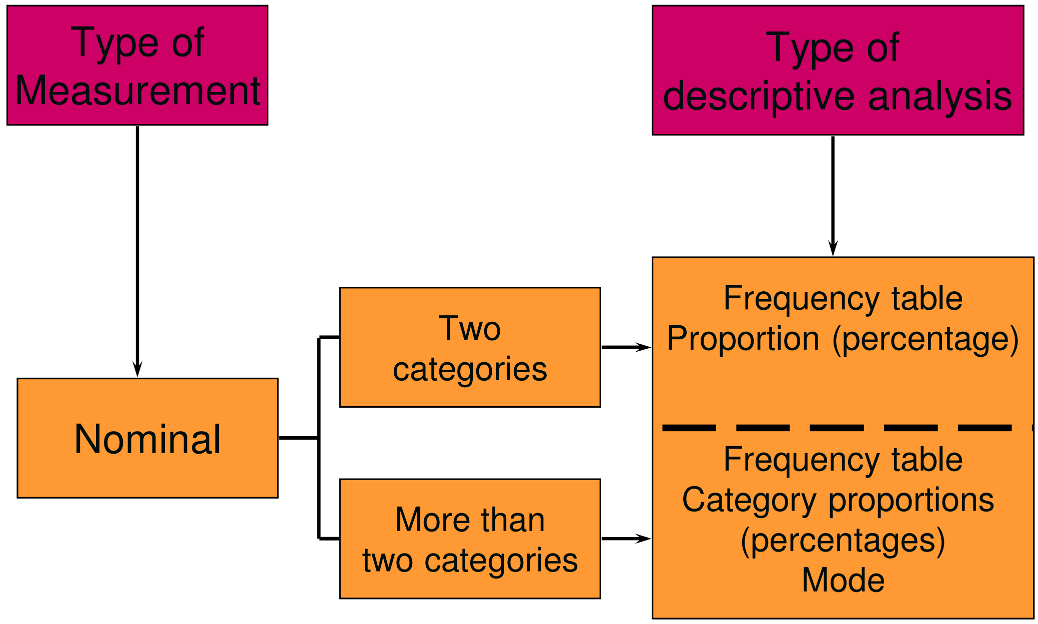
Research Methods

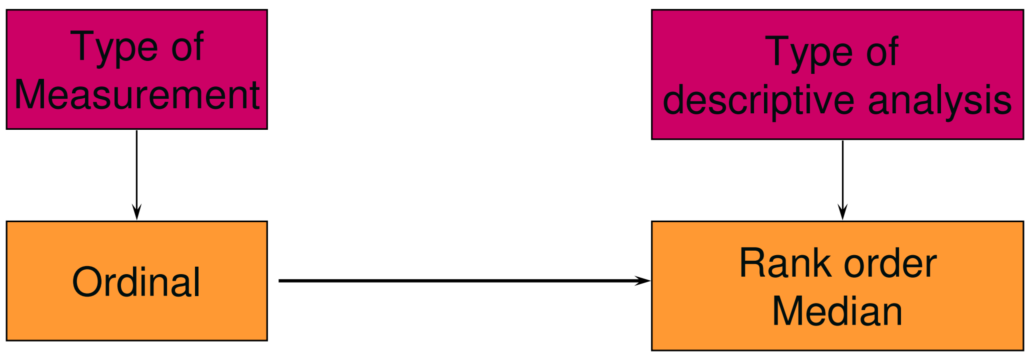
William G. Zikmund

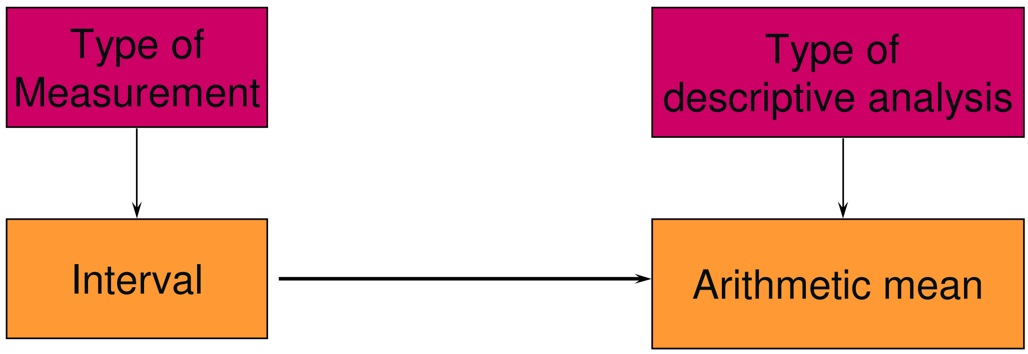
Basic Data Analysis: Descriptive Statistics

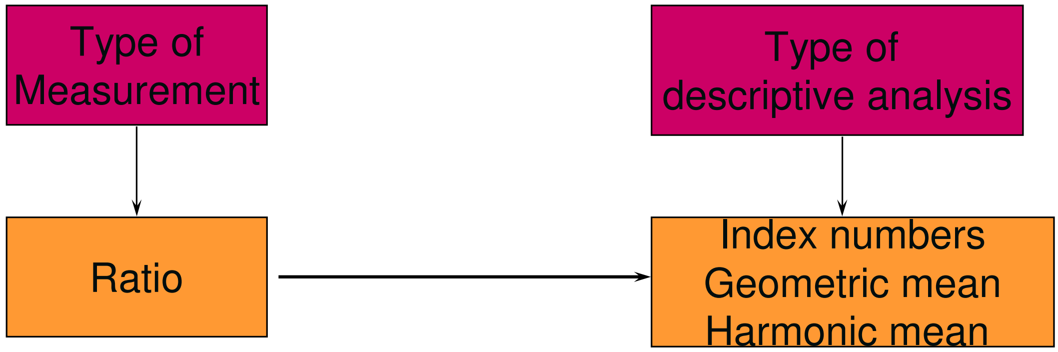
Descriptive Analysis

- The transformation of raw data into a form that will make them easy to understand and interpret; rearranging, ordering, and manipulating data to generate descriptive information


Type of
Measurement
Type of
descriptive analysis
Nominal
Two
categories
More than
two categories
Frequency table
Proportion (percentage)
Frequency table
Category proportions
(percentages)
Mode


Type of
Measurement
Type of
descriptive analysis
Ordinal
Rank order
Median


Type of Measurement
Interval
Type of descriptive analysis
Arithmetic mean


Type of Measurement
Ratio
Type of descriptive analysis
Index numbers
Geometric mean
Harmonic mean

Tabulation

- Tabulation - Orderly arrangement of data in a table or other summary format
- Frequency table
- Percentages

Frequency Table

- The arrangement of statistical data in a row-and-column format that exhibits the count of responses or observations for each category assigned to a variable

Central Tendency

Type of Scale	Measure of Central Tendency	Measure of Dispersion
Nominal	Mode	None
Ordinal	Median	Percentile
Interval or ratio	Mean	Standard
deviation		

Cross-Tabulation

- A technique for organizing data by groups, categories, or classes, thus facilitating comparisons; a joint frequency distribution of observations on two or more sets of variables
- Contingency table- The results of a cross-tabulation of two variables, such as survey questions

Cross-Tabulation

- Analyze data by groups or categories
- Compare differences
- Contingency table
- Percentage cross-tabulations

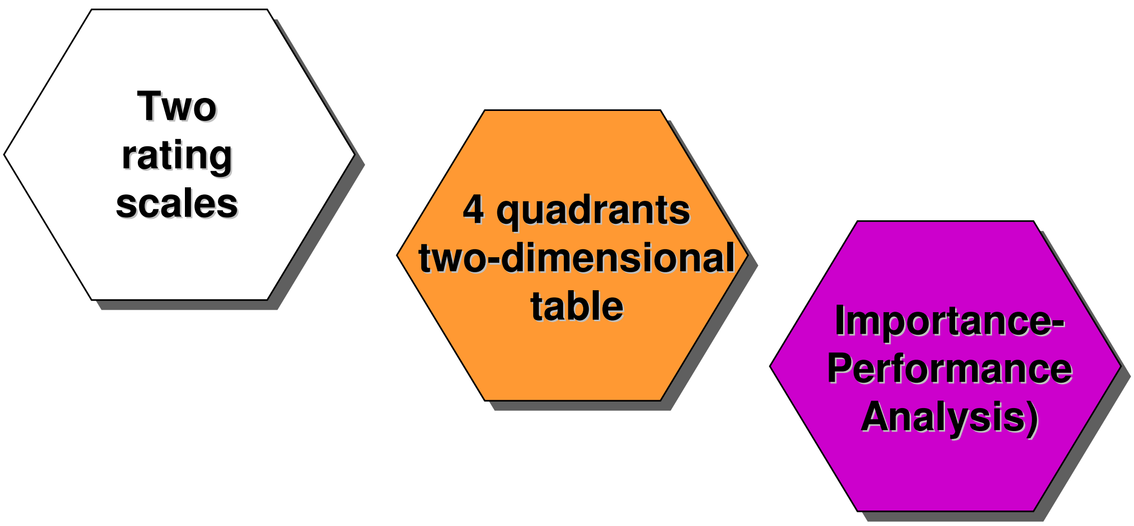
Base

- The number of respondents or observations (in a row or column) used as a basis for computing percentages

Elaboration and Refinement

- Moderator variable
  – A third variable that, when introduced into an analysis, alters or has a contingent effect on the relationship between an independent variable and a dependent variable.
  – Spurious relationship
    - An apparent relationship between two variables that is not authentic.

Quadrant Analysis



Data Transformation

- Data conversion
- Changing the original form of the data to a new format
- More appropriate data analysis
- New variables

Data Transformation

Summative Score =

VAR1 + VAR2 + VAR 3

Health Economics Research Method 2003/2

Collapsing a Five-Point Scale

- Strongly Agree
- Agree
- Neither Agree nor Disagree
- Disagree
- Strongly Disagree

- Strongly Agree/Agree
- Neither Agree nor Disagree
- Disagree/Strongly Disagree

Index Numbers

- Score or observation recalibrated to indicate how it relates to a base number
- CPI - Consumer Price Index

Calculating Rank Order

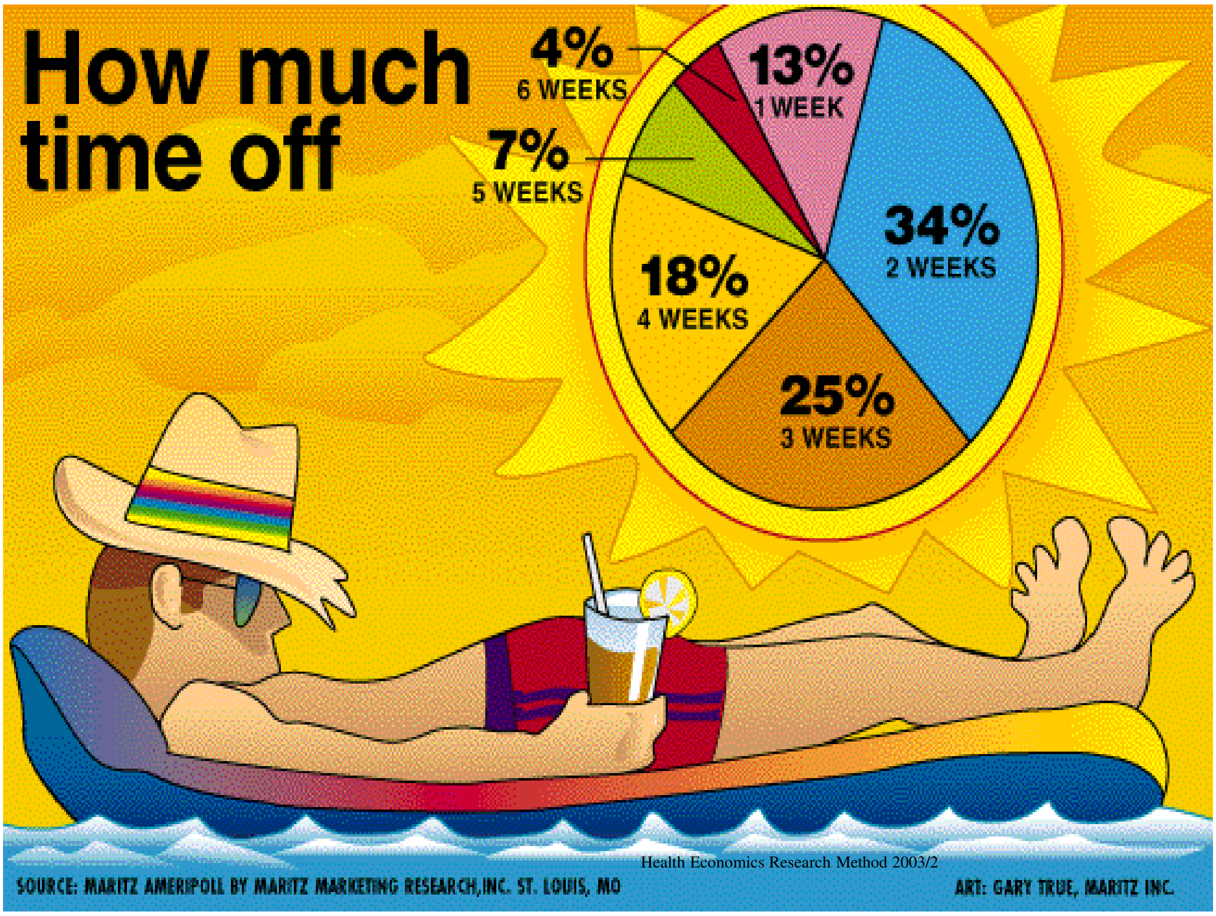
- Ordinal data
- Brand preferences

Tables

- Bannerheads for columns
- Studheads for rows

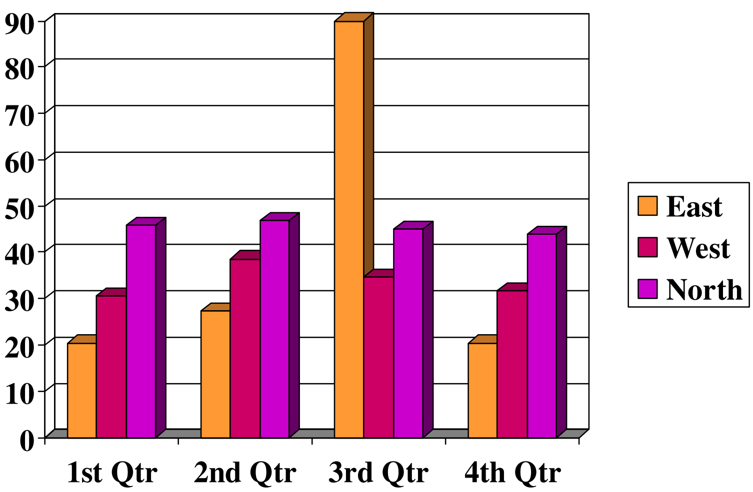
Charts and Graphs

- Pie charts
- Line graphs
- Bar charts
  - Vertical
  - Horizontal

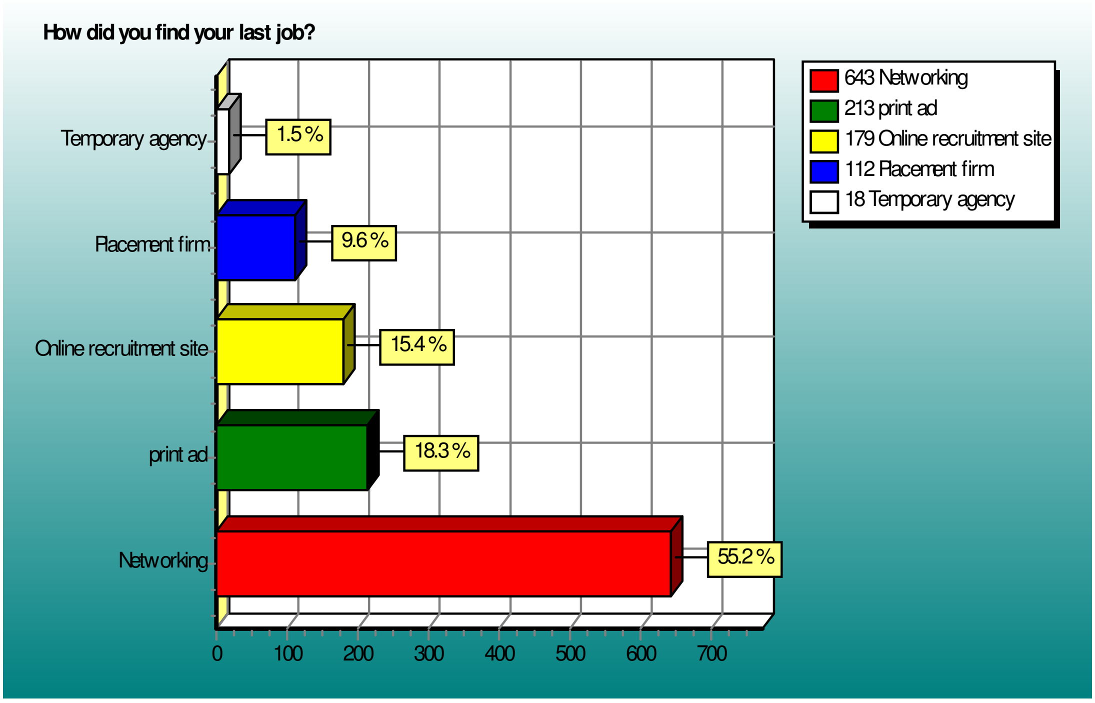


Line Graph

Bar Graph



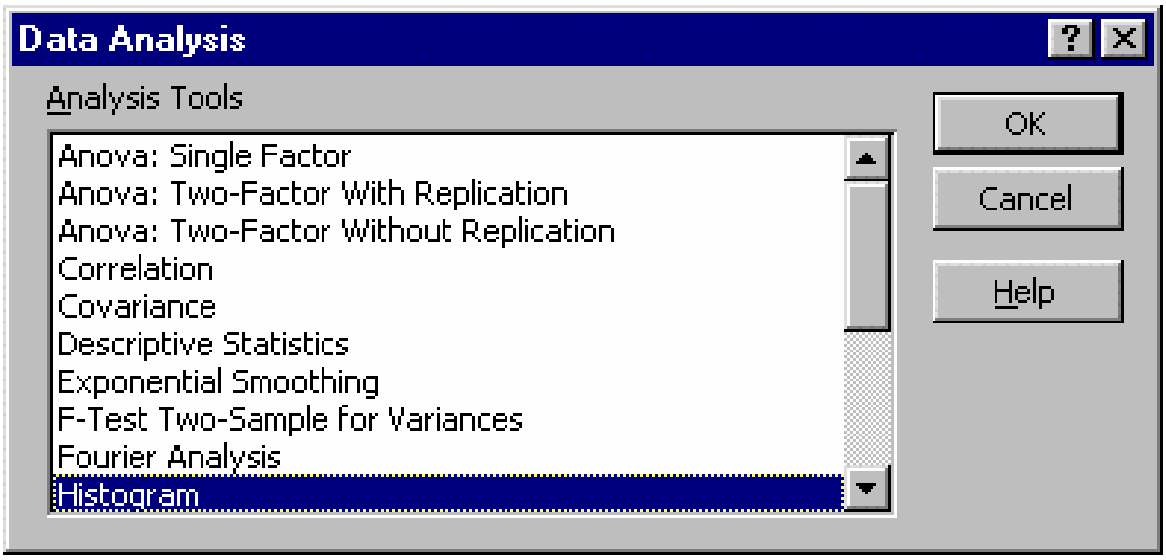
WebSurveyor Bar Chart



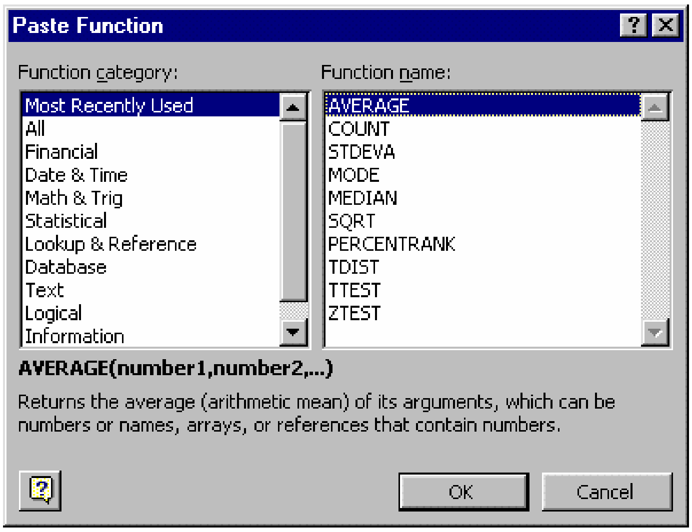
Computer Programs

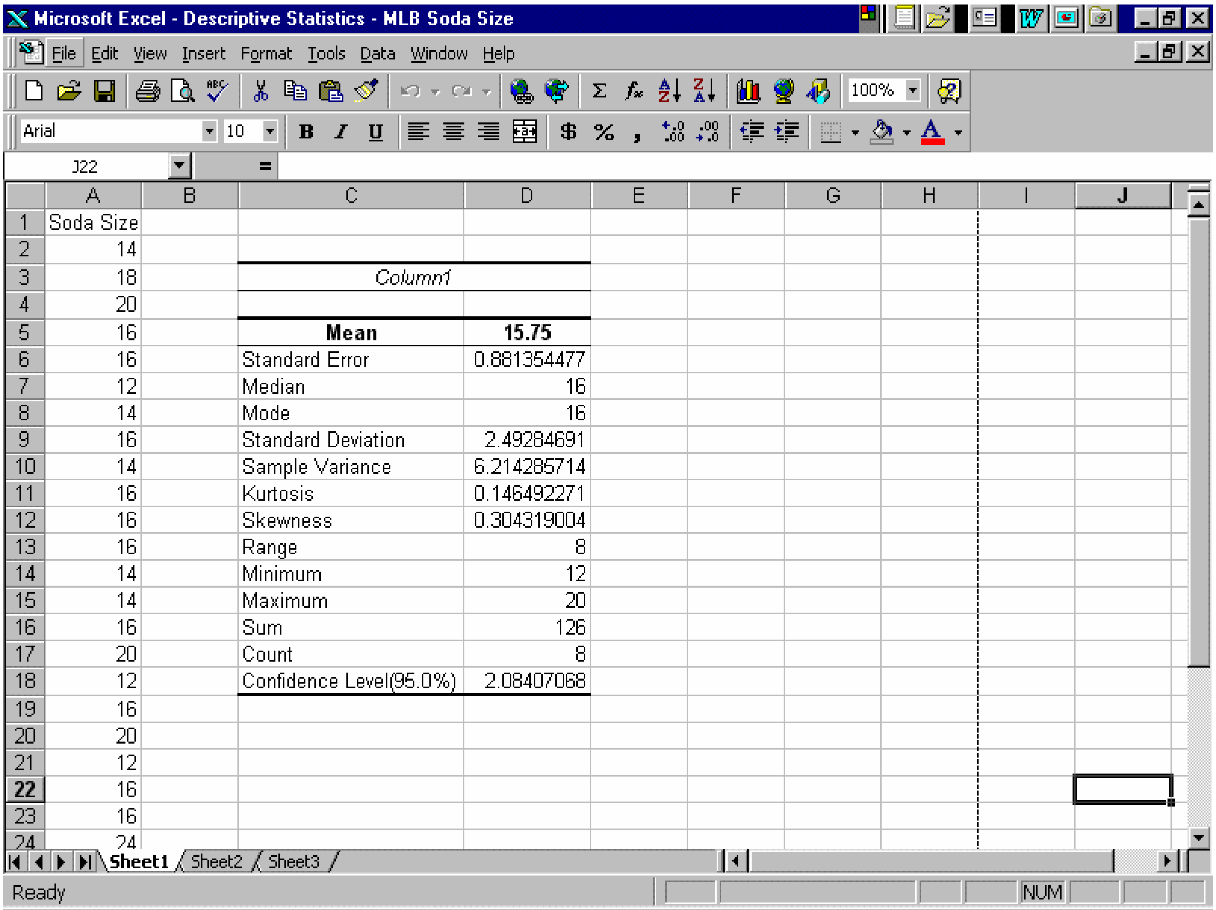
- SPSS
- SAS
- SYSTAT
- Microsoft Excel
- WebSurveyor

Microsoft Excel -Data Analysis



The Paste Function Provides Numerous Statistical Operations





	A	B	C	D
1	Soda Size			
2	14			
3	18		*Column1*	
4	20			
5	16		**Mean**	**15.75**
6	16		Standard Error	0.881354477
7	12		Median	16
8	14		Mode	16
9	16		Standard Deviation	2.49284691
10	14		Sample Variance	6.214285714
11	16		Kurtosis	0.146492271
12	16		Skewness	0.304319004
13	16		Range	8
14	14		Minimum	12
15	14		Maximum	20
16	16		Sum	126
17	20		Count	8
18	12		Confidence Level(95.0%)	2.08407068
19	16			
20	20			
21	12			
22	16			
23	16			
24	24			

Computer Programs

- Box and whisker plots
- Interquartile range - midspread
- Outlier

Interpretation

- The process of making pertinent inferences and drawing conclusions
- concerning the meaning and implications of a research investigation

Research Methods

William G. Zikmund

Univariate Statistics

Univariate Statistics

- Test of statistical significance
- Hypothesis testing one variable at a time

Hypothesis

- Unproven proposition
- Supposition that tentatively explains certain facts or phenomena
- Assumption about nature of the world

Hypothesis

- An unproven proposition or supposition that tentatively explains certain facts or phenomena
  - Null hypothesis
  - Alternative hypothesis

Null Hypothesis

- Statement about the status quo
- No difference

Alternative Hypothesis

- Statement that indicates the opposite of the null hypothesis

Significance Level

- Critical probability in choosing between the null hypothesis and the alternative hypothesis

Significance Level

- Critical Probability
- Confidence Level
- Alpha
- Probability Level selected is typically .05 or .01
- Too low to warrant support for the null hypothesis

The null hypothesis that the mean is equal to 3.0:

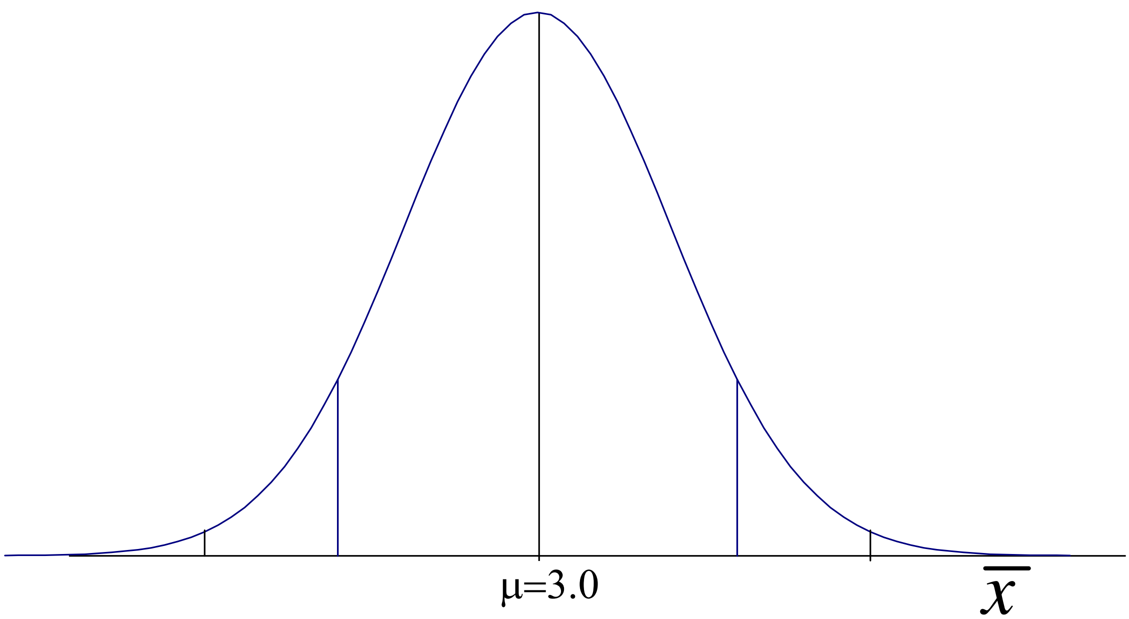
$$H_o : \mu = 3.0$$

Health Economics Research Method 2003/2

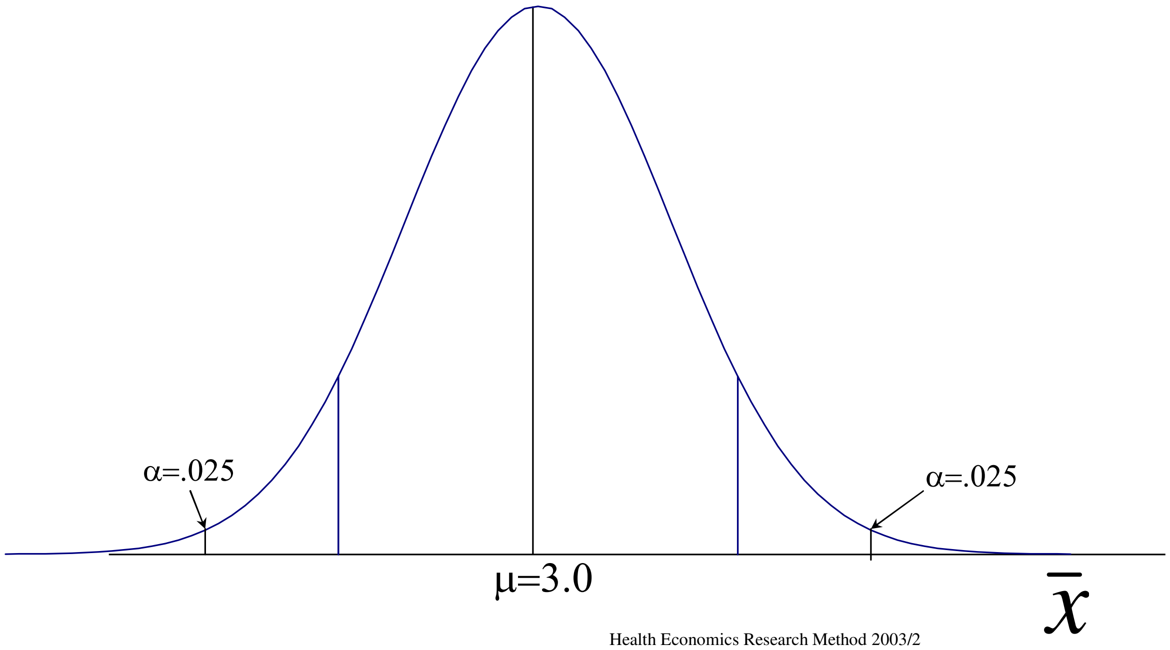
The alternative hypothesis that the mean does *not* equal to 3.0:

$$H_1 : \mu \neq 3.0$$

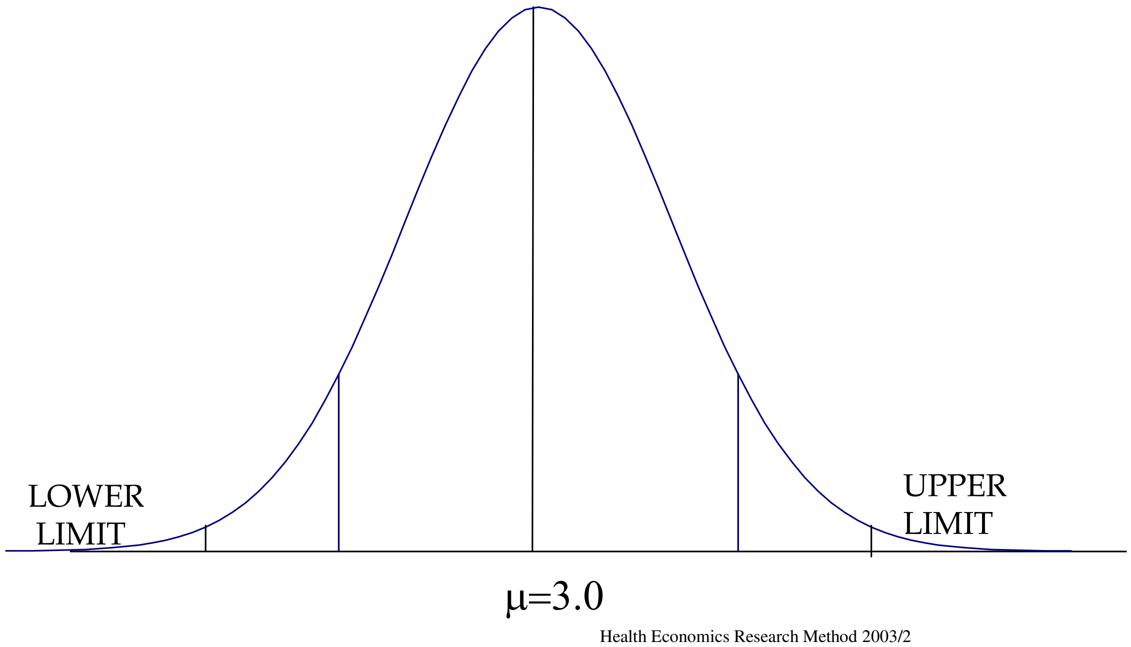
A Sampling Distribution



A Sampling Distribution



A Sampling Distribution



Critical values of μ

Critical value - upper limit

$$= \mu + ZS_{\overline{X}} \quad \text{or} \quad \mu + Z\frac{S}{\sqrt{n}}$$

$$= 3.0 + 1.96\left(\frac{1.5}{\sqrt{225}}\right)$$

Critical values of μ

$$= 3.0 + 1.96(0.1)$$

$$= 3.0 + .196$$

$$= 3.196$$

Critical values of μ

Critical value - lower limit

$$= \mu - ZS_{\bar{X}} \quad \text{or} \quad \mu - Z\frac{S}{\sqrt{n}}$$

$$= 3.0 - 1.96\left(\frac{1.5}{\sqrt{225}}\right)$$

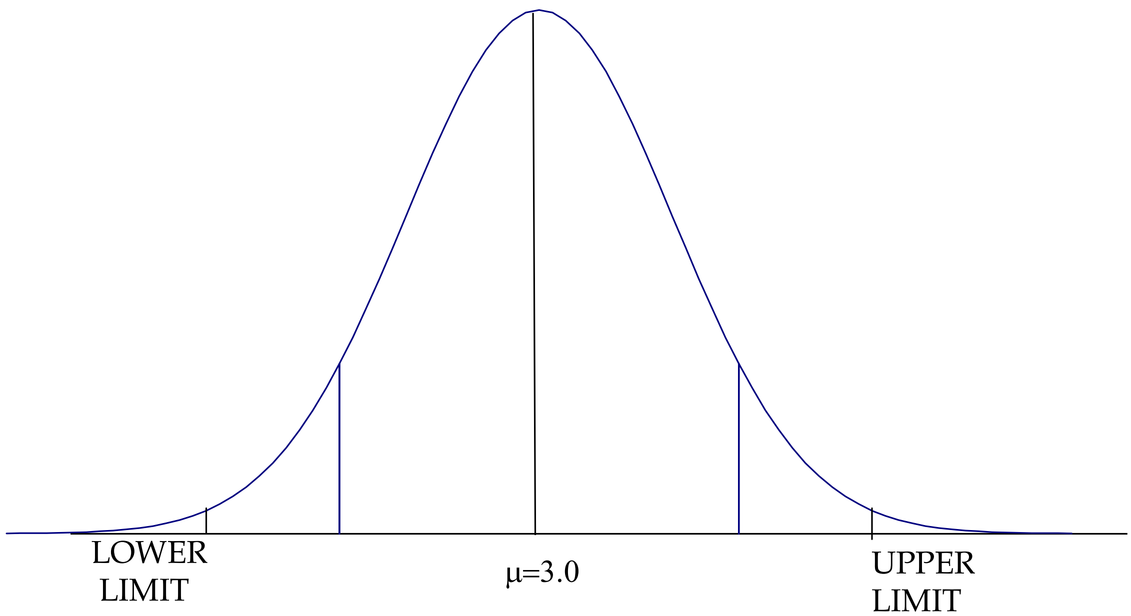
Critical values of μ

$$= 3.0 - 1.96(0.1)$$

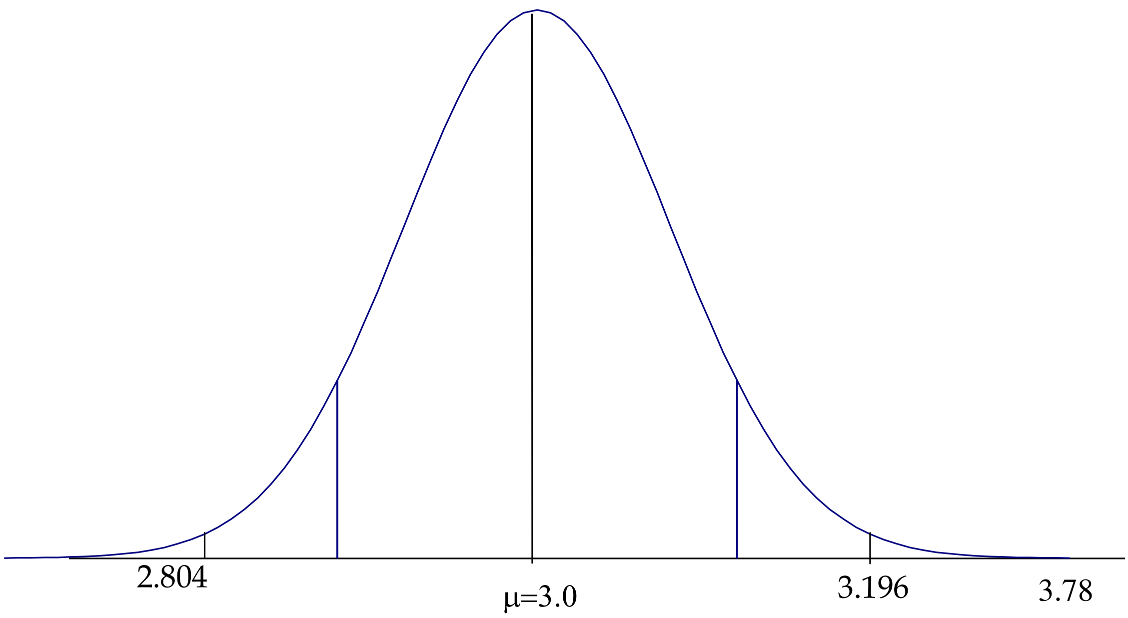
$$= 3.0 - .196$$

$$= 2.804$$

Region of Rejection



Hypothesis Test μ =3.0



Type I and Type II Errors

	Accept null	Reject null
Null is true	Correct-no error	Type I error
Null is false	Type II error	Correct-no error

Type I and Type II Errors in Hypothesis Testing

State of Null Hypothesis in the Population	Decision: Accept Ho	Decision: Reject Ho
Ho is true	Correct--no error	Type I error
Ho is false	Type II error	Correct--no error

Calculating Z_{obs}

$$Z_{obs} = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

Alternate Way of Testing the Hypothesis

$$Z_{obs} = \frac{\bar{X} - \mu}{S_{\bar{X}}}$$

Alternate Way of Testing the Hypothesis

$$Z_{obs} = \frac{3.78 - \mu}{S_{\overline{X}}} = \frac{3.78 - 3.0}{.1}$$

$$= \frac{0.78}{.1} = 7.8$$

Choosing the Appropriate Statistical Technique

- Type of question to be answered
- Number of variables
  - Univariate
  - Bivariate
  - Multivariate
- Scale of measurement

PARAMETRIC STATISTICS

NONPARAMETRIC STATISTICS

t-Distribution

- Symmetrical, bell-shaped distribution
- Mean of zero and a unit standard deviation
- Shape influenced by degrees of freedom

Degrees of Freedom

- Abbreviated d.f.
- Number of observations
- Number of constraints

Confidence Interval Estimate Using the t-distribution

$$\mu = \overline{X} \pm t_{c.l.} S_{\overline{X}}$$

or

$$\text{Upper limit} = \overline{X} + t_{c.l.} \frac{S}{\sqrt{n}}$$

$$\text{Lower limit} = \overline{X} - t_{c.l.} \frac{S}{\sqrt{n}}$$

Confidence Interval Estimate Using the t-distribution

μ = population mean

$\overline{X}$ = sample mean

$t_{c.l.}$ = critical value of t at a specified confidence level

$S_{\overline{X}}$ = standard error of the mean

S = sample standard deviation

n = sample size

Confidence Interval Estimate Using the t-distribution

$$\mu = \overline{X} \pm t_{cl} s_{\bar{x}}$$

$$\overline{X} = 3.7$$

$$S = 2.66$$

$$n = 17$$

$$\text{upper limit} = 3.7 + 2.12(2.66\sqrt{17})$$
$$= 5.07$$

$$\text{Lower limit} = 3.7 - 2.12(2.66\sqrt{17})$$
$$= 2.33$$

Hypothesis Test Using the t-Distribution

Univariate Hypothesis Test Utilizing the t-Distribution

Suppose that a production manager believes the average number of defective assemblies each day to be 20. The factory records the number of defective assemblies for each of the 25 days it was opened in a given month. The mean $\bar{X}$ was calculated to be 22, and the standard deviation, S ,to be 5.

$$H_0 : \mu = 20$$

$$H_1 : \mu \neq 20$$

$$S_{\overline{X}} = S / \sqrt{n}$$

$$= 5 / \sqrt{25}$$

$$= 1$$

Univariate Hypothesis Test Utilizing the t-Distribution

The researcher desired a 95 percent confidence, and the significance level becomes .05.The researcher must then find the upper and lower limits of the confidence interval to determine the region of rejection. Thus, the value of *t* is needed. For 24 degrees of freedom (*n*-1, 25-1), the *t*-value is 2.064.

$$\text{Lower limit:}$$

$$\begin{aligned} \mu - t_{c.l.} S_{\bar{X}} &= 20 - 2.064\left(5/\sqrt{25}\right) \\ &= 20 - 2.064(1) \\ &= 17.936 \end{aligned}$$

Upperlimit:

$$\mu + t_{c.l.} S_{\overline{X}} = 20 + 2.064\left(5/\sqrt{25}\right)$$
$$= 20 + 2.064(1)$$
$$= 20.064$$

Univariate Hypothesis Test t-Test

$$t_{obs} = \frac{\overline{X} - \mu}{S_{\overline{X}}} = \frac{22 - 20}{1}$$
$$= \frac{2}{1}$$
$$= 2$$

Testing a Hypothesis about a Distribution

- Chi-Square test
- Test for significance in the analysis of frequency distributions
- Compare observed frequencies with expected frequencies
- "Goodness of Fit"

Chi-Square Test

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Health Economics Research Method 2003/2

Chi-Square Test

x^2 = chi-square statistics
O_i = observed frequency in the i^{th} cell
E_i = expected frequency on the i^{th} cell

Chi-Square Test
Estimation for Expected Number for Each Cell

$$E_{ij} = \frac{R_i C_j}{n}$$

Chi-Square Test
Estimation for Expected Number for Each Cell

R_i = total observed frequency in the i^{th} row
C_j = total observed frequency in the j^{th} column
n = sample size

Univariate Hypothesis Test
Chi-square Example

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Univariate Hypothesis Test Chi-square Example

$$X^2 = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50}$$

$$= 4$$

Hypothesis Test of a Proportion

π is the population proportion

p is the sample proportion

π is estimated with p

Hypothesis Test of a Proportion

$$H_0 : \pi = .5$$

$$H_1 : \pi \neq .5$$

$$S_p = \sqrt{\frac{(0.6)(0.4)}{100}} = \sqrt{\frac{.24}{100}}$$

$$= \sqrt{.0024} \qquad = .04899$$

$$Zobs = \frac{p - \pi}{S_p} = \frac{.6 - .5}{.04899}$$

$$= \frac{.1}{.04899} = 2.04$$

Hypothesis Test of a Proportion: Another Example

$$n = 1{,}200$$

$$p = .20$$

$$S_p = \sqrt{\frac{pq}{n}}$$

$$S_p = \sqrt{\frac{(.2)(.8)}{1200}}$$

$$S_p = \sqrt{\frac{.16}{1200}}$$

$$S_p = \sqrt{.000133}$$

$$S_p = .0115$$

Hypothesis Test of a Proportion: Another Example

$$n = 1{,}200$$
$$p = .20$$
$$S_p = \sqrt{\frac{pq}{n}}$$
$$S_p = \sqrt{\frac{(.2)(.8)}{1200}}$$
$$S_p = \sqrt{\frac{.16}{1200}}$$
$$S_p = \sqrt{.000133}$$
$$S_p = .0115$$

Hypothesis Test of a Proportion: Another Example

$$Z = \frac{p - \pi}{S_p}$$
$$Z = \frac{.20 - .15}{.0115}$$
$$Z = \frac{.05}{.0115}$$
$$Z = 4.348$$

The Z value exceeds 1.96, so the null hypothesis should be rejected at the .05 level. Indeed it is significant beyond the .001