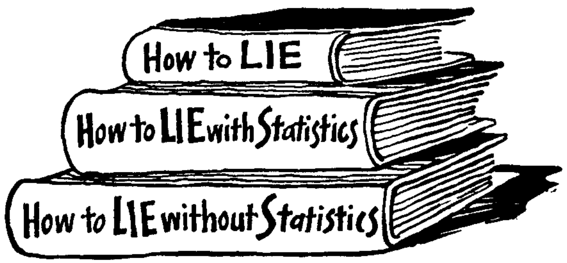
Introduction To Statistics

„There are three kinds of lies: lies, damned lies, and statistics.“ (B.Disraeli)



Why study statistics?

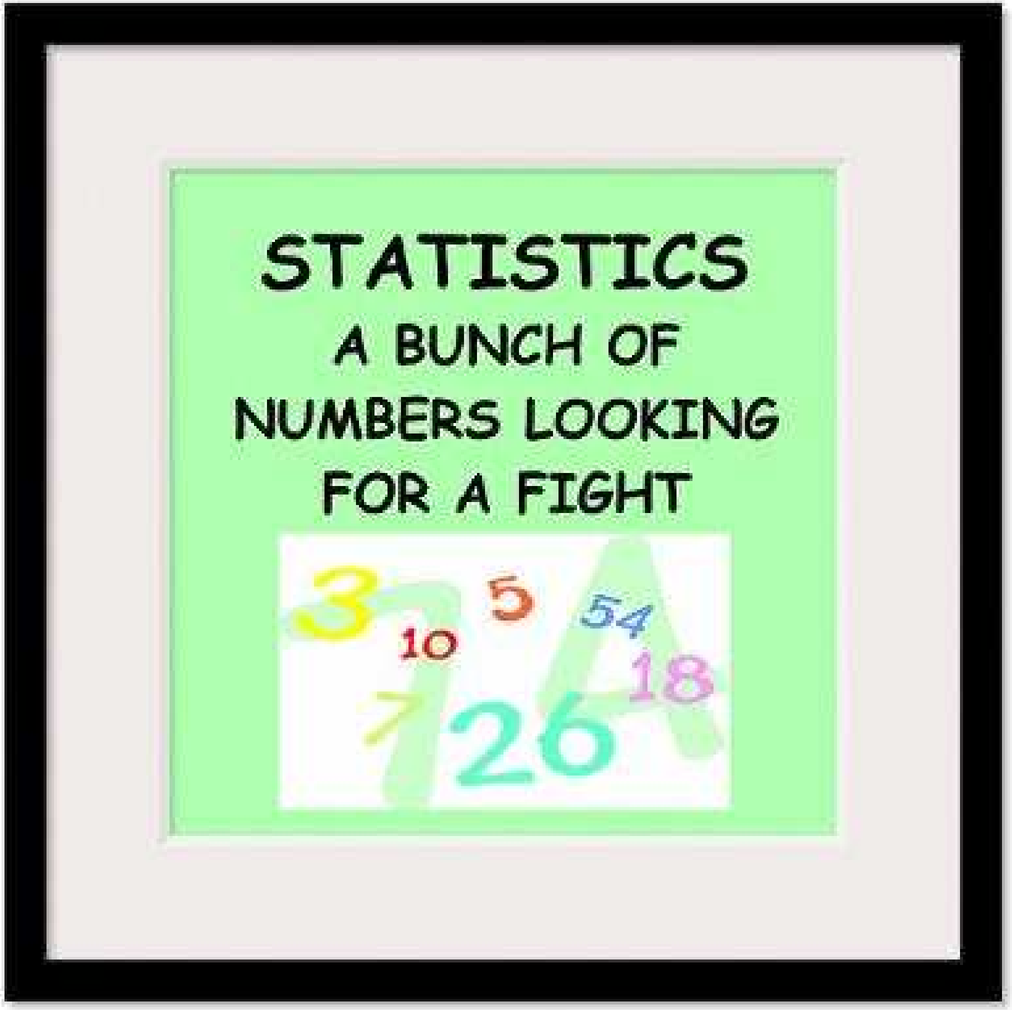
1. Data are everywhere
2. Statistical techniques are used to make many decisions that affect our lives
3. No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions efectively

Applications of statistical concepts in the business world

- Finance – correlation and regression, index numbers, time series analysis
- Marketing – hypothesis testing, chi-square tests, nonparametric statistics
- Personel – hypothesis testing, chi-square tests, nonparametric tests
- Operating management – hypothesis testing, estimation, analysis of variance (ANOVA), time series analysis

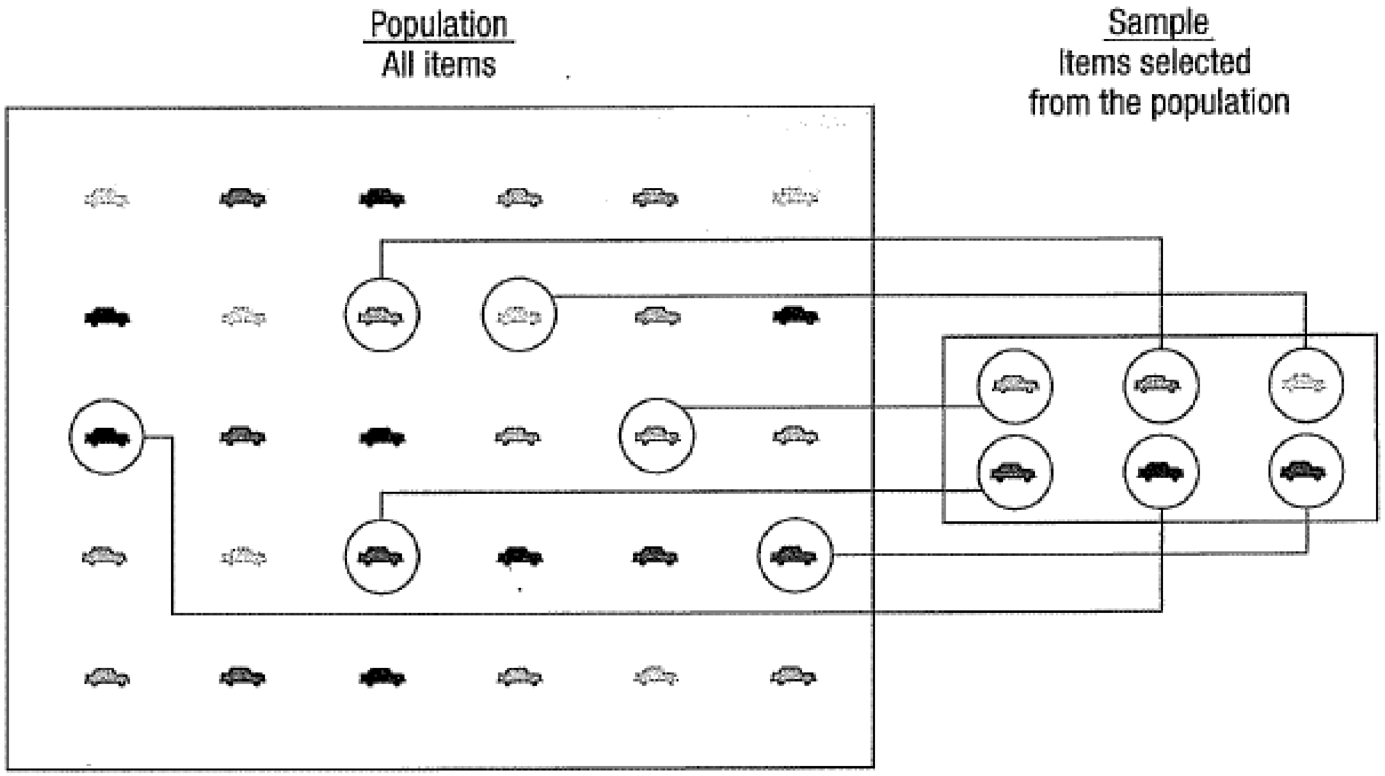
Statistics

- The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions
- Statistical analysis – used to manipulate, summarize, and investigate data, so that useful decision-making information results.


STATISTICS
A BUNCH OF
NUMBERS LOOKING
FOR A FIGHT
3
10
5
54
18
7
26

Types of statistics

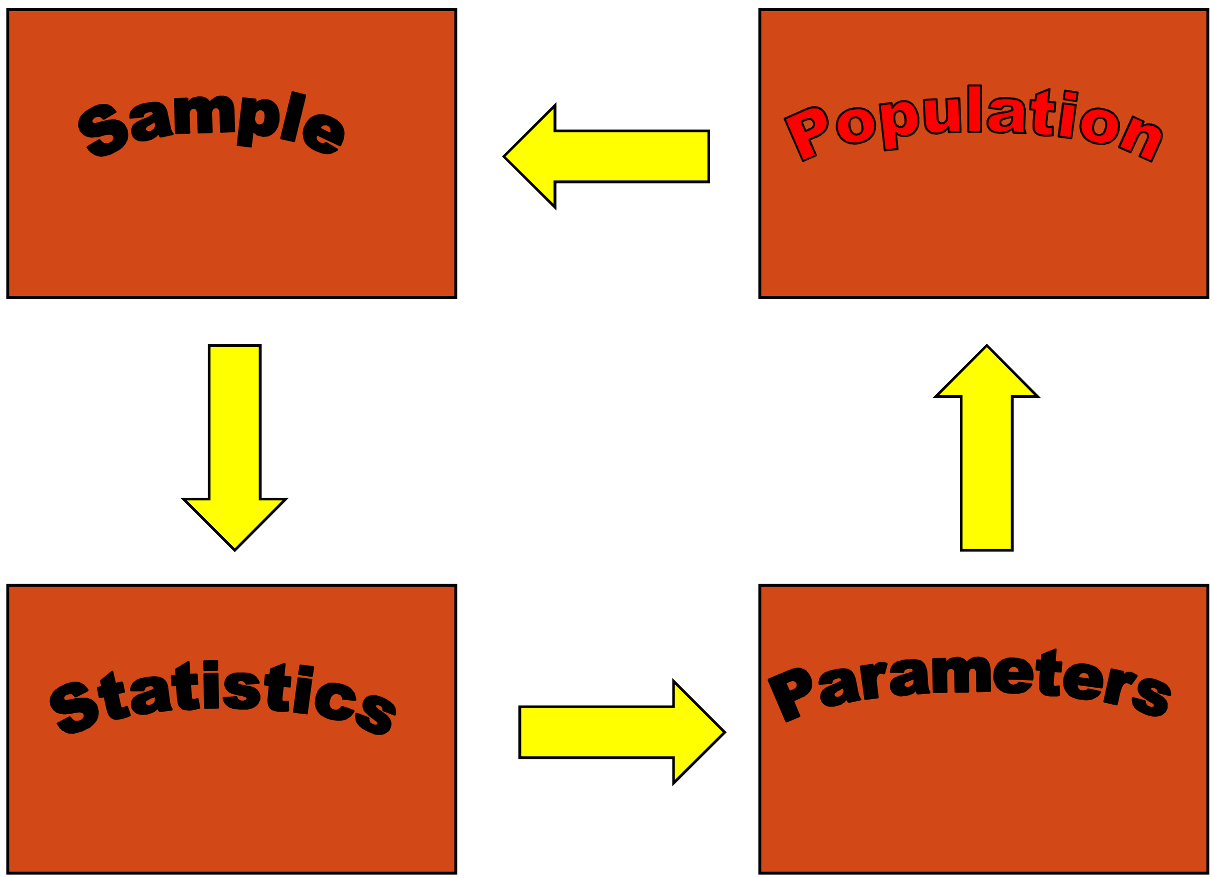
- **Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way
- **Inferential statistics** – The methods used to determine something about a population on the basis of a sample
  - Population – The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
  - Sample – A portion, or part, of the population of interest


Population
All items
Sample
Items selected
from the population

Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kg

Inference is the process of drawing conclusions or making decisions about a population based on sample results


Sample
Population
Statistics
Parameters

Review of Basic Statistical Concepts

- Introduction to Inference
  - The purpose of inference is to draw conclusions from data.
  - Conclusions take into account the natural variability in the data, therefore formal inference relies on probability to describe chance variation.
  - We will go over the two most prominent types of formal statistical inference
    - *Confidence Intervals* for estimating the value of a population parameter.
    - *Tests of significance* which asses the evidence for a claim.
  - Both types of inference are based on the *sampling distribution* of *statistics*.

Review of Basic Statistical Concepts

- Parameters and Statistics
  - A *parameter* is a number that describes the *population*.
    - A parameter is a fixed number, but in practice we do not know its value.
  - A *statistic* is a number that describes a *sample*.
    - The value of a statistic is known when we have taken a sample, but it can change from sample to sample.
    - We often use statistic to estimate an unknown parameter.

Sampling

a sample should have the same characteristics as the population it is representing.

Sampling can be:

- **with replacement**: a member of the population may be chosen more than once (picking the candy from the bowl)
- **without replacement**: a member of the population may be chosen only once (lottery ticket)

Sampling methods

Sampling methods can be:

- **random** (each member of the population has an equal chance of being selected**)**
- **nonrandom**

The actual process of sampling causes **sampling errors**. For example, the sample may not be large enough or representative of the population. Factors not related to the sampling process cause nonsampling errors. A defective counting device can cause a nonsampling error.

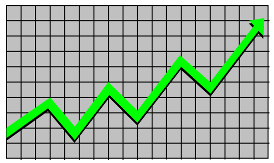
Random sampling methods

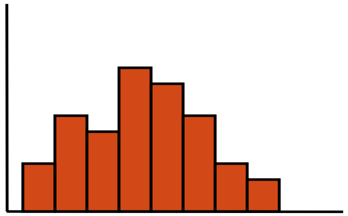
- **simple random sample** (each sample of the same size has an equal chance of being selected)
- **stratified sample** (divide the population into groups called strata and then take a sample from each stratum)
- **cluster sample** (divide the population into strata and then randomly select some of the strata. All the members from these strata are in the cluster sample.)
- **systematic sample** (randomly select a starting point and take every n-th piece of data from a listing of the population)
- Convenience sample – medical field – volunteers
- Judgment sample – Expert selects sample

Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs





- Summarize data
  - e.g., Sample mean = $\dfrac{\sum X_i}{n}$

Statistical data

- The collection of data that are relevant to the problem being studied is commonly the most difficult, expensive, and time-consuming part of the entire research project.
- Statistical data are usually obtained by counting or measuring items.
  - **Primary data** are collected specifically for the analysis desired
  - **Secondary data** have already been compiled and are available for statistical analysis
- A **variable** is an item of interest that can take on many different numerical values.
- A **constant** has a fixed numerical value.

Data

Statistical data are usually obtained by counting or measuring items. Most data can be put into the following categories:

- **Qualitative -** data are measurements that each fail into one of several categories. (hair color, ethnic groups and other attributes of the population)
- **quantitative** - data are observations that are measured on a numerical scale (distance travelled to college, number of children in a family, etc.)

Qualitative data

Qualitative data are generally described by words or letters. They are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense to find an average hair color or blood type.

Qualitative data can be separated into two subgroups:

- **dichotomic** (if it takes the form of a word with two options (gender - male or female)
- **polynomic** (if it takes the form of a word with more than two options (education - primary school, secondary school and university).

Quantitative data

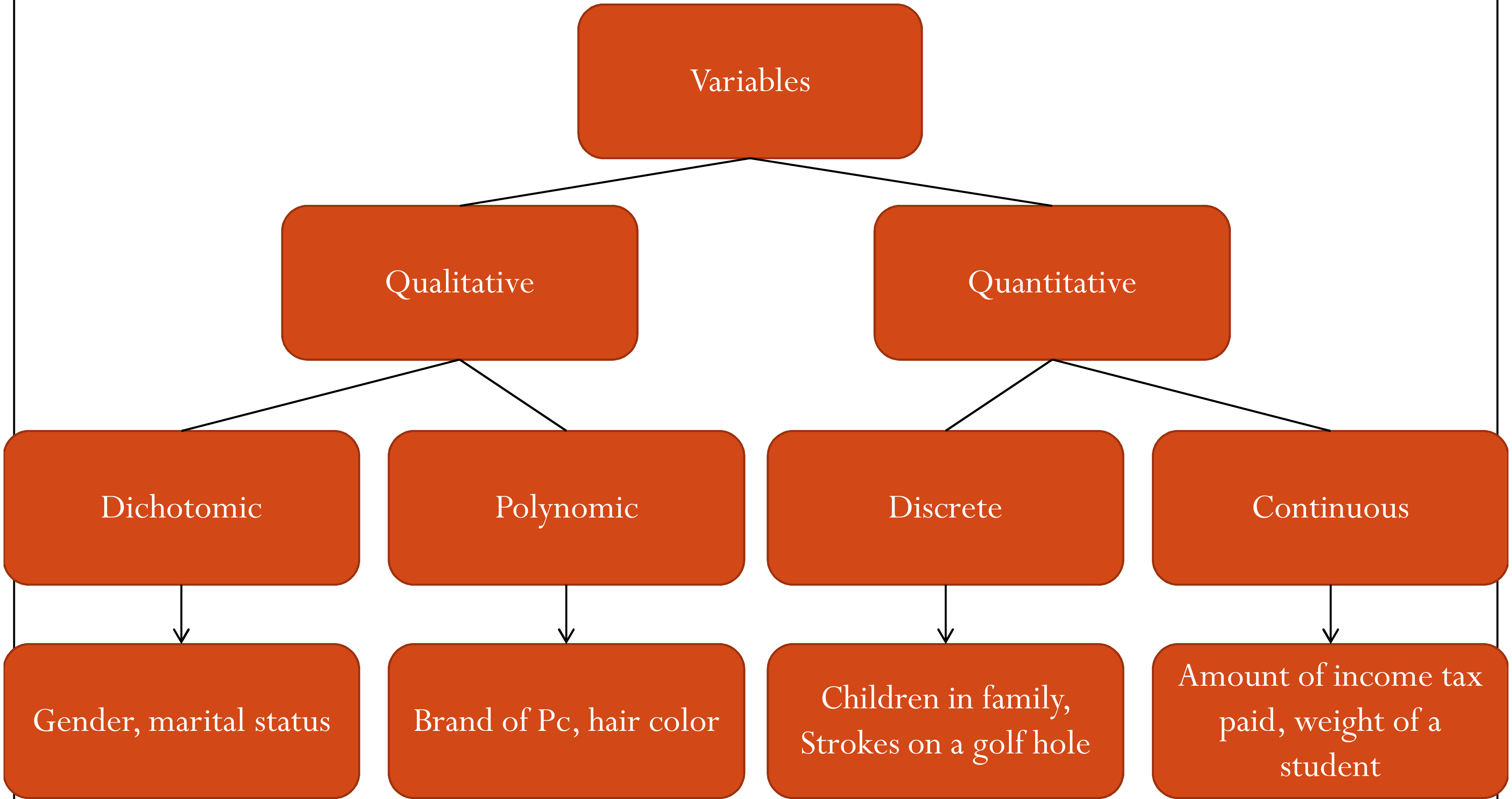
Quantitative data are always numbers and are the **result of counting or measuring** attributes of a population.

Quantitative data can be separated into two

subgroups:

- **discrete** (if it is the result of *counting* (the number of students of a given ethnic group in a class, the number of books on a shelf, …)
- **continuous** (if it is the result of *measuring* (distance travelled, weight of luggage, …)

Types of variables



Numerical scale of measurement:

- **Nominal** – consist of categories in each of which the number of respective observations is recorded. The categories are in no logical order and have no particular relationship. The categories are said to be ***mutually exclusive*** since an individual, object, or measurement can be included in only one of them.
- **Ordinal** – contain more information. Consists of distinct categories in which order is implied. Values in one category are larger or smaller than values in other categories (e.g. rating-excellent, good, fair, poor)
- **Interval** – is a set of numerical measurements in which the distance between numbers is of a known, constant size.
- **Ratio** – consists of numerical measurements where the distance between numbers is of a known, constant size, in addition, there is a nonarbitrary zero point.

Numerical presentation of qualitative data

- **pivot table** (qualitative dichotomic statistical attributes)
- **contingency table** (qualitative statistical attributes from which at least one of them is polynomic)

You should know how to convert absolute

values to relative ones (%).

Pivot table

A3 Sum of Amount

Category	(All)							
Sum of Amount	Product							
Country	Apple	Banana	Beans	Broccoli	Carrots	Mango	Orange	Grand Total
Australia	20634	52721	14433	17953	8106	9186	8680	131713
Canada	24867	33775		12407		3767	19929	94745
France	80193	36094	680	5341	9104	7388	2256	141056
Germany	9082	39686	29905	37197	21636	8775	8887	155168
New Zealand	10332	40050		4390			12010	66782
United Kingdom	17534	42908	5100	38436	41815	5600	21744	173137
United States	28615	95061	7163	26715	56284	22363	30932	267133
Grand Total	191257	340295	57281	142439	136945	57079	104438	1029734

Contingency table

	Dog	Cat	Total
Male	42	10	52
Female	9	39	48
Total	51	49	100

Frequency distributions – numerical presentation of quantitative data

- Frequency distribution – shows the frequency, or number of occurences, in each of several categories. Frequency distributions are used to summarize large volumes of data values.
- When the raw data are measured on a qunatitative scale, either interval or ration, categories or classes must be designed for the data values before a frequency distribution can be formulated.

Steps for constructing a frequency distribution

1. Determine the number of classes $m = \sqrt{n}$
2. Determine the size of each class $h = \frac{(\max - \min)}{m}$
3. Determine the starting point for the first class
4. Tally the number of values that occur in each class
5. Prepare a table of the distribution using actual counts and/ or percentages (relative frequencies)

Frequency table

- **absolute frequency "n_i"** (Data Tab→Data Analysis→Histogram)
- **relative frequency "f_i"**

Cumulative frequency distribution shows the total number of occurrences that lie above or below certain key values.

- **cumulative frequency "N_i"**
- **cumulative relative frequency "F_i"**

Frequency table

Degree	Frequency	Relative Frequency	Percentage
High School	2	0.050	5.0
Bachelor's	7	0.175	17.5
MBA	20	0.500	50.0
Master's	3	0.075	7.5
Law	4	0.100	10.0
PhD	4	0.100	10.0
	40		

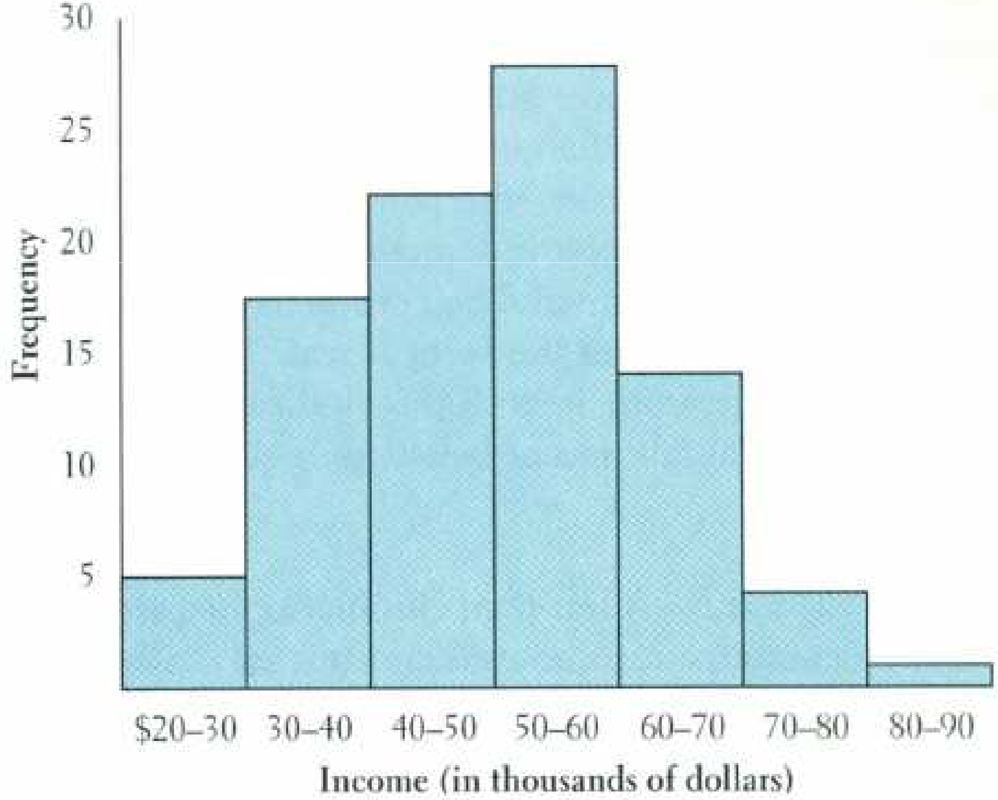
Charts and graphs

- Frequency distributions are good ways to present the essential aspects of data collections in concise and understable terms
- Pictures are always more effective in displaying large data collections

Histogram

- Frequently used to graphically present interval and ratio data
- Is often used for interval and ratio data
- The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes

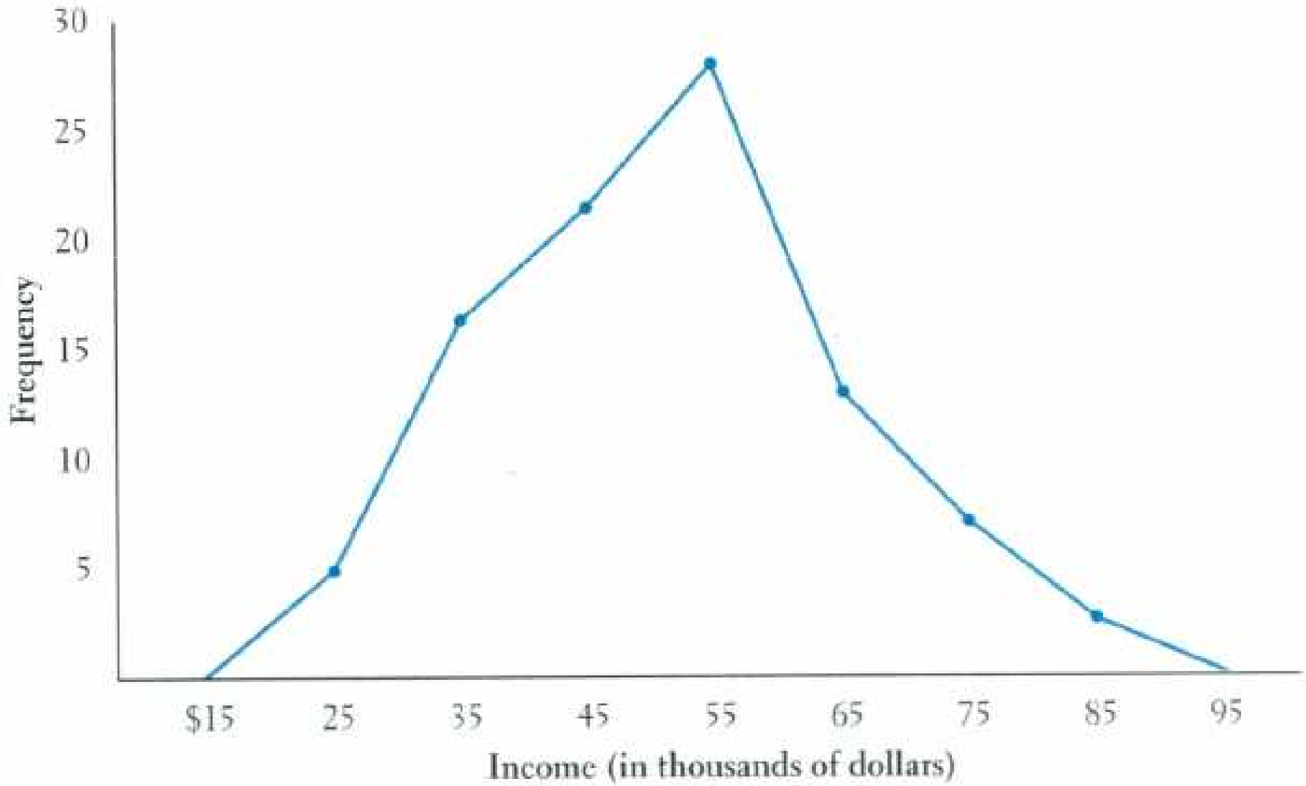
FIGURE 3.7 Histogram—Executive Incomes for the Sunrunner Corporation



Frequency polygon

- Another common method for graphically presenting interval and ratio data
- To construct a frequency polygon mark the frequencies on the vertical axis and the values of the variable being measured on the horizontal axis, as with the histogram.
- If the purpose of presenting is comparation with other distributions, the frequency polygon provides a good summary of the data

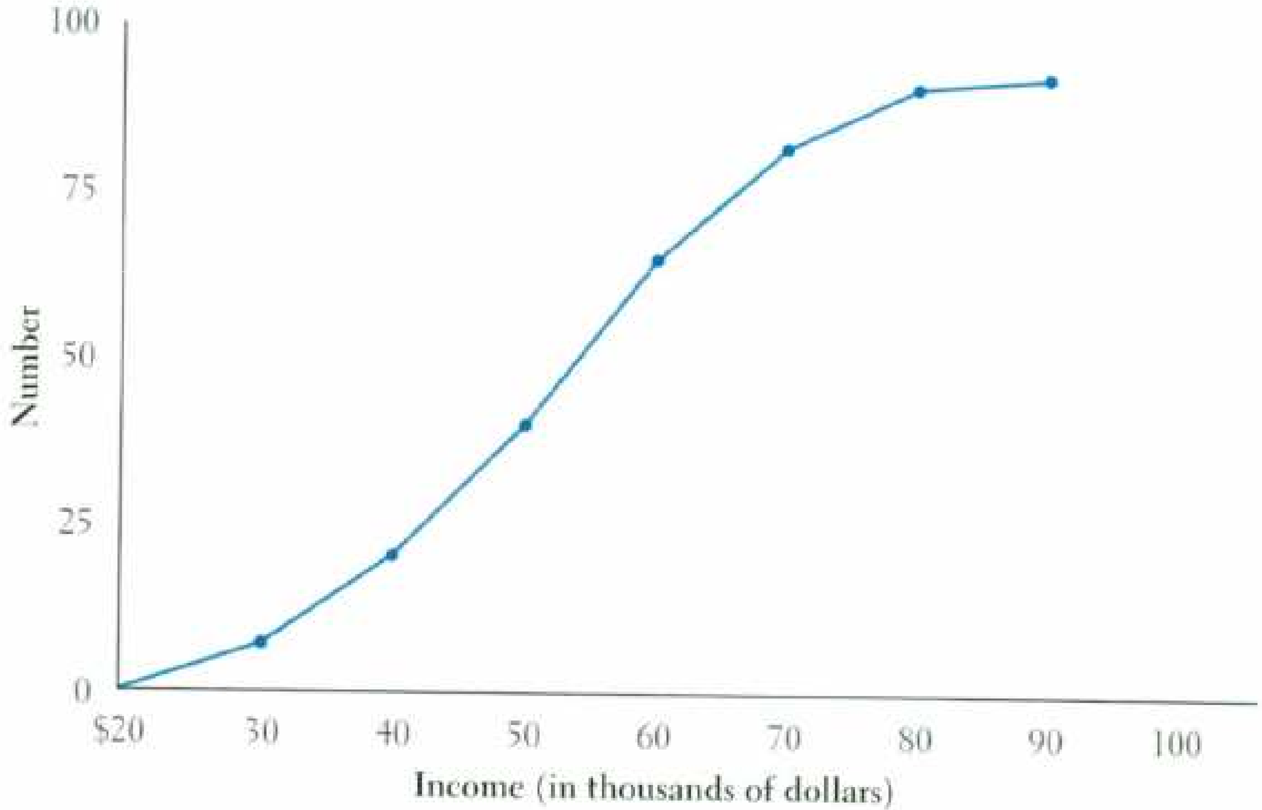
FIGURE 3.8 Frequency Polygon—Executive Incomes



Ogive

- A graph of a cumulative frequency distribution
- Ogive is used when one wants to determine how many observations lie above or below a certain value in a distribution.
- First cumulative frequency distribution is constructed
- Cumulative frequencies are plotted at the upper class limit of each category
- Ogive can also be constructed for a relative frequency distribution.

FIGURE 3.9 Ogive—Executive Incomes (frequencies)

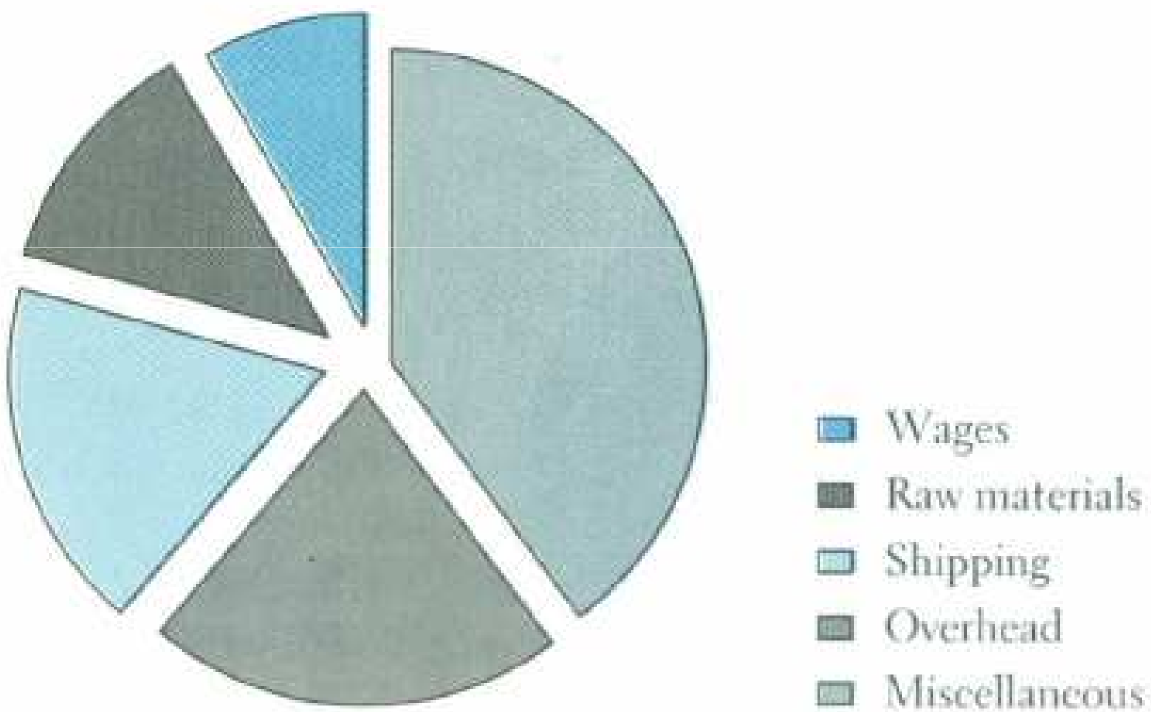


Pie Chart

- The pie chart is an effective way of displaying the percentage breakdown of data by category.
- Useful if the relative sizes of the data components are to be emphasized
- Pie charts also provide an effective way of presenting ratio- or interval-scaled data after they have been organized into categories

Pie Chart

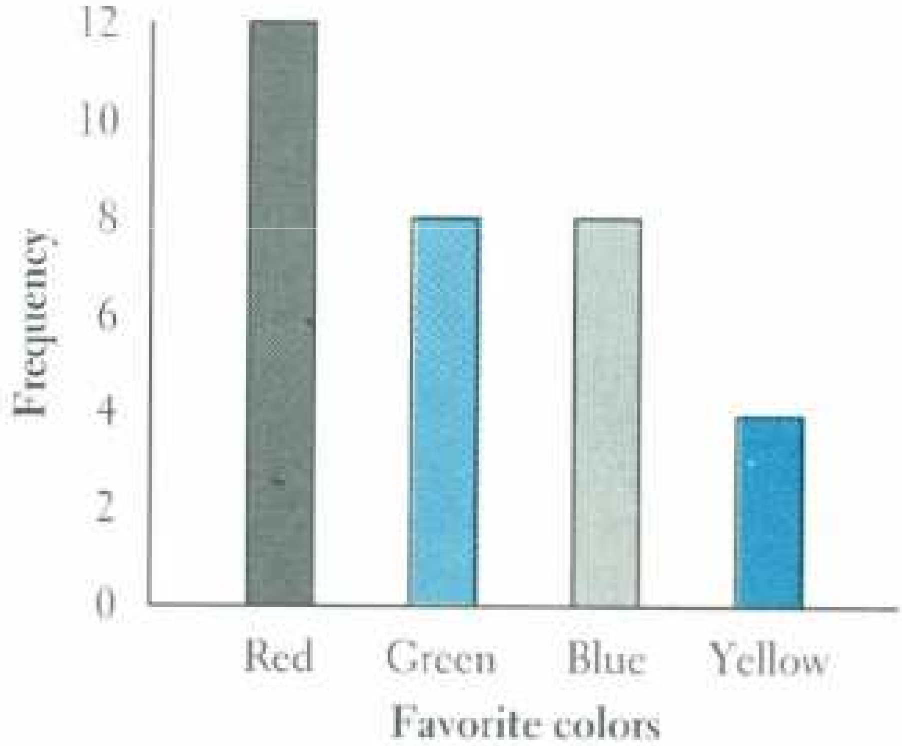
FIGURE 3.3 Pie Chart—Expenditures of Funds for Itrex Company



Bar chart

- Another common method for graphically presenting nominal and ordinal scaled data
- One bar is used to represent the frequency for each category
- The bars are usually positioned vertically with their bases located on the horizontal axis of the graph
- The bars are separated, and this is why such a graph is frequently used for nominal and ordinal data – the separation emphasize the plotting of frequencies for distinct categories

FIGURE 3.4 Bar Chart—Favorite Colors of 32 People



Time Series Graph

- The time series graph is a graph of data that have been measured over time.
- The horizontal axis of this graph represents time periods and the vertical axis shows the numerical values corresponding to these time periods

FIGURE 3.13 Time Series Graph—Corporate Revenue, Flightcraft Corp.

Two Broad Areas of Statistics

- Descriptive Statistics
  - Numerical descriptors
  - Graphical devices
  - Tabular displays
- Inferential Statistics
  - Estimation & Hypothesis testing
  - Confidence intervals
  - Model building/selection

Descriptive Statistics

- When computed for a *population* of values, numerical descriptors are called
  - Parameters
- When computed for a *sample* of values, numerical descriptors are called
  - Statistics

Descriptive Statistics

- Two important aspects of any population
- Magnitude of the responses
- Spread among population members

Descriptive Statistics

- Measures of Central Tendency (magnitude)
- Mean - most widely used
- - uses all the data
- - *best* statistical properties
- - susceptible to outliers
- Mode
- Median - does not use all the data
- - resistant to outliers

Descriptive Statistics

- Measures of Spread (variability)

- range – simple to compute
- – does not use all the data

- variance – uses all the data
- – *best* statistical properties
- – measures average distance of values from a reference point
- Standard deviation
- Quartiles

Frequency Distribution

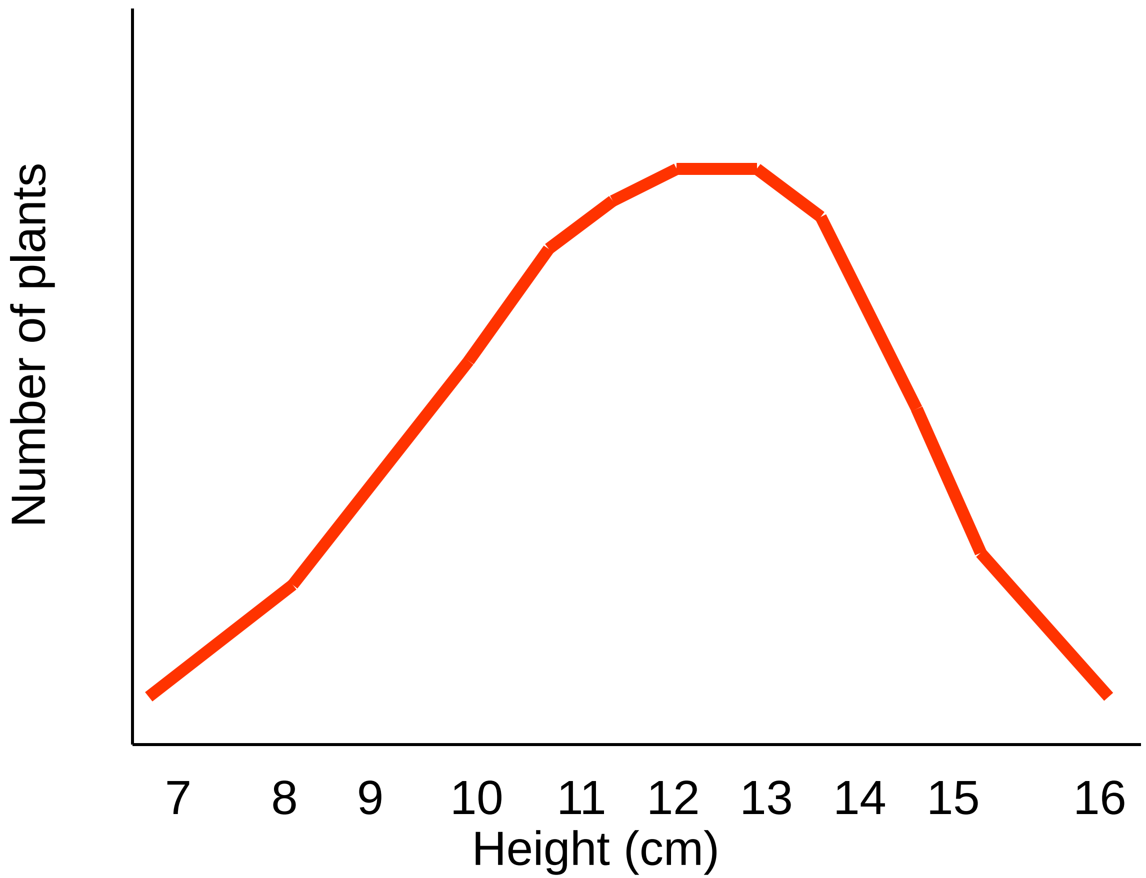
Data is collected and often organized into formats that are interpreted easily.

Example: Plant height due to the application of fertilizers. Height is given in centimeters (cm.)

fertilizer

plants	10	14	11	12	15
	15	12	13	14	13
	12	8	12	9	10
	13	11	12	8	10
	9	16	7	11	9

Frequency Distribution (cont.)



Measures of Central Tendancy

- Median- The middle number in a set of data.
- Mode- The number within the set of data that appears the most frequently.
- Mean- The average

  a. Denoted by $\bar{x}$

  b. Calculated by the following formula

$$\bar{x} = \frac{\Sigma x}{n}$$

Measures of Dispersion

- Variance- Determined by averaging the squared difference of all the values from the mean.
  - symbolized by σ^2

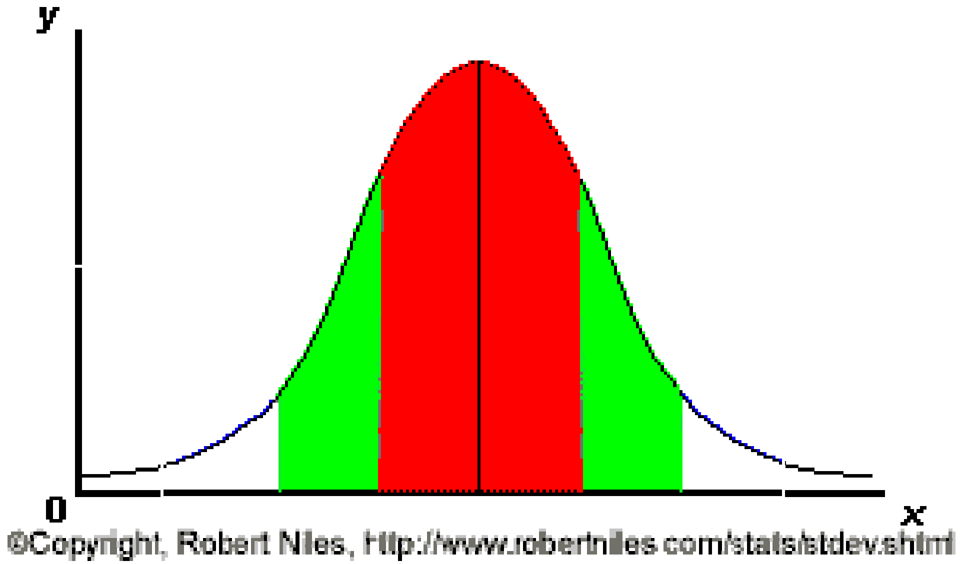
$$\sigma^2 = \frac{\Sigma\,(\bar{x} - x)^2}{n-1}$$

Measure of Dispersion (cont.)

- Standard Deviation- Is a measure of dispersion that defines how an individual entry differs from the mean.
- - calculated by finding the square root of the variance.
- Defines the shape of the normal distribution curve

$$\sigma = \sqrt{\sigma^2}$$

- Range
- Quartiles



- The red area represents the first standard deviant.
- **68%** of the data falls within this area.
- Calculated by **x ± δ**
- The green area represents the second standard deviant.
- **95%** of the data falls within the green PLUS the red area.
- Calculated by **x ± 2δ**
- The blue area represents the third standard deviant.
- **99%** of the data falls within blue PLUS the green PLUS the red area.
- Calculated by **x ± 3δ**

Sample Mean and Standard Deviation

For a series of N observations, the most probable estimate of the mean μ is the average $\bar{x}$ of the observations. We refer to this as the sample mean $\bar{x}$ to distinguish it from the parent mean μ.

$$\mu \cong \bar{x} \equiv \frac{1}{N}\sum x_i$$ **Sample Mean**

Our best estimate of the standard deviation σ would be from:

$$\sigma^2 \cong \frac{1}{N}\sum\left(x_i - \mu\right)^2 = \frac{1}{N}\sum x_i^2 - \mu^2$$

But we cannot know the true parent mean μ so the best estimate of the sample variance and standard deviation would be:

$$\sigma^2 \cong s^2 \equiv \frac{1}{N-1}\sum\left(x_i - \bar{x}\right)^2$$ **Sample Variance**

Properties of Statistics

- **Unbiasedness** - On target
- **Minimum variance** - Most reliable- efficient

- If an estimator possesses both properties then it is a **MINVUE** = MINimum Variance Unbiased Estimator

- ***Sample Mean*** and ***Variance*** are **UMVUE** = Uniformly MINimum Variance Unbiased Estimator

Inferential Statistics

- – Hypothesis Testing
- – Interval Estimation

Hypothesis Testing

- Specifying hypotheses:
- H_0: "null" or no effect hypothesis
- H_1: research or alternative hypothesis
- *Note:* Only H_0 (null) is tested.

Errors in Hypothesis Testing

Reality → ↓ Decision	H_0 True	H_0 False
Fail to Reject H_0	√	β(Type II err) Consumer's risk
Reject H_0	α(Type I err) Producer's risk	√

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR}-}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Type I and Type II errors

Table of error types Chart David .2004		Null hypothesis (H_0) is	
		Valid/True	**Invalid/False**
Judgement of Null Hypothesis (H_0)	**Reject**	Type I error (False Positive)	Correct inference (True Positive)
	Fail to reject	Correct inference (True Negative)	Type II error (False Negative)
<<Memory formulas>> **Type-1 = True H_0 but reject it (False Positive)** **Type-2 = False H_0 but accept it (False Negative)**			

Type I and Type II errors

- **Example 1**
- *Hypothesis:* "Adding water to toothpaste protects against cavities."
- *Null hypothesis:* "Adding water to toothpaste has no effect on cavities."
- This null hypothesis is tested against experimental data with a view to nullifying it with evidence to the contrary.
- A type I occurs when detecting an effect (adding water to toothpaste protects against cavities) that is not present. The null hypothesis is true (i.e., it is true that adding water to toothpaste has no effect on cavities), but this null hypothesis is rejected based on bad experimental data.
- **Example 2**
- *Hypothesis:* "Adding fluoride to toothpaste protects against cavities."
- *Null hypothesis:* "Adding fluoride to toothpaste has no effect on cavities."
- This null hypothesis is tested against experimental data with a view to nullifying it with evidence to the contrary.
- A type II error occurs when failing to detect an effect (adding fluoride to toothpaste protects against cavities) that is present. The null hypothesis is false (i.e., adding fluoride is actually effective against cavities), but the experimental data is such that the null hypothesis cannot be rejected.

Predictor Error Measures

- Measure predictor accuracy: measure how far off the predicted value is from the actual known value
- **Loss function**: measures the error bet. y_i and the predicted value y_i'
  - Absolute error: $| y_i - y_i' |$
  - Squared error: $(y_i - y_i')^2$
- Test error (generalization error): the average loss over the test set
  - Mean absolute error: $\dfrac{\sum_{i=1}^{d} | y_i - y_i' |}{d}$ Mean squared error: $\dfrac{\sum_{i=1}^{d} (y_i - y_i')^2}{d}$
  - Relative absolute error: $\dfrac{\sum_{i=1}^{d} | y_i - y_i' |}{\sum_{i=1}^{d} | y_i - \bar{y} |}$ Relative squared error: $\dfrac{\sum_{i=1}^{d} (y_i - y_i')^2}{\sum_{i=1}^{d} (y_i - \bar{y})^2}$

The mean squared-error exaggerates the presence of outliers

Popularly use (square) root mean-square error, similarly, root relative squared error

Hypothesis Testing

- In parametric tests, actual parameter values are specified for H_0 and H_1.
- H_0: μ <= 120
- H_1: μ > 120
- Two tailed test – RR in 2 sides
- One tailed test – RR in one side based on

μ > 120 or μ < 120

Hypothesis Testing

- Another example of explicitly
- specifying H_0 and H_1.
- H_0: $\beta = 0$
- H_1: $\beta \neq 0$

Hypothesis Testing

- General framework:

- Specify null & alternative hypotheses
- Specify test statistic
- Find the Critical value and State rejection rule (RR)
- Compute test statistic and compare to RR
- State conclusion

Statistical Significance

- Statistical significance is calculated by determining:
  - if the probability differences between sets of data occurred by chance
  - or were the result of the experimental treatment.
- Two hypotheses need to be formed:
  - Research hypothesis- the one being tested by the researcher.
  - Null hypothesis- the one that assumes that any differences within the set of data is due to chance and is not significant.

Statistical Significance (cont.)

- Example of Null Hypothesis:

  The mean weight of college football players is not significantly different from professional football players.

$$\mu_{cf} = \mu_{pf}$$

μ, 'mu' symbol for

Null Hypothesis

Common Statistical Tests

- Large Sample test (n >30)
  - Test for single proportion (out of 250 coin toss 130 are head)
  - Test for comparison of 2 sample proportions (2 city wheat consumption or 2 brand products are compared)
  - Test for single mean (average mark of a subject in university)
  - Test for comparison of 2 sample mean (analyze the effect of chemical on 2 sets of people like controlled and exposed)
- Small sample test
  - Test for single mean
  - Test for comparison of 2 sample mean

Common Statistical Tests

Test Name	Purpose
One-sample (z) t-test	Test value of a mean
Two-sample (z) t-test	Compare two means
Paired t-test	Compare difference in means (compare re-lated means)
ANOVA	Test for differences in 2 or more means
F test or variance test	
Chisquare test	

Common Statistical Tests (cont.)

Test	Purpose
Test on binomial proportion(s)	Test whether binomial proportions $=\pi_0$, or each other.
Test on correlation coefficient(s)	Test whether correlation coefficient $=\rho_0$, or each other.
Regression	Test whether slope $=\beta_0$
RxC contingency table analysis	Test whether two categorical variables are related

Advanced Topics

Test	Purpose
Multivariate Tests e.g., MANOVA	Test value of several parameters simultaneously
Repeated Measures / Crossovers	Test means when subjects repeatedly measured
Survival Analysis	Estimate and compare survival probabilities for one or more groups
Nonparametric Tests	Many analogous to standard parametric tests

T Test

- Statistical test that helps to show if there is a real difference between different treatments being tested in a controlled scientific trial.
- The **Student *t* test** is used to determine if the two sets of data from a sample are really different?
  - The uncorrelated *t* test is used when no relationship exist between measurements in the two groups.

T Test

- Two basic formulas for calculating an uncorrelated *t* test.

Equal sample size

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s^2_1 + s^2_2}{n}}}$$

Unequal sample size

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{(n_1 - 1)s^2_1 + (n_2 - 1)s^2_2}{n_1 + n_2 - 2} \cdot \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

Degree of Freedom

- Represents the number of independent observations in a sample.
- Is a measure that states the number of variables that can change within a statistical test.
- Calculated by n-1 (sample size – 1)

Level of Signifance

- Is determined by the researcher.
- Symbolized by α
- Is affected by the sample size and the nature of the experiment.
- Common levels of significance are

  .05, .01, .001
- Indicates probability that the researcher made an error in rejecting the null hypothesis.

Finding critical value of t

- A probability table is used
- First determine degrees of freedom
- Decide the level of significance

Sampling Distribution for *t* test

Sampling distribution for *t* test				
Degrees of freedom	Probability (Level of significance)			
	0.1	0.05	0.01	0.001
1	6.314	12.706	63.657	636.619
2	2.920	4.303	9.925	31.598
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.864

Finding critical value of t

- Example: degrees of freedom= 4

  α= .05

Sampling Distribution for *t* test

Sampling distribution for *t* test				
Degrees of freedom	Probability (Level of significance)			
	0.1	0.05	0.01	0.001
1	6.314	12.706	63.657	636.619
2	2.920	4.303	9.925	31.598
3	2.353	3.182	5.841	12.924
4	2.132	$\boxed{2.776}$	4.604	8.610
5	2.015	2.571	4.032	6.864

- The critical value of t= 2.776

Finding critical value of t

- If the calculated value of *t* is less than the critical value of *t* obtained from the table, the null hypothesis is not rejected.
- If the calculated value of *t* is greater than the critical value of *t* from the table, the null hypothesis is rejected.

Filling out summary table

- The following information is needed in a summary table

Descriptive statistics		
Mean Variance Standard deviation 1SD (68% Band) 2 SD (95% Band) 3 SD (99% Band) Number		
Results of *t* test		

summary table

- Example: Data obtained from a experiment comparing the number of un-popped seeds in popcorn brand A and popcorn brand B.

A	B
26	32
22	35
30	20
34	33

Is the difference significant?

- H_0 – Null hypothesis : $\overline{x}_A = \overline{x}_B$
- H_1 – Alternate hypothesis : $\overline{x}_A \mathrel{!=} \overline{x}_B$

- Determine mean, variance and standard deviation of samples.

$$\text{Mean } \bar{x}_A = \frac{\Sigma x}{n} = \frac{26+22+30+34}{4} = \mathbf{28}$$

$$\text{Mean } \bar{x}_B = \frac{\Sigma x}{n} = \frac{32+35+20+33}{4} = \mathbf{30}$$

variance

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

$$\text{Popcorn A} = \frac{(26-28)^2 + (22-28)^2 + (30-28)^2 + (34-28)^2}{3}$$

$$= \mathbf{26.6}$$

$$\text{Popcorn B} = \frac{(30-30)^2 + (35-30)^2 + (20-30)^2 + (33-30)^2}{3}$$

$$= \frac{0 + 25 + 100 + 9}{3} = \mathbf{44.67}$$

Standard deviation: $s = \sqrt{s^2}$

popcorn A

$$\sqrt{26.6} = \mathbf{5.16}$$

Popcorn B

$$\sqrt{44.67} = \mathbf{6.68}$$

Finding Calculated *t*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2_1 + s^2_2}{n}}}$$

$$t = \frac{28 - 30}{\sqrt{\frac{26.6 + 44.67}{4}}}$$

$$= \frac{-2}{4.22} = \mathbf{-0.47}$$

Determine critical value of *t*

- Select level of significance

α=.01

- Determine degrees of freedom

  degrees of freedom of A= 3 (i.e n-1)

  degrees of freedom of B= 3 (i.e n-1)

  total degrees of freedom = 6

- Critical value of *t* = 3.707

Calculated value of *t =-0.47* is less than critical value of *t* from the table, 3.707.

The null hypothesis is not rejected.

Filling out summary table

Descriptive statistics	popcorn A	popcorn B
Mean Variance Standard deviation	28 26.6 5.16	30 44.67 6.68
Number	4	
Results of *t* test	*t*= -0.47 *t* of -0.47 < 3.707	df=6 α=.01

Writing Conclusion

- Write a topic sentence stating the independent and dependent variables and a reference to a table or graph.
- Write sentences comparing the measures of central tendency of the groups.
- Write sentences describing the statistical tests, levels of significance, and the null hypothesis.
- Write sentences comparing the calculated value with the required statistical value. Make a statement about rejection of the null hypothesis.
- Write a sentence stating support of the research hypothesis by the data.