Unit – 3

UNIVARIATE ANALYSIS
(Analysis of single variable)

1. **Introduction to Single variable:**
2. **Distribution Variables**
3. **Numerical Summaries of Level and Spread**
4. **Scaling and Standardizing**
5. **Inequality.**

1. Introduction to Single variable:

How much alcohol do men and women drink each week?

How many households have no access to a car?

What is a typical household income in Britain?

Which country in Europe has the longest working hours?

To answer these kinds of questions we need to collect information from a large number of people.

Univariate Analysis

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

Univariate data can be described through:

∅ Frequency Distribution Tables

The frequency distribution table reflects how often an occurrence has taken place in the data. It gives a brief idea of the data and makes it easier to find patterns.

Example:

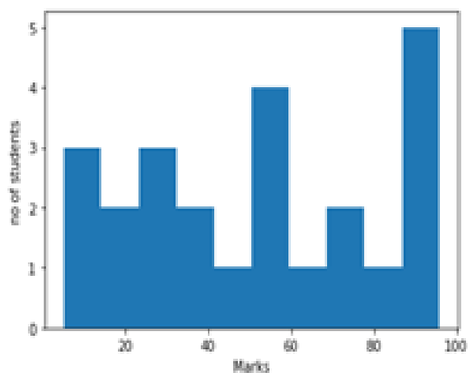
The list of IQ scores is: 118, 139, 124, 125, 127, 128, 129, 130, 130, 133, 136, 138, 141, 142, 149, 130, 154.

IQ Range Number
118-125 3
126-133 7
134-141 4
142-149 2
150-157 1

∅ Histograms

Histograms are similar to bar charts and display the same categorical variables against the category of data. Histograms display these categories as bins which indicate the number of data points in a range. It is best for visualizing continuous data.

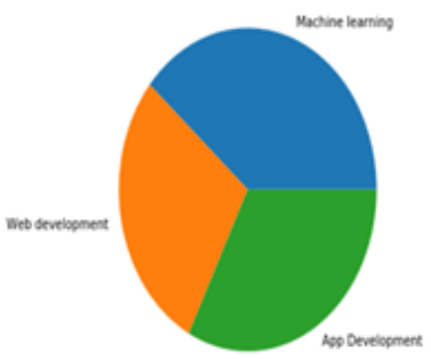
```
In [17]: import matplotlib.pyplot as plt
         import numpy as np
         fig,ax = plt.subplots(1,1)
         a = np.array([22,87,5,43,56,73,55,54,11,20,51,5,79,31,27,63,71,90,92,95,96,32,37,40])
         plt.hist(a)
         ax.set_xlabel('Marks')
         ax.set_ylabel('no of students')
         plt.show()
```



∅ Pie Charts

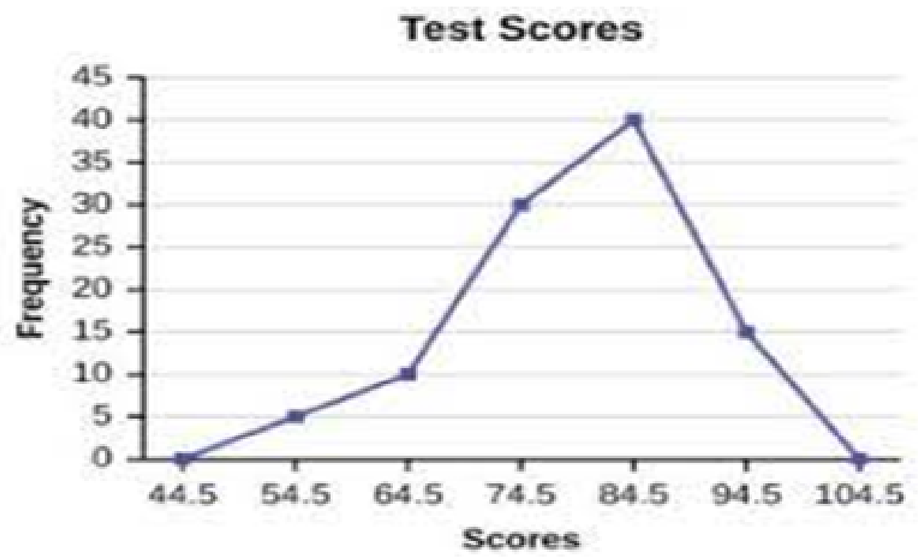
Pie charts are mainly used to comprehend how a group is broken down into smaller pieces. The whole pie represents 100 percent, and the slices denote the relative size of that particular category.

```
In [18]: import matplotlib.pyplot as plt
         fig=plt.figure()
         ax=fig.add_axes([0,0,1,1])
         courses=['Machine learning','Web development','App Development']
         students_enrolled=[50,37,42]
         ax.pie(students_enrolled, labels=courses)
         plt.show()
```



Ø Frequency Polygons

Similar to histograms, a frequency polygon is used for comparing datasets or displaying the cumulative frequency distribution.



Overview:

1. Introduction to problem statement
2. Hypothesis generation with respect to problem statement
3. Introduction to dataset
4. Importing dataset and first impressions
5. Variable Identification and Typecasting
6. Univariate Analysis : Numerical Variables
7. Univariate Analysis : Categorical Variables
8. Univariate Analysis : Missing Values
9. Univariate Analysis : Oulier Values
10. Summary of Univariate Analysis

- The main purpose of univariate analysis is to take data,
    - **summarize that data**, and **find patterns** among the values.
- It **doesn't deal with** causes **or relationships between the values**.
- Several techniques that describe the patterns found in univariate data include
    - **Central tendency** (that is the mean, mode, and median)
    - **Dispersion** (that is, the range, variance, maximum and minimum quartiles (including the interquartile range),
    - **Standard deviation**

https://www.slideshare.net/drswaroopsoumya/univariate-analys

Three types of analysis

- Univariate analysis
  - the examination of the distribution of cases on only one variable at a time (e.g., weight of college students)

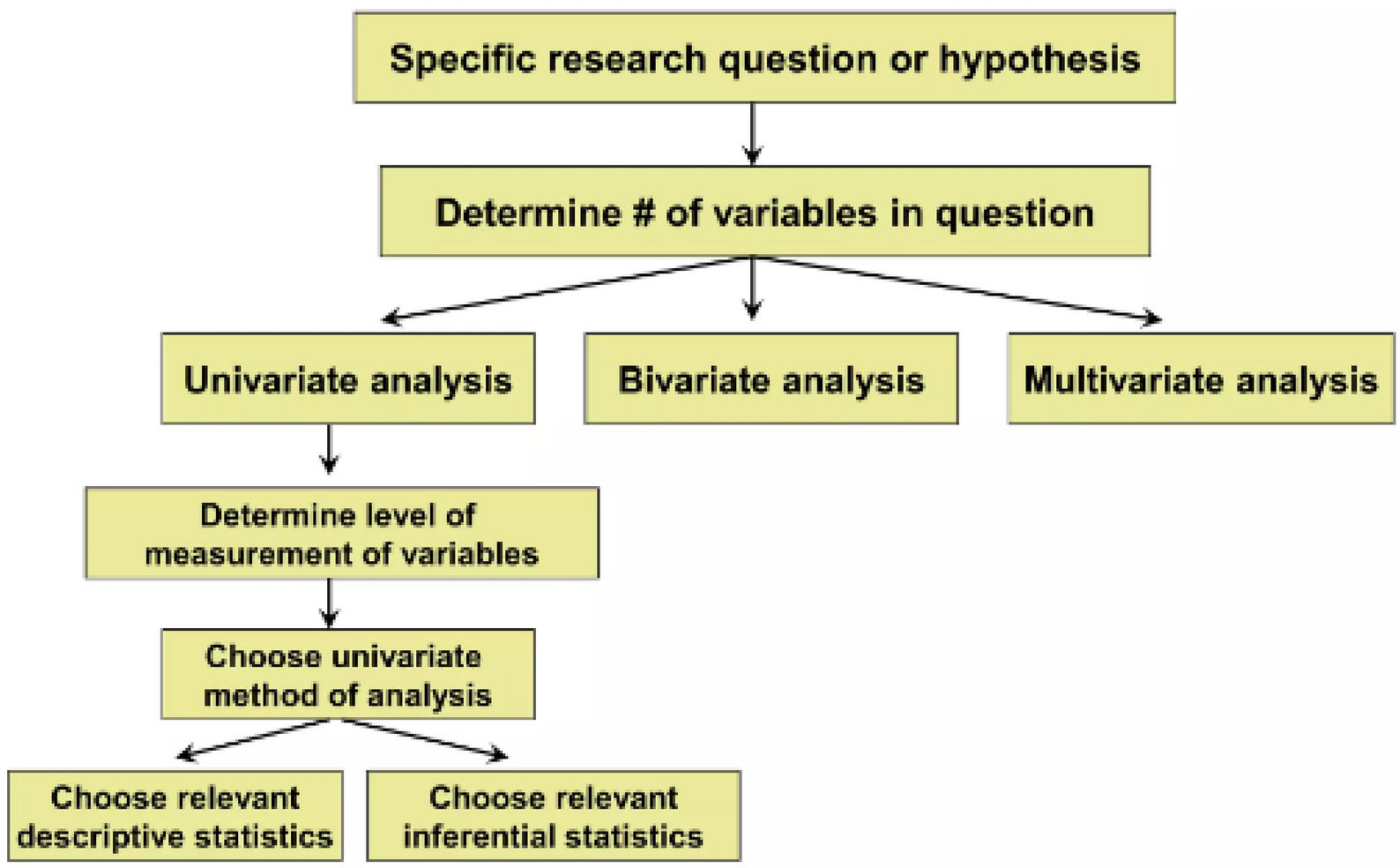
- Bivariate analysis
  - the examination of two variables simultaneously (e.g., the relation between gender and weight of college students)

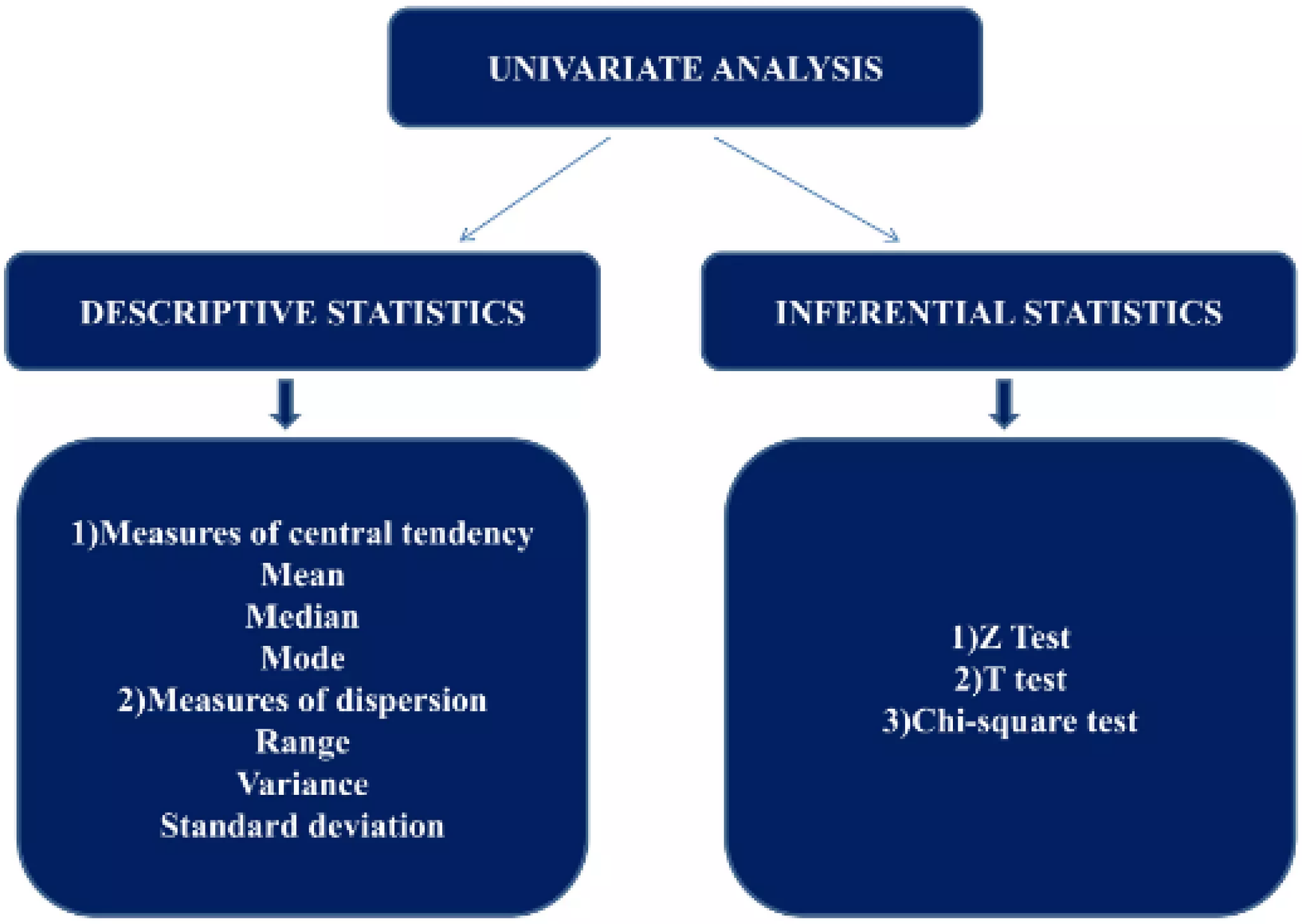
- Multivariate analysis
  - the examination of more than two variables simultaneously (e.g., the relationship between gender, race and weight of college students)

Purpose of diff. types of analysis

- Univariate analysis
  - Purpose: mainly description
- Bivariate analysis
  - Purpose: determining the empirical relationship between the two variables
- Multivariate analysis
  - Purpose: determining the empirical relationship among multiple variables

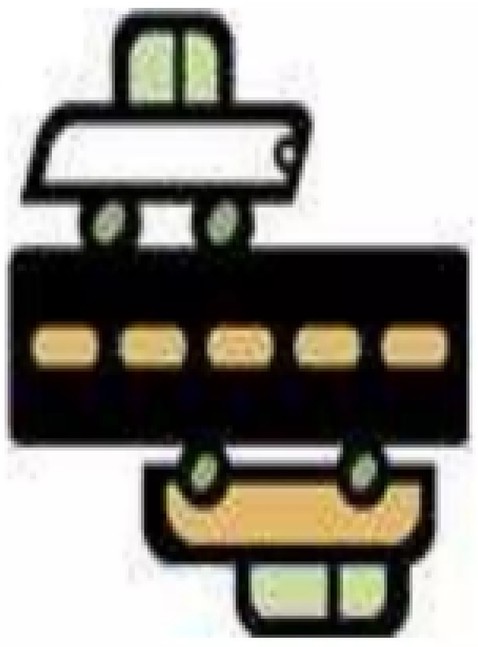
Choosing the Statistical Technique





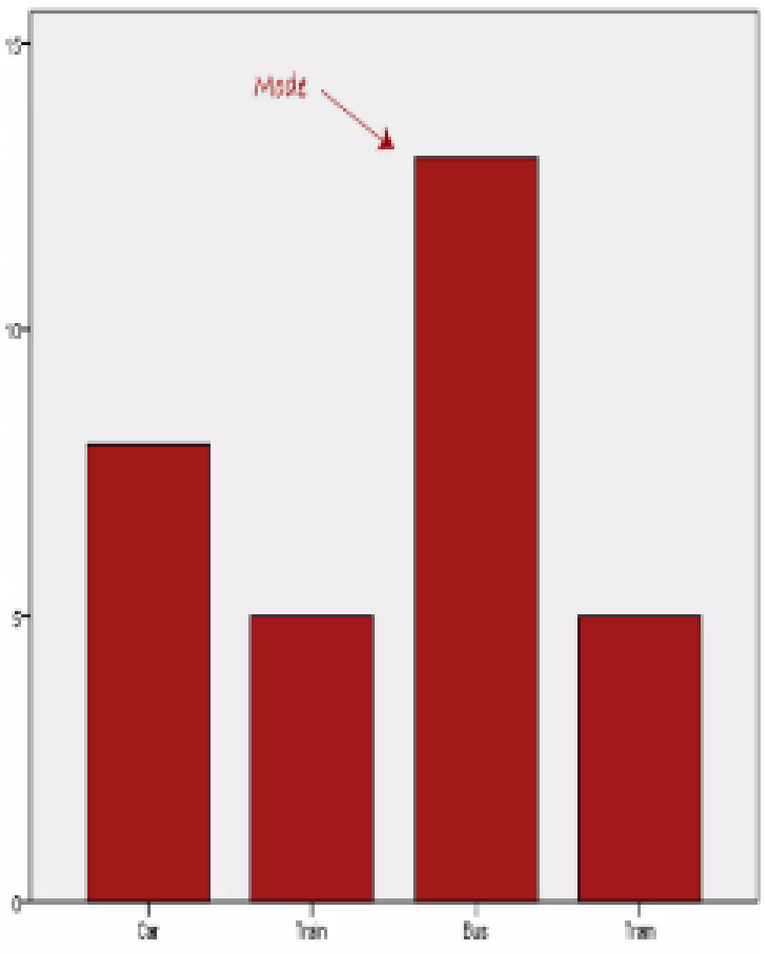
Median

- The median is the middle observation, that is, the point at which half the observations are smaller and half are larger.
- Used in place of mean when the data is skewed (not sensitive to outliers)
- E.g –median price of housing in real estate, median household income

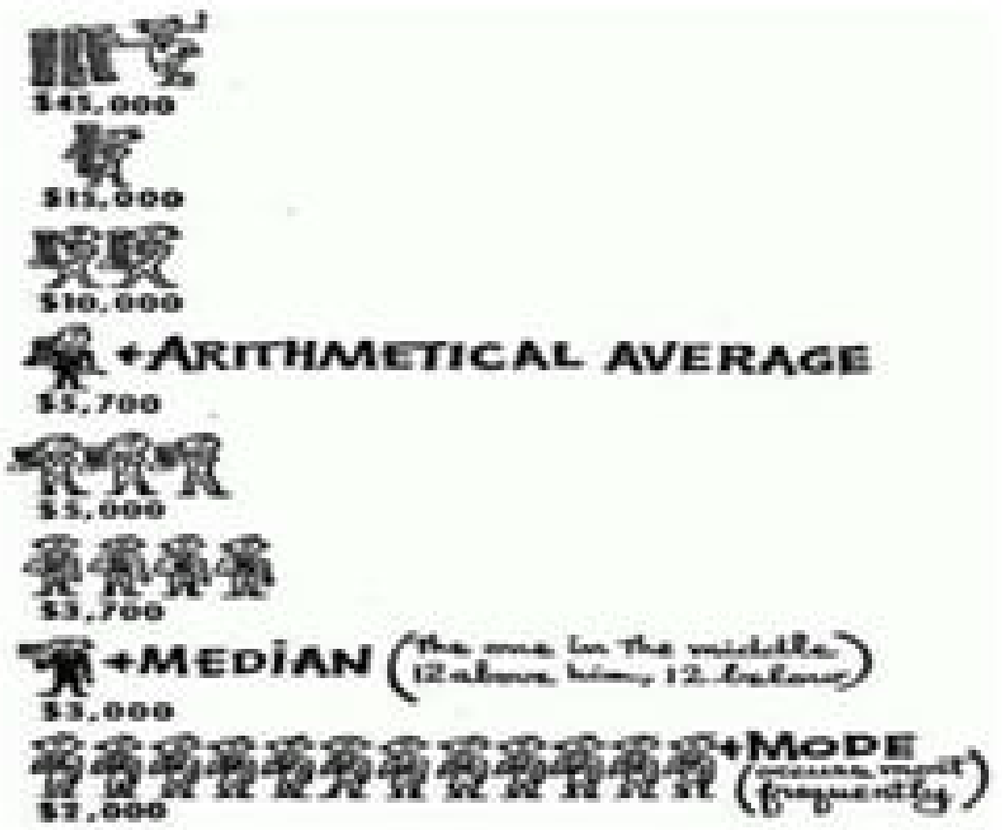


Mode

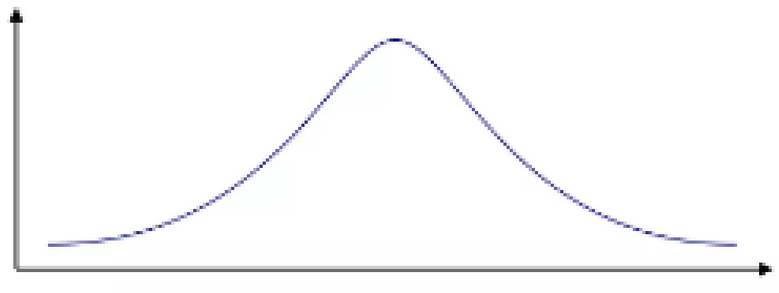
- Value that occurs most frequently.
- Commonly used for a large number of observations when the researcher wants to designate the value that occurs most often.
- Bimodal: When a set of data has two modes
- For frequency tables mode is estimated by the modal class.



Example of mean median mode related to incomes



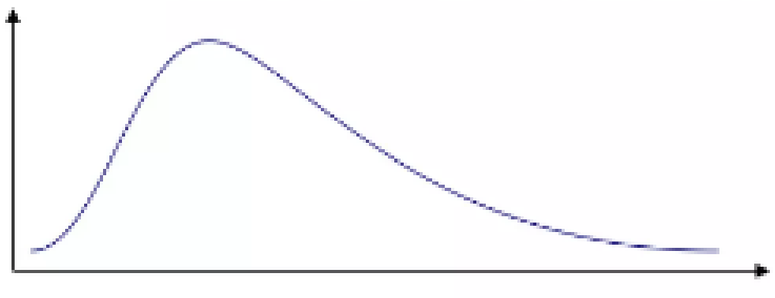
Shapes of distribution



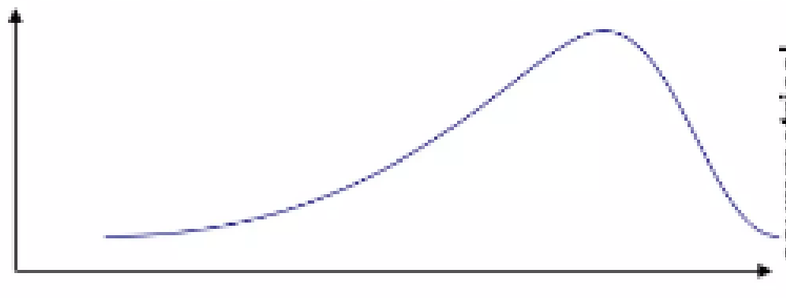
←**Normal distribution:**
symmetrical Bell-shaped curve

symmetrical

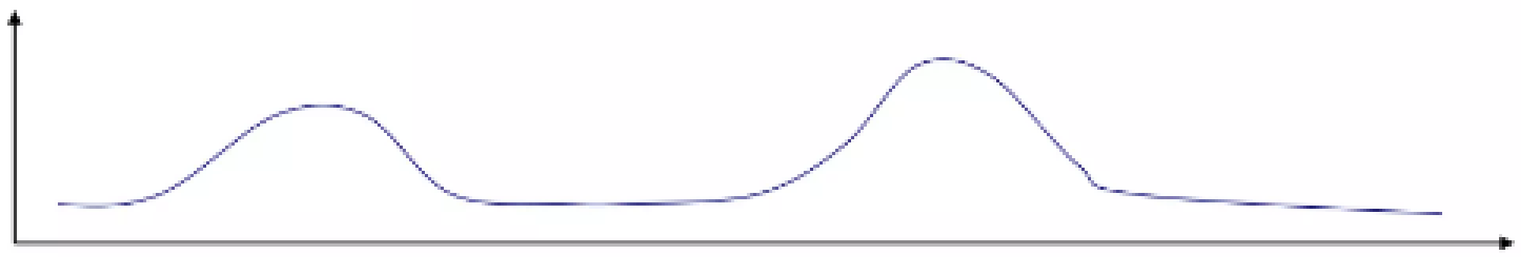
asymmetrical



Positively skewed:
tail on the right, cluster towards low end of the variable

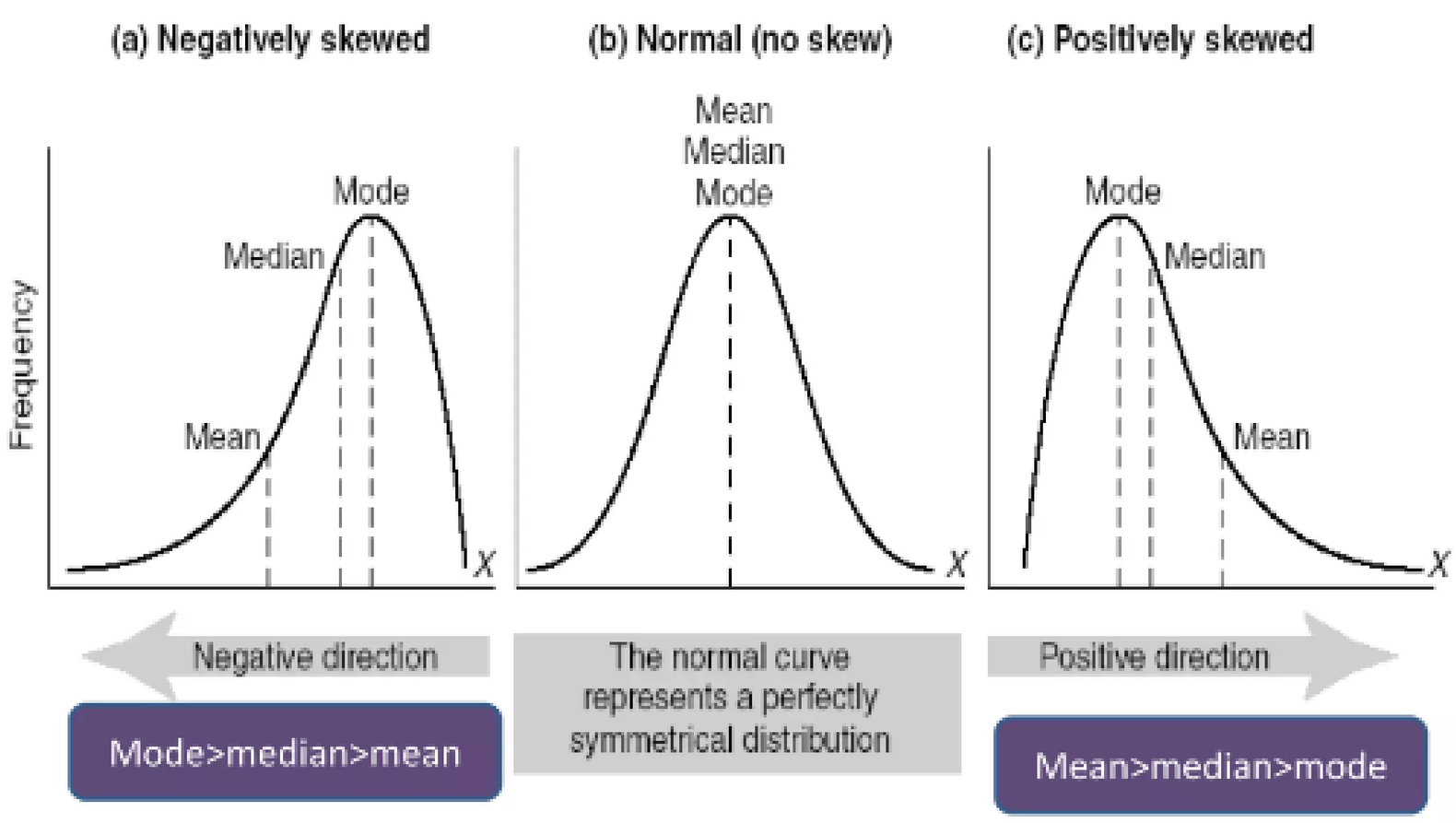


Negatively skewed:
tail on the left, cluster towards high-end of the variable



Bimodality: A double peak

Position of mean median mode



Averages: advantages and disadvantages

Types of average	Advantage	Disadvantage
Mean	1.Uses all data values 2.Algebraically defined	1.Distorted by outliers 2.Distorted by skewness
Median	1.Not distorted by outliers 2.Not distorted by skewness	1.Ignores most of the information 2.Not Algebraically defined
Mode	1.Easily determined for categorical data	1.Not Algebraically defined
Geometric mean	1.Appropriate for skewed data	1.Only appropriate if log transformation produces a symmetrical distribution

Range

- Difference between minimum and maximum value in a data set
- Larger range usually (but not always) indicates a large spread or deviation in the values of the data set.
  - (73, 66, 69, 67, 49, 60, 81, 71, 78, 62, 53, 87, 74, 65, 74, 50, 85, 45, 63, 100)
- Range- 100-45 =55
- Range defines the normal limits of a biological characteristic.
- e. g- systolic BP – 100-140 mm of Hg
  diastolic BP - 80-90 mm of Hg
  Urea - 15-40 mg

Variance

- A measure of how data points differ from the mean
  - Measure of dispersion for a given score, The larger the variance is, the more the scores deviate, on average, away from the mean

- **Average of the squared differences from the mean**

For population variance, $\sigma^2 = \frac{\sum\left(x-\bar{X}\right)^2}{N}$

For sample variance, $s^2 = \frac{\sum\left(x-\bar{X}\right)^2}{n-1}$

- **difficult to interpret because it is in the units of square of variables**

Standard Deviation

- A measure of variation that gives spread of data about the mean
- **Square root of the variance**
- higher standard deviation indicates higher spread, less consistency, and less clustering.

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

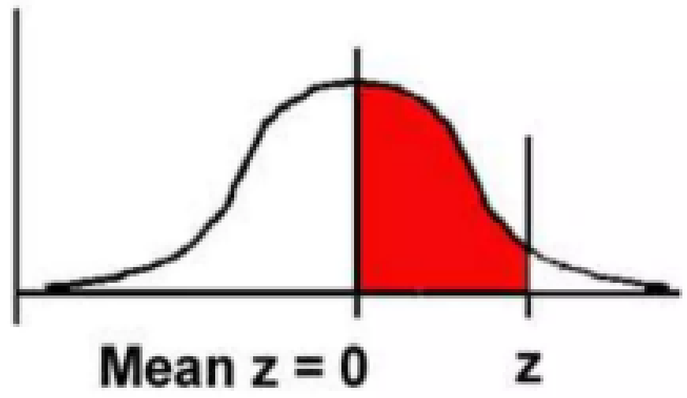
- sample standard deviation:

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

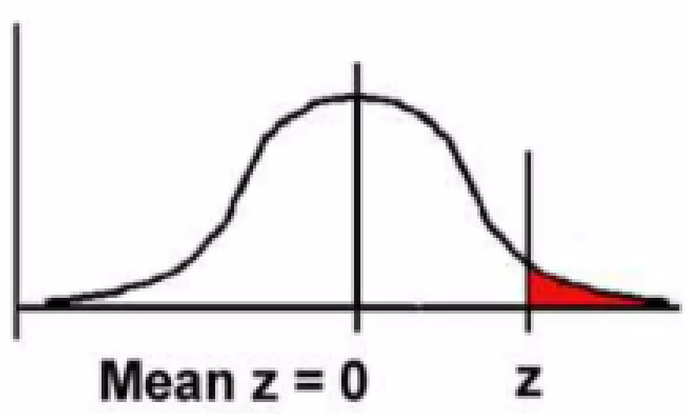
- population standard deviation:
- **measured in the *same units as the data*, making it easy to interpret.**

Relative or Standard Normal Deviate or Variate (Z)

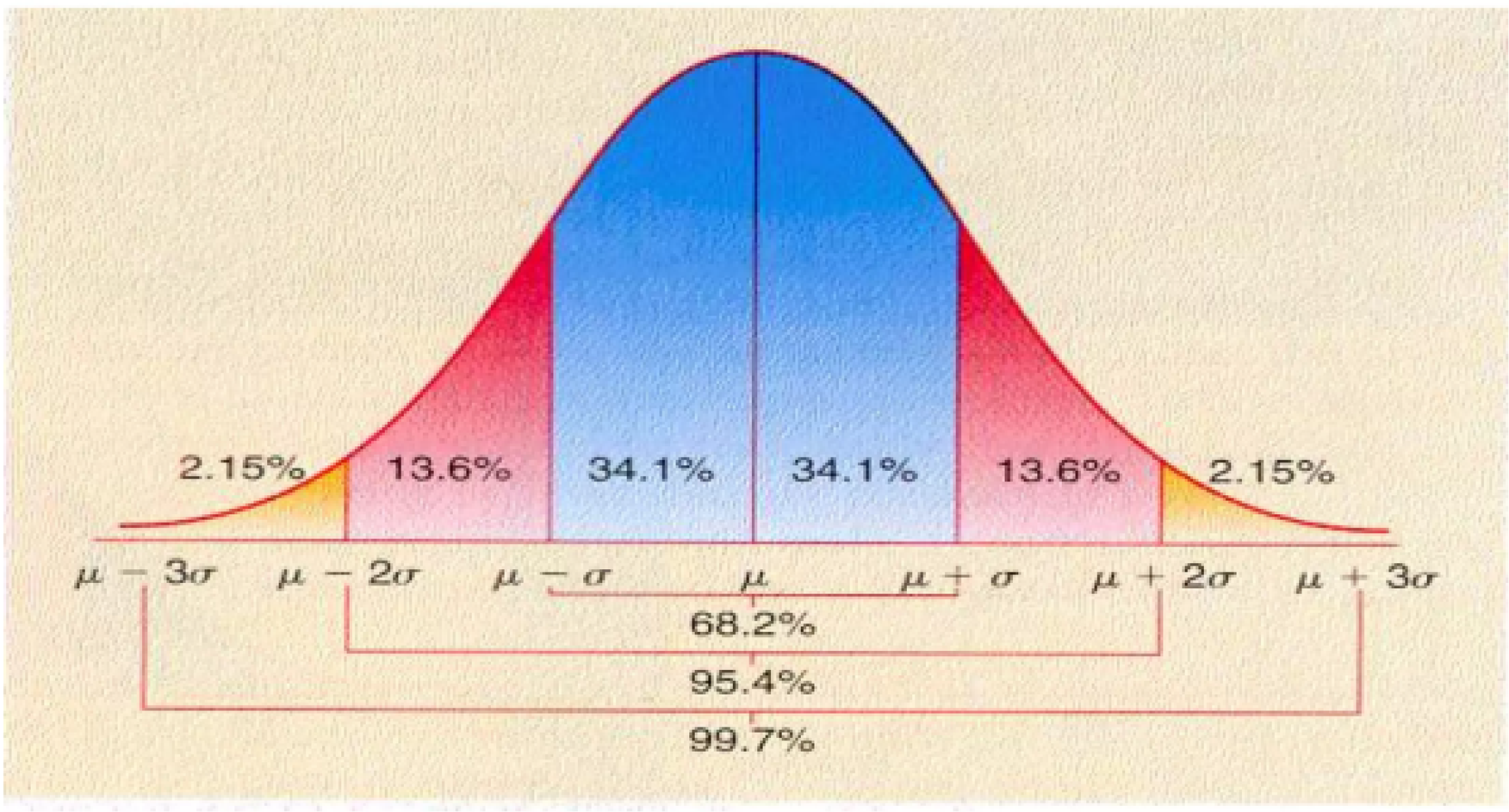
- **Deviation from the mean in a normal distribution or curve**
- **Given by symbol "Z"**
- $Z = \frac{observation - Mean}{SD}$
- **Indicates how much an observation is bigger or smaller than mean in units of SD**



- **95% confidence level: Z = 1.96**
- **99% confidence level: Z = 2.575**
- **99.9 confidence level: Z = 3.27**

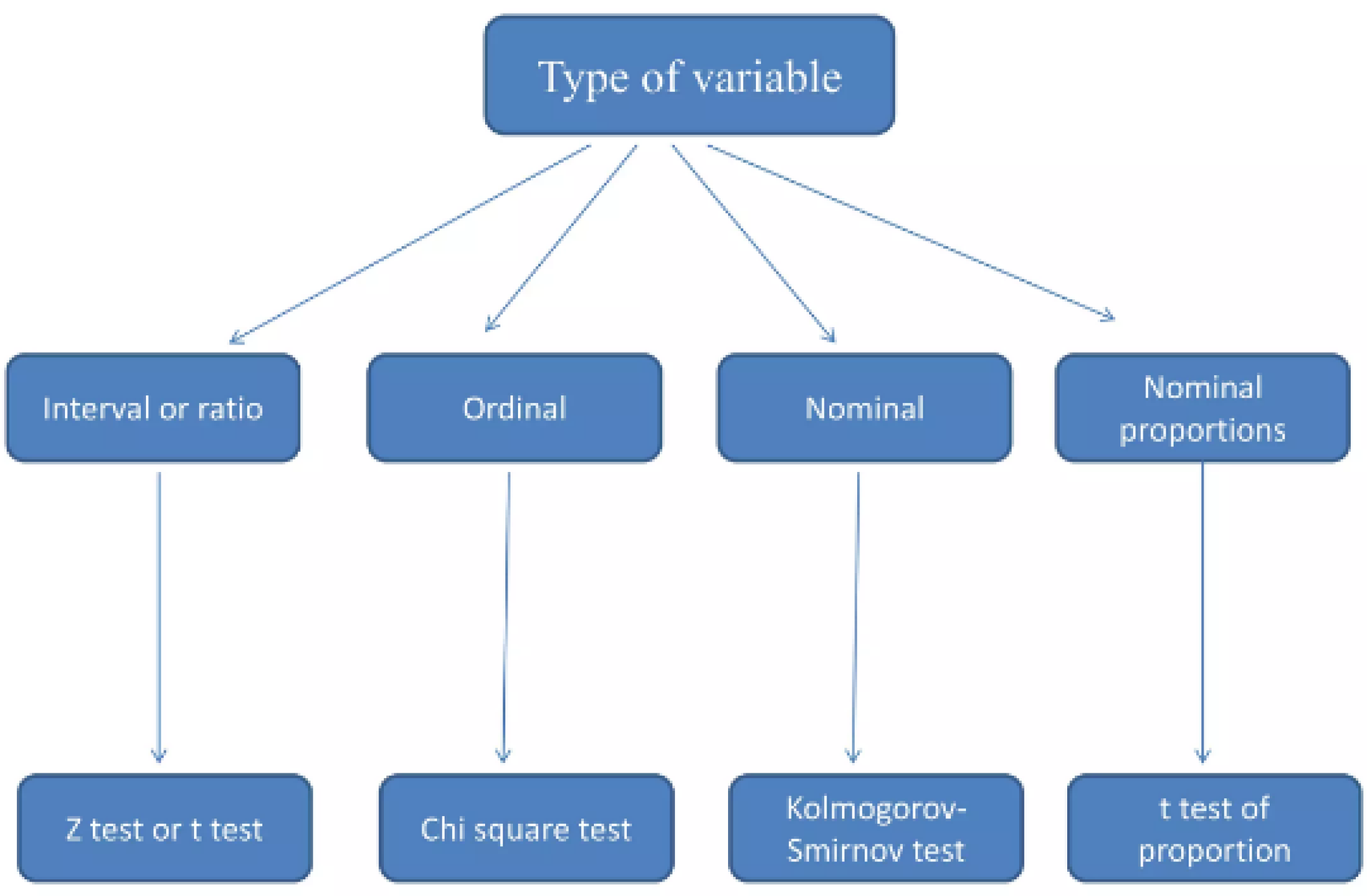


Normal distribution curve with SD

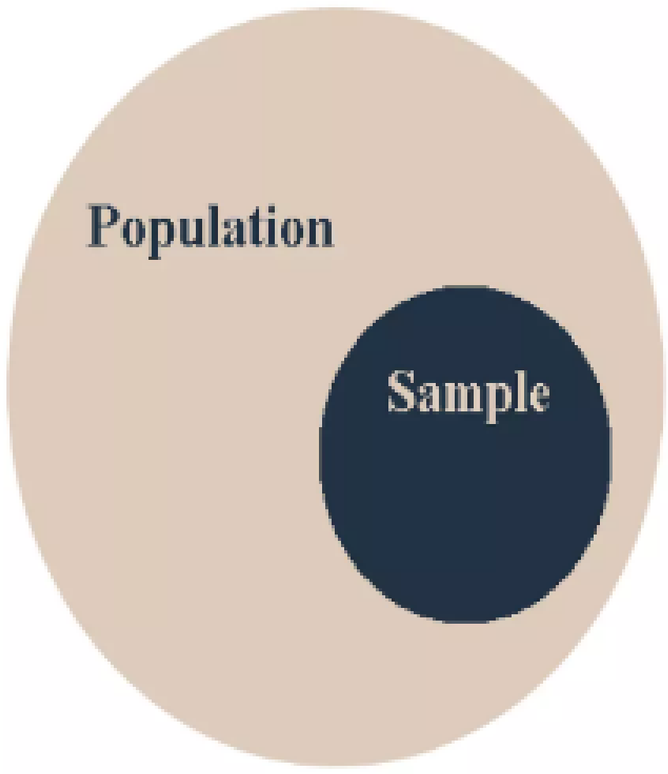


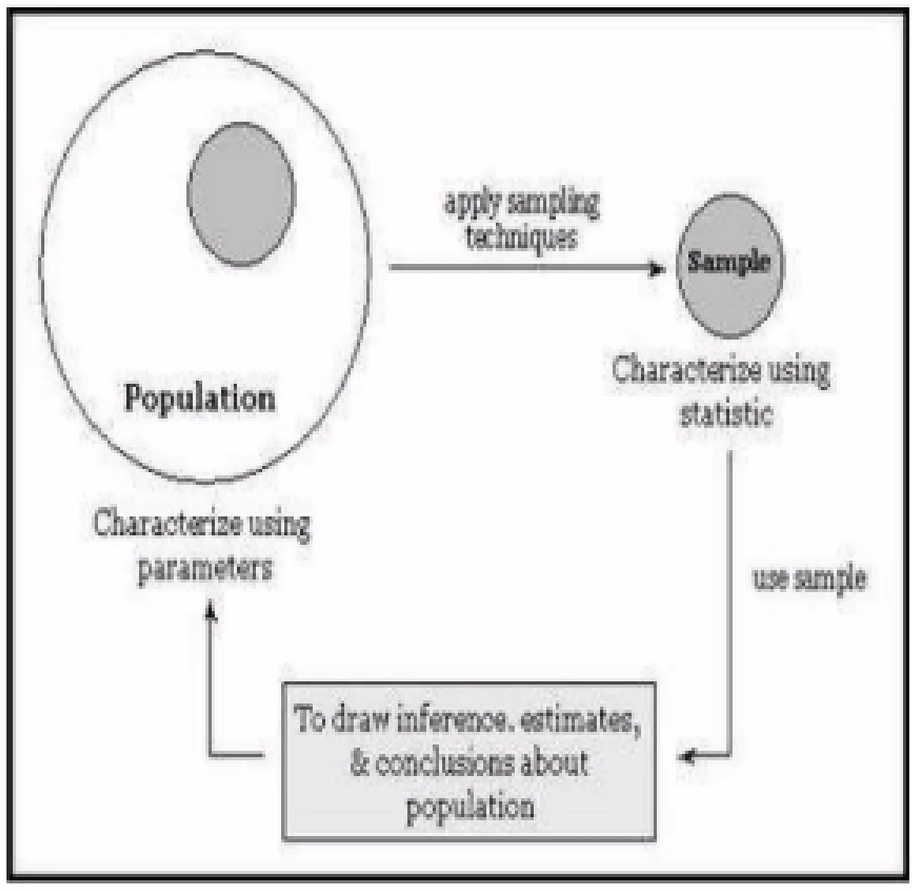
Inferential univariate analysis components

Test Description	Test Statistic
Compare an Observed Mean with Some Predetermined Value	*Z or t-test*
Compare an Observed Frequency with a Predetermined Value	X^2
Compare an Observed Proportion with Some Predetermined Value	*Z or t-test for Proportions*



Drawing sample from population





Z Test determines if there is a significant difference between sample and population means.

> **Z Test = $(\bar{x} - \mu) / (\sigma / \sqrt{n})$**

Here,

- $\bar{x}$ = Mean of Sample
- μ = Mean of Population
- σ = Standard Deviation of Population
- **n** = Number of Observation

Z Test = (x̄ – μ) / (σ / √n)

	A	B
1	Sample Mean (x̄)	112.5
2	Population Mean (μ)	100
3	Standard Deviation of Population (σ)	15
4	Number of Observation (n)	30
5		
6	Z Test Statistics is calculated using the formula given below	
7	**Z Test = (x̄ - μ) / (σ / √n)**	
8		
9	**Z Test Statistics Formula**	**=(B1-B2)/(B3/SQRT(B4))**
10	**Z Test Statistics**	**4.56**
11		

- Z Test = (112.5 – 100) / (15 / √30)
- Z Test **= 4.56**

UNIVARIATE ANALYSIS

Dr. Soumya Swaroop Sahoo
JR, Community Medicine
PGIMS Rohtak

Contents

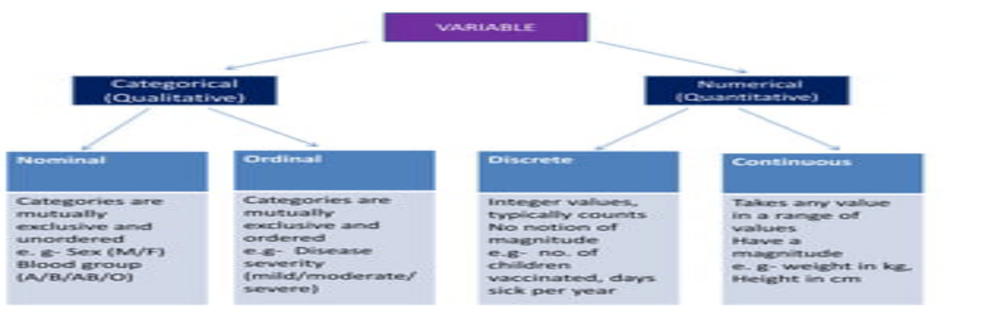
- Introduction
- Variables
- Types of variables
- Scales of measurement
- Types of analysis
- Components of univariate analysis
- Advantages and limitations

Introduction

- The word "statistics" has several meanings: data or numbers, the process of analyzing the data, and the description of a field of study.
- It is derived from the Latin word status, meaning "manner of standing" or "position."
- Statistics were first used by tax assessors to collect information for determining assets and assessing taxes.

Introduction

- Variable -any character, characteristic or quality that varies is termed a variable.
- E.g. - to collect basic clinical and demographic information on patients with a particular illness. The variables of interest may include the sex, age and height of the patients.



Scales of measurement

- Nominal or categorical
- Ordinal
- Interval
- Ratio

Nominal scale

- Simplest level of measurement when data values fit into categories.
- Observations are dichotomous or binary in that the outcome can take on only one of two values: yes or no.
- Mutually exclusive.
- E.g sex of patient(M/F), nationality

Ordinal scale

- When an inherent order occurs among the categories, the observations are said to be measured on an ordinal scale.
- Clinicians often use ordinal scales to determine a patient's amount of risk or the appropriate type of therapy.
- E.g socio-economic class, rank order of a class(1^{st},2^{nd}, 3^{rd}) VAS(visual analog scale) for pain

Scales of measurement

- Interval Scale
  - Data classified by ranking.
  - Quantitative classification .
  - Zero point of scale is arbitrary (differences are meaningful).
  - Fahrenheit temp. scale , Time
- Ratio Scale
  - Data classified as the ratio of two numbers.
  - Quantitative classification .
  - Zero point of scale is absolute (data can be added, subtracted, multiplied, and divided).
  - E. g- Kelvin temp. scale, Weight, Height

Descriptive statistics

- Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that patterns might emerge from the data.
- Does not allow us to make conclusions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made.
- Simply a way to describe the data.

Inferential statistics

- Inferential statistics is concerned with making predictions or inferences about a population from observations and analysis of a sample.
- We can take the results of an analysis using a sample and can generalize it to the larger population that the sample represents.

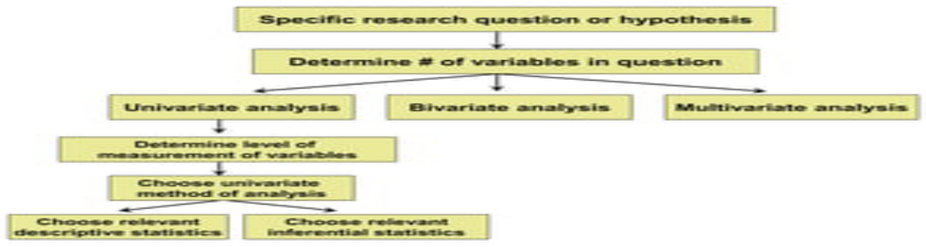
Three types of analysis

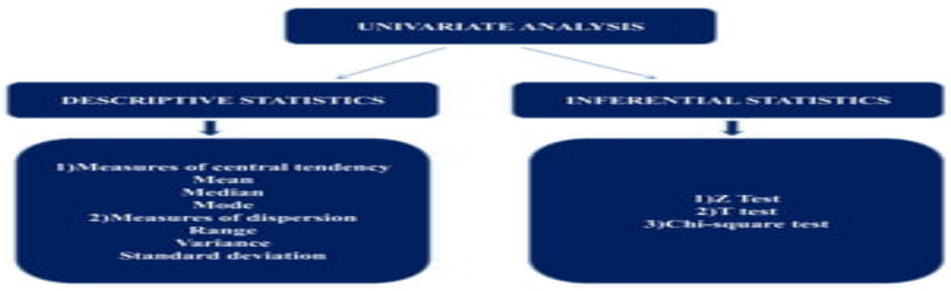
- Univariate analysis
  - the examination of the distribution of cases on only one variable at a time (e.g., weight of college students)
- Bivariate analysis
  - the examination of two variables simultaneously (e.g., the relation between gender and weight of college students)
- Multivariate analysis
  - the examination of more than two variables simultaneously (e.g., the relationship between gender, race and weight of college students)

Purpose of diff. types of analysis

- Univariate analysis
  - Purpose: mainly description
- Bivariate analysis
  - Purpose: determining the empirical relationship between the two variables
- Multivariate analysis
  - Purpose: determining the empirical relationship among multiple variables

Choosing the Statistical Technique

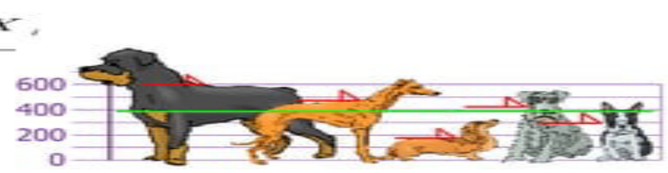




Mean

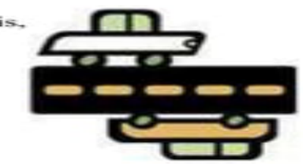
- Arithmetic, or simple, mean is used most frequently in statistics.
- Arithmetic average of the observations.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$



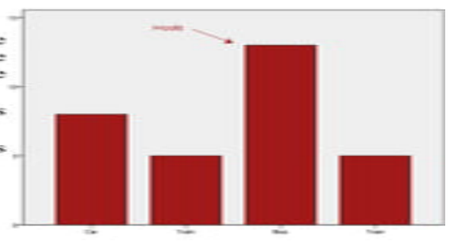
Median

- The median is the middle observation, that is, the point at which half the observations are smaller and half are larger.
- Used in place of mean when the data is skewed (not sensitive to outliers)
- E.g –median price of housing in real estate, median household income



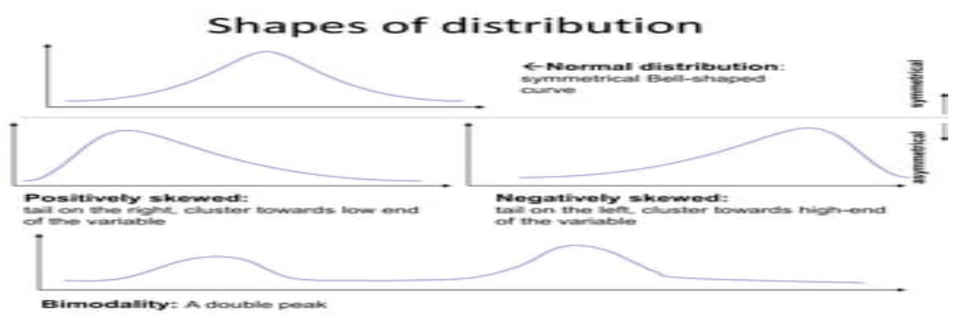
Mode

- Value that occurs most frequently.
- Commonly used for a large number of observations when the researcher wants to designate the value that occurs most often.
- Bimodal: When a set of data has two modes
- For frequency tables mode is estimated by the modal class.

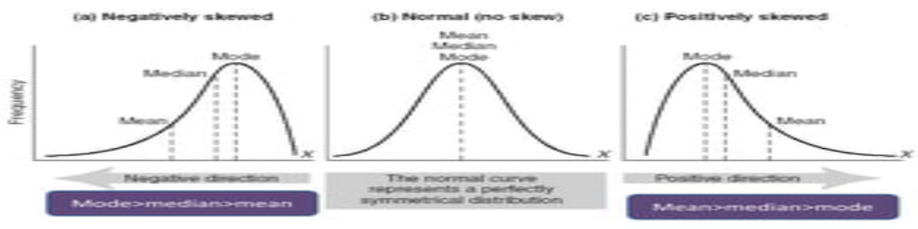


Example of mean median mode related to incomes

Shapes of distribution



Position of mean median mode



Averages: advantages and disadvantages

Types of average	Advantage	Disadvantage
Mean	1.Uses all data values 2.Algebraically defined	1.Distorted by outliers 2.Distorted by skewness
Median	1.Not distorted by outliers 2.Not distorted by skewness	1.Ignores most of the information 2.Not Algebraically defined
Mode	1.Easily determined for categorical data	1.Not Algebraically defined
Geometric mean	1.Appropriate for skewed data	1.Only appropriate if log transformation produces a symmetrical distribution

Range

- Difference between minimum and maximum value in a data set
- Larger range usually (but not always) indicates a large spread or deviation in the values of the data set.
  - (73, 66, 69, 67, 49, 60, 81, 71, 78, 62, 53, 87, 74, 65, 74, 50, 85, 45, 63, 100)
  - Range- 100-45 =55
  - Range defines the normal limits of a biological characteristic.
  - e. g- systolic BP – 100-140 mm of Hg
    diastolic BP - 80-90 mm of Hg
    Urea - 15-40 mg

Variance

- A measure of how data points differ from the mean
  - Measure of dispersion for a given score. The larger the variance is, the more the scores deviate, on average, away from the mean
- **Average of the squared differences from the mean**

For population variance, $\sigma^2 = \frac{\Sigma(x-\bar{X})^2}{N}$

For sample variance, $s^2 = \frac{\Sigma(x-\bar{X})^2}{n-1}$

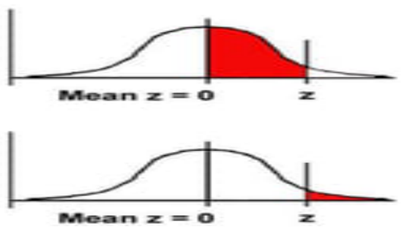
- **difficult to interpret because it is in the units of square of variables**

Standard Deviation

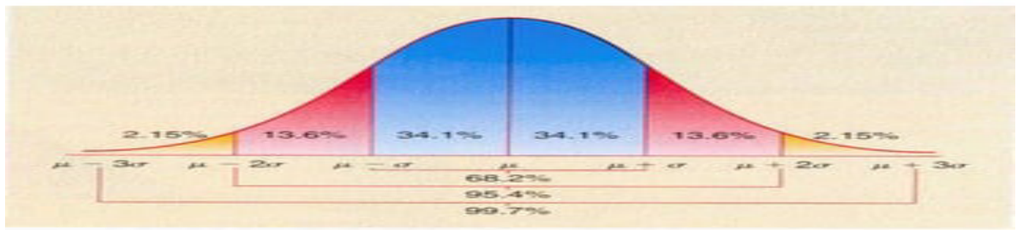
- A measure of variation that gives spread of data about the mean
- **Square root of the variance**
- higher standard deviation indicates higher spread, less consistency, and less clustering.
- sample standard deviation: $s = \sqrt{\frac{\Sigma(x-\bar{X})^2}{n-1}}$
- population standard deviation: $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$
- **measured in the *same units as the data*, making it easy to interpret.**

Relative or Standard Normal Deviate or Variate (Z)

- **Deviation from the mean in a normal distribution or curve**
- **Given by symbol "Z"**
- $Z = \frac{observation - Mean}{SD}$
- **Indicates how much an observation is bigger or smaller than mean in units of SD**
- **95% confidence level: Z = 1.96**
- **99% confidence level: Z = 2.575**
- **99.9 confidence level: Z = 3.27**

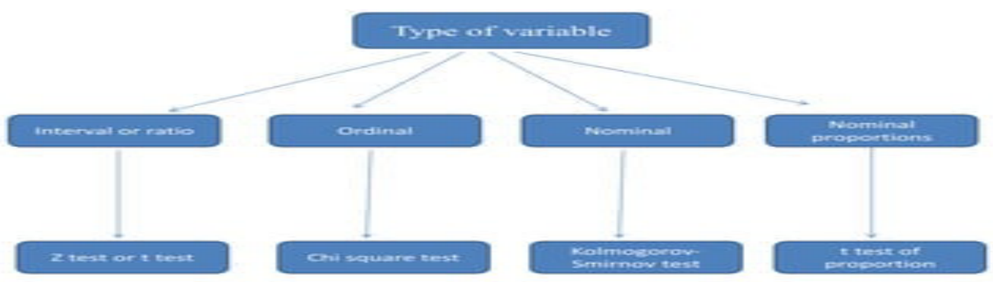


Normal distribution curve with SD

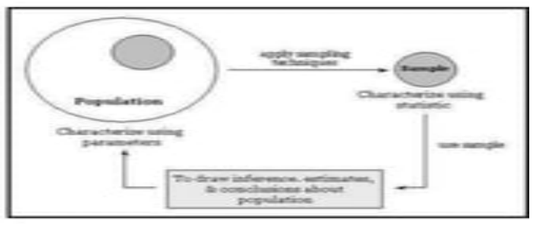


Inferential univariate analysis components

Test Description	Test Statistic
Compare an Observed Mean with Some Predetermined Value	*Z or t-test*
Compare an Observed Frequency with a Predetermined Value	X^2
Compare an Observed Proportion with Some Predetermined Value	*Z or t-test for Proportions*



Drawing sample from population



Interval estimation
Confidence interval (CI)

Provide us with a range of values that we believe, with a given level of confidence, containes a true value

CI for the population means

$$95\%CI = \bar{x} \pm 1.96\,SEM$$
$$99\%CI = \bar{x} \pm 2.58\,SEM$$
$$SEM = \frac{SD}{\sqrt{n}}$$

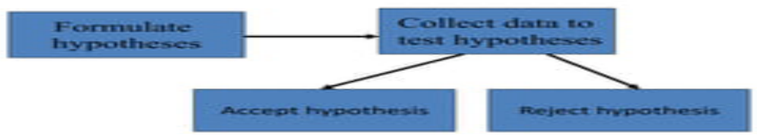
Testing of hypotheses

learning objectives:

- to understand the role of significance
- to distinguish the null and alternative hypotheses
- to interpret p-value, type I and II errors

Hypothesis testing

- Hypotheses are defined as formal statements of explanations stated in a testable form.
- To test statistical hypotheses two presumptions are made to draw the inference from sample value.
- Logic- designed to detect *significant differences*: differences that did *not* occur by random chance.



Null and alternate hypothesis

1. **Null Hypothesis (H_0)**
   - The difference is caused by random chance.
   - The H_0 always states there is "no significant difference." it means that there is no significant difference between the population mean and the sample mean.
2. **Alternate hypothesis (H_1)**
   - "The difference is real".
   - (H_1) always contradicts the H_0.
- One (and only one) of these explanations *must* be true.

Testing Hypotheses: The Five Step Model

1. Make Assumptions and meet test requirements.
2. State the null hypothesis.
3. Select the sampling distribution and establish the critical region.
4. Compute the test statistic.
5. Make a decision and interpret results.

Two-tailed vs. One-tailed Tests

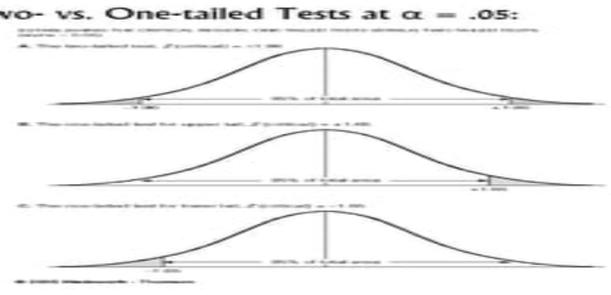
- In a two-tailed test, the direction of the difference is not predicted.
- A two-tailed test splits the critical region equally on both sides of the curve.
- In a one-tailed test, the researcher predicts the direction (i.e. greater or less than) of the difference.
- All of the critical region is placed on the side of the curve in the direction of the prediction.

The Curve for Two- vs. One-tailed Tests at $\alpha = .05$:

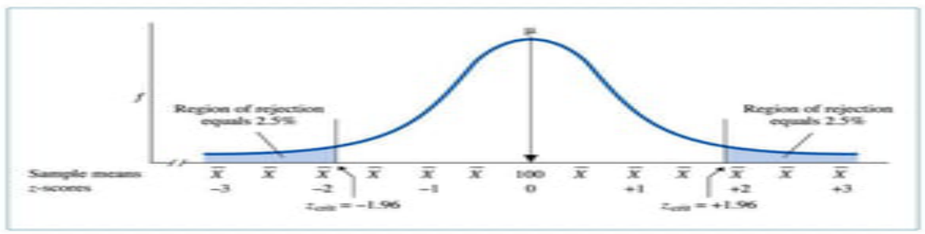
Two-tailed test:
"is there a **significant difference**?"

One-tailed tests:
"is the sample mean **greater** than μ or P_u?"

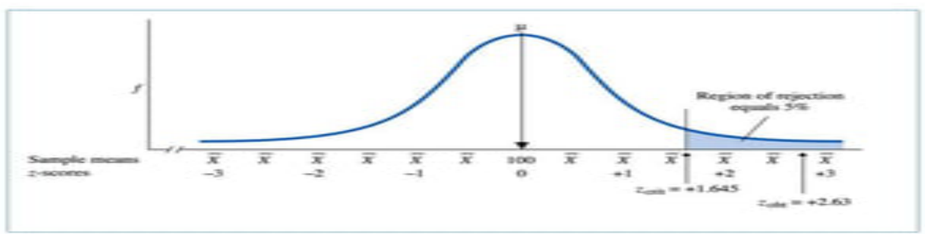
"is the sample mean **less** than μ or P_u?"



Two tailed test



One tailed test



Type I and Type II Errors

- Type I, or alpha error:
  - Rejecting a true null hypothesis
- Type II, or beta error:
  - Failing to reject a false null hypothesis

TABLE 8.4 DECISION MAKING AND THE NULL HYPOTHESIS

The H_0 Is Actually:	Decision	
	Reject	Fail to Reject
True	Type I, or α, error	OK
False	OK	Type II, or β, error

Testing of hypotheses
Type I and Type II Errors. Example

Suppose there is a test for a particular disease.

If the disease really exists and is diagnosed early, it can be successfully treated

If it is not diagnosed and not treated, the person will become severely disabled

If a person is erroneously diagnosed as having the disease and treated, no physical damage is done.

To which type of error we are willing to risk ?

Testing of hypotheses
Type I and Type II Errors. Example.

Decision	No disease	Disease
Not diagnosed	OK	Type II error
Diagnosed	Type I error	OK

treated but not harmed by the treatment

irreparable damage would be done

Inference: to avoid Type error II, have high level of significance

Z-test

- For testing the significance of differences between two means of samples
- *It calculates the probability that the two samples could have drawn from populations of the same mean, differences arising merely from sampling variability.*
- ***Assumptions for z-test***
- Population approximately Normal or large sample. **Sample size must be larger than 30**
- **Data must be quantitative**
- **The variable is assumed to follow normal distribution in the population**

Example (z test of proportion)

- In a recent survey, 55% of the population were found to be aware of modes of HIV transmission. A random sample of 150 urban persons showed that 49% of them were aware of the same. Is the difference significant?

- Using the formula for proportions and 5 step method to solve…

Solution:

Step 1:
Random sample
The sample is large (>30)

Step 2:
H_0: $P_u = .55$ (converting % to proportion)
(H_0: $P_s = P_u$)
H_1: $P_u \neq .55$

Step 3:
The sample is large, so we use Z distribution
Alpha (α) = .05
Critical Z = ±1.96

Solution (cont.)

Step 4

$$Z = \frac{P_s - P_u}{\sqrt{P_u(1-P_u)/n}} = \frac{.49 - .55}{\sqrt{.55(1-.55)/150}} = -1.48$$

Step 5

- Z (obtained) < Z (critical)
- Fail to reject H_0. There is no significant difference between the state population and the urban sample.

t test

- When the sample size is small (approximately < 30) then the Student's *t* distribution should be used
- The test statistic is known as "*t*".
- The curve of the *t* distribution is flatter than that of the Z distribution but as the sample size increases, the *t*-curve starts to resemble the Z-curve
- t ⟶ Z as n increases,
  If n>100, t approaches Z.

Example of t test

- A random sample of 26 sociology graduates scored 458 on the advanced sociology test with a standard deviation of 20. Is this significantly different from the population average (μ = 440)?

Solution (using five step model)

- Step 1: Make Assumptions and Meet Test Requirements:

- 1. Random sample
- 2. Level of measurement is interval-ratio
- 3. The sample is small (<30)

Solution (cont.)

Step 2: State the null and alternate hypotheses.

H_0: null hypothesis $\mu = 440$ (or H_0: $\bar{X} = \mu$)

H_1:alternate hypothesis $\mu \neq 440$

Solution (cont.)

- Step 3: Select Sampling Distribution and Establish the Critical Region

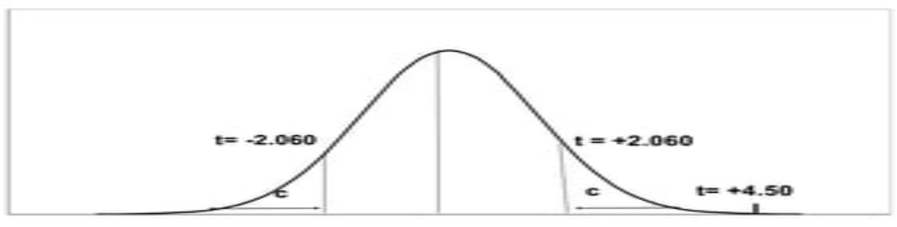
1. Small sample, I-R scale of measurement, so *t* test to be used.
2. Alpha (α) = .05
3. Degrees of Freedom = n-1 = 26-1 = 25
4. Critical t = ±2.060

Solution (cont.)

- Step 4: Using Formula to Compute the Test Statistic

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n-1}} = \frac{458 - 440}{20/\sqrt{26-1}} = 4.5$$

Looking at the curve for the t distribution
Alpha (α) = .05



Step 5 Make a Decision and Interpret Results

- The obtained t score fell in the Critical Region, so we *reject* the H_0 { t (obtained) > t (critical) }
  - If the H_0 were true, a sample outcome of 458 would be unlikely.
  - Therefore, the H_0 is false and must be rejected.
- Sociology graduates have a score that is significantly different from the general student body at ($t = 4.5$, $df = 25$, $\alpha = .05$).

Main Considerations in Hypothesis Testing:

- Sample size
  - Use Z for large samples, t for small (<100)
- There are two other choices to be made:
  - One-tailed or two-tailed test
    - "Is there a difference?" = 2-tailed test
    - "Is the difference less than or greater than?" = 1-tailed test
- Alpha (α) level
  - .05, .01, or .001? (α=.05 is most common)

Selected nonparametric tests
Chi-Square goodness of fit test.

To determine whether a variable has a frequency distribution comparable to the one expected

$$\chi^2 = \sum_i \frac{1}{f_{ei}} (f_{oi} - f_{ei})^2$$

expected frequency can be based on
- theory
- previous experience
- comparison groups

Selected nonparametric tests
Chi-Square goodness of fit test. Example

The average prognosis of total hip replacement in relation to pain reduction in hip joint is
excellent - 80%
good - 10%
medium - 5%
bad - 5%
expected

In our study of we had got a different outcome
excellent - 95%
good - 2%
medium - 2%
bad - 1%
observed

Do observed frequencies differ from expected ?

Selected nonparametric tests
Chi-Square goodness of fit test. Example

$f_{e1} = 80$, $f_{e2} = 10$, $f_{e3} = 5$, $f_{e4} = 5$;
$f_{o1} = 95$, $f_{o2} = 2$, $f_{o3} = 2$, $f_{o4} = 1$;

$\chi^2 = 14.2$, df=3 (4-1)

$0.001 < p < 0.01$

$\chi^2 > 9.84$ $p < 0.05$
$\chi^2 > 11.34$ $p < 0.01$
$\chi^2 > 16.27$ $p < 0.001$

Null hypothesis is rejected

Advantages

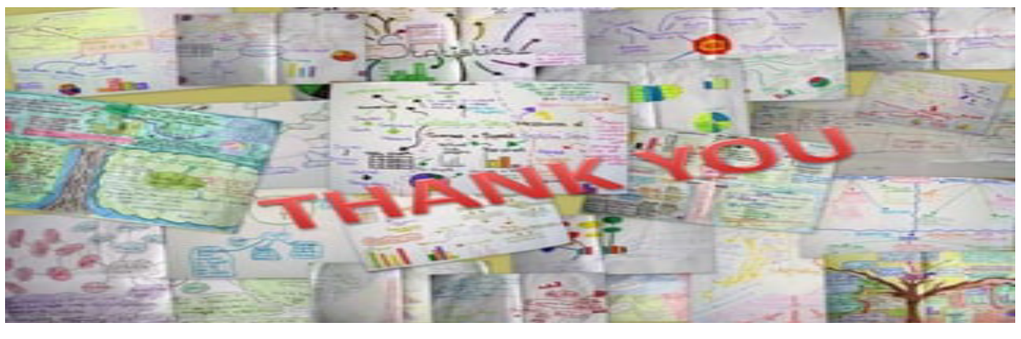
- Simpler model
  - Easier to build, test and understand than other models
- More reliable
  - More reliable as only one variable is used.
- Descriptive method
  - A univariate model is a strong descriptive method. Analysts can change one variable each time the model is run to obtain results that show "what if" scenarios. For example, changing the variables from age to income can show different results which describe what happens when one factor changes within the model.

Disadvantages

- Not comprehensive
  - A univariate model is less comprehensive compared to multivariate models. In the real world, there is often more than just one factor at play and a univariate model is unable to take this into account due to its inherent limitations.
- Does not establish relationships
  - As only one variable can be changed at a time, univariate models are unable to show relationships between different factors.

References

- Methods in Biostatistics *by* BK Mahajan
- Statistical Methods *by* SP Gupta
- Basic & Clinical Biostatistics *by* Dawson and Beth
- Park's textbook of preventive and social medicine 22nd edition



Testing of hypotheses

learning objectives:

- to understand the role of significance
- to distinguish the null and alternative hypotheses
- to interpret p-value, type I and II errors