WELLCOME

STATISTIC WITH PYTHON

EMBARKING ON A JOURNEY INTO DATA SCIENCE

YA MANON

PROBABILITY REVIEW

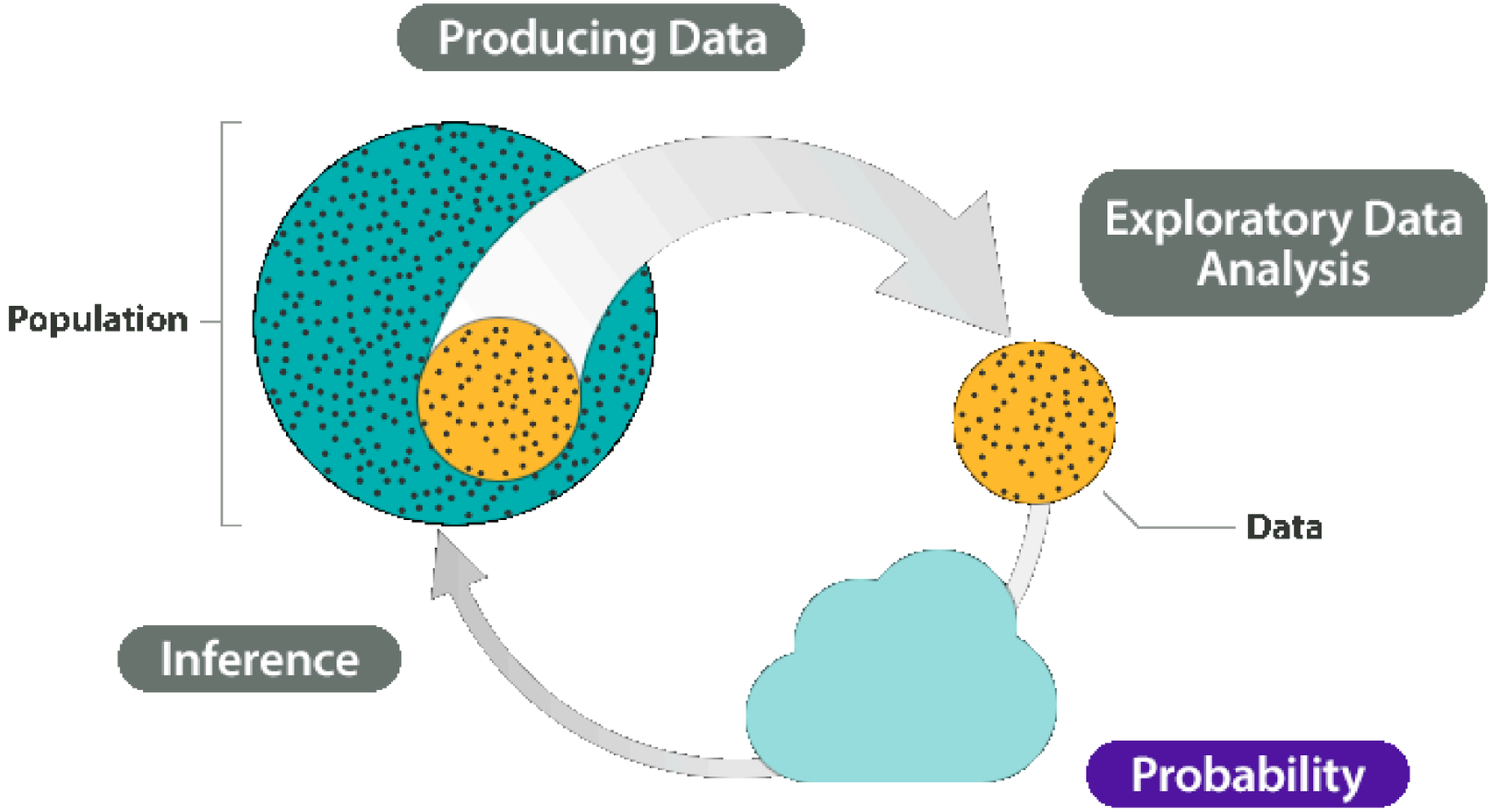
PROBABILITY REVIEW

TOPICS WE'LL COVER:

- **WHY Probability?**
- **Calculate Probab**
- **Random**
- **Rule of Probability**

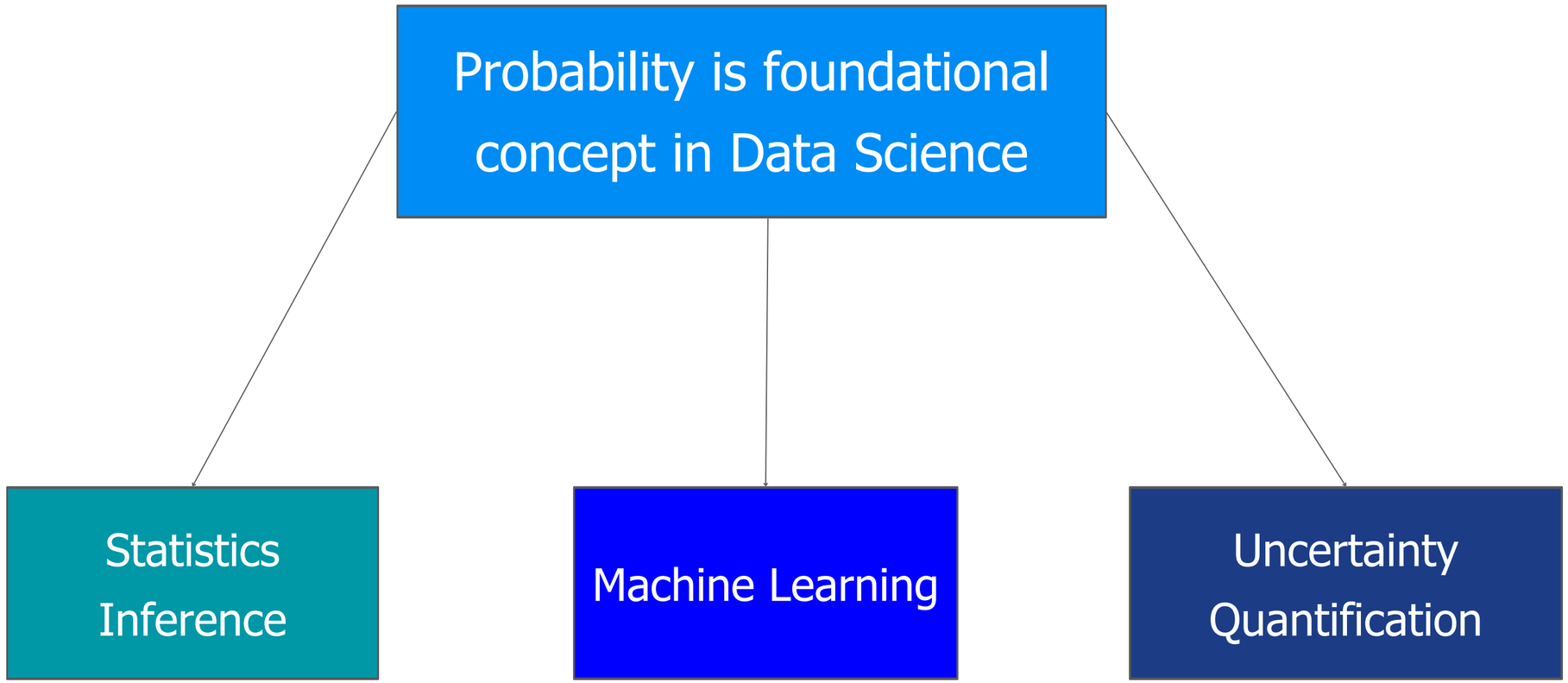
WHY PROBABILITY?

- Why Probability?
- Calculate Probab
- Normal Distribution
- Rule of Probability



WHY PROBABILITY?

- **Why Probability?**
- Calculate Probab
- Normal Distribution
- Rule of Probability



WHY PROBABILITY?

Statement “There is a 20% probability of rain today.”

Interpretation 1 It will rain for 20% of the day 4.8 hours.

Interpretation 2 There is a 1 in 5 chance that it will rain.

Interpretation 3 We can be 20% confident that it will rain today.

PROBABLITY CALCULATION

$$P(E) = n(E) / n(S)$$

Sample space of an experiment, denoted by ***S***, is the set of all possible outcomes of that experiment.

Example: - Sample space for single coin toss = {Heads, Tails}

- Sample space for single die roll ={1, 2,3, 4, 5,6}

Event *A* is any collection (subset) of outcomes contained in the sample space ***S.***

That is, if ***A*** is an event then ***A =*** { ***A***: in(ϵ) ***S*** }.

RANDOM

- Why Probability?
- Calculate Probab
- Random
- Rule of Probability

RULE OF PROBABILITY

- Why Probability?
- Calculate Probab
- Random
- Rule of Probability

PROBABILITY DISTRIBUTIONS

PROBABILITY DISTRIBUTIONS

In this section we'll cover modeling data with **probability distributions**, and use the normal distribution to calculate probabilities and make estimates about normal populations

TOPICS WE'LL COVER:

- **Distribution Basics**
- **Distribution Tyes**
- **Normal Distribution**
- **Z-Scores**
- **Probabilities**
- **Values Estimates**

GOALS FOR THIS SECTION:

- *Understand the concept of a probability distribution, and its relationship with frequency distributions*
- *Learn about the different types of probability distributions, and their main differences*
- *Identify the properties of the normal distribution*
- *Calculate probabilities, values, and z-scores from normal distributions using Excel functions*

PROBABILITY DISTRIBUTIONS

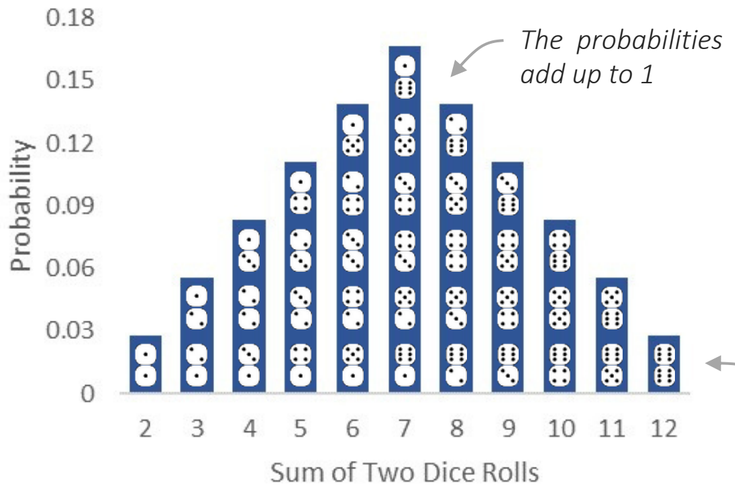
- **Distribution Basics**
- Distribution Tyes
- Normal Distribution
- Z-Scores
- Probabilities
- Values Estimates

A **probability distribution** represents a variable's idealized frequency distribution It shows all the possible values a variable can take, and their chances of occurring

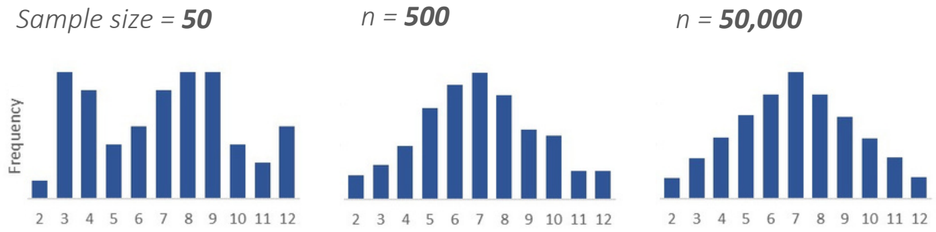
Frequencies in a sample are based on the underlying probabilities of those values occurring

EXAMPLE *Results of rolling two dice*

PROBABILITY DISTRIBUTION *(Population)**:***



FREQUENCY DISTRIBUTION *(Sample)**:***



In an infinite sample, a variable's relative frequency distribution is equal to its probability distribution!

*This is known as a **binomial distribution**, and it can be used to calculate probabilities on the outcome of rolling two dice (without rolling them fifty thousand times!)*

TYPES OF PROBABILITY DISTRIBUTIONS

There are two **types of probability distributions**: Discrete & Continuous

1)**Discrete** probability distributions

Uniform

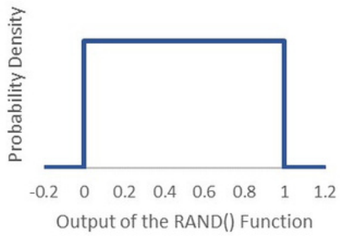
Binomial

Poisson

The height of each bar is its probability There are "gaps" between the numbers

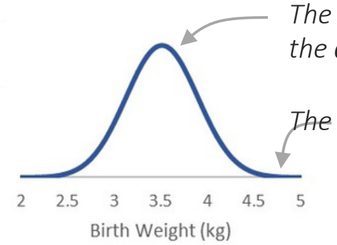
2) **Continuous** probability distributions

Uniform



Exponential

Normal



*The height of the curve is NOT its probability, the **area under the curve** is (more on this later!)*

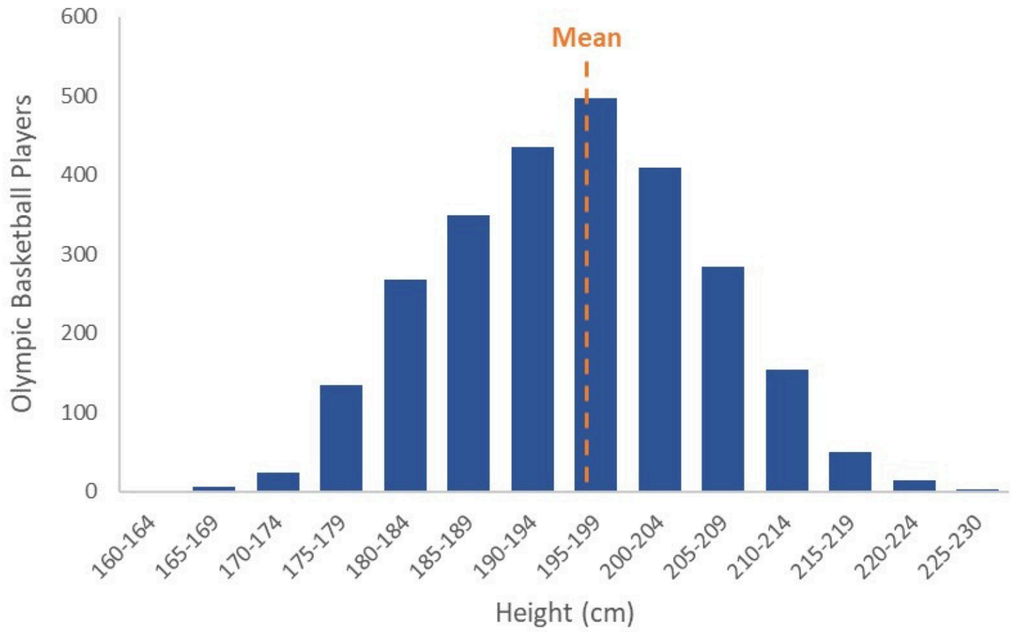
The numbers can take any value

THE NORMAL DISTRIBUTION

Many numerical variables naturally follow a **normal distribution**, or “bell curve”

- Normal distributions are symmetrical around the mean and have no skew *(mean = median)*, with most data concentrated around its center and flaring out in “tails” on both ends

EXAMPLE | *Olympic Basketball Player Heights*



HEY THIS IS IMPORTANT!

Since they are so common, many statistical tests are designed for normally distributed populations, which is why we’ll mostly focus on the normal distribution in the course

THE NORMAL DISTRIBUTION

- Distribution Basics
- Distribution Tyes
- **Normal Distribution**
- Z-Scores
- Probabilities
- Values Estimates

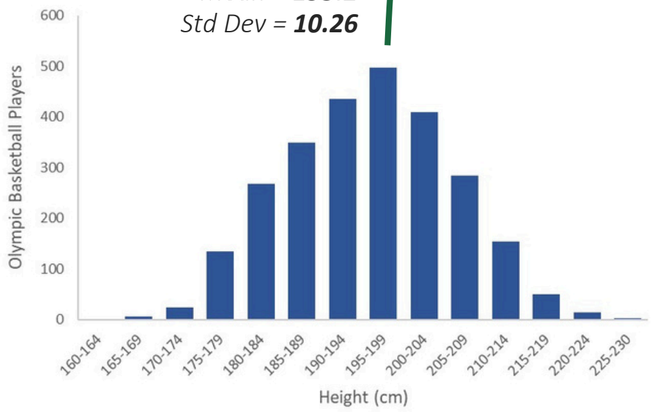
The normal distribution is described by two values: the **mean** & **standard deviation**

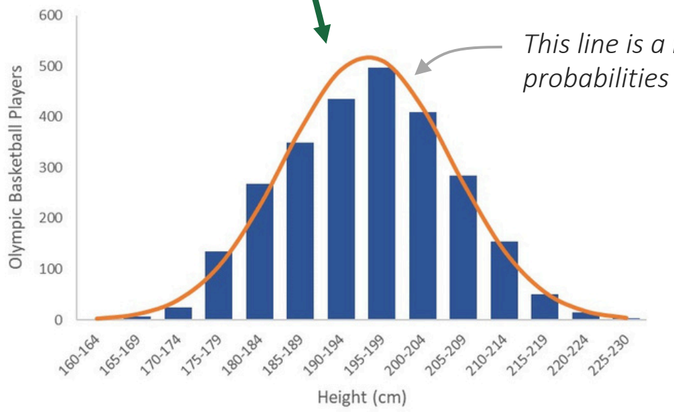
$$\frac{1}{\sigma 2\sqrt{\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

*This is the **probability density function** for the normal distribution, which determines the height of the normal curve at any value (x) given a mean (μ) and a standard deviation (σ)*

(don't worry, there's an Excel function for it!)

*Mean = **195.2***
*Std Dev = **10.26***



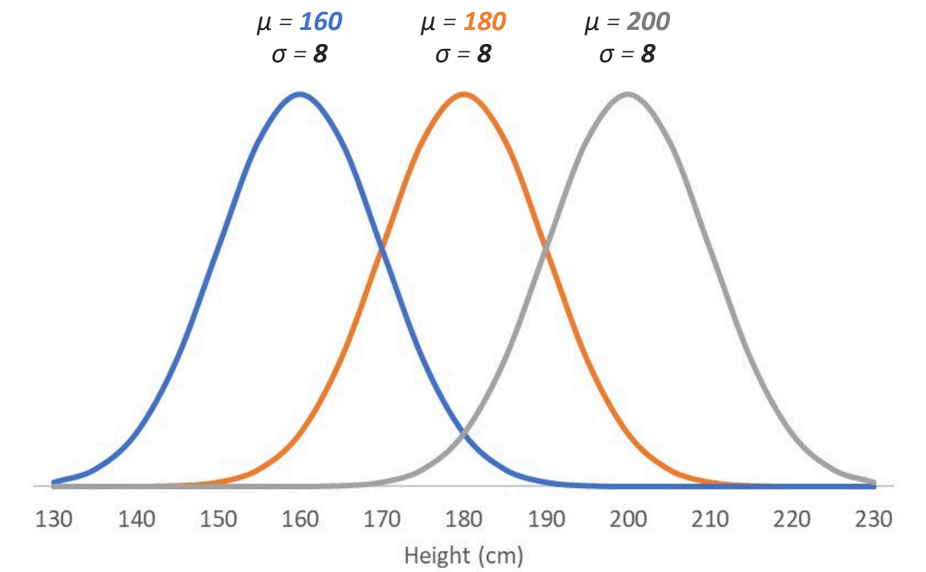


This line is a model that we can use to calculate probabilities of Olympic Basketball player heights!

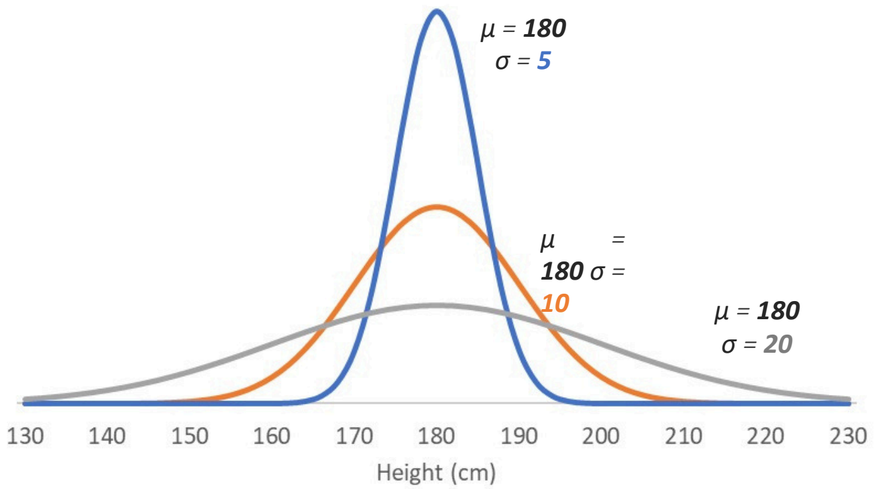
THE NORMAL DISTRIBUTION

The normal distribution is described by two values: the **mean** & **standard deviation**

- The mean determines the *center* of the distribution, and the standard deviation its *width*



*Changing the mean **shifts** the curve along the x axis*



*Changing the standard deviation **squeezes** or **stretches** the curve*

Z-SCORES

A **z-score** indicates how many standard deviations away from the mean a value lies

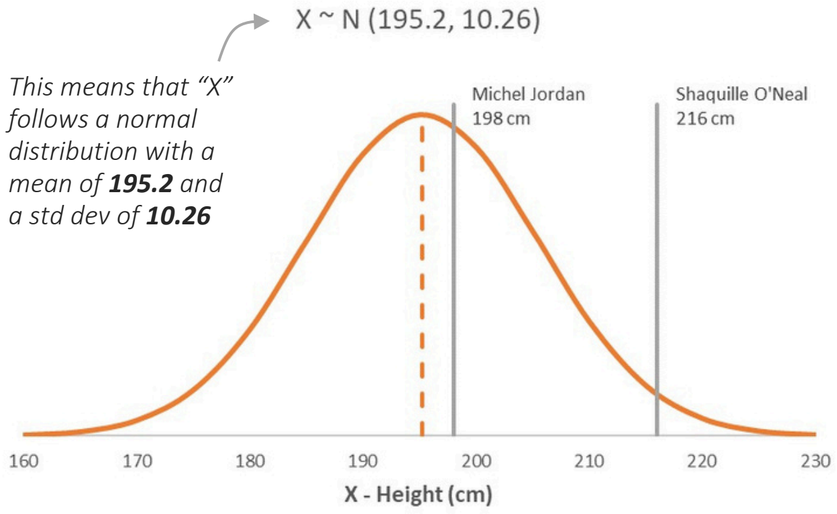
$$z = \frac{x - \mu}{\sigma}$$

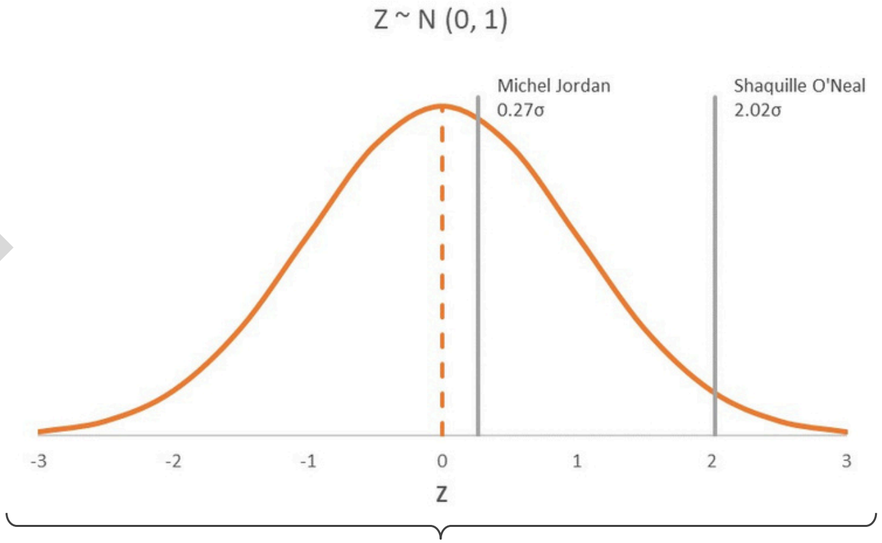
To calculate a z-score for a value, simply subtract the mean and divide by the standard deviation

(or use the STANDARDIZE function)

$$z = \frac{198 - 195.2}{10.26} = 0.27$$

*This means that "X" follows a normal distribution with a mean of **195.2** and a std dev of **10.26***



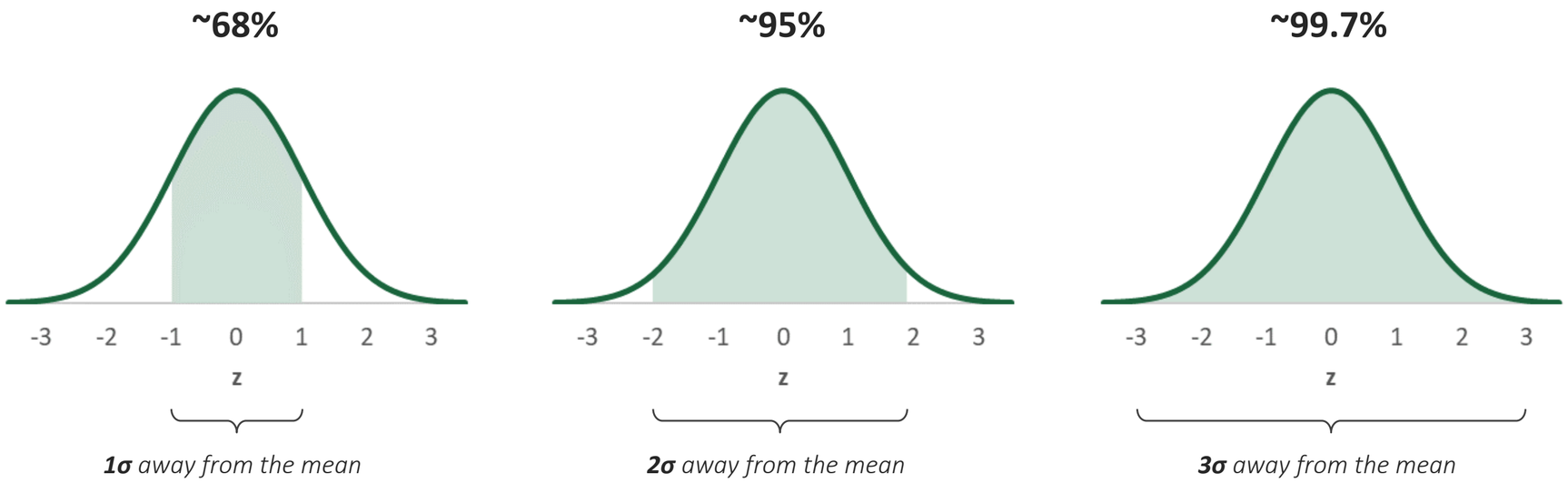


*This is known as the **standard normal distribution**, or z-distribution, and has a mean of 0 and a standard deviation of 1*

THE EMPIRICAL RULE

- Distribution Basics
- Distribution Tyes
- Normal Distribution
- Z-Scores
- **Probabilities**
- Values Estimates

The **empirical rule** outlines where most values fall in a normal distribution



PRO TIP: Beyond using a histogram to determine whether your data is distributed normally, check if it follows the empirical rule

EXCEL NORMAL DISTRIBUTION FUNCTIONS

- Distribution Basics
- Distribution Tyes
- Normal Distribution
- Z-Scores
- **Probabilities**
- Values Estimates

These **Excel functions** help make calculations related to the normal distribution:

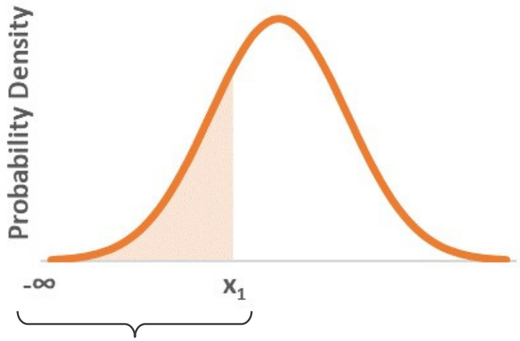
NORM.DIST()	*Returns the cumulative probability or the probability density at an x value from a given normal distribution*	=**NORM.DIST**(x, μ, σ, cumulative)
NORM.INV()	*Returns the x value in a given normal distribution at a specified cumulative probability*	=**NORM.INV**(probability, μ, σ)
STANDARDIZE()	*Returns the z-score for a specified x value in a given normal distribution*	=**STANDARDIZE**(x, μ, σ)
NORM.S.DIST()	*Returns the cumulative probability or the probability density at a z-score from the standard normal distribution*	=**NORM.S.DIST**(z, cumulative)
NORM.S.INV()	*Returns the z-score in the standard normal distribution at a specified cumulative probability*	=**NORM.S.INV**(probability)

CALCULATING PROBABILITIES

If a variable follows a normal distribution, you can **calculate the probability** of randomly obtaining a value within a specified range

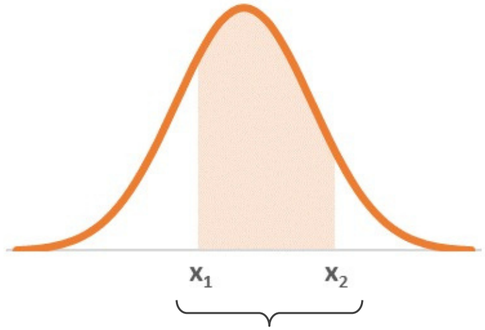
- This is determined by the area under the curve in that range

Less than or equal to "x1"



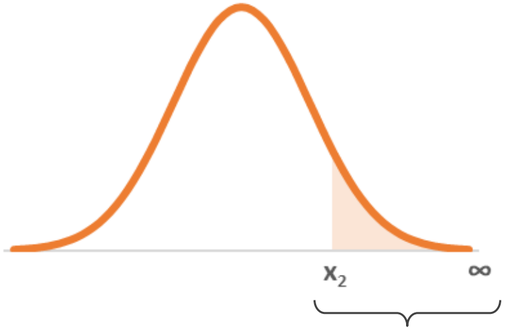
*The area from negative infinity to "x1" is the **cumulative probability***

Between "x1" and "x2"



This is the cumulative probability of "x2" minus the cumulative probability of "x1"

Greater than or equal to "x2"



This is 1 (the entire area under the curve) minus the cumulative probability of "x2"

HEY THIS IS IMPORTANT!

You CANNOT calculate the probability of obtaining an x value *exactly* – there's no area under a single point!

THE NORM.DIST FUNCTION

NORM.DIST()

Returns the cumulative probability or the probability density at "x" from a normal distribution

=**NORM.DIST**(x, mean, standard_dev, cumulative)

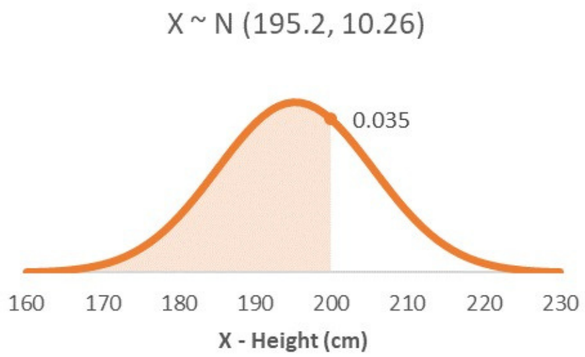
*The **value** to calculate the probability for*

*The **mean** & **standard deviation** for the normal distribution of the population*

***TRUE**: The area under the curve **FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **2 meters tall or shorter**?"*



=**NORM.DIST**(200, 195.2, 10.26, TRUE) **= 0.68**

This is the probability!

=**NORM.DIST**(200, 195.2, 10.26, FALSE) **= 0.035**

This is just the height of the curve

THE NORM.DIST FUNCTION

- Distribution Basics
- Distribution Tyes
- Normal Distribution
- Z-Scores
- **Probabilities**
- Values Estimates

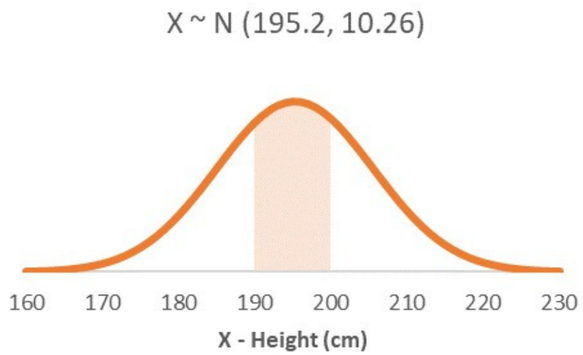
NORM.DIST() — *Returns the cumulative probability or the probability density at "x" from a normal distribution*

=**NORM.DIST**(x, mean, standard_dev, cumulative)

- *The **value** to calculate the probability for*
- *The **mean** & **standard deviation** for the normal distribution of the population*
- ***TRUE**: The area under the curve **FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **between 1.9 and 2 meters tall**?"*



=**NORM.DIST**(200, 195.2, 10.26, TRUE) **= 0.68**

=**NORM.DIST**(190, 195.2, 10.26, TRUE) **= 0.3061**

=0.68-0.306 **= 0.3739**

This is the probability!

THE NORM.DIST FUNCTION

- Distribution Basics
- Distribution Tyes
- Normal Distribution
- Z-Scores
- **Probabilities**
- Values Estimates

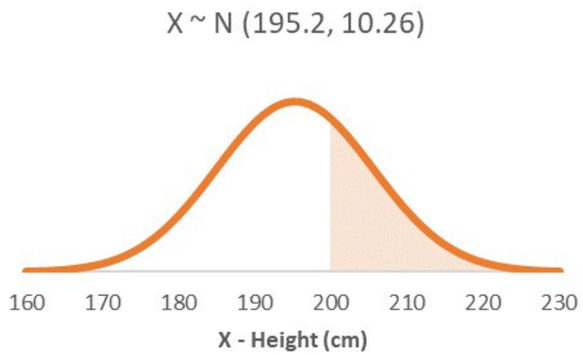
NORM.DIST() — *Returns the cumulative probability or the probability density at "x" from a normal distribution*

=**NORM.DIST**(x, mean, standard_dev, cumulative)

- *The **value** to calculate the probability for*
- *The **mean** & **standard deviation** for the normal distribution of the population*
- ***TRUE**: The area under the curve **FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **at least 2 meters tall**?"*



=**NORM.DIST**(200, 195.2, 10.26, TRUE) **= 0.68**

=1-**NORM.DIST**(190, 195.2, 10.26, TRUE) **= 0.32**

This is the probability!

The cumulative probability under the entire curve is equal to 1

(it's every value possible!)

THE NORM.S.DIST FUNCTION

- Distribution Basics
- Distribution Tyes
- Normal Distribution
- Z-Scores
- **Probabilities**
- Values Estimates

NORM.S.DIST()

Returns the cumulative probability or the probability density at "z" from the z-distribution

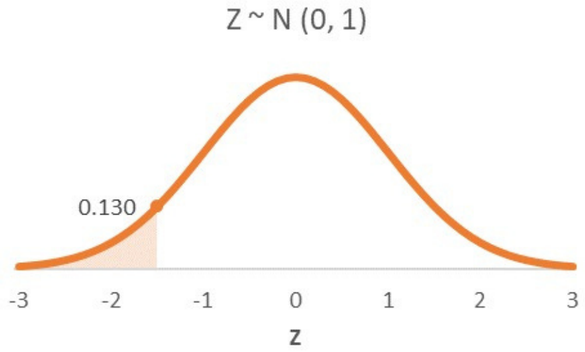
=**NORM.S.DIST**(z, cumulative)

*The **z-score** to calculate the probability for*

***TRUE**: The area under the curve **FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **at least 1.5 standard deviations shorter than the mean**?"*



=**NORM.S.DIST**(-1.5, TRUE) **= 0.066**

This is the probability!

=**NORM.S.DIST**(-1.5, FALSE) **=**

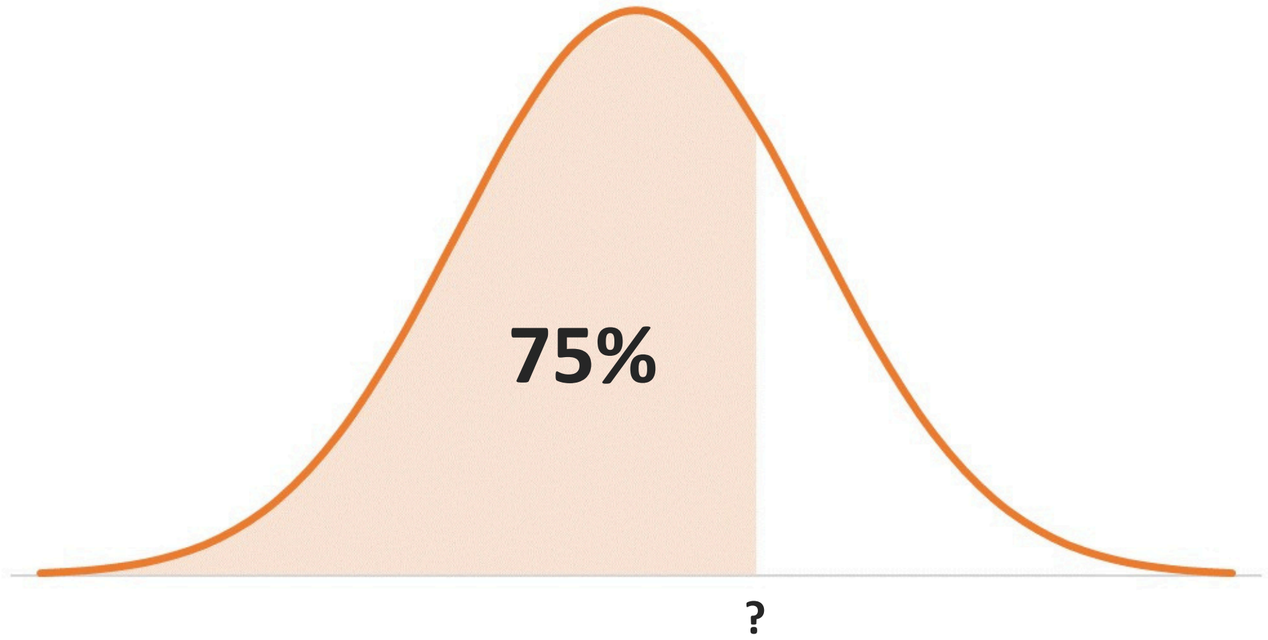
This is just the height of the curve

0.130

ESTIMATING VALUES

- Distribution Basics
- Distribution Tyes
- Normal Distribution
- Z-Scores
- Probabilities
- **Values Estimates**

If a variable follows a normal distribution, you can **estimate the value of “x” or “z”** at a specified cumulative probability



THE NORM.INV FUNCTION

- Distribution Basics
- Distribution Tyes
- Normal Distribution
- Z-Scores
- Probabilities
- **Values Estimates**

NORM.INV()

Returns the x value in a normal distribution at a specified cumulative probability

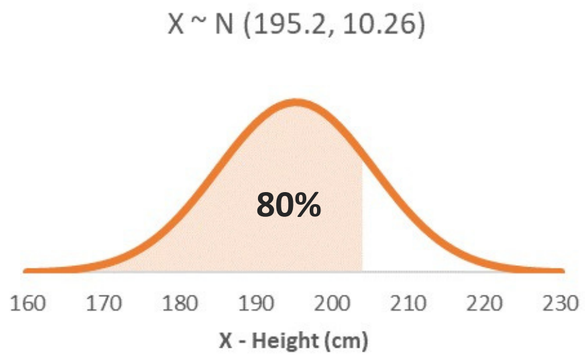
=**NORM.INV**(probability, mean, standard_dev)

*The **cumulative probability** for the value you want*

*The **mean** & **standard deviation** for the normal distribution of the population*

Possible question:

*"How tall do you need to be to be **taller than 80%** of Olympic Basketball Players?"*



=**NORM.INV**(0.8, 195.2, 10.26) **= 203.8** cm

THE NORM.S.INV FUNCTION

- Distribution Basics
- Distribution Tyes
- Normal Distribution
- Z-Scores
- Probabilities
- **Values Estimates**

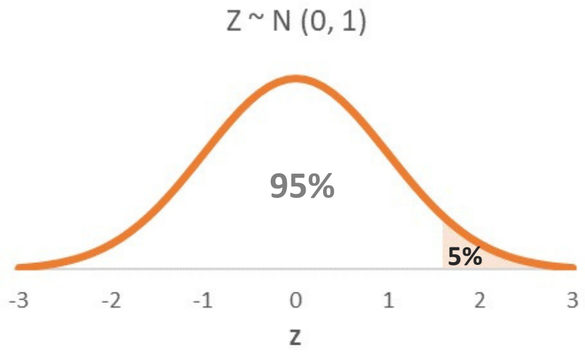
NORM.S.INV() — *Returns the z-score in the standard normal distribution at a specified cumulative probability*

=**NORM.S.INV**(probability)

*The **cumulative probability** for the z-score you want*

Possible question:

*"The **top 5%** of Olympic Basketball Players are how many standard deviations taller than the mean?"*



=**NORM.S.INV**(1-0.05) **= 1.64** σ

Remember that the cumulative probability starts from negative infinity, so for the "top 5%" the probability is 95% (1-5%)

KEY TAKEAWAYS: PROBABILITY DISTRIBUTIONS

A probability distribution is an **idealized frequency distribution**

- *It shows all the possible values the variable can take, and the probability of each value occurring*

Many variables naturally follow a **normal distribution**

- *The data is symmetrical around its mean, and flares out in "tails" (the width depends on the standard deviation)*

The probability in a normal distribution is the **area under its curve**

- *It can only be calculated in intervals, not for exact values!*

There are **Excel functions** to solve normal probability problems

- *NORM.DIST and NORM.S.DIST let you calculate the probability of randomly obtaining values in specified ranges*
- *NORM.INV and NORM.S.INV let you estimate values or z-scores based on their cumulative probabilities*

MAVEN MEDICAL CENTER | PROJECT BRIEF

You are a Data Analyst at the **Maven Medical Center** in Springfield, MA and just received a project request from the chief gynecologist

From: **Betty Birth** *(Chief Gynecologist)*

Subject: **Need some probability figures**

Good morning!

We've had over 30% of the babies born this year weigh under 2.5kg, which is considered low. The percentage itself seems a little high to me though. Is there any way you could check what the probability of a baby weighing under 2.5kg is with the data we have?

I could also use the number of births we've had so far in the top & bottom 1% if possible.

Thank you!

Birth_Weights.xlsx

Reply Forward

Key Objectives

1. Check if the weights can be assumed to follow a normal distribution
2. If so, calculate the probability of a baby weighing 2.5kg or less
3. Estimate the values at the 1% and 99% cumulative probabilities
4. Count the number of births under and over those thresholds