Basic Statistics

Agenda

- Statistics
- Sampling
- Parts of the statistical process

What is the need?

- To evaluate printed numerical facts.
- To interpret the results of sampling or to perform statistical analysis in your work.
- To make inference about the population using information collected from the sample.

What Do Statisticians Do?

- Gather data
- Summarize data
- Analyze data
- Draw conclusions and report the results of their analysis

Collecting Data

Collecting Data

Methods	Challenges
Personal Interview	People usually respond when asked by a person but their answers may be influenced by the interviewer.
Telephone Interview	Cost-effective but need to keep it short since respondents tend to be impatient.
Self-Administered Questionnaires	Cost-effective but the response rate is lower and the respondents may be a biased sample.
Direct Observation	For certain quantities of interest, one may be able to measure it from the sample.
Web-Based Survey	Can only target the population who uses the web.

Strategies for Collecting Data

There are two types of methods for collecting data:

1. Non-probability methods
2. probability methods

Some Definitions

- POPULATION: Population is the collection of the elements which has some or the other characteristic in common. Number of elements in the population is the size of the population.
- SAMPLE: Sample is the subset of the population. The process of selecting a sample is known as sampling. Number of elements in the sample is the sample size.

Non-probability Methods

1. Convenience sampling

For instance, surveying students as they pass by in the university's student union building

2. Gathering volunteers

using an advertisement in a magazine or on a website inviting people to complete a form or participate in a study.

Probability Methods

1. Simple random sample

making selections from a population where each subject in the population has an equal chance of being selected.

2. Stratified random sample

where you have first identified population of interest, you then divide this population into strata or groups based on some characteristic (e.g. sex, geographic region), then perform simple random sample from each strata.

Probability Methods Contd..

3. Cluster sample

- Where a random cluster of subjects is taken from population of interest. An example might be grabbing handful of M&M's from a large jar of M&M's. Our entire population is divided into clusters or sections and then the clusters are randomly selected. All the elements of the cluster are used for sampling. Clusters are identified using details such as age, sex, location etc.
- Cluster sampling can be done in following ways:
- **Single Stage Cluster Sampling**
- Entire cluster is selected randomly for sampling.
- **Two Stage Cluster Sampling**
- Here first we randomly select clusters and then from those selected clusters we randomly select elements for sampling

Probability Methods Contd..

4. Multi Stage sample

- It is the combination of one or more methods described above.
- Population is divided into multiple clusters and then these clusters are further divided and grouped into various sub groups (strata) based on similarity. One or more clusters can be randomly selected from each stratum. This process continues until the cluster can’t be divided anymore. For example country can be divided into states, cities, urban and rural and all the areas with similar characteristics can be merged together to form a strata.

Airline Company Survey of Passengers..

- **Simple Random Sampling**:

You randomly select a set of passengers flying on your airline and question those that you have selected.

- **Stratified Sampling**:

You stratify your passengers by the class they fly (first, business, economy), and then take a random sample from each of these strata.

Airline Company Survey of Passengers..

- **Cluster Sampling**:

You stratify your passengers by class they fly (first, business, economy) and randomly select such classes from various flights and survey each passenger in that that class and flight selected.

Example: Airline Company Survey of Passengers

- Let's say that you are the owner of a large airline company and you live in Los Angeles. You want to survey your L.A. passengers on what they like and dislike about traveling on your airline.

Non Probabilistic Sampling

- **Convenience Sampling**:

Here the samples are selected based on the availability. This method is used when the availability of sample is rare and also costly. So based on the convenience samples are selected.

Since you live in L.A. you go to airport and just interview passengers as they approach your ticket counter.

- **Volunteer Sampling**:

You have your ticket counter personnel distribute a questionnaire to each passenger requesting they complete the survey and return it at end of flight.

Airline Company Survey of Passengers..

- **Purposive Sampling**

- This is based on the intention or the purpose of study. Only those elements will be selected from the population which suits the best for the purpose of our study.

- **Quota Sampling**

- This type of sampling depends of some pre-set standard. It selects the representative sample from the population. Proportion of characteristics/ trait in sample should be same as population. Elements are selected until exact proportions of certain types of data is obtained or sufficient data in different categories is collected.

Types of Studies

1. **Observational -** these studies show that there is a relationship.

2. **Experimental** – this involves some random assignment of a treatment; researchers can draw cause and effect (or causal) conclusions.

Example: Class Quizzes

- In an **observational study** we may find that better students tend to take the quizzes and do better on exams. Consequently, we might conclude that there may be a relationship between quizzes and exam scores.

- In an **experimental study** we would randomly assign quizzes to specific students to look for improvements. In other words, we would look to see whether taking quizzes causes higher exam scores.

Observational Studies Versus Scientific Studies

Observational Studies	Researcher observes the data and has no control over which subject takes which treatment
Scientific Studies	The researcher assigns randomly treatments to each subject - there is

Question??

- Decide which one is more effective in reducing fever:

  - **Paracetamol**
  - **Citrizine**

Question??..

Method 1: Ask the subjects which one they use and ask them to rate the effectiveness.

Question: Is this an observational study or scientific study?

Answer: This is an observational study since we just observe the data and have no control on which subject to use what type of treatment.

Question??..

- ***Method 2:*** Randomly assign half of the subjects to take Paracetamol and the other half to take Citrizine. Ask the subjects to rate the effectiveness.
- ***Question:*** Is this an observational study or scientific study?
- ***Answer:*** This is a scientific study since we can decide which subject to use what type of treatment. Thus the self selection bias will be eliminated.

Principles of Experimental Design

The following principles of experimental design have to be followed to enable a researcher to conclude that differences in the results of an experiment, not reasonably attributable to chance, are likely caused by the treatments.

Control	Need to control for effects due to factors other than the ones of primary interest.
Randomization	Subjects should be randomly divided into groups to avoid unintentional selection bias in the groups.

Types of Bias

- **Non-response** – Large percentage of those sampled do not to respond or participate.

- **Response** – When study participants either do not respond truthfully or give answers they feel the researcher wants to hear. For example, when students are asked if they ever cheated on an exam even those who have would respond with "no".

Types of Bias Contd..

- **Selection**

This bias occurs when the sample selected does not reflect the population of interest. For instance, you are interested in the attitude of female students regarding campus safety but when sampling you also include males. In this case your population of interest was female students however your sample included subject not in that population (i.e. males).

Graphing Data

Variables

Categorical and Numerical

Types of Variable

1. **Qualitative/Categorical**
2. **Quantitative**

Qualitative/Categorical Variable

- Data that serves the function of a name only. For example, for coding purposes, you may assign Male as 0, Female as 1. The numbers 0 and 1 stand only for the two categories and there is no order between them. Categorical values may be:
- Binary – where there are two choices, e.g. Male and Female;
- Ordinal – where the names imply levels with hierarchy or order of preference, e.g. level of education
- Nominal – where no hierarchy is implied, e.g. political party affiliation.

Quantitative

- Data that takes on numerical values that has a measure of distance between them. Quantitative values can be discrete, or “counted” as in the number of people in attendance, or continuous or “measured” as in the weight or height of a person.

Questions??

Q1: Number of females in this class

Ans: Quantitative, Discrete

Q2: Nationality

Ans: Categorical, nominal

Q3: Amount of milk in a 1 gallon container

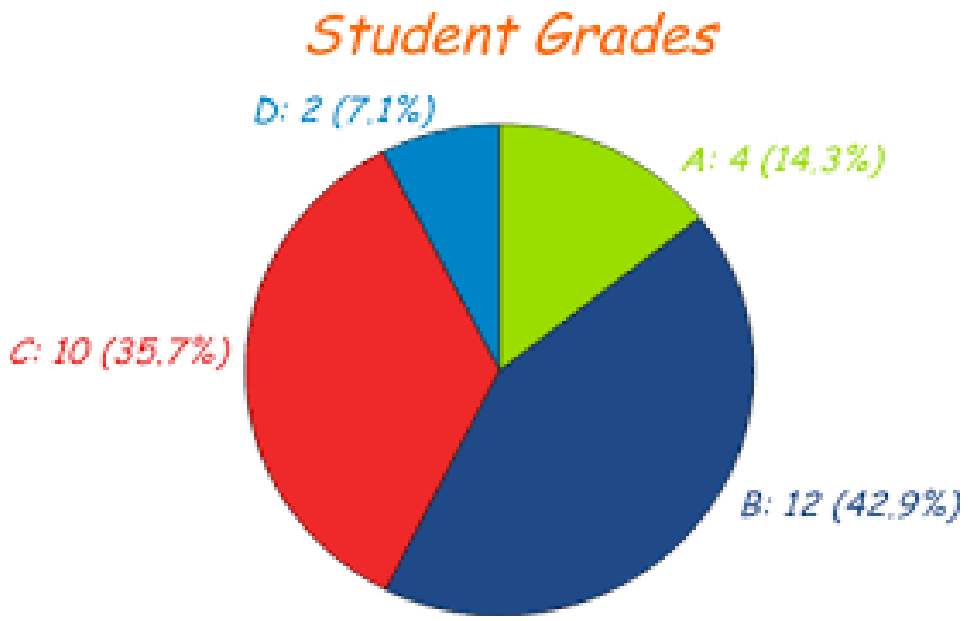
Ans: Quantitative, Continuous

Q4: Sex of students (even if coded as $M = 0$, $F = 1$)

Ans: Categorical, Binary

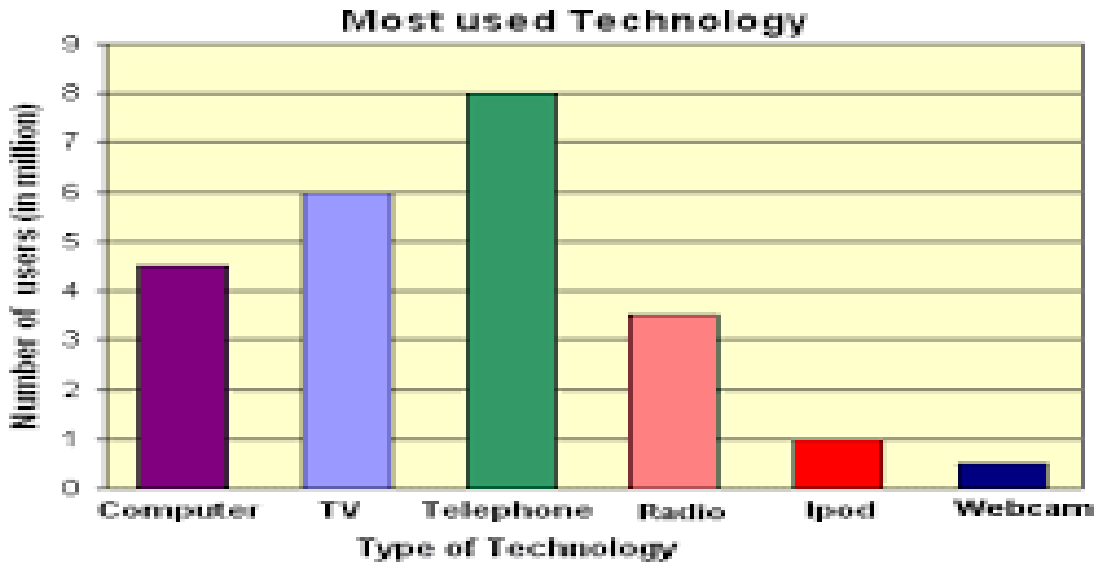
Graphs - Categorical Variable

1. **Pie Chart:** Area of the pie represents the percentage of that category.



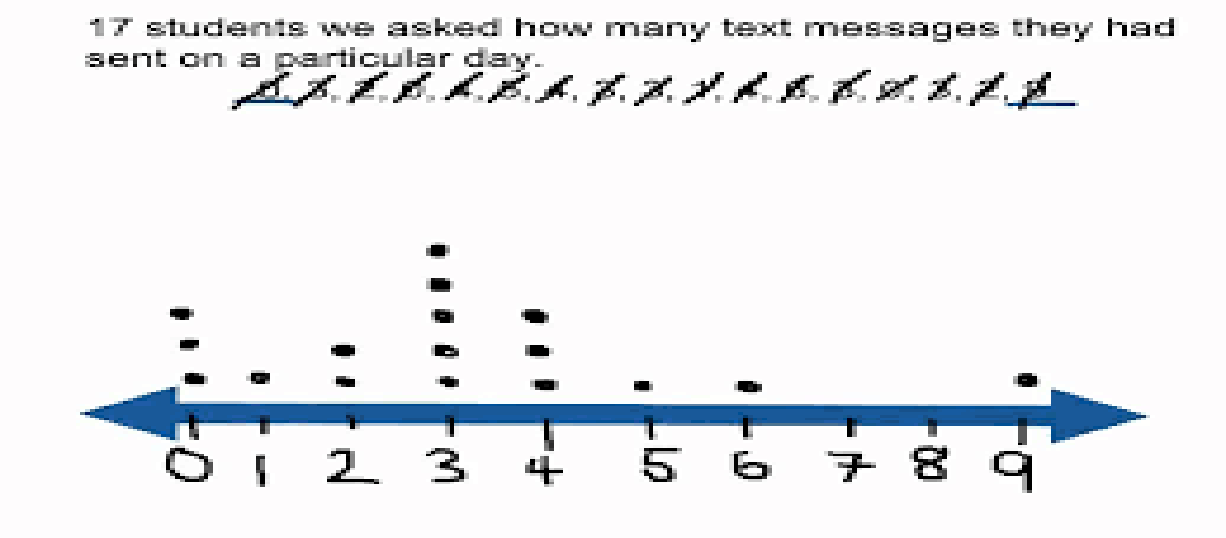
Graphs - Categorical Variable Contd..

2. Bar Chart: The height of the bar for each category is equal to the frequency (number of observations) in the category. Leave space in between the bars to emphasize that there is no ordering in the classes.



Graphs - Single Quantitative Variable

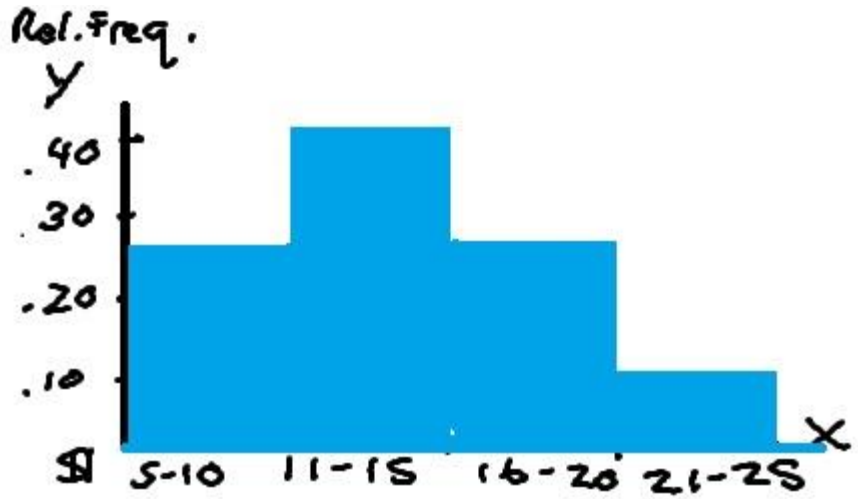
1. Dotplot: Useful to show the relative positions of the data.



Graphs - Single Quantitative Variable Contd..

2. Frequency Histogram & Relative Frequency Histogram: If there are many data points and we would like to see the distribution of the data, we can represent the data by a frequency histogram or a relative frequency histogram.

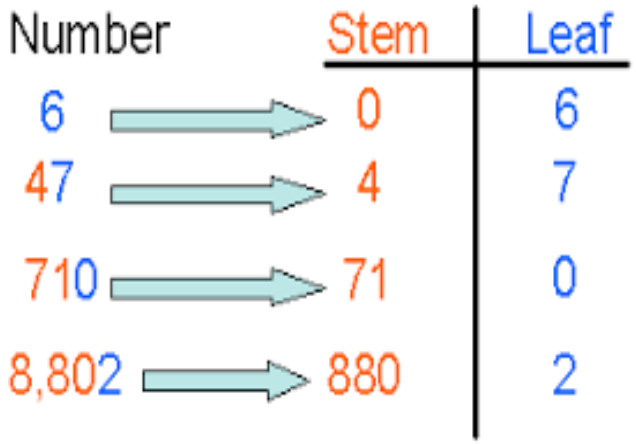
- Group the data into about 5-20 class intervals and show the frequency or relative frequency of data in each interval.



Graphs - Single Quantitative Variable Contd..

3. Stem-and-Leaf Diagram: Group the data and still keep the number.

One can recover the original data (except the order the data is taken) from the diagram. The stem represents the major groupings of the data. The leaves represent the last digit. For example, the first value (also smallest value) is 132, with 13 as the stem and 2 as the leaf.



Summarizing Data

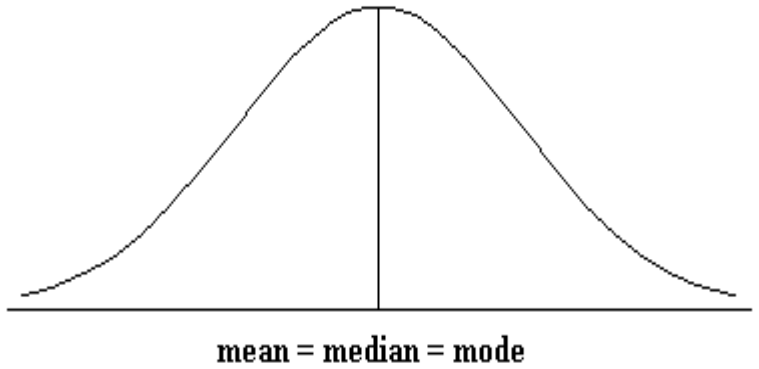
Measures of Central Tendency

1. Mean	the average of the data
2. Median	the middle value of the ordered data
3. Mode	the value that occurs most often in the data

Skewness

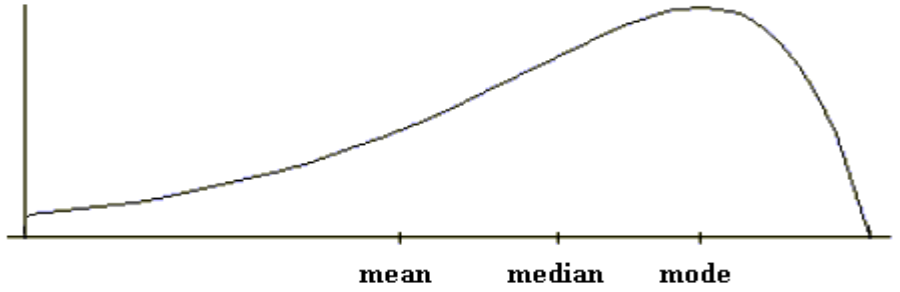
Skewness is a measure of degree of asymmetry of the distribution.

1. Symmetric - Mean, median, and mode are all the same here; the distribution is mound shaped, and no skewness is apparent. The distribution is described as symmetric.

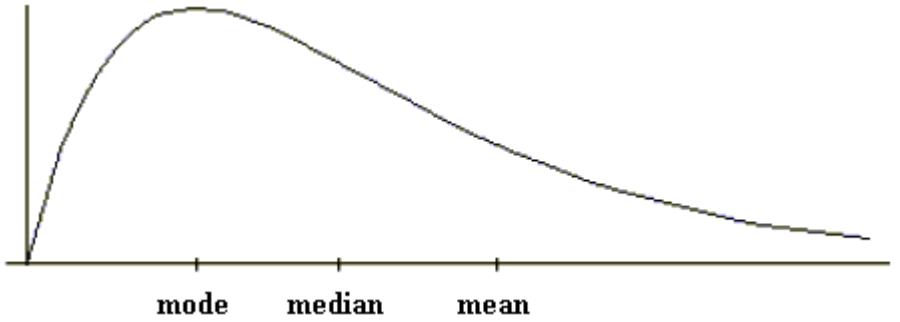


Skewness Contd..

2. Skewed Left - Mean to the left of the median, long tail on the left.



3. Skewed Right - Mean to the right of the median, long tail on the right.



Measures of Variability

There are many ways to describe variability including:

- Range
- Interquartile range (IQR)
- Variance and Standard deviation

Range

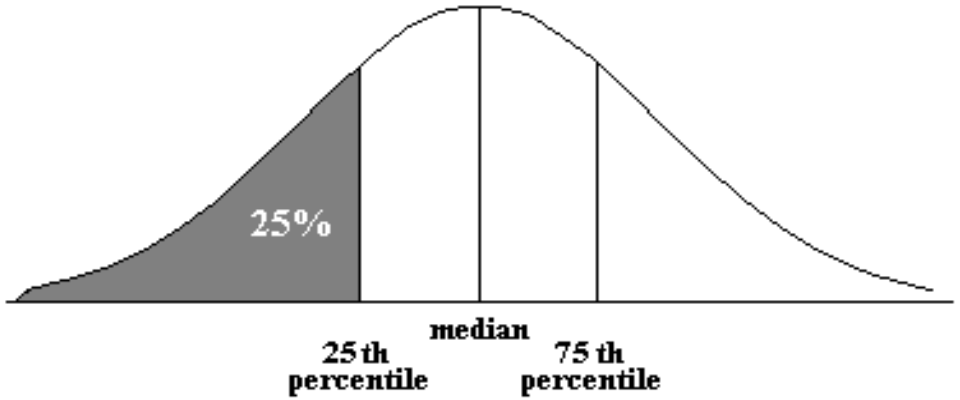
Range – R= Maximum - Minimum

- Easy to calculate
- Very much affected by extreme values (range is not a resistant measure of variability)

Interquartile range (IQR)

The **interquartile range** is the difference between upper and lower quartiles and denoted as **IQR**.

- IQR = Q_3 - Q_1 = upper quartile - lower quartile = 75th percentile - 25th percentile.
  - where Q_1 = lower quartile = the 25th percentile
  - Q_3 = upper quartile = the 75th percentile.



Variance

- Variance is the average squared distance from the mean.
- Population Variance is defined as: $$\sigma^2 = \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{N}$$
- In this formula μ is the population mean and the summation is over all possible values of the population. *N* is the population size.
- The sample variance that is computed from the sample and used to estimate σ^2 is:

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

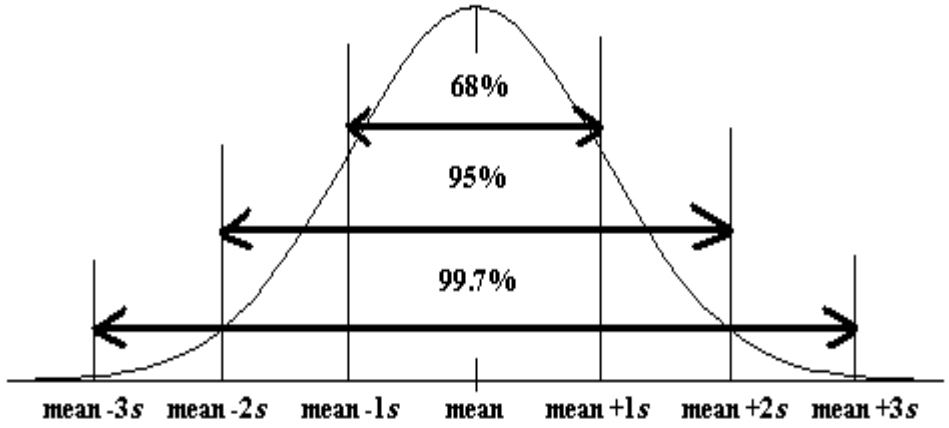
Standard Deviation

The standard deviation is approximately the average distance the values of a data set are from the mean, and is a very useful measure. One reason is that it has the same unit of measurement as the data itself (e.g. if a sample of student heights were in inches then so, too, would be the standard deviation.

Empirical Rule

Empirical Rule is sometimes referred to as the 68-95-99.7% Rule. If the set of measurements follows a bell-shaped distribution, then

$\bar{y} \pm s$	contains about 68% of data
$\bar{y} \pm 2s$	contains about 95% of data
$\bar{y} \pm 3s$	contains about all of data



Coefficient of Variation

In the last few slides, we considered three measures of variation:

1. Range
2. Interquartile Range (IQR)
3. Variance & Standard Deviation

These are all measures we can calculate from one quantitative variable e.g. height, weight.

But how can we compare dispersion (i.e. variability) of data from two or more distinct populations that have vastly different means? – Answer is *Coefficient of Variation* or CV

Coefficient of Variation Contd...

- This is a unit-free statistic and one where the higher the value the greater the dispersion.
- The calculation of CV is:

CV = Standard Deviation / Mean

Z-value, Z-score, or Z

- Z-value, or sometimes referred to as Z-score or simply Z, represents the number of standard deviations an observation is from the mean for a set of data.
- To find the z-score for a particular observation we apply the following formula:

$$Z = (\text{observed value} - \text{mean}) / SD$$

Example of Z Score

- For a recent final exam the mean was 68.55 with a standard deviation of 15.45

  - If you scored an 80%: Z = (80 - 68.55) / 15.45 = 0.74, which means your score of 80 was 0.74 SD above the mean.

  - If you scored a 60%: Z = (60 - 68.55) / 15.45 = -0.55, which means your score of 60 was 0.55 SD below the mean.

Box Plots

- Box plots can be use, when we want a graph that is not as detailed as a histogram, but still shows:

    1. the skewness of the distribution
    2. the central location
    3. the variability

Five Number Summary

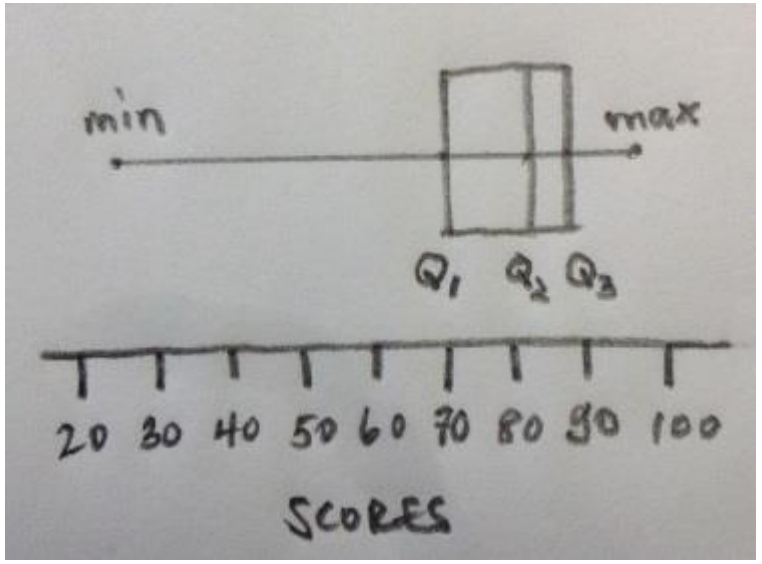
- To create this box plot we need the following:
  - minimum value,
  - Q_1 (lower quartile),
  - Q_2 (median),
  - Q_3 (upper quartile), and
  - maximum value.
- This list is also called the **five number summary**.

Five Number Summary Contd..

- Assume that the five number summary of final exam scores of 18 students is:

min	Q_1	*Median (Q_2)*	Q_3	max
24	70	82.5	89	97

- Using the five number summary, one can construct a skeletal box plot.
- Mark the five number summary above the horizontal axis with vertical lines.
- Connect Q_1, Q_2, Q_3 to form a box, then connect the box to min and max with a line to form the whisker.



Presenting data graphically

- We cannot do any analysis of the following data:

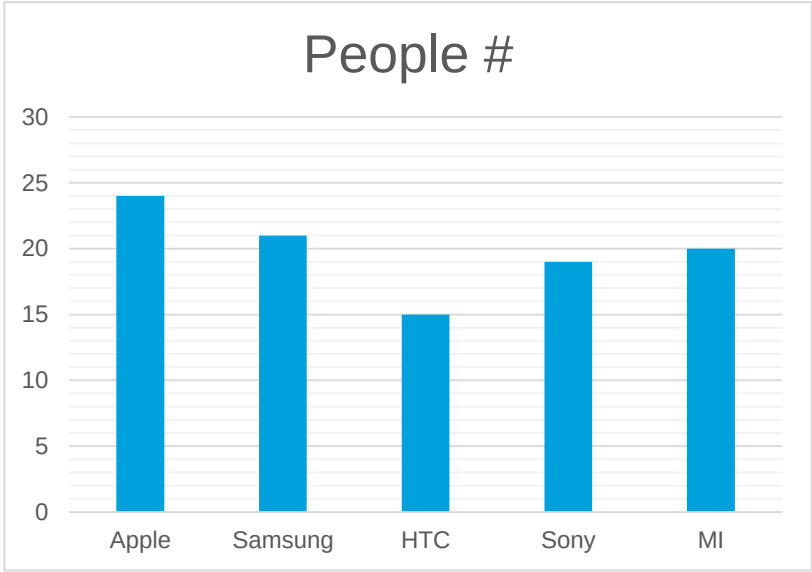
MOBILE PHONE SALES								
Samsung	HTC	Samsung	Sony	Sony	Apple	HTC	Sony	MII
Sony	Apple	Samsung	Apple	Sony	HTC	MII	Apple	Apple
HTC	MII	Sony	MII	Sony	HTC	Samsung	Apple	Samsung
Samsung	Apple	Samsung	HTC	Sony	MII	Samsung	Samsung	HTC
MII	Samsung	Sony	Apple	MII	Apple	MII	Apple	Sony
MII	Sony	MII	Apple	Samsung	HTC	MII	MII	Sony
MII	Samsung	Apple	Apple	Apple	Apple	Apple	Sony	Samsung
HTC	Apple	HTC	Samsung	Sony	Apple	Samsung	MII	Apple
MII	MII	MII	MII	Samsung	Sony	HTC	Apple	Apple
Samsung	Apple	Samsung	HTC	HTC	HTC	Sony	Apple	MII
Samsung	Sony	Sony	Sony	MII	Samsung	HTC	Samsung	Apple

Data needs to be transformed for analysis

Just create a table containing the counts of the people buying these phones .

Mobile	People #
Apple	24
Samsung	21
HTC	15
Sony	19
MI	20

It would be better if we can present this data in the form of a graph as below:



Hypothesis Testing

- A *hypothesis* is something that has not yet been proven to be true.
- It is an assertion/statement .
- Eg it can rain today, the college gives you 100% placement etc.
- Hypothesis testing is the process of determining whether or not a given hypothesis is true.

Null Hypothesis

- A null *hypothesis* is something that is the situation of status quo or no change
- A null hypothesis is an assertion about the value of a population parameter. It is an assertion that we hold as true unless we have sufficient statistical evidence to conclude otherwise.
- Hypothesis testing is the process of determining whether or not a given hypothesis is true.

Hypothesis Testing

- In the legal system, the accused is assumed innocent until proved guilty “beyond a reasonable
- doubt.” We will call this the null hypothesis—the hypothesis that the accused is innocent.
- We will hold the null hypothesis as true until a time when we can prove, beyond a
- reasonable doubt, that it is false and that the alternative hypothesis—the hypothesis that
- the accused is guilty—is true. We want to have a small probability (preferably zero) of
- convicting an innocent person, that is, of rejecting a null hypothesis when the null
- hypothesis is actually true

Errors in Hypothesis testing

- We often have to make an accept–reject type of decision based on incomplete data.
- Eg. 1. A recruiter has to accept or reject a job applicant, usually based on evidence gathered from a résumé and interview.

  2.A bank manager has to accept or reject a loan application, usually based on financial data on the application.

No error is committed when a good prospect is accepted or a bad one is rejected. But there is a small chance that a bad prospect is accepted or a good one is rejected. Of course, we would like to minimize the chances of such errors.

Type 1 & type 2 Error in Hypothesis testing

- TYPE 1 ERROR: Rejecting a true null hypothesis .
- TYPE 2 ERROR: Failing to Reject a false null hypothesis .

Instances of Type I and Type II Errors

	H_0 True	H_0 False
Accept H_0	No error	Type II error
Reject H_0	Type I error	No error

Type 1 & type 2 Error in Hypothesis testing

- TYPE 1 ERROR: Rejecting a true null hypothesis .
- TYPE 2 ERROR: Failing to Reject a false null hypothesis .
- Let us see if its possible to reduce Type 1 error.
- Is it possible, even with imperfect sample evidence, to reduce the probability of type I error all the way down to zero?
- The answer is yes. Just accept the null hypothesis no matter what the evidence is.
- But doing this would be foolish. Suppose you get a false null hypothesis, and now in order to reduce type 1 error, you have accepted it. Therefore now Type 2 error will occur.

Type 1 & type 2 Error in Hypothesis testing

- TYPE 1 ERROR: Rejecting a true null hypothesis .
- TYPE 2 ERROR: Failing to Reject a false null hypothesis .
- Let us see if its possible to reduce Type 2 error.
- Reject all null hypothesis, therefore we will never have a case where we failed to reject a null hypothesis when it was false.
- But we immediately realize that now you will have the Type 1 error , which is rejecting the null hypothesis when it is true.
- Therefore we cannot completely avoid Type 1 and Type 2 errors, we should look for an optimal solution.

Type 1 & type 2 Error in Hypothesis testing

- TYPE 1 ERROR: Rejecting a true null hypothesis .
- TYPE 2 ERROR: Failing to Reject a false null hypothesis .
- How to reduce Type 1 error.

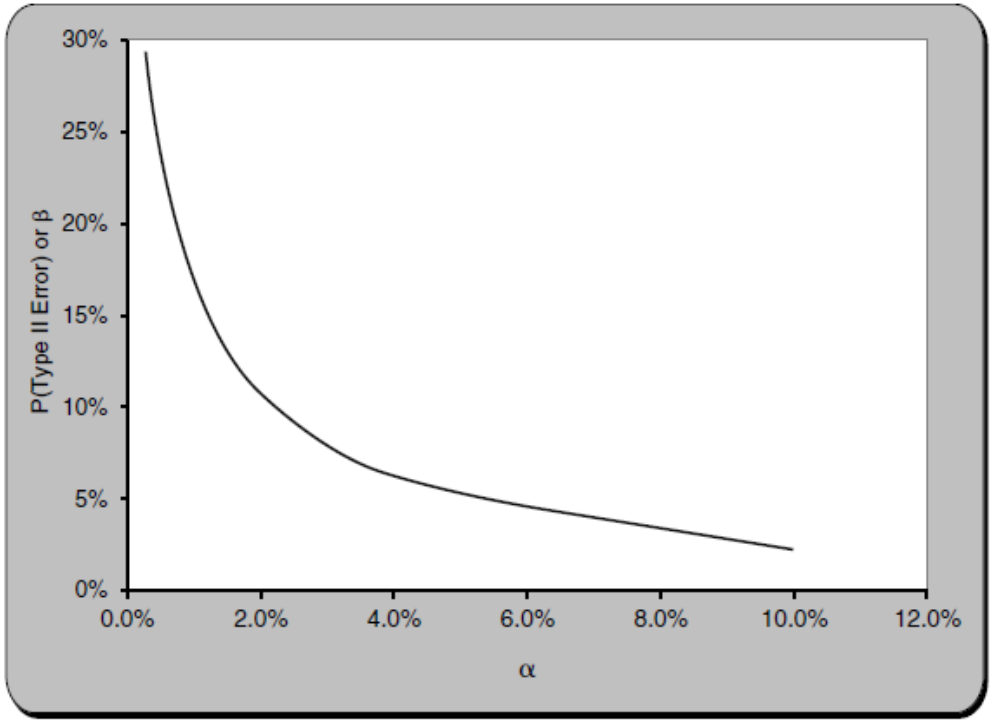
  Reduce the risk factor or the level of significance.
- How to reduce Type 2 error.

  Increase the sample size.

Type 1 & type 2 Error in Hypothesis testing

- Relationship between Type 1 and Type 2 error

FIGURE 7–1 Probability of Type II Error versus α for the Case H_0: $\mu \geq 1{,}000$, $\sigma = 10$, $n = 30$, Assumed μ for Type II Error = 994



Type 1 & type 2 Error in Hypothesis testing

- Which error to reduce?
- Suppose we are making some fasteners and calculating the average strength of them, in order to see if they are above a minimum standard.

Here Type 1 error will be rejecting good fasteners. Cost of Type 1 error will be cost of fasteneres.

Type 2 error:

It will be failing to reject bad bolts. Cost will depend on what kind of use we are putting these fasteners to.

Eg. If we are using the bolts for fastening garbage covers then we can reduce alpha, so we reduce Type 1 error. (alpha =1%)

However, if we are using them in pacemakers, then we need to increase alpha (say 10%) .

When we do not know about the relative cost, we keep it as 5%

ANOVA

- This is used for comparing different population means
- Hence we have to study different variances, hence we have analysis of variance.
- Eg , we want to check if fertilizers A, B, C and D are applied on a plant, which is the best ?
- We conduct the ANOVA test in these cases.
- The required assumptions of ANOVA:

1. We assume *independent random sampling* from each of the populations.

2. We assume that the *r* populations under study are *normally distributed,* with means that may or may not be equal, but with *equal variances*

Thank You!