Introduction to

Statistical Data Analysis I

JULY 2011

Afsaneh Yazdani

Preface

What is Statistics?

Preface

What is Statistics?

Science of:

- designing studies or experiments, collecting data
- Summarizing/modeling/analyzing data for the purpose of decision making/scientific discovery
- when the available information is both limited and variable.

Preface

What is Statistics?

Science of:

- designing studies or experiments, collecting data
- Summarizing/modeling/analyzing data for the purpose of decision making & scientific discovery
- when the available information is both limited and variable.

Statistics is the science of Learning from Data

Preface

Learning from Data

Preface

Learning from Data

Qualitative or quantitative attributes of a variable

Data are typically the results of measurements or observations of a variable

Data are often viewed as the lowest level of abstraction from which information and then knowledge are derived.

Preface

Learning from Data

Four-step process by which we can learn from data:

1. Defining the Problem
2. Collecting the Data
3. Summarizing the Data
4. Analyzing Data, Interpreting the Analyses, and Communicating the results

Preface

Defining the Problem

- Understanding the problem being addressed
- Specifying the objective of the study
- Identifying the variables of interest
  - Reviewing the previous studies
  - Brainstorming of the experts
  - Importance rating of the factors

Preface

Data Gathering

The most appropriate method to collect the data, should be selected. Data collection processes include:

Preface

Data Gathering

The most appropriate method to collect the data, should be selected. Data collection processes include:

- Surveys

Surveys are passive. The goal of the survey is to gather data on existing conditions, attitudes, or behaviors, without any interference.

Preface

Data Gathering

The most appropriate method to collect the data, should be selected. Data collection processes include:

- Surveys
- Experiments

Experimental studies, tend to be more active. The person conducting the study, varies the experimental conditions to study the effect of the conditions on the outcome of the experiment.

Preface

Data Gathering

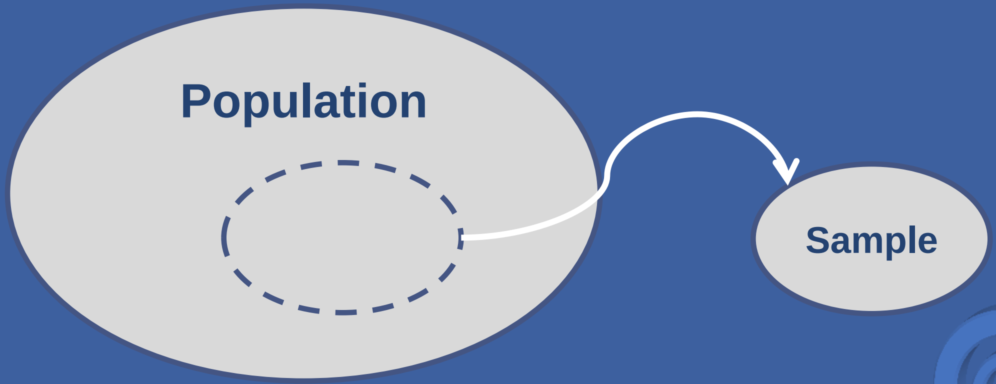
The most appropriate method to collect the data, should be selected. Data collection processes include:

- Surveys
- Experiments
- Examination of existing data from business records, censuses, government records, and previous studies

Preface

Using Surveys to gather data

The manner in which the sample is selected from the population **(sampling design)** must be determined, so that the sample accurately reflects the population as a whole **(representative sample)**



Preface

What can affect **representativeness** of a sample?

- Sampling Design
- Nonresponse
- Measurement Problems
  - Inability to recall answers to questions
  - Leading questions
  - Unclear wording of questions, …

Preface

Data Collection Techniques

- Personal Interview
- Telephone Interview
- Self-Administered Questionnaire
- Direct Observation

Preface

Data Collection Techniques

- Personal Interview
- Telephone Interview
- Self-Administered Questionnaire
- Direct Observation

Advantages:
Higher Response Rate

Disadvantages:
Cost
Trained Interviewer

Preface

Data Collection Techniques

- Personal Interview
- Telephone Interview
- Self-Administered Questionnaire
- Direct Observation

Advantages:
Low Cost
Easier to Monitor Interviewer

Disadvantages:
Unavailability of a list that closely corresponds to the population,
People who screen calls before answering,
Interview must be short

Preface

Data Collection Techniques

- Personal Interview
- Telephone Interview
- Self-Administered Questionnaire
- Direct Observation

Advantages:
Cost

Disadvantages:
Low Response Rate
Difficult to word questionnaire

Preface

Data Collection Techniques

- Personal Interview
- Telephone Interview
- Self-Administered Questionnaire
- Direct Observation

Advantages:
Data not-affected by respondent

Disadvantages:
Possibility of error in observation
Time-consuming

Preface

Using Experiments to gather data

In experimental studies, researcher should follow a systematic plan (experiment plan) prior to running the experiment. The plan includes:

- Research objectives of the experiment
- Selection of the factors that will be varied (treatments)

Preface

Using Experiments to gather data

In experimental studies, researcher should follow a systematic plan (experiment plan) prior to running the experiment. The plan includes:

- Identification of extraneous factors that may be present in the experimental units or in the environment of the experimental setting (blocking factors)

Preface

Using Experiments to gather data

In experimental studies, researcher should follow a systematic plan (experiment plan) prior to running the experiment. The plan includes:

- Characteristics to be measured on the experimental units (response variable)

Preface

Using Experiments to gather data

In experimental studies, researcher should follow a systematic plan (experiment plan) prior to running the experiment. The plan includes:

- Method of randomization, either “randomly selecting from treatment populations” or the “random assignment of experimental units to treatments”

Preface

Using Experiments to gather data

In experimental studies, researcher should follow a systematic plan (experiment plan) prior to running the experiment. The plan includes:

- Procedures to be used in recording the responses from the experimental units

Preface

Using Experiments to gather data

In experimental studies, researcher should follow a systematic plan (experiment plan) prior to running the experiment. The plan includes:

- Selection of the number of experimental units assigned to each treatment may require designating the level of significance and power of tests or the precision and reliability of confidence intervals

Preface

Experimental Designs

- Completely Randomize Design
- Randomized Block Design
- Latin Square Design
- Factorial Treatment Structure

Preface

Experimental Designs

- Completely Randomize Design

- When comparing *t* treatments (*t* levels of a single factor)
- An SRS sample of observations for each treatment
- Sample size for each treatment can be different

Comparing Tire Wear for 4 Brands of Tire

Car 1	Car 2	Car 3	Car 4
Brand B	Brand A	Brand A	Brand D
Brand B	Brand A	Brand B	Brand D
Brand B	Brand C	Brand C	Brand D
Brand C	Brand C	Brand A	Brand D

Preface

Experimental Designs

- Completely Randomize Design
- Randomized Block Design (similar to stratified random sample)

> ➢ When comparing *t* treatments
> ➢ Each treatment is assigned to a block (homogenous group)

Comparing Tire Wear for 4 Brands of Tire

Car 1	Car 2	Car 3	Car 4
Brand A	Brand A	Brand A	Brand A
Brand B	Brand B	Brand B	Brand B
Brand C	Brand C	Brand C	Brand C
Brand D	Brand D	Brand D	Brand D

Preface

Experimental Designs

- Completely Randomize Design
- Randomized Block Design (similar to stratified random sample)
- Latin Square Design

> When comparing *t* treatments
> Having two blocking variables

Comparing Tire Wear for 4 Brands of Tire

Position	Car 1	Car 2	Car 3	Car 4
Right/Front	Brand A	Brand B	Brand C	Brand D
Left/Front	Brand B	Brand C	Brand D	Brand A
Right/Back	Brand C	Brand D	Brand A	Brand B
Left/Back	Brand D	Brand A	Brand B	Brand C

Preface

Experimental Designs

- Completely Randomize Design
- Randomized Block Design (similar to stratified random sample)
- Latin Square Design
- Factorial Treatment Structure

➢ several factors rather than just *t* levels of a single factor whether or not interaction exists

Summarizing Data

Major branches of Statistics:

- Descriptive Statistics

  - the set of measurements available is frequently the entire population
  - making sense of the data by reducing a large set of measurements to a few summary measures which reflect good picture of the whole

Summarizing Data

Major branches of Statistics:

- Descriptive Statistics

- Inferential Statistics

- Measurements of all units in the population are not available.
- A sample is drawn to draw conclusions about the population

Summarizing Data

Major branches of Statistics:

- Descriptive Statistics

- Inferential Statistics

For both, an insightful description of the data is an important step in drawing conclusions from it.

Summarizing Data

Major Methods of Data Describing:

- Graphical Techniques
- Numerical Descriptive Techniques

More commonly used Soft-wares:
SAS, SPSS,
Minitab,
STATA,
R

Summarizing Data

Describing Data on a Single Variable: Graphical Methods

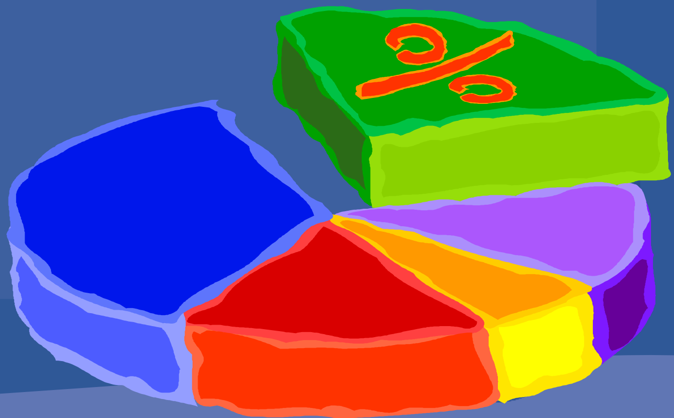
- Categorized/Nominal/Ordinal Measurement
- Continuous Measurement

Summarizing Data

Graphs

1. Pie Chart:

Used to display the percentage of the total number of measurements falling into each category of the variable by partitioning a circle (similar to slicing a pie).

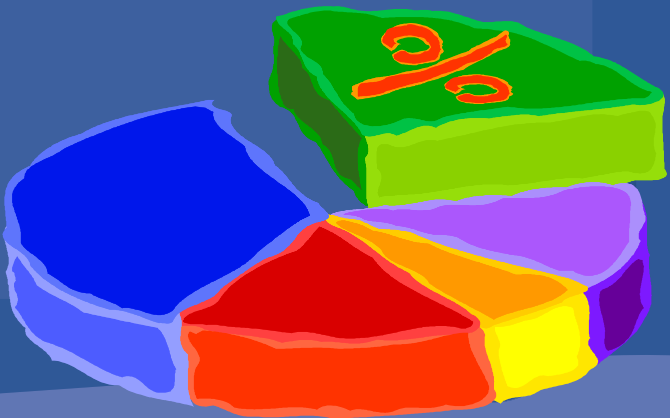


Summarizing Data

Graphs

Guidelines for Constructing Pie Charts:

1. Choose a small number of categories for the variable because too many make the pie chart difficult to interpret.
2. Whenever possible, sort the percentages in either ascending or descending order.

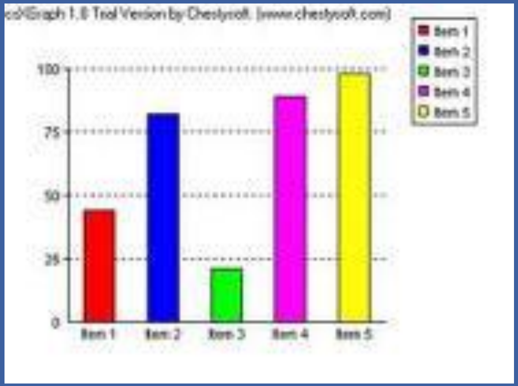


Summarizing Data

Graphs

2. Bar Chart:

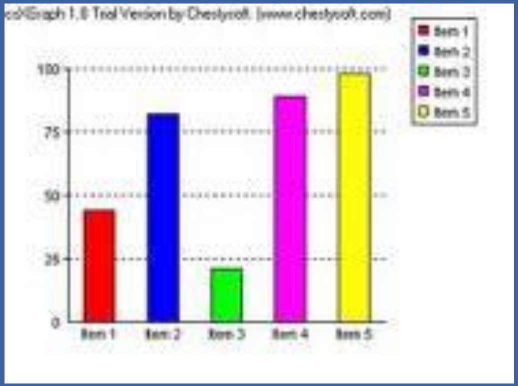
Used to display the frequency or value of measurements of each category of variable (sometimes across time)



Summarizing Data
Graphs

Guidelines for Constructing Bar Charts:

1. Label values on one axis and categories of the variable on the other axis.
2. The height of the bar is equal to the frequency or value.
3. Leave a space between each category for distinction.

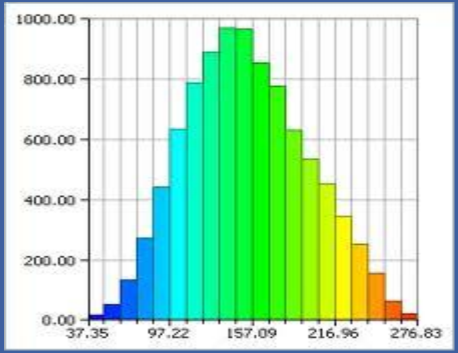


Summarizing Data

Graphs

3. Histogram:

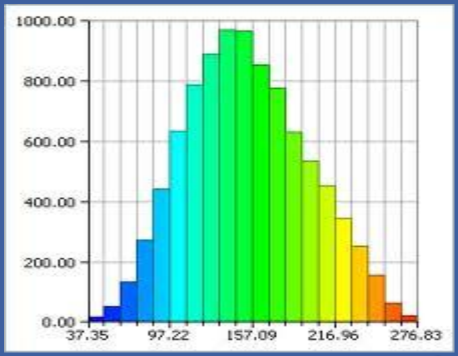
Frequency/Relative Frequency Histogram, applicable only to **quantitative/continuous** measurements, used to show the distribution of a variable



Summarizing Data
Graphs

Guidelines for Constructing Histograms:

1. Divide the range of the measurements by the approximate number of class intervals desired. (generally 5-20), then round it to get common width for class intervals.

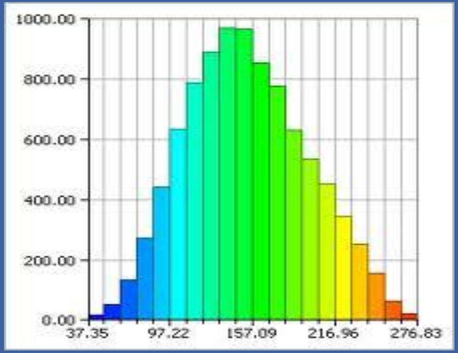


Summarizing Data
Graphs

Guidelines for Constructing Histograms:

1. Divide the range of the measurements by the approximate number of class intervals desired. (generally 5-20), then round it to get common width for class intervals.

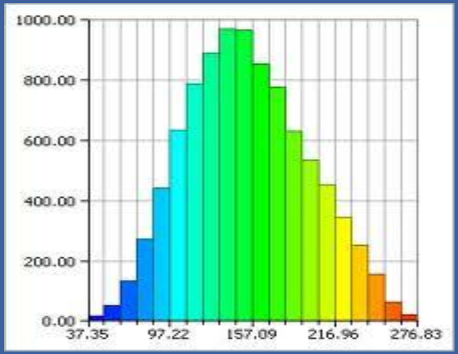
Too small number, can hide most of the patterns or trends in the data



Summarizing Data
Graphs

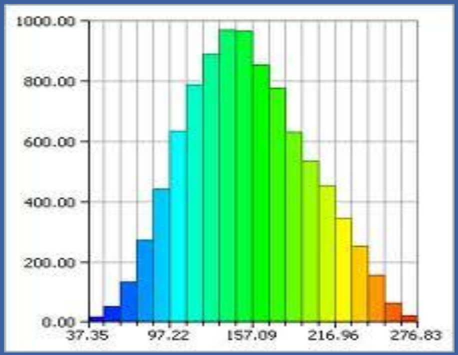
Guidelines for Constructing Histograms:

2. Choose the first class interval so that it contains the smallest measurement
3. Construct the frequency/ relative frequency table



Guidelines for Constructing Histograms:

4. Construct the histogram based on the frequency/ relative frequency of class intervals.



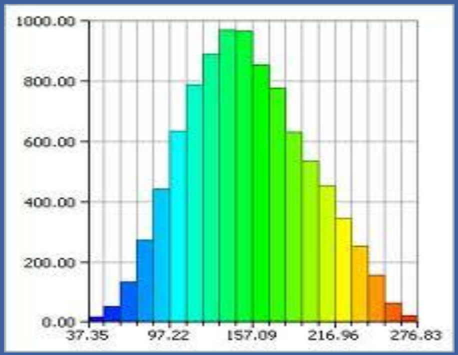
Summarizing Data

Graphs

Guidelines for Constructing Histograms:

4. Construct the histogram based on the frequency/ relative frequency of class intervals.

each rectangle is constructed over each class interval with a height equal to the class frequency or relative frequency



4. Stem-and-leaf Plot:

A clever, simple device for constructing a histogram-like picture of a frequency distribution

8.	0	0						
9.	0							
10.	0	0						
11.	0	0	5					
12.	0	0	0	2				
13.	2	5	8	8				
14.	0	0	0	0	4	6	8	
15.	0	0	5					
16.	0	2	6	8				
17.	0	0	5					
18.	0	2	5					
19.	0	5						
20.	0	5						

Guidelines for Constructing Stem-and-leaf Plot:

1. Split each score or value into two sets of digits. The first or leading set of digits is the stem and the second or trailing set of digits is the leaf.
2. List all possible stem digits from lowest to highest.
3. For each score in the mass of data, write the leaf values on the line labeled by the appropriate stem number.

Summarizing Data

Graphs

Guidelines for Constructing Successful Graphics:

- Before constructing a graph, set your priorities.
- Choose the appropriate type of graph.
- Pay attention to the titles.
- Use the colors cleverly.

Summarizing Data

Describing Data on a Single Variable: Numerical Measures

Numerical descriptive measures are commonly used to convey a mental image of pictures, objects, and other phenomena. The two most common numerical descriptive measures are:

- Measures of central tendency (location/position)
- Skewness
- Measures of variability

Summarizing Data

Measures of Central Tendency

Describe center of the distribution of measurements

Mode

Mode of a set of Measurements is defined to be the measurement that occurs most often (with the highest frequency).

Distributions can be Unimodal or bi-modal

Mode of grouped data

Modal Interval is the interval with the highest frequency.

Mode is the mid-point of the modal interval

Summarizing Data

Measures of Central Tendency

Describe center of the distribution of measurements

Median

The median of a set of measurements is defined to be the middle value when the measurements are arranged from lowest to highest.

The median for an even number of measurements is the average of the two middle values.

Median of grouped data

L: Lower limit of the class interval containing the median

n: Total frequency

cfb: Cumulative frequency for all class intervals before median class interval

fm: Frequency of class interval containing the median

w: interval width

$$med. = L + \frac{w}{fm}(0.5n - cfb)$$

Summarizing Data

Measures of Central Tendency

Describe center of the distribution of measurements

Mean

The arithmetic mean or mean of a set of measurements is defined to be the sum of measurements divided by the total number of measurements. The population (sample) mean is denoted by 'μ' ($\bar{y}$)

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N}, \; \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Mean of grouped data

y_i: Mid-point of the i-th class interval

f_i: Frequency of the i-th class interval

n: Total number of measurements ($\sum_{i=1}^{k} f_i$)

$$\bar{y} = \frac{\sum_{i=1}^{k} f_i y_i}{n}$$

Large number of class intervals, closer to actual mean.

Summarizing Data

Measures of Central Tendency

Describe center of the distribution of measurements

Geometric Mean

Geometric mean of a set of measurements is:

$$G = \sqrt[n]{(X_1 X_2 \ \dots \ X_n)}$$

Appropriate for ratios,
Only for data sets containing positive observations

Harmonic Mean

Harmonic mean of a set of measurements is:

$$H = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \ \dots + \frac{1}{X_n}}$$

Appropriate for rates of changes
Only for data sets containing positive observations

Major Characteristics of "Mode"

1. It is the most frequent or probable measurement in the data set.
2. There can be more than one mode for a data set.
3. It is not influenced by extreme measurements.
4. Modes of subsets cannot be combined to determine the mode of the complete data set.
5. For grouped data its value can change depending on the categories used.
6. It is applicable for both qualitative/quantitative data.

Summarizing Data

Measures of Central Tendency

Major Characteristics of "Median"

1. It is the central value; 50% of the measurements lie above it and 50% fall below it.
2. There is only one median for a data set.
3. It is not influenced by extreme measurements.
4. Medians of subsets cannot be combined to determine the median of the complete data set.
5. For grouped data, its value is rather stable even when the data are organized into different categories.
6. It is applicable to quantitative data only.

Major Characteristics of "Mean"

1. It is the arithmetic average of the measurements in a data set.
2. There is only one mean for a data set.
3. Its value is influenced by extreme measurements; trimming can help to reduce the degree of influence. (50% trimmed mean is the median)
4. Means of subsets can be combined to determine the mean of the complete data set.
5. It is applicable to quantitative data only.

Major Characteristics of "Mean"

1. ...is the arithmeti... age of th... urements in a data ...
2. Ther...

(3...

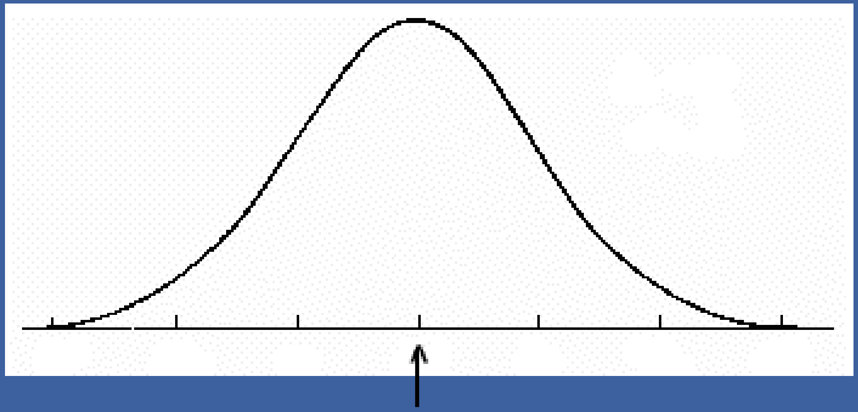
4. ... the mean of the co...
5. It is app... ativ... ly.

Measures of central tendency do not provide a complete picture of the frequency distribution for a set of measurements.

Summarizing Data
Skewness

Skewness is:

How mode, median, and mean are related for a set of measurements?



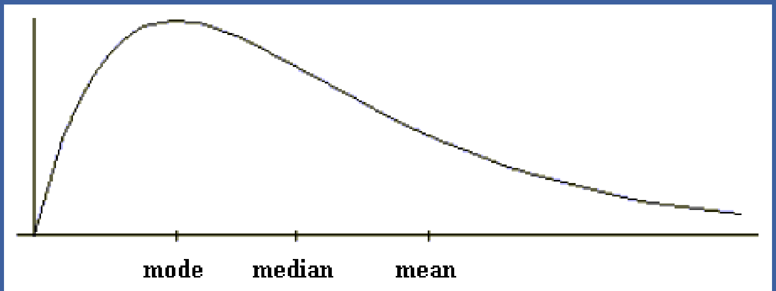
Mode = Median = Mean

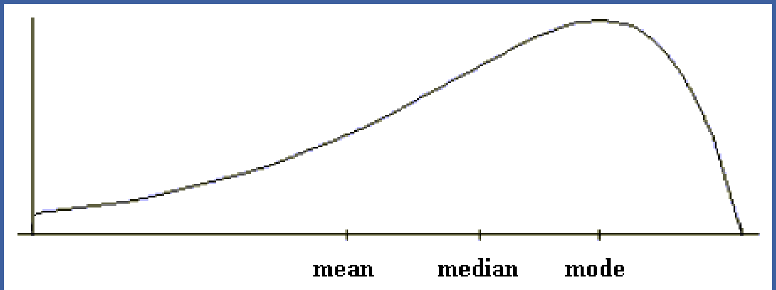
Summarizing Data

Skewness

Skewness is:

How mode, median, and mean are related for a set of measurements?





Mode < Median < Mean

Mean < Median < Mode

Summarizing Data
Skewness

Skewness is:

How mode, median, and mean are related for a set of measurements?

$$m_3 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^3}{n-1}$$

$$standardized\ skewness\ measure = \frac{m_3}{s^3}$$

Describe variability of measurements

Range

The range of a set of measurements is defined to be the difference between the largest and the smallest measurements of the set.

- Easy to compute
- Sensitive to outliers
- Does not give much information about the pattern of variability

Range of grouped data

The difference between the upper limit of the last interval and the lower limit of the first interval.

Summarizing Data

Measures of Variability

Describe variability of measurements

Percentiles

The p-th percentile of a set of n measurements arranged in order of magnitude is that value that has at most p% of the measurements below it and at most (100 p)% above it.

25th, 50th, and 75th percentiles, often called: the lower quartile, the middle quartile (median), and the upper quartile, respectively

Percentiles of grouped data

L: Lower limit of the class interval containing the percentile

n: Total frequency

cfb: Cumulative frequency for all class intervals before percentile class interval

fm: Frequency of class interval containing the percentile

w: interval width

$$P_{90} = L + \frac{w}{fm}(0.9n - cfb)$$

Describe variability of measurements

Interquartile Range

The interquartile range (IQR) of a set of measurements is defined to be the difference between the upper and lower quartiles; that is,

IQR=75th percentile - 25th percentile

IQR=3rd quartile – 1st quartile

Interquartile Range

1. Less Sensitive to extreme measurements
2. Misleading when data concentrates about the median
3. Covers only the variability of 50% of measurements
4. Less useful for a single set of measurements, quite useful for comparing variability of two or more data sets

Summarizing Data

Measures of Variability

Describe variability of measurements

Variance

The variance of a set of *n* measurements $y_1, y_2, \ldots. y_n$ with mean $\bar{y}$ is :

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

The population (sample) variance is denoted by σ^2 (s^2)

Variance of grouped data

y_i: Mid-point of the *i*-th class interval

f_i: Frequency of the *i*-th class interval

n: Total number of measurements $(\sum_{i=1}^{k} f_i)$

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

Large number of class intervals, closer to actual sample variance

Summarizing Data

Measures of Variability

Describe variability of measurements

Standard Deviation

The standard deviation of a set of measurements is defined to be the positive square root of the variance

It yields a measure of variability having the same units of measurements as the original data.

Standard Deviation is appealing because:

❶ We can compare the variability of two or more sets of data using the standard deviation

Summarizing Data

Measures of Variability

Describe variability of measurements

Standard Deviation

The standard deviation of a set of measurements is defined to be the positive square root of the variance

It yields a measure of variability having the same units of measurements as the original data.

Standard Deviation is appealing because:

❷ We can use the results of the "Empirical Rule" that follows to interpret the standard deviation of a single set of measurements

Summarizing Data

Measures of Variability

Empirical Rule:

Given a set of n measurements possessing a mound-shaped histogram, then:

- the interval $\bar{y} \pm s$ contains approximately 68% of the measurements
- the interval $\bar{y} \pm 2s$ contains approximately 95% of the measurements
- the interval $\bar{y} \pm 3s$ contains approximately 99.7% of the measurements.

Summarizing Data

Measures of Variability

Empirical Rule:

Given a set of n measurements possessing a mound-shaped histogram, then:

- the interval $\bar{y} \pm s$ contains approximately **68%** of the measurements
- the interval $\bar{y} \pm 2s$ contains approximately **95%** of the measurements
- the interval $\bar{y} \pm 3s$ cont… of the measurements.

Approximate value for '*s*':

$$s = \frac{\text{Range}}{4}$$

(better to overestimate)

Summarizing Data

Measures of Variability

Describe variability of measurements

Coefficient of Variability

In a process or population with mean μ and standard deviation σ, the coefficient of variation is defined as:

$CV = \frac{\sigma}{|\mu|}$, provided $\mu \neq 0$,

sometimes expressed as a percentage $CV = 100\frac{\sigma}{|\mu|}\%$

CV is unit-free so it can be used to:

Compare the variability in two considerably different processes or populations.

If CV is 15%, the standard deviation of the population is 15% of its mean

Describe variability of measurements

Coefficient of Variability

In a process or population with mean μ and standard deviation σ, the coefficient of variation is defined as:

$CV = \frac{\sigma}{|\mu|}$, provided $\mu \neq 0$,

sometimes expressed as a percentage $CV = 100\frac{s}{|\mu|}\%$

CV is unit-free so it can be used to:

as an index of population variability

i.e.
populations with similar CVs, have similar variability

Summarizing Data

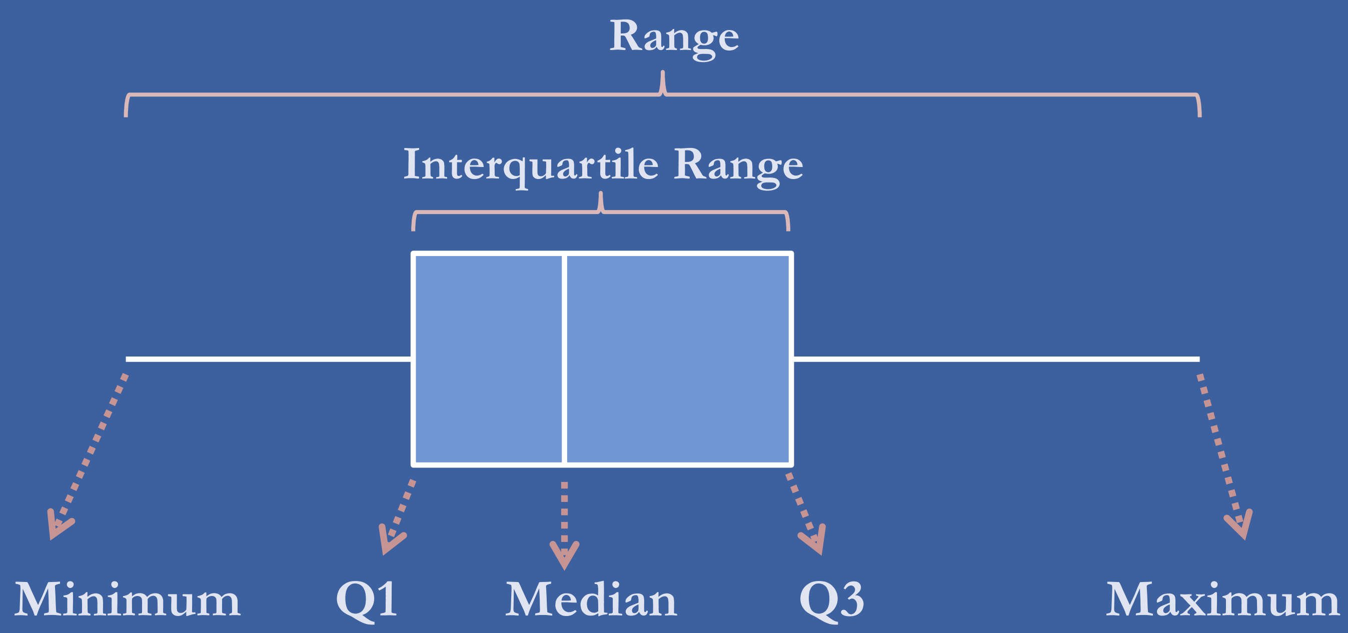
Measures of Variability

Summary Values Required for minimal description:

Minimum, Lower Quartile,
Median,
Higher Quartile, Maximum

Summarizing Data

5. Boxplot:



Summarizing Data

5. Boxplot:

With a quick glance, it gives an impression about:

1. The lower and upper quartiles, Q1 and Q3
2. The interquartile range (IQR)
3. The most extreme (lowest and highest) values
4. The symmetry or asymmetry of the distribution

	Lower	Upper
Inner Fence	Q1-1.5 IQR	Q3+1.5 IQR
Outer Fence	Q1-3 IQR	Q3+3 IQR

Summarizing Data

Summarizing Data from More Than One Variable: Graphs and Correlation

- Contingency table
- Graphs
- Correlation Coefficient

Summarizing Data from More Than One Variable: Contingency Table

Cross-tabulation to develop percentage comparisons to be used to describe the relationship between two qualitative variables.

Summarizing Data from More Than One Variable: Graphs

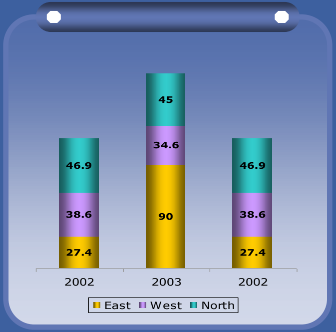
- Stacked bar graph
- Cluster bar graph
- Scatter plot
- Side-by-side boxplots

Summarizing Data

Summarizing Data from More Than One Variable: Graphs

- Stacked bar graph
- Cluster bar graph
- Scatter plot
- Side-by-side boxplots

Extension of bar chart, for a pair of qualitative variables



Summarizing Data from More Than One Variable: Graphs

- Stacked bar graph
- Cluster bar graph
- Scatter plot
- Side-by-side boxplots

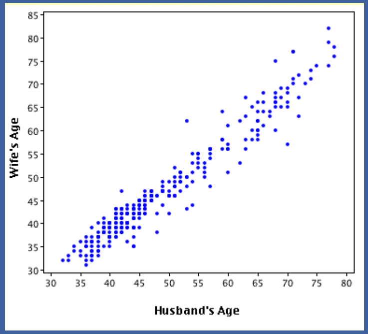
Extension of bar chart, for a single quantitative and a qualitative variable

Summarizing Data

Summarizing Data from More Than One Variable: Graphs

- Stacked bar graph
- Cluster bar graph
- Scatter plot
- Side-by-side boxplots

Visual assessment of relationship between two quantitative variables

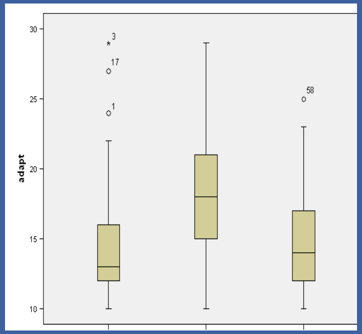


Summarizing Data

Summarizing Data from More Than One Variable: **Graphs**

- Stacked bar graph
- Cluster bar graph
- Scatter plot
- Side-by-side boxplots

Visual assessment of distribution of quantitative variables.



Summarizing Data from More Than One Variable: Correlation Coefficient

The correlation coefficient measures the strength of the linear relationship between two quantitative variables. The correlation coefficient is usually denoted as ‘r’.

Summarizing Data

Summarizing Data from More Than One Variable: Correlation Coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Summarizing Data from More Than One Variable: Correlation Coefficient Properties

- Value of ‘r’ is a number between -1 and 1. Closer value to ± 1 means stronger association between variables.
- A positive (negative) value for ‘r’ indicates a positive (negative) association between the two variables.

Summarizing Data from More Than One Variable: Correlation Coefficient Properties

- ‘r’ is a unit-free measure of association.
- ‘r’ measures the degree of straight line relationship between two variables not a curved relationship, no matter how strong the relationship is.