Using Pearson's Correlation Coefficient

I. Correlation Coefficients

-Correlation coefficients in are a class of statistical tests that provide standardized estimates of the strength of the relationship between 2 variables.

-The type of data (nominal, ordinal, interval/ratio) determines the statistic that you use.

- For example Cramer's *V* (a.k.a, phi) is used for discrete data

- In general they represent the ratio of the variance shared between two variables (Covariance) divided by the total variability in both measures.

- Squaring a correlation coefficient tell us the % of the variance in the Dependent Variable that is explained by / predicted by / accounted for by variability in the Independent Variable

II. Pearson's *r*

- Used with two continuous variables
- Estimates the Strength and Direction of the Relationship between the IV and DV
- Scores range from -1 to 1
    - the sign of the coefficient indicates the direction of the relationship
        - -1 = perfect negative correlation (inverse relationship)
            - As one variable increase the other variable decreases.
        - 1 = perfect positive correlation (direct relationship)
            - As one variable increase the other variable increases.
    - the closer the absolute value of *r* is to 1 (regardless of wether it is positive or negative), the stronger the relationship between the two variables
    - the closer the value of *r* is to 0, the weaker the relationship is between the two variables.

<u>Statistical Hypotheses:</u> Ho : $r = 0$

Ha : *r* =\= 0

- we are testing whether or not the strength of the relationship between the IV and the DV differs from what we would expect by chance alone. That is, does it significantly differ from 0.

<u>Example Research Question:</u> What is the strength and direction of the association between the number of Vienna Sausages that a person eats in 24 hr period (measured in grams, range = 0-900) and the amount of time they spend doing impersonations of south park characters in that same 24 hr. period (measured in minutes, range = 0-1440).

III. Calculating Pearson's *r*.

<u>Pearson's *r*</u> : Allows us to test the strength of the association between two continuous variables. It represents a ratio of the Covariance (variance shared by two variables) and the total variance (covariance + unique variance). **<u>df = n-2</u>**

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}} \quad or \quad r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}}$$

<u>r^2</u> : The Coefficient of Determination : *r* represents a ratio of covariance to total variance, but if we want to know what percent of the variance in the DV is explained by variance in the IV, then we can square *r* to find out.

$1-r^2$ = the amount of variance (as a percentage) that is not accounted for. How much residual (left over) variance there is. How much variance is due to error.

Error = The influence on the dependent variable that is attributed to sources other than the independent variable. Also called residual variance and unexplained variance.

IV. How Does *r* Work.

-What the r statistic does is divide up the variance or the deviation in the data set. The deviation of each observation (subject score) is made up of two parts.

Total Variance = (Variance Shared by Variables) + (Variance due to Error)
Covariance

Covariance = the amount of variance in (Y) that is explained (accounted for) by the variance in (X).

$$\text{Covariance}_{xy} = \frac{\Sigma (X - X)(Y - Y)}{n-1} \quad \text{Or} \quad \frac{\Sigma XY - [(\Sigma X)(\Sigma Y)/n]}{n - 1}$$

Variance due to Error = the amount of variance in (Y) that is not explained (shared with / accounted for) by the variance in (X).

What the r formula does is determine the total amount of Covariance and then divide this value by the total amount of variance.

-Pearson's r Formula

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}}$$

$$r = \frac{\text{Covariance}}{\text{Total Variance for Each Variable}}$$

V. Significance of *r*.

- Compare the absolute value of the *r* - obtained, with the *r* - critical, for *df* = n – 2, a the *p* = .05 level.
  - The table below presents the critical values of Pearson's *r*.
- If r - obtained > r - critical$_{(df = n-2,\ p = .05,\ two\ tailed)}$, then reject H_o and Fail to reject H_A
- If r - obtained < r - critical$_{(df = n-2,\ p = .05,\ two\ tailed)}$, then reject H_A and Fail to reject H_o

- The table below presents One and Two tailed tests. For our purposes, we are really only interested in a Two Tailed test.
  - There are occasions where a One Tailed test is permissible, but they are relatively rare and we don't need to worry about them at this point.

df (= N-2) (N= number of pairs)	Level of significance for one-tailed test			
	.05	.025	.01	.005
	Level of significance for two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.628
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.495
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.456
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.284
90	.173	.205	.242	.267
100	.164	.195	.230	.254