Unit-II
Statistical Concepts
Continued

Measures of Shape

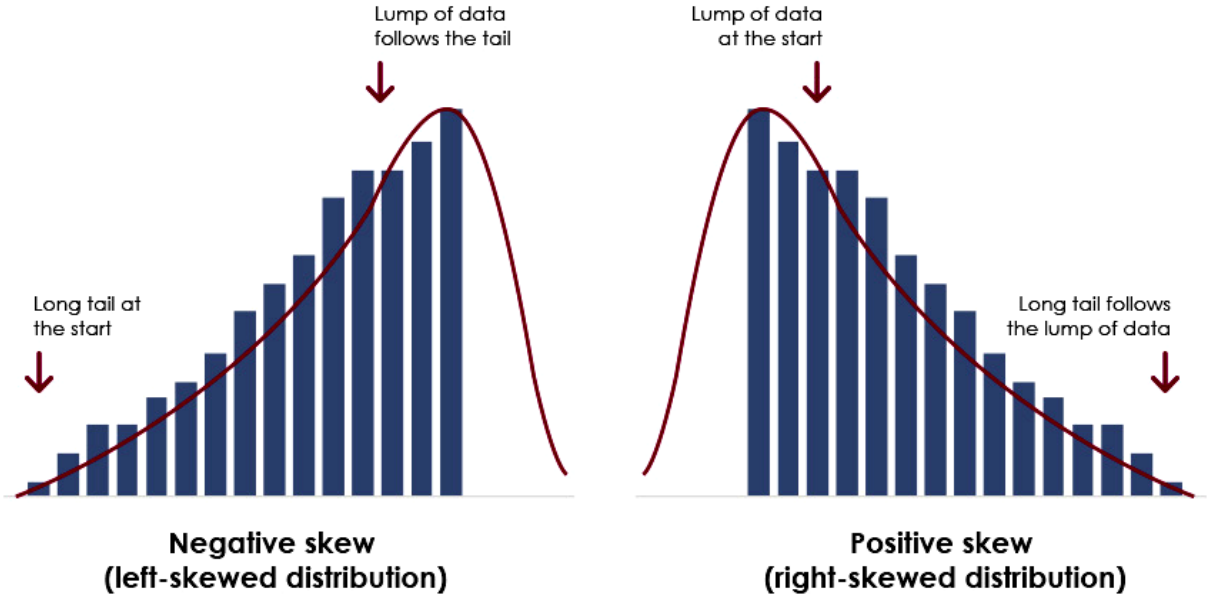
Measures of Shape

There are two final measures of a distribution you will hear occasionally:

- **skewness** and
- **kurtosis**.

Skewness

- Skewness refers to the **degree of symmetry,** or more precisely, the degree of lack of symmetry.
- It is the measure of asymmetry that occurs when our data deviates from the norm.



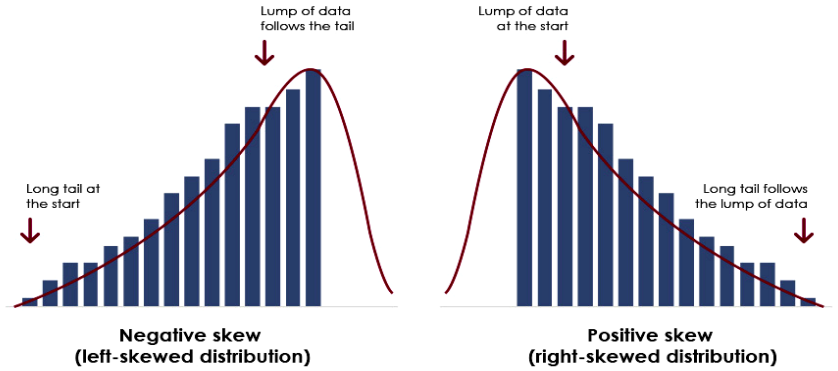
Measures of Shape (cont..) Skewness

How these **central tendency measures tend to spread when the normal distribution** is **distorted**.

For the nomenclature just follow the direction of the tail

— For the **right graph** has the **tail to the right**, so it is **right-skewed (positively skewed)** and

— For the **left graph** since the **tail is to the left, it is left-skewed (negatively skewed).**



Pearson's Coefficient of Skewness : This method is most frequently used for measuring skewness.

The formula for measuring coefficient of skewness is given by

$$\text{Pearson's First Coefficient} = \frac{Mean - Mode}{Standard\ deviation}$$

$$\text{Skewness} = \frac{\sum (x - \bar{x})^3}{(n-1) \cdot S^3}$$

Where:

S: standard deviation

$\bar{X}$: Mean

The value of this coefficient would be **zero** in a ***symmetrical distribution***. If **mean** is **greater** than **mode**, coefficient of skewness would be **positive** otherwise **negative**. The value of the Pearson's coefficient of skewness usually lies between ±1 for moderately skewed distribution.

Measures of Shape (cont..)

Skewness

$$\text{Pearson's First Coefficient} = \frac{Mean - Mode}{Standard\ deviation}$$

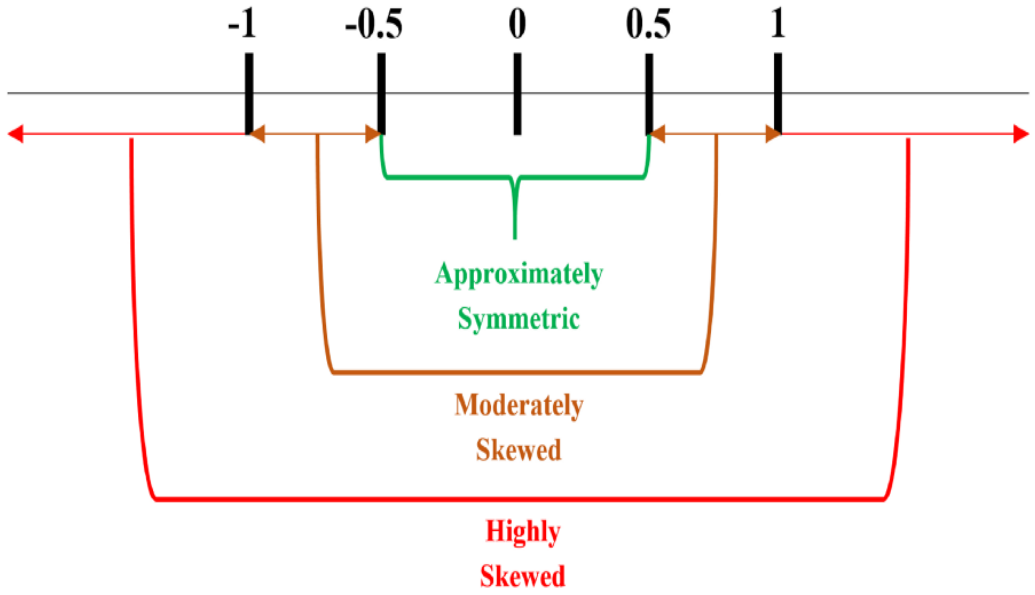
If mode is not well defined, we use the formula

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$

Substituting this in Pearson's first coefficient gives us Pearson's **second coefficient** and the formula for skewness:

$$\text{Pearson's Second Coefficient} = \frac{3(Mean - Median)}{Standard\ deviation}$$

Scale of Skewness:



(-0.5, 0.5) = Low

(-1, -0.5) U (0.5, 1) = Moderate

(-1 & beyond) U (1 & beyond) = High

Measures of Shape (cont..)

What Is Kurtosis?

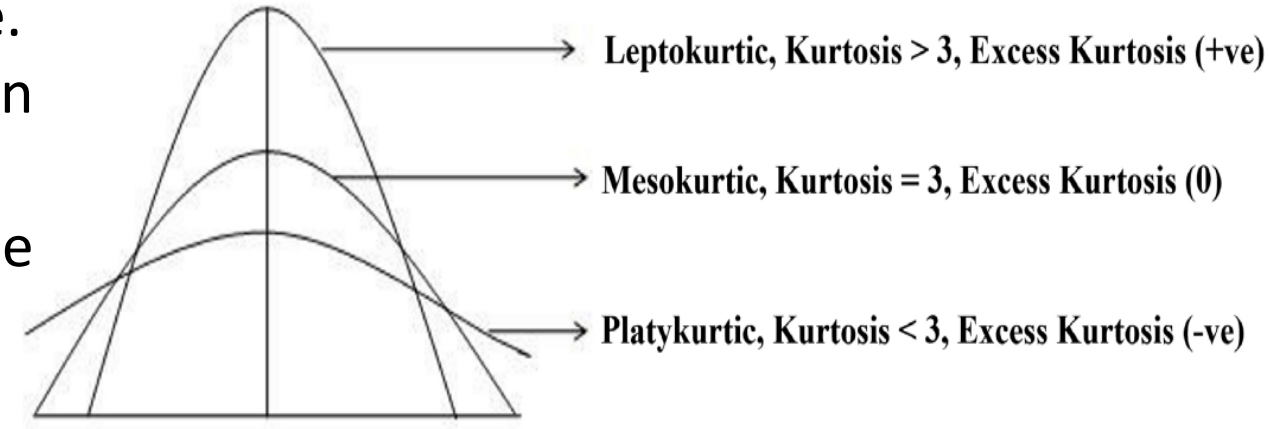
- Kurtosis gives a **measure of flatness** of distribution.
- We need to know another **measure to get the complete idea about the shape** of the distribution which can be studied with the **help of Kurtosis**.
- Kurtosis is associated with the ***"movement of probability mass from the shoulders of a distribution into its center and tails."***
- The degree of kurtosis of a distribution is measured relative to that of a normal curve.
- The curves with **greater peakedness** than the normal curve are called ***"Leptokurtic"***.
- The curves which are **more flat than** the normal curve are called ***"Platykurtic"***.
- The **normal curve** is called ***"Mesokurtic."***

For sample:

$$\text{Kurtosis} = \frac{\sum (x - \bar{x})^4}{(n - 1) \cdot S^4}$$

Where:

S: standard deviation $\quad \bar{X}$: Mean



Measures of Sample Skewness and Kurtosis

Examples: Calculate Sample Skewness and Sample Kurtosis from the following grouped data

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

Solution:

Classes	Mid value (x)	f	$f \cdot x$	$(x - \bar{x})$	$f \cdot (x - \bar{x})^2$	$f \cdot (x - \bar{x})^3$	$f \cdot (x - \bar{x})^4$
2 - 4	3	3	3×3= 9	3-5.2=-2.2	3×-2.2×-2.2=14.52	14.52×-2.2= -31.944	70.27
4 - 6	5	4	4×5= 20	5-5.2=-0.2	4×-0.2×-0.2=0.16	0.16×-0.2= -0.032	0.0064
6 - 8	7	2	2×7= 14	7-5.2=1.8	2×1.8×1.8=6.48	6.48×1.8=11.664	20.98
8 - 10	9	1	1×9= 9	9-5.2=3.8	1×3.8×3.8=14.44	14.44×3.8= 54.872	208.5
---	---	---	---	---	---	---	
-TOTAL-	--	n=10	$\sum f \cdot x$=52	`--	=35.6	=34.56	=299. 79

$$\text{Mean} = \sum \boldsymbol{f} \cdot \boldsymbol{x} = \frac{\sum f \cdot x}{\sum f} = \frac{52}{10} = 5.2$$

Standard deviation (S.D)

$$S.D = \sqrt{\frac{\sum_{i=1}^{n} fi(Xi - \bar{X})^2}{\sum_{i=1}^{n} fi}}$$

$$S.D = \sqrt{\frac{35.6}{10}} = 1.88$$

Calculate the Skewness

$$\text{Skewness} = \frac{\sum (x - \bar{x})^3}{(n - 1) \cdot S^3}$$

$$\text{Skewness} = \frac{34.56}{9*(1.88)^3} = 0.48$$

Calculate the Kurtosis:

$$\text{Kurtosis} = \frac{\sum (x - \bar{x})^4}{(n - 1) \cdot S^4}$$

$$\text{Kurtosis} = \frac{299.79}{9*(1.88)^4} = 2.12$$

Outliers and Missing values

Outliers

- An **outlier** is a value or an entire observation (row) that lies well outside of the norm.
- Even if values are not unusual by themselves, there still might be unusual *combinations* of values.
- When dealing with outliers, it is best to run the analyses two ways: with the outliers and without them.
- Let's just agree to define outliers as ***extreme*** values, and then for any particular data set, you can decide how ***extreme*** a value needs to be to qualify as an outlier.

For example, let us consider a row of data [10,15,22,330,30,45,60]. In this dataset, we can easily conclude that 330 is way off from the rest of the values in the dataset, thus 330 is an outlier. It was easy to figure out the outlier in such a small dataset, but when the dataset is huge, we need various methods to determine whether a certain value is an outlier or necessary information.

Outliers (cont..) Types of outliers :

There are three types of outliers

- **Global Outliers:** *The data point or points whose values are **far outside everything else** in the dataset are global outliers. Suppose we look at a taxi service company's number of rides every day. The **rides suddenly dropped to zero due to the pandemic-induced lockdown.** This sudden decrease in the number is a global outlier for the taxi company.*

- **Collective Outliers:** *Some data points collectively as a whole deviates from the dataset. These data points individually may not be a global or contextual outlier, but they behave as **outliers when aggregated together.** For example, **closing all shops in a neighborhood is a collective outlier** as individual shops keep on opening and closing, but all shops together never close down; hence, this scenario will be considered a collective outlier.*

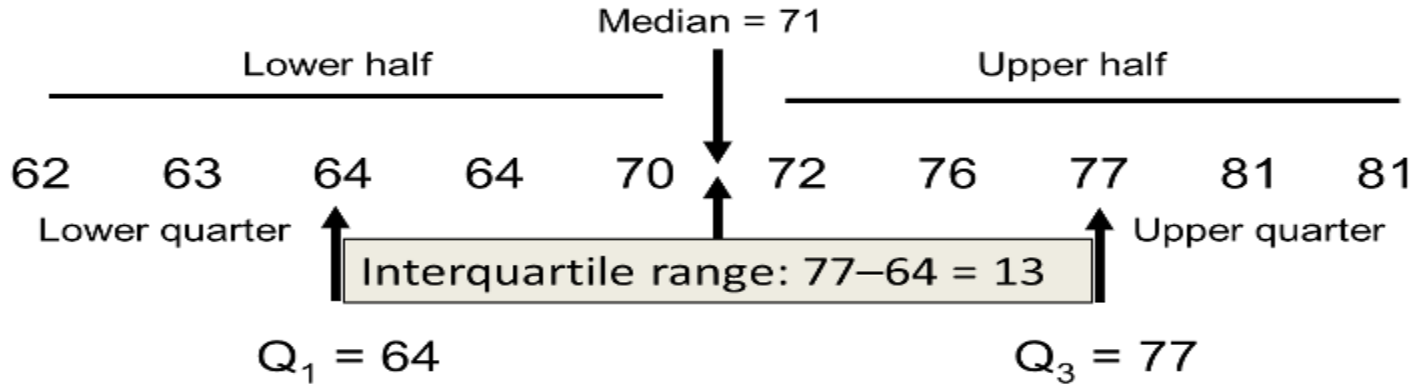
- **Contextual Outliers:** *Contextual outliers are those values of data points that deviate quite a lot from the rest of the data points that are in the same context, however, in a **different context**, it may not be an outlier at all. For example, **a sudden surge in orders for an e-commerce site at night** can be a contextual outlier.*

Outliers can lead **to vague or misleading predictions** while using machine learning models. Specific models like linear regression, logistic regression, and support vector machines are susceptible to outliers. Outliers decrease the mathematical power of these models, and thus the output of the models becomes unreliable.

Outliers (cont..): 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.
- Outliers are values **below Q1-1.5(Q3-Q1)** or **above Q3+1.5(Q3-Q1)** or equivalently, values **below Q1-1.5 IQR** or above **Q3+1.5 IQR**.

- In previous example, for the diastolic blood pressures, the lower limit is 64 - 1.5(77-64) = 44.5 and the upper limit is 77 + 1.5(77-64) = 96.5. The diastolic blood pressures range from 62 to 81. Therefore there are no outliers.

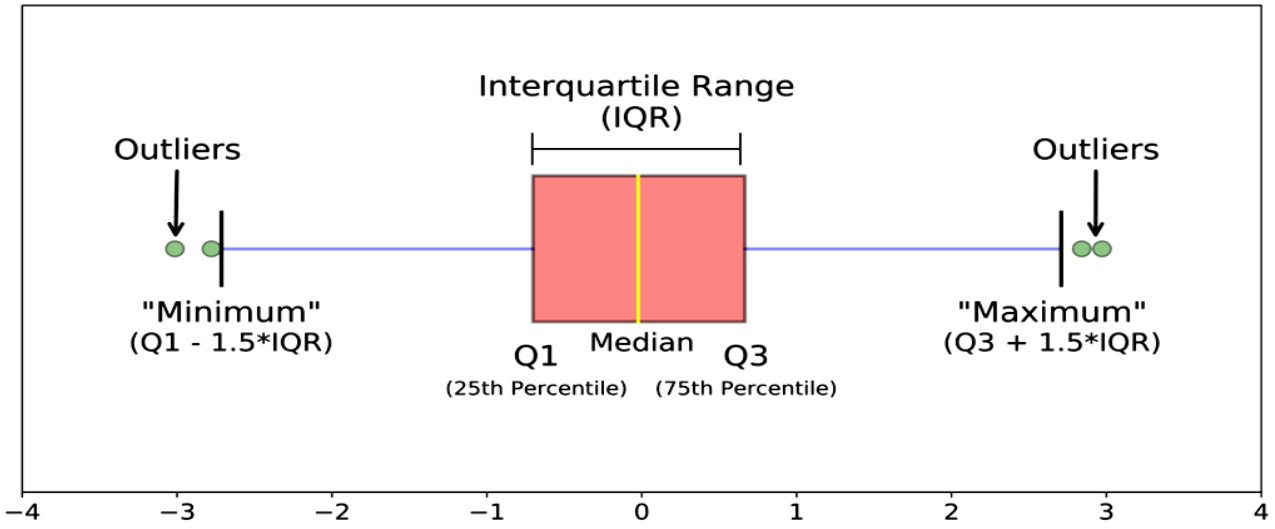


Outliers (cont..)

Boxplot Analysis

Box plots are a simple way to **visualize data through quantiles** and detect outliers. **A boxplot incorporates the five-number summary as follows: Minimum, Q1, Median, Q3, Maximum**

- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the **first** and **third quartiles**, i.e., the height of the box is IQR. IQR(Interquartile Range) is the basic mathematics behind boxplots.
  - The median is marked by a line within the box
  - Two lines (called **whiskers**) outside the box extend to the smallest (Minimum) and largest (Maximum) observations. The top and bottom whiskers can be understood as the boundaries of data, and any data lying outside it will be an outlier.
  - **Outliers:** points beyond a specified outlier threshold, plotted individually



Statistical detection :*Removing and modifying the outliers using statistical detection techniques is a widely followed method.*

- *Z-Score*
- *Density-based spatial clustering*
- *Regression Analysis*
- *Proximity-based clustering*
- *IQR Scores*

Outliers (cont..)

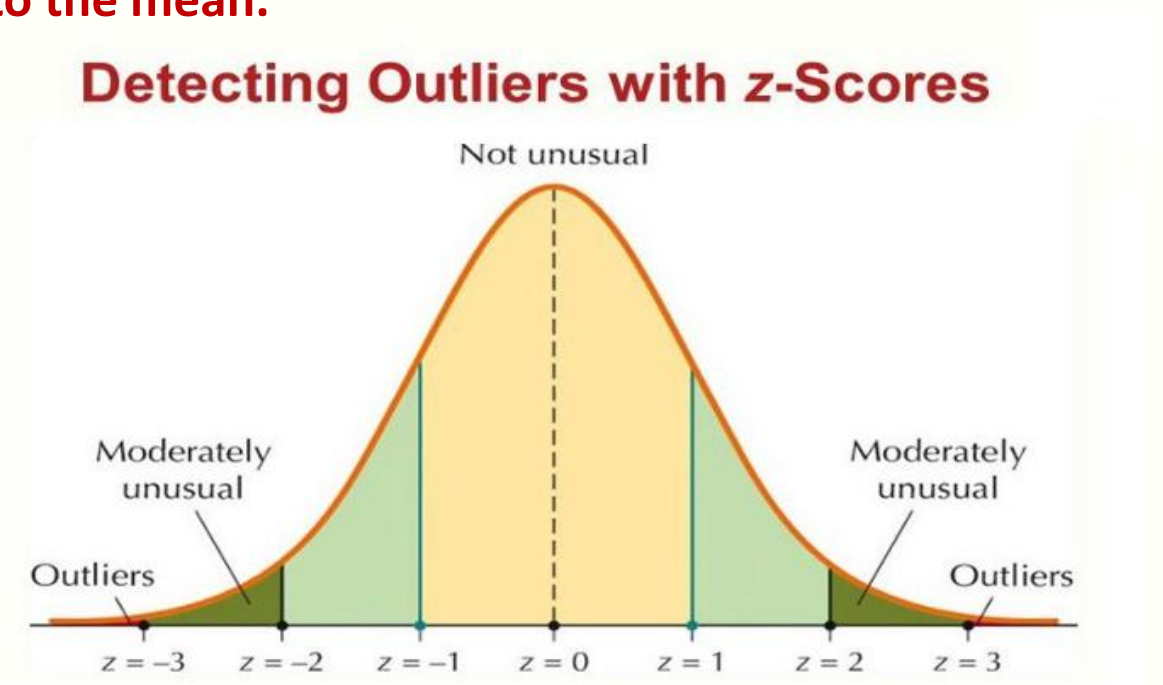
- **Z-Score** for Outlier Detection.

While **data points** are referred to as **x** in a normal distribution, they are called z or z scores in the z distribution. A z score is a standard score that tells you how many standard deviations away from the mean an individual value (x) lies:

- **A positive z score means that your x value is greater than the mean.**
- **A negative z score means that your x value is less than the mean.**
- **A z score of zero means that your x value is equal to the mean.**

$$Z\ Score = \frac{(\text{Observation} — \text{Mean})}{\text{Standard Deviation}}$$

$$Z\ Score = \frac{X - \mu}{\sigma}$$



Outliers and Missing values

Missing Values Detection and Handling.

- Related to Pre-processing.
- Consumes most of the time in Data Analytics.
- Handling missing values is one of the challenges of data analysis

- **Reasons for Missing Values.**
  - Improper maintenance of past data.
  - Observations are not recorded for certain fields due to some reasons.
  - Failure in recording the values due to human error.
  - The user has not provided the values intentionally.
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
  - Information is not collected (e.g., people decline to give their age and weight)

- **Handling missing values**
  - Eliminate data objects or variables
  - Estimate missing values
  - Ignore the missing value during analysis
  - Replace with all possible values (weighted by their probabilities)

Types of Missing Values

Some definitions are based on representation: Missing data is the lack of a recorded answer for a particular field.

- **Missing completely at random (MCAR)**
- **Missing at Random (MAR)**
- **Missing Not at Random (MNAR)**

Types of Missing Values

Missing Completely at Random (MCAR)

- Missingness of a value is independent of attributes
  - Fill in values based on the attribute
  - Analysis may be unbiased overall
- The missingness on the variable is completely unsystematic.

Example when we take a random sample of a population, where each member has the same chance of being included in the sample.

ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

When we make this assumption, we are assuming that whether or not the person has missing data is completely unrelated to the other information in the data.

*When data is missing completely at random, it means that **we can undertake analyses using only observations that have complete data** (provided we have enough of such observations).*

Types of Missing Values

Missing at Random (MAR)

- Missingness is related to other variables
- Fill in values based other values
- Almost always produces a bias in the analysis

Example of MAR is when we take a sample from a population, where the probability to be included depends on some known property.

A simple predictive model is that income can be predicted based on gender and age. Looking at the table, we note that our missing value is for a Female aged 30 or more, and observations says the other females aged 30 or more have a High income. As a result, we can predict that the missing value should be High.

ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

*There is a **systematic relationship between the inclination of missing values and the observed data**, but not the missing data. All that is required is a probabilistic relationship*

Types of Missing Values (cont..)

Missing not at Random (MNAR) - Nonignorable

- When data are MNAR, the fact that the **data are missing is systematically** related to the **unobserved data**, that is, the **missingness is related to events or factors which are not measured by the researcher.**

MNAR means that the probability of being missing varies for reasons that are unknown to us.

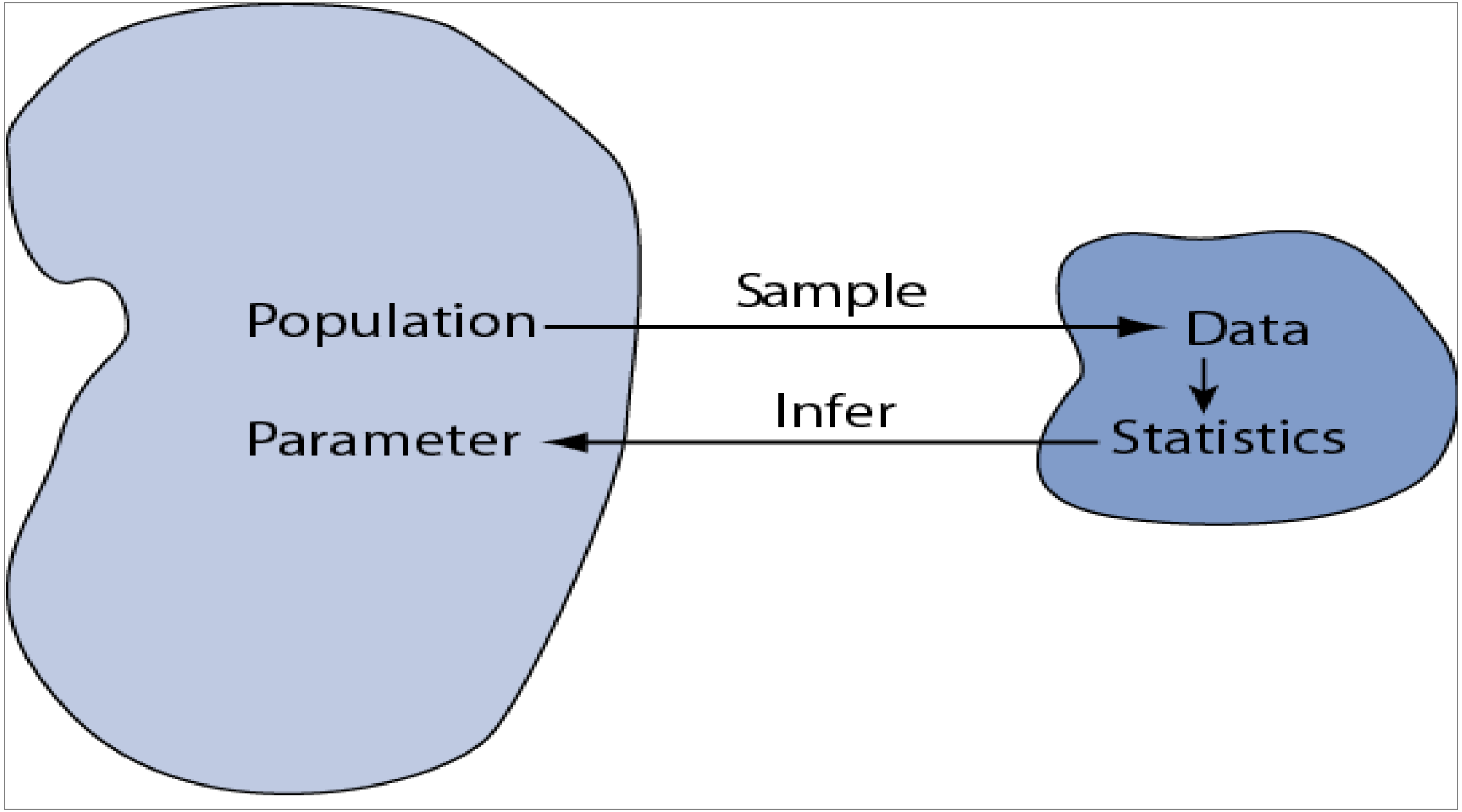
Study looking at homelessness : Data was obtained from **31 women**, of whom **14** were located six months later. Of these, **03** had **exited** from homelessness, so the estimated proportion to have **exited homelessness** is 3/14 = **21%.** As there is no data for the **17** women who could not be contacted (i.e., 31 – 14), it is possible that **none, some, or all of these 17** may have exited from homelessness. This means that potentially the proportion to have exited from homelessness in the sample is between 3/31 = 10% and 20/31 = 65% (17+3). As a result, reporting 21% as being the correct result is misleading. In this example the missing data is nonignorable.

Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Terms Introduce in Prior Chapter

- **Population** ≡ all possible values
- **Sample** ≡ a portion of the population
- **Statistical inference** ≡ generalizing from a sample to a population with calculated degree of certainty
- Two forms of statistical inference
  - **Hypothesis testing**
  - **Estimation**
- **Parameter** ≡ a characteristic of population, e.g., population mean μ
- **Statistic** ≡ calculated from data in the sample, e.g., sample mean (X bar)

$\bar{x}$


Population
Sample
Data
Infer
Parameter
Statistics

Hypothesis Testing

Hypothesis testing is a **formal procedure** for **investigating** our ideas about the world using statistics. It is most often used by scientists to **test specific predictions, called hypotheses**, that arise from theories.

There are 5 main steps in hypothesis testing:

1. **State** your **research hypothesis** as a **null hypothesis (Ho)** and **alternate hypothesis (Ha or H1).**
2. **Collect data** in a way designed to **test the hypothesis.**
3. Perform an **appropriate statistical test.**
4. **Decide** whether to **reject** or **fail** to **reject** your null **hypothesis.**
5. **Present** the **findings in your results** and discussion section.

Step 1: Formulate the Hypothesis

- A **null hypothesis** is a statement of the existing condition, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made.
- An **alternative hypothesis** is one in which some difference or effect is expected.
- The null hypothesis refers to a specified value of the population parameter (e.g., $\mu, \sigma \pi$), not a sample statistic (e.g., $\bar{X}$).

Step 1: State your null and alternate hypothesis

After **developing your initial research hypothesis** (the prediction that you want to investigate), it is important to **restate it as a null (Ho) and alternate (Ha) hypothesis** so that you can test it mathematically.

Hypothesis testing example

You want to test whether there is a **relationship between gender and height**. Based on your knowledge of human physiology, **you formulate a hypothesis that men are, on average, taller than women.** To test this hypothesis, you restate it as:

H0: Men are, on average, not taller than women.

Ha: Men are, on average, taller than women.

Step 2: Collect data

For a **statistical test to be valid**, perform **sampling** and **collect data in** a way that is designed **to test your hypothesis**. If your data are **not representative**, then you cannot make **statistical inferences** about the population you are interested in.

Hypothesis testing example

To test differences in **average height between men and women**, your **sample should have an equal proportion of men and women**, and **cover a variety of socio-economic classes** and **any other control variables** that might influence average height.

Step 3: Perform a statistical test

There are a **variety of statistical tests** available, **but they are all based on the comparison of within-group variance** (how spread out the data is within a category) **versus between-group variance** (how different the categories are from one another).

If the **between-group variance is large enough** that there is **little or no overlap between groups**, then your **statistical test will reflect that by showing a low p-value.** This **means it is unlikely that the differences between these groups came about by chance.**

Alternatively, **if there is high within-group variance** and **low between-group variance**, then your statistical test will **reflect that with a high p-value**. This **means it is likely that any difference you measure between groups is due to chance.**

Choice of **statistical test** will be **based on the type of variables and the level of measurement** of your **collected data.**

Hypothesis testing example

Based on the type of data you collected, perform a **one-tailed t-test (a statistical test that compares the means of two samples)** to test whether men are in fact taller than women.

This test gives you:

an estimate of the **difference in average height between the two groups.**

a **p-value showing how likely you are to see this difference if the null hypothesis of no difference is true.**

Your **t-test shows an average height of 175.4 cm for men and an average height of 161.7 cm for women**, with an estimate of the **true difference ranging from 10.2 cm to infinity. The p-value is 0.002.**

Step 4: Decide whether to reject or fail to reject your null hypothesis

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

In **most cases you will use the p-value** generated by your statistical test **to guide your decision**. And in most cases, your **predetermined level of significance for rejecting the null hypothesis will be 0.05** – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.

Hypothesis testing example

In **your analysis of the difference in average height between men and women, you find that the p-value of 0.002 is below your cutoff of 0.05, so you decide to reject your null hypothesis of no difference.**

Step 5: Present your findings

The **results of hypothesis testing will be presented in the results and discussion sections** of your research paper, dissertation or thesis.

Give a brief summary of the data and a summary of the results of your statistical test (for example, the estimated difference between group means and associated p-value).