HIROSHIMA UNIVERSITY

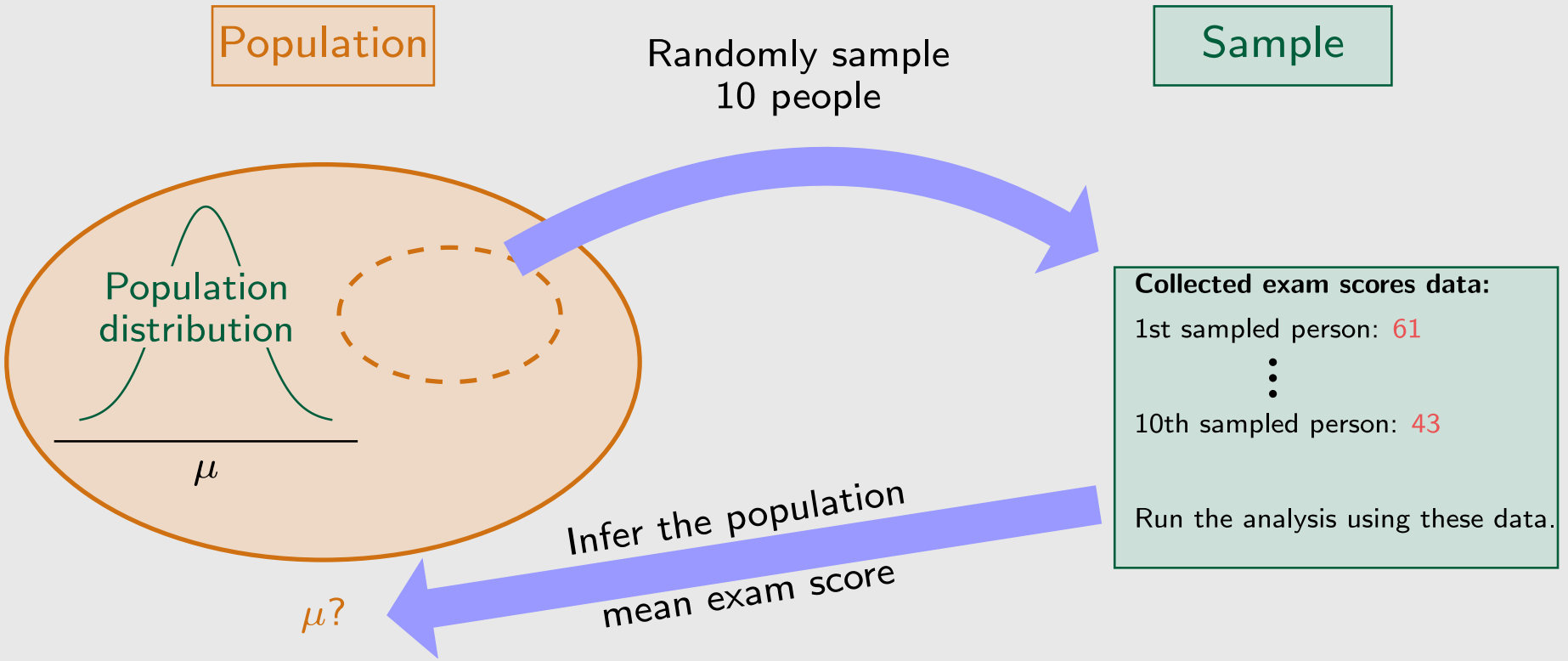
Fundamental Data Science (30104001)

Lecture 15 — Interval estimation

Jorge N. Tendeiro

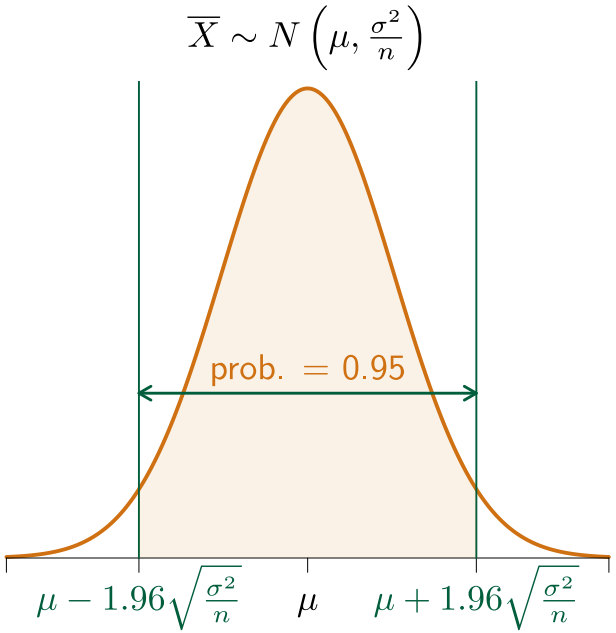
Hiroshima University

Review — Interval estimation



- We assume that there are random variables $X_1, X_2, \ldots$ all following the population distribution.
- We further assume that each measurement acts as if it were randomly sampled from the random variable's population distribution.

Review — Interval estimation



Assuming that the population distribution is $N(\mu, \sigma^2)$ with variance σ^2 known, then the 95% confidence interval for μ is

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

In Excel, use the command `CONFIDENCE.NORM`.

But, what do we do in case:

- The population is normally distributed, but the variance σ^2 is **unknown**?
- The population is **not** normally distributed?

Interval estimate
(when variance σ^2 is unknown)

Interval estimation when σ^2 is unkown — What changes

Changes compared to the previous method:

1. Replace σ^2 by an unbiased estimate of variance $\widehat{\sigma}^2$:

$$\widehat{\sigma}^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

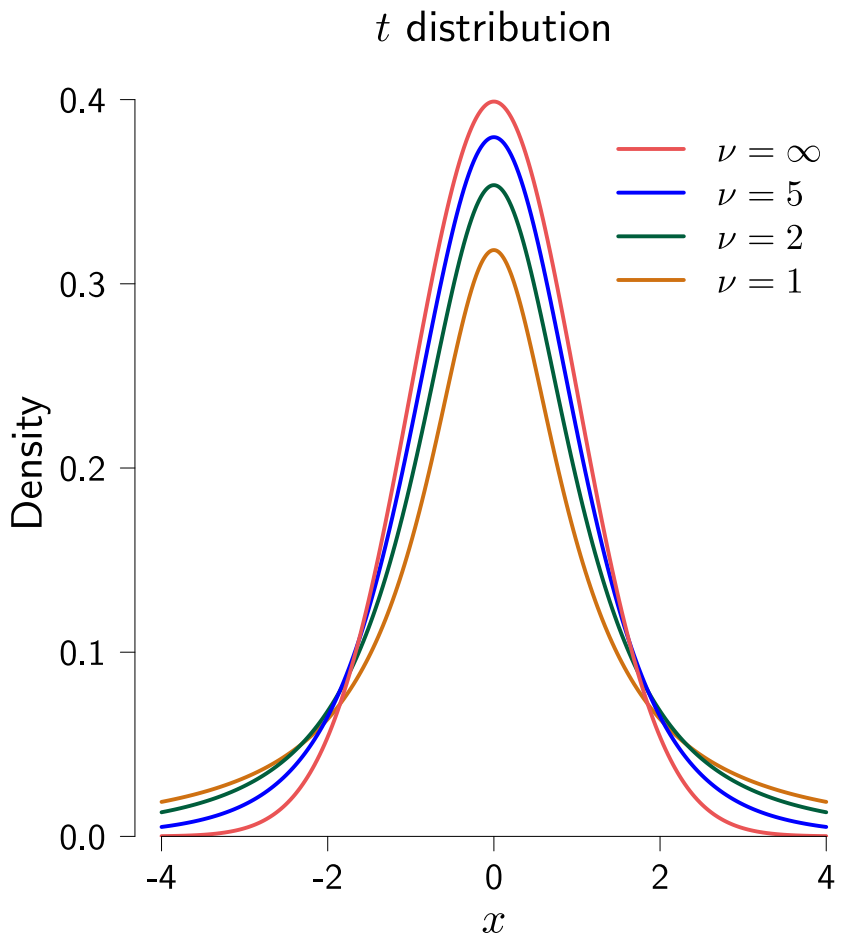
- $x_i = i$th observation
- $\overline{x} =$ observed sample mean
- $n =$ sample size

2. Replace the normal distribution by the t distribution with $(n-1)$ degrees of freedom.
The t distribution looks similar to the normal distribution but it has heavier tails.

Conclusion:
The 95% confidence interval for the mean μ when σ^2 is known is

$$\overline{X} - t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}} \leq \mu \leq \overline{X} + t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}}$$

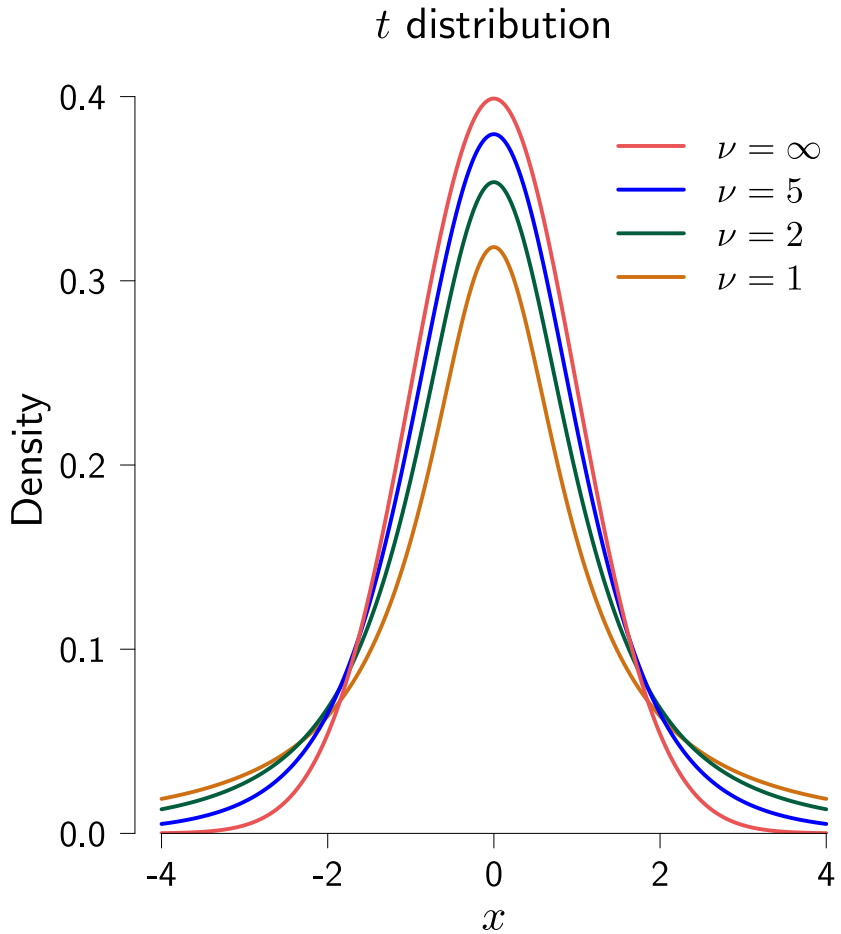
t distribution



The t distribution has a parameter denoted ν.

- Read ν as 'nyu'.
- ν can be any non-negative real number.
- ν is known as the number of degrees of freedom.
- We sometimes write t_ν to specify the t distribution with ν degrees of freedom.
- t has heavier tails than the normal distribution.
  - Easier to get large values in absolute size in comparison to the normal distribution.
  - Takes into account that estimating σ^2 increases our uncertainty.
- $t_\infty = N(0, 1)$

t distribution



For interval estimation, use the t distribution with $\nu = n - 1$ degrees of freedom.

In particular, this means that the specific t distribution to be used in interval estimation varies with sample size.

t distribution — critical values

$$\overline{X} - t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}} \leq \mu \leq \overline{X} + t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}}$$

Also note that the value of t_{n-1} depends on the desired confidence level.
The value of t_{n-1} can vary as illustrated below:

	Confidence level		
n	90%	95%	99%
5	2.1318	2.7764	4.6041
10	1.8331	2.2622	3.2498
15	1.7613	2.1448	2.9768

You don't need to worry about this though:
We will use Excel for the computations.

Interval estimation when σ^2 is unkown — Summary

When the population is normally distributed with mean μ and unknown variance σ^2 (i.e., $N(\mu, \sigma^2)$), we use the following formula to compute confidence intervals for μ:

$$\overline{X} - t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}} \leq \mu \leq \overline{X} + t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}}$$

Technically, this formula is based on the mathematical fact that

$$T = \frac{\overline{X} - \mu}{\sqrt{\widehat{\sigma}^2/n}} \sim t_{n-1}.$$

The exact value of t_{n-1} depends on both:

- The sample size, via the degrees of freedom ($\nu = n - 1$).
- The confidence level.

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, and the (unbiased) variance estimate was 90.*

Population distribution: $N(\mu, \sigma^2)$, with σ^2 unknown.

Estimate the population mean μ via a 95% confidence interval:

$$\overline{X} - t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}} \leq \mu \leq \overline{X} + t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}}$$

$$60 - 2.26\sqrt{\frac{90}{10}} \leq \mu \leq 60 + 2.26\sqrt{\frac{90}{10}}$$

$$53.22 \leq \mu \leq 66.78$$

Note: The value of t_{n-1} was found using software (see the table shown before).

Example 1

$$95\%\ \text{CI for}\ \mu = (53.22, 66.78)$$

Interpretation:

> *We estimate the mean score μ to be between 53.22 and 66.78.*
> *The estimation is based on a procedure which leads to a good answer 95% of the times.*

Alternative interpretation:

> *The mean score, μ, is estimated to be between 53.22 and 66.78 with 95% confidence level.*

When you read this, always keep the following in mind:

> *The mentioned 'confidence' pertains to the procedure used to compute the CI.*
> *The 'confidence' does **not** pertain to the CI itself.*

Confidence interval — In Excel

The procedure in Excel is the same as before, except for the function to compute the margin of error:

$$\overline{X} - \underbrace{t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}}}_{\text{margin of error}} \leq \mu \leq \overline{X} + \underbrace{t_{n-1}\sqrt{\frac{\widehat{\sigma}^2}{n}}}_{\text{margin of error}}$$

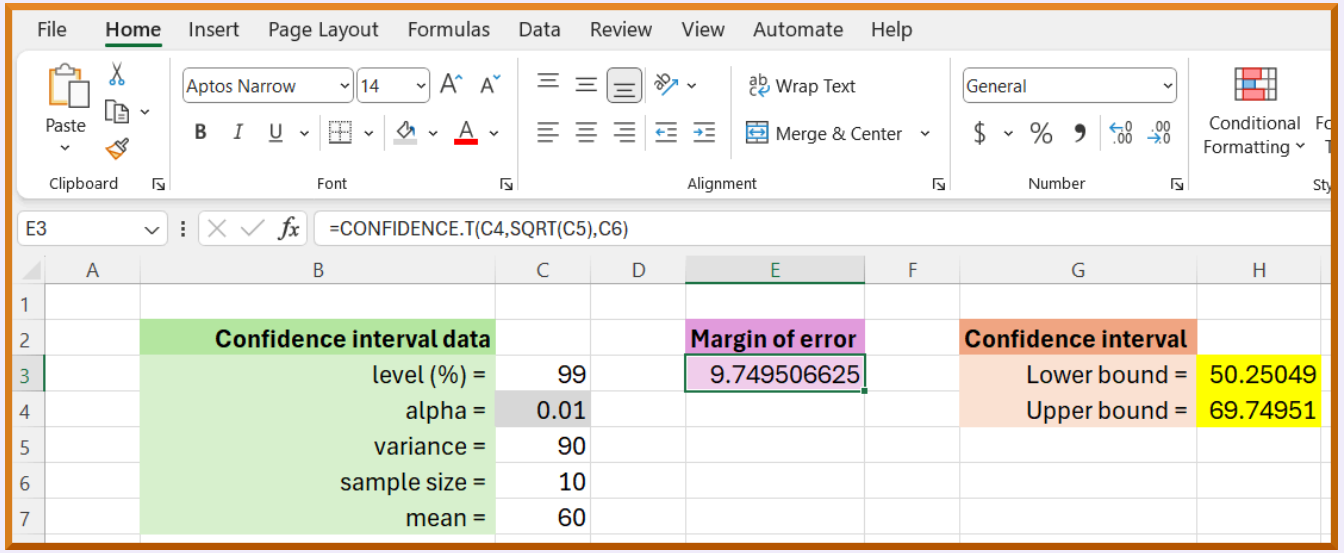
In Excel, use the `CONFIDENCE.T` command to compute the margin of error:

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

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

Exercise 1

> *Under the same setting as Example 1, calculate the 99% confidence interval of μ. Provide your answer by rounding to the second decimal place.*



The estimated 99% confidence interval is $(50.25, 69.75)$.

Remember:
When you know that the population distribution follows a normal distribution but σ^2 is unknown, use the t distribution for interval estimation.

Interval estimate (when the population is **not** normally distributed)

When the population is not normally distributed

There are various strategies to deal with this problem.

We will only consider the simpler strategy:

> *Increase the sample size.*

The following result is one of the most famous theorems in mathematical statistics:

Central limit theorem

> *Suppose the population distribution has any shape, as long as both its mean μ and variance σ^2 are finite. Given a random sample of size n from this population $\{X_1, \ldots, X_n\}$, the distribution of the sample mean $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ approaches $N\left(\mu, \frac{\sigma^2}{n}\right)$ as n increases.*

The main idea is this:
For large enough n, the sampling distribution of the sample mean approximately follows a normal distribution.

We can therefore still use the confidence interval that we learned before!

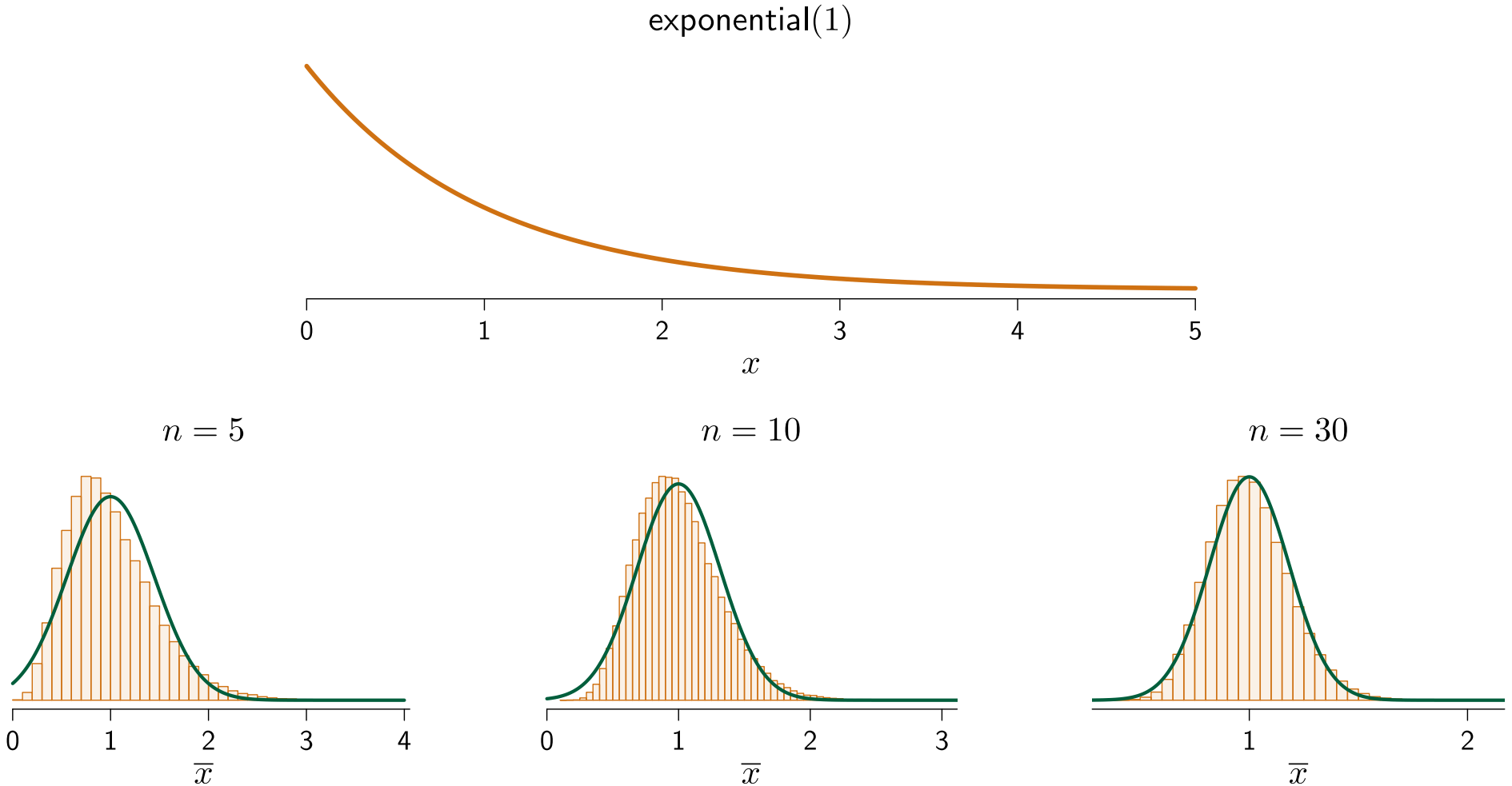
The central limit theorem at work

Let's see how extremely *abnormal* (rather, not-normal) population distributions can still have distributions of the sample mean that are close to being normally distributed.

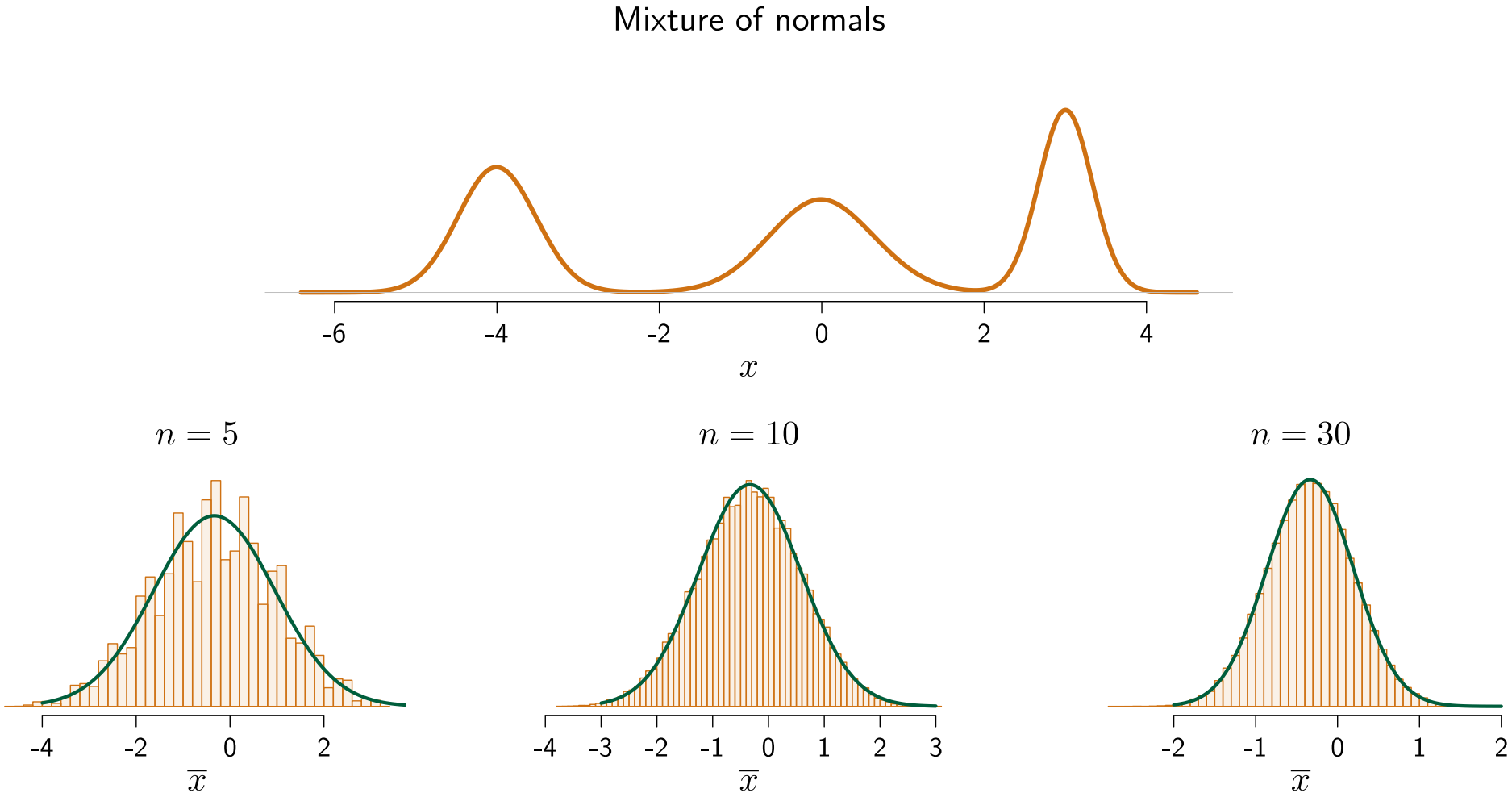
The approximation improves as n increases.

The distribution of sample means (histograms) is based on 100,000 random draws.

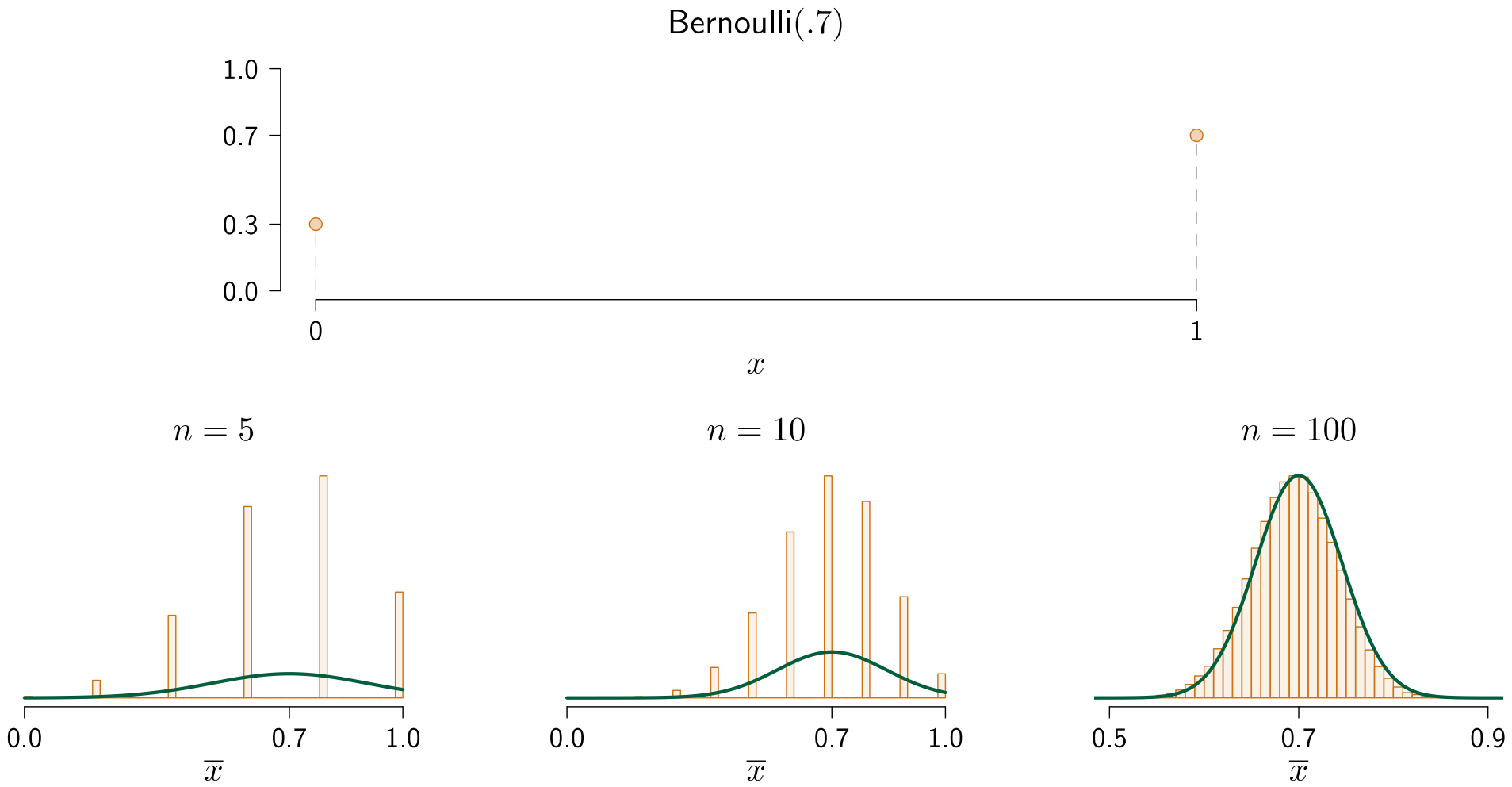
The central limit theorem at work



The central limit theorem at work



The central limit theorem at work



When the population is not normally distributed — Summary

Regardless of the shape of the population distribution, and under mild conditions, when the sample size n is large enough and for a random sample:

When σ^2 is known, the 95% confidence interval is:

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

When σ^2 is unknown, the 95% confidence interval is:

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

where $\widehat{\sigma}^2$ is an unbiased estimate for the population variance σ^2.

When the population is not normally distributed — Summary

In Excel:

When σ^2 is known, use

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

to compute the margin of error.

- `alpha` = 0.05 for a 95% confidence level
- `standard deviation` = $\sqrt{\sigma^2}$
- `sample size` = n

When σ^2 is unknown, use

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

to compute the margin of error.

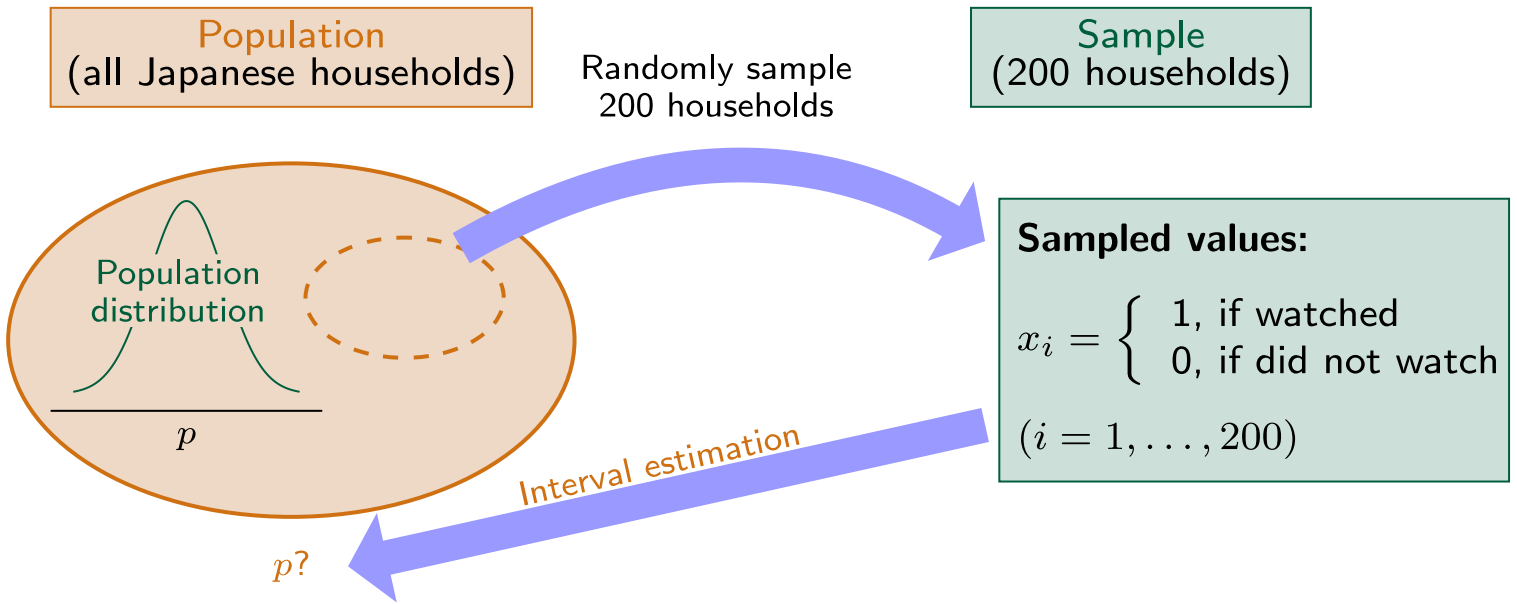
- `alpha` = 0.05 for a 95% confidence level
- `standard deviation` = $\sqrt{\widehat{\sigma}^2}$
- `sample size` = n

Note:
Interval estimates get more accurate as the sample size n increases.
This is due to the CLT: the larger n is, the better the approximation to the normal distribution.

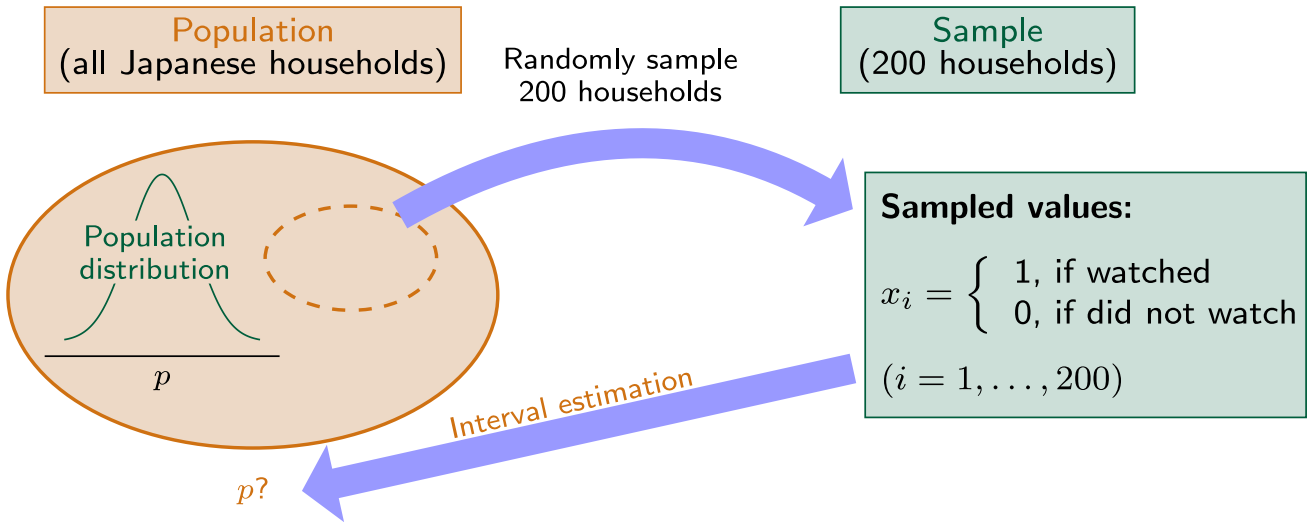
Example 2

Suppose you want to know the ratings of a TV show among all Japanese households.

You decide to compute an interval estimate for the population rating.



Example 2



Here the population distribution is assumed to be a Bernoulli distribution $Be(p)$ (recall Lecture 11), where

$$p = \text{proportion of all Japanese households viewing the TV show.}$$

We can compute a confidence interval for p, noting that $p =$ expected value of $Be(p)$.

Example 2

Interval estimation for p:
The 95% confidence interval is given by

$$\overline{x} - 1.96\sqrt{\frac{\widehat{\sigma}^2}{n}} \leq p \leq \overline{x} + 1.96\sqrt{\frac{\widehat{\sigma}^2}{n}}.$$

Suppose that 20 out of 200 households in the sample watch the show.
Then:

- $\overline{x} = \frac{20}{200} = 0.1$
- $\widehat{\sigma}^2 = 0.1 \times (1 - 0.1) = 0.09.$

Example 2

The 95% confidence interval is therefore:

$$\overline{x} - 1.96\sqrt{\frac{\widehat{\sigma}^2}{n}} \leq p \leq \overline{x} + 1.96\sqrt{\frac{\widehat{\sigma}^2}{n}}$$

$$0.1 - 1.96\sqrt{\frac{0.09}{200}} \leq p \leq 0.1 + 1.96\sqrt{\frac{0.09}{200}}$$

$$0.06 \leq p \leq 0.14$$

Interpretation:

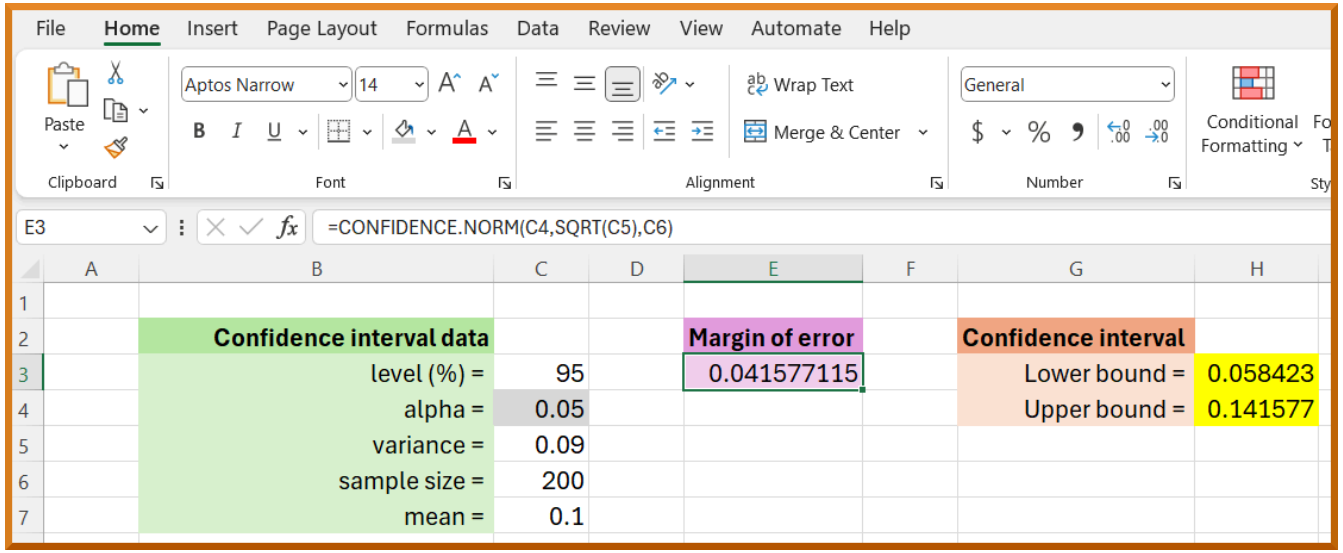
> *We estimate the mean viewership rate p to be between 6% and 14%.*
> *The estimation is based on a procedure which leads to a good answer 95% of the times.*

Or, if you prefer,

> *The estimated mean viewership rate p is between 6% and 14% with confidence level 95%,*

where 'confidence' means 'confidence in the procedure used to compute the CI'.

Example 2 — In Excel



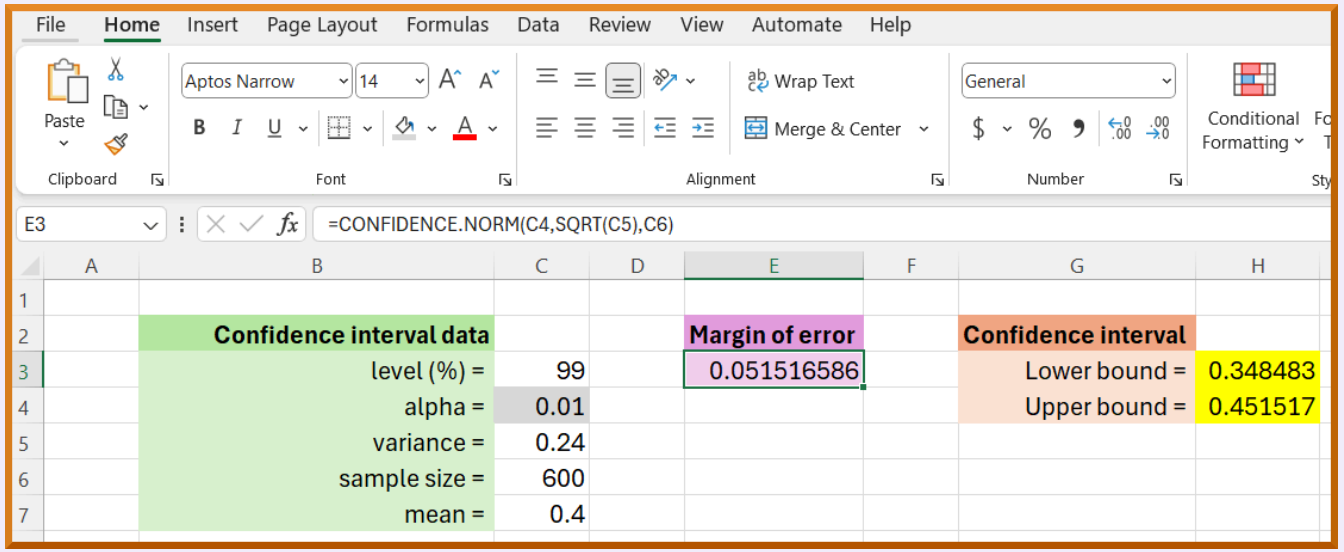
Note:
When the population variance σ^2 is unknown, use an unbiased *sample* estimate of the population variance.

Exercise 2

> *In a constituency with few thousands of voters, we do a poll to determine the approval rating p of party A. We randomly sampled 600 people and scored each person with 1 (supporting party A) or 0 (not supporting party A). It was found that 240 people supported party A, Furthermore, the unbiased sample estimate of the population variance was 0.24 ($= \frac{240}{600}\left(1 - \frac{240}{600}\right)$). Estimate the approval rate p with a 99% confidence interval. Calculate the interval by rounding to the second decimal place.*

Exercise 2 — ANSWER

- Assume that the population distribution is $Be(p)$.
- mean $= \frac{240}{600} = 0.4$



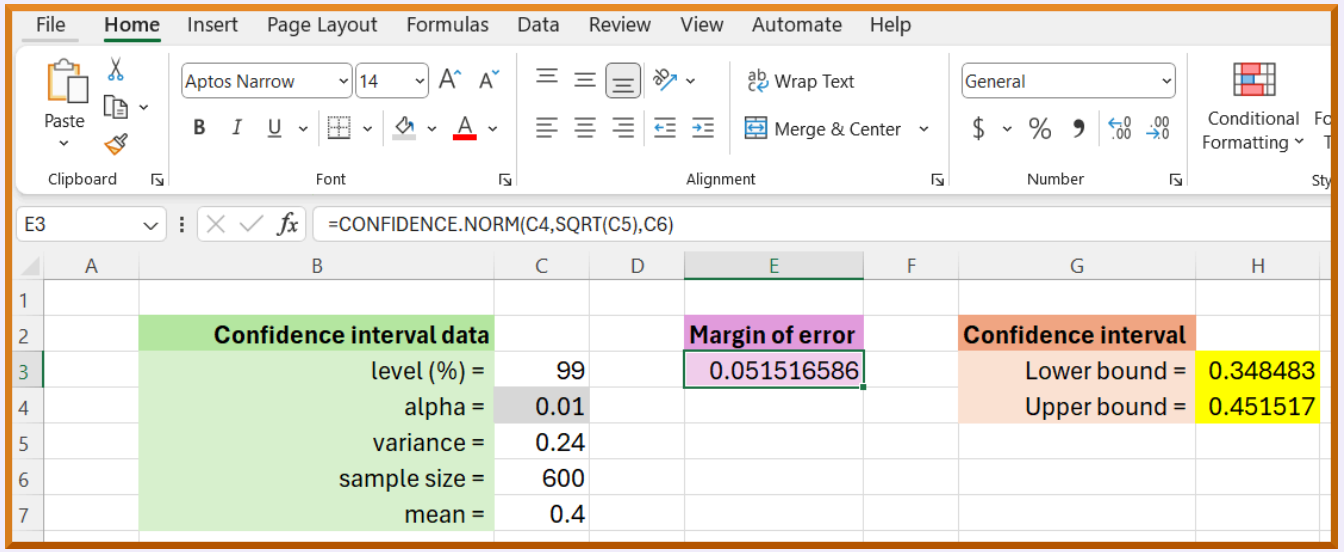
Interpretation:

We estimate the approval rate p to be between 35% and 45%.

The estimation is based on a procedure which leads to a good answer 99% of the times.

Exercise 2 — ANSWER

- Assume that the population distribution is $Be(p)$.
- mean $= \frac{240}{600} = 0.4$



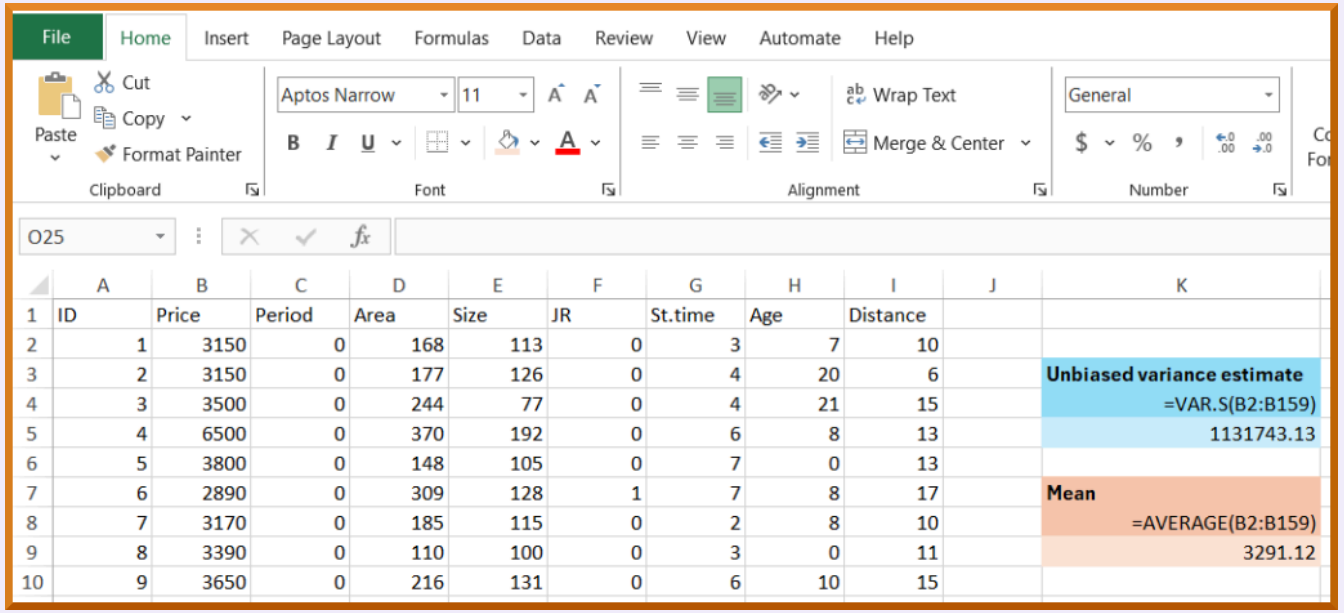
Alternative interpretation:
We estimate the approval rate p to be between 35% and 45% with confidence 99%.
Here, 'confidence' means 'confidence in the procedure used to compute the CI'.

Exercise 3

> *Calculate the 95% confidence interval estimate for the mean of* `Price` *in the data file "HOUSE.csv".*
> *Assume that the sample size is large enough.*
> *Get the upper and lower interval bounds by rounding to the second decimal place.*
> *(Use* `VAR.S` *in Excel to calculate the unbiased variance estimate.)*

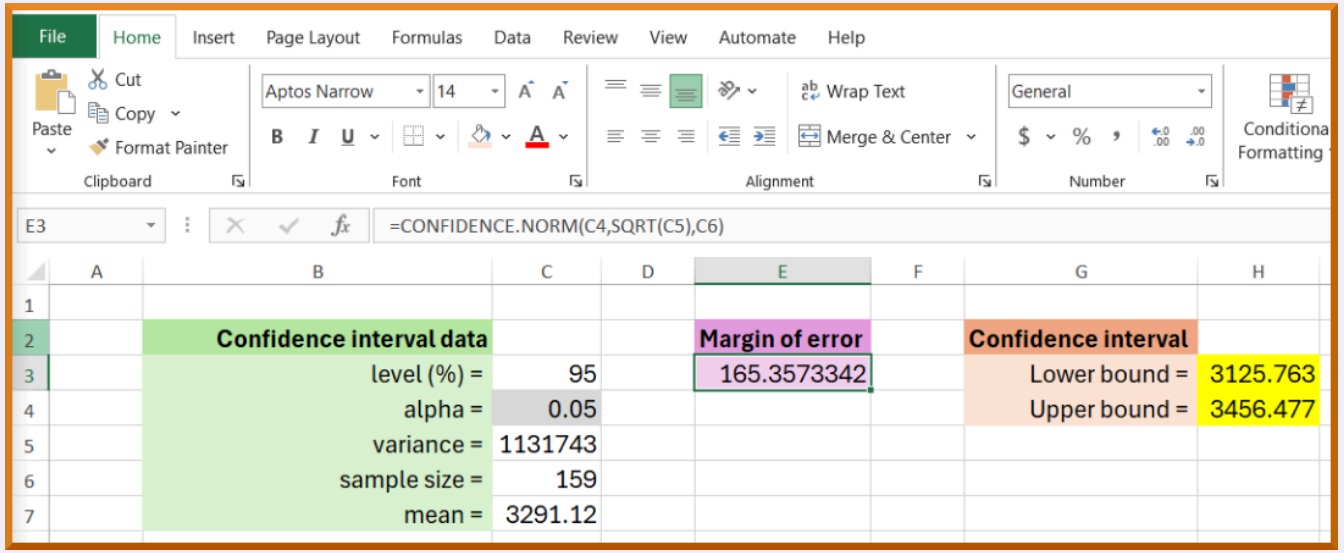
Exercise 3 — ANSWER

First compute the sample mean and variance for `Price`:



	A	B	C	D	E	F	G	H	I	J	K
1	ID	Price	Period	Area	Size	JR	St.time	Age	Distance		
2	1	3150	0	168	113	0	3	7	10		
3	2	3150	0	177	126	0	4	20	6		**Unbiased variance estimate**
4	3	3500	0	244	77	0	4	21	15		=VAR.S(B2:B159)
5	4	6500	0	370	192	0	6	8	13		1131743.13
6	5	3800	0	148	105	0	7	0	13		
7	6	2890	0	309	128	1	7	8	17		**Mean**
8	7	3170	0	185	115	0	2	8	10		=AVERAGE(B2:B159)
9	8	3390	0	110	100	0	3	0	11		3291.12
10	9	3650	0	216	131	0	6	10	15		

Exercise 3 — ANSWER



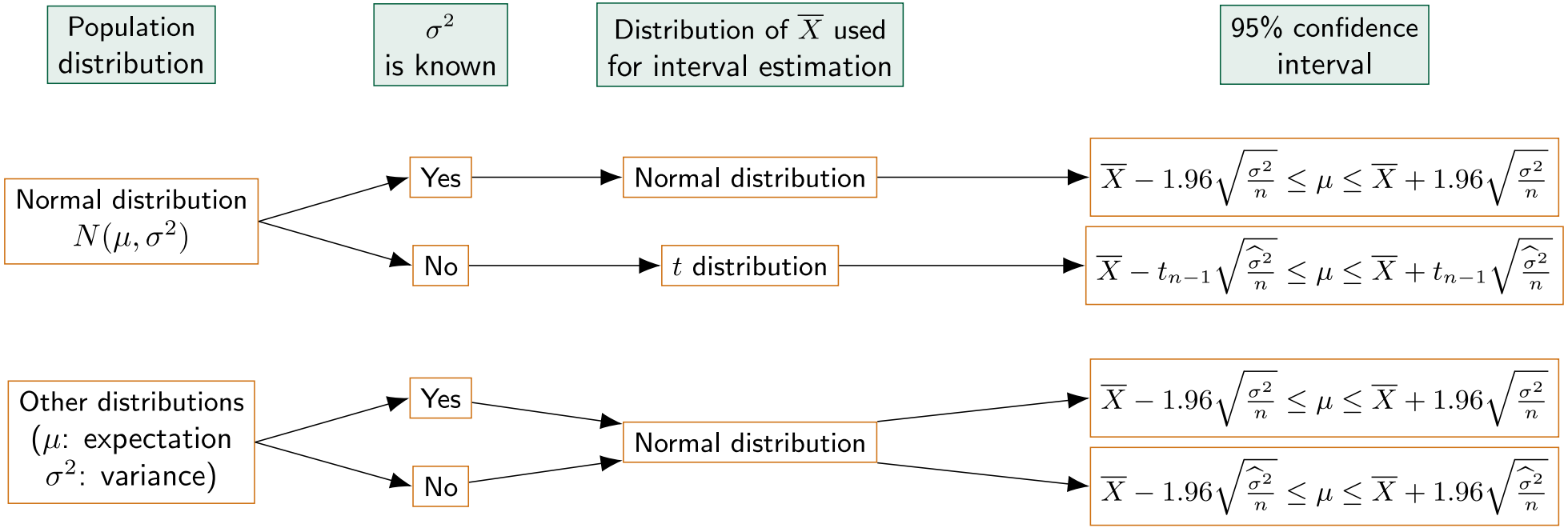
	A	B	C	D	E	F	G	H
1								
2		Confidence interval data			Margin of error		Confidence interval	
3		level (%) =	95		165.3573342		Lower bound =	3125.763
4		alpha =	0.05				Upper bound =	3456.477
5		variance =	1131743					
6		sample size =	159					
7		mean =	3291.12					

Interpretation:
We estimate the house price to be between 3126 and 3456 (in $10,000$ yen units).
The estimation is based on a procedure which leads to a good answer 95% of the times.

Alternative interpretation:
We estimate the house price to be between 3126 and 3456 (in $10,000$ yen units) with confidence 95%.
Here, 'confidence' means 'confidence in the procedure used to compute the CI'.

Interval estimation — Summary

Interval estimation of the population mean μ:



The results above hold under random sampling, and for a large sample size n for non-normal distributions.

Closing remarks

- We covered some basic issues about of data science, especially statistics, in this course.
- The main focus of the course is to understand a rough idea of the contents.
  Thus, some parts were explained roughly or not precisely.
  - It is enough for now to just have an idea.
  - Using formulas would deepen your understanding.
- There are countless statistical methods which were not covered in this course.
- When you study about data science in the future, do remember what was mentioned in this course!