

Stats Review
Data Analysis as a Decision Making Process

I Levels of Measurement

NOIR (See Whitley, 2001, pp. 350-351, for details)

Nominal = Categories with Names; Yes vs. No (don't ask sometimes vs. never), Sex, Religion, Relationship Status, Political Affiliation, Experimental Group Membership (Control Group, Manipulation Group, Comparison Group).

Numbers represent groups (1 = Female, 2 = Male).

Order is arbitrary.

Ordinal = Nominal Categories with a Logical Order; Class Rank, Height from tallest to shortest, Responses on a Numerical Rating Scale (1 = strongly agree, 7 = Strongly disagree).

Intervals between numbers is not standard from unit to unit.

Interval = Numerical scales with logically ordered units that are equidistant, but the zero is artificial.

E.g., Temperature in Centigrade and Fahrenheit (Zero does not represent an absence of temperature). Time of day (no 0 o'clock). Calendar Dates. Calendar Years.

Ratio = Numerical scales with logically ordered units that are equidistant and have a true Zero (the zero represents a lack of that which is being measured). E.g., Elapsed Time, Temperature in Kelvin (0 degrees Kelvin = -273.15 degrees celcius), Length, Mass. Because it uses a true zero, numerical values can be used to define ratios: 5 inches is five times more length than 1 inch. 10 inches is twice as long as 5 inches.

Continuous Vs. Discrete Variables

Discrete Variables = Mutually Exclusive/Exhaustive Numerical Categories that can't be broken down in to finer units (e.g., if sex is represented by 1: male and 2: female, there is no 1.5).

All Nominal and Ordinal Variables are Discrete. However, many Ordinal variables will be treated as continuous (e.g., the Numerical rating scales are often averaged to form a single score which is treated as continuous).

Continuous Variables = Numerical systems where there are an infinite number of possible points between each unit. Also the measurements can be broken down into finer units (e.g., elapsed time : Years, Months, Days, Hours, Minutes, Seconds, Milliseconds, Nanoseconds, etc..).

II. Choosing your Statistics

Knowing which statistic to use to test the relationship between each variable depends on the type of data you have (and sometimes the type of question you want to answer)

A. Single Discrete Variable

Goodness of Fit χ^2 = Allows us to test whether the group frequencies differ from chance patterns (base rate frequencies : the frequency instances naturally occur in the environment). ($df = k - 1$, where k = number of groups)

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad \text{where } O_i = \text{observed frequency for each separate group}$$

E_i = expected group frequencies (based on chance)

Statistical Hypotheses: $H_o : O_f = E_f$
 $H_a : O_f \neq E_f$

Example Research Question: Does number of people who say they like cheezy poofs

(Yes = 1) vs. those who do not like cheezy poofs (No = 0), differ significantly from the number expected by chance alone?

B. Discrete X Discrete

Pearson's χ^2 (AKA: Test of Independence)= Allows us to test whether the cross tabulation pattern of two nominal variables differs from the patterns expected by chance. If one variable is ordinal then t or F are normally used.

($df = (R-1)(C-1)$) where $R = \#$ of rows & $C = \#$ of columns.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{where } i = \text{the different groups for Variable 1}$$

$j = \text{the different groups for Variable 2}$

$$E_{ij} = \frac{R_i C_j}{N} \quad \text{where } R_i = \text{Row total of row } i$$

$C_j = \text{Column total of column } j$

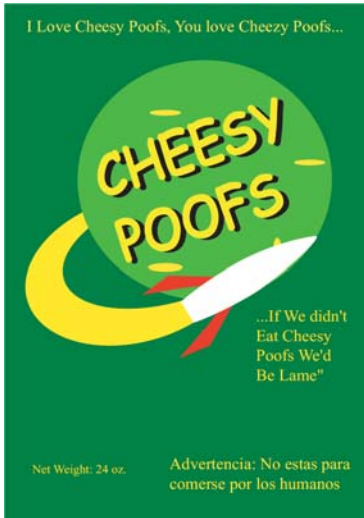
Statistical Hypotheses:
 Ho : $O_f = E_f$
 Ha : $O_f \neq E_f$

Example Research Question: Does the number of people who think they are Eric Cartman (yes = 1, no = 0), relative to whether or not they eat cheezy poofs, significantly differ from the frequencies that are expected by chance alone.

Note: a significant Pearson's chi square will not tell you which cells are different. The crosstabulation matrix must be examined to determine this. "Eye-balling" the standardized, corrected residuals seems to be the most useful.

Limitations on χ^2 :

- 1) Responses must be independent and mutually exclusive and exhaustive. Each case from the sample should fit into one and only one cell of the cross tab matrix.
- 2) Low expected Frequencies limit the validity of χ^2 . If $df = 1$ (e.g., 2x2 matrix), then no expected frequency can be less than 5. Also, If $df = 2$, all expected frequencies should exceed 2. If $df=3$ or greater, then all expected frequencies except one should be 5 or greater and the one cell needs to have an expected frequency of 1 or greater.



Phi Coefficient (if 2X2 matrix) correlation coefficient that estimates the strength of the relationship between two dichotomous nominal variables. Note: Phi can not estimate the direction (e.g., positive linear vs. negative linear) of the relationship between 2 nominal variables because the numerical values are arbitrary (direction is meaningless).

-This correlation coefficient can be calculated exactly like Pearson's r (below) or can be estimated using the χ^2 statistic. Thus any χ^2 can be converted to Phi or Phi can be converted to χ^2 . (significance should be determined using χ^2 tables)

$$\phi = \sqrt{\frac{\chi^2}{N}} \quad \text{and} \quad \chi^2 = \phi^2 N$$

Statistical Hypotheses:
 Ho : $\Phi = 0$
 Ha : $\Phi \neq 0$

Example Research Question: What is the strength of the relationship between whether

Significance (regardless of the number of subjects) is tested using the same Z test as for Rho, only substitute τ for r_s . Use the same Critical Values of Z: If Z greater than 1.96, then alpha < .05. If Z greater than 2.85, then alpha < .01.

Statistical Hypotheses: Ho : Tau = 0
Ha : Tau \neq 0

Example Research Question: What is the strength and direction of the relationship between Eric Cartman's rankings of 20 participant's suitability as a mate (range = 1-20) and participants rankings with respect to how many Cheezy Poofs they eat per-week (range = 1 - 20).



C. Discrete X Continuous

If Discrete Variable is Dichotomous (only 2 levels)

z-test compares a single sample mean to the population mean, when the population standard deviation is known

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \quad \text{where } \mu \text{ \& } \sigma = \text{population mean and standard deviation, respectively}$$

Significance Use the Critical Values of Z: If Z greater than 1.96, then alpha < .05. If Z greater than 2.85, then alpha < .01.

Statistical Hypotheses: Ho : Group 1 mean = Population Mean
Ha : Group 1 mean \neq Population Mean

Example Research Question: Does the average number of Cheezy Poofs eaten per day by students enrolled in Graduate Research Statistics (range = 0-600) differ from the average number of Cheezy Poofs eaten per day in the general population of Graduate Students.

t-test Significance : If t obtained exceeds the t critical (see any t table for critical values) for a given df at the .05 alpha level, then the groups are significantly different.

Single Sample t-test = compare a sample mean to a population mean when the only the sample standard deviation is known. **df = n-1**

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \quad \text{where } s = \text{sample standard deviation}$$

Statistical Hypotheses: Ho : Group 1 mean = Population Mean
Ha : Group 1 mean \neq Population Mean

Example Research Question: Does the average number of Cheezy Poofs eaten per day by students enrolled in Graduate Research Statistics (range = 0-600) differ from the average number of Cheezy Poofs eaten per day in the United States.

Independent Sample t-test = compare two the means of two unrelated groups. **df = n-2**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Statistical Hypotheses: Ho : $\bar{x}_1 = \bar{x}_2$
 Ha : $\bar{x}_1 \neq \bar{x}_2$

Example Research Question: Does a randomly assigned group exposed to 37 hours of South Park (group = 1) reruns report significantly more positive or negative attitudes toward Research Methods (measured using a 5 item questionnaire employing a 7 point rating scale: item averages range from 1-7) compared to a randomly assigned comparison condition exposed to 37 hours of Sally Struthers Feed the Children commercials (group = 2).

Repeated Measure / Matched Sample t-test = Repeated Measure: test the significance of the averaged difference in scores between time 1 and time 2. Matched Sample: compare averaged difference in scores between group 1 and group 2 when the subjects from each group have been matched on some variable (e.g. age, intelligence, etc.). **df = n-1**

$$t_{dif} = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{\frac{\sum (X_{t1} - X_{t2})^2 - \frac{(\sum (X_{t1} - X_{t2}))^2}{n}}{n-1}}}$$

Statistical Hypotheses: Ho : $\bar{x}_{t1} = \bar{x}_{t2}$
 Ha : $\bar{x}_{t1} \neq \bar{x}_{t2}$

Example Research Question: After being exposed to 37 hours of South Park reruns (Time = 2), do participants report significantly more positive or negative attitudes toward Research Methods (measured using a 5 item questionnaire employing a 7 point rating scale: item averages range from 1-7) compared to their pre-exposure scores (Time = 1).

Biserial vs. Point Biserial (Artificial Dichotomies vs. Natural Dichotomies).

Point Biserial Correlation Coefficient r_{pb} = Same formula as Person's r (see below), only one variable is a natural dichotomy often designated by 0 & 1. This correlation coefficient will indicate the strength of the relationship between category membership and the continuous score. Note that like Phi, the direction of the relationship (positive vs. negative) is arbitrarily based on the numerical labels assigned to the groups. Examination of the means is necessary to determine the direction of the group differences. **df = n-2**

Statistical Hypotheses: Ho : $r_{pb} = 0$
 Ha : $r_{pb} \neq 0$

Example Research Question: What is the strength of the relationship between being randomly assigned to a group exposed to 37 hours of South Park reruns (group = 1) vs. being randomly assigned to a comparison condition exposed to 37 hours of Sally Struthers Feed the Children commercials (group = 0) and self-report attitudes toward Research Methods (measured using a 5 item questionnaire employing a 7 point rating scale: item averages range from 1-7).

Biserial Correlation Coefficient r_b = Used when a Dichotomy is developed from a continuous variable (e.g. mean split, or median split methods). Groups are often designated by 0 & 1. r_b is estimated from a normal Pearson's r (see below). This

statistic will tell you the strength of the association between category membership (based on an artificial dichotomy) and a continuous score. Again, direction of the relationship will be arbitrary depending on the numerical category labels (however since they are based on continuous scores the numerical label with greater value should be given to the upper end of the continuum, making interpretation easier) **df = n-2**

$$r_b = \frac{r_{pearson} \sqrt{\%_{belowcp} \%_{abovecp}}}{\frac{X_{cp} - \mu}{\sigma}}$$

where X_{cp} = the raw score cut point (raw score used to split distribution into 2 groups)

Statistical Hypotheses: Ho : $r_b = 0$
Ha : $r_b \neq 0$

Example Research Question: What is the strength and direction of the relationship between watching more than 10 hrs. per week of South Park (Group = 1) vs. watching 10 or fewer hours of South Park per week (group = 0) and self-report attitudes toward Research Methods (measured using a 5 item questionnaire employing a 7 point rating scale: item averages range from 1-7), where 10 hrs. per week is the mean of self-reported South Park viewing habits.

If Discrete Variable has more than 2 Levels (more than 2 groups).

One Way ANOVA : Anova tests whether 2 or more group means are significantly different.

-When using 2 groups $F = t^2$.

-See ANOVA handout for details on One Way Anova

Statistical Hypotheses : Ho : Mean grp1 = Mean grp 2 = Mean grp j (for j groups)
Ha : At least one group mean significantly different from one other group mean.

Example Research Question: Are there any significant difference between three randomly assigned groups (1, exposed to 37 hours of South Park Episodes; 2, exposed to 37 hours of Sally Struthers Feed the Children commercials; & 3, no TV control condition) with respect to their self-report attitudes toward Research Methods (measured using a 5 item questionnaire employing a 7 point rating scale: item averages range from 1-7).

If Discrete X Discrete X Continuous (Where both discrete variables are predictors)

Two Way ANOVA : If we have 2 independent variables (or 1 IV and a Blocking Variable) then Two Way Anova (Or Factorial ANOVA) is called for. Again, this will tell us if one group mean (or matrix cell mean) is significantly different from one other group mean (or cell mean). Also, Factorial Anova can handle More than 2 IV's (and or Blocking Variables). (See Two Way Anova Handout for Details).

Moderation - when we find a significant interaction between two predictor variables we can say that one predictor moderates the relationship between the other predictor and the outcome (DV). Our decision about which predictor is the IV and which is the Moderating Variable is based on our theoretical perspective.

Statistical Hypotheses

- The two way ANOVA actually tests several hypotheses at once.

1) Main Effects

IV1 : Ho : Mean Group 1. = Mean Group i. (for i groups)

Ha : At least one group mean significantly different from one other group mean.

IV2 : Ho : Mean Group .1 = Mean Group .j (for j groups)

Ha : At least one group mean significantly different from one other group mean.

2) Interaction effects (moderation effects)

IV1 : Ho : Mean grp11 = Mean grp 21 = Mean grp12 = Mean grp ij
(for i and j groups)

Ha : At least one group mean significantly different from one other group mean.

Example Research Question: Do males (sex = 0) have significantly more positive or negative attitudes toward Research Methods (measured using a 5 item questionnaire employing a 7 point rating scale: item averages range from 1-7) compared to females (sex =1). Also, does a randomly assigned group exposed to 37 hours of South Park (group = 1) reruns report significantly more positive or negative attitudes toward Research Methods (measured using a 5 item questionnaire employing a 7 point rating scale: item averages range from 1-7) compared to a randomly assigned comparison condition exposed to 37 hours of Sally Struthers' Feed the Children commercials (group = 2). Does participant sex (m = 0, f = 2) moderate the relationship between south park exposure (IV: random assignment to condition: 1) exposed to 37 hours of South Park Episodes; 2) exposed to 37 hours of Sally Struthers' Feed the Children commercials.) and attitude toward Research Methods (measured using a 5 item questionnaire employing a 7 point rating scale: item averages range from 1-7).

ANCOVA : Analysis of Covariance : Sometimes we want to remove the effects of a third variable (Covariate: Can be continuous or discrete). We may want to remove the effects of a nuisance variable that is correlated with our main variables of interest or we may want to test a mediational hypothesis (that the 3rd variable explains the relationship between the IV and DV. that is, the 3rd variable accounts for all the shared variance between the IV & DV). (see a good stats book for details)

Statistical Hypotheses: Ho : Mean Group1 = Mean Groupj (for j groups)

Ha : At least one group mean significantly different from one other group mean.

Example Research Question: Do males (Sex = 0) have significantly more positive or negative attitudes toward Cheezy Poofs (measured using a 6 item questionnaire employing a 5 point rating scale: item averages range from 1-5) compared to females (Sex = 1) after the effects associated with IQ (range 70-150) are removed.

NOTE - Multiple Regression : Anything Anova and Ancova can do, Multiple regression can do as well through the use of dummy coding, effects coding, and contrast coding.

D. Continuous X Continuous

1. Pearson's r : 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). **df = n-2**

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

Statistical Hypotheses: Ho : $r = 0$
Ha : $r \neq 0$

Example Research Question: 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).

r^2 : The Coefficient of Determination : r represents a ratio of covariance to total variance, but if we want to know how much variance in the DV is explained by variance in the IV, then we can square r to find out.

NOTE : for all of the correlation based statistics above (Rho, Tau, Biserial, Point-Biserial) if the sample size exceeds 30 then the Pearson r is robust enough that it can will provide a reasonable approximation of any of these statistics without significantly inflating the Type I error rate.

Significance of r : significance tables for r are available in most stats books. If r obtained exceeds the critical value of r for a given df at the .05 alpha level, then there is a significant association between the variables of interest. If r tables are not available you can use t tables, as r tables are based on t conversions.

$$t = \sqrt{\frac{n - 2}{1 - r^2}}$$

2. Partial Correlation pr : Tests the association between two continuous variables with the effects of a third variable removed. We may want to remove the effects of a nuisance variable that is correlated with our main variables of interest or we may want to test a mediational hypothesis (that the 3rd variable explains the relationship between the IV and DV. that is, the 3rd variable accounts for all the shared variance between the IV & DV). (see a good stats book for details)

-Note: The covariate does not necessarily have to be continuous (especially if your N is larger than 30).

Statistical Hypotheses: Ho : $pr = 0$
Ha : $pr \neq 0$

Example Research Question: 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) when the variance in attitudes toward Vienna Sausages (measure using 4 item measure employing a 7 point numerical rating scale, item averages rang from 1 to 7) is removed.

3. Regression: When you have a single variable, regression is essentially the same as correlation. Instead of r , you calculate for b (beta), where b reflects the slope of line that

best fits the data (least-squares regression coefficient). Strength of association is represented as the change in Y (DV) that results from a 1 unit change in X (IV).

Equation for a Straight Line

(Least-Squares Regression) $Y = a + bX$

Where:

$Y = \text{Dependent Variable}$: seen as a function of (or predicted by) the independent variable (X);

$X = \text{Independent Variable}$: the dimension or characteristic that is seen as the determinant or cause of the dependent variable (Y);

$b = \text{Slope of the Line}$: rise (or drop) divided by run;

$a = \text{Y-intercept}$: where the value of $X = 0$ and the line intercepts the Y axis.

Formula for Calculating a and b

$$b = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} \quad a = \bar{y} - b\bar{x}$$

Once a and b are known they can be plugged into the regression line formula ($Y = a + bX$) and the predicted value of Y can be estimated for any value of X.

Example Research Question- For a given slope (b) of .6300 and a Y-intercept (a) of 3.2, then how much time would a person be expected to spend impersonating Southpark characters (measured in minutes, range 0-1440) in 24 hour period after eating 300 grams of Vienna Sausages?

E. Multiple Continuous Independent Variables and Single Continuous Dependent Variable

Multiple Regression - Like Factorial Anova, Multiple Regression can deal with multiple Independent Variables. Also, as a form of regression, multiple regression can be used for prediction (predicting Y based on the values of the IVs).

$R = \text{eta} = \text{The Multiple Correlation Coefficient}$

- indicates the total association between the Predictors (IVs) and the Criterion (DV).

$R^2 = \text{eta squared} = \text{indicates the \% of the variance in the DV that is accounted for by the IVs.}$

$F = \text{used to test the significance of } R.$

$b = \text{The multiple regression coefficient}$

- indicates the increase in Y resulting from a 1 unit increase in Y, when all other IVs are held at the constant of 0. That is, it tells us about the unique effect a given predictor has on the criterion

$\beta = \text{Beta} = \text{The standardized multiple regression coefficient}$

- same as b only the units are standardized in Z-score units.

$t = (b/\text{standard error for } b) = \text{used to test the significance of the regression coefficient.}$

- The research questions that you can ask with Multiple regression are quite flexible.

- **Single Step** - identifies the unique association of each predictor with a criterion

- **Hierarchical Regression** (Multiple Steps) - identifies the unique contribution of a

single variable or group of variables to the multiple correlation coefficient ($R^2\Delta$)

- **Mediation Analyses** - (Because) A third variable accounts for/Explains the relationship between X and Y. Why is X related to Y, because of Z.

- This is the ultimate goal of Science.

- **Moderation Analyses** / Interaction effects - (It Depends) A third variable influences the strength and/or direction of the relationship between an IV and DV.

What influence does X have on Y? It depends on Z.

F. Multiple Dependent Variables

1. Discrete IVs and Multiple Continuous DV's

MANOVA : Multivariate Analysis of Variance - When you have multiple continuous outcome variables that reflect a related set of constructs and you want to test the association with 1 or more discrete IV's you can use Multiple Analysis of Variance.

- Returns a single F that indicates whether the IV (or IVs) are significantly associated with the DVs as a group.

- This is most useful for keeping the Type I error rate down when conducting multiple analyses.

- If it is significant then it is usually followed up with Univariate tests assessing one dependent variable at a time.

- (see a good stats book for details)

MANCOVA : Manova with a Covariate (a variable that is having its influence removed from the test). (see a good stats book for details)

2. One or more Continuous Independent Variables and Multiple Continuous Dependent Variables

- **Canonical Correlation or Set Correlation** - Returns a single correlation coefficient that is the best fitting correlation between set 1 (IVs) and set 2 (DV's) determined through multiple iterations.

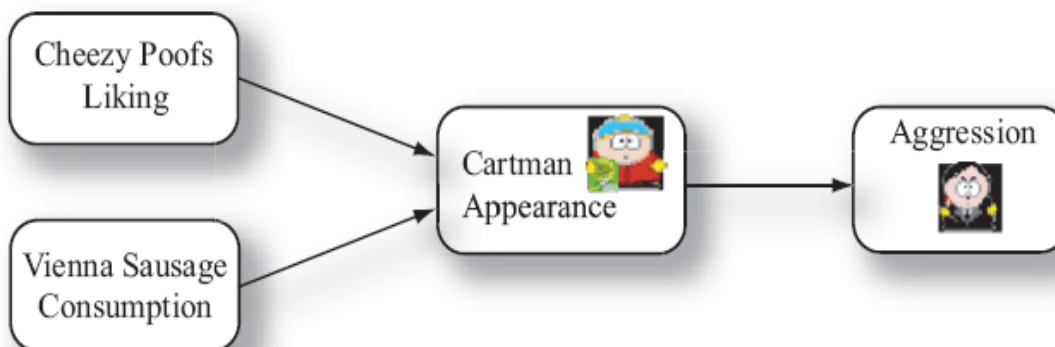
- **Path Analysis / Structural Equation Modeling**

- Theory/Model Testing Procedures -

- Allows you to determine the degree to which causal relationship predicted by theory fit with the data; Goodness of Fit, which is expressed as a chi-square and other fit indices.

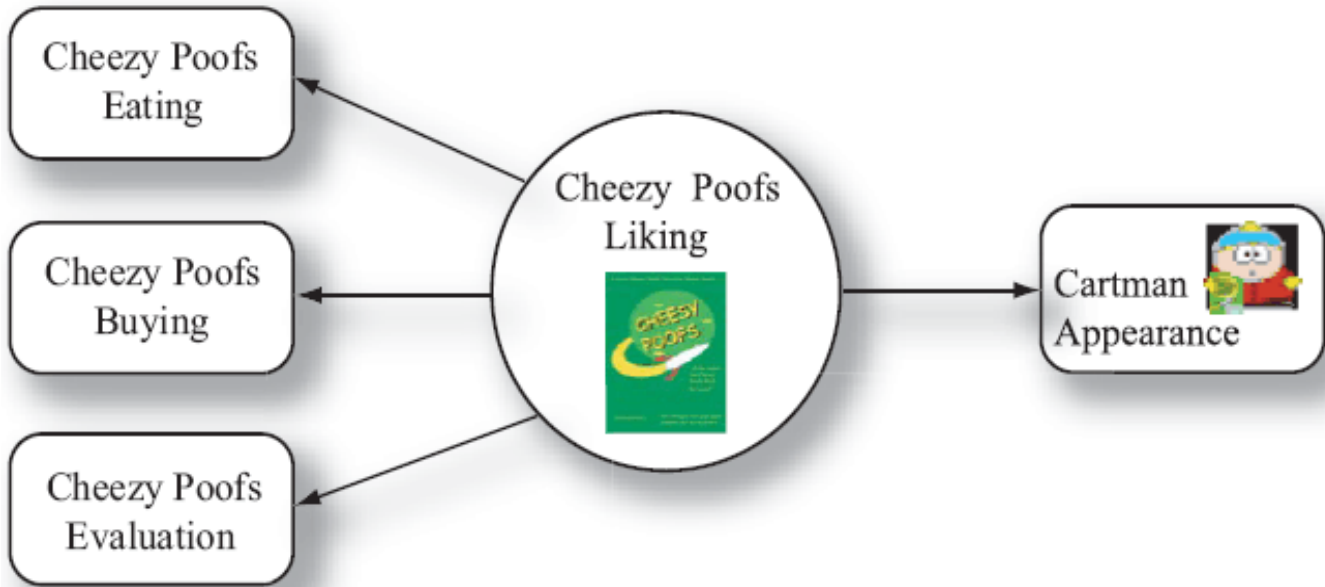
- **Path Analysis / Causal Modeling** - only considers Manifest Variables, which are directly measured variables.

Sample Path Model



- **Structural Equation Modeling / Latent Variable Models** - include Latent Variables, which are not measured directly. For example SES, is not determined by a single indicator. It is a latent variable made up of manifest variables like income, education, job prestige, and obtained wealth.

Sample Structural Model



G. Continuous Predictors and Categorical Dependent Variables

- Logistical Regression

- Allows you to ask a variety of research questions with minimal statistical assumptions (e.g., assumptions regarding normal distributions).
- The “most important” of which would be Strength of Association between predictors and outcomes and Prediction (predicting outcomes/group membership for future cases).
- Logistic regression can handle multiple predictors that either continuous or discrete and combinations of both.

G. Data Reduction

- Reducing a larger number of variables down to more manageable set.
- Selecting Items for scale development
- Factor Analysis / Principle Component Analysis
 - Exploratory -
 - Identifies groups of variables that have been responded to in similar way when no a priori groups have been identified.
 - Confirmatory -
 - Similar to SEM and Path Analysis, identifies the “goodness of fit” between priori groups of items/variables and the data.

I. There are others..... Many Many Others.....