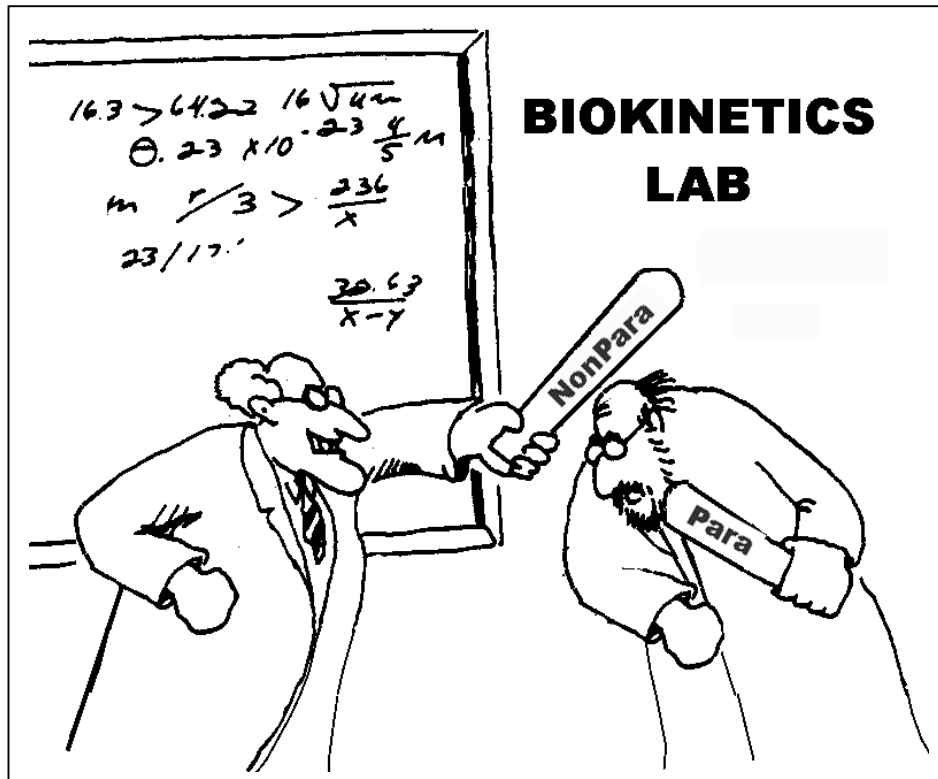


# Statistics 101: An Overview for the Gait & Movement Analysis Community (Parametric Statistics)

by

**Gregory S. Rash, EdD**



After hours of debate, Drs. Rash & Quesada finally decide on the best method for determining which statistical approach to use...

## Statistics 101: An Overview for the Gait & Movement Analysis Community (Parametric)

Gregory S. Rash, EdD

Research data can be obtained in a variety of formats. Each of these formats has unique characteristics that impact on the types of statistical techniques that can be appropriately applied. This tutorial is an overview of common statistics used for each of these types of data. It is not intended to give everyone all the tools necessary to handle all of their statistical needs, but to give them an overview of statistical options when dealing with the different types of data produced in a gait lab environment. By having a better understanding of the statistical options available in the planning stages, one is able to design a better study and have data that can be analyzed. This tutorial describe several non-parametric and parametric statistical techniques, as well as identifies common indications for their applications.

### Tutorial Outline

#### I. Types of data

- Nominal - No arithmetic relationship or order between different classifications. **Examples:** Occupation (Clerk, Police Officer, Teacher, ...); Gender (Male, Female)
- Ordinal - Data can be ordered into discrete categories, but categories have no arithmetic relationship. **Examples:** Survey Data (5 - Strongly Agree, 4 - Somewhat Agree, 3 - Neither Agree Nor Disagree, 2 - Somewhat Disagree, 1 - Strongly Disagree); Manual Muscle Test (5 - Full Range of Motion with Maximal Resistance, 4 - Full Range of Motion with Resistance, 3 - Full Range of Motion Against Gravity, 2 - Full Range of Motion without Gravity, 1 - Partial Range of Motion without Gravity, 0 - None or Trace Movement)
- Interval - Data on a measurement scale with an arbitrary zero point in which numerically equal intervals at different locations on the scale reflect the same quantitative difference. **Examples:** Temperature (Fahrenheit or Celsius)
- Ratio - Data on a measurement scale with an absolute zero point in which numerically equal intervals at different locations on the scale reflect the same quantitative difference. **Examples:** Height, Weight, Pressure, Temperature (Kelvins)

#### II. Types of Variables

- Independent – A variable that is manipulated (the treatment variable, the cause).
- Dependent – A variable that is measured (the outcome, the effect).
- Categorical – A classification variable that is analyzed (e.g. gender, race).
- Control – A characteristic that is restricted in the study, but not compared (e.g. only stroke pts).
- Extraneous – A variable that affects the dependent variable, but is not part of the design, is not controlled. (e.g. Amount of sleep).
- Confounding – When an extraneous variable is systematically related to the independent variable. (e.g. Vertical GRF & ankle compression force).
- Predictor – Another name for the independent variable in regression.
- Response - Another name for the dependent variable in regression. Sometimes called the Criterion.
- Dummy – Variables constructed to allow analysis within a specific models framework.
- Endogenous – Variables not affected by other variables in the study.
- Exogenous - Variables that are affected by other variables in the study.

#### III. Description

- A. Central Tendency
  1. Mean =  $\Sigma x/n$
  2. Median = Middle Score
  3. Mode = Most frequent score
- B. Variation
  1. Variance =  $(\Sigma x^2 - ((\Sigma x)^2/n))/n-1$
  2. Standard Deviation =  $\sqrt{(\Sigma x^2 - ((\Sigma x)^2/n))/n-1}$
  3. Minimum = Lowest Score

4. Maximum = Highest Score
5. Range = Lowest Score to Highest Score
6. Quartiles = Used in sales and survey data to divide populations into groups [0% (minimum), 25%, 50% (median), 75%, 100%(maximum)].

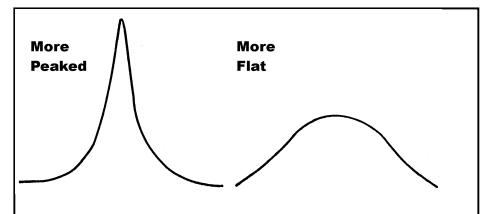
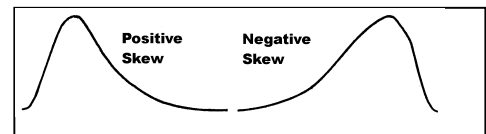
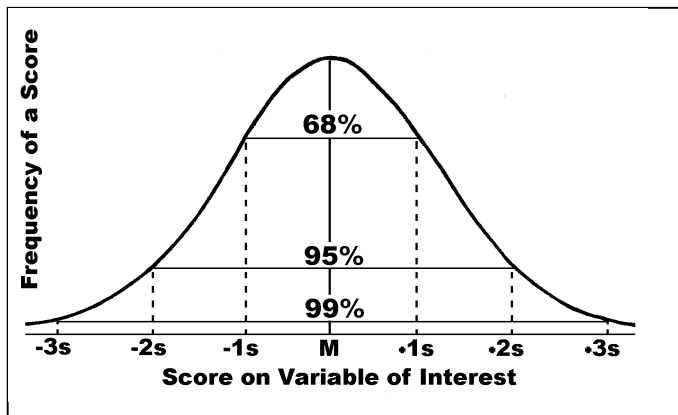
**Example:** Scores of x: 85, 70, 82, 75, 83, 81, 79, 86, 74, 77, 71, 73

- List in order
- 70
  - 71 Range 70-86
  - 73 Mode = No mode, or all are mode
  - 74 Minimum = 70
  - 75 25% = 73.75
  - 77 Median = 78
  - 79 Add a 90 to end & median would be 79
  - 81 75% = 82.25
  - 82 Maximum = 86
  - 83
  - 85
  - 86

x	x <sup>2</sup>	
85	7225	For Mean: $\Sigma x = 936, n = 12$
70	4900	$936/12 = 78$
82	6724	
75	5625	$s_x = \sqrt{((\Sigma x^2 - ((\Sigma x)^2/n))/n-1)}$
83	6889	$s_x = \sqrt{(73336 - ((936)^2/12))/12-1)}$
81	6561	$s_x = \sqrt{(73336 - (876096/12))/11)}$
79	6241	$s_x = \sqrt{(73336 - 73008)/11)}$
86	7396	$s_x = \sqrt{328/11}$
74	5476	$s_x = \sqrt{29.82}$ ← variance
77	5929	$s_x = 5.46$
71	5041	
73	5329	

$\Sigma x = 936$      $\Sigma x^2 = 73336$

III. **Parametric Statistics** - These methods are based on the assumptions that the parameters, mean and standard deviation describe a normally distributed population. In normal curve → Mean, median, mode are the same and  $\pm 1$  SD = 68% of scores,  $\pm 2$  SD = 95% of scores, &  $\pm 3$  SD = 99% of scores. However, these assumptions are robust. Can look at skewness (direction of tail) and kurtosis (height of bump) to verify normality. However, most of the parametric statistics are robust to departures from normality, although the data should be symmetric.



A. **Correlation** - Assessment of linear association between two or more variables (Pearson Product-Moment). **Association NOT Causation!!!**

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2) * (n\sum y^2 - (\sum y)^2)}}$$

Where: x = Independent variable  
y = Dependent variable  
n = Pairs of scores

Use the x scores from before & the scores below for y

y	y <sup>2</sup>	xy	
90	8100	7650	$r = \frac{(12*82710)-(936)(1058)}{\sqrt{(12*73336)-(936)^2} * \sqrt{(12*93466)-(1058)^2}}$
85	7225	5950	
90	8100	7380	$r = \frac{(992520)-(990288)}{\sqrt{(880032-876096)} * \sqrt{(1121592-1119364)}}$
85	7225	6375	
97	9409	8051	
90	8100	7290	
90	8100	7110	
92	8464	7912	$r = \frac{2232}{\sqrt{(3936)} * \sqrt{(2228)}}$
87	7569	6438	
82	6724	6314	
85	7225	6035	$r = \frac{2232}{62.74 * 47.20}$
85	7225	6205	$r = \frac{2232}{2961.43}$

$r = 0.753$

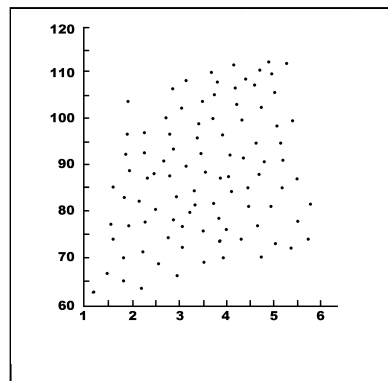
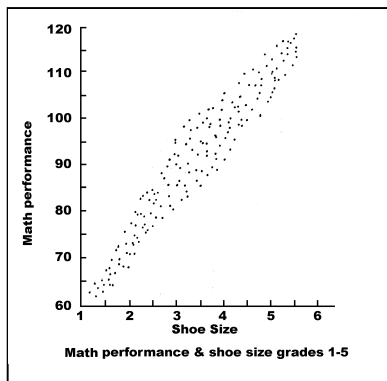
$\sum y = 1058$     $\sum y^2 = 93466$     $\sum xy = 82710$

Crit  $r_{22df, \alpha=0.01} \approx 0.487$ , thus significant. However, a significant correlation is not a big deal!!! All significance means is that the slope is not zero. Look at the Coefficient of Determination or  $r^2 = 0.753^2$     $r^2 = 0.567$

The descriptive meaning is “the variable y explains 56.7% of the variance in variable x.

With a df of 100 a r of 0.1638 is significant →  $r^2 = 0.026$ . Only explains 2.6% of the variability...

**Partial Correlation:** The correlation between the dependent variable and an independent variable when the linear effects of the other independent variables in the model have been removed from both. Craig et al found a correlation of 0.91 when looking at the relationship of math achievement with shoe size in grades 1-5. However, when done as a partial correlation holding age constant ( $r_{12 \cdot 3}$ ), he found  $r = 0.36$ .



B. **Regression** - Prediction of dependent variables based on correlations with independent variables (Linear, Non-linear, Logistic, Multiple). Goal is to develop a regression equation in the form:

$$Y = a + bx \text{ or from algebra } y = mx + b$$

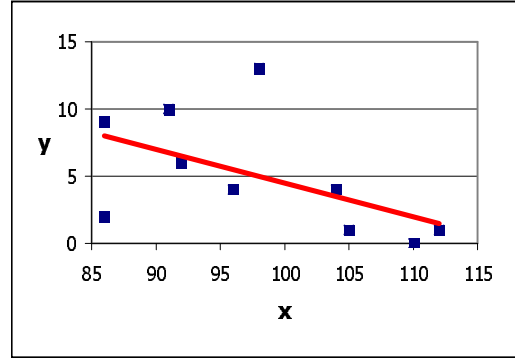
Where: a = Intercept, b = Slope, y = Predicted Score, x = Predictor variable

Use least squares fit to find b →  $b = r (s_y/s_x)$

Where:  $r$  = Correlation     $s_y$  = SD of  $y$      $s_x$  = SD of  $x$   
 $a$  = is value of  $y$  when  $x$  = zero ( $a = y - bx$ )

**Example:** Try to use body weight as a predictor of the number of pull-ups for grades 1-6.

$x$  = (Body wt)     $y$  = (pull-ups)  
 $x_{bar} = 98$      $y_{bar} = 5$   
 $s_x = 9.44$      $s_y = 4.40$   
 $r = -0.54$      $r^2 = 0.29$   
 $b = r (s_y/s_x) \rightarrow -0.54 * (4.4/9.44) \rightarrow b = -0.252$   
 $a = y_{bar} - bx_{bar} \rightarrow 5 - (-.252 * 98) \rightarrow a = 29.696$



Prediction equation  $\rightarrow y = 29.696 - 0.252x$

If you  $\uparrow r \rightarrow \downarrow$  error of prediction.  
 If you  $\downarrow$  SD of criterion  $\rightarrow \downarrow$  error.

If  $r \neq 1$  or  $-1$ , then get best fit & will have ERROR. Look at Coefficient of Determination or  $r^2$  as indicator. Interpret like  $r^2$  for correlation.

Thus if a child weighs 100 lbs, the equation will predict that they can do 4 pull-ups & the equation accounts for 29% of the variance. Crit  $r_{18df, \alpha=0.05}=0.444$ , thus significant, but not a great predictor...

**Multiple Regression:**     $y' = a + b_1 x_1 + b_2 x_2 + \dots b_n x_n$

Two or more predictor variables (independent) & one dependent variable. Using more Predictors  $\uparrow$  accuracy (the multiple  $r^2$  will always go up). Many ways to calculate which variables are the best to keep: stepwise is most common, but none of them work better than a through knowledge of the subject matter & understanding how the variables are related and interact.

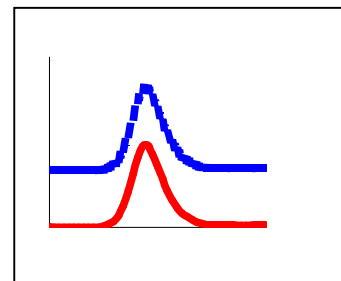
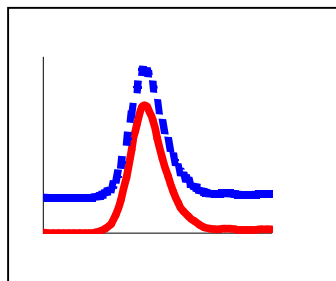
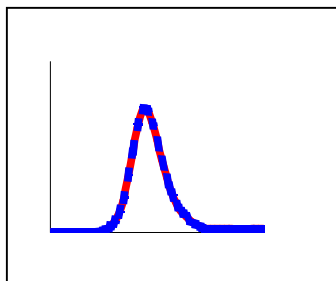
Good example of a non linear multiple regression equation is a body density equation or a %fat skinfold prediction equation:

$$\text{Body Density} = 1.112 - 0.0004349 * (\sum 7 \text{ skinfolds}) + 0.00000055 * (\sum 7 \text{ skinfolds})^2 - 0.000288 * (\text{age})$$

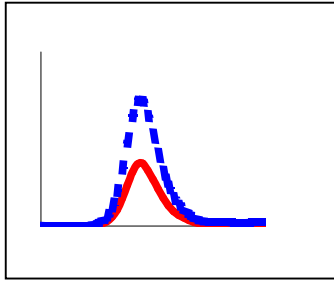
$$\% \text{fat} = 0.2928 * (\sum 4 \text{ skinfolds}) - 0.0005 * (\sum 4 \text{ skinfolds})^2 + 0.15845 * (\text{age})$$

C. **Time Series Data (Comparison of Waveforms)** - Time-series waveforms typically cannot be analyzed by standard statistical methods. Therefore, a statistical measure to compare waveforms is useful in our profession. The methods that one sees used most are the Pearson product moment correlation coefficient, the Concordance Correlation Coefficient (CCC), Coefficient of multiple Correlation (CMC) & Intra Class Correlation (ICC). There are a couple of others that mathematician's say may work better, but I've not seen them applied to our type data in the literature. My preference is the CMC, also known as the adjusted coefficient of multiple determination ( $R_a^2$ ). A  $R_a^2$  of 1 indicates that the waveforms are identical, whereas a  $R_a^2$  of 0 denotes complete dissimilarity.

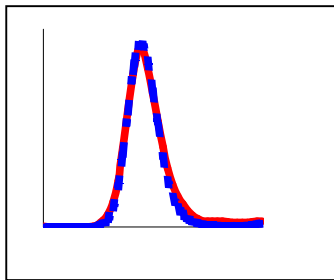
Identical  $\rightarrow$   $CMC^2 = 1$      $+2 \rightarrow CMC^2 = 0.6423$      $+5 \rightarrow CMC^2 = 0.1467$   
 $ICC^2 = 1$      $ICC^2 = 1$      $ICC^2 = 1$



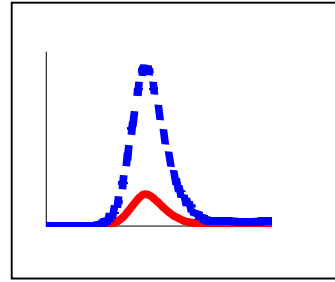
X2 →  $CMC^2 = 0.7306$   
 $ICC^2 = 0.7901$



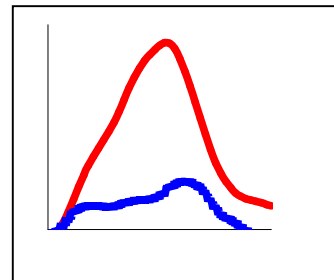
**Normal Patient Bilateral Eye**  
 Blink →  $CMD = 0.9955$   
 $ICC^2 = 0.9970$



X5 →  $CMC^2 = 0.2351$   
 $ICC^2 = 0.3087$



**Bell's Palsy Bilateral Eye**  
 Blink →  $CMD = 0.0681$   
 $ICC^2 = 0.3672$



In the review process for the Rash et al., 1999 paper a reviewer suggested we use ICC instead of CMC. Their reasoning was that the CMC was not normally distributed (I've not seen this in print). I reanalyzed all the data using ICC and found no difference in the comparisons except at the 3<sup>rd</sup> or 4<sup>th</sup> decimal place. It should be noted that all the curves in this particular paper had CMD values above 0.90. The paper was published using CMC's. One can see by the comparisons above that the ICC can't handle DC shifted data, although it does pick up the differences with scaled data. The ICC yielded a higher value in all cases so one might deduce that the ICC is more liberal than the CMC. Are any of these the best for time series? I must apply a quote from Thomas Payne: "Things are not necessarily right, just no one has taken the time to show it's wrong."

#### D. Inferential Statistics

**Errors In Inferential Testing:** There are two types of errors that can be made when testing a hypothesis. These are:

- Type I errors ( $\alpha$ ) - Rejecting the null hypothesis when in fact it is true.
- Type II errors ( $\beta$ ) - Accepting the null hypothesis when in fact it should have been rejected.

Type I ( $\alpha$ ) errors are typically caused by two main problems: **Small sample sizes** or **too many analyses**.

One of the most common problems leading to Type I errors is **small sample size**. The best way to demonstrate this is to use baseball hitting as an analogy. At the beginning of the year there may be a great number of batters who are batting either way above or way below their normal average. In fact, in the early season there may be several .400 hitters.

In the early season players have had only a few "at bats". If they have been "hot" early, they may have a very high batting average; if they've been "cold" early, they may have a very low average. On the basis of the early season averages one might be tempted to say that a particular batter is hitting "significantly" higher or lower than his average in the previous year. In fact, as the year goes along most of these extremes in averages will even out and .300 hitters will be hitting around .300, the .200

hitters will be hitting around .200s, etc. The problem that led one to believe that a batter had significantly changed was a low sample size, i.e., too few times at bat. As the season wears on and the times at bat rise, this sampling error is corrected and the results are no longer significant.

Other contributors to Type I errors are variations on the theme of **too many analyses**: analyzing too many variables, too many subgroups, or analyzing the data too often.

Many times we try to look at too many variables at the same time. Assume that we are looking at the difference between a group that gets surgery and another that receives non-surgical options. By the time we include all the kinematic, kinetic and ADL variables that we feel are important, we've obtained a very large list of variables. The more variables we look at, the greater the chance that one will show significance, when in fact the differences are may be due to sampling variations. Thus, the more variables we look at, the greater the chance of committing a Type I error.

A related problem is when we have **too many subgroups**. Let's say we are comparing the surgery and non-surgery groups again. We may not find differences in our patients taken as a group. However, if we start dividing them up into subgroups we may find differences in a particular subgroup. Yet this may not be a true, reproducible difference. It may have been found because of sampling variation and analysis of many subgroups. As we increase the number of subgroups, we increase the chance that one of the subgroups will show a Type I error.

Be aware that some studies look at large numbers of variables and many subgroups, and show still some legitimate significant differences. If the sample size involved in the variable or subgroup population is sufficient, the significance of the study may be real. If the subgroup is too small, the difference has a greater chance of being a Type I error. It may require follow-up studies of the specific subgroups to make this determination.

Another related cause of Type I errors is **analysis that is too frequent**. When the data are analyzed too often, what frequently happens is that the study is stopped as soon as significance is found. Remember that, like a batting average, the result of an experimental study may fluctuate greatly with small sample size. If one analyzes a study frequently and stops as soon as it reaches significance, the chances of the results being a Type I error are increased. It is important that sample size be determined prior to a study and the entire study carried out before any meaningful analysis is attempted. Although batting averages are calculated daily, if we stop calculating batting averages when a hitter is significantly different from past years, we may have a Type I error.

Type II ( $\beta$ ) errors also occur primarily because of 2 main problems: **Small sample size & wide variability of the data**.

As with Type I errors, Type II (beta) errors can occur because of small sample size. In fact, small sample size is the most common cause of Type II errors. Also, as with Type I errors, the larger the sample size is, the less chance there is of a Type II error.

Another aspect of small sample size has to do with the concept of statistical difference and clinical importance. Let's say a study showed a mortality rate of 50% with traditional therapy, which was reduced to 40% with a new therapy. Clinically, this reduction of 10% is an important reduction in mortality. However, due to the small size of the sample, a statistically significant difference was not found. In this example the difference was clinically important, but because of small sample size, was not found to be statistically significant. This points out the importance of assuring that the sample size of a study is large enough to show differences we feel are clinically significant. In addition to small sample size, Type II errors can be caused by wide variability in the data.

A new surgical technique reduced post surgical therapy time from a mean of 60 days with conventional technique to a mean of 45 days with the new technique. This is a 20% reduction. However, the range of data for the conventional group was 25-95 days and 20-75 days for the new technique. This wide variance in the data may eliminate any chance of finding statistical significance. Wide variability of data is most often the result of a heterogeneous population or error in technique.

Because some error will always creep into any study, we specify the amount of error that we are willing to tolerate. For Type I ( $\alpha$ ) error, it is generally considered that a 5% error is acceptable. This means that we're willing to accept the idea that our claim of a statistically significant difference will be wrong 5% of the time. This is not necessarily our choice, but dictated by what the journals are willing to publish.

For Type II ( $\beta$ ) error, in medicine it is generally considered that an error of between 10% - 20% is acceptable. What this means is that if you report no difference between a treatment group and the control group, there is only a 10%- 20% chance that you would have missed an improvement of some specific difference.

**Power:** Closely related to the Type II errors and relates to the studies ability to detect differences. (e.g. a study might be "powerful" enough to detect a difference of 30 percent change in ROM, but not powerful enough to detect a difference of 20 percent. The power of a study is calculated by the formula: **Power = 1 -  $\beta$  error** (e.g. with  $\beta$  error of 10%, the power of a study = 1 - 10% = 90%).

Suppose we have a study in which we are looking for a 25% improvement: however, our study does not find it. Our  $\beta$  error was 20%: thus, the power of our study was 80%. This means that if a 25% improvement actually did exist, we would have detected it 80 times out of 100. Therefore, the study had acceptable power (80%) to detect a 25% improvement, but was not of adequate power to detect a lesser difference, say 20% or 10%. Although in our example the null hypothesis was not rejected, one can't rule out the possibility that a 20% or 10% change could have occurred and they just missed it because of a 20%.

A significant point here is that a negative study becomes much more meaningful with an understanding of the power of the study. This approach also gives more meaning to clinical significance over statistical significance. A 10% improvement may be very important clinically, but it may have failed to show statistical significance.

We all want our conclusions to be correct and reproducible by other investigators. The most important things you can do to reduce Type I and Type II errors are:

- Use respectable sample size (small sample size leads to both Type I and Type II errors).
- Don't overload your study with comparisons or subgroups and don't take too many "peeks" at the data.
- Design your study to obtain "clean" data with as little variability as possible

**t-test:** (Independent or Dependent)

Independent: Compares means for two groups tested on the same independent variable. Ideally, for this test, the subjects should be randomly assigned to two groups, so that any difference in response is due to the treatment (or lack of treatment) and not to other factors.

**Example:** We can calculate the independent "t" statistic to test the  $H_0$  that there is no significant difference between x & y for the data from pages 3 & 4. Use  $\alpha = 0.05$ .

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(s_x^2/n) + (s_y^2/n)}} = \frac{78 - 88.2}{\sqrt{5.46^2/12 + 4.11^2/12}} = \frac{-10.2}{\sqrt{29.8/12 + 16.9/12}}$$

$$t = \frac{-10.2}{\sqrt{2.48 + 1.41}} = \frac{-10.2}{\sqrt{3.89}} = \frac{-10.2}{1.97} = \boxed{-5.18}$$

$$df=(n_1+n_2)-2 \rightarrow df=(12+12)-2 \rightarrow df=22$$

Critical  $t_{(22)} @ .05 = 2.074$ . Thus reject  $H_0$  & accept  $H_a$  that they are significantly different.

Dependent: The two groups are related in one of 2 general relations. Most commonly it compares the means of two variables for a single group (Pre/Post test). It computes the differences between values of the two variables for each case and tests whether the average differs from 0. However, it is also often used when groups are matched and felt to not be independent.

**Example:** We can calculate the dependent "t" statistic to test the  $H_0$  that there is no significant difference between x (pre test) & y (posttest). Use  $\alpha = 0.01$ . First calculate the difference (D) &  $D^2$ .

D	$D^2$			
-5	25			
-5	25			
-8	64			
-10	100			
-14	196			
-9	81			
-11	121			
-6	36			
-13	169			
-5	25			
-14	196			
-12	144			
$\Sigma$ -112	1182			

	$\frac{\Sigma D}{n}$	$\frac{-112}{12}$		
$t = \sqrt{\frac{(n \Sigma D^2 - (\Sigma D)^2)}{(n-1)}}$			$= \sqrt{\frac{(12 * 1182 - (-112)^2)}{(12-1)}}$	
	$\frac{-112}{11}$		$= \sqrt{\frac{1640}{11}}$	$= \sqrt{149.09}$
	$t = 12.21$		$t = -9.17$	
	$df = n_{pairs} - 1$	$\rightarrow$	$df = 12 - 1$	$\rightarrow$
			$df = 11$	

Critical  $t_{(11)} @ .01 = 3.106$ . Therefore reject  $H_0$  & accept  $H_a$  that they are significantly different

**ANOVA:** An extension of independent t test or the t-test is a special case of ANOVA. It is a method of testing the null hypothesis that several group means are equal in the population, by comparing the sample variance estimated from the group means to that estimated within the groups. Uses F statistic & only tells us that there is significance somewhere in the model. If we find an overall significance, we need to then find between which means. Often used follow up tests: Scheffe (most conservative), LSD, Tukey's, Newman-Keuls, Duncan (most liberal), etc. All try to keep the  $\alpha$  at its original value, if not the  $\alpha$  degrades with each comparison [e.g.  $(\alpha_r = 1 - (1 - \alpha)^r)$  using  $\alpha = 0.05$  & 10 comparisons without correction  $\alpha_{10} = 0.4$ . Thus 4 out of the 10 comparisons have probability of committing Type I errors]. Planned comparisons are best (a priori)  $\rightarrow$  must have idea of comparisons at the beginning.

**Example:** Using the ABC method we will construct a one way ANOVA table to test the  $H_0$  that there is no significant difference between x, y (from pg 3 & 4) and z (Use  $\alpha = 0.01$ ). Use the following 12 scores as a third variable, the z variable (87, 78, 85, 78, 90, 85, 85, 79, 80, 77, 78, 90. Sum of Z = 992, sum of  $Z^2 = 82266$ ).

$A = \Sigma X^2$      $A = 73336 + 93466 + 82266$      $A = 249068$   
 $B = (\Sigma X)^2 / N$      $B = (936 + 1058 + 992)^2 / 36$      $B = 2986^2 / 36$      $B = 8916196 / 36$      $B = 247672.11$   
 $C = (\Sigma X_1)^2 / n_1 + (\Sigma X_2)^2 / n_2 + (\Sigma X_3)^2 / n_3$      $C = (936^2 / 12) + (1058^2 / 12) + (992^2 / 12)$   
 $C = (876096 / 12) + (1119364 / 12) + (984064 / 12)$   
 $C = (73008) + (93280.3) + (82005.3)$      $C = 248293.6$

Source	SS	df	MS	F
Between (True)	C-B	k-1	$SS_B / df_B$	$MS_B / MS_W$
Within (Error)	A-C	N-k	$SS_W / df_W$	
Total	A-B	N-1		

Where: X=subjects score    N=total # subjects    n=# subjects in group    k=#groups  
 SS= Sum of Squares    MS=Mean Square    df=Degrees of Freedom

Source	SS	df	MS	F
Between (True)	248293.6-247672.11	3-1	$SS_B / df_B$	$MS_B / MS_W$
Within (Error)	249068-248293.6	36-3	$SS_W / df_W$	
Total	249068-247672.11	36-1		

Source	SS	df	MS	F
Between (True)	621.49	2	-621.49/2	$MS_B / MS_W$
Within (Error)	774.4	33	774.4/33	
Total	1395.89	35		

Source	SS	df	MS	F
Between (True)	621.49	2	310.75	$MS_B/MS_W$
Within (Error)	774.4	33	23.46	
Total	1395.89	35		

Source	SS	df	MS	F
Between (True)	621.49	2	310.75	13.25*
Within (Error)	774.4	33	23.46	
Total	1395.89	35		

Critical  $F_{(2, 33)} @ .01 = 6.78$  Therefore, reject  $H_0$  & accept  $H_a$  that there is at least one group significantly different than one of the others. We need a follow up test to find out which means are different.

**Factorial Design:** When you manipulate more than one independent variable and evaluate the effects on a single dependent variable. We can look at the main effects & interactions.

**Example:** Using the ABC method & the data to the right, we will construct a two way (2X3) ANOVA table to test the  $H_0$  that there is no significant difference between frequency of exercise (2days/week vs 3 days/wk) & intensity of exercise (40%, 60% or 80% max HR). We'll use the  $\alpha = 0.05$ .

**Intensity of Exercise**

		Frequency of Exercise				
		2 days/week		3 days/week		
		$X_{r1c1}$	$X_{r1c1}^2$	$X_{r1c2}$	$X_{r1c2}^2$	
40%		2940	8643600	2980	8880400	$X_{r1}=30285$
		3070	9424900	3160	9985600	
		3100	9610000	3025	9150625	
		2925	8555625	3045	9272025	
		3050	9302500	2990	8940100	
60%		$X_{r2c1}$	$X_{r2c1}^2$	$X_{r2c2}$	$X_{r2c2}^2$	$X_{r2}=33510$
		3150	9922500	3720	13838400	
		3020	9120400	3630	13176900	
		2990	8940100	3570	12744900	
		3050	9302500	3690	13616100	
80%		$X_{r3c1}$	$X_{r3c1}^2$	$X_{r3c2}$	$X_{r3c2}^2$	$X_{r3}=35620$
		3170	10048900	3920	15366400	
		3120	9734400	4040	16321600	
		3050	9302500	4110	16892100	
		3110	9672100	4005	16040025	
	3105	9641025	3990	15920100		
		$X_{c1}=45830$		$X_{c2}=53585$		

$A = \sum X^2 = 8643600 + 9424900 + \dots + 15920100 = \boxed{A=249068}$

$B = (\sum X) / N = (2940 + 3070 + 3100 + \dots + 3990) / 30 = 99415^2 / 36 = 9883342225 / 30 = \boxed{B=329444741}$

$C = [(\sum X_{r1})^2 + (\sum X_{r2})^2 + (\sum X_{r3})^2] / n_{r1} = (30285^2) + (33510^2) + (35620^2)$

$C = (917181225) + (1122920100) + (1268784400) / 10 = 3308885725 / 10 = \boxed{C=330888572.5}$

$D = (\sum X_{c1})^2 + (\sum X_{c2})^2 / n_{c1} = [45830^2 + 53585^2] / 15 = [2100388900 + 2871352225] / 15$

$D = 4971741125 / 15 = \boxed{D=331449408}$

$E = [(\sum X_{r1c1})^2 + (\sum X_{r1c2})^2 + \dots + (\sum X_{r3c2})^2] / n_{r1c1} = [15085^2 + 15200^2 + \dots + 20065^2] / 5$

$E = [227557225 + 231040000 + \dots + 402604225] / 5 = 1669517975 / 5 = \boxed{E=333903595}$

Source	SS	df	MS	F
Row (Intensity)	C-B	r-1	$SS_r/df_r$	$MS_r/MS_e$
Columns (Freq)	D-B	c-1	$SS_c/df_c$	$MS_c/MS_e$
R X C	(E-B)-(C-B)+(D-B)	(r-1)(c-1)	$SS_{rc}/df_{rc}$	$MS_{rc}/MS_e$
Error	(A-B)-(E-B)	(N-1)-(r-1)+(c-1)+ (r-1)(c-1)	$SS_e/df_e$	
Total	A-B	N-1		

Where: X=subjects score N=total # subjects n=# subjects in group r=#rows c=#columns  
 SS= Sum of Squares MS=Mean Square df=Degrees of Freedom

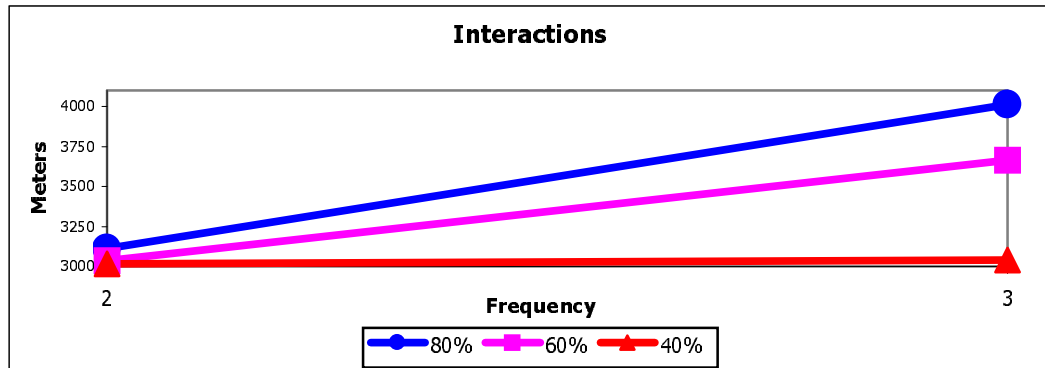
Source	SS	df	MS	F
Row (Intensity)	330888573-329444741	3-1	$SS_r/df_r$	$MS_r/MS_e$
Columns (Freq)	331449408-329444741	2-1	$SS_c/df_c$	$MS_c/MS_e$
R X C	(333903595-329444741)- ( 330888573-329444741)+ ( 331449408-329444741)	(3-1)(2-1)	$SS_{rc}/df_{rc}$	$MS_{rc}/MS_e$
Error	(249068-329444741)- ( 333903595-329444741)	(30-1)-(3-1)+ (2-1)+ (3-1)(2-1)	$SS_e/df_e$	
Total	249068-329444741	30-1		

Source	SS	df	MS	F
Row (Intensity)	1443832	2	1443832/2	$MS_r/MS_e$
Columns (Freq)	2001667	1	2001667/1	$MS_c/MS_e$
R X C	1010355	2	1010355/2	$MS_{rc}/MS_e$
Error	107230	24	107230 /24	
Total	4566084	29		

Source	SS	df	MS	F
Row (Intensity)	1443832	2	721916	721916/4468
Columns (Freq)	2001667	1	2001667	2001667/4468
R X C	1010355	2	505177	505177/4468
Error	107230	24	4468	
Total	4566084	29		

Source	SS	df	MS	F
Row (Intensity)	1443832	2	721916	161.57*
Columns (Freq)	2001667	1	2001667	448.00*
R X C	1010355	2	505177	113.07*
Error	107230	24	4468	
Total	4566084	29		

Start from bottom up in a factorial design. For Interactions, Critical  $F_{(1, 24)} @ .05 = 3.40$  Therefore reject  $H_0$  & accept  $H_a$  that there is a significant difference with respect to interactions. Need a follow up test to determine which. Since the interactions are significant, we ignore the main effects even though significant (Intensity, Critical  $F_{(2, 24)} @ .05 = 3.40$  & Frequency, Critical  $F_{(1, 24)} @ .05 = 4.26$ ).



Repeated Measure Design: When you take multiple measurements of the dependent variable, typically when looking at the affect of the independent on the dependent variable over time.

**Example:** Using the ABC method & the data to the right, we will

Subjects	$X_0$	$X_0^2$	$X_w$	$X_w^2$	$X_m$	$X_m^2$
1	5	25	7	49	9	81
2	3	9	4	16	4	16
3	4	16	6	36	7	49
4	5	25	5	25	7	49
5	6	36	8	64	9	81
$\Sigma$	23	111	30	190	36	276
$X_{bar}$	4.6		6.0		7.2	

construct a Repeated Measure ANOVA table to test the  $H_0$  that there is no significant difference between subjects or time post Botox injection (same day, 1 week, 1 month) with respect to the amount of dorsi flexion measured at the ankle. We'll use the  $\alpha = 0.05$ .

$$A = \sum x^2 = 111 + 190 + 276 \quad \boxed{A=557}$$

$$B = (\sum x)^2 / N = (23 + 30 + 36)^2 / 15 = 89^2 / 15 = 7921 / 15 \quad \boxed{B=528.07}$$

$$C = [(\sum x_{r1})^2 + (\sum x_{r2})^2 + \dots + (\sum x_{r5})^2] / n_{r1} = [21^2 + 11^2 + \dots + 23^2] / 3 = [441 + 121 + \dots + 529] / 3 = 1669 / 3 \quad \boxed{C=556.33}$$

$$D = (\sum x_{c1})^2 + (\sum x_{c2})^2 + (\sum x_{c3})^2 / n_{c1} = [23^2 + 30^2 + 36^2] / 5 = [529 + 900 + 1296] / 5 = 2725 / 5 \quad \boxed{D=545}$$

Source	SS	df	MS	F
Row (Subjects)	C-B	r-1	$SS_r / df_s$	$MS_s / MS_r$
Columns (Time)	D-B	c-1	$SS_t / df_t$	$MS_t / MS_r$
Residual	(A-B)-(C-B)+(D-B)	(r-1)(c-1)	$SS_r / df_r$	
Total	A-B	N-1		

Where: X=subjects score N=total # subjects n=# subjects in group r=#rows c=#columns  
 SS= Sum of Squares MS=Mean Square df=Degrees of Freedom

Source	SS	df	MS	F
Row (Subjects)	556.33-528.07	5-1	$SS_r / df_s$	$MS_s / MS_r$
Columns (Time)	545-528.07	3-1	$SS_t / df_t$	$MS_t / MS_r$
Residual	(557-528.07)- (556.33-528.07)+ (545-528.07)	(5-1)(3-1)	$SS_r / df_r$	
Total	557-528.07	15-1		

Source	SS	df	MS	F
Row (Subjects)	28.27	4	28.27/4	$MS_s / MS_r$
Columns (Time)	16.93	2	16.93/2	$MS_t / MS_r$
Residual	3.73	8	3.73/8	
Total	48.93	14		

Source	SS	df	MS	F
Row (Subjects)	28.27	4	7.07	7.07/0.46
Columns (Time)	16.93	2	8.47	8.47/0.46
Residual	3.73	8	0.46	
Total	48.93	14		

Source	SS	df	MS	F
Row (Subjects)	28.27	4	7.07	15.36*
Columns (Time)	16.93	2	8.47	18.41*
Residual	3.73	8	0.46	
Total	48.93	14		

We don't care about differences in subjects, but for Time: Critical  $F_{(2, 8) @ .05} = 4.46$  Therefore reject  $H_0$  & accept  $H_a$  that there is a significant difference with respect to time. Need a follow-up test to know if it's between initial & 1 week, initial & 1 month or 1 week & 1 month.

Follow-up test: Already stated that A prior test are more powerful statistic, but out of the scope of this tutorial. Thus, we need Post Hoc follow-up tests. Many to choose from and which follow-up test you choose is a tutorial in itself. Most common are:

Least Significant Difference (LSD) & Protected LSD – Basically multiple t-test

HSD – Tukey's – Ok, but lose power

Sheffee' - Ok

Newman-Keuls – Seldom use

Duncan's Multiple Range – Best because it keeps track of the number of means you are comparing and modifies the  $\alpha$  for various p values.

**ANCOVA** The analysis of covariance (ANCOVA) is a technique that combines the features of an ANOVA and regression. In a one-way ANOVA, you typically are looking at the means of the dependent variable with respect to some treatment variable (plus the residuals). But suppose that on each unit we have also measured another variable that is linearly related to our means? We can now have a regression coefficient of Y on X. This is basically the ANCOVA model. If X and Y are closely related, we may expect this model to fit the values better than an ANOVA model. The model extends easily to more complex situations (two-way, randomized blocks, etc.).

**Example:** There are many variables in biology & medicine that are not fully meaningful by themselves (e.g. O<sub>2</sub> consumption is more meaningful on a per kilogram basis than as an actual number). Let's assume that we want to look at O<sub>2</sub> consumption in premature infants after giving one of three drugs. Since the value for O<sub>2</sub> consumption is affected by gestational age & weight we would want to correct any measures of O<sub>2</sub> consumption for gestational age & weight. Without doing so, the values would NOT be meaningful across the wide range of premature infants. The same can be said for analyzing joint moments across the wide range of heights & weights if the kinetics are standardized to weight or weight & height.

Gestational age & weight would be our covariates. What the ANCOVA does is adjusts each raw O<sub>2</sub> consumption score for gestational age & weight of each infant and then conduct an ANOVA. The ANCOVA simply corrects the raw data for the covariates, giving the corrected values used in the ANOVA a common base.

**MANOVA** Using more than 1 dependent variable. The factor variables divide the population into groups allowing you to test null hypotheses about the effects of factor variables on the means of various groupings of a joint distribution of dependent variables. You can investigate interactions between factors as well as the effects of individual factors. In addition, the effects of covariates and covariate interactions with factors can be included. There is more power with univariate statistics. So if you can combine factors without losing information, you can show significant differences easier. You must still have a normal distribution with each variable. You look at a vector of dependent variables. To be considered multivariate normal, each variable must be univariate normal and any linear combination of variables must be univariate normal. Hard to meet & even harder to test. You typically look at each variable separately. However, this can't show you that it is multivariate normal, but can show you if it isn't univariate normal, thus it isn't multivariate normal.

We generalize the t test for multivariate:  $t^2 = ((\bar{x}_{var} - \mu) / (s / \sqrt{n}))^2 \approx (t_{n-1})^2 \approx F_{n-1} = T^2$  (Hotelling's T)

**Example:** I used a 4 X 3 factorial MANOVA to look at the maximum flexion angle, maximum extension angle, maximum angular velocity & maximum extension moment at the ankle, knee & hip during walking.

Source	Hotel T <sup>2</sup>	Exact F	H <sub>0</sub> df	ER df	Prob
Joint	220.0	1653.79	2	15	<0.000
Kin	159.29	743.37	3	14	<0.000
J X K	746.98	1369.46	6	11	<0.000

Since I had a significant interaction, I didn't look at the main effects. I performed follow up tests to find which interactions were significant (I followed up with A Priori Bonferonni adjusted paired t test on all possible interactions) Schutz & Gessaroli (1987).

I can also use a multivariate approach for repeated measure analysis. In the previous repeated measure example, below is the Multivariate results:

**MULTIVARIATE ANALYSIS**

Effect	Value	F	Hyp df	Error df	Sig.
DORSI Pillai' s T	0.867	9.750(a)	2.000	3.000	0.049
Wilks' Λ	0.133	9.750(a)	2.000	3.000	0.049
Hotelling' s T	6.500	9.750(a)	2.000	3.000	0.049
Roy' s Largest Root	6.500	9.750(a)	2.000	3.000	0.049

a Exact statistic

Multivariate Regression: Regression analysis of multiple dependent variables by one or more factor variables or covariates. For regression analysis, the independent (predictor) variables are specified as covariates. The procedure produces a prediction equation of the form:

$$Y_{ij} = B_{0j} + B_{1j}Z_{1i} + B_{2j}Z_{2i} + \dots + B_{rj}Z_{ri} + e_{ij}$$

Where: each B is a (r+1) X m matrix  
 each Z is a r+1 vector (age, ht, wt, etc)  
 each e is a r+1 error vector

Principal Component Analysis:

A principal component (PC) analysis is concerned with explaining the variance-covariance structure through linear combinations of the original variables. Its general objectives are (1) data reduction, and (2) interpretation.

Although all components are required to reproduce the total variability of the system, often much of this variability can be accounted for by a small number of the principal components. In many cases, there is almost as much information in the smaller number of components as there was using all components. The smaller number of components can then replace the original data set.

An analysis of PC often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result. PC analyses are more of a means to an end rather than an end in themselves because they frequently serve as intermediate steps in much larger investigations. For example, PCs may be inputs to a multiple regression or cluster analysis. Also, scaled PCs are one "factoring" of the covariance matrix for factor analysis discussed later.

Algebraically, PCs are particular linear combinations of the variables. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system. The new axes represent the directions with maximum variability and provide a simpler and more meaningful description of the covariance structure.

PCs depend solely on the covariance matrix (correlation matrix). Their development does not require a multivariate normal assumption. On the other hand, additional inferences can be made from the sample components when the population is multivariate normal

Factor Analysis

Factor analysis has provoked rather turbulent controversy throughout its history. Its modern beginnings lie in attempts to define and measure "intelligence." Thus, primarily scientists interested in psychometric measurement developed factor analysis. Arguments over several early interpretations and the lack of powerful computing facilities hindered its initial development as a statistical method. The advances in computer technology have helped its growth the last decade. Most of the original techniques have been discarded and the early controversies resolved. Each application of the technique must be examined on its own merits to determine its success in that field of study.

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors. Basically, the factor model is motivated by trying to show that variables can be grouped by their correlations. That is, all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. If this is true, it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations. For example, correlations from a group of ADL task scores may suggest an underlying "ROM" factor. A second group of variables, representing physical-activity scores, if available, might correspond to another factor. It is this type of structure that factor analysis seeks to confirm.

Factor analysis can be considered as an extension of PC analysis. Both can be viewed as attempts to approximate the covariance matrix. However, the approximation based on the factor analysis model is more elaborate. The primary question in factor analysis is whether the data are consistent with a prescribed structure.

There are many decisions that must be made in any factor analysis study. Probably the most important decision is the choice of the number of common factors. Although a large sample test of model adequacy is available for a given number of common factors, it is suitable only for data that are approximately normally distributed. Also, the test will most assuredly reject model adequacy for a small number of common factors if the number of variables and observations is large. Yet this is when factor analysis provides the most useful approximation. Most often, the final choice of the number of common factors is based on some combination of (1) the proportion of sample variance explained, (2) subject matter knowledge, and (3) the "reasonableness" of the results. Thus, the choice of solution method and type of rotation are less crucial decisions.

Factor analysis still maintains the flavor of an art form over a science. However, Johnson & Wichern suggest the following approach.

1. Perform a principal component factor analysis. This method is particularly appropriate for a first pass through the data.
  - (a) Look for suspicious observations by plotting the factor scores. Also calculate standardized scores for each observation and squared distances.
  - (b) Try a varimax rotation.
2. Perform a maximum likelihood factor analysis including a varimax rotation.
3. Compare the factor analyses solutions.
  - (a) Do the loadings group in the same manner?
  - (b) Plot factor scores obtained for principal components against scores from the maximum likelihood analysis.
4. Repeat the first three steps for other numbers of common factors. Do extra factors necessarily contribute to the understanding and interpretation of the data?
5. For large data sets, split them in half and perform a factor analysis on each part. Compare the two solutions with each other and that obtained from the complete data set to check solution stability. (The data might be divided at random or by placing odd-numbered cases in one group and even-numbered cases in the other group.)

### Discrimination and Classification

Discriminant analysis and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a separation procedure, it is often employed on a one-time basis in order to investigate observed differences when causal relationships are not well understood. Classification procedures are less exploratory in the sense that they lead to well-defined rules, which can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination. The immediate goals of each are:

**Discrimination:** To describe either graphically (in three or fewer dimensions) or algebraically, the differential features of objects (observations) from several known collections (populations). Want to find "discriminants" whose numerical values are such that the collections are separated as much as possible. A more descriptive term for this goal is separation.

**Classification :** To sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign a new object to the labeled classes. This is sometimes called allocation

A function that separates may sometimes serve as an allocator, and, conversely, an allocatory rule may suggest a discriminatory procedure. In practice, separation and allocation frequently overlap and the distinction between them becomes blurred.

### Clustering

Searching the data for a structure of "natural" groupings (or clusters) is an important exploratory technique. Groupings can provide an informal means for assessing dimensionality, identifying outliers and suggesting interesting hypotheses concerning relationships. Exploratory procedures are often quite helpful in understanding the complex nature of multivariate relationships.

Clustering is distinct from the classification methods discussed above. Classification pertains to a known number of groups, and the objective is to assign new observations to one of these groups. Cluster analysis is a more primitive technique in that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or dissimilarities. The inputs required are similarity measures or data from which similarities can be computed.

To illustrate the nature of the difficulty in defining a natural grouping, consider sorting the 16 face cards in an ordinary deck of playing cards into clusters of similar objects. It is immediately clear that meaningful partitions depend on the definition of similar (same suit, same color, same value).

In most practical applications of cluster analysis, the investigator knows enough about the problem to distinguish "good" groupings from "bad" groupings. Why not enumerate all possible groupings and select the "best" ones for further study?

For the playing card example, there is one way to form a single group of 16 face cards; there are 32,767 ways to partition the face cards into two groups (of varying sizes); there are 7,141,686 ways to sort the face cards into three groups (of varying sizes), and so on.' Obviously time constraints make it impossible to determine the best groupings of similar objects from a list of all possible structures. Even large computers are easily overwhelmed by the typically large number of cases, so one must settle for algorithms that search for good, but not necessarily the best, groupings.

Thus, the basic objective in cluster analysis is to discover natural groupings of the variables. In turn, we must first develop a quantitative scale on which to measure the association (similarity) between objects. Even without the precise notion of a natural grouping, we are often able to cluster objects in two- or three-dimensional scatter plots by eye. To take advantage of the mind's ability to group similar objects, several graphical procedures are available for depicting high-dimensional observations in two dimensions.

## REVIEW OF AVAILABLE STATISTICAL TESTS

This tutorial has discussed many different statistical tests. To select the right test, ask yourself two questions: What type of data do you have? What is your goal? Then refer to the table that follows. Most of the tests described in this tutorial & the table can be performed by most advanced statistical packages.

### Review Of Nonparametric Tests

Choosing the right test to compare measurements is tricky, as you must choose between two families of tests (parametric and nonparametric). Many statistical tests are based upon the assumption that the data are sampled from a Normal distribution. These tests are referred to as parametric tests. Commonly used parametric tests are listed in the 2nd column of the table (e.g. t test & ANOVA).

Tests that do not make assumptions about the population distribution are referred to as nonparametric tests. We covered some of these nonparametric tests in the tutorial. All commonly used nonparametric tests rank the outcome variable from low to high and then analyze the ranks. These tests are listed in the 3<sup>rd</sup> column of the table (e.g. Wilcoxon, Mann-Whitney test, and Kruskal-Wallis tests). These tests are also called distribution-free tests.

### Choosing Between Parametric And Nonparametric Tests: The Easy Ones

Choosing between the two types of tests is sometimes easy. Definitely choose a parametric test if you are sure that your data were sampled from a population that follows a Normal distribution (at least approximately). Definitely select a nonparametric test in three situations:

- The outcome is a rank or a score and the population is clearly not Normal. Examples include class ranking of students, the Apgar score, the visual analogue score for pain (measured on a continuous scale where 0 is no pain and 10 is unbearable pain), or a manual muscle test (measured on a continuous scale where 0 is no movement and 5 is basically normal).
- Some values are "off the scale," that is, too high or too low to measure. Even if the population is Normal, it is impossible to analyze such data with a parametric test since you don't know all of the values. Using a nonparametric test with these data is simple. Assign values too low to measure an arbitrary very low value and

assign values too high to measure an arbitrary very high value. Then perform a nonparametric test. Since the nonparametric test only knows about the relative ranks of the values, it won't matter that you didn't know all the values exactly.

- The data are measurements, and you are sure that the population is not distributed in a Normal manner. If the data are not sampled from a Normal distribution, consider whether you can transform the values to make the distribution become Normal (e.g. take the logarithm or reciprocal of all values). There are often biological or chemical reasons (as well as statistical ones) for performing a particular transform.

### Choosing Between Parametric And Nonparametric Tests: The Hard Ones

It is not always easy to decide whether a sample comes from a Normal population. Consider these points:

- If you collect many data points (over a hundred or so), you can look at the distribution of data and it will be fairly obvious whether the distribution is approximately bell shaped. A formal statistical test (Kolmogorov-Smirnoff test, not explained in this tutorial) can be used to test whether the distribution of the data differs significantly from a Normal distribution. With few data points, it is difficult to tell whether the data are Normal by inspection, and the formal test has little power to discriminate between Normal and non-Normal distributions.
- You should look at previous data as well. Remember, what matters is the distribution of the overall population, not the distribution of your sample. In deciding whether a population is Normal, look at all available data, not just data in the current experiment.
- Consider the source of scatter. When the scatter comes from the sum of numerous sources (with no one source contributing most of the scatter), you expect to find a roughly Normal distribution.

When in doubt, some people choose a parametric test (because they aren't sure the Normal assumption is violated), and others choose a nonparametric test (because they aren't sure the Normal assumption is met).

### Choosing Between Parametric And Nonparametric Tests: Does It Matter?

Does it matter whether you choose a parametric or nonparametric test? The answer depends on sample size. Here are four situations to give some insight:

- Large sample. What happens when you use a parametric test with data from a non-Normal population? The central limit theorem ensures that parametric tests work well with large samples even if the population is non-Normal. In other words, parametric tests are robust to deviations from Normal distributions, so long as the samples are large. The snag is that it is impossible to say how large is large enough, as it depends on the nature of the particular non-Normal distribution. Unless the population distribution is really weird, you are probably safe choosing a parametric test when there are at least two-dozen data points in each group.
- Large sample. What happens when you use a nonparametric test with data from a Normal population? Nonparametric tests work well with large samples from Normal populations. The P values tend to be a bit too large, but the discrepancy is small. In other words, nonparametric tests are only slightly less powerful than parametric tests with large samples.
- Small samples. What happens when you use a parametric test with data from non-Normal populations? You can't rely on the central limit theorem, so the P value may be inaccurate.
- Small samples. When you use a nonparametric test with data from a Normal population, the P values tend to be too high. The nonparametric tests lack statistical power with small samples.

Thus, large data sets present no problems. It is usually easy to tell if the data come from a Normal population, but it doesn't really matter because the nonparametric tests are so powerful and the parametric tests are so robust. It is the small data sets that present a dilemma. It is difficult to tell if the data come from a Normal population, but it matters a lot. The nonparametric tests are not powerful and the parametric tests are not robust.

### One- Or Two-Sided P Value?

With many tests, you must choose whether you wish to calculate a one- or two-sided P value (one- or two-tailed). Let's review the difference in the context of a t test. The P value is calculated for the null hypothesis that the two population means are equal, and any discrepancy between the two sample means is due to chance. If this null hypothesis is true, the one-sided P value is the probability that two sample means would differ as much as was

observed (or further) in the direction specified by the hypothesis just by chance, even though the means of the overall populations are actually equal. The two-sided P value also includes the probability that the sample means would differ that much in the opposite direction (i.e., the other group has the larger mean). The two-sided P value is twice the one-sided P value.

A one-sided P value is appropriate when you can state with certainty (and before collecting any data) that there either will be no difference between the means or that the difference will go in a direction you can specify in advance (i.e. you have specified which group will have the larger mean). If you cannot specify the direction of any difference before collecting data, then a two-sided P value is more appropriate. If in doubt, select a two-sided P value. Most recommend that you always calculate a two-sided P value.

### **Paired Or Unpaired Test?**

When comparing two groups, you need to decide whether to use a paired test. When comparing three or more groups, the term paired is not appropriate and the term repeated measures is used instead.

Use an unpaired test to compare groups when the individual values are not paired or matched with one another. Select a paired or repeated-measures test when values represent repeated measurements on one subject (before and after an intervention) or measurements on matched subjects. The paired or repeated-measures tests are also appropriate for repeated laboratory experiments run at different times, each with its own control.

You should select a paired test when values in one group are more closely correlated with a specific value in the other group than with random values in the other group. It is only appropriate to select a paired test when the subjects were matched or paired before the data were collected. You cannot base the pairing on the data you are analyzing.

### **Fisher's Test Or The Chi-Square Test?**

When analyzing contingency tables with two rows and two columns, you can use either Fisher's exact test or the chi-square test. The Fisher's test is the best choice as it always gives the exact P value. The chi-square test is simpler to calculate but yields only an approximate P value. If a computer is doing the calculations, you should choose Fisher's test unless you prefer the familiarity of the chi-square test. You should definitely avoid the chi-square test when the numbers in the contingency table are very small (any number less than about six). When the numbers are larger, the P values reported by the chi-square and Fisher's test will be very similar. The chi-square test calculates approximate P values, and the Yates' continuity correction is designed to make the approximation better. Without the Yates' correction, the P values are too low. However, the correction goes too far, and the resulting P value is too high. Statisticians give different recommendations regarding Yates' correction. With large sample sizes, the Yates' correction makes little difference. If you select Fisher's test, the P value is exact and Yates' correction is not needed and is not available.

### **Regression Or Correlation?**

Linear regression and correlation are similar and easily confused. In some situations it makes sense to perform both calculations. Calculate linear correlation if you measured both X and Y in each subject and wish to quantify how well they are associated. Select the Pearson (parametric) correlation coefficient if you can assume that both X and Y are sampled from Normal populations. Otherwise choose the Spearman nonparametric correlation coefficient. Don't calculate the correlation coefficient (or its confidence interval) if you manipulated the X variable.

Calculate linear regressions only if one of the variables (X) is likely to precede or cause the other variable (Y). Definitely choose linear regression if you manipulated the X variable. It makes a big difference which variable is called X and which is called Y, as linear regression calculations are not symmetrical with respect to X and Y. If you swap the two variables, you will obtain a different regression line. In contrast, linear correlation calculations are symmetrical with respect to X and Y. If you swap the labels X and Y, you will still get the same correlation coefficient.

### Selecting a Statistical Test for Common Situations

Goal	Type of Data			
	Measurement from Normal Population	Rank, Score, or Measure from Non-Normal Population	Binomial (Two Possible Outcomes)	Survival Time
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample t test	Wilcoxon test	Chi-square or Binomial test	-
Compare two unpaired groups	Unpaired t test	Mann-Whitney test, Nominal data: Fisher's Exact (small sample), Chi-square (large sample). Ordinal Data: Wilcoxon Rank Sum	Fisher's (small sample), Chi-square (large sample)	Log-rank test or Mantel-Haenszel
Compare two paired groups	Paired t test	Wilcoxon Signed Rank, Sign test (small sample), McNemar's test (large sample)	Sign test (small sample), McNemar's test (large sample)	Conditional proportional hazards regression
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression
Compare three or more groups (matched or unmatched)	Repeated-measures ANOVA or MANOVA	Friedman test	Cochrane Q	Conditional proportional hazards regression
Compare groups with known association to other variables	ANCOVA, MANOVA (Principle Components & Factor Analysis)	-	-	-
Quantify association between two variables	Pearson's r	Nominal: Relative Risk Odds Ratio. Ordinal: Spearman Rho, Kendall's Tau	Contingency coefficients	-
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression	Simple logistic regression	Cox proportional hazard regression
Predict value from several measured or binomial variables	Multiple linear regression or Multiple nonlinear regression	-	Multiple logistic regression	Cox proportional hazard regression
<i>Developed from: Intuitive Biostatistics, H.J. Motulsky, Ch. 37, Oxford University Press, 1995. &amp; Hermansen, M. Biostatistics: Some Basic Concepts. Caduceus Medical Publishers. 1990.</i>				

Statistical Tables

Distribution of t (two tail)		
df	0.05	0.01
1	12.706	63.657
∴	∴	∴
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
∴	∴	∴
21	2.080	2.831
22	2.074	2.819
23	2.069	2.807
∴	∴	∴
100	1.982	2.626
120	1.980	2.617
∞	1.960	2.576

Correlation Significance Levels		
df	0.05	0.01
3	0.878	0.959
∴	∴	∴
18	0.444	0.561
19	0.433	0.549
20	0.423	0.537
25	0.381	0.487
∴	∴	∴
100	0.195	0.254
300	0.113	0.148
500	0.088	0.115

Distribution of F					
		df Numerator			
		1	2	3	4
df Denominator	7	5.59	4.74	3.35	4.12
		<b>12.25</b>	<b>9.55</b>	<b>8.45</b>	<b>7.85</b>
	8	5.32	4.46	4.07	3.84
		<b>11.26</b>	<b>8.65</b>	<b>7.59</b>	<b>7.01</b>
	9	5.12	4.26	3.86	3.63
		<b>10.56</b>	<b>8.02</b>	<b>6.99</b>	<b>6.42</b>
	∴	∴	∴		
	24	4.26	3.4		
		<b>7.82</b>	<b>5.61</b>		
	25	4.24	3.38		
		<b>7.77</b>	<b>5.57</b>		
	26	4.22	3.37		
		<b>7.72</b>	<b>5.53</b>		
	27	4.21	3.35		
		<b>7.68</b>	<b>5.49</b>		
	28	4.2	3.34	∴	∴
		<b>7.64</b>	<b>5.45</b>		
	29	4.18	3.33		
		<b>7.6</b>	<b>5.42</b>		
	30	4.17	3.32		
		<b>7.56</b>	<b>5.39</b>		
	32	4.15	3.3		
		<b>7.5</b>	<b>5.34</b>		
	34	4.13	3.28		
<b>7.44</b>		<b>5.29</b>			
∴	∴	∴			
100	3.94	3.09	2.7	2.46	
	<b>6.9</b>	<b>4.82</b>	<b>3.98</b>	<b>3.51</b>	
400	3.86	3.02	2.62	2.39	
	<b>6.7</b>	<b>4.66</b>	<b>3.83</b>	<b>3.36</b>	

### References

- Anderson, T. An Introduction to Multivariate Statistical Analysis. (2<sup>nd</sup> Ed) John Wiley & Sons. 1984
- Chau, T. A review of analytical techniques for gait data. Part I: fuzzy, statistical & fractal methods. *Gait & Posture* 13:49-66.
- Derrick TR, et al., Evaluation of time series data sets using the Pearson product moment correlation coefficient. *Med Sci Sport Exer* 1994. 26 (7) 919-928.
- Fisher, R. Statistical Methods for Research Workers. (4<sup>th</sup> Ed) Oliver & Boyd, Edinburgh.
- Growney et al, Repeated measures of adult normal walking using a video tracking system. *Gait & Posture* 6:147-162 (1997).
- Hermansen, M. Biostatistics: Some Basic Concepts. Caduceus Medical Publishers. 1990.
- Johnson, R., & Wichern, D. Applied Multivariate Statistical Analysis. Prentice-Hall, 1982.
- Kadaba, M.P., et al., 1989. Repeatability of Kinematic, Kinetic, and Electromyographic Data in Normal Adult Gait. *J Orthopaedic Research* 7, 849-860.
- Lin et al, *Biometrics* 45: 255-268 (1989).
- Motulsky, H. Intuitive Biostatistics, Oxford University Press, 1995.
- Rash, G., et al. "Validation of a 3D Video Motion Analysis Method for Measuring Finger Motion". *Journal of Biomechanics*, Vol 32(12), pp. 1337-1341, 1999.
- Schutz, R., & Gessaroli, M. (1987). The analysis of repeated measures designs involving multiple dependent variables. *Research Quarterly*. 58: 132-149.
- Snedecor, G., & Cochran, W. Statistical Methods (7<sup>th</sup> Ed) Iowa State University Press, 1980.
- Somia, N., Rash, G., et al. "Computerized Eyelid Motion Analysis". *Clinical Biomechanics*, Vol 15, pp. 766-771, December, 2000.
- Winer, B., 1971. Statistical principles in experimental design. McGraw-Hill, New York, pp. 261-288.
- Younger, M. A First Course in Linear Regression (2<sup>nd</sup> Ed) Duxbury Press, 1985.
- Zanchi, Papic, Cecic, Quantitative human gait analysis, Modeling and Simulation in Biology and Medicine, Simulation Practice and Theory vol. 8, (Nos. 1-2), p.p. 127-140, April, 2000.