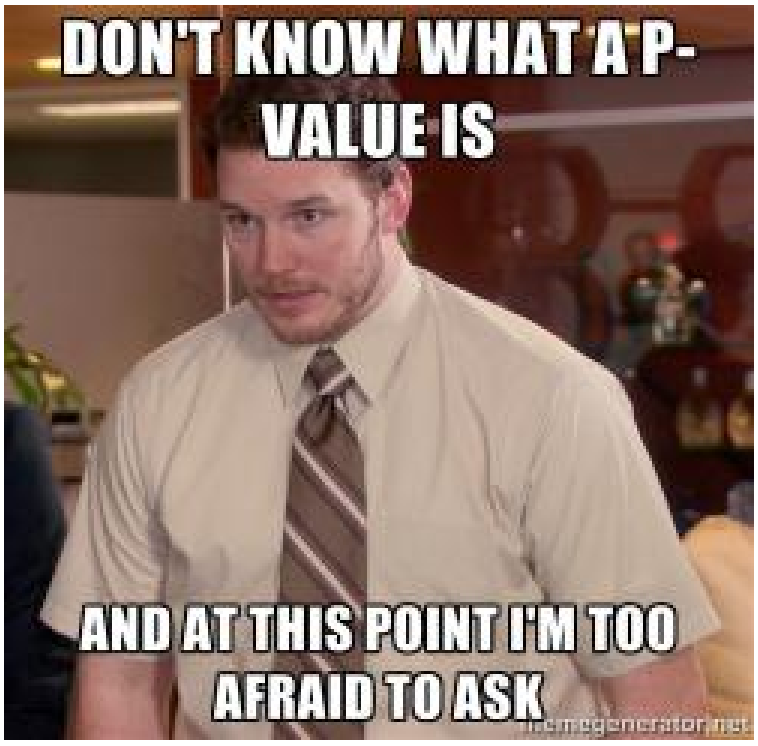
STA258 Master Notes

Haris Aljic

2023-11-27



Contents

Module 1 - "Exploring Categorical Data"

- Categorical variables represent groups/categories like colour/gender/etc
- They can be nominal (non-ordered) or ordinal (ordered)
  - This is a "scale of measure" for the data
- They're graphical displays are most often bar graphs and pie charts

Frequency Distributions

- A "distribution" is a variables *pattern of variation*
- A frequency distribution orders a set of scores from highest to lowest
- Can be modeled as either a table or graph, but regardless, the *same two elements* are presented
  - Set of categories that make up the original measurement scale
  - A record of the frequency, or number of individuals in each category
- So a frequency distribution presents a **picture** of how individual scores/observations are distributed on the measurement scale

Describing Categorical Data

- Frequency (counts)
- Relative frequency (proportion)
  - Count of category divided by total count
- Percentage (proportion times 100)

This information is often *derived from* or *given* in a Frequency Distribution Table (and graphically displayed as either a bar graph or pie chart)

Consider the Titanic Example

```
titanic <- read.csv("Titanic.csv")
attach(titanic)
str(titanic)
```

```
## 'data.frame':    2201 obs. of  4 variables:
##  $ Survived: chr  "Dead" "Dead" "Dead" "Dead" ...
##  $ Age     : chr  "Child" "Child" "Child" "Child" ...
##  $ Sex     : chr  "Male" "Male" "Male" "Male" ...
##  $ Class   : chr  "Third" "Third" "Third" "Third" ...
```

We can see two types of distribution tables here for the "Survived" variable:

```
# See the Frequency Distribution Table for the Survived variable
addmargins(table(Survived))
```

```
## Survived
## Alive  Dead   Sum
##   711  1490  2201
```

This table tells us precisely HOW MANY (frequency/count) people survived or not

```r
# See the RELATIVE Frequency Distribution Table for the Survived variable
addmargins(prop.table(table(Survived)))
```

```
## Survived
##    Alive     Dead      Sum
## 0.323035 0.676965 1.000000
```

This table tells us the relative frequency (proportion) of survivors to the total population

We can extend this further to say that $\approx 32\%$ survived, for example.

Variable Roles

A variable can be classified as **Response (outcome, dependent)**, or **Explanatory (predictor, independent)**

A study on two variable where one is a **response variable**, the other **explanatory**, we observe the outcomes of the response *depends* on the values of the explanatory

Contingency Tables and Marginal Distribution Tables

Consider the Ticket Classes for the Titanic dataset

```
## Class
##   Crew  First Second  Third  Total
##    885    325    285    706   2201
```

And define the contingency table as follows:

```r
Contingency.Table <- table(Survived, Class)
addmargins(Contingency.Table)
```

```
##          Class
## Survived Crew First Second Third  Sum
##    Alive  212   203    118   178  711
##    Dead   673   122    167   528 1490
##    Sum    885   325    285   706 2201
```

(Note this table is still classified as 2x4)

Notice how for any single cell, three different percentages exist.

For example, take the number of First Class passengers that are alive (203):

- Row percentage: $203/711 = 28.6\%$ of survivors in first-class
- Column percentage: $203/325 = 62.5\%$ of first-class passengers survived
- Overall percentage: $203/2201 = 9.2\%$ of passengers were in first-class

Marginal Distributions and Joint Distributions

The sum for each row or for each column give totals. These are the *margins*. In the Ticket Class example, we have two variables - **Survived** (row) and **Ticket Class** (column)

Marginal distributions can be displayed as counts or percentages.

```
## Survived
##    Alive     Dead      Sum
## 0.323035 0.676965 1.000000
```

Similar to Marginal Distributions are Joint Distributions, where each cell of the dataset sums together to the overall total, like so:

```
##         Class
## Survived       Crew      First     Second      Third        Sum
##    Alive 0.09631985 0.09223080 0.05361199 0.08087233 0.32303498
##    Dead  0.30577010 0.05542935 0.07587460 0.23989096 0.67696502
##    Sum   0.40208996 0.14766015 0.12948660 0.32076329 1.00000000
```

Comparing Two Conditional Proportions (Difference and Ratio of Proportions)

The **Difference of Proportions** is simply subtracting one proportion from another to determine which has higher **percentage points**

For example, with the ticket class data set, the first-class passengers who survived (62%), is 37 **percentage points higher** than the third-class passengers (25%)

It is incredibly important to note that the percentage of first-class survivors is NOT 37% higher than the percentage of third-class survivors. It is 37 **percentage points** higher.

The **Ratio of Proportions**, however, tells us much larger or smaller one proportion is to another.

Following the same example from earlier, $0.62/0.25 = 2.48$ implies that the proportion of survived is 2.48 times larger for first-class passengers than for third-class (or that first-class passengers were 2.48 times more likely to have survived than third-class)

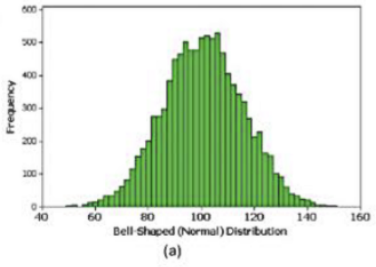
Interpreting Ratios

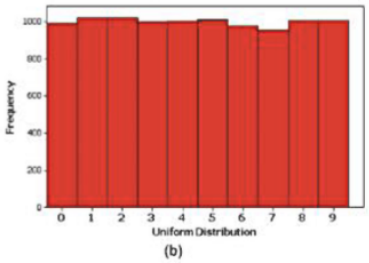
- When proportions are (nearly) identical, their difference is zero, ratio is one, and are said to have **no association** - The greater their difference, the stronger their association.
- Ratio of proportions is often preferred to the difference of proportions when the proportions themselves are both small.
- It is common to speak of *percent increases* when you have a ratio in between 1 and 2
  - For example, a ratio $\frac{a}{b} = 1.09$ implies that proportion a is 1.09 times larger than b, and hence is 9% higher.
- Conversely, a ratio below 1 implies a percent decrease - $\frac{b}{a} = 0.91 \implies 100\% - 91\% = 9\%$ decrease (note this is just the reciprocal of the earlier fraction)

Module 2 - "Exploring Quantitaive Data"

Shapes of Histograms

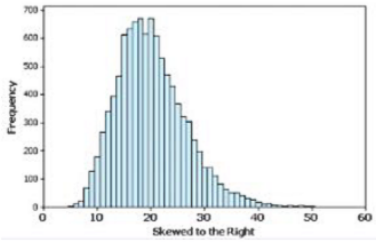
One peak (Unimodal)
Approx. Bell-shaped
and Symmetric.

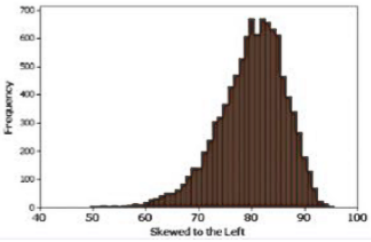




No peak
Uniform Distribution

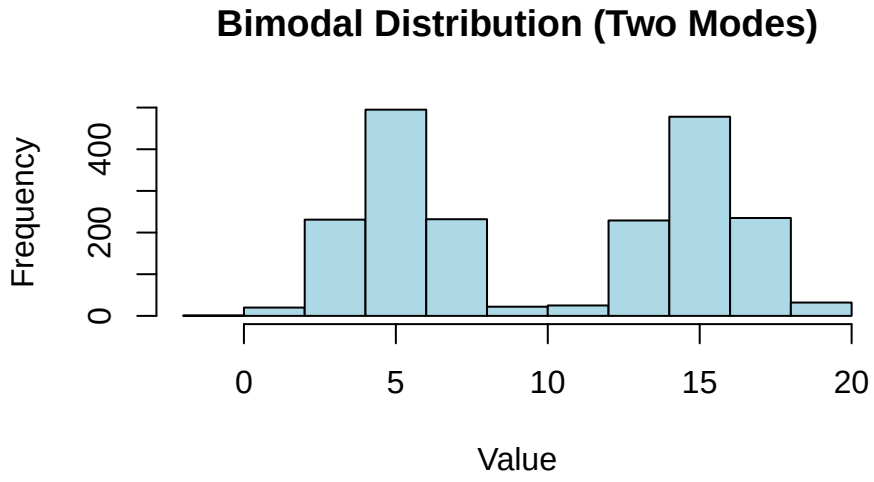
Right Skewed:
Tail to the right;
most of the
observations in the
data are to the left.





Left Skewed:
Tail to the left;
most of the
observations in the
data are to the right.

An easy way to remember, is that the type of skewness corresponds to the tail. For example, if most of the observations are on the right side, that means there's a tail to the left, which means the distribution is **left-tailed** so it is also **left-skewed**.

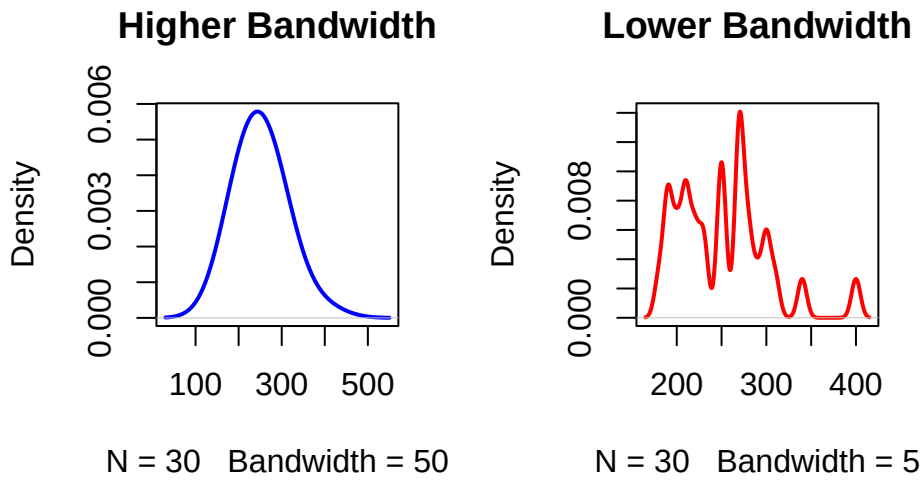


Density Plot

A **Density Plot** is a smooth, continuous function over an interval to visualize the distribution of data. The advantage of density plots over histograms is that they are unaffected by the number of bins (bars) used, and can have a shape or form with lower bin counts that histograms typically would not.

Another detail to consider when discussing density plots is their **bandwidth selection**. Higher bandwidth

smooths out the curve, and will likely leaves out smaller variations. Lower bandwidth, however, shows the much finer details and fluctuations. Take a look below to see the difference bandwidth makes on the *same* dataset:



Centre and Central Tendency

The **centre** is defined to be a value in the data, whereas the **Central Tendency** is a statistical measure (like the mean, median, or mode) used to determine a single score that *defines* the centre of a distribution.

Our goal surrounidng the central tendency is to find a single score that best represents the data of the entire group.

Mean, Median, and Mode

Mean

The mean can be described as a balance point, located in between the lowest and highest points of data. The total distance below the mean is equal to to total distance aboove the mean.

Note: The mean is sensitive to extremely large and small cases.

Formula: Sum all observations from the variable and divide by the total count

We discuss two different types of means - the **population mean** and the **sample mean**

The **population mean** (denoted by μ), is found by adding every single value in an entire population, and dividing it by the total population size

The **sample mean** (denoted by $\bar{x}$), however, is found by adding only a selection of the values from the population, and dividing by the size of said collection. This selection is called a *sample*.

Median

The median is the middle value of the data, when sorted from smallest to largest. It is the midpoint of the list. Median can also be denoted the **50th percentile** - a point on a scale such that 50% of the values are below it, and 50% of the values are above it.

Note: The median is resistant to extremely large and extremely small data points. This is because the median is not found by taking literal data points into account, unlike the mean - **We will report the median for skewed distribution**

Formula:

- Case 1: Odd number of data
  - Order the data from smallest to largest
  - The data point in the $\frac{n+1}{2}$ position (middle) is the median.
- Case 2: Even number of data
  - Order the data from smallest to largest
  - Calculate the average (mean) of the values in the $\frac{n}{2}$ and $\frac{n}{2}+1$ positions (i.e. $\frac{entry\ at(\frac{n}{2})+entry\ at(\frac{n}{2}+1)}{2}$) to find the median.

Note: Do NOT rely on mean and median values to determine the shape of the data. It can be that the mean and median are approximately or nearly equal but data is still skewed.

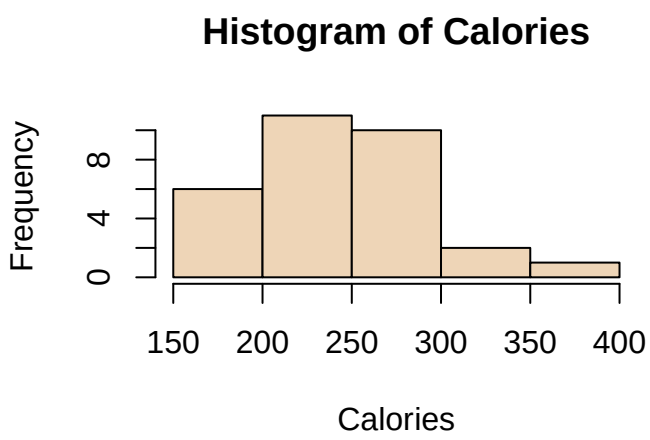
Consider the Tim Hortons data.

```
# Find the Mean, then Median
mean(Calories)
```

```
## [1] 251
```

```
median(Calories)
```

```
## [1] 250
```



The histogram is right skewed, despite the mean (251) and median (250) are very close in value.

Interpretation of the mean and median for calorie counts in donuts:

- Mean: On average, the donuts have 251 calories.
- Median: Half of the donuts have at least 250 calories, where as the other half have at most 250 calories

Mode

The mode of a variable refers to the most frequently appearing data point. In a frequency distribution graph, the modes are the peaks of the graph (plural because there may be multiple modes - two or more different values could each share the same largest frequency)

Items of note:

- Symmetrical Distributions
  - The mean and median are equal and exactly at the centre
  - If the distribution is only approximately symmetric, the mean and median are clsoe to each other
  - Unimodal distributions (one mode) with mean = median = mode implies the distribution is bell-shaped symmetric
  - Bimodal distributions (two modes) can still be symmetric
- Skewed Distributions
  - Occurs when there exists an observation(s) that deviate from the general pattern of data.
  - This effects where you can find the mean, median, and mode.
    - Left/Negative skewed: Mean first, then median, then mode
    - Right/Positive skewed: Mode first, then median, then mean
  - In other words:
    - Mean $<$ Median $\implies$ left skewed
    - Median $<$ Mean $\implies$ right skewed
  - In general, just looking at the values is not always good enough to determine the shape of the data (for example, the distribution could just be bimodal)

Spread - Sample Variance (and Sample Standard Deviation)

The spread of a set of data tells us how much data varies around its centre. In other words, how far from the mean/median do observations tend to be?

The **Variability** describes the distribution. It is a quantitative measure of the differences between observations, describing **how** spread out or **how** grouped together these data points are. It measures how well either a single score, or group of scores represents the entire distribution, as well as how much *error* to expect when using a sample to represent a population.

Describing a data set numerically typically requires a report of its spread, and the centre.

The variance will help us (indirectly) determine the spread of the data in a sample.

Sample Variance

Sample variance measures the spread about the sample mean, $\bar{x}$

Formula: $S^2 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Taking the square root leaves us with **sample standard deviation**, which is the value that will tell us roughly on average how much values differ from the mean.

Here:

- $x_i = i^{th}$ observation in the data
- $\bar{x}$ = sample mean
- n = sample size

Note: The sample variance has (n - 1) degrees of freedom since one value out of the sample is dependent on all the others to determine the sample variance - (n - 1) values are "free".

Population Variance

Formula: $\sigma^2 = \frac{\Sigma_{i=1}^{N}(x_i - \mu)^2}{N}$

Taking the square root leaves us with **population standard deviation**, which is the value that will tell us roughly on average how much values differ from the mean.

Here:

- $x_i = i^{th}$ observation in the data
- μ = Population mean
- n = Population size

Note: Unlike sample variance, population variance divides by the total population size, N.

It it important to see that the standard deviation is always greater than or equal to zero. $(\sigma, S \geq 0)$, and the *larger* the standard deviation, the *greater* the variability.

(Standard deviation also rescales when the data is rescaled)

Percentiles, Quartiles, and Interquartile Range

Distributions can also be described with a measure of position. For example, **range = max data point - min data point**. However, some measures describe **centre**, and some **variability**.

Percentiles

The p^{th} percentile is a point in a data set where $p\%$ of the observations fall under it (or equivalently, $(100 - p)\%$ fall above it).

For example, consider the 50^{th} percentile. This is equivalent to the *median*, since half the values are above it, and the other half below.

We should familiarize ourselves with some common terminology surrounding percentiles before proceeding further.

Quartile corresponds to a quarter of the data, so it is only used when discussing 0, 25, 50, 75, and 100 percent of the data.

Quantile is a decimal value between zero and one used as a decimal representation for the percentage p

Percentile is how we defined our p, to be the point where $p\%$ of the observations fall under it

For example, let p = 25, then:

- "0.25 quantile" = "25^{th} percentile" = "First quartile"/"1 quartile/Q1"

Now let p = 33, then:

- "0.33 quantile" = "33^{rd} percentile" = "quantile not defined for this value"

For the remainder of these notes, I will try to primarily refer to first, second, and third quartiles as Q1, Q2, and Q3 respectively.

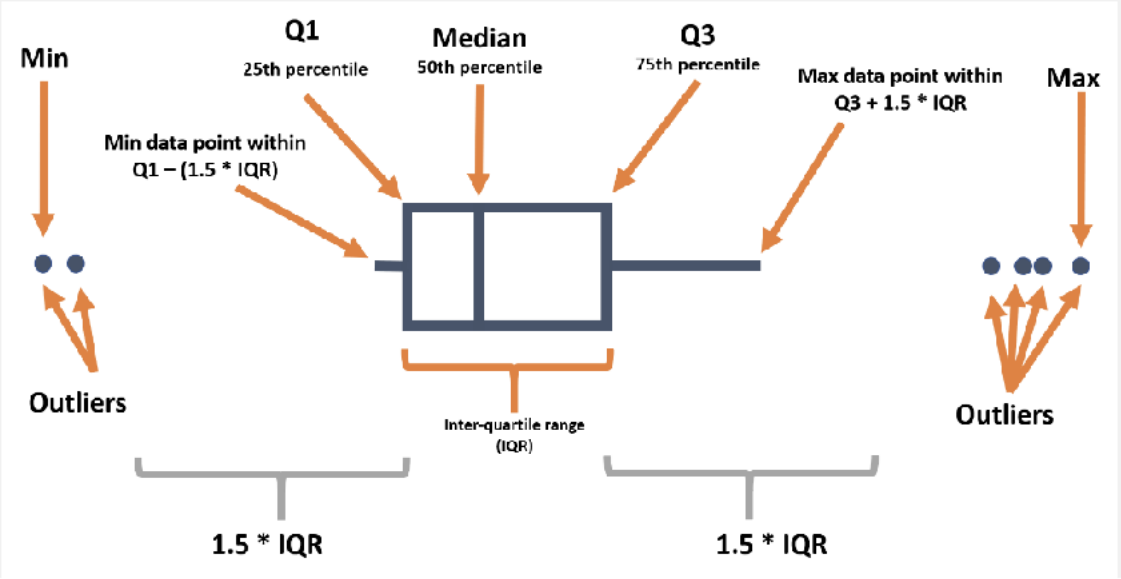
Special Definition: (Sample Percentile) * The i^{th} smallest observation in an ordered list is called the **sample** $[\frac{100(i-0.5)}{n}]$**th percentile**

Interquartile Range (IQR)

The IQR is defined to be the middle half of the data, where 50% of the data falls.

Since Q1 contains the first 25% of the data below it, and Q3 contains the last 25% above it, the IQR is simply Q3 - Q1. In other words, the 75^{th} percentile - 25^{th} percentile. The IQR is less affected by outliers (sometimes it's not affected at all).

See the below illustration for how the IQR works with boxplots



For boxplots, we determine shape based on the position of the median (centre line in the box).

- If the line is further to the **right/top**, it is **left skewed**.
- Conversely, if the line is further to the **left/bottom**, it is **right skewed**.
  - You can think of it as "the tyoe of skewness corresponds to the side with more space".
  - In the above illustration, the data would have a *right skew*.
- Otherwise, a median line in the middle of the box implies the data is approximately symmetric.

Centre is reported as the position of the median line in the box.

Spread is reported in terms of the range of the box - ie, Q3 - Q1.

Z-Scores

A **Z-Score** refers to a **standardized data value**, representing how far an observation is away from the mean **in terms of standard deviations**

$Z = \frac{observation - mean}{standard\ deviation} = \frac{x-\mu}{\sigma}$

Note this example uses population mean and standard deviation. For a sample, simply swap μ for $\bar{x}$ and σ for S.

- If your Z-score is positive, the observation lies **above** the mean
- If your Z-score is negative, the observation lies **below** the mean
- A data point is more unusual in the collection the greater the Z-score is in magnitude
  - it is very rare for a bell-shaped symmetric distribution to have values falling more than 3 standard deviations above or below the mean

Standardizing into Z-scores shifts the data to have the mean centred at zero, and rescales for the standard deviation to become 1. This does not change the shape of the distribution.

Empirical Rule

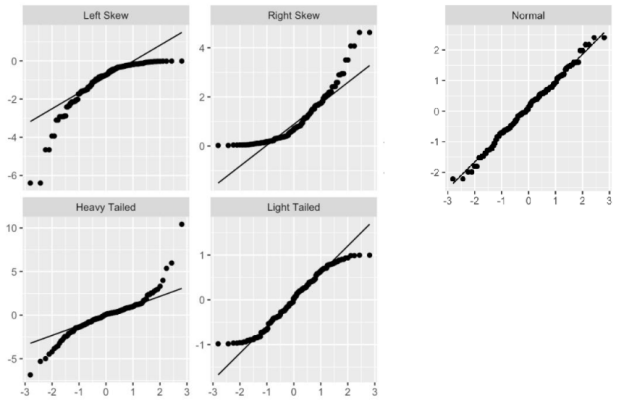
The empirical rule states that for distributions that are approximately bell-shaped symmetric, without outliers, certain fixed (approximate) percentages of the sample fall within certain ranges of the graph.

- $\approx 68\%$ of the data falls within one SD (standard deviation) of the mean - $\mu \pm \sigma$
- $\approx 95\%$ of the data falls within two SD's of the mean - $\mu \pm 2\sigma$
- $\approx 99.7\%$ of the data falls within three SD's of the mean - $\mu \pm 3\sigma$

Different Plots and how they relate to Normal Distributions

You can tell if a boxplot is bell-shaped symmetric if $\frac{IQR}{S} \approx 1.33$

Below are some probability plots for different shapes of distributions:



Module 3 - "Sampling Distributions Related to a Normal Population"

Facts about the Normal Distribution ($X \sim N(\mu, \sigma)$)

- They are bell-shaped symmetric, smooth curves (this is not unique to normal distributions)
- The mean is located at the centre of the curve
- The mean, median, and mode are equal
- The standard deviation control the spread of the curve
- It models most real-world distributions

From a more mathematical perspective

- The maximum is located at $x = \mu$
- Inflection points occur at $x = \mu \pm \sigma$
- It is symmetric about $x = \mu$
- The x-axis is a horizontal asymptote

Moments

Let Y be a random variable.

The Expected Value ($E[Y]$) is called the "first moment" of Y, $E[Y^2]$ is the second moment, and so on.

Generalized, $E[Y^k]$ is the k^{th} moment of Y.

Moreover, $E[(Y - E[Y])] = 0$ is defined as the "first **central** moment" of Y, and $E[(Y - E[Y])^2]$ as the second central moment.

Take note that that second central moment is equivalent to the definition of the variance for the random variable Y: $Var(Y) = E[(Y - E[Y])^2]$

Module 4

Module 5

Module 6

Module 7

Module 8

Module 9

Module 10

Module 11