Introduction to Statistics

Bob Conatser
Irvine 210
Research Associate
Biomedical Sciences
conatser@oucom.ohiou.edu

Statistics - Definition

- The scientific study of numerical data based on variation in nature. (Sokal and Rohlf)
- A set of procedures and rules for reducing large masses of data into manageable proportions allowing us to draw conclusions from those data. (McCarthy)

Basic Terms

- Measurement – assignment of a number to something
- Data – collection of measurements
- Sample – collected data
- Population – all possible data
- Variable – a property with respect to which data from a sample differ in some measurable way.

Types of Measurements

- Ordinal – rank order, (1^{st},2^{nd},3^{rd},etc.)
- Nominal – categorized or labeled data (red, green, blue, male, female)
- Ratio (Interval) – indicates order as well as magnitude. An interval scale does not include zero.

Types of Variables

- Independent Variable – controlled or manipulated by the researcher; causes a change in the dependent variable. (x-axis)
- Dependent Variable – the variable being measured (y-axis)
- Discreet Variable – has a fixed value
- Continuous Variable - can assume any value

Types of Statistics

- Descriptive – used to organize and describe a sample
- Inferential – used to extrapolate from a sample to a larger population

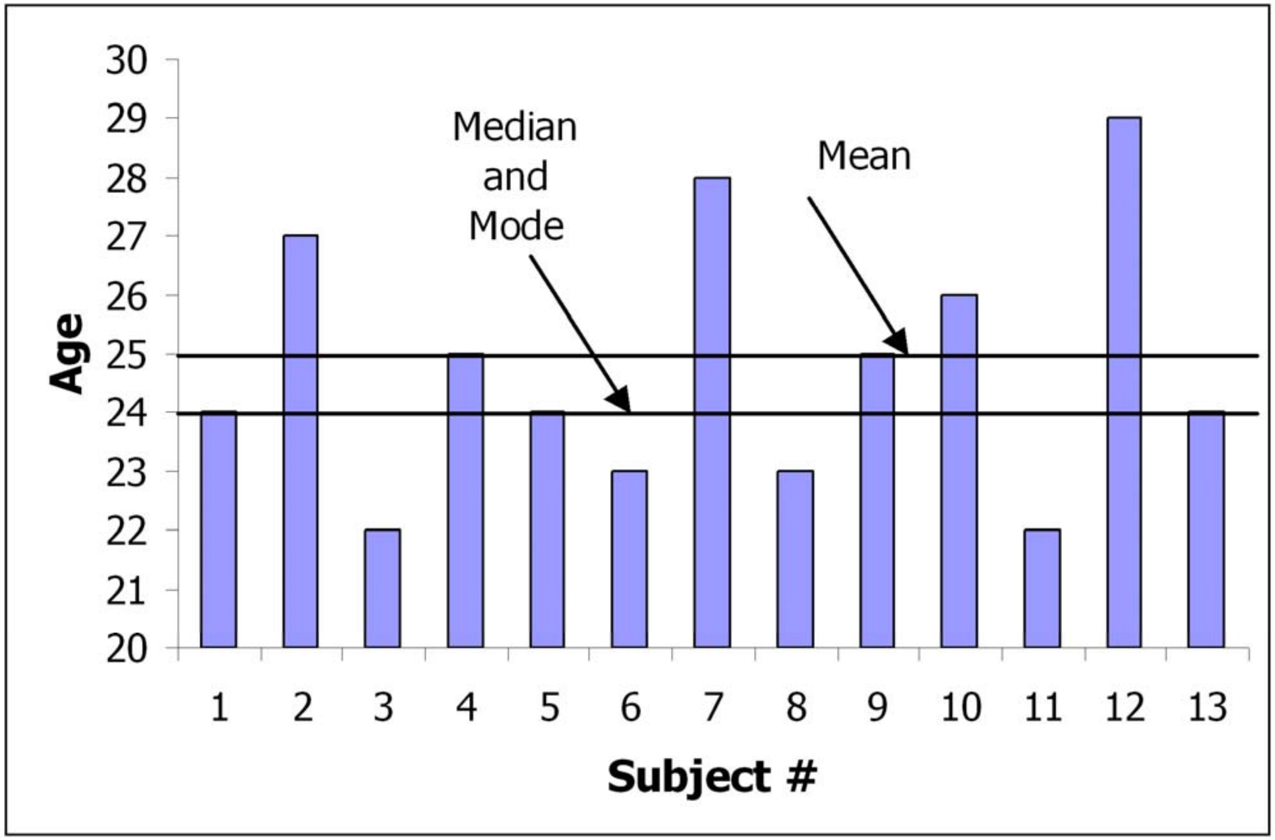
Descriptive Statistics

- Measures of Central Tendency
  - Mean (average)
  - Median (middle)
  - Mode (most frequent)
- Measures of Dispersion
  - variance
  - standard deviation
  - standard error
- Measures of Association
  - correlation

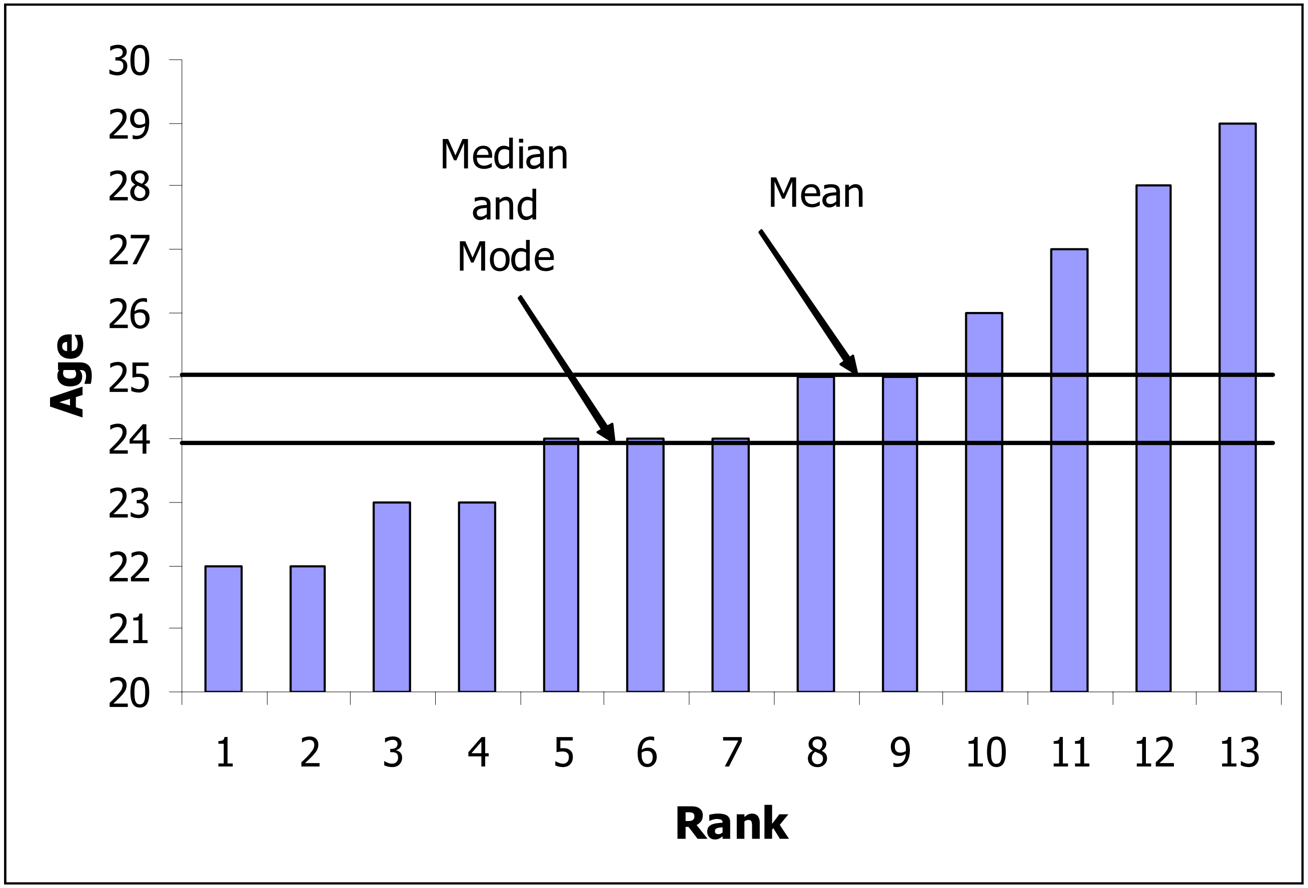
Descriptive Stats
Central Tendency

	A	B	C	D
1	Subject #	Age		**mean** - add up all ages and divide by the toal
2	1	24		Excel command is =average(b2:b25)
3	2	27		(24+27+22+25+24+23+28+23+25+26+22+29+24) / 13
4	3	22		25
5	4	25		
6	5	24		**median** - halfway point, equal number of variables on both sides
7	6	23		Excel command is =median(b2:b23)
8	7	28		22,22,23,23,24,24,24,25,25,26,27,28,29
9	8	23		24
10	9	25		
11	10	26		**mode** - most frequent
12	11	22		Excel command is =mode(b2:b23)
13	12	29		22,22,23,23,24,24,24,25,25,26,27,28,29
14	13	24		24

Descriptive Stats
Central Tendency



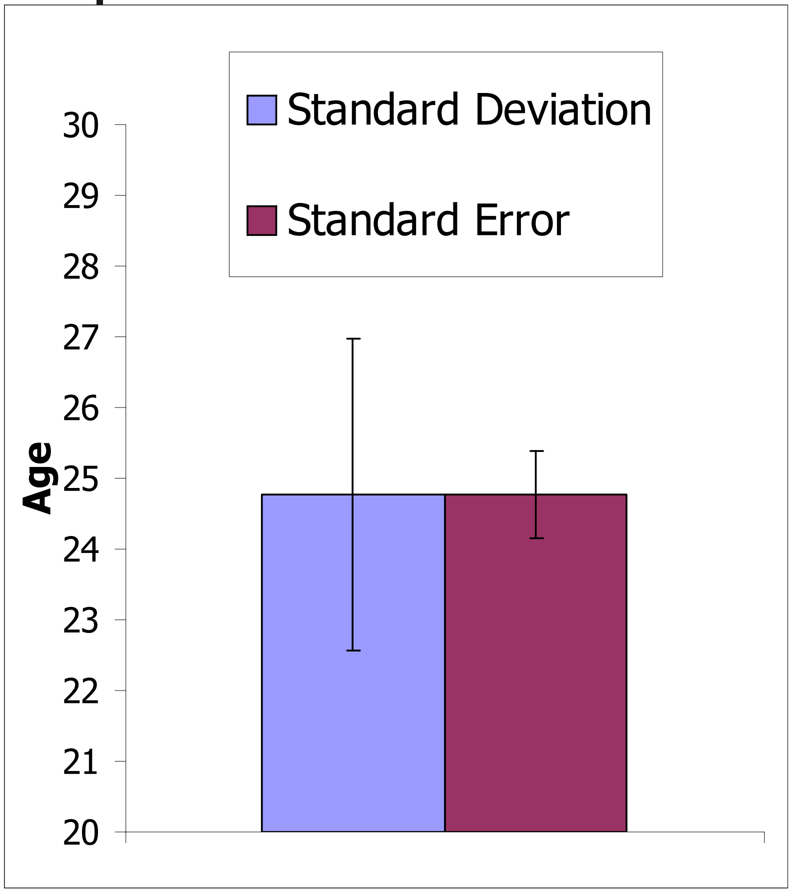
Descriptive Stats Central Tendency

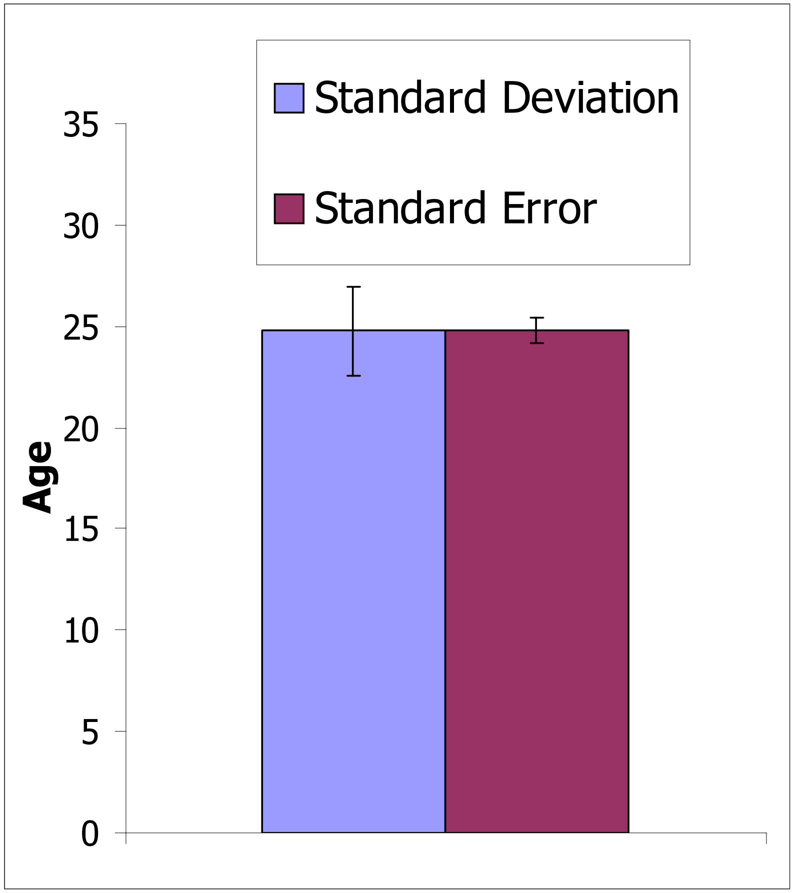


Descriptive Stats
Dispersion

	A	B	C	D
1	Subject #	Age		**mean** - add up all ages and divide by the total
2	1	24		Excel Command is =average(b2:b14)
3	2	27		(24+27+22+25+24+23+28+23+25+26+22+29+24)/13 =
4	3	22		25
5	4	25		
6	5	24		**Standard Deviation** - square root of the sum of the squared
7	6	23		individual differences with the mean divided by
8	7	28		the total number of data points minus 1.
9	8	23		S.D. = $\sqrt{[\Sigma(yi - ymean)2/(N - 1)]}$
10	9	25		Excel command is = stdev(b2:b14)
11	10	26		2.2
12	11	22		
13	12	29		**Standard Error** - Represents the spread in means if many
14	13	24		samples of the same size are taken from the population.
15				S.E. = S.D. / $\sqrt{N}$

Descriptive Stats Dispersion



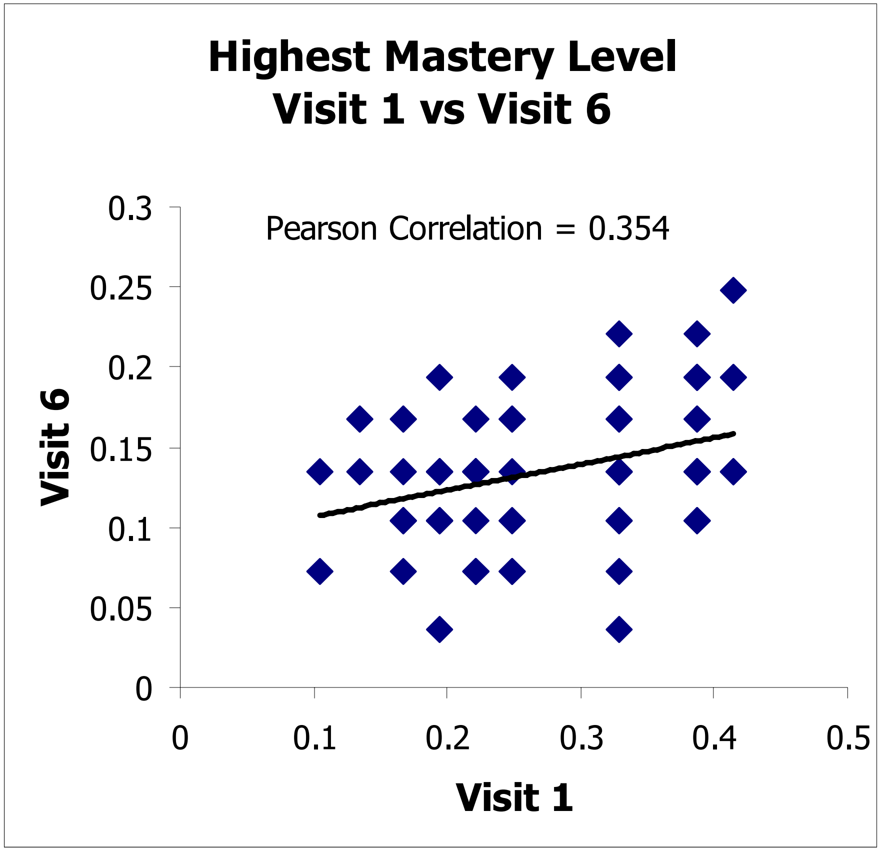


Descriptive Stats
Association

Correlations

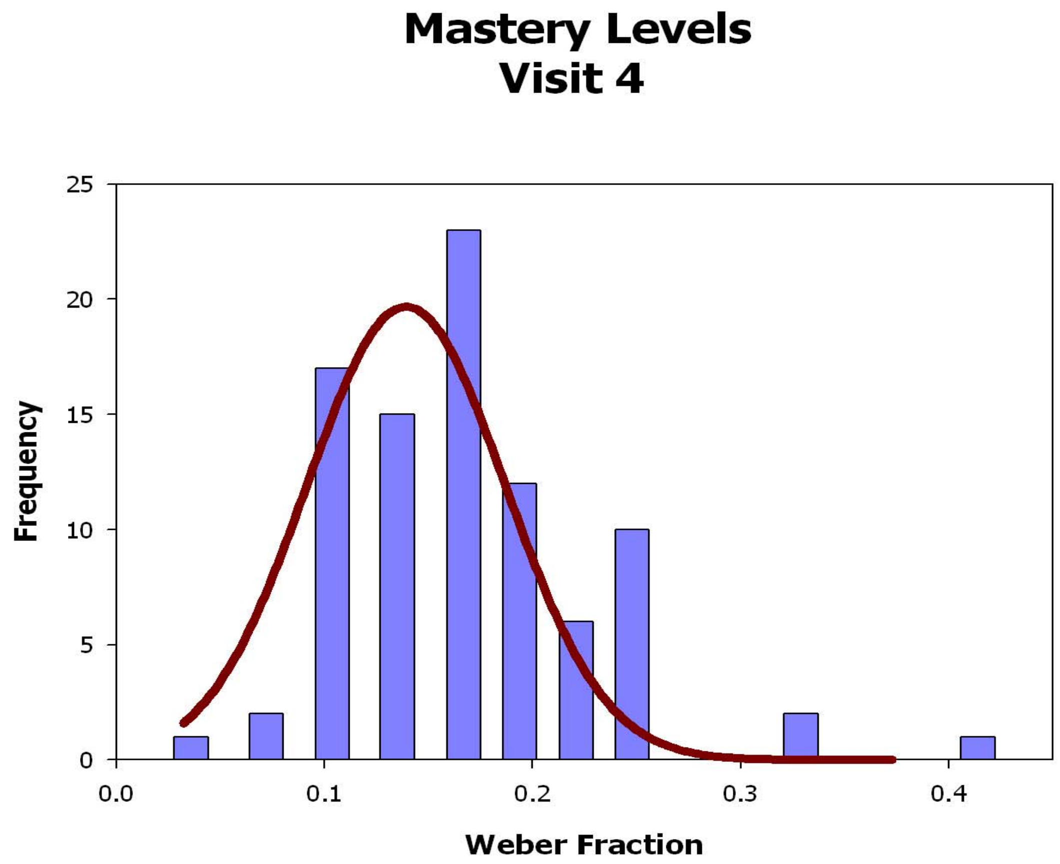
		Visit_1	Visit_6
Visit_1	Pearson Correlation	1	.354**
	Sig. (2-tailed)		.001
	N	83	83
Visit_6	Pearson Correlation	.354**	1
	Sig. (2-tailed)	.001	
	N	83	83

**. Correlation is significant at the 0.01 level

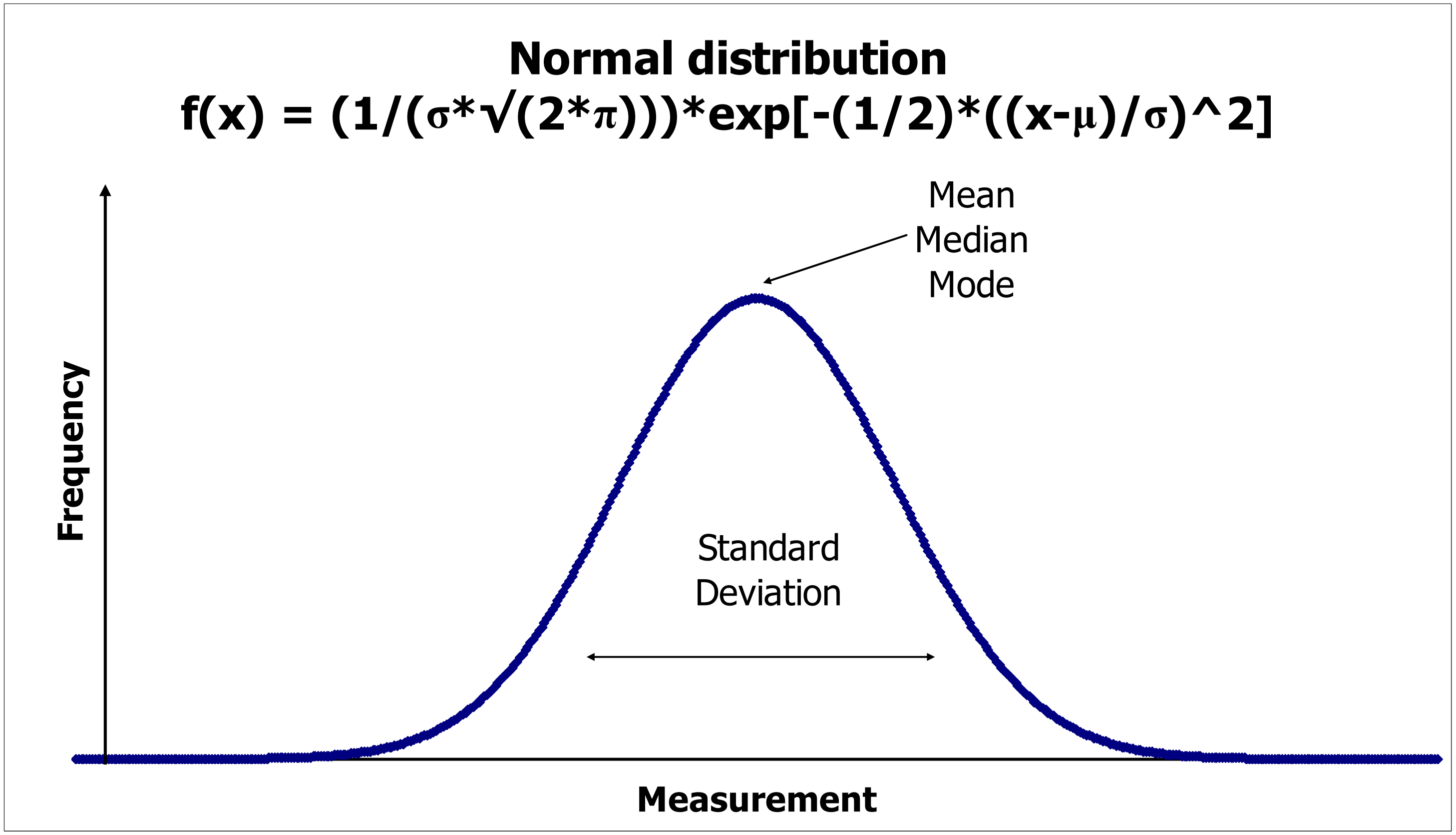


Descriptive Statistics

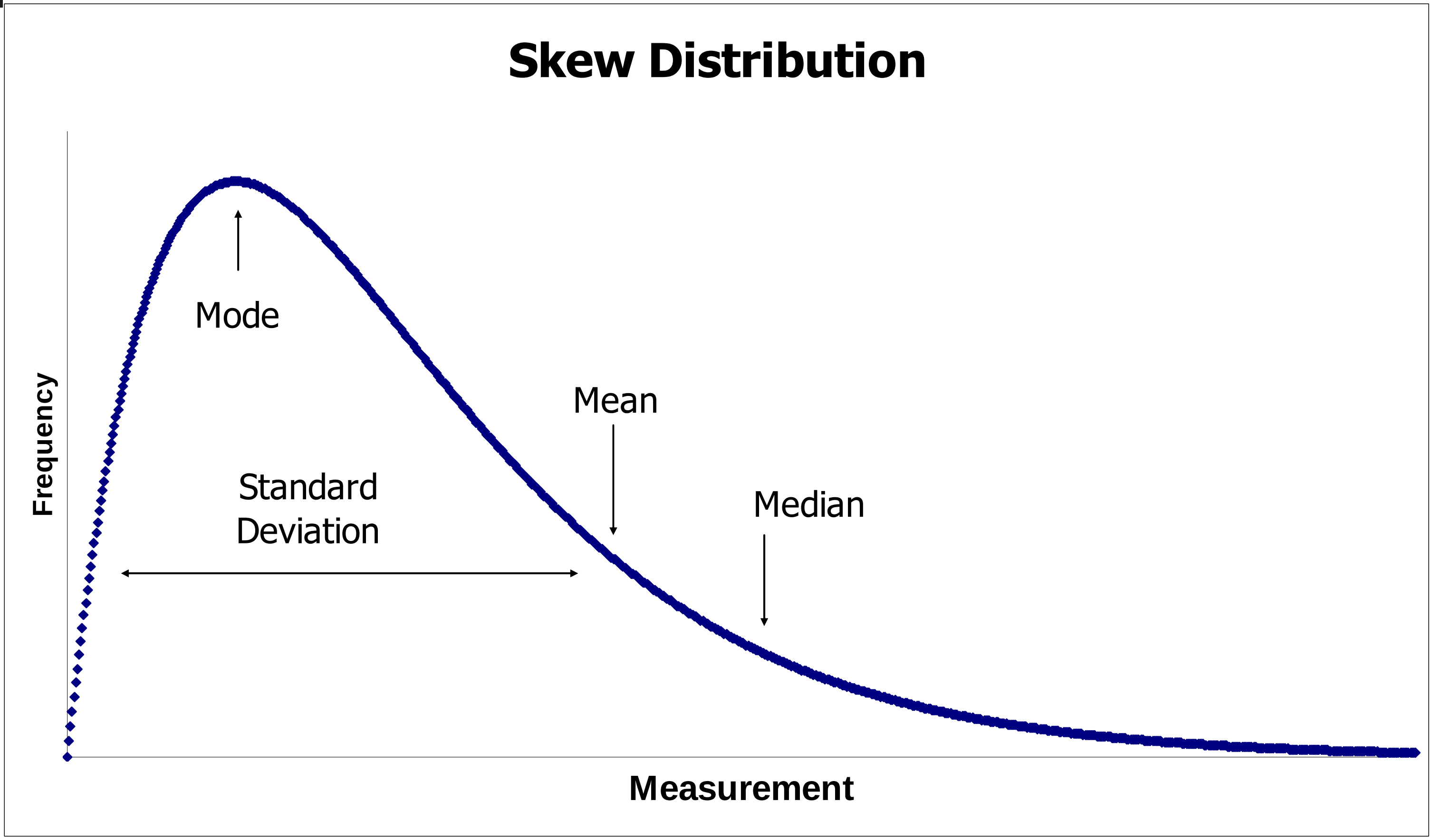
- Data can usually be characterized by a normal distribution.
- Central tendency is represented by the peak of the distribution.
- Dispersion is represented by the width of the distribution.



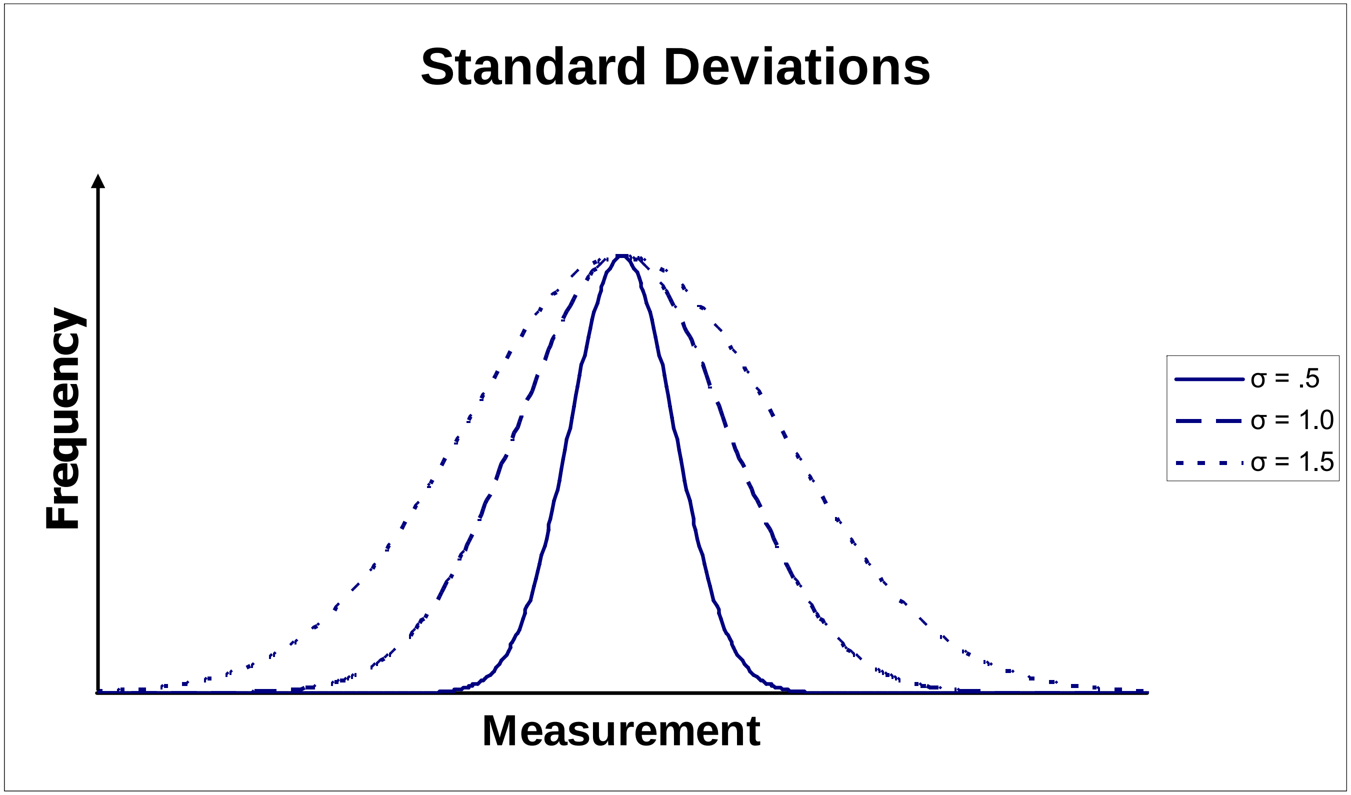
Descriptive Statistics



Descriptive Statistics

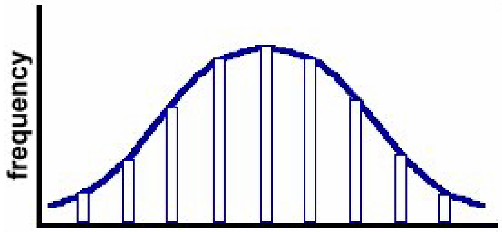


Descriptive Statistics



Inferential Statistics

Can your experiment make a statement about the general population?



Two types

1. Parametric
   - Interval or ratio measurements
   - Continuous variables
   - Usually assumes that data is normally distributed

2. Non-Parametric
   - Ordinal or nominal measurements
   - Discreet variables
   - Makes no assumption about how data is distributed

Inferential Statistics

Null Hypothesis

Statistical hypotheses usually assume no relationship between variables.

- There is no association between eye color and eyesight.

If the result of your statistical test is significant, then the original hypothesis is false and you can say that the variables in your experiment are somehow related.

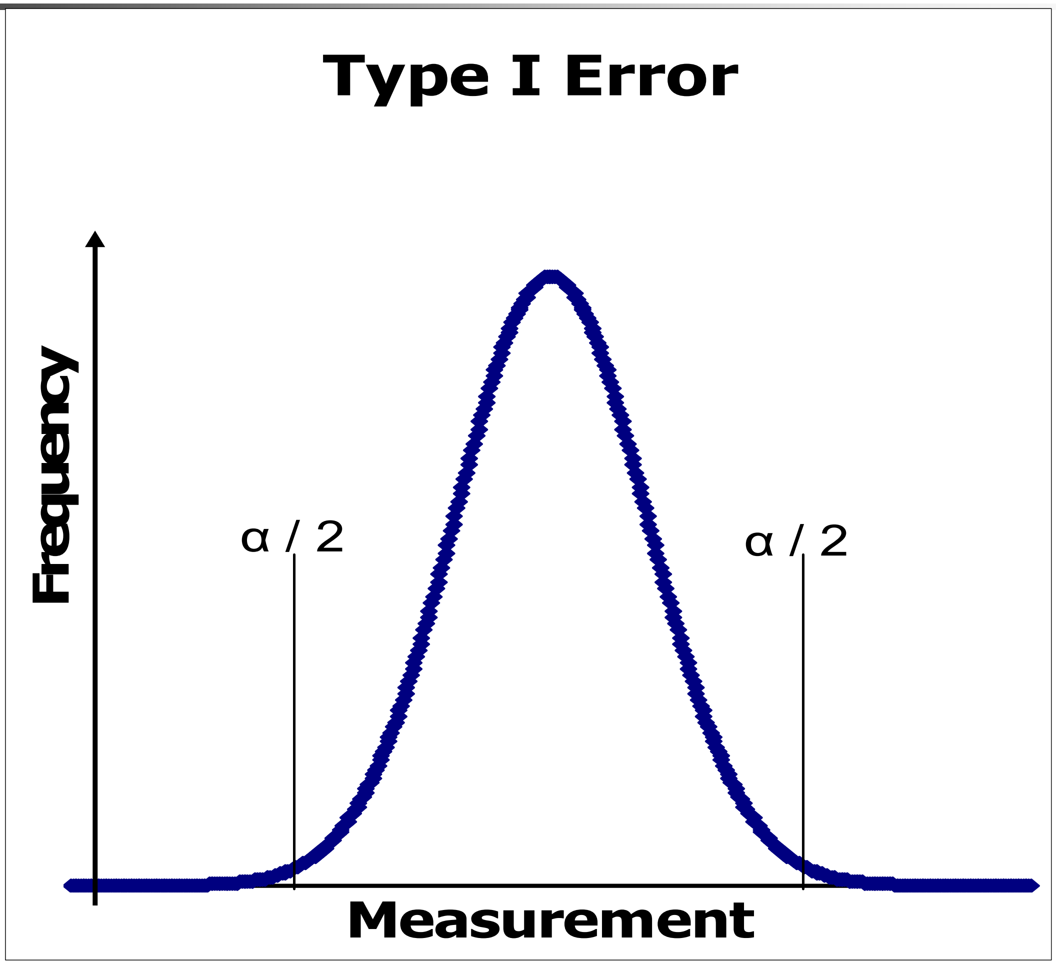
Inferential Statistics - Error

Type I – false positive, α
Type II – false negative, β

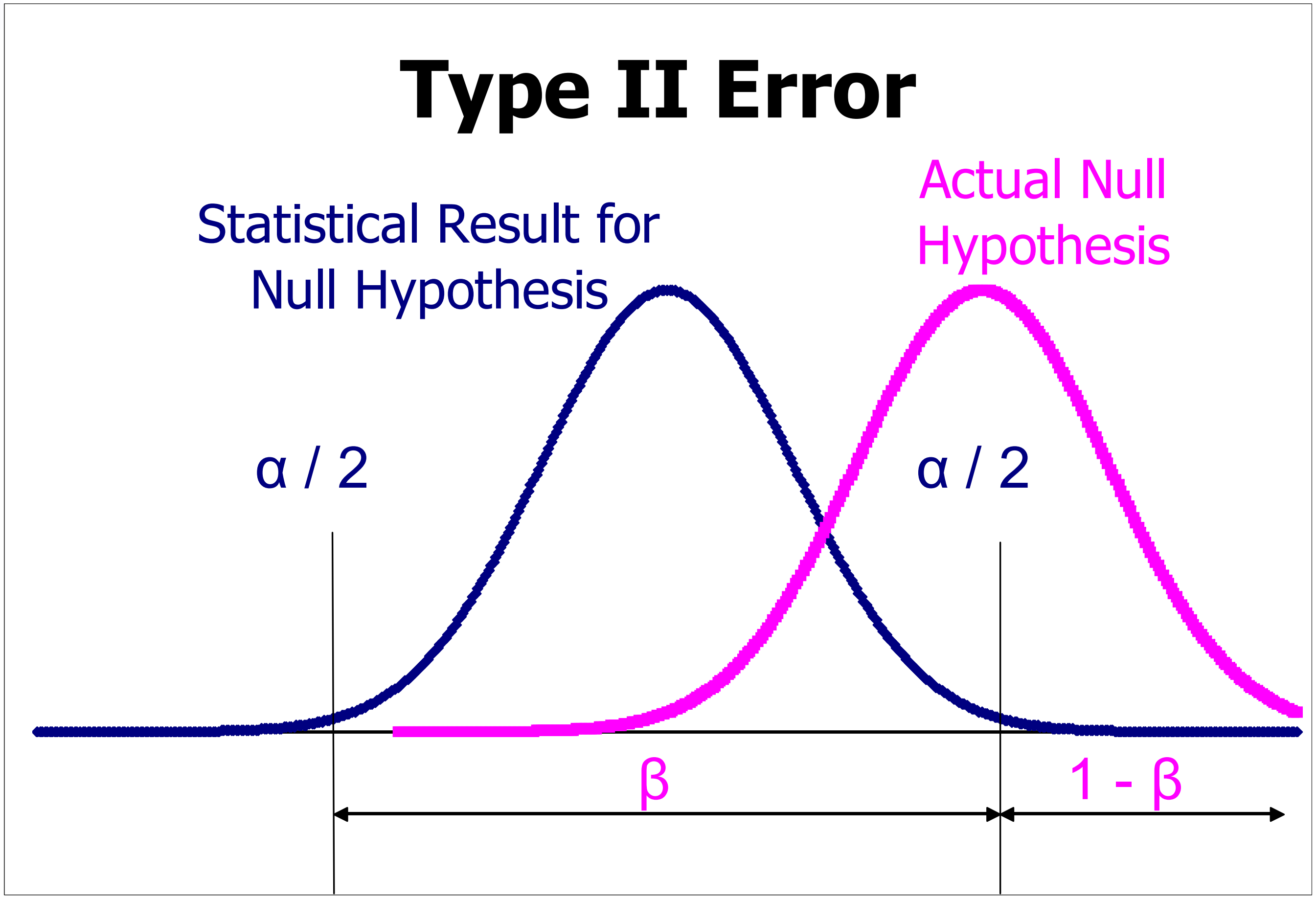
		Statistical Result for Null Hypothesis	
		Accepted	Rejected
Actual Null	TRUE	Correct	Type I Error
Hypothesis	FALSE	Type II Error	Correct

Unfortunately, α and β cannot both have very small values. As one decreases, the other increases.

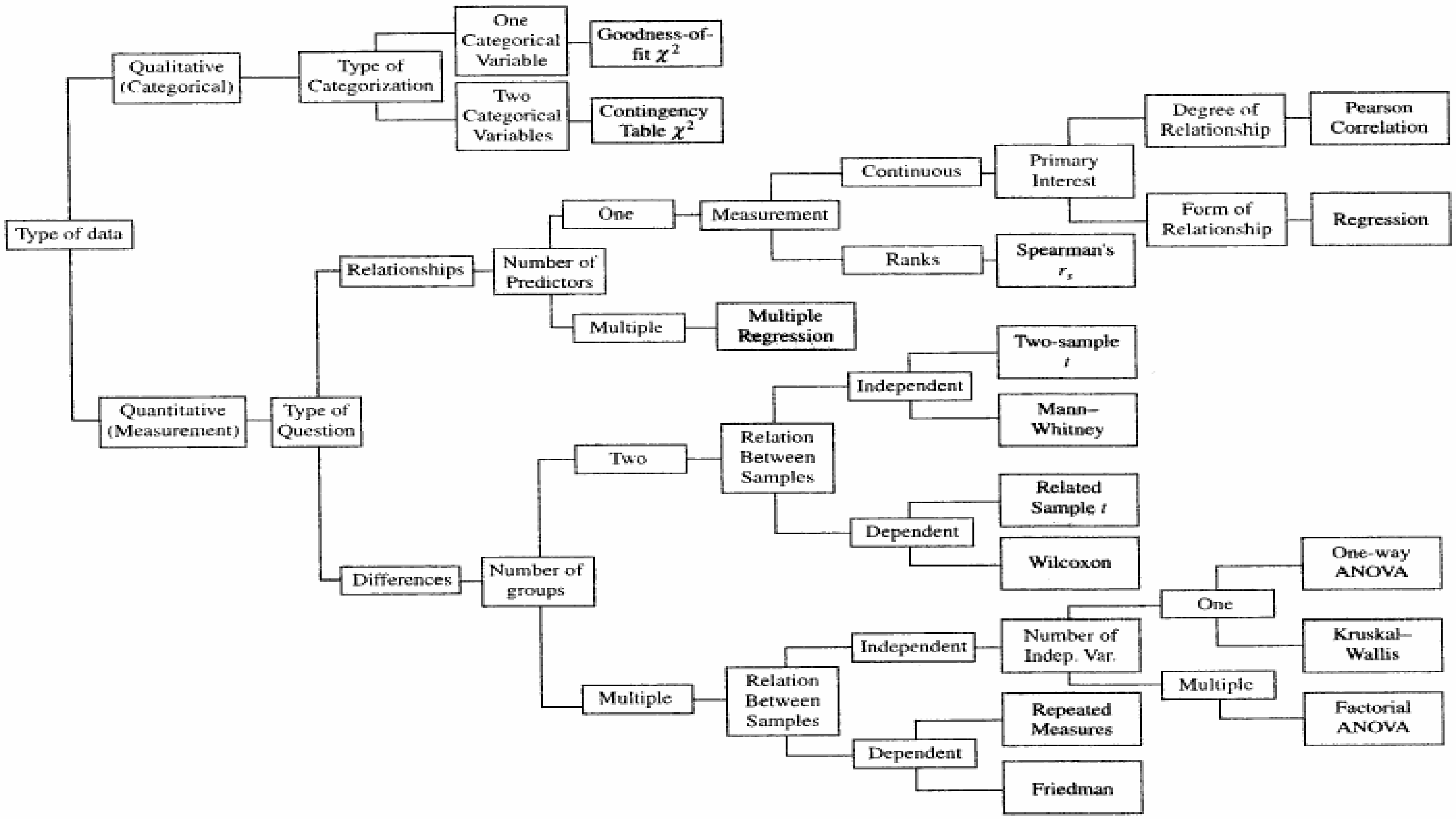
Inferential Statistics - Error



Inferential Statistics - Error



Decision Tree



Howell, D.C. Fundamental Statistics for the Behavioral Sciences. Belmont, CA: Brooks/Cole, Thomson Learning, 2004.

from
Dr. Mann
201 Grosvenor

Inferential Statistics - Power

- The ability to detect a difference between two different hypotheses.
- Complement of the Type II error rate ($1 - \beta$).
- Fix α (= 0.05) and then try to minimize β (maximize $1 - \beta$).

Inferential Statistics - Power

Power depends on:

- sample size
- standard deviation
- size of the difference you want to detect

The sample size is usually adjusted so that power equals .8.

Inferential Statistics

Effect Size

- Detectable difference in means / standard deviation
- Dimensionless
- ~ 0.2 – small (low power)
- ~ 0.5 – medium
- ~ 0.8 – large (powerful test)

Inferential Statistics – T-Test

- Are the means of two groups different?
- Groups assumed to be normally distributed and of similar size.

$\mathbf{t_{\alpha,\nu} = (Y_1 - Y_2) / \sqrt{[(\sigma_1^2 - \sigma_2^2) / n]}}$ (equal sample sizes)

$\mathbf{Y_1}$ and $\mathbf{Y_2}$ are the means of each group
σ_1 and σ_2 are the standard deviations
n is the number of data points in each group
α is the significance level (usually 0.05)
ν is the degrees of freedom (2 * (n – 1)) (Sokal & Rohlf)

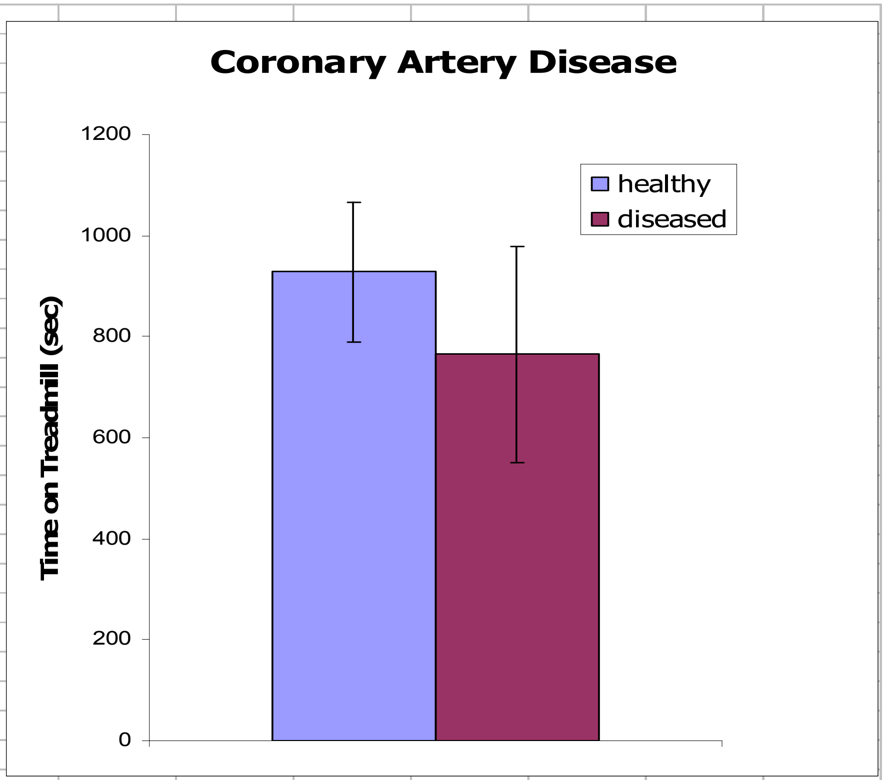
Inferential Statistics – T-Test

	Two Tailed Significance (α)			
ν	0.2	0.1	0.05	0.01
2	1.89	2.92	4.30	9.92
3	1.64	2.35	3.18	5.84
4	1.53	2.13	2.78	4.60
5	1.48	2.02	2.57	4.03
6	1.44	1.94	2.45	3.71
7	1.41	1.89	2.36	3.50
8	1.40	1.86	2.31	3.36
9	1.38	1.83	2.26	3.25
10	1.37	1.81	2.23	3.17

Compare calculated $\mathbf{t}_{\alpha,\nu}$ value with value from table.
If calculated value is larger, the null hypothesis is false. (Lentner, C., 1982, *Geigy Scientific Tables vol. 2*, CIBA-Geigy Limited, Basle, Switzerland)

T-Test Example

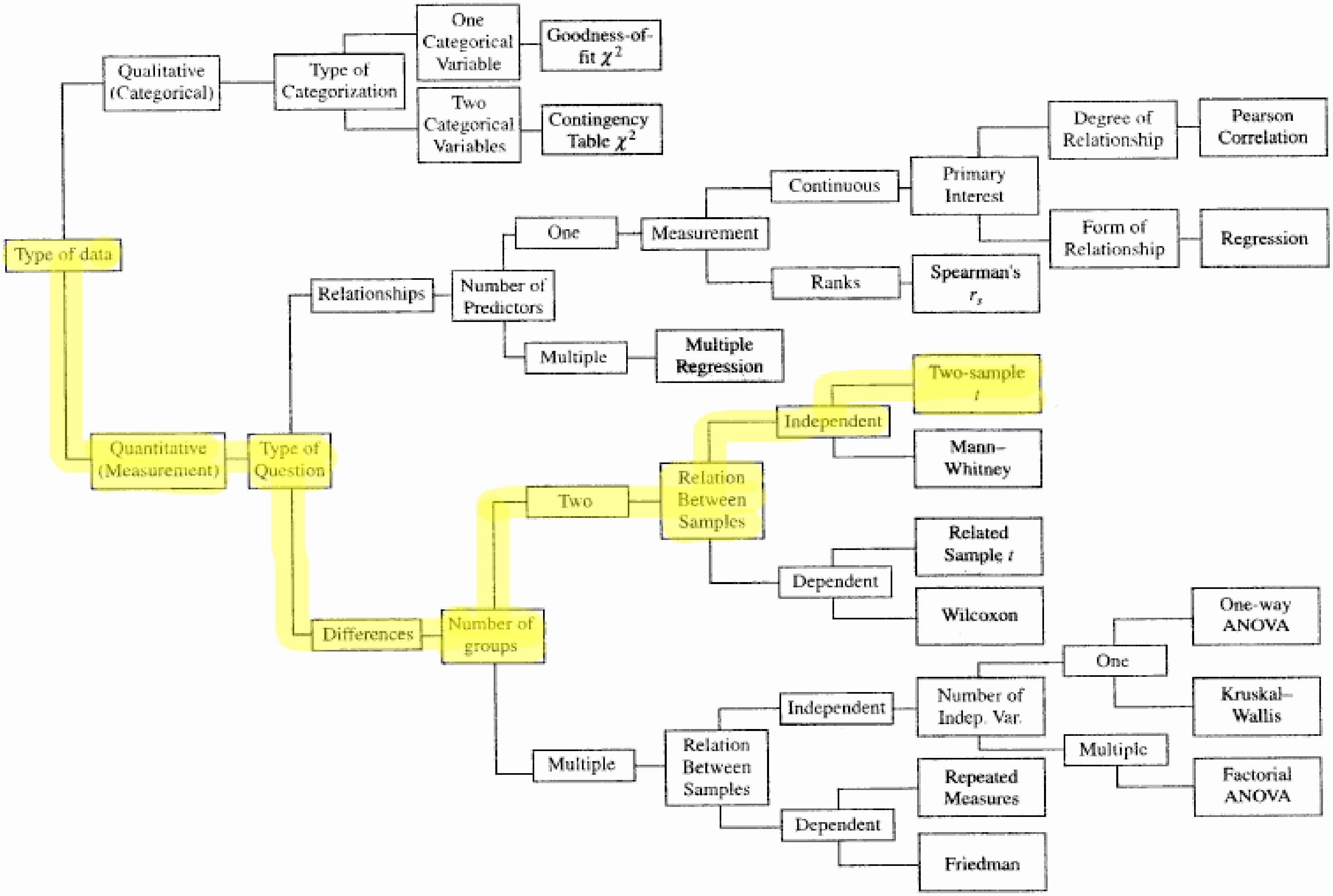
Group 1 - healthy 2 - diseased	Treadmill time in seconds
1	1014
1	684
1	810
1	990
1	840
1	978
1	1002
1	1110
2	864
2	636
2	638
2	708
2	786
2	600
2	1320
2	750
2	594
2	750
Group1	
mean	928.5
StDev	138.121
Group2	
mean	764.6
StDev	213.75



Example data from SPSS

Null Hypothesis - There is no difference between healthy people and people with coronary artery disease in time spent on a treadmill.

Decision Tree



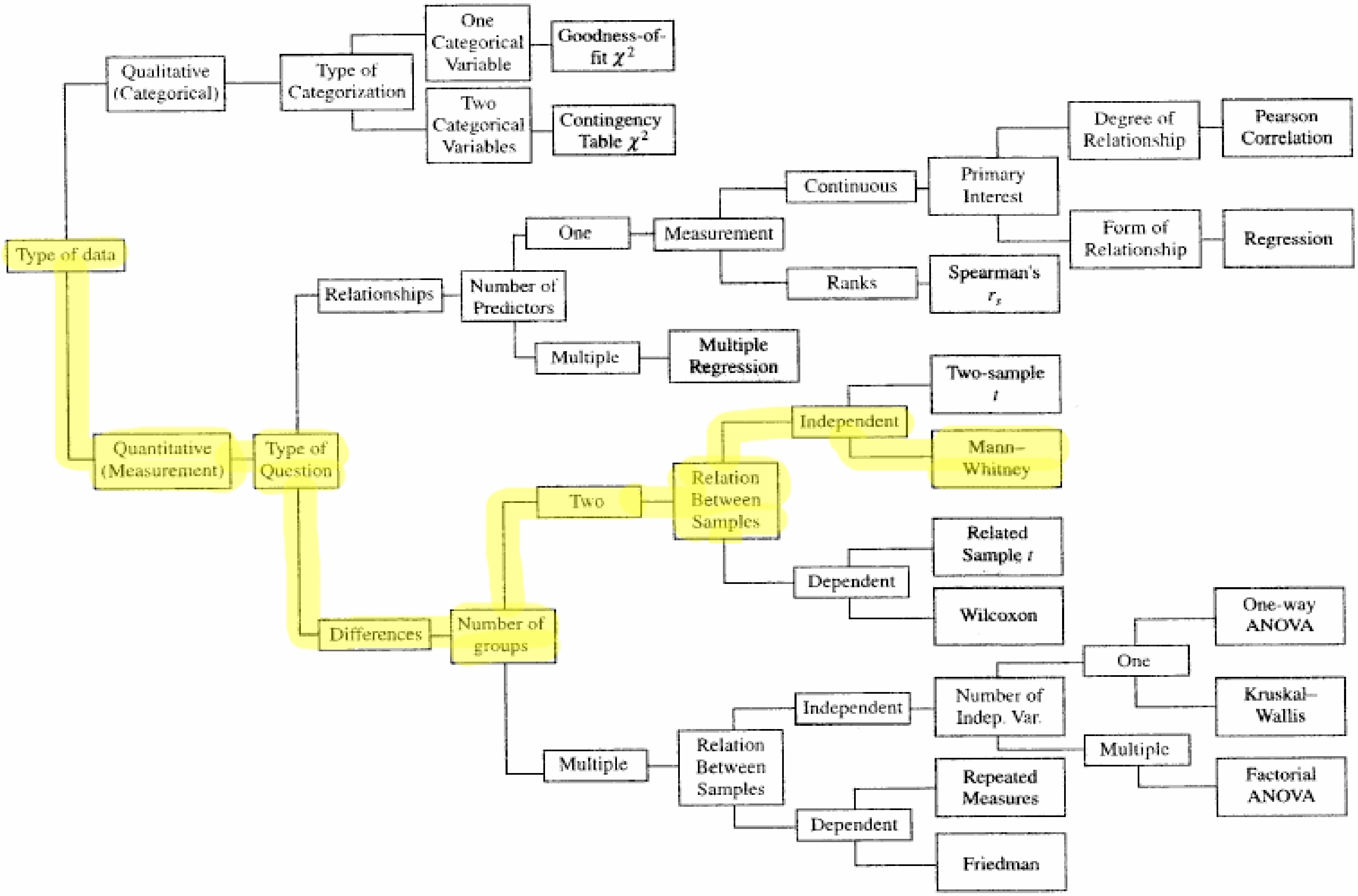
T-Test Example (cont.)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
Treadmill time in seconds	Equal variances assumed	.137	.716	1.873	16	.080	163.900	87.524	-21.642	349.442
	Equal variances not assumed			1.966	15.439	.068	163.900	83.388	-13.398	341.198

Null hypothesis is accepted because the results are not significant at the 0.05 level.

Decision Tree



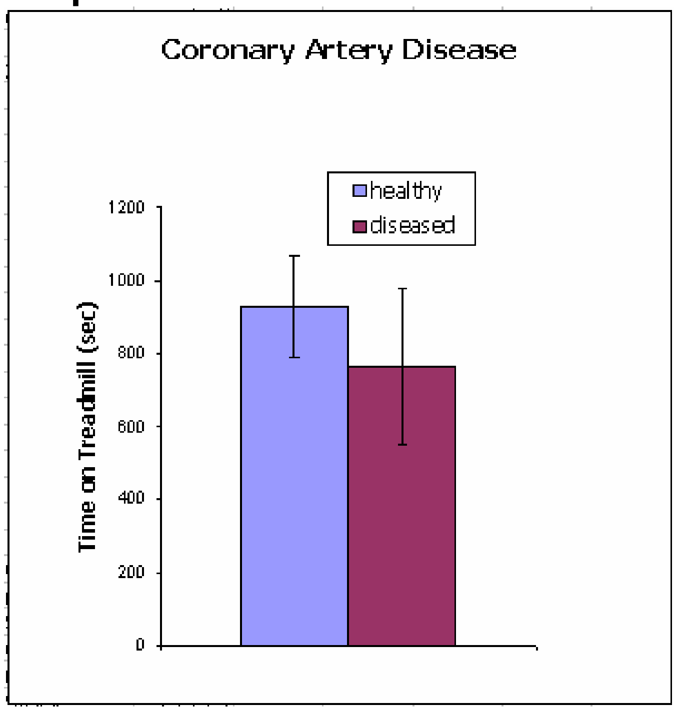
Non-Parametric Statistics

- Makes no assumptions about the population from which the samples are selected.
- Used for the analysis of discreet data sets.
- Also used when data does not meet the assumptions for a parametric analysis ("small" data sets).

Non-Parametric Example I

Mann-Whitney

Most commonly used as an alternative to the independent samples T-Test.



Test Statistics[b]

	Treadmill time in seconds
Mann-Whitney U	15.000
Wilcoxon W	70.000
Z	-2.222
Asymp. Sig. (2-tailed)	.026
Exact Sig. [2*(1-tailed Sig.)]	.027[a]

a. Not corrected for ties.

b. Grouping Variable: group

Note difference in results between this test and T-Test.

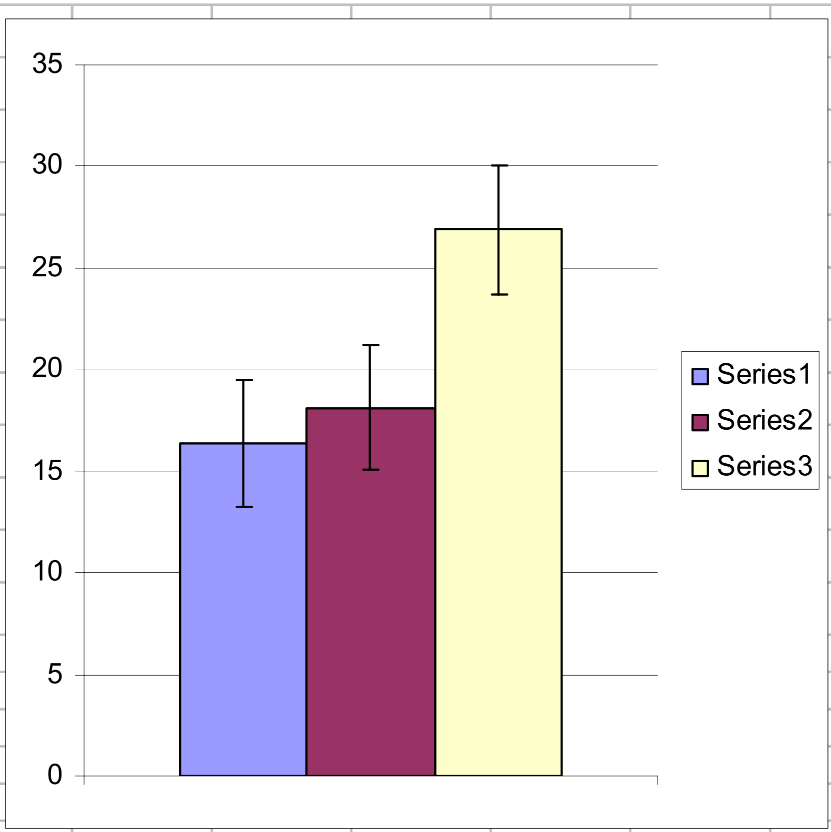
Inferential Statistics - ANOVA

ANOVA – **<u>An</u>**alysis **<u>o</u>**f **<u>Va</u>**riance

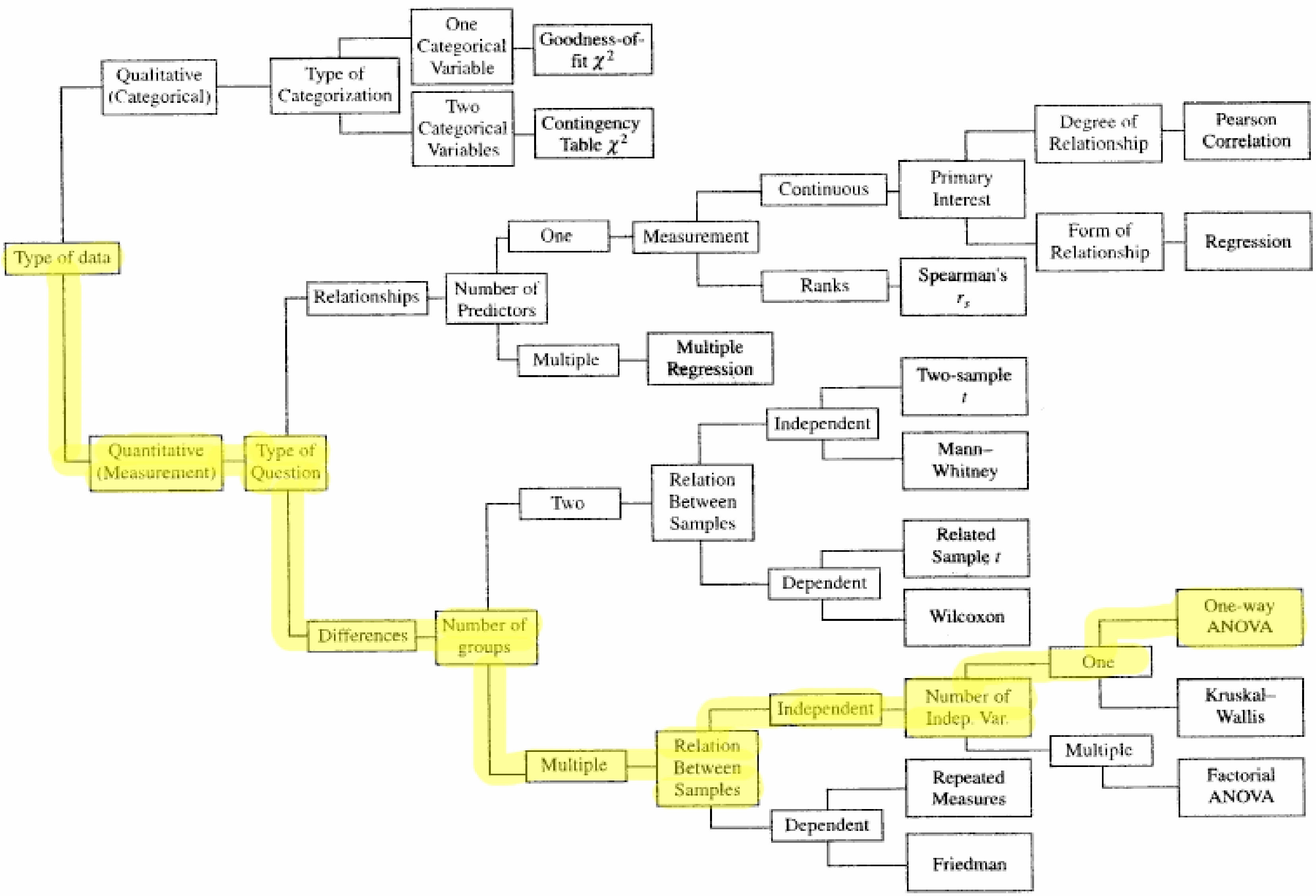
- Compares the means of 3 or more groups
- Assumptions:
  - Groups relatively equal.
  - Standard deviations similar. (Homogeneity of variance)
  - Data normally distributed.
  - Sampling should be randomized.
  - Independence of errors.
- Post-Hoc test

ANOVA - Example

	12	15	23
	13	15	23
	15	16	25
	15	17	26
	17	17	27
	19	21	30
	20	21	30
	20	23	31
Mean	16.38	18.13	26.88
StDev	3.114	3.091	3.182
StErr	0.794	0.747	0.626



Decision Tree



ANOVA - Results

1.

Test of Homogeneity of Variances

VAR00001

Levene Statistic	df1	df2	Sig.
.001	2	21	.999

3.

Multiple Comparisons

Dependent Variable: VAR00001
Tukey HSD

(I) VAR00004	(J) VAR00004	Mean Difference (I-J)	Std. Error	Sig.
1.00	2.00	-1.75000	1.56458	.514
	3.00	-10.50000*	1.56458	.000
2.00	1.00	1.75000	1.56458	.514
	3.00	-8.75000*	1.56458	.000
3.00	1.00	10.50000*	1.56458	.000
	2.00	8.75000*	1.56458	.000

*. The mean difference is significant at the .05 level.

2. **ANOVA**

VAR00001

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	506.333	2	253.167	25.855	.000
Within Groups	205.625	21	9.792		
Total	711.958	23			

ANOVA – Example 2

	1	322.000	2	322.007	3	321.986	4	321.994
	1	322.005	2	322.031	3	321.990	4	322.003
	1	322.022	2	322.011	3	322.002	4	322.006
	1	321.991	2	322.029	3	321.984	4	322.003
	1	322.011	2	322.009	3	322.017	4	321.986
	1	321.995	2	322.026	3	321.983	4	322.002
	1	322.006	2	322.018	3	322.002	4	321.998
	1	321.976	2	322.007	3	322.001	4	321.991
	1	321.998	2	322.018	3	322.004	4	321.996
	1	321.996	2	321.986	3	322.016	4	321.999
	1	321.984	2	322.018	3	322.002	4	321.990
	1	321.984	2	322.018	3	322.009	4	322.002
	1	322.004	2	322.020	3	321.990	4	321.986
	1	322.000	2	322.012	3	321.993	4	321.991
	1	322.003	2	322.014	3	321.991	4	321.983
	1	322.002	2	322.005	3	322.002	4	321.998
Mean		321.999		322.014		321.998		321.995
StDev		0.011		0.011		0.010		0.007

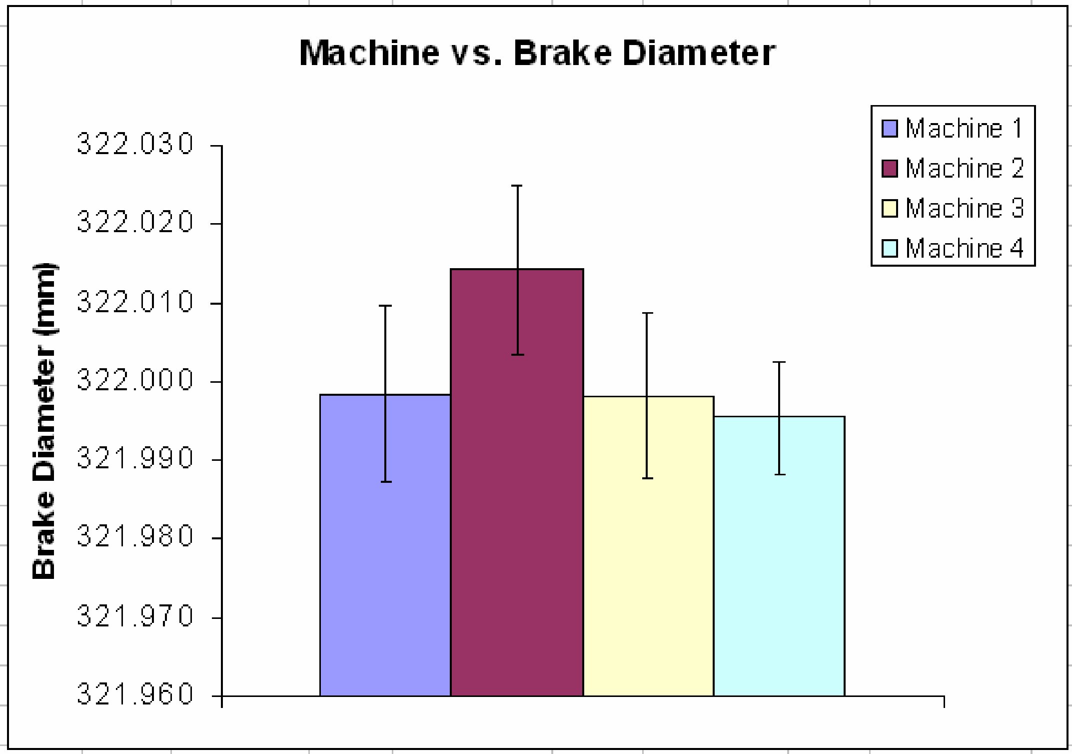
SPSS example
Machine type vs. brake diameter

4 group ANOVA

Null Hypothesis –

There is no difference among machines in brake diameter.

ANOVA – Example 2



ANOVA – Example 2 - Results

2.

ANOVA

Disc Brake Diameter (mm)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.004	3	.001	11.748	.000
Within Groups	.006	60	.000		
Total	.009	63			

Null hypothesis is rejected because result is highly significant.

1.

Test of Homogeneity of Variances

Disc Brake Diameter (mm)

Levene Statistic	df1	df2	Sig.
.697	3	60	.557

3.

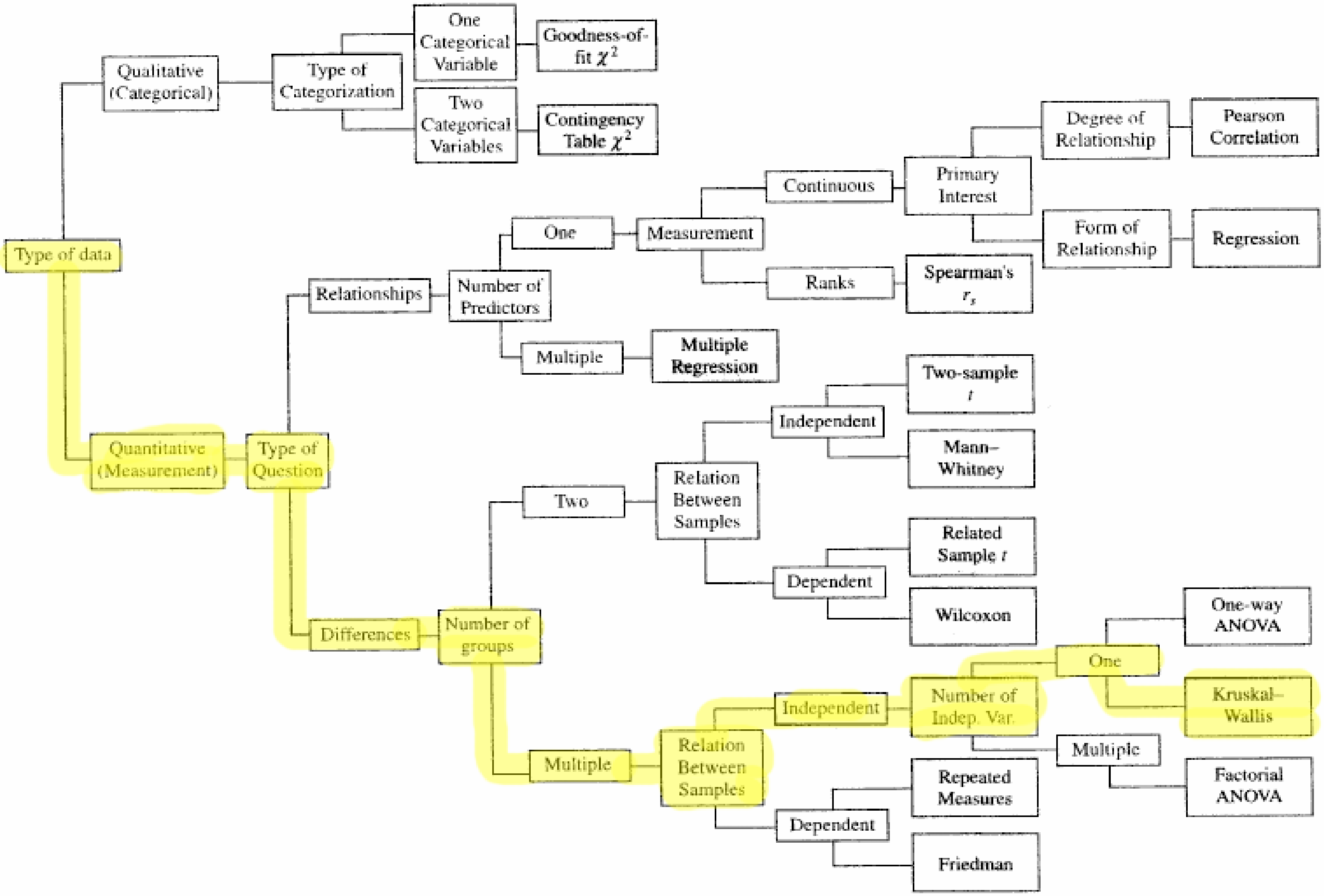
Multiple Comparisons

Dependent Variable: Disc Brake Diameter (mm)
Tukey HSD

(I) Machine Number	(J) Machine Number	Mean Difference (I-J)	Sig.
1	2	-.0157487*	.000
2	1	.0157487*	.000
	3	.0159803*	.000
	4	.0188277*	.000
3	2	-.0159803*	.000
4	2	-.0188277*	.000

*. The mean difference is significant at the .05 level.

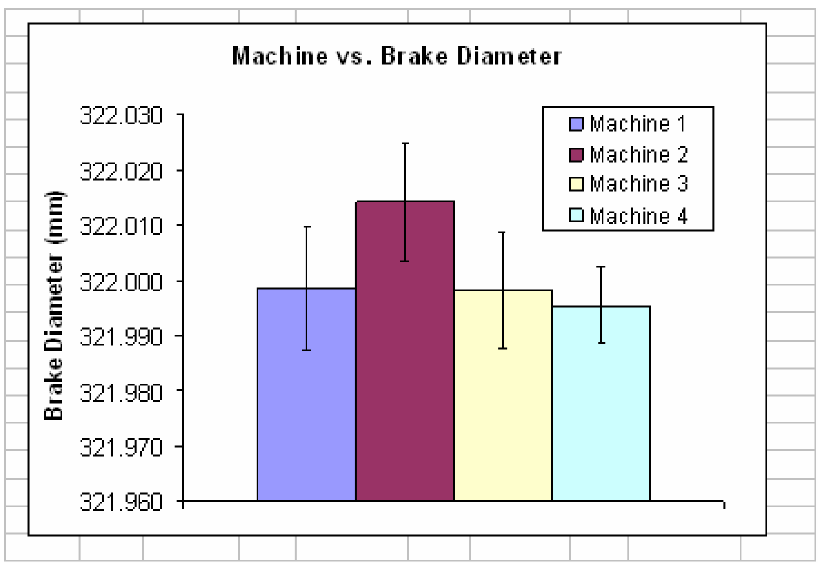
Decision Tree



Non-Parametric ANOVA Example II

Kruskal-Wallis

The Kruskal-Wallis test is a non-parametric alternative to one-way analysis of variance.



The test result (shown below) is highly significant. A post hoc test (multiple Mann-Whitney tests) would be done to determine which groups were different.

Test Statistics[a,b]

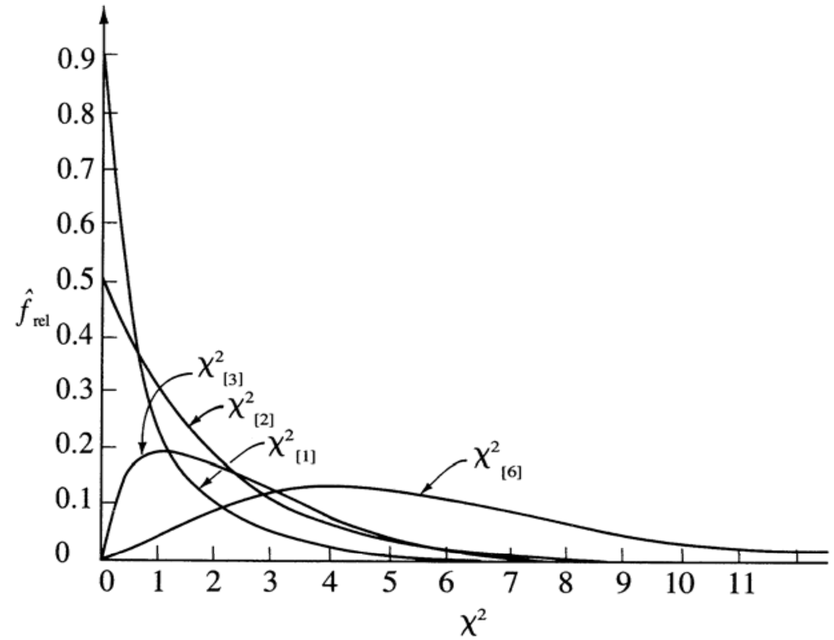
	Brake_Dia
Chi-Square	23.563
df	3
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: Machine

Chi-Square Test

- used with categorical data
- two variables and two groups on both variables
- results indicate whether the variables are related



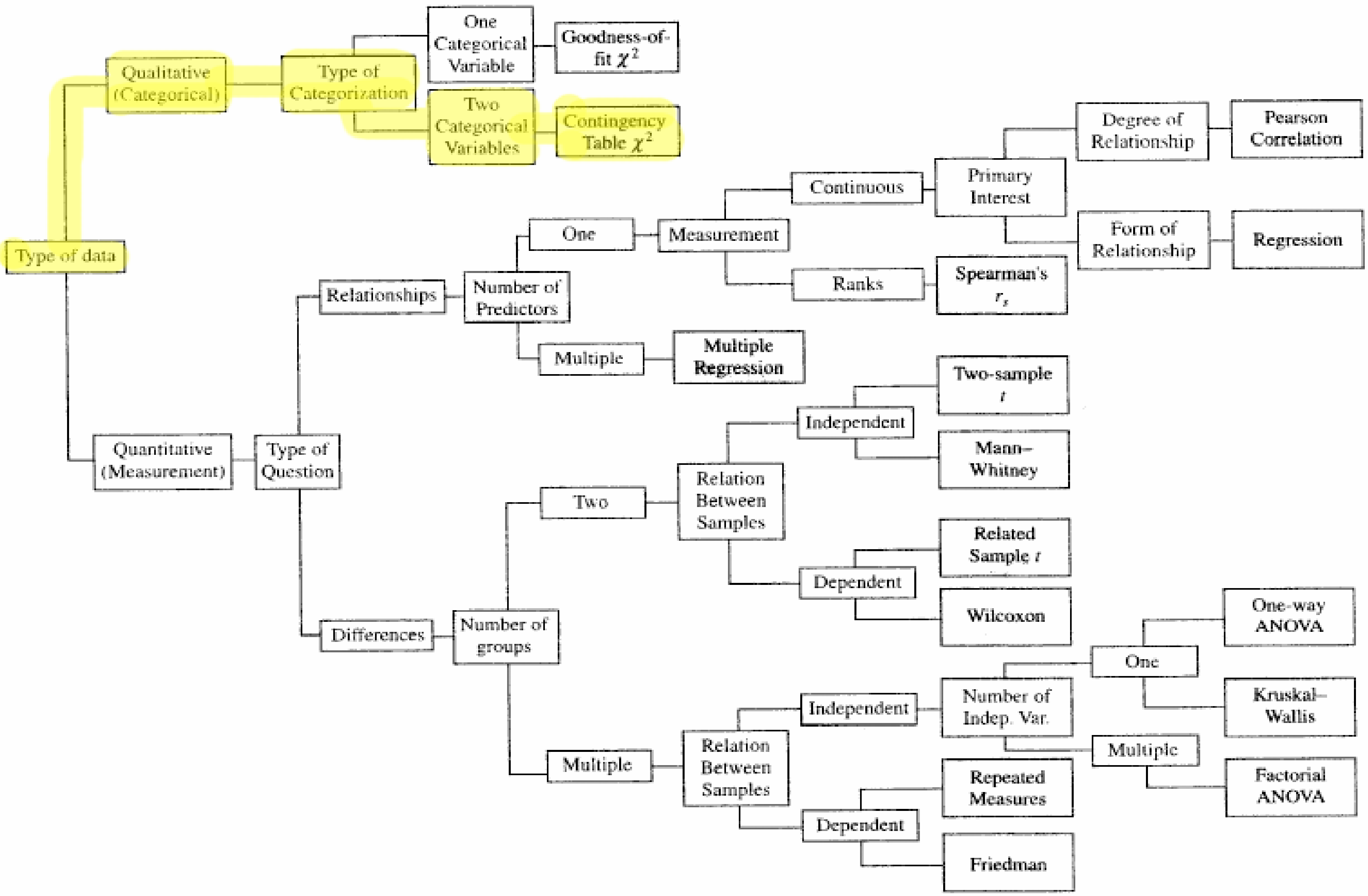
Frequency curves of χ^2-distribution for 1, 2, 3, and 6 degrees of freedom.

Figure from *Geigy Scientific Tables vol. 2*

Chi-Square Test

- Assumptions:
  - observations are independent
  - categories do not overlap
  - most expected counts > 5 and none < 1
- Sensitive to the number of observations
- Spurious significant results can occur for large n.

Decision Tree



Chi-Square Example

A 1991 U.S. general survey of 225 people asked whether they thought their most important problem in the last 12 months was health or finances.

Null hypothesis – Males and females will respond the same to the survey.

1 - males	1 - health
2 - females	2 - finances
2	1
2	1
2	1
2	1
...	...
2	2
1	2
2	2
2	2

Chi-Square Example

Cross-tabulation table shows how many people are in each category.

The non-significant result signifies that the null hypothesis is accepted.

problem * group Crosstabulation

Count

		group		Total
		Males	Females	
problem	Health	35	57	92
	Finance	56	77	133
Total		91	134	225

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.372[b]	1	.542		
Continuity Correction[a]	.223	1	.637		
Likelihood Ratio	.373	1	.541		
Fisher's Exact Test				.582	.319
Linear-by-Linear Association	.371	1	.543		
N of Valid Cases	225				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 37.21.

Chi-Square Example II

Chi-square test can be extended to multiple responses for two groups.

problem * group Crosstabulation

Count

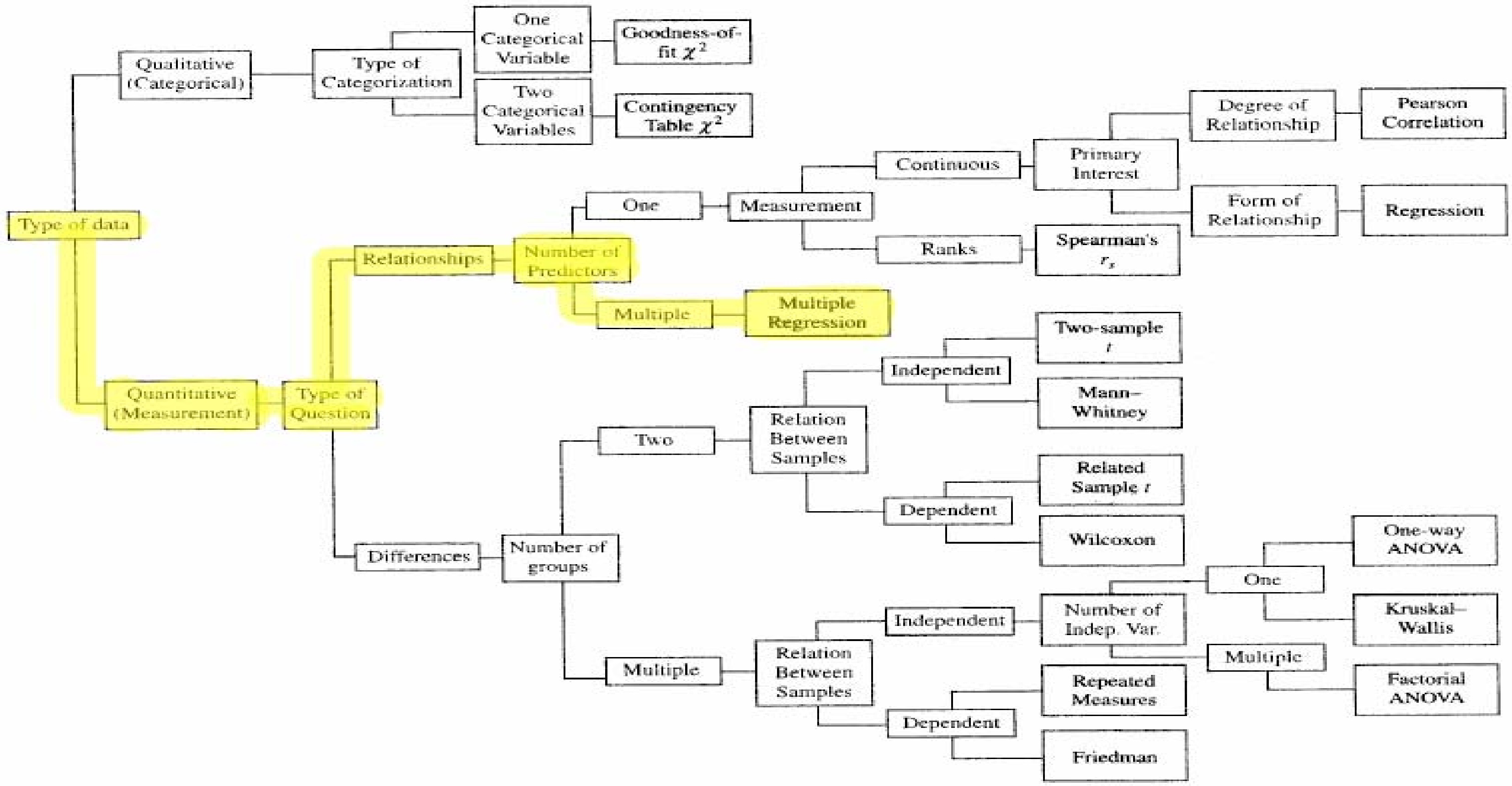
		group		Total
		Males	Females	
problem	Health	35	57	92
	Finance	56	77	133
	Family	15	33	48
	Personal	9	10	19
	Miscellaneou	15	25	40
Total		130	202	332

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Squa	2.377[a]	4	.667
Likelihood Ratio	2.400	4	.663
Linear-by-Linear Association	.021	1	.885
N of Valid Cases	332		

a. 0 cells (.0%) have expected count less than 5. T minimum expected count is 7.44.

Decision Tree



Howell, D.C. Fundamental Statistics for the Behavioral Sciences. Belmont, CA: Brooks/Cole, Thomson Learning, 2004.

from
Dr. Mann
201 Grosvenor

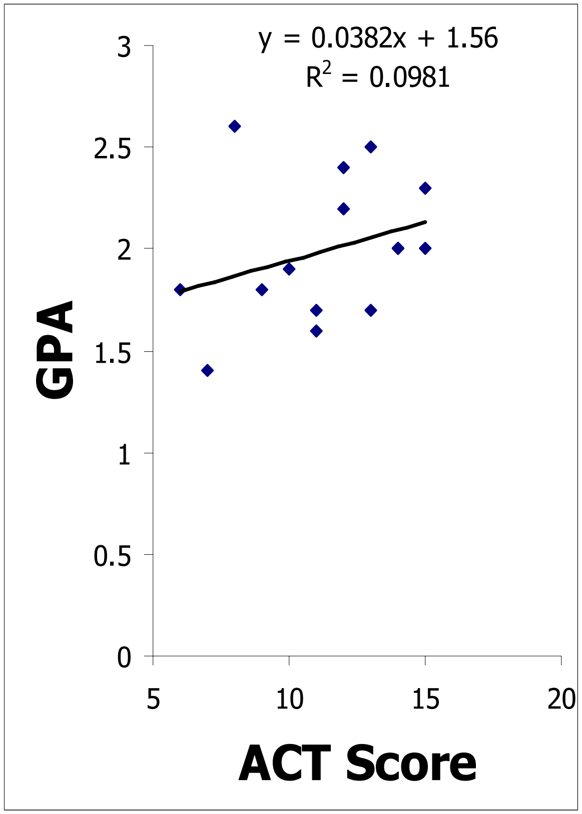
Multiple Regression

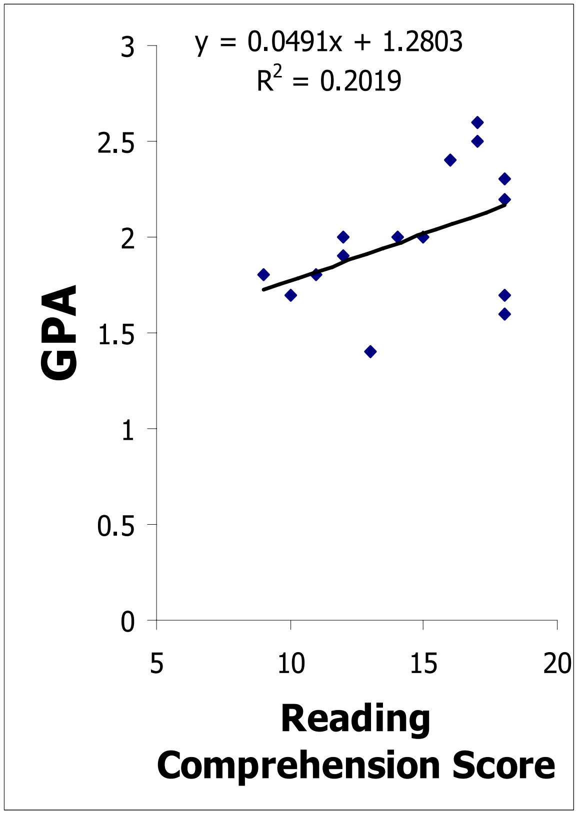
- Null Hypothesis – GPA at the end of Freshman year cannot be predicted by performance on college entrance exams.

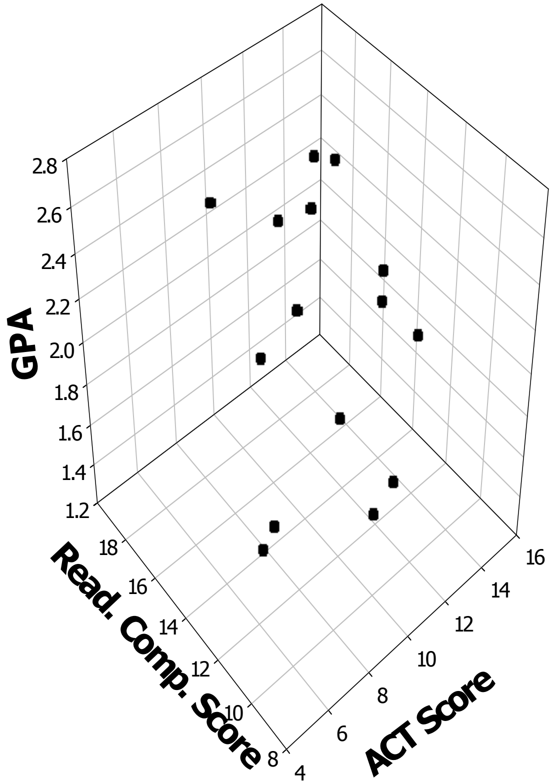
- GPA = α * ACT score + β * Read. Comp. score

	GPA	ACT	comp
1	2.20	12.00	18.00
2	1.40	7.00	13.00
3	1.80	9.00	9.00
4	1.60	11.00	18.00
5	2.50	13.00	17.00
6	1.90	10.00	12.00
7	2.00	14.00	12.00
8	2.40	12.00	16.00
9	2.60	8.00	17.00
10	1.80	6.00	11.00
11	1.70	13.00	18.00
12	2.00	15.00	15.00
13	1.70	11.00	10.00
14	2.00	14.00	14.00
15	2.30	15.00	18.00

Multiple Regression







Multiple Regression

The analysis shows no significant relationship between college entrance tests and GPA.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.373	2	.186	1.699	.224[a]
	Residual	1.317	12	.110		
	Total	1.689	14			

a. Predictors: (Constant), Reading Comprehension score, ACT score

b. Dependent Variable: Grade Point Average in first year of college

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.172	.460		2.551	.025
	ACT score	.018	.034	$\alpha =$ .151	.537	.601
	Reading Comprehension score	.042	.031	$\beta =$ .386	1.374	.195

a. Dependent Variable: Grade Point Average in first year of college

MANOVA

Multivariate **AN**alysis of **VA**riance (MANOVA)

MANOVA allows you to look at differences between variables as well as group differences.

- assumptions are the same as ANOVA
- additional condition of multivariate normality
- also assumes equal covariance matrices (standard deviations between variables should be similar).

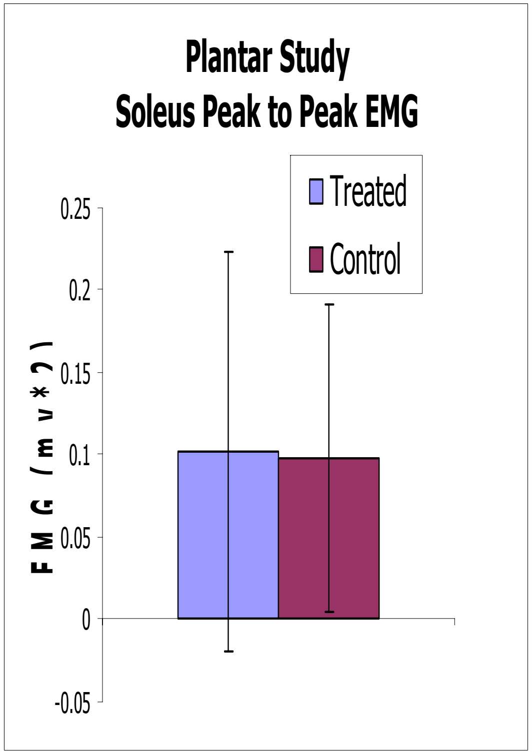
MANOVA Example

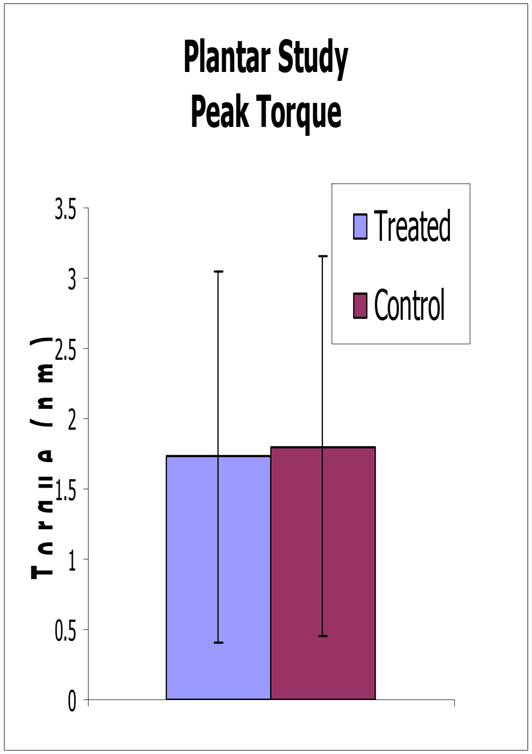
Subset of plantar fasciitis dataset.

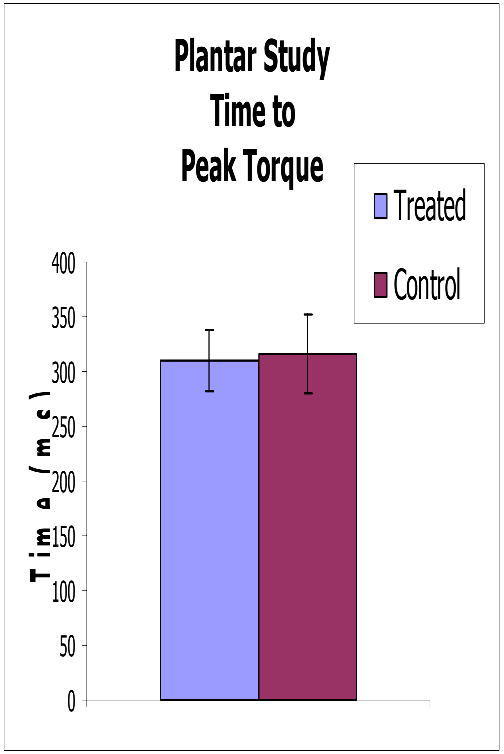
Null Hypothesis
There is no difference in soleus emg activity, peak torque, or time to peak torque for quick stretch measurements in people with plantar fasciitis who receive counterstrain treatment compared with the same group of people receiving a placebo treatment.

	Treatment Group 1 - Treated 2 - Control	Peak to Peak Soleus EMG Response millivolt*2	Peak Quick Stretch Torque newton-meter	Time to Peak Torque milliseconds
	1	0.0706	0.883	322.56
	1	0.0189	0.347	329.28
	1	0.0062	0.388	319.2
	1	0.0396	1.104	325.92
	1	0.0668	3.167	315.84
	1	0.2524	2.628	248.64
	1	0.0183	0.346	336
	1	0.0393	1.535	332.64
	1	0.1319	3.282	292.32
	1	0.3781	3.622	278.88
	2	0.039	0.557	299.04
	2	0.074	0.525	372.96
	2	0.0396	1.400	362.88
	2	0.0143	0.183	295.68
	2	0.076	3.074	322.56
	2	0.2213	3.073	258.72
	2	0.0196	0.271	346.08
	2	0.0498	2.278	302.4
	2	0.155	3.556	309.12
	2	0.2887	3.106	292.32
Treated				
Mean		0.10221	1.73017	310.128
StDev		0.12151083	1.31802532	28.15859087
Control				
Mean		0.09773	1.80212	316.176
StDev		0.09324085	1.35575411	35.22039409

MANOVA Example







MANOVA Results

Box's Test of Equality of Covariance Matrices[a]

Box's M	5.165
F	.703
df1	6
df2	2347.472
Sig.	.647

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept+group

Box's test checks for equal covariance matrices. A non-significant result means the assumption holds true.

Levene's Test of Equality of Error Variances[a]

	F	df1	df2	Sig.
Soleus peak to peak emg (mv*2)	.349	1	18	.562
Peak quick stretch torque (nm)	.078	1	18	.783
Time to peak torque (milliseconds)	.550	1	18	.468

Tests the null hypothesis that the error variance of the dependent v is equal across groups.

a. Design: Intercept+group

Levene's tests checks for univariate normality. A non-significant result means the assumption holds true.

MANOVA Results

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.996	1207.992[a]	3.000	16.000	.000
	Wilks' Lambda	.004	1207.992[a]	3.000	16.000	.000
	Hotelling's Trace	226.499	1207.992[a]	3.000	16.000	.000
	Roy's Largest Root	226.499	1207.992[a]	3.000	16.000	.000
group	Pillai's Trace	.020	.109[a]	3.000	16.000	.954
	Wilks' Lambda	.980	.109[a]	3.000	16.000	.954
	Hotelling's Trace	.020	.109[a]	3.000	16.000	.954
	Roy's Largest Root	.020	.109[a]	3.000	16.000	.954

a. Exact statistic

b. Design: Intercept+group

The non-significant group result indicates that the null hypothesis is true. If the result had been significant, you would need to do post hoc tests to find out which variables were significant.

References

Bruning, James L. and Kintz, B.L. *Computational handbook of statistics (4th Edition)*. Massachusetts: Addison Wesley Longman, Inc., 1997.

Research Resource Guide 3rd Edition 2004-2005, http://ohiocoreonline.org, go to Research->Research Education.

Norusis, Marija J. *SPSS 9.0 Guide to data analysis.* New Jersey: Prentice-Hall, Inc., 1999.

Sokal, Robert R. and Rohlf, James F. *Biometry (2nd Edition).* New York: W.H. Freeman, 1981.

Stevens, James. *Applied Multivariate Statistics for the Social Sciences*. New Jersey: Lawrence Erlbaum Associates, Inc., 1986.

Lentner, Cornelius. *Geigy Scientific Tables vol. 2.* Basle, Switzerland: Ciba-Geigy Limited, 1982.