1.1 **STATISTICS**

Statistics is the science of data. This involves collecting, summarizing, organizing, and analyzing data in order to draw meaningful conclusions about the universe from which the data is obtained.

Statistics

Descriptive Statistics Inferential statistics

Descriptive Statistics

Descriptive Statistics is the science of collecting, organizing, and summarizing data, in addition to utilizing numerical and graphical methods to present and reveal information about the data.

Inferential Statistics

A statistical inference is an estimate or prediction or some other generalization about a population based on information contained in a sample.

Measure of reliability

A measure of reliability is a statement (usually quantified) about the degree of uncertainty associated with a statistical inference.

Example

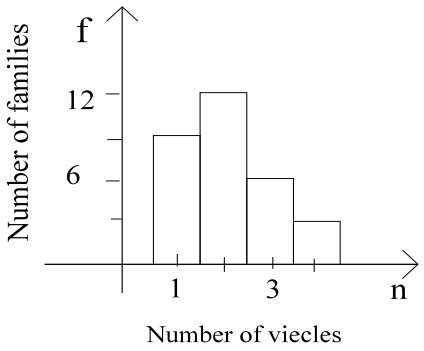
How many vehicles does a typical family own? Thirty families were polled and the following results are obtained

1 3 4 2 1 1 3 2 2 2
2 3 1 1 4 3 4 2 2 1
2 2 1 1 2 2 3 2 1 3

Before

Vehicle	Freq	Rel. Freq.
1	9	0.3
2	12	0.4
3	6	0.2
4	3	0.1

Frequency distribution



After

The Fundamental Elements of Statistics

Descriptive

population
A population is a set of units (equally people, objects, transactions or events) that we are interested in studying.

Variable
A variable is a characteristic or property of an individual population unit.

Sample
A sample is a subset of the population.

Parameter
A statement or measurable characteristic pertaining to the Population.

Statistic
A statement or measurable characteristic pertaining to the sample.

Inferential

Statistical inference
A statistical inference is an estimate or prediction or some other generalization about a population based on information contained in a sample.

Measure of reliability
A measure of reliability is a statement (usually quantified) about the degree of uncertainty associated with a statistical inference.

"The Uncertainty when cast in statistical terms, turns into variations."

The methods of statistics applied to data reveal variation inherent in the data which can be used to summerize and predict outcomes form representative subsets of the population from which the data is obtained.

Example:
To be able to predict the amount of rain fall during rain fall season in the South West USA, a sample of 500 days is randomly selected from the last 100 years of rain fall season.

Population
The days of the year during rain fall season for the last century.

Variable
The number of inches of rain fall each day in the Southwest USA during rain fall season.

Sample
Simple random sample of 500 measurements of rain fall in inches, on per day.

Descriptive Statistics

A Frequency Distribution

A bar chart

The average daily rain fall in inches during rain fall season

Inferential Statistics

The average rain fall in inches per day during a rain fall season in the South West USA next year.

The error in the prediction of the average. rain fall per day next year.

1.2 The Need to Model Uncertainty

Deterministic outcomes

The outcome is determine by methods that produce certain outcome, up to error in measurements.

Uncertain outcomes

The outcomes can not be determined with any degree of certainty**.**

Statistical outcomes

Outcomes that depend on the analysis of data.

Probabilistic Outcomes

Outcomes that relate to chance occurrence**.**

1.3 Random Variables And Distributions

Variables whose values can not be predicted with certainty are identified as random variables.

Examples:

- The weight of a new born baby.
- The closing price of an index in a financial market.
- The number of minutes a client waits in a teller line in a bank.
- The outcome of coin flip. (Head or Tail)

1.4 **DATA TYPES**

Qualitative

A qualitative data are measurements that can not be measured on a natural numerical scale; they can only be classified into one of a group of categories.

Examples

The inventory of a set of rain coats at a department store are classified by size as Small (S), Medium (M), Large(L) or Extra Large(XL)

The classification of group of 10 people as male or female.

M M F F M F F F M M

Quantitative or numerical:

Discrete

The data consists of counts

Continuous

The data consists of measurements that are recorded on a naturally occurring numerical scale.

Example Discrete

A sample of data that gives the number of school age children per family in santa Clara County.
A sample of twenty families yielded the following:

1 1 2 3 2 1 4 3 2 1 1 3 2 2 2 4 1 3 1 3

Example Continuous

The high daily temperature in Fahrenheit in Down Town Los Angeles during the month of July in 1998 .

101 98 99 94 90 85 84 90 91 92 88 89 92 ....

Data Types

Qualitative

A qualitative data are measurements that can not be measured on a natural numerical scale; they can only be classified into one of a group of categories

Quantitative or numerical:

Discrete

The data consists of counts

Continuous

The data consists of measurements that are recorded on a naturally occurring numerical scale.

Levels of measurement

Nominal scale level

Qualitative: Categories, colors, names, labels and favorite foods along with yes or no responses are examples

Nominal scale data are not ordered

Cell phone manufacturers: Sony, Motorola, Nokia, Samsung and Apple

Ordinal scale level

Same as Nominal but it can be ordered although there is no way to measure differences between the ranking or order.

Classifying Test performances as "excellent," "good," "satisfactory" and "unsatisfactory."

Interval scale level

Same as Ordinal level, but the differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F), there is no lowest temperature or a starting point.

Ratio scale level

Ratio scale data is like interval scale data but, in addition, it has a 0 point and ratios can be calculated.

For example, four final exam scores are 80, 68, 20 and 92 (out of a possible 100 points)..

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points.

Ratios can be calculated. The smallest score for ratio data is 0. So 80 is 4 times 20. The score of 80 is 4 times better than the score of 20.

1.5 # Producing Data

1. Sampling
   A process by which a subset of a population is selected.

2. Experiments
   A process by which an outcome is obtained.

3. Surveys
   A process by which data is collected by soliciting responses from objects under study.

4. Census
   A survey in which every member of the population contribute the responses.

5. Simulations
   A process in which assumptions about the type of outcome is made and these outcomes are produced by some form of a technology such as flipping a coin or using a random integer number generator that produces integers from o to n, for some number n with equally likelihood.

Examples

Samples

Population:

Deanza College Students from the time of its inception to previous year.

Sample:

Randomly selecting 500 Denza College students from the population of students.

Population:

All restaurants registered and open for business in the city of Cupertino the year 2014.

Sample:

Randomly select 30 restaurants from the population

Experiments

Flip a coin 3 times

population:

All possible outcomes of three flips of the coin

HHH HHT HTT TTT HTT HHT HTH HTH

Experiment

The result of repeating the experiment 4 times.

TTH HHT HTH TTH

Census

The USA Census Year 2010

Simulation

Use the number generator in your calculator to simulate flipping a coin 50 times.
Randint(0,1,10), where H=1, 0=T
0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1 or T, H, H, H, T, T, H, H, T, H, H

Random samples

- Simple random sampling
  In a sample of size n, every element in the population has equal chance of being in the sample and every sample of size n hs an equal chance of being the ample.

- Systematic sampling
  A sample in which repeatedly every mth data point, beginning the count from 1 directly after the last choice, is included in the sample while the population is arranged in some order.

- Stratified sampling
  A sample in which the population is divided int strata, (order), then chose a simple random sample of size n from each strata

- Proportionally stratified sampling
  Same as Stratified sampling with the exception that the number of data points from each strata follow a set rule given in proportions of the date in the sample, such as (½, 1/4, 1/4) with three different strata.

- Cluster sampling
  Divide the population in two clusters, then chose every member in a simple random sample of clusters.

- Convenience sampling
  A sample in which data is collected at the convenience of the data collector.

1.6 Statistical Experiments

Components of Comparative Experiments:

Experimental units

Explanatory Variable (Factor)

Levels or Categories

Response Variable

Responses to the different levels or categories

Randomization

A random assignment mechanism by which the experimental units are assigned to subgroups within the sample.

Replication

The idea of reducing chance variation in the response to treatment by the administration of the treatment to a large number of experimental units.

Treatment

That which is given to the experimental units and produces a response.

Placebo

That which is given to the experimental units, but it is not responsible for the response

Treatment Group

The group that receive the treatment

Control Group

The group that receive the placebo

Lurking Variables

Unaccounted for variables that may affect the response of the experimental units to the treatment.

Confounding

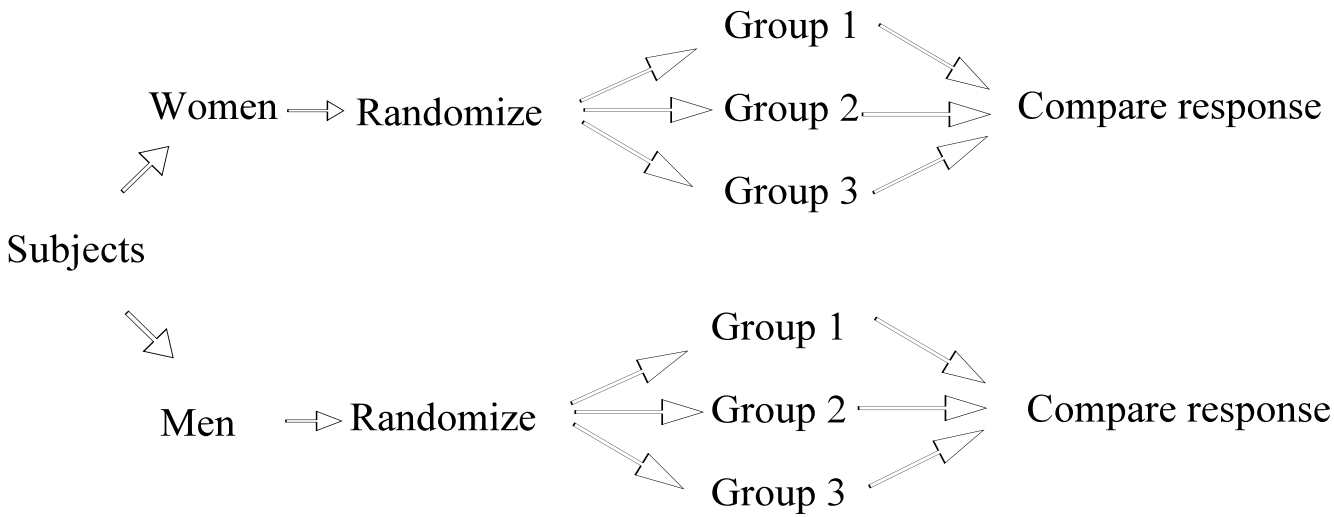
The affect of variables that are not considered in the experiment but may partially influence the responses.

Block

A block is a group of experimental units known to be similar in how they affect the response to the treatment**.**

Blocking

Dividing the experimental units into blocks that may affect the responses in similar ways and conducting the same experiment on each block.


Women
Randomize
Group 1
Group 2
Group 3
Compare response
Subjects
Men
Randomize
Group 1
Group 2
Group 3
Compare response