

Lab 07 Microsoft Excel

Probability

This is an introduction to basic concepts in probability specific to data analysis in Excel. A definition of probability,

“Probability is a numerical measure of the frequency of occurrence of an event.”

Probability can be described as the likelihood or chance of an event occurring. It is the number of ways of achieving success. The total number of possible outcomes.

- Chance of rain today
- Likelihood of my favourite team winning

Managers often wish to know such things as the likelihood that a new product will be profitable, or the chances that a project will be completed on time.

There are some very different schools of thought as to the definition of probability. For simplicity we will stick to the frequencies' notion of probability. The basic idea behind the frequency approach is straightforward...

“The probability 0.5 of getting Heads on a coin toss can be thought of as the long run frequency of multiple coin tosses”

- The chance or probability of getting heads is 50%

“The probability is 1/6 of getting the number 6 on rolling a fair dice can be considered as a long run frequency of multiple dice rolls”

- Hence, the probability of a six turning up on a roll of a dice, is 1/6, 16.67%.

Concepts of Probability- Random Experiment and Associated Random Variable

Random Experiment

One of the important concepts of probability is that of a random experiment.

A random experiment is simply any situation where a process leads to more than one possible outcome. For example,

- A coin toss
- Roll of a dice
- A company declaring its earnings
- Closing value of the stock market tomorrow
- Bonus that you get at the end of the year

Random Variable

A random variable is an associated concept. It is a variable that takes on values, determined by the outcome of a random experiment.

This random variable, the outcome, can take on two possible values in the case of a coin toss and six possible values in the case of rolling a die. Similarly, the random variable associated with a company about to declare its earnings can be earnings per share (EPS) and in this case, it can take on multiple possible values.

Random Experiment

Coin Toss



Roll of a Dice



Company declaring
Its Earning



Random Variable

the "Outcome"

the "Outcome"

the "EPS"

Viewing business processes as a random experiment with an associated random variable is helpful in characterising them and making predictions about the outcome. To do this, we need to introduce statistical distributions.

Statistical Distribution

A statistical distribution is a tool to help us 'characterise' or 'model' the random variable.

Remember the histogram we used for salaries, CEO's for small businesses. The histograms told us how CEOs salaries varied. Some CEO's have high salaries, while a majority of them have a salary around \$200,000 to \$450,000. A few CEO's have much lower salaries. This is how the salaries are distributed:

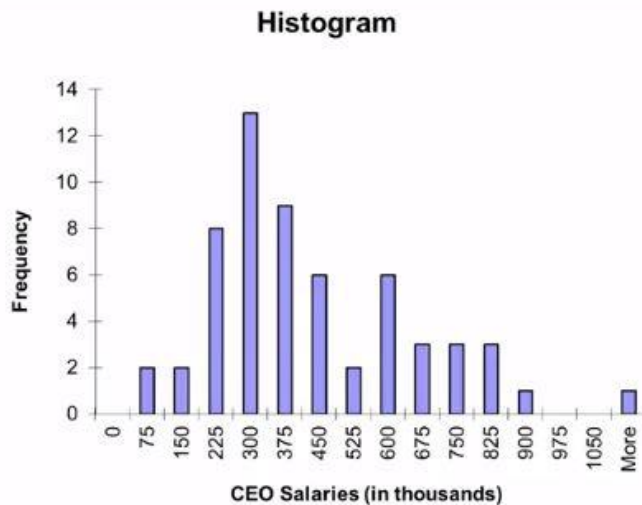


Figure 1: Histogram

We can view this as a random experiment. Because, as we saw a random experiment is any situation where, in the process, leads to more than one possible outcome. Here, different CEOs have different salaries. There are multiple outcomes for salary.

Random Experiment

Multiple possibilities of
CEO salary



Random Variable

the "Salary"

The associated random variable for this random experiment becomes the "Salary".

This will then allow us to understand how serious "Salary" would occur perhaps across a different set of small forms. Or maybe how they would occur in some different country and so on. It is a tool that will help us 'characterise' or 'model' the salary for CEOs across small forms. The benefit of doing so would be that, then we can make predictions about CEO salaries for small firms across a different set of firms not included in the data from which the histogram was created. The better we are able to model the salary the better we are at predictions.

Business statistics has a handful of statistical distributions to choose from:

- Beta
- Binomial
- Gamma
- Poisson
- Normal
- T distribution

The aim is to choose the most appropriate distribution to approximate a model the random variability. One of the most important and popular statistical distribution is the normal distribution, also known as the **bell curve**.

Statistical distributions are broadly categorised into two categories, discrete distributions and continuous distributions.

Discrete distribution:

- A statistical distribution used for discrete data

Example:

- number of students in a class, 25, 30
- the number of patients admitted to a hospital in a day
- the number of companies with revenue > \$1 billion,

Test of Discreteness:

- If we take any two possible realisations of the random variable, and ask ourselves, how many possible values exist between the two realizations, if the answer is finite, then the data is discrete. If the answer is infinite, the data is continuous

Continuous distribution:

- A statistical distribution used for continuous data

Example:

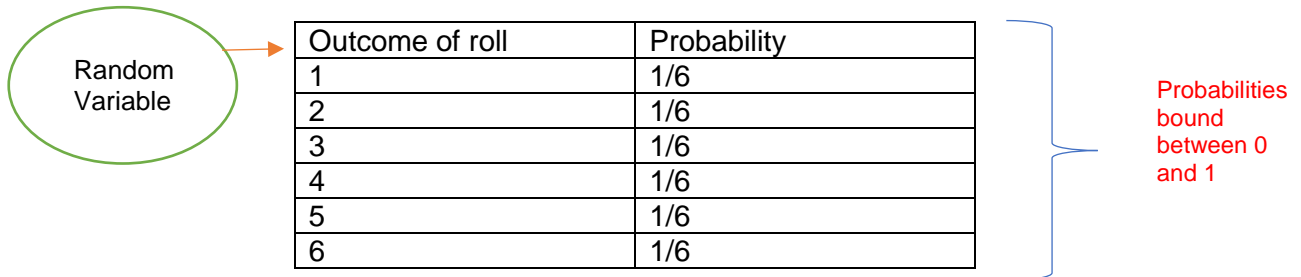
- Heights of men and women – there could exist infinite number of possibilities, someone may have a height of 151.cm, 150.1140cm

Probability Mass Function

The Probability Mass Function (PMF) associated with the outcome of a Roll of a Dice is shown in the table. Here again the Random variable can be termed as the Outcome. Note two important aspects of this Probability Mass Function.

Firstly, the probability associated with any particular outcome or realisation of the random variable is always bound between 0 and 1. Secondly and more importantly, the sum total of all probabilities is 1.

That is the function covers all possible values of the Random variable.



Outcome of roll	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Probabilities bound between 0 and 1

Table 1: PMT – All Values of the Random variable

It is easy to write down the PMF of,

- a coin toss
- roll of a dice

What about the PMF of a more complicated process?

For example,

Number of customers that will arrive at the checkout counter of a grocery store in one hour?

- we approximate this process with a statistical distribution and
- use the PMF of that statistical distribution.

customers arriving → 0 1 2 3 4 5 6 ...

The PMF will assign probabilities to different possible outcomes.

0 customers arriving at the checkout counter, 1 customer arriving, 2 customers arriving, 3, 4, 5 and so on. Note that here the number of customers arriving at the check-out counter in one hour is discrete data. That is you cannot have, e.g., four and a half customers. The PMF of the discrete

distribution would assign some probability to, say, 4 customers arriving, some other probability to 5 customers arriving and so on. There will be 0 probability attached to any number between, for example, 4 and 5 customers arriving.

When using a continuous distribution...

- PMF is called the Probability Density Function (PDF) (definition is the same)
 - It is a rule that assigns probabilities to various possible values that a random variable takes.
 - Probability of a particular outcome is always zero

Example, the possible heights of men and women in your neighbourhood.

(heights of men and women)



Probability (height = 5'2") = ?
= 0

Figure 2: Continuous distribution – heights of men and women

This is continuous data. What is the probability of someone's height being 5'2"?

The answer is 0, because even if your friend has a height of 5'2", you get a better measuring instrument if the height is not 5' 2", but say 5' 2.01".

Such kind of argument can be given for any height that you come up with, thus implying that the probability of someone having a particular height is always 0.

When using a continuous distribution,

Probability of a particular outcome is always zero...

... hence we always consider ranges of outcomes

For example,

- What is the probability that someone's height is between 5'2" and 5'5"?
- What is the probability that someone's height is less than 5' feet?
- What is the probability that someone's height is greater than 5' feet?

Graph:

This is explained here in the graph:

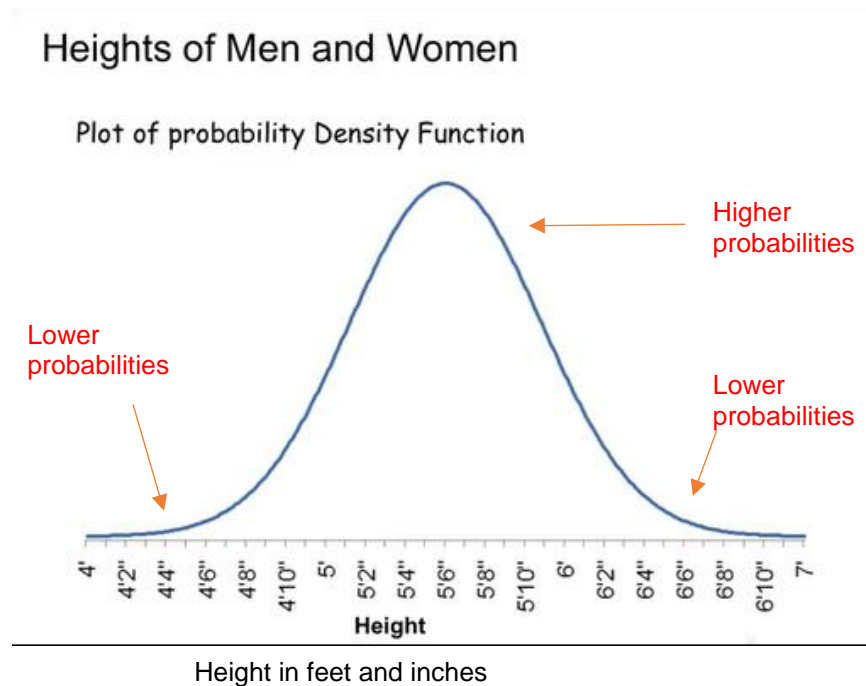


Figure 3: Probabilities of height

The curve is a continuous distribution used to approximate the very heights that are distributed in your neighbourhood. This curve is the plot of PDF for that continuous distribution we are using to approximate the occurrence of the random variable height

Next we will see alternative probabilities of heights...

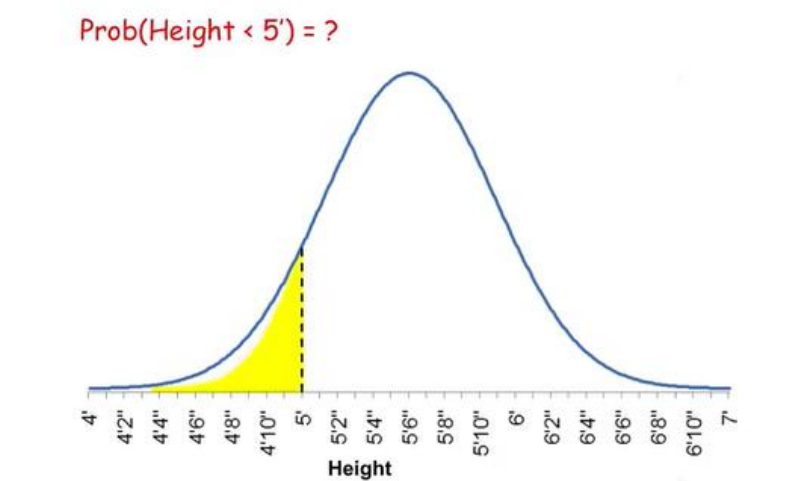


Figure 4: Height <5

$\text{Prob}(\text{Height} > 5' 10") = ?$

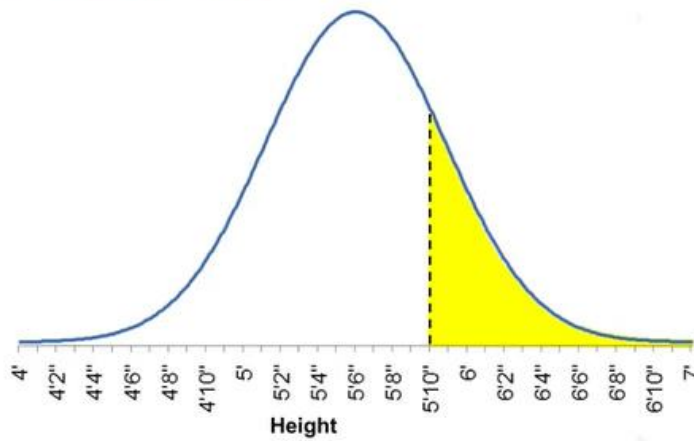


Figure 5: Height > 5'10"

$\text{Prob}(5' < \text{Height} < 5' 6") = ?$

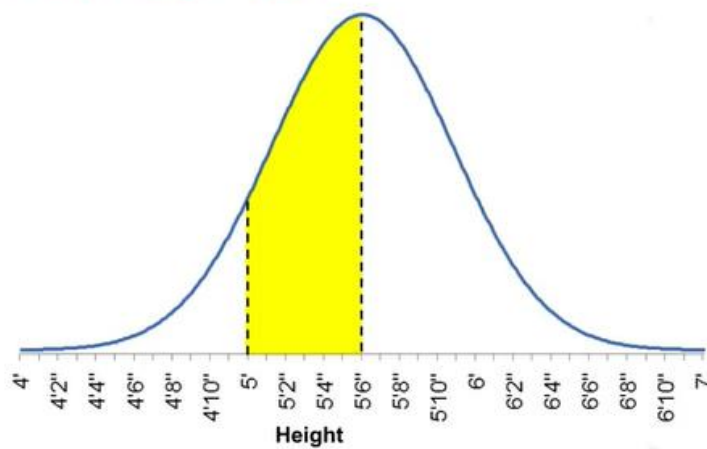


Figure 6: Between 5' and 5'6"

How do we measure area under the curve? This is where we make use of Excel formulas.

$$\text{Prob}(\text{Height} = 5') = 0$$

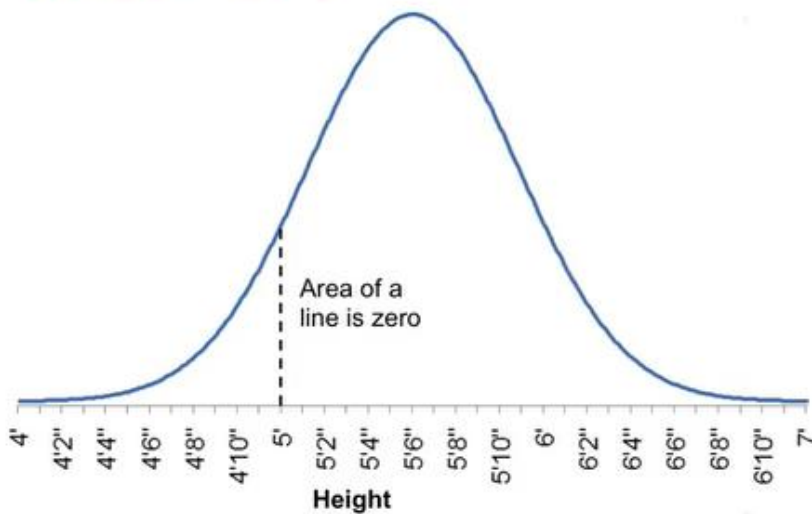


Figure 7: Height = 5'

Someone's height is exactly 5'?

It is area under the curve at the point 5. This would be area of a straight line, which is 0. Hence the statement that probability of someone's height being exactly 5' is 0. That is in a continuous distribution, the probability that the random variable takes our particular value is always 0.

We introduce the notion of area under the curve being a measure of probability. Further notice that since the probability cannot be created in one, and the distribution assigns probabilities to all possible outcomes of the random variable. Thus, the area under the entire curve has to be one.

Area under the entire curve = 1

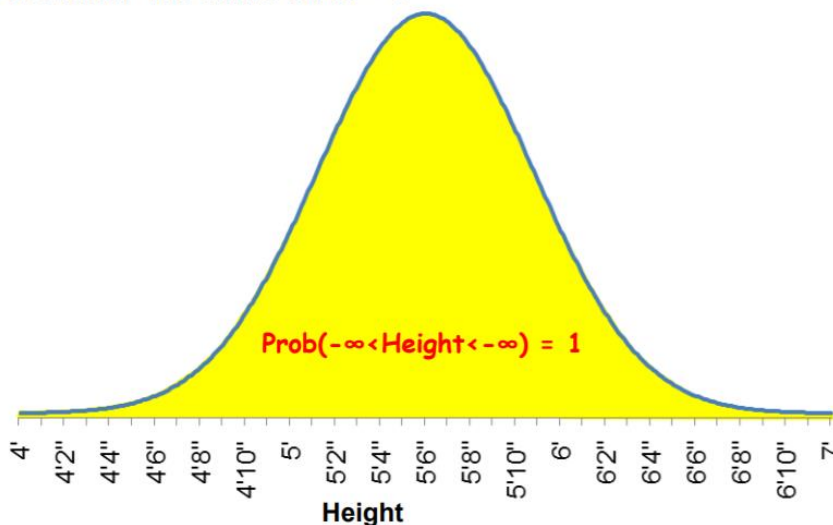


Figure 8: Area under the entire curve

Normal Distribution a.k.a the Bell Curve

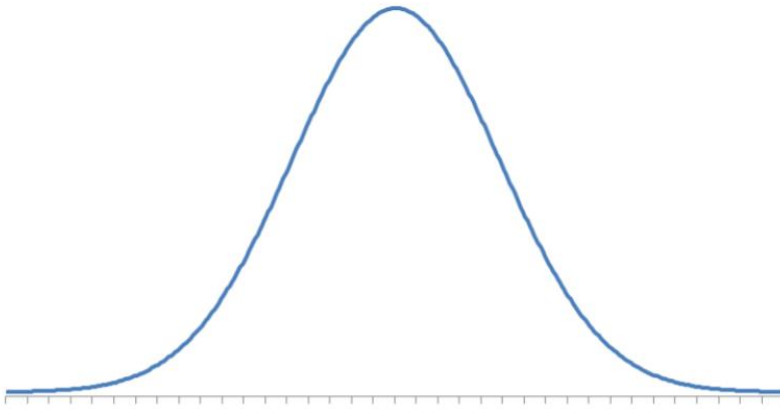


Figure 9: The Bell Curve- normal distribution

The normal distribution also known as the Bell Curve. It is the most widely used distribution, and also the most important. The normal distribution is a symmetric distribution, which we saw in the previous example. So if we were to vertically slice this distribution we would get two halves,

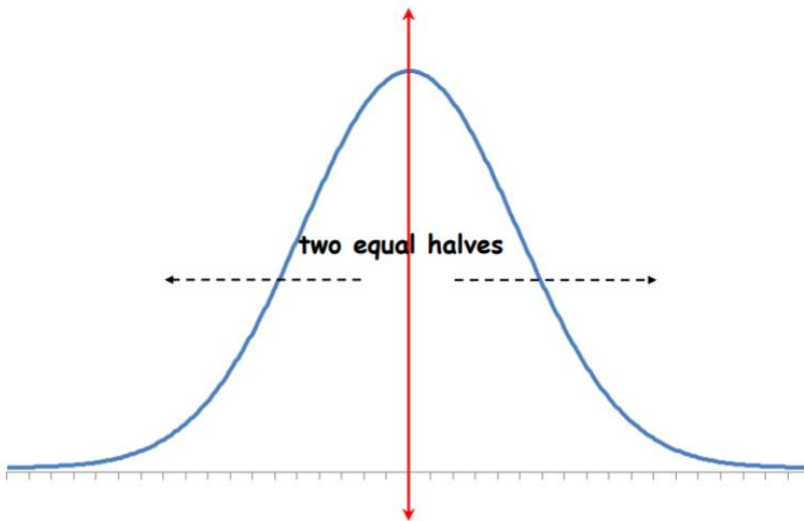


Figure 10: The Bell Curve – equal halves

A distribution gets identified by where on the horizontal axis it is placed and how spread out or peaked it is. In the case of a normal distribution, you can get many forms of the distribution, depending on where on the horizontal axis it is centred and how spread out or peaked it is:

Normal Distribution

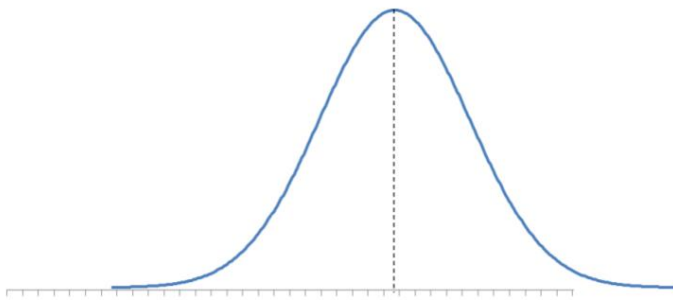


Figure 11: The Bell Curve

Normal Distribution

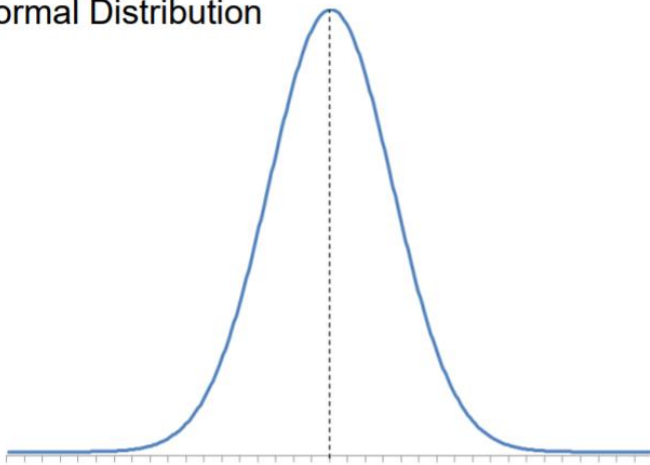


Figure 12: The Bell Curve

Normal Distribution

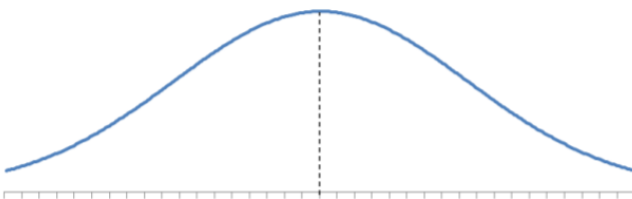


Figure 13: The Bell Curve

A particular normal distribution gets uniquely defined by two parameters. A location parameter, also known as the **mean** of the distribution. And secondly, a spread parameter also known as the **standard deviation** of the normal distribution. Depending on the values of these two parameters, we can have a multitude of normal distributions. The mean can take any value from negative infinity to positive infinity, while the standard deviation can take on only positive values.

The mean and the standard deviation numbers uniquely define a particular normal distribution. And this distribution is symmetric around the mean value.

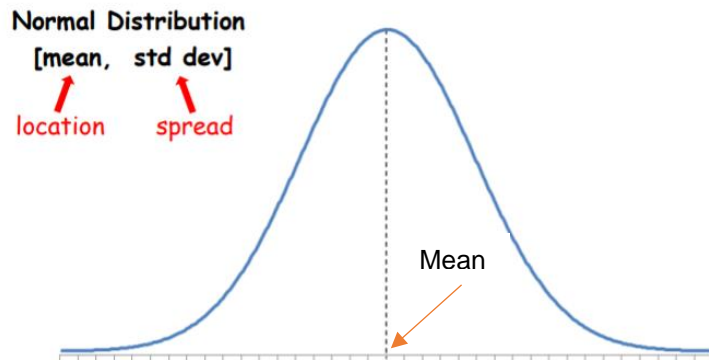


Figure 14: The Bell Curve - mean

Excel

In Excel, here we have the two parameters, the mean and standard deviation. The mean is 2 and standard deviation is 1.2. Notice the distribution is symmetric about the mean value.

That is, if I slice the distribution at the mean value, I'll get two equal halves.

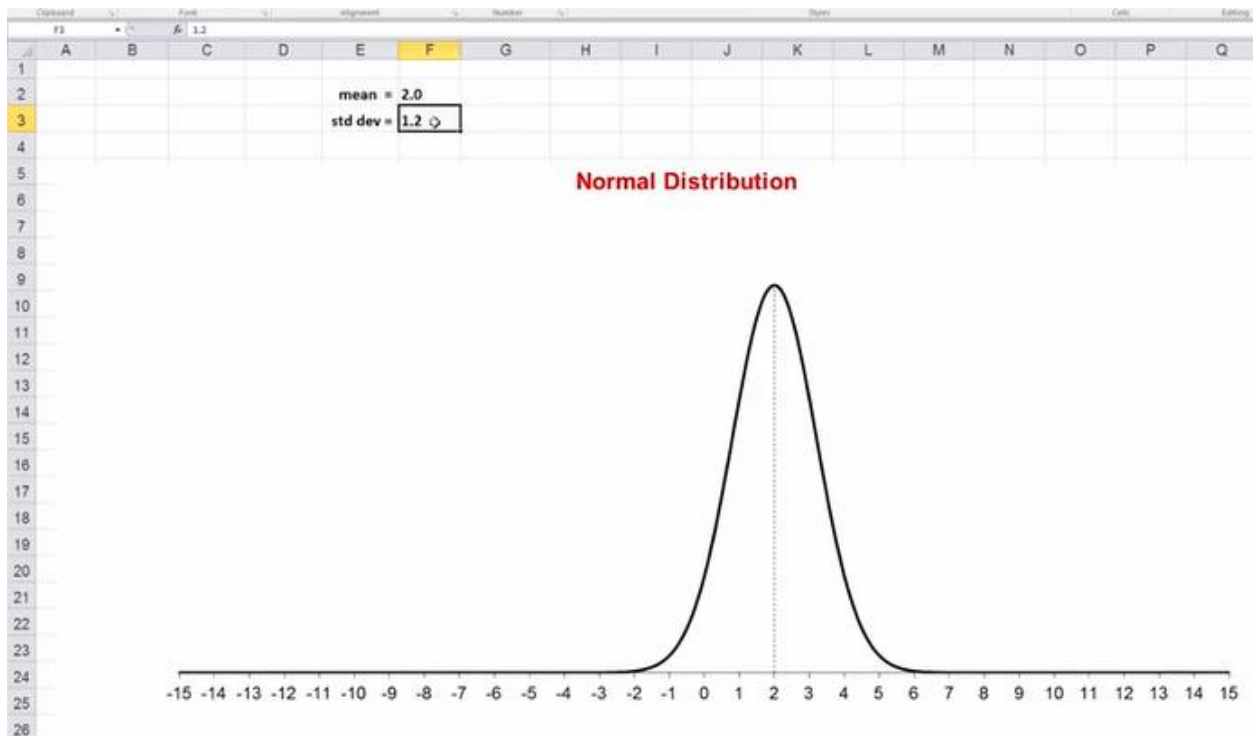


Figure 15: Excel

If I change the mean of the first parameter of the distribution. If I change it with three, the distribution shifts towards the right,

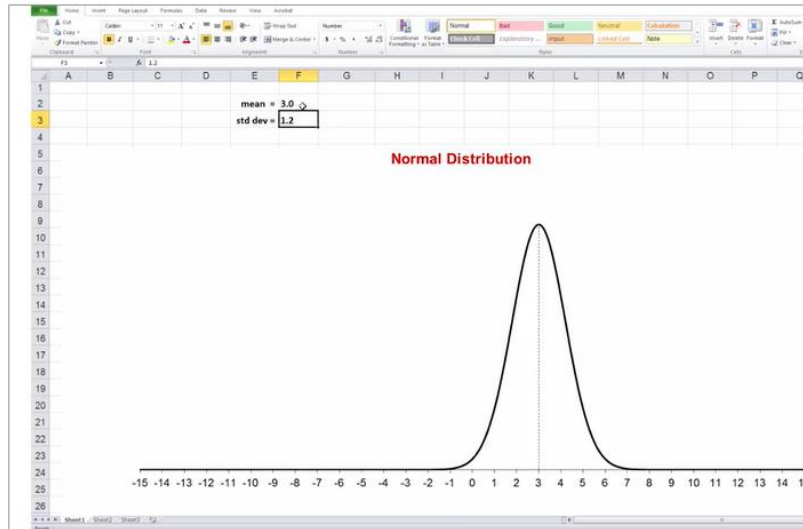


Figure 16: Excel

Change it to four, maybe six and the distribution is shifting towards the right.

What if I reduce the mean 5, 4, maybe 2, 1?

I can also go negative, -1, -2, so the distribution is shifting towards the left. That is the reason why the mean of the first parameter is also called the location parameter. It locates a distribution.

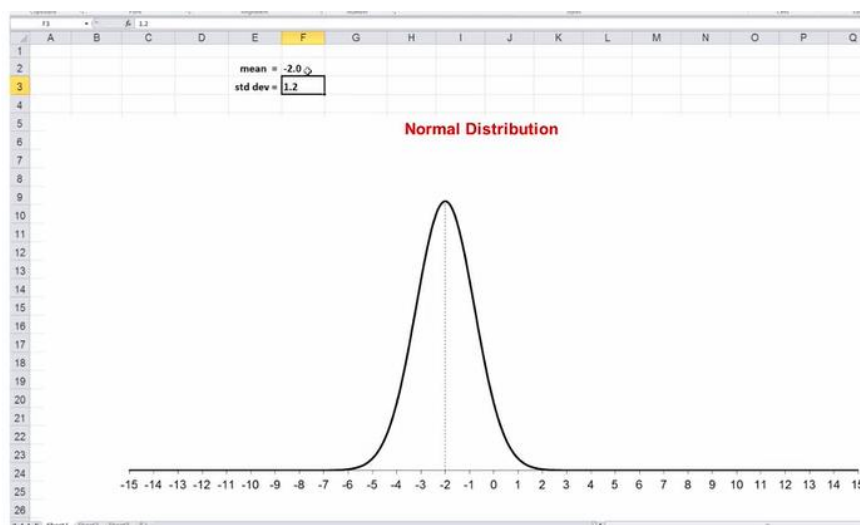


Figure 17: Excel

The second parameter, the standard deviation, if this was changed. e.g. reduce standard deviation to 1.0, note the distribution is getting more peaked.

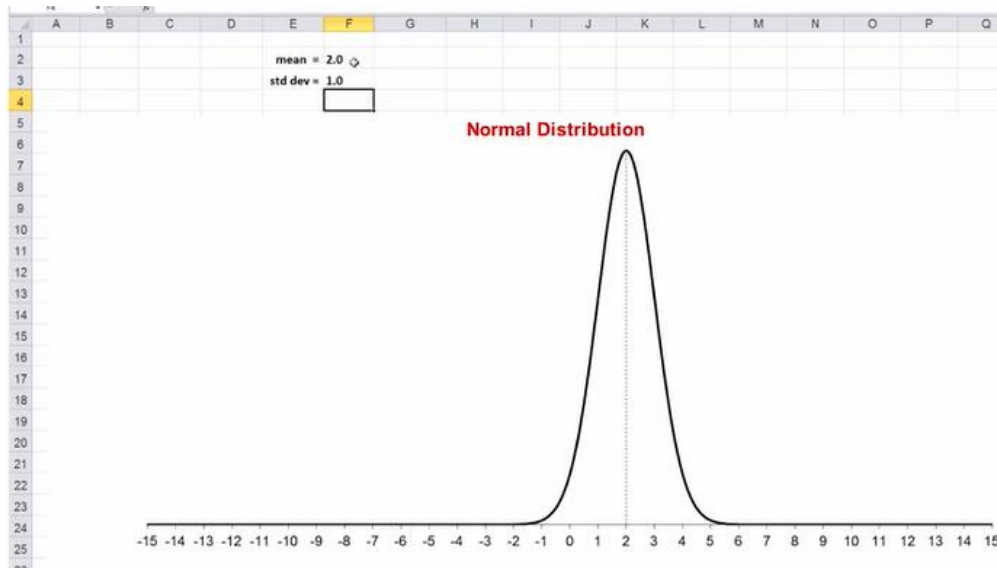


Figure 18: Excel

If I increase the standard deviation maybe 1.2 or to 1.7, the distribution gets more spread out or flattened:

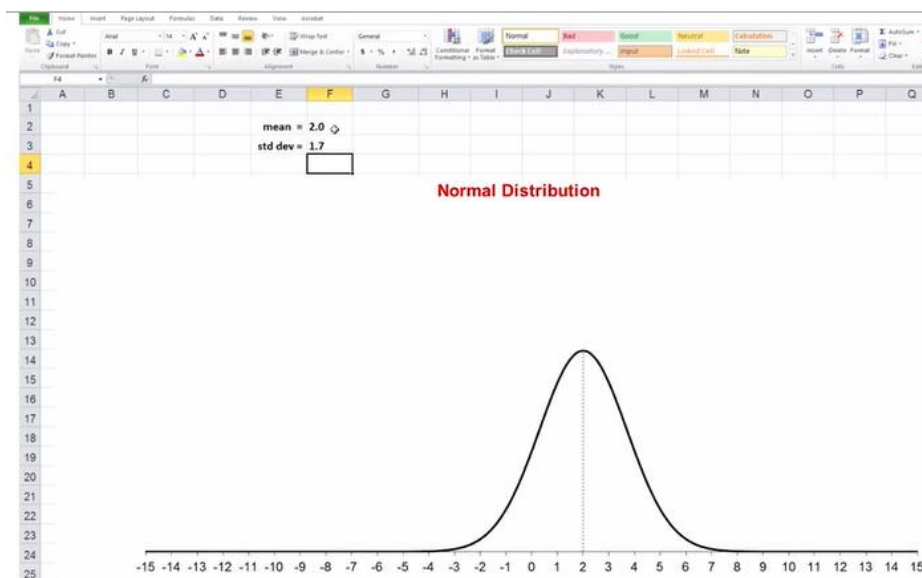


Figure 19: Excel

The distribution remains at the same place. That is the centred and the mean value. And that's why the standard deviation or the second parameter, the normal distribution is sometimes also referred to as the split parameter.

Another thing to note is that in a normal distribution, all outcomes from negative infinity to positive infinity are possible. Or in other words, the two tails. The left tail and the right tail go all the way to negative and positive infinity, respectively. The probabilities beyond a certain range get so small, that they're insignificant.

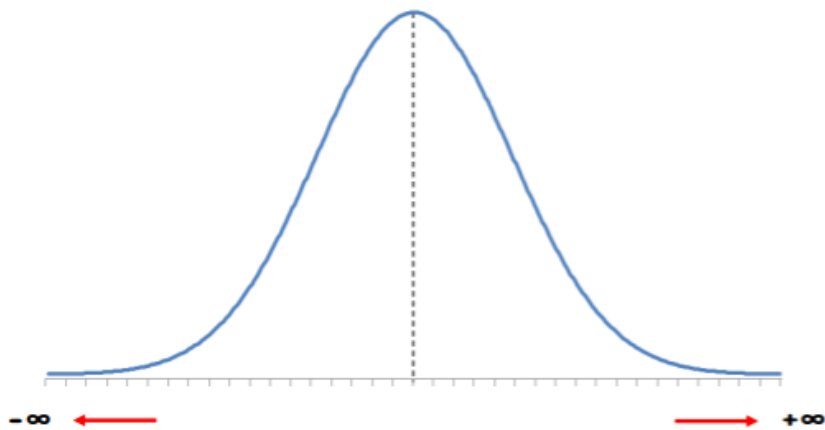


Figure 20: left tail and right tail (negative and positive)

The PDF has an equation, however we will focus on this through Excel using functions for the normal distribution.

Normal Distribution, the NORM.DIST function

Example,

- A bread producing company produces whole wheat loaves of bread in its factory.
- The company observes that on average every day 85 loaves of bread get discarded on account of being defective (due to packaging, spoilages etc). The mean value is 85.
- The standard deviation of number of defectives is 9 loaves.

This standard deviation indicates the spread in the number of defective loaves day to day, around the mean value of 85.

Given this data, the company wishes to predict:

that in the production run tomorrow, what are the chances or probability that less than 70 loaves of bread will get discarded on being defective?

The random variable, defective, follows a normal distribution with mean 85, standard deviation 9, we wish to calculate the probability that is, random variable is less than 70.

Looking at the plot of probability density function of the normal, we're interested in calculating area under the curve to the left of 70. In more technical terms, we wish to integrate the PDF of the normal between negative infinity and 70. Remember, the range of normal distribution varies from a negative infinity to positive infinity.

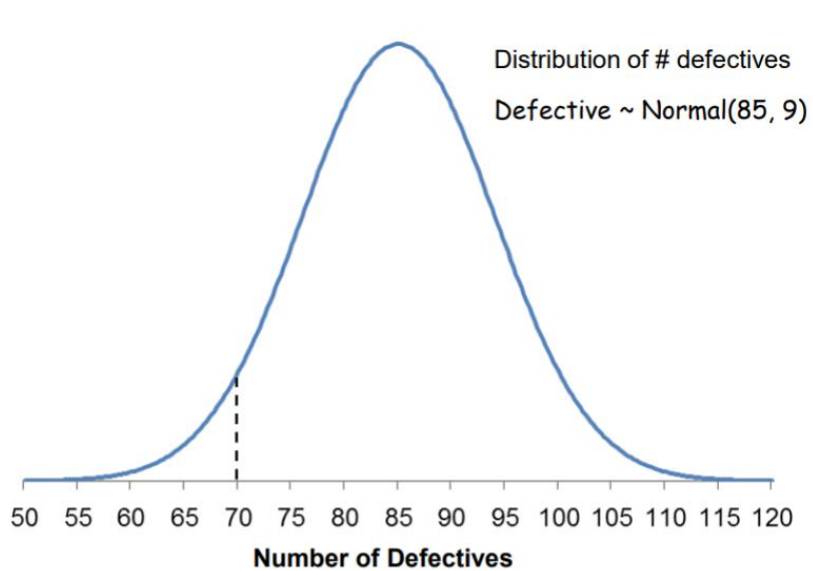


Figure 21: Probability of defective loaves <70

This data is discrete, because you cannot have half a loaf discarded.

To calculate the area under the curve, we use the NORM.DIST function in Excel.

Syntax:

=NORM.DIST(x, mean, std, TRUE)

The value of x is 70, mean is 85 and std dev is 9.

=NORM.DIST(70, 85, 9, TRUE)

In Excel,

	A	B	C	D	E	F	G	H
1								
2								
3								
4			Defective ~ Normal(85, 9)					
5								
6			Prob(Defective < 70) = ?				0.04779	
7								
8								
9								
10								
11								
12								

Figure 22: Excel – NORM function <70

The answer is 0.04779. Or in other words, there is a 4.779% chance that in tomorrow's production run, the number of defective loaves would be less than 70.

Defective ~ Normal(85, 9)	
Prob(Defective > 95) = ?	0.13326

[illegible]

Figure 23: Excel – NORM function >95

What is the probability that the number of defective loaves produced tomorrow are in excess of 95?

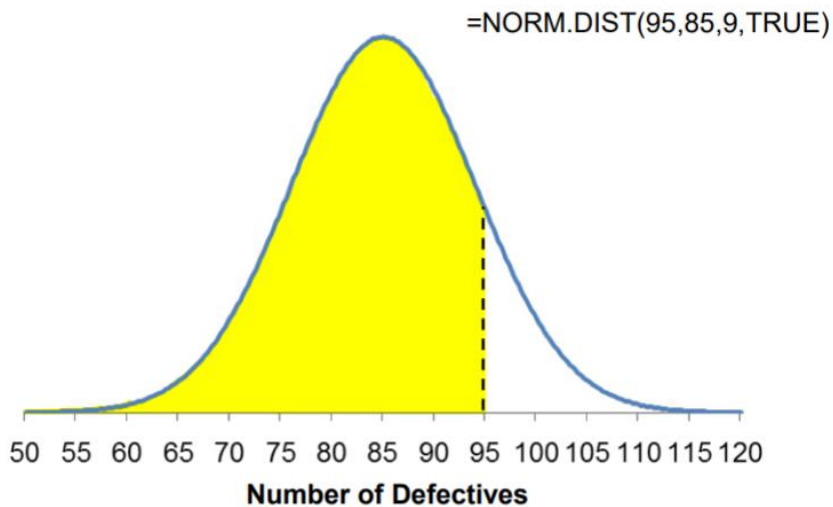


Figure 24: Defective loaves (<95)

We want to calculate the curve to the right of 95. The area to the right of 95 is 1 minus area to the left of 95.

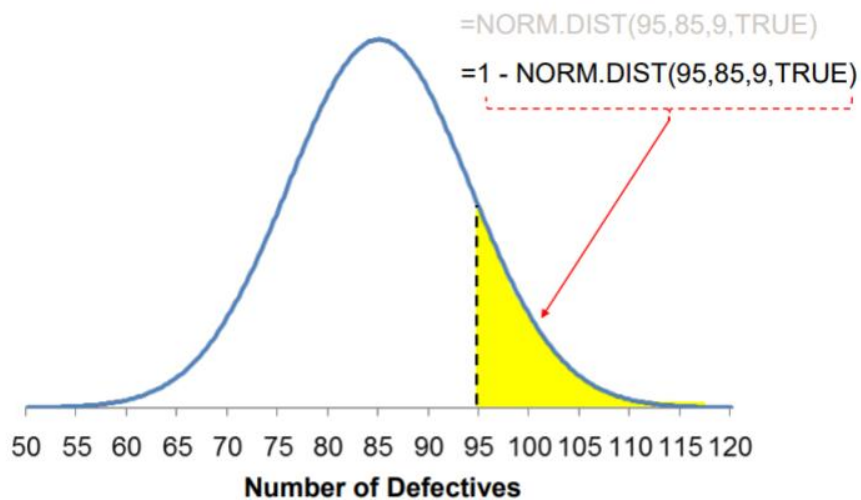


Figure 25: Defective loaves (area to right of 95)

Defective ~ Normal(85, 9)	
Prob(Defective > 95) = ?	<code>=NORM.DIST(95,85,9,TRUE)</code> <small>(NORM.DIST, mean, standard dev, cumulative)</small>

The answer is 0.13326 or approximately 13.33% chance that the number of defectives in excess of 95.

What is the probability that the number of defective loaves produced is between 75 and 80?

Here the answer is the area under the curve, between 75 and 80:

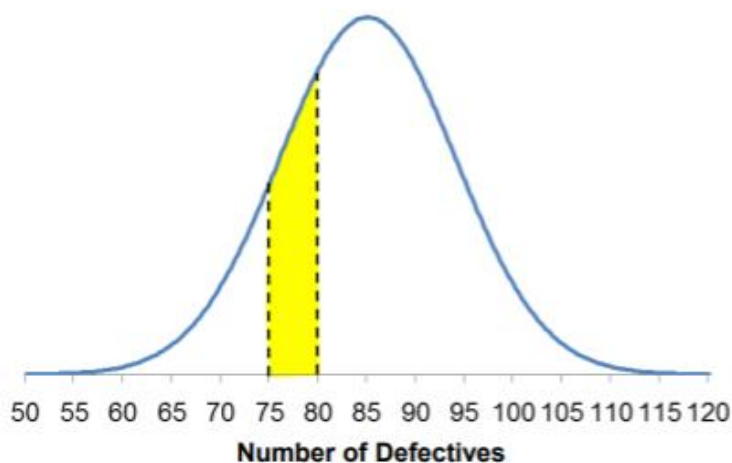


Figure 26: Defective loaves between 75 and 80

Using the Norm.Dist function it is calculated as shown:

$$\text{Prob}(75 < \text{Defective} < 80) = \text{NORM.DIST}(80, 85, 9, \text{TRUE}) - \text{NORM.DIST}(75, 85, 9, \text{TRUE})$$

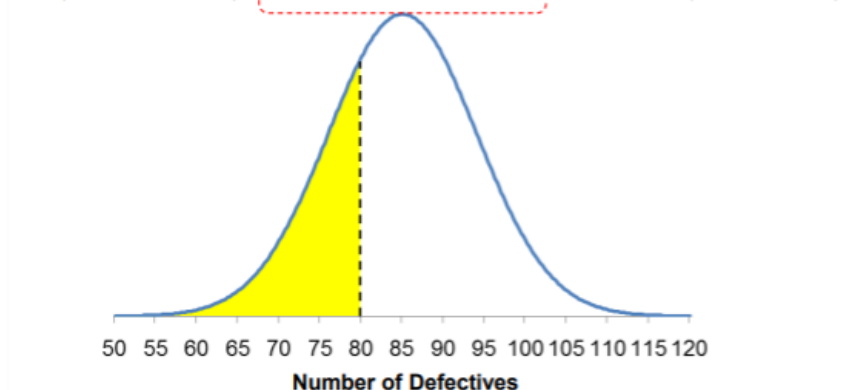


Figure 27: Defective loaves (left of 80)

The first part of the expression `norm.dist 80, 85, 9, TRUE`, calculates a probability of number of defectives being less than 80. That is, area under the curve to the left of 80. However, we do not want this entire probability. We only want the probability of the random variable, the number of defectives being less than 80, but greater than 75. So the second expression subtracts out the probability of the random variable being less than 75 resulting in our desired probability.

$$\text{Prob}(75 < \text{Defective} < 80) = \text{NORM.DIST}(80, 85, 9, \text{TRUE}) - \text{NORM.DIST}(75, 85, 9, \text{TRUE})$$

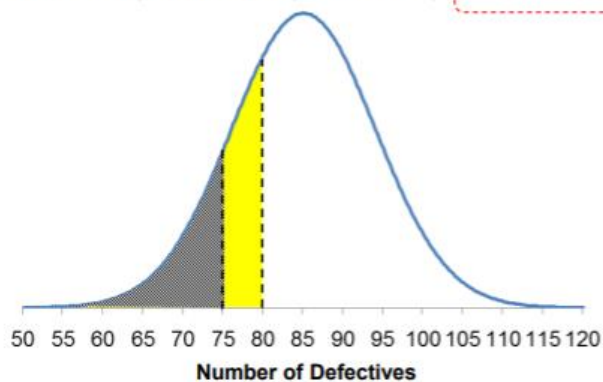


Figure 28: Defective loaves 75 and 80

In Excel,

Defective ~ Normal(85, 9)	
Prob(75 < Defective < 80) = ?	<code>=NORM.DIST(80,85,9,TRUE) - NORM.DIST(75,85,9,TRUE)</code> <small>NORM.DIST(x, mean, standard_dev, cumulative)</small>

The probability is 0.1560.

Defective ~ Normal(85, 9)	
Prob(75 < Defective < 80) = ?	0.1560

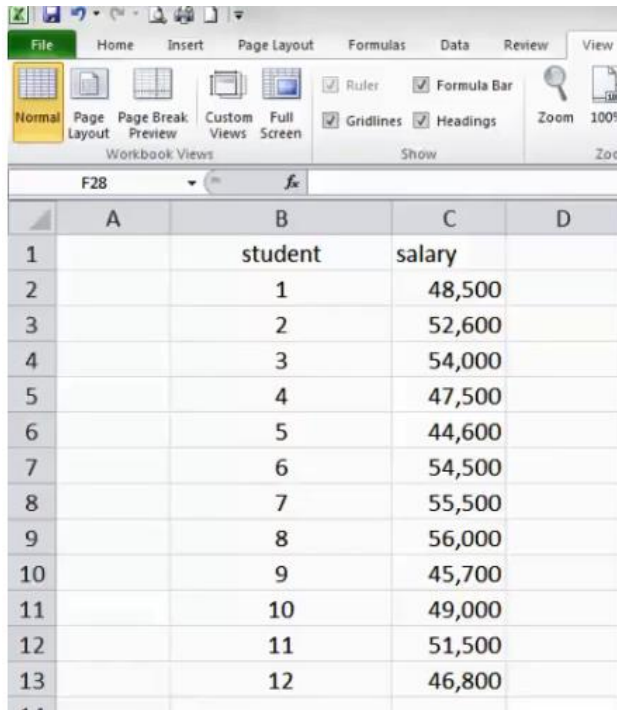
Or in other words, there's a 15.6% chance that the number of defective loaves produced is less than 80 but greater than 75.

Task:

Consider the following:

- Agnes Hammer is a senior majoring in management science. She is curious about what starting salary offers she might receive. There are 140 seniors in the graduating class for her major, and more than half have received job offers. She asked 12 of her classmates at random what their annual starting salary offers were, and she received the following responses:
- \$48,500, \$52,600, \$54,000, \$47,500, \$44,600, \$54,500, \$55,500, \$56,000, \$45,700, \$49,000, \$51,500, and \$46,800
- Assume that starting salaries are normally distributed.
 - a. Compute the mean and standard deviation for these data.
 - b. Determine the probability that Agnes will receive a salary offer of less than \$50,000.
 - c. What is the probability that Agnes will receive a salary offer between \$45,000 and \$55,000?

Type the following into Excel and answer the questions:



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D
1		student	salary	
2		1	48,500	
3		2	52,600	
4		3	54,000	
5		4	47,500	
6		5	44,600	
7		6	54,500	
8		7	55,500	
9		8	56,000	
10		9	45,700	
11		10	49,000	
12		11	51,500	
13		12	46,800	