MIT-WPU

॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥

Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Data Science

By
Shilpa Sonawani

SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY

Basic statistics

- Statistics: A bunch of mathematics used to summarize, analyze, and interpret a group of numbers or observations.

  *It is a tool.

  *Cannot replace your research design, your research questions, and theory or model you want to use.

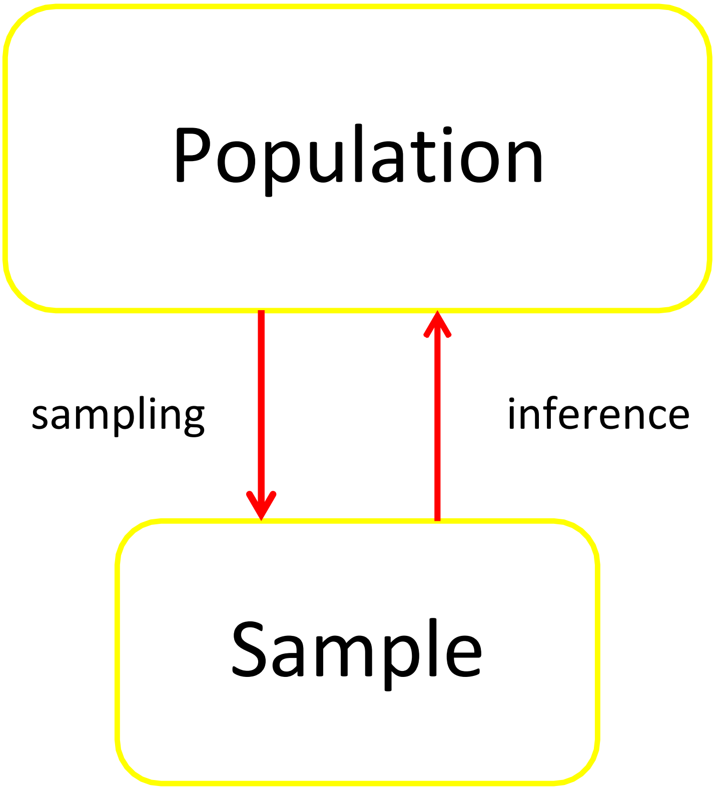
Population and sample

- Population: any group of interest or any group that researchers want to learn more about.
  - Population parameters (unknown to us): characteristics of population
- Sample: a group of individuals or data are drawn from population of interest.
  - Sample statistics: characteristics of sample

Population and sample

- We are much more interested in the population from which the sample was drawn.
  - Example: 30 GPAs as a representative sample drawn from the population of GPAs of the freshmen currently in attendance at a certain university or the population of freshmen attending colleges similar to a certain university.

Population and sample



Primary & Secondary Data

- **Raw or Primary data:** when data collected having lot of unnecessary, irrelevant & un wanted information

- **Treated or Secondary data:** when we treat & remove this unnecessary, irrelevant & un wanted information

- **Cooked data:** when data collected not genuinely and is false and fictitious

Ungrouped & Grouped Data

Ungrouped data: when data presented or observed individually.

For example if we observed no. of children in 6 families

2, 4, 6, 4, 6, 4

Grouped data: when we grouped the identical data by frequency.

For example above data of children in 6 families can be grouped as:

No. of children	**Families**
2	**1**
4	**3**
6	**2**

or alternatively we can make classes:

No. of children	**Frequency**
2 - 4	**4**
5 - 7	**2**

Variable

A variable is something that can be changed, such as a characteristic or value. For example age, height, weight, blood pressure etc

Types of Variable

Independent variable: is typically the variable representing the value being manipulated or changed.

- The independent variable is the cause. Its value is independent of other variables in study.

Dependent variable: is the observed result of the independent variable being manipulated.

- The dependent variable is the effect. Its value depends on changes in the independent variable

Confounding variables: Are those that affect other variables in a way that produces spurious or distorted associations between two variables..

For example,

You collect data on sunburns and ice cream consumption. You find that higher ice cream consumption is associated with a higher probability of sunburn. Does that mean ice cream consumption causes sunburn? Here, the confounding variable is temperature: hot temperatures cause people to both eat more ice cream and spend more time outdoors under the sun, resulting in more sunburns.

Types of measurement

- Discrete: Quantitative data are called discrete if the sample space contains a finite or countably infinite number of values.
  - How many days did you go to Gym during the last 7 days

Types of measurement

- Continuous: Quantitative data are called continuous if the sample space contains an interval or continuous span of real numbers.
  - Weight, height, temperature
  - Height: 1.72 meters, 1.7233330 meters

Types of measurement

- Nominal
  - Categorical variables. Numbers that are simply used as identifiers or names represent a nominal scale of measurement such as female vs. male.

Types of measurement

- Ordinal
  - An ordinal scale of measurement represents an ordered series of relationships or rank order. Likert-type scales (such as "On a scale of 1 to 10, with one being no pain and ten being high pain, how much pain are you in today?") represent ordinal data.

Types of variable measurement scales

- Interval Scale:

- Don’t have a true zero

A true zero has no value - there is none of that thing. But 0 degrees C definitely has a value: it's quite chilly. You can also have negative numbers.

If you don't have a true zero, you can't calculate ratios. This means addition and subtraction work, but division and multiplication don't.

Types of variable measurement scales

- Ratio Scale
- Has all properties of interval-scaled
- Have a true zero

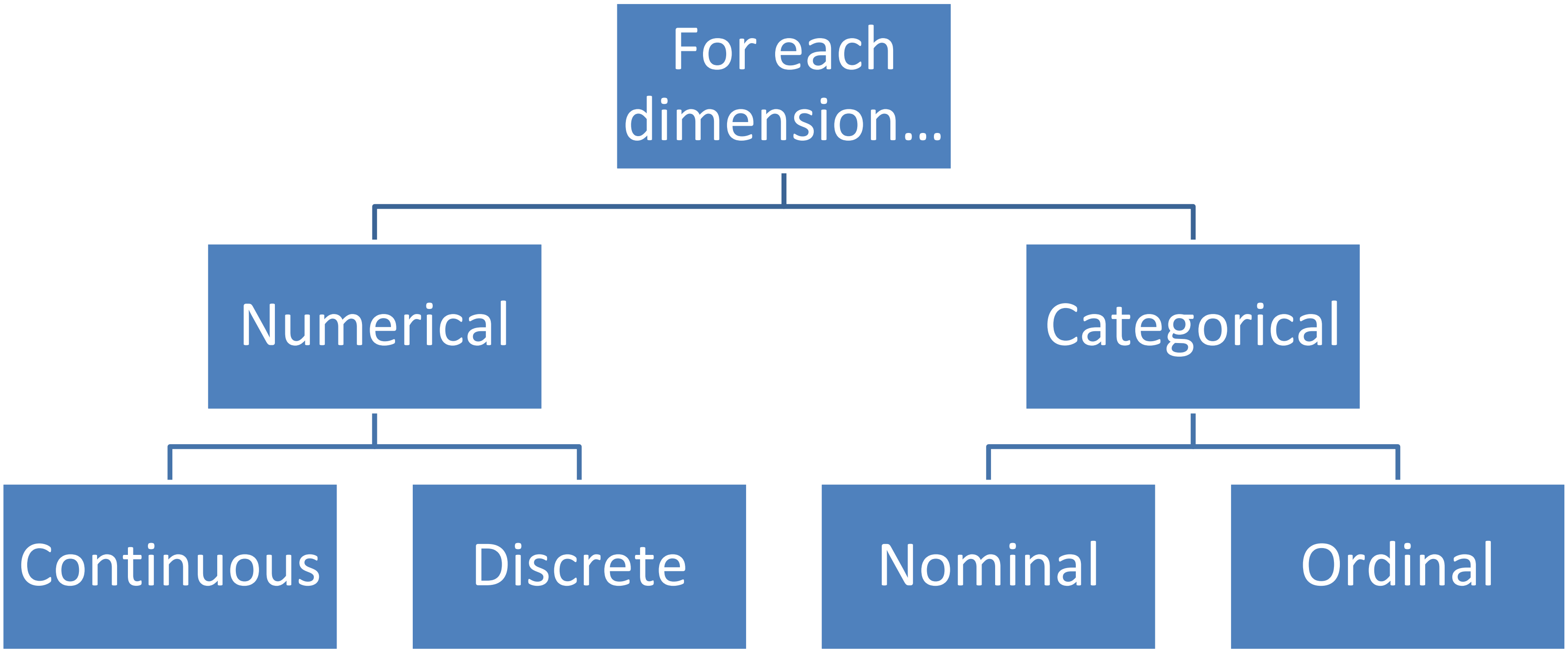
  A good example of ratio data is weight in kilograms. If something weighs zero kilograms, it truly weighs nothing—compared to temperature (interval data), where a value of zero degrees doesn't mean there is "no temperature," it simply means it's extremely cold!

- Values can be added, subtracted, multiplied & divided
- e.g. Weight, height, etc

Types of measurement

- Qualitative vs. Quantitative variables
  - Qualitative variables: values are texts (e.g., Female, male), we also call them string variables.
  - Quantitative variables: are numeric variables.

Data Types



For each dimension…

Numerical

Categorical

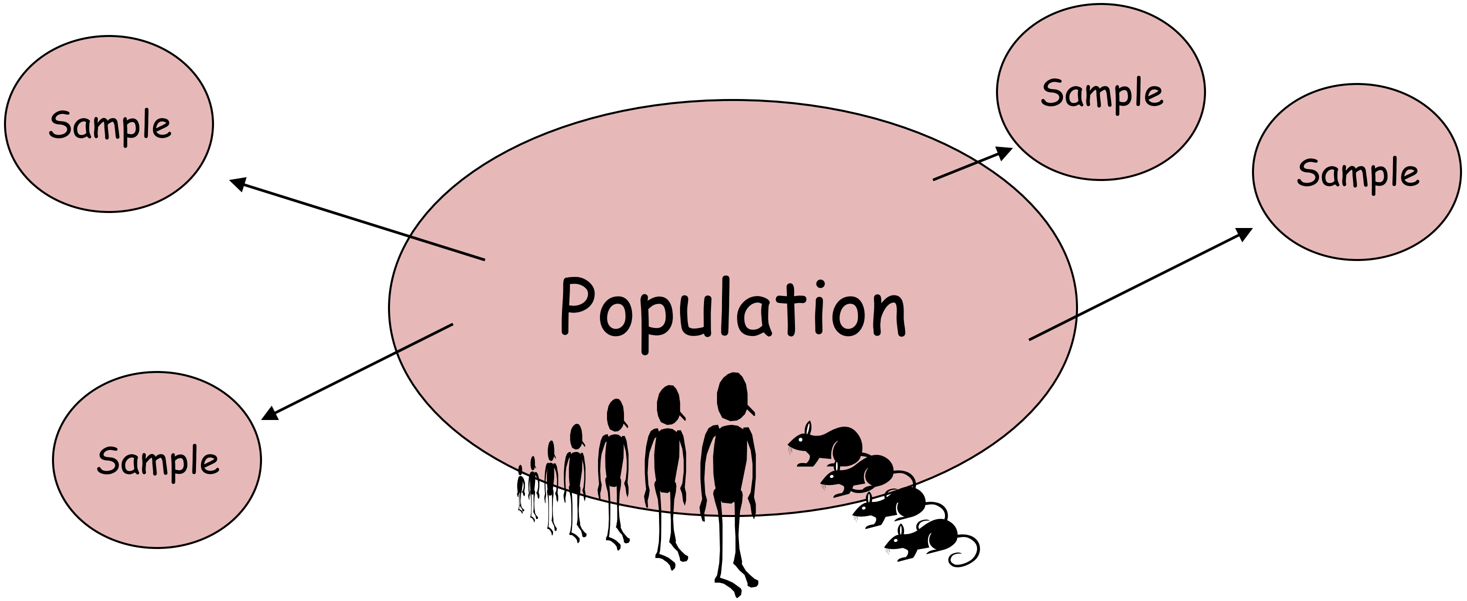
Continuous

Discrete

Nominal

Ordinal

What is a Statistic????



Parameter: value that describes a population

Statistic: a value that describes a sample
PSYCH → always using samples!!!

Basic Statistics

- Two types of statistics
  - Descriptive statistics
  - Inferential statistics

These two distinct approaches help us to make a sense of data and draw conclusions

Descriptive & Inferential Statistics

Aspect	**Descriptive Statistics**	**Inferential Statistics**
Purpose	Summarize and describe data	Draw conclusions or predictions
Data Sample	Analyzes the entire dataset	Analyzes a sample of the data
Examples	Mean, Median, Range, Variance	Hypothesis testing, Regression
Scope	Focuses on data characteristics	Makes inferences about populations
Goal	Provides insights and simplifies data	Generalizes findings to a larger population
Assumptions	No assumptions about populations	Requires assumptions about populations
Common Use Cases	Data visualization, data exploration	Scientific research, hypothesis testing

Descriptive Statistics

Key Aspects of Descriptive Statistics:

- **Measures of Central Tendency:** Descriptive statistics include calculating the mean, median, and mode, which offer insights into the center of the data distribution.

- **Measures of Dispersion:** Variance, standard deviation, and range help us understand the spread or variability of the data.

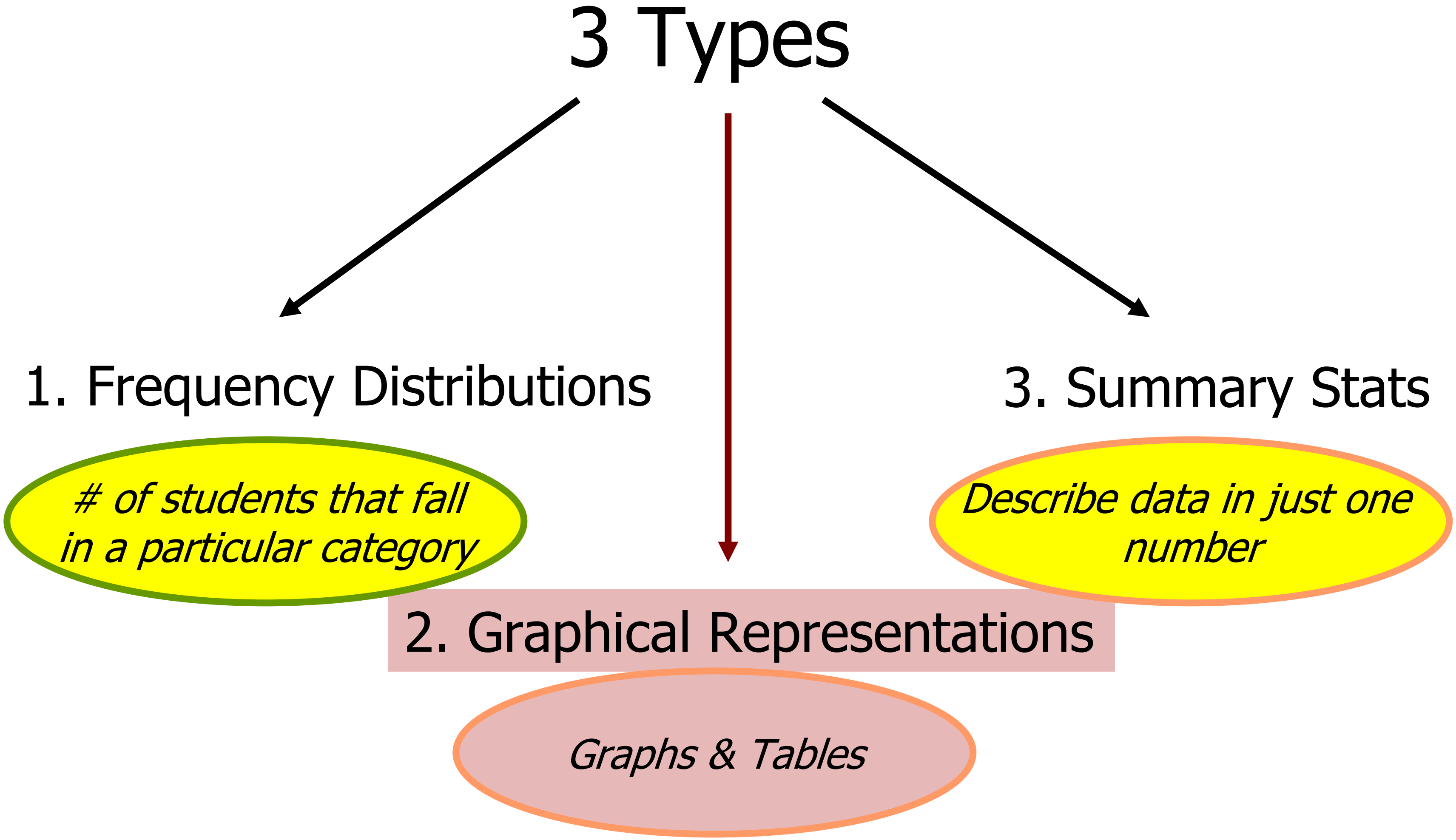
- **Visualizations:** Creating graphs, histograms, bar charts, and pie charts visually represent the data’s distribution and characteristics.

Inferential Statistics

Key Aspects of Inferential Statistics:

- **Sampling Techniques:** Relies on carefully selecting representative samples from a population to make valid inferences.
- **Hypothesis Testing:** This process involves setting up hypotheses about population characteristics and using sample data to determine if these hypotheses are statistically significant.
- **Confidence Intervals:** These provide a range of values within which we're confident a population parameter lies based on sample data.
- **Regression Analysis:** Inferential statistics also encompass techniques like regression analysis to model relationships between variables and predict outcomes.

Descriptive Statistics



3 Types

1. Frequency Distributions

# of students that fall in a particular category

2. Graphical Representations

Graphs & Tables

3. Summary Stats

Describe data in just one number

1. Frequency Distributions

# of students that fall in a particular category

How many males and how many females are in our class?

	Male	Female	total
Frequency	?	?	
(%)	**?/tot x 100**	**?/tot x 100**	
	-----%	**------%**	

1. Frequency Distributions

# of students that fall in a particular category

Categorize on the basis of more that one variable at same time

CROSS-TABULATION

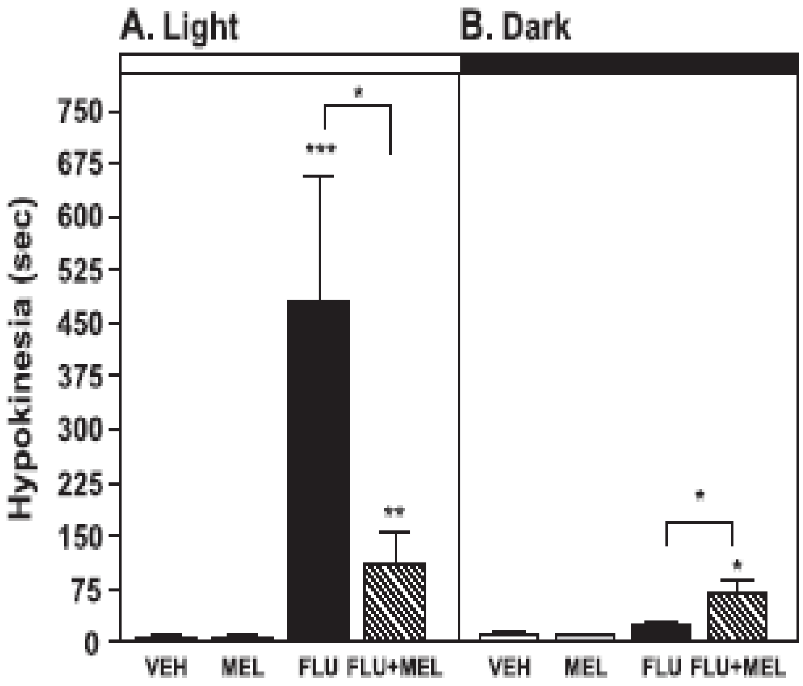
	Male	Female	total
Democrats	24	1	25
Republican	19	6	25
Total	43	7	50

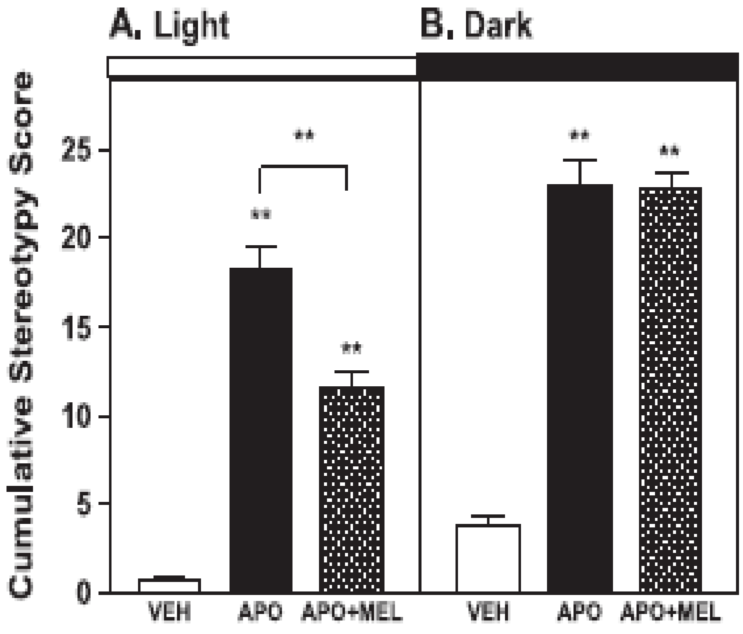
2. Graphical Representations

Graphs & Tables

Bar graph (ratio data - quantitative)

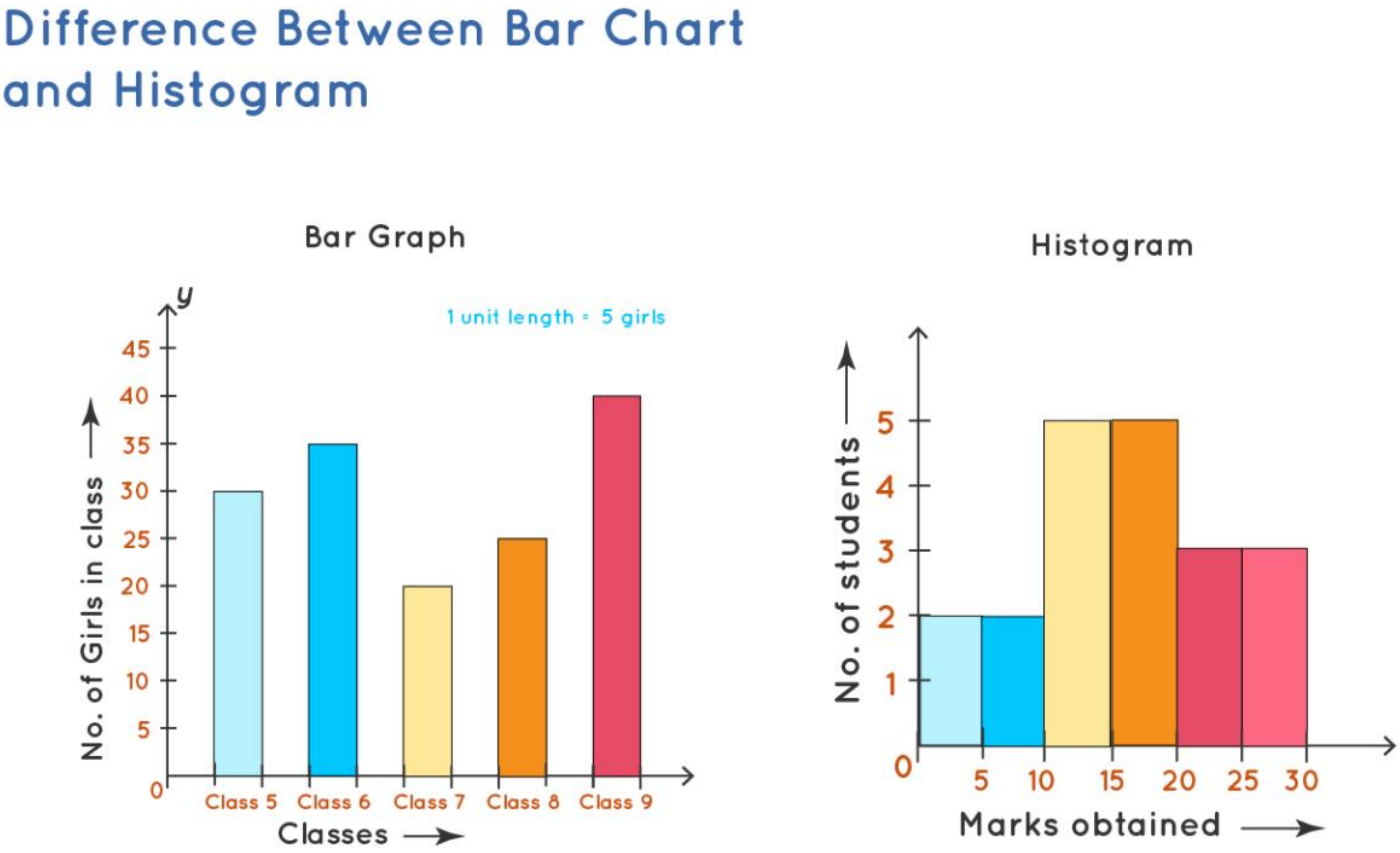
730 I.C. Sumaya et al. / Pharmacology, Biochemistry and Behavior 78 (2004) 727–733





2. Graphical Representations

Histogram of the categorical variables



Difference Between Bar Chart and Histogram

2. Graphical Representations

Line Graph

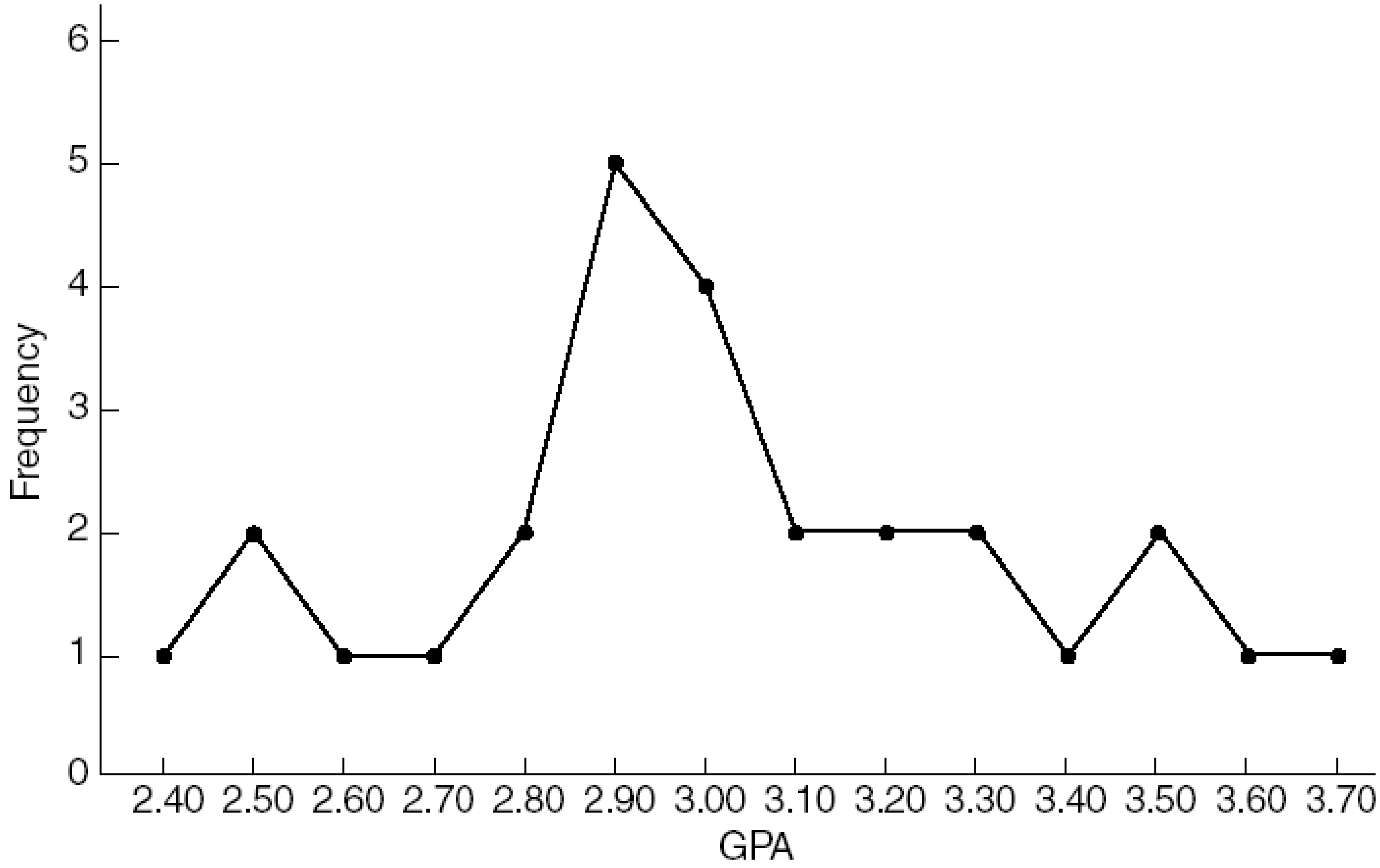


FIGURE 15.4 A line graph of grade point average

2. Graphical Representations

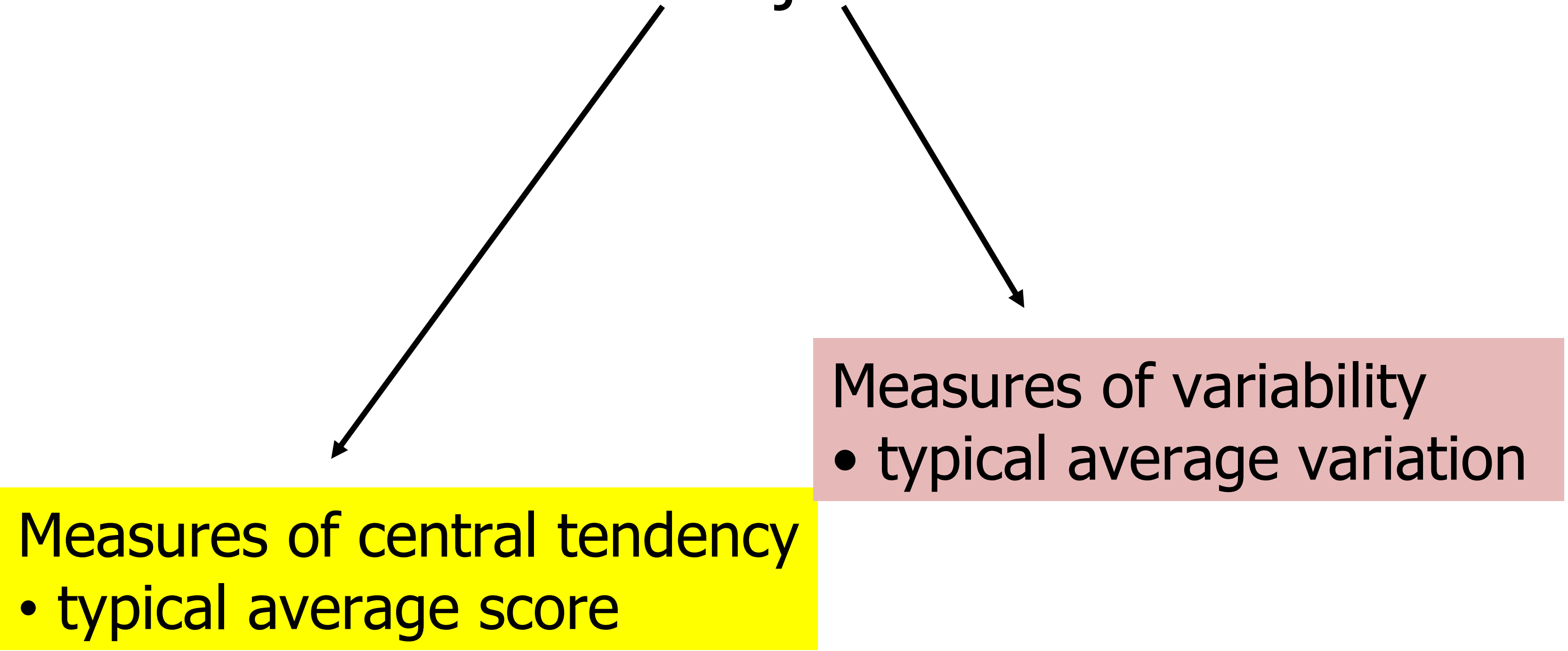
Graphs & Tables

How many brothers & sisters do you have?
Lets plot class data: HISTOGRAM

# of bros & sis	Frequency
7	?
6	?
5	?
4	?
3	?
2	?
1	?
0	?

Summary Statistics

describe data in just 2 numbers



Measures of central tendency

- typical average score

Measures of variability

- typical average variation

Measures of Central Tendency

- We use statistical measures to locate a single score that is most representative of all scores in a distribution
- Quantitative data:

  ↑ ↓

  - Mode – the most frequently occurring observation
  - Median – the middle value in the data (50 50)
  - Mean – arithmetic average
- Qualitative data:
  - Mode – always appropriate
  - Mean – never appropriate

Mean

- The most common and most useful average
- Mean = $\frac{\text{sum of all observations}}{\text{number of all observations}}$
- Observations can be added in any order.

$\mu_x = \frac{\Sigma X}{N}$ **mean of a population**

$\bar{X} = \frac{\Sigma X}{n}$ **mean of a sample**

Notation

- Sample vs population
- Sample mean = $\bar{X}$
- Population mean = μ
- Summation sign = $\sum$
- Sample size = n
- Population size = N

Special Property of the Mean
Balance Point

- The sum of all observations expressed as positive and negative deviations from the mean always equals zero!!!!
  - The mean is the single point of equilibrium (balance) in a data set
- The mean is affected by all values in the data set
  - If you change a single value, the mean changes.

Descriptive statistics

- Example 1: we want to know how 25 students performed in math tests.
- Data are in the next slide.

Descriptive statistics

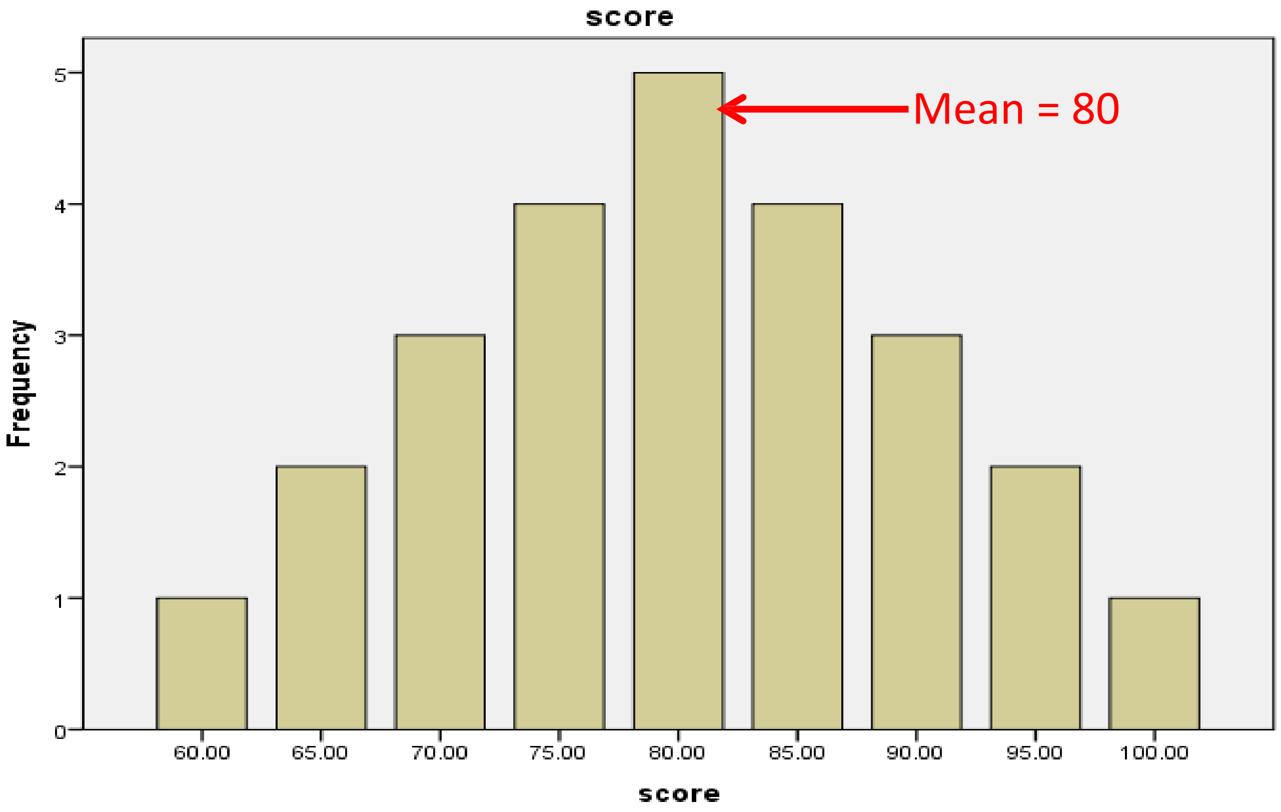
Score (X)	Frequency (f)	fX
60	1	60
65	2	130
70	3	210
75	4	300
80	5	400
85	4	340
90	3	270
95	2	190
100	1	100
Sum	25	2000

Descriptive statistics

- How to calculate mean for those 25 scores?

- $X = {}^{\Sigma f x\,/\,n} \quad = {}^{2000/25} = 80.00$

Descriptive statistics



Descriptive statistics

- Median
  - Data: 2, 3, 4, 5, 7, 10, 80.
  - Mean of those scores is 15.86.
  - 80 is an outlier.
  - Mean fails to reflect most of the data. We use median instead of mean to remove the influence of an outlier.
  - Median is the middle value in a distribution of data listed in a numeric order.

Descriptive statistics

- Median:

- For odd numbered sample size, position of median = (n +1)/2

- For odd numbered sample size: 3,6,5,3,8,6,7.

- First place each score in numeric order: 3,3,5,6,6,7,8.

- Position 4. median = 6

Descriptive statistics

- Median
  - For even-numbered sample size: 3,6,5,3,8,6. First place each score in numeric order: 3,3,5,6,6,8. Position

    3.5. Median = $\frac{5+6}{2}$ = 5.5

Descriptive statistics

- Example 2: we want to know average salary of 36 cases.

Salary	Frequency
$20k	1
$25k	2
$30k	3
$35k	4
$40k	5
$45k	6
$50k	5
$55k	4
$200k	3
$205k	2
$210k	1
Total	36

Descriptive statistics

- Median = ?
- Position 18.5
- Which number is at position 18.5?
- Median = $45k

Descriptive statistics

- Mode
  - The value in a data set that occurs most often or most frequently.
  - Example: 2,3,3,3,4,4,4,4,7,7,8,8,8. Mode = 4

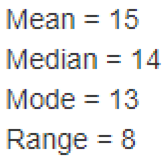
Excercise

- Find the mean, median, mode, and range for the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

Mean, Median, Mode, Range

- **Question:** Find the mean, median, mode, and range for the following list of values:

  13, 18, 13, 14, 13, 16, 14, 21, 13

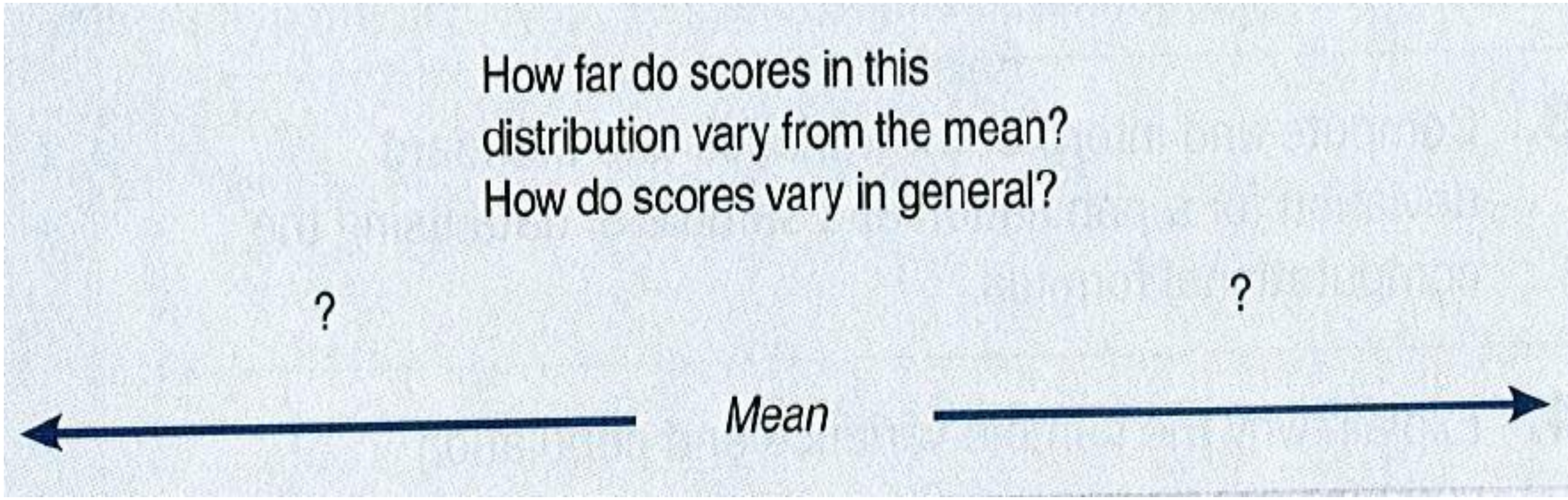


Mean = 15
Median = 14
Mode = 13
Range = 8

- **Create sample data in python using randint containing values between 7- 9 and size 20. Calculate mean, median and mode of data in Python notebook**

Descriptive statistics

- Dispersion (Variability): a measure of the spread of scores in a distribution.

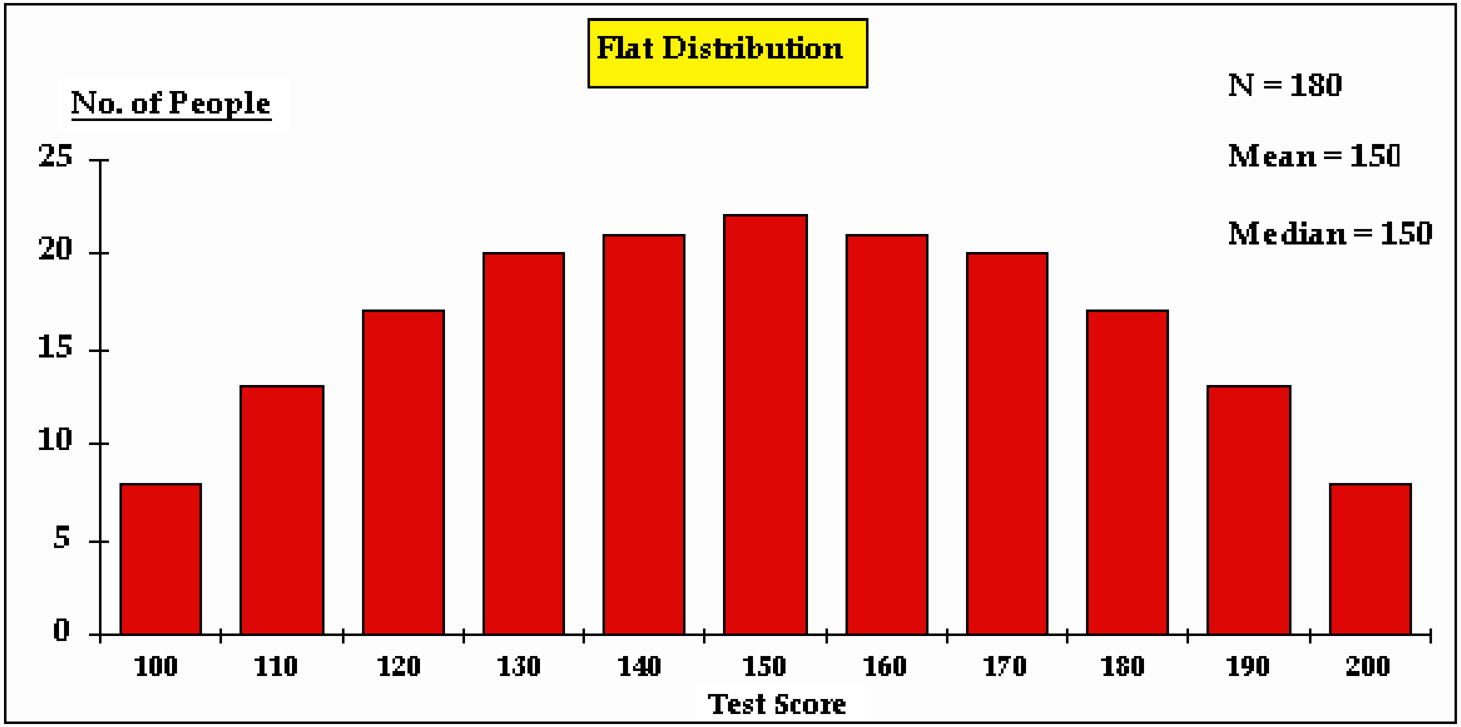


How far do scores in this distribution vary from the mean? How do scores vary in general?

- Lower Variability: Better for Prediction (Close by samples)
- Higher Variability: Harder to predict (Values may less consistent)

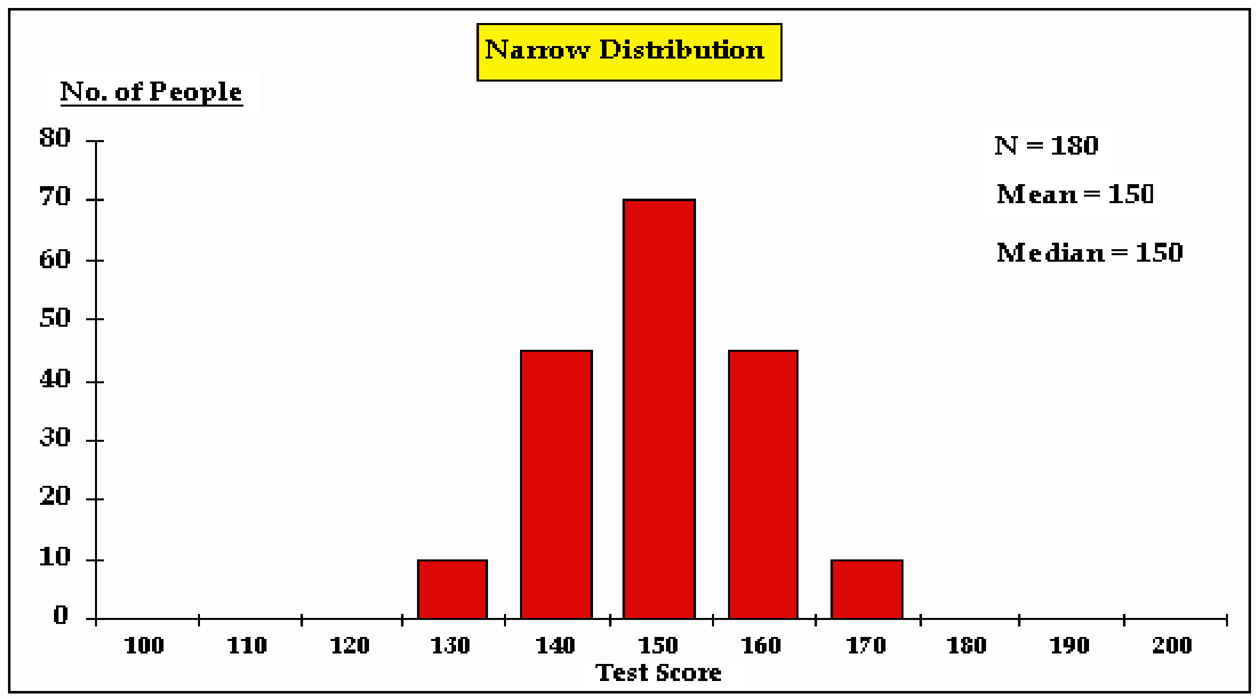
Descriptive statistics

- Compare different normal distributions



Descriptive statistics

- Compare different normal distributions



- Two sets of data have the same sample size, mean, and median.
- But they are different in terms of variability.

Descriptive statistics

- Variability commonly measured with the following:
  - Range
    - Diff. betn highest and lowest values
  - Inter Quartile Range(IQR)
    - Range of the middle half of a distribution
  - Standard deviation
    - Average distance from the mean
  - Variance
    - Average of squared distances from the mean

Descriptive statistics

- Range
  - It is the difference between the largest value and smallest value.
  - It is informative for data without outliers.

Range

- Max –Min

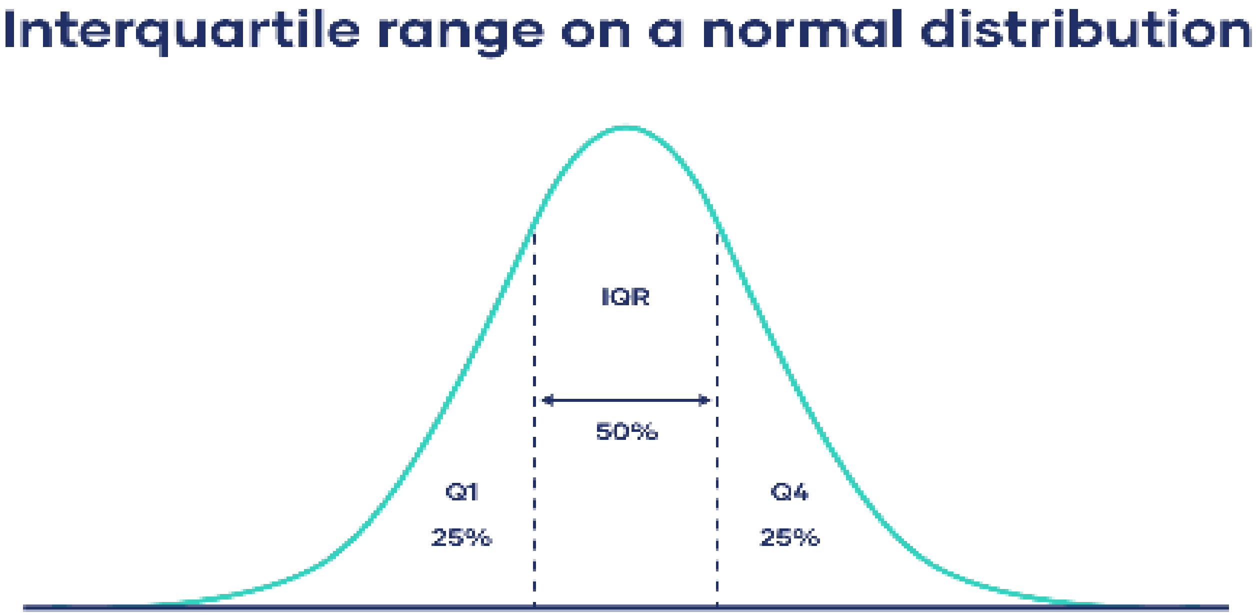
R: range(x)

unit 1	unit 2
9.7	9.0
11.5	11.2
11.6	11.3
12.1	11.7
12.4	12.2
12.6	12.5
13.1	13.2
13.5	13.8
13.6	14.0
14.8	15.5
16.3	15.6
26.9	16.2
	16.4

Descriptive statistics

- Inter Quartile Range(IQR)
  - Gives the spread of the middle of distribution.
  - For any distribution that’s ordered from low to high, the interquartile range contains half of the values.
  - First quartile (Q1) contains the first 25% of values, the fourth quartile (Q4) contains the last 25% of values.

Descriptive statistics (IQR)

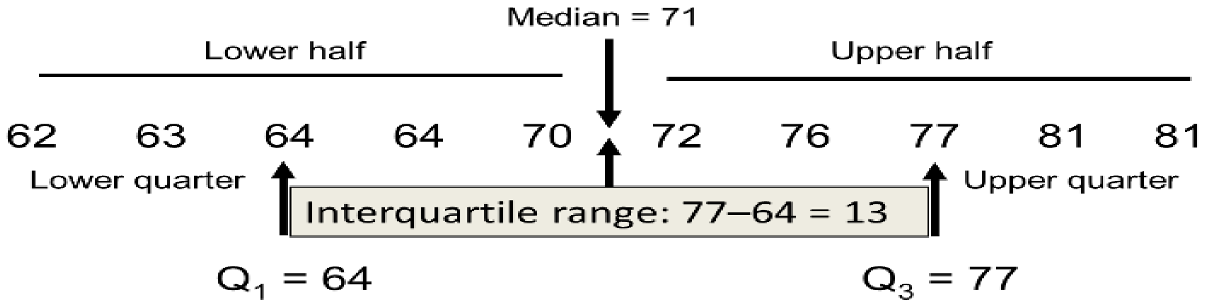


- **Interquartile range (IQR) = Value of third quartile (Q3) – Value of first quartile (Q1)**

- IQR is less affected by outliers
- It gives a consistent measure of variability for skewed as well as normal distributions

Descriptive statistics (IQR)

IQR with Even Sample Size



Median = 71

Lower half

Upper half

62 63 64 64 70 72 76 77 81 81

Lower quarter

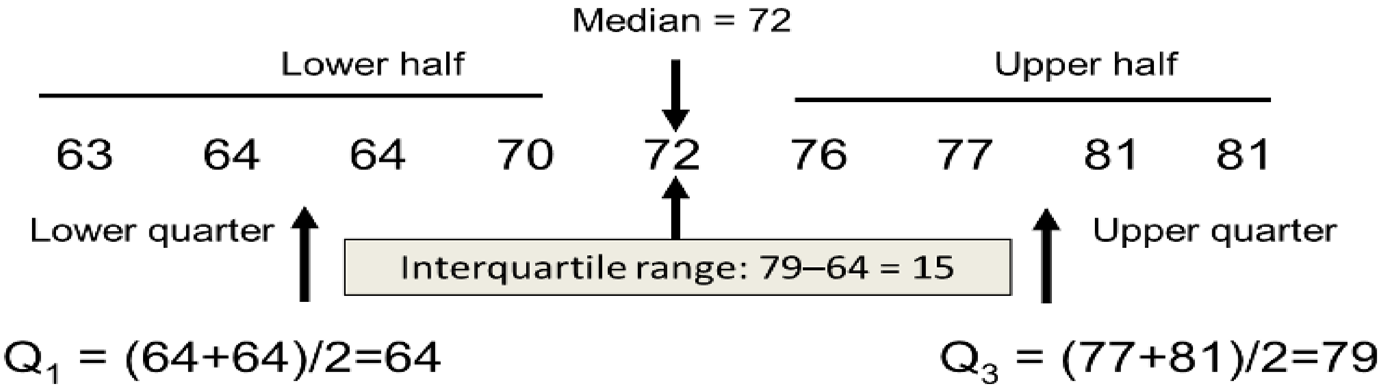
Upper quarter

Interquartile range: 77–64 = 13

Q_1 = 64

Q_3 = 77

IQR with Odd Sample Size



Median = 72

Lower half

Upper half

63 64 64 70 72 76 77 81 81

Lower quarter

Upper quarter

Interquartile range: 79–64 = 15

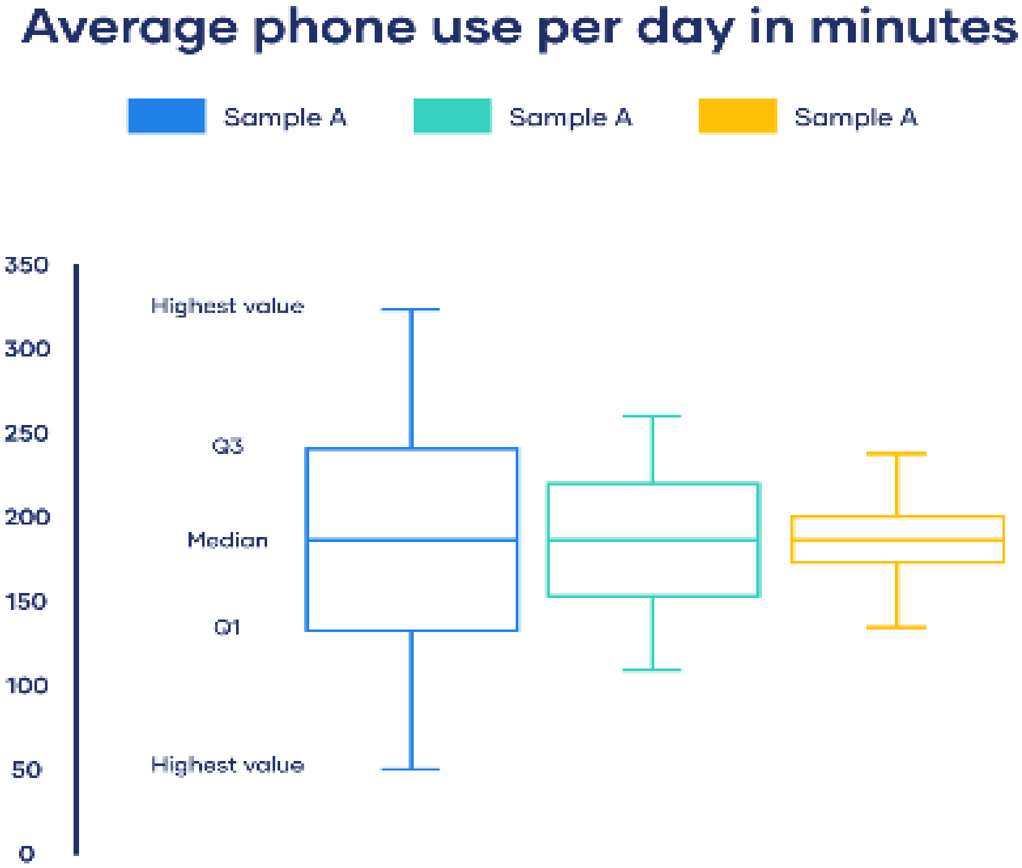
Q_1 = (64+64)/2=64

Q_3 = (77+81)/2=79

Inter Quartile Range (IQR)

Visualize the interquartile range in boxplots

For each of our samples, the horizontal lines in a box show Q1, the median and Q3, while the whiskers at the end show the highest and lowest values

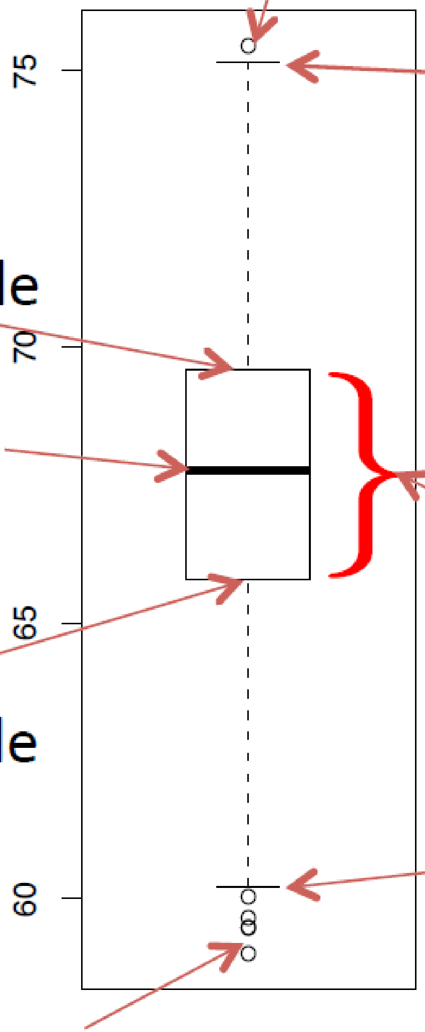


Outliers and Tukey Fences:

- When there are no outliers in a sample,
  - the mean and standard deviation are used to summarize a typical value and the variability in the sample, respectively.

- When there are outliers in a sample,
  - the median and interquartile range are used to summarize a typical value and the variability in the sample, respectively.
- Tukey Fences
  - There are several methods for determining outliers in a sample.
  - Outliers are values below Q_1-1.5(Q_3-Q_1) or
    above Q_3+1.5(Q_3-Q_1) or equivalent

Possible outliers



Largest data value inside fences

Upper Fence = $X_{0.75} + 1.5 * IQR$

Upper quartile

Median

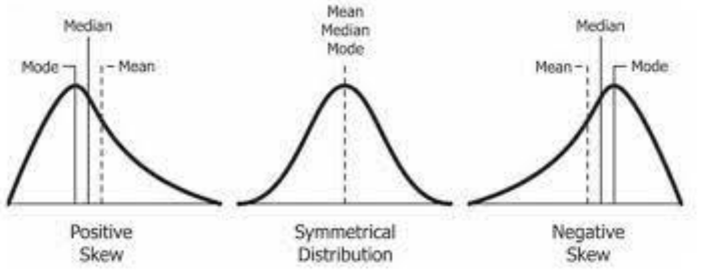
Inter-quartile Range (IQR)

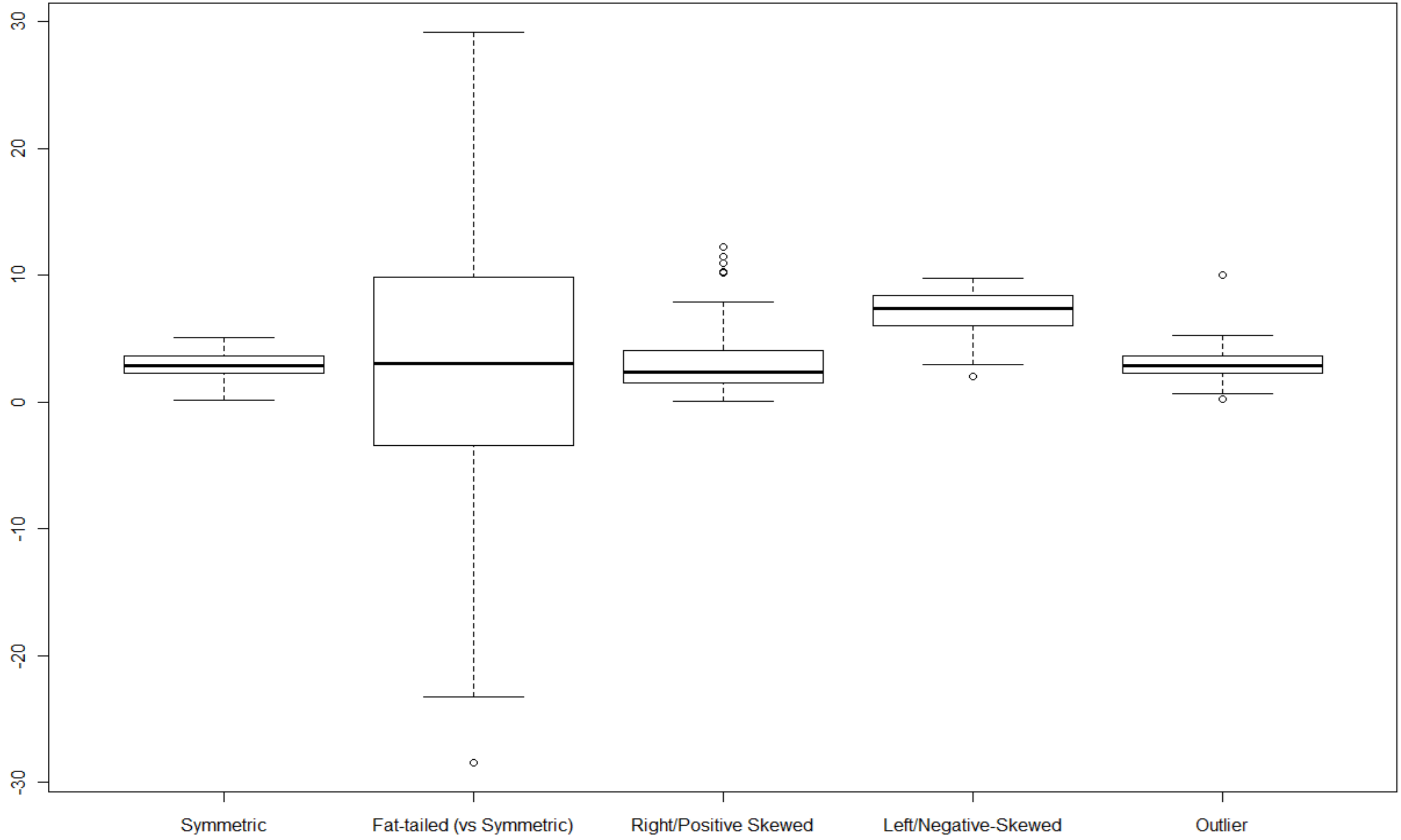
Lower quartile

Smallest data value inside fences

Lower Fence = $X_{0.25} - 1.5 * IQR$

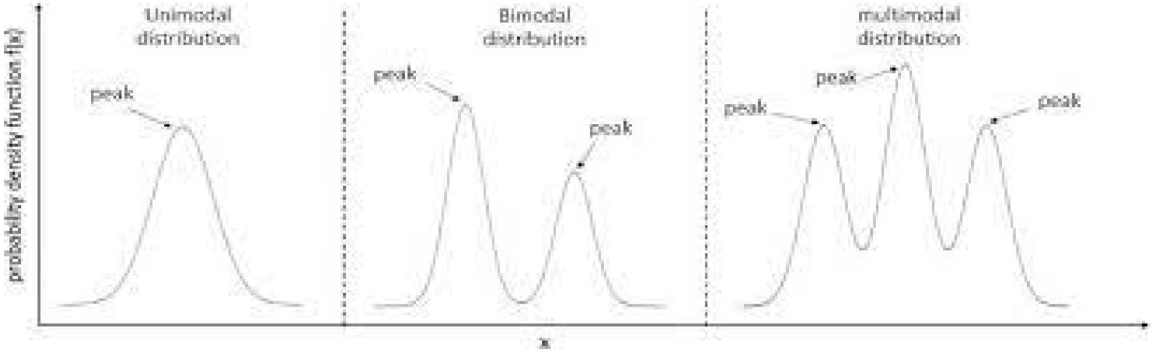
Possible outliers

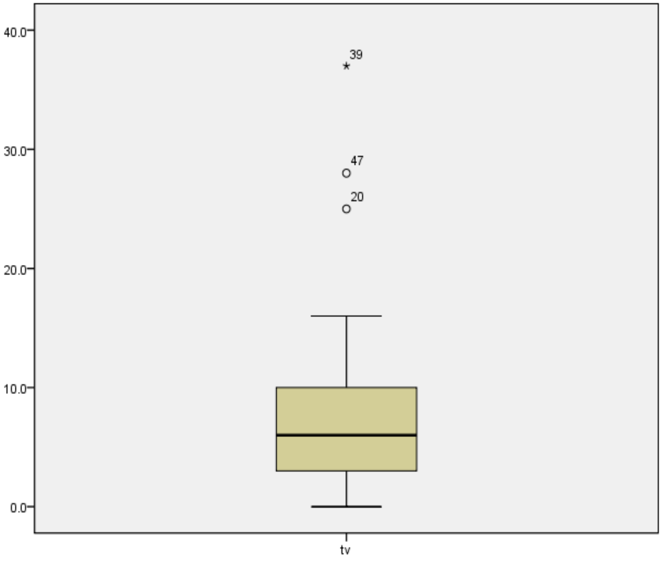

Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew


30
20
10
0
-10
-20
-30
Symmetric
Fat-tailed (vs Symmetric)
Right/Positive Skewed
Left/Negative-Skewed
Outlier

Boxplots

- For numerical data
- However, boxplots cannot identify modes (e.g. unimodal, bimodal, etc.)
- **A unimodal distribution only has one peak in the distribution**, a bimodal distribution has two peaks, and a multimodal distribution has three or more peaks.



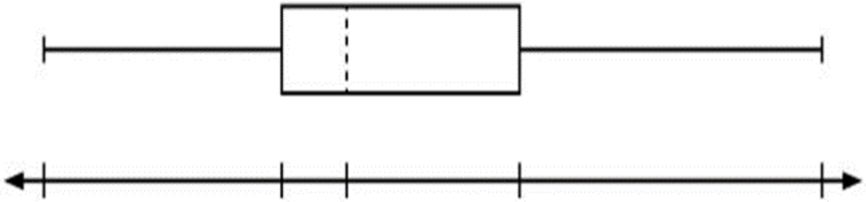


Box Plot calculation and interpretation

The following data are the heights of 40 students in a statistics class.
59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value =
- Maximum value =
- Q_1, First quartile =
- Q_2, Second quartile or median=
- Q_3, Third quartile =

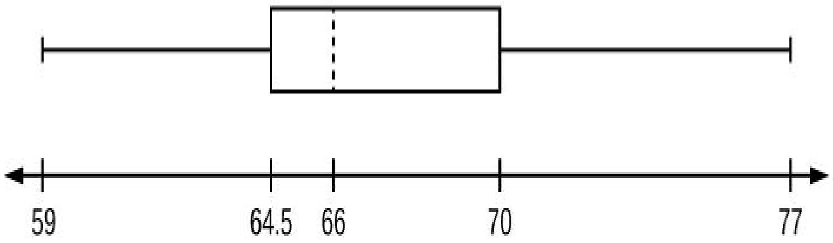


1. **Each quarter has approximately 25% of the data.**
2. **The spreads of the four quarters are …. – ….. = …… (first quarter), …… – …… = …… (second quarter), …… – ……. = ….. (third quarter), and …… – …… = ….. (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.**
3. **Range = maximum value – the minimum value = …… – …… = ……**
4. **Interquartile Range: *IQR* = *Q*3 – *Q*1 = …… – ……. = …….**
5. **The interval …… – ……. has more than 25% of the data so it has more data in it than the interval …… – ……. which has 25% of the data.**
6. **The middle 50% (middle half) of the data has a range of …….. inches.**

Box Plot calculation and interpretation

The following data are the heights of 40 students in a statistics class.
59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77
Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- Q_1, First quartile = 64.5
- Q_2, Second quartile or median= 66
- Q_3, Third quartile = 70



1. **Each quarter has approximately 25% of the data.**
2. **The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.**
3. **Range = maximum value – the minimum value = 77 – 59 = 18**
4. **Interquartile Range: *IQR* = *Q*3 – *Q*1 = 70 – 64.5 = 5.5.**
5. **The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.**
6. **The middle 50% (middle half) of the data has a range of 5.5 inches.**

Percentiles

- *p*th percentile:

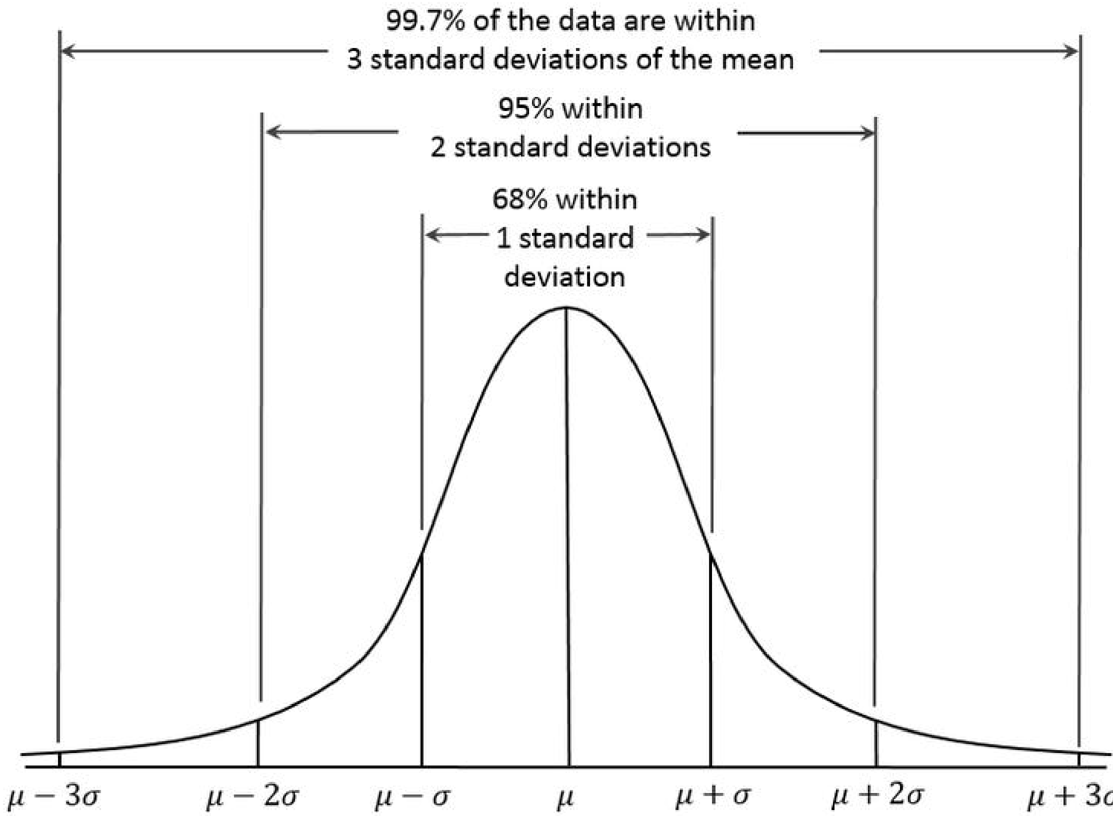
  - *p* percent of observations below it

  - (100 - *p*)% above it.

- Like 95% of CAT percentile means 95% are below it and 5% are above it
- 1,2,3,4,5,6,7,8,9,10 - What is 25th percentile?
- 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20 - What is 25th percentile? What is 80th percentile?

Standard Deviation

- Is the average amount of variability in dataset.
- It tells on an average, how far each score lies from the mean.
- The larger the standard deviation, more variable the data set is.



Standard deviation formula for population

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

σ = population standard deviation

X = each value

μ = population mean

N = number of values in the population

Standard deviation formula for sample

$$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$$

s = population standard deviation

X = each value

$\bar{x}$ = population mean

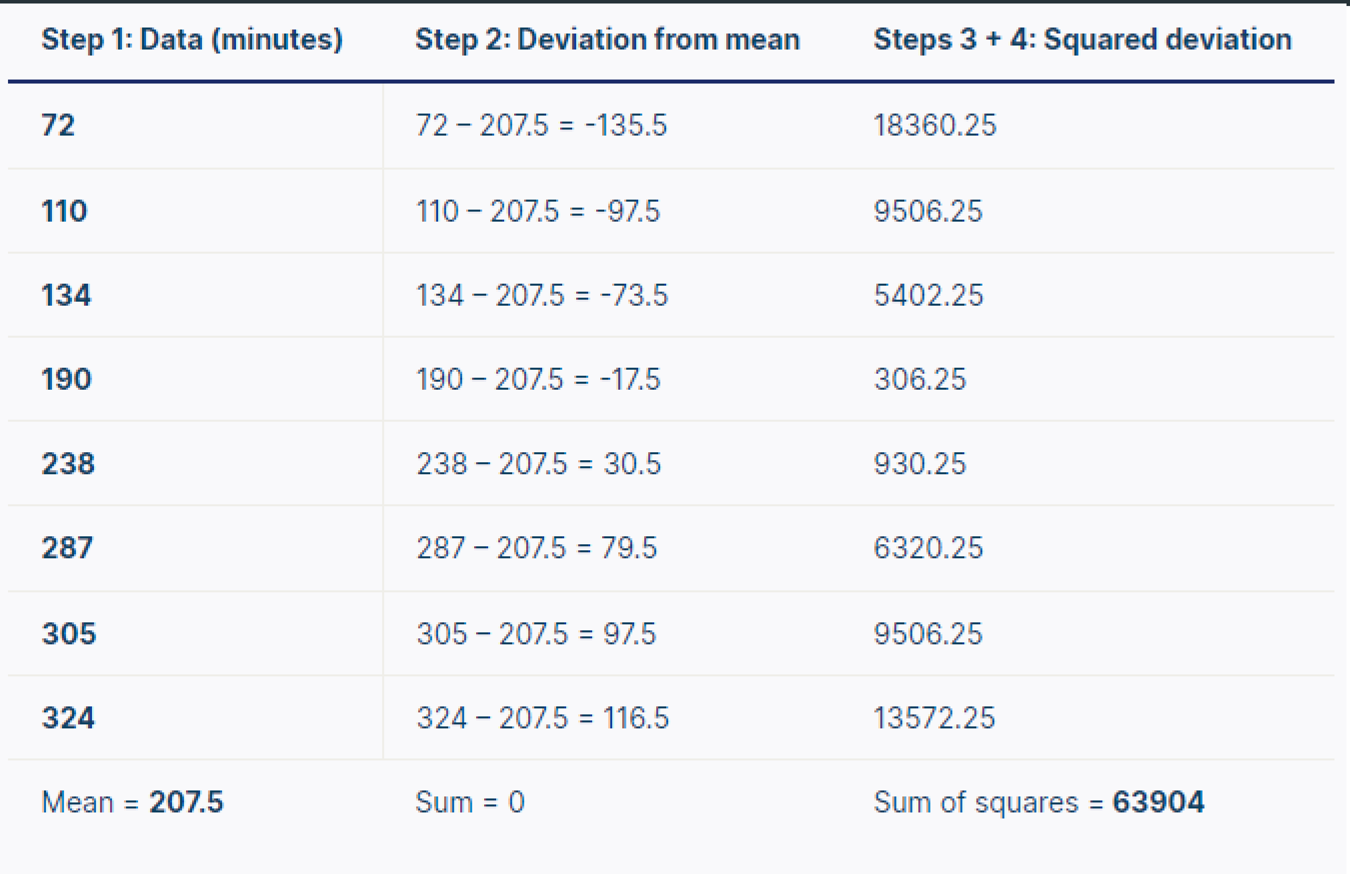
n = number of values in the population

- For population (full) data, we get an exact value for standard deviation.
- For sample data, we get an estimate of the population standard deviation (approx.)
- If we use *n* in the formula tends to give you a biased estimate that consistently underestimates variability.
- Reducing the sample *n* to *n* – 1 makes the standard deviation artificially large, gives a conservative estimate of variability.
- it is better to overestimate rather than underestimate variability in samples.

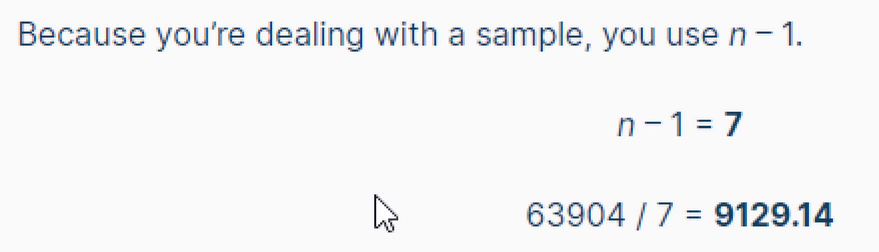
Standard Deviation Calculation

- There are six steps for finding the standard deviation by hand:
  - List each score and find their mean.
  - Subtract the mean from each score to get the deviation from the mean.
  - Square each of these deviations.
  - Add up all of the squared deviations.
  - Divide the sum of the squared deviations by *n* – 1 (for a sample) or *N* (for a population).
  - Take the square root of that number

Standard Deviation Calculation



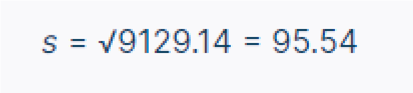
Step 1: Data (minutes)	Step 2: Deviation from mean	Steps 3 + 4: Squared deviation
72	72 − 207.5 = -135.5	18360.25
110	110 − 207.5 = -97.5	9506.25
134	134 − 207.5 = -73.5	5402.25
190	190 − 207.5 = -17.5	306.25
238	238 − 207.5 = 30.5	930.25
287	287 − 207.5 = 79.5	6320.25
305	305 − 207.5 = 97.5	9506.25
324	324 − 207.5 = 116.5	13572.25
Mean = **207.5**	Sum = 0	Sum of squares = **63904**



Because you're dealing with a sample, you use $n - 1$.

$n - 1 = \mathbf{7}$

$63904 / 7 = \mathbf{9129.14}$



$s = \sqrt{9129.14} = 95.54$

This means that on average, each score deviates from the mean by 95.54 points.

Variance

- Is the average of squared deviations from the mean (square of the standard deviation).
- A deviation from the mean is how far a score lies from the mean
- Sample variance: s^2
- Population variance: σ^2

Descriptive statistics

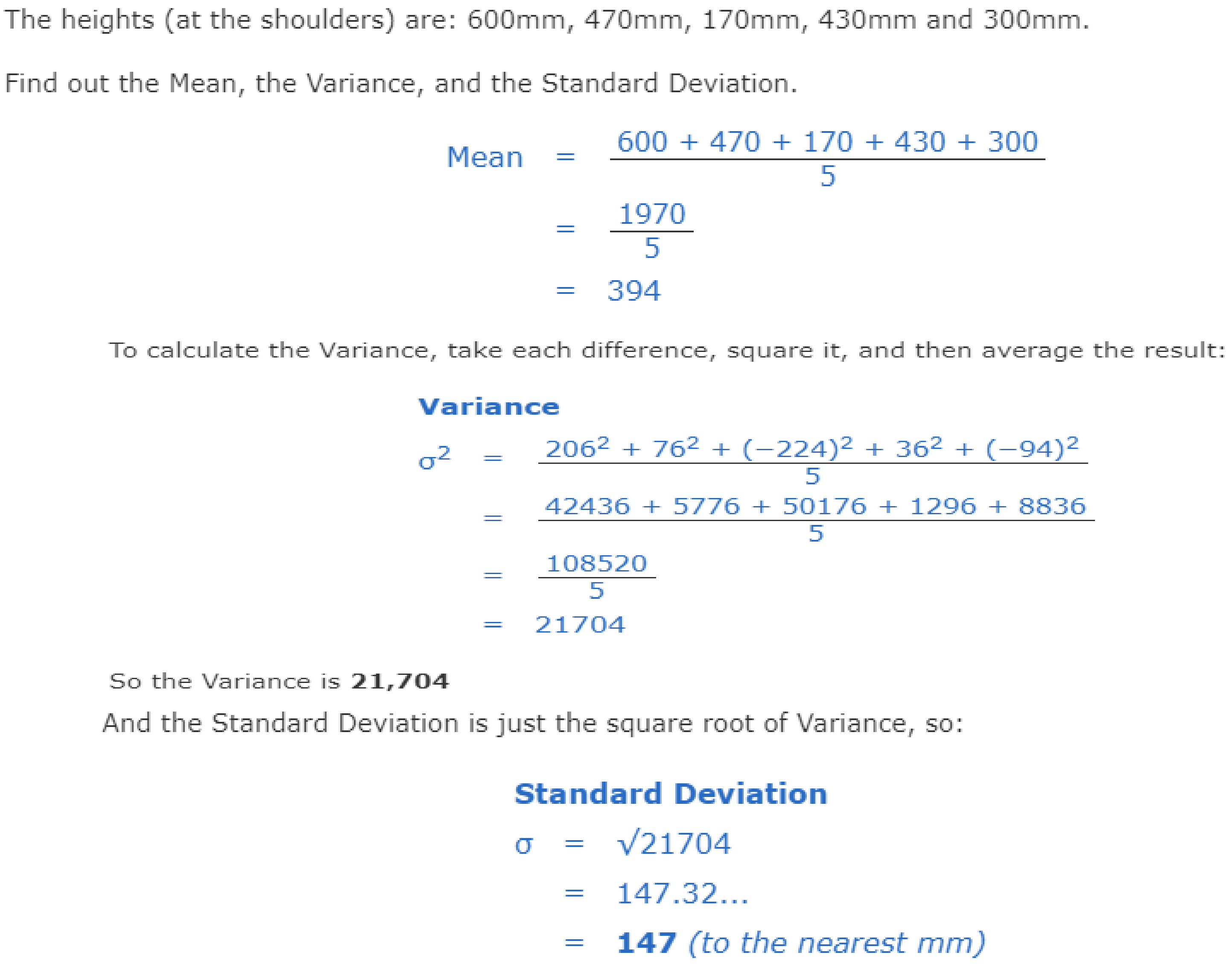
- **Variance for Populations:**

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

- **Variance for Samples:**

$$s^2 = \frac{\sum(X - \bar{x})^2}{n - 1}$$

Example with Solution



The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

$$\begin{aligned}\text{Mean} &= \frac{600 + 470 + 170 + 430 + 300}{5} \\ &= \frac{1970}{5} \\ &= 394\end{aligned}$$

To calculate the Variance, take each difference, square it, and then average the result:

Variance

$$\begin{aligned}\sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\ &= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5} \\ &= \frac{108520}{5} \\ &= 21704\end{aligned}$$

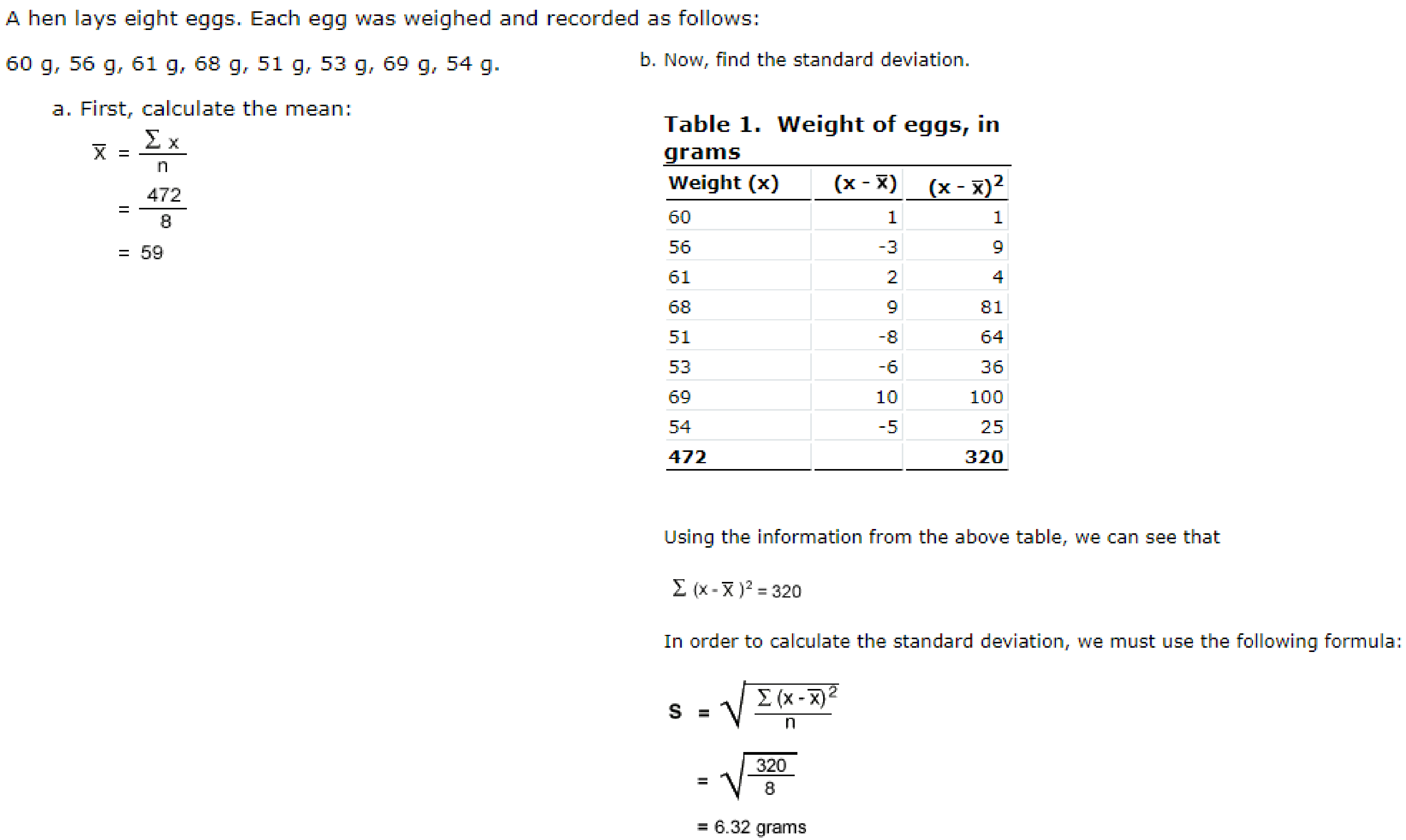
So the Variance is **21,704**

And the Standard Deviation is just the square root of Variance, so:

Standard Deviation

$$\begin{aligned}\sigma &= \sqrt{21704} \\ &= 147.32... \\ &= \mathbf{147} \text{ \textit{(to the nearest mm)}}\end{aligned}$$

Example with Solution



A hen lays eight eggs. Each egg was weighed and recorded as follows:

60 g, 56 g, 61 g, 68 g, 51 g, 53 g, 69 g, 54 g.

a. First, calculate the mean:

$$\overline{x} = \frac{\Sigma x}{n} = \frac{472}{8} = 59$$

b. Now, find the standard deviation.

Table 1. Weight of eggs, in grams

Weight (x)	$(x - \overline{x})$	$(x - \overline{x})^2$
60	1	1
56	-3	9
61	2	4
68	9	81
51	-8	64
53	-6	36
69	10	100
54	-5	25
472		**320**

Using the information from the above table, we can see that

$$\Sigma (x - \overline{x})^2 = 320$$

In order to calculate the standard deviation, we must use the following formula:

$$s = \sqrt{\frac{\Sigma (x - \overline{x})^2}{n}} = \sqrt{\frac{320}{8}} = 6.32 \text{ grams}$$

Descriptive statistics

- Normal distribution
  - –Probability: the frequency of times an outcome is likely to occur divided by the total number of possible outcomes.
    - It varies between 0 and 1.
    - Example (next slide)

Descriptive statistics

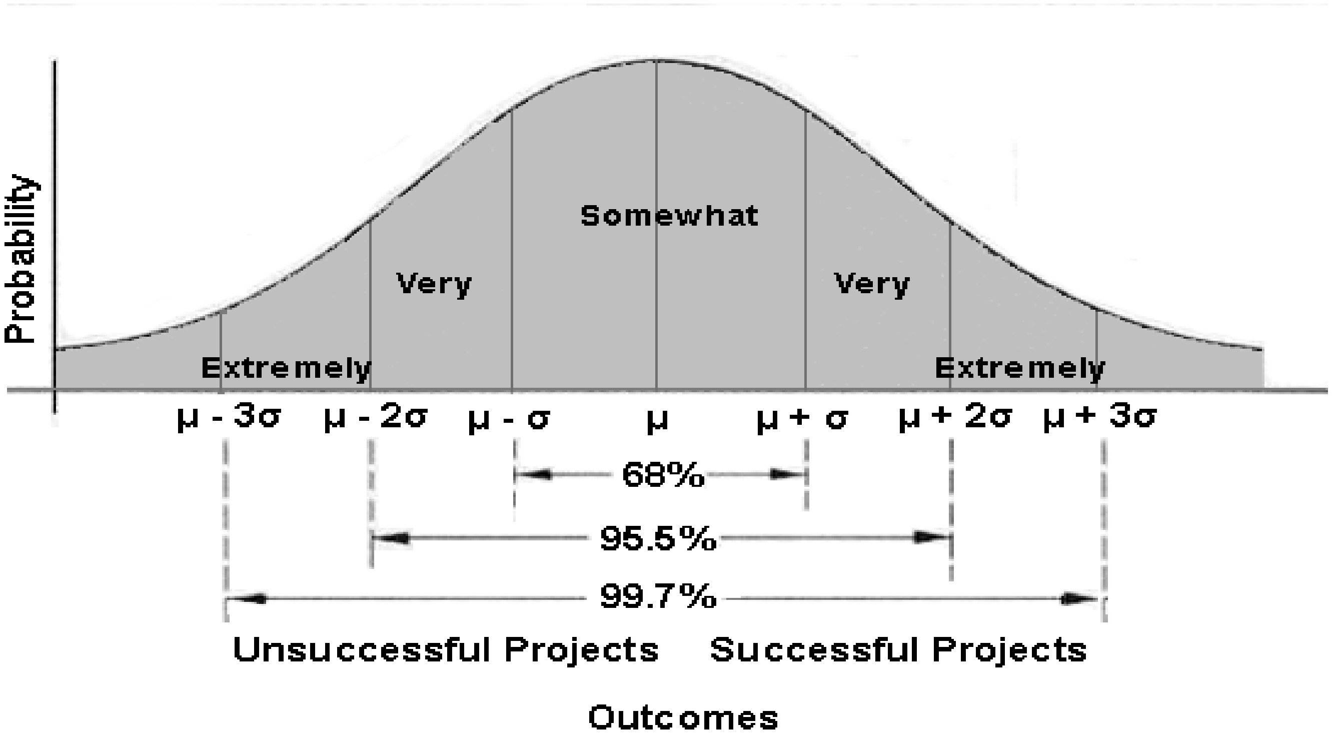
- Probability

	Fail	Pass	Total
Male	3	2	5
Female	1	4	5
Total	4	6	10

1. What is the probability of Fail? 4/10 =0.4
2. What is the probability of Pass? 6/10 = 0.6
3. What is the probability of Fail among males? 3/5 = 0.6
4. What is the probability of Pass among females? 4/5 =0.8

Descriptive statistics

- Normal Distribution/Normal Curve
  - –Data are symmetrically distributed around mean, median, and mode.
  - –Also called the symmetrical, Gaussian, or bell-shaped distribution.



Descriptive statistics

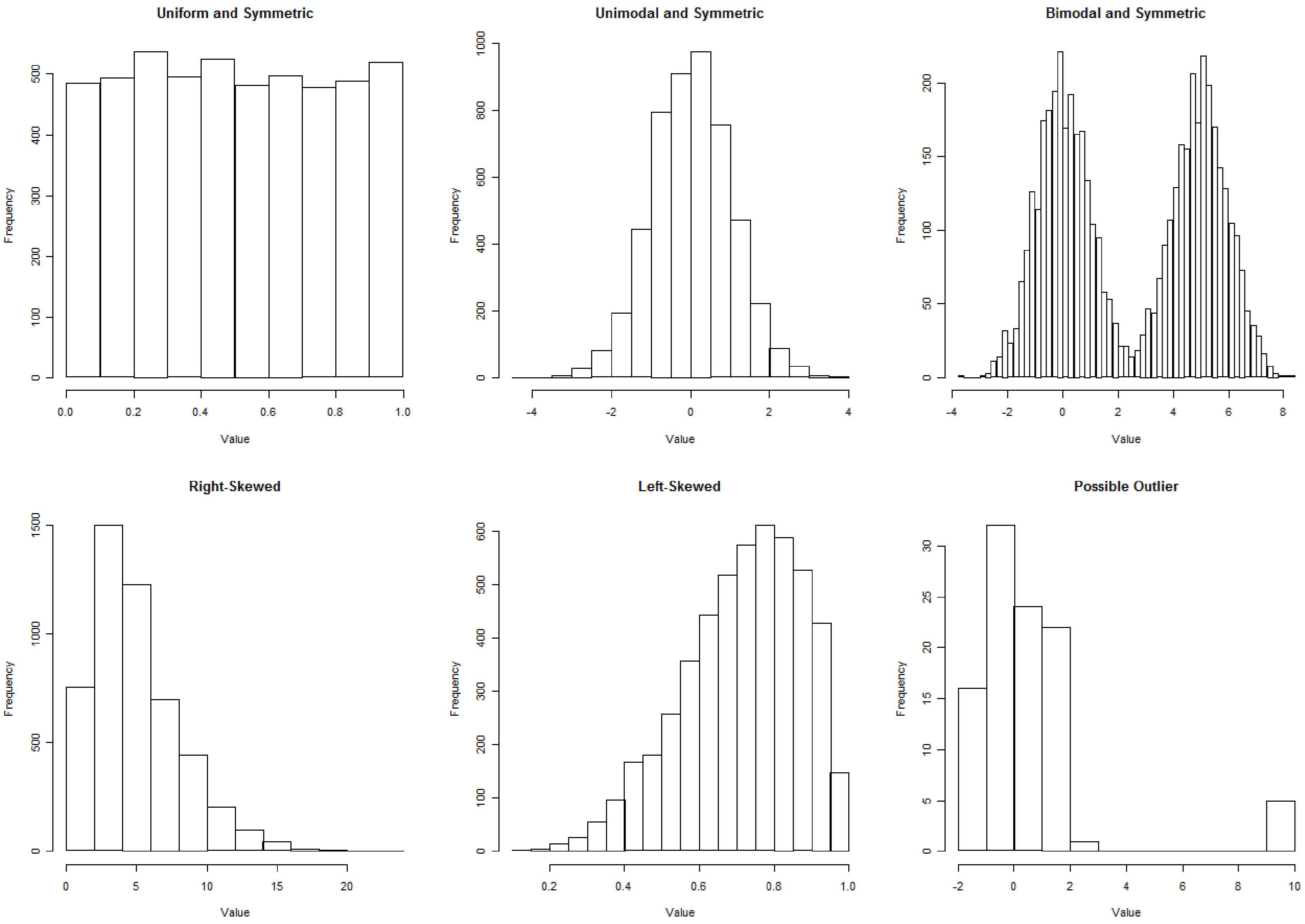
- Characteristics of normal distribution
  - The normal distribution is mathematically defined
  - The normal distribution is theoretical.
  - The mean, median, and mode are all the same values at the center of the distribution.
  - The normal distribution is symmetrical.
  - The form of a normal distribution is determined by its mean and standard deviation.
  - Standard deviation can be any positive value.
  - The total area under the curve is equal to 1.
  - The tails of normal distribution are always approaching to x-axis, but never touch it.

Histograms

- For numerical data

- A method to show the "shape" of the data by tallying frequencies of the measurements in the sample

- Characteristics to look for:
  - Modality: Uniform, unimodal, bimodal, etc.
  - Skew: Symmetric (no skew), right/positive-skewed, left/negative-skewed distributions
  - Quantiles: Fat tails/skinny tails
  - Outliers



Uniform and Symmetric

Unimodal and Symmetric

Bimodal and Symmetric

Right-Skewed

Left-Skewed

Possible Outlier

Pandas/.hist()

- **It does the grouping.**
  When using `.hist()` there is no need for the initial `.groupby()` function! `.hist()` automatically groups your data into bins. (By default, into 10 bins.)
  Note: again, “grouping into bins” is not the same as “grouping by unique values” — as a bin usually contains a range of values.

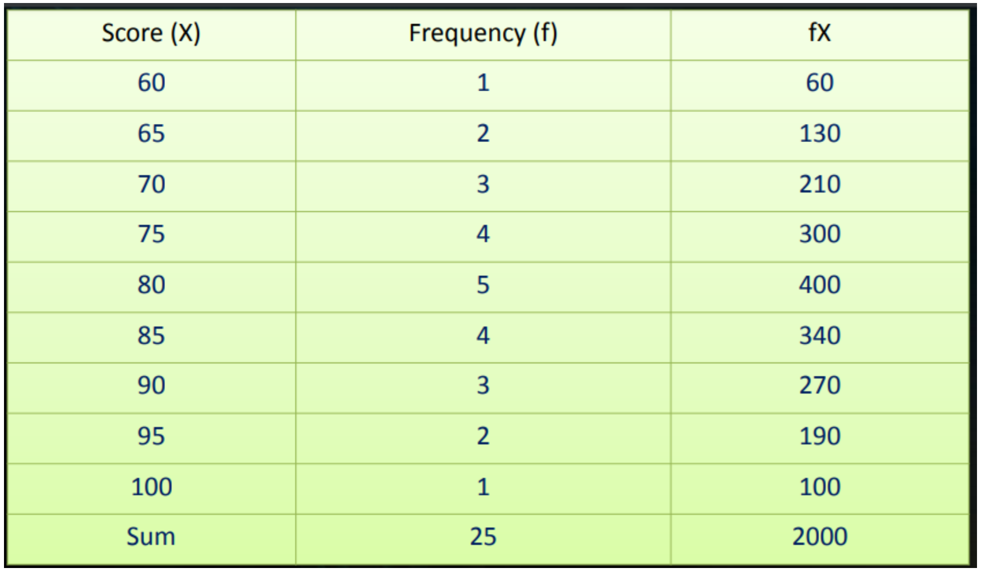
1. **It does the counting.** (No need for `.count()` function either.)

2. **It plots a histogram for each column** in your dataframe that has numerical values in it.

- **So plotting a histogram (in Python, at least) is definitely a very convenient way to visualize the distribution of your data.**

Histogram

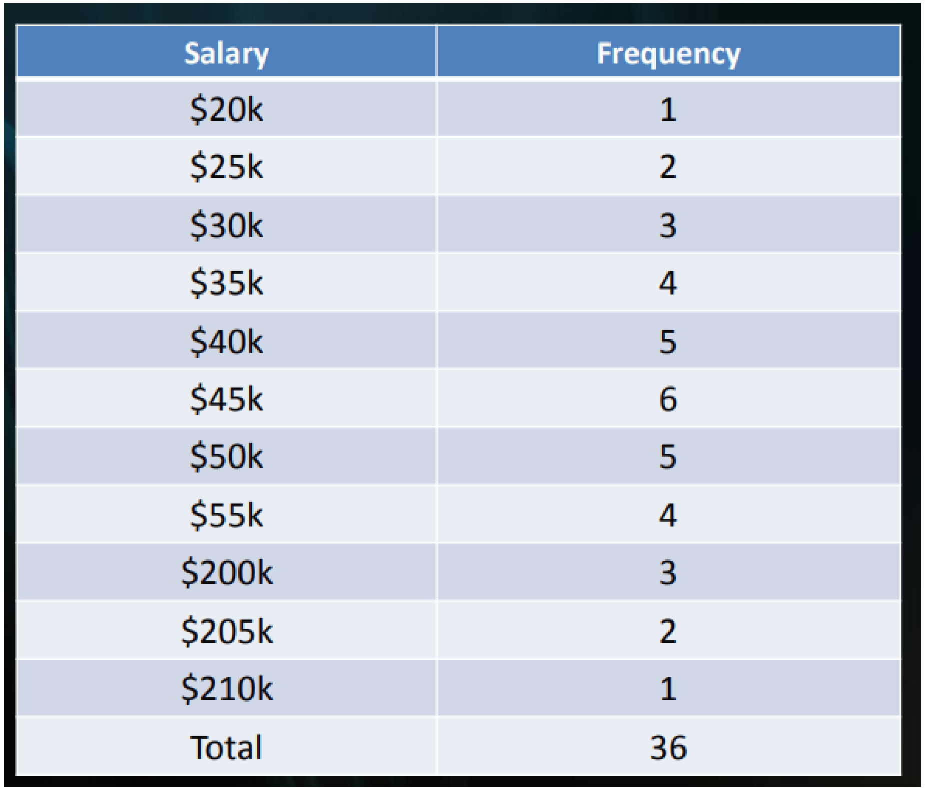
- Calculate Mean, Median using formula.
- Plot Histogram of scores and find mean and median



Score (X)	Frequency (f)	fX
60	1	60
65	2	130
70	3	210
75	4	300
80	5	400
85	4	340
90	3	270
95	2	190
100	1	100
Sum	25	2000

Histogram

- Find mean, median of the given data using formula.
- Plot histogram and find mean and median.



Salary	Frequency
$20k	1
$25k	2
$30k	3
$35k	4
$40k	5
$45k	6
$50k	5
$55k	4
$200k	3
$205k	2
$210k	1
Total	36

Sample data for test

- Create an array containing 10000 random floats between 0 and 5:

  Between 0 and 5

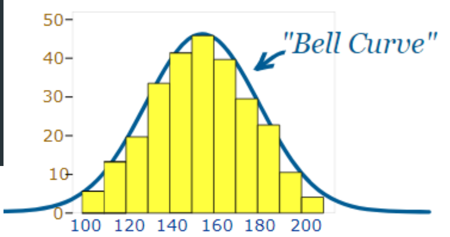
  - numpy.random.uniform(0.0, 5.0, 100000) , Hist?

- Data from *normal data distribution*, or the *Gaussian data distribution*
  - numpy.random.normal(5.0, 1.0, 100000) , Hist?

  mean

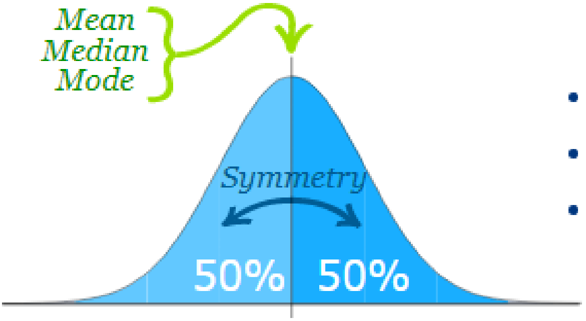
  sd

Normal Distribution



A Normal Distribution

- Normal distribution/Normal curve
  - Also called the symmetrical, Gaussian, or bell-shaped distribution.
- Characteristics of normal distribution
  - The normal distribution is mathematically defined.
  - The normal distribution is symmetrical
  - The form of a normal distribution is determined by its mean and standard deviation.
  - Standard deviation can be any positive value.
  - The tails of normal distribution are always approaching to x axis, but never touch it.
  - The total area under the curve is equal to 1.
  - We use normal distribution to locate probabilities for scores.
  - The area under the curve can be used to determine the probabilities at different points.
  - **Example: About 95% of all scores lie within two standard deviation of the mean (Normal scores: close to the mean). we have 95% chance of selecting a score that is within 2 standard deviation of mean. 68-95-99.7 rule.**



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

Normal Distribution to Standard Normal Distribution

- Convert a value to a Standard Score ("z-score"):

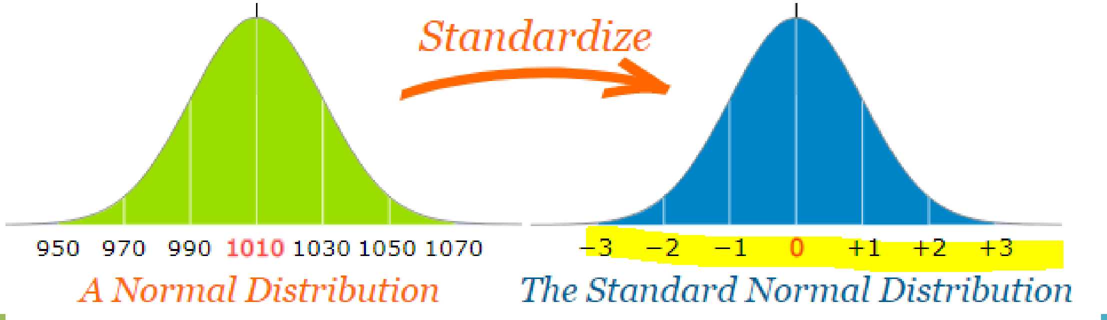
  $z = (x – \mu) / \sigma$

  – **x observation**

  – **μ mean**

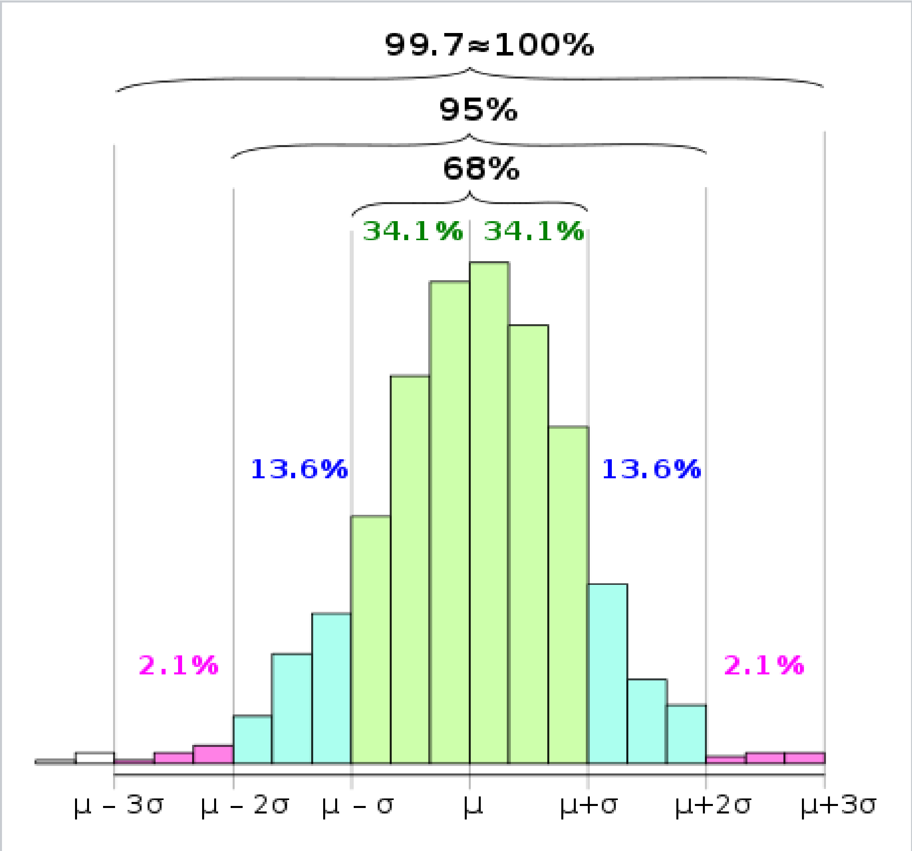
  – **σ standard deviation**

  - first subtract the mean,
  - then divide by the Standard Deviation
  - doing that is called "**Standardizing**"



Standard normal distribution or Z distribution

- A normal distribution with mean = 0, and standard deviation = 1.
- A Z score is a value on the x-axis of a standard normal distribution.
- We can take any Normal Distribution and convert it to The Standard Normal Distribution.



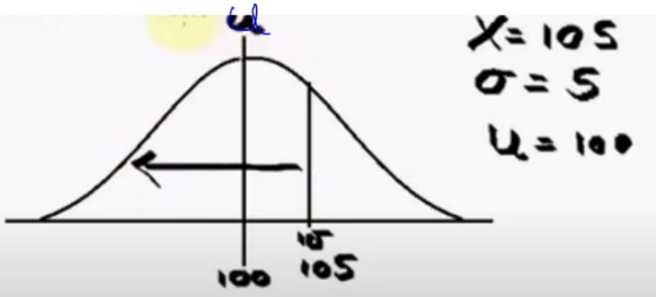
For an approximately normal data set, the values within one standard deviation of the mean account for about 68% of the set; while within two standard deviations account for about 95%; and within three standard deviations account for about 99.7%. Shown percentages are rounded theoretical probabilities intended only to approximate the empirical data derived from a normal population.

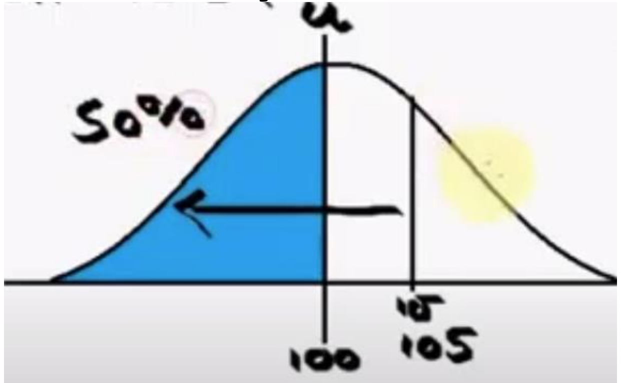
Z Score

- **z = (x – μ) / σ**
  - **x observation**
  - **μ mean**
  - **σ standard deviation**

What is the probability that any observed value is less than 105? Greater than 105?

x=105, μ=100, σ =5, Find Z score & Probability under the curve



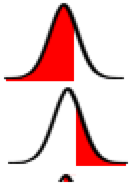


Z = 105-100/5=**1**

Refer Z table : P(x<Z) = 0.84134 i. e **84 %**
P(x>Z) = 0.15866 i. e 15%

P(x<Z) = 0.84134

P(x>Z) = 0.15866



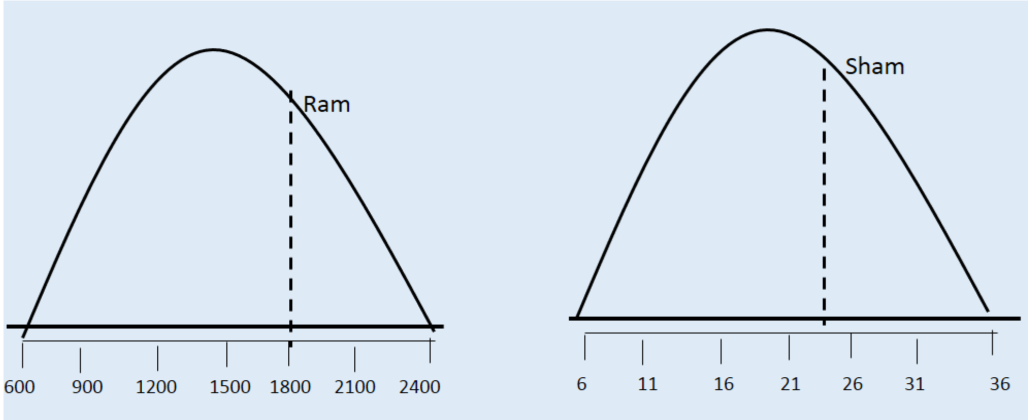
https://www.mathsisfun.com/data/standard-normal-distribution-table.html

Z-score Example

A person is having two sons. He wants to know who scored better on their standardized test with respect to the other test takers. Ram who earned an 1800 on his SAT or Sham who scored a 24 on his ACT Exam ?

SAT Score ~ N(mean = 1500, Sd= 300)

ACT Score ~ N(mean = 21, Sd= 5)



Here we cannot simply compare and tell who has done better as they are measured in different scale.

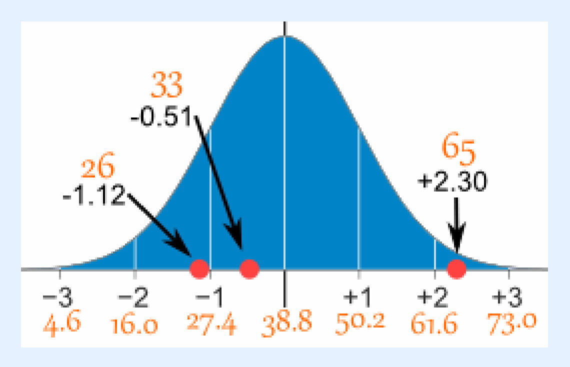
So, his father will be interested to observe how many standard deviation of their respective mean of their distribution Ram and Sham score.
Ram = (1800- 1500) / 300 =1 standard deviation above the mean
Sham = (24 – 21) / 5= 0.6 standard deviation above the mean
Now his father can conclude Ram indeed did a better score than Sham.

Example

- A survey of daily travel time had these results (in minutes):

  26, 33, 65, 28, 34, 55, 25, 44, 50, 36, 26, 37, 43, 62, 35, 38, 45, 32, 28, 34

- The **Mean is 38.8 minutes**, and the **Standard Deviation is 11.4 minutes**.
- Convert the values to the Standard Normal Distribution and plot histogram.

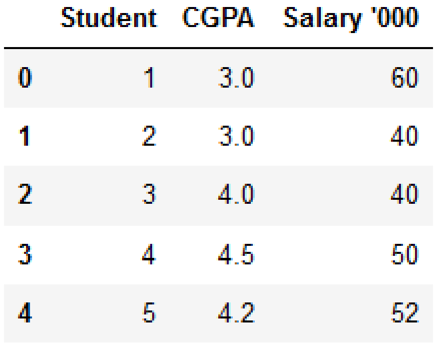


Why **Standardization?**

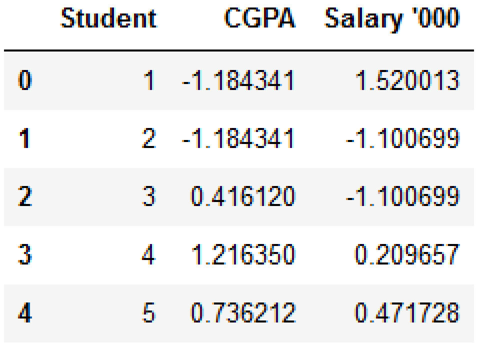
- Datasets that have multiple features spanning varying degrees of magnitude, range, and units.
  - This is a significant obstacle as a few machine learning algorithms are highly sensitive to these features.
  - For example, one feature is entirely in kilograms while the other is in grams, another one is liters, and so on.
- Feature Scaling: normalization and standardization

Feature Scaling

- CGPA scores of students (ranging from 0 to 5)
- future incomes (in thousands Rupees):
- Since both the features have different scales, there is a chance that higher weightage is given to features with higher magnitude.
  - – This will impact the performance of the machine learning algorithm
  - – **Scale your Data**



	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52



	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

Feature Scaling

- **Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.**

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation**

$$X' = \frac{X - \mu}{\sigma}$$

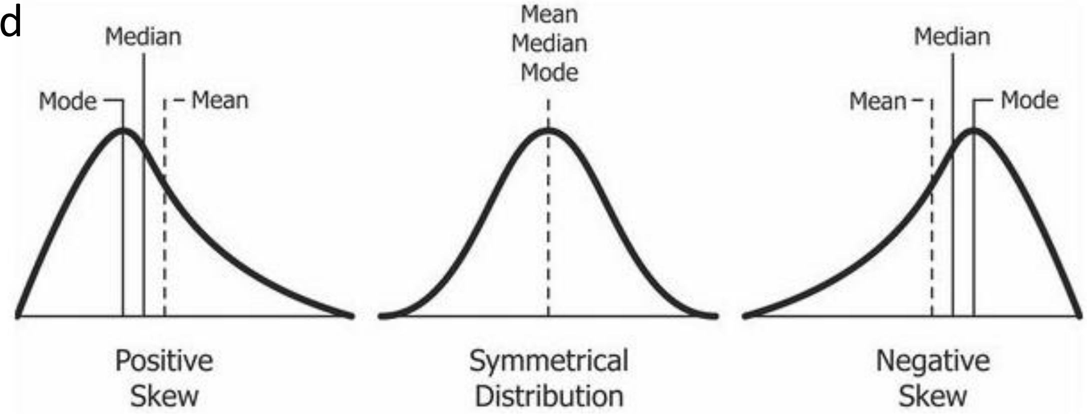
- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution

Skewness

- Skewness is a number that indicates to what extent a variable is asymmetrically distributed.
- It is the degree of distortion from the symmetrical bell curve or the normal distribution.
- A symmetrical distribution will have a skewness of 0.

For a right skewed distribution, **Mean >= Median >= Mode**



For a left skewed distribution, **Mode >= Median >=Mean**

What to do when data is skewed?

- It is very difficult to interpret and analyze the data which is skewed.

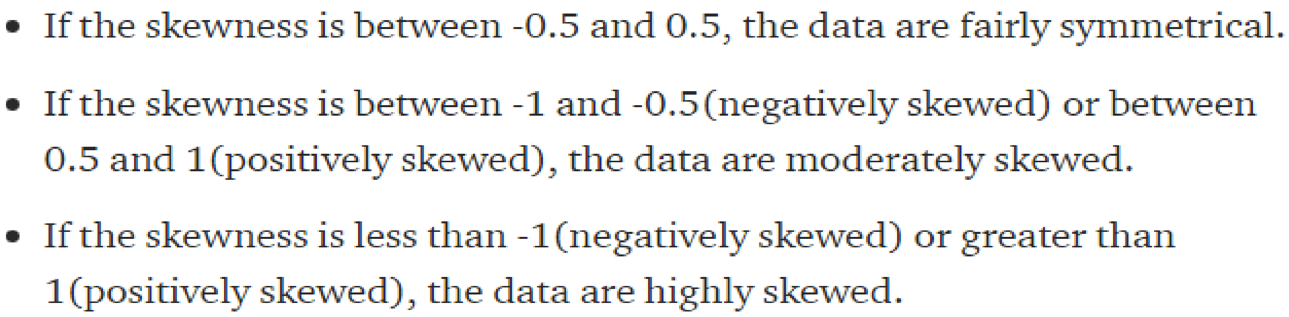
- Transformations to be applied in the data so that its information will be preserved and at the same time data will be get plotted under a symmetrical curve.

- Transformation is decided based on the characteristics of the data

Transformation

- Taking the **square root /cube root/logarithm / reciprocal** of each data point and plotting it again.
- There are several techniques to calculate skewness. One of the techniques is:

By using mean, Sk = (Mean – Mode)/Std Dev $-1 < sk <1$



- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
- If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.

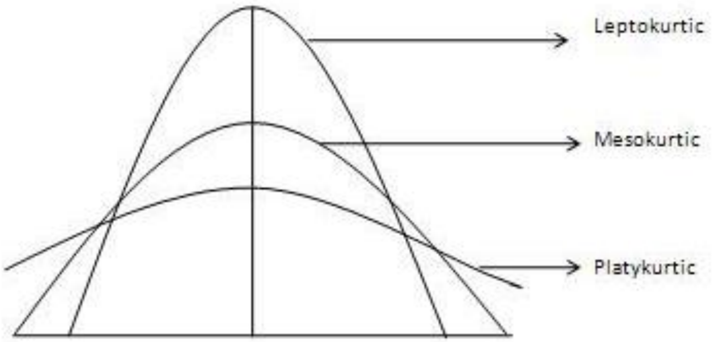
If mode is not defined, we can use median,

Sk = 3(Mean – Median)/Std Dev $-3< sk <3$

Kurtosis

- Kurtosis is all about the tails of the distribution
- It is actually the *measure of outliers* present in the distribution.
- **High kurtosis** in a data set is an indicator that data has heavy tails or outliers.
  - investigate why do we have so many outliers.
- **Low kurtosis** in a data set is an indicator that data has light tails or lack of outliers
  - need to investigate and trim the dataset of unwanted results

Kurtosis



- **Mesokurtic**: This distribution has kurtosis statistic similar to that of the normal distribution (***Kurtosis = 0***).
- **Leptokurtic (*Kurtosis > 3*):** Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-tailed or more outliers.
- **Platykurtic: (*Kurtosis < 3*):** Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers.

Lab : Histogram

https://www.spss-tutorials.com/skewness/
https://brownmath.com/stat/shape.htm

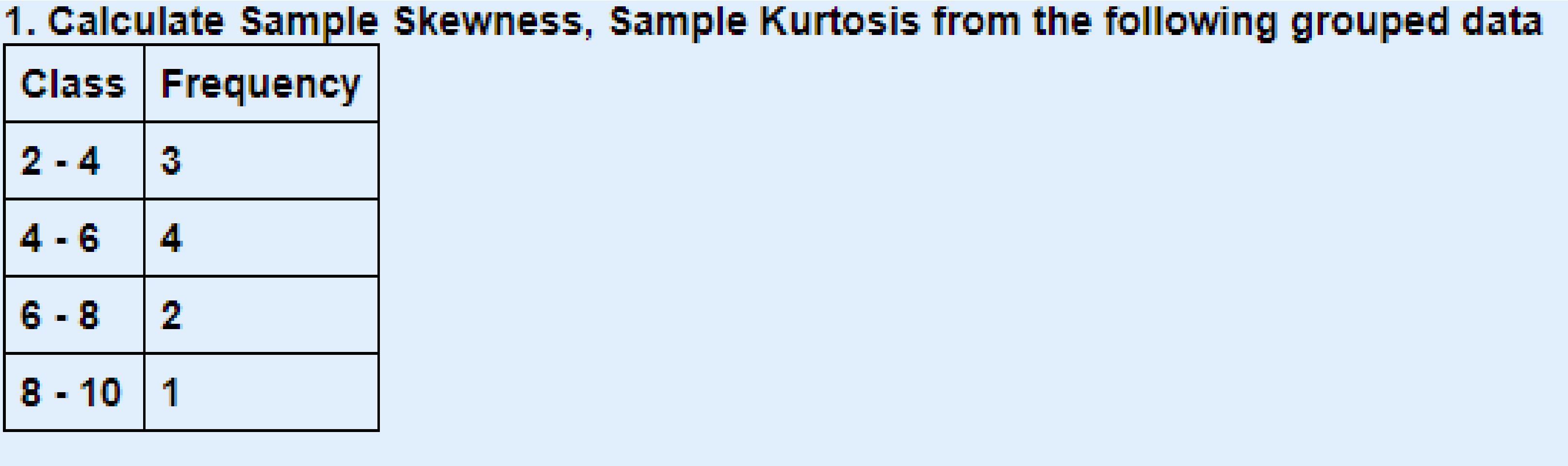
- Create a histogram on variable ‘actual’ in prdsale data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?
- Create a histogram on variable ‘msrp’ in cars data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?
- Create a histogram on variable ‘weight’ in cars data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?

Compare the above three histograms.

Lab

- What is the mean of ‘msrp’ in cars data?
- Is it reflecting the average value of price?
- What is median of ‘msrp’ in cars data?
- Is it reflecting the average value of price?
- Run Proc Univariate on weight varaibale in cars data. Find mean, Median & Mode.

Excercise



1. Calculate Sample Skewness, Sample Kurtosis from the following grouped data

Class	Frequency
2 - 4	3
4 - 6	4
6 - 8	2
8 - 10	1

Contingency Tables

- Cross classifications of categorical variables in which rows (typically) represent categories of explanatory variable and columns represent categories of response variable.
- Counts in “cells” of the table give the numbers of individuals at the corresponding combination of levels of the two variables

Example: Happiness and Family Income of 1993 families (GSS 2008 data: “happy,” “finrela”)

	Happiness			
Income	Very	Pretty	Nottoo	Total
Above Aver.	164	233	26	423
Average	293	473	117	883
Below Aver.	132	383	172	687
Total	---- 589	1089	315	1993

Contingency tables

- Example: Percentage “very happy” is
  - 39% for above average income (164/423 = 0.39)
  - 33% for average income (293/883 = 0.33)
  - What percent for below average income?

	Happiness			
Income	Very	Pretty	Not oo	Tota
				l
Above	164 (39%)	233 (55%)	26 (6%)	423
Average	293 (33%)	473 (54%)	117 (13%)	883
Below	132 (19%)	383 (56%)	172 (25%)	687

- What can we conclude? Is happiness depending on Income? Or Happiness is independent of Income?
- Inference questions for later chapters?

Correlation

- **Correlation** describes strength of association between two variables
- Falls between -1 and +1, with sign indicating direction of association (formula & other details later)
- The larger the correlation in absolute value, the stronger
  the association (in terms of a straight line trend)
- **Examples**:   (positive or negative, how strong?)
  - Mental impairment and life events, correlation =
  - GDP and fertility, correlation =
  - GDP and percent using Internet, correlation =

Calculating Correlation

The most widely used formula to compute correlation coefficient is Pearson's ’r’ for Sample

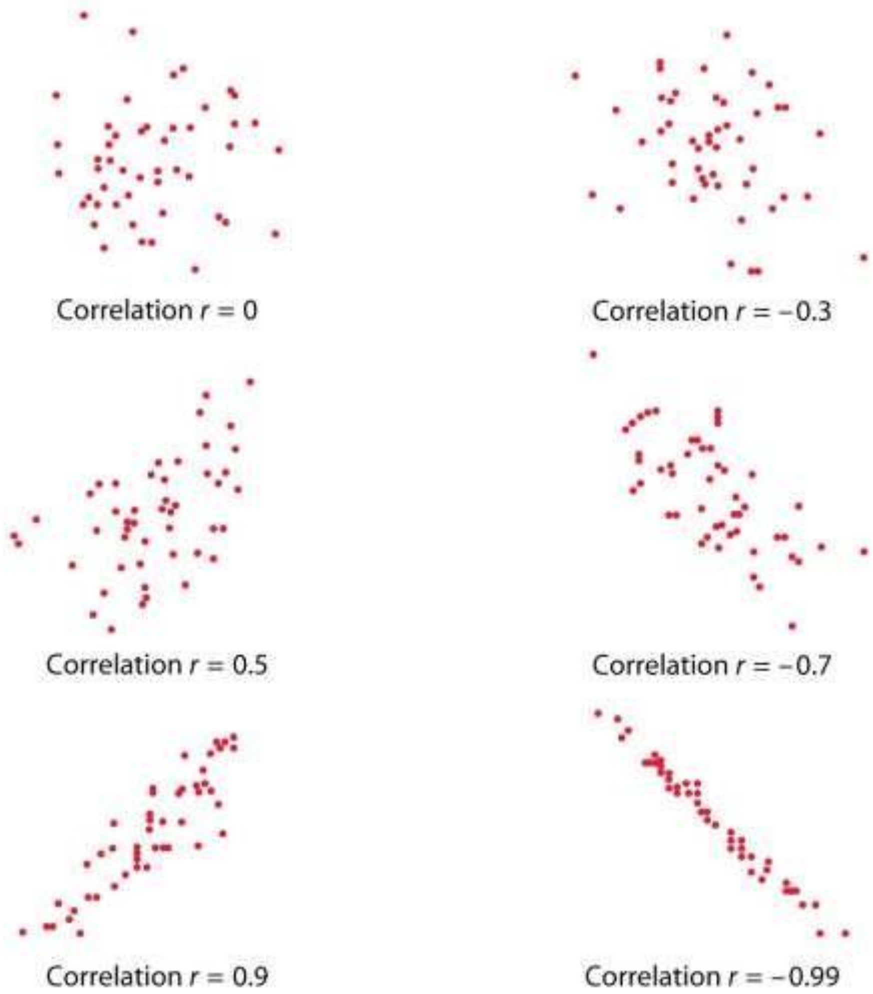
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

In the above formula,

- x_i, y_i - are individual elements of the x and y series
- The numerator corresponds to the covariance
- The denominators correspond to the individual standard deviations of x and y

Strength of Association

- Correlation 0 No linear association
- Correlation 0 to 0.25 Negligible positive association
- Correlation 0.25-0.5 Weak positive association
- Correlation 0.5-0.75 Moderate positive association
- Correlation >0.75 Very Strong positive association
- What are the limits for negative correlation



Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$

Regression

- **Regression analysis** gives line predicting *y* using *x*(algorithm & other details later)

- *y* = college GPA, *x* = high school GPA
- Predicted y = 0.234 + 1.002(x)

Calculating Covariance

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

In the above formula,

- x_i, y_i - are individual elements of the x and y series
- $\bar{x}$, $\bar{y}$ - are the mathematical means of the x and y series
- N - is the number of elements in the series

The denominator is N for a whole dataset and N - 1 in the case of a sample. As our dataset is a small sample of the entire Iris dataset, we use N - 1.

Covariance and Correlation

- Both covariance and correlation are about the relationship between the variables.
- Covariance defines the *directional association* between the variables.
- Covariance values range from -*inf* to *+inf* where a positive value denotes that both the variables move in the same direction and a negative value denotes that both the variables move in opposite directions.

Covariance and Correlation

- Correlation is a standardized statistical measure that expresses the extent to which two variables are linearly related (meaning how much they change together at a constant rate).
- The *strength and directional association* of the relationship between two variables are defined by correlation and it ranges from -1 to +1.
- Similar to covariance, a positive value denotes that both variables move in the same direction whereas a negative value tells us that they move in opposite directions.

Covariance and Correlation

- Both covariance and correlation are vital tools used in data exploration for feature selection and multivariate analyses.
- For example, an investor looking to spread the risk of a portfolio might look for stocks with a high covariance, as it suggests that their prices move up at the same time.
- However, a similar movement is not enough on its own. The investor would then use the correlation metric to determine how strongly linked those stock prices are to each other.

Lab

- Use corrwith() function to find the correlation among two dataframe objects along the **column axis**

Quantile-Quantile Plots (QQ Plots)

- When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile–quantile plot or qqplot.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

Quartile versus Quantile

- **Quartiles**
  - **First quartile:** Also known as Q1 (the number halfway between the lowest number and the middle number).
  - **Second quartile:** Also known as Q2 or the median (the middle number halfway between the lowest number and the highest number).
  - **Third quartile:** Also known as Q3, or the upper quartile (the number halfway between the middle number and the highest number).

- **Quantile**
  - Are values that split sorted data or a probability distribution into equal parts (In general, q-quantile divides sorted data into q parts).
  - A quartile is a type of quantile.
    - Quartiles (4-quantiles): Three quartiles split the data into four parts.
    - Deciles (10-quantiles): Nine deciles split the data into 10 parts.
    - Percentiles (100-quantiles): 99 percentiles split the data into 100 parts.
  - There is always one fewer quantile than there are parts created by the quantiles.

Quantile-Quantile Plots (QQ Plots)

- For numerical data: visually compare collected data with a known distribution

- Most common one is the Normal QQ plots
  - We check to see whether the sample follows a normal distribution
  - **This is a common assumption in statistical inference that your sample comes from a normal distribution**

- Summary: If your scatterplot “hugs” the line, there is good reason to believe that your data follows the said distribution.

Steps for drawing a Normal QQ plot

1.Collect the Data: Gather the dataset for which you want to create the Q-Q plot. Ensure that the data are numerical and represent a random sample from the population of interest.

2.Sort the Data: Arrange the data in either ascending or descending order. This step is essential for computing quantiles accurately.

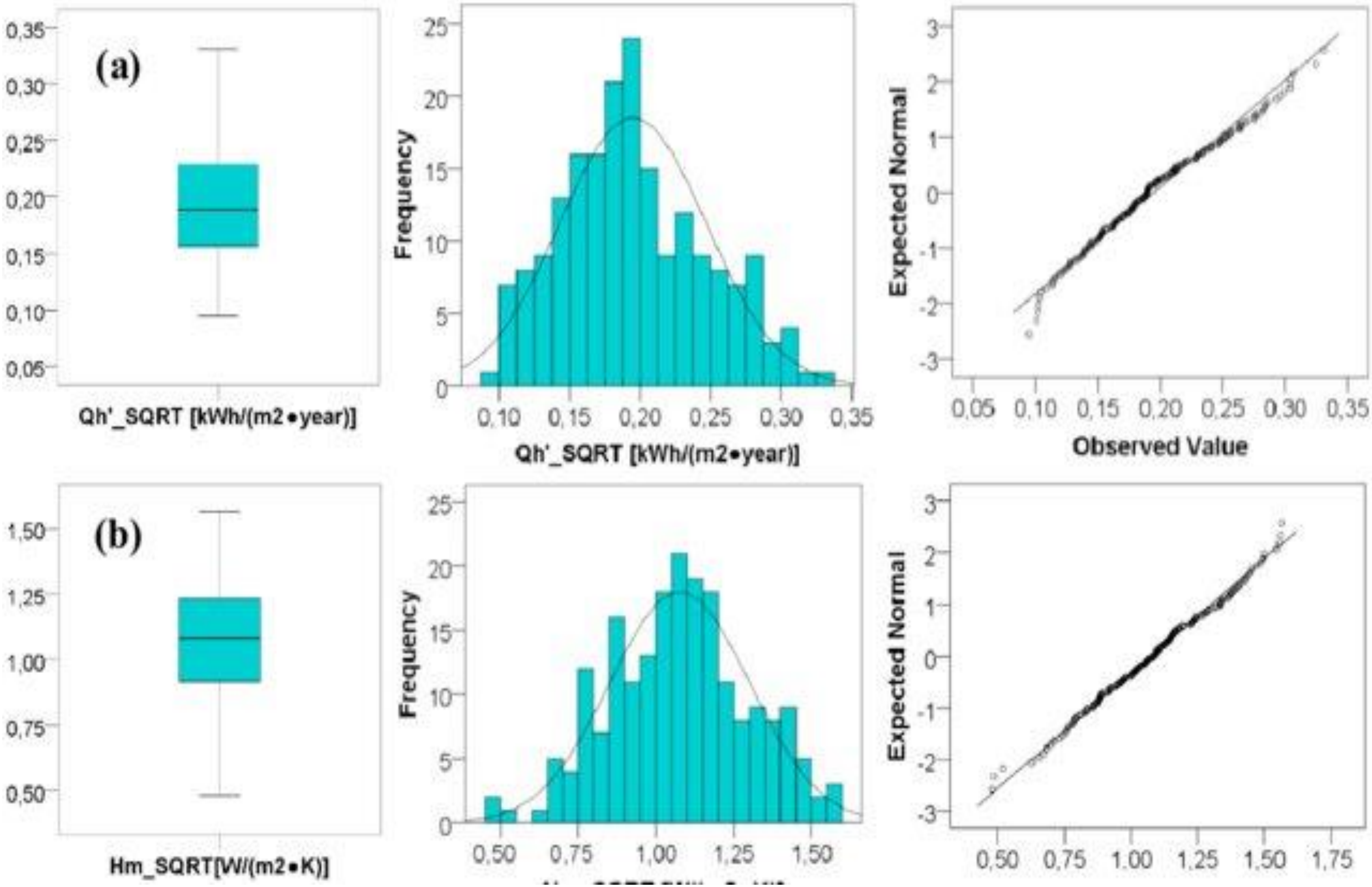
3.Choose a Theoretical Distribution: Determine the theoretical distribution against which you want to compare your dataset. Common choices include the normal distribution, exponential distribution, or any other distribution that fits your data well.

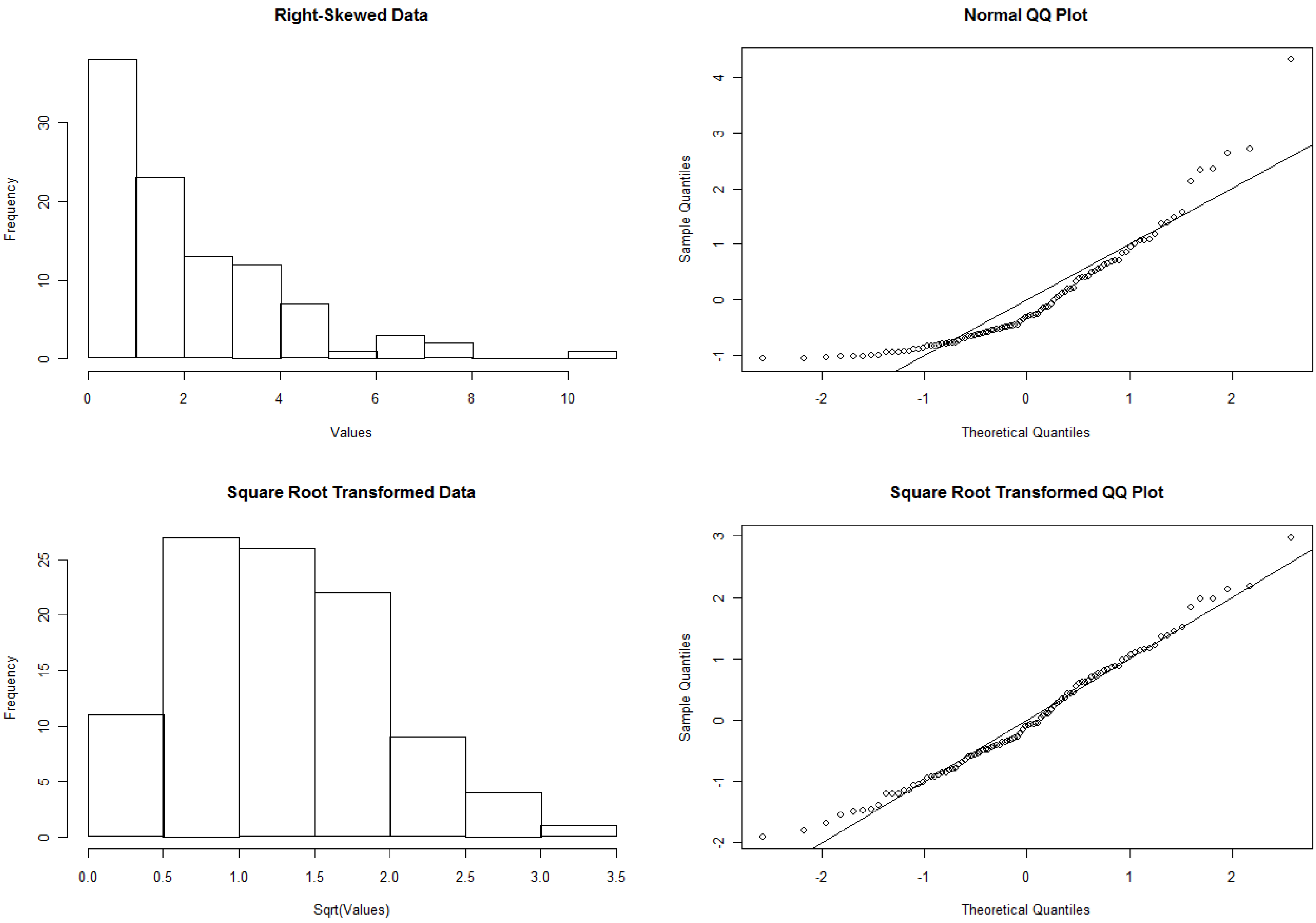
4.Calculate Theoretical Quantiles: Compute the quantiles for the chosen theoretical distribution. For example, if you’re comparing against a normal distribution, you would use the inverse cumulative distribution function (CDF) of the normal distribution to find the expected quantiles.

5.Plotting:

5. Plot the sorted dataset values on the x-axis.
6. Plot the corresponding theoretical quantiles on the y-axis.
7. Each data point (x, y) represents a pair of observed and expected values.
8. Connect the data points to visually inspect the relationship between the dataset and the theoretical distribution.

Box Plot - Histogram - QQ Plots




Right-Skewed Data

Normal QQ Plot

Square Root Transformed Data

Square Root Transformed QQ Plot

Comparing the three visual techniques

- Histograms

- Advantages:
  - With properly-sized bins, histograms can summarize any shape of the data (modes, skew, quantiles, outliers)
- Disadvantages:
  - Difficult to compare side-by-side (takes up too much space in a plot)
  - Depending on the size of the bins, interpretation may be different

- Boxplots

- Advantages:
  - Can identify whether the data came from a certain distribution
  - Don’t have to tweak with “graphical” parameters (i.e. bin size in histograms)
  - Summarize quantiles
- Disadvantages:
  - Difficult to compare side-by-side
  - Difficult to distinguish skews, modes, and outliers

QQ Plots

- Advantages:
  - Don’t have to tweak with “graphical” parameters (i.e. bin size in histograms)
  - Summarize skew, quantiles, and outliers
  - Can compare several measurements side-by-side
- Disadvantages:
  - Cannot distinguish modes!

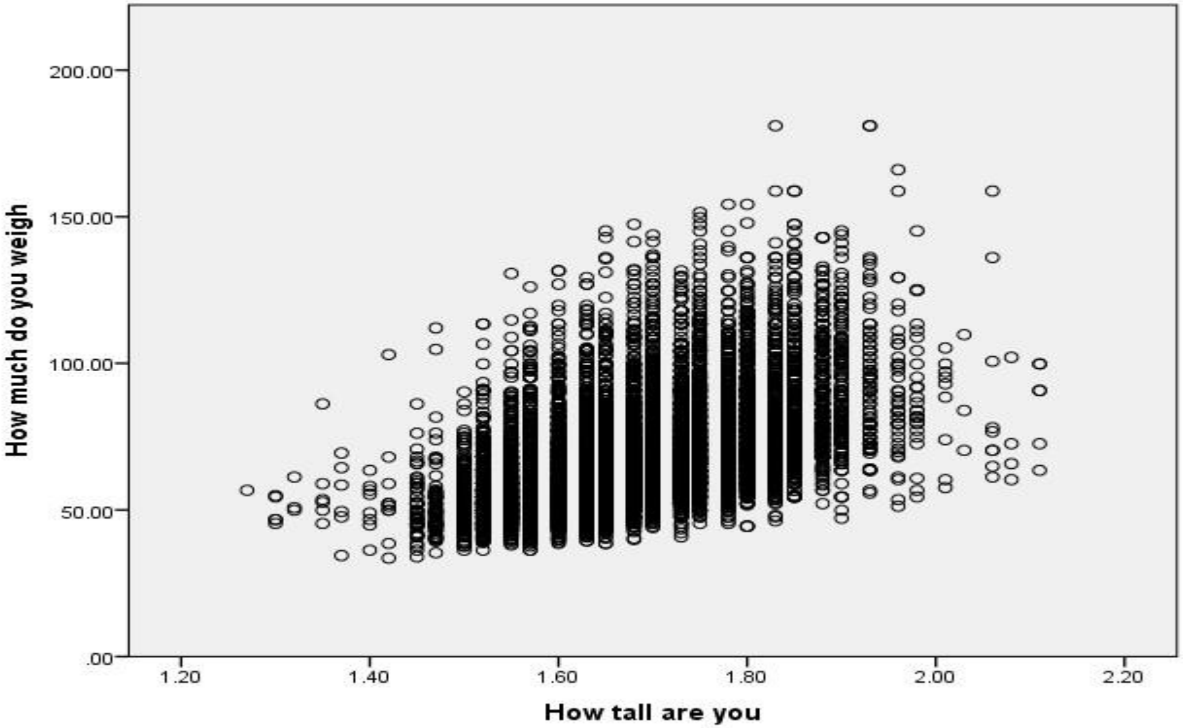
Scatterplots

- For multidimensional, numerical data: $\mathrm{X_i} = (X_{i1}, X_{i2}, \dots, X_{ip})$

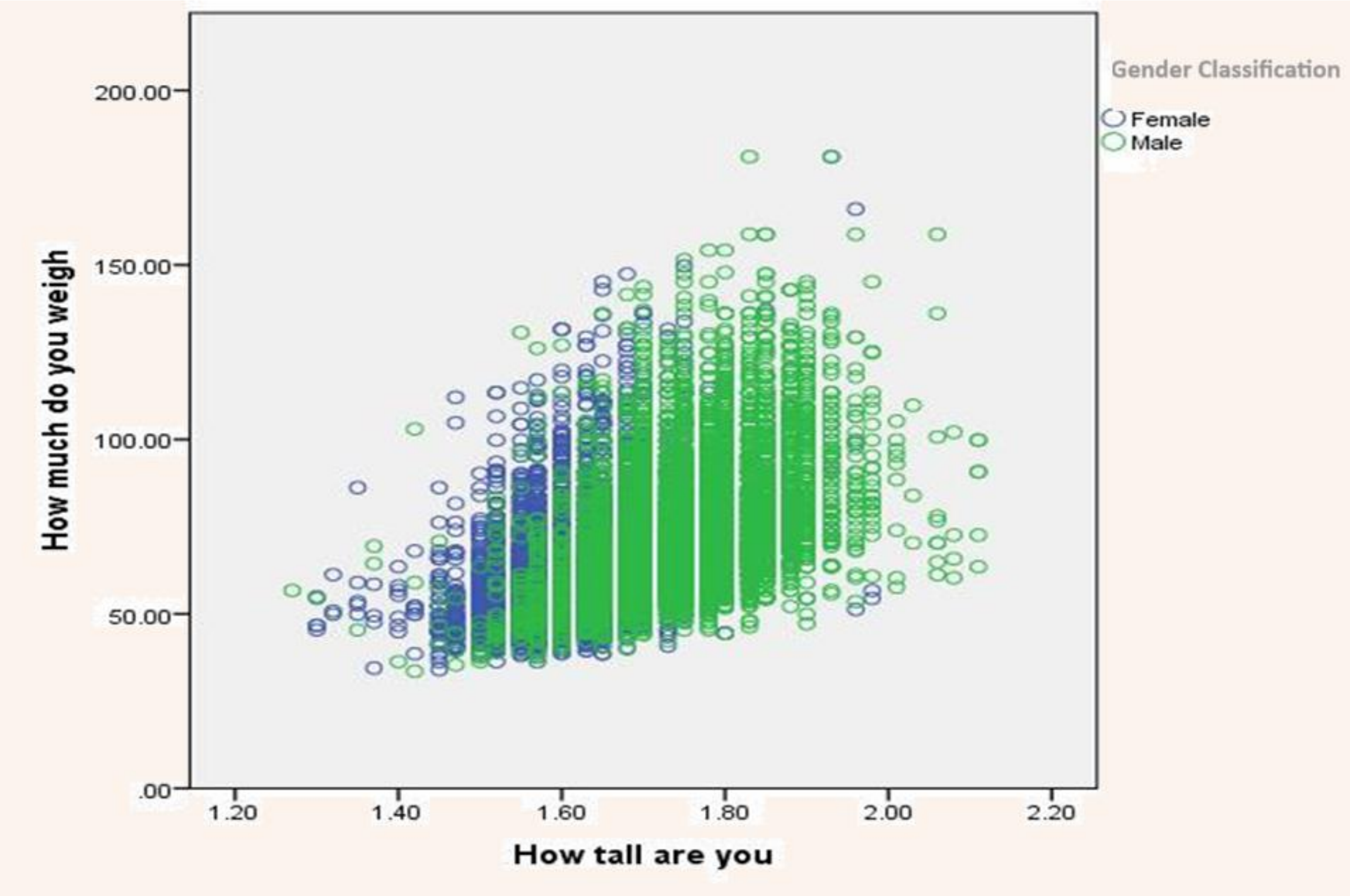
- Plot points on a p dimensional axis

- Characteristics to look for:
  - Clusters
  - General patterns

Scatter Plot



Scatter Plot



Box Plot

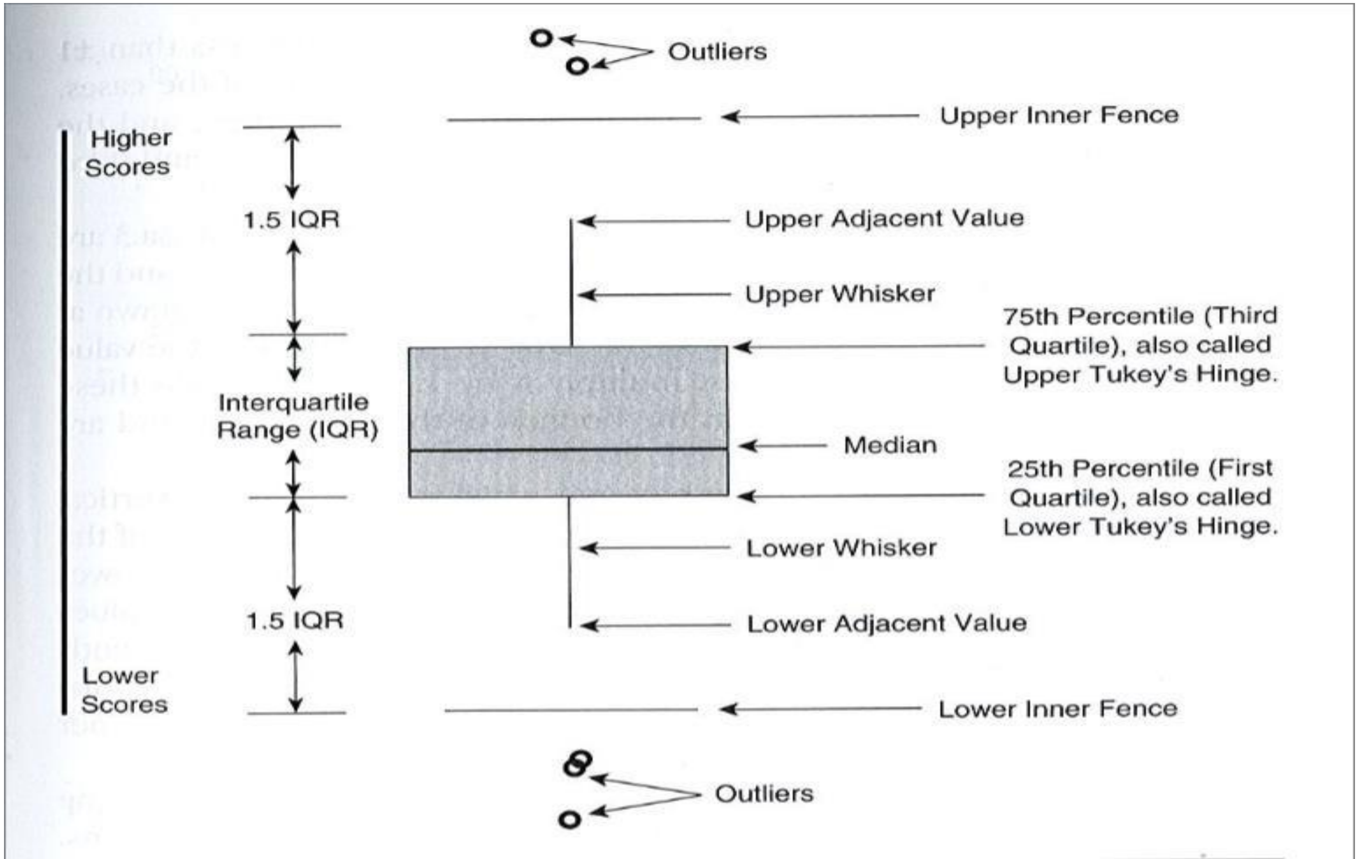
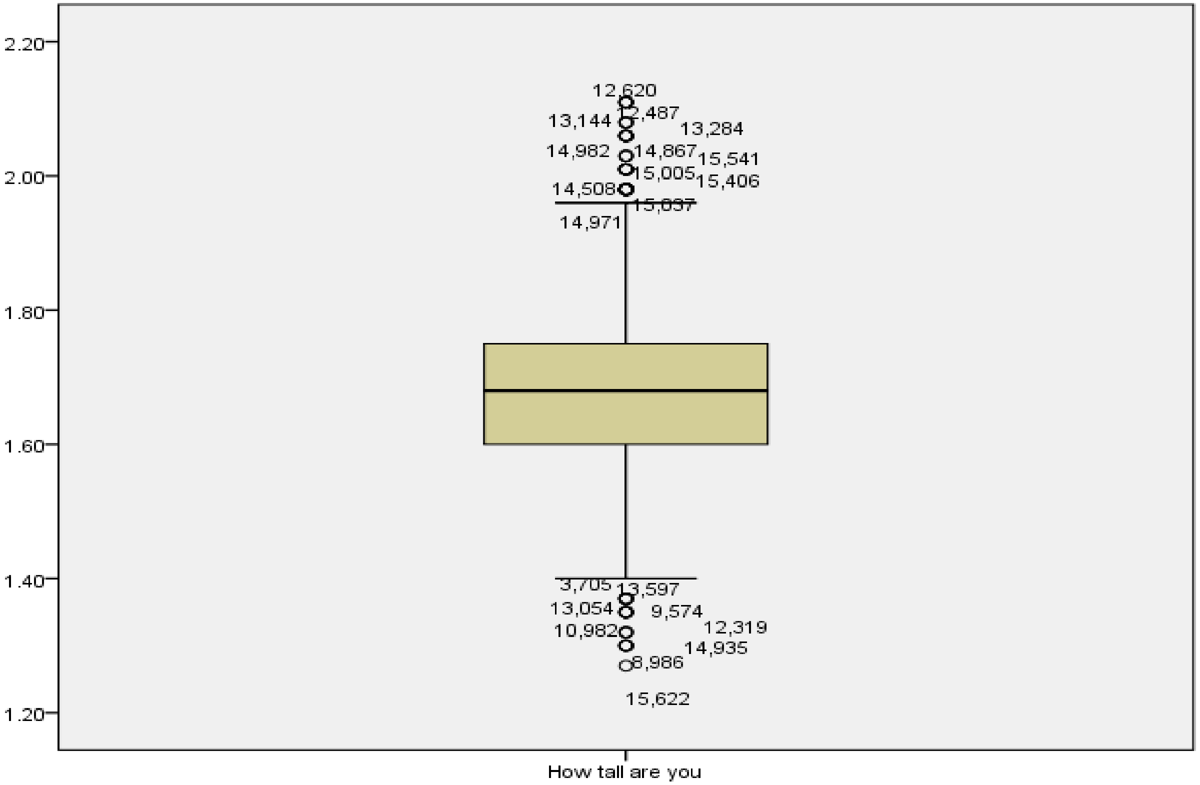


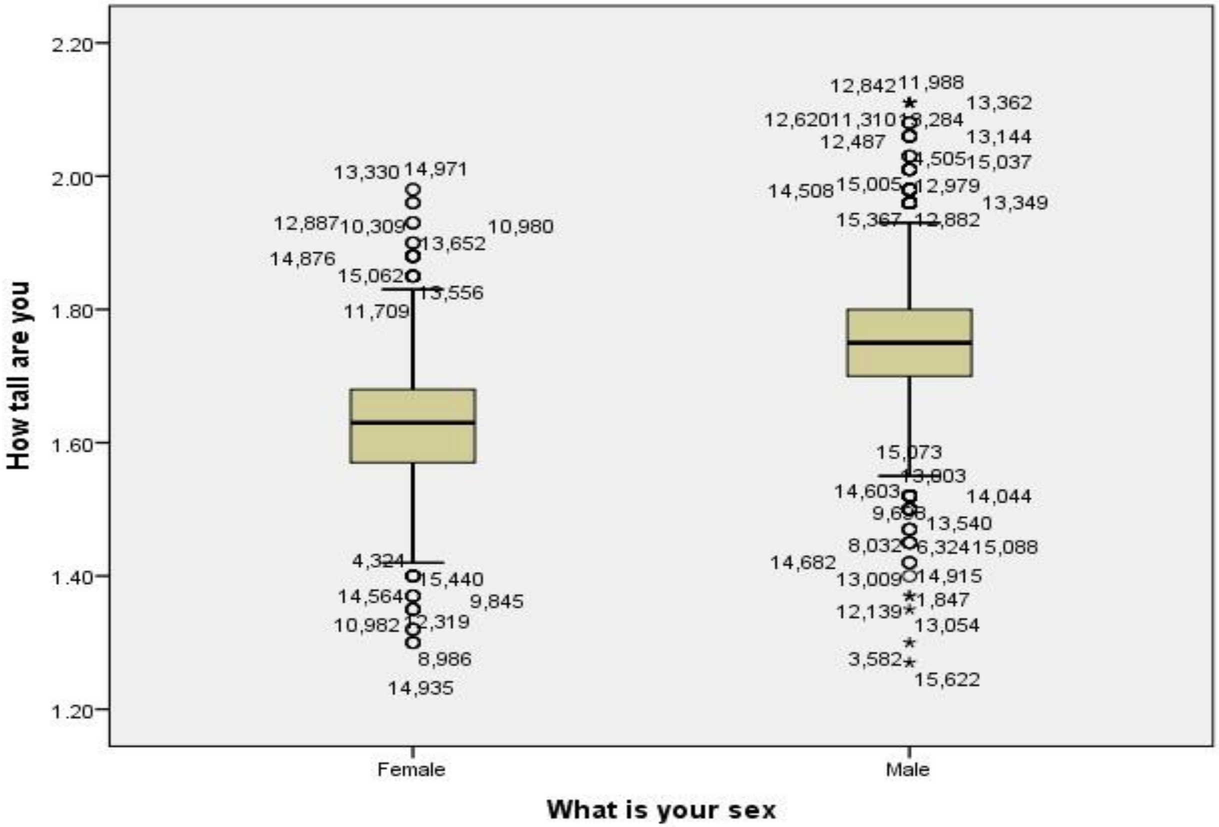
Figure 3a.3 The General Form of a Box and Whiskers Plot Based on Cohen's (1996) Description

Box Plot



2.20
2.00
1.80
1.60
1.40
1.20
12,620
12,487
13,144
13,284
14,982
14,867
15,541
15,005
15,406
14,508
15,637
14,971
3,705
13,597
13,054
9,574
10,982
12,319
14,935
8,986
15,622
How tall are you

Box Plot


2.20
2.00
1.80
1.60
1.40
1.20
How tall are you
13,330
14,971
12,887
10,309
10,980
13,652
14,876
15,062
13,556
11,709
4,324
15,440
14,564
9,845
10,982
12,319
8,986
14,935
12,842
11,988
13,362
12,620
11,310
13,284
12,487
13,144
14,505
15,037
14,508
15,005
12,979
13,349
15,367
12,882
15,073
13,803
14,603
14,044
9,698
13,540
8,032
6,324
15,088
14,682
13,009
14,915
1,847
12,139
13,054
3,582
15,622
Female
Male
What is your sex

Binomial distribution

- Binomial distribution is a type of discrete probability distribution representing probabilities of different values of the binomial random variable (X) in repeated independent N trials in an experiment.

- Thus, in an experiment comprising of tossing a coin 10 times (n), the binomial random variable (number of heads represented as successes) could take the value of 0-10.

- The binomial probability distribution is the probability distribution representing the probabilities of a random variable taking the value of 0-10.

- The probability that a random variable X with binomial distribution B(n,p) is equal to the value k, where k = 0, 1,....,n , is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

- The mean and the variance of the binomial distribution of an experiment with **n** number of trials and the probability of success in each trial is **p** is follows:

- Mean $\mu_X = np$
- Variance $\sigma_X^2 = np(1-p)$

Binomial distribution

- The necessary conditions and criteria to use binomial distributions:
- **Rule 1:** Situation where there are only two possible mutually exclusive outcomes (for example, yes/no survey questions).
- **Rule2:** A fixed number of repeated experiments and trials are conducted (the process must have a clearly defined number of trials).
- **Rule 3:** All trials are identical and independent (identical means every trial must be performed the same way as the others; independent means that the result of one trial does not affect the results of the other subsequent trials).
- **Rule: 4:** The probability of success is the same in every one of the trials.

Examples of binomial distribution problems:

- The number of defective/non-defective products in a production run.
- Yes/No Survey (such as asking 150 people if they watch ABC news).
- Vote counts for a candidate in an election.
- The number of successful sales calls.
- The number of male/female workers in a company

Binomial distribution

- Let’s say that 80% of all business startups in the IT industry report that they generate a profit in their first year. If a sample of 10 new IT business startups is selected, find the probability that exactly seven will generate a profit in their first year.
- First, do we satisfy the conditions of the binomial distribution model?
- There are only two possible mutually exclusive outcomes – to generate a profit in the first year or not (yes or no).
- There are a fixed number of trails (startups) – 10.
- The IT startups are independent and it is reasonable to assume that this is true.
- The probability of success for each startup is 0.8.

Binomial distribution

n = 10, p=0.80, q=0.20, x=7

The probability of 7 IT startups to generate a profit in their first year is:

$$P(x=7)=\frac{10!}{7!(10-7)!}0.80^{7}(1-0.80)^{10-7}$$

This is equivalent to:

$$P(x=7)=\frac{10(9)(8)(7)(6)(5)(4)(3)(2)(1)}{[7(6)(5)(4)(3)(2)(1)]\ [(3)(2)(1]}\ 0.80^{7}(1-0.80)^{10-7}=0.2013$$

Interpretation/solution: There is a 20.13% probability that exactly 7 of 10 IT startups will generate a profit in their first year when the probability of profit in the first year for each startup is 80%.

Poisson distribution

Poisson distribution is actually another probability distribution formula. As per binomial distribution, we won’t be given the number of trials or
the probability of success on a certain trail. The average number of successes will be given in a certain time interval. The average number of successes is called “Lambda” and denoted by the symbol “λ”.

The formula for Poisson Distribution formula is given below:

Poisson Distribution Formula

$$P(x) = \frac{e^{-\lambda} * \lambda^{x}}{x!}$$

Here,
λ is the average number
x is a Poisson random variable.
e is the base of logarithm and e = 2.71828 (approx).

Poisson distribution

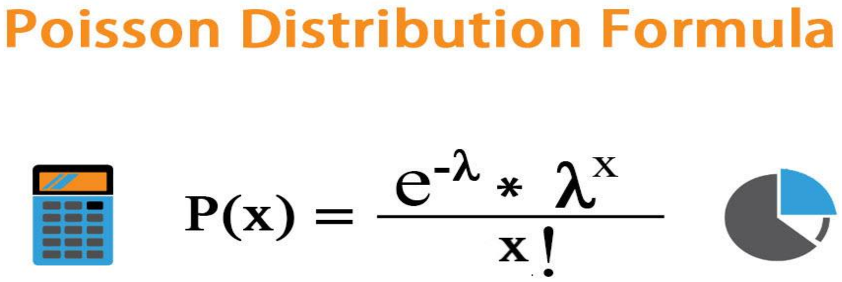
Question: As only 3 students came to attend the class today,
find the probability for exactly 4 students to attend the classes tomorrow.
Solution:
Given,
Average rate of value(λ) =
Poisson random variable(x) =
Poisson distribution = P(X = x) =
P(X=)=
P(X=)=

Poisson Distribution Formula

$$P(x) = \frac{e^{-\lambda} * \lambda^{x}}{x!}$$

Poisson distribution

Question: As only 3 students came to attend the class today,
find the probability for exactly 4 students to attend the classes tomorrow.
Solution:
Given,
Average rate of value(λ) = 3
Poisson random variable(x) = 4
Poisson distribution = P(X = x) =
P(X=4)=e−3·3**4/4!
P(X=4)=0.16803135574154



Poisson Distribution Formula

$$P(x) = \frac{e^{-\lambda} * \lambda^{x}}{x!}$$

Poisson probability distribution

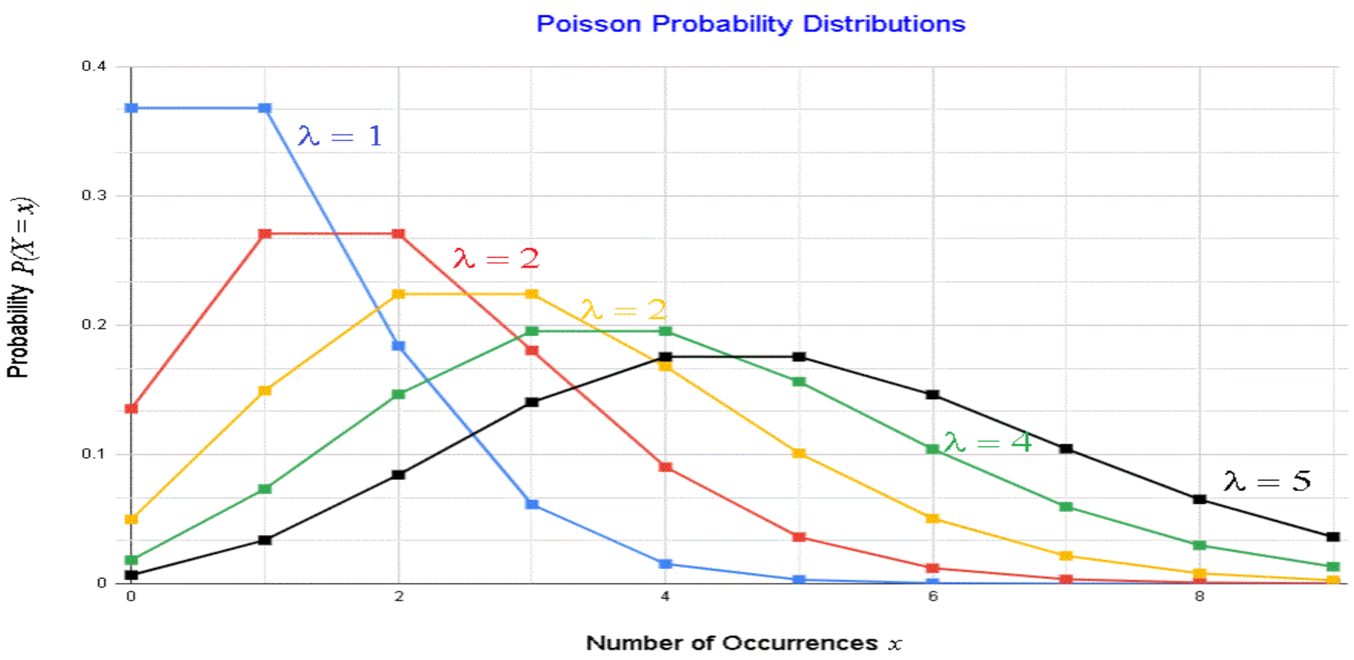
- Poisson probability distribution is used in situations where events occur randomly and independently a number of times on average during an interval of time or space.
- The random variable **XX** associated with a Poisson process is discrete and therefore the Poisson distribution is discrete.
- These are examples of events that may be described as Poisson processes:
  - My computer crashes on average once every 4 months.
  - Hospital emergencies receive on average 5 very serious cases every 24 hours.
  - The number of cars passing through a point, on a small road, is on average 4 cars every 30 minutes.
  - I receive on average 10 e-mails every 2 hours.
  - Customers make on average 10 calls every hour to the customer help center

Poisson distribution

- Conditions for a Poisson distribution are
  - Events are discrete, random and independent of each other.
  - The average number of times of occurrence of the event is constant over the same period of time.
  - Probabilities of occurrence of event over fixed intervals of time are equal.
  - Two events cannot occur at the same time; they are mutually exclusive.

Poisson distribution

the graph of P(X) for several values of the average λ and we note that the probability is maximum for xx close to the average λ and decreases as x takes larger values which makes sense.



Central Limit Theorem

- The Central Limit Theorem states that the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger, no matter what the shape of the data distribution.
- An essential component of the Central Limit Theorem is the average of sample means will be the population mean.
- Similarly, if you find the average of all of the standard deviations in your sample, you will find the actual standard deviation for your population.
- Mean of sample is same as the mean of the population.
- The standard deviation of the sample is equal to the standard deviation of the population divided by the square root of the sample size.
- Central limit theorem is applicable for sufficiently large sample sizes ($n \geq 30$). The formula for central limit theorem can be stated as follows:

Central Limit Theorem

$$\mu_{\overline{x}} = \mu$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Where,
μ = Population mean
σ = Population standard deviation
μx— = Sample mean
σx— = Sample standard deviation
n = Sample size

Example

Question: The record of weights of the male population follows the normal distribution. Its mean and standard deviations are 70 kg and 15 kg respectively. If a researcher considers the records of 50 males, then what would be the mean and standard deviation of the chosen sample?

Solution:
Mean of the population μ = 70 kg
Standard deviation of the population = 15 kg
sample size n = 50
Mean of the sample is given by:

$$\mu_{\overline{x}} = \mu$$

$\mu_{\overline{x}}$ = 70 kg
Standard deviation of the sample is given by:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma_{\overline{x}}$ = 15/√50
$\sigma_{\overline{x}}$ = 2.122 = 2.1 kg (approx)

Confidence Interval

- **Confidence**, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.
- Your desired confidence level is usually one minus the alpha (*a*) value you used in your statistical test:
  - **Confidence level** = 1 − *a*
- So if you use an alpha value of $p < 0.05$ for statistical significance, then your confidence level would be 1 − 0.05 = 0.95, or 95%.

Confidence Interval

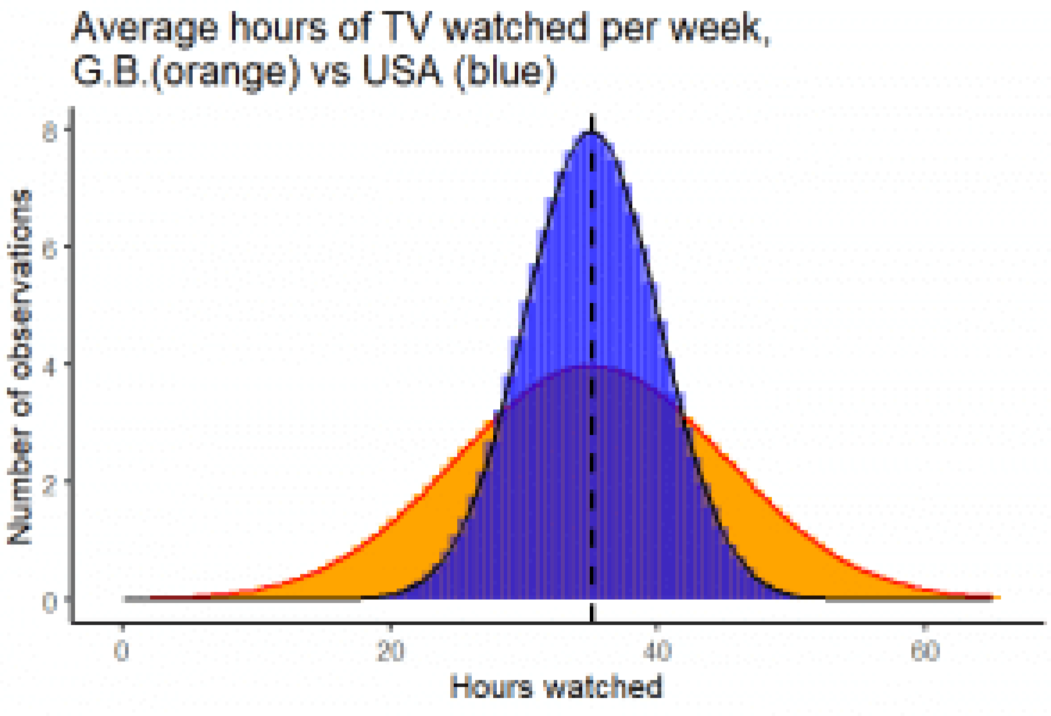
- You can calculate confidence intervals for many kinds of statistical estimates, including:
- Proportions
- Population means
- Differences between population means or proportions
- Estimates of variation among groups

Confidence Interval

- You survey 100 Britsh and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week.
- However, the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.
- Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.

Confidence Interval

- the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.
- Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.



Confidence Interval

A **confidence interval for a mean** is a range of values that is likely to contain a population mean with a certain level of confidence.

It is calculated as:

Confidence Interval = x +/- t*(s/√n)

where:

- **x:** sample mean
- **t:** t-value that corresponds to the confidence level
- **s:** sample standard deviation
- **n:** sample size
- For a two-tailed 95% confidence interval, the alpha value is 0.025, and the corresponding critical value is 1.96.

Confidence level	90%	95%	99%
alpha for one-tailed CI	0.1	0.05	0.01
alpha for two-tailed CI	0.05	0.025	0.005
z-statistic	1.64	1.96	2.57

Example Confidence Interval

In the survey of Americans’ and Brits’ television watching habits, we can use the sample mean, sample standard deviation,
and sample size in place of the population mean, population standard deviation, and population size.
To calculate the 95% confidence interval, we can simply plug the values into the formula.
For the USA:
In the TV-watching example, the point estimate is the mean number of hours watched: 35.
You survey 100 Brits and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week.

$$
\begin{aligned}
CI &= 35 \pm 1.96 \frac{5}{\sqrt{100}} \\
&= 35 \pm 1.96(0.5) \\
&= 35 \pm 0.98
\end{aligned}
$$

So for the USA, the lower and upper bounds of the 95% confidence interval are 34.02 and 35.98.
For GB:

$$
\begin{aligned}
CI &= 35 \pm 1.96 \frac{10}{\sqrt{100}} \\
&= 35 \pm 1.96(1) \\
&= 35 \pm 1.96
\end{aligned}
$$

So for the GB, the lower and upper bounds of the 95% confidence interval are 33.04 and 36.96.