HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 14 — Point estimation and interval estimation

Jorge N. Tendeiro

Hiroshima University

Inferential statistics

Inferential statistics:

> *Infer characteristics of a whole population from the information available in a sample.*

Example:
Learn about the mean test score of the entire population of exam takers.

Since there are so many exam takers, we may not have access to everyone's score.

Instead, we can infer the mean score for the population based on a part of the exam takers.

Today

We will discuss two inferential methods:

- Point estimation:
  Based on finding one value.

  For example, we will be able to say things like this:

  > *The mean score of all exam takers is inferred (estimated) to be* 62.

- Interval estimation:
  Based on finding a range of values.

  For example, we will be able to say things like this:

  > *The mean score of all exam takers is inferred (estimated) to be between* 55 *and* 60.
  > *The estimation is based on a procedure which leads to a good answer* 95% *of the times.*

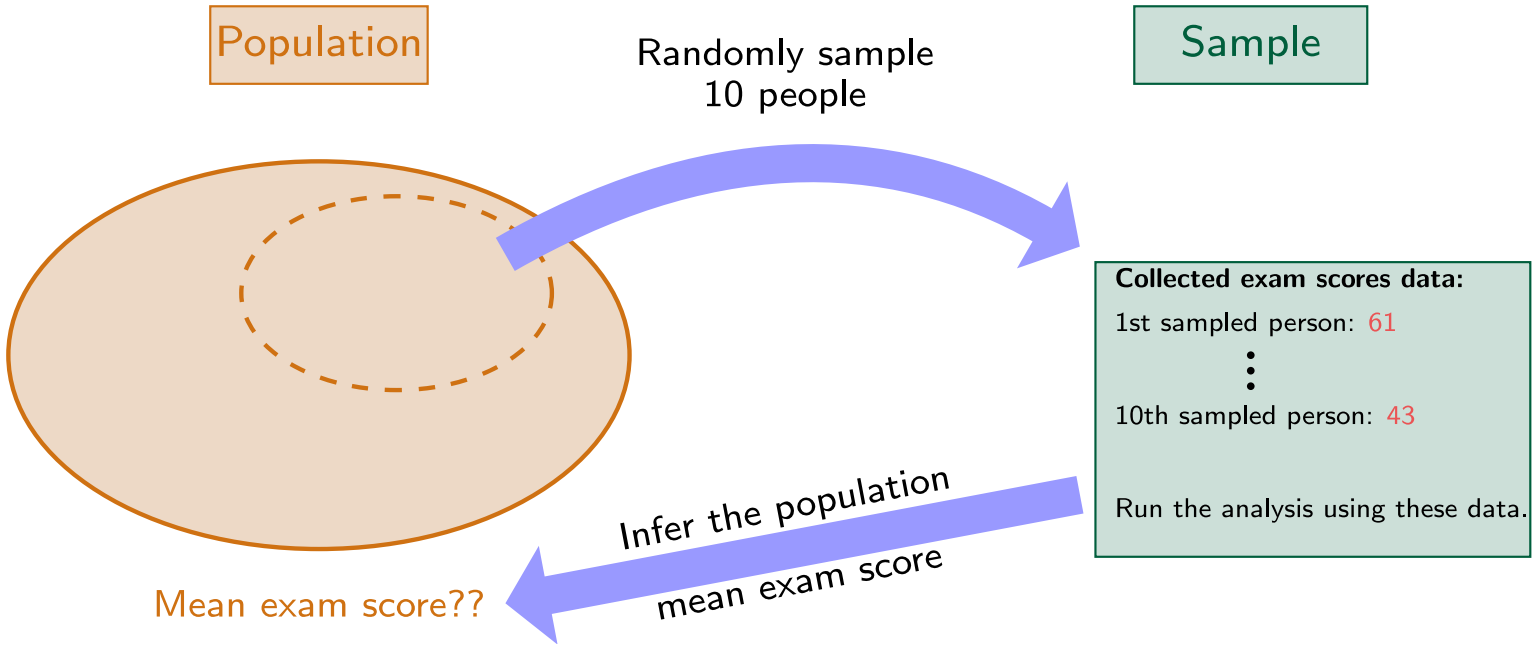
We will use Excel to assist with the computations.

Terminology, point estimation

Population, sample, random sampling

Population: The entire set of objects that we want to study.
Sample: A part of the population that is extracted.
Random sampling: Picking a sample at random (so that *population* and *sample* look alike as much as possible).

Consider the example of inferring the mean score of all exam takers:

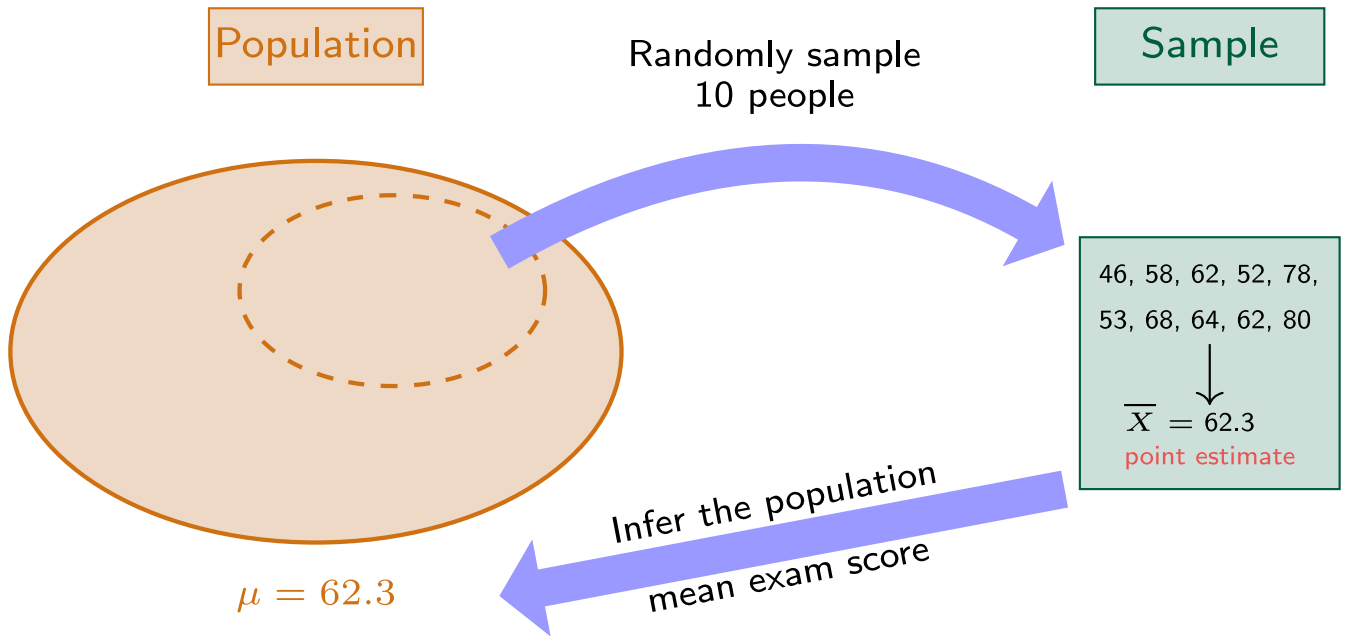


Point estimate

Point estimate:

> *Infer (estimate) characteristics of the population by means of one value.*

Example: Use the sample mean as a point estimate to infer the population mean.



Point estimate — Some notes

The good part of a point estimate is that a characteristic of the population can be estimated with one value.
Use point estimation when you need one value as a good summarizer.

Example:

> *I want to know how many defective products are produced in a factory per month.*

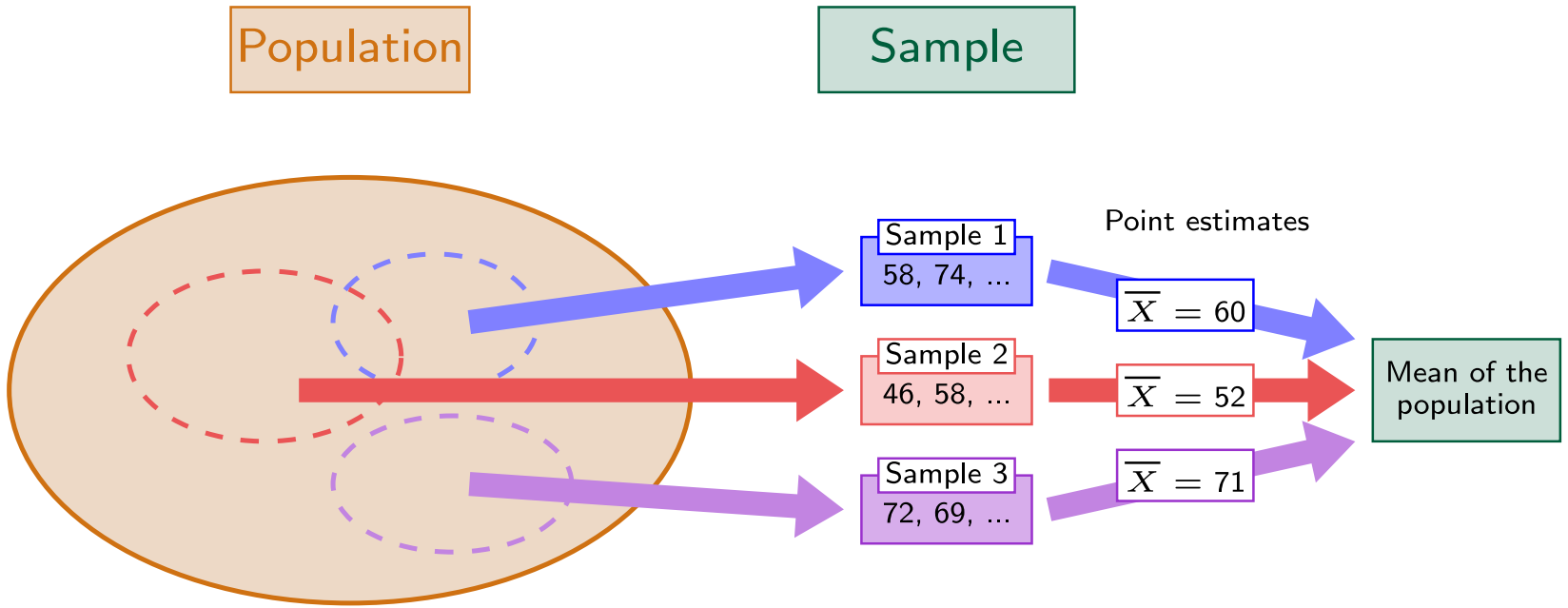
Here is one possible way to address this problem:

1. We collect data for several months.
2. We fit a Poisson distribution with rate parameter λ (recall Lecture 11).
3. As for the value of λ, we use the mean number of defective products per month in our data as the point estimate of λ.

Point estimate — Some notes

- In general, it is **impossible** that a point estimate is perfectly equal to the population feature it attempts to estimate.
  Perfect estimation would typically require that the sample *is* the entire population.

- Samples differ across sampling.
  As a result, point estimates typically differ from sample to sample.
  And, we do not know which point estimates are closer to the population feature.



Terminology, interval estimation

Interval estimate

Interval estimate:

> *Infer (estimate) characteristics of the population by means of a numerical range, probabilistically.*

Use interval estimation when you want to add credibility to your estimates.

Example:
The mean score of all exam takers is inferred (estimated) to be between 55 and 60.
The estimation is based on a procedure which leads to a good answer with probability $.95$.

Important:
Very often, you will find interval estimates being reported as follows:

> *The mean score of all exam takers is estimated to be between* 55 *and* 60 *with confidence* 95%.

You can do this as long as you **always** remember the following:

> ✓ *'Confidence' = confidence in the procedure used to compute the CI.*
> ✗ *'Confidence' = confidence that the population mean is inside this interval.*

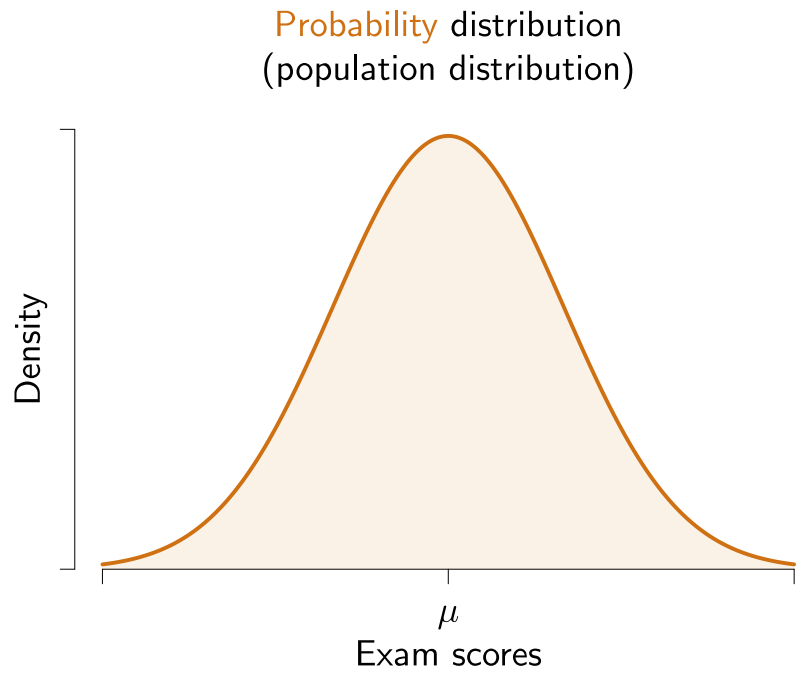
Interval estimate

Characteristics of an interval estimate:

- Does **not** estimate population characteristics by means of one value.
  Instead, it does so by means of a range of values.
- Gives probabilistic credibility to an estimate outcome, through the confidence level.
  Moreover, the confidence level can be set freely by yourself.

Random variable, probability distribution

1. Assume that there is a probability distribution modeling the uncertainty of the feature we are studying in the population (say, exam scores):



2. Further assume that each sampled value is *as if* it were sampled from the probability distribution.

If X_i is the random variable denoting the exam score of the ith person then we say that

$$X_i \sim \text{probability distribution}$$

(read: X_i 'follows' the probability distribution).

For example, if

$$X_1 = 61$$

then we are saying that the exam score of person 1 was probabilistically determined through the probability distribution.

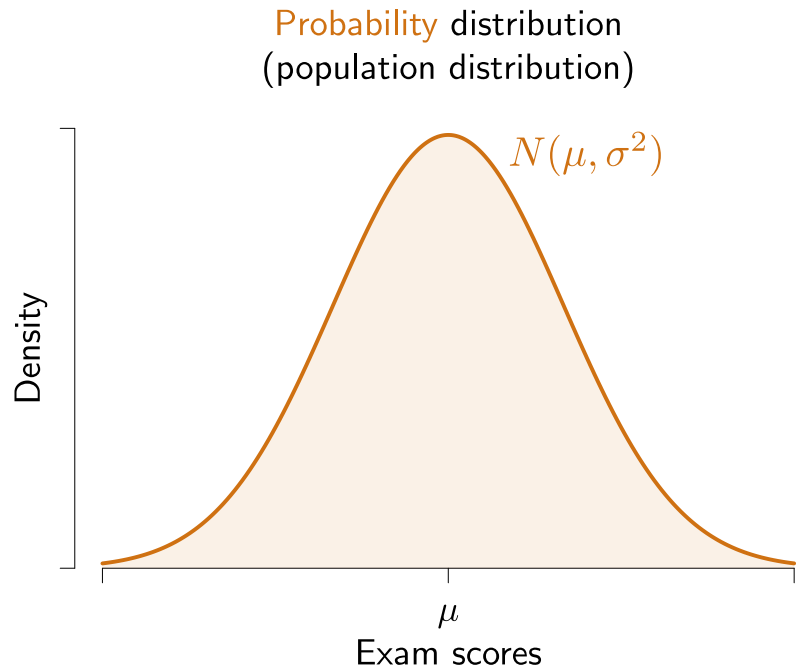
The big challenge now is:
How to choose a 'good' probability distribution?

Framework for interval estimation

Set the probability distribution

What we typically do is to choose a family of probability distributions that is known to work well in many settings.

One such family is the normal distribution:



Thus, we assume that the exam scores in the population are normally distributed, with mean μ and variance σ^2:

$$\text{exam scores} \sim N(\mu, \sigma^2)$$

Observe that

$$\mu = \text{population mean score.}$$

This is a way of expressing our research question into a parameter (μ).

Our goal is to use the sample to infer what the population mean value μ might be.

For that we will use interval estimation.

Interval estimation framework — Assuming variance σ^2 known

Use the following fact:

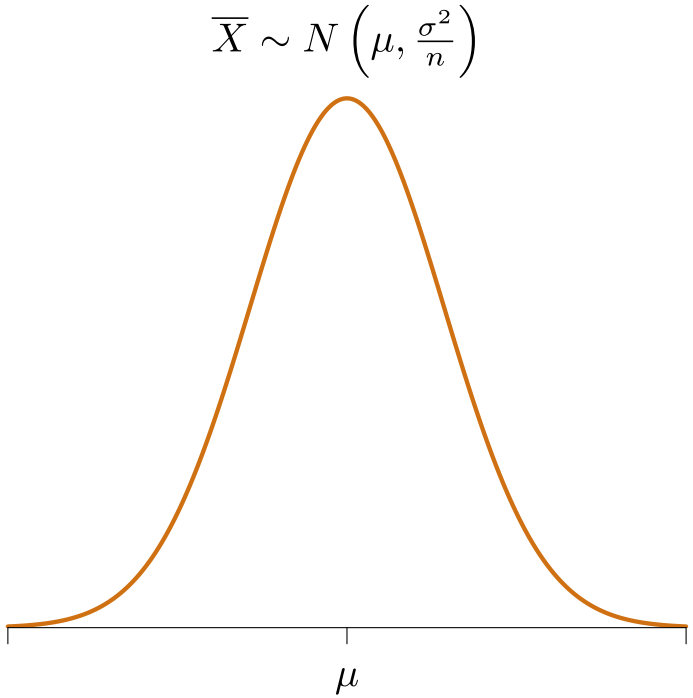
> *Given random variable X following a normal distribution $N(\mu, \sigma^2)$, the mean of a random sample (sample mean) with sample size n,*
>
> $$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$
>
> *verifies*
>
> $$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

In simple terms, this results tells us how we can 'transfer' our uncertainty about the value of X onto our uncertainty about the value of the sample mean $\overline{X}$.

Just like X is a random variable (i.e., its value changes on repeated sampling), also the sample mean $\overline{X}$ is a random variable (its value also changes on repeated sampling). But, it changes in a predictable way!

Interval estimation framework — Assuming variance σ^2 known

Let's learn a procedure that computes an interval that includes μ with probability $.95$, in the long run.

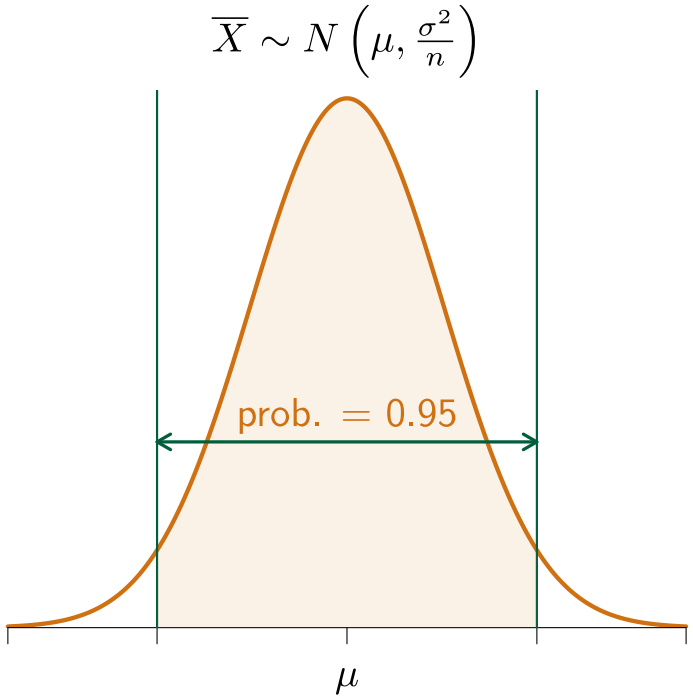


1. Consider the distribution of the mean of a random sample of size n.

This is known as the sampling distribution of the mean.

Interval estimation framework — Assuming variance σ^2 known

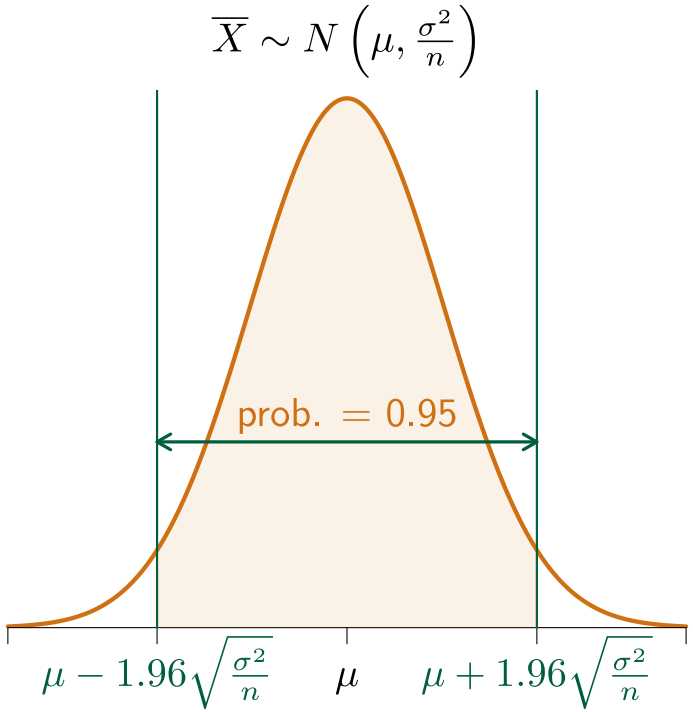
Let's learn a procedure that computes an interval that includes μ with probability $.95$, in the long run.



2. Find the central, symmetric around μ, interval of the sampling distribution of the mean that contains $\overline{X}$ with probability $.95$.

Interval estimation framework — Assuming variance σ^2 known

Let's learn a procedure that computes an interval that includes μ with probability $.95$, in the long run.



3. It can be shown that

$$\mu - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \overline{X} \leq \mu + 1.96\sqrt{\frac{\sigma^2}{n}}$$

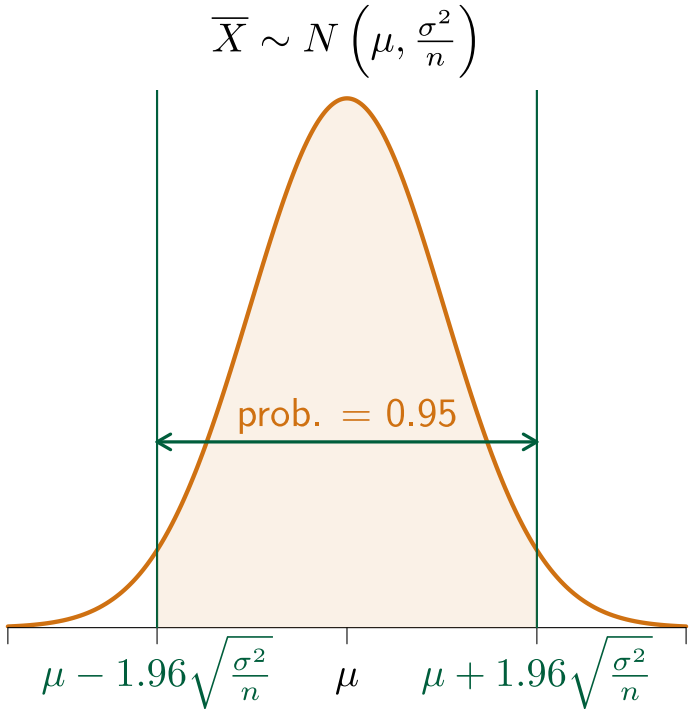
This interval contains $\overline{X}$ with probability $.95$.
Notice that the value 1.96 is an approximation.

We are almost there... But not yet.

The above interval is an interval for $\overline{X}$.
However, we are looking for an interval for μ.

Interval estimation framework — Assuming variance σ^2 known

Let's learn a procedure that computes an interval that includes μ with probability $.95$, in the long run.



4. With a little algebra, we can further 'invert' the previous interval and see that

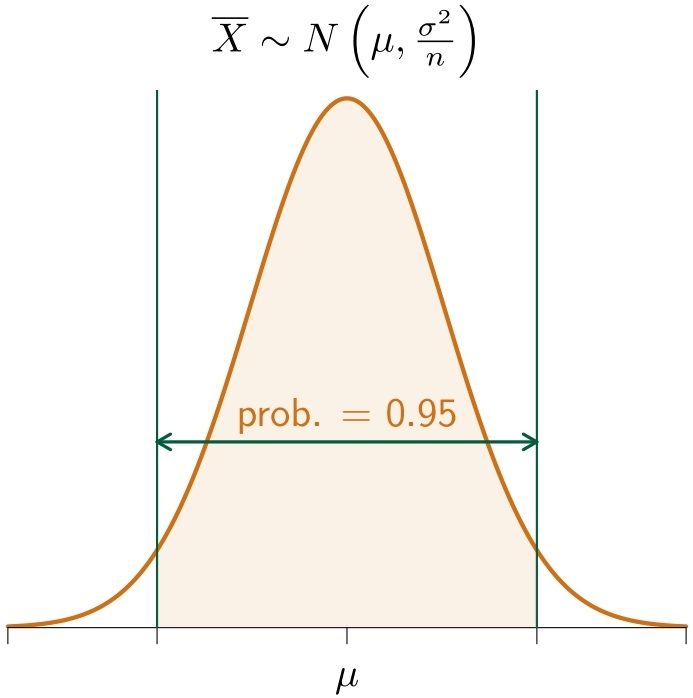
$$\overline{X} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

And this is it!
This interval contains μ with probability $.95$
(at least before $\overline{X}$ is known).

The interval above is known as the 95% confidence interval for μ.

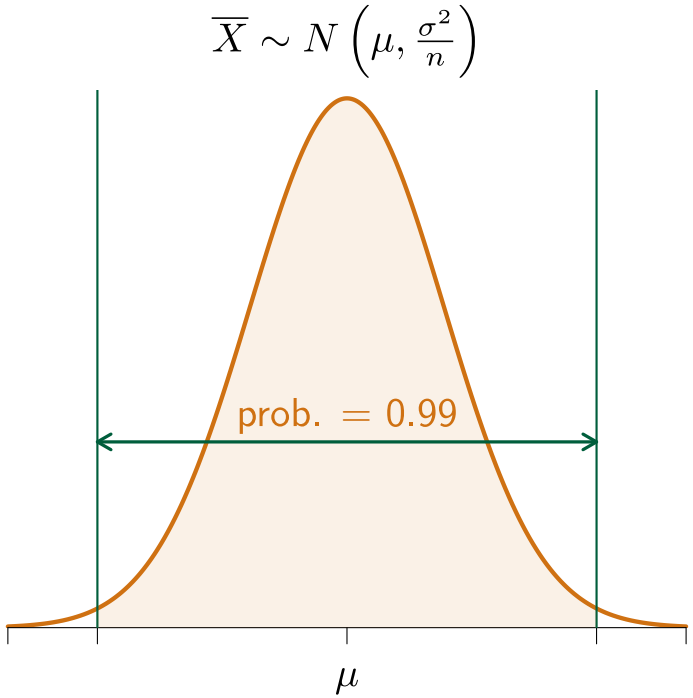
Interval estimation framework — Assuming variance σ^2 known

What if we rather compute an interval that includes μ with probability $.99$, in the long run?



Interval estimation framework — Assuming variance σ^2 known

What if we rather compute an interval that includes μ with probability $.99$, in the long run?



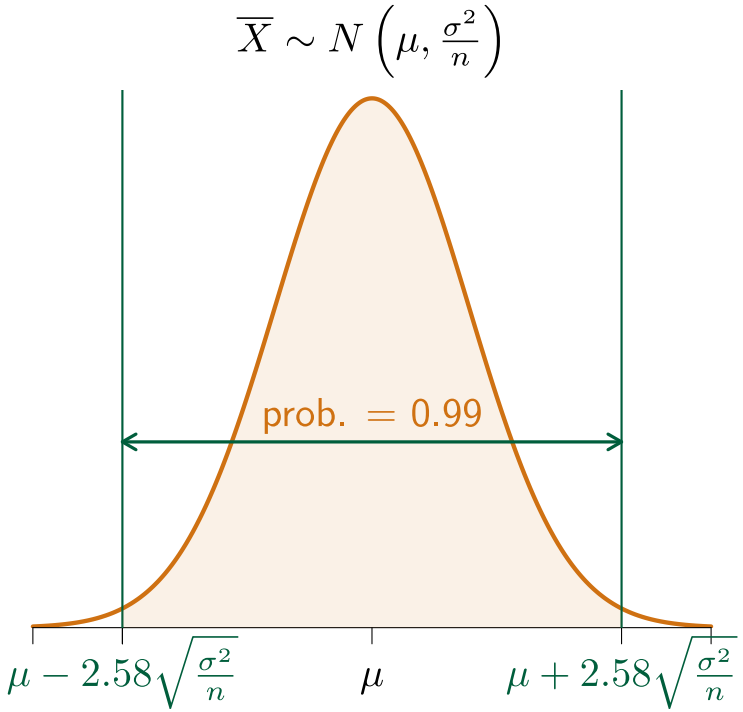
Increasing the probability inside the interval implies choosing a wider interval.

In other words, there is a relation between the confidence level and the width of the confidence interval:

> *The larger the confidence level, the wider the confidence interval.*

Interval estimation framework — Assuming variance σ^2 known

What if we rather compute an interval that includes μ with probability $.99$, in the long run?



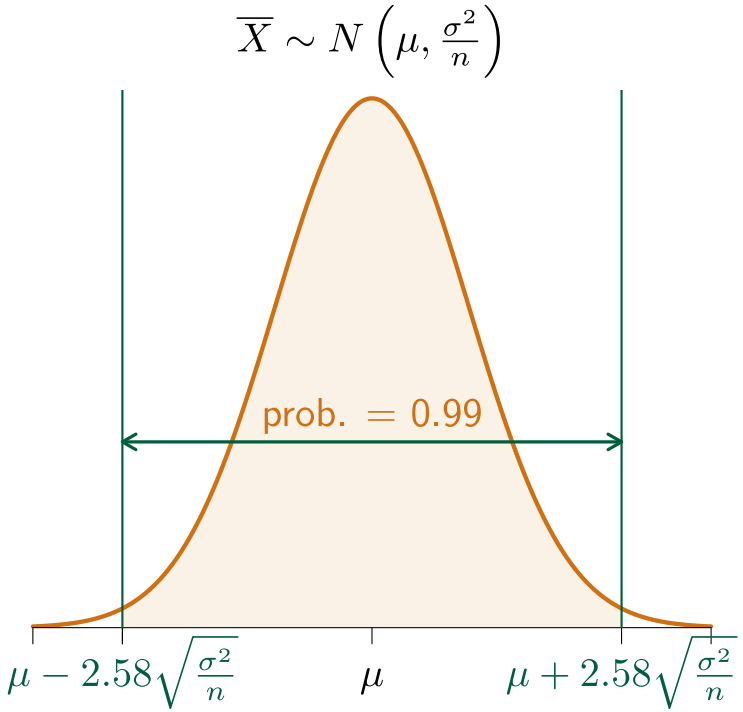
It can be shown that

$$\mu - 2.58\sqrt{\frac{\sigma^2}{n}} \leq \overline{X} \leq \mu + 2.58\sqrt{\frac{\sigma^2}{n}}$$

This interval contains $\overline{X}$ with probability $.99$.

Interval estimation framework — Assuming variance σ^2 known

What if we rather compute an interval that includes μ with probability $.99$, in the long run?



And finally

$$\overline{X} - 2.58\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X} + 2.58\sqrt{\frac{\sigma^2}{n}}$$

This is the 99% confidence interval for μ.

It contains μ with probability $.99$
(at least before $\overline{X}$ is known).

Confidence interval for μ, variance σ^2 known — Summary

When random variable X follows a normal distribution $N(\mu, \sigma^2)$ then:

- The 95% confidence interval for μ is given by

$$\overline{X} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X} + 1.96\sqrt{\frac{\sigma^2}{n}}.$$

  This is an interval estimate which contains μ with probability $.95$, in the long run.

- The 99% confidence interval for μ is given by

$$\overline{X} - 2.58\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X} + 2.58\sqrt{\frac{\sigma^2}{n}}.$$

  This is an interval estimate which contains μ with probability $.99$, in the long run.

Confidence interval for μ, variance σ^2 known — Example

> *Suppose that $\sigma^2 = 100$, $n = 5$, and the mean value of the data is $\overline{X} = 68$.*
> *Compute the 95% confidence interval.*

In this case we have that

$$\overline{X} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

$$68 - 1.96\sqrt{\frac{100}{5}} \leq \mu \leq 68 + 1.96\sqrt{\frac{100}{5}}$$

$$59.23 \leq \mu \leq 76.77$$

The 95% confidence interval is $(59.23, 76.77)$.

Confidence interval — Interpretation (important!!)

How can you report a specific 95% confidence interval such as $(59.23, 76.77)$?

Well, you *can* say the following:

> *The interval* $(59.23, 76.77)$ *was computed through a procedure that, in the long run, produces intervals that include the population mean* μ 95% *of the times.*

Yuck!
Can't I just say this?:

> *The population mean* μ *lies in the interval* $(59.23, 76.77)$ *with probability* $.95$.

Well, **no, you cannot**...

Confidence interval — Interpretation (important!!)

Let's look carefully at the 95% confidence interval for μ:

$$\left(\overline{X} - 1.96\sqrt{\frac{\sigma^2}{n}}, \overline{X} + 1.96\sqrt{\frac{\sigma^2}{n}}\right)$$

This interval depends on the sample mean $\overline{X}$:

$$\text{different samples} \implies \text{different } \overline{X} \text{ values} \implies \text{different } 95\% \text{ confidence intervals.}$$

However, the population parameter that we are estimating, μ, is considered to be fixed (i.e., it does **not change**).

Confidence interval — Interpretation (important!!)

- The 'confidence' from a 95% confidence interval is **not** about a particular interval containing the population parameter.
- Instead, the 'confidence' pertains to the procedure used to compute the interval:

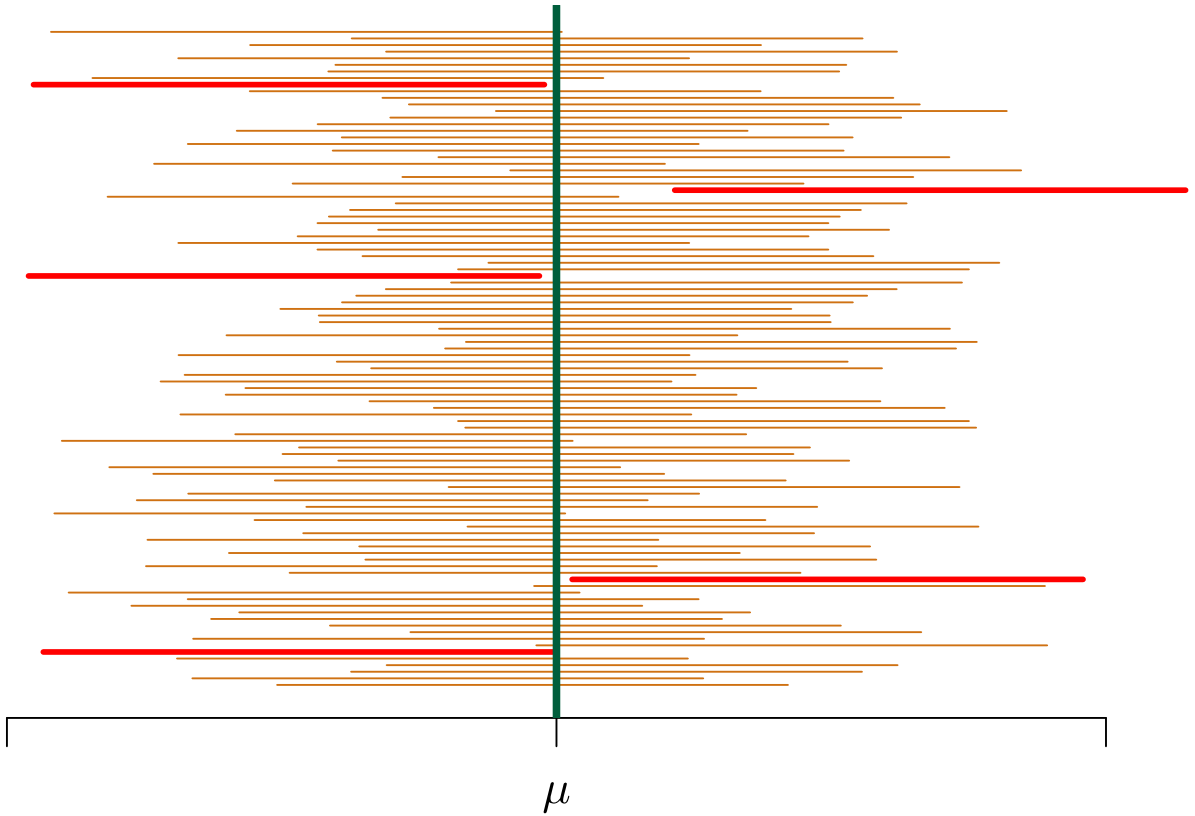
> *We are confident in the sense that we trust that using such a procedure gives us a 'good' interval 95% of the times, in the long run.*

To see the 95% probability of confidence intervals at work, we need to sample repeatedly under the same conditions (i.e., from the same population, with the same sample size).

Let's do this, say, 100 times.

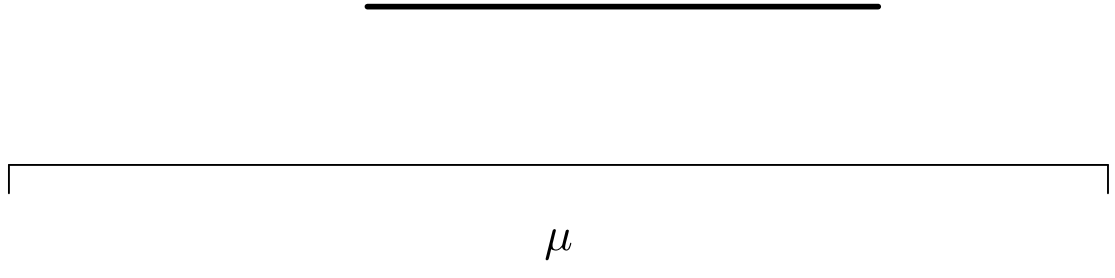
Confidence interval — Interpretation (important!!)

In theory: About 95% of the confidence intervals include μ.



Confidence interval — Interpretation (important!!)

In practice: We only compute one confidence interval.
It either includes or excludes μ: We **don't know** for sure because we do not know the true value of μ!



Confidence interval — Example 1

> *We are interested in the mean score of all exam takers of a test. Here, we set the entire exam takers as the population and randomly sampled 10 people from the population. The mean score of these 10 people was 60.*

Assume that the population of all exam scores is normally distributed with mean μ and variance $\sigma^2 = 90$ (i.e., assume variance known).

Compute the 99% confidence interval for μ:

$$\overline{X} - 2.58\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X} + 2.58\sqrt{\frac{\sigma^2}{n}}$$

$$60 - 2.58\sqrt{\frac{90}{10}} \leq \mu \leq 60 + 2.58\sqrt{\frac{90}{10}}$$

$$52.26 \leq \mu \leq 67.74$$

Confidence interval — Example 1

$$\boxed{52.26 \leq \mu \leq 67.74}$$

Interpretation:

> *We estimate the population mean score of all exam takers to be between* 52.26 *and* 67.74*. The estimation is based on a procedure that, in the long run, produces intervals that include the population mean* μ 99% *of the times.*

An alternative, very popular, way of reporting the result is the following:

> *We estimate the population mean score of all exam takers to be between* 52.26 *and* 67.74 *with confidence level* 99%*.*

This is fine as long as you keep in mind that

> *'confidence' means 'confidence in the procedure used to compute the CI'.*

Confidence interval — In Excel

Given random variable X following a normal distribution with mean μ and known variance σ^2, the $(1-\alpha)100\%$ confidence interval for the population mean μ can be calculated in Excel like this:

1. Calculate the so-called confidence interval's margin of error:

For a 95% confidence interval ($\alpha = .05$):

$$\left(\overline{X} - \underbrace{1.96\sqrt{\frac{\sigma^2}{n}}}_{\text{margin of error}}, \overline{X} + \underbrace{1.96\sqrt{\frac{\sigma^2}{n}}}_{\text{margin of error}}\right)$$

For a 99% confidence interval ($\alpha = .01$):

$$\left(\overline{X} - \underbrace{2.58\sqrt{\frac{\sigma^2}{n}}}_{\text{margin of error}}, \overline{X} + \underbrace{2.58\sqrt{\frac{\sigma^2}{n}}}_{\text{margin of error}}\right)$$

In Excel, use the `CONFIDENCE.NORM` command to compute the margin of error:

```
=CONFIDENCE.NORM(alpha, standard deviation, sample size)
```

Note: The standard deviation, σ, is equal to the square root of the variance: $\sigma = \sqrt{\sigma^2}$.

Confidence interval — In Excel

Given random variable X following a normal distribution with mean μ and known variance σ^2, the $(1-\alpha)100\%$ confidence interval for the population mean μ can be calculated in Excel like this:

2. Once the margin of error is computed, simply compute the desired confidence interval as

$$(\overline{X} - (\text{margin of error}), \overline{X} + (\text{margin of error})).$$

8. Compute 99 percent confidence interval

Jorge Tendeiro

01:34

Link

Exercise 1

> *We are interested in the mean score of all exam takers of a test. Here, we set the entire exam takers as the population and randomly sampled 10 people from the population. The mean score of these 10 people was 60.*

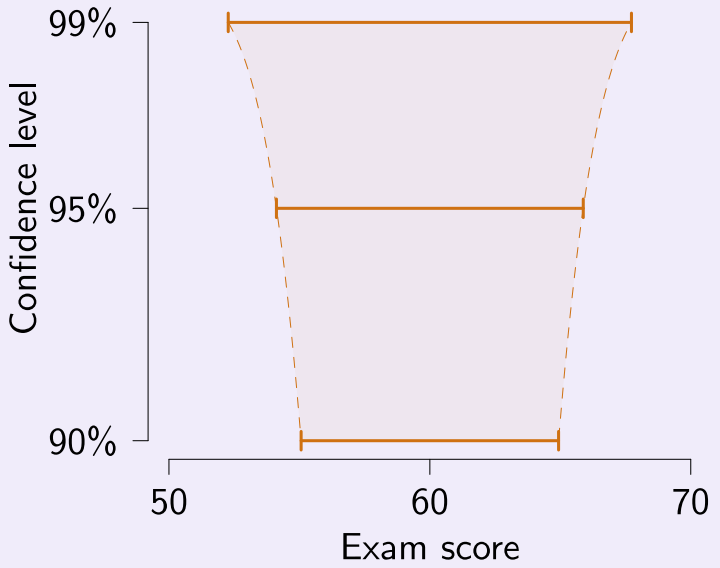
What happens to the confidence interval if we change the confidence level?

Confidence interval data		Margin of error	Confidence interval	
level (%) =	99	7.727487911	Lower bound =	52.27251
alpha =	0.01		Upper bound =	67.72749
variance =	90			
sample size =	10			
mean =	60			

Confidence interval data		Margin of error	Confidence interval	
level (%) =	95	5.879891954	Lower bound =	54.12011
alpha =	0.05		Upper bound =	65.87989
variance =	90			
sample size =	10			
mean =	60			

Confidence interval data		Margin of error	Confidence interval	
level (%) =	90	4.934560881	Lower bound =	55.06544
alpha =	0.1		Upper bound =	64.93456
variance =	90			
sample size =	10			
mean =	60			

The lower the confidence level, the narrower the width of the confidence interval.



Summary

We learned about:

- Point estimation — Using one value to infer about a population parameter:

  > *The mean score of all exam takers is inferred (estimated) to be* 62.

- Interval estimation — Using a range of values to probabilistically infer about a population parameter:

  > *The mean score of all exam takers is inferred (estimated) to be between* 55 *and* 60.
  > *The estimation is based on a procedure which leads to a good answer* 95% *of the times.*

Summary

Random sampling is a major premise for today's topic.

Without random sampling...

- The sampling distribution may not be as expected (for example, normal distribution in case of μ).
- The coverage probability of, say, 95% confidence intervals may be quite different from 95%, in the long run.

Next lecture:

> *What can we do in case σ^2 is unknown (which is almost always the case!), or when the population distribution is not normal?*