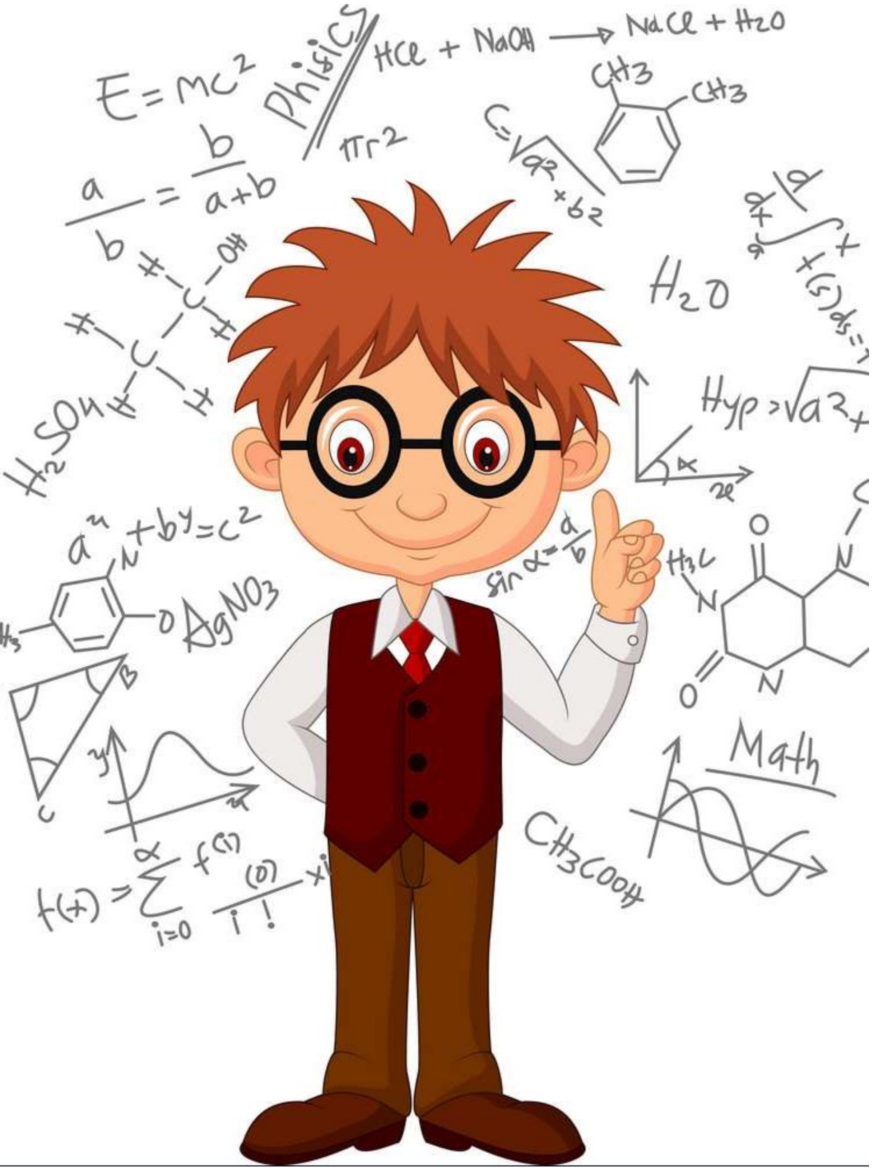
# Statistics for Data Analytics

Lecturer: Marina Iantorno
E-mail: miantorno@cct.ie

CCT College Dublin
Computing • IT • Business




# In today’s class we will cover:

Continuous Distribution:

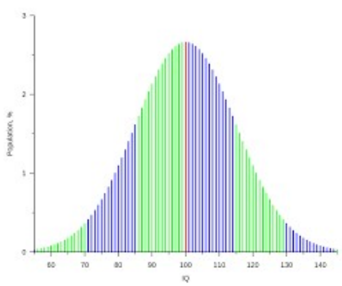
❑Normal Distribution

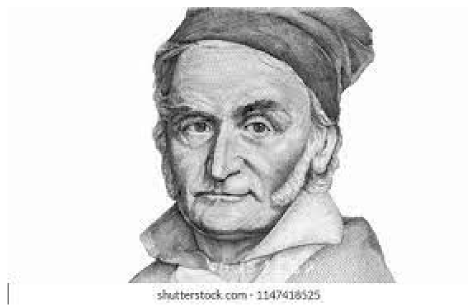
# Normal Distribution

Normal distribution is also known as Gaussian distribution. This is a continuous distribution and one of the most important in Statistics.

Initially, Abraham DeMoivre (1700) was solving a gamble issue using the Binomial distribution and he tried to transform it in a continuous variable. Later on, about 1800, Carl Gauss took that idea, and he created the Normal distribution. The main idea was to prove that it was possible to create a formula that includes an infinite number of data points.

A particular characteristic of this distribution is that it is represented by a graph that looks like a bell, a reason why it is commonly called  “the Gauss Bell Curve”.





# Normal Distribution

Some characteristics about this Distribution.

- Although it takes into consideration an infinite number of data points, the expected value and the variance are finite numbers.
- This distribution calculates a cumulation of probabilities, therefore we cannot calculate the probability to get an exact value, it will be always greater or less than, but never equals to.
- This distribution is always symmetric, which means that the curve will never be skewed to one side, and the expected value (average) is always in the middle of the bell curve.
- Some discrete distributions, when we analyse them in big samples, tend to become Normal (see Central Limit Theory for more information).

# Normal Distribution

It is known that the IQ of the kids at a certain school is normally distributed with an average of 100 points and a standard deviation of 10 points. If we randomly pick one of the kids:

a) What is the probability that s/he has an IQ lower than 90 points?

b) What is the probability that s/he has an IQ lower than 105 points?

c) What is the probability that s/he has an IQ between 90 and 105 points?

d) What is the probability that s/he has an IQ greater than 110?

e) If we picked 3 kids, what is the probability to find 2 of them with an IQ level greater than 110 points?

# Normal Distribution

It is known that the IQ of the kids at a school is normally distributed with an average of 100 points a standard deviation of 10 points.

X = IQ level of the students (in points)

The unit of measurement will always be something that counts infinite data points

The exercise usually says that the variable is normally distributed, follows a normal distribution, follows an approximately normal distribution, etc.

$\mu = 100$ → Average

$\sigma = 10$ → Standard Deviation

# Normal Distribution

a) What is the probability that he has an IQ lower than 90 points?

X = IQ level of the students (in points)

$\mu = 100$

$\sigma = 10$

$P( X < 90) =$

To calculate this probability, we need to go through a process called "standardization" in which we will need to calculate the requested data point in another plane. Let's do this together!

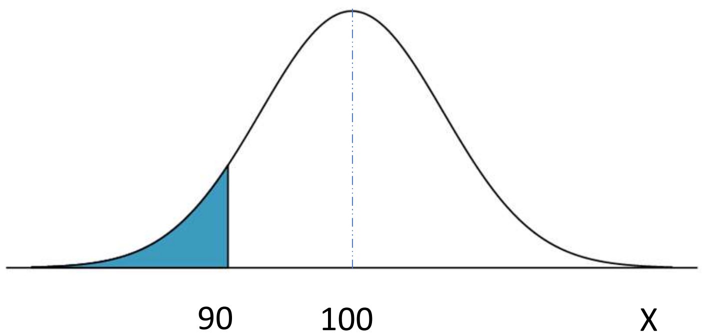
# Normal Distribution

X = IQ level of the students (in points)

$\mu = 100$

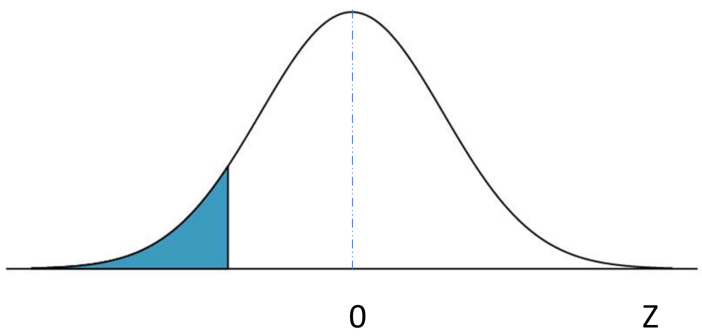
$\sigma = 10$

P( X < 90) =




Standardization: We need to assign to the value in X a value in Z. In other words, we need to transform 90 in a possible value for Z. We do this using the formula below:

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$



# Normal Distribution

X = IQ level of the students (in points)

$\mu = 100$

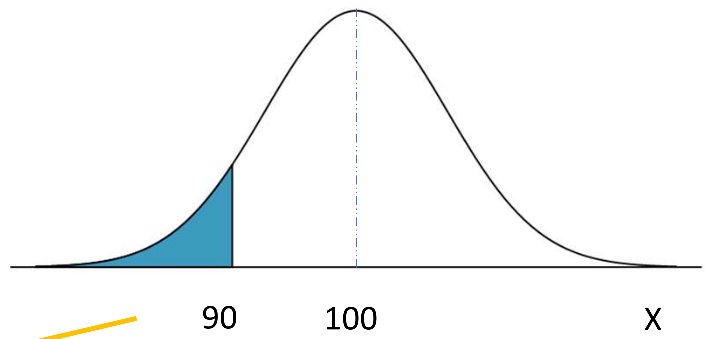
$\sigma = 10$

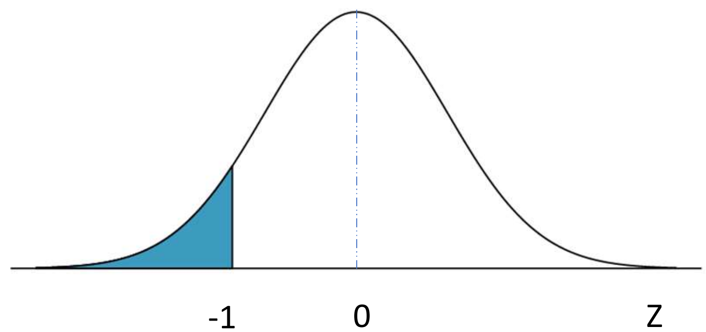
$P( X < 90) =$

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

$$Z = \frac{90 - 100}{10} = -1$$

Once we have the value of Z, we can find the probability using the tables.

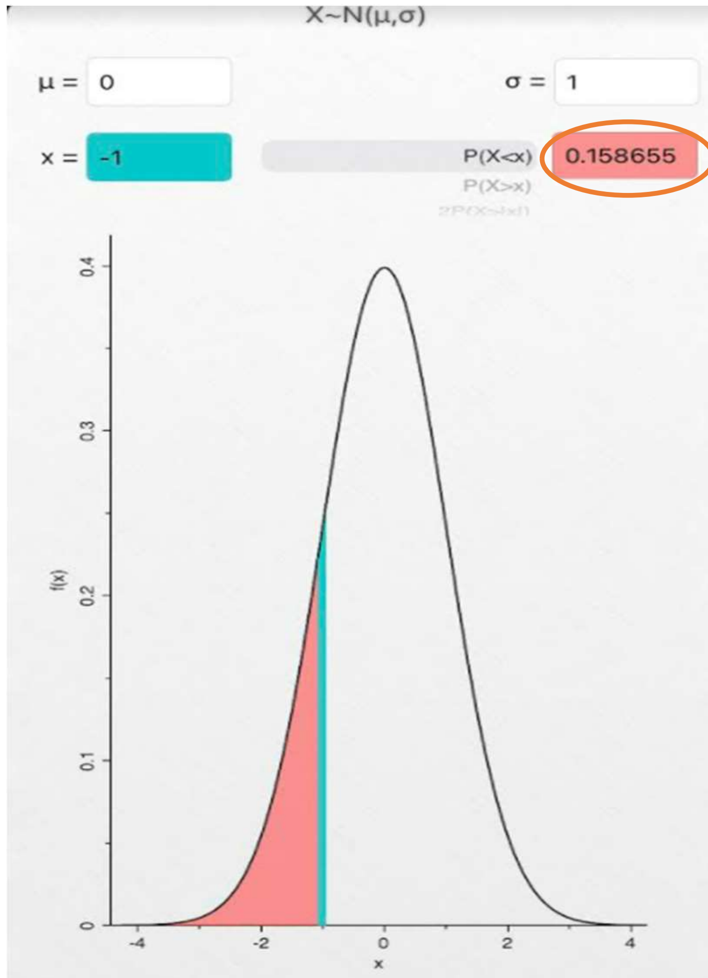


# Normal Distribution

Do not change this

Add z value here




Do not change this

Read your result here

# Normal Distribution

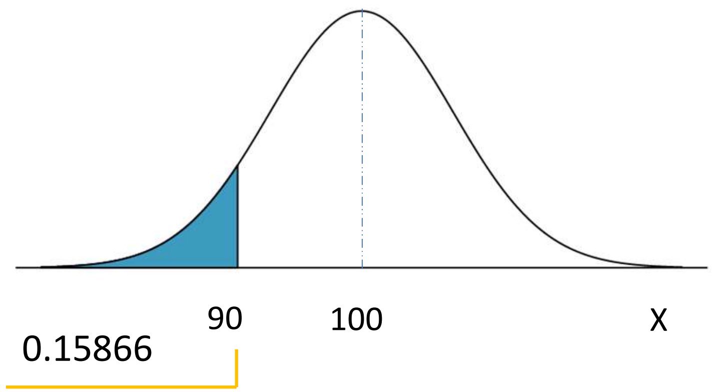
X = IQ level of the students (in points)

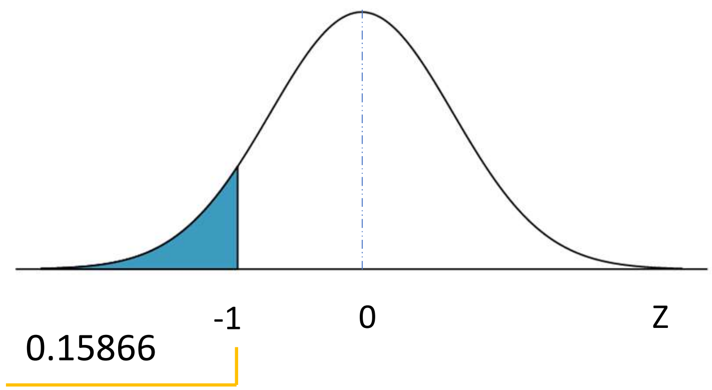
$\mu = 100$

$\sigma = 10$

$P(X < 90) = \boxed{0.15866}$

Answer: the probability that a kid has an IQ level lower than 90 points is 0.15866 or 15,86%



# Normal Distribution

b) What is the probability that he has an IQ lower than 105 points?

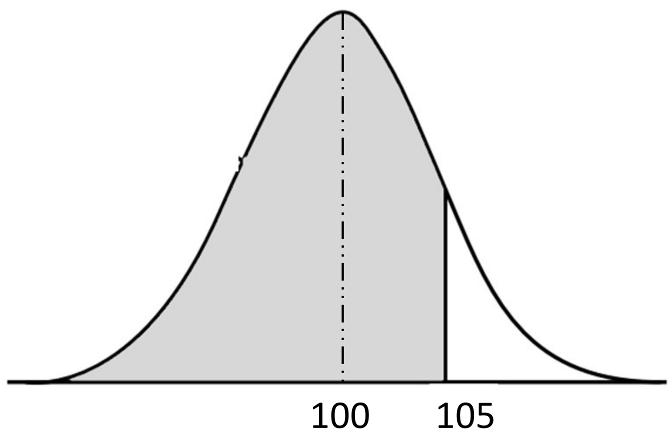
X = IQ level of the students (in points)

$\mu = 100$

$\sigma = 10$

$P(X < 105) =$

Here we have an example very similar to the previous one



# Normal Distribution

X = IQ level of the students (in points)

$\mu = 100$

$\sigma = 10$

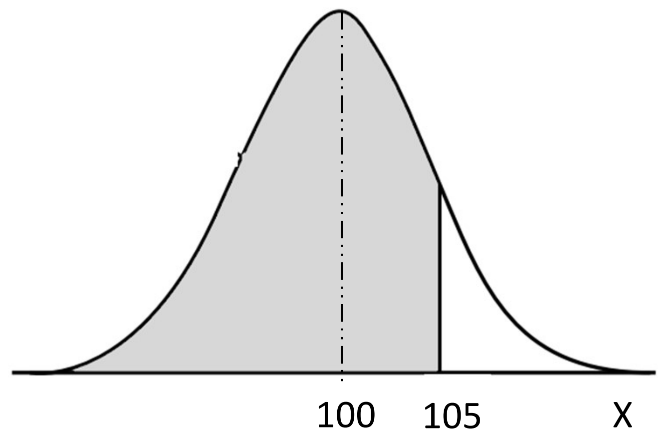
P( X < 105) =

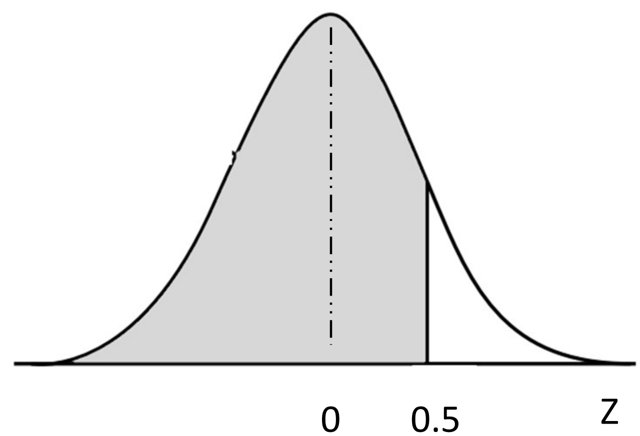
Let's use the formula to standardize

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

$$Z = \frac{105 - 100}{10} = 0.5$$

Now we can look for the probability!



# Normal Distribution

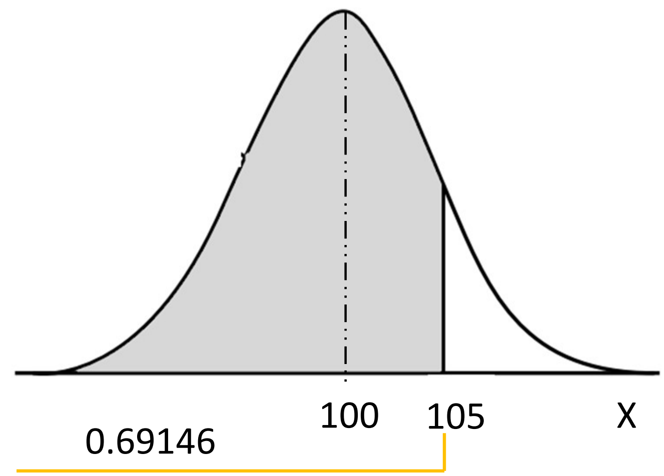
X = IQ level of the students (in points)

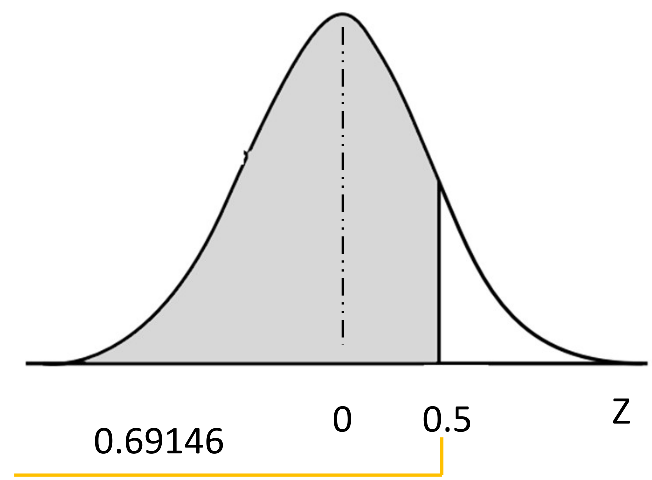
μ = 100

σ = 10

P( X < 105) = 0.69146

Answer: The probability that a kid has an IQ level lower than 105 points is 0.69146



# Normal Distribution

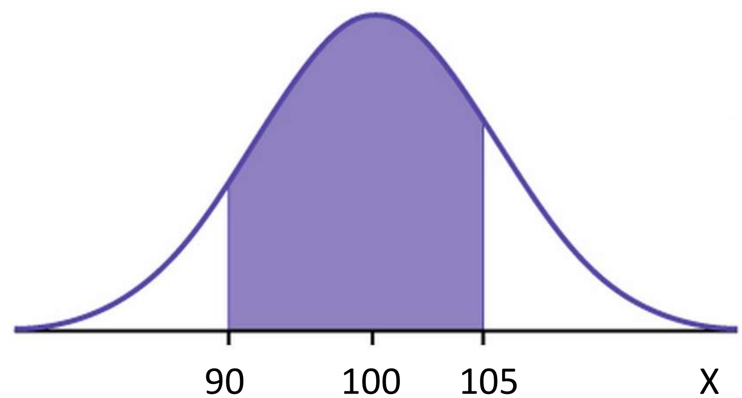
c) What is the probability that he has an IQ between 90 and 105 points?

X = IQ level of the students (in points)

$\mu = 100$

$\sigma = 10$

$P( 90 < X < 105) =$


In this case, we have to calculate the probabilities between two points. If the distribution reads the cumulated probabilities only, how can we do here?

# Normal Distribution

c) What is the probability that he has an IQ between 90 and 105 points?

X = IQ level of the students (in points)

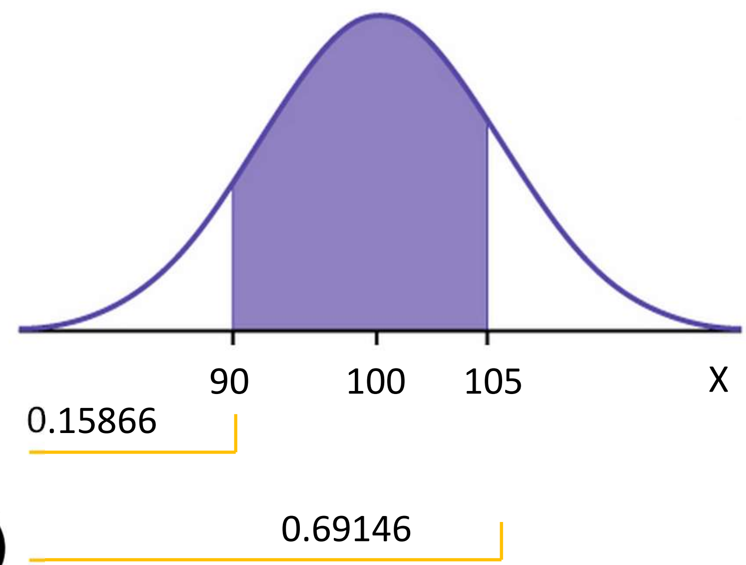
$\mu = 100$

$\sigma = 10$

$P( 90 < X < 105) =$

Let's think it together:

We previously calculated the probability that a kid has an IQ lower than 90 and an IQ lower than 105

# Normal Distribution

c) What is the probability that he has an IQ between 90 and 105 points?

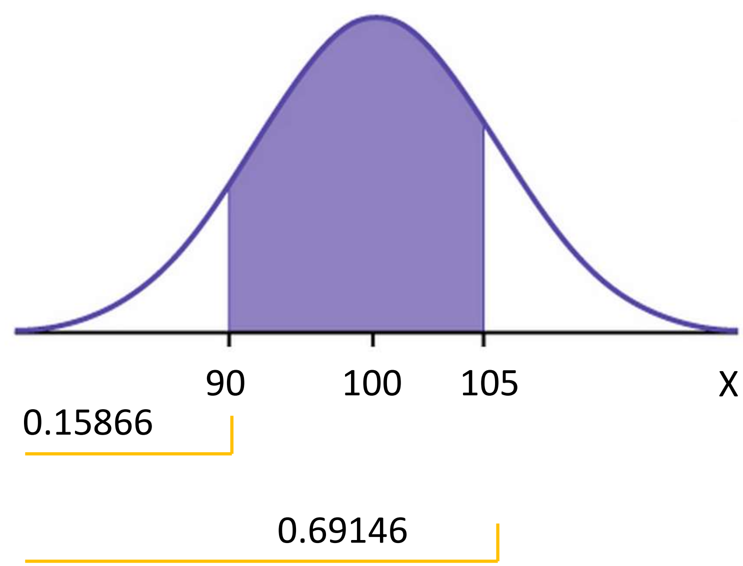
X = IQ level of the students (in points)

$\mu = 100$

$\sigma = 10$

P( 90 < X < 105) = 0.69146 – 0.15866

P( 90 < X < 105) = $\boxed{0.5328}$




Answer: The probability that a kid has an IQ level between 90 and 105 points is 0.5328

# Normal Distribution

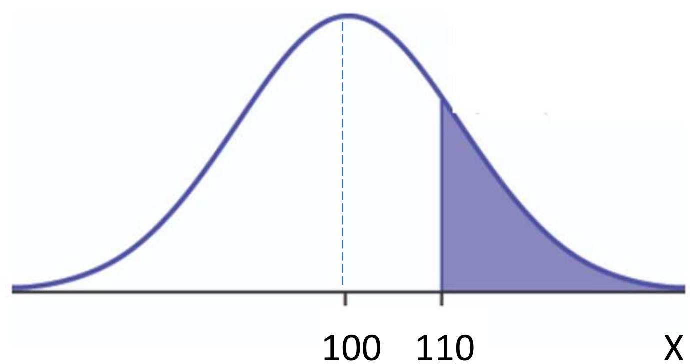
d) What is the probability that he has an IQ greater than 110?

X = IQ level of the students (in points)

$\mu = 100$

$\sigma = 10$

$P( X > 110) =$




This exercise is different than the previous ones. In this case we need to know the cumulated probabilities to the right.

# Normal Distribution

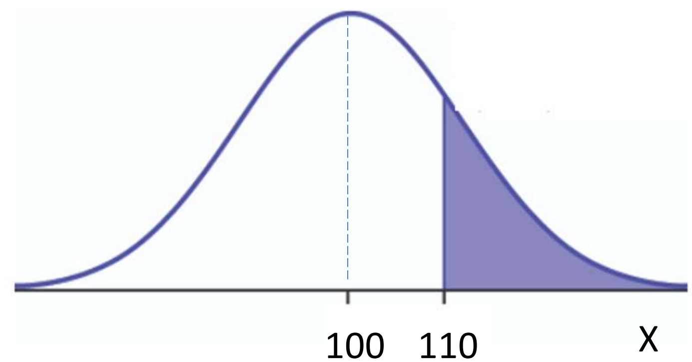
d) What is the probability that he has an IQ greater than 110?

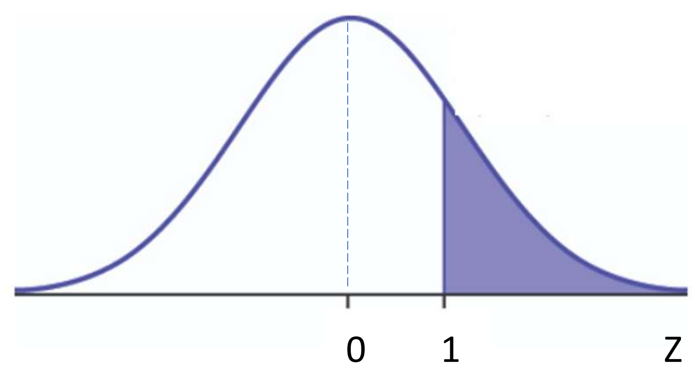
X = IQ level of the students (in points)

$\mu = 100$

$\sigma = 10$

P( X > 110) =




$$Z = \frac{110 - 100}{10} = 1$$

# Normal Distribution

d) What is the probability that he has an IQ greater than 110?

X = IQ level of the students (in points)

$\mu = 100$

$\sigma = 10$

P( X > 110) = 1 – 0.84134

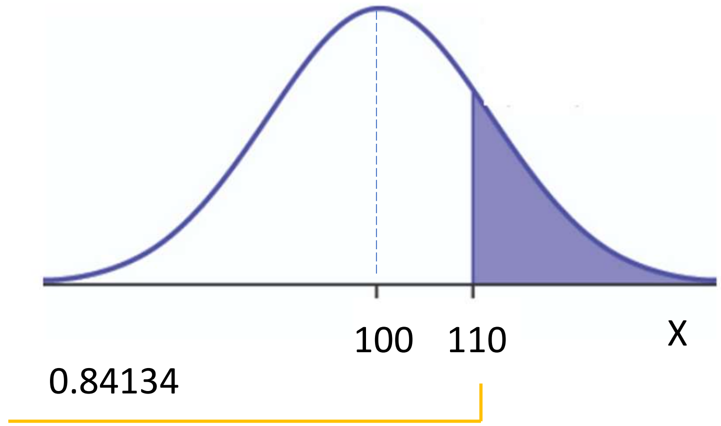
P( X > 110) = $\boxed{0.15866}$

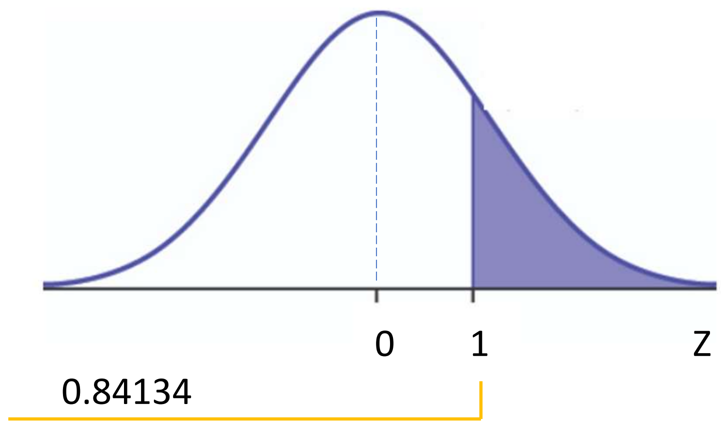
Now we have to use the complementary probability:

If we have the cumulated probabilities to the left, we

can have the same to the right. Or else, we can

look for the probability on the app changing the sign.



# Normal Distribution

d) If we picked 3 kids, what is the probability to find 2 of them with an IQ level greater than 110 points?

CAREFUL!

Before doing anything, let’s read the question and understand the variable in place

# Normal Distribution

d) If we picked 3 kids, what is the probability to find 2 of them with an IQ level greater than 110 points?

You are right!

We are counting kids, so our variable changed.

$$Y = \underbrace{\text{number of kids}}_{\textit{Elements}}\ \underbrace{\text{with an IQ level greater than 110 points}}_{\textit{Characteristic}}\ \underbrace{\text{(within 3 kids)}}_{\textit{Limit}}$$

What distribution do we have here?

$n = 3$

$p = 0.15866 \rightarrow p = 0.16$

# Normal Distribution

d) If we picked 3 kids, what is the probability to find exactly 2 of them with an IQ level greater than 110 points?

You are right!

Y is a Binomial distribution! Now we need the parameters

Y = number of kids with an IQ level greater than 110 points (within 3 kids)

$n = 3$

$p = 0.16$

In this case, p comes from the exercise “e”

$P( Y = 2) = \boxed{0.0645}$


Practice
Time

# Normal Distribution

## Exercise

The height of the plants in a greenhouse follows a normal distribution with an average of 70 cm and a variance of 100 $cm^2$. If we pick a plant:

a) Calculate the probability that the height of the plant is between 58 and 65 cm.

b) Calculate the probability that the height of the plant is less than 75 cm.

c) Calculate the probability that the height of the plant is more than 60 cm.

d) What is the height of the 20% of the smaller plants?

THAT'S ALL FOR TODAY

# THANK YOU