Advanced Statistics Course

Bloom Business School

By:
Ashraf Shaarawy

- Introduction to Quantitative Analysis
- Descriptive Analysis using SPSS
- Correlation/Inferential Statistics using SPSS

Session 1

- Statistical Thinking
- Types of Data
- Critical Thinking
- Summarizing and Graphing Data
- Frequency Distributions
- Histograms
- Statistical Graphics
- Measures of Center
- Measures of Variation
- Measures of Relative Standing and Boxplots
- Outliers

What is Statistics

- **Statistics**

  is the science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data

What is Statistics

- **Data**

  collections of observations (such as measurements, genders, survey responses)

- **Population**

  The complete collection of all individuals (scores, people, measurements, and so on) to be studied.

- **Census**

  Collection of data from every member of a population.

- **Sample**

  Sub-collection of members selected from a population.

What is Statistics

- **Parameter**

  a numerical measurement describing some characteristic of a **population.**

- **Statistic**

  a numerical measurement describing some characteristic of a **sample**.

Statistical Significance

- Consider the likelihood of getting the results by chance.
- If results could easily occur by chance, then they are *not statistically significant*.
- If the likelihood of getting the results is so small, then the results are *statistically significant*.

Data Types

- **Quantitative (or numerical) data**

It consists of numbers representing counts or measurements.

Example: Weights, Ages of respondents

- **Categorical (or qualitative or attribute) data**

It consists of names or labels (representing categories).

Example: Genders - Athletes shirt numbers

Levels of Measurement

Another way to classify data is to use levels of measurement. Four of these levels are used in measurement.

- Nominal - categories only
- Ordinal - categories with some order
- Interval - differences but no natural starting point
- Ratio - differences <u>and</u> a natural starting point

Sampling Design

- There are two major types of sampling design: **probability** and **nonprobability** sampling.
- **Probability sampling**: the elements in the population have some known, nonzero chance or probability of being selected as sample subjects.
- N**onprobability sampling**: the elements do not have a known or predetermined chance of being selected as subjects.

Sampling Design

- Probability sampling designs are used when the representativeness of the sample is important for generalizability.
- Nonprobability sampling is used when time or other factors, rather than generalizability, become critical.

Methods of Sampling

Probability Sampling

- Random
- Systematic
- Stratified

Non Probability Sampling

- Convenience
- Cluster
- Purposive (Judgmental or Expert)
- Quota

Error Types

No matter how well you plan and execute the sample collection process, there is likely to be some error in the results.

- **Sampling error**

  the difference between a sample result and the true population result; it results from chance sample fluctuations

- **Non-sampling error**

  sample data incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly)

Important Characteristics of Data

1. **Center**: A representative or average value that indicates where the middle of the data set is located.
2. **Variation**: A measure of the amount that the data values vary.
3. **Distribution**: The nature or shape of the spread of data over the range of values (such as bell-shaped, uniform, or skewed).
4. **Outliers**: Sample values that lie very far away from the vast majority of other sample values.

Frequency Distribution

When working with large data sets, it is often helpful to organize and summarize data by constructing a table called a frequency distribution.

The importance of them is what they tell us about data sets.

It helps us understand the nature of the *distribution* of a data set.

Frequency Distribution

Frequency Distribution
(or Frequency Table)

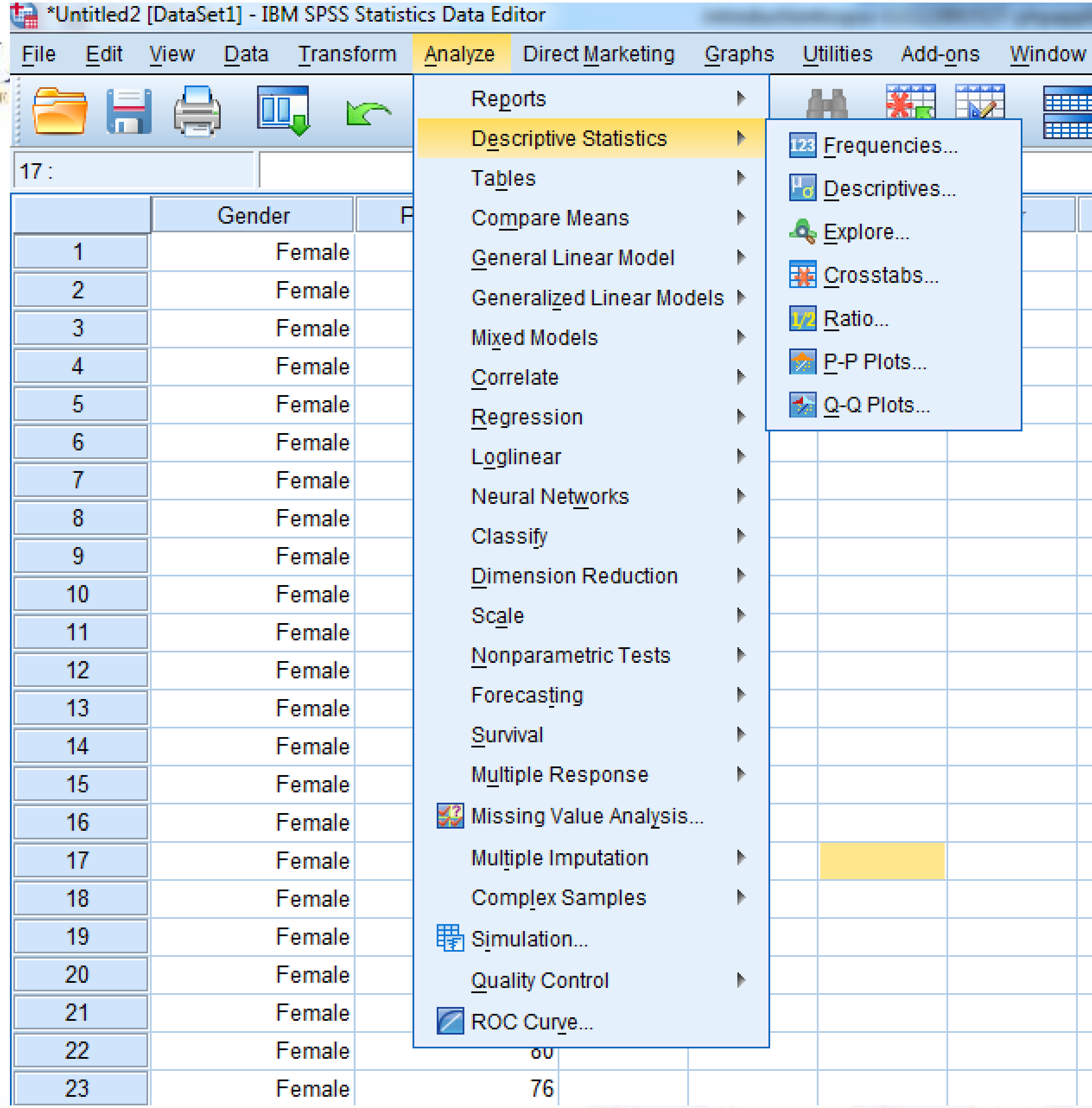
shows how a data set is partitioned among all of several categories (or classes) by listing all of the categories along with the number of data values in each of the categories.

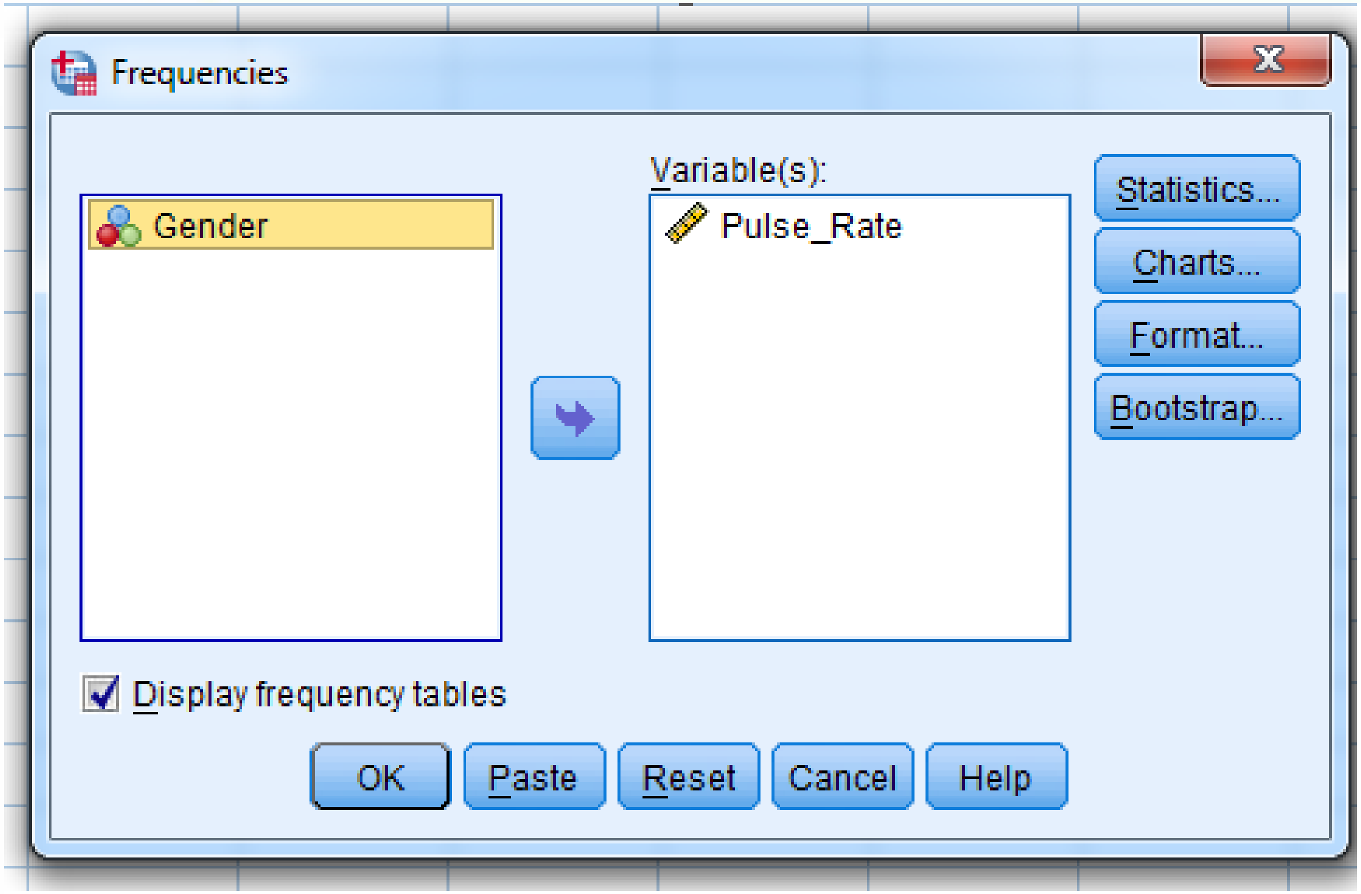
Pulse Rates of Females and Males Example 1

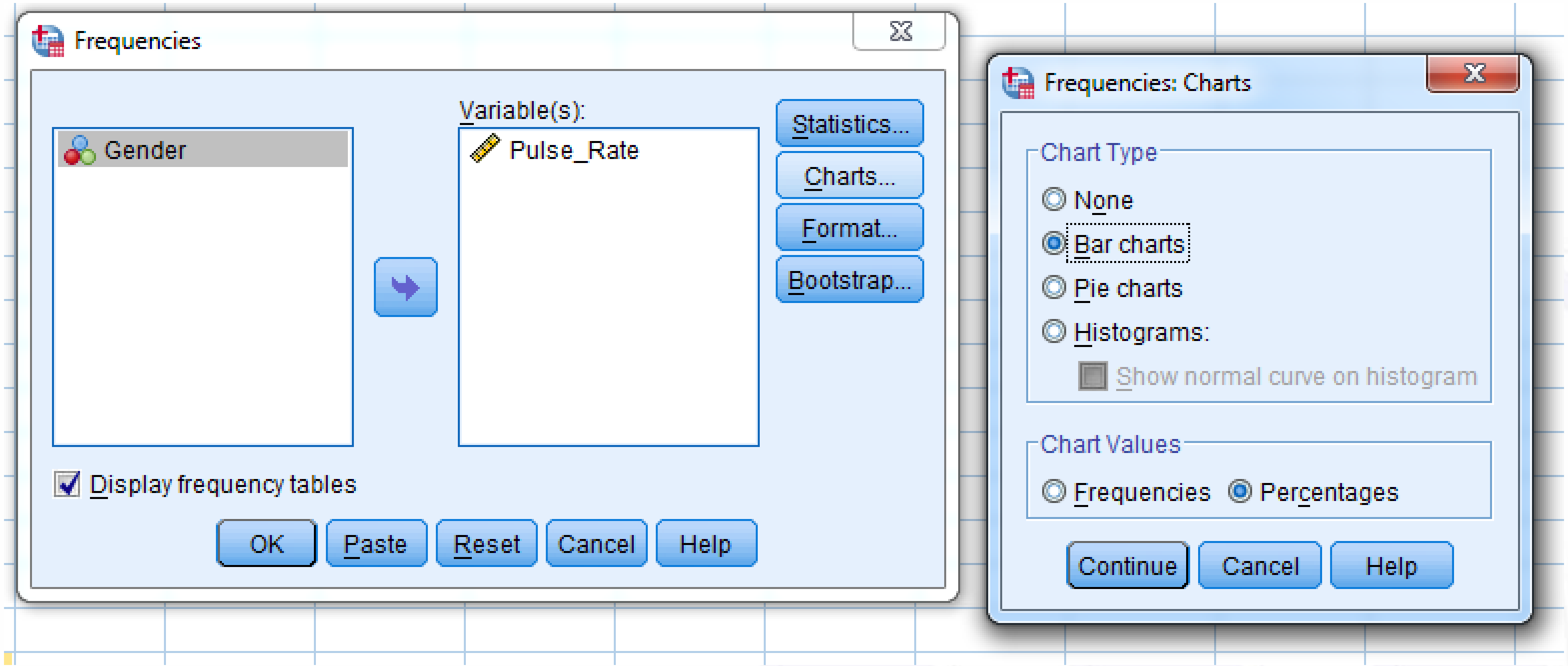
Original Data

Pulse Rates (beats per minute) of Females and Males

Females																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	96	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
Males																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72


*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
17 :
Gender
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
1 Female
2 Female
3 Female
4 Female
5 Female
6 Female
7 Female
8 Female
9 Female
10 Female
11 Female
12 Female
13 Female
14 Female
15 Female
16 Female
17 Female
18 Female
19 Female
20 Female
21 Female
22 Female
23 Female 76


Frequencies
Variable(s):
Gender
Pulse_Rate
Statistics...
Charts...
Format...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help


Frequencies
Variable(s):
Gender
Pulse_Rate
Statistics...
Charts...
Format...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help
Frequencies: Charts
Chart Type
None
Bar charts
Pie charts
Histograms:
Show normal curve on histogram
Chart Values
Frequencies
Percentages
Continue
Cancel
Help

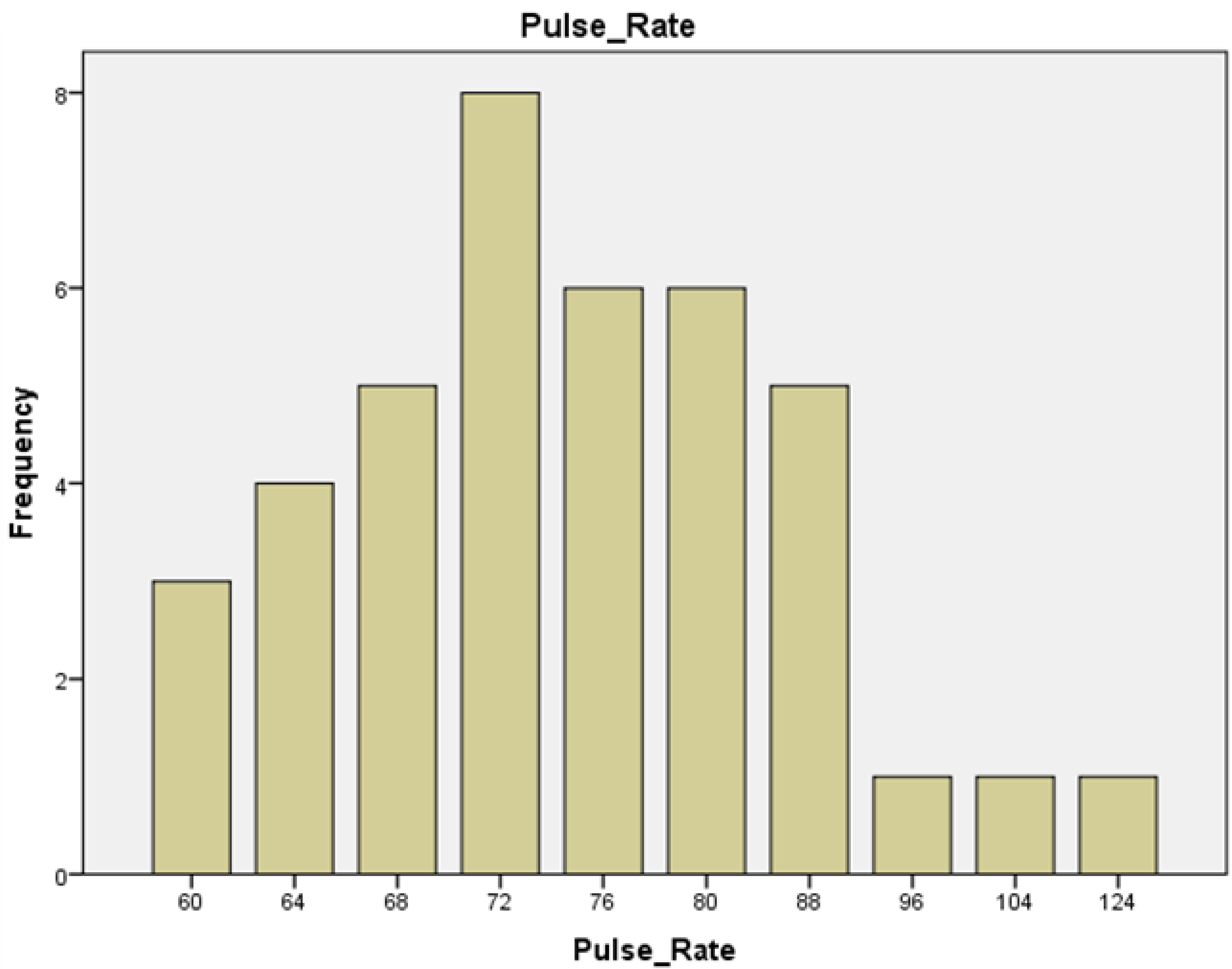
Frequency Distribution Tables

Pulse_Rate

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	60	3	7.5	7.5	7.5
	64	4	10.0	10.0	17.5
	68	5	12.5	12.5	30.0
	72	8	20.0	20.0	50.0
	76	6	15.0	15.0	65.0
	80	6	15.0	15.0	80.0
	88	5	12.5	12.5	92.5
	96	1	2.5	2.5	95.0
	104	1	2.5	2.5	97.5
	124	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

Bar Chart

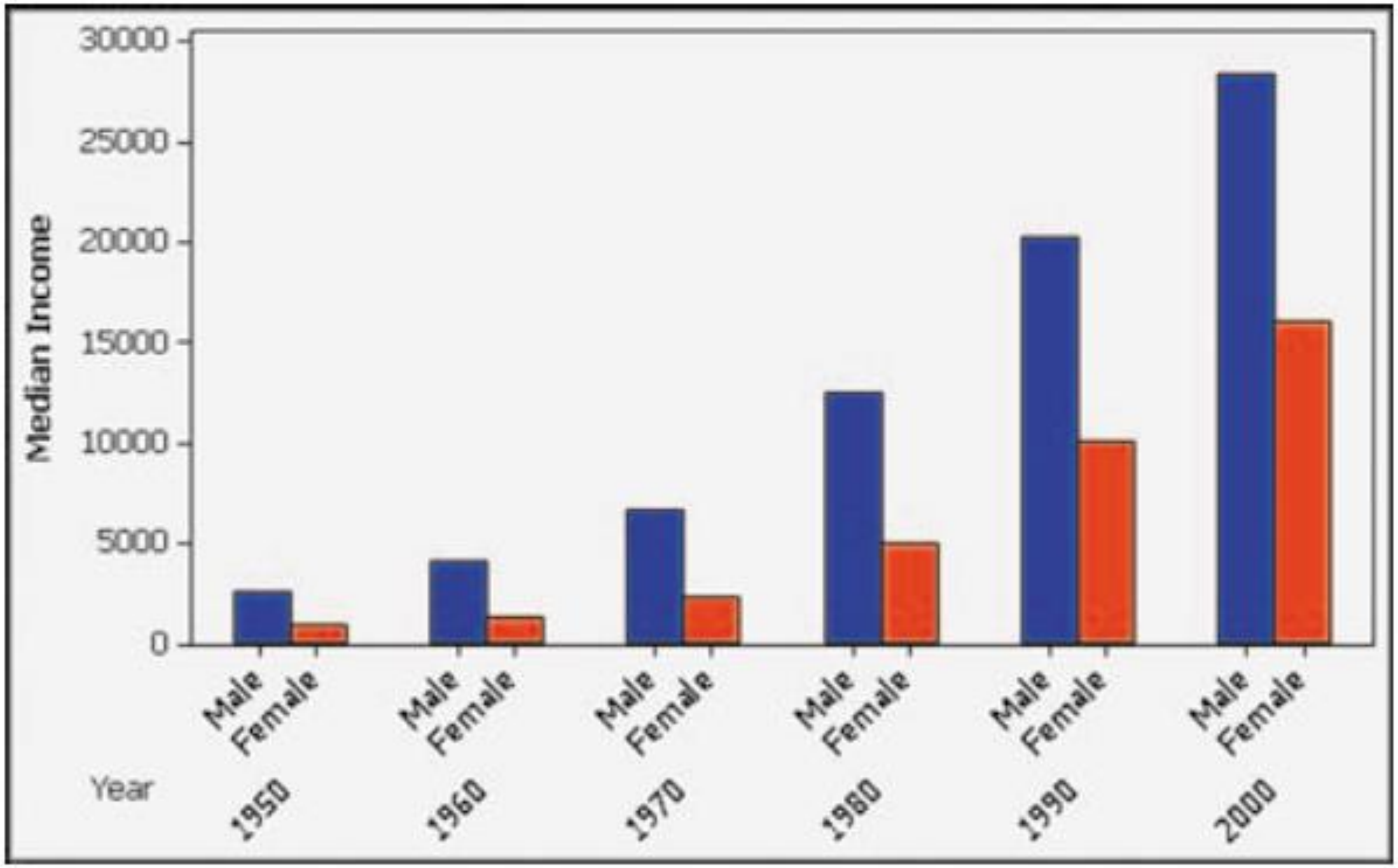
a graphic version of a frequency distribution.



Bar Graph

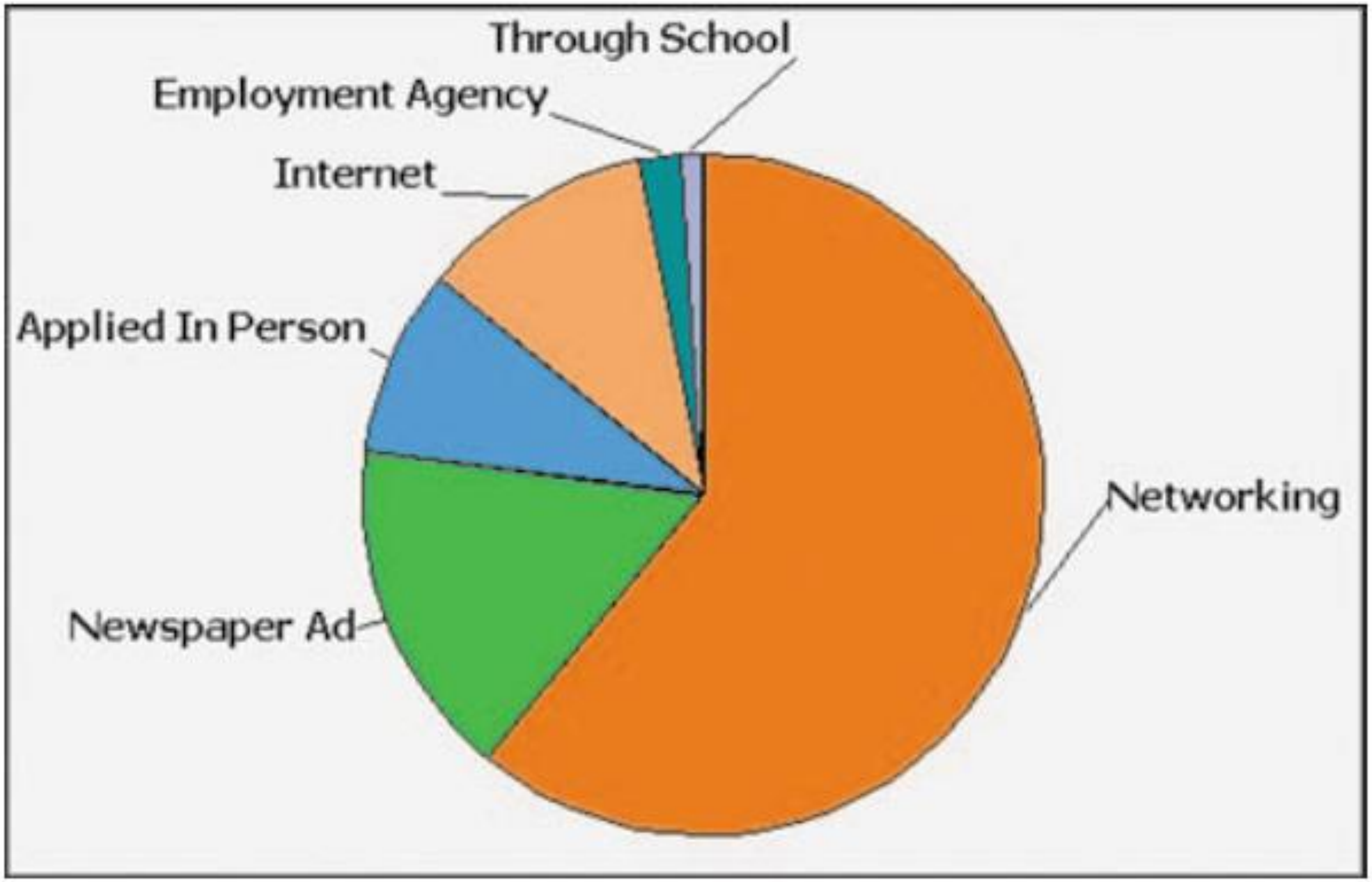
Uses bars of equal width to show frequencies of categories of qualitative data.

Median Income of Males and Females



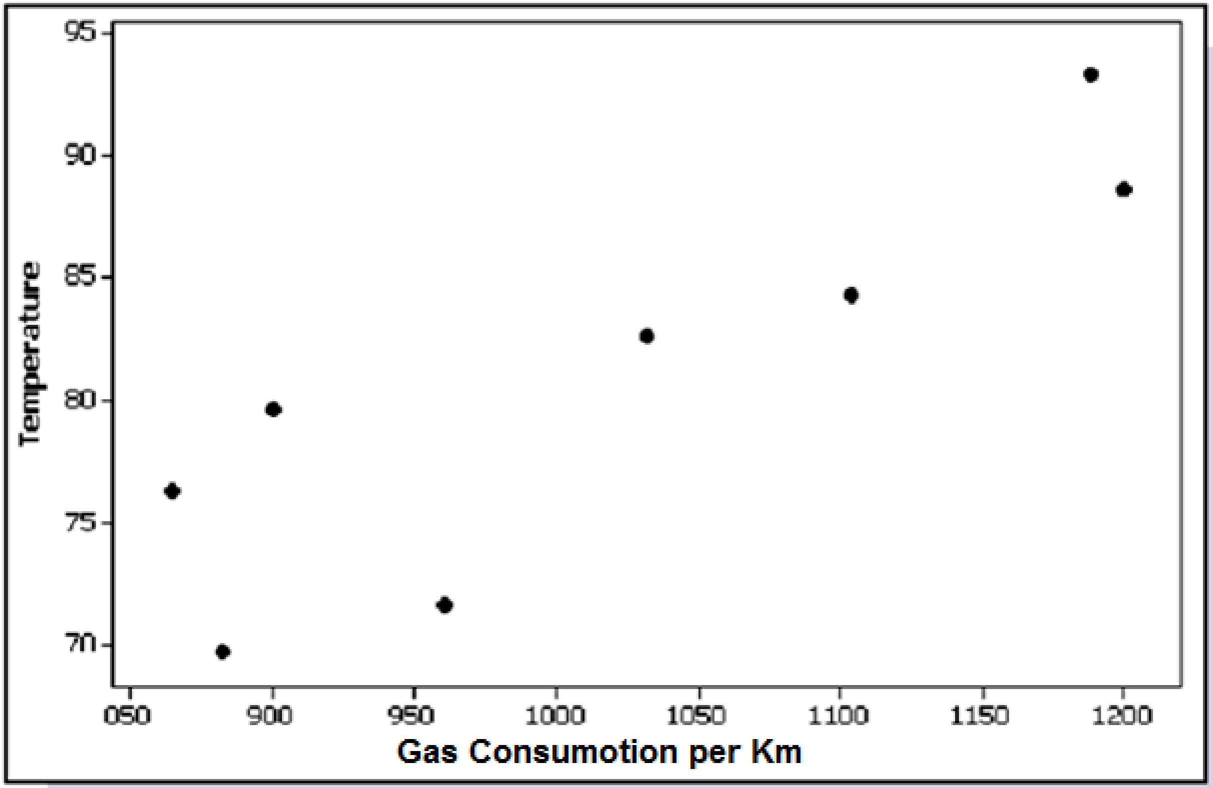
Pie Chart

A graph depicting qualitative data as slices of a circle, size of slice is proportional to frequency count



Scatter Plot (or Scatter Diagram)

A plot of paired (*x*,*y*) data with a horizontal *x*-axis and a vertical *y*-axis. Used to determine whether there is a relationship between the two variables



Activity 1

Please Open "Activity 1" file

Produce Pie Charts – Education-Age-Gender-Management Level

Bar Graphs

Histogram of the Activity 1 file

Measures of Center

Arithmetic Mean

- **Arithmetic Mean (Mean)**

  the measure of center obtained by adding the values and dividing the total by the number of values

  Most people call it "*Average*".

Arithmetic Mean

$\overline{x}$ is pronounced 'x-bar' and denotes the mean of a set of sample values

$$\overline{x} = \frac{\sum x}{n}$$

μ is pronounced 'mu' and denotes the mean of all values in a population

$$\mu = \frac{\sum x}{N}$$

Notation

Σ denotes the sum of a set of values.

x is the variable usually used to represent the individual data values.

n represents the number of data values in a sample.

N represents the number of data values in a population.

Mean

- **Advantages**

  Is relatively reliable, means of samples drawn from the same population don't vary as much as other measures of center

  Takes every data value into account

- **Disadvantage**

  Is sensitive to every data value, one extreme value can affect it dramatically; is not a resistant measure of center.

Median

- Median

  the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude

- often denoted by $\tilde{x}$ (pronounced 'x-tilde')

- is not affected by an extreme value - is a

  resistant measure of the center

Finding the Median

First *sort* the values (arrange them in order), the follow one of these

1. If the number of data values is odd, the median is the number located in the exact middle of the list.
2. If the number of data values is even, the median is found by computing the mean of the two middle numbers.

5.40 1.10 0.42 0.73 0.48 1.10

0.42 0.48 0.73 1.10 1.10 5.40

(in order - even number of values – no exact middle shared by two numbers)

$$\frac{0.73 + 1.10}{2}$$

MEDIAN is 0.915

5.40 1.10 0.42 0.73 0.48 1.10 0.66

0.42 0.48 0.66 $\boxed{0.73}$ 1.10 1.10 5.40

(in order - odd number of values)

exact middle **MEDIAN is 0.73**

Mode

- Mode
  the value that occurs with the greatest frequency

- Data set can have one, more than one, or no mode

Mode is the only measure of central tendency that can be used with nominal data

Mode - Examples

a. 5.40 1.10 0.42 0.73 0.48 1.10

b. 27 27 27 55 55 55 88 88 99

c. 1 2 3 6 7 8 9 10

Mode is 1.10

Bimodal - 27 & 55

No Mode

Skewed and Symmetric

- **Symmetric**

  distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half

- **Skewed**

  distribution of data is skewed if it is not symmetric and extends more to one side than the other

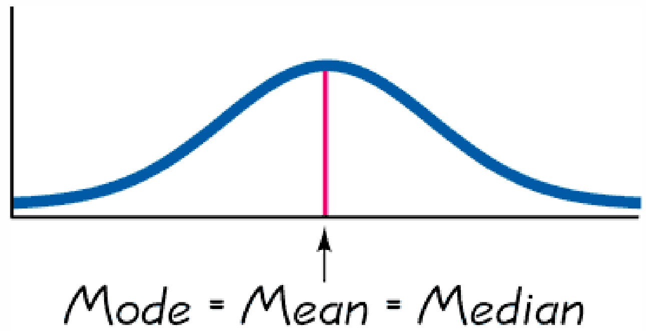
- **Skewed to the left**

  (Negatively skewed) have a longer left tail, The mean is to the left of the median.

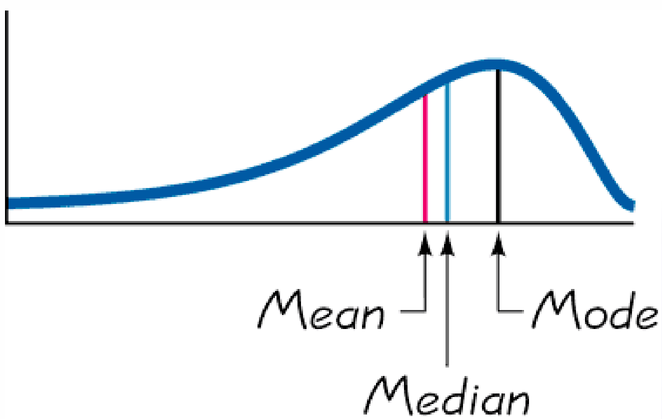
- **Skewed to the right**

  (Positively skewed) have a longer right tail, The mean is to the right of the median.

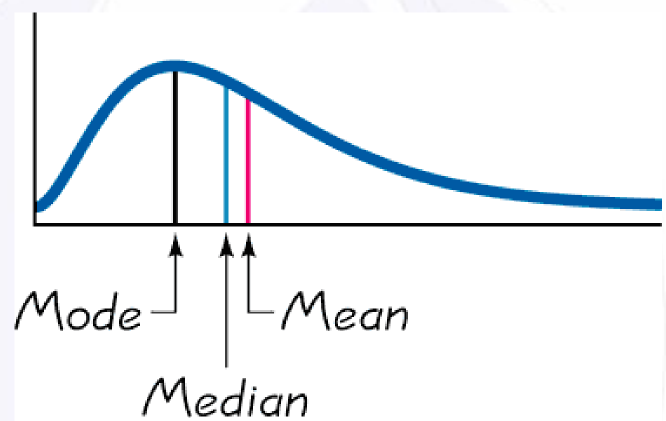
Skewness



(b) Symmetric



(a) Skewed to the Left
(Negatively)



(c) Skewed to the Right
(Positively)

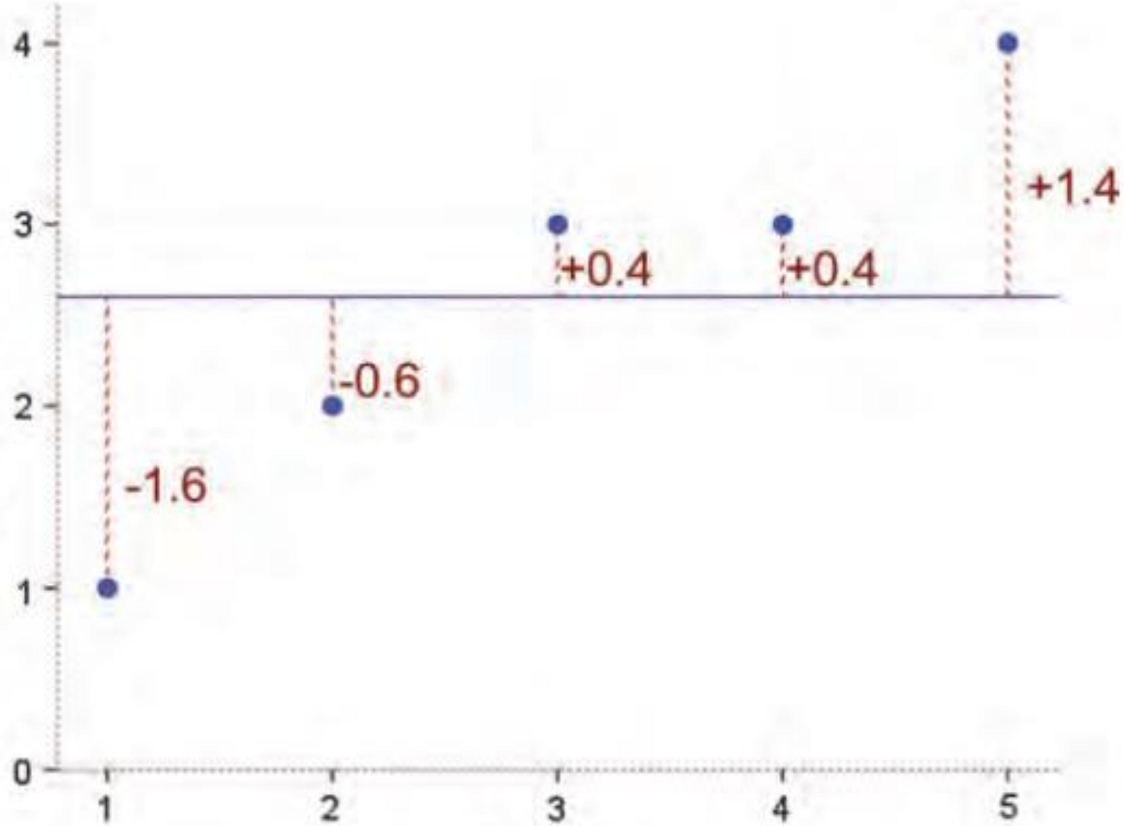
Measures of Variation

Standard Deviation

The **standard deviation** of a set of sample values, denoted by *s*, is a measure of variation of values about the mean.

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

Standard Deviation



$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Variance

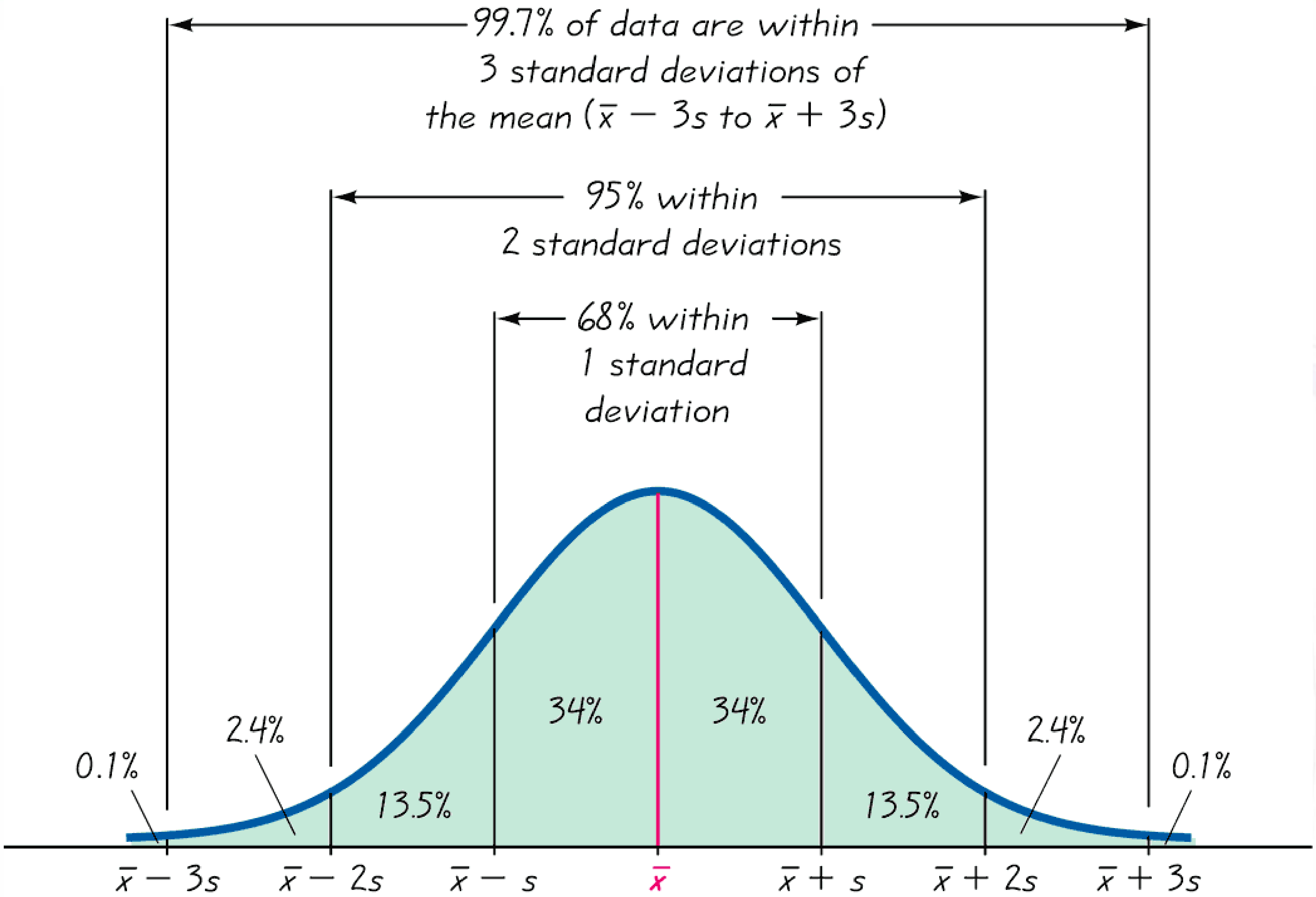
- The variance of a set of values is a measure of variation equal to the square of the standard deviation.
- Sample variance: s^2 - Square of the sample standard deviation s
- Population variance: σ^2 - Square of the population standard deviation σ

Empirical (or 68-95-99.7) Rule

For data sets having a distribution that is approximately bell shaped, the following properties apply:

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.

The Empirical Rule

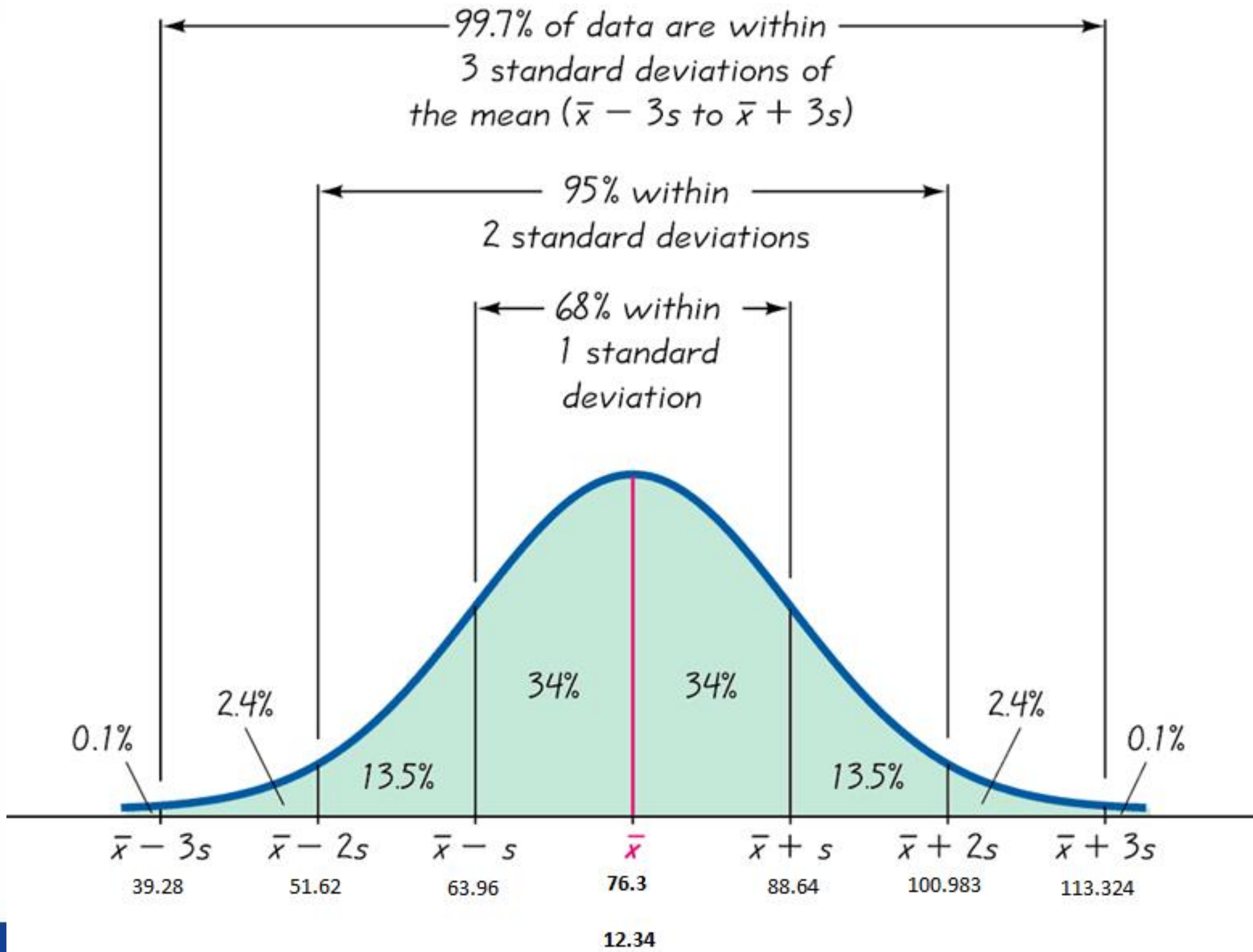


Session 2

- The Standard Normal Distribution
- Applications of Normal Distributions
- Sampling Distributions and Estimators
- The Central Limit Theorem
- Assessing Normality

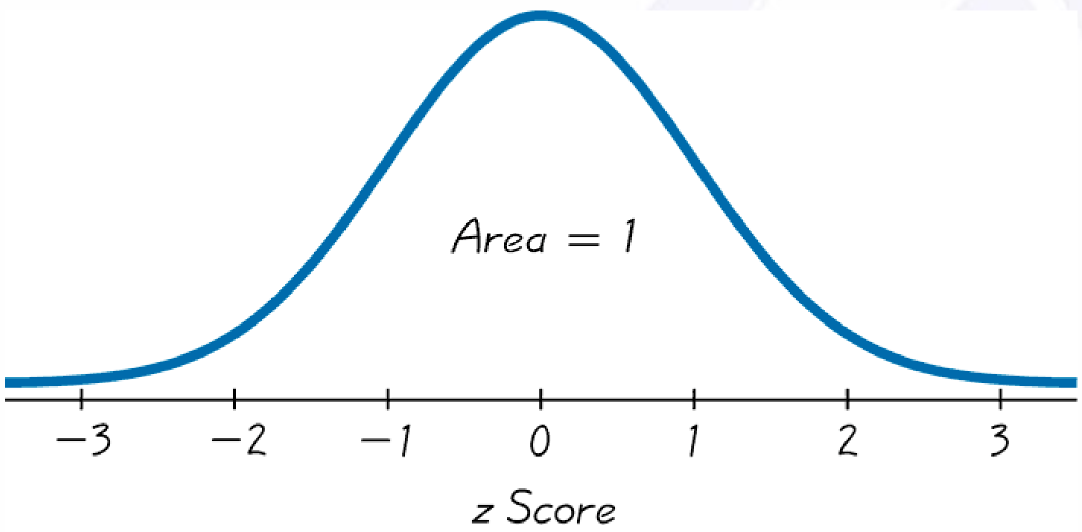
Basics of *z* Scores

The Empirical Rule



Standard Normal Distribution

The standard normal distribution is a normal probability distribution with $\mu = 0$ and $\sigma = 1$. The total area under its density curve is equal to 1.



Z - Score

- *z* Score (or standardized value)

It is the number of standard deviations that a given value *x* is above or below the mean

Sample

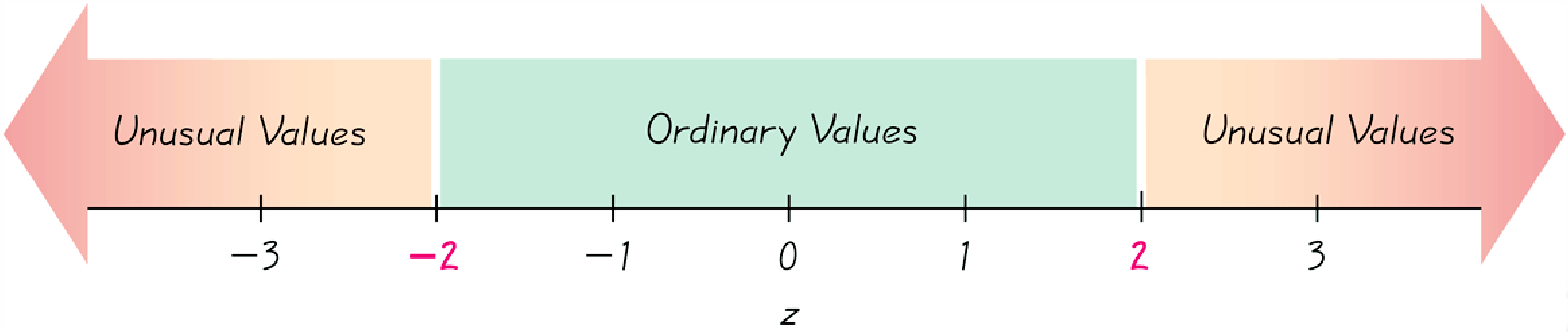
$$z = \frac{x - \bar{x}}{s}$$

Population

$$z = \frac{x - \mu}{\sigma}$$

Round *z* scores to 2 decimal places

Interpreting Z Scores

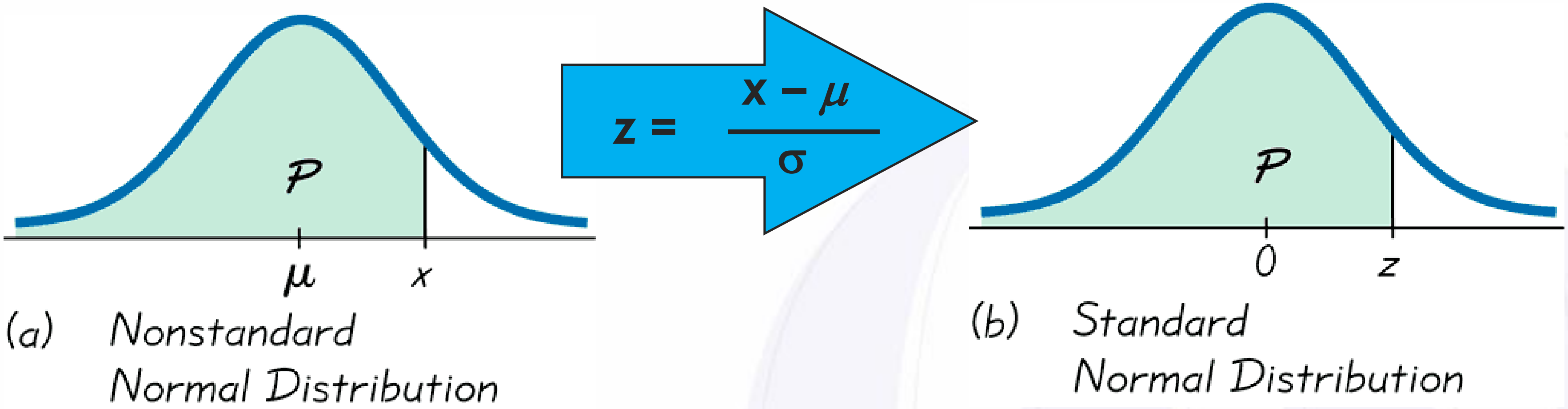


Whenever a value is less than the mean, its corresponding z score is negative

Ordinary values: $-2 \leq$ z score ≤ 2

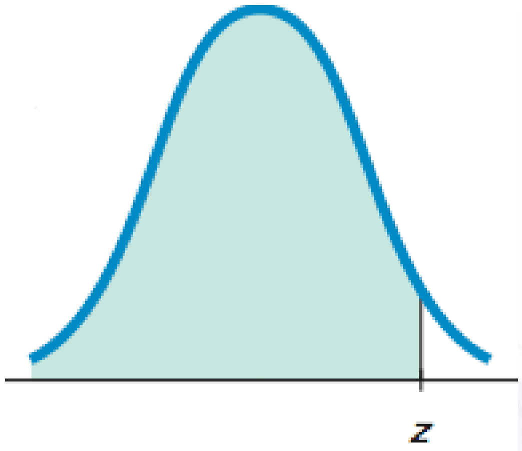
Unusual Values: z score < -2 or z score > 2

Converting to a Standard Normal Distribution



Methods for Finding Normal Distribution Areas

It is not easy to find areas in the adjacent Figure, so mathematicians have calculated many different areas under the curve, and those areas are included in the Table in the next slide



Z Score Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Example

- The Precision Scientific Instrument Company manufactures thermometers that are supposed to give readings of 0 C at the freezing point of water. Tests on a large sample of these instruments reveal that at the freezing point of water, some thermometers give readings below 0 (denoted by negative numbers) and some give readings above 0 (denoted by positive numbers).
- Assume that the mean reading is 0 C and the standard deviation of the readings is 1.00 C. Also assume that the readings are normally distributed. If one thermometer is randomly selected, find the probability that, at the freezing point of water, the reading is less than 1.27 .

Example

- The following example requires that we find the probability associated with a *z* score less than 1.27.
- Begin with the *z* score of 1.27 by locating 1.2 in the left column; next find the value in the adjoining row of probabilities that is directly below 0.07, as shown in the following excerpt from Table.

$$z = \frac{x - \mu}{\sigma}$$

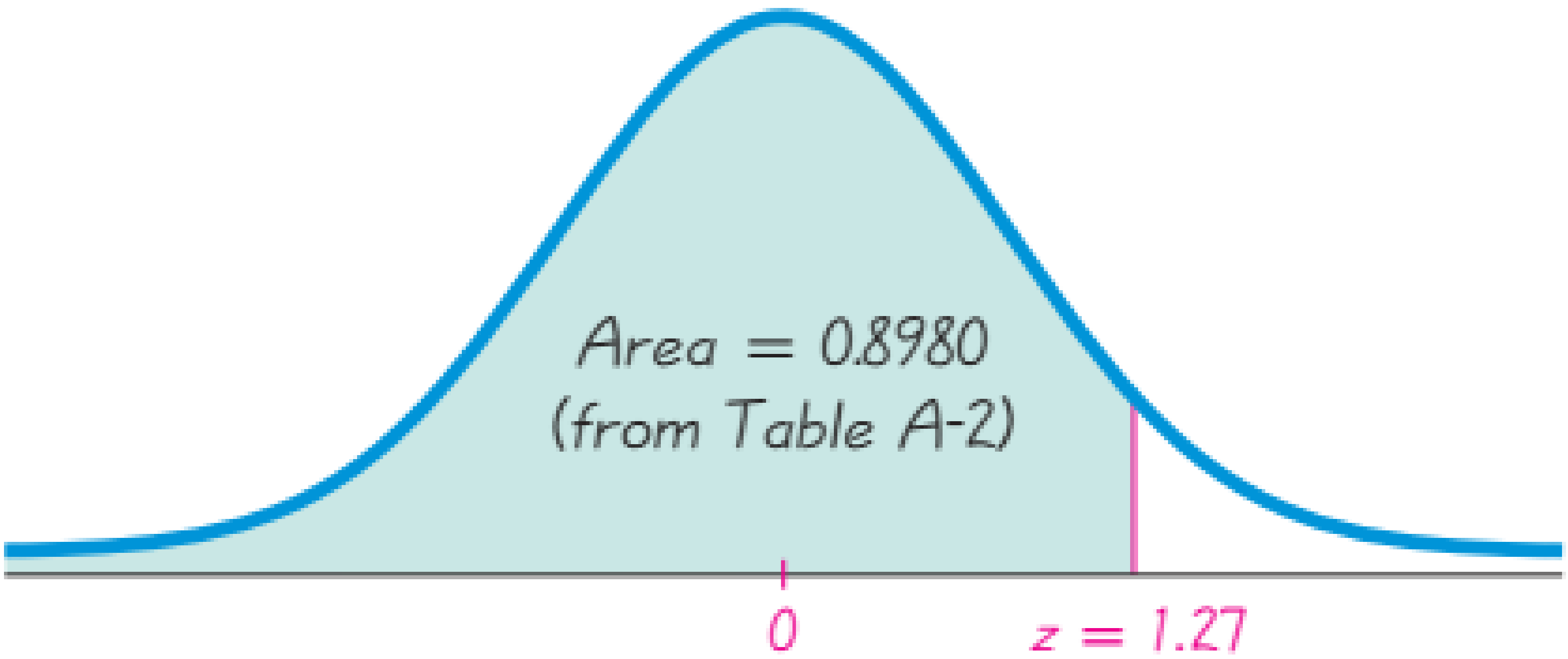
Round z scores to 2 decimal places

Example

TABLE A-2 (*continued*) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

Example

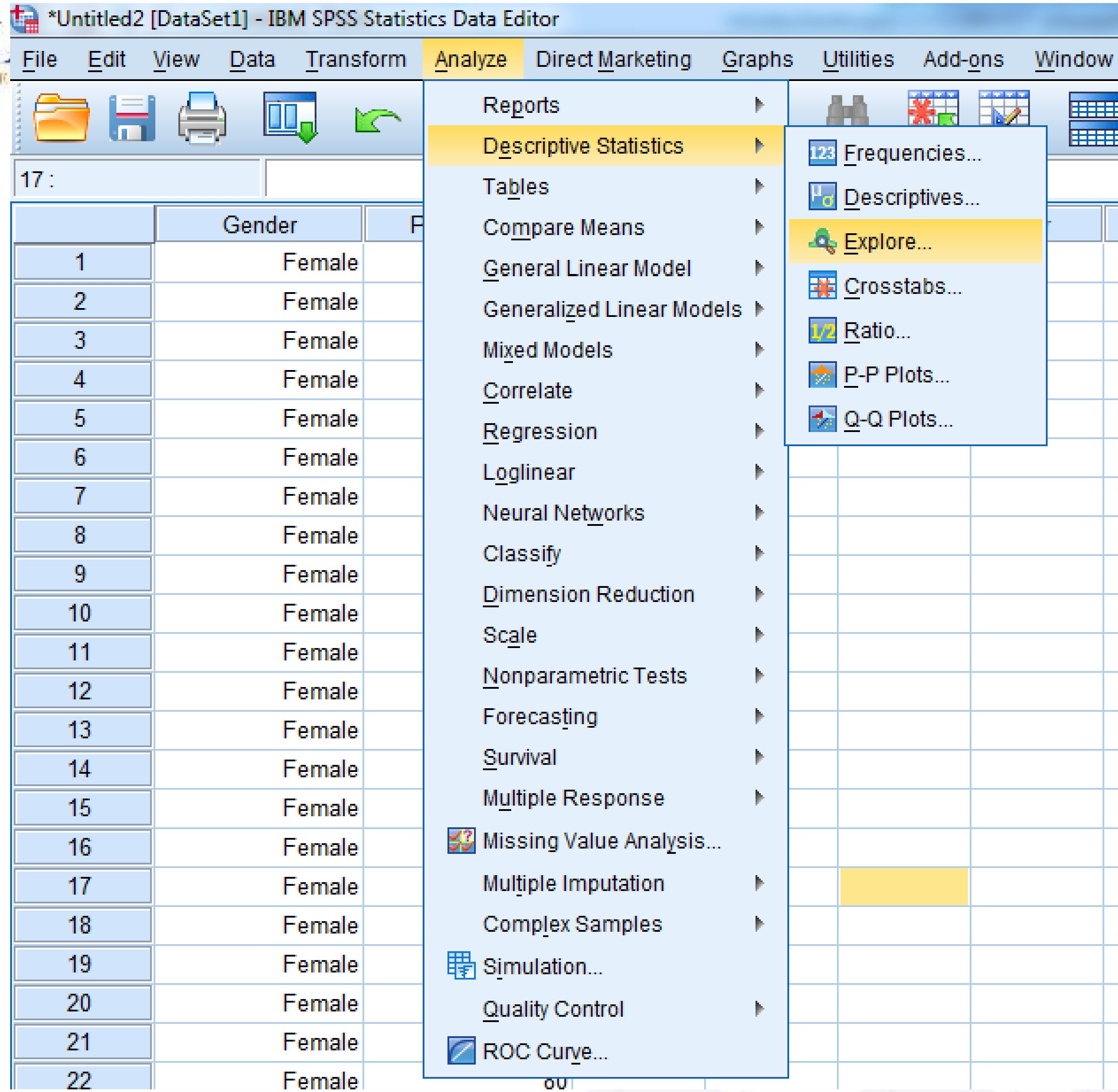


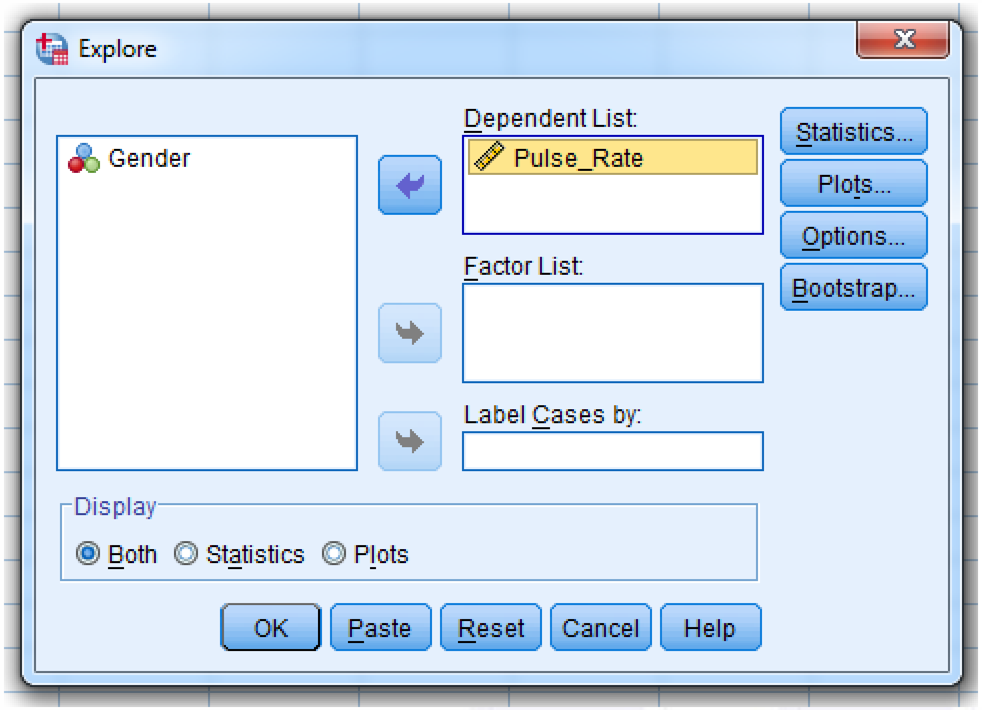
Assessing Normality

Assessing Normality

This section presents criteria for determining whether the requirement of a normal distribution is satisfied.

The criteria involve visual inspection of a histogram to see if it is roughly bell shaped, identifying any outliers, and constructing a graph called a normal quantile plot.


*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
17 :
Gender
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
1 Female
2 Female
3 Female
4 Female
5 Female
6 Female
7 Female
8 Female
9 Female
10 Female
11 Female
12 Female
13 Female
14 Female
15 Female
16 Female
17 Female
18 Female
19 Female
20 Female
21 Female
22 Female


Explore
Dependent List:
Pulse_Rate
Gender
Factor List:
Label Cases by:
Statistics...
Plots...
Options...
Bootstrap...
Display
Both
Statistics
Plots
OK
Paste
Reset
Cancel
Help

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Pulse_Rate	.184	40	.002	.866	40	.000

a. Lilliefors Significance Correction

- If the **Sig.** value of the Shapiro-Wilk Test is greater than 0.05, the data is normal.

- If it is below 0.05, the data significantly deviate from a normal distribution.

Activity 3

- Assessing the Normality of Activity 1 File

Sampling distribution of a statistic

The sampling distribution of a statistic (such as the sample mean or sample standard deviation) is the distribution of all values of the statistic when all possible samples of the same size *n* are taken from the same population. (The sampling distribution of a statistic is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Sampling distribution of the mean

The **sampling distribution of the mean** is the distribution of sample means, with all samples having the same sample size *n* taken from the same population.

- **Properties**
  - Sample means target the value of the population mean. (That is, the mean of the sample means is the population mean.)
  - The distribution of the sample means tends to be a normal distribution.

Sampling distribution of the variance

The **sampling distribution of the variance** is the distribution of sample variances, with all samples having the same sample size *n* taken from the same population.

- **Properties**
  - Sample variances target the value of the population variance. (That is, the mean of the sample variances is the population variance.)
  - The distribution of the sample variances tends to be a distribution skewed to the right.

The Central Limit Theorem

Central Limit Theorem

The *Central Limit Theorem* tells us that for a population with *any* distribution, the distribution of the sample means approaches a normal distribution as the sample size increases.

The procedure in this section form the foundation for estimating population parameters and hypothesis testing.

Central Limit Theorem

Given:

1. The random variable x has a distribution (which may or may not be normal) with mean μ and standard deviation σ.
2. Simple random samples all of size n are selected from the population. (The samples are selected so that all possible samples of the same size n have the same chance of being selected.)

Central Limit Theorem – cont.

Conclusions:

1. The distribution of sample $\bar{x}$ will, as the sample size increases, approach a normal distribution.

2. The mean of the sample means is the population mean μ.

3. The standard deviation of all sample means is $\sigma/\sqrt{n}$.

Practical Rules Commonly Used

1. For samples of size *n* larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets closer to a normal distribution as the sample size *n* becomes larger.

2. If the original population is *normally distributed*, then for any sample size *n*, the sample means will be normally distributed (not just the values of *n* larger than 30).

Practical Rules Commonly Used

the mean of the sample means

$$\mu_{\bar{x}} = \mu$$

the standard deviation of sample mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(often called the standard error of the mean)

Estimating a Population Mean

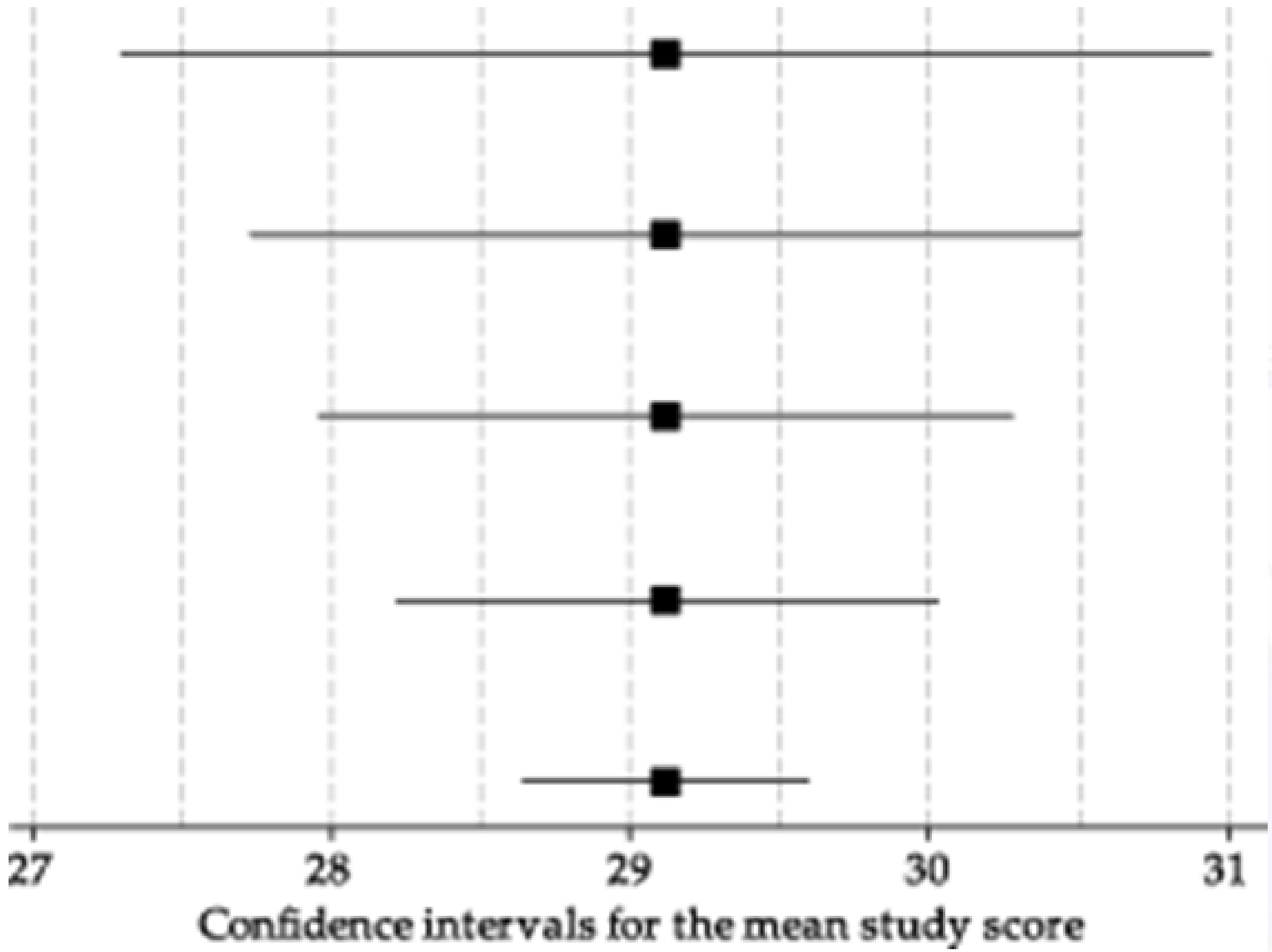
Point estimate

- A point estimate is a single value (or point) used to approximate a population parameter.
- The sample mean is the best point estimate of the population mean.

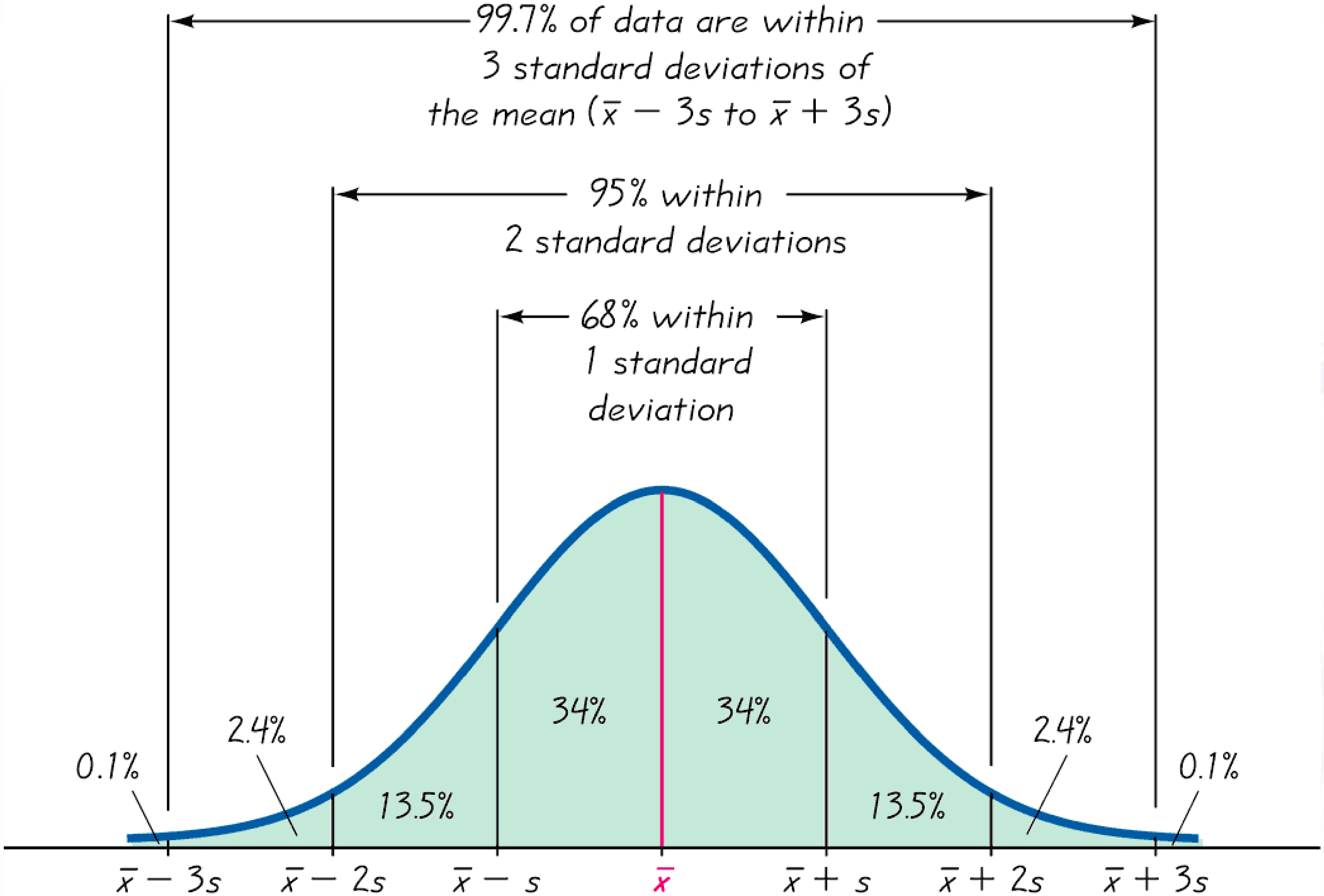
Confidence Interval

A confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter. A confidence interval is sometimes abbreviated as CI.

Confidence Interval



The Empirical Rule



Significance Level

The significance level (denoted by α) is the probability that the test statistic (mean) will fall in the unlikely region.
Common choices for α are 0.10, 0.05 and 0.01.

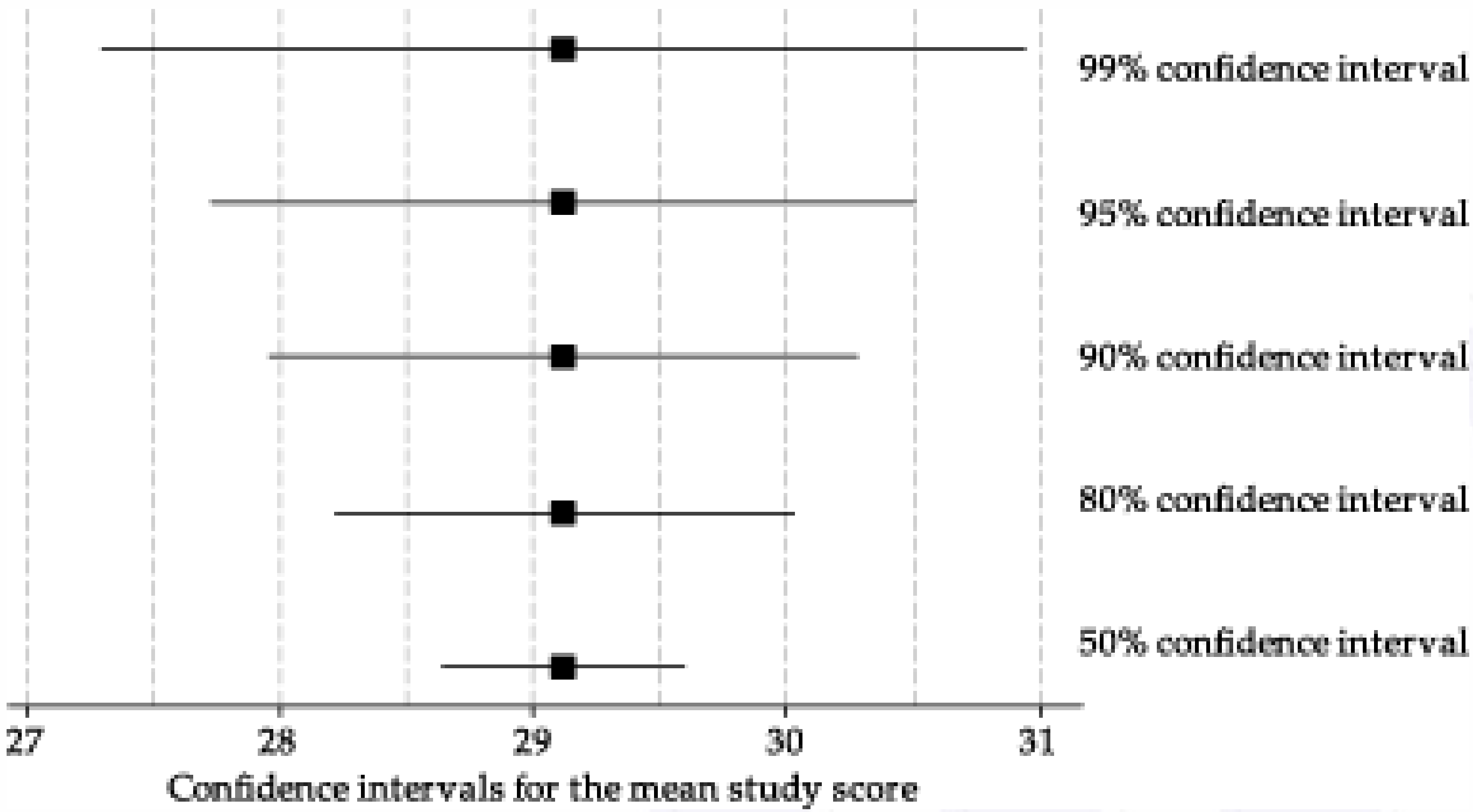
Confidence Level

A confidence level is the probability $1 - \alpha$ (often expressed as the equivalent percentage value) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. (The confidence level is also called degree of confidence, or the confidence coefficient.)

Most common choices are 90%, 95%, or 99%.

(α = 10%), (α = 5%), (α = 1%)

Confidence Level & Confidence Interval



Interpreting a Confidence Interval

We must be careful to interpret confidence intervals correctly. There is a correct interpretation and many different and creative incorrect interpretations of the confidence interval $0.677 < \mu < 0.723$.

"We are 95% confident that the interval from 0.677 to 0.723 actually does contain the true value of the population mean μ."

This means that if we were to select many different samples of size ***n*** and construct the corresponding confidence intervals, 95% of them would actually contain the value of the population mean μ.

(Note that in this correct interpretation, the level of 95% refers to the success rate of the process being used to estimate the proportion.)

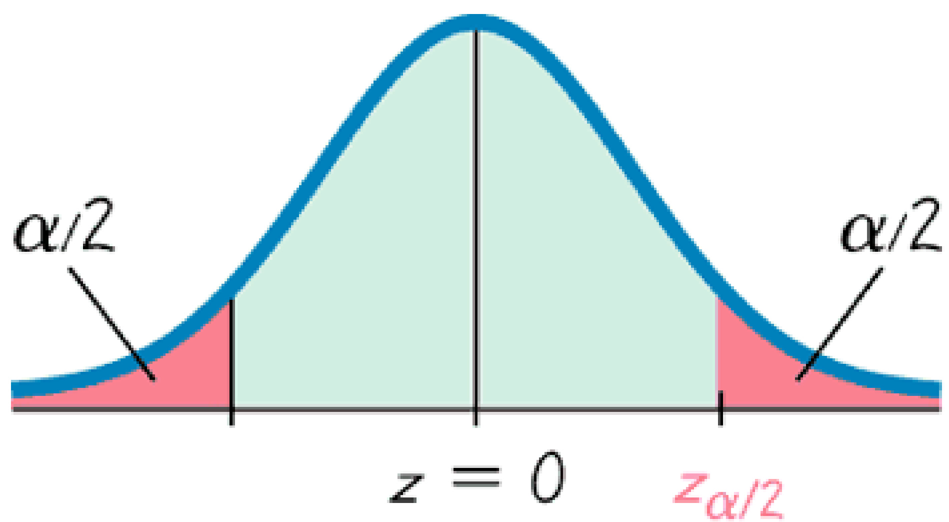
Critical Value

A critical value is the number on the borderline separating sample statistics that are likely to occur from those that are unlikely to occur. The number $z_{\alpha/2}$ is a critical value that is a z score with the property that it separates an area of $\alpha/2$ in the right tail of the standard normal distribution.

Critical Values

A standard *z* score can be used to distinguish between sample statistics that are likely to occur and those that are unlikely to occur. Critical values are based on the following observations:

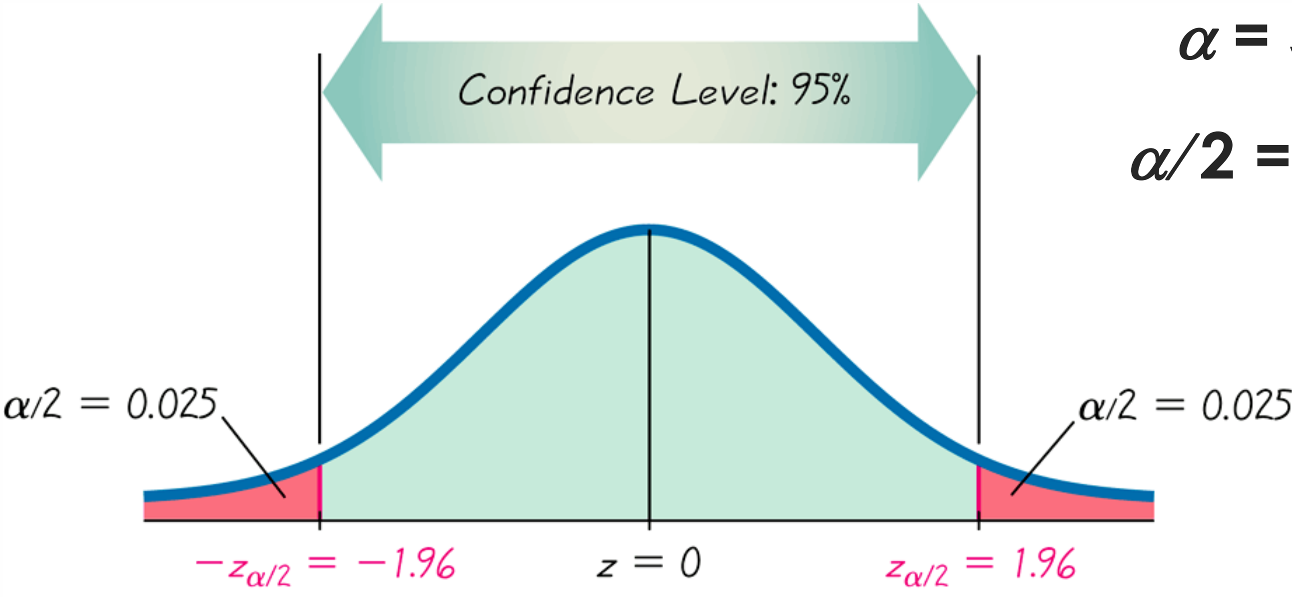
1. **Under certain conditions, the sampling distribution of sample means can be approximated by a normal distribution.**
2. **A z score associated with a sample mean has a probability of $\alpha/2$ of falling in the right tail.**



Critical Values

3. The z score separating the right-tail region is commonly denoted by $z_{\alpha/2}$ and is referred to as a critical value because it is on the borderline separating z scores from sample proportions that are likely to occur from those that are unlikely to occur.

Finding $z_{\alpha/2}$ for a 95% Confidence Level



$\alpha = 5\%$

$\alpha/2 = 2.5\% = .025$

$-z_{\alpha/2}$ $z_{\alpha/2}$

Critical Values

Estimating a Population Mean: σ Known

Z - Score

z Score (or standardized value)

It is the number of standard deviations that a given value x is above or below the mean

Population

$$z = \frac{x - \mu}{\sigma}$$

Round z scores to 2 decimal places

Practical Rules Commonly Used

the mean of the sample means

$$\mu_{\bar{x}} = \mu$$

the standard deviation of sample mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(often called the standard error of the mean)

Confidence Interval for Estimating a Population Mean (with σ Known)

μ = population mean

σ = population standard deviation

$\bar{x}$ = sample mean

n = number of sample values

E = margin of error

$z_{\alpha/2}$ = z score separating an area of $\alpha/2$ in the right tail of the standard normal distribution

Z - Score

- *z* Score (or standardized value)

It is the number of standard deviations that a given value *x* is above or below the mean

$$z = \frac{\bar{x} \pm \mu_x}{\sigma_x}$$

$$\bar{x} \pm \mu = z\sigma_x$$

$$\mu = z\,\sigma_x \pm \bar{x}$$

Confidence Interval for Estimating a Population Mean (with σ Known)

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{where} \quad E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

or $$\bar{x} \pm E$$

or $$(\bar{x} - E, \bar{x} + E)$$

The two values x – *E* and x + *E* are called confidence interval limits.

Session 3

- Basics of Hypothesis Testing
- Testing a Claim About a Mean: σ Known
- Testing a Claim About a Mean: σ Not Known
- Inferences About Two Means: Independent Samples
- Inferences from Dependent Samples
- Analysis of Variance

Estimating a Population Mean: σ Not Known

t-Distribution

The *t*-distribution, also known as the Student's *t*-distribution, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails. *t*-distributions have a greater chance for extreme values than normal distributions, hence the fatter tails.

t-Distribution

The *t*-distribution is a continuous probability distribution of the z-score (*t* -Values) when the estimated standard deviation is used in the denominator rather than the true standard deviation.

The *t*-distribution, like the normal distribution, is bell-shaped and symmetric, but it has heavier tails, which means it tends to produce values that fall far from its mean.

Margin of Error E for Estimate of μ (With σ Not Known)

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ has $n - 1$ degrees of freedom.

$$\bar{x} - E < \mu < \bar{x} + E$$

Notation

μ = population mean

$\bar{x}$ = sample mean

s = sample standard deviation

n = number of sample values

E = margin of error

$t_{\alpha/2}$ = critical t value separating an area of $\alpha/2$ in the right tail of the t distribution

Degrees of freedom

The number of **degrees of freedom** for a collection of sample data is:

The number of sample values that can vary after certain restrictions have been imposed on all data values. The degree of freedom is often abbreviated df.

degrees of freedom = n – 1

in this section.

Degrees of freedom Example

- Consider a data sample consisting of, for the sake of simplicity, five positive integers. The values could be any number with no known relationship between them. This data sample would, theoretically, have five degrees of freedom.

Degrees of freedom Example

- Four of the numbers in the sample are {3, 8, 5, and 4} and the average of the entire data sample is revealed to be 6.
- This must mean that the fifth number has to be 10. It can be nothing else. It does not have the freedom to vary.
- So the Degrees of Freedom for this data sample is 4.

Basics of Hypothesis Testing

Hypothesis Testing

In statistics, a hypothesis is a claim or statement about a property of a population.

A hypothesis test (or test of significance) is a standard procedure for testing a claim about a property of a population.

Components of a Formal Hypothesis Test

Null Hypothesis: H_0

- The null hypothesis (denoted by H_0) is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is equal to some claimed value.
- We test the null hypothesis directly.
- Either reject H_0 or fail to reject H_0.

Alternative Hypothesis: H_1

- The alternative hypothesis (denoted by H_1 or H_a or H_A) is the statement that the parameter has a value that somehow differs from the null hypothesis.

- The symbolic form of the alternative hypothesis must use one of these symbols: $\neq$, $<$, $>$.

Note about Forming Your Own Claims (Hypotheses)

If you are conducting a study and want to use a hypothesis test to support your claim, the claim must be worded so that it becomes the alternative hypothesis.

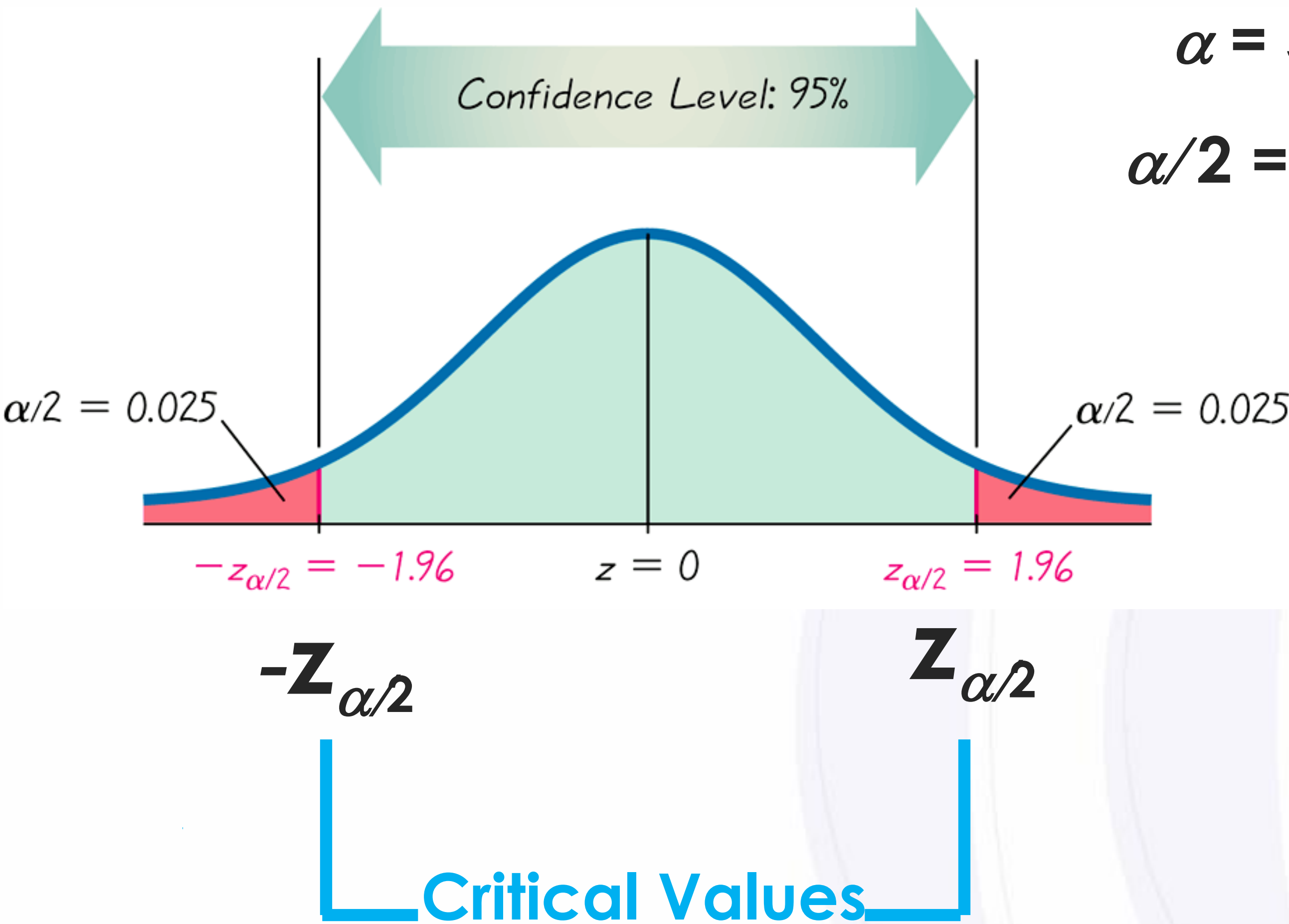
Significance Level

The significance level (denoted by α) is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true.
Common choices for α are 0.05, 0.01, and 0.10.

Significance Level

$\alpha = 5\%$

$\alpha/2 = 2.5\% = .025$



P-Value

The *P*-value (or *p*-value or probability value) is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the null hypothesis is true.

Types of Hypothesis Tests: Two-tailed, Left-tailed, Right-tailed

The tails in a distribution are the extreme regions bounded by critical values.

Determinations of *P*-values and critical values are affected by whether a critical region is in two tails, the left tail, or the right tail.

Non directional hypothesis test is two-tailed. Directional Hypothesis is one-tailed.

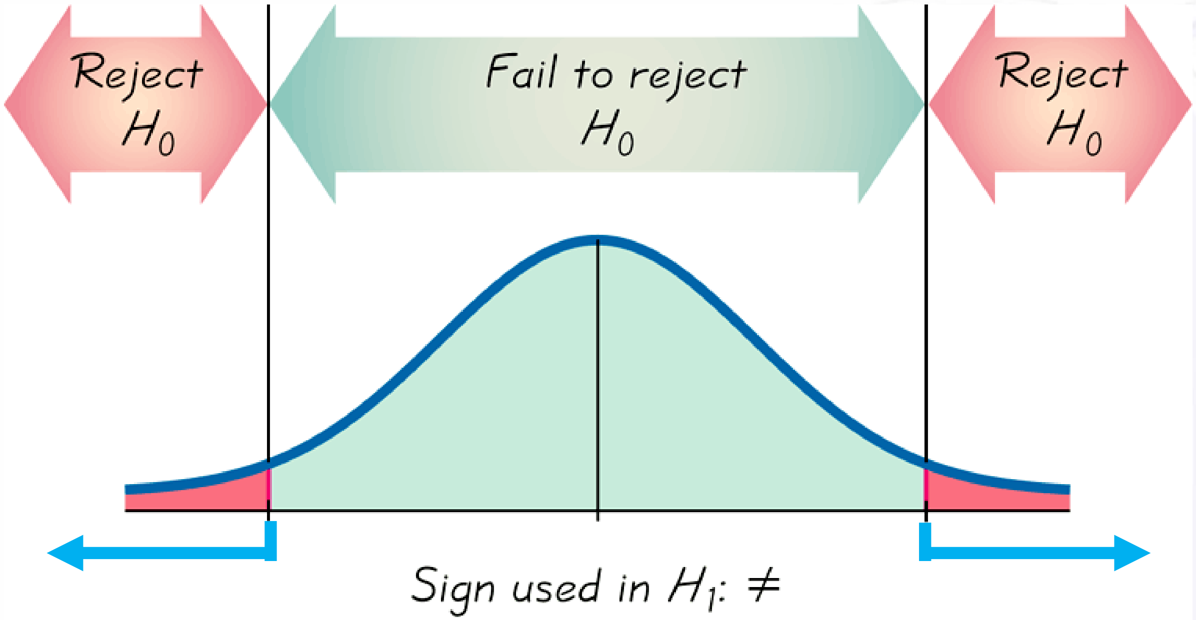
Two-tailed Test

$H_0: =$

α **is divided equally between the two tails of the critical region**

$H_1: \neq$

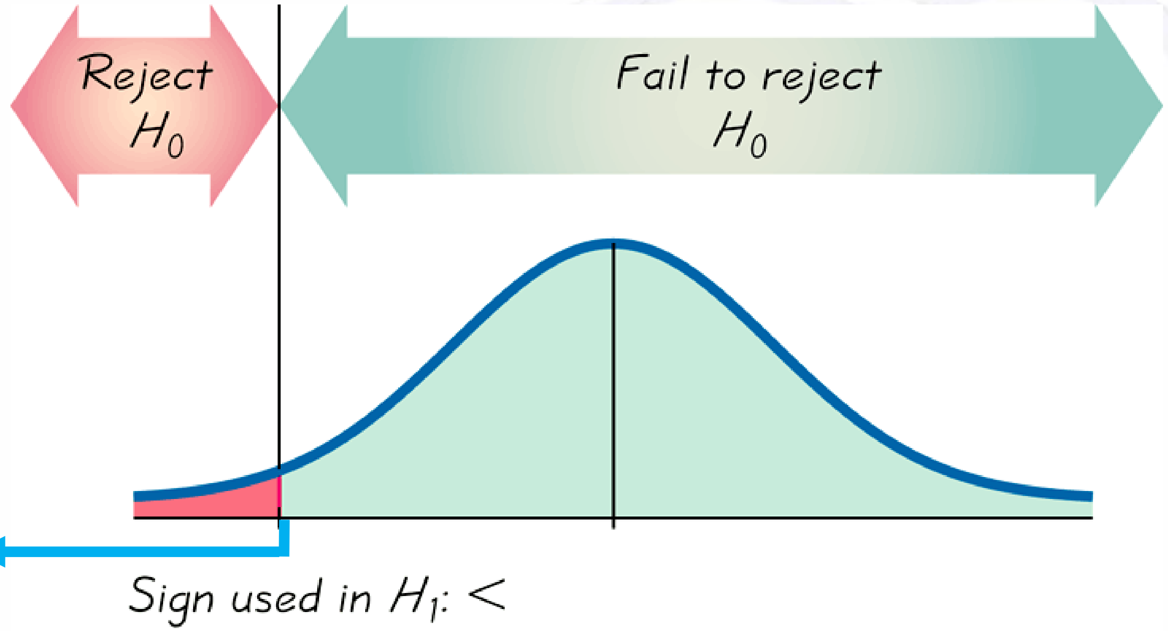
Means less than or greater than



Left-tailed Test

H_0: =

H_1: < **Points Left**



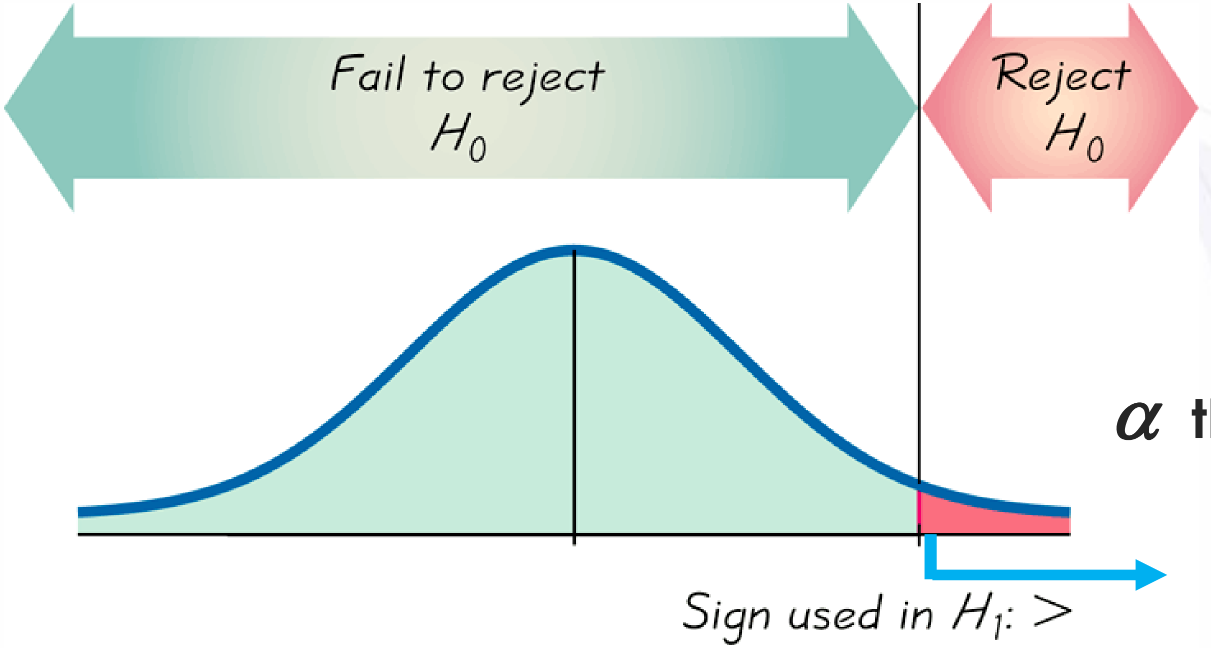
α **the left tail**

Right-tailed Test

H_0: =

H_1: >

Points Right



α the right tail

P-Value

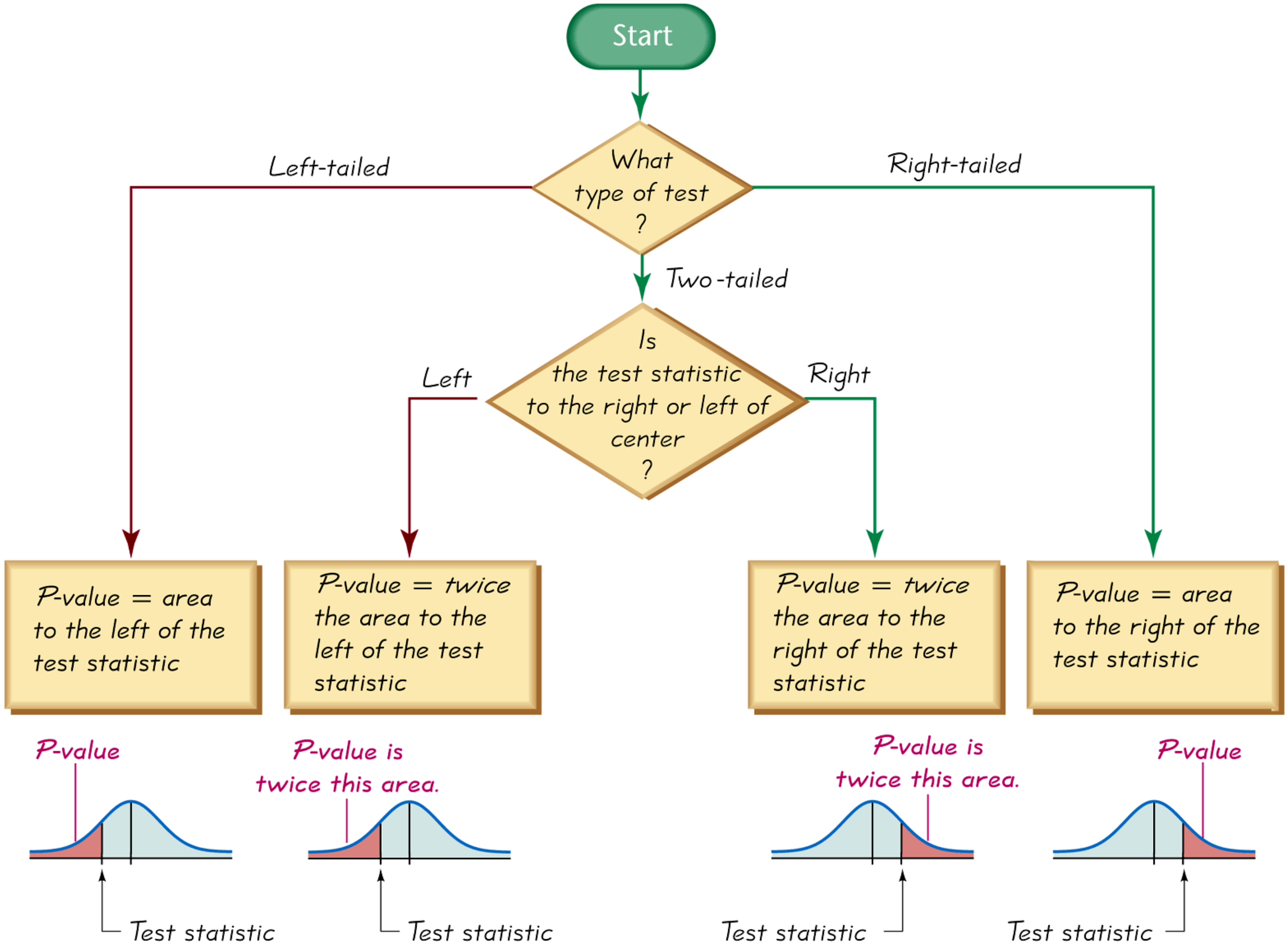
Critical region in the left tail: **P-value = area to the left of the test statistic**

Critical region in the right tail: **P-value = area to the right of the test statistic**

Critical region in two tails: **P-value = twice the area in the tail beyond the test statistic**

The null hypothesis is rejected if the P-value is very small, such as 0.05 or less.

Procedure for Finding *P*-Values



- We always test the null hypothesis. The initial conclusion will always be one of the following:
- 1. Reject the null hypothesis.
- 2. Fail to reject the null hypothesis.

Decision Criterion

P-value method:
Using the significance level α:

If *P*-value $\leq \alpha$, reject H_0 (Accept Ha).

If *P*-value $> \alpha$, fail to reject H_0

(Accept H0).

Inferences About Two Means

Two samples are independent if the sample values selected from one population are not related to or somehow paired or matched with the sample values from the other population.

Two samples are dependent if the sample values are *paired*. (That is, each pair of sample values consists of two measurements from the same subject (such as before/after data).

Inferences About Two Means: Independent Samples

This section presents methods for using sample data from two independent samples to test hypotheses made about two population means.

Notation

μ_1 = population mean

σ_1 = population standard deviation

n_1 = size of the first sample

$\bar{X}_1$ and $\bar{X}_2$ = sample mean

s_1 = sample standard deviation

Corresponding notations for μ_2, σ_2, s_2, and n_2 apply to population 2.

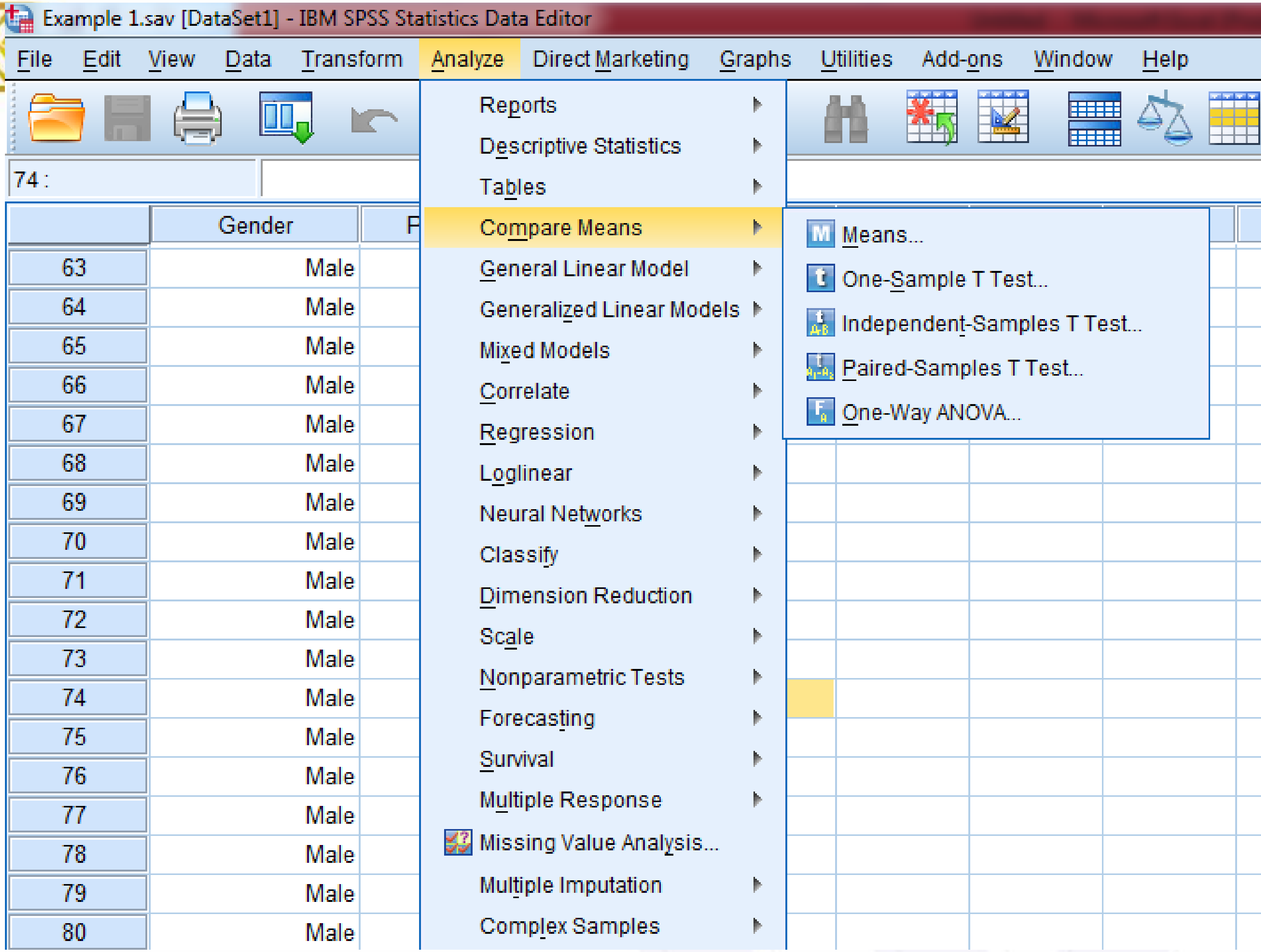
Requirements

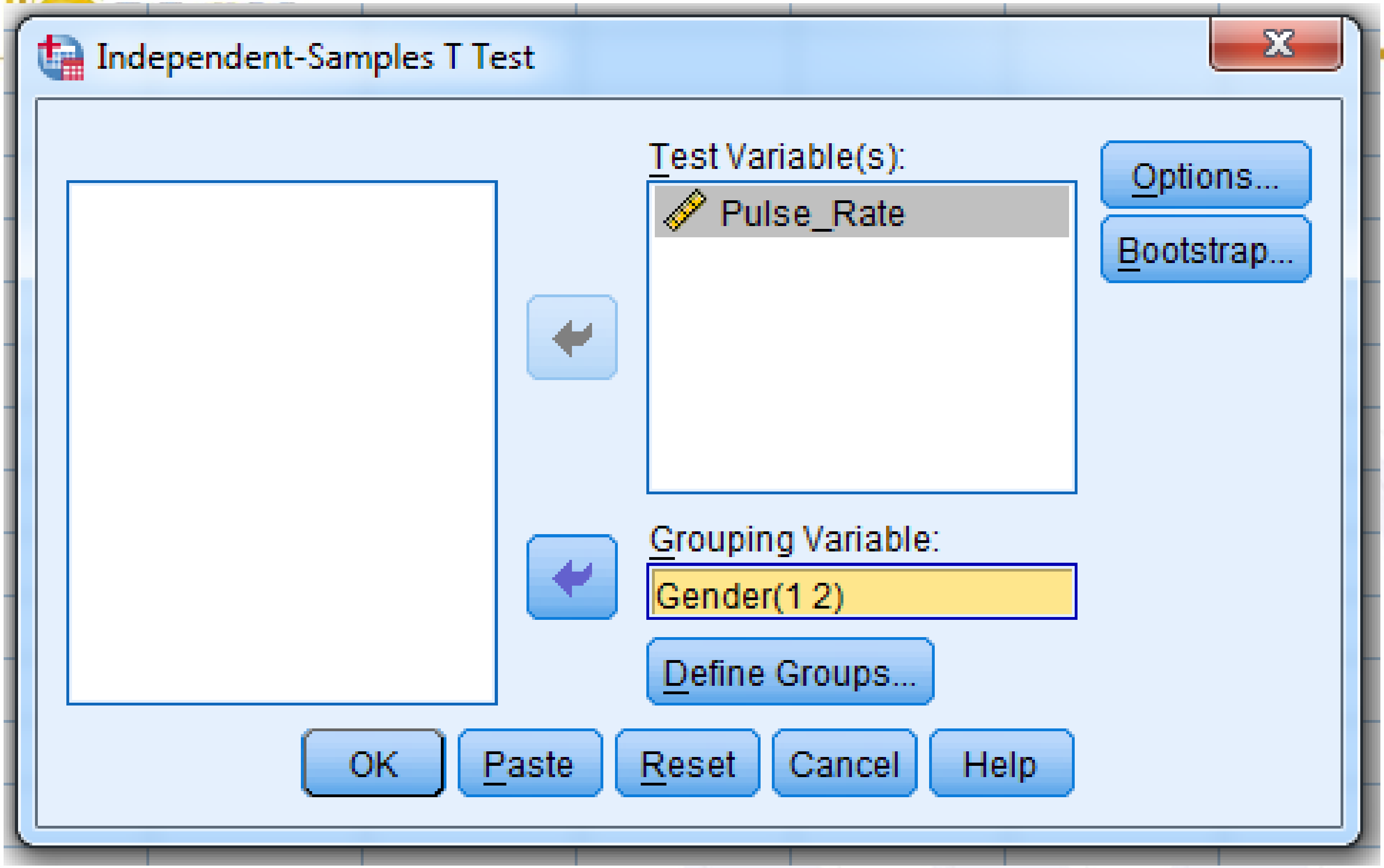
1. σ_1 an σ_2 are unknown and no assumption is made about the equality of σ_1 and σ_2 .

2. The two samples are independent.

3. Both samples are simple random samples.

satisfied: 4. Either or both of these conditions are
large (with $n_1 > 30$ The two sample sizes are both
samples come from and $n_2 > 30$) or both
normal distributions. populations having


Example 1.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
74 :
Gender
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
63 Male
64 Male
65 Male
66 Male
67 Male
68 Male
69 Male
70 Male
71 Male
72 Male
73 Male
74 Male
75 Male
76 Male
77 Male
78 Male
79 Male
80 Male


Independent-Samples T Test
Test Variable(s):
Pulse_Rate
Options...
Bootstrap...
Grouping Variable:
Gender(1 2)
Define Groups...
OK
Paste
Reset
Cancel
Help

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Pulse_Rate	Female	40	76.30	12.499	1.976
	Male	40	69.40	11.297	1.786

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									Interval of the	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pulse_Rate	Equal variances assumed	.135	.714	2.590	78	.011	6.900	2.664	1.597	12.203
	Equal variances not assumed			2.590	77.217	.011	6.900	2.664	1.596	12.204

Inferences from Matched Pairs

In this section we develop methods for testing hypotheses and constructing confidence intervals involving the mean of the differences of the values from two dependent populations.

With dependent samples, there is some relationship whereby each value in one sample is paired with a corresponding value in the other sample.

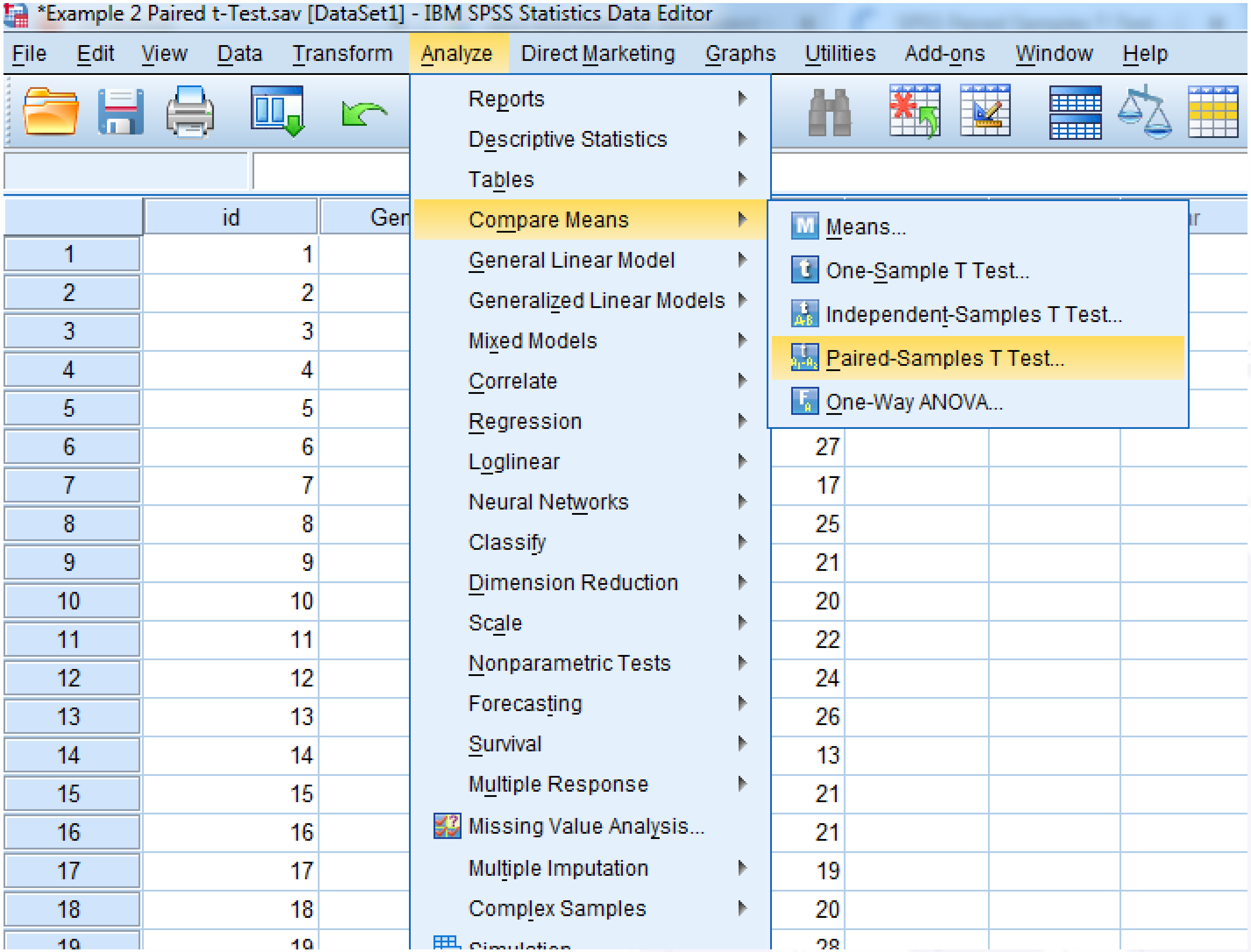
Requirements

1. The sample data are dependent.

2. The samples are simple random samples.

3. Either or both of these conditions is satisfied: The number of pairs of sample data is large ($n > 30$) or the pairs of values have differences that are from a population having a distribution that is approximately normal.

Paired t-Test Example – Example 2



Paired t-Test Example – Example 2

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	Interval of the				
				Lower	Upper			
Exam Version 1 - Correct Answers - Exam Version 2 - Correct Answers	-.579	2.524	.579	-1.795	.637	-1.000	18	.331

- t-Test for Activity File based on Gender

Analysis of Variance
One-Way ANOVA

Definition

One-way analysis of variance (ANOVA) is a method of testing the equality of three or more population means by analyzing sample variances. One-way analysis of variance is used with data categorized with *one* treatment (or factor), which is a characteristic that allows us to distinguish the different populations from one another.

Procedure for testing

H0: μ1 = μ2 = μ3 = . . .

If the *P*-value $\leq \alpha$, reject the null hypothesis of equal means and conclude that at least one of the population means is different from the others.

If the *P*-value $> \alpha$, fail to reject the null hypothesis of equal means.

Requirements

1. The populations have approximately normal distributions.

2. The populations have the same variance s^2 (or standard deviation *s*).

3. The samples are simple random samples.

4. The samples are independent of each other.

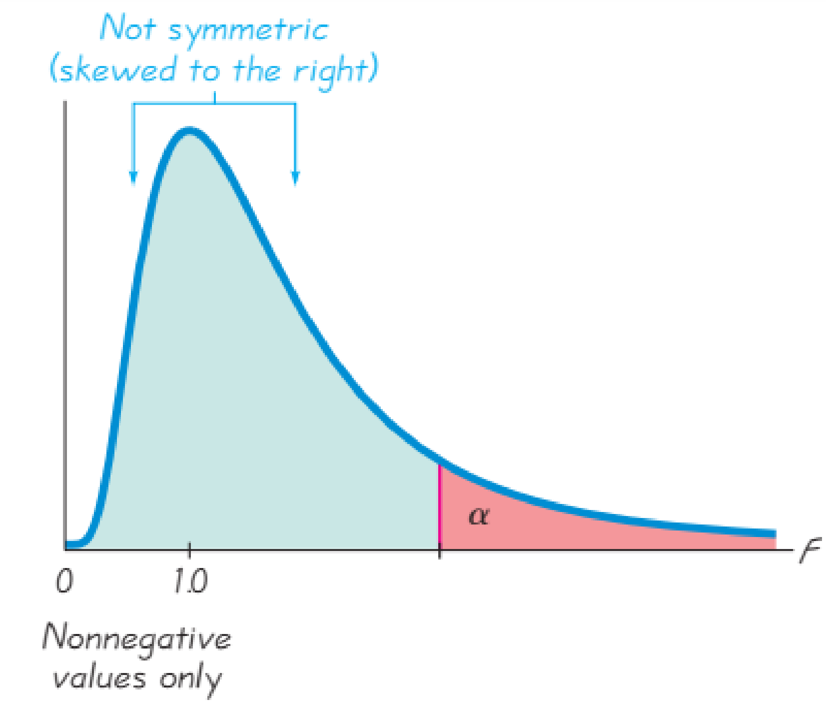
5. The different samples are from populations that are categorized in only one way.

Test Statistic for One-Way ANOVA

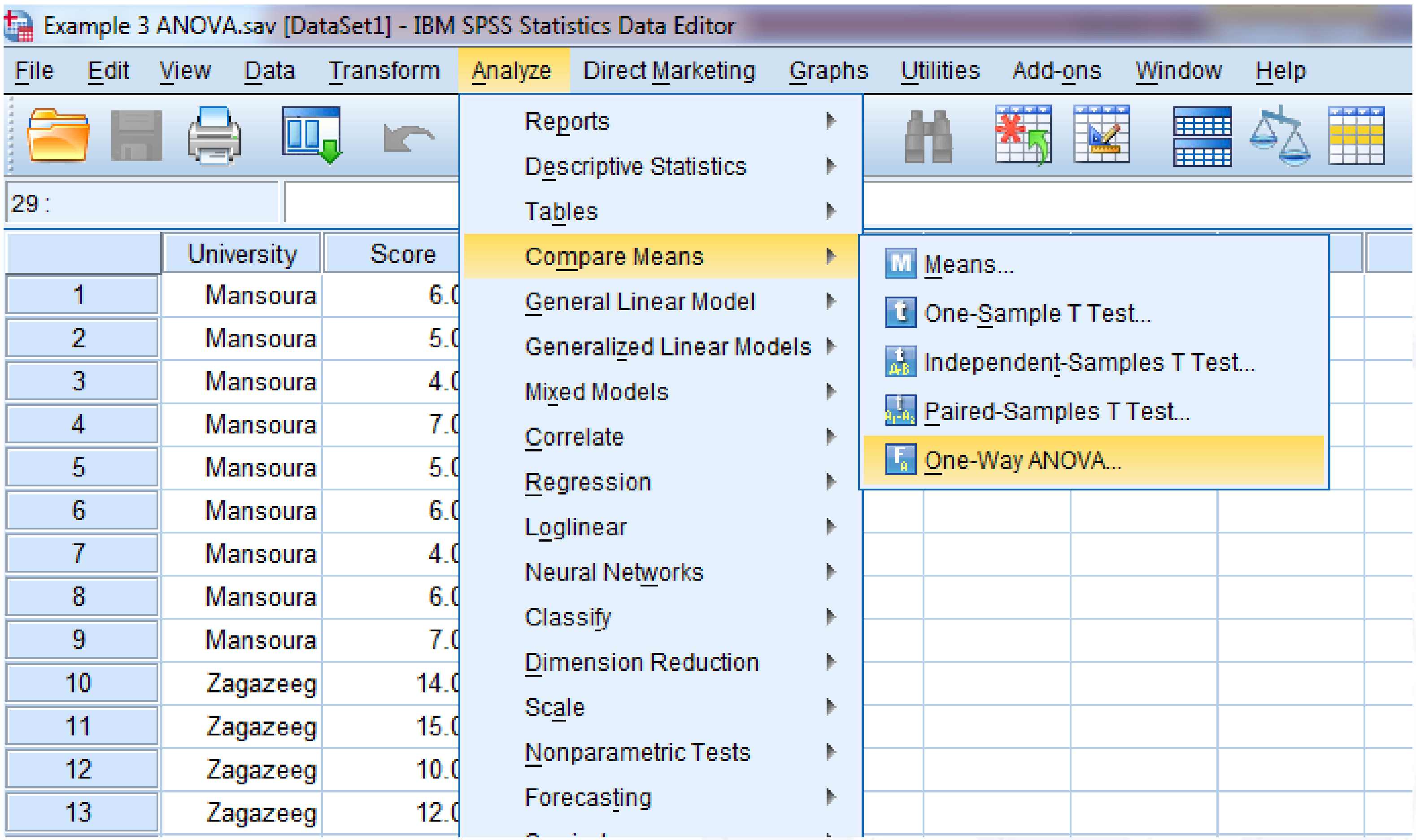
$$F = \frac{\text{variance between samples}}{\text{variance within samples}}$$

***F* Distribution**

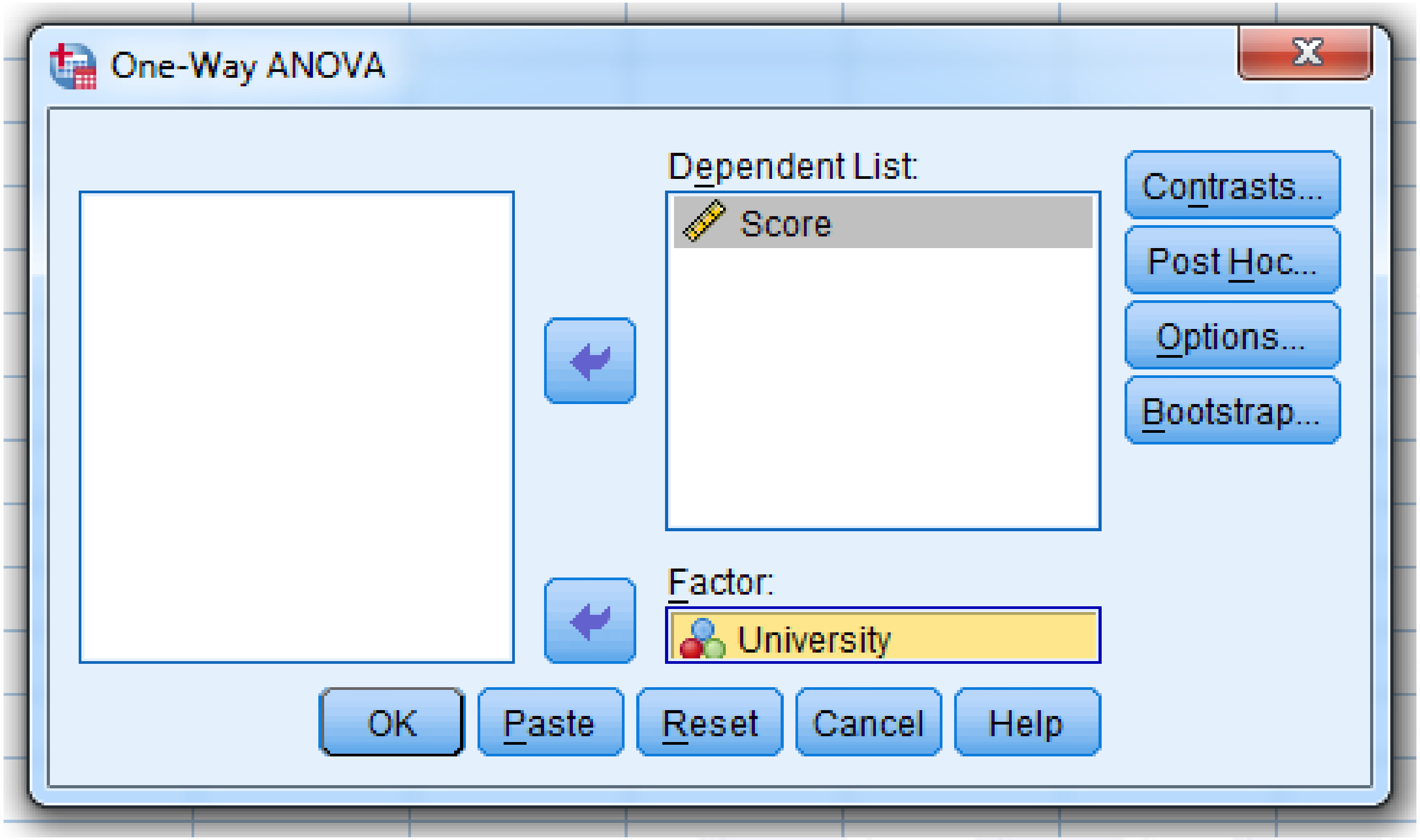
There is a different *F* distribution for each different pair of degrees of freedom for numerator and denominator.



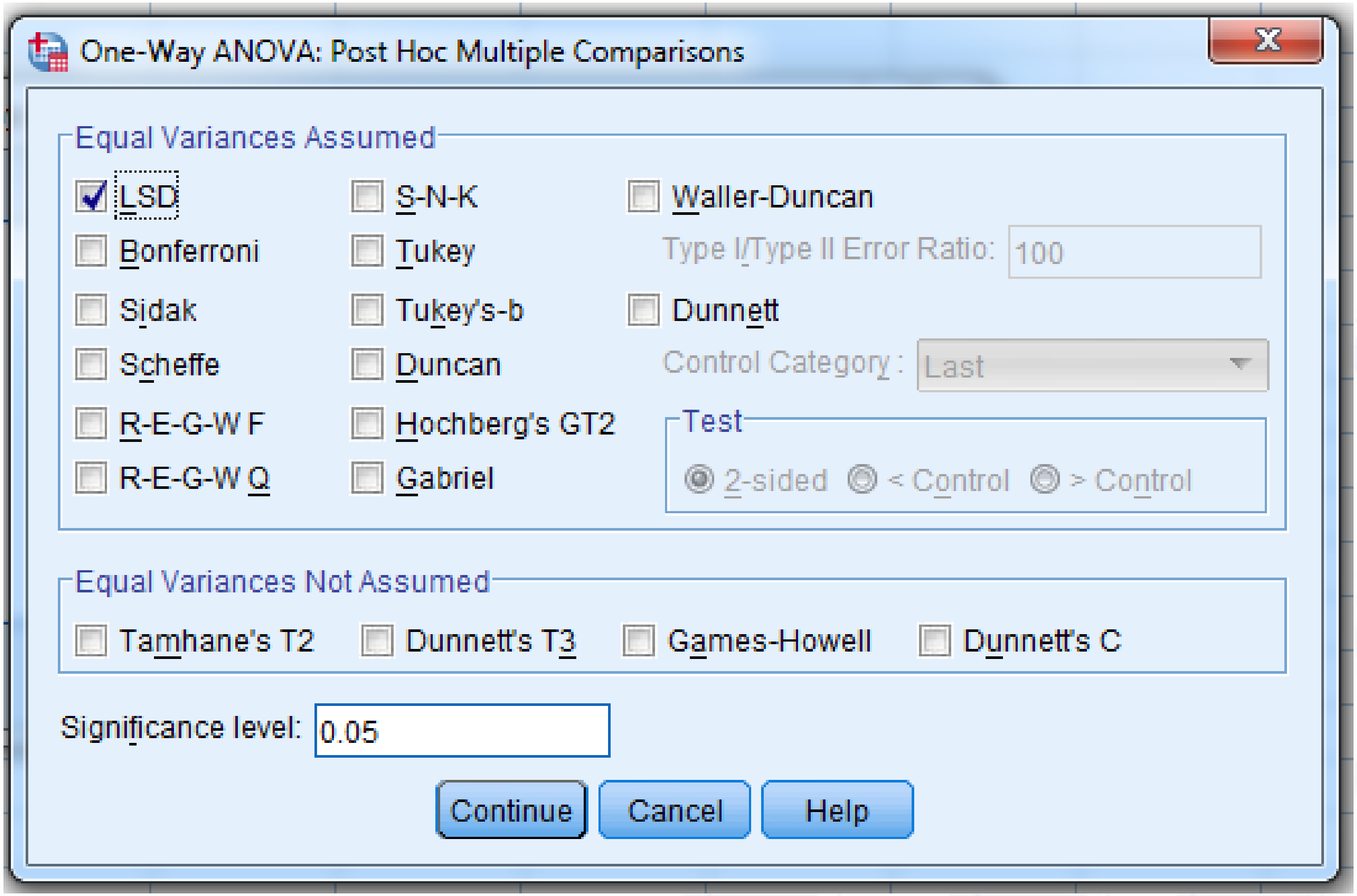
Example 3 - One-Way ANOVA



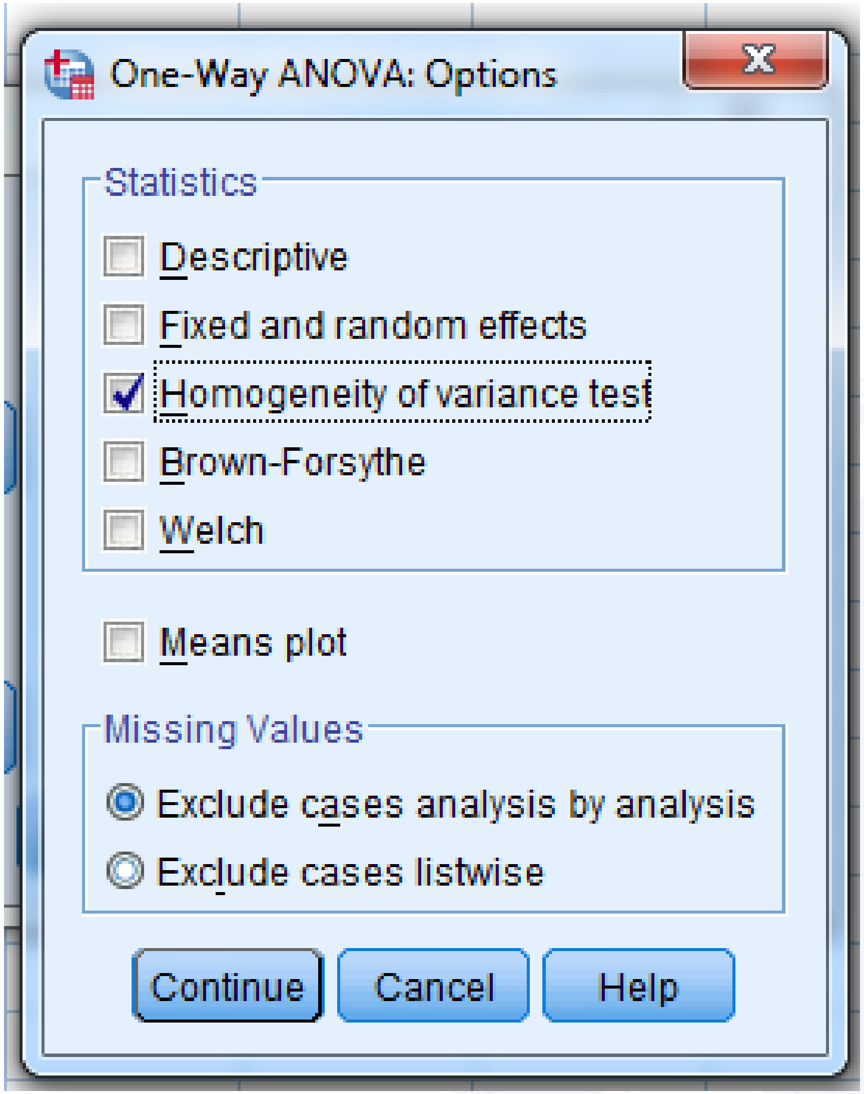
Example 3- One-Way ANOVA



Example 3 - One-Way ANOVA



Example 3 - One-Way ANOVA



One-Way ANOVA

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	261.113	2	130.557	12.477	.000
Within Groups	261.601	25	10.464		
Total	522.714	27			

Multiple Comparisons

Dependent Variable: Score

LSD

(I) University	(J) University	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Mansoura	Zagazeeg	-6.47222*	1.57184	.000	-9.7095	-3.2350
	Menoufia	-6.58586*	1.45394	.000	-9.5803	-3.5914
Zagazeeg	Mansoura	6.47222*	1.57184	.000	3.2350	9.7095
	Menoufia	-.11364	1.50309	.940	-3.2093	2.9820
Menoufia	Mansoura	6.58586*	1.45394	.000	3.5914	9.5803
	Zagazeeg	.11364	1.50309	.940	-2.9820	3.2093

*. The mean difference is significant at the 0.05 level.

Activity 5

- One Way ANOVA for Activity File based on Educaiton

Quartiles

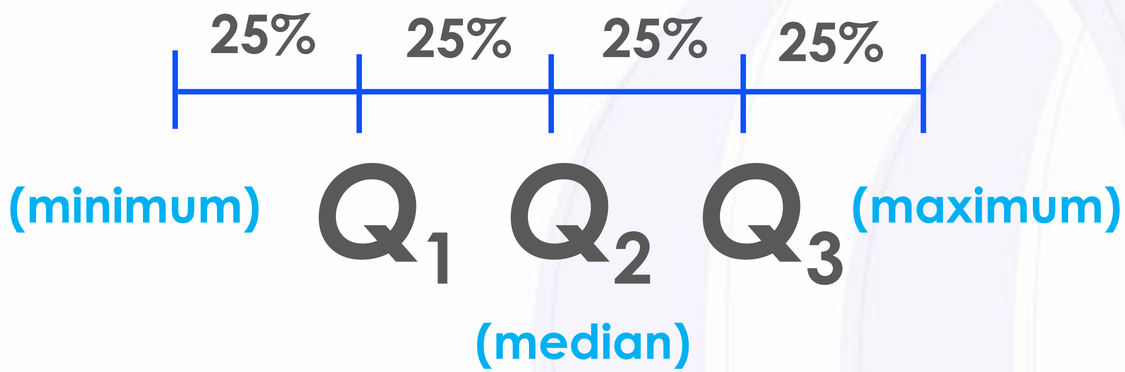
Are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of data into four groups with 25% of the values in each group.

- Q_1 **(First Quartile)** separates the bottom 25% of sorted values from the top 75%.
- Q_2 **(Second Quartile)** same as the median; separates the bottom 50% of sorted values from the top 50%.
- Q_3 **(Third Quartile)** separates the bottom 75% of sorted values from the top 25%.

Quartiles

Q_1, Q_2, Q_3

divide **data points** into four equal parts



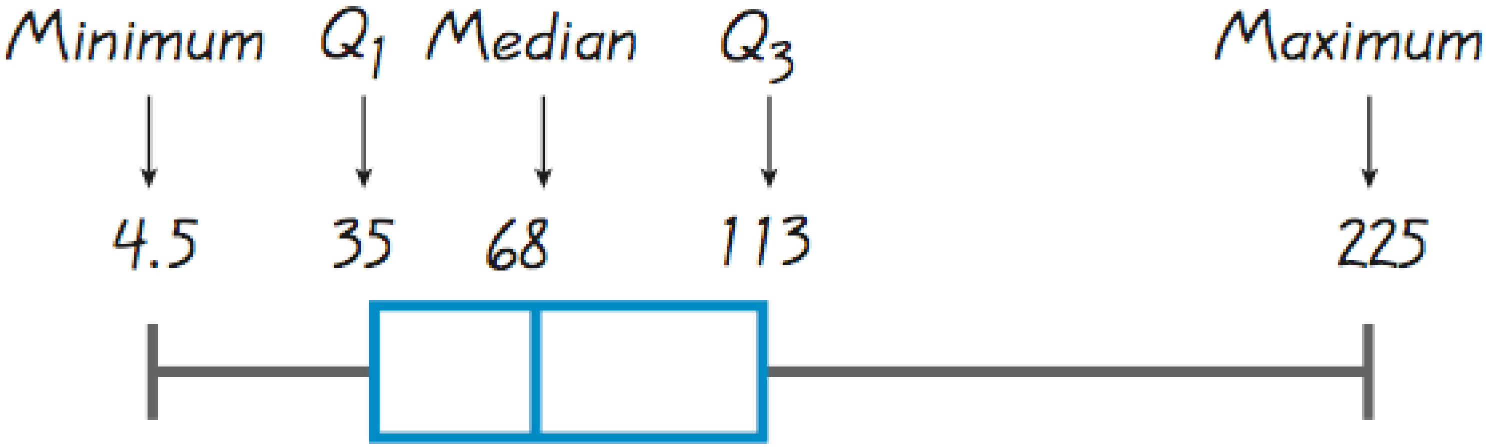
5-Number Summary

- For a set of data, the 5-number summary consists of:
- The minimum value;
- The first quartile Q_1;
- The second quartile Q_2;
- The third quartile, Q_3;
- The maximum value.

Boxplot

- A boxplot (or box-and-whisker-diagram) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, Q_1; the median or Q2; and the third quartile, Q_3.

Boxplots



Boxplot of Data Points

Outliers

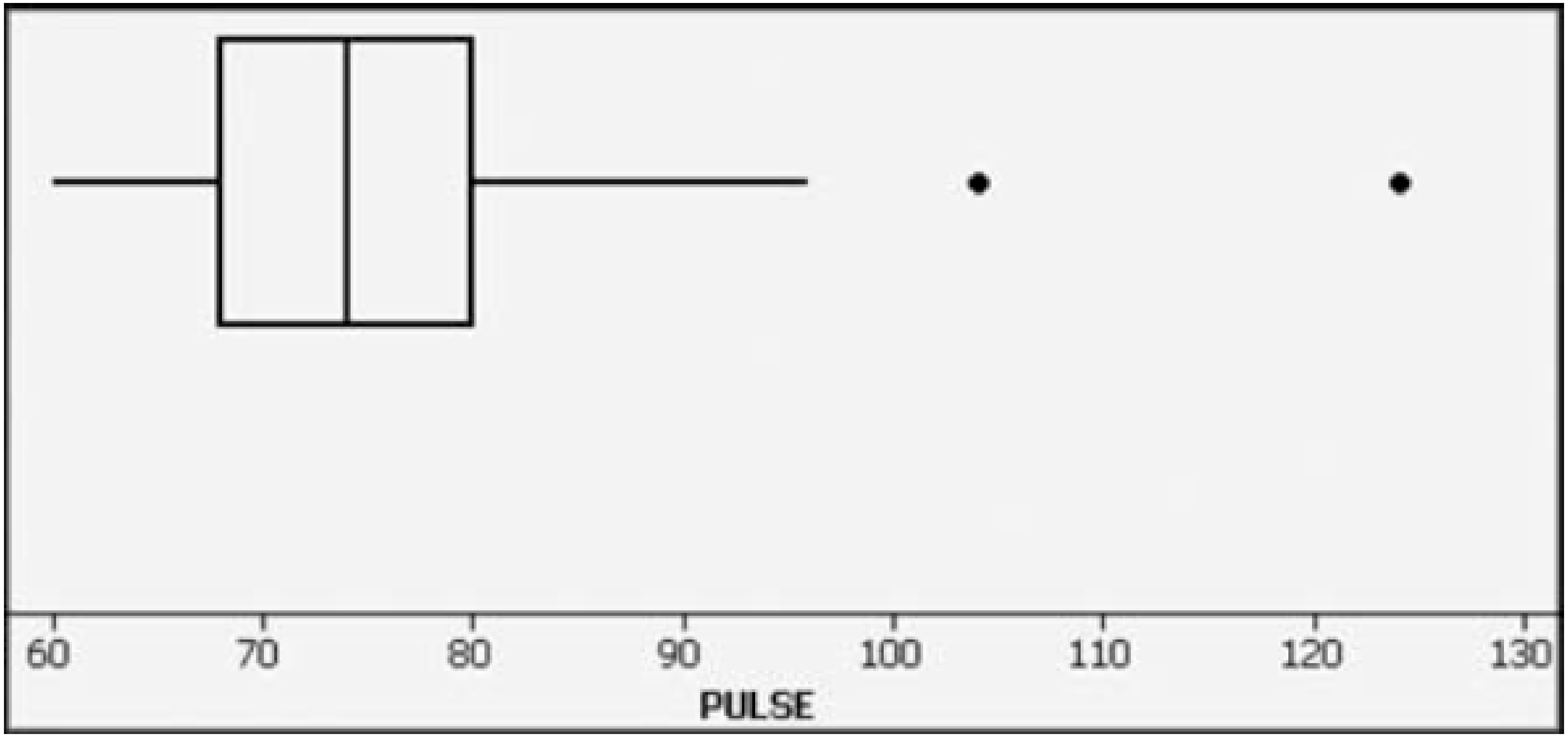
- An **outlier** is a value that lies very far away from the the vast majority of the other values in a data set.

- An outlier can have a dramatic effect on the mean.

- An outlier can have a dramatic effect on the standard deviation.

- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured.

Modified Boxplots - Example



Pulse rates of females listed in Fig XXX

Outliers for Modified Boxplots

we can consider outliers to be data values meeting specific criteria.

- Interquartile Range (or IQR): Q3 – Q1

In modified boxplots, a data value is an outlier if it is . . .

above Q_3 by an amount greater than $1.5 \times$ IQR

or below Q_1 by an amount greater than $1.5 \times$ IQR

- Demo on How to check for the Outliers using SPSS on Activity 2 file Female Pulse Rate

Session 4

- Correlation
- Regression
- Multiple Regression
- Heteroscedasticity, Autocorrelation, Multicoliniarity
- Logit and Probit Regression
- Non Parametric Tests

Correlation

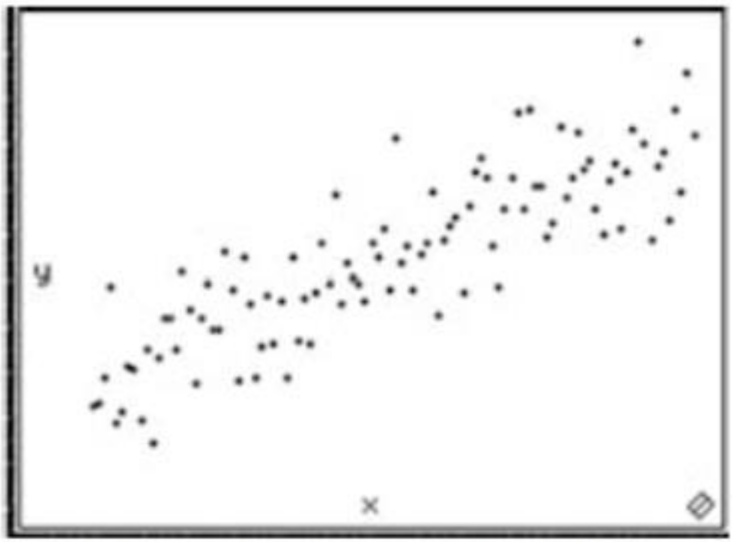
Correlation Analysis

A correlation exists between two variables when the values of one are somehow associated with the values of the other in some way.

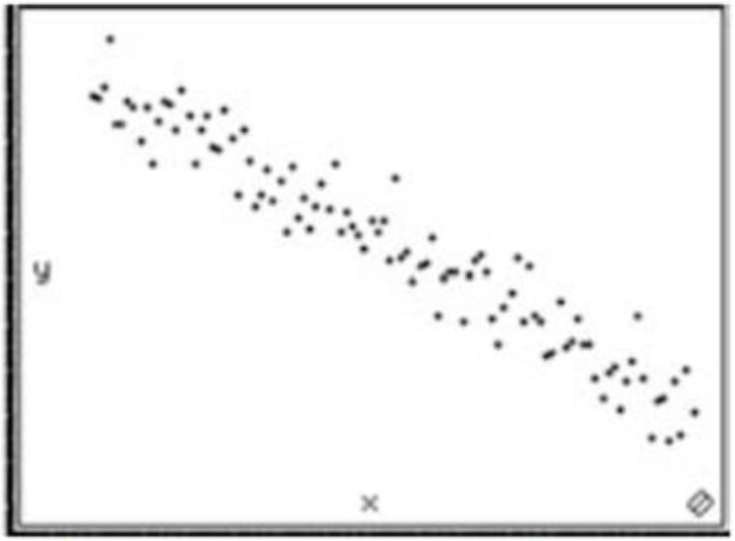
The linear correlation coefficient r measures the strength of the linear relationship between the paired quantitative x- and y-values in a sample.

Exploring the Data

We can often see a relationship between two variables by constructing a scatterplot.

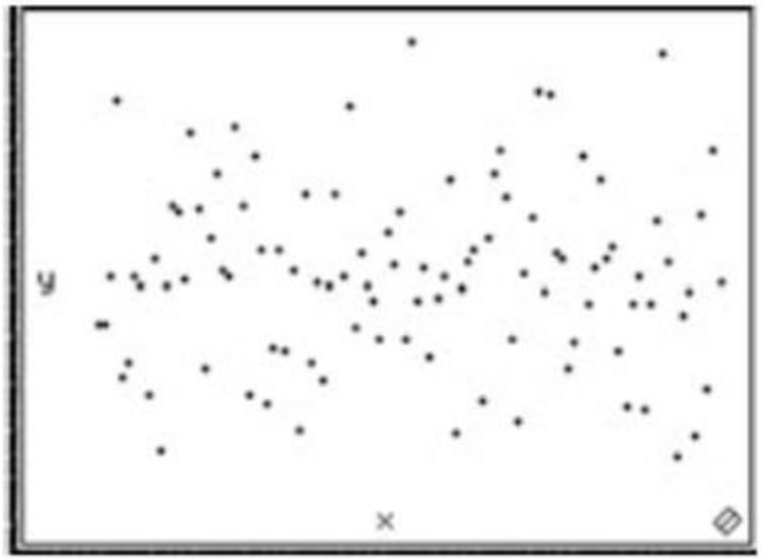


(a) Positive correlation:
$r = 0.851$



(b) Negative correlation:
$r = -0.965$

Scatterplots of Paired Data



(c) No correlation: $r = 0$

Figure 10-2

Requirements

1. The sample of paired (*x, y*) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating *r* with and without the outliers included.

Interpreting *r*

Using Software: **If the computed *P*-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.**

Properties of the Linear Correlation Coefficient r

1. $-1 \le r \le 1$
2. if all values of either variable are converted to a different scale, the value of r does not change.
3. The value of r is not affected by the choice of x and y. Interchange all x- and y-values and the value of r will not change.
4. r measures strength of a linear relationship.
5. r is very sensitive to outliers, they can dramatically affect its value.

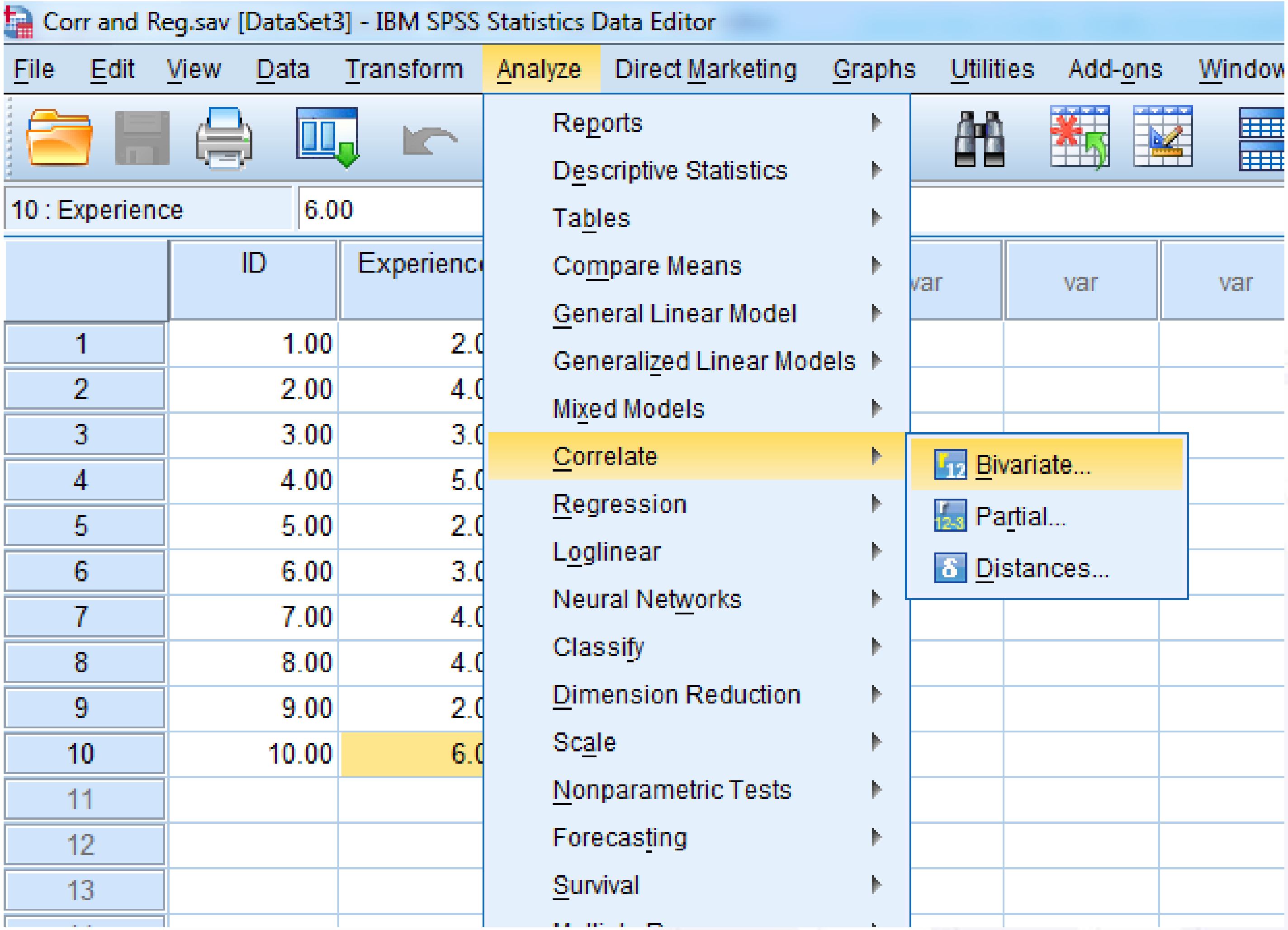
Interpreting r: Explained Variation

The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y.

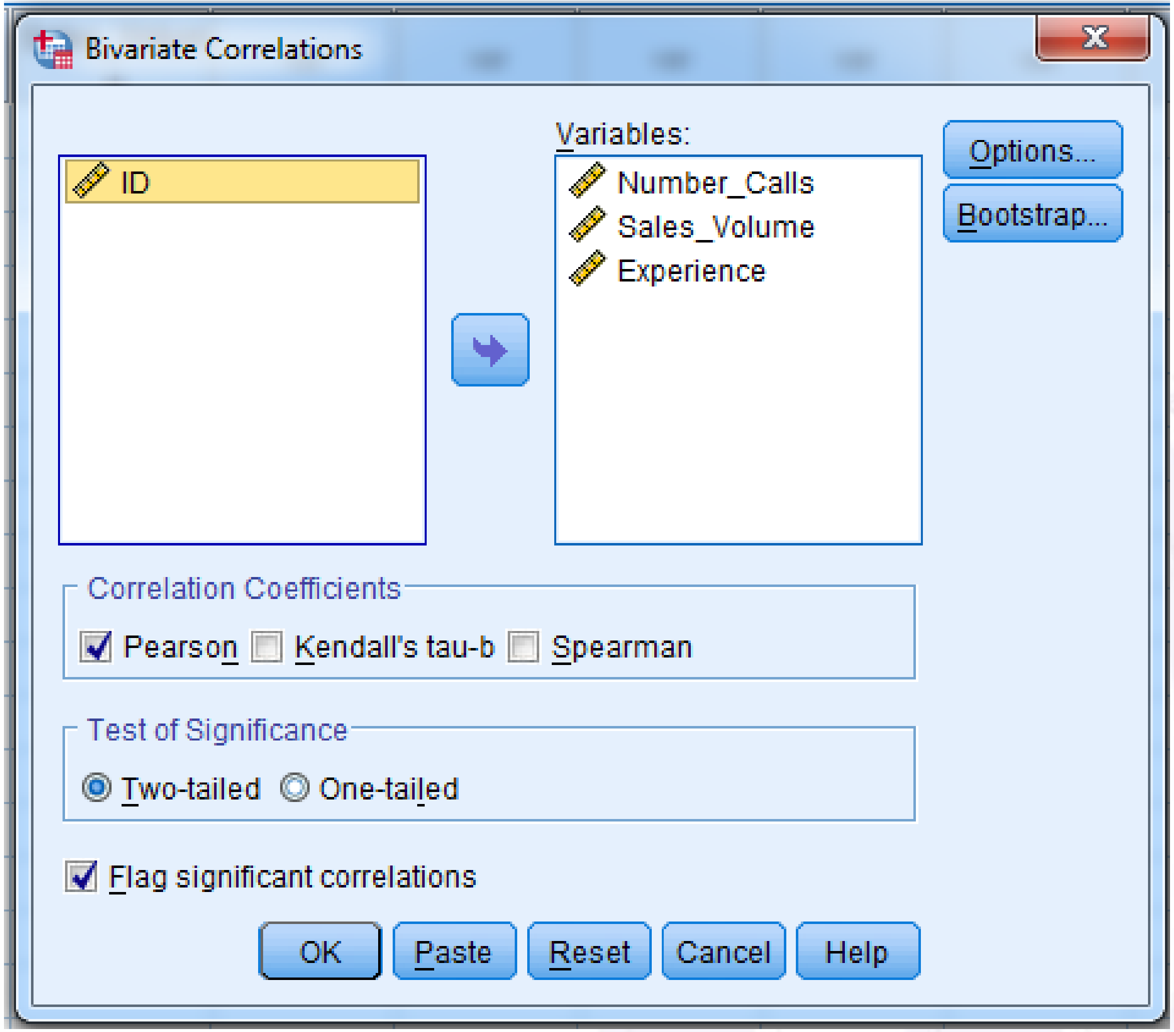
Formal Hypothesis Test

We wish to determine whether there is a significant linear correlation between two variables.

Example 4 - Correlation



Example 4 - Correlation



Example 4 - Correlation

Correlations

		Number_Calls	Experience	Sales_Volume
Number_Calls	Pearson Correlation	1	.625	.759*
	Sig. (2-tailed)		.053	.011
	N	10	10	10
Experience	Pearson Correlation	.625	1	.944**
	Sig. (2-tailed)	.053		.000
	N	10	10	10
Sales_Volume	Pearson Correlation	.759*	.944**	1
	Sig. (2-tailed)	.011	.000	
	N	10	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

- Correlation of the 4 independent file and the dependent Variable of Activity 1 File

Basic Concepts of Regression

Regression Analysis

The regression equation expresses a relationship between x (called the explanatory variable, predictor variable or independent variable), and y (called the response variable or dependent variable).

The typical equation of a straight line
$\hat{y} = mx + b$ is expressed in the form
$y = b_0 + b_1x$, where b_0 is the y-intercept and b_1 is the slope.

Definitions

- **Regression Equation**

Given a collection of paired data, the regression equation

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the relationship between the two variables.

- **Regression Line**

The graph of the regression equation is called the regression line (or line of best fit, or least squares line).

Notation for Regression Equation

	Population Parameter	Sample Statistic
y-intercept of regression equation	β_0	b_0
Slope of regression equation	β_1	b_1
Equation of the regression line	$y = \beta_0 + \beta_1 x$	$y = b_0 + b_1 x$

Requirements

1. The sample of paired (x, y) data is a random sample of quantitative data.

2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.

3. Any outliers must be removed if they are known to be errors. Consider the effects of any outliers that are not known errors.

Using the Regression Equation for Predictions

1. **Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.**

2. **Use the regression equation for predictions only if the linear correlation coefficient r indicates that there is a linear correlation between the two variables.**

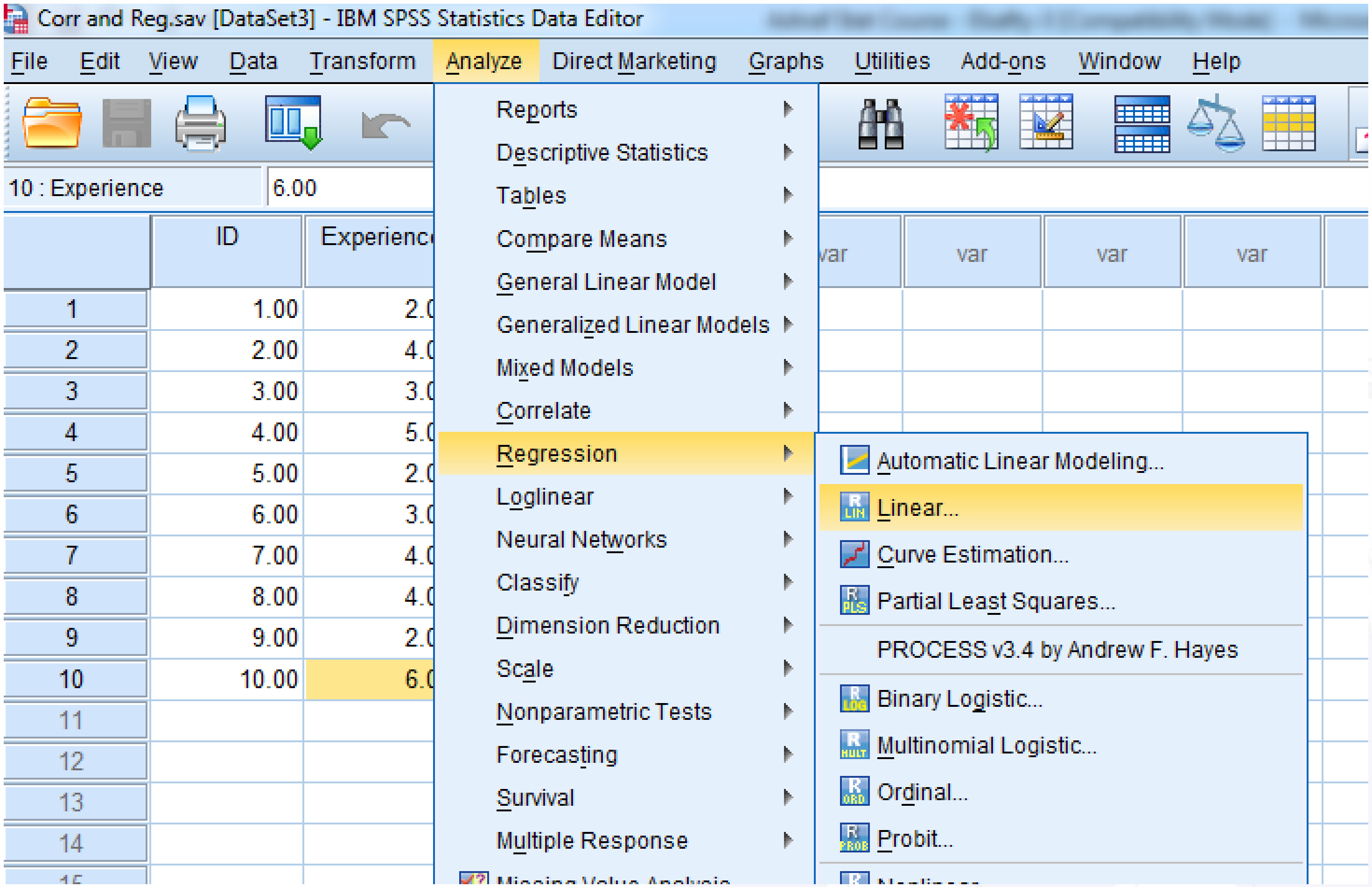
Complete Regression Analysis

1. Construct a scatterplot and verify that the pattern of the points is approximately a straight-line pattern without outliers. (If there are outliers, consider their effects by comparing results that include the outliers to results that exclude the outliers.)
2. Construct a residual plot and verify that there is no pattern (other than a straight-line pattern) and also verify that the residual plot does not become thicker (or thinner).

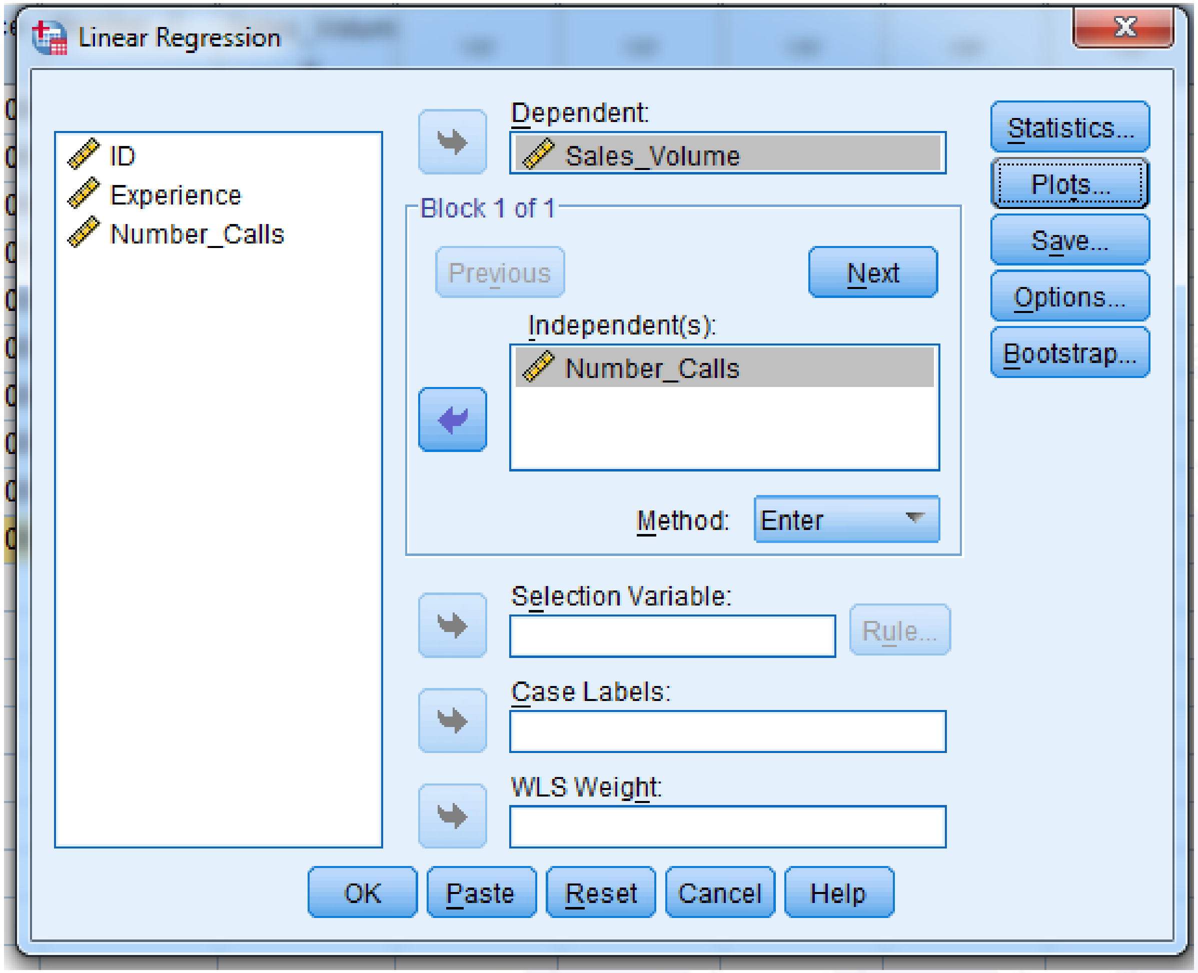
Complete Regression Analysis

3. Use a histogram and/or normal quantile plot to confirm that the values of the residuals have a distribution that is approximately normal.

Example 5- Simple Regression



Example 5- Simple Regression



Example 5- Simple Regression

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.759[a]	.576	.523	9.90082	2.159

a. Predictors: (Constant), Number_Calls

b. Dependent Variable: Sales_Volume

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1065.789	1	1065.789	10.872	.011[b]
	Residual	784.211	8	98.026		
	Total	1850.000	9			

a. Dependent Variable: Sales_Volume

b. Predictors: (Constant), Number_Calls

Example 5- Simple Regression

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	18.947	8.499		2.229	.056
	Number_Calls	1.184	.359	.759	3.297	.011

a. Dependent Variable: Sales_Volume

In working with two variables related by a regression equation, the marginal change in the variable (Sales Volume) is the amount (slope b_1) that it changes when the other variable (Number of Sales Call) changes by exactly one unit.

The slope b_1 in the regression equation represents the marginal change in y that occurs when x changes by one unit.

Beta Coefficient

In working with two variables related by a regression equation, the marginal change in a variable is the amount that it changes when the other variable changes by exactly one unit. The slope b_1 in the regression equation represents the marginal change in y that occurs when x changes by one unit.

Activity 7

- Regression Analysis for one independent variable and the Dependent Variable of Activity File1 the dependent Variable of Activity File 1

Logit and Probit Regression

- Logit and probit differ in how they define the function.
  - The logit model uses something called the cumulative distribution function of the logistic distribution.
  - The probit model uses something called the cumulative distribution function of the standard normal distribution to define.
- Both functions will take any number and rescale it to fall between 0 and 1.
- Any function that would return a value between zero and one would do the task.

Logit and Probit Regression

- The logistic turn out to be convenient mathematically and are programmed into just about any general purpose statistical package.
- Is logit better than probit, or vice versa?
- Both methods will yield similar (though not identical) inferences:
  - Logit – also known as logistic regression – is more popular in social sciences.
  - Probit models are used in some contexts by economists and political scientists

Nonparametric tests

- There tests are called **distribution-free tests** because they are based on fewer assumptions (e.g., they do not assume that the outcome is approximately normally distributed).
- Parametric tests involve specific probability distributions (e.g., the normal distribution) and the tests involve estimation of the key parameters of that distribution (e.g., the mean or difference in means) from the sample data.

Nonparametric tests

- Nonparametric tests are generally less powerful than their parametric counterparts
  - **Mann-Whitney test**. Use this test to compare differences between two independent groups when dependent variables are either ordinal or continuous.
  - **Kruskal-Wallis test**. Use this test instead of a one-way ANOVA to find out if two or more medians are different. Ranks of the data points are used for the calculations, rather than the data points themselves.
  - **Spearman Rank Correlation.** Use when you want to find a correlation between two sets of data.