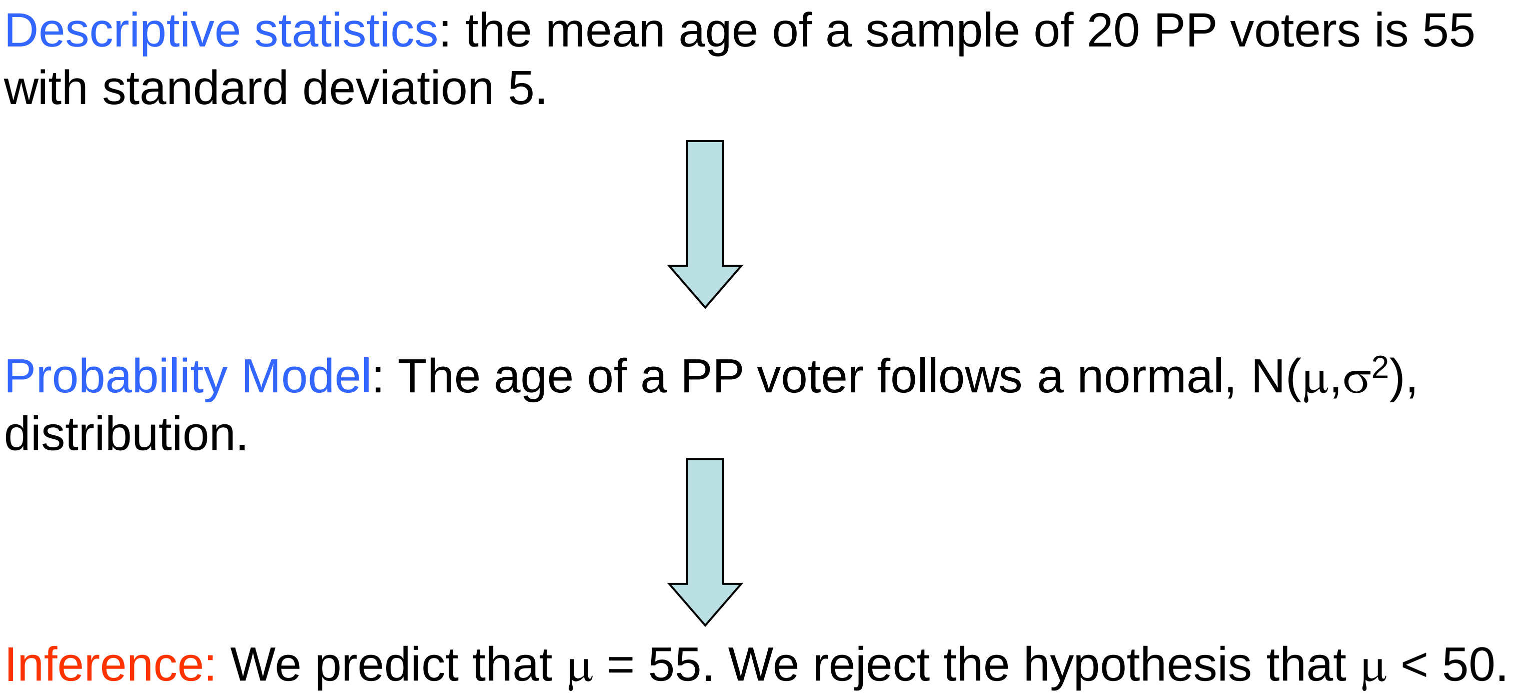
Chapter 5: Introduction to statistical inference

1. Outline and objectives
2. Statistics and sampling distribution
3. Point estimation
4. Interval estimation
5. Hypothesis tests: means, proportions, independence

Recommended reading:

- You tube videos on confidence intervals, hypothesis tests …

5.1 Outline and objectives



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(μ,σ^2), distribution.

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

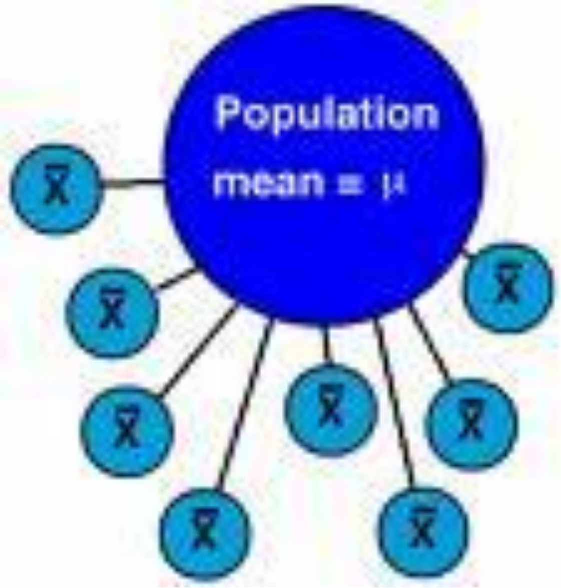
5.2 Statistics and the sampling distribution

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 the following page:
http://www.stat.tamu.edu/~west/ph/sampledist.html

5.3 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.

http://www.ruf.rice.edu/~lane/stat_sim/conf_interval/index.html

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)

$$\begin{aligned}
\bar{X} &\sim N\left(\mu, \frac{\sigma^2}{N}\right) \\
\frac{\bar{X}-\mu}{\sigma/\sqrt{N}} &\sim N(0,1) \\
P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} < 1.96\right) &= 0.95 \\
P\left(\bar{X} - 1.96\sigma/\sqrt{N} < \mu < \bar{X} + 1.96\sigma/\sqrt{N}\right) &= 0.95
\end{aligned}$$

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

$$(\bar{x} - 1.96\sigma/\sqrt{N}, \bar{x} + 1.96\sigma/\sqrt{N})$$

Why 1.96?

What would a 90% confidence interval look like?

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.

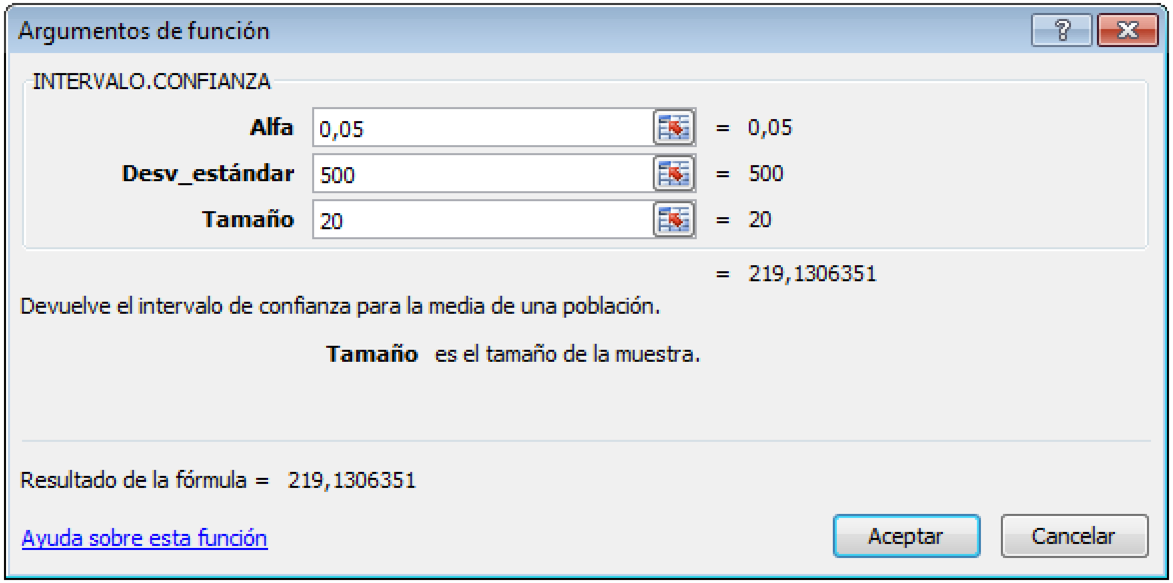
In a sample of 10 politics students, the mean height was 170cm. If the standard deviation of the heights of Spanish adults is 5cm, calculate a 99% confidence interval for the true mean Spanish height.

Computation in Excel

n	20		Data
mean	2000		
sd	500		
alpha	0,05		Computación de z
alpha/2	0,025		
1-alpha/2	0,975		
z	1,96		DISTR.NORM.ESTAND.INV(0,975)
z*sigma/root(n)	219,13		B8*B3/RAIZ(B1)
interval	1780,87	2219,13	
	B2-B10	B2+B10	

Is there a faster way to do this?

In Excel 2010 you can use INTERVALO.CONFIANZA.NORM



Argumentos de función

INTERVALO.CONFIANZA

Alfa 0,05 = 0,05

Desv_estándar 500 = 500

Tamaño 20 = 20

= 219,1306351

Devuelve el intervalo de confianza para la media de una población.

Tamaño es el tamaño de la muestra.

Resultado de la fórmula = 219,1306351

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

A 95% confidence interval for a proportion

$$\begin{aligned} X &\sim Bi(N,p) \Rightarrow \\ X &\approx N\left(Np, Np(1-p)\right) \\ \hat{p}=\frac{X}{N} &\approx N\left(p, \frac{p(1-p)}{N}\right) \end{aligned}$$

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

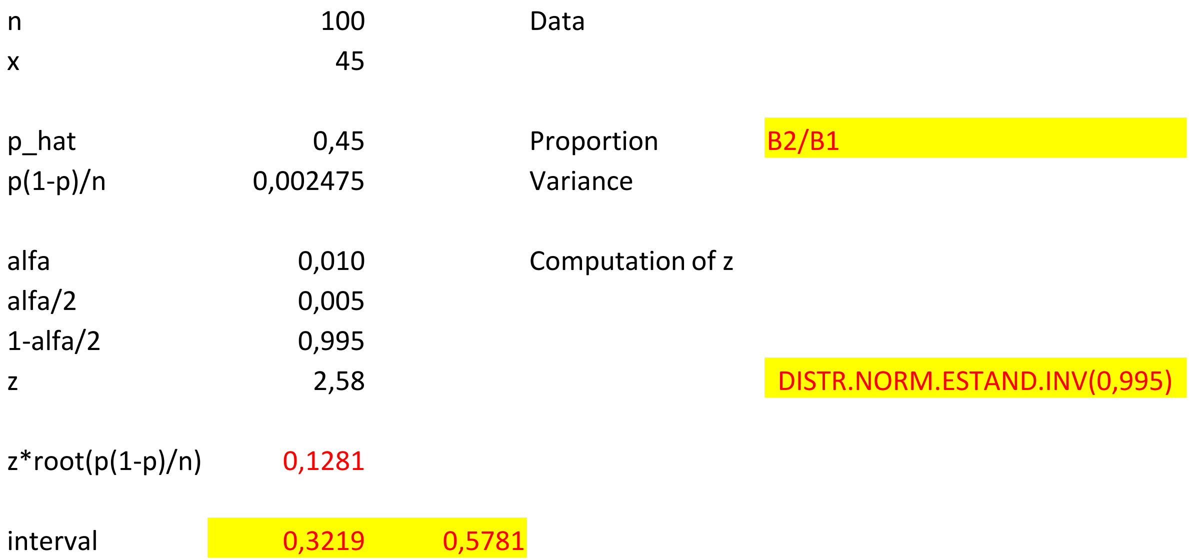
$$\left(\hat{p}-1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}, \hat{p}+1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right)$$

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.

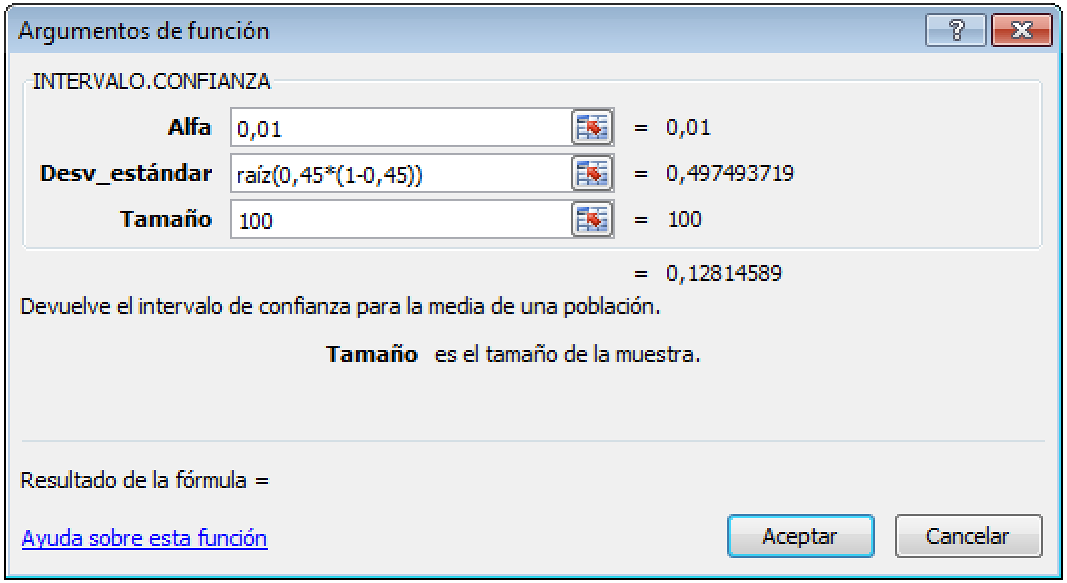
Computation en Excel



n	100	Data	
x	45		
p_hat	0,45	Proportion	B2/B1
p(1-p)/n	0,002475	Variance	
alfa	0,010	Computation of z	
alfa/2	0,005		
1-alfa/2	0,995		
z	2,58		DISTR.NORM.ESTAND.INV(0,995)
z*root(p(1-p)/n)	0,1281		
interval	0,3219	0,5781	

Could we use INTERVALO.CONFIANZA?

Yes!



Argumentos de función

INTERVALO.CONFIANZA

Alfa 0,01 = 0,01

Desv_estándar raíz(0,45*(1-0,45)) = 0,497493719

Tamaño 100 = 100

= 0,12814589

Devuelve el intervalo de confianza para la media de una población.

Tamaño es el tamaño de la muestra.

Resultado de la fórmula =

The standard deviation is √ (p^ x (1-p^)).

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

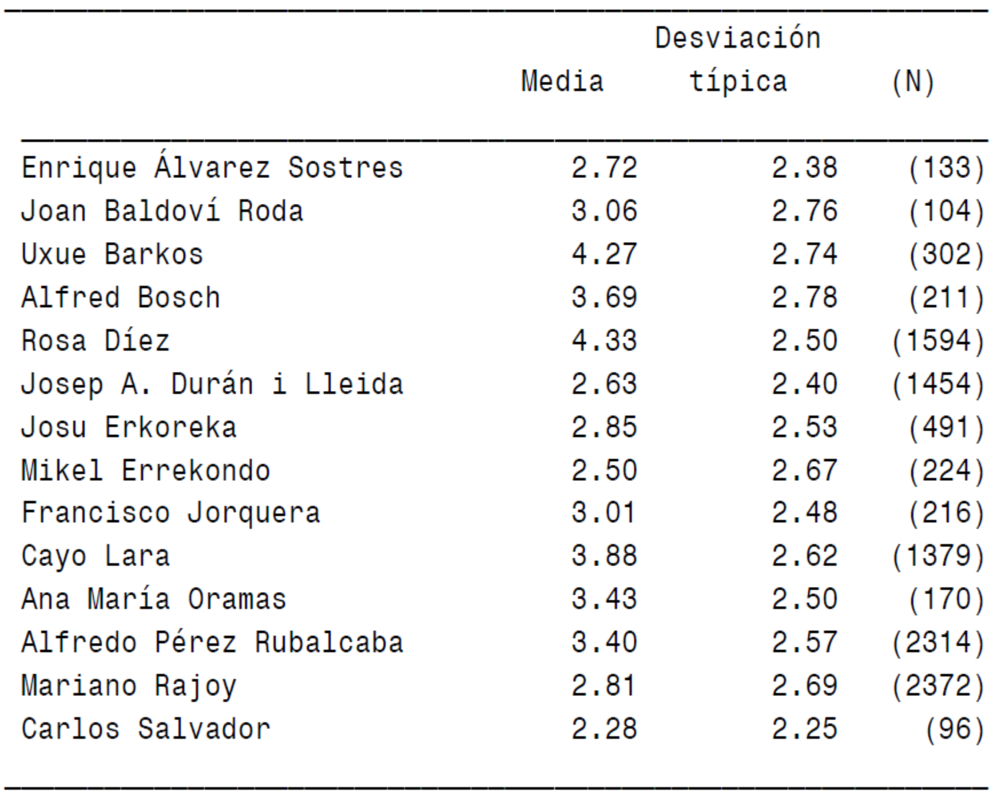
Example

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?

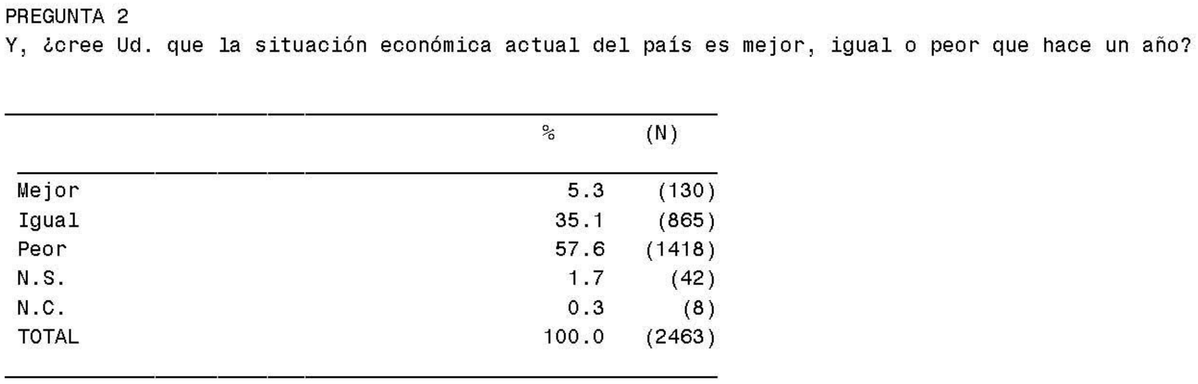
Why?



	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.

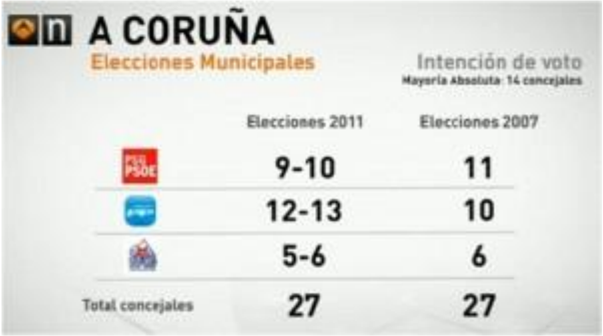
Example

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**.



A CORUÑA
Elecciones Municipales

Intención de voto
Mayoría Absoluta: 14 concejales

	Elecciones 2011	Elecciones 2007
PSOE	9-10	11
	12-13	10
	5-6	6
Total concejales	27	27

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

Additional Material

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.

If the sample is small, we can use a *Student’s t interval*.

What is t?

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1}(0,975)$$

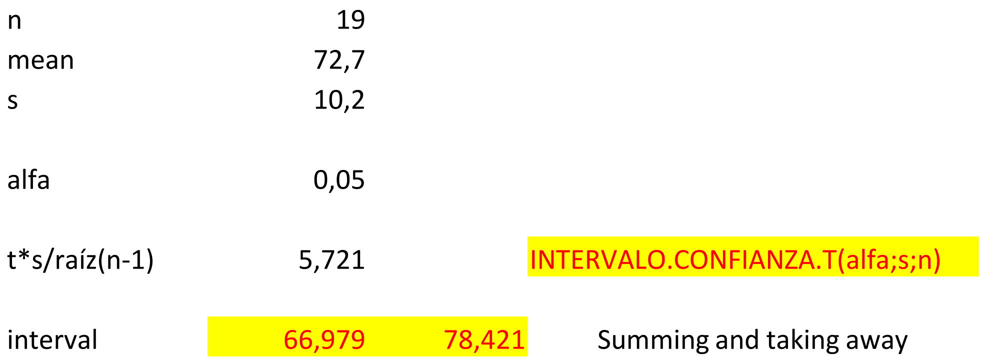
This looks tough, but is easy with Excel 2010 …

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

We can use the function INTERVALO.CONFIANZA.T



n	19		
mean	72,7		
s	10,2		
alfa	0,05		
t*s/raíz(n-1)	5,721		INTERVALO.CONFIANZA.T(alfa;s;n)
interval	66,979	78,421	Summing and taking away

The interval is from 66.98 to 78.42 months.

With the original data it is even easier …

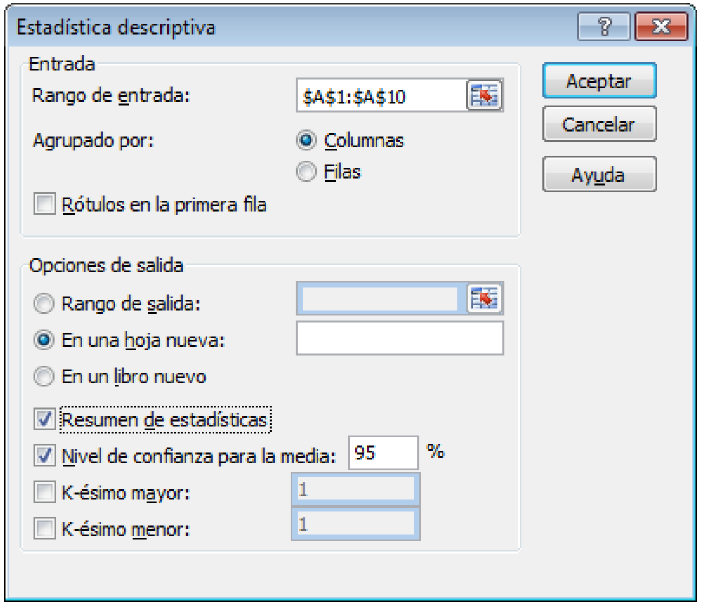
Example

A small survey was carried out in order to estimate the mean wage of Spanish bankers. A sample of 10 bankers gave the following results (in thousands of euros).

1200, 1000, 1500, 800, 750, 2400, 1000, 1600, 700, 600

Calculate a 95% confidence interval for the true mean wage of Spanish bankers.

We can use the Descriptive Statistics option in Data Analysis in Excel.



Estadística descriptiva

Entrada

Rango de entrada: $A$1:$A$10

Agrupado por: ● Columnas ○ Filas

□ Rótulos en la primera fila

Opciones de salida

○ Rango de salida:

● En una hoja nueva:

○ En un libro nuevo

☑ Resumen de estadísticas

☑ Nivel de confianza para la media: 95 %

□ K-ésimo mayor: 1

□ K-ésimo menor: 1

Columna1	
Media	1155,00
Error típico	173,92
Mediana	1000,00
Moda	1000,00
Desviación estándar	549,97
Varianza de la muestra	302472,22
Curtosis	1,95
Coeficiente de asimetría	1,40
Rango	1800,00
Mínimo	600,00
Máximo	2400,00
Suma	11550,00
Cuenta	10,00
Nivel de confianza(95,0%)	393,43

Summing and taking away gives the interval.

(€761570, €1548430)

761,57	1548,43

A 95% interval for the difference between two normal means (paired data)

What are paired data?

Example

	Año	
Banker	2012	2013
1	1300	1200
2	1100	1000
3	1200	1500
4	900	800
5	800	750
6	2000	2400
7	1100	1000
8	1500	1600
9	700	700
10	500	600

We have the wages for the same bankers in 2012 and 2013. Suppose that we wish to estimate the average increase of bankers' wages in this period.

	Year		
Banker	2012	2013	Difference
1	1300	1200	-100
2	1100	1000	-100
3	1200	1500	300
4	900	800	-100
5	800	750	-50
6	2000	2400	400
7	1100	1000	-100
8	1500	1600	100
9	700	700	0
10	500	600	100
	1110	1155	45

Calculate the average wage each year and calculate the difference:

1155-1110 = 45

or calculate the wage increases and calculate the mean:

(-100 -100 + 300 + … + 100)/10 = 45

A reasonable point estimate of the average increase in bankers’ wages is €45000.
How can we calculate a confidence band?

Banker	Difference
1	-100
2	-100
3	300
4	-100
5	-50
6	400
7	-100
8	100
9	0
10	100
	45

Just look at the sample of differences.

We have a single sample and we can just use a Student’s t interval.

Columna1	
Media	45,00
Error típico	56,98
Mediana	-25,00
Moda	-100,00
Desviación estándar	180,20
Varianza de la muestra	32472,22
Curtosis	0,21
Coeficiente de asimetría	1,15
Rango	500,00
Mínimo	-100,00
Máximo	400,00
Suma	450,00
Cuenta	10,00
Nivel de confianza(95,0%)	128,91

The interval is 45 ± 128,91 thousands of euros, i.e. (- € 83910, €173910).

It seems plausible that there has been no real changes in bankers’ mean wages in this period.