Basic Statistics

Agenda – Basic Statistics

1. Data Types – Continuous, Discrete, Nominal, Ordinal, Interval, Ratio, Random Variable, Probability, Probability Distribution
2. First, second, third & fourth moment business decisions
3. Graphical representation – Barplot, Histogram, Boxplot, Scatter diagram
4. Hypothesis Testing
5. Simple Linear Regression

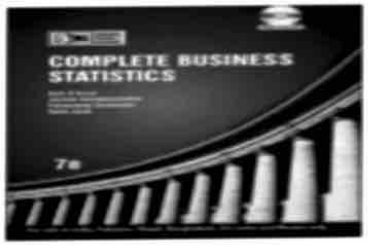
STATISTICS

Statistics is the science of data. It involves

- ✓collecting,
- ✓classifying,
- ✓summarizing,
- ✓analyzing,
- ✓and interpreting numerical information.

Statistics is used in several different disciplines (both scientific and non-scientific) to make decisions and draw conclusions based on data.

Book for the Course



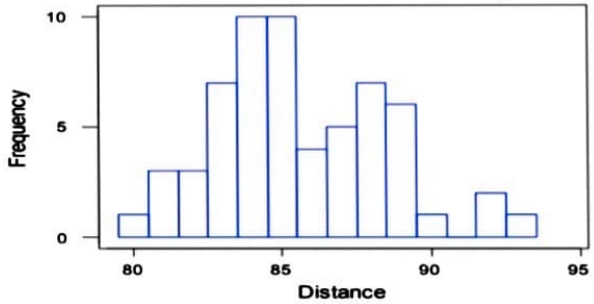
- Textbook is much more exhaustive than what we will cover in five weeks
- The best use of the book is as a reference, go to specific sections of chapter where you need more clarity
- First solve the exercises from the textbook before thinking of more practice problems

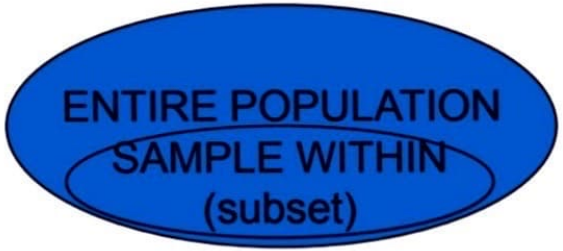
Software for the Course

python™

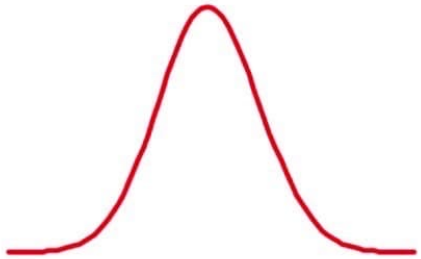
spyder

SAMPLE





POPULATION



Sample Statistics

A sample is a set of n observations actually obtained and a statistic is a numerical value that describes the sample.

$\bar{X}$ = Sample Mean
s^2 = Sample Variance
s = Sample Standard Deviation

Population Parameters

a hypothetical set of N observations from which the sample is obtained (typically N very large)

μ = Population mean
σ^2 = Population Variance
σ = Population Standard Deviation

STATISTICS

There are two types of statistics that are often referred to when making a statistical decision or working on a statistical problem.

Descriptive Statistics : Descriptive statistics utilize numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present the information in a convenient form that individuals can use to make decisions. The main goal of descriptive statistics is to describe a data set. Thus, the class of descriptive statistics includes both numerical measures (e.g. the mean or the median) and graphical displays of data
(e.g. pie charts or bar graphs).

Inferential Statistics: Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data. Some examples of inferential statistics might be a z statistics or a t-statistics

STATISTICS

Inferential Statistics

The main goal of inferential statistics is to make a conclusion about a population based off of a sample of data from that population. One of the most commonly used inferential techniques is *hypothesis testing*

Ex: New drug tests

Data come in many flavors ...

Type of data	Definition	Example
Nominal	Categories	Your previous degree
Ordinal	Can be ranked / ordered but not measured	Business school rankings
Interval scale	Intervals are meaningful but not ratios	Temperature in Fahrenheit or Celsius
Ratio scale	Ratios are meaningful	Sales of a new product

Source of data	Definition	Example
Observational	Analyst does not control data generating process	Stock returns on BSE
Experimental	Analyst has good control over data generation	Drug efficacy in clinical trials

Data Types – Preliminaries

Nominal

- Merely labels. No further information can be gleaned.
- Example: "Coke" and "Pepsi".

Ordinal

- Conveys only upto preference information. Direction alone.
- Example: "I prefer Coke to Pepsi".

Interval

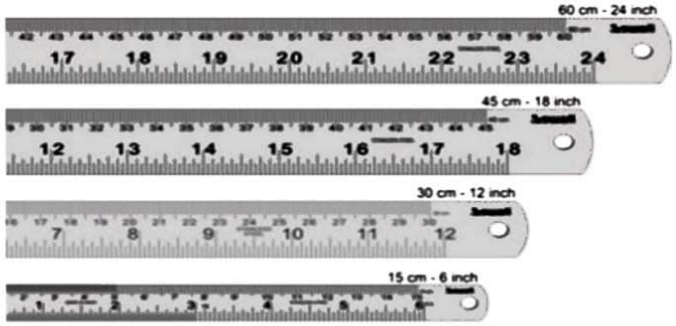
- Conveys relative magnitude information, in addition to preference.
- Example: "I rate Coke a 7 and Pepsi a 4 on a scale of 10".

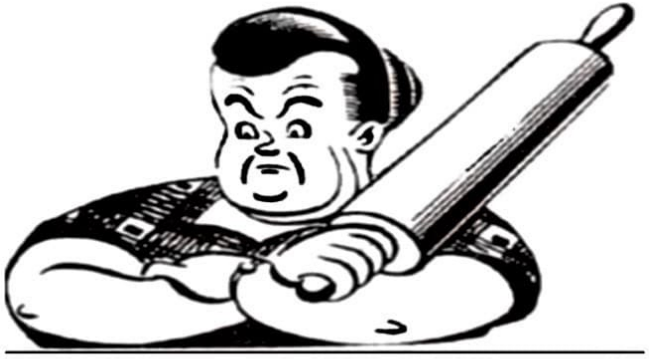
Ratio

- Conveys information on an absolute scale.
- Example: "I paid Rs 11 for Coke and Rs 12 for Pepsi".

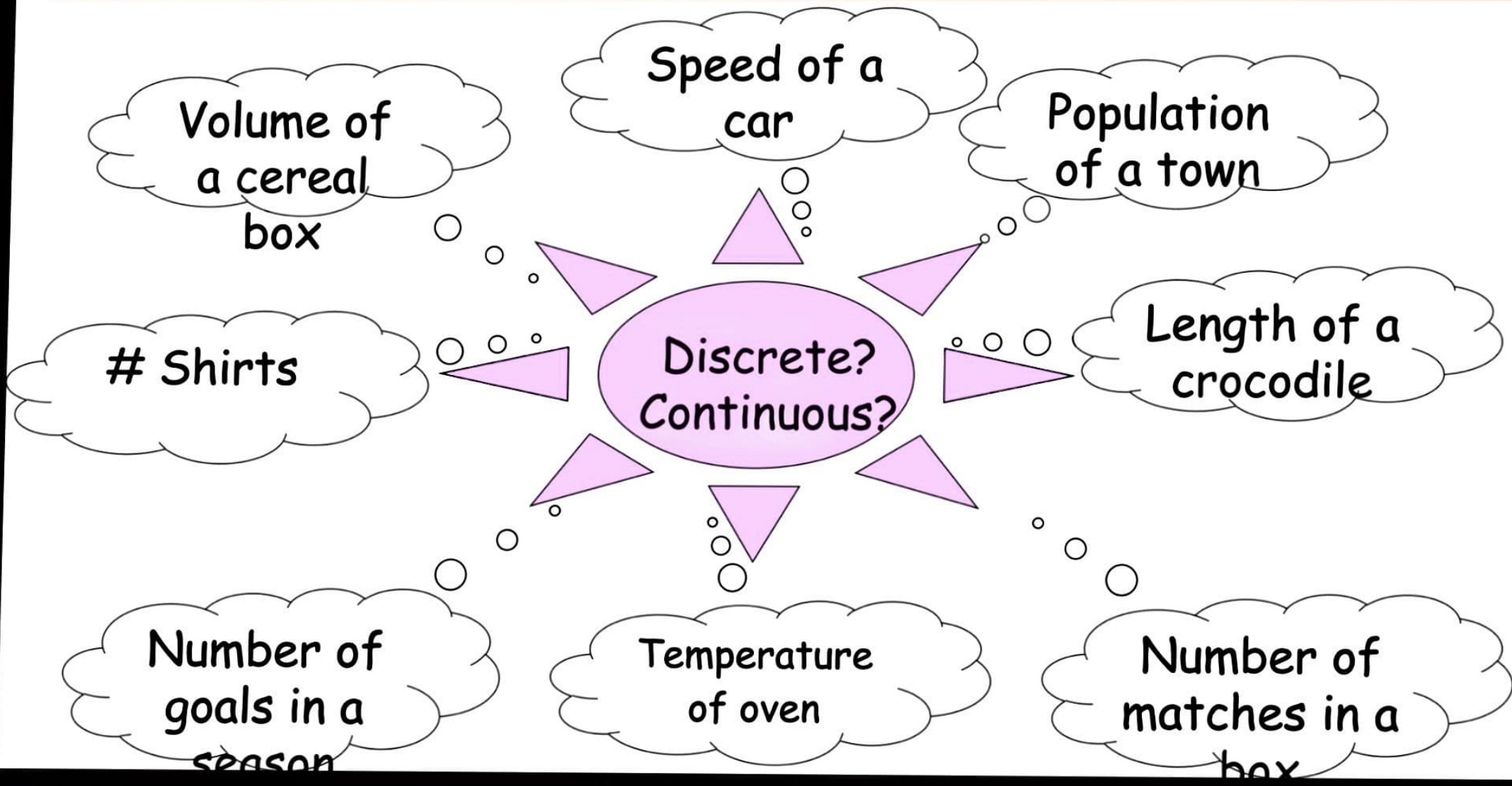
NOMINAL	ORDINAL	INTERVAL	RATIO
Mode	Mode	Mode	Mode
Frequencies	Median	Median	Median
Percentages	Frequencies	Mean	Mean
	Percentages	Frequencies	Frequencies
	Some Statistical Analysis	Percentages	Percentages
		Variance	Variance
		Standard Deviation	Standard Deviation
		Most Statistical Analysis	Ratio of numbers
			All Statistical Analysis

Data Types – Continuous & Discrete



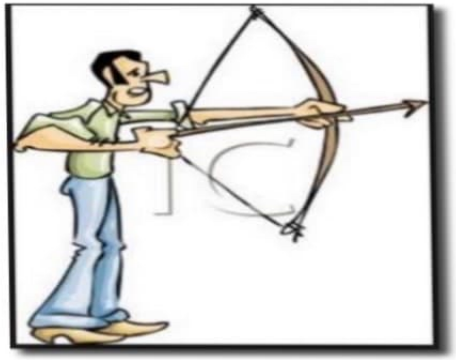


Group the following as either discrete or continuous data.



Measures of Central Tendency

Central Tendency	Population	Sample
Mean / Average	$\mu = \frac{\Sigma(x_i)}{N}$	$\bar{x} = \frac{\Sigma(x_i)}{n}$
Median	Middle value of the data	
Mode	Most occurring value in the data	



"Every American should have above average income, and my Administration is going to see they get it." – American President

Sample Mean for a Distribution

Σ y means, "Add up all the Y's"

For a discrete function

$$\bar{X} = \hat{\mu} = \sum_{i=1}^{N} x_i / N = \frac{X_1 + X_2 + \ldots . X_N}{N}$$

Examples:

Coating weights: 8.47, 8.67, 9.34, 7.99

$$\text{Coating AVERAGE} = \frac{8.47 + 8.67 + 9.34 + 7.99}{4} = 8.62$$

Batting Performance: 0, 0, 1, 0, 1 (0= no hit, 1=hit)

$$\text{BATTING AVERAGE} = \frac{0 + 0 + 1 + 0 + 1}{5} = 0.400$$

Mean = Average

Sample Median

Assume that x1, x2, …xn is a list of sample data sorted in ascending order.
Then…

$$\tilde{X} = \begin{cases} \text{middle value, if n is odd} \\ \text{the average of the two middle values, if n is even} \end{cases}$$

Find the sample mean and median for the two data sets below:

X: Data Set 1 : 10, 12, 11, 14, 11, 13, 12, 14, 16, 13

$\bar{X}$ = $\tilde{X}$ =

Y: Data Set 2: 10, 12, 11, 14, 11, 13, 12, 14, 44, 13

$\bar{Y}$ = $\tilde{Y}$ =

Mode

The modal value of a set of data is the most frequently occurring value.

Find the mode for: 2, 6, 3, 9, 5, 6, 2, 6

It can be seen that the most frequently occurring value is 6.
(There are 3 of these)

Bi model and Multi model

Mode for:

1) 1,2,3,3,3,4,4,4,5,6,7
2) 2,2,3,10,11,17,3,10

Measures of Variability

The mean, mode, and median do a nice job in telling where the center of the data set is, but often we are interested in more…

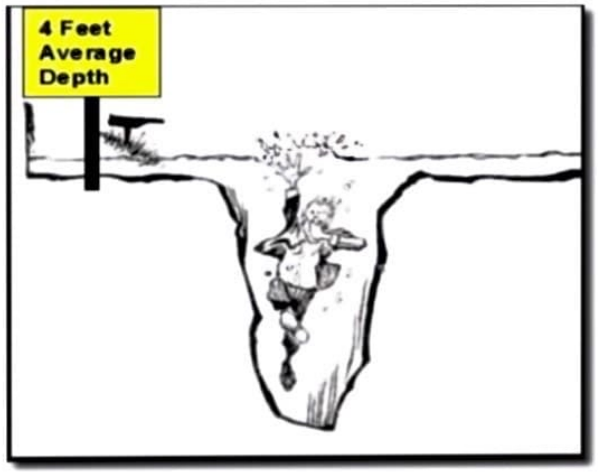
For example, a pharmaceutical engineer develops a new drug that regulates sugar in the blood.
Suppose she finds out that the average sugar content after taking the medication is the optimal level.

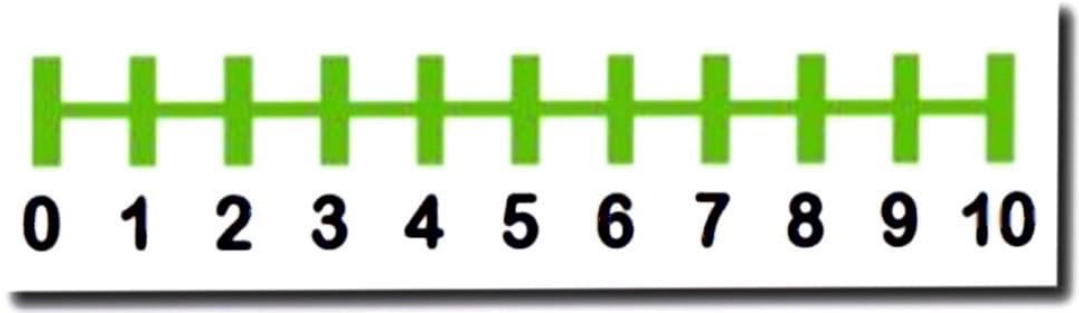
This does not mean that the drug is effective. There is a possibility that half of the patients have dangerously low sugar content while the other half has dangerously high content. Instead of the drug being an effective regulator, it is a deadly poison.

What the pharmacist needs is a measure of how far the data is spread apart. This is what the variance and standard deviation do

Measures of Dispersion

Dispersion	Population	Sample
Variance	$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$	$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$
Standard Deviation	$= \sqrt{\frac{\Sigma (X_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$
Range	Max – Min	





We define the *variance* to be

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

standard deviation to be

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

Range

The "Range" for a data set is the difference between the largest value and smallest value contained in the data set. First reorder the data set from smallest to largest then subtract the first element from the last element

Data Set = 2, 5, 9, 3, 5, 4, 7

Reordered = 2, 3, 4, 5, 5, 7, 9

Range = (9 - 2) = 7

In class exercise

Find std and Var

44, 50, 38, 96, 42, 47, 40, 39, 46, 50