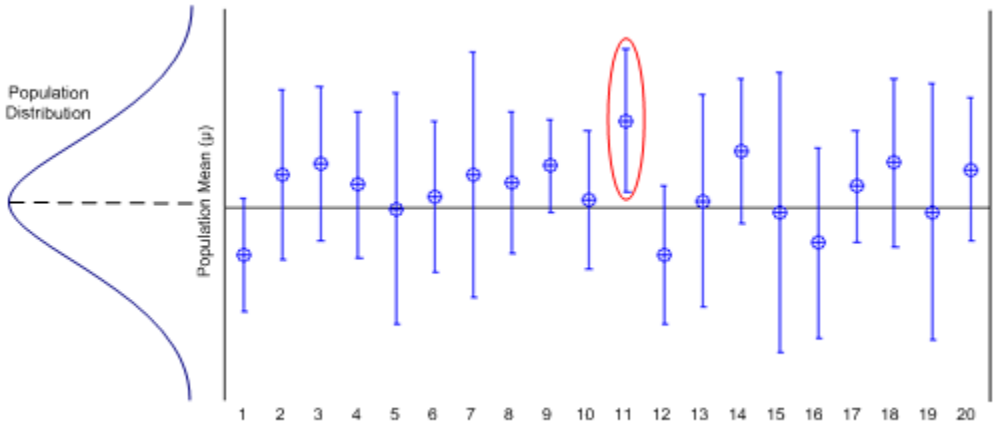
Class 11: Introduction to statistical inference: point and interval estimation



Objective

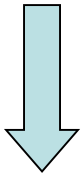
Introduce the basic ideas of statistical inference and point and interval estimation, using a sample to estimate the characteristics of a wider population.

Recommended reading:

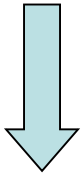
- Chapters 20 and 21 of Peña and Romo (1997)

Statistical inference

Descriptive statistics: the mean age of a sample of 20 PP voters is 55 with standard deviation 5.



Probability Model: The age of a PP voter follows a normal, $N(\mu,\sigma^2)$, distribution.



Inference: We predict that $\mu = 55$. We reject the hypothesis that $\mu < 50$.

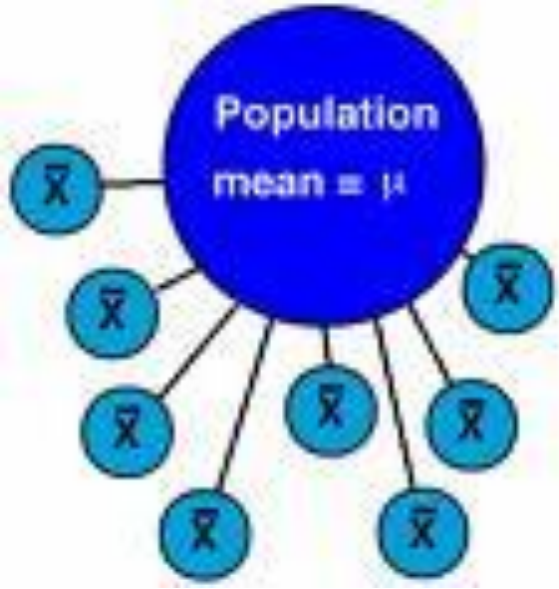
The sampling distribution

Different samples have different means. Before the sample is taken, the sample mean is a variable.

The mean and variance of the sample mean are

$$E[\bar{X}] = \mu \qquad V[\bar{X}] = \sigma^2/N$$

If N is big enough, the sample mean follows a normal distribution.



Have a look at this page which shows the distribution of the sample mean for different sample sizes and different population types

Point estimation

The sample mean $\overline{X}$ is a good estimator of the population mean μ.

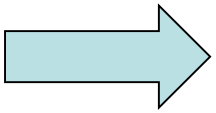
Given a sample, $\overline{x}$ is a point estimate of μ.

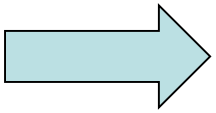
The sample mean has good statistical properties: unbiased, maximum likelihood, etc.

S^2 is also a reasonable estimator of σ^2.

5.4 Interval estimates

We want to find an interval that we are reasonably sure will contain μ.

Wide interval  very imprecise

Narrow interval  more chance of making a mistake

Probability based approach:

- choose a confidence level, e.g. 95% (or 90% or 99%)
- choose variables $L(X_1,\ldots,X_N)$, $U(X_1,\ldots,X_N)$ such that $P(L < \mu < U) = 95\%$
- given the sample data, the 95% confidence interval is

$$(L(x_1,\ldots,x_N),\ U(x_1,\ldots,x_N))$$

Interpretation

If we construct many 95% confidence intervals this way in lots of experiments, 95% of these intervals will contain the parameter that we want to estimate.

This page illustrates the idea.

If we have calculated a 95% confidence interval, it is not true to say that the probability that μ lies in this interval is 0.95.

A 95% confidence interval for a normal mean (known variance or large sample)

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

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

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} < 1.96\right) = 0.95$$

$$P\left(\bar{X} - 1.96\sigma/\sqrt{N} < \mu < \bar{X} + 1.96\sigma/\sqrt{N}\right) = 0.95$$

Given a sample, $x_1,\ldots x_N$, a 95% confidence interval for μ is:

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{N}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{N}}\right)$$

Why 1.96?

Would a 99% confidence interval be wider or narrower?

Examples

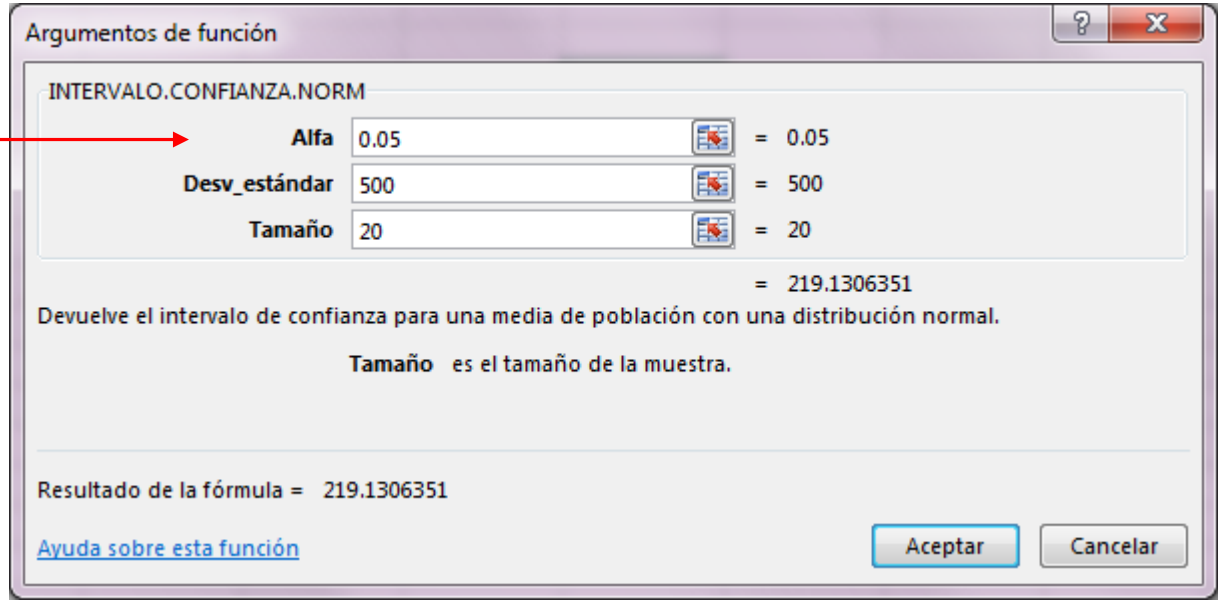
In a sample of 20 Catalans, the mean monthly wage was € 2000. Supposing that the standard deviation of monthly wages is Cataluña is € 500, calculate a 95% confidence interval for the true mean wage.

In a sample of 10 politics students, the mean height was 170cm. If the standard deviation of the heights of Spanish adults is 5cm, calculate a 99% confidence interval for the true mean Spanish height.

Can we do these with Excel?

Calculation via Excel

1 – 0.95



Excel won't calculate the whole interval, just the value of $1.96\frac{\sigma}{\sqrt{N}}$.

We just have to subtract (and add) this to the mean to calculate the interval.

$$2000 \pm 219.13 = (1780.87, 2219.13)$$

A 95% confidence interval for a proportion

Given a sample of size N with sample proportion $\hat{p}$, a 95% confidence interval for the population proportion p is:

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right)$$

This looks like an even more horrible formula!

Examples

In a sample of 100 voters, 45 of them voted for the PSOE in the last elections. Use this information to estimate the true proportion of PSOE voters in these. Give a point estimate and a 95% confidence interval.

20 out of a sample of 30 Americans were in favour of the death penalty. Estimate the true proportion of Americans who are in favour and give a 90% interval.

If we can't do these with Excel I'll be very depressed.

Computation en Excel

Excel doesn’t have a function for calculating confidence intervals for a proportion ...

Computation en Excel

... but we can use the same Excel function as previously with a little trick!

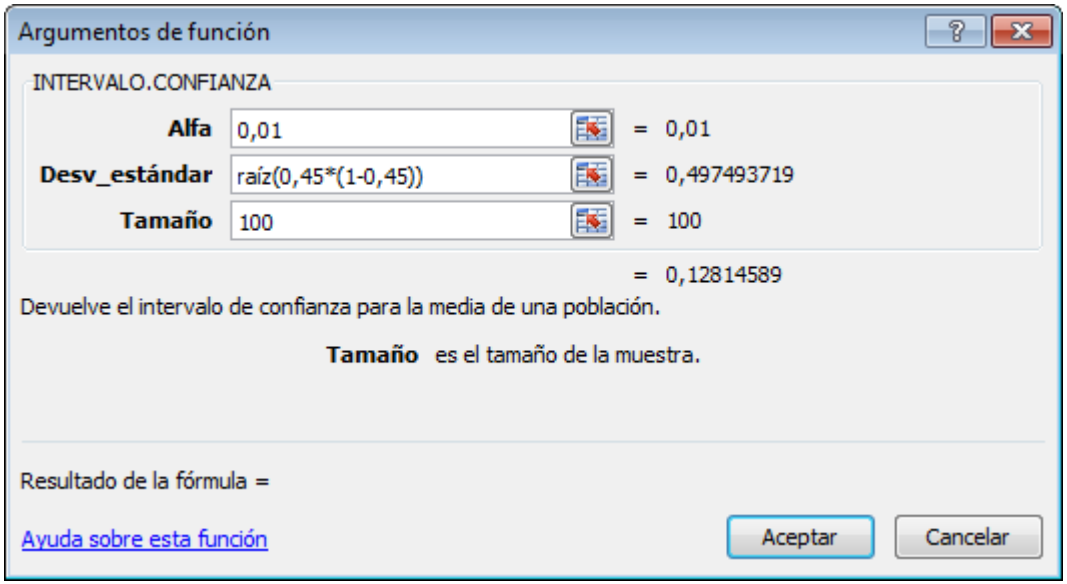
Compare the two formulae:

$$1.96\frac{\sigma}{\sqrt{N}}$$

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{N}}$$

In the 2nd formula, σ is replaced by $\sqrt{\hat{p}(1-\hat{p})}$.

Computation en Excel



Subtracting and summing this to 0.45, gives the confidence interval.

Example

The following data come from the last CIS barometer. The ratings are assumed to come from normal distributions with standard deviations as in the table.

Calculate 95% confidence intervals for the true mean ratings of Alfredo Pérez Rubalcaba and Mariano Rajoy.

Is it reasonable to assume that these are the same?

Why?

	Media	Desviación típica	(N)
Enrique Álvarez Sostres	2.72	2.38	(133)
Joan Baldoví Roda	3.06	2.76	(104)
Uxue Barkos	4.27	2.74	(302)
Alfred Bosch	3.69	2.78	(211)
Rosa Díez	4.33	2.50	(1594)
Josep A. Durán i Lleida	2.63	2.40	(1454)
Josu Erkoreka	2.85	2.53	(491)
Mikel Errekondo	2.50	2.67	(224)
Francisco Jorquera	3.01	2.48	(216)
Cayo Lara	3.88	2.62	(1379)
Ana María Oramas	3.43	2.50	(170)
Alfredo Pérez Rubalcaba	3.40	2.57	(2314)
Mariano Rajoy	2.81	2.69	(2372)
Carlos Salvador	2.28	2.25	(96)

Example

The following table comes from the CIS barometer of 2011.

PREGUNTA 2
Y, ¿cree Ud. que la situación económica actual del país es mejor, igual o peor que hace un año?

	%	(N)
Mejor	5.3	(130)
Igual	35.1	(865)
Peor	57.6	(1418)
N.S.	1.7	(42)
N.C.	0.3	(8)
TOTAL	100.0	(2463)

Calculate a 95% confidence interval for the true proportion of Spanish adults who think that the economic situation worsened over this year.

Example

The following news item was reported in The Daily Telegraph online on 8th May 2010.

General Election 2010: half of voters want proportional representation

Almost half of all voters believe Britain should conduct future general elections under proportional representation, a new poll has found.

The ICM survey for The Sunday Telegraph revealed that 48 per cent backed PR – a key demand of the Liberal Democrats. Some 39 per cent favoured sticking with the current "first past the post system" for electing MPs.
The public was split when asked how they wanted Britain to be governed after Thursday's general election resulted in a hung parliament, with the Conservatives, on 306 seats, the largest party.
Some 33 per cent wanted a coalition government between the Tories and the Liberal Democrats, while 32 per cent thought Nick Clegg's party should team up with Labour.
Just 18 per cent favoured a minority Tory government.
…
*ICM Research interviewed a random sample of 532 adults aged 18+ by telephone on 8 May 2010.

Calculate a 95% confidence interval for the true proportion of adults who are in favour of proportional representation.

Example

The following is taken from *Electrometro.com: La web de encuestas electorales en España*.

The PSdG could renew its coalition with BNG in A Coruña (Antena 3)

Lunes 9 Mayo 2011

According to the results of the survey carried out by TNS-Demoscopia for Antena 3 and Onda Cero, the **PP** will get **38.7%** of the votes in **A Coruña**, which will give them **12-13 councilmen** as opposed to the 10 they have at the moment. On the other hand, the **PSdG** will lose 5.6 point with respect to the previous elections and will obtain **29,4%** of the votes which will give them **9 or 10 councilmen.** The **BNG** will obtain **5 or 6 councilmen** by getting **17.7%** of the votes, 3 points less than four years ago.
FICHA TÉCNICA: 500 interviews carried out on **3rd and 4th of May** by **TNS-Demoscopia** for **Antena 3** and **Onda Cero**.

Calculate a 95% confidence interval for the percentage of votes that the Partido Popular (PP) will obtain in A Coruña, given the survey results..

A 95% confidence interval for a normal mean (unknown variance)

Until now, we have assumed a known variance when constructing a confidence interval. In practice, this may be unrealistic.

What should we do?

If the sample size is large (> 30), we can construct the same, normal, confidence interval as earlier, simply substituting the true standard deviation by the sample standard deviation.

A 95% confidence interval for a normal mean (unknown variance)

If the sample is small, we can use a *Student's t interval*:

$$\left(\bar{x} - t_{N-1}(0.975)\frac{s}{\sqrt{N}},\ \bar{x} + t_{N-1}(0.975)\frac{s}{\sqrt{N}}\right)$$

Another repulsive formula but fortunately, we can still do the calculations with Excel.

Example

Data are available on the prison sentences of 19 murderers in Spain. The mean and standard deviation of the prison sentences are 72.7 and 10.2 months respectively.

Calculate a 95% interval for the mean duration of murder sentences in Spain

Calculation via Excel

As with the known variance case, simply add and subtract the result from the sample mean to get the interval.