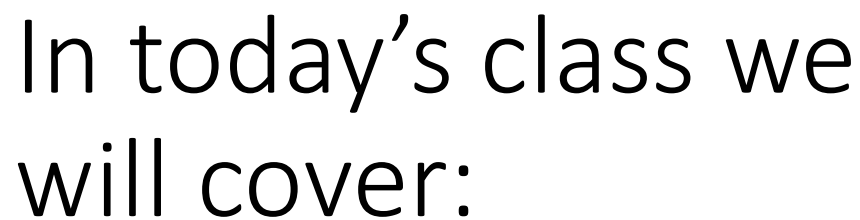


Statistics for Data Analytics

Lecturer: Marina Iantorno

E-mail: miantorno@cct.ie





- ❑ Random Variables
- ❑ Binomial Distribution
- ❑ Practice time



Random Variables

Random Variables

Since we started the module, we talk about events. There is a study in Statistics that is in charge of measuring the result of a chance event, and this is called “Random Variable”. You can also find this in books with other names such as random quantity, aleatory variable, or stochastic variable.

As its name indicates it, a Random Variable is a variable, and we intend to represent the possible outcomes of each value that this variable might take.

Let’s see an example:

Suppose that you want to analyse the number that you get when you roll a dice. The first thing we need to do is to define the name of our variable.



Random Variables

Keep in mind that we will find two possible variables:

- **Discrete variables:** These are the variables that count only whole numbers, as it would result when rolling the dice. Other examples could be people, books, coins, chairs, etc.
- **Continuous variables:** These are the variables that take into consideration all the possible numbers between two points, like litres of water a person drinks per day. Other examples could be kilometres, centimetres, scores, kilograms, etc.



Random Variables

X = number that shows up when we roll the dice

1, 2, 3, 4, 5, 6 } These are the sides of the dice, and we do not have another option, and this is the *Domain* of this random variable

X	$P(X)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$

What we have to do, is to build a table that contains all the possible outcomes and their probabilities. To find the probabilities we recall the basic formula that we used in Probabilities:

$$P(A) = \frac{\text{Favorable cases of the event A}}{\text{Total cases}}$$

If the dice has 6 faces and each number has the same probability of occurrence, then all of them will have the probability $1/6$.

Random Variables

After rolling the dice:

a) What is the probability of getting a 4?

$$P(X = 4) = 1/6 = 0.1666$$

X	P(X)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

The only thing we do to answer this question is having a look at the table and find the number.

Random Variables

After rolling the dice:

b) What is the probability of getting a number greater than 3?

$$P(X > 3) = P(X = 4) \cup P(X = 5) \cup P(X = 6)$$

$$P(X > 3) = 1/6 + 1/6 + 1/6 = 0.50$$

X	P(X)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Random Variables

After rolling the dice:

b) What is the probability of getting a number smaller than 3?

$$P(X < 3) = P(X = 2) + P(X = 1)$$

$$P(X < 3) = 1/6 + 1/6 = 0.3333$$

X	P(X)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Random Variables

C) What is the **expected** number?

The expected number is the *mean* of this variable, and that is why this is the number we can expect when rolling the dice.

X	P(X)	X * P(X)
1	1/6	1 * 1/6
2	1/6	2 * 1/6
3	1/6	3 * 1/6
4	1/6	4 * 1/6
5	1/6	5 * 1/6
6	1/6	6 * 1/6
		3.5

Answer: The expected number is 3.5



Even though we are working with a discrete variable, we do not round this number. Whatever we get, this is the answer and will be considered the expected value. Formula:

$$E(X) = \sum X * P(X)$$

Random Variables

d) What is the **variance** of this random variable?

The variance is known as a disperse measure. This particular measure does not have a real interpretation.

We represent the variance as S^2 .

X	P(X)	E(X)	X – E(X)	[X – E(X)] ²	P(X) * [X-E(X)] ²
1	1/6	3.5	-2.5	6.25	1.0416
2	1/6	3.5	-1.5	2.25	0.375
3	1/6	3.5	-0.5	0.25	0.0416
4	1/6	3.5	0.5	0.25	0.0416
5	1/6	3.5	1.5	2.25	0.375
6	1/6	3.5	2.5	6.25	1.0416
					2.9164

Random Variables

e) What is the **standard deviation** of this random variable?

This is also a disperse measure. The standard deviation could be considered as an error margin from the expected value, and we just need to apply a squared root to the variance.

$$S^2 = 2.9164$$

$$S = \sqrt{2.9164} = 1.70$$

Answer: The expected value when we roll the dice is 3.5 ± 1.70 .



The unit of measurement of the variance is expressed squared (i.e, books², people², chairs², etc, and this is why it does not have an interpretation. If we want to give an interpretation, we need to calculate the standard deviation

Random Variables

Making all these calculation would take a lot of time. We could do this quicker using Microsoft Excel.

The first thing is writing on an Excel sheet the known values.

A	B	C	D	E	F
X	P(X)	E(X)	$X - E(X)$	$[X - E(X)]^2$	$P(X) * [X - E(X)]^2$
1	1/6	3.5			
2	1/6	3.5			
3	1/6	3.5			
4	1/6	3.5			
5	1/6	3.5			
6	1/6	3.5			

Random Variables

Now we will use the formulas to proceed with the calculations registered on each header. First, we will do $X - E(X)$, and once it is done, we will drag the formula till the bottom of the column.

	A	B	C	D	E	F
1	X	P(X)	E(X)	$X - E(X)$	$[X - E(X)]^2$	$P(X) * [X - E(X)]^2$
2	1	1/6	3.5	=A2-C2		
3	2	1/6	3.5			
4	3	1/6	3.5			
5	4	1/6	3.5			
6	5	1/6	3.5			
7	6	1/6	3.5			
8						

Random Variables

We will continue with $[X - E(X)]^2$. Do not forget to put the cell in brackets to include the sign of the numbers before the squared.

	A	B	C	D	E	F
	X	P(X)	E(X)	X - E(X)	$[X - E(X)]^2$	P(X) * $[X - E(X)]^2$
1	1	1/6	3.5	-2.5	<code>=(D2)^2</code>	
2	2	1/6	3.5	-1.5		
3	3	1/6	3.5	-0.5		
4	4	1/6	3.5	0.5		
5	5	1/6	3.5	1.5		
6	6	1/6	3.5	2.5		

Random Variables

Finally, we will proceed with $P(X) * [X-E(X)]^2$. Again, drag the formula till the bottom of the column.

A	B	C	D	E	F
X	P(X)	E(X)	$X - E(X)$	$[X - E(X)]^2$	$P(X) * [X - E(X)]^2$
1	1/6	3.5	-2.5	6.25	1.0417
2	1/6	3.5	-1.5	2.25	0.3750
3	1/6	3.5	-0.5	0.25	0.0417
4	1/6	3.5	0.5	0.25	0.0417
5	1/6	3.5	1.5	2.25	0.3750
6	1/6	3.5	2.5	6.25	1.0417
					2.9167

Random Variables

Finally, we will proceed with $P(X) * [X - E(X)]^2$. Again, drag the formula till the bottom of the column.

A	B	C	D	E	F
X	P(X)	E(X)	$X - E(X)$	$[X - E(X)]^2$	$P(X) * [X - E(X)]^2$
1	1/6	3.5	-2.5	6.25	1.0417
2	1/6	3.5	-1.5	2.25	0.3750
3	1/6	3.5	-0.5	0.25	0.0417
4	1/6	3.5	0.5	0.25	0.0417
5	1/6	3.5	1.5	2.25	0.3750
6	1/6	3.5	2.5	6.25	1.0417
					2.9167

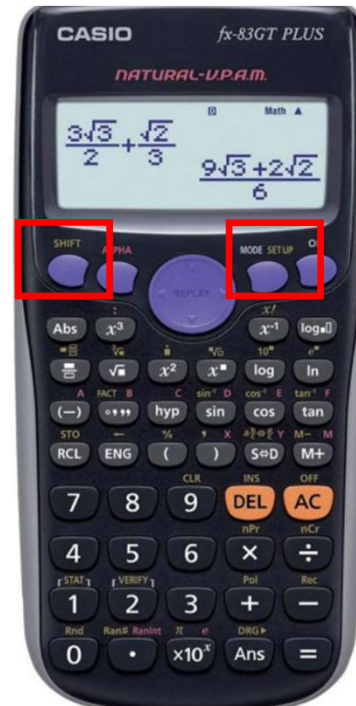


Here is the variance

Random Variables

We can also use our scientific calculator to get these results.

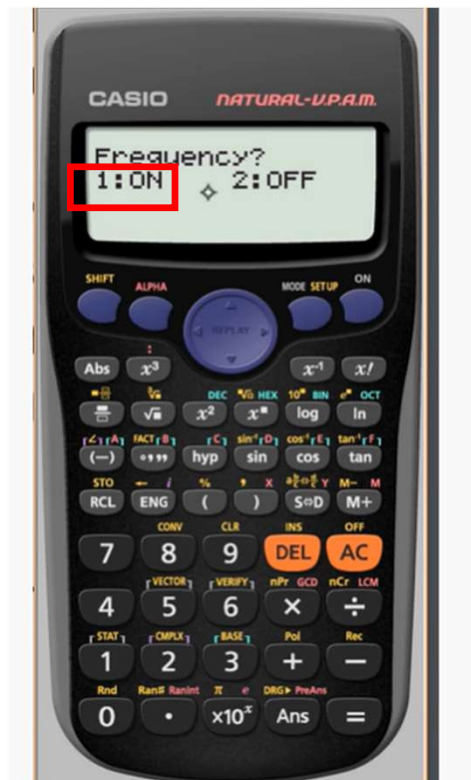
First, you need to press SHIFT + SETUP and look for the option STAT. If you do not find it in the first screen scroll down until you see it. In my calculator it is the option number 4 in the second screen, but it may change from model by model.



Random Variables

We can also use our scientific calculator to get these results.

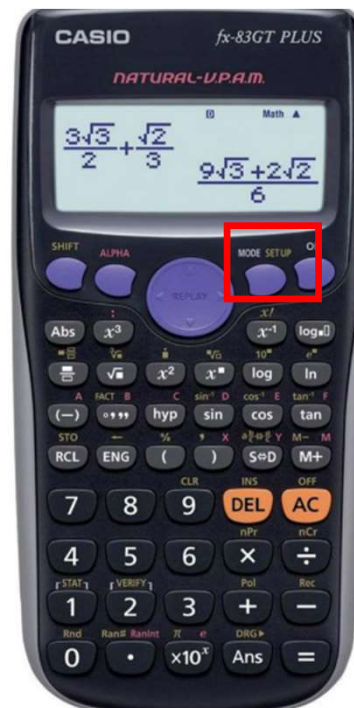
Now you need to press the option that will turn the frequency ON. It is usually 1, same as you see in this screen.



Random Variables

We can also use our scientific calculator to get these results.

Now press MODE and select the option STAT. In my calculator is the option 2, but it may change from model by model.



Random Variables

We can also use our scientific calculator to get these results.

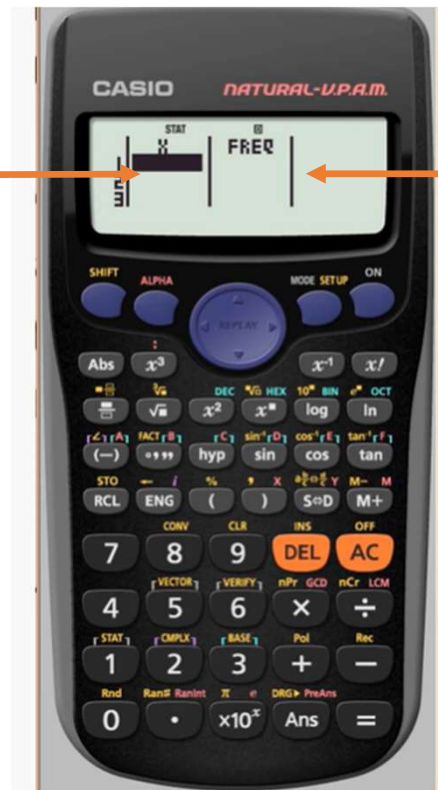
Now choose the option 1-VAR. In my calculator is the option 1, but it may change from model by model.



Random Variables

We can also use our scientific calculator to get these results.

You will add in X all the values of
the domain in X



You will add in FREQ the
probability of that value to happen

Random Variables

We can also use our scientific calculator to get these results.

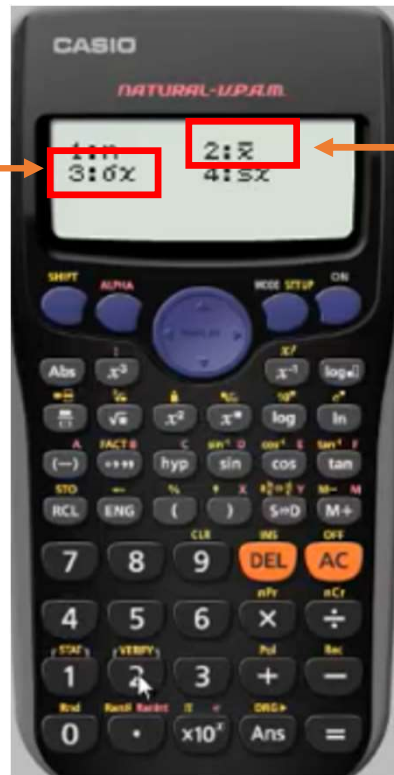
Once you uploaded all the data in the calculator, you will press SHIFT + STAT and you will choose the option VAR. In my calculator is the option number 5, but it may change according to the model of your calculator



Random Variables

We can also use our scientific calculator to get these results.

This is the standard deviation



This is the expected value

Random Variables

Properties of the Expected Value

- 1) $E(a) = a$
- 2) $E(aX) = a * E(X)$
- 3) $E(X + Y) = E(X) + E(Y)$
- 4) $E(XY) = E(X) * E(Y)$

Note:

- *“a” is a constant value*
- *Property 4 applies only if the variables are independent*

Random Variables

Properties of the Variance

- 1) $\text{Var}(X) \geq 0$
- 2) $\text{Var}(a) = 0$
- 3) $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
- 4) $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$
- 5) $\text{Var}(XY) = \text{Var}(X) * \text{Var}(Y)$
- 6) $\text{Var}(aX) = a^2 \text{Var}(X)$
- 7) $\text{Var}(X + a) = \text{Var}(X)$

Note:

- *“a” is a constant value*
- *Property 3 and 5 apply only if the variables are independent*

Binomial Distribution

$$P(X) = \binom{n}{x} p^x q^{n-x}$$

$$\binom{n}{r} = nC_r = \frac{n!}{(n-r)! r!}$$

$$q = 1-p$$

Binomial Distribution

Probability Distribution

A probability distribution help us to know all the possible outcomes of a particular experiment. Through this, we will have the probabilities of the occurrence of some events.

It is important to understand that :

- Each variable has a domain;
- The sum of all the probabilities within that domain will result 1;
- Probability distributions are random variables.



Probability Distribution

Why do we work with distributions instead of random variables?

The principle is the same, and if we created a table with the domain of the variable and its probabilities, we would have the same characteristics as we saw in random variables. However, we can adapt the random variables to different techniques according to the data in study, and it will make our job easier.

Let's get started!



Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

- a) What is the probability to find exactly 3 people who smoke?
- b) What is the probability to find less than 3 people who smoke?
- c) What is the probability to find more than 3 people who smoke?
- d) What is the probability to find between 3 and 5 people who smoke?
- e) What is the probability that none of them smoke?
- f) How many people do we expect to be smokers within these 10? Calculate Variance and Standard Deviation.

Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people.

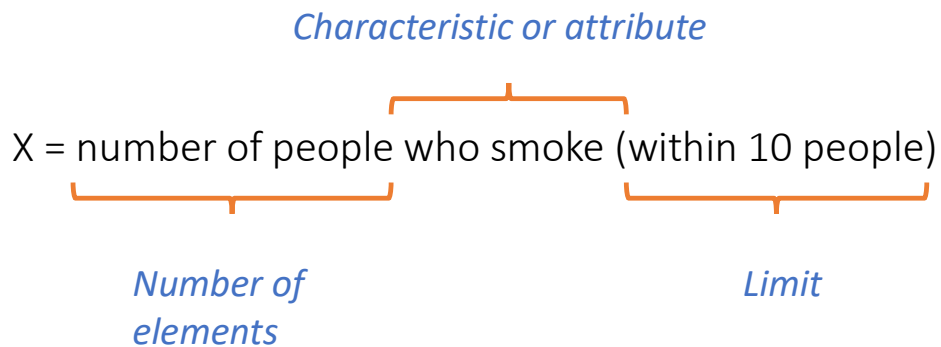
Structure of the random variable

As this is a random variable, we need a definition.

Characteristic or attribute

X = number of people who smoke (within 10 people)

Number of elements *Limit*



The structure of this variable is **ALWAYS** as follow:

X = number of elements with a characteristic/attribute (within a limit)


Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

Domain of the random variable

Earlier we saw the data displayed on a table, in which we saw all the possible values of the variable and their probabilities of occurrence. In this case, we need to define the domain. Remember that it includes *ALL* the possible results when we choose 10 people.

Domain: { 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 }



The last value of the domain will **always** be the limit

Binomial Distribution

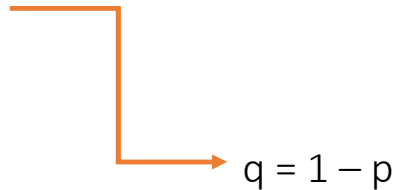
It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

Parameters of this distribution

The Binomial distribution has two parameters:

n = number of my sample (limit)

p = probability of a random element to have the characteristic or attribute



By default we can get the probability of random element not to have the characteristic or attribute.

Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

a) What is the probability to find exactly 3 people who smoke?

X = number of people who smoke (within 10 people)

$n = 10$

$p = 0.20$

$q = 0.80$

Formula



$$P_{(x)} = \binom{n}{x} p^x q^{n-x}$$

$$\binom{n}{x}$$



This is called “Combinatory Analysis”. The main idea is to find all the possible combinations in which the person who smokes could show up



...ETC

Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

a) What is the probability to find exactly 3 people who smoke?

X = number of people who smoke (within 10 people)

$n = 10$ $p = 0.20$ $q = 0.80$

$P(X = 3) =$

0	1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	---	----

We will look for this value on the Probability Distributions App!

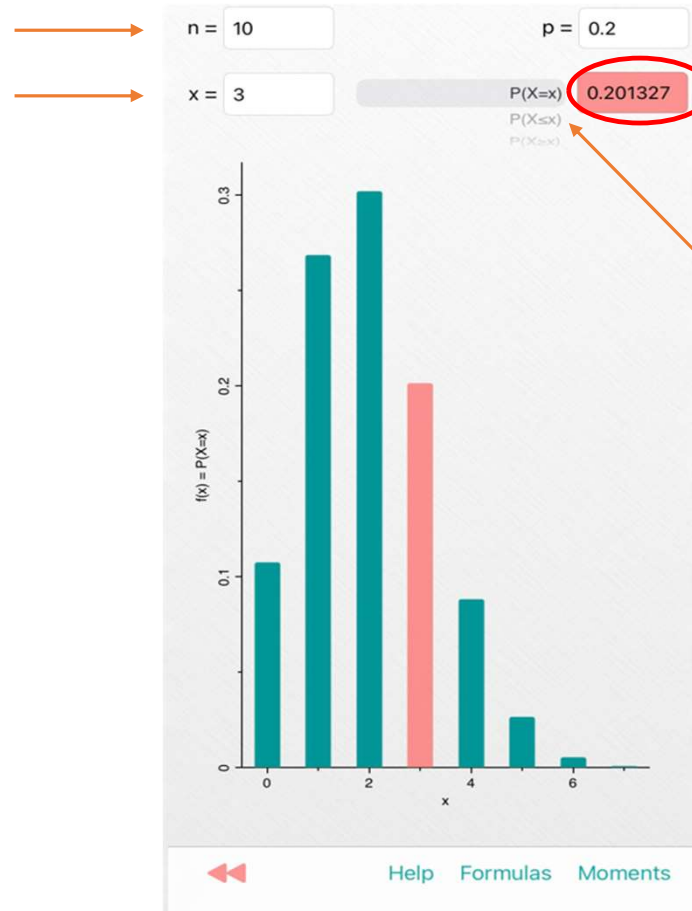


Binomial Distribution

How to use it?

Here we add n

Here we add the
value that we want



Here we add p

Here we read the result

Here we choose
the sign:

=
 \leq
 \geq

Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

a) What is the probability to find **exactly** 3 people who smoke?

X = number of people who smoke (within 10 people)

$$n = 10$$

$$p = 0.20$$

$$q = 0.80$$

$$P(X = 3) = 0.2013$$

0	1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	---	----

Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

b) What is the probability to find **less** than 3 people who smoke?

X = number of people who smoke (within 10 people)

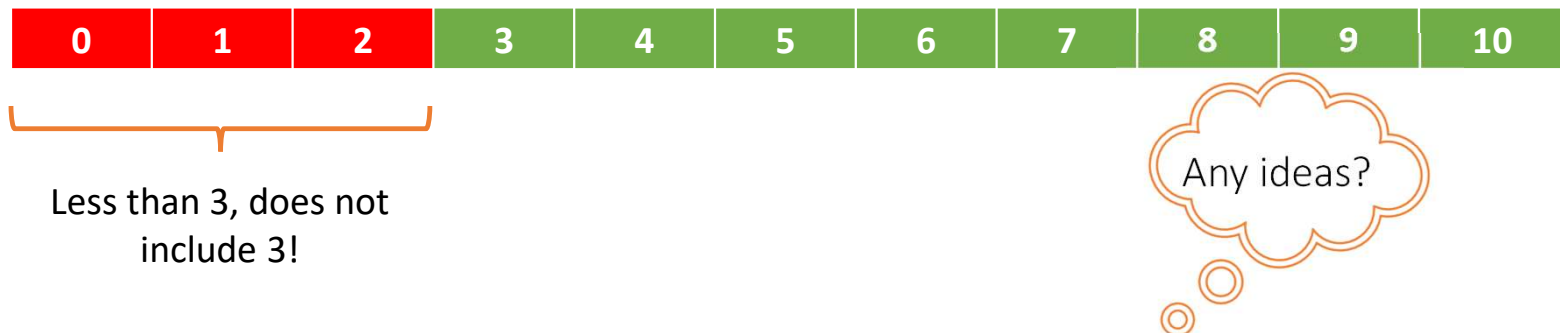
$n = 10$

$p = 0.20$

$q = 0.80$

In this case we need to remember the probability rule:

We can find 0 people who smoke **OR** one person who smokes **OR** two people who smoke



Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

b) What is the probability to find **less** than 3 people who smoke?

X = number of people who smoke (within 10 people)

$$n = 10$$

$$p = 0.20$$

$$q = 0.80$$

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) \quad \leftarrow \text{We can find this as } P(X \leq 2)$$

$$P(X < 3) = 0.1073 + 0.2685 + 0.3019 = 0.6777$$



Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

c) What is the probability to find **more** than 3 people who smoke?

X = number of people who smoke (within 10 people)

$$n = 10$$

$$p = 0.20$$

$$q = 0.80$$

$$P(X > 3) = 1 - F(3)$$



Now that we know the column $F(r)$

$$P(X > 3) = 1 - 0.8791 = 0.1209$$

You can also find it on the APP as $P(X \geq 4)$. However, we can always follow the principle of $1 - F(x)$ to find the complement



More than 3, does not include 3!



Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

d) What is the probability to find between 3 and 5 people who smoke **inclusive**?

X = number of people who smoke (within 10 people)

$n = 10$ $p = 0.20$ $q = 0.80$

$$P(3 \leq X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5)$$

$$P(3 \leq X \leq 5) = 0.2013 + 0.0881 + 0.0264 = 0.3158$$



Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

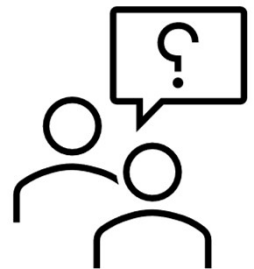
e) What is the probability that none of them smoke?

X = number of people who smoke (within 10 people)

$n = 10$ $p = 0.20$ $q = 0.80$

If we want none of them to smoke, what is the probability that we have to look for?

0	1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	---	----



Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

e) What is the probability that none of them smoke?

X = number of people who smoke (within 10 people)

$n = 10$ $p = 0.20$ $q = 0.80$

$$P(X = 0) = 0.1074$$



Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

f) How many people do we **expect** to be smokers within those 10? Calculate **Variance** and **Standard Deviation**.

X = number of people who smoke (within 10 people)

$$n = 10$$

$$p = 0.20$$

$$q = 0.80$$

When we talk about the expected value, we talk about the average. In Binomial distribution, we calculate it as follow:

$$E(X) = n * p$$

$$E(X) = 10 * 0.20 = 2 \text{ people}$$

Answer: We expect to find 2 people who smoke in this group

Binomial Distribution

It is known by previous studies that 20% of the population in Dublin smokes. If we randomly choose 10 people:

f) How many people do we **expect** to be smokers within those 10? Calculate **Variance** and **Standard Deviation**.

X = number of people who smoke (within 10 people)

$$n = 10 \quad p = 0.20 \quad q = 0.80$$

Formula

$$V(X) = n * p * q \quad S(X) = \sqrt{n * p * q}$$

$$V(X) = 10 * 0.20 * 0.80 = 1.6 \text{ people}^2$$

$$S(X) = \sqrt{10 * 0.20 * 0.80} = 1.2649 \text{ people}$$

Remember that the unit of measurement of the variance is expressed squared

Answer: We expect to find 2 people who smoke +/- 1.2649 people.

THAT'S ALL FOR TODAY

THANK YOU

