Transforming Data

$\bar{X} = \frac{\sum x_i}{N}$

Look at below Question.

$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$

→ Sample data

1. Below are the weights of 5 persons. Calculate Mean, Standard Deviation :

105, 156, 145, 172, 100

$\boxed{\bar{X} = 135.6}$, $\boxed{\sigma = 31.75}$

2. Suppose each one of them gained extra 5 Kg. weight during winters. Can you calculate the new Mean and Standard deviation?

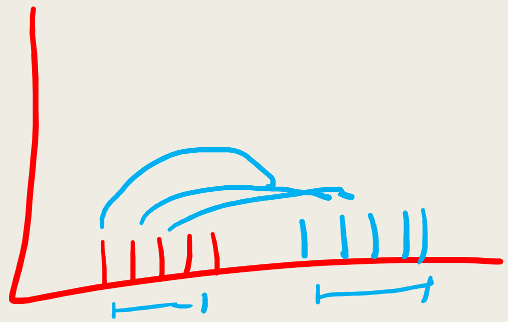
New data → 110, 161, 150, 177, 105

New $\bar{X}$ → 140.6

New σ → 31.75

Comparison Old & New - $\bar{X}$ changed , σ remain same

↓ 5 units



Same spread

→ Additive term

(+) (−) in our data

then only central Tendency will be affected

No change in spread

Transforming Data

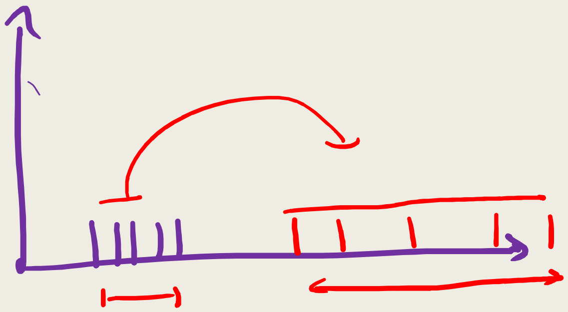
100 kg * 2.5ml = 250 ml + 750 ml = 1000 ml

Example 2:

1. Considering the same set of people from previous example, Suppose that these persons are advised to drink 2.5 ml of water for every Kg they weigh plus 750 ml of water everyday.

105, 156, 145, 172, 100

What is the mean and STD for the amount of water consumed everyday?



ml of water => (2.5 * weight) + 750

1012.5, 1140, 1112.5, 1180, 1000

$\bar{X}_{ml} \rightarrow 1089 \rightarrow 135.6 * 2.5 + 750$

$\sigma_{ml} \rightarrow 79.38 \rightarrow 31.75 * 2.5$

Central Tendency → Additive | Multiplicative term

+ − * /

Spread → Multiplicative term

TRANSFORMING DATA

GUIDELINES

$\log(x)$

x^2

$x-1$

$\frac{1}{x}$

20

400

MEASURES OF CENTRE

AFFECTED BY:

+ − × ÷

MODE, MEDIAN, MEAN

MEASURES OF SPREAD

AFFECTED BY:

× ÷

RANGE, STANDARD DEVIATION

SUPPOSE THAT IN ORDER TO STAY HYDRATED, THESE STUDENTS DRINK 2.5mL OF WATER FOR EVERY POUND THEY WEIGH; PLUS 750mL OF WATER A DAY. WHAT IS THE MEAN AND STANDARD DEVIATION FOR THE AMOUNT OF WATER CONSUMED EVERY DAY?

105

156

145

172

100

$\bar{x} = 135.6$ $\times$ 2.5 $+$ 750

$\bar{x}_{NEW} = (135.6)(2.5) + 750 = 1089$

$s = 31.75$ $\times$ 2.5

$s_{NEW} = (31.75)(2.5) = 79.38$

σ^2 = Variance $(31.75)^2 \rightarrow$ Variance

Multiplicative term   Additive term

1089 = (135.6) * 2.5 + 750

MEASURES OF CENTRE

$CENTRE_{NEW} = (CENTRE_{OLD})(X) + B$

79.38   31.75 * 2.5

MEASURES OF SPREAD

$SPREAD_{NEW} = (SPREAD_{OLD})(X)$

Multiplicative term

Bank loan

A

B

1 cr

50 Lakh

10000

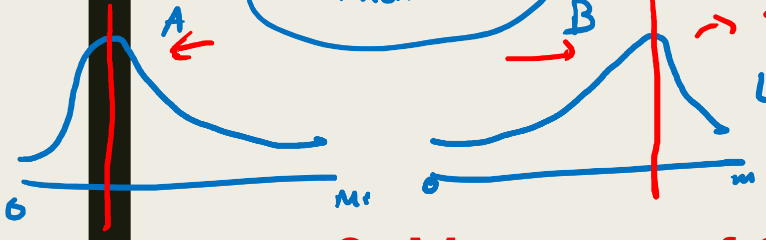
80000

$100,000,000,000

Outstanding Amount

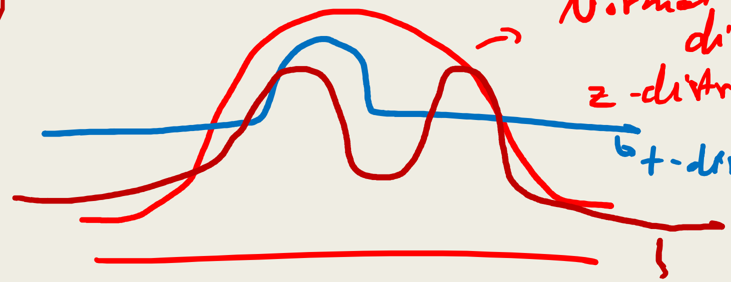
Symmetric vs Asymmetry

SHAPE is Same



→ Skewed data

L ← from the center → R



Normal distribution

z-distribution

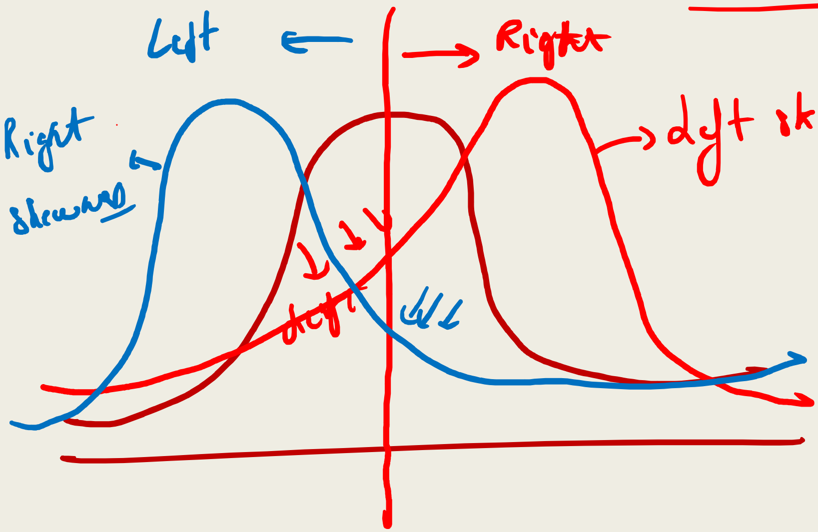
t-distribution

Bimodal distribution

Accumulation of values before / after the Center

3. Measure of Symmetricity & Shape – Skewness and Kurtosis

Left ← → Right



z-dist → Normal Distribution.

t-dist.

hypothesis testing

bell curve

→ Normal dist

measure of symmetricity

1. Skewness

Normal distn

+ve → Left skewness
-ve → Right skewness

Skewness is usually described as a measure of a dataset's symmetry – or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0. The normal distribution has a skewness of 0. Skewness is calculated as:

```
import numpy as np
from scipy.stats import skew
x = np.random.normal(0, 2, 10000) # create random values based on a normal distribution
print(skew(x))
```

→ [-0.5, +0.5] $\boxed{0}$

....

Mathematically:

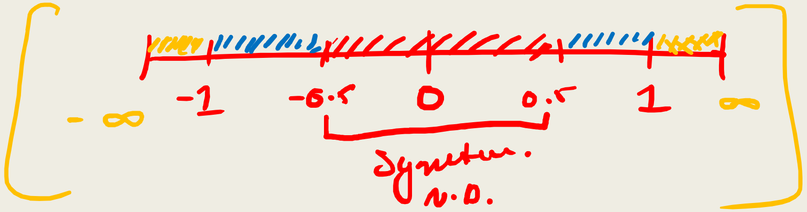
$$\frac{\sum x_i^2}{N} \qquad (x_i - \bar{x})^2$$

$$a_3 = \sum \frac{(X_i - \bar{X})^3}{ns^3}$$

→ you will never be using the formula

where n is the sample size, X_i is the i^{th} X value, X-Bar is the average and s is the sample standard deviation. Note the exponent in the summation. It is "3". The skewness is referred to as the "third standardized central moment for the probability model."

Skewness



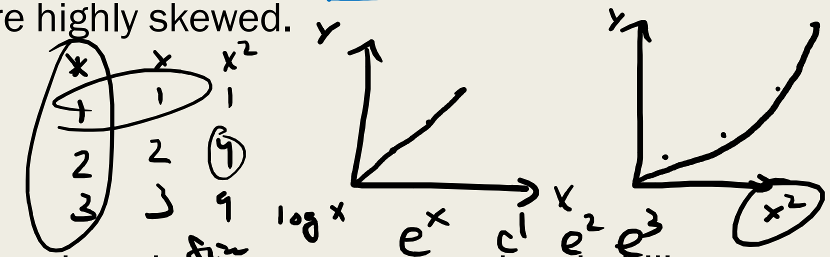
So, when is the skewness too much? The rule of thumb seems to be:

1. If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
2. If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.
3. If the skewness is less than -1 or greater than 1, the data are highly skewed.

Importance of Skewness:

ILATE

$\frac{1}{x}$

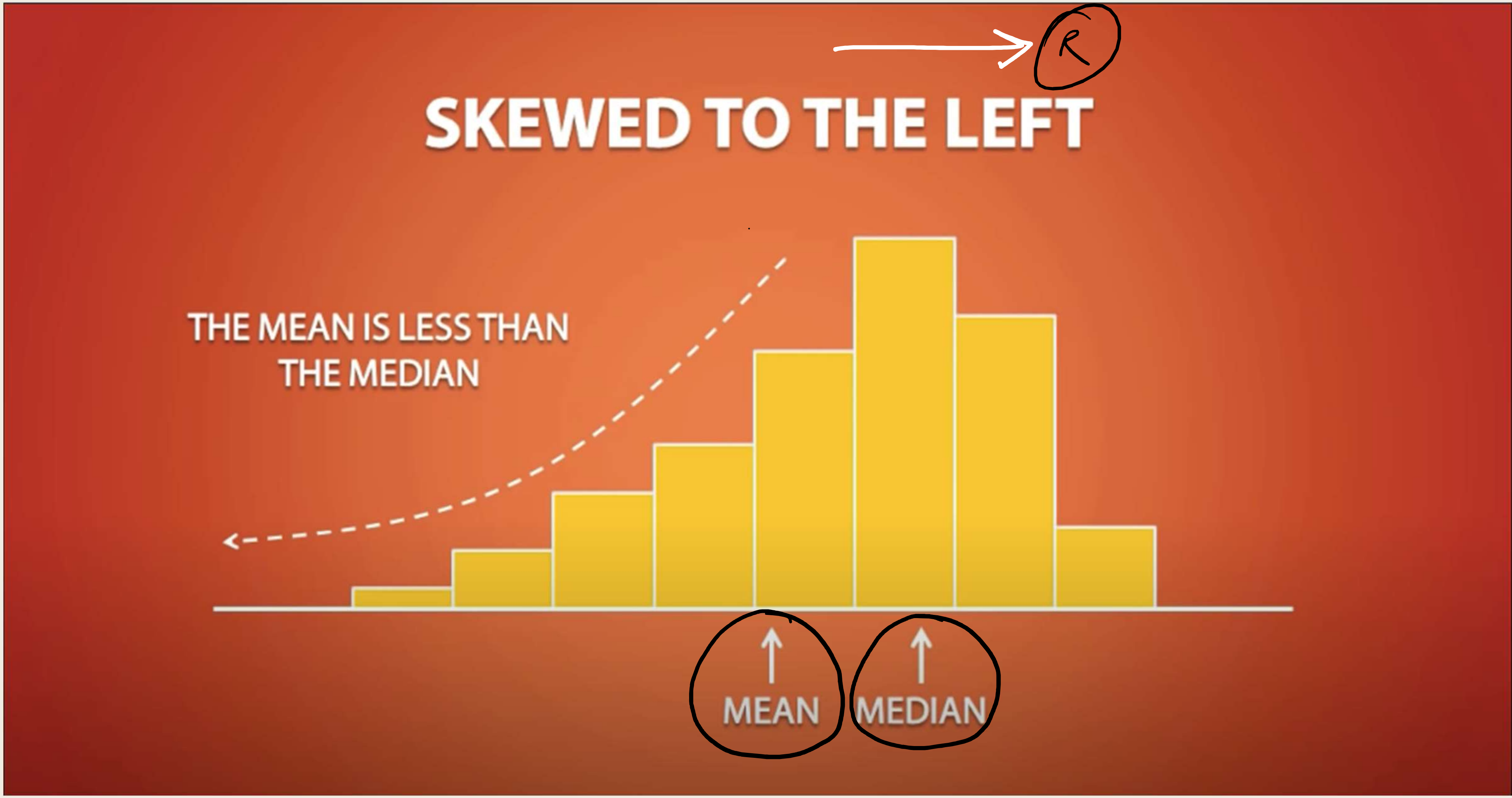


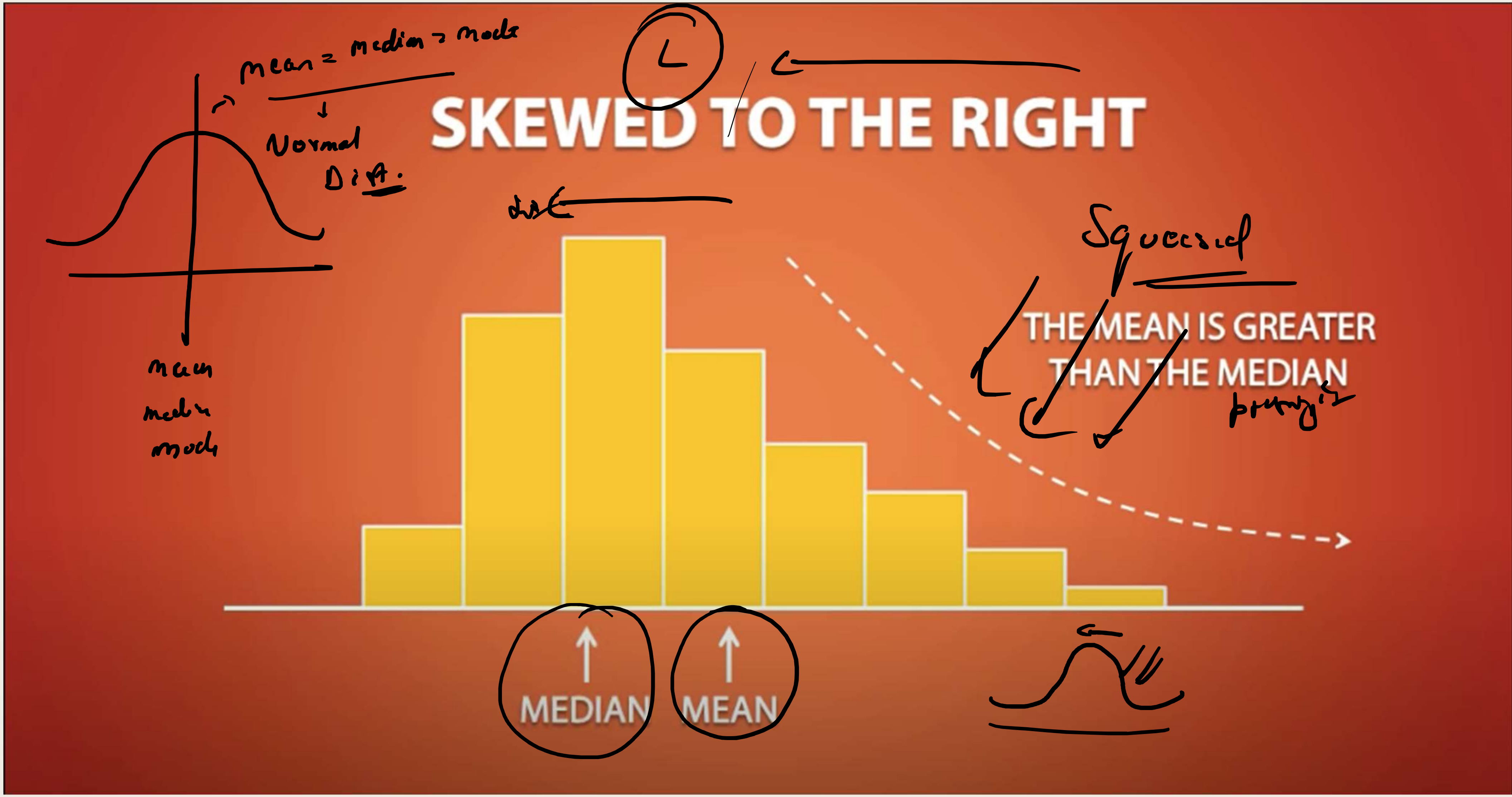
Measures of asymmetry like skewness are the link between central tendency measures and probability theory, which ultimately allows us to get a more complete understanding of the data we are working with.

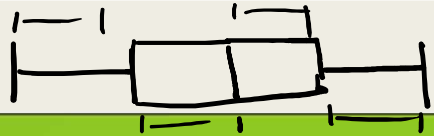
Knowing that the market has a 70% probability of going up and a 30% probability of going down may appear helpful if you rely on normal distributions. However, if you were told that if the market goes up, it will go up 2% and if it goes down, it will go down 10%, then you could see the skewed returns and make a better informed decision.

$$E(r) = 0.7*0.02 + 0.3*\text{-}0.1 = \text{-}0.014$$


SKEWNESS
REFERS TO ASYMMETRY
SKEWED TO THE LEFT
SKEWED TO THE RIGHT
R
L


R
SKEWED TO THE LEFT
THE MEAN IS LESS THAN THE MEDIAN
MEAN
MEDIAN


SKEWED TO THE RIGHT
THE MEAN IS GREATER THAN THE MEDIAN
MEDIAN
MEAN
mean = median = mode
Normal Dist.
mean
median
mode
Squeezed

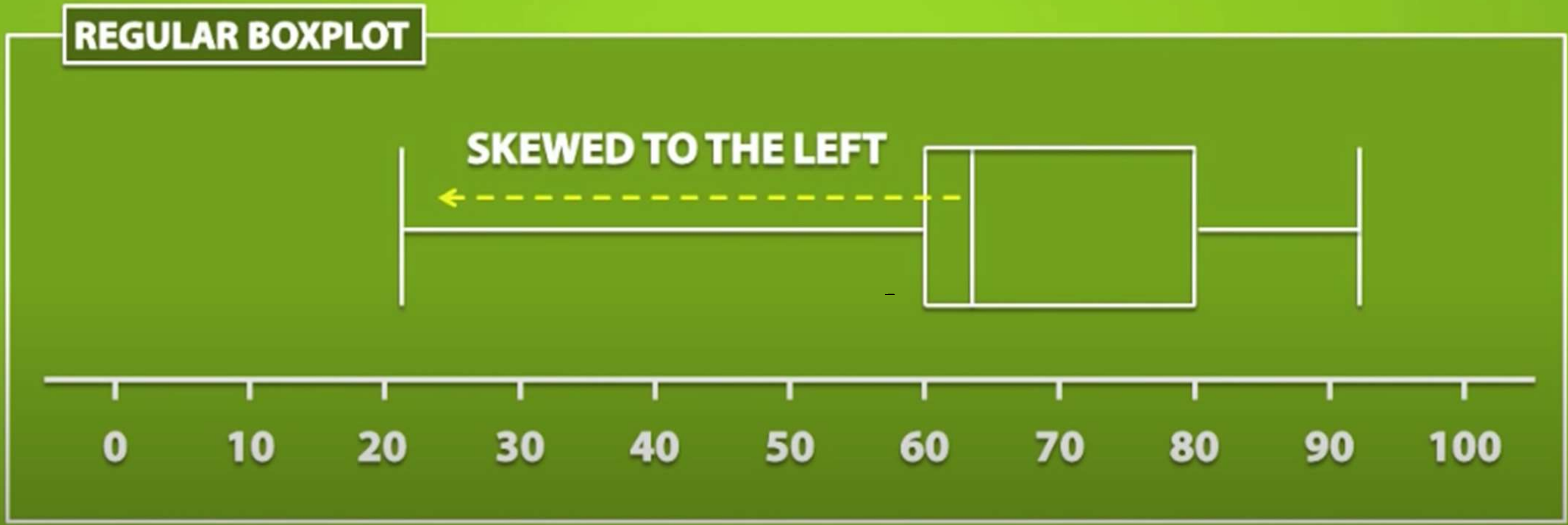


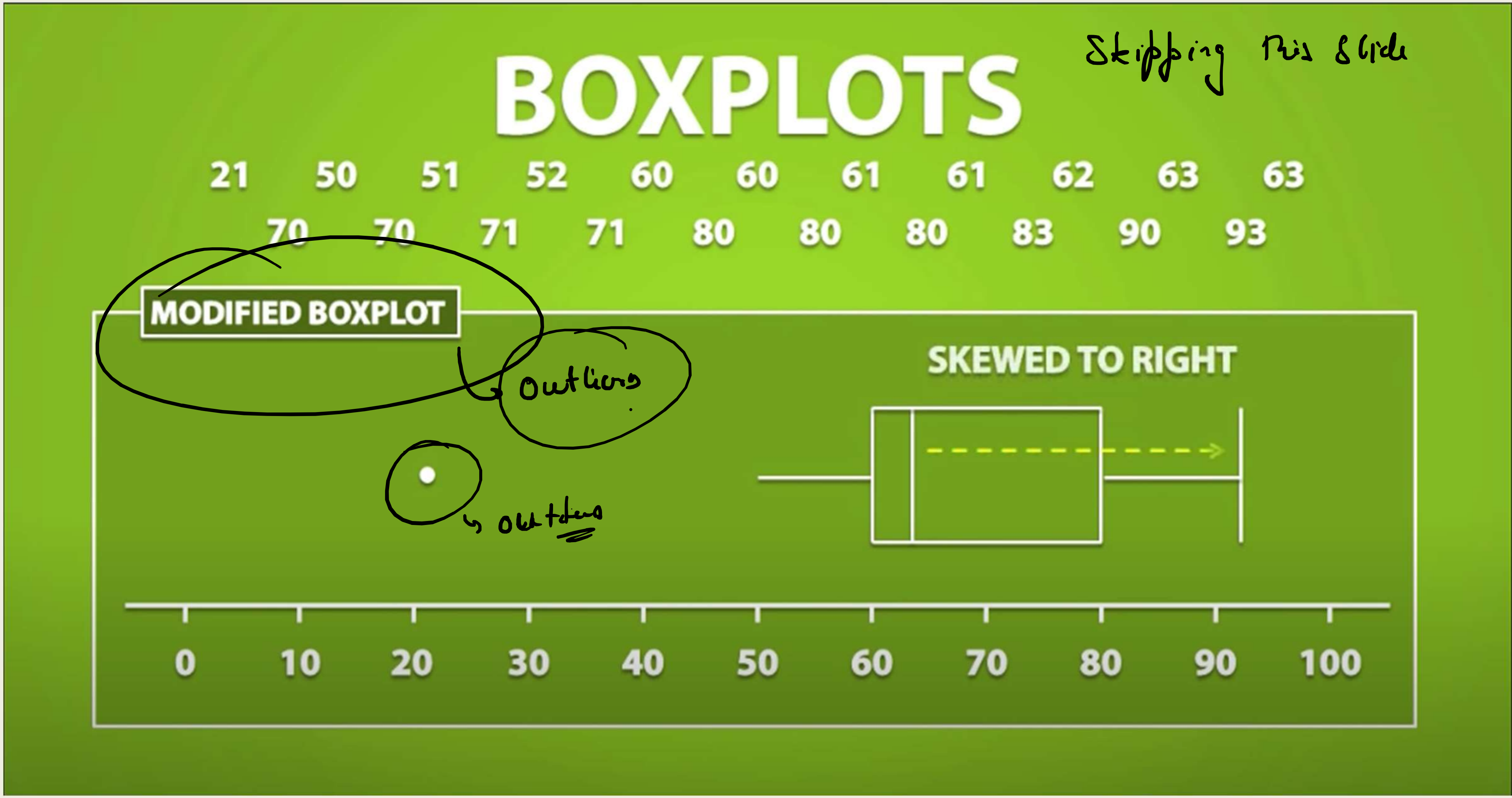
→ N.D.

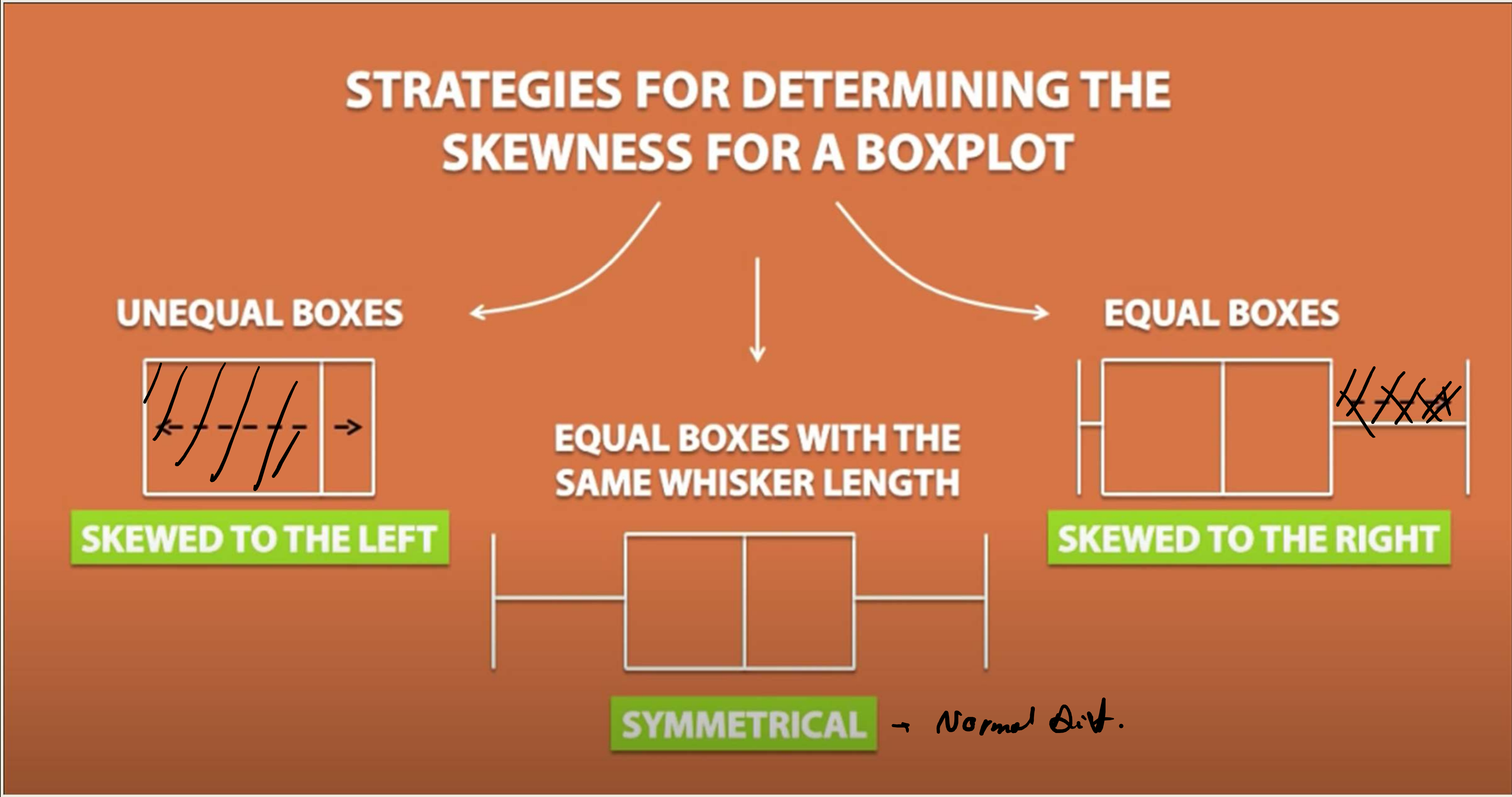
BOXPLOTS

21 50 51 52 60 60 61 61 62 63 63

70 70 71 71 80 80 80 83 90 93



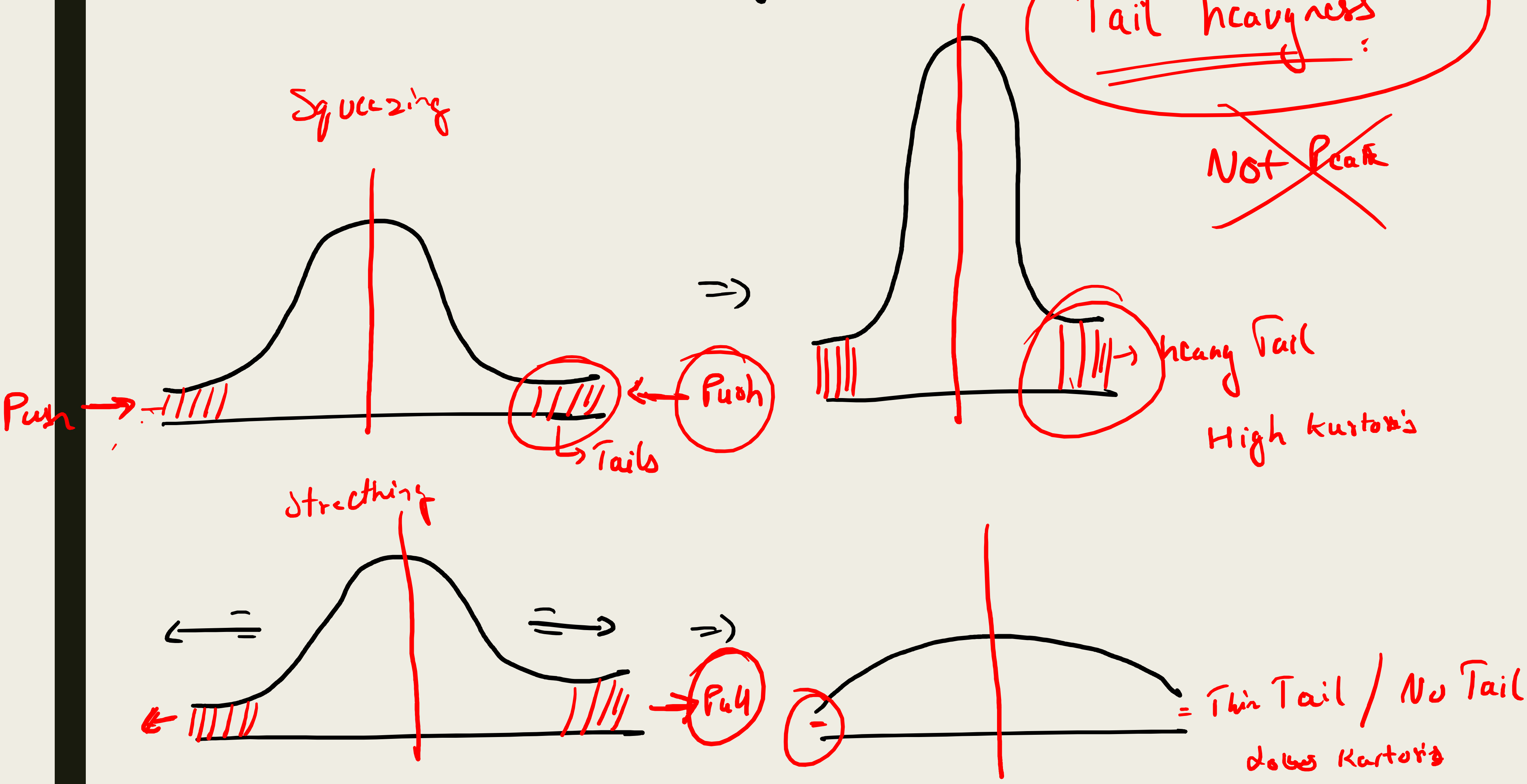

BOXPLOTS
Skipping this slide
21 50 51 52 60 60 61 61 62 63 63
70 70 71 71 80 80 80 83 90 93
MODIFIED BOXPLOT
Outliers
outliers
SKEWED TO RIGHT
0 10 20 30 40 50 60 70 80 90 100


STRATEGIES FOR DETERMINING THE SKEWNESS FOR A BOXPLOT
UNEQUAL BOXES
SKEWED TO THE LEFT
EQUAL BOXES WITH THE SAME WHISKER LENGTH
SYMMETRICAL
→ Normal Dist.
EQUAL BOXES
SKEWED TO THE RIGHT

Kurtosis = Shape Abnormality

Tail heaviness :

Not Peak



2. Kurtosis

N.D. kurtosis = 3

0

3

Kurtosis is all about the tails of the distribution – not the peakness or flatness. It measures the **tail-heaviness** of the distribution. Kurtosis is calculated as:

3

2.2

5.6

import numpy as np
from scipy.stats import kurtosis
x = np.random.normal(0, 2, 10000) # create random values based on a normal distribution
print(kurtosis(x))

Excess Kurtosis = (Kurtosis) − 3

N.D. (EK) = 0

Mathematically:

$$a_4 = \sum \frac{(X_i - \bar{X})^4}{ns^4}$$

> 0

< 0

where n is the sample size, X_i is the i^{th} X value, X-Bar is the average and s is the sample standard deviation. Note the exponent in the summation. It is “4”. The kurtosis is referred to as the “fourth standardized central moment for the probability model.”

Note: Kurtosis calculated by Excel or through Python/R is actually excess kurtosis, which is (Kurtosis – 3)

The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, often the **excess kurtosis** is presented: excess kurtosis is simply **kurtosis−3**. For example, the "kurtosis" reported by Excel or any statistical library is actually the excess kurtosis.

1. A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈3 (excess ≈0) is called **mesokurtic**.

2. A distribution with kurtosis <3 (excess kurtosis <0) is called **platykurtic**. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.

3. A distribution with kurtosis >3 (excess kurtosis >0) is called **leptokurtic**. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.