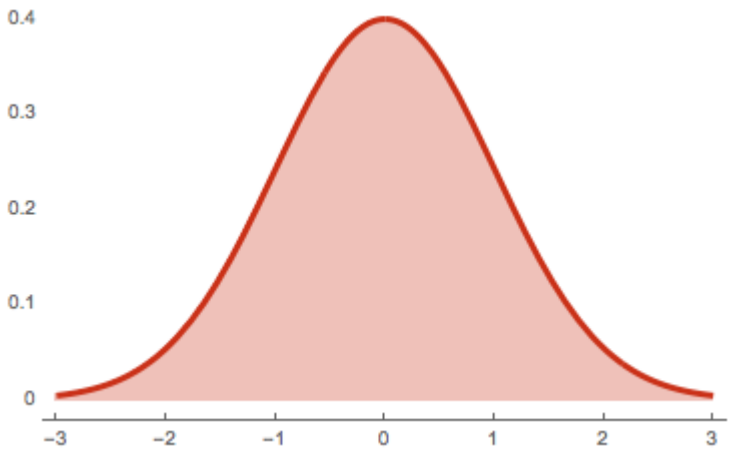
Class 10: Random variables and probability models



Objective

Introduce the concept of random variables and some classes of variable which occur frequently in many real life situations.

Random variables

In the previous questions, we asked questions about specific events:

What is the chance that a Madrileño has never voted in a municipal election?

Now we want to change the question:

What is the distribution of the number of times that Madrileños have voted?

On average, how many times do Madrileños vote in municipal elections?

These are questions about random variables.

Frequency tables and probability distributions

In class 3, we took a sample of 60 Madrileños and looked at how many times they voted.

We represented the results in this table.

What if we looked at the whole population of Madrileños?

Times voted	Absolute frequency (n)	Cumulative frequency (N)	Relative frequency (f)	Cumulative relative frequency (F)
0	4	4	4/60 = 0,0667	0,0667
1	10	4+10 = 14	0,1667	14/60 = 0,2333
2	12	4+10+12 = 26	0,2000	0,4333
3	15	41	0,2500	0,6833
4	11	52	0,1833	0,8667
5	5	57	0,0833	0,9500
6	1	58	0,0167	0,9667
7	1	59	0,0167	0,9833
8	1	60	0,0167	1,0000
>8	0	60	0,0000	1,0000
Total	60		1,0000	

Frequency tables and probability distributions

What if we looked at the whole population of Madrileños and asked what the chance is of a randomly chosen person having voted x times.

This is like making a frequency table where we sample everyone in a very big (essentially infinite) population.

Times voted (X)	Probability	Cumulative probability
0	$P(X = 0)$	$P(X \leq 0)$
1	$P(X = 1)$	$P(X \leq 1)$
2	$P(X = 2)$	$P(X \leq 2)$
3	$P(X = 3)$	$P(X \leq 3)$
4	$P(X = 4)$	$P(X \leq 4)$
5	$P(X = 5)$	$P(X \leq 5)$
6	$P(X = 6)$	$P(X \leq 6)$
7	$P(X = 7)$	$P(X \leq 7)$
8	$P(X = 8)$	$P(X \leq 8)$
> 8	$P(X > 8)$	1
Total	1	---

Medians, means, standard deviations etc.

Before, we asked, “On average, how many times have the people in the sample voted”. Now we ask about the population.

The population mean is written μ or E[X]:

$\mu = 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2) + \ldots$

The median is the first point such that $P(X \leq x)$ is at least 50%.

We could also calculate the variance (σ^2 or V[X]) or standard deviation (σ or SD[X]).

Times voted (X)	Probability	x P(X = x)
0	P(X = 0)	0
1	P(X = 1)	P(X=1)
2	P(X = 2)	2P(X=2)
3	P(X = 3)	3P(X=3)
4	P(X = 4)	4P(X=4)
5	P(X = 5)	5P(X=5)
6	P(X = 6)	6P(X=6)
7	P(X = 7)	7P(X=7)
8	P(X = 8)	8P(X=8)
> 8	P(X > 8)	⋮
Total	1	μ

Some typical probability models

Discrete models	Continuous models
Coin tossing models: Bernoulli, geometric and binomial distributions.	The normal distribution.

Bernoulli trials

A Bernoulli model is an experiment with the following characteristics:

- In each trial, there are only two possible results, success ($B = 1$) and failure ($B = 0$).
- The result obtained in each trial is statistically independent of the previous results.
- The probability of success is constant, $P(B=1) = p$, and does not change from one trial to the next.

The geometric distribution

Suppose we have a Bernoulli model. What is the distribution of the number of failures, F, before the first success?

- P($F=0$) = P(0 failures before the 1st success) = p
- P($F=1$) = P(failure, success) = $(1-p)p$
- P($F=2$) = P(failure, failure, success) = $(1-p)^2\, p$
- P($F=f$) = P(f failures before the 1st success) = $(1-p)^f\, p$ for f = 0, 1, 2, …

The distribution of F is called the geometric distribution with parameter p.

$$E[F] = 1-p/p \qquad V[F] = (1-p)/p^2$$

Example

On average, one in every ten members of the CCOO union is a delegate.

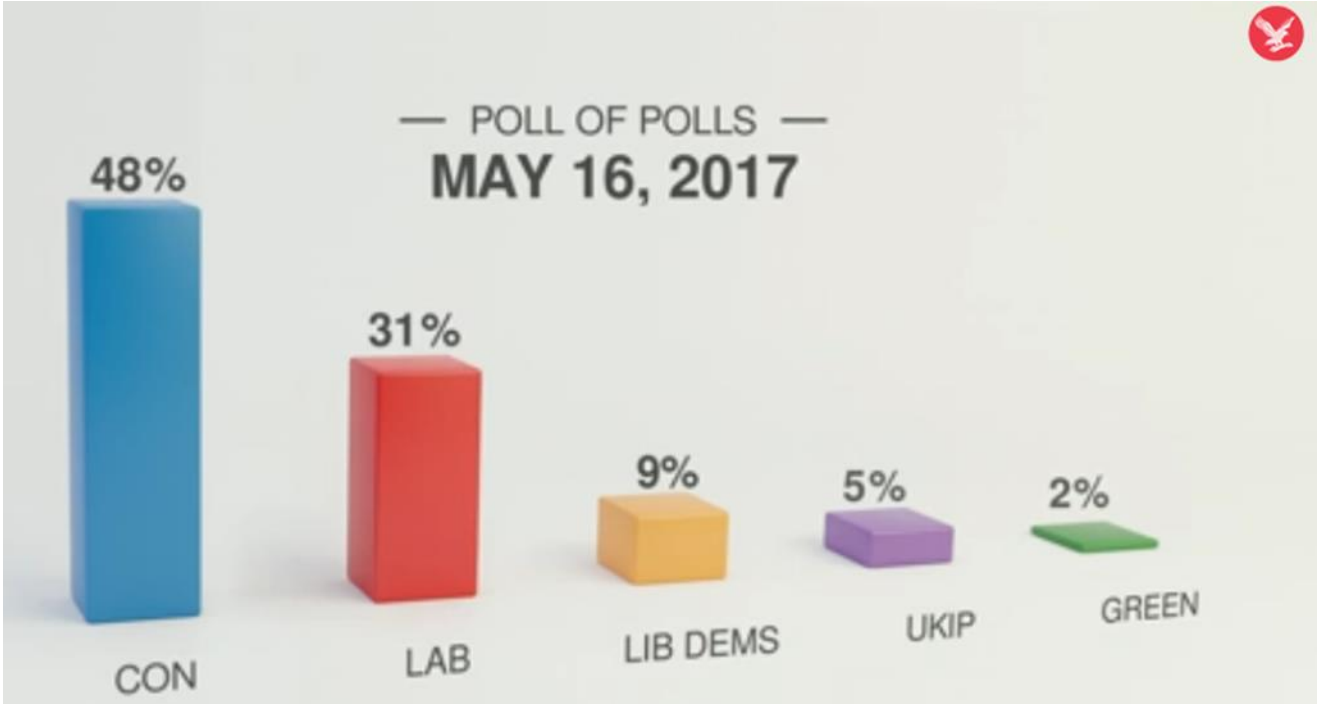
In independent interviews with randomly chosen CCOO members, what is the probability that the first delegate will be the fourth person interviewed?

0.9 x 0.9 x 0.9 x 0.1

Exercise

The following graphic taken from **The Independent** newspaper reflects the results of a latest poll of polls (May 16th 2017) for voting intentions in the next UK elections of June 8th 2017.

Assuming these results are representative of public opinion, what is the probability that when people are consulted independently, the first Labour voter is the third person consulted?



The binomial distribution

Suppose we have a Bernoulli model. What is the distribution of the number of successes, X, in n trials?

$$P(X = x) = C_x^n p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n$$

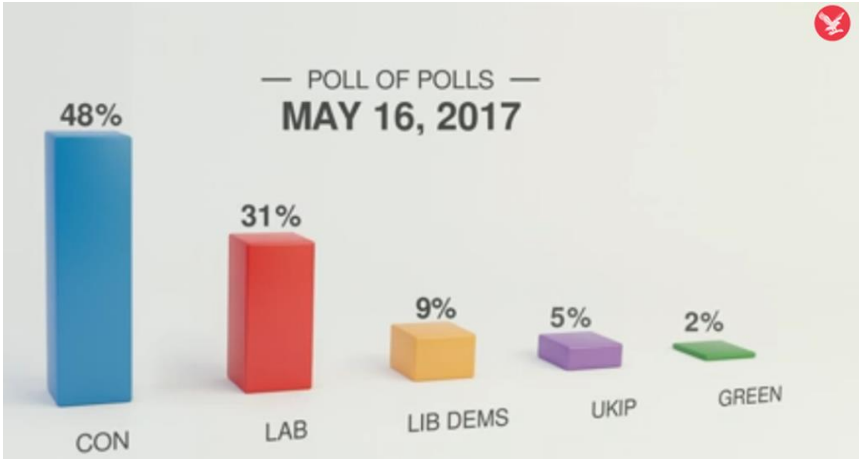
Just ignore this formula. It's only there to frighten you!

The distribution of X is called the binomial distribution with parameters n and p.

$E[X] = np$ $\quad$ $V[X] = npq$

Example

If we consult 100 voters at random, How many Conservatives would we expect to see?



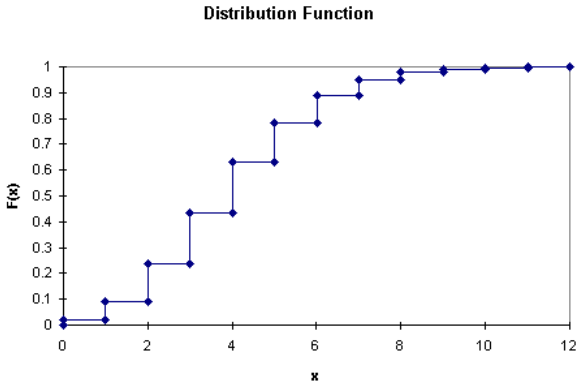
Out of 4 randomly chosen voters, what is the probability that they all vote Conservative?

Don’t try to use the formula! It is incomprehensible.

Continuous Random Variables

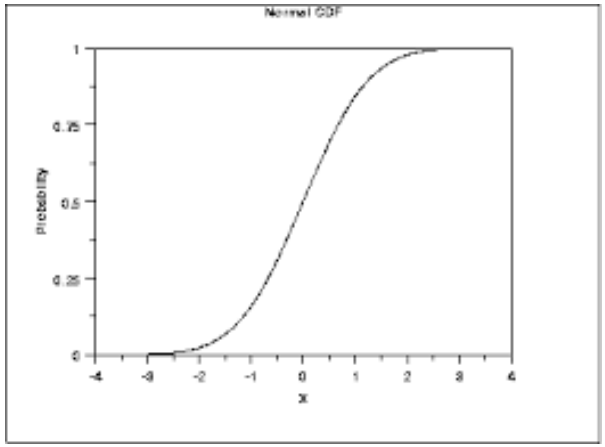
For a discrete variable X, the *cumulative distribution function*, $F(x) = P(X \leq x)$, is a step function:

$$F(x) = \Sigma_{i:x_i \leq x} P(X=x_i)$$



For a continuous variable, the cdf is a smooth, non decreasing function.

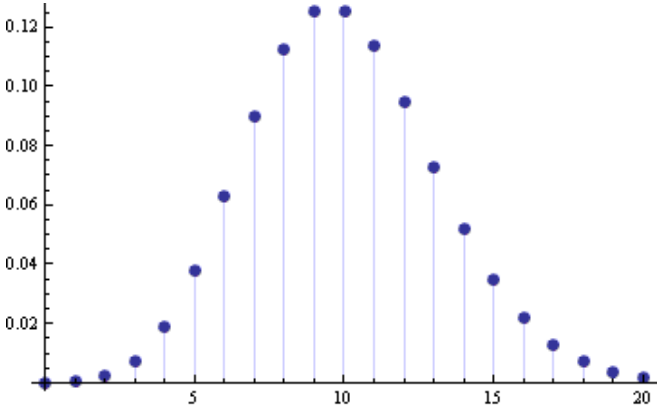
- $0 \leq F(x) \leq 1$
- $F(-\infty) = 0$
- $F(x) \leq F(x+h)$ for $h > 0$
- $F(\infty) = 1$



The density function

For a discrete variable X, the *probability mass function is P(X = x),* which is positive at a discrete set of values x_1, x_2, …

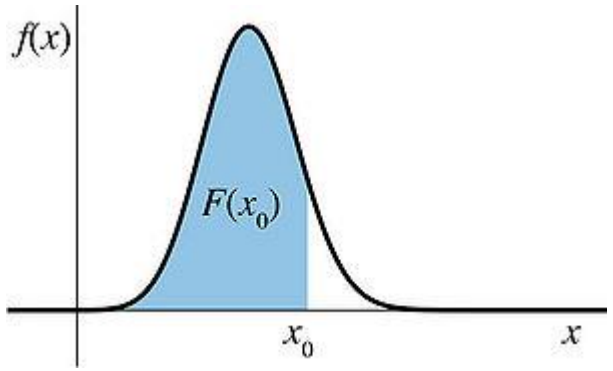
$$0 \le P(X=x) \le 1, \quad \Sigma_i P(X=x_i) = 1$$



For a strictly continuous variable, *P(X=x) = 0*!

Instead we have a *density function*, *f(x).*

- $0 \le f(x)$
- The area under the density up to x is the same as $F(x) = P(X \le x)$.
- The area under the whole density is 1.



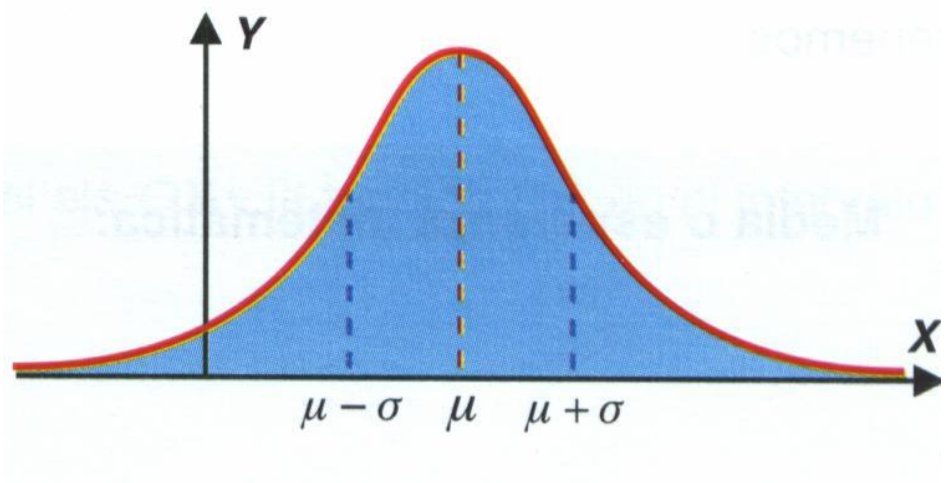
The normal or gaussian distribution

Many variables have a bell shaped density.

Examples:

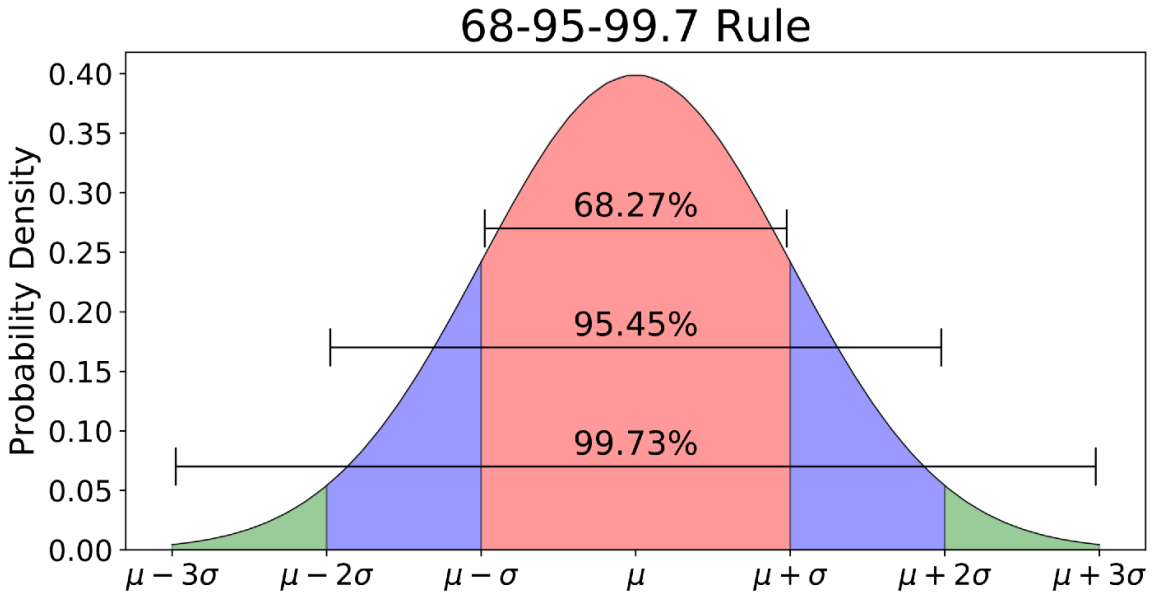
- Weights of a population of the same age and sex.
- Heights of the same population.
- The grades in a course (*urban myth*).

To say that a continuous variable X, has a normal distribution with mean μ and standard deviation σ , we write:



$X \sim N(\mu,\sigma^2)$

Properties of the normal distribution



For any normal distribution, the chance that an observation is less than 2 standard deviations from the mean is 95.45%.

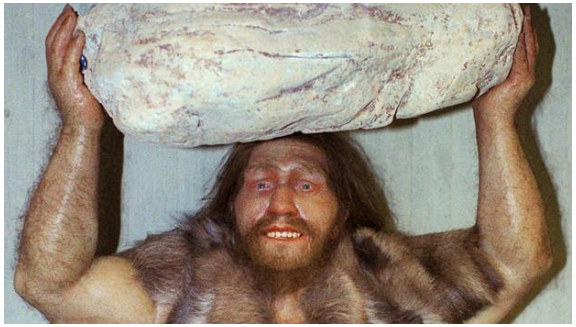
Remember our empirical rule for determining outliers.

Calculating probabilities for the normal distribution

In the old days, we would have to transform to a standard normal distribution:

$$Z = (X-\mu)/\sigma$$

Then probabilities were calculated using tables …



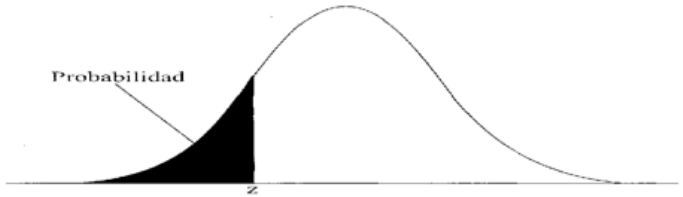


Tabla 3. *Probabilidad de que una variable normal de media cero y desviación típica u un valor menor que z*

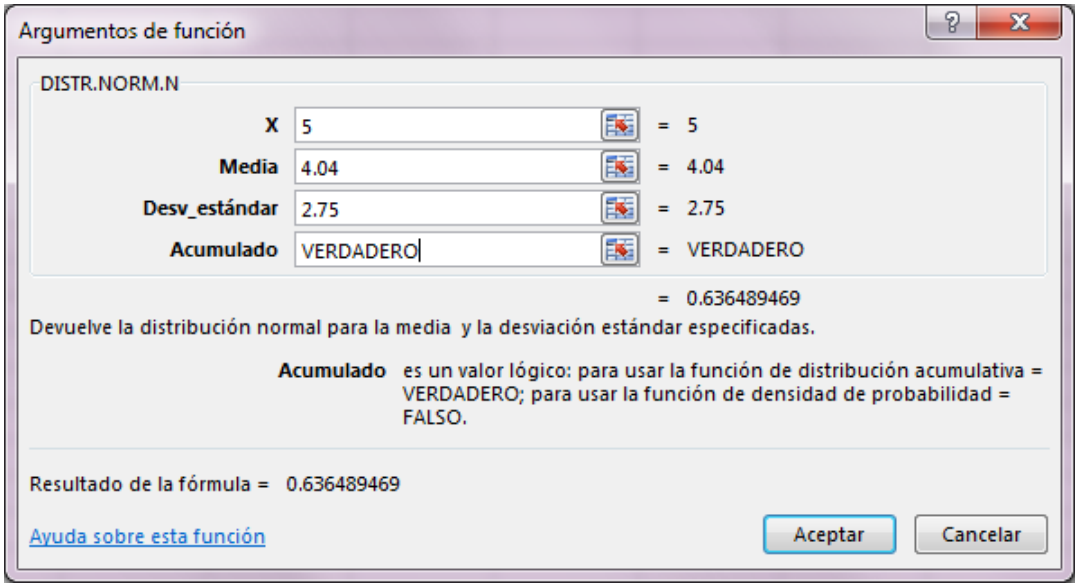
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
—3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003
—3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004
—3,2	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005
—3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007
—3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010
—2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014
—2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020
—2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027
—2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037
—2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049
—2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066
—2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087
—2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113
—2,1	0,0179	0,0174	0,0170	0,0166	0,016	0,0158	0,0154	0,0150	0,0146
—2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188
—1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239
—1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301
—1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375
—1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465

Calculating probabilities for the normal distribution

Nowadays, we simply use a computer package like Excel.

According to the CIS barometer of July 2018, Pedro Sanchez obtained a mean rating of 4.04 with a standard deviation of 2.75. Assuming that the ratings follow a normal distribution, what is the probability that a random chosen Spaniard gives Pedro Sanchez a “pass” mark of at least 5?

p = 1 – 0.6365 = 0.3635.



Exercise

According to the last CIS survey, the mean level of satisfaction with Mariano Rajoy is 3.09 with standard deviation 2.5. If these evaluations follow a normal distribution and a person is chosen at random, then the probability that they give Rajoy a rating of less than 3.09 is:

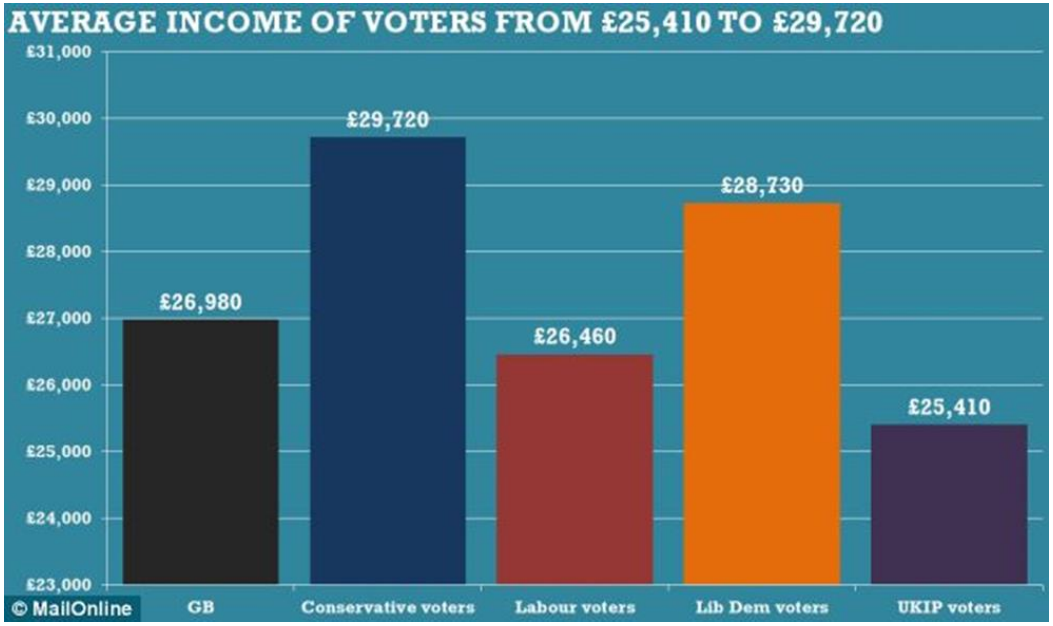
a) 0.5.
b) 0.
c) 1.236
d) 1.

What is the probability his rating is exactly 3.09?

You can do this without either Excel or tables!

Exercise

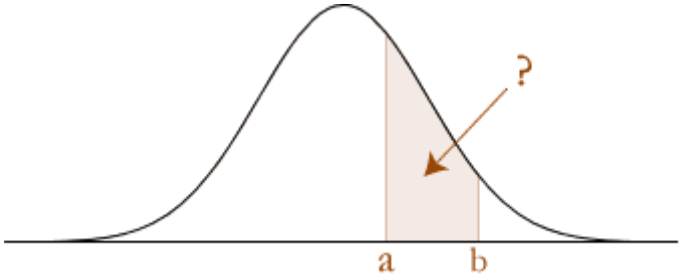
The following graphic comes from the 3rd March 2014 edition of The Mail.

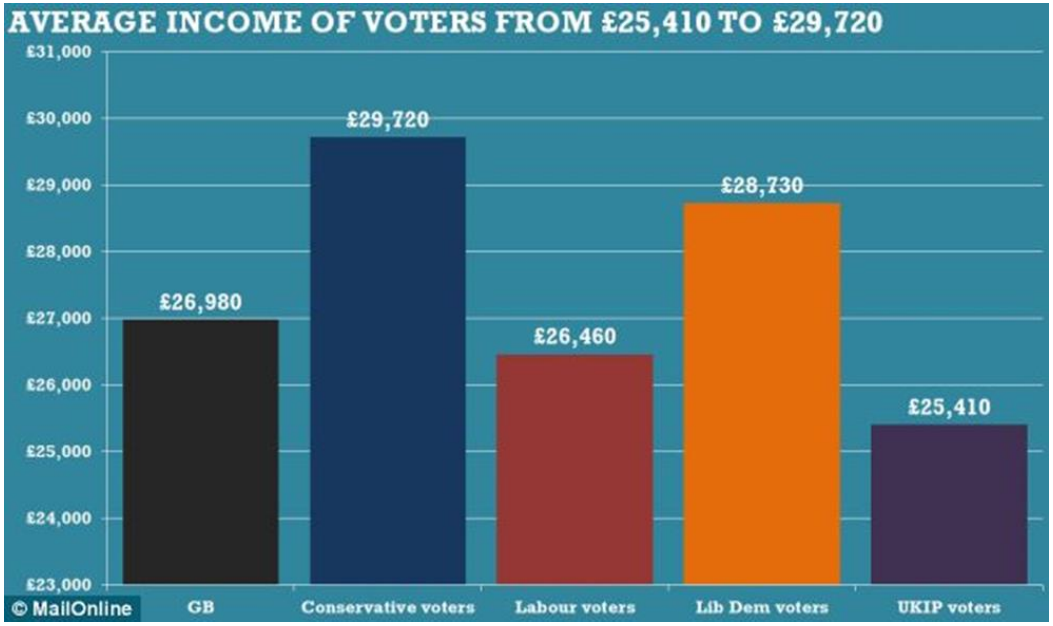


Assuming that income of voters in the different parties follows a normal distribution with mean as given in the chart and standard deviation £2000, what is the chance that a randomly chosen Labour voter earns more than the average wage of a UKIP voter and less than the median wage of a Liberal Democrat voter?

Exercise

We want P(25410 < X < 28730) for X a normal variable with mean 26460 and s.d. 2000.





Excel will give P(X < b) and P(X < a).

Exercise

P(X < 28730) = 0.8718.
P(X < 25410) = 0.2998.
P(25410 < X < 28730) = 0.5720.