the result random?

- Yes, the result is random, but not useless

The estimate is random but it is likely to be close to the population parameter

Estimator, estimate, sampling distribution

An estimate is the result of a random experiment:

i. A sample is randomly extracted,
ii. The estimator (e.g. the sampling proportion) is applied,
iii. The estimate (the value of the estimator for that sample) is obtained.

If I were to extract a new sample, I would obtain a new estimate.

Therefore, is the result random?

- Yes, the result is random, but not useless

The estimate is random but it is likely to be close to the population parameter

(If we are doing everything right)

In practice

To understand how it works, we will try to extract many samples.

The population:

- We have a population of 100 units, represented by 100 slips of paper
- A number is written on each slip
- The parameter is the mean of these numbers

In practice

To understand how it works, we will try to extract many samples.

The population:

- We have a population of 100 units, represented by 100 slips of paper
- A number is written on each slip
- The parameter is the mean of these numbers

The sample:

- Extract 5 slips of paper from the population, this is the sample
- Write the numbers extracted on the sheet
- Compute the mean of the 5 numbers, this is the estimate

In practice

To understand how it works, we will try to extract many samples.

The population:

- We have a population of 100 units, represented by 100 slips of paper
- A number is written on each slip
- The parameter is the mean of these numbers

The sample:

- Extract 5 slips of paper from the population, this is the sample
- Write the numbers extracted on the sheet
- Compute the mean of the 5 numbers, this is the estimate

Let's put the results together

Simulations using R I

- The population, that is, the data relating to all individuals of the population, is stored in the variable of the same name
- A sample of size 10 is randomly selected with

```
sample=sample(population,10)
sample
```

```
## [1]  30  7 32 53 43 50 23 38 76 35
```

- The sample mean is

```
m=mean(sample)
m
```

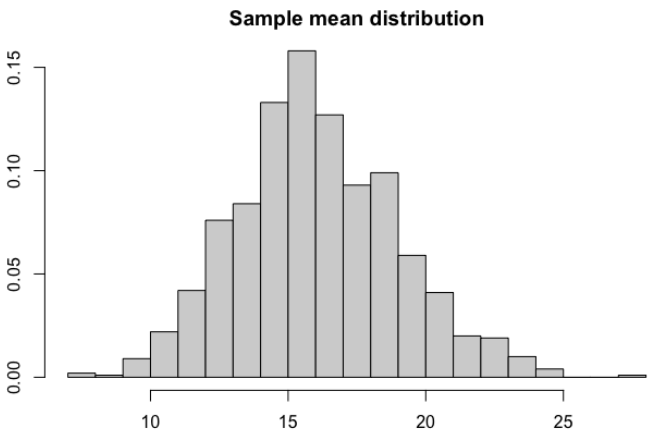
```
## [1] 38.7
```

Simulations using R II

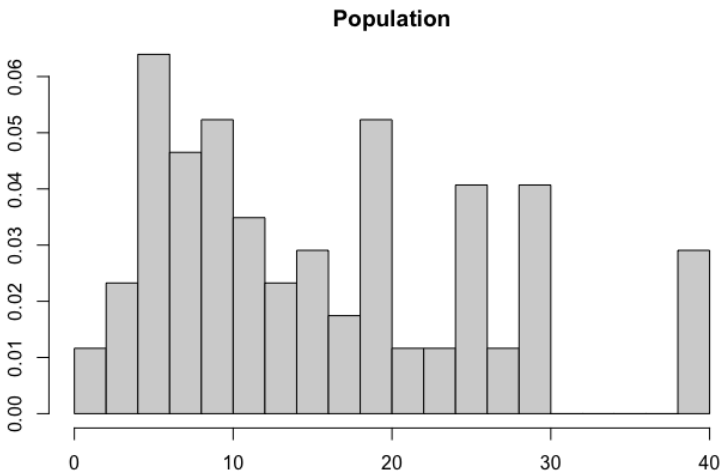
- Now we repeat the sampling step and computation of the estimator 1000 times

Simulations using R III

Check the results

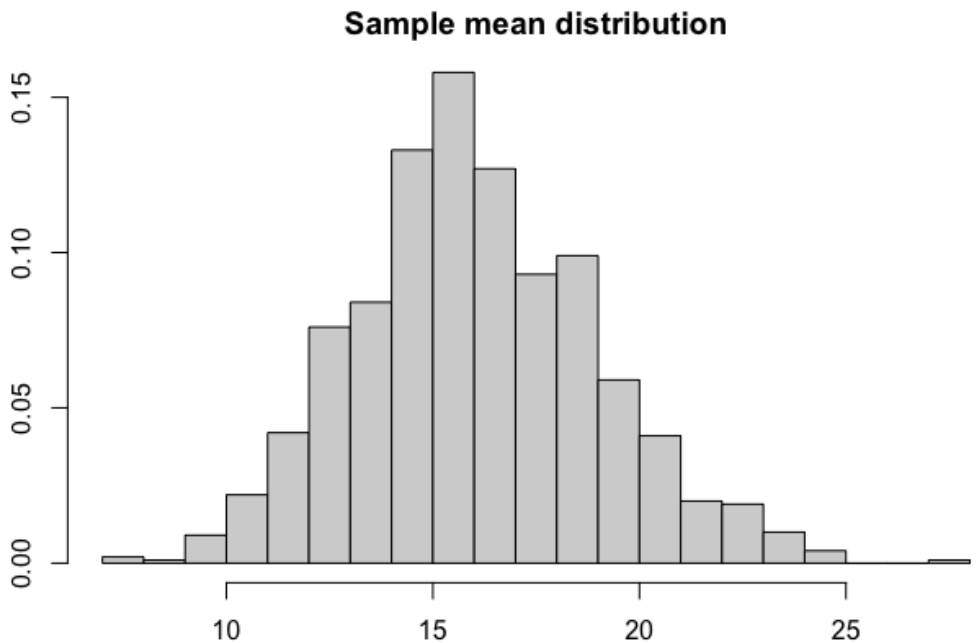


Compare it with the population



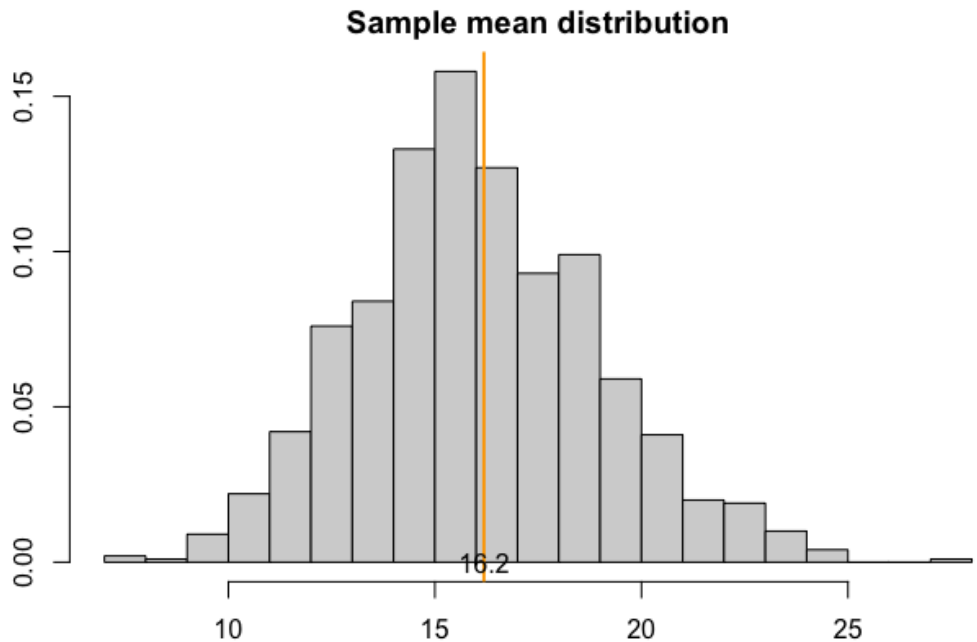
Sampling distribution

The 1000 means derived from our 1000 samples are distributed as follows



Sampling distribution

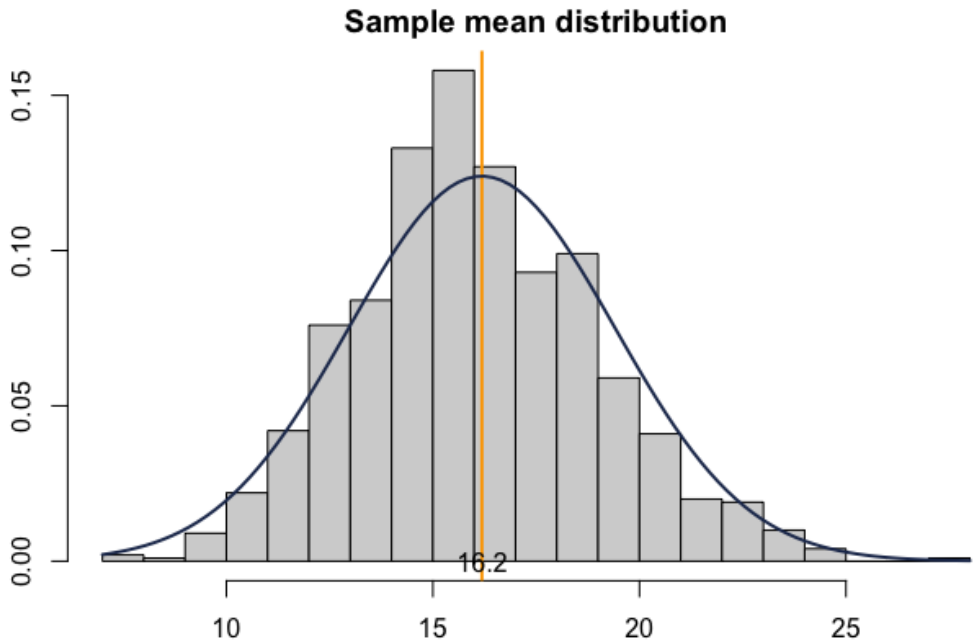
The 1000 means derived from our 1000 samples are distributed as follows



They concentrate around the true value of the parameter, in the sense that values close to the population mean, 16.2, are more likely to be observed than values far away

Sampling distribution

The 1000 means derived from our 1000 samples are distributed as follows

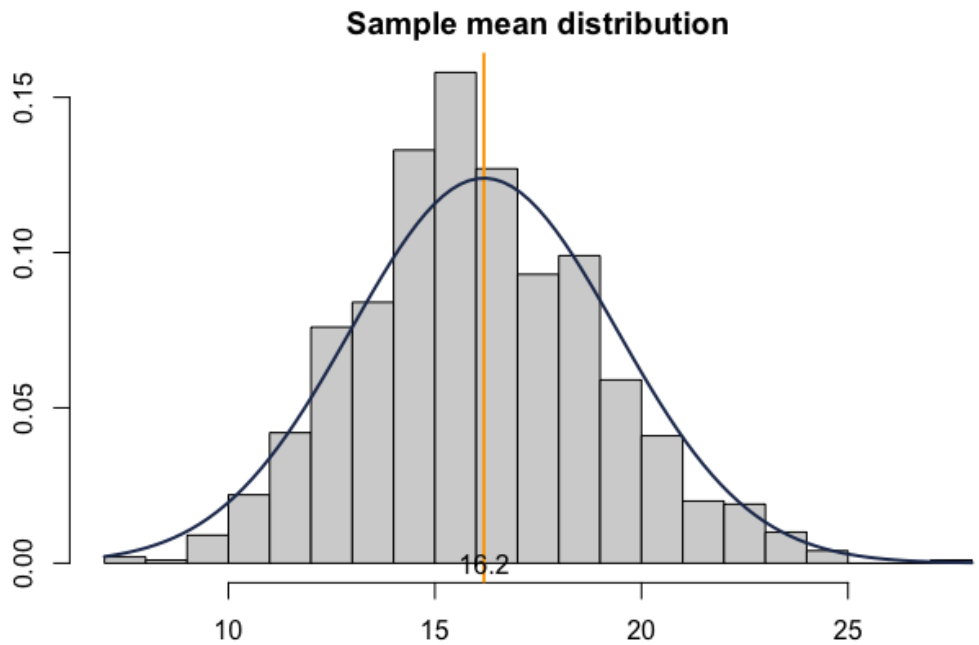


They concentrate around the true value of the parameter, in the sense that values close to the population mean, 16.2, are more likely to be observed than values far away

Imagine we have an infinite number of samples, we can move on to the density function

Sampling distribution

The 1000 means derived from our 1000 samples are distributed as follows



In real life however

- we do not know the population mean
- we observe only one sample

Sampling distribution

We know the distribution but we don't know where it is

We observe a sample and its mean

We can compare the sample with the possible sample mean distributions

$$\bar{X}_n \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$$

Sampling distribution

We know the distribution but we don't know where it is

We observe a sample and its mean

We can compare the sample with the possible sample mean distributions

$$\bar{X}_n \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$$

The most likely value for the population mean is therefore equal to the sample mean, this is why we use it as an estimate

This reasoning also allows us to define ranges of values calculated based on the sample which with a certain probability contains the parameter

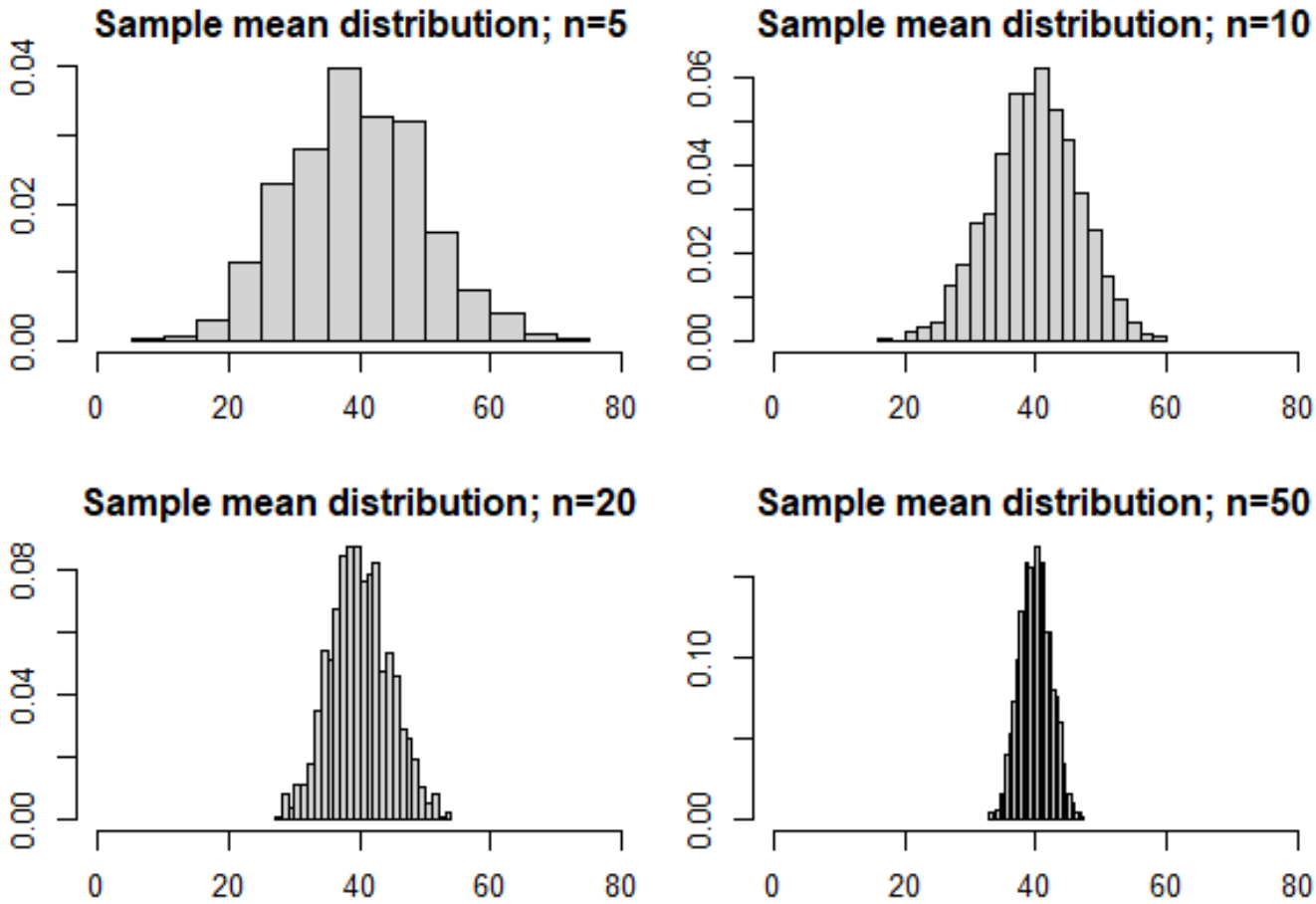
General simulation

Let us explore how the sample size n and the population size N affect the sample mean distribution

We write the instructions so we can easily change some parameters

```
# Subjects in the population
N=10000
# Subjects in the sample
n=10
population=rnorm(N,40,10)
m=vector(mode="numeric",1000)
for (i in 1:1000){
  sample=sample(population,n)
  m[i]=mean(sample)
}
```

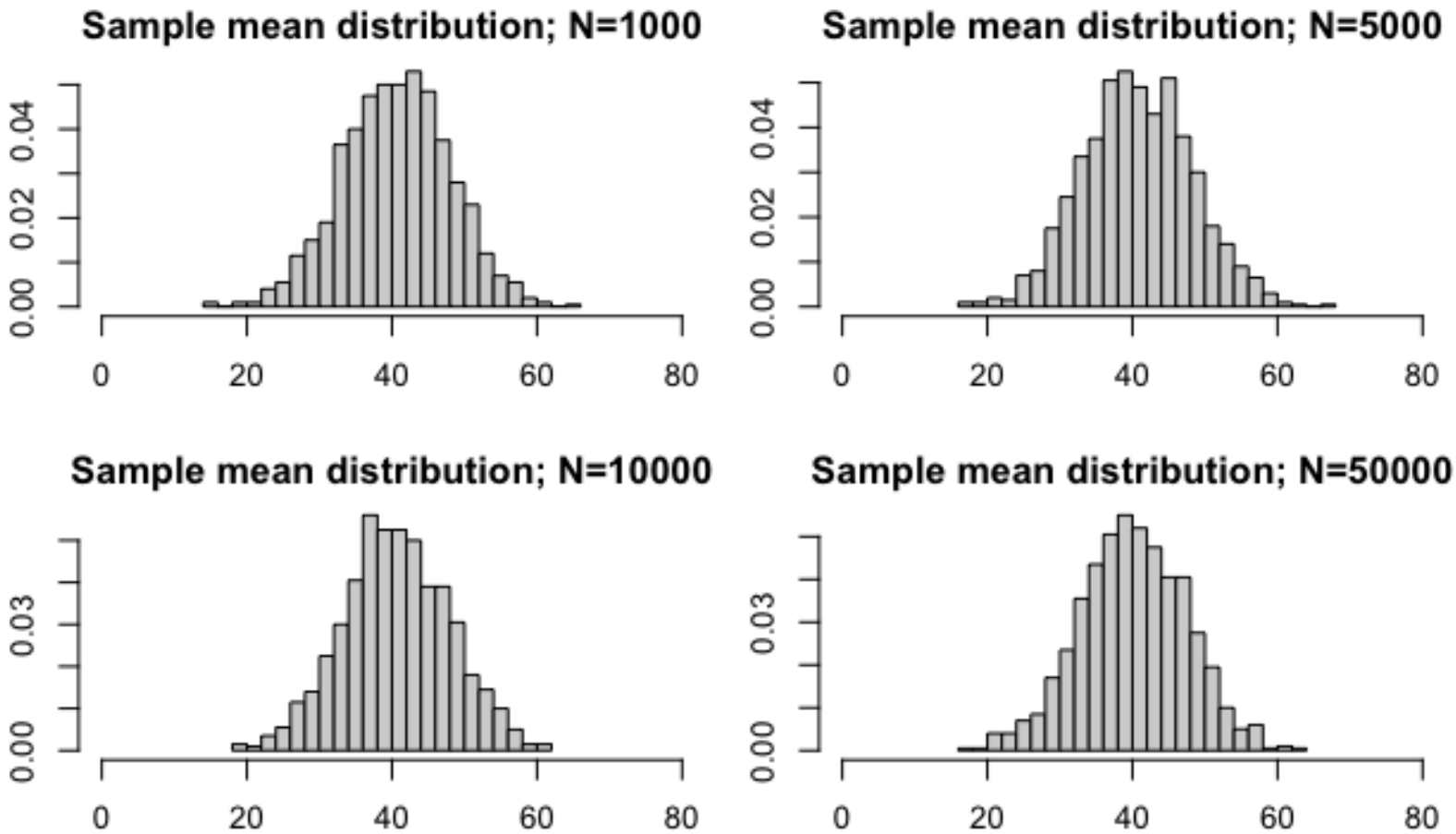
What happens if the sample size n increases?



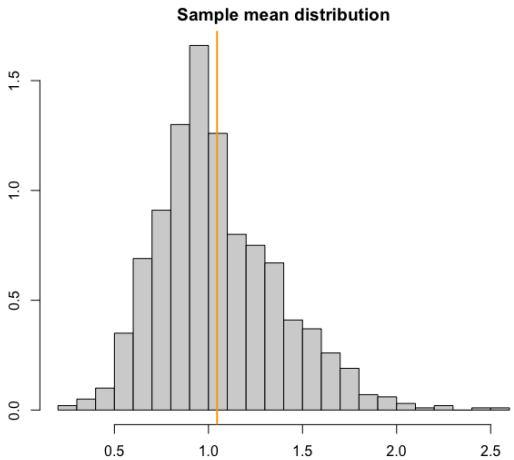
The larger the sample, the more concentrated the estimates around the mean

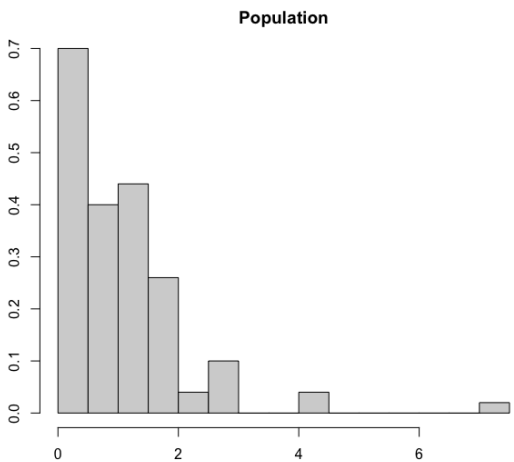
$$\bar{X}_n \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$$

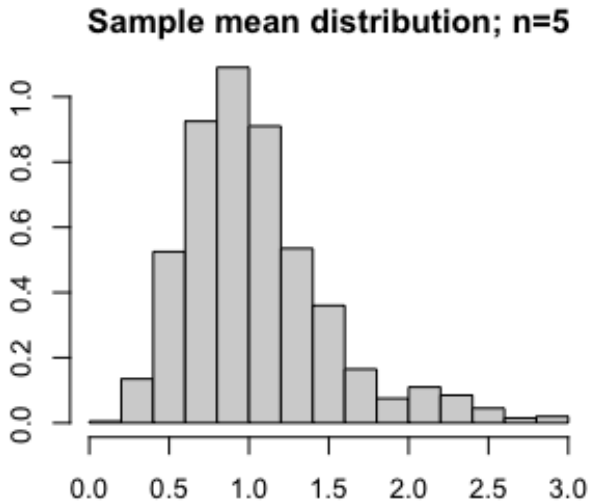
What happens if the N population size increases?

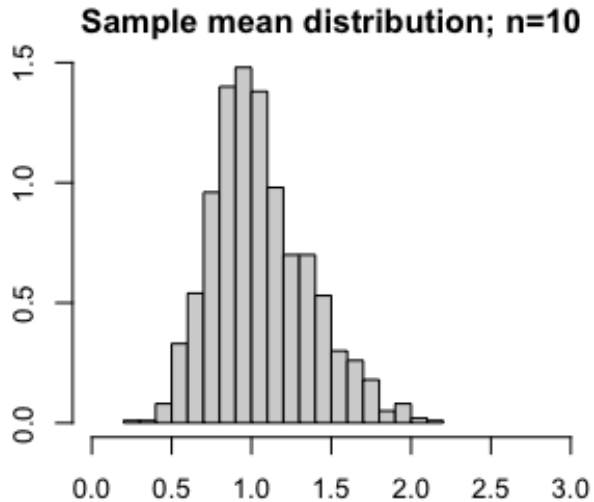


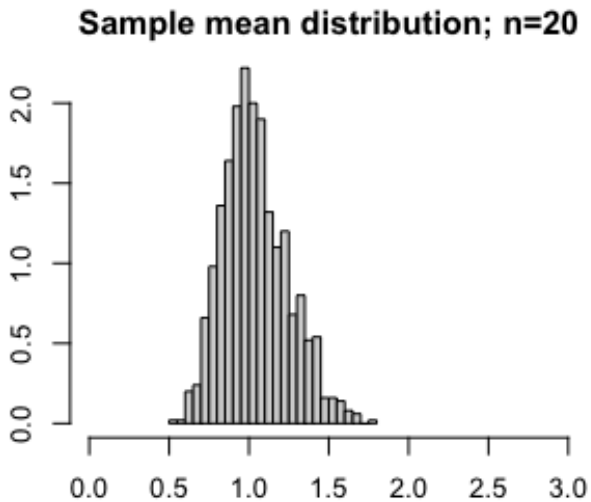
What is the impact of the population distribution?

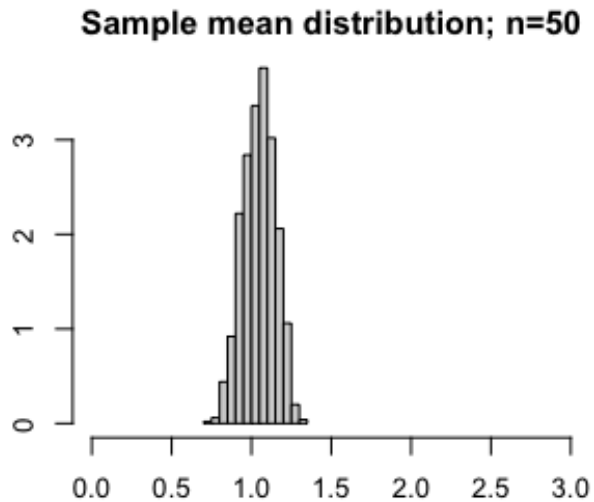












Example

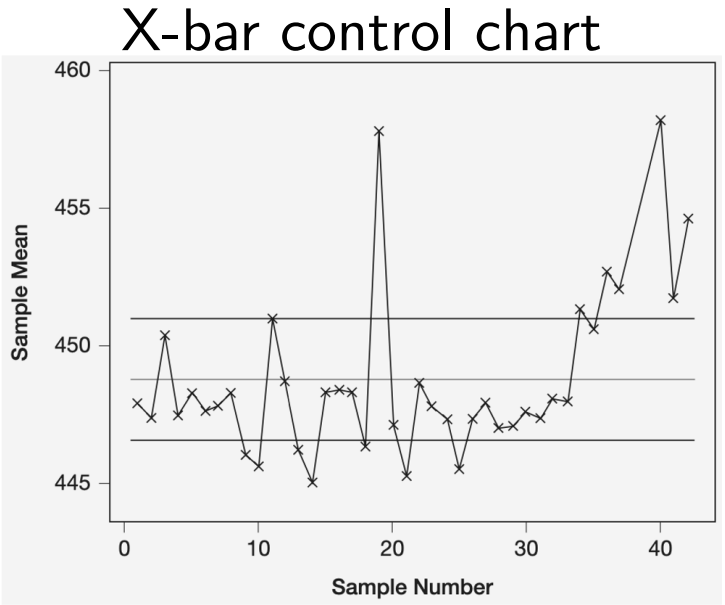
Prairie View Cereals. Inc. is concerned about maintaining correct package weights at its cereal-packaging facility. The package label weight is 440 grams, and company officials are interested in monitoring the process to ensure that package weights are stable

A random sample of five packages is collected every 30 minutes, and each package is weighed electronically

The CLT provides the rationale for using the normal distribution to establish limits for the small sample means

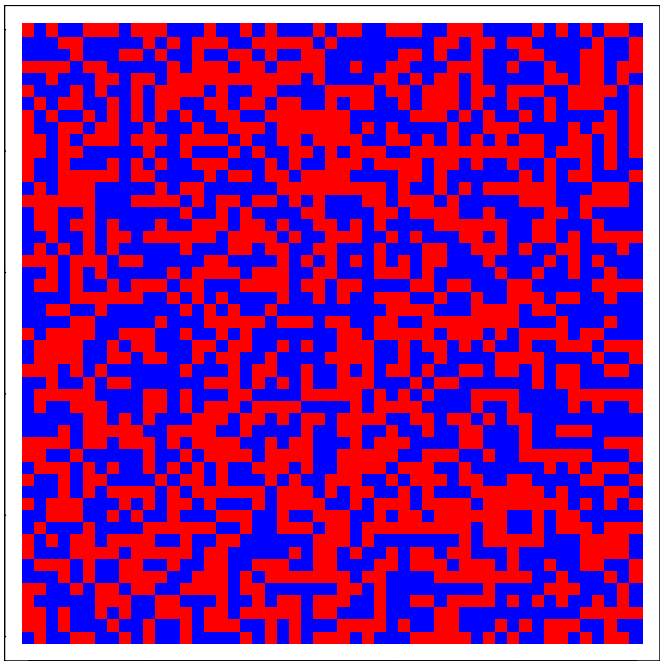
We thus expect that the interval $\bar{X} \pm 3\sigma_{\bar{X}}$ contains almost all the sample means under the normal distribution

Given the number of sample means falling outside the interval, the process should be adjusted



An important statistical theory drives a key management process

Example: probability point of view

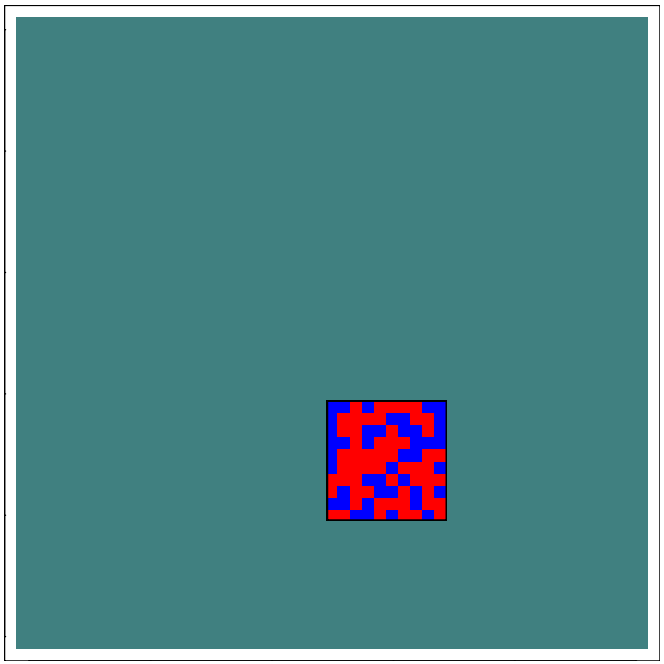


The population of red and blue points

The population is known: $\pi = 50\%$ of the points is red
(Note that for the parameter we use the Greek letters, here π)

Knowing that $\pi = 50\%$ is red

Example: probability point of view



The population of red and blue points

The population is known: $\pi = 50\%$ of the points is red
(Note that for the parameter we use the Greek letters, here π)

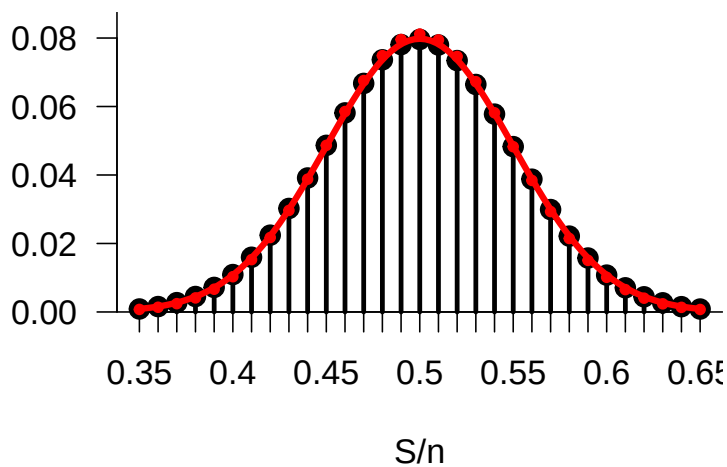
Knowing that $\pi = 50\%$ is red

↓

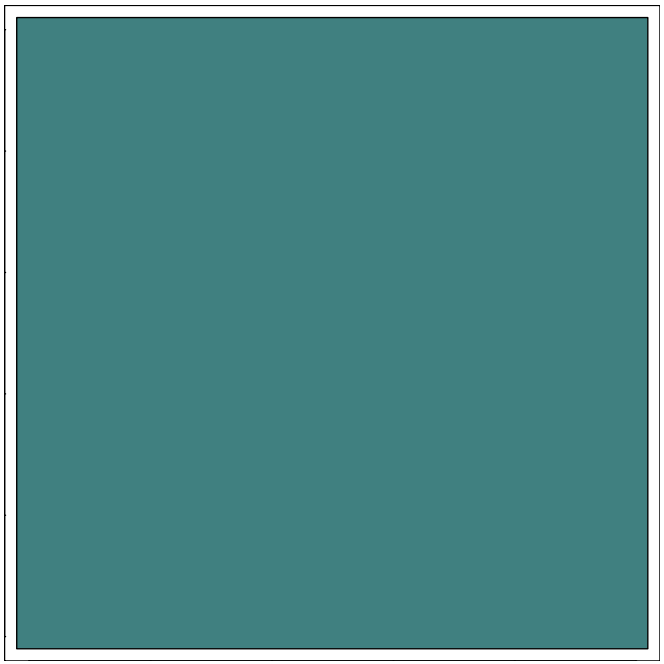
Extracting at random a portion of the population, e.g. $n = 100$ points, we observe a random number of red dots

Using the CLT, we can compute the probability that a proportion of points S/n is red, approximately

$$\bar{S} = S/n \sim N(\pi \frac{\pi(1-\pi)}{n})$$



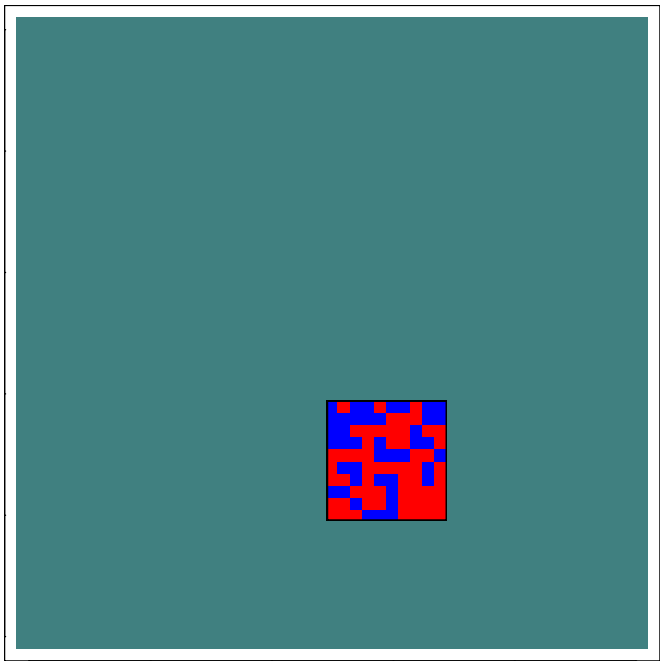
Example: inferential point of view



The population of red and blue points

The population is unknown: $\pi =?$ (proportion of red points)

Example: inferential point of view

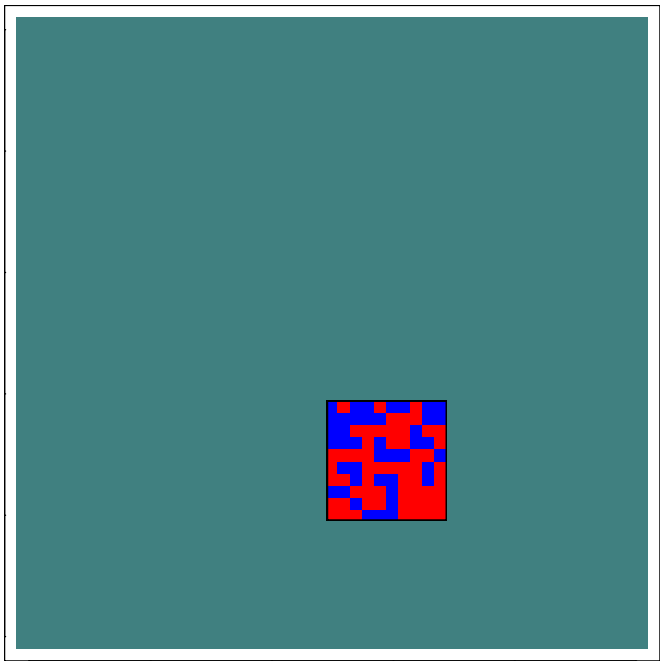


The population of red and blue points

The population is unknown: $\pi =?$ (proportion of red points)

We observe only the sample $S = 54$ over $n = 100$.

Example: inferential point of view



The population of red and blue points

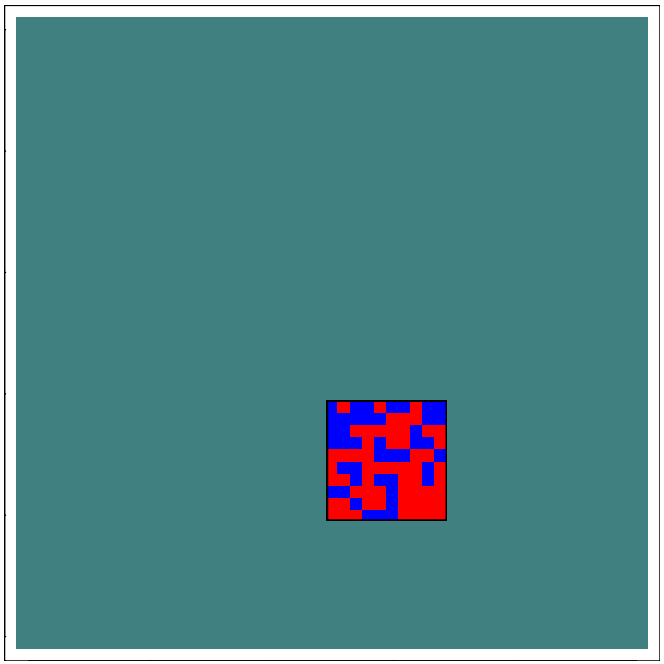
The population is unknown: $\pi =?$ (proportion of red points)

Sampling distribution

$$\bar{S} = S/n \sim N(\pi \frac{\pi(1-\pi)}{n})$$

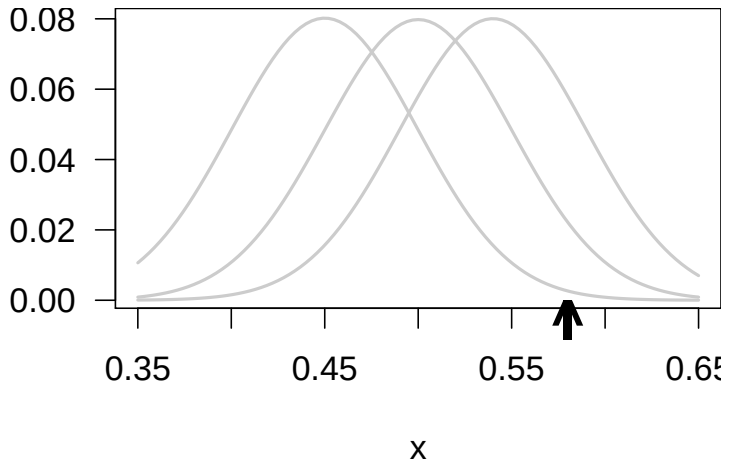
We observe only the sample $S = 54$ over $n = 100$.

Example: inferential point of view



The population of red and blue points

The population is unknown: $\pi =?$ (proportion of red points)



$\bar{S}$ is "probably" close to π (LLN), it will be our estimate of π: $\hat{\pi} = \bar{S}_n$.

↑

Sampling distribution

$$\bar{S} = S/n \sim N(\pi \frac{\pi(1-\pi)}{n})$$

We observe only the sample $S = 54$ over $n = 100$.

Sampling strategies

Almost all statistical methods are based on the idea of randomness

Even in observational studies, we use random samples

However large it may be, a not representative sample does not allow for generalizations (unless additional assumptions are made)

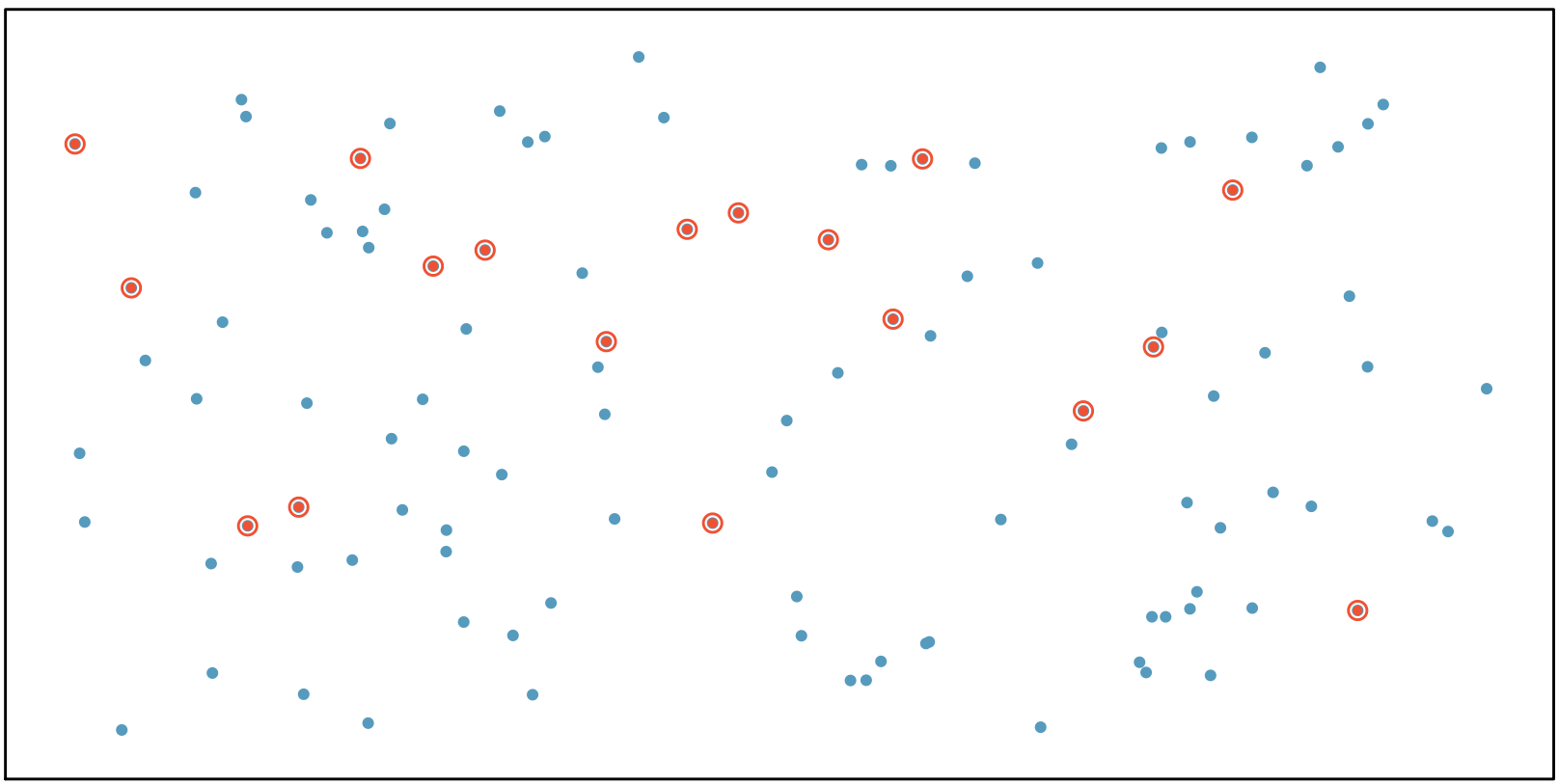
The most common sampling technique is simple random sampling

also the stratified, and cluster random samplings are usually applied (and combined) to

- improve representativeness
- simplify the procedure

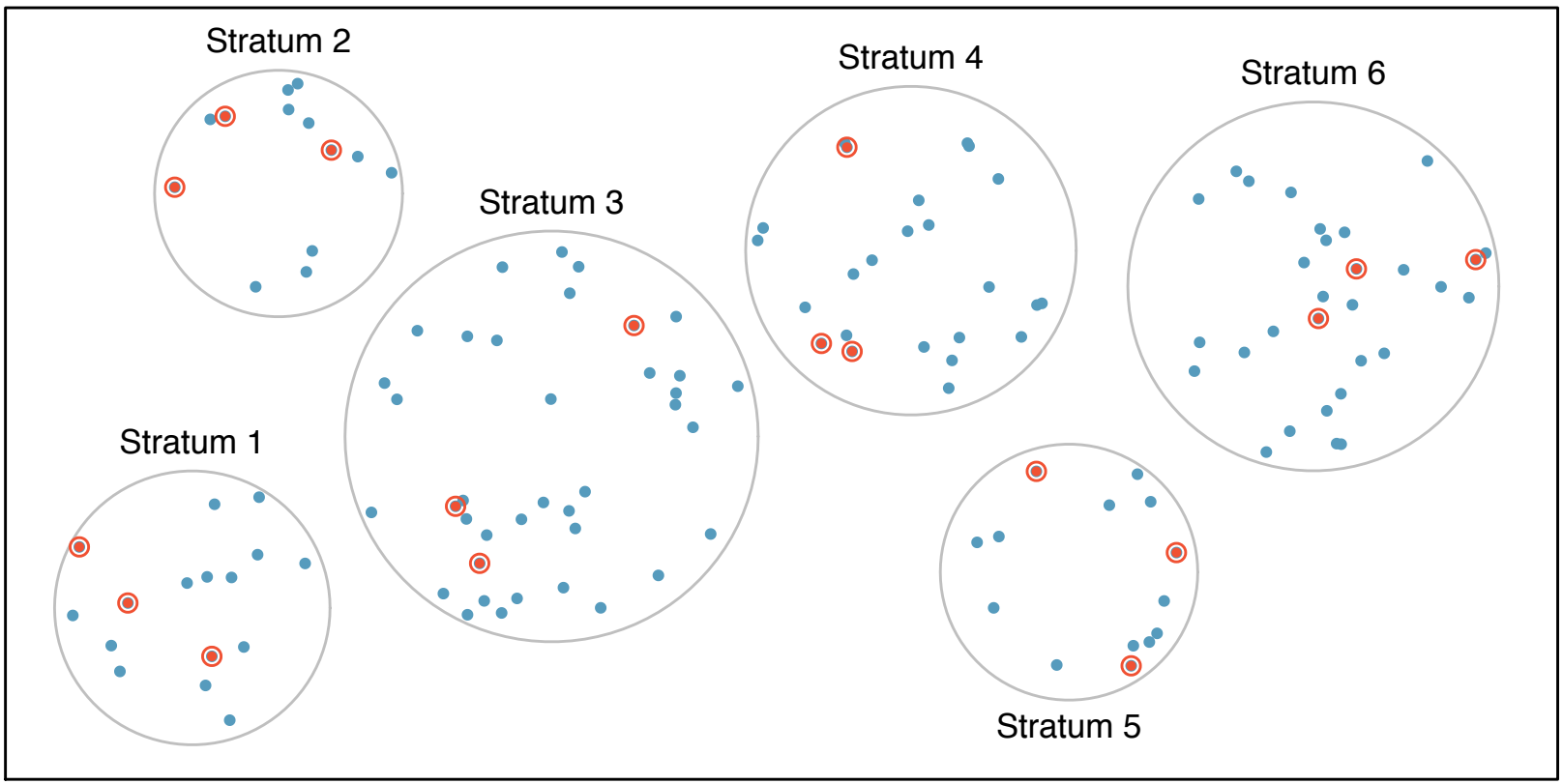
Simple random sample [srs]

The n units are independent and each unit in the population has the same probability of being extracted



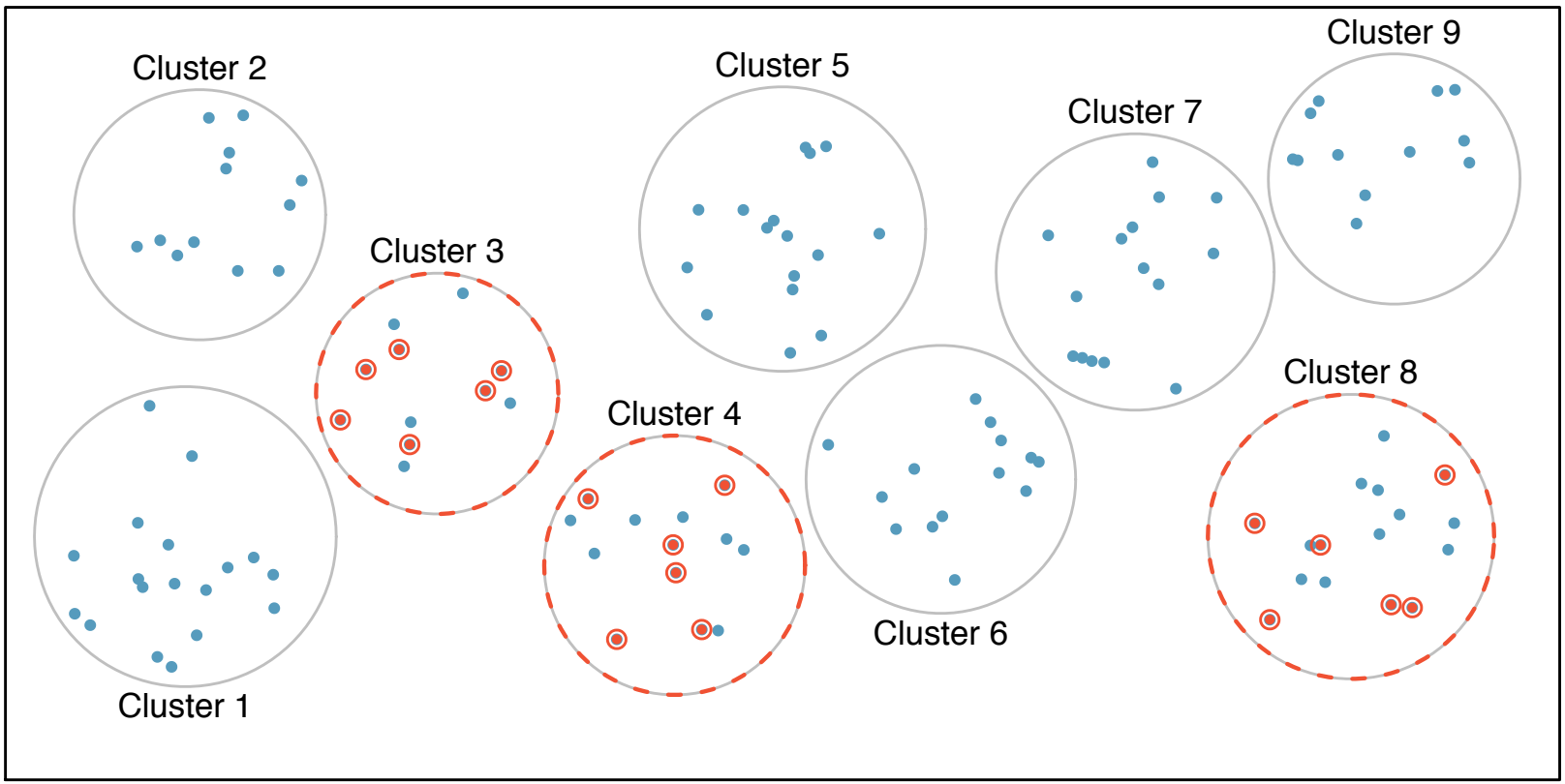
Stratified sample

We divide the population into strata of similar units. Then we extract a srs from each stratum



Cluster sample

The population is divided into clusters, which usually consist of non-homogeneous observations One srs of clusters is taken



In multi-staged sampling (single, double, . . .), first we extract a srs of clusters, then we extract a srs from each cluster

Sampling in electoral polls

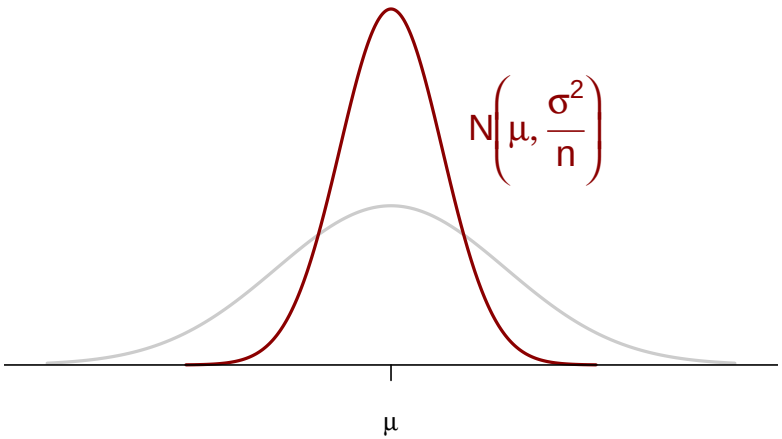
The electoral polls are done with a different logic, and the results that are provided are also based on political considerations and pose some problems

- does not cover the entire population but only people who can be reached by telephone (often only landlines);
- the non-answer could be connected with the vote (it is often stated that those who vote on the right respond less frequently);
- time of the survey temporally distant from the election (people can change their mind).
- Often, quota sampling is used:
  - The population is divided into strata (eg by age and sex).
  - Within the strata, the sample is chosen for convenience.
  - The sample will resemble the frequency as far as the frequency of characteristics subject to stratification
  - There may be selection bias.
  - Generalization is not guaranteed.

Estimator precision - sample mean, known variance

The sampling distribution of an estimator allows to evaluate the error that is committed

Let us consider the sample mean estimator, $\bar{X}$ for the mean of a normal distribution



$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

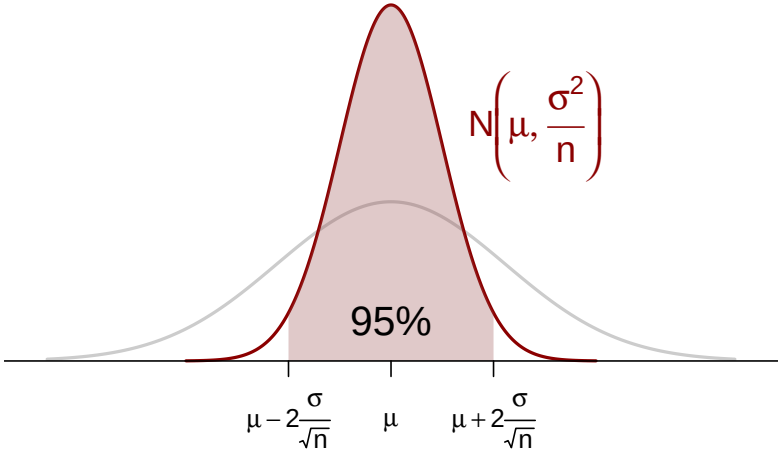
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

The distribution is centered around μ, values close to μ are more likely than those far from

Estimator precision - sample mean, known variance

The sampling distribution of an estimator allows to evaluate the error that is committed

Let us consider the sample mean estimator, $\bar{X}$ for the mean of a normal distribution



$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

The distribution is centered around μ, values close to μ are more likely than those far from

We can be more precise, indeed from the normal distribution properties we know that the observed sample mean $\bar{X}$ falls in a range defined by $\pm 2\frac{\sigma}{\sqrt{n}}$ from the population mean μ with probability 0.95

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu - 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

then we can write

$$P\left(|\bar{X} - \mu| \leq 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

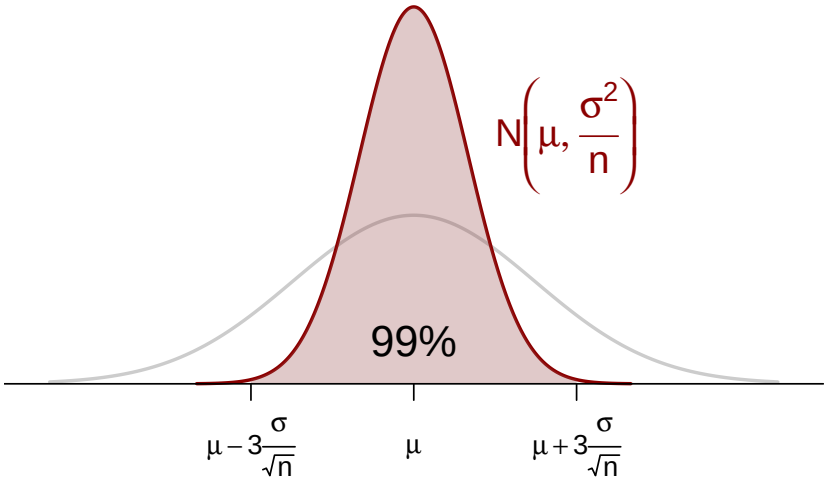
that is, the margin of error (ME) is smaller than $2\frac{\sigma}{\sqrt{n}}$ at 95%

Also, we can say that with a probability of about 0.95 the (unknown) population mean μ falls in the interval

Estimator precision - sample mean, known variance

The sampling distribution of an estimator allows to evaluate the error that is committed

Let us consider the sample mean estimator, $\bar{X}$ for the mean of a normal distribution



$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

The distribution is centered around μ, values close to μ are more likely than those far from

We can increase our "certainty" level but at the price of a bigger margin of error

$$P\left(|\bar{X} - \mu| \leq 3\frac{\sigma}{\sqrt{n}}\right) \approx 0.99$$

that is, the margin of error is at 99% smaller than $3\frac{\sigma}{\sqrt{n}}$

Example

It is considered an industrial process of packaging a food product

Assume that the X weight of a package is normally distributed with standard deviation $\sigma = 5$ around an average μ
We observe a sample of size 10 and the estimate is $\bar{X} = 98$, what can we say about the margin of error?

With probability of 95%, the margin of error is smaller than

$$2\frac{5}{\sqrt{10}} = 2 \times 1.58 = 3.16$$

and with a probability of 99%, the margin of error is smaller than

$$3\frac{5}{\sqrt{10}} = 3 \times 1.58 = 4.74$$

Note that the margin of error does not depend on the value of $\bar{X}$

Example

It is considered an industrial process of packaging a food product

Assume that the X weight of a package is normally distributed with standard deviation $\sigma = 5$ around an average μ
We observe a sample of size 10 and the estimate is $\bar{X} = 98$, what can we say about the margin of error?

Increasing the sample size, the margin of error decreases
With a sample of size 20, at 95% the margin of error is smaller than

$$2\frac{5}{\sqrt{20}} = 2 \times 1.12 = 2.24$$

From here we can easily derive the sample size needed to obtain an estimate with a given margin of error of ± 1 at 95%

$$\frac{\sigma}{\sqrt{n}} = 0.5 \Leftrightarrow \sqrt{n} = \frac{1}{0.5}\sigma = 10 \Rightarrow n = 100$$

Sample size for a given margin of error

Normal sample

In the normal case, we have,

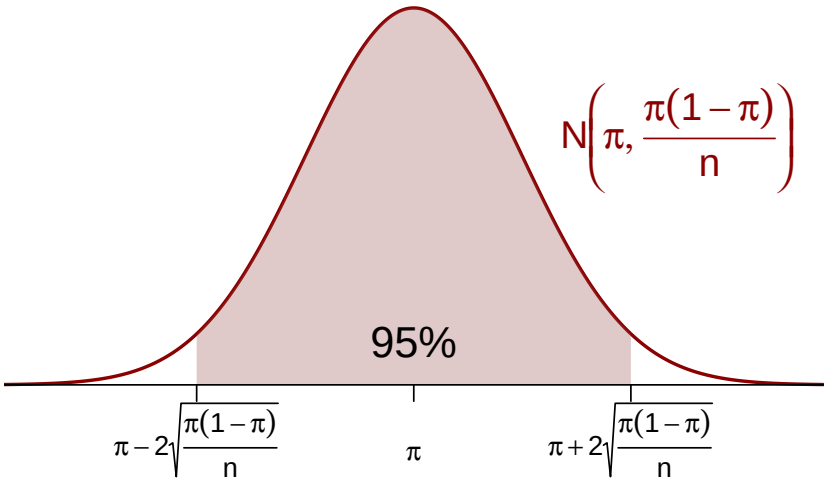
$$P\left(|\bar{X} - \mu| \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

that is, with probability $1 - \alpha$ the margin of error is smaller than

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Estimator precision - sample proportion

Let us consider the sample proportion, $\hat{\pi}$ for the mean of a binomial



$$\hat{\pi} \dot{\sim} \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \dot{\sim} \mathcal{N}(0,1)$$

The approximate distribution is centered around π, values close to π are more likely than those far from π

By the normal distribution properties, we know that the observed sample proportion $\hat{\pi}$ falls in a range defined by $\pm 2\sqrt{\frac{\pi(1-\pi)}{n}}$ from the population proportion π with probability 0.95

$$P\left(\pi - 2\sqrt{\frac{\pi(1-\pi)}{n}} \leq \hat{\pi} \leq \pi + 2\sqrt{\frac{\pi(1-\pi)}{n}}\right) \approx 0.95$$

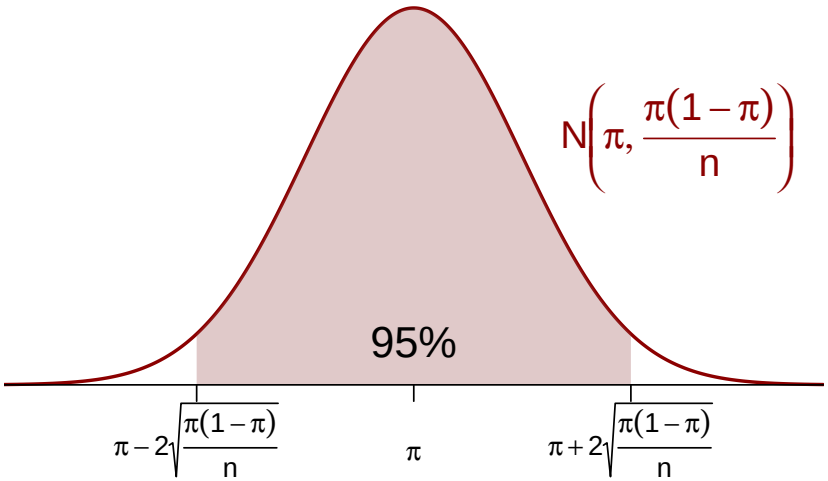
then we can write that at 95% the margin of error is smaller than $2\sqrt{\frac{\pi(1-\pi)}{n}}$

$$P\left(|\hat{\pi} - \pi| \leq 2\sqrt{\frac{\pi(1-\pi)}{n}}\right) \approx 0.95$$

Also, we can say that with a probability of about 0.95 the (unknown) population proportion π falls in the interval (substituting $\hat{\pi}$ for π in the variance)

Estimator precision - sample proportion

Let us consider the sample proportion, $\hat{\pi}$ for the mean of a binomial



$$\hat{\pi} \dot{\sim} \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \dot{\sim} \mathcal{N}(0,1)$$

The approximate distribution is centered around π, values close to π are more likely than those far from π

Equivalently, at 99% the margin of error is smaller than $3\sqrt{\frac{\pi(1-\pi)}{n}}$

$$P\left(|\hat{\pi} - \pi| \leq 3\sqrt{\frac{\pi(1-\pi)}{n}}\right) \approx 0.99$$

Also, we can say that with a probability of about 0.99 the (unknown) population proportion π falls in the interval

$$\left[\hat{\pi} - 3\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + 3\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Example: Binomial

Assuming we want to estimate a proportion π, we observe a sample of size 10 and a sample proportion of $\hat{\pi} = 0.6$, what can we say about the margin of error?

Note that to compute the formula we should know π, but it is unknown. We substitute π with the sampling value $\hat{\pi}$, thus, at 95%, the margin of error is smaller than

$$2\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 2\sqrt{\frac{0.6(1-0.6)}{10}} = 2 \times 0.155 = 0.31$$

Also in this case, we can evaluate an appropriate sample size to obtain a given margin of error, say ± 0.1, at 95%

$$0.1 = 2\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq 2\sqrt{\frac{0.5(1-0.5)}{n}}$$

then

$$\sqrt{n} \geq \frac{0.5}{0.05} = 10 \Rightarrow n \geq 100$$

Sample size for a given margin of error

Binomial sample

In the binomial case, we have,

$$P\left(|\hat{\pi} - \pi| \leq z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1 - \alpha$$

that is, with probability $1 - \alpha$ the margin of error is smaller than

$$z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

Errors in polls

In the poll conducted by Research and Co., the support in favour of Republicans was estimated to be 42%, what does this mean?

- The sample size was $n = 1025$ (1000-1500 is a common sample size for electoral pools).
- Assume that the extracted sample was representative of the population: subjects are randomly selected and they were honest about their preference (usually this is not true in dealing with electoral polls).

Observing the 42% means that, the margin of error at 95% is not greater than

$$1.96\sqrt{\frac{0.42(1-0.42)}{1025}} = 1.96 \times 0.015 \approx 0.03$$

then we are "sure at 95%" that the true percentage is in the interval

$$[0.42 - 0.03, 0.42 + 0.03] \quad \rightarrow \quad [0.39, 0.45]$$

Errors in polls

In the poll conducted by Research and Co., the support in favour of Republicans was estimated to be 42%, what does this mean?

- The sample size was $n = 1025$ (1000-1500 is a common sample size for electoral pools).
- Assume that the extracted sample was representative of the population: subjects are randomly selected and they were honest about their preference (usually this is not true in dealing with electoral polls).

If we want to be "sure at 99%" we compute the margin of error

$$2.57\sqrt{\frac{0.42(1-0.42)}{1025}} = 0.04$$

then, the interval will be wider

$$[0.42 - 0.04, 0.42 + 0.04] \quad \rightarrow \quad [0.38, 0.46]$$

How do we choose n

The first point in selecting n is fixing the acceptable margin of error

We compute the margins of error related to different n for a fixed "confident" level, say 95%

$$n = 1000 \quad \rightarrow \quad 1.96\sqrt{\frac{\pi(1-\pi)}{1000}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{1000}} = 0.031$$

How do we choose n

The first point in selecting n is fixing the acceptable margin of error

We compute the margins of error related to different n for a fixed "confident" level, say 95%

$$n = 500 \quad \rightarrow \quad 1.96\sqrt{\frac{\pi(1-\pi)}{500}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{500}} = 0.043$$

$$n = 1500 \quad \rightarrow \quad 1.96\sqrt{\frac{\pi(1-\pi)}{1500}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{1500}} = 0.025$$

$$n = 2000 \quad \rightarrow \quad 1.96\sqrt{\frac{\pi(1-\pi)}{2000}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{2000}} = 0.021$$

Note that it decreases slower than it increases

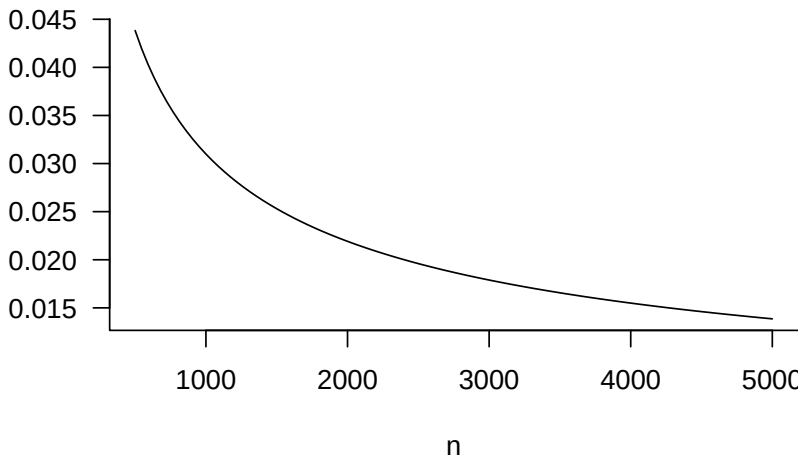
How do we choose n

The first point in selecting n is fixing the acceptable margin of error

We compute the margins of error related to different n for a fixed "confident" level, say 95%

We can plot the errors as a function of n

$$1.96\sqrt{\frac{0.5(1-0.5)}{n}}$$



Then, to halve the margin of error we need to sample quadruple!

Variance estimator

Let us consider again the case $X \sim \mathcal{N}(\mu, \sigma^2)$. We defined an estimator for μ ($\bar{X}$, the sample mean) but not for σ^2, when it is unknown

We can use the sample variance

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

but this is a biased estimator, for this reason, we prefer to use

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

called unbiased sample variance

Point estimate and confidence intervals

In the point estimate, we compute a likely value for the parameter from the sample

What is the probability that we obtain exactly the parameter value?

$$P(\bar{X} = \mu) = ?$$

Point estimate and confidence intervals

In the point estimate, we compute a likely value for the parameter from the sample

What is the probability that we obtain exactly the parameter value?

$$P(\bar{X} = \mu) = ?$$

$\bar{X}$ is a continuous random variable, then

$$P(\bar{X} = \mu) = 0$$

So with probability equal to 1, we are wrong

We know that the estimate is likely close to the parameter value

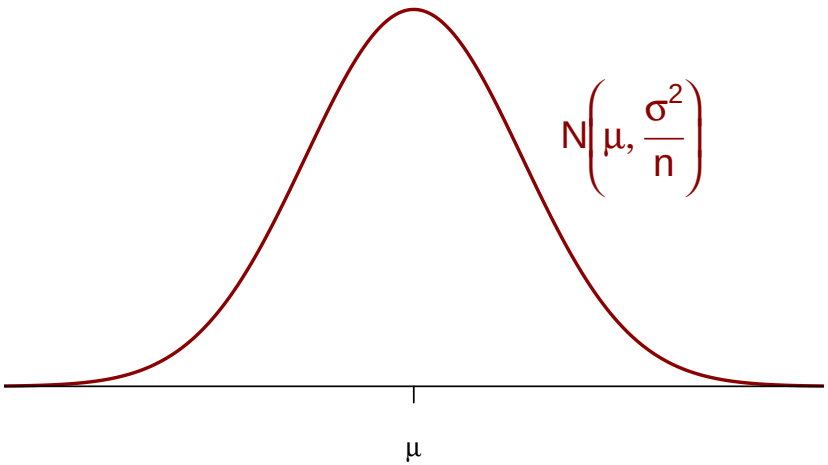
And we can evaluate how far away it is, using the sampling distribution

We can use this information to get an interval estimate (actually, we have not done it yet!)

An interval is a way of expressing the estimate and the uncertainty together

CI for the mean of a normal [variance known]

Consider the estimator $\bar{X}$ for the mean of a Normal

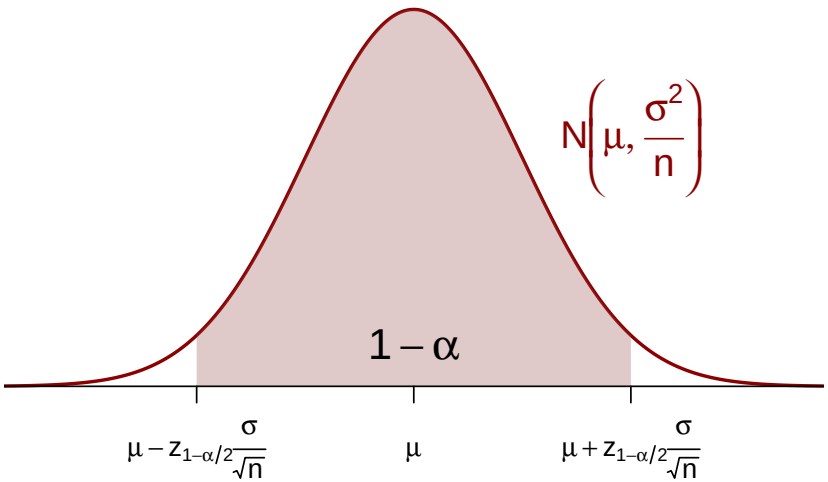


We start from

$$D = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

CI for the mean of a normal [variance known]

Consider the estimator $\bar{X}$ for the mean of a Normal



We start from

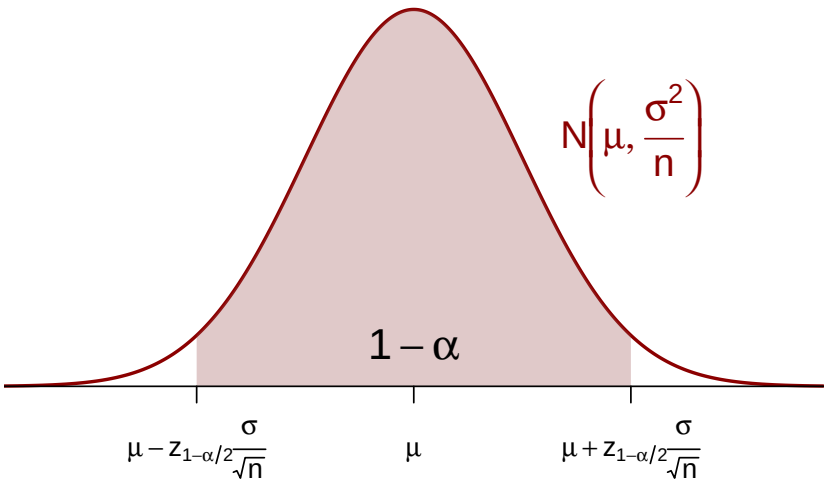
$$D = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

from which

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

CI for the mean of a normal [variance known]

Consider the estimator $\bar{X}$ for the mean of a Normal



We start from

$$D = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

from which

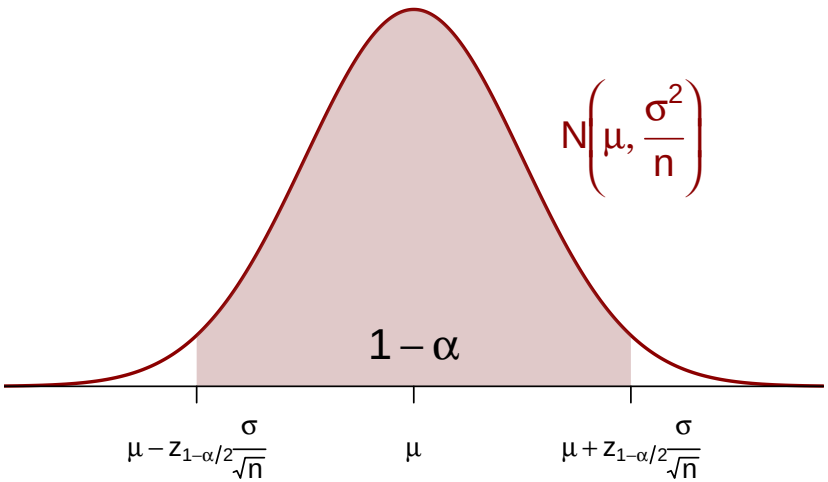
$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

equivalently we can write

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

CI for the mean of a normal [variance known]

Consider the estimator $\bar{X}$ for the mean of a Normal



We start from

$$D = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

from which

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

equivalently we can write

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

The random interval with bounds

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

contains the population mean μ with probability $1 - \alpha$

Confidence interval for the mean [known variance]

From

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

we know that for all μ and σ

$$P\left(|\bar{X}-\mu| < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

that is, the interval

$$\left[\bar{X}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

contains, with probability $1-\alpha$ the parameter value μ (for each μ and σ^2)

Confidence interval for μ

The interval

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

that contains, with probability $1-\alpha$ the parameter value μ (for each μ and σ^2) is called confidence interval for μ of level $1-\alpha$

Instead of a single number ($\bar{X}$), we estimate μ with an interval of values, this accounts for the uncertainty around μ that (partially) remains also after observing the sample (the interval has random bounds that depends on the sample)

Example: beans

In the food industry, the machine that canning beans produces packages whose actual weight is distributed according to a normal with an average μ and standard deviation equal to 3 grams, we want to estimate the mean weight μ

We weigh 10 packages and the sample mean results are equal to 101

We want to obtain a confidence interval at 95% for μ

The relevant quantile of the normal distribution is

$$z_{1-0.05/2} = z_{0.975} = 1.96$$

The CI bounds are

$$101 \pm 1.96 \times \frac{3}{\sqrt{10}} = 101 \pm 1.86$$

Example: beans

In the food industry, the machine that canning beans produces packages whose actual weight is distributed according to a normal with an average μ and standard deviation equal to 3 grams, we want to estimate the mean weight μ

We weigh 10 packages and the sample mean results are equal to 101

We want to obtain a confidence interval at 99% for μ

The relevant quantile of the normal distribution is

$$z_{1-0.01/2} = z_{0.995} = 2.57$$

The CI bounds are

$$101 \pm 2.57 \times \frac{3}{\sqrt{10}} = 101 \pm 2.44$$

Example: beans

In the food industry, the machine that canning beans produces packages whose actual weight is distributed according to a normal with an average μ and standard deviation equal to 3 grams, we want to estimate the mean weight μ

We want to obtain a CI at 95% with width lower or equal to 1gr, what the sample size should be?

The width of the interval is

$$2z_{0.975}\frac{3}{\sqrt{n}}$$

if we want to fix it equal to 1

$$2z_{0.975}\frac{3}{\sqrt{n}} \leq 1 \Rightarrow n \geq (6z_{0.975})^2 = 138.29$$

then we have to increase the sample size up to 139 units

CI width and sample size

The CI width is

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

then, if you want to obtain a CI of level $1-\alpha$ with width lower or equal to w, we have to write

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq w$$

then

$$n \geq 4\left(z_{1-\alpha/2}\frac{\sigma}{w}\right)^2$$

Example

Q: Let $\sigma = 5$, $n = 100$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.95$?

Example

Q: Let $\sigma = 5$, $n = 100$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.95$?

A: It is the interval

$$\left[6 - 1.96\frac{5}{10}, 6 + 1.96\frac{5}{10}\right] \quad \rightarrow \quad [5.02, 6.98]$$

Example

Q: Let $\sigma = 5$, $n = 100$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.95$?
A: It is the interval

$$\left[6 - 1.96\frac{5}{10}, 6 + 1.96\frac{5}{10}\right] \quad \rightarrow \quad [5.02, 6.98]$$

Q: Let $\sigma = 5$, $n = 100$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.98$?

Example : different α

Q: Let $\sigma = 5$, $n = 100$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.95$?
A: It is the interval

$$\left[6 - 1.96\frac{5}{10}, 6 + 1.96\frac{5}{10}\right] \quad \rightarrow \quad [5.02, 6.98]$$

Q: Let $\sigma = 5$, $n = 100$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.98$?
A: We have $\alpha = 0.02$, thus $1 - \alpha/2 = 0.99$, we have to find $\Phi^{-1}(0.99) = 2.33$ then the interval is

$$\left[6 - 2.33\frac{5}{10}, 6 + 2.33\frac{5}{10}\right] \quad \rightarrow \quad [4.83, 7.16]$$

We have higher "confidence" in the second interval, which is wider

Example: different n

Q: Let $\sigma = 5$, $n = 100$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.95$?
A: It is the interval

$$\left[6 - 1.96\frac{5}{10}, 6 + 1.96\frac{5}{10}\right] \quad \rightarrow \quad [5.02, 6.98]$$

Q: Let $\sigma = 5$, $n = 200$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.95$?

Example: different n

Q: Let $\sigma = 5$, $n = 100$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.95$?
A: It is the interval

$$\left[6 - 1.96\frac{5}{10}, 6 + 1.96\frac{5}{10}\right] \quad \rightarrow \quad [5.02, 6.98]$$

Q: Let $\sigma = 5$, $n = 200$, and $\bar{X} = 6$ be the standard deviation, the sample size, and the sample mean, respectively, what is the C.I. of level $1 - \alpha = 0.95$?
A: It is the interval

$$\left[6 - 1.96\frac{5}{\sqrt{200}}, 6 + 1.96\frac{5}{\sqrt{200}}\right] \quad \rightarrow \quad [5.31, 6.69]$$

We got the same estimate from two samples of different sizes, but the uncertainty is lower

Summary

The interval

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

that contains with probability $1-\alpha$ the parameter μ (for each μ and σ^2) is called confidence interval for μ of level $1-\alpha$

- When $1-\alpha$ increases α decreases, $1-\alpha/2$ increases, $z_{1-\alpha/2}$ increases, the CI increases: to have a higher probability to include the value of μ we have to increase the interval
- When σ^2 increases, the CI increases: the variability around $\bar{X}$ increases
- When n increases, the CI decreases: the variability around $\bar{X}$ decreases

CI for the mean [unknown variance]

We estimate σ^2 with its unbiased sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

we replace

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

that we cannot evaluate, with

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

that does not follow a Normal distribution, but a Student t with $n-1$ degrees of freedom

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

CI for the mean [unknown variance]

From

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

we know that

$$P\left(|\bar{X}-\mu| < t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} < \bar{X}-\mu < t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X}-t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X}+t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

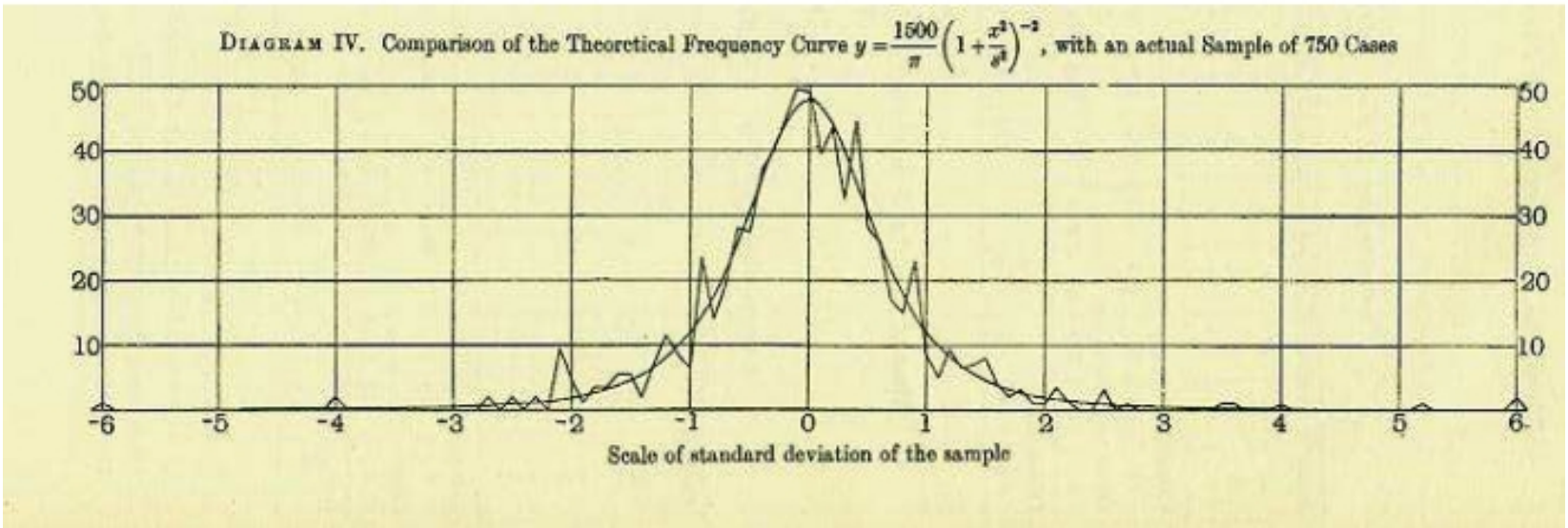
that is, the interval

$$\left[\bar{X}-t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X}+t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

contains, with probability $1-\alpha$ the parameter value μ (for each μ and σ^2)

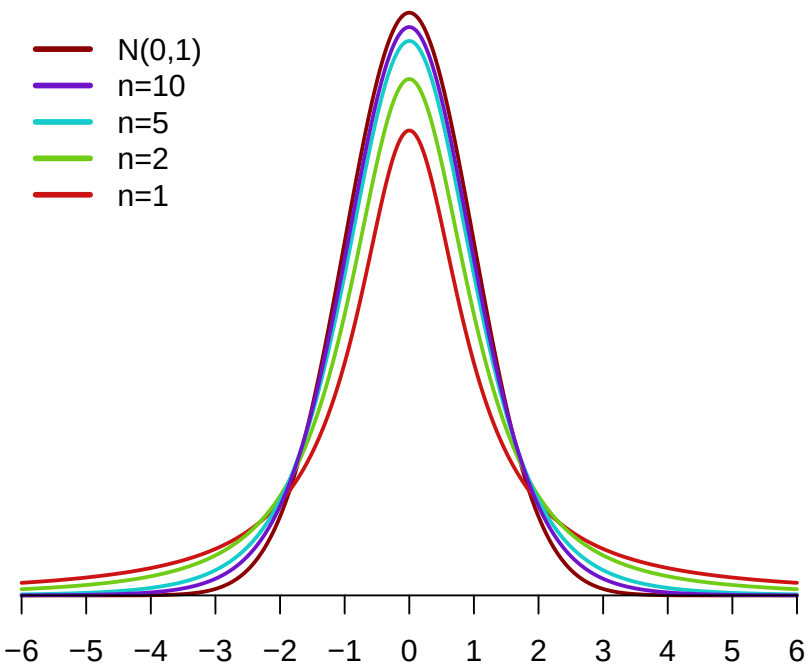
Student t

The Student t distribution was introduced by William Gossett (alias Student, 1876-1937), in charge of the brewing for Guinness to investigate the quality of beer productions comparing beer samples



Student t

The Student t distribution was introduced by William Gossett (alias Student, 1876-1937), in charge of the brewing for Guinness to investigate the quality of beer productions comparing beer samples



The Student t with $n \in \mathbb{N}$ degrees of freedom, t_n, has probability density function

$$f_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

It is bell-shaped as the Normal distribution, however the smaller the degrees of freedom n the heavier its tails

Student t table

Differently from the normal distribution table, the Student t table indicates the quantiles for different probability levels

On the rows, we have the degrees of freedom

On the columns, we have the probabilities

n	...	...	p	...
...	...	...	...	...
k	...	...	$t_{k,p}$	...
...	...	...	...	...

$t_{k,p}$ is the Student t quantile of order p with k degrees of freedom

	Probability					
n	0.9	0.95	0.975	0.990	0.995	0.999
1	3.08	6.31	12.71	31.82	63.66	318.30
2	1.89	2.92	4.30	6.96	9.93	22.33
3	1.64	2.35	3.18	4.54	5.84	10.21
4	1.53	2.13	2.78	3.75	4.60	7.17
5	1.48	2.02	2.57	3.37	4.03	5.89
6	1.44	1.94	2.45	3.14	3.71	5.21
7	1.42	1.90	2.37	3.00	3.50	4.79
8	1.40	1.86	2.31	2.90	3.35	4.50
9	1.38	1.83	2.26	2.82	3.25	4.30
10	1.37	1.81	2.23	2.76	3.17	4.14
11	1.36	1.80	2.20	2.72	3.11	4.03
12	1.36	1.78	2.18	2.68	3.06	3.93
13	1.35	1.77	2.16	2.65	3.01	3.85
14	1.34	1.76	2.14	2.62	2.98	3.79
15	1.34	1.75	2.13	2.60	2.95	3.73
16	1.34	1.75	2.12	2.58	2.92	3.69
17	1.33	1.74	2.11	2.57	2.90	3.65
18	1.33	1.73	2.10	2.55	2.88	3.61
19	1.33	1.73	2.09	2.54	2.86	3.58
20	1.32	1.73	2.09	2.53	2.85	3.55
21	1.32	1.72	2.08	2.52	2.83	3.53
22	1.32	1.72	2.07	2.51	2.82	3.50
23	1.32	1.71	2.07	2.50	2.81	3.48
24	1.32	1.71	2.06	2.49	2.80	3.47
25	1.32	1.71	2.06	2.48	2.79	3.45
26	1.31	1.71	2.06	2.48	2.78	3.44
27	1.31	1.70	2.05	2.47	2.77	3.42
28	1.31	1.70	2.05	2.47	2.76	3.41
29	1.31	1.70	2.04	2.46	2.76	3.40
30	1.31	1.70	2.04	2.46	2.75	3.38

Example

From a normal population $\mathcal{N}(\mu, \sigma^2)$ we extract an IID sample of size $n = 10$

$$(7.3, 4, 5.7, 6.4, 3.1, 4.3, 5.9, 4.3)$$

We compute the sample mean and unbiased sample variance

$$\bar{X} = 5.125; \quad S^2 = 2.002$$

Example

From a normal population $\mathcal{N}(\mu, \sigma^2)$ we extract an IID sample of size $n = 10$

$$(7.3, 4, 5.7, 6.4, 3.1, 4.3, 5.9, 4.3)$$

We compute the sample mean and unbiased sample variance

$$\bar{X} = 5.125; \quad S^2 = 2.002$$

To get an interval that contains the mean μ at 95%, we consider the quantile 0.975 of the Student t distribution with $n - 1 = 7$ degrees of freedom:

$$t_{7,0.975} = 2.37$$

then we obtain the interval

$$5.125 \pm 2.37 \cdot \sqrt{\frac{2.002}{8}} \to [3.93, 6.31]$$

CI for the proportion

From

$$\hat{\pi} \dot{\sim} \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

follows that

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \dot{\sim} \mathcal{N}(0,1)$$

then

$$P\left(\left|\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right| < z_{1-\alpha/2}\right) = 1 - \alpha$$

thus the confidence interval

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}},\ \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right]$$

contains, with probability $1-\alpha$ the parameter value π

CI for the proportion

From

$$\hat{\pi} \dot{\sim} \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

follows that

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \dot{\sim} \mathcal{N}(0,1)$$

then

$$P\left(\left|\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right| < z_{1-\alpha/2}\right) = 1 - \alpha$$

thus the confidence interval

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}},\ \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right]$$

contains, with probability $1 - \alpha$ the parameter value π

CI for the proportion

From

$$\hat{\pi} \dot{\sim} \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

follows that

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \dot{\sim} \mathcal{N}(0,1)$$

then

$$P\left(\left|\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right| < z_{1-\alpha/2}\right) = 1 - \alpha$$

thus the confidence interval

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\ \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

contains, with probability $1 - \alpha$ the parameter value π

From the CI to data: example 1

Assume you extracted a sample for a proportion π and a confidence interval of level 95% with bounds 0.227 and 0.493 has been observed, find $\hat{\pi}$ and the sample size n

From the CI to data: example 1

Assume you extracted a sample for a proportion π and a confidence interval of level 95% with bounds 0.227 and 0.493 has been observed, find $\hat{\pi}$ and the sample size n

Since we are dealing with a proportion, the formula we have to refer to is

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\ \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

the interval is centered around $\hat{\pi}$, thus

$$\hat{\pi} = \frac{1}{2}(0.227 + 0.493) = 0.36$$

From the CI to data: example 1

Assume you extracted a sample for a proportion π and a confidence interval of level 95% with bounds 0.227 and 0.493 has been observed, find $\hat{\pi}$ and the sample size n

Since we are dealing with a proportion, the formula we have to refer to is

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\ \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

To find the sample size we can notice that

$$2z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.493 - 0.227 = 0.266$$

$$2\times 1.96\sqrt{\frac{0.36(1-0.36)}{n}} = 2\times 1.96\frac{0.48}{\sqrt{n}} = 0.266 \Rightarrow n = 7.07^2 = 49.9849$$

From the CI to data: example 2

Assume you extracted a sample from a normal population with known variance and the CI at 95% for the mean μ has bounds 9.898 and 10.68. Find $\hat{\mu}$ and the sample variance

From the CI to data: example 2

Assume you extracted a sample from a normal population with known variance and the CI at 95% for the mean μ has bounds 9.898 and 10.68. Find $\hat{\mu}$ and the sample variance

Since it is a mean, the formula we have to refer to is

$$\left[\hat{\mu} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

the CI is centered around $\hat{\mu}$, thus

$$\hat{\mu} = \frac{1}{2}(9.898 + 10.68) = 10.29$$

From the CI to data: example 2

Assume you extracted a sample from a normal population with known variance and the CI at 95% for the mean μ has bounds 9.898 and 10.68. Find $\hat{\mu}$ and the sample variance

Since it is a mean, the formula we have to refer to is

$$\left[\hat{\mu} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

We can't find the sample size, indeed

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 10.68 - 9.898 = 0.782$$

from which

$$\frac{\sigma}{\sqrt{n}} = \frac{0.782}{2z_{1-\alpha/2}} = 0.1995 \Rightarrow n = \frac{\sigma^2}{0.0398}$$

From the CI to data: example 3

Assume you extracted a sample of size $n = 10$ from a normal population and the CI at 95% for the mean μ has bounds 9.26 and 12.52. What is the unbiased sample variance S^2?

From the CI to data: example 3

Assume you extracted a sample of size $n = 10$ from a normal population and the CI at 95% for the mean μ has bounds 9.26 and 12.52. What is the unbiased sample variance S^2?

Since it is a mean, the formula we have to refer to

$$\left[\hat{\mu} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \hat{\mu} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

To find S^2 we write

$$2t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} = 12.52 - 9.26 = 3.26$$

thus

$$s = \sqrt{n}\frac{3.26}{2t_{1-\alpha/2,n-1}} = \sqrt{10}\frac{3.26}{2 \times 2.262} = 2.279 \Rightarrow S^2 = 5.193$$

Example: newborn

We have data related to the city of Muggia, where over a $n = 85$ newborns $x = 38$ are males,

Example: newborn

We have data related to the city of Muggia, where over a $n = 85$ newborns $x = 38$ are males,
then we have

$$\hat{\pi} = \frac{38}{85} = 0.447$$

a CI at 95% is then

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

$$\left[0.447 - z_{0.975}\sqrt{\frac{0.447(1-0.447)}{85}}, 0.447 + z_{0.975}\sqrt{\frac{0.447(1-0.447)}{85}}\right]$$

$$[0.447 - z_{0.975}0.0539, 0.447 + z_{0.975}0.0539]$$

$$[0.341, 0.553]$$

Example: newborn

Data are related to the FVG region, where over $n = 10337$ newborns $x = 5286$ are males, then we have

$$\hat{\pi} = \frac{5286}{10337} = 0.511$$

a CI at 95% is then

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

$$\left[0.511 - z_{0.975}\sqrt{\frac{0.511(1-0.511)}{10337}}, 0.511 + z_{0.975}\sqrt{\frac{0.511(1-0.511)}{10337}}\right]$$

$$[0.511 - z_{0.975}0.00492, 0.511 + z_{0.975}0.00492]$$

$$[0.501, 0.521]$$

Example: newborn

The table below shows the CI for the probability that a male will be born, computed using the samples of Muggia, FVG and Italy

x	n	$\hat{\pi}$	$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	CI 95%		CI 98%	
38	85	0.447	0.053900	0.341	0.553	0.308	0.586
5286	10337	0.511	0.004920	0.501	0.521	0.498	0.524
289185	561944	0.515	0.000667	0.514	0.516	0.513	0.517

Recap: confidence intervals

Normal, σ known	$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
Normal, σ unknown	$\left[\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right]$
Proportion	$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	
1	Tail	

We lost, let's do another round

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	
1	Tail	
2	Tail	

We lost, let's do another round

We lost again, let's do another round

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	
1	Tail	We lost, let's do another round
2	Tail	We lost again, let's do another round
3	Tail	Oh no, again!

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	
1	Tail	We lost, let's do another round
2	Tail	We lost again, let's do another round
3	Tail	Oh no, again!
4	Tail	Another one

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	
1	Tail	We lost, let's do another round
2	Tail	We lost again, let's do another round
3	Tail	Oh no, again!
4	Tail	Another one
5	Tail	Oh, come on!

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome		
1	Tail		We lost, let's do another round
2	Tail		We lost again, let's do another round
3	Tail		Oh no, again!
4	Tail		Another one
5	Tail		Oh, come on!
6	Tail		
7	Tail		
8	Tail		
10	Tail		What!?!

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	
1	Tail	
2	Tail	
3	Tail	
4	Tail	
5	Tail	
6	Tail	
7	Tail	
8	Tail	
10	Tail	
11	Tail	
12	Tail	
13	Tail	
14	Tail	

We lost, let's do another round

We lost again, let's do another round

Oh no, again!

Another one

Oh, come on!

What!?!

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome		
1	Tail		We lost, let's do another round
2	Tail		We lost again, let's do another round
3	Tail		Oh no, again!
4	Tail		Another one
5	Tail		Oh, come on!
6	Tail		
7	Tail		
8	Tail		
10	Tail		What!?!
11	Tail		
12	Tail		
13	Tail		
14	Tail		
15	Tail		Maybe we should stop...

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome		
1	Tail		We lost, let's do another round
2	Tail		We lost again, let's do another round
3	Tail		Oh no, again!
4	Tail		Another one
5	Tail		Oh, come on!
6	Tail		
7	Tail		
8	Tail		
10	Tail		What!?!
11	Tail		
12	Tail		
13	Tail		
14	Tail		
15	Tail		Maybe we should stop...

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	Probability	
1	Tail	0.5	We lost, let's do another round
2	Tail	0.25	We lost again, let's do another round
3	Tail	0.125	Oh no, again!
4	Tail	0.062	Another one
5	Tail	0.031	Oh, come on!
6	Tail	...	
7	Tail	...	
8	Tail	...	
10	Tail	0.001	What!?!
11	Tail	...	
12	Tail	...	
13	Tail	...	
14	Tail	...	
15	Tail	3.0517578×10^{-5}	Maybe we should stop...

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	Probability	
1	Tail	0.5	Initially, we played because we believed that the coin was fair
2	Tail	0.25	
3	Tail	0.125	
4	Tail	0.062	
5	Tail	0.031	
6	Tail	...	
7	Tail	...	
8	Tail	...	
10	Tail	0.001	What!?!
11	Tail	...	
12	Tail	...	
13	Tail	...	
14	Tail	...	
15	Tail	3.0517578×10^{-5}	Maybe we should stop...

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	Probability
1	Tail	0.5
2	Tail	0.25
3	Tail	0.125
4	Tail	0.062
5	Tail	0.031
6	Tail	...
7	Tail	...
8	Tail	...
10	Tail	0.001
11	Tail	...
12	Tail	...
13	Tail	...
14	Tail	...
15	Tail	3.0517578×10^{-5}

Initially, we played because we believed that the coin was fair

At the first trial, we had no reason to think that the coin is not fair, since we assumed a success and a failure to be equally likely

What!?!

Maybe we should stop...

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	Probability
1	Tail	0.5
2	Tail	0.25
3	Tail	0.125
4	Tail	0.062
5	Tail	0.031
6	Tail	...
7	Tail	...
8	Tail	...
10	Tail	0.001
11	Tail	...
12	Tail	...
13	Tail	...
14	Tail	...
15	Tail	3.0517578×10^{-5}

Initially, we played because we believed that the coin was fair

At the first trial, we had no reason to think that the coin is not fair, since we assumed a success and a failure to be equally likely

As we go on, we suspect that there is something wrong: the sequence we observe is highly unlikely to be observed for a fair coin

Okay, it is enough: the coin is not far!

The logic behind hypothesis testing

We play a game in which we win 1 if a coin tosses head, and we lose 1 if a tail is thrown

Toss	Outcome	Probability
1	Tail	0.5
2	Tail	0.25
3	Tail	0.125
4	Tail	0.062
5	Tail	0.031
6	Tail	...
7	Tail	...
8	Tail	...
10	Tail	0.001
11	Tail	...
12	Tail	...
13	Tail	...
14	Tail	...
15	Tail	3.0517578×10^{-5}

Initially, we played because we believed that the coin was fair

At the first trial, we had no reason to think that the coin is not fair, since we assumed a success and a failure to be equally likely

As we go on, we suspect that there is something wrong: the sequence we observe is highly unlikely to be observed for a fair coin

We started playing under the hypothesis that the 'coin is fair', but after several trials, we reject this hypothesis

Null hypothesis

Assume we have a population with an associated parameter (that is, a parameter describes a characteristic of the distribution):

$$\theta : \text{parameter}; \quad (\text{probability, mean})$$

We make an hypothesis about the parameter value, that is called the null hypothesis H_0 and that can be of different types

- $H_0 : \theta = \theta_0$ (in the coin example was $\theta = 0.5$), simple hypothesis;
- $H_0 : \theta \geq \theta_0$ composite hypothesis;
- $H_0 : \theta \leq \theta_0$ composite hypothesis;

We observe a sample and our goal is to evaluate whether the null hypothesis H_0 is compatible with the observed sample or not

In other words, the sample observation may not support my hypothesis (as in the coin example)

Example: newborn in Muggia

We want to test the null hypothesis as *in the human population, the probability of a male newborn is 0.5*, i.e.

$$H_0 : \pi = \pi_0 = 0.5$$

Example: newborn in Muggia

We want to test the null hypothesis as *in the human population, the probability of a male newborn is 0.5*, i.e.

$$H_0 : \pi = \pi_0 = 0.5$$

To perform the hypothesis testing we use the data about newborns in Muggia, recall that we observed 38 males on 85 newborns, thus, $\hat{\pi} = 0.447$

We know that the sample proportion $\hat{\pi}$ follows approximately a normal distribution

$$\hat{\pi} \dot{\sim} \mathcal{N}(\pi, \pi(1-\pi)/n)$$

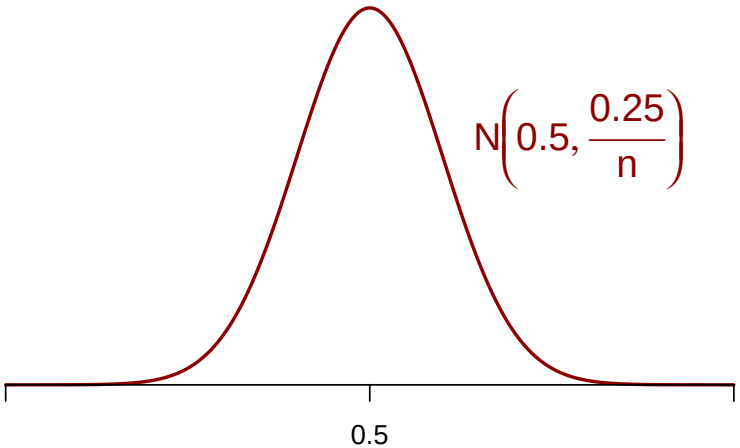
where $\pi =$ Probability of a male newborn

If the null hypothesis H_0 is true

$$\hat{\pi} \dot{\sim} \mathcal{N}(0.5, 0.25/n)$$

So we compare the evidence from the sample with the $\hat{\pi}$ distribution under the null hypothesis, that is, assuming that the null hypothesis is true

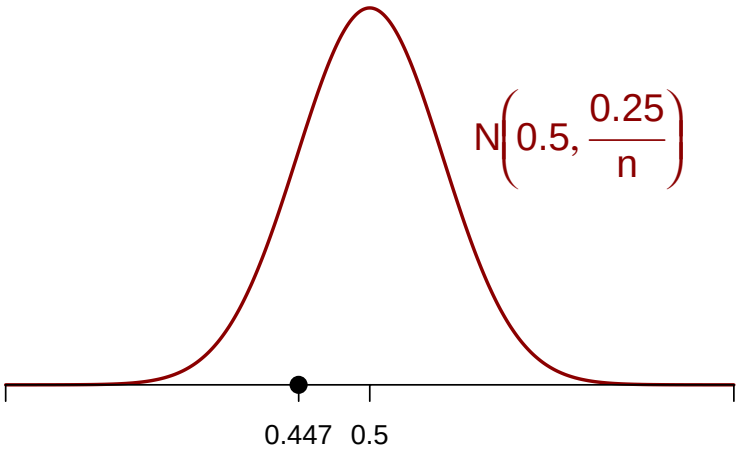
Example: newborn



If the null hypothesis is true, the sample proportion for a sample of size n follows

$$\hat{\pi} \dot{\sim} \mathcal{N}\left(0.5, \frac{0.25}{n}\right)$$

Example: newborn



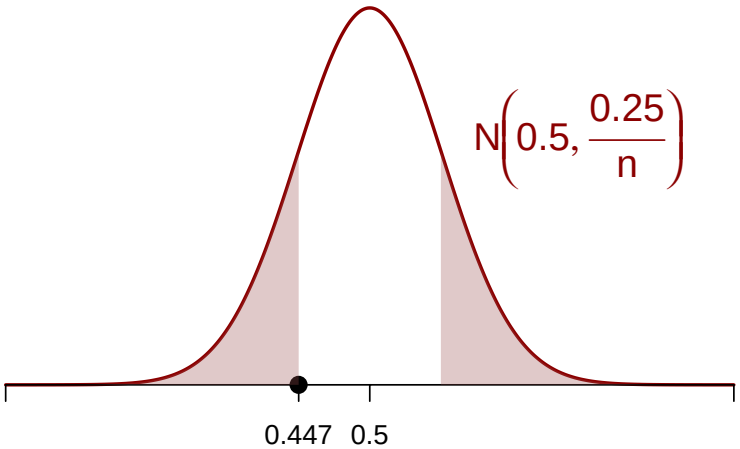
If the null hypothesis is true, the sample proportion for a sample of size n follows

$$\hat{\pi} \dot{\sim} \mathcal{N}\left(0.5, \frac{0.25}{n}\right)$$

We observed

$$\hat{\pi} = \frac{38}{85} \approx 0.447$$

Example: newborn



If the null hypothesis is true, the sample proportion for a sample of size n follows

$$\hat{\pi} \dot{\sim} \mathcal{N}\left(0.5, \frac{0.25}{n}\right)$$

We observed

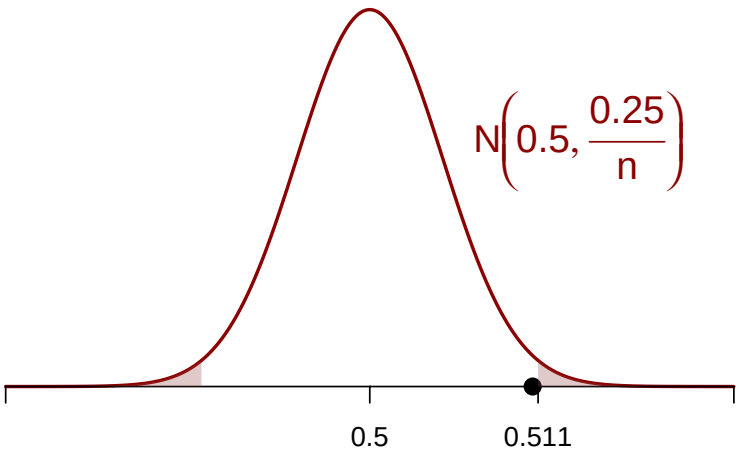
$$\hat{\pi} = \frac{38}{85} \approx 0.447$$

We measure how compatible this estimate is with the null hypothesis by computing the probability of observing the same deviation from 0.5 ($|\hat{\pi} - 0.5|$) assuming that the null hypothesis is true

$$P(|\hat{\pi} - 0.5| \geq |0.447 - 0.5|) = P\left(\left|\frac{\hat{\pi} - 0.5}{0.5/\sqrt{n}}\right| \geq \frac{0.0529}{0.5/\sqrt{n}}\right) = 2\left(1 - \Phi\left(0.976\right)\right) = 0.329$$

Since the probability is not small, we evaluate that the sample is compatible with the hypothesis made
In other words, the observations do not allow to exclude the validity of the hypothesis

Example: newborn in FVG



Now we repeat the steps above considering the same H_0 but a larger sample:
in the FVG region, we observed 5286 males over 10337 newborns, thus

$$\hat{\pi} = \frac{5286}{10337} \approx 0.511$$

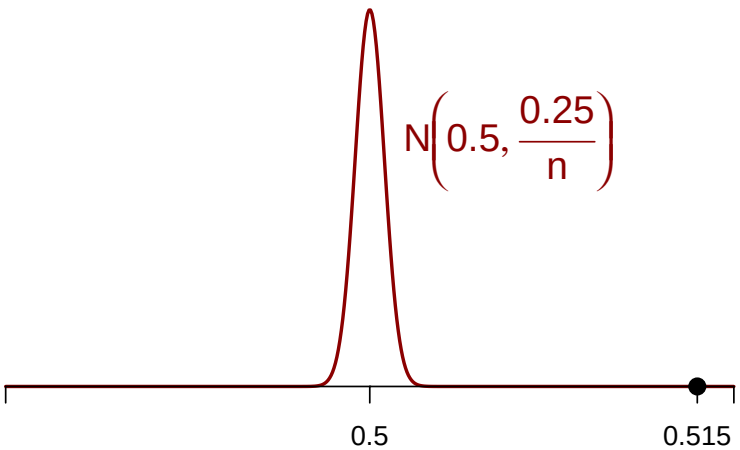
We compute the probability of observing the same deviation

$$P(|\hat{\pi} - 0.5| \geq |0.511 - 0.5|) = P\left(\left|\frac{\hat{\pi} - 0.5}{0.5/\sqrt{n}}\right| \geq \frac{0.0114}{0.5/\sqrt{n}}\right) = 2\left(1 - \Phi\left(2.31\right)\right) = 0.02081$$

The probability is now very small, we may bring the hypothesis into question

We could collect a larger sample

Example: newborn in Italy



We consider an even larger sample: in Italy, we observed 289185 males over 561944 newborns, thus

$$\hat{\pi} = \frac{289185}{561944} \approx 0.515$$

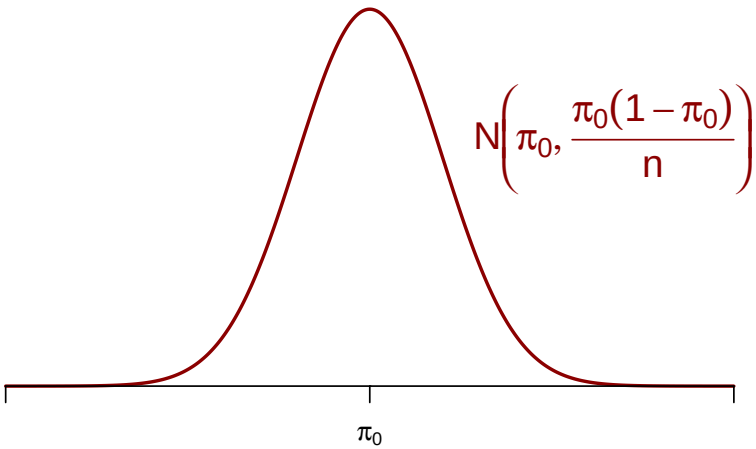
As before we compute

$$P(|\hat{\pi} - 0.5| \geq |0.5 - 0.515|) = P\left(\left|\frac{\hat{\pi} - 0.5}{0.5/\sqrt{n}}\right| \geq \frac{0.0146}{0.5/\sqrt{n}}\right) = 2\left(1 - \Phi\left(21.9\right)\right) \approx 0$$

The probability is extremely small ($1.9889688 \times 10^{-106}$, exactly) and the null hypothesis is rejected

Reference material about the F/M ratio: wikipedia.

Significance test for a proportion



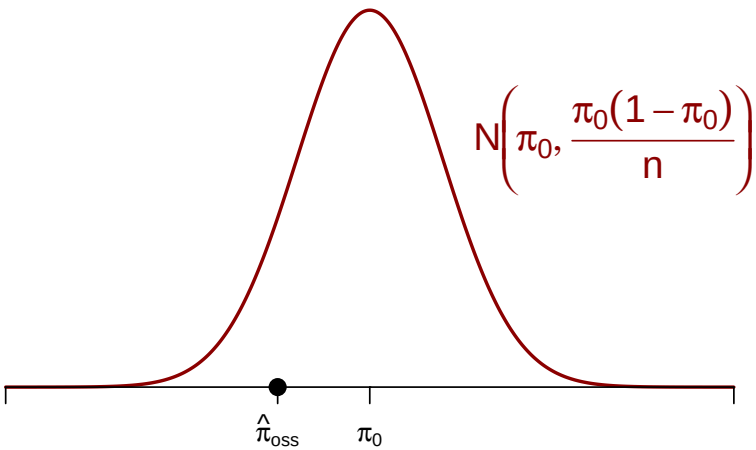
In general terms, for a null hypothesis as

$$H_0 : \pi = \pi_0$$

Under the assumption that the H_0 is true, the sample proportion computed on a sample of size n is approximately Normal

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

Significance test for a proportion



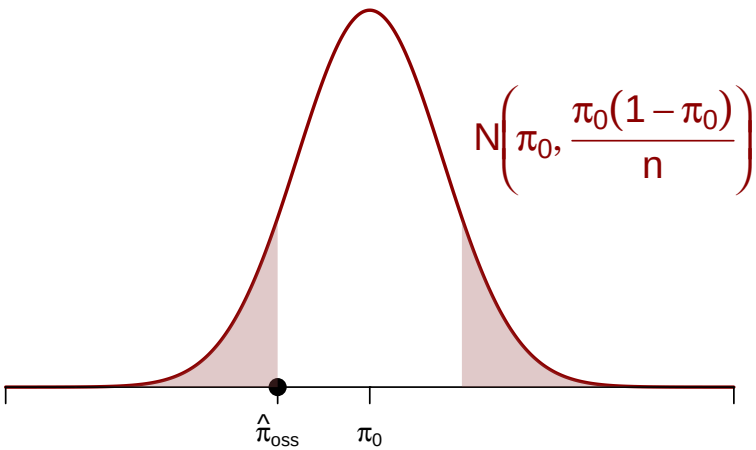
In general terms, for a null hypothesis as

$$H_0 : \pi = \pi_0$$

Under the assumption that the H_0 is true, the sample proportion computed on a sample of size n is approximately Normal

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

Significance test for a proportion



In general terms, for a null hypothesis as

$$H_0 : \pi = \pi_0$$

Under the assumption that the H_0 is true, the sample proportion computed on a sample of size n is approximately Normal

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

We observe $\hat{\pi} = \hat{\pi}_{\text{obs}}$ and we evaluate how compatible this is with the null hypothesis computing the probability of observing an equally large deviation from π_0 if the null hypothesis is true, called *p*-value

$$P(|D_0| > |d_0|) = P\left(\left|\frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right| > \left|\frac{\hat{\pi}_{\text{obs}} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right|\right) = 2\left(1 - \Phi\left(\left|\frac{\hat{\pi}_{\text{obs}} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right|\right)\right)$$

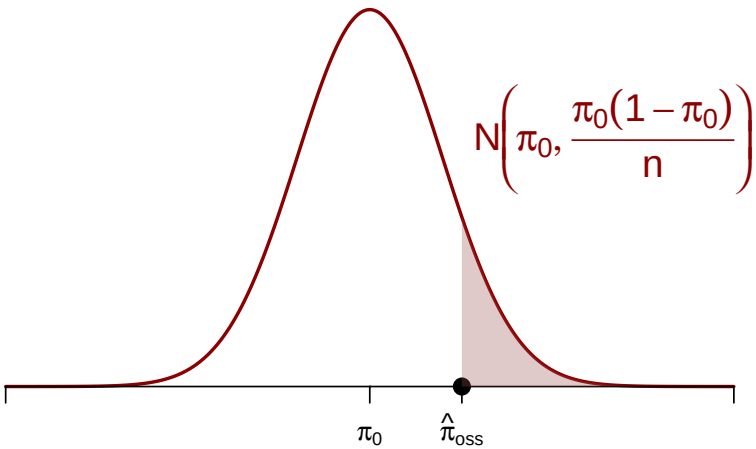
If this probability is small (that is the p-value is small), we conclude that we don't have evidence from the sample in favour of the null hypothesis

Generalization

We have several hypothesis types and several models

- Hypothesis:
  - simple $H_0 : \pi = \pi_0$
  - composite $H_0 : \pi \leq \pi_0$
  - composite $H_0 : \pi \geq \pi_0$
- Model:
  - proportion
  - mean of a Normal, variance known
  - mean of a Normal, variance unknown
- Then we move to the two populations' case

Significance test for a proportion: composite hp



If the null hypothesis is

$$H_0 : \pi \leq \pi_0$$

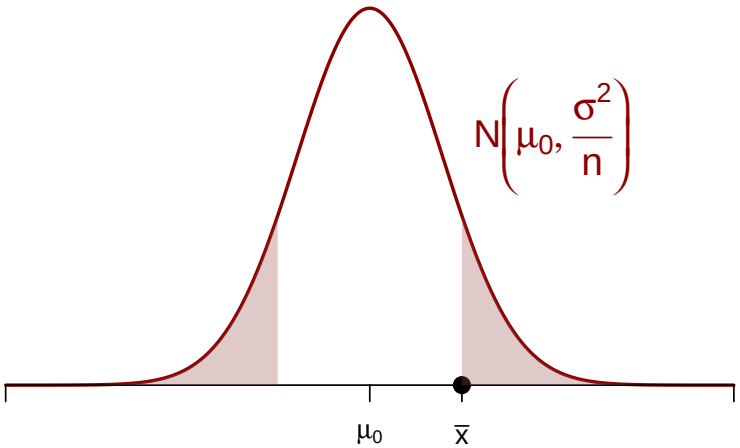
and we observe $\hat{\pi} = \hat{\pi}_{obs}$, we may reject H_0 if $\hat{\pi}_{obs}$ is greater than π_0

We measure how compatible this is with the hypothesis based on the probability of observing a larger deviation from π_0 under the null hypothesis

$$P(D_0 > d_0) = P\left(\frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > \frac{\hat{\pi}_{obs} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right) = 1 - \Phi\left(\frac{\hat{\pi}_{obs} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right)$$

What is the setting for $H_0 : \pi \geq \pi_0$?

Significance test for the Normal [variance known]



If the null hypothesis is

$$H_0 : \mu = \mu_0$$

the sample mean distribution follows a Normal, then

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

The measure of the deviation between the null hypothesis and the observation $\bar{x}$ is

$$P(|D_0| > |d_0|) = P\left(\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right) = 2\left(1 - \Phi\left(\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right)\right)$$

What if $H_0 : \mu \leq \mu_0$ or $H_0 : \mu \geq \mu_0$?

Significance test for the Normal [variance unknown]

If the null hypothesis is

$$H_0 : \mu = \mu_0$$

the quantity to evaluate is

$$D_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

Let $\bar{x}$ be the estimate, the deviation measure between the null hypothesis and estimate is

$$P(|D_0| > |d_0|) = P\left(\left|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right| > \left|\frac{\bar{x} - \mu_0}{S/\sqrt{n}}\right|\right)$$

and we compute it using the cumulative distribution function of the Student t distribution with $n - 1$ degrees of freedom

Summary: significance test

The significance test procedure is

a. set a null hypothesis
b. collect a sample
c. evaluate the deviation between the sample and the hypothesis computing the probability of observing a sample as far as the observed one assuming the null hypothesis is true

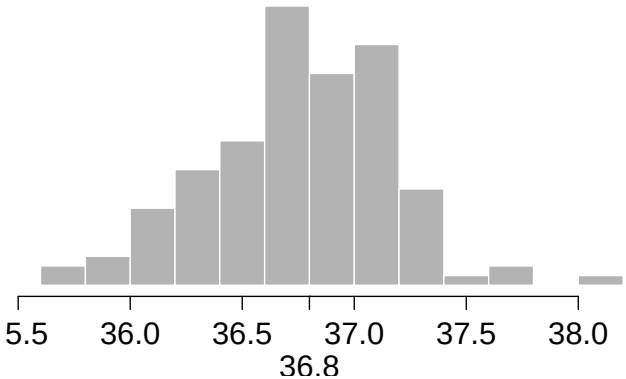
This measure is called *p*-value and tells us how similar the observed sample is to the null hypothesis

Summary: p-value

			Null hypothesis		
		d_0	$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$
Normal	σ known	$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$2\left(1-\Phi\left(\lvert d_0\rvert\right)\right)$	$1-\Phi\left(d_0\right)$	$\Phi\left(d_0\right)$
	σ unknown	$\frac{\bar{x}-\mu_0}{S/\sqrt{n}}$	$2\left(1-F_{t_{n-1}}\left(\lvert d_0\rvert\right)\right)$	$1-F_{t_{n-1}}\left(d_0\right)$	$F_{t_{n-1}}\left(d_0\right)$
			$H_0 : \pi = \pi_0$	$H_0 : \pi \leq \pi_0$	$H_0 : \pi \geq \pi_0$
Proportion		$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	$2\left(1-\Phi\left(\lvert d_0\rvert\right)\right)$	$1-\Phi\left(d_0\right)$	$\Phi\left(d_0\right)$

With $F_{t_{n-1}}$ is the cumulative distribution function of the Student t distribution with $n-1$ degrees of freedom

Example: body temperature

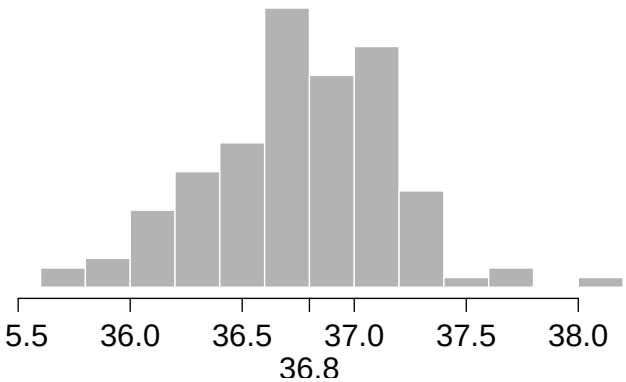


$$n = 130$$

$$\bar{x} = 36.8$$

$$S^2 = 0.166$$

Example: body temperature



$$n = 130$$

$$\bar{x} = 36.8$$

$$S^2 = 0.166$$

The null hypothesis is

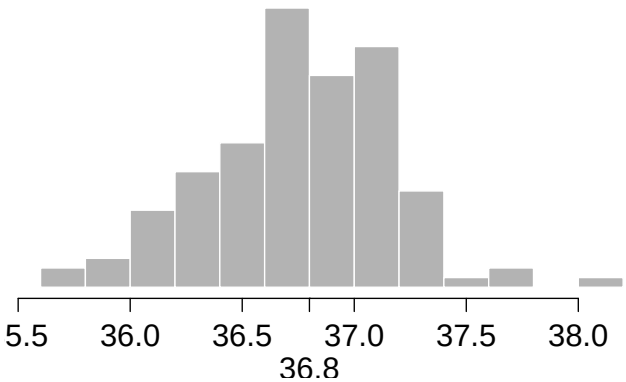
$$H_0 : \mu = 37$$

the variance is unknown, we evaluate the quantity

$$D_0 = \frac{\bar{x} - 37}{S/\sqrt{n}} = \frac{36.8 - 37}{0.407/\sqrt{130}} = \frac{36.8 - 37}{0.036} = -5.556$$

We refer to the Student t with 129 degrees of freedom, which is very close to the standard normal

Example: body temperature



$$n = 130$$

$$\bar{x} = 36.8$$

$$S^2 = 0.166$$

The null hypothesis is

$$H_0 : \mu = 37$$

the variance is unknown, we evaluate the quantity

$$D_0 = \frac{\bar{x} - 37}{S/\sqrt{n}} = \frac{36.8 - 37}{0.407/\sqrt{130}} = \frac{36.8 - 37}{0.036} = -5.556$$

We refer to the Student t with 129 degrees of freedom, which is very close to the standard normal

Therefore we compute the p-value through the cumulative distribution function of the standard normal distribution

$$2\left(1 - \Phi\left(\left|\frac{\bar{x} - 37}{S/\sqrt{n}}\right|\right)\right) = 2(1 - \Phi(5.556)) \approx 0$$

Neyman-Pearson approach

The Neyman-Pearson approach refers to a system of hypotheses

- null hypothesis, H_0
- alternative hypothesis, H_1

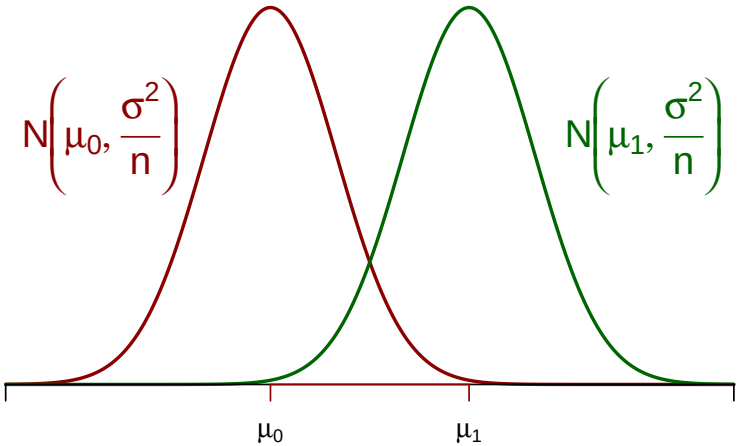
Before observing the sample, we want to formulate a rule according to which accept or reject the null hypothesis once the sample is extracted

The rule is a set of sample values $\mathcal{R}$, called rejection region so

- if $X \in \mathcal{R}$ the null hypothesis is rejected;
- if $X \notin \mathcal{R}$ the null hypothesis is accepted;

The complementary of the rejection region is also called the acceptance region

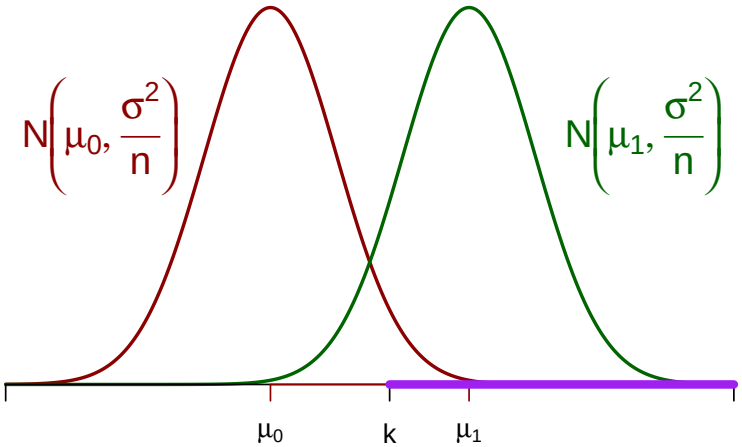
Rejection region: mean with known variance



Assume the following system of hypotheses

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

Rejection region: mean with known variance



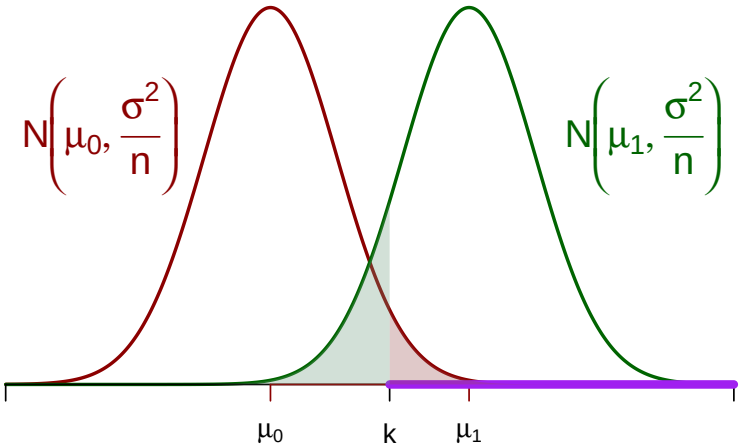
Assume the following system of hypotheses

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

and consider the rejection region

$$\mathcal{R} = \{\bar{X} > k\}$$

Rejection region: mean with known variance



Assume the following system of hypotheses

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

and consider the rejection region

$$\mathcal{R} = \{\bar{X} > k\}$$

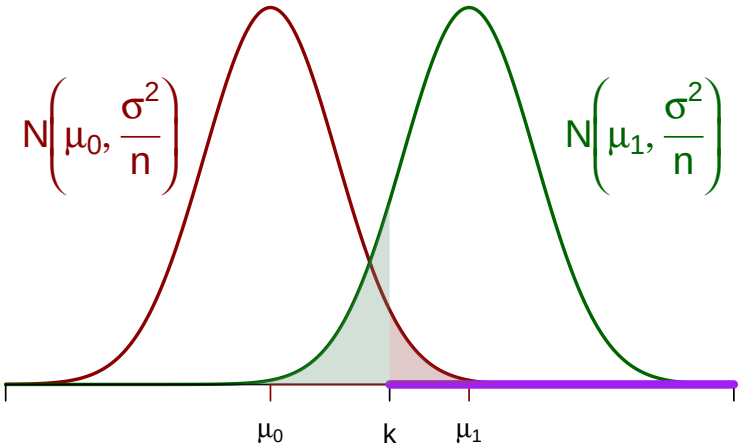
To choose k we refer to the probability to wrongly reject H_0 when it is true

		States of nature	
		H_0 true	H_0 false
Decision on H_0	Fail to reject: $\bar{X} \notin \mathcal{R}$	Correct	Type II error prob
	Reject: $\bar{X} \in \mathcal{R}$	Type I error prob	Correct

In the picture above

- red: $P(\bar{X} \in \mathcal{R}; H_0) = P(\bar{X} > k; H_0) = P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{k-\mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{k-\mu_0}{\sigma/\sqrt{n}}\right)$
- green: $P(\bar{X} \notin \mathcal{R}; H_1) = P(\bar{X} \leq k; H_1) = P\left(\frac{\bar{X}-\mu_1}{\sigma/\sqrt{n}} \leq \frac{k-\mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{k-\mu_1}{\sigma/\sqrt{n}}\right)$

Rejection region: mean with known variance



Assume the following system of hypotheses

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

and consider the rejection region

$$\mathcal{R} = \{\bar{X} > k\}$$

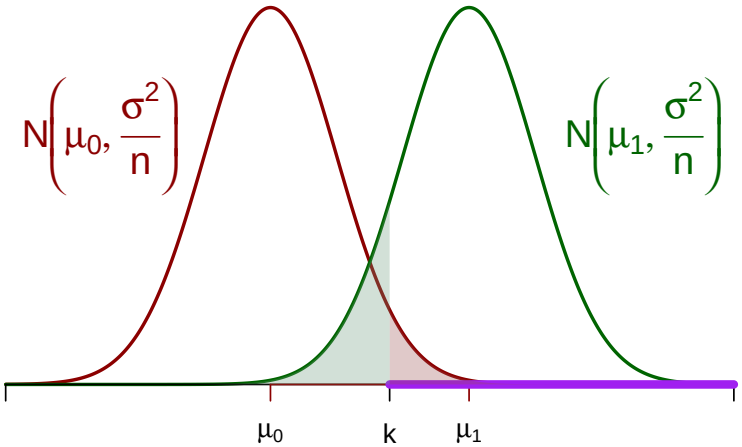
To choose k we refer to the probability to wrongly reject H_0 when it is true

		States of nature	
		H_0 true	H_0 false
Decision on H_0	Fail to reject: $\bar{X} \notin \mathcal{R}$	Correct	$\Phi\left(\frac{k-\mu_1}{\sigma/\sqrt{n}}\right)$
	Reject: $\bar{X} \in \mathcal{R}$	$1 - \Phi\left(\frac{k-\mu_0}{\sigma/\sqrt{n}}\right)$	Correct

Ideally, we would like to achieve low type I and type II error probabilities, but when k increases

- the Type I error probability decreases
- the Type II error probability increases

Rejection region: mean with known variance



Assume the following system of hypotheses

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

and consider the rejection region

$$\mathcal{R} = \{\bar{X} > k\}$$

The k value is determined by fixing the Type I error probability (wrongly reject H_0 when it is true)

$$\alpha = 1 - \Phi\left(\frac{k - \mu_0}{\sigma/\sqrt{n}}\right) \quad \rightarrow \quad k = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

The rejection region with Type I error probability at most equal to α is called the region of level α

The complement of the Type II error probability is called power of the test

$$1 - P(X \in \mathcal{R}; H_1) = 1 - \Phi\left(\frac{k - \mu_1}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right)$$

Hypothesis testing, summary

To test a hypothesis system with the Neyman-Pearson paradigm

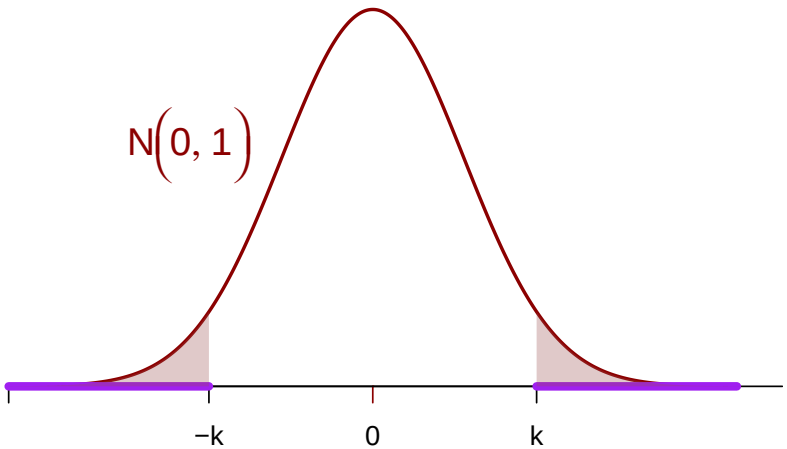
(1) Formulate the two hypotheses
(2) Determine the rejection region
    - In the cases that we explore it will be natural
(3) Choose a particular rejection region such that the Type I error probability is α
    - The probability of type II error comes as a consequence, essentially the choice is made to keep Type I error under control (the idea is that it is " more serious " of type II).
    - Note that to obtain the rejection region we refer to the sample distribution under the null hypothesis (the alternative hypothesis is important to determine the Type II error probability)

Rejection regions: two-sided case

Assume the following two-sided alternative hypothesis

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Rejection regions: two-sided case



Assume the following two-sided alternative hypothesis

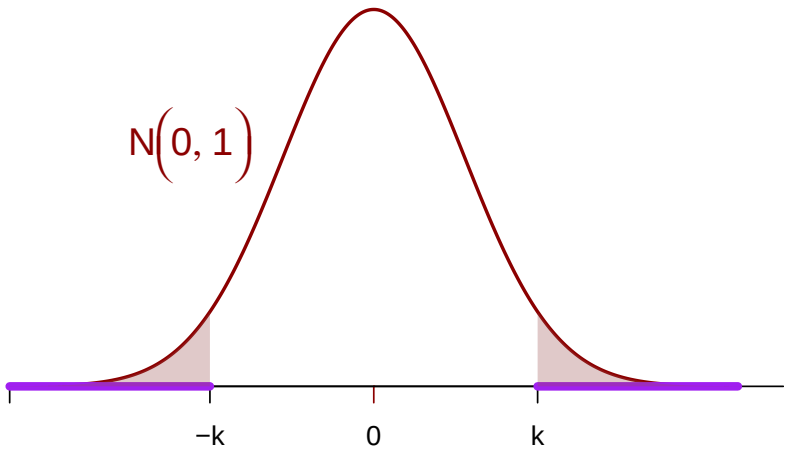
$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

It is easy if we refer to the quantity

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

that, under the null hypothesis H_0 follows a standard normal distribution,

Rejection regions: two-sided case



Assume the following two-sided alternative hypothesis

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

It is easy if we refer to the quantity

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

that, under the null hypothesis H_0 follows a standard normal distribution, Consider the rejection region

$$\mathcal{R} = \{|D_0| > k\} = \left\{ \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > k > 0 \right\}$$

The Type I error probability is $2(1 - \Phi(k))$, then

$$\alpha = 2(1 - \Phi(k)) \Leftrightarrow k = z_{1-\alpha/2}$$

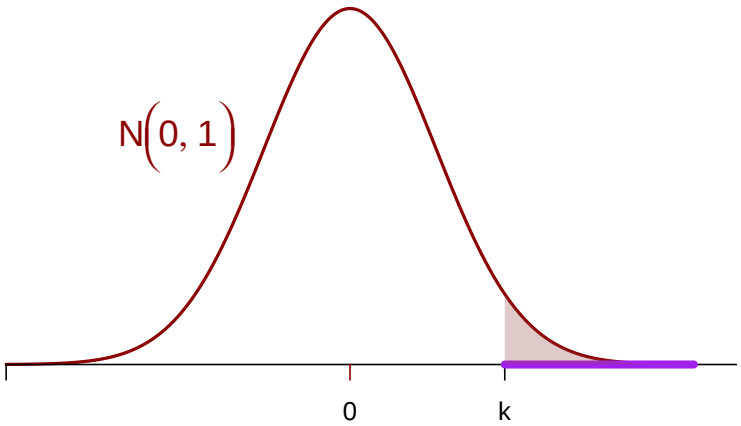
Rejection regions: one-sided case

Assume the following one-sided alternative hypothesis

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

In this case, with a composite null hypothesis, we don't know the distribution of $\bar{X}$ under the null hypothesis

Rejection regions: one-sided case



Assume the following one-sided alternative hypothesis

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

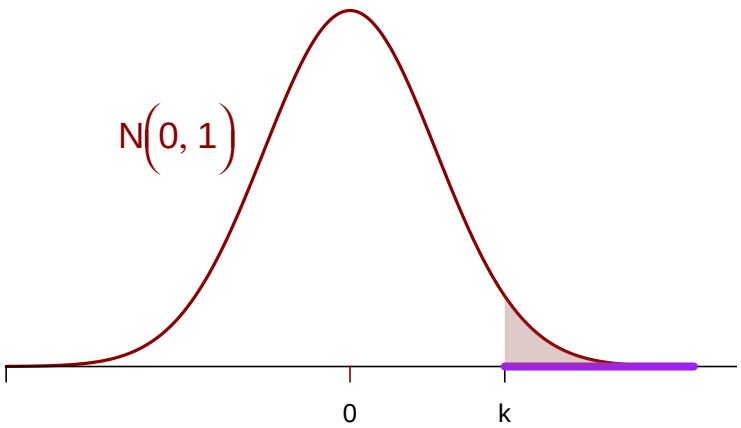
In this case, with a composite null hypothesis, we don't know the distribution of $\bar{X}$ under the null hypothesis

but again we refer to the quantity

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

that, if $\mu = \mu_0$, follows a standard normal distribution.

Rejection regions: one-sided case



Assume the following one-sided alternative hypothesis

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

In this case, with a composite null hypothesis, we don't know the distribution of $\bar{X}$ under the null hypothesis

but again we refer to the quantity

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

that, if $\mu = \mu_0$, follows a standard normal distribution.
Consider the rejection region

$$\mathcal{R} = \{D_0 > k\} = \left\{ \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > k > 0 \right\}$$

For $k = z_{1-\alpha}$, the type I error probability, which we cannot compute exactly in this case, is not greater than α:

$$\text{if } \mu \leq \mu_0 \text{ then } P(D_0 > z_{1-\alpha}) \leq \alpha$$

it can be easily proved that if $\mu \geq \mu_0$ then $P(D_0 < -z_{1-\alpha}) \leq \alpha$

Rejection regions: proportion

In dealing with a proportion, we can use the approximate result that

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

from which we can compute the rejection regions

Hypotheses		
$H_0 : \pi = \pi_0$ $H_1 : \pi \neq \pi_0$	$H_0 : \pi \leq \pi_0$ $H_1 : \pi > \pi_0$	$H_0 : \pi \geq \pi_0$ $H_1 : \pi < \pi_0$
$\left\lvert \frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \right\rvert > z_{1-\alpha/2}$	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > z_{1-\alpha}$	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < -z_{1-\alpha}$

Rejection regions: mean with unknown variance

To perform a hypothesis testing on the mean with unknown variance, we use the result that if $\mu = \mu_0$ then it is approximately true that

$$D_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

then the rejection regions

Hypotheses		
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$
$\left\lvert \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \right\rvert > t_{n-1,1-\alpha/2}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{n-1,1-\alpha}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{n-1,1-\alpha}$

Rejection regions: summary

	Hypotheses		
	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$
Normal, σ known	$\left\lvert \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \right\rvert > z_{1-\alpha/2}$	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha}$
Normal, σ unknown	$\left\lvert \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \right\rvert > t_{n-1,1-\alpha/2}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{n-1,1-\alpha}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{n-1,1-\alpha}$
	$H_0 : \pi = \pi_0$ $H_1 : \pi \neq \pi_0$	$H_0 : \pi \leq \pi_0$ $H_1 : \pi > \pi_0$	$H_0 : \pi \geq \pi_0$ $H_1 : \pi < \pi_0$
Proportion	$\left\lvert \frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \right\rvert > z_{1-\alpha/2}$	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > z_{1-\alpha}$	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < -z_{1-\alpha}$

Example

A machine for canning beans produces packages whose weight is distributed according to a normal with mean μ and variance $\sigma^2 = 9$

The packages have declared weight 100 g, and this is how the machine is set, however with use the machine could produce with different mean, in which case it is necessary to suspend production and register the machine

To decide when it is necessary to register the machine, a sample of n packs is periodically extracted and weighed, if the average weight differs too much from 100, the production is suspended

To decide when the average weight is too different from 100, a 5% rejection region is used

Determine the region of rejection assuming that you observe samples of $n = 10$ packs

The hypothesis system is

$$H_0 : \mu = 100, \quad H_1 : \mu \neq 100$$

hence the rejection region is

$$\left| \frac{\bar{x} - 100}{3/\sqrt{10}} \right| > z_{0.975} = 1.96$$

in other words, the production is suspended if the average weight of the 10 packs falls outside the range

$$100 \pm 1.96 \times 0.948 \rightarrow [98.14, 101.86]$$

Note that this interval is the acceptance region

Assuming that the machine perfectly works, i.e. the mean is always equal to 100, how many times do we expect to (wrongly) suspend the production?

Because of the definition of the rejection region, we expect to be wrong 5 % of the time

If the machine produces packages with an average weight of 98, what is the probability that the production will be stopped?

Assuming that the machine perfectly works, i.e. the mean is always equal to 100, how many times do we expect to (wrongly) suspend the production?

Because of the definition of the rejection region, we expect to be wrong 5 % of the time

If the machine produces packages with an average weight of 98, what is the probability that the production will be stopped?

This is the probability that we reject H_0 observing a mean of 98

$$\Phi\left(\frac{98.14-98}{3/\sqrt{10}}\right)+\left(1-\Phi\left(\frac{101.86-98}{3/\sqrt{10}}\right)\right)\approx 0.559$$

How do the answers to the previous questions change if you look at $n = 30$ packs?

- the rejection region at 5 % is the complementary of

$$100 \pm 1.96 \times 3/\sqrt{30} = 100 \pm 1.96 \times 0.548 \rightarrow [98.93, 101.07]$$

- the frequency with which the machine stops if it works perfectly is always 5 %
- the probability of rejecting if the mean is 98 is

$$\Phi\left(\frac{98.93 - 98}{3/\sqrt{30}}\right) + \left(1 - \Phi\left(\frac{101.07 - 98}{3/\sqrt{30}}\right)\right) \approx 0.955$$

Note that, once the type I error probability is fixed, increasing the sample size decreases the probability of type II error, that is, the power of the test increases

Assuming the variance is unknown, determine the rejection region by supposing to observe samples of $n = 10$ packs
The hypotheses do not change

$$H_0 : \mu = 100, \quad H_1 : \mu \neq 100$$

however the rejection region is

$$\left| \frac{\bar{x} - 100}{S/\sqrt{10}} \right| > t_{9,0.975} = 2.26$$

For example, if a sample were observed with $S^2 = 9$, that is, the same value of the known variance in the previous example, the production would be suspended if the average weight of the 10 packages fell outside the range

$$100 \pm 2.26 \times 0.948 \rightarrow [97.86, 102.14]$$

this interval is wider than the one obtained in the case of known variance $\sigma^2 = 9$ (uncertainty is higher)

Example

Assume you observed a sample of size $n = 9$ from a Normal population with mean μ and variance $\sigma^2 = 4$

Determine a rejection region at level $\alpha = 10\%$ for the following hypotheses

$$H_0 : \mu = 1; \qquad H_1 : \mu \neq 1$$

We use the quantity:

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - 1}{2/3}$$

thus the rejection region for two-sided alternative hypotheses is

$$\mathcal{R} = \{|D_0| > k\} = \left\{ \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > k \right\}$$

Assuming that H_0 is true, $D_0 \sim N(0, 1)$, thus for a fixed level $\alpha = 0.1$

$$P(|D_0| > k; H_0) = \alpha \to k = z_{1-\alpha/2} = z_{0.95} = 1.64$$

$$
\begin{aligned}
R &= \{|D_0| > z_{1-\alpha/2}\} \\
&= \left\{ \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2} \right\} \\
&= \left\{ \left| \bar{X} - \mu_0 \right| > z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} \\
&= \left\{ \bar{X} > \mu_0 + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{X} < \mu_0 - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} \\
&= \left\{ \bar{X} > 1 + 1.64 \frac{2}{\sqrt{3}} \quad \text{or} \quad \bar{X} < 1 - 1.64 \frac{2}{\sqrt{3}} \right\} \\
&= \left\{ \bar{X} > 2.09 \quad \text{or} \quad \bar{X} < -0.09 \right\}
\end{aligned}
$$

Let the sample mean be $\bar{X} = 0$, what is the p-value for the hypothesis $H_0 : \mu = 1$?

We use the quantity:

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{0-1}{2/3} = -1.5$$

Under the null hypothesis, $D_0 \sim \mathcal{N}(0,1)$, what is the probability of observing a value as extreme as the observed one?

$$P(|D_0| > 1.5) = P(D_0 < -1.5 \cup D_0 > 1.5)$$

$$\begin{aligned}
\text{p-value} &= P(|D_0| > 1.5) \\
&= P(D_0 < -1.5 \cup D_0 > 1.5) \\
&= P(D_0 < -1.5) + P(D_0 > 1.5) \\
&= 2P(D_0 > 1.5) \\
&= 2(1 - \Phi(1.5)) = 0.133
\end{aligned}$$

Example: cats

In a cat population a percentage π has black fur. We observe a sample of size $n = 100$.

Compute the rejection region at 5% for the hypotheses

$$H_0 : \pi \leq 0.3; \qquad H_1 : \pi > 0.3$$

We use the quantity:

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{\hat{\pi} - 0.3}{\sqrt{\frac{0.3(1-0.3)}{100}}}$$

thus the rejection region for one-sided alternative hypotheses is

$$\mathcal{R} = \{D_0 > k\} = \left\{ \frac{\hat{\pi} - 0.3}{\sqrt{\frac{0.3(1-0.3)}{100}}} > k \right\}$$

Assuming that H_0 is true, $D_0 \sim N(0, 1)$, thus for a fixed level $\alpha = 0.05$

$$P(D_0 > k; H_0) = \alpha \rightarrow k = z_{1-\alpha} = z_{0.95} = 1.64$$

Thus, the rejection region at 5% is

$$R = \{\hat{\pi} > 0.3 + 1.64 \cdot 0.046 = 0.375\}$$

Compute the p-value for the null hypothesis $H_0 : \pi \leq 0.3$, given that $\hat{\pi} = 0.4$

$$P(\hat{\pi} > 0.4; \pi = 0.3) = 1 - \Phi\left(\frac{0.4 - 0.3}{\sqrt{0.3 \cdot 0.7/100}}\right) = 1 - \Phi(2.18) = 0.015$$

If $\hat{\pi} = 0.4$, what is the interval at 95%?

$$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{100}}$$

$$\left[0.4 \pm 1.96\sqrt{\frac{0.4 \cdot 0.6}{100}}\right] = [0.3; 0.5]$$

Hypothesis testing and CI

The three procedures

- p-value
- Neyman-Pearson hypothesis testing (acceptance and rejection region)
- confidence interval

are linked since

- they are all based on the repeated sampling principle
- they can be derived from the same quantity, that is, the pivotal quantity

Pivotal quantity

Given

- a parameter θ
- a sample $X_1, \ldots, X_n \sim \text{IID}(f(x;\theta))$

we define an estimator

$$\hat{\theta} = g(X_1, \ldots, X_n)$$

that follows a certain sampling distribution

A pivotal quantity is a function of the sample and of the parameter, whose distribution does not depend on the parameter (sometimes approximately)

$$D = \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}} \sim F \quad (\mathcal{N}(0,1) \text{ o } t_{n-1})$$

Pivotal quantity

$$D = \frac{\left[\begin{array}{c}\text{Estimator}\\ \text{(from the sample)}\end{array}\right] - \left[\begin{array}{c}\text{Parameter}\\ \text{(of the population)}\end{array}\right]}{\sqrt{\left[\begin{array}{c}\text{Variance}\\ \text{estimator}\end{array}\right]}} = \begin{cases} \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \\ \\ \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \\ \\ \frac{\hat{\pi}-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim \mathcal{N}(0,1) \end{cases}$$

Confidence interval

We call $100(1-\alpha)\%$ confidence interval of the parameter θ an interval with random bounds $\hat{\theta}_1$, $\hat{\theta}_2$ (functions of the sample), such that

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$$

(according to the sampling distribution)

The quantity $100(1-\alpha)\%$ is called the confidence level of the interval

Pivotal quantity $\rightarrow$ Confidence interval

We call $100(1-\alpha)\%$ confidence interval of the parameter θ an interval with random bounds $\hat{\theta}_1$, $\hat{\theta}_2$ (functions of the sample), such that

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$$

(according to the sampling distribution)

From the pivotal quantity, we can compute the confidence interval

$$P\left(F^{-1}(\alpha/2) \leq \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}} \leq F^{-1}(1-\alpha/2) \right) = 1 - \alpha$$

Pivotal quantity $\rightarrow$ Confidence interval

We call $100(1-\alpha)\%$ confidence interval of the parameter θ an interval with random bounds $\hat{\theta}_1$, $\hat{\theta}_2$ (functions of the sample), such that

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$$

(according to the sampling distribution)

From the pivotal quantity, we can compute the confidence interval

$$P\left(F^{-1}(\alpha/2) \leq \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}} \leq F^{-1}(1-\alpha/2) \right) = 1 - \alpha$$

then

$$P\left(\hat{\theta} + F^{-1}(\alpha/2)\sqrt{\hat{V}(\hat{\theta})} \leq \theta \leq \hat{\theta} + F^{-1}(1-\alpha/2)\sqrt{\hat{V}(\hat{\theta})} \right) = 1 - \alpha$$

Pivotal quantity $\rightarrow$ Confidence interval

We call $100(1-\alpha)\%$ confidence interval of the parameter θ an interval with random bounds $\hat{\theta}_1$, $\hat{\theta}_2$ (functions of the sample), such that

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$$

(according to the sampling distribution)

From the pivotal quantity, we can compute the confidence interval

$$P\left(F^{-1}(\alpha/2) \leq \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}} \leq F^{-1}(1-\alpha/2) \right) = 1 - \alpha$$

then

$$P\left(\hat{\theta} + F^{-1}(\alpha/2)\sqrt{\hat{V}(\hat{\theta})} \leq \theta \leq \hat{\theta} + F^{-1}(1-\alpha/2)\sqrt{\hat{V}(\hat{\theta})} \right) = 1 - \alpha$$

and the confidence interval of level $1-\alpha$

$$\left[\hat{\theta} + F^{-1}(\alpha/2)\sqrt{\hat{V}(\hat{\theta})},\ \ \hat{\theta} + F^{-1}(1-\alpha/2)\sqrt{\hat{V}(\hat{\theta})} \right]$$

Pivotal quantity $\rightarrow$ Confidence interval

From the pivotal quantity, we can compute the confidence interval

$$P\left(F^{-1}(\alpha/2) \leq \frac{\hat{\theta}-\theta}{\sqrt{\hat{V}(\hat{\theta})}} \leq F^{-1}(1-\alpha/2)\right) = 1-\alpha$$

and the confidence interval of level $1-\alpha$

$$\left[\hat{\theta} + F^{-1}(\alpha/2)\sqrt{\hat{V}(\hat{\theta})},\ \ \hat{\theta} + F^{-1}(1-\alpha/2)\sqrt{\hat{V}(\hat{\theta})}\right]$$

If the population is repeatedly sampled a very large number of times (obtaining several samples and thus several intervals), a proportion $100(1-\alpha)\%$ of the computed CIs would contain the true parameter value θ

Rejection region A rejection region of level α is a subset of the $\mathcal{R}_\alpha$ sampling space such that

$$P_{H_0}(X \in \mathcal{R}_\alpha) \leq \alpha$$

Pivotal quantity $\rightarrow$ Rejection region A rejection region of level α is a subset of the $\mathcal{R}_\alpha$ sampling space such that

$$P_{H_0}(X \in \mathcal{R}_\alpha) \leq \alpha$$

As an example, from the following hypotheses

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$

we can compute a region $\mathcal{R}$ of level α from the pivotal quantity
Since F is symmetric around 0

$$P_{H_0}\left(\left|\frac{\hat{\theta} - \theta_0}{\sqrt{\hat{V}(\hat{\theta})}}\right| > F^{-1}(1 - \alpha/2)\right) = \alpha$$

Pivotal quantity $\rightarrow$ Rejection region A rejection region of level α is a subset of the $\mathcal{R}_\alpha$ sampling space such that

$$P_{H_0}(X \in \mathcal{R}_\alpha) \leq \alpha$$

As an example, from the following hypotheses

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$

we can compute a region $\mathcal{R}$ of level α from the pivotal quantity
Since F is symmetric around 0

$$P_{H_0}\left(\left|\frac{\hat{\theta} - \theta_0}{\sqrt{\hat{V}(\hat{\theta})}}\right| > F^{-1}(1 - \alpha/2)\right) = \alpha$$

If the population is repeatedly sampled a very large number of times (obtaining several samples and thus several $\hat{\theta}$), and under the assumption that the null hypothesis is true, for a proportion at most equal to $\alpha\%$, the value $\hat{\theta}$ would fall into the rejection region, that is, we would be wrong in rejecting the null hypothesis $\alpha\%$ of times (Type I error)

Confidence intervals $\leftrightarrow$ Rejection regions

There exists a link between CIs and rejection regions for two-sided hypothesis

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$

Given θ_0, the acceptance region of a hypothesis test of level α includes the values of $\hat{\theta}$ such that

$$-F^{-1}(1-\alpha/2) \leq \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{V}(\hat{\theta})}} \leq F^{-1}(1-\alpha/2)$$

however, for a fixed $\hat{\theta}$, this is the condition that defines the CI of level $100(1-\alpha)\%$, i.e.

for a fixed $\hat{\theta}$, the CI of level $100(1-\alpha)\%$ includes those values of θ that would be accepted in a two-sided hypothesis test of level α.

Example

The production manager of Circuits Unlimited has asked for your assistance in analyzing a production process

This process involves drilling holes whose diameters are normally distributed with a population standard deviation of 0.06 inch

A random sample of 9 measurements had a sample mean of 1.95 inches

Determine the 95% confidence interval of the parameter μ

$$\left[\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[1.95 \pm 1.96\frac{0.06}{3}\right] = [1.818; 2.082]$$

For a fixed $\bar{x} = 1.95$, the CI contains all the values of μ that I would fail to reject at level 95% considering a two-sided alternative hypothesis e.g.

$$H_0 : \mu = 2; \quad H_1 \neq 2 \rightarrow R = \left\{\bar{x} > 2 + 1.96\frac{0.06}{3} \text{ or } \bar{x} < 2 - 1.96\frac{0.06}{3}\right\}$$

$$H_0 : \mu = 1; \quad H_1 \neq 1 \rightarrow R = \left\{\bar{x} > 1 + 1.96\frac{0.06}{3} \text{ or } \bar{x} < 1 - 1.96\frac{0.06}{3}\right\}$$

p-value

The $p-$ value under the null hypothesis is the probability of observing a value at least as extreme as the observed one if the null hypothesis is true

p-value

The $p-$ value under the null hypothesis is the probability of observing a value at least as extreme as the observed one if the null hypothesis is true

Under the null hypothesis

$$H_0 : \theta = \theta_0$$

if $\hat{\theta}_{\text{obs}}$ is the estimate of $\hat{\theta}$, the $p-$value is

$$P_{H_0}\left(\left|\frac{\hat{\theta}-\theta}{\sqrt{\hat{V}(\hat{\theta})}}\right| > \left|\frac{\hat{\theta}_{\text{obs}}-\theta}{\sqrt{\hat{V}(\hat{\theta})}}\right|\right) = 2\left(1 - F\left(\left|\frac{\hat{\theta}_{\text{obs}}-\theta}{\sqrt{\hat{V}(\hat{\theta})}}\right|\right)\right)$$

$p-$value $\leftrightarrow$ Rejection regions

For the hypotheses

$$H_0 : \theta = \theta_0; \quad H_1 : \theta \neq \theta_0$$

the rejection region is determined choosing d_α such that

$$P_{H_0}(|D| > d_\alpha) = \alpha \quad \left(\Rightarrow d_\alpha = F^{-1}(1 - \alpha/2)\right)$$

However, the $p-$value under $H_0 : \theta = \theta_0$ is (by symmetry around 0 of F)

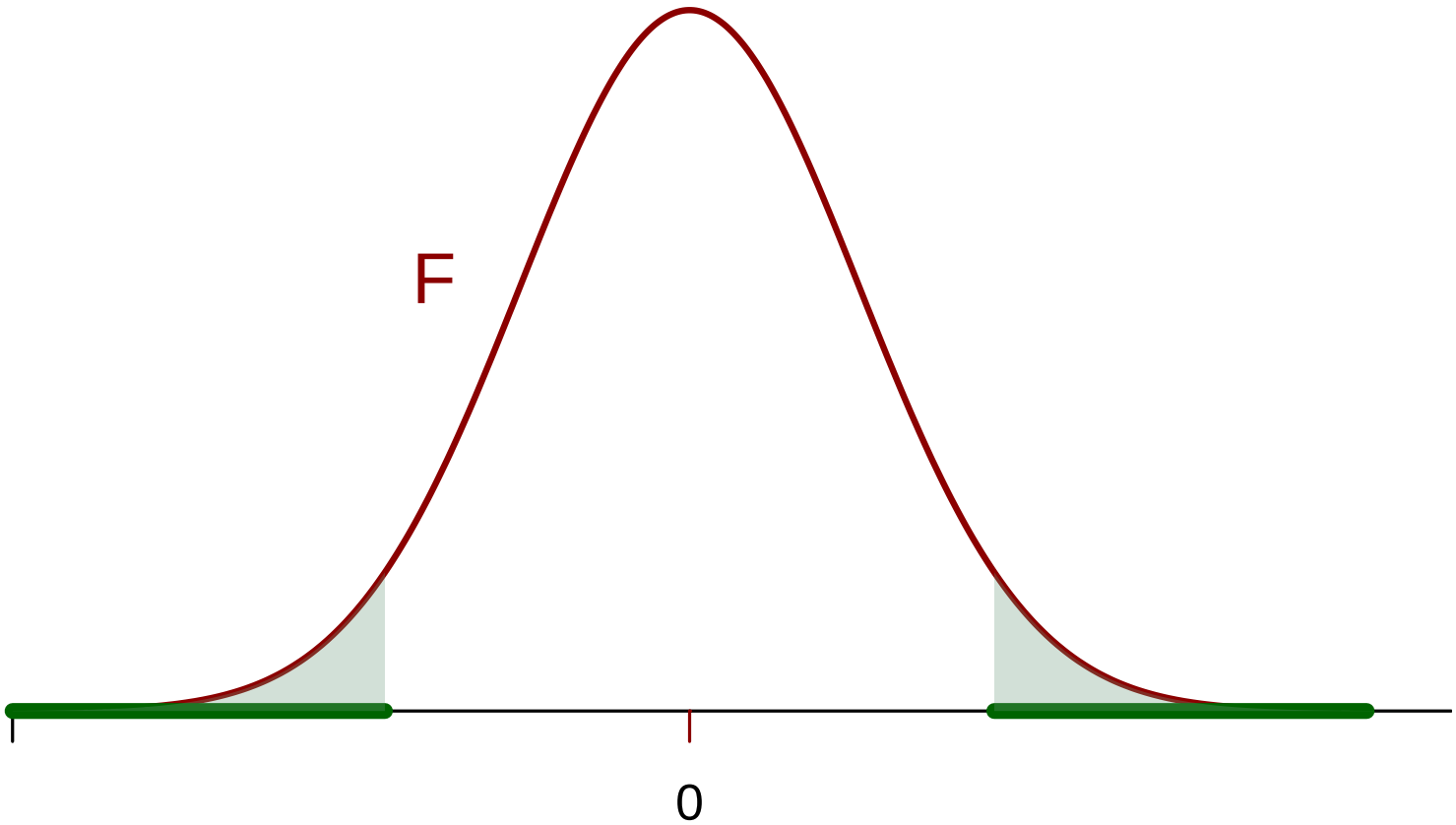
$$p = P_{H_0}(|D| > |d_{\text{obs}}|) = 2(1 - F(|d_{\text{obs}}|))$$

This means that, if the $p-$value is smaller than α, the observed sample (from which d_{obs} is derived) falls in the rejection region of level α:

$$|d_{\text{obs}}| = F^{-1}(1 - p/2) > F^{-1}(1 - \alpha/2)$$

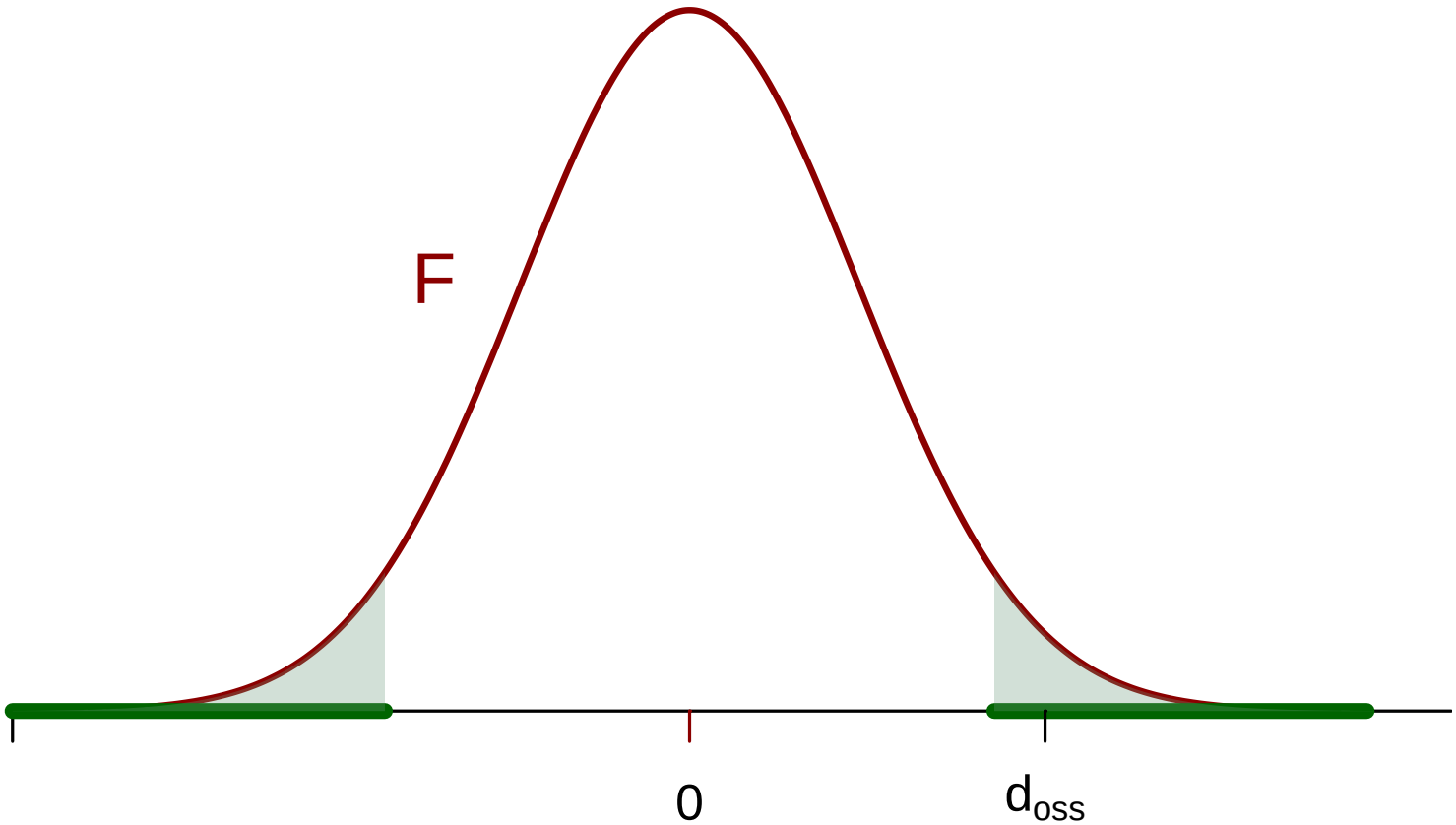
in other words, the $p - value$ is the minimum significance level such that we reject the null hypothesis in a Neyman-Person hypothesis testing, i.e. is the maximum significance level at which we accept the null hypothesis in a Neyman-Pearson test

$p-$value $\leftrightarrow$ Rejection regions



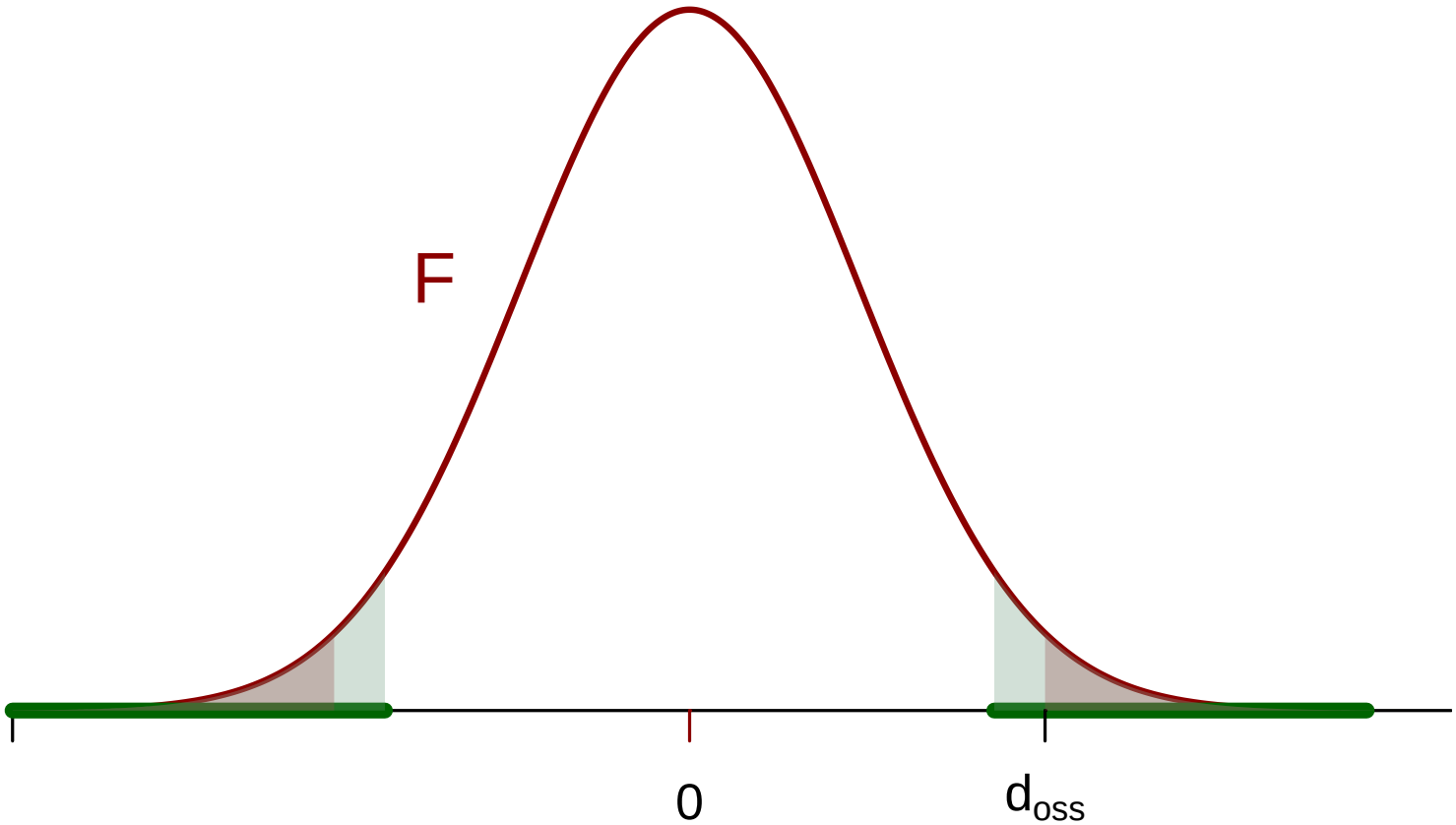
The green area represents a rejection region of level α

$p-$value $\leftrightarrow$ Rejection regions



Assume we observe a value d_{obs} that is in the rejection region

$p-$value $\leftrightarrow$ Rejection regions



What is the p-value for d_{obs}?

$$p - value < \alpha \rightarrow d_{obs} \in \mathcal{R}$$

Example

The production manager of Circuits Unlimited has asked for your assistance in analyzing a production process

This process involves drilling holes whose diameters are normally distributed with a population standard deviation of 0.06 inch

A random sample of nine measurements had a sample mean of 1.95 inches

What is the $p-$value for $H_0 = 2$ given a sample mean of $\bar{x} = 1.95$?

$$P\left(|D_0| > \left|\frac{1.95-2}{0.06/3}\right|\right) = 2(1-\Phi(2.5)) = 2(1-0.994) = 0.012$$

Compute a rejection region at level $\alpha = 1.2\%$

$$z_{1-\alpha/2} = 2.5$$

$$\rightarrow$$

$$R = \left\{\bar{x} > 2 + 2.5\frac{0.06}{3} \text{ or } \bar{x} < 2 - 2.5\frac{0.06}{3}\right\} = \{\bar{x} > 2.05 \text{ or } \bar{x} < 1.95\}$$

Example

Grand Junction Vegetables is a producer of a wide variety of frozen vegetables. The company president has asked you to determine if the weekly sales of 16-ounce packages of frozen broccoli has increased. The mean weekly number of sales per store has been 2,400 packages over the past 6 months

You have obtained a random sample of sales data from 134 stores for your study with mean $\bar{x} = 3,593$ and unbiased sample standard deviation $S = 4,919$

$$H_0 : \mu = 2400; \quad H_1 : \mu > 2400$$

$$P\left(D_0 > \frac{3593 - 2400}{4919/\sqrt{134}}\right) = 1 - P\left(D_0 < 2.81\right) = 1 - 0.9971487 = 0.002$$

At a significance level of $\alpha = 5\%$, the rejection region is

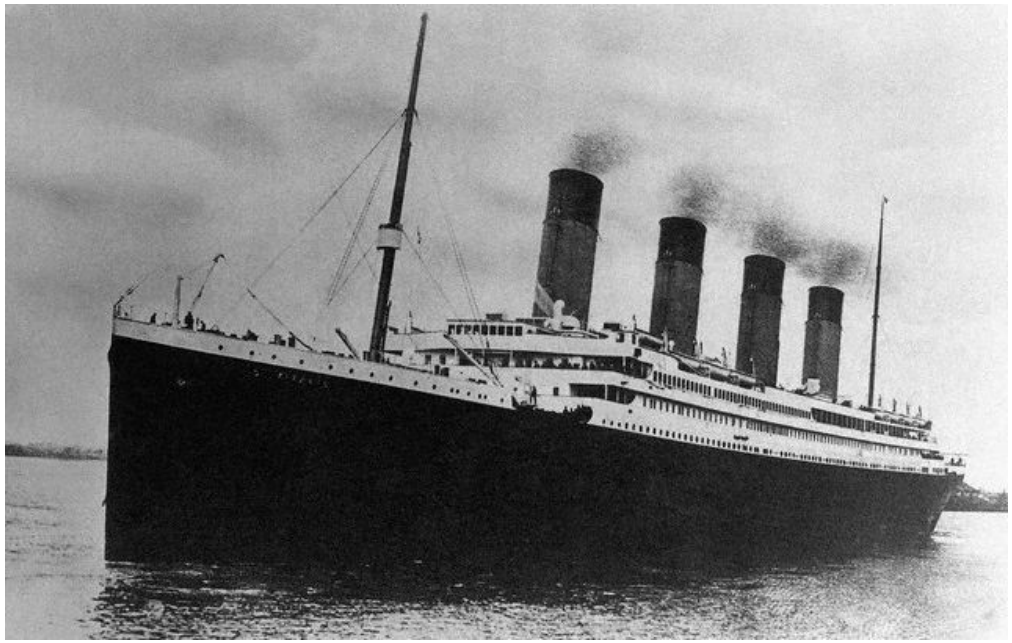
$$R = \left\{\bar{x} > \mu_0 + t_{n-1,1-\alpha} S/\sqrt{n}\right\} = \left\{\bar{x} > 2400 + 1.65 \cdot 4919/\sqrt{134}\right\} = \{\bar{x} > 3101.146\}$$

Based on the result, we reject the null hypothesis and conclude that the mean sales increased

Note that no assumptions have been made on the population distribution. But the sample size is large, thus we can assume by the CLT that the sampling distribution for the sample mean is normal

Titanic

The British ocean liner RMS Titanic sinks following a collision with an iceberg on the night between 14 and 15 April 1912



Of the 2201 people on board between passengers and crew, only 711 survived

Among the controversies following the shipwreck, some argue that third-class passengers were neglected in the evacuation operations, giving preference to the "rich"

Titanic

The contingency table describing the data about the shipwreck according to the passengers "classes" is

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

Is the chance of survival the same among the passengers?

In other words, we want to see if the conditional distributions on having survived or not are the same (if the two groups are homogeneous)

The problem can also be seen in terms of independence: we test whether the class and the outcome are independent

Mutual independence

Recall that in a contingency table

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_t	total
y_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	n_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{it}	n_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	n_{s1}	$\cdots$	n_{sj}	$\cdots$	n_{st}	n_{s0}
total	n_{01}	$\cdots$	n_{0j}	$\cdots$	n_{0t}	N

Y is not associated with X if, for $i = 1, \ldots, s$,

$$\frac{n_{i1}}{n_{01}} = \frac{n_{i2}}{n_{02}} = \cdots = \frac{n_{ij}}{n_{0j}} = \cdots = \frac{n_{it}}{n_{0t}}$$

that is if $n_{ij} = \hat{n}_{ij}$ with

$$\hat{n}_{ij} = \frac{n_{i0} n_{0j}}{N}$$

Then Y and X are mutually independent

Pearson X^2 test

If the two variables are independent, the expected frequency for the categories i and j is

$$\hat{n}_{ij} = \frac{n_{i0} n_{0j}}{N}$$

We measure how much the observed frequencies differ from the theoretical ones with

$$X^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

that you have already knew from the first part of the course

Pearson's X^2 can be used also to make inference, indeed

- if the variables are mutually independent,
- n is large

the X^2 is approximately distributed as a $\chi^2_{(s-1)(t-1)}$

$p-$value and rejection region

Observations are 'far' from the independence hypothesis if the X^2 statistic is large, so the $p-$value is

$$P(\chi^2_{(s-1)(t-1)} > X^2) = 1 - F_{\chi^2_{(s-1)(t-1)}}(\chi^2)$$

where $F_{\chi^2_{(s-1)(t-1)}}$ is the cdf of a χ^2 with $(s-1)(t-1)$ degrees of freedom

Moreover, if $\chi^2_{(s-1)(t-1),1-\alpha}$ is the quantile $1-\alpha$ of a distribution $\chi^2_{(s-1)(t-1)}$, the region

$$R = \{X^2 > \chi^2_{(s-1)(t-1),1-\alpha}\}$$

is a rejection region of level α

Titanic

Observed contingency table

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

We compute the differences

$n_{ij} - \hat{n}_{ij}$

Expected contingency table

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

Titanic

Observed contingency table

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

We compute the differences

$n_{ij} - \hat{n}_{ij}$

we square them

$(n_{ij} - \hat{n}_{ij})^2$

Expected contingency table

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

	1st	2nd	3rd
No	6362.7	98.7	8046.2
Yes	6362.7	98.7	8046.2

Titanic

Observed contingency table

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

We compute the differences

$n_{ij} - \hat{n}_{ij}$

we square them

$(n_{ij} - \hat{n}_{ij})^2$

and we divide them by the expected frequencies

$(n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$

Expected contingency table

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

	1st	2nd	3rd
No	6362.7	98.7	8046.2
Yes	6362.7	98.7	8046.2

	1st	2nd	3rd
No	31.5	0.6	18.4
Yes	51.6	0.9	30.1

Titanic

Observed contingency table

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

Expected contingency table

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

We compute the differences

$n_{ij} - \hat{n}_{ij}$

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

we square them

$(n_{ij} - \hat{n}_{ij})^2$

	1st	2nd	3rd
No	6362.7	98.7	8046.2
Yes	6362.7	98.7	8046.2

and we divide them by the expected frequencies

$(n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$

	1st	2nd	3rd
No	31.5	0.6	18.4
Yes	51.6	0.9	30.1

their sum is equal to 133.05, we conclude that...

Titanic

Observed contingency table

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

Expected contingency table

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

We compute the differences

$n_{ij} - \hat{n}_{ij}$

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

we square them

$(n_{ij} - \hat{n}_{ij})^2$

	1st	2nd	3rd
No	6362.7	98.7	8046.2
Yes	6362.7	98.7	8046.2

and we divide them by the expected frequencies

$(n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$

	1st	2nd	3rd
No	31.5	0.6	18.4
Yes	51.6	0.9	30.1

their sum is equal to 133.05, we conclude that...

Although the use of the chi-square test for association may indicate that there is a relationship between two variables, this procedure does not indicate the direction or strength of the relationship.

Exercise

The ability in mathematics and the interest in statistics for a group of people are recorded to evaluate if these two skills are independent

	Low Ability	Average Ability	High Ability
Low Interest	63	42	15
Average Interest	58	61	31
High Interest	14	47	29

test at the $\alpha = 0.01$ significance level whether a person's ability in mathematics is independent of his/her interest in statistics

The null hypothesis H_0 represents no association between X and Y

In order to test the null hypothesis, we use the Pearson's X^2 statistic

$$X^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

	Low Ability	Average Ability	High Ability
Low Interest	63	42	15
Average Interest	58	61	31
High Interest	14	47	29

Under the null hypothesis

	Low Ability	Average Ability	High Ability
Low Interest	45	50	25
Average Interest	56.25	62.5	31.25
High Interest	33.75	37.5	18.75

We now use the expected frequencies to calculate the test statistic, which is:

$$X^2 = \frac{(63-45)^2}{45} + \frac{(42-50)^2}{42} + \cdots + \frac{(29-18.75)^2}{29} = 32.14$$

The null hypothesis H_0 represents no association between X and Y

The χ^2 test for a contingency table with s rows and t columns uses critical values from the χ^2 distribution with degrees of freedom

$$(s-1)\cdot(t-1)=4$$

The null hypothesis H_0 represents no association between X and Y

The χ^2 test for a contingency table with s rows and t columns uses critical values from the χ^2 distribution with degrees of freedom

$$(s-1)\cdot(t-1)=4$$

Using the χ^2 table, we found the critical value $\chi^2_{4,0.01}=$

The null hypothesis H_0 represents no association between X and Y

The χ^2 test for a contingency table with s rows and t columns uses critical values from the χ^2 distribution with degrees of freedom

$$(s-1)\cdot(t-1)=4$$

Using the χ^2 table, we found the critical value $\chi^2_{4,0.01}=13.277$
The rejection rule at 1% is:

$$R=\{X^2>\chi^2_{4,0.01}=13.277\}$$

The X^2 statistic was 32.14, therefore the null hypothesis of no association can be rejected at 1% level

Tuberculosis

In 1948, Austin Bradford Hill (1897-1991) tested the efficacy of streptomycin for TB on a sample of 107 tuberculous patients, randomly assigning the treatments (the idea comes from agriculture where it replaced historical comparisons)

- 55, chosen at random, are treated with streptomycin and rest
- The other 52 are treated with rest alone

After the experiment

	Dead	Survived
Treated (streptomycin)	4	51
Controls	14	38

Because of the random setup of the experiment, we can assume that the difference between the two groups is due either to chance or to streptomycin (tertium non datur)

Credit risk assessment

A bank's loan department wants to assess the likelihood that potential customers will repay the money lent based on the outcomes of past years' loans

	clients	insolvents
autonomous	528	45
employees	1021	75
retirees	754	44
Total	2303	164

1) Estimate the probability that the loan will not be repaid for employees
2) Compute a 98% CI for the probability that the loan will not be repaid by an employee
3) Tell if, and to what extent, the data suggest that the probability of repayment is different for employees and retirees

Credit risk assessment

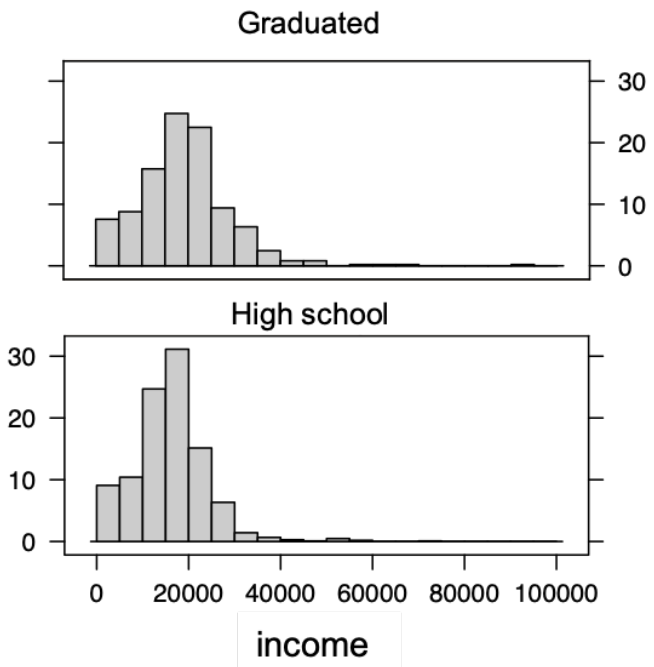
A bank's loan department wants to assess the likelihood that potential customers will repay the money lent based on the outcomes of past years' loans

	clients	ins	perc	se
autonomous	528	45	8.52	0.012
employees	1021	75	7.35	0.008
retirees	754	44	5.84	0.009
Total	2303	164	7.12	0.005

1) Estimate the probability that the loan will not be repaid for employees
2) Compute a 98% CI for the probability that the loan will not be repaid by an employee
3) Tell if, and to what extent, the data suggest that the probability of repayment is different for employees and retirees

Income data by education

To evaluate the effect of education on the employees' income, we consider a sample made up of employees with high school qualifications and graduates, all aged between 30 and 35 years



	highschool	graduates
n	1057.00	489.00
mean	15844.74	19027.29
sd	8000.36	10166.73
min	81.00	428.00
max	73292.00	94729.00

1) What is the evidence from the data?
2) Estimate the difference with a 95% CI