STATISTICS ONE – NOTES AND FORMULAES

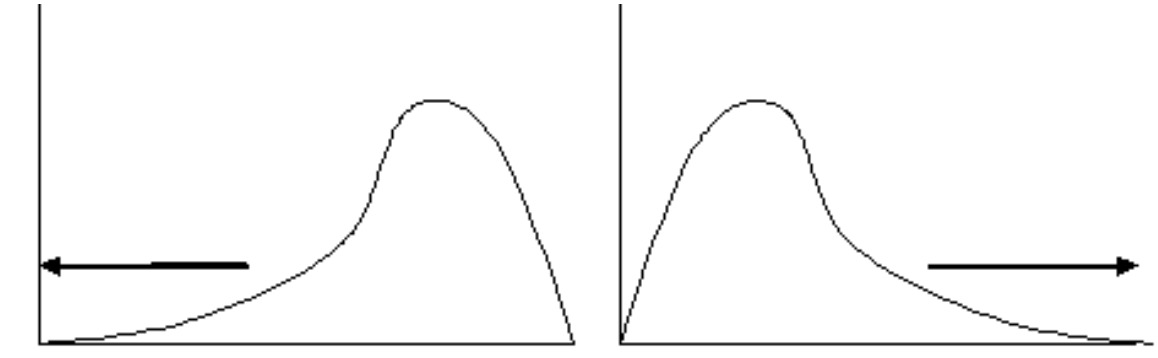
1) Descriptive Statistics

- L1a: Randomized Studies

Definitions:
• Population: the entire collection of cases to which we want to generalize (e.g., all children in the US)
• Sample: a subset of the population
• Parameter: a numerical measure that describes a characteristic of a population
• Statistic: a numerical measure that describes a characteristic of a sample
• Descriptive statistics: procedures used to summarize, organize, and simplify data
• Inferential statistics: techniques that allow for generalizations about population parameters based on sample statistics
• Independent variable: a variable manipulated by the experimenter
– aka treatment, e.g., polio vaccine
• Dependent variable: a variable that represents the aspect of the world that the experimenter predicts will be affected by the independent variable
– aka response, e.g., rate of polio

- L2a: Histograms

Four "moments" of the mean:
• Central tendency (mean/median/mode)
• Variability
• Skew
• Kurtosis



Negative Skew
Elongated tail at the **left**
More data in the left tail than would be expected in a normal distribution

Positive Skew
Elongated tail at the **right**
More data in the right tail than would be expected in a normal distribution

- **L2b: Summary Statistics**

Measures of central tendency

- Mean: the average, M = (ΣX)/N
- Median: the middle score (the score below which 50% of the distribution falls)
- Mode: the score that occurs most often

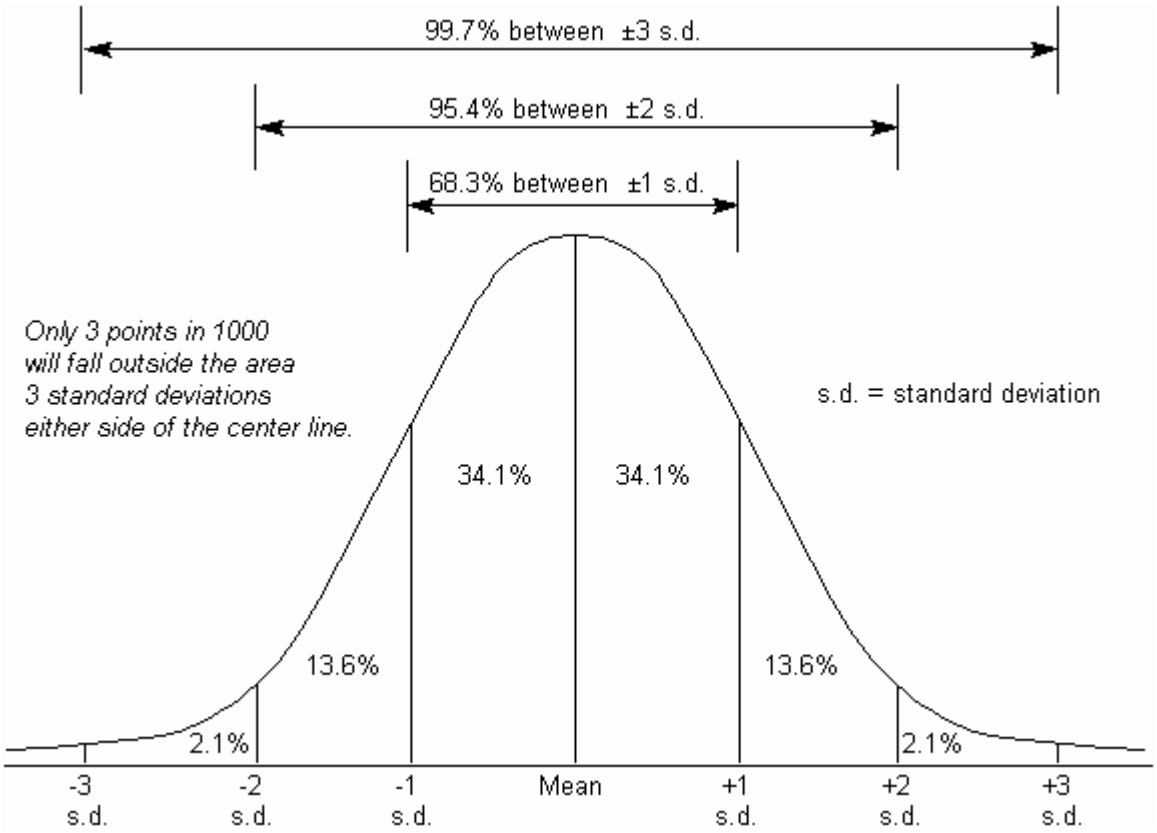
Mean => best when normal distribution
Median => preferred when extreme scores

Variability: A measure that describes the range and diversity of scores in a distribution

- Standard deviation (SD): average deviation from the mean in a distribution
- Variance = SD^2

$SD^2 = [\Sigma(X - M)^2] / N$ (population)
$SD^2 = [\Sigma(X - M)^2] / N$ (sample)

- **L2c: Tools for Inferential Statistics**



Z-scores: A standardized unit of measurement

- Convert raw scores to z-scores:

$Z = (X - M) / SD$

Percentile rank: The percentage of scores that fall at or below a given score

2) Correlation & Measurement

- **L4a: Examples of Correlations**

Correlation: A statistical procedure used to measure and describe the relationship between two variables

=> Correlations can range between +1 and -1

- +1 is a perfect positive correlation
- -1 is a perfect negative correlation

- **L4b: Calculating Correlations:**

Correlation coefficient (r)

- Pearson product-moment correlation coefficient
- r = the degree to which X and Y vary together, relative to the degree to which X and Y vary independently
- r = (covariance of X & Y) / (variance of X & Y)

SP = Sum of cross Products

Review: To calculate SS

– For each subject in the sample, calculate their deviation score

- $(X - Mx)$

– Square the deviation scores

- $(X - Mx)^2$

– Sum the squared deviation scores

- $SSx = \Sigma[(X - Mx)^2] = \Sigma[(X - Mx) \times (X - Mx)]$

To calculate SP

– For each subject in the sample, calculate their deviation scores on both X and Y

- $(X-Mx)$
- $(Y - My)$

– Then, for each subject, multiply the deviation score on X by the deviation score on Y

- $(X - Mx) \times (Y - My)$

– Then sum the "cross products"

- $SP = \Sigma[(X - Mx) \times (Y - My)]$

Formulae to calculate r

Raw score formula:

$r = SPxy / \sqrt{SSxSSy}$

$SSx = \Sigma(X - Mx)^2 = \Sigma[(X - Mx)(X - Mx)]$

$SSy = \Sigma(Y - My)^2 = \Sigma[(Y - My)(Y - My)]$

$SPxy = \Sigma[(X - Mx)(Y - My)]$

$r = SPxy / \sqrt{SSxSSy}$

$r = \Sigma[(X - Mx)(Y - My)] / \sqrt{\Sigma(X - Mx)^2 \Sigma(Y - My)}$

Z-score formula:

$r = \Sigma(zxzy) / N$

$zx = (X - Mx) / SDx$

$zy = (Y - My) / SDy$

$SDx = \sqrt{\Sigma(X-Mx)^2 / N}$

$SDy = \sqrt{\Sigma(Y - My)^2 / N}$

$zx = (X - Mx) / SDx$

$zy = (Y - My) / SDy$

$SDx = \sqrt{\Sigma(X-Mx)^2 / N}$

$SDy = \sqrt{\Sigma(Y - My)^2 / N}$

Proof of equivalence:

$zx = (X - Mx) / \sqrt{\Sigma(X - Mx)^2 / N}$

$zy = (Y - My) / \sqrt{\Sigma(Y - My)^2 / N}$

$r = \Sigma \{ [(X - Mx) / \sqrt{\Sigma(X - Mx)^2 / N}] [(Y - My) / \sqrt{\Sigma(Y - My)^2 / N}] \} / N$

$r = \Sigma \{ [(X - Mx) / \sqrt{\Sigma(X - Mx)^2 / N}] [(Y - My) / \sqrt{\Sigma(Y - My)^2 / N}] \} / N$

$r = \Sigma [(X - Mx)(Y - My)] / \sqrt{\Sigma(X - Mx)^2 \Sigma(Y - My)^2}$

$r = SPxy / \sqrt{SSxSSy}$

<u>Variance and covariance</u>

- Variance = MS = SS / N
- Covariance = COV = SP / N
- Correlation is standardized COV
  - Standardized so the value is in the range -1 to 1

<u>Note on the denominators</u>

- Correlation for descriptive purposes
  - Divide by N
- Correlation for inferential purposes
  - Divide by N-1

- **L4c: Interpreting correlations:**

<u>Assumptions for correlation</u>

- Normal distributions for X and Y
- Linear relationship between X and Y
- Homoskedasticity

- To detect violations, examine scatterplots and plot a histogram of residuals
- In a scatterplot the distance between a dot and the regression line reflects the amount of prediction error
- Homoskedasticity means that the distances (the errors, or residuals) are not related to the variable plotted on the X axis (they are not a function of X)

Reliability of a correlation

- Does the correlation reflect more than just chance covariance?
- One approach to this question is to use NHST
  - H_0 = null hypothesis: e.g., r = 0
  - H_A = alternative hypothesis: e.g., r > 0

Truth	*Decision =>*	**Retain H_0**	**Reject H_0**
H_0 true		Correct Decision p = (1 - α)	Type I error (False alarm) p = α
H_0 false		Type II error (Miss) p = β (1 - POWER)	Correct Decision p = (1 - β) POWER

NHST

- p = P(D|H0)
- Given that the null hypothesis is true, the probability of these, or more extreme data, is p
- NOT: The probability of the null hypothesis being true is p
- In other words, P(D|H0) <> P(H0|D)

NHST can be applied to:

- *r* (Is the correlation significantly different from zero?)
- *r*1 vs. *r*2 (Is one correlation significantly larger than another?)

There are other correlation coefficients:

- Point biserial r => When 1 variable is continuous and 1 is dichotomous
- Phi coefficient => When both variables are dichotomous
- Spearman rank correlation => When both variables are ordinal (ranked data)

- **L5a: Reliability & Validity**

Reliability

- Classical test theory

Raw scores (X) are not perfect
They are influenced by bias and chance error
In a perfect world, we would obtain a "true score"

X = true score + bias + error
Also known as “true score theory”
A measure (X) is considered to be reliable as it approaches the true score
The problem is we don’t know the true score
So, we estimate reliability

- Methods to estimate reliability
    - Test / re-test

=> Measure everyone twice (X1 X2)
=> The correlation between X1 and X2 is an estimate of reliability
However, if the bias is uniform then we won’t detect it with the test / re-test method

- Parallel tests

=> Measures with two different ways (X1 X2)
=> The correlation between X1 and X2 is an estimate of reliability
AND, now the bias of the wand will be revealed

- Inter-item

=> Inter-item is the most commonly used method in the social sciences
Test / re-test and parallel tests are time consuming
Inter-item is therefore more cost efficient

<u>Validity</u>

- Construct

An ideal “object” that is not directly observable
As opposed to “real” observable objects
For example, “intelligence” is a construct

- Construct validity
    - Content validity

Does the test consist of words that children should know?

- Convergent validity

Does the test correlate with other, established measures of verbal ability?
For example, reading comprehension

- Divergent validity

Does the test correlate less well with measures designed to test a different type of ability?
For example, spatial reasoning

- Nomological validity

Are scores on the test consistent with more general theories, for example, of child development and neuroscience
For example, a child with neural damage or disease to brain regions associated with language development should score lower on the test

- **L5b: Sampling**

<u>Sampling error</u>

- The difference between the population and the sample

- BUT, We typically don't know the population parameters
- Sampling error clearly depends on the size of the sample, relative to the size of the population
- Also depends on the variance in the population
- We therefore estimate sampling error from the size of the sample and the variance in the sample
- Under the assumption that the sample is random and representative of the population

Standard Error

- Standard error is an estimate of amount of sampling error
- SE = SD / SQRT(N)

With SE: Standard error
SD: Standard deviation of the sample
N: Size of the sample

Probability histogram

- A distribution of sample means
- Standard error is the standard deviation of the probability histogram

Distribution of sample means

Characteristics

- It is hypothetical, i.e., we don't *know* the dimensions of the distribution as we do with a distribution of individual scores (we estimate the dimensions)
- The mean of the distribution of sample means should be the same as the mean of the population of individuals
- The variance of the distribution of sample means is less than the variance in the population of individuals
- The shape of the distribution of sample means is approximately normal
    - $\sigma^2M = \sigma^2 / N$
    - σ^2M is the variance of the distribution of sample means
    - σM is the standard deviation of the distribution of sample means (standard error)
    - σ^2 is the variance of the population
    - σ is the standard deviation of the population
    - N is the sample size

Central Limit Theorem

- Three principles
    - The mean of the distribution of sample means is the same as the mean of the population
    - The standard deviation of the distribution of sample means is the square root of the variance of the distribution of sample means, $\sigma^2M = \sigma^2 / N$

- The shape of the distribution of sample means is approximately normal if either (a) N >= 30 or (b) the shape of the population distribution is normal

3) Regression & Hypothesis testing

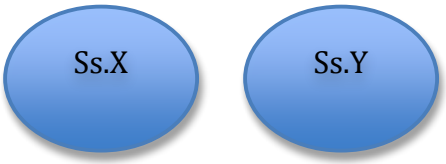
- L7a: Introduction to Regression

What is a regression?
A statistical analysis used to predict scores on an outcome variable, based on scores on one or more predictor variables
For example, we can predict how many runs a baseball player will score (Y) if we know the player's batting average (X)

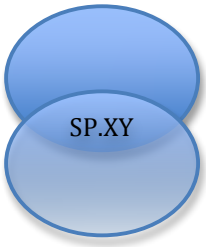
Regression equation
$Y = B0 + B1X1 + e$
$\hat{Y} = B0 + B1X1$ # $\hat{Y}$ is the predicted score on Y
$Y - \hat{Y} = e$ # e is the prediction error (residual)

Estimation of coefficients
The values of the coefficients (B) are estimated such that the model yields optimal predictions.
Minimize the residuals!
The sum of the squared (SS) residuals is minimized
$SS.RESIDUAL = \Sigma(\acute{Y} - Y)2$
ORDINARY LEAST SQUARES estimation

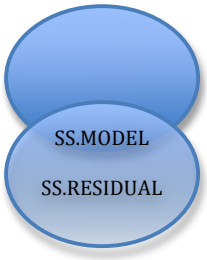
Sum of Squared deviation scores (SS) in variable X = SS.X; in Y = SS.Y



Sum of Cross Products (SP.XY) (Also called SS.MODEL)



SS.Y = SS.MODEL + SS.RESIDUAL



How to calculate B (unstandardized)

$B = r \times (SD_y / SD_x)$

Standardized regression coefficient = $\beta = r$
If X and Y are standardized then:
$SD_y = SD_x = 1$
$B = r \times (SD_y / SD_x)$
$\beta = r$

- **L7b: A Closer Look at NHST**

H_0 = null hypothesis: e.g., $r = 0$, $B = 0$
H_A = alternative hypothesis: e.g., $r > 0$, $B > 0$
Assume H_0 is true, then calculate the probability of observing data with these characteristics, given that H_0 is true
Thus, $p = P(D| H_0)$
If $p < \alpha$ then Reject H_0, else Retain H_0

- $t = B / SE$

B is the unstandardized regression coefficient
SE = standard error
$SE = \sqrt{}$ SS.RESIDUAL / (N – 2)

Problems

- Biased by N

p-value is based on t-value
$t = B / SE$
$SE = \sqrt{}$ SS.RESIDUAL / (N – 2)

- Binary outcome

Technically speaking, one must Reject or Retain the Null Hypothesis
What if p = .06?

- Null "model" is a weak hypothesis

Demonstrating that your model does better than NOTHING is not very impressive

Alternatives to NHST

- Effect size

Correlation coefficient (r)
Standardized regression coefficient (B)
Model R2

- Confidence intervals

Sample statistics are "point estimates"
    Specific to the sample
    Will vary as a function of sampling error
Instead report "interval estimates"
    Width of interval is a function of standard error

- Model comparison

Propose multiple models
    Model A
    Model B
Compare Model R^2

- **L8a: Introduction to Multiple Regression**

<u>Simple vs. multiple regression</u>
Simple regression =>Just one predictor (X)
Multiple regression => Multiple predictors (X1, X2, X3, ...)

<u>Multiple regression equation</u>
Just add more predictors (multiple Xs)
$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + ... + B_kX_k$
$\hat{Y} = B_0 + \Sigma(B_kX_k)$

$\hat{Y}$ = predicted value on the outcome variable Y
B_0 = predicted value on Y when all X = 0
X_k = predictor variables
B_k = unstandardized regression coefficients
$Y - \hat{Y}$ = residual (prediction error)
k = the number of predictor variables

<u>Model R and R^2</u>
R = multiple correlation coefficient
$R = r\hat{Y}Y$
The correlation between the predicted scores and the observed scores

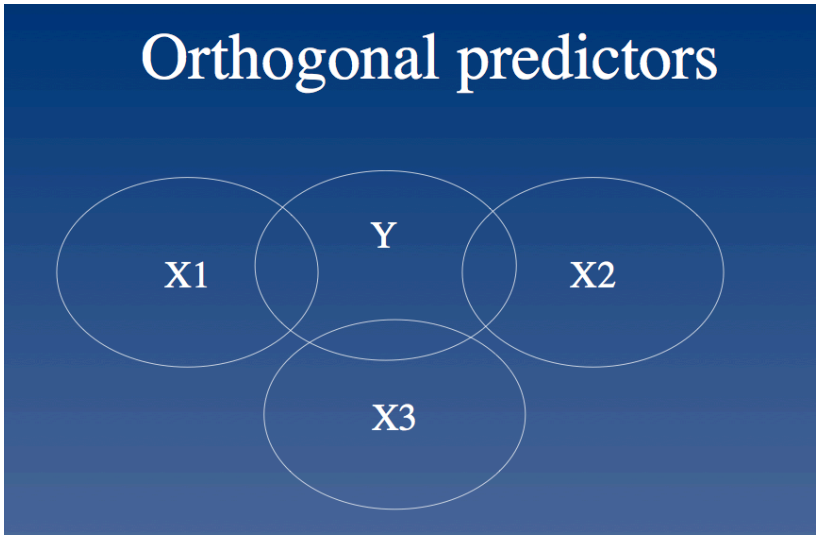
R^2
The percentage of variance in Y explained by the model

<u>Types of multiple regression</u>

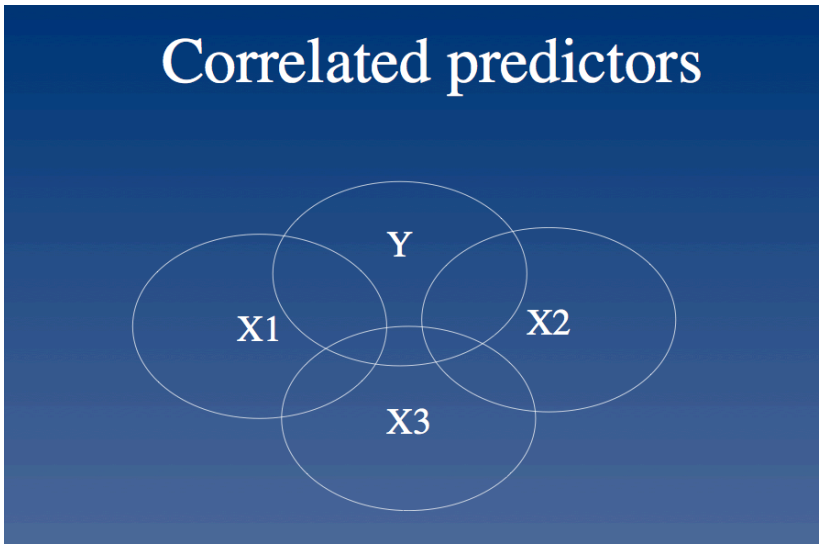
Standard
Sequential (aka hierarchical)

- ⇨ The difference between these approaches is how they handle the correlations among predictor variables

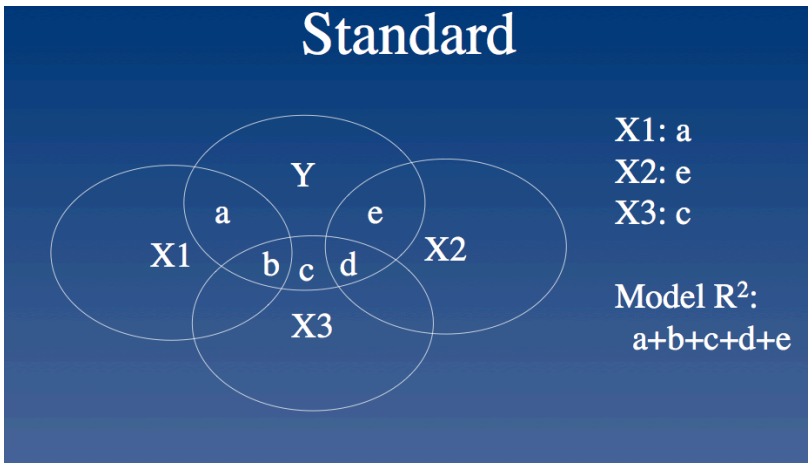
If X1, X2, and X3 are not correlated then type of regression analysis doesn't matter



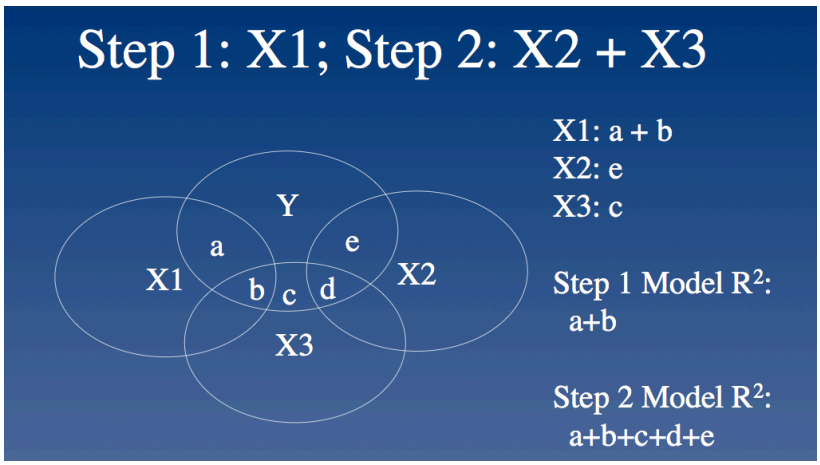
If predictors are correlated then different methods will return different results



- Standard
    - All predictors are entered into the regression equation at the same time
    - Each predictor is evaluated in terms of what it adds to the prediction of Y that is different from the predictability offered by the others
    - Overlapping areas are assigned to R2 but not to any individual B



- Sequential
    - Predictors are entered into the regression equation in ordered steps; the order is specified by the researcher
    - Each predictor is assessed in terms of what it adds to the equation *at its point of entry*
    - Often useful to assess the change in R2 from one step to another



- **L8b: Matrix Algebra**
- A matrix is a rectangular table of known or unknown numbers, e.g.,

$$M = \begin{matrix} 1 & 2 \\ 5 & 1 \\ 3 & 4 \\ 4 & 2 \end{matrix}$$

- The size, or *order*, of a matrix is given by identifying the number of rows and columns, e.g., the order of matrix M is 4x2
- The *transpose* of a matrix is formed by rewriting its rows as columns

$$M = \begin{matrix} 1 & 2 \\ 5 & 1 \\ 3 & 4 \\ 4 & 2 \end{matrix}$$

$$M^T = \begin{matrix} 1 & 5 & 3 & 4 \\ 2 & 1 & 4 & 2 \end{matrix}$$

- Two matrices may be added or subtracted only if they are of the same order
- Two matrices may be multiplied when the number of columns in the first matrix is equal to the number of rows in the second matrix. If so, then we say they are conformable for matrix multiplication.
- $R = M^T * N$ $\qquad R_{ij} = \sum (M^T_{ik} * N_{kj})$

$$R = M^T * N = \begin{pmatrix} 1 & 5 & 3 & 4 \\ 2 & 1 & 4 & 2 \end{pmatrix} * \begin{pmatrix} 2 & 3 \\ 4 & 5 \\ 1 & 2 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 37 & 38 \\ 18 & 21 \end{pmatrix}$$

- A square matrix has the same number of rows as columns
- A square symmetric matrix is such that $D = D^T$

$$D = \begin{pmatrix} 17 & 13 & 18 \\ 13 & 25 & 32 \\ 18 & 32 & 9 \end{pmatrix}$$

- Diagonal matrices are square matrices with zeroes in all off-diagonal cells

$$D = \begin{pmatrix} 17 & 0 & 0 \\ 0 & 25 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

- The inverse of a matrix is similar to the reciprocal of a scalar

e.g., the inverse of 2 is 1/2 and their product = 1
Inverses only exist for square matrices and not necessarily for all square matrices

- An inverse is such that D * D-1 = I where I is the identity matrix

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- The determinant of a matrix is a scalar derived from operations on a square matrix. For example, for a 2x2 matrix A the determinant is denoted as |A| and is obtained as follows:

$|A| = a_{11}*a_{22} - a_{12}*a_{21}$

- A vector is a matrix with only one row or only one column

A row vector is a row of vector elements
A column vector is a column of vector elements

$$R = \begin{pmatrix} 4 & 7 & 5 & 3 \end{pmatrix} \qquad C = \begin{pmatrix} 4 \\ 7 \\ 5 \\ 3 \end{pmatrix}$$

- Raw data matrix

Subjects as rows, variables as columns

$$X_{np} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix}$$

- Row vector of sums (totals)
-

$$T_{1p} = 1_{1n} * X_{np} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} * \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix} = \begin{pmatrix} 34 & 35 & 34 \end{pmatrix}$$

- Row vector of means

$M_{1p} = T_{1p} * N^{-1} = (34\ 35\ 34) * 10^{-1} = (3.4\ 3.5\ 3.4)$

- Matrix of means

$$M_{np} = 1_{n1} * M_{1p} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} * \begin{pmatrix} 3.4 & 3.5 & 3.4 \end{pmatrix} = \begin{pmatrix} 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \end{pmatrix}$$

- Matrix of deviation scores

$$D_{np} = X_{np} - M_{np} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix} - \begin{pmatrix} 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \end{pmatrix} = \begin{pmatrix} -.4 & -1.5 & -.4 \\ -.4 & -1.5 & -.4 \\ -1.4 & .5 & .6 \\ .6 & -.5 & .6 \\ .6 & .5 & -.4 \\ 1.6 & .5 & -.4 \\ -1.4 & 1.5 & .6 \\ -.4 & -.5 & -1.4 \\ 1.6 & -.5 & .6 \\ -.4 & 1.5 & .6 \end{pmatrix}$$

- Sum of squares and cross-products matrix

$$S_{xx} = D_{pn}^{T} * D_{np} = \begin{pmatrix} -.4 & -.4 & -1.4 & .6 & .6 & 1.6 & -1.4 & -.4 & 1.6 & -.4 \\ -1.5 & -1.5 & .5 & -.5 & .5 & .5 & 1.5 & -.5 & -.5 & 1.5 \\ -.4 & -.4 & .6 & .6 & -.4 & -.4 & .6 & -1.4 & .6 & .6 \end{pmatrix} * \begin{pmatrix} -.4 & -1.5 & -.4 \\ -.4 & -1.5 & -.4 \\ -1.4 & .5 & .6 \\ .6 & -.5 & .6 \\ .6 & .5 & -.4 \\ 1.6 & .5 & -.4 \\ -1.4 & 1.5 & .6 \\ -.4 & -.5 & -1.4 \\ 1.6 & -.5 & .6 \\ -.4 & 1.5 & .6 \end{pmatrix}$$

$$= \begin{pmatrix} 10.4 & -2.0 & -.6 \\ -2.0 & 10.5 & 3.0 \\ -.6 & 3.0 & 4.4 \end{pmatrix}$$

- Variance-covariance matrix

$$C_{xx} = S_{xx} * N^{-1} = \begin{pmatrix} 10.4 & -2.0 & -.6 \\ -2.0 & 10.5 & 3.0 \\ -.6 & 3.0 & 4.4 \end{pmatrix} * 10^{-1} = \begin{pmatrix} 1.04 & -.20 & -.06 \\ -.20 & 1.05 & .30 \\ -.06 & .30 & .44 \end{pmatrix}$$

- Diagonal matrix of standard deviations

$$S_{xx} = (\text{Diag}(C_{xx}))^{1/2} = \begin{pmatrix} 1.02 & 0 & 0 \\ 0 & 1.02 & 0 \\ 0 & 0 & .66 \end{pmatrix}$$

- Correlation matrix

$$R_{xx} = S_{xx}^{-1} * C_{xx} * S_{xx}^{-1} =$$

$$\begin{pmatrix} 1.02^{-1} & 0 & 0 \\ 0 & 1.02^{-1} & 0 \\ 0 & 0 & .66^{-1} \end{pmatrix} * \begin{pmatrix} 1.04 & -.20 & -.06 \\ -.20 & 1.05 & .30 \\ -.06 & .30 & .44 \end{pmatrix} * \begin{pmatrix} 1.02^{-1} & 0 & 0 \\ 0 & 1.02^{-1} & 0 \\ 0 & 0 & .66^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} 1.00 & -.19 & -.09 \\ -.19 & 1.00 & .44 \\ -.09 & .44 & 1.00 \end{pmatrix}$$

- **L8c: Estimation of Coefficients**
- Still ORDINARY LEAST SQUARES estimation, but using matrix algebra
- The values of the coefficients (B) are estimated such that the model yields optimal predictions.
    - Minimize the residuals!
    - The sum of the squared (SS) residuals is minimized
    - SS.RESIDUAL = $\Sigma(\acute{Y} - Y)^2$
    - ORDINARY LEAST SQUARES estimation

- Regression equation
    - $\hat{Y} = B0 + B1X1$ # $\hat{Y}$ is the predicted score on Y
    - $Y - \hat{Y} = e$ # e is the prediction error (residual)

- Regression equation, matrix form
    - $\hat{Y} = BX$

- $\hat{Y}$ is a [N x 1] vector (N = number of cases)
- B is a [(1+k) x 1] vector (k = number of predictors)
- X is a [N x (1+k)] matrix

- Make X square and symmetric
  - To do this, pre-multiply by the transpose of X, X'
  - $X'\hat{Y} = X'XB$
- To solve for B, get rid of X'X
  - To do this, pre-multiply by the inverse, $(X'X)^{-1}$

$(X'X)^{-1}X'\hat{Y} = (X'X)^{-1}X'XB$
$(X'X)^{-1}X'X = I$
$IB = B$
$(X'X)^{-1}X'\hat{Y} = B$

4) Mediation & Moderation

Specific types of regression analyses that are popular in the social sciences:

- Mediation

Study the relation between two variables (X, Y) and try to find a third variable (M) that mediates this relation
Popular because social sciences rely heavily on observational studies, raising concerns over causation. Mediation helps answer some of these concerns.

- Moderation (lect. 11)

Variable z (moderator) influences the relation between two variables (X, Y). Z might cancel or enhance the relationship.

Two approaches to mediation:
- Regression method
- Path analysis method

- **L10a: Regression method for Mediation**

An example:
X: Psychological trait (extraversion)
Y: Behavioral outcome (happiness)
M: Mechanism (diversity of life experience)
Z: Moderator (Socio-economic status, SES)

Mediation hypothesis:
Extrovert people might have more diverse experiences, which in turn make you happier

Moderation hypothesis:

SES moderates the relationship between psychological trait and behavioral outcome (eg. True for high SES, but not for lower SES)

A mediation analysis is typically conducted to better understand and observed correlation between X and Y
Eg. Why is extraversion correlated with happiness?

We know from simple regression analysis that if X and Y are correlated then we can use regression to predict Y from X

$Y = B_0 + B_1X + e$

Now if X and Y are correlated BECAUSE of the mediator M, then (X=>M=>Y):

$Y = B_0 + B_1M + e$ (M=>Y)

&

$M = B_0 + B_1X + e$ (X=>M)

In a single equation:

$Y = B_0 + B_1M + B_2X + e$

What will happen to the predictive value of X?
In other words, will B_2 be significant?

⇨ If there is full mediation, B_2 should no longer be significant

It is possible, however, that B_2 will remain significant. In that case, we will have partial mediation

A mediator variable (M) accounts for some or all of the relationship between X and Y:
Some => partial mediation
All => full mediation

CAUTION!

- Correlation does not imply causation!
- In other words, there is a BIG difference between statistical mediation (based on cross-sectional data) and true causal mediation (based on an experimental design)

<u>How to test for mediation</u>

- Run three regression models:
  - lm(Y~X)
  - lm(M~X)
  - lm(Y~X+M)

To get full mediation:

- lm(Y~X)

Regression coefficient for X should be significant

- lm(M~X)

Regression coefficient for X should be significant

- lm(Y~X+M)

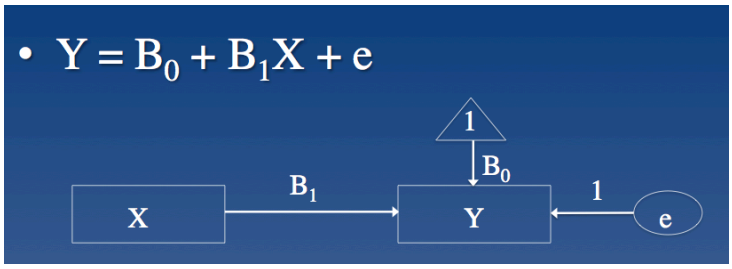
Regression coefficient for M should be significant
Regression coefficient for X?
(If X becomes ns => full mediation, if X remains significant => partial mediation)

Eg.

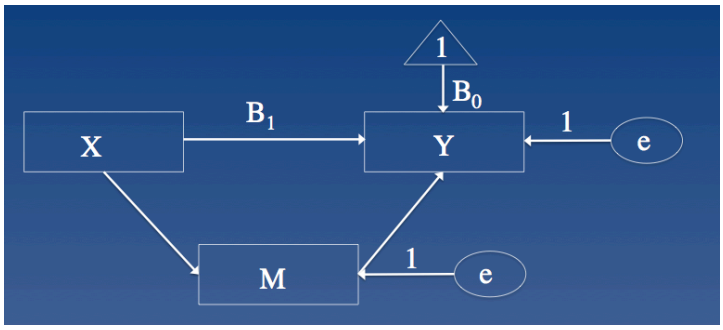
- Assume N = 188
- Participants surveyed and asked to report:
  - Happiness (happy)
  - Extraversion (extra)
  - Diversity of life experiences (diverse)

- **L10b: Path Analysis Method for Mediation**

Mediation analyses are typically illustrated using "path models"
Rectangles: Observed variables (X, Y, M)
Circles: Unobserved variables (e)
Triangles: Constants
Arrows: Associations (more on these later)

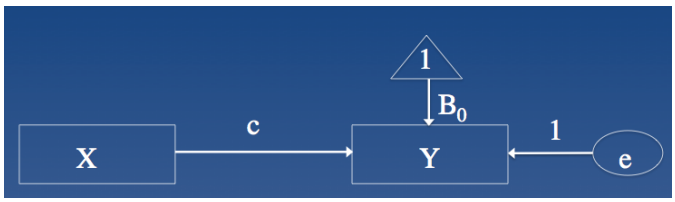


Path model with a mediator

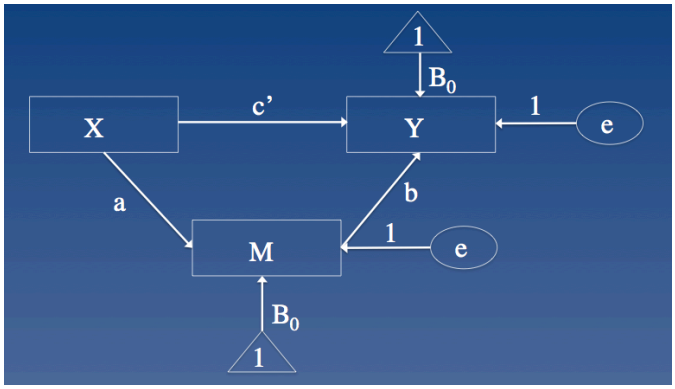


To avoid confusion, let's label the paths
a: Path from X to M
b: Path from M to Y
c: Direct path from X to Y (before including M)
c': Direct path from X to Y (after including M)

Note: (a*b) is known as the indirect path



Path model with a mediator



<u>How to test for mediation</u>
Three regression equations can now be re-written with new notation:
$Y = B_0 + cX + e$
$Y = B_0 + c'X + bM + e$
$M = B_0 + aX + e$

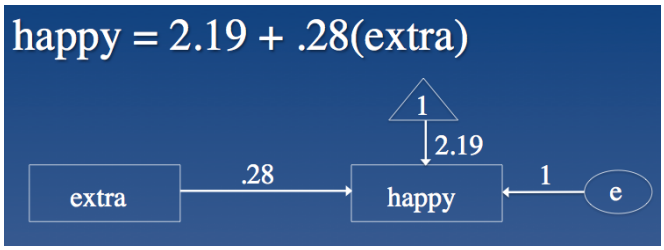
- The Sobel test

$z = (B_a * B_b) / \sqrt{(B_a^2 * SE_b^2) + (B_b^2 * SE_a^2)}$

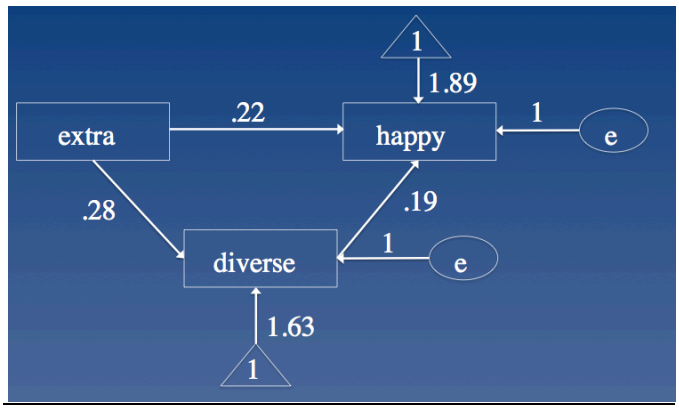
  - The null hypothesis

The indirect effect is zero
$(B_a * B_b) = 0$

<u>Results in path model</u>



Path model with a mediator



- Three regression equations:

happy = 2.19 + .28(extra) + e
diverse = 1.63 + .28(extra) + e
happy = 1.89 + .22(extra) +.19(diverse) + e

Sobel z = +1.98, p = .04

=> Interpretation
- Partial not full, mediation
- Partial mediation because the direct effect (extra) is still significant after adding the mediator (diverse) into the regression equation
- According to the Sobel test, the indirect effect is statistically significant

Mediation: Final comments
- Here we used path analysis to *illustrate* the mediation analysis
- It is also possible to test for mediation using a statistical procedure called: Structural Equation Modeling (SEM)

Also:
Causality!
The example here was weak in terms of ability to make causal statements
Mediation analysis is more powerful with:

- True independent variables
- The incorporation of time

- **L11a: Moderation Example**

- KISS! Keep It Simple Stupid!
- Only 4 variables!

X: Predictor variable (could be an IV)
Y: Outcome variable (could be a DV)
M: Mediator variable
Z: Moderator variable

<u>An example</u>
X: Psychological trait = Extraversion
Y: Behavioral outcome = Happiness
M: Mechanism = Diversity of life experience
Z: Moderator (ZAP! or ZING!) = Socio-Economic-Status (SES)

A moderator variable (Z) has influence over the relationship between X and Y
=> For example, suppose X and Y are positively correlated
The moderator (Z) can change that (ZAP, ZING)

If X and Y are correlated then we can use regression to predict Y from X
$Y = B_0 + B_1X + e$
CAUTION!
If there is a moderator, Z, then B1 will depend on Z
The relationship between X and Y changes as a function of Z

- Assume N = 188
  Participants surveyed and asked to report:

Happiness (happy)
Extraversion (extra)
Diversity of life experiences (diverse)
  All scales 1-5

- To simplify, let's make SES categorical

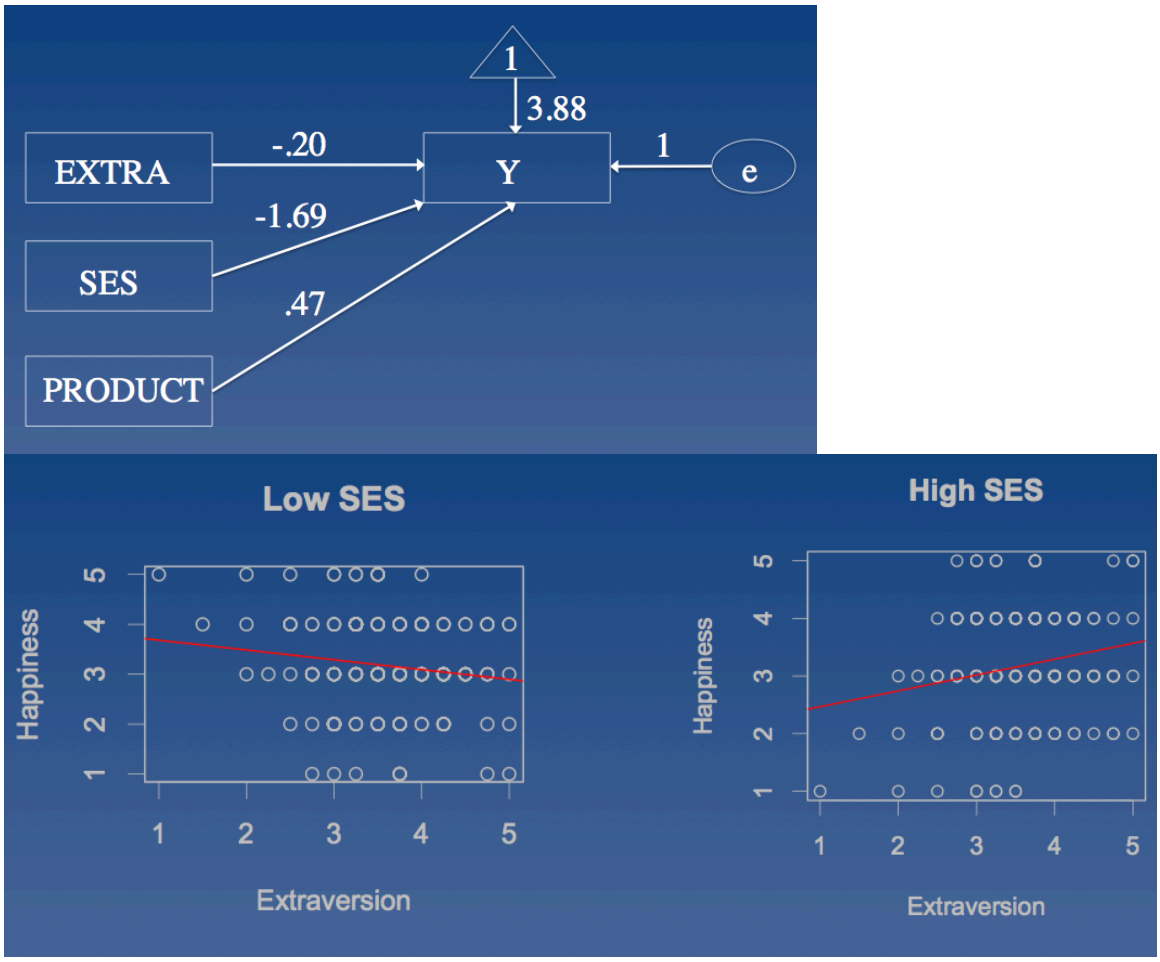
SES = 1 = HIGH SES
SES = 0 = LOW SES

- Results: Before adding PRODUCT

$\hat{Y}= B_0(1) + B_1(EXTRA) + B_2(SES)$
$\hat{Y}= 3.04 + .039\ (EXTRA) + 0.00(SES)$

- Results: After adding PRODUCT

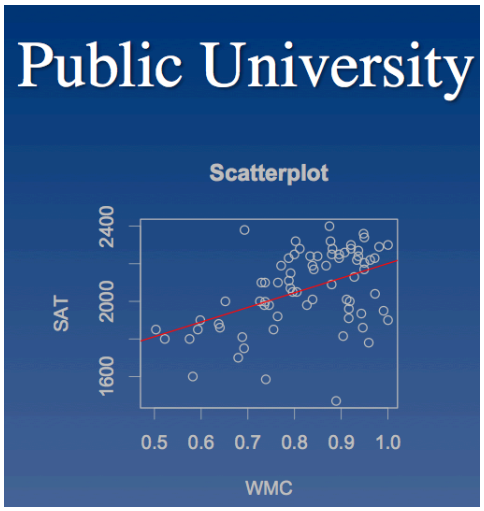
$\hat{Y}= B0(1) + B1(EXTRA) + B2(SES) + B3(PRODUCT)$
$\hat{Y}= 3.88 + -0.20(EXTRA) + -1.69(SES) + 0.47(PRODUCT)$

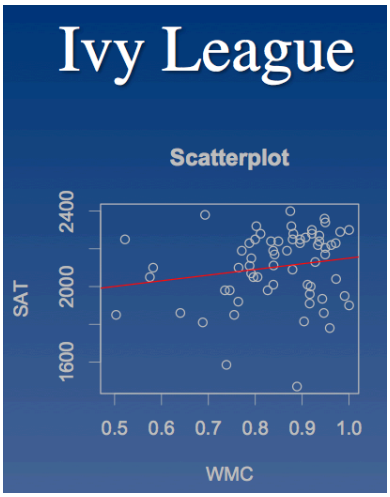


=> Interpretation
SES moderates the relationship between extraversion and happiness
Moral of the story:
The picture can change, literally, when you consider a new variable

<u>Quick example:</u>
Working memory capacity (X)
SAT (Y)
Type of University (Z)

- Large Public State University
- Ivy League (ZAP!)



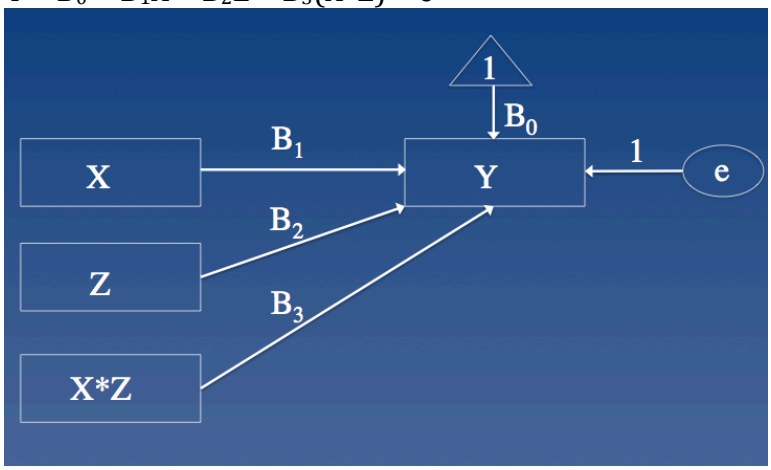


=> Interpretation

Type of University moderates the relationship between WMC and SAT

Moderation model

$Y = B_0 + B_1X + B_2Z + B_3(X*Z) + e$



How to test for moderation

Run just one regression model

lm(Y~X + Z + X*Z)

- Need to create new column for (X*Z)
- Let's call it PRODUCT

- **L11b: Data Centering and Dummy Coding**

<u>Centering predictors</u>

- To center means to put in deviation form

XC = X - M

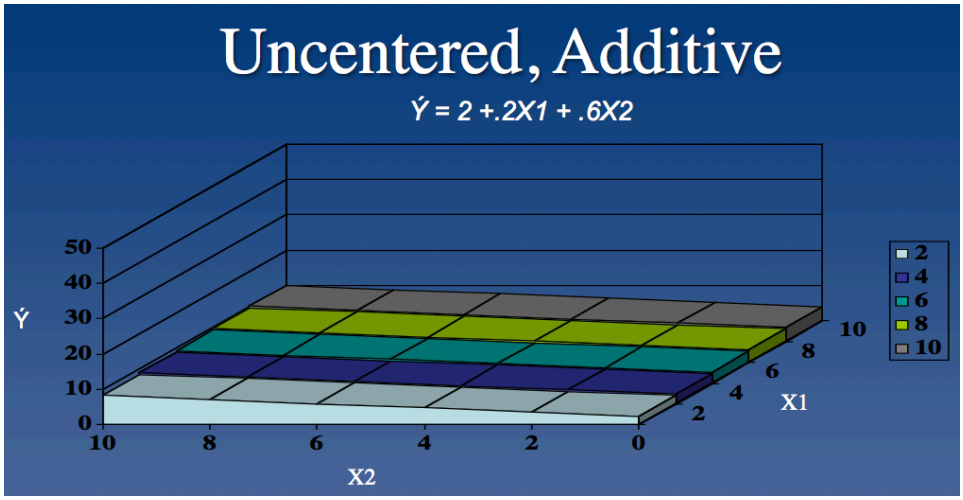
- Why center?

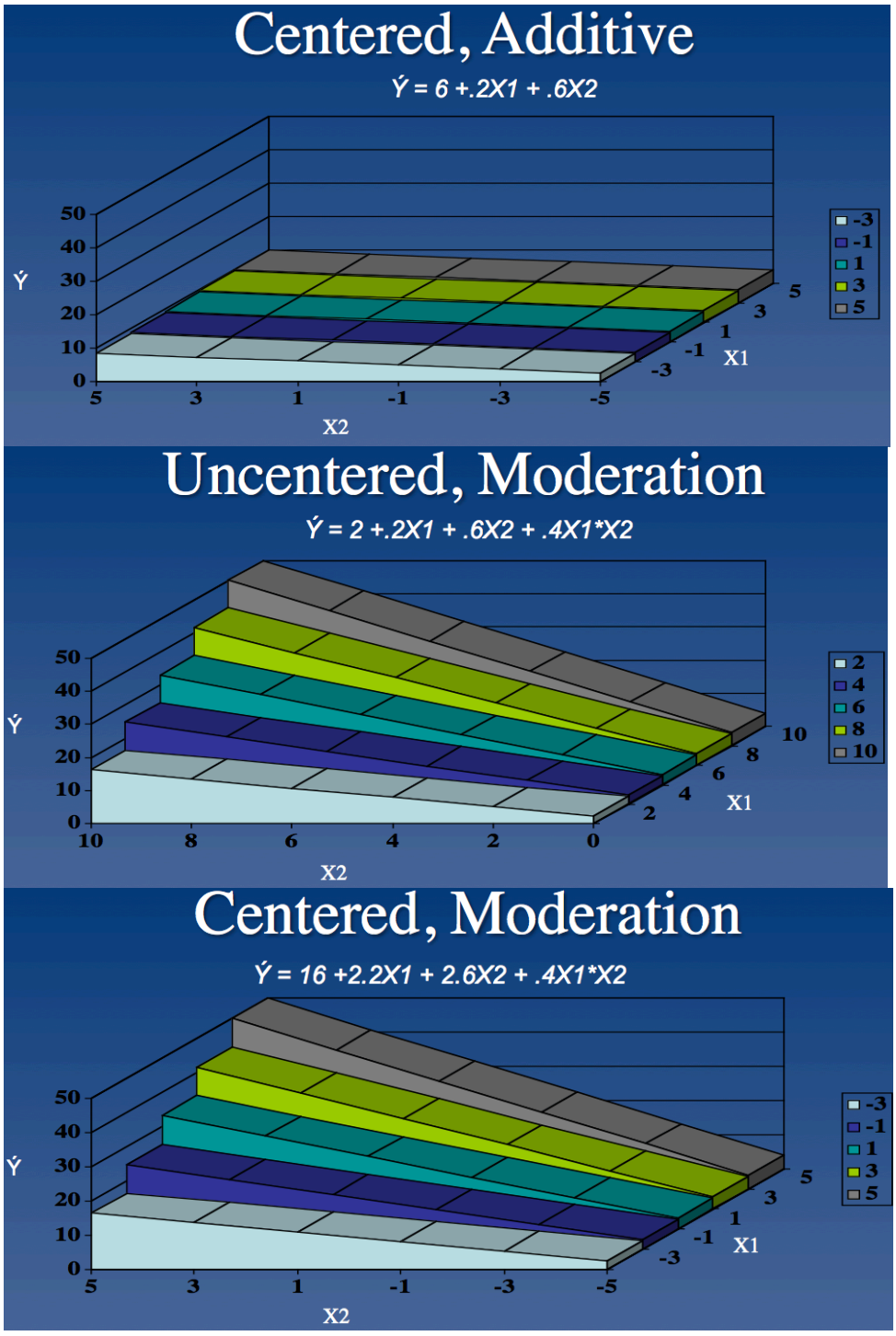
Two reasons:

  - Conceptual reason

Suppose
Y = child's language development
X1 = mother's vocabulary
X2 = child's age

The intercept, B_0, is the predicted score on Y when all X are zero
If X = zero is meaningless, or impossible, then B_0 will be difficult to interpret
If X = zero is the average then B_0 is easy to interpret
The regression coefficient B_1 is the slope for X_1 assuming an average score on X_2
No moderation implies that B_1 is consistent across the entire distribution of X_2
However, moderation implies that B_1 is NOT consistent across the entire distribution of X_2
Where in the distribution of X_2 is B_1 most representative?
Let's look at this graphically





- Statistical reason

The predictors, X1 and X2, can become highly correlated with the product, X1*X2

⇨ Can result in multi-colinearity

<u>Centering for moderation: Summary</u>

Center predictors
Run sequential regression (2 steps)
- Step 1: Main effects
- Step 2: Moderation effect
  - Evaluate B for PRODUCT or ΔR^2 from Model 1 to Model 2

<u>Dummy coding</u>

A system to code categorical predictors in a regression analysis
Example
- IV: Area of research
  - Cognitive
  - Social
  - Neuroscience
  - Cognitive neuroscience
- DV: # of publications

Dummy coding

	C1	C2	C3
Cognitive	0	0	0
Social	1	0	0
Neuro	0	1	0
Cog neuro	0	0	1

Data file

Case	Group	DV	C1	C2	C3
1	Cog	61	0	0	0
2	Soc	78	1	0	0
3	Neuro	47	0	1	0
4	CN	65	0	0	1
…					

Regression model

$\hat{Y} = B_0 + B_1(C1) + B_2(C2) + B_3(C3)$

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	93.308	6.495		14.366	.000
	Social (C1)	-32.641	10.155	-.514	-3.214	.003
	Neuro (C2)	10.192	11.558	.138	.882	.384
	Cog Neuro (C3)	-23.183	10.523	-.351	-2.203	.035

Descriptive Statistics
Dependent Variable: Publications

Area	Mean	Std. Deviation	N
Cognitive	93.3077	29.48272	13
Cog Neuro	70.1250	21.82029	8
Neuro	103.5000	23.64530	6
Social	60.6667	11.12430	9
Total	81.6944	27.88017	36

Unweighted effects coding

	C1	C2	C3
Cognitive	-1	-1	-1
Social	1	0	0
Neuro	0	1	0
Cog neuro	0	0	1

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	81.900	4.055		20.198	.000
	Social (C1)	-21.233	6.849	-.598	-3.100	.004
	Neuro (C2)	21.600	7.883	.550	2.740	.010
	Cog Neuro (C3)	-11.775	7.122	-.322	-1.653	.108

Weighted effects coding

	C1	C2	C3
Cognitive	$-n_s/n_c$	$-n_n/n_c$	$-n_{cn}/n_c$
Social	1	0	0
Neuro	0	1	0
Cog neuro	0	0	1

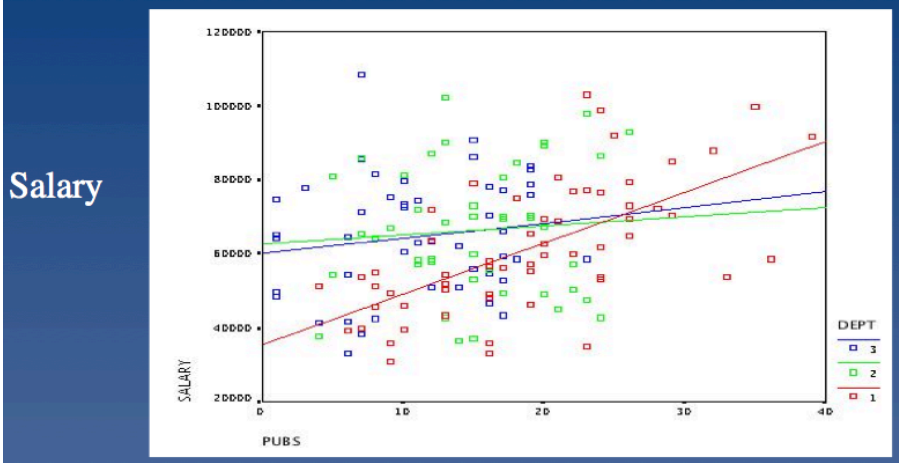
- **L11c: Moderation Example 2**

DV = salary
IVs
– # of publications
– Department
  Psychology
  Sociology
  History

Steps of the analysis:
1/Center continuous predictor (here publications
2/Dummy code categorical predictor (create dummy codes for department, with Psych as the reference group)
3/Create moderation terms
4/Run sequential regression in 2 steps
  Step 1: Main effects
  Step 2: Moderation effect

DEPT		PUBS	SALARY
Psychology	Mean	18.983	61,718.5327
	N	60	60
	Std. Deviation	8.0811	17,589.31294
Sociology	Mean	15.227	66,523.7720
	N	44	44
	Std. Deviation	5.6067	17,530.14462
History	Mean	11.174	64,937.3416
	N	46	46
	Std. Deviation	5.9603	16,001.40284
Total	Mean	15.487	64,115.1710
	N	150	150
	Std. Deviation	7.5064	17,110.14759

Salary as a function of publications and department



Regression model: Before moderation

$\hat{Y} = B_0 + B_1(PUBS.C) + B_2(C1) + B_3(C2)$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	58,482.406	2167.838		26.977	.000
	PUBS.C	925.577	193.424	.406	4.785	.000
	History(C1)	10,447.030	3472.340	.282	3.009	.003
	Sociology(C2)	8281.763	3248.812	.221	2.549	.012

Interpretation of results

The estimated salary for a Psychologist with 15.5 pubs is **58,482**

The average return per publication across all three departments is **926**

When taking into account publications, Historians earn **10,447** more than Psychologists

When taking into account publication rate, Sociologists earn **8,282** more than Psychologists

Regression model: Moderation

$\hat{Y} = B_0 + B_1(PUBS.C) + B_2(C1) + B_3(C2) + B_4(C1*PUBS.C) + B_5(C2*PUBS.C)$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
2	(Constant)	56,918.347	2207.376		25.786	.000
	PUBS.C	1,372.920	252.472	.602	5.438	.000
	History(C1)	9,795.738	3615.187	.265	2.710	.008
	Sociology(C2)	9,672.411	3235.203	.258	2.990	.003
	History(PRODUCT)	-960.978	466.235	-.215	-2.061	.041
	Sociology(PRODUCT)	-1,115.009	495.413	-.196	-2.251	.026

Interpretation of results

The estimated salary for a Psychologist with 15.5 pubs is **56,918** (taking into account the rate of return for Psychologists)

The average return per pub for Psychology is **1,373**

The difference in salary between Psychology and History is **9,796** (for a person with 15.5 pubs, taking into account rate of return)

The difference in salary between Psychology and Sociology is **9,672** (for a person with 15.5 pubs, taking into account rate of return)

The difference in the pubs by salary slope between Psychology and History is **-961**

The difference in the pubs by salary slope between Psychology and Sociology is **-1,115**

Further questions

Is the History slope significant?
Is the Sociology slope significant?
Is the difference in slope between History and Sociology significant?
Re-code to make a different reference group and re-run the analysis

Test of simple slopes
Don't enter the main effect of publications
Create moderation terms that represent the slope for each group

Simple slopes coding

Case	Dept	Pubs	Pubs.c	SS1	SS2	SS3
1	Psy	17	-2	-2	0	0
2	Soc	21	6	0	6	0
3	His	19	8	0	0	8
…						

Note:
Pubs_c = Publications, centered
SS1 = G1xPubs_c, SS2 = G2xPubs_c, SS3 = G3xPubs_c
G1 = Group 1 (Psy), G2 = Group 2 (Soc), G3 = Group 3 (Hist)

Simple slopes: Regression terms

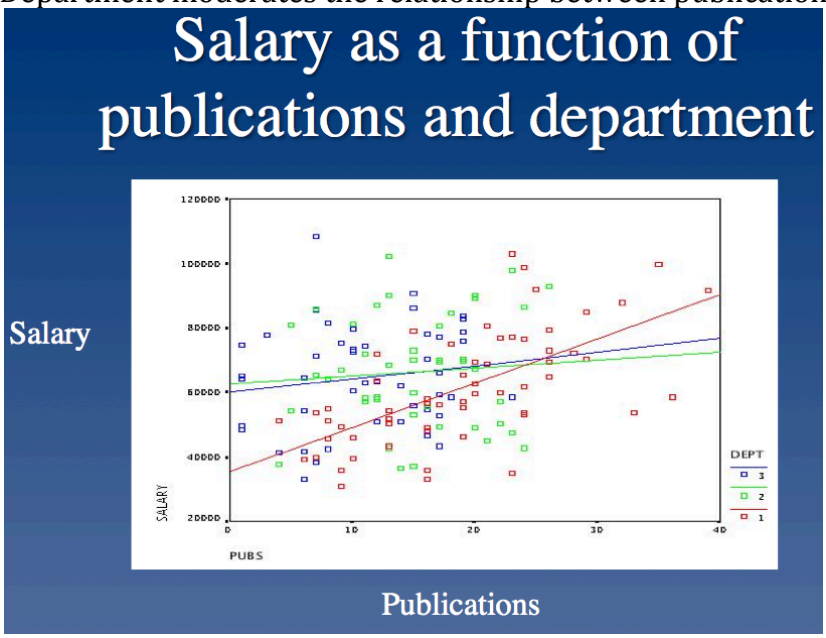
Enter DV (salary)
Enter IVs
- Main effect of department
  - C1, C2
- Three moderation terms
  - SS1, SS2, SS3

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	56,918.347	2207.376		25.786	.000
	History (C1)	9,795.738	3615.187	.265	2.710	.008
	Sociology (C2)	9,672.411	3235.203	.258	2.990	.003
	Psychology (SS1)	1,372.920	252.472	.431	5.438	.000
	History (SS2)	411.942	391.960	.092	1.051	.295
	Sociology (SS3)	257.911	426.253	.045	.605	.546

Interpretation of results

The B's for the moderation terms are the simple slopes
- Psychology is significant
- History is not significant
- Sociology is not significant

Department moderates the relationship between publications and salary



5) Student's t-test & ANOVA

- **L13a: Overview of Students t-test**

z = (observed – expected) / SE
t = (observed – expected) / SE

When to use z and t?

- z

When comparing a sample mean to a population mean and the standard deviation of the population is known

- Single sample t

When comparing a sample mean to a population mean and the standard deviation of the population is not known

- Dependent samples t

When evaluating the difference between two related samples

- Independent samples t

When evaluating the difference between two independent samples

Notation
σ: population standard deviation
μ: population mean
SD: sample standard deviation
M: sample mean
SE: standard error
SE_M: standard error for a mean
SE_{MD}: standard error for a difference (dependent)
$SE_{Difference}$: standard error for a difference (independent)

p values for z and t
Exact p value depends on:
Directional or non-directional test?
df

Different t-distributions for different sample sizes

z : NA
t (single sample) : N-1
t (dependent) : N-1
t (independent) : (N1-1) + (N2-1)

Single sample t
Compare a sample mean to a population mean

$t = (M - \mu) / SE_M$
$SE^2M = SD^2/N$
$SE_M = SD/SQRT(N)$
$SD^2 = \Sigma(X - M)^2 / (N - 1) = SS/df = MS$

Example:
Suppose it takes rats just 2 trials to learn how to navigate a maze to receive a food reward
A researcher surgically lesions part of the brain and then tests the rats in the maze. Is the number of trials to learn the maze significantly more than 2?

Single sample t

Rat #	X	X-M	$(X-M)^2$
1	8	2	4
2	6	0	0
3	4	-2	4
4	9	3	9
5	3	-3	9
			26

$SD^2 = \Sigma(X - M)^2 / (N - 1) = SS/df = 26 / 4 = 6.5$
$SE^2M = SD^2/N = 6.5 / 5 = 1.3$
$SE_M = 1.14$
$t = (M - \mu) / SE_M = (6 - 2) / 1.14 = 3.51$

Effect size (Cohen's *d*)

- $d = (M - \mu) / SD = (6 - 2) / 2.55 = 1.57$

For a directional test with alpha = .05, df = 4, p = .012
⇨ Reject H_0

- **L13b: Dependent & Independent t-tests**

Dependent means t

The formulae are actually the same as the single sample t but the raw scores are difference scores, so the mean is the mean of the difference scores and SEM is based on the standard deviation of the difference scores

Suppose a researcher is testing a new technique to help people quit smoking. The number of cigarettes smoked per day is measured before and after treatment. Is the difference significant?

Subject #	X1	X2	D
1	19	12	-7
2	35	36	1
3	20	13	-7
4	31	24	-7

Subject #	D	$(D - M_D)$	$(D - M_D)^2$
1	-7	-2	4
2	1	6	36
3	-7	-2	4
4	-7	-2	4

$SD^2 = \Sigma(D - MD)^2 / (N - 1) = SS/df = 48 / 3 = 16$
$SE^2_{MD} = SD^2/N = 16 / 4 = 4$
$SE_{MD} = 2$
$t = (M_D - \mu) / SE_{MD} = (-5 - 0) / 2 = -2.5$
$t = M_D / SE_{MD} = -5 / 2 = -2.5$
For a directional test with alpha = .05, df = 3, p = .044
=> Reject H0

Effect size
$d = (M_D - \mu) / SD = -5/4 = -1.25$
Note: $\mu = 0$

<u>Independent means t</u>

Compares two independent groups
For example, males and females, control and experimental, patients and normals, etc.

$t = (M_1 - M_2) / SE_{Difference}$
$SE^2_{Difference} = SE^2_{M1} + SE^2_{M2}$
$SE^2_{M1} = SD2_{Pooled} / N1$
$SE^2_{M2} = SD2Pooled / N2$
$SD^2_{Pooled} = df_1/df_{Total}(SD^2_1) + df_2/df_{Total}(SD^2_2)$
Notice that this is just a weighted average of the sample variances

Group 1 (young adults)
$M_1 = 350$
$SD_1 = 20$
$N_1 = 100$

Group 2 (elderly adults)
$M_2 = 360$
$SD_2 = 30$
$N_2 = 100$

Null hypothesis: $\mu_1 = \mu_2$ Alternative hypothesis: $\mu_1 < \mu_2$

$SD^2_{Pooled} = df_1/df_{Total} (SD^2_1) + df_2/df_{Total} (SD^2_2)$
$SD^2_{Pooled} = 99/198(400) + 99/198(900)$
$SD2_{Pooled} = 650$

$SE^2_{M1} = SD^2_{Pooled} / N1 = 650 / 100 = 6.5$
$SE^2_{M2} = SD^2_{Pooled} / N2 = 650 / 100 = 6.5$
$SE^2_{Difference} = SE^2_{M1} + SE^2_{M2} = 13$
$SE_{Difference} = \sqrt{SE^2_{Difference}} = 3.61$

$t = (M_1 - M_2) / SE_{Difference}$
$t = (350 - 360) / 3.61 = -2.77$
$p = .003$ (based on df = 198, $\alpha = .05$, directional test)
⇨ Reject H0

- **L14a: GLM**

GLM is the mathematical framework used in many common statistical analyses, including multiple regression and ANOVA
ANOVA is typically presented as distinct from multiple regression but it IS a multiple regression

<u>Characteristics of GLM</u>
Linear: pairs of variables are assumed to have linear relations
Additive: if one set of variables predict another variable, the effects are thought to be additive
BUT! This does not preclude testing non-linear or non-additive effects
GLM can accommodate such tests, for example,
Transformation of variables
Transform so non-linear becomes linear
Moderation analysis
Fake the GLM into testing non-additive effects

<u>Examples</u>
Simple regression $Y = B_0 + B_1X + e$
Multiple regression $Y = B_0 + B_1X + B_2X + B_3X + e$
One-way ANOVA $Y = B_0 + B_1X + e$
Factorial ANOVA $Y = B_0 + B_1X + B_2X + B_3X + e$

ANOVA: Appropriate when the predictors (IVs) are all categorical and the outcome (DV) is continuous

- ⇨ Most common application is to analyze data from randomized experiments
- ⇨ More specifically, randomized experiments that generate more than 2 means

(If only 2 means then use t-tests)

NHST may accompany ANOVA

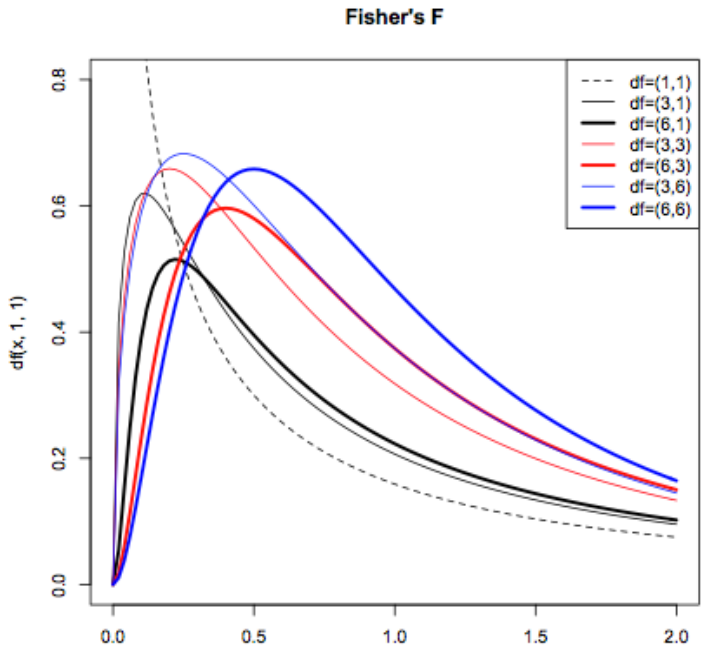
The test statistic is the F-test

F = systematic variance / unsystematic variance

Like the t-test and it's family of t distributions, the F-test has a family of F distributions, depending on:

Number of subjects per group

Number of groups



- **L14b: One-way ANOVA**

F ratio

F = systematic variance / unsystematic variance

F = between-groups variance / within-groups variance

$F = MS_{Between} / MS_{Within}$

$F = MS_A / MS_{S/A}$

With $MS_A = SS_A / df_A$

$MS_{S/A} = SS_{S/A} / df_{S/A}$

$SS_A = n \Sigma(Y_j - Y_T)^2$

Y_j are the treatment means

Y_T is the grand mean

$SS_{S/A} = n \Sigma(Y_j - Y_T)^2$

Y_{ij} are individual scores
Y_j are the treatment means

$df_A = a - 1$
$df_{S/A} = a(n - 1)$
$df_{Total} = N - 1$

Summary Table

Source	SS	df	MS	F
A	$n \Sigma(Y_j - Y_T)^2$	a - 1	SS_A/df_A	$MS_A /MS_{S/A}$
S/A	$\Sigma(Y_{ij} - Y_j)^2$	a(n -1)	$SS_{S/A}/df_{S/A}$	-----
Total	$\Sigma(Y_{ij} - Y_T)^2$	N - 1	-----	-----

<u>Effect size</u>

$R^2 = \eta^2$ (eta-squared)
$\eta^2 = SS_A / SS_{Total}$

<u>Assumptions</u>

DV is continuous
DV is normally distributed
Homogeneity of variance
Within-groups variance is equivalent for all groups
Levene's test (If Levene's test is significant then homogeneity of variance assumption has been violated) => Conduct comparisons using a restricted error term

- **L14c: Factorial ANOVA**

Two IVs (treatments)
One continuous DV (response)
Three F ratios:
F_A
F_B
F_{AxB}

- Main effect: the effect of one IV averaged across the levels of the other IV
- Interaction effect: the effect of one IV depends on the other IV (the simple effects of one IV change across the levels of the other IV)
- Simple effect: the effect of one IV at a particular level of the other IV

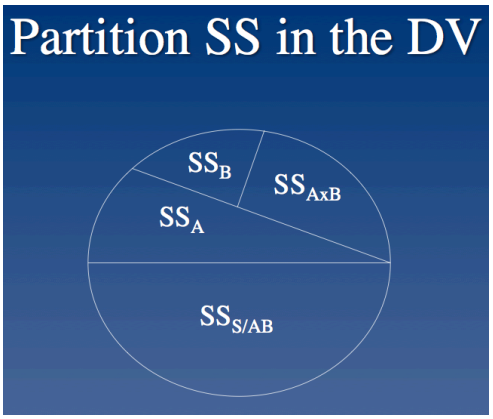
Main effects and interaction effect are independent from one another

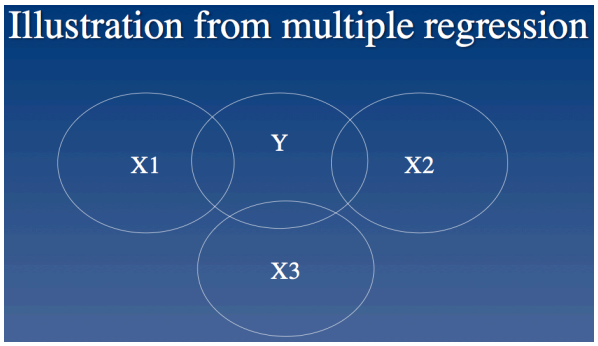
Note that this is different from studies that don't employ an experimental design

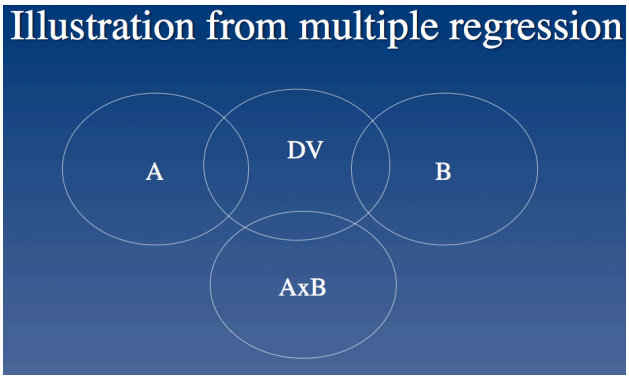
For example, in MR, when predicting faculty salary, the effects of publications and years since the Ph.D. were correlated

Factorial ANOVA is just a special case of multiple regression.

It is a multiple regression with perfectly independent predictors (IVs).







<u>F ratios</u>

$F_A = MS_A / MS_{S/AB}$

$F_B = MS_B / MS_{S/AB}$

$F_{AxB} = MS_{AxB} / MS_{S/AB}$

<u>MS</u>

$MS_A = SSA / dfA$

$MS_B = SSB / dfB$

$MS_{AxB} = SS_{AxB} / df_{AxB}$

$MS_{S/AB} = SS_{S/AB} / df_{S/AB}$

<u>df</u>

$df_A = a - 1$

$df_B = b - 1$

$df_{AxB} = (a -1)(b - 1)$

$df_{S/AB} = ab(n - 1)$

$df_{Total} = abn - 1 = N – 1$

<u>Follow-up tests</u>

Main effects

- Post-hoc tests

Interaction

- Analysis of simple effects
  - Conduct a series of one-way ANOVAs
    - For example, we could conduct 3 one-way ANOVAs comparing high and low spans at each level of the other IV

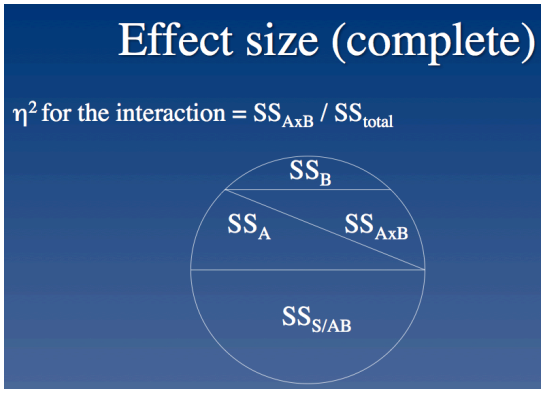
<u>Effect size</u>

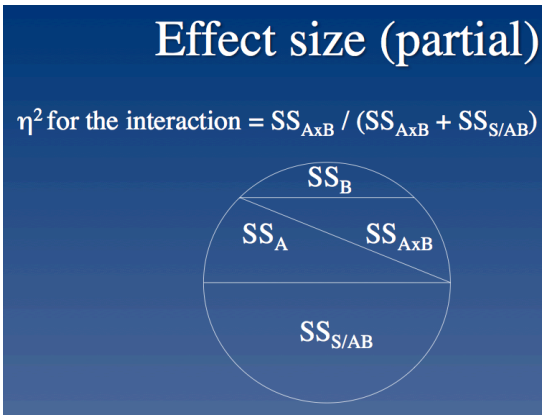
Complete η^2

- $\eta^2 = SS_{effect} / SS_{total}$

Partial η^2

- $\eta^2 = SS_{effect} / (SS_{effect} + SS_{S/AB})$





Assumptions

Assumptions underlying the factorial ANOVA are the same as for the one-way ANOVA

- DV is continuous
- DV is normally distributed
- Homogeneity of variance

6) Factorial ANOVA & Model Comparison

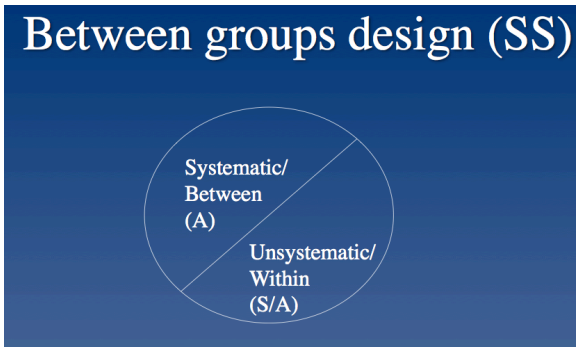
- **L16a: Benefits of Repeated Measures ANOVA**

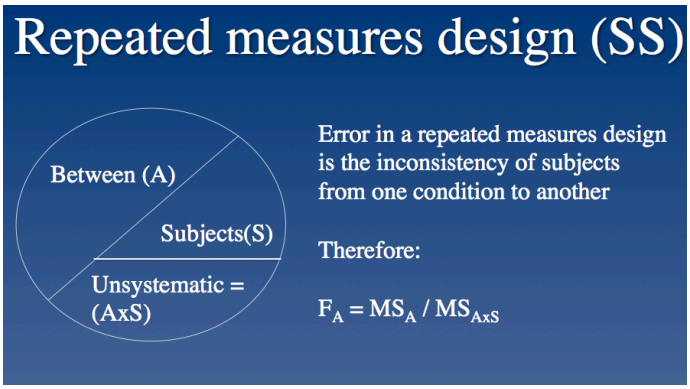
Benefits

Less cost (fewer subjects required)

More statistical power

- Variance across subjects may be systematic
- If so, it will not contribute to the error term





<u>MS and F</u>
$MS_A = SS_A / df_A$
$MS_{AxS} = SS_{AxS} / df_{AxS}$
$F = MS_A / MS_{AxS}$

<u>Post-hoc tests</u>
The error term MSAxS is *NOT* appropriate
Need to calculate a new error term based on the conditions that are being compared
$F_{\psi A} = MS_{\psi A} / MS_{\psi AxS}$
$MS_{\psi A} = SS_{\psi A} / df_{\psi A}$
$MS_{\psi AxS} = SS_{\psi AxS} / df_{\psi AxS}$

Correct for multiple comparisons
Bonferroni

<u>Sphericity assumption</u>
Homogeneity of variance
Homogeneity of correlation
r12 = r13 = r23

How to test?
Mauchly's test
If significant then report the p value from one of the corrected tests
Greenhouse-Geisser
Huyn-Feldt

- **L16b: Risks of Repeated Measures ANOVA**

- <u>Order effects</u>

- <u>Counterbalancing</u>
  - Consider a simple design with just two conditions, A1 and A2

One approach is a Blocked Design

Subjects are randomly assigned to one of two
"order" conditions
A1, A2
A2, A1

- Consider a simple case with just two conditions, A1 and A2

Another approach is a Randomized Design
Conditions are presented randomly in a mixed fashion
A2, A1, A1, A2, A2, A1, A2.....

Now suppose a = 3 and a blocked design
There are 6 possible orders (3!)
A1, A2, A3
A1, A3, A2
A2, A1, A3
A2, A3, A1
A3, A1, A2
A3, A2, A1
To completely counterbalance, subjects would be randomly assigned to one of 6 order conditions
The number of conditions needed to completely counterbalance becomes large with more conditions
4! = 24
5! = 120

With many levels of the IV a better approach is to use a "Latin Squares" design

Latin Squares designs aren't completely counterbalanced but every condition appears at every position at least once

For example, if a = 3, then
A1, A2, A3
A2, A3, A1
A3, A1, A2

<u>Missing data</u>

Two issues to consider

- *Relative amount* of missing data
How much is a lot?
No hard and fast rules
A rule of thumb is
Less than 10% on any one variable, OK
Greater than 10%, not OK

- *Pattern* of missing data

Is the pattern random or lawful?

This can easily be detected

For any variable of interest (X) create a new variable (XM)

XM = 0 if X is missing

XM = 1 if X is not missing

Conduct a t-test with XM as the IV and X as the DV

If significant then pattern of missing data *may be* lawful

- *Remedies*

Drop all cases without a perfect profile

Drastic

Use only if you can afford it

Keep all cases and estimate the values of the missing data points

There are several options for how to estimate values

Estimation methods

Insert the mean

Conservative

Decreases variance

Regression-based estimation

More precise than using the mean but

Confusion often arises over which variables to use as predictors in the regression equation

- **L17a: Mixed Factorial ANOVA**

<u>Design</u>

One IV is manipulated between groups

One IV is manipulated within groups

Repeated measures

<u>What's new?</u>

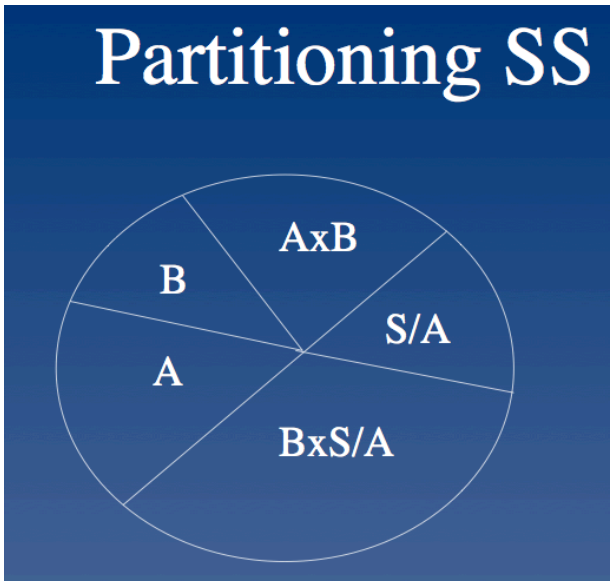
Partitioning SS

Formulae for F_A, F_B, F_{AxB}

Error term for post-hoc tests

Approach to simple effects analyses

Assumptions



df

$df_A = a - 1$
$df_B = b - 1$
$df_{AxB} = (a - 1)(b - 1)$
$df_{S/A} = a(n - 1)$
$df_{BxS/A} = a(b - 1)(n - 1)$
$df_{Total} = (a)(b)(n) - 1$

MS

$MS_A = SS_A / df_A$
$MS_B = SS_B / df_B$
$MS_{AxB} = SS_{AxB} / df_{AxB}$
$MS_{S/A} = SS_{S/A} / df_{S/A}$
$MS_{BxS/A} = SS_{BxS/A} / df_{BxS/A}$

F

$F_A = MS_A / MS_{S/A}$
$F_B = MS_B / MS_{BxS/A}$
$F_{AxB} = MS_{AxB} / MS_{BxS/A}$

Post-hoc tests on main effects

Post-hoc tests on the between-groups IV are performed in the same way as with a one-way ANOVA

TukeyHSD

Post-hoc tests on the repeated measures IV are performed in the same way as with a one-way repeated measures ANOVA

Pairwise comparisons with Bonferroni correction

Simple effects analyses

Must choose one approach or the other (to report both is redundant)

Simple effects of the between groups IV

Or Simple effects of the repeated IV

Subject	A	DV
1	1	B1
1	1	B2
1	1	B3
…	…	
3	2	B1
3	2	B2
3	2	B3

Simple effects of the between groups IV

Simple effect of A at each level of B

$F_{A.at.b1} = MS_{A.at.b1} / MS_{S/A.at.b1}$

Simple comparisons use the same error term

$MS_{S/A.at.b1}$

Simple effect on the repeated measures IV

Simple effect of B at each level of A

$F_{B.at.a1} = MS_{B.at.a1} / MS_{BxS/A.at.a1}$

<u>Assumptions</u>

Each subject provides b scores

Therefore, there are

b variances

(b*(b+1) / 2) - b covariances (correlations)

e.g., if b = 3 then 3 covariances

e.g., if b = 4 then 6 covariances

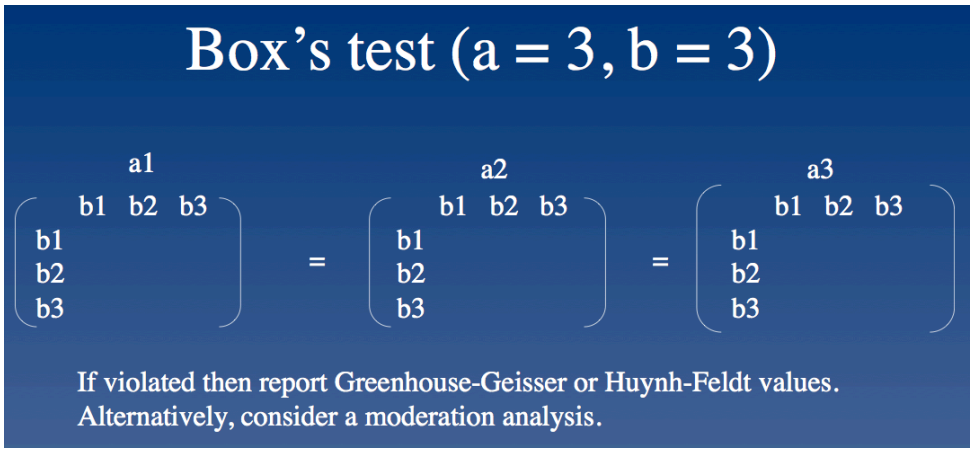
- Between-groups assumptions

The variances do not depend upon the group

Levene's test (If violated then calculate a new restricted error term)

The covariances do not depend upon the group

Box's test of equality of covariance matrices



- Within-subjects assumptions

Sphericity: the variances of the different treatment scores (b) are the same and the correlations among pairs of treatment means are the same

If violated then report Greenhouse-Geisser or Huynh-Feldt values