DSC Program 2020-2021

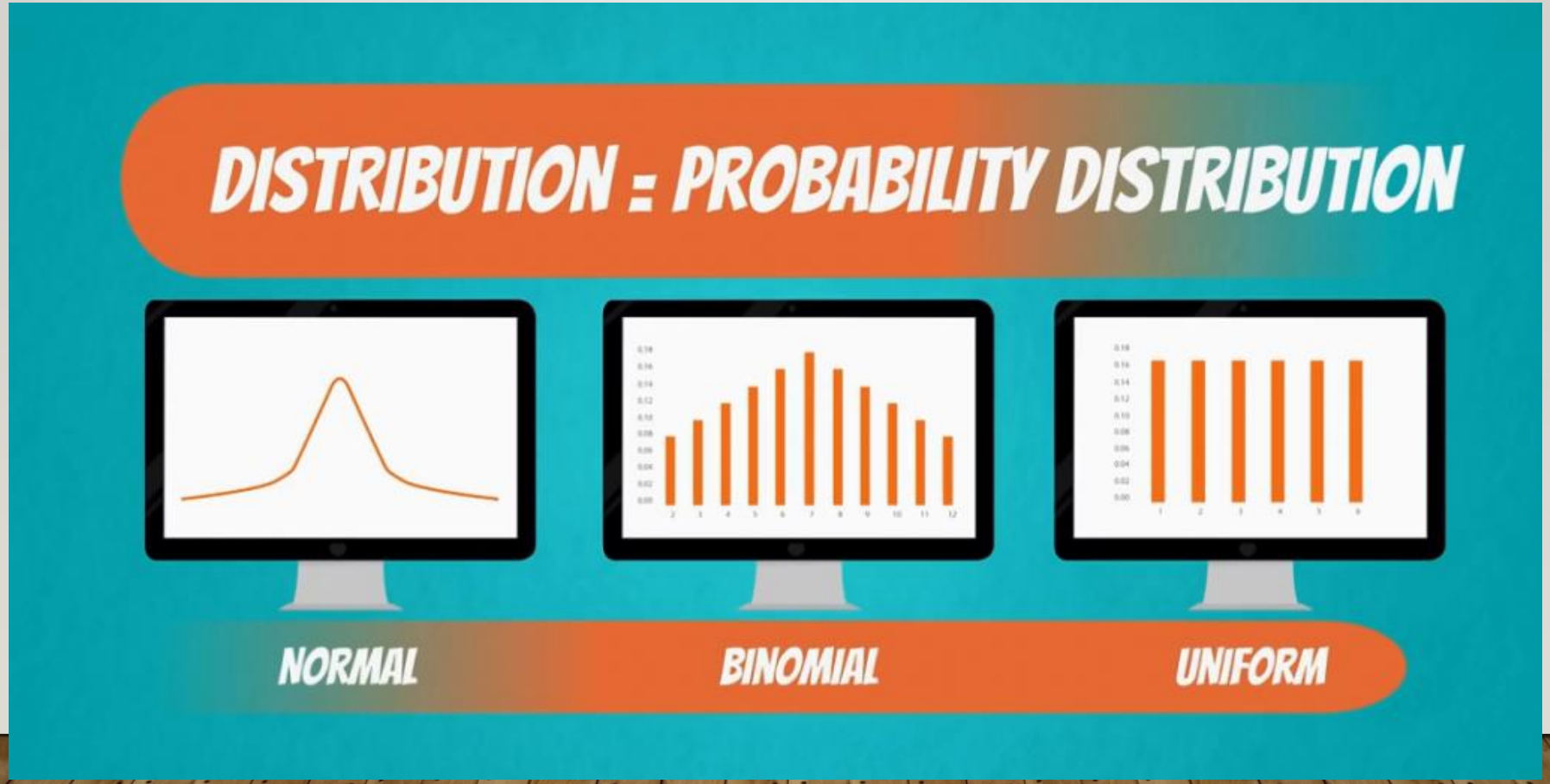
STATISTICS

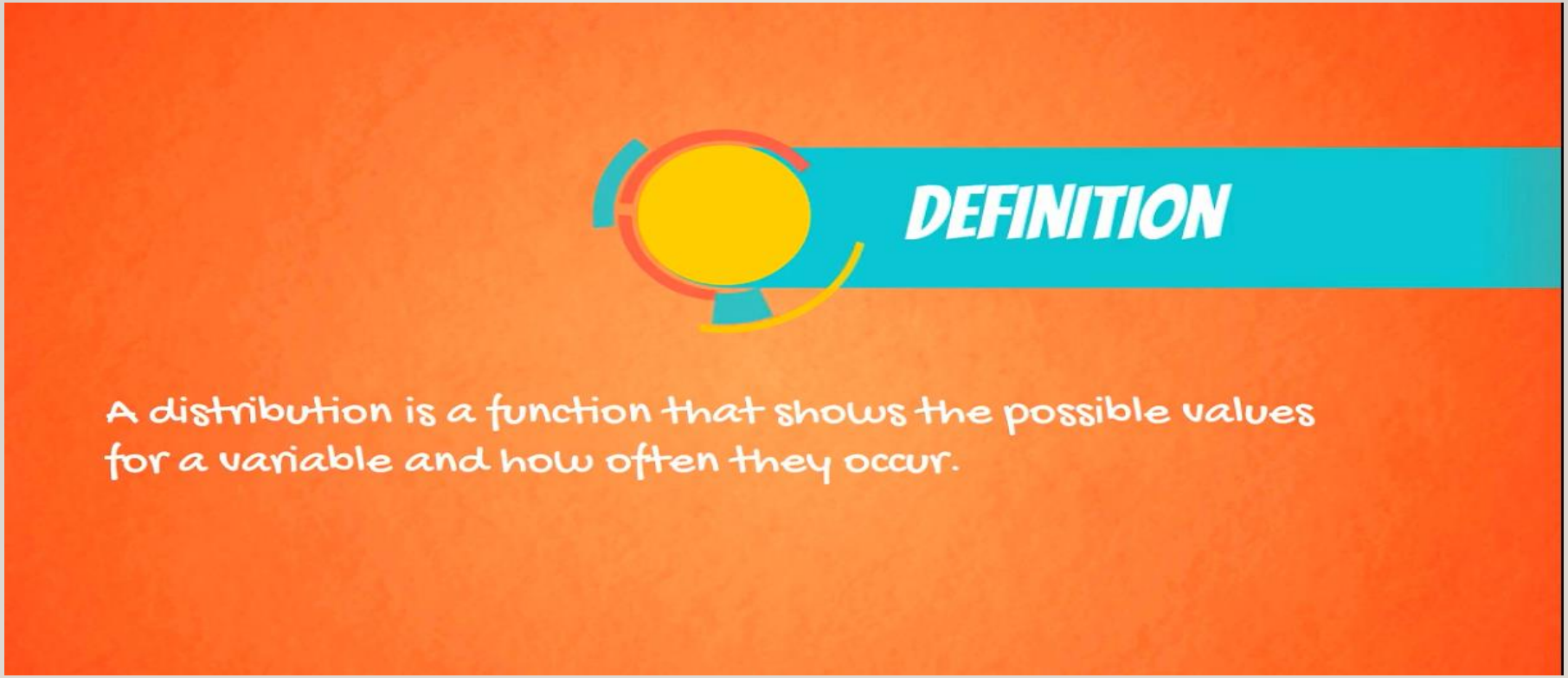
Session 2			
8	The Normal Distribution	PPT Jupyter	
9	Central limit theorem	PPT Jupyter	
10	Confidence Intervals	PPT Jupyter	
11	Statistical Hypothesis Testing	PPT Jupyter	

- **Descriptive vs. Inferential Statistics**
- **Descriptive Statistics**
- Descriptive statistics **is about describing our collected data**
- **Inferential Statistics**
- Inferential Statistics **is about using our collected data to draw conclusions to a larger population**.
- We looked at specific examples that allowed us to identify the
- **Population** - our entire group of interest.
- **Parameter** - numeric summary about a population
- **Sample** - subset of the population
- **Statistic** numeric summary about a sample

Remember that all **parameters** pertain to a population, while all **statistics** pertain to a sample.

Parameter	Statistic	Description
μ	$\bar{x}$	"The mean of a dataset"
π	p	"The mean of a dataset with only 0 and 1 values - a proportion"
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	"The difference in means"
$\pi_1 - \pi_2$	$p_1 - p_2$	"The difference in proportions"
β	b	"A regression coefficient - frequently used with subscripts"
σ	s	"The standard deviation"
σ^2	s^2	"The variance"
ρ	r	"The correlation coefficient"


DISTRIBUTION = PROBABILITY DISTRIBUTION
NORMAL
BINOMIAL
UNIFORM


DEFINITION
A distribution is a function that shows the possible values for a variable and how often they occur.

ROLLING A DIE

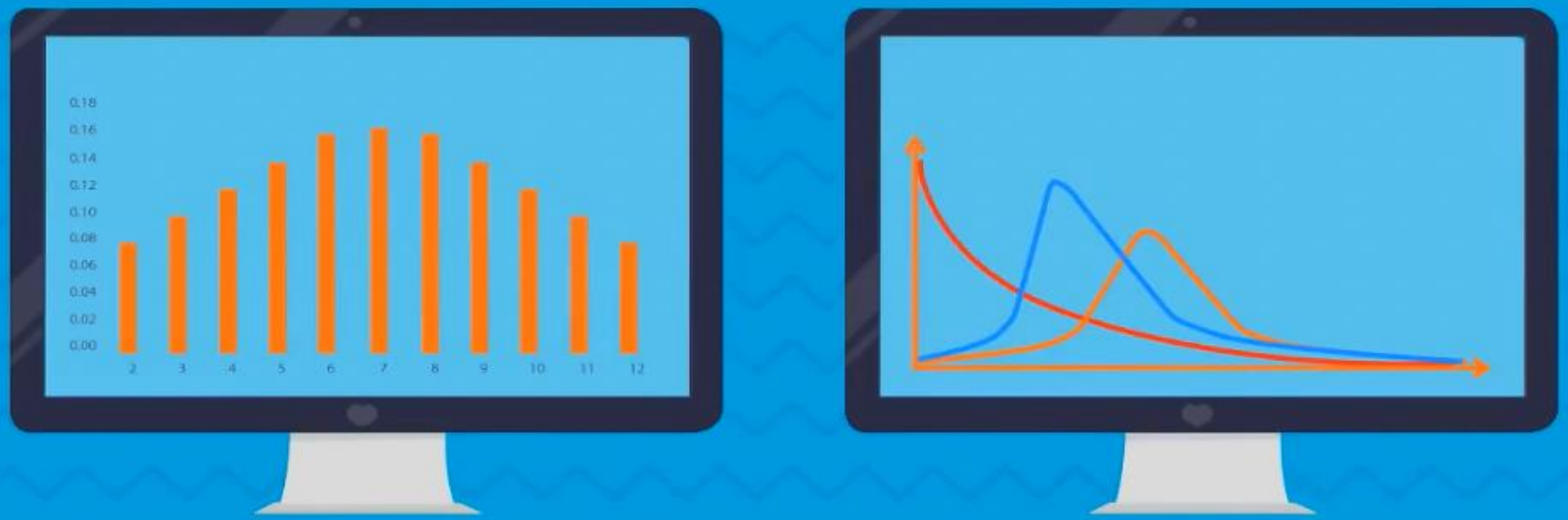
OUTCOME	PROBABILITY
1; 2; 3; 4; 5; 6	0.17
All else	0

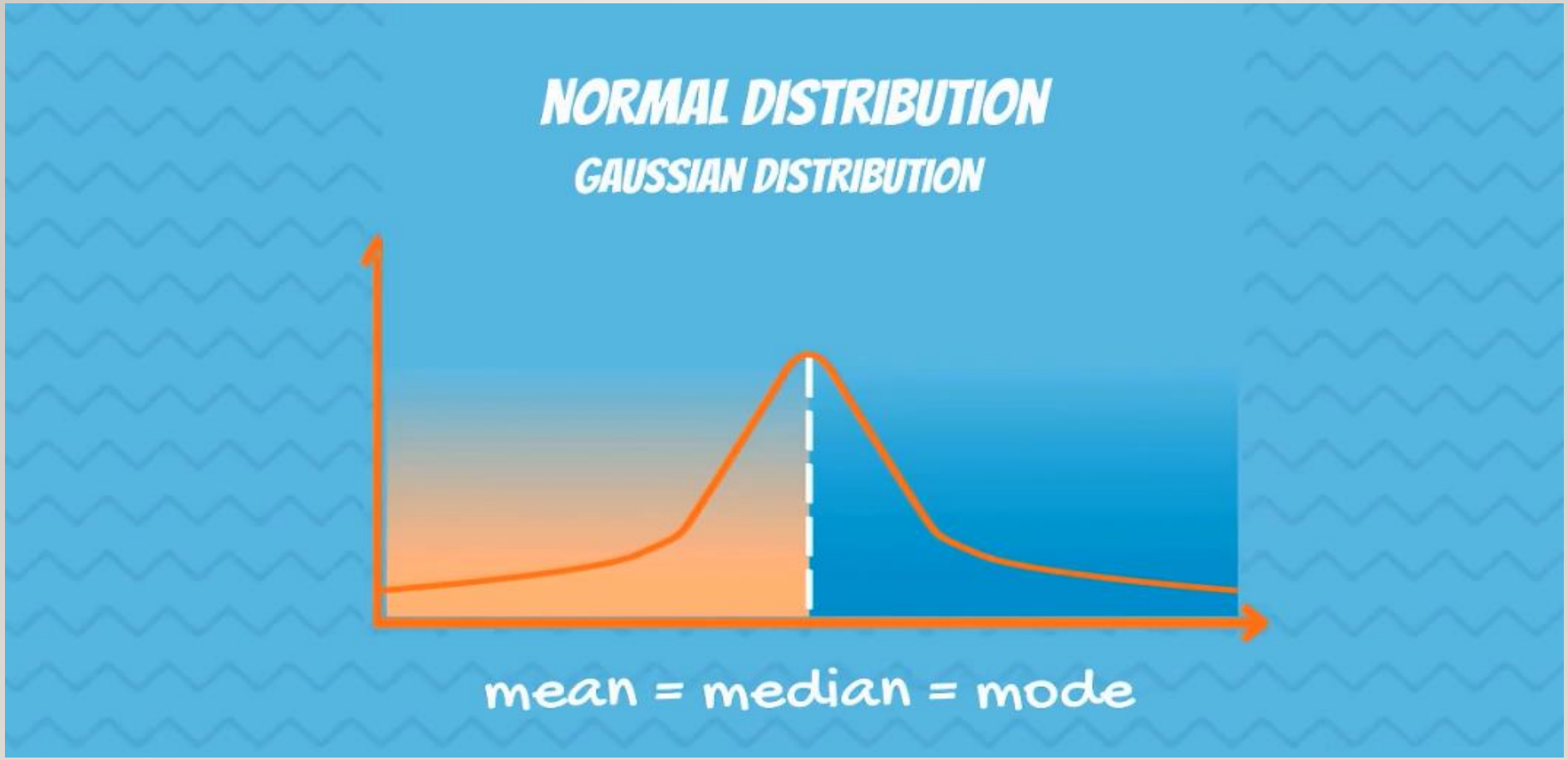
DISCRETE UNIFORM DISTRIBUTION

ROLLING TWO DICE

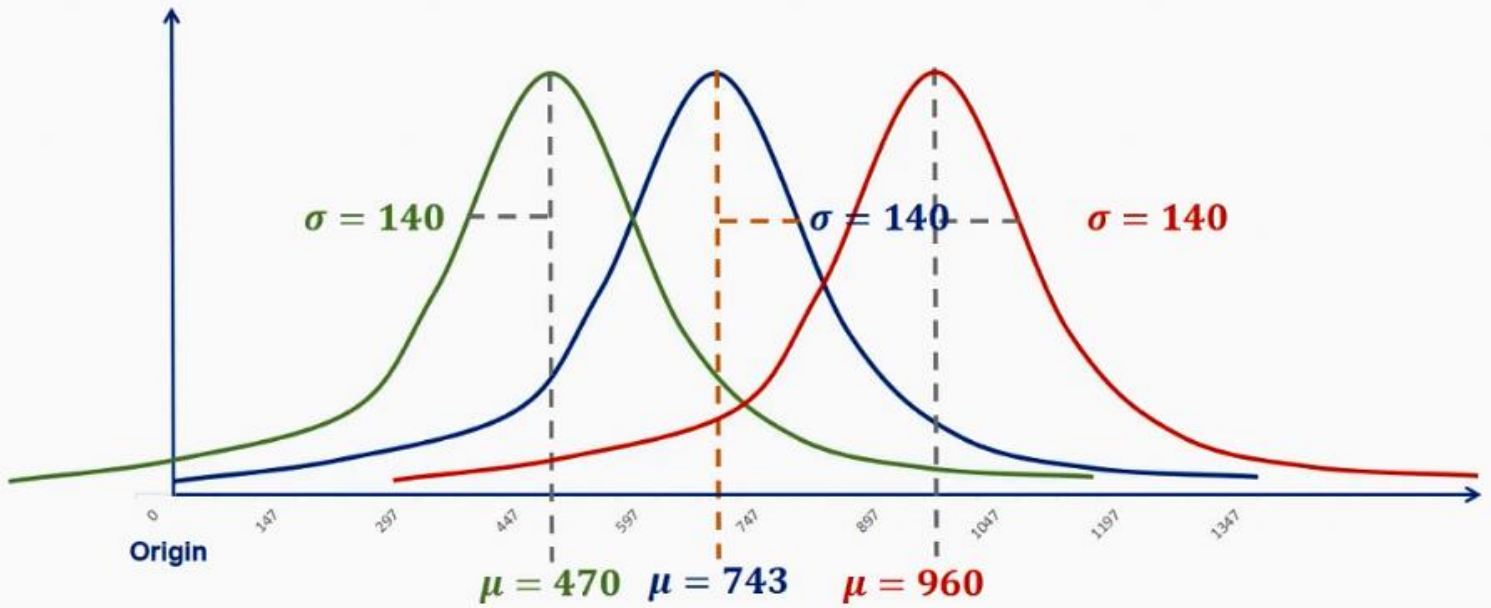
OUTCOME	PROBABILITY
2	0.03
3	0.06
4	0.08
5	0.11
6	0.14
7	0.17
8	0.14
9	0.11
10	0.08
11	0.06
12	0.03
All else	0

THE DISTRIBUTION OF A DATASET

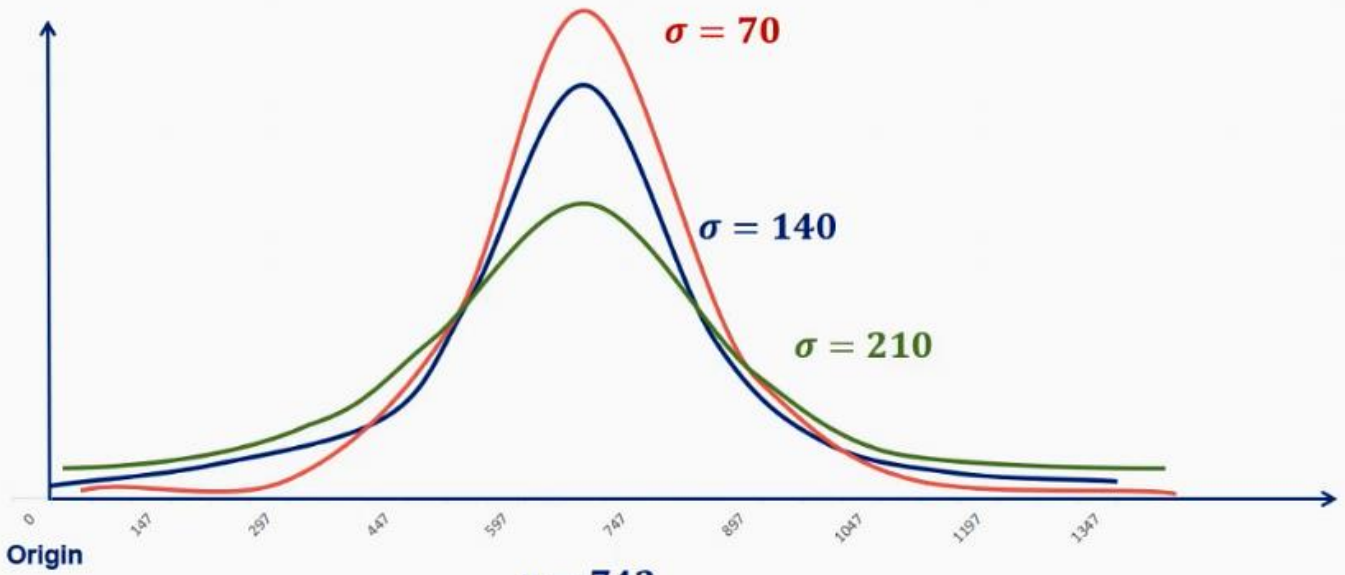



NORMAL DISTRIBUTION
GAUSSIAN DISTRIBUTION
mean = median = mode

Normal distribution. Controlling for standard deviation



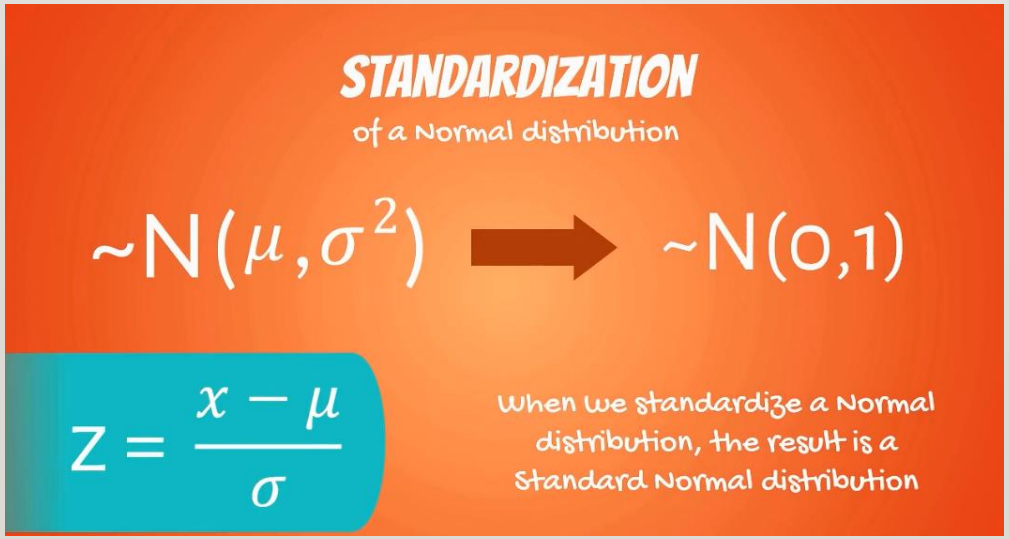
Normal distribution. Controlling for the mean



$\mu = 743$

$\mu = 743$

$\mu = 743$


STANDARDIZATION
of a Normal distribution
~N(μ,σ²)
~N(0,1)
Z = (x − μ) / σ
When we standardize a Normal distribution, the result is a Standard Normal distribution

Standard normal distribution

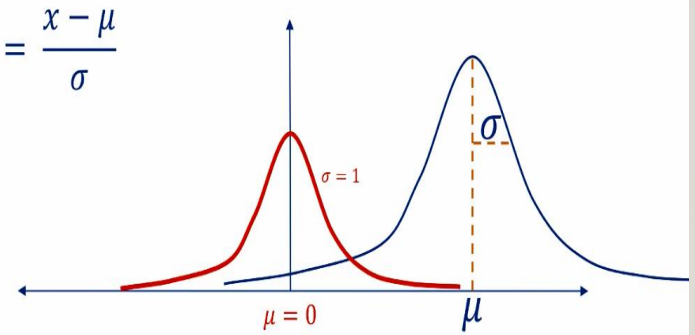
Standardization

Original dataset		Subtract mean		Divide by std	
1		-2		-1.63	
2	Mean 3	-1	Mean 0	-0.82	Mean 0.00
2	St. dev 1.22	-1	St. dev 1.22	-0.82	St. dev 1.00
3		0		0.00	
3	N~(3,1.49)	0	N~(0,1.49)	0.00	N~(0,1)
3		0		0.00	
4		1		0.82	
4		1		0.82	
5		2		1.63	

$$x \qquad x - \mu \qquad \frac{x - \mu}{\sigma}$$

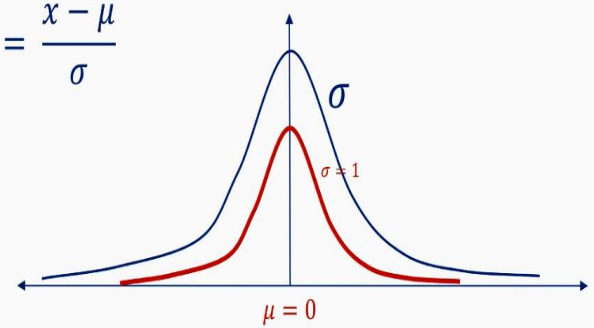
STANDARDIZATION

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



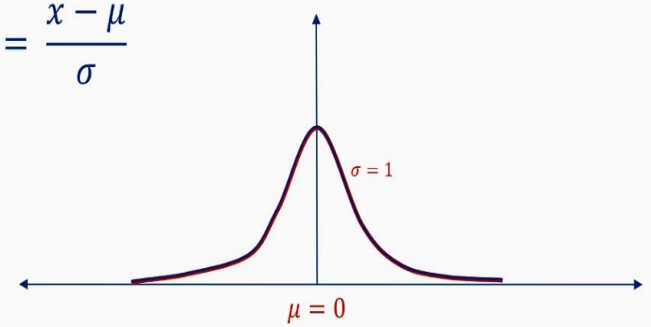
STANDARDIZATION

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



STANDARDIZATION

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



A **sampling distribution** is the distribution of a statistic. Here we looked the distribution of the proportion for samples of 5 students. This is key to the ideas covered not only in this lesson, but in future lessons.

JUPYTER

SAMPLING DISTRIBUTIONS

Sampling Distributions Notes

We have already learned some really valuable ideas about sampling distributions:

First, we have defined sampling distributions as the distribution of a statistic.

This is fundamental

- I cannot stress the importance of this idea. We simulated the creation of sampling distributions in the previous ipython notebook for samples of size 5 and size 20, which is something you will do more than once in the upcoming concepts and lessons.

- Second, we found out some interesting ideas about sampling distributions that will be iterated later in this lesson as well.

We found that for proportions (and also means, as proportions are just the mean of 1 and 0 values), the following characteristics hold.

1-The sampling distribution is centered on the original parameter value.

2-The sampling distribution decreases its variance depending on the sample size used. Specifically, the variance of the sampling distribution is equal to the variance of the original data divided by the sample size used. This is always true for the variance of a sample mean!

JUPYTER

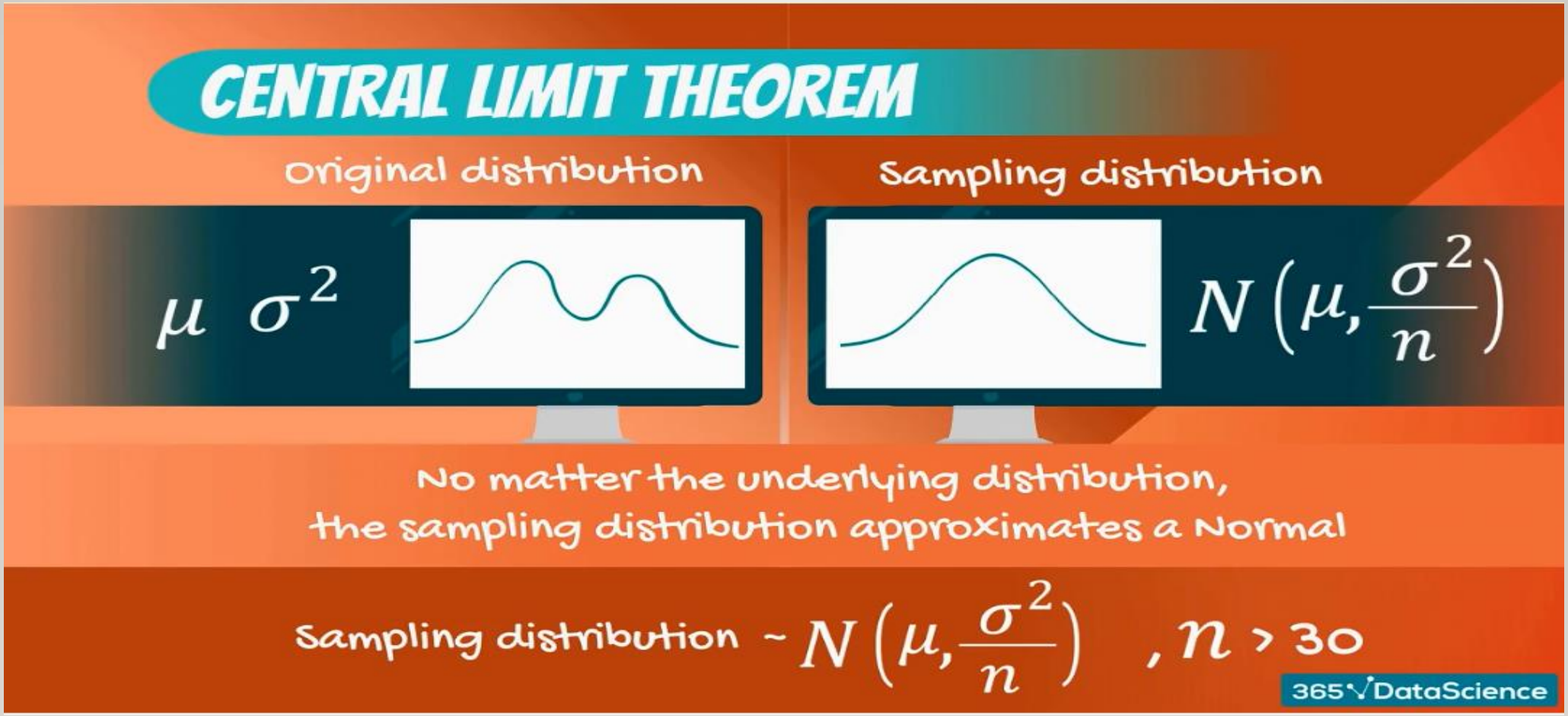
WHAT NOW

Two important mathematical theorems for working with sampling distributions

LAW OF LARGE NUMBERS

OUR SAMPLE SIZE INCREASES, THE SAMPLE MEAN GETS CLOSER TO THE POPULATION MEAN.

JUPYTER
LOW OF LARGE NUMBERS



it applies for additional statistics, but it doesn't apply for all statistics! .

JUPYTER
SAMPLING DISTRIBUTIONS
-CENTRAL LIMIT THEOREM -MEAN

JUPYTER
SAMPLING DISTRIBUTION
CENTRAL LIMIT THEOREM - VARIANCE

CONFIDENCE INTERVALS


Confidence Interval Lower Bound
Confidence Interval Upper Bound
99%
.05%
.05%

$\bar{x}$ p S b


Confidence Interval Lower Bound
Confidence Interval Upper Bound
95%
2.5%
2.5%

$\bar{x}$ p s b

JUPYTER
CONFIDENCE INTERVALS