

Quantitative Research Methods

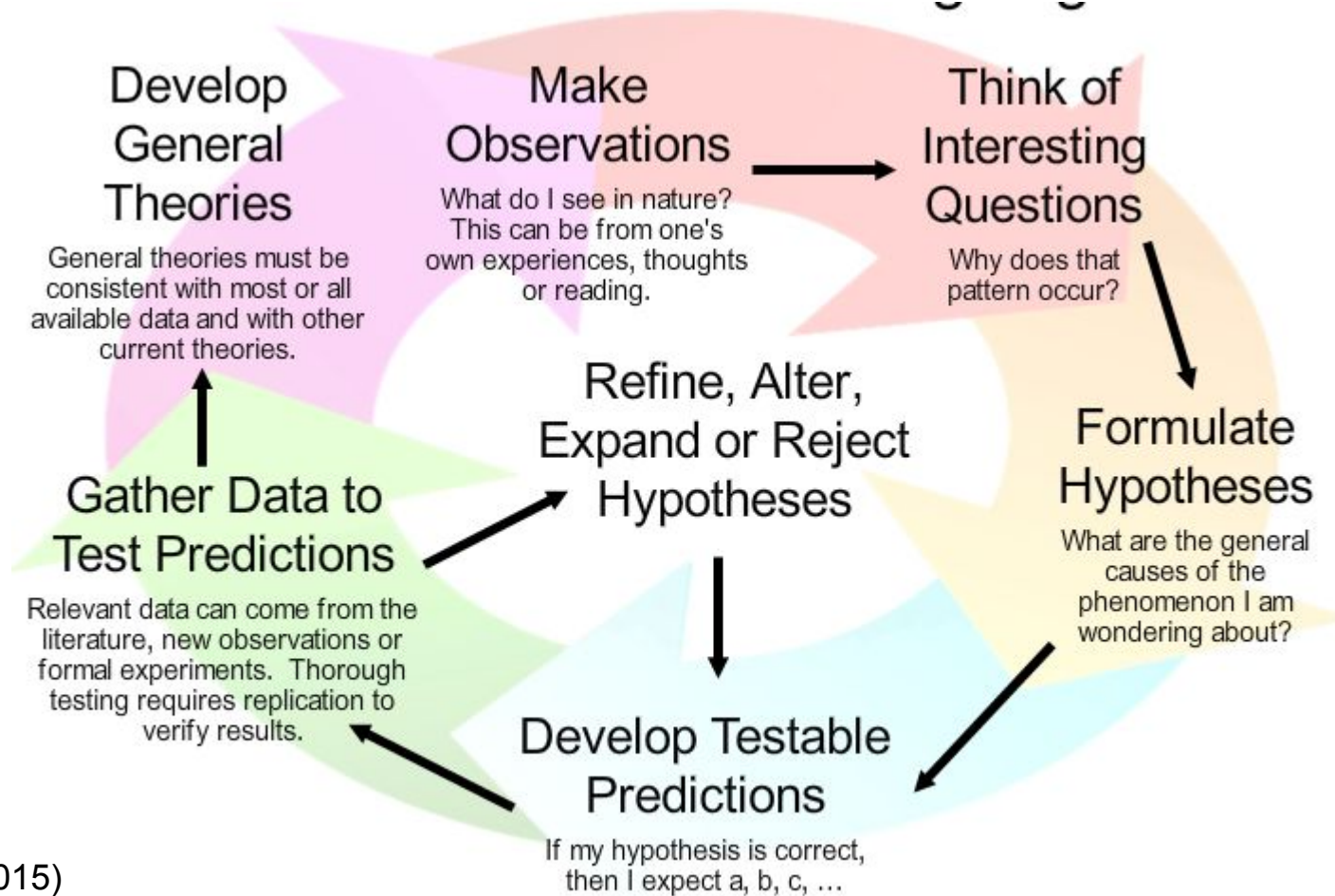
Probability and Statistics for Data Science
CSE594 - Spring 2016

Quantitative Research Methods

Probability and Statistics for Data **Science**

CSE594 - Spring 2016

The Scientific Method



(Garland, 2015)

Hypothesis Testing

Hypothesis -- something one asserts to be true.

Classical Approach:

H_0 : *null hypothesis* -- some “default” value (usually that one’s hypothesis is false)

H_1 : *the alternative* -- usually that one’s “hypothesis” is true

Hypothesis Testing

Hypothesis -- something one asserts to be true.

Classical Approach:

H_0 : *null hypothesis* -- some “default” value (usually that one’s hypothesis is false)

H_1 : *the alternative* -- usually that one’s “hypothesis” is true

Goal: Use probability to determine if we can “reject the null” (H_0) in favor of H_1 .
“There is less than a 5% chance that the null is true” (i.e. 95% alternative is true).

Hypothesis Testing

Hypothesis -- something one asserts to be true.

Classical Approach:

H_0 : *null hypothesis* -- some “default” value (usually that one’s hypothesis is false)

H_1 : *the alternative* -- usually that one’s “hypothesis” is true

Goal: Use probability to determine if we can “reject the null” (H_0) in favor of H_1 .
“There is less than a 5% chance that the null is true” (i.e. 95% alternative is true).

Example: Hypothesize a coin is biased.

H_0 : the coin is not biased (i.e. flipping n times results in a Binomial(n , 0.5))



Hypothesis Testing

H_0 : *null hypothesis* -- some “default” value (usually that one’s hypothesis is false)

H_1 : *the alternative* -- usually that one’s “hypothesis” is true

More formally: Let X be a random variable and let R be the range of X . $R_{\text{reject}} \subset R$ is the *rejection region*. If $X \in R_{\text{reject}}$ then we reject the null.

Example: Hypothesize a coin is biased.

H_0 : the coin is not biased (i.e. flipping n times results in a Binomial(n , 0.5))



Hypothesis Testing

H_0 : *null hypothesis* -- some “default” value (usually that one’s hypothesis is false)

H_1 : *the alternative* -- usually that one’s “hypothesis” is true

More formally: Let X be a random variable and let R be the range of X . $R_{\text{reject}} \subset R$ is the *rejection region*. If $X \in R_{\text{reject}}$ then we reject the null.

in the example, if $n = 1000$, then then $R_{\text{reject}} = [0, 469] \cup [531, 1000]$

Example: Hypothesize a coin is biased.

H_0 : the coin is not biased (i.e. flipping n times results in a Binomial($n, 0.5$))



Hypothesis Testing

Example: Communities with higher population have different amounts of violent crimes (per capita) than those with lower population.

Assignment 1, Programming Problem “C) 9.”



Hypothesis Testing

Example: Communities with higher population have different amounts of violent crimes (per capita) than those with lower population.

Assignment 1, Programming Problem “C) 9.”



$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent,
same variance)
t statistic for 2 samples

Hypothesis Testing

Degrees of Freedom: the number of values that are free to vary

The number of observations available to measure a parameter in a distribution. In other words, what is the minimum i , such that given i observations one could determine the parameter?

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent,
same variance)
t statistic for 2 samples

Hypothesis Testing

Degrees of Freedom: the number of values that are free to vary

The number of observations available to measure a parameter in a distribution. In other words, what is the minimum i , such that given i observations one could determine the parameter?

Statistical test is asking about generalizability to the population (or if we had infinite data).

Examples: mean, variance

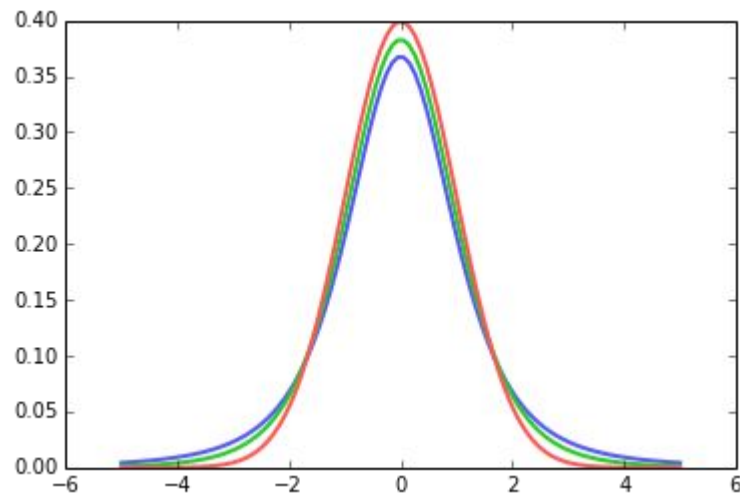
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent,
same variance)
t statistic for 2 samples

Hypothesis Testing

t-test: comparing means of distributions

Remember, t identifies an x
in a distribution
(Student's t distribution)
 $P(T < t; df)$



$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent,
same variance)

Hypothesis Testing

t-test: comparing means of distributions

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{(s_1^2/n_1)^2}{df_1} + \frac{(s_1^2/n_1)^2}{df_2}}$$

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

(assuming independent,
different variance)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent,
same variance)

Hypothesis Testing

t-test: comparing means of distributions

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{(s_1^2/n_1)^2}{df_1} + \frac{(s_1^2/n_1)^2}{df_2}}$$

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

(assuming independent,
different variance)

$$t = \frac{\bar{X}_1 - \mu_0}{\frac{s_1}{\sqrt{n_1}}}$$

(compared to
theoretical mean)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent,
same variance)

Hypothesis Testing

Important logical question:

Does failure to reject the null mean the null is true?



Hypothesis Testing

Important logical question:

Does failure to reject the null mean the null is true?

no.



Hypothesis Testing

Important logical question:

Does failure to reject the null mean the null is true?

no.

Traditionally, one of the most common reasons to fail to reject the null:
n is too small (not enough data)

Thought experiment: If we have infinite data, can the null ever be true?



Hypothesis Testing

Important logical question:

Does failure to reject the null mean the null is true?

no.

Traditionally, one of the most common reasons to fail to reject the null:
n is too small (not enough data)

Thought experiment: If we have infinite data, can the null ever be true?

Big Data problem: “everything” is significant. Thus, consider “effect size”



Type I, Type II Errors

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

Type I, Type II Errors

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

Which is worse?



Quantitative Research Review: 3-1

- The Scientific Method
- Null Hypotheses, Alternative Hypotheses
- Defining a rejection region based on hypothesis
- T-tests
- Degrees of Freedom
- Error types

Type I, Type II Errors

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

Type I, Type II Errors

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

	H_0	H_A
Reject H_0	P(Reject H_0 H_0)	P(Reject H_0 H_1)
'Accept' H_0	P(Fail to Reject H_0 H_0)	P(Fail to Reject H_0 H_1)

Type I, Type II Errors

	H_0	H_A
<u>Reject H_0</u>	P(Reject H_0 H_0)	P(Reject H_0 H_1)

Power

significance level (“p-value”) = $P(\text{type I error}) = \mathbf{P(\text{Reject } H_0 \mid H_0)}$
(probability we are incorrect)

power = $1 - P(\text{type II error}) = \mathbf{P(\text{Reject } H_0 \mid H_1)}$
(probability we are correct)

	H_0	H_A
<u>Reject H_0</u>	$\mathbf{P(\text{Reject } H_0 \mid H_0)}$	$\mathbf{P(\text{Reject } H_0 \mid H_1)}$

Power

significance level (“p-value”) = $P(\text{type I error}) = \mathbf{P(\text{Reject } H_0 \mid H_0)}$
(probability we are incorrect)

power = $1 - P(\text{type II error}) = \mathbf{P(\text{Reject } H_0 \mid H_1)}$
(probability we are correct)

Formally, a power function of a test with rejection region, R , is:

$$\beta(\theta) = P_{\theta}(X \in R)$$

where θ is the parameters of the distribution over which R is defined.
(e.g. p, n for a binomial distribution)

Multi-test Correction

If $\alpha = .05$, and I run 40 variables through significance tests, then, by chance, how many are likely to be significant?



Multi-test Correction

If $\alpha = .05$, and I run 40 variables through significance tests, then, by chance, how many are likely to be significant?

2 (5% any test rejects the null, by chance)



Multi-test Correction

How to fix?



2 (5% any test rejects the null, by chance)

Multi-test Correction

How to fix?



What if all tests are independent?

=> “Bonferroni Correction” (α/m)



Multi-test Correction

How to fix?



What if all tests are independent?

=> “Bonferroni Correction” (α/m)

But this may over-correct.

Multi-test Correction

Benjamini-Hochberg Correction Procedure

1. Let $P_{(1)} < \dots < P_{(m)}$ denote ordered p-values

2. Define:

$$\ell_i = \frac{ia}{C_m m}, \text{ and } R = \max \{i : P_{(i)} < \ell_i\}$$

where $C_m = 1$ if p-values are independent,

$$C_m = \sum_{i=1}^m \frac{1}{i} \quad \text{otherwise}$$

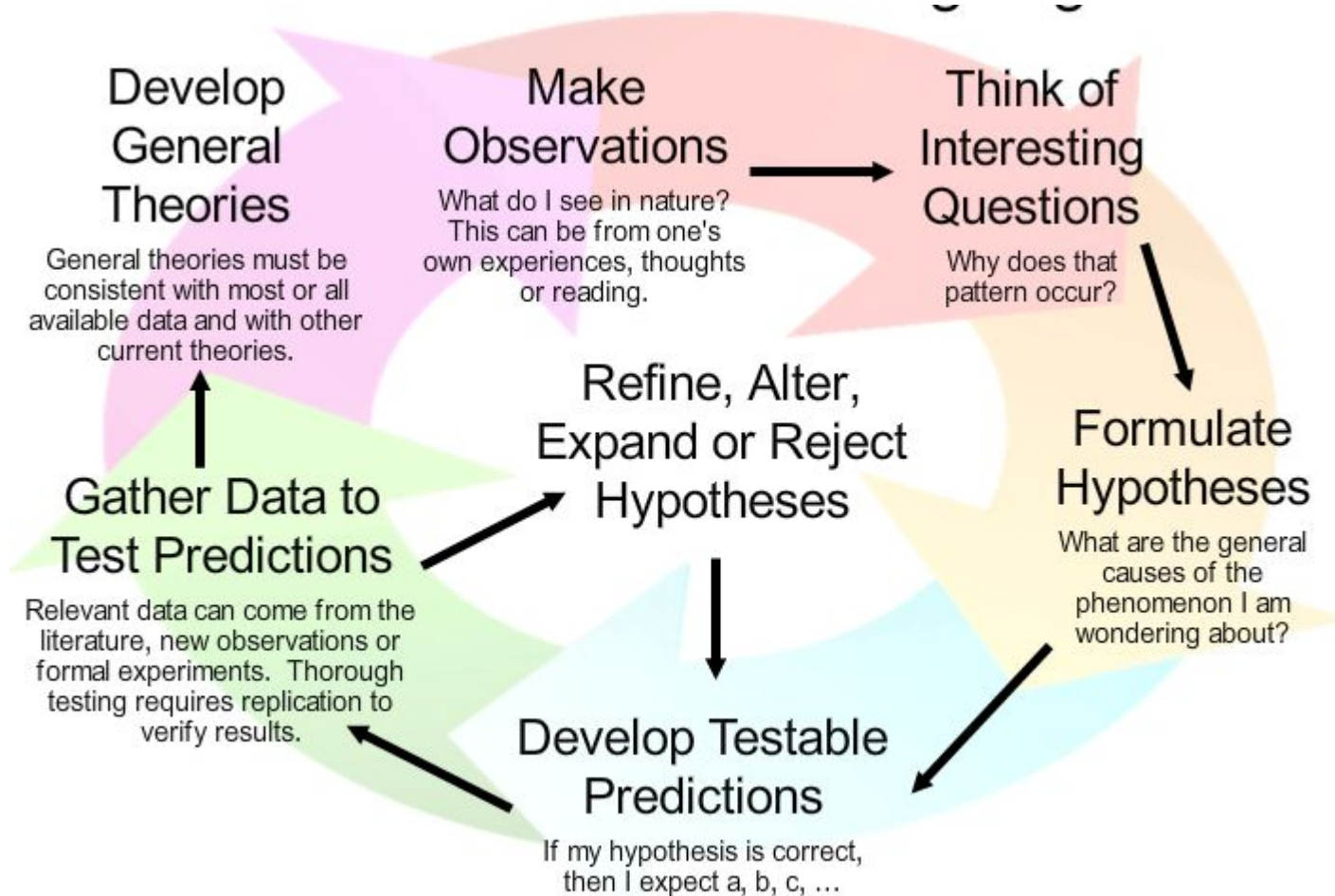
3. Let $T = P_{(R)}$, the “rejection threshold”

4. Reject all $H_{(i)}$ for which $P_i \leq T$

(Weiss, 2005)

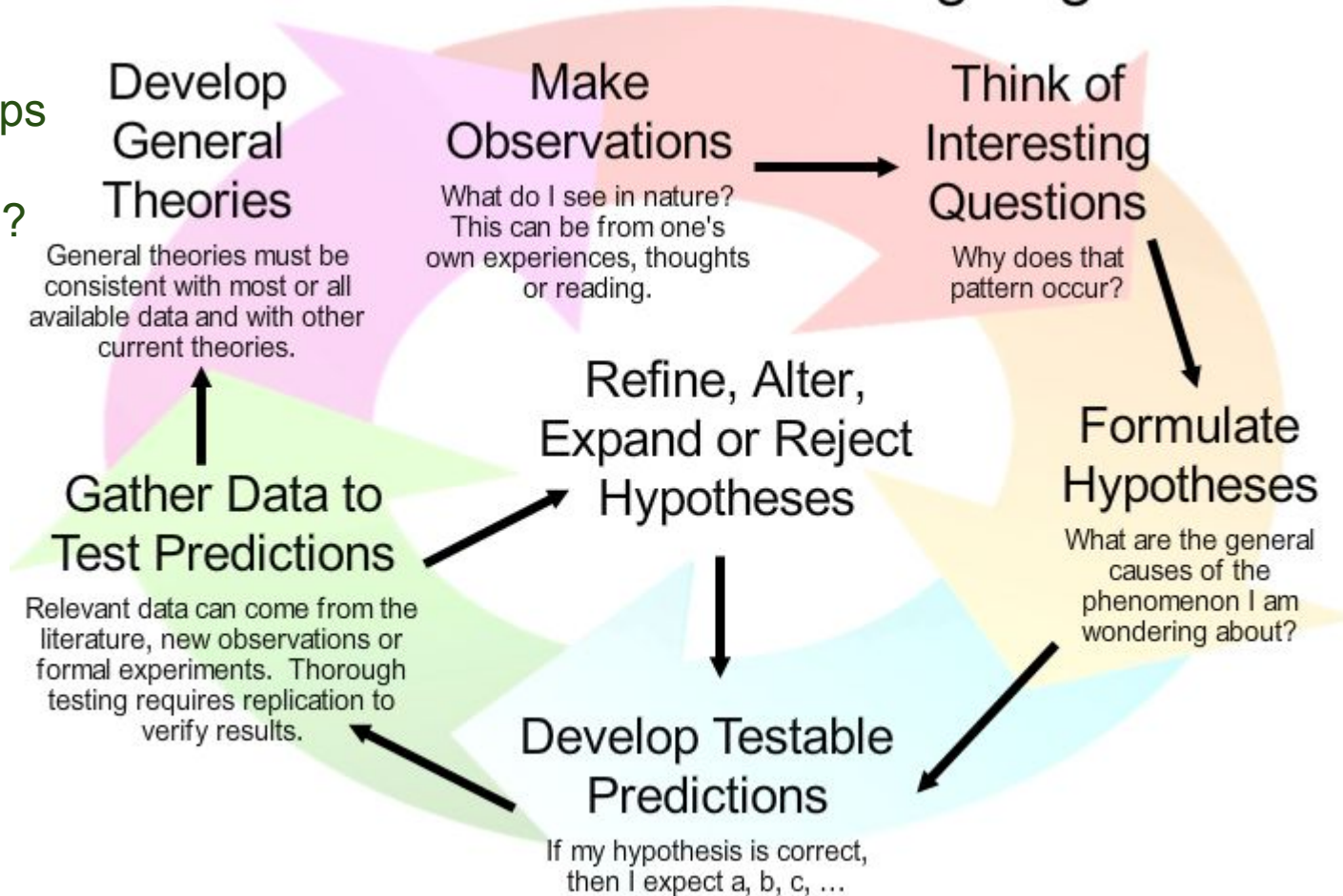
But this may over-correct.

The Scientific Method



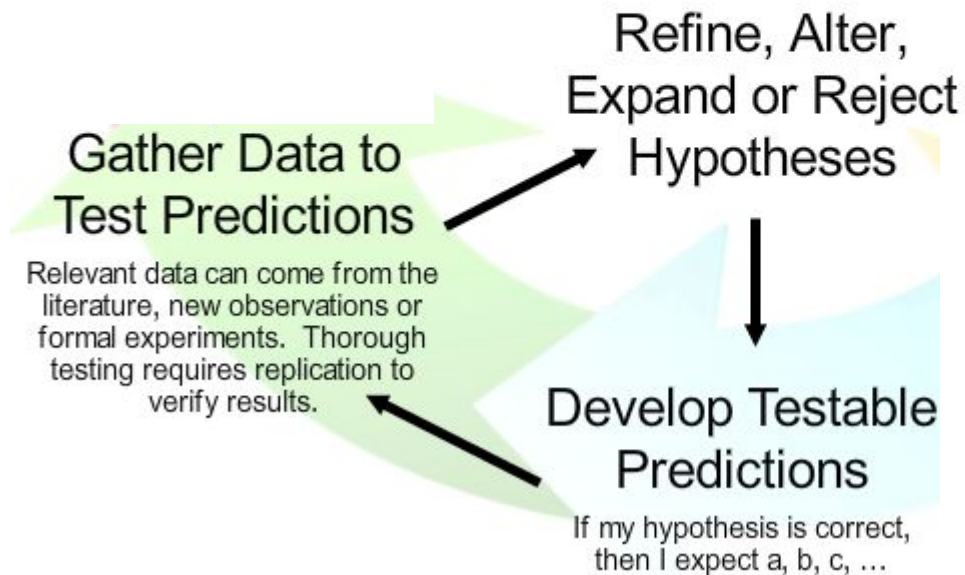
The Scientific Method

Which steps are most subjective?



The Scientific Method

Potential Effect from Big Data



Hypothesis Testing

Terminology: “tails” -- is the rejection region made up of one or two sides of the rejection region?

Example: Comparing two means:

- **two-tailed p-value:** $P(T > |t| \text{ or } T < -|t|) = 2 * P(T > |t|)$?
(when there is no assumption of direction of difference)
- **one-tailed p-value:** $P(T > t)$? (when H_a posits the second mean is greater)
 $P(T < t)$? (when H_a posits the second mean is less)

Resampling Techniques

“nonparametric” tests

The permutation test:

- t_{obs} = Compute observed score
- passes = 0
- for 1 to B :
 - randomly permute the data, keeping the same sizes per class
 - t_B = compute score on permuted data
 - if $t_B >$ (or $<$) t_{obs} : passes+=1
- p_value = passes/ B

Application: comparing two distributions, especially when they are unknown.



Linear Regression

Finding a linear function based on X to best yield Y .

X = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

Y = “response variable” = “outcome” = “dependent variable”

Regression: $r(x) = E(Y|X = x)$

goal: estimate the function r

Linear Regression

Finding a linear function based on X to best yield Y .

X = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

Y = “response variable” = “outcome” = “dependent variable”

Regression: $r(x) = E(Y|X = x)$

goal: estimate the function r

Linear Regression (univariate version): $r(x) = \beta_0 + \beta_1 x$

goal: find β_0, β_1 such that $r(x) \approx E(Y|X = x)$

Linear Regression

Simple Linear Regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
where $\mathbf{E}(\epsilon_i|X_i) = 0$ and $\mathbf{V}(\epsilon_i|X_i) = \sigma^2$

$$r(x) = \beta_0 + \beta_1 x$$

Linear Regression

Simple Linear Regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

where $\mathbf{E}(\epsilon_i|X_i) = 0$ and $\mathbf{V}(\epsilon_i|X_i) = \sigma^2$

intercept slope error expected variance

Linear Regression

Simple Linear Regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $E(\epsilon_i|X_i) = 0$ and $V(\epsilon_i|X_i) = \sigma^2$

intercept

slope

error

expected variance

Estimated intercept and slope: $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{Y}_i = \hat{r}(X_i)$$

Residual: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$



Linear Regression

Simple Linear Regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $E(\epsilon_i|X_i) = 0$ and $V(\epsilon_i|X_i) = \sigma^2$

intercept slope error expected variance

Estimated intercept and slope: $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{Y}_i = \hat{r}(X_i)$$

Residual: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$: which minimizes the residual sum of squares:

$$RSS = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

Linear Regression

via Gradient Descent

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:

Calculate all \hat{Y}_i

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left(\sum_{i=1}^n \hat{Y}_i - Y_i \right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha \left(\sum_{i=1}^n X_i (\hat{Y}_i - Y_i) \right)$$

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the residual sum of squares:

$$RSS = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

Linear Regression

via Gradient Descent

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:

Calculate all \hat{Y}_i

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left(\sum_{i=1}^n \hat{Y}_i - Y_i \right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha \left(\sum_{i=1}^n X_i (\hat{Y}_i - Y_i) \right)$$

Learning rate

Based on derivative of *RSS*

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the residual sum of squares:

$$RSS = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

Linear Regression

via Gradient Descent

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:

Calculate all \hat{Y}_i

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left(\sum_{i=1}^n \hat{Y}_i - Y_i \right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha \left(\sum_{i=1}^n X_i (\hat{Y}_i - Y_i) \right)$$

via Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the residual sum of squares:

$$RSS = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}((X - \bar{X})(Y - \bar{Y})) \end{aligned}$$

**via Direct Estimates
(normal equations)**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X, Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}((X - \bar{X})(Y - \bar{Y})) \end{aligned}$$

Correlation

$$\begin{aligned} r = r_{X,Y} &= \frac{Cov(X, Y)}{s_X s_Y} \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \end{aligned}$$

via Direct Estimates
(normal equations)

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X, Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}((X - \bar{X})(Y - \bar{Y})) \end{aligned}$$

Correlation

$$\begin{aligned} r = r_{X,Y} &= \frac{Cov(X, Y)}{s_X s_Y} \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \end{aligned}$$

via Direct Estimates (normal equations)

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

If one standardizes X and Y (i.e. subtract the mean and divide by the standard deviation) before running linear regression, then: $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = r$

Multiple Linear Regression

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{m1} + \epsilon_i$

If we include and $X_{0i} = 1$ for all i (i.e. adding the intercept to X). Then we can say:

$$Y_i = \sum_{j=0}^m \beta_j X_{ij} + \epsilon_i$$

Multiple Linear Regression

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$

If we include $X_{0i} = 1$ for all i . Then we can say:

$$Y_i = \sum_{j=0}^m \beta_j X_{ij} + \epsilon_i$$

Or in vector notation
across all i : $Y = X\beta + \epsilon$

Where β and ϵ are vectors and
 X is a matrix.

Multiple Linear Regression

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{m1} + \epsilon_i$

If we include and $X_{0i} = 1$ for all i . Then we can say:

$$Y_i = \sum_{j=0}^m \beta_j X_{ij} + \epsilon_i$$

Or in vector notation
across all i : $Y = X\beta + \epsilon$

Where β and ϵ are vectors and
 X is a matrix.

Estimating β :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Multiple Linear Regression

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$

If we include $X_{0i} = 1$ for all i . Then we can say:

$$Y_i = \sum_{j=0}^m \beta_j X_{ij} + \epsilon_i$$

To test for significance of individual Coefficient, j :

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{s^2}{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}}}$$

Or in vector notation

across all i : $Y = X\beta + \epsilon$

Where β and ϵ are vectors and X is a matrix.

Estimating β :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv \underbrace{\mathbf{P}(Y_i = 1 | X = x)} = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

Note: this is a probability here.

In simple linear regression we wanted an expectation:

$$r(x) = E(Y | X = x)$$

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv \underbrace{\mathbf{P}(Y_i = 1 | X = x)} = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

Note: this is a probability here.

In simple linear regression we wanted an expectation:

$$r(x) = E(Y | X = x)$$

(i.e. if $p > 0.5$ we can confidently predict $Y_i = 1$)

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

$$\text{logit}(p_i) = \log \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$$

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

$$\text{logit}(p_i) = \log \left(\frac{p_i}{\boxed{1 - p_i}} \right) = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$$

$\text{P}(Y_i = 0 | X = x)$

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$$

To estimate β ,
one can use

reweighted least squares:

set $\hat{\beta}_0 = \dots = \hat{\beta}_m = 0$ (remember to include an intercept)

1. Calculate p_i and let W be a diagonal matrix

where $\text{element}(i, i) = p_i(1 - p_i)$.

2. Set $z_i = \text{logit}(p_i) + \frac{Y_i - p_i}{p_i(1 - p_i)} = X\hat{\beta} + \frac{Y_i - p_i}{p_i(1 - p_i)}$

3. Set $\hat{\beta} = (X^T W X)^{-1} X^T W z$ // weighted lin. reg. of Z on Y .

4. Repeat from 1 until $\hat{\beta}$ converges.

(Wasserman, 2005; Li, 2010)

Review: 3-3, 3-10

- Power
- Multi-test Correction: Bonferroni and Benjamini-Hochberg
- The Permutation Test
- “Tails”
- Regression goal and terminology
- Simple Linear Regression (Residual Sum of Squares)
- Multiple Linear Regression
- P-values for linear regression coefficients
- Logistic Regression: reweighted least squares

A lot can be answered with multiple linear regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$$

Mediation

Path Analyses (a type of “structured equation modeling”)

How much does **M** mediate the effect of **X** on **Y**?

$$Y = \beta_0 + c'X + bM + \epsilon$$

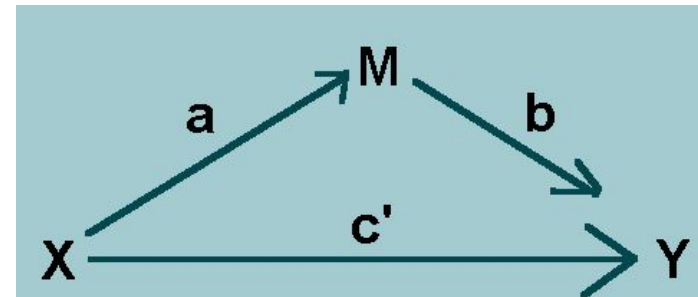
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$$

Mediation

Path Analyses (a type of “structured equation modeling”)

How much does **M** mediate the effect of **X** on **Y**?

$$Y = \beta_{0c} + cX + \epsilon_c$$
$$X = \beta_{0a} + aM + \epsilon_a$$
$$Y = \beta_{0c'b} + c'X + bM + \epsilon_{c'b}$$



(Kenney, 2015)

<http://davidakenny.net/cm/mediate.htm>

Mediation

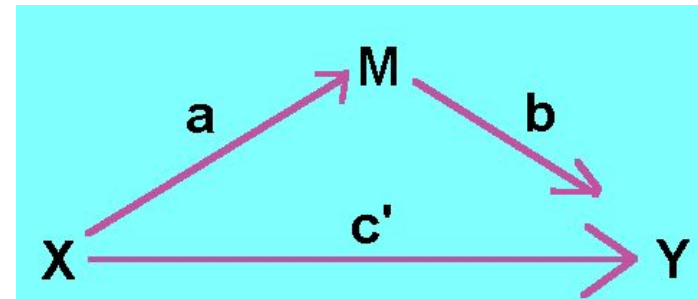
Path Analyses (a type of “structured equation modeling”)

How much does **M** mediate the effect of **X** on **Y**?

$$Y = \beta_{0c} + cX + \epsilon_c$$
$$X = \beta_{0a} + aM + \epsilon_a$$
$$Y = \beta_{0c'b} + c'X + bM + \epsilon_{c'b}$$

Effect size: often reported as $c - c'$.

Used for *basic* causal inference.

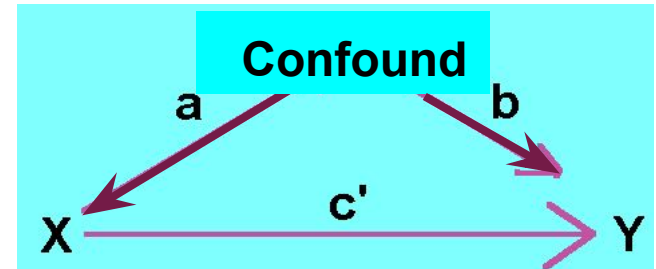


(Kenney, 2015)

<http://davidakenny.net/cm/mediate.htm>

Hierarchical Linear Models (HLM)

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.



Hierarchical Linear Models (HLM)

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.

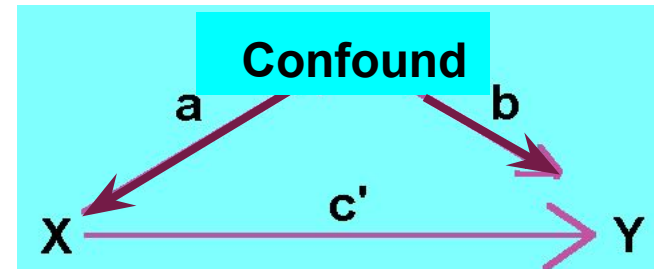
Examples:

Pot heads are more likely to say “hella”
but really californians are more like to say “hella” and be potheads.

X = use of “hella”

Y = pot-head or not

Confound = from california?



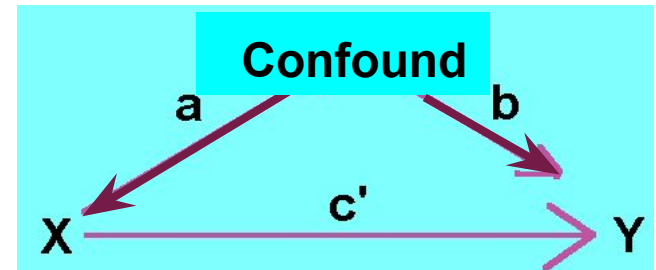
Hierarchical Linear Models (HLM)

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.

Examples:

Pot heads are more likely to say “hella”
but really californians are more like to say “hella” and be potheads.

Females are more likely to post pictures of food
but really both food posts and females are more common on Pinterest.



Hierarchical Linear Models (HLM)

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.

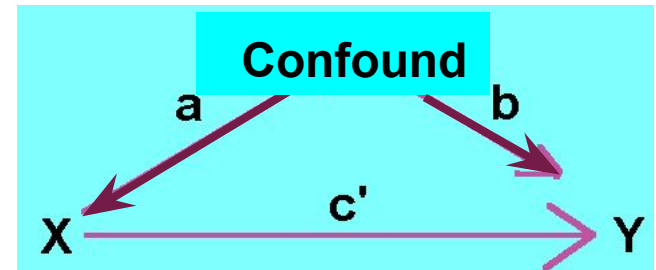
Examples:

Pot heads are more likely to say “hella”
but really californians are more like to say “hella” and be potheads.

Females are more likely to post pictures of food
but really both food posts and females are more common on Pinterest.

Solution: include aggregate confounding variable as a covariate in multiple linear regression. (also useful for prediction)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{m1} + \epsilon_i$$



Hierarchical Linear Models (HLM)

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.

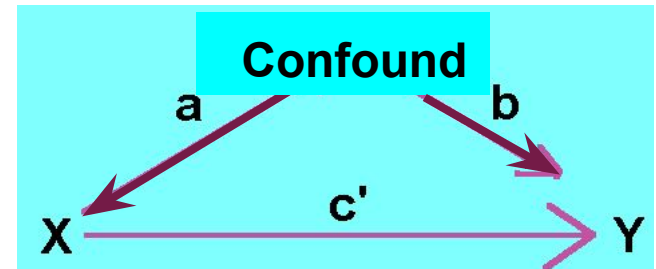
Examples:

Pot heads are more likely to say “hella”
but really californians are more like to say “hella” and be potheads.

Females are more likely to post pictures of food
but really both food posts and females are more common on Pinterest.

Solution: include aggregate confounding variable as a covariate in multiple linear regression. (also useful for prediction)

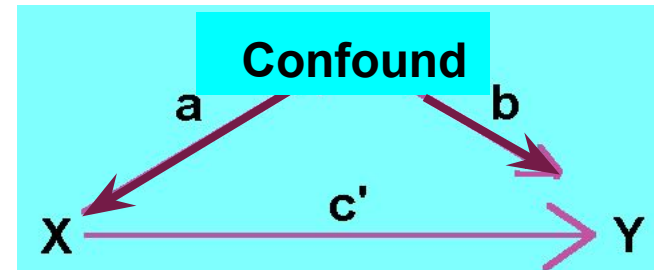
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{m1} + \epsilon_i$$



Hierarchical Linear Models (HLM)

Problem: Sometimes variables at one unit of analysis are *confounded* by a variable at another level.

$$Y = \beta_0 + \beta_1 X_1 + \beta_A A + \epsilon$$



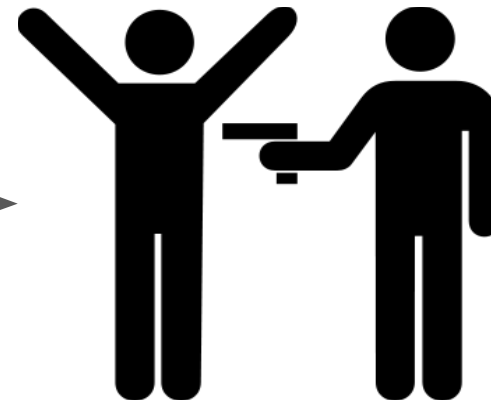
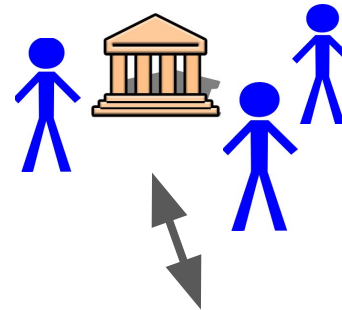
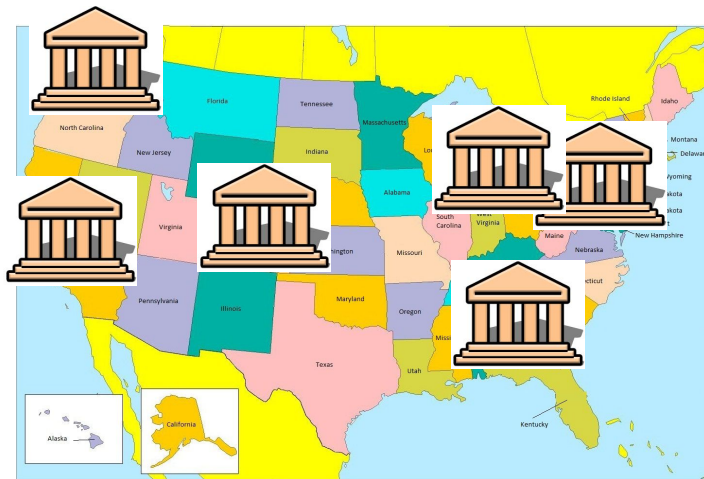
A : aggregate indicator variable (is in region or not? Pinterest usage).

Solution: include aggregate confounding variable as a covariate in multiple linear regression. (also useful for prediction)

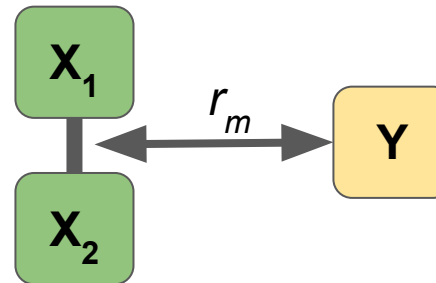
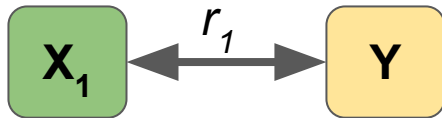
Ecological Fallacy

The assumption that an effect at one unit of analysis will hold for a smaller or larger unit of analysis.

Example:



Moderation (interaction)



When $r_1 \neq r_m$, X_2 moderates the relationship between X_1 and Y .

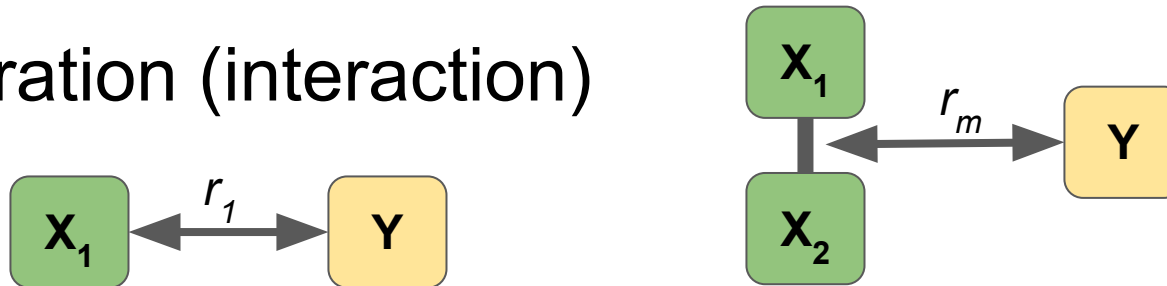
Examples:

Y : Attend church? X_1 : Agreeableness, X_2 : From US?

Movie Reviews:

Y : Rated Depressing, X_1 : "death" in review, X_2 : Silly Horror Movie?

Moderation (interaction)



When $r_1 \neq r_m$, X_2 moderates the relationship between X_1 and Y .

More precisely moderation analyses fit the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_M X_1 X_2 + \beta_2 X_2 + \epsilon$$

$X_1 X_2$: The interaction term. (Element-wise multiplication)

β_M can then be tested for significance using the same t-test we use for any individual coefficient in multiple linear regression

Mediation, Moderation Code Examples



Review: 3-22

- Mediation, path models
- HLM
- Moderation
- Ecological Fallacy

Discrete Variable Comparison Metrics

Examples:

Single class:

- X_1 : Smoker or not(0/1) X_2 : has cancer? (0/1)
- Y : Picture of goat? (0/1) \hat{Y} : prediction from a logistic model (0/1)
or any model (e.g. a *gradient boosting deep bayes neural forest*)

Multi-class

- Y : word is subject, direct object, or indirect object (1, 2, or 3 but order means nothing)
 \hat{Y} : prediction from a multi-class model
(a “multinomial” distribution)

Discrete Variable Comparison Metrics

- Chi-Square test for independence
- (true|false) (positive|negative) based metrics:

Discrete Variable Comparison Metrics

Single class:

- X_1 : Smoker or not(0/1) X_2 : has cancer? (0/1)

N = 100 people sampled from cancer screening center population

	no cancer	cancer	
not smoker	60	10	
smoker	22	8	

Discrete Variable Comparison Metrics

Single class:

- X_1 : Smoker or not(0/1) X_2 : has cancer? (0/1)

N = 100 people sampled from cancer screening center population

	no cancer	cancer	
not smoker	60	10	
smoker	22	8	

Chi-Squared Test for Independence

H_0 : Y and Z are independent

H_1 : Y and Z are dependent

$$U = \sum_{i=0}^{classes_{x1}} \sum_{j=0}^{classes_{x2}} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

where $E_{ij} = \frac{X_{i*} X_{*j}}{n}$

	no cancer	cancer	
not smoker	60	10	70
smoker	22	8	30
	82	18	100

Chi-Squared Test for Independence

H_0 : Y and Z are independent

H_1 : Y and Z are dependent

$$U = \sum_{i=0}^{classes_{x1}} \sum_{j=0}^{classes_{x2}} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

where $E_{ij} = \frac{X_{i*} \cdot X_{*j}}{n}$

	no cancer	cancer		<i>Expected distribution</i>	
not smoker	60	10	70	$70 \cdot 82 / 100 = 57.4$	12.6
smoker	22	8	30	24.6	5.4
	82	18	100		

Chi-Squared Test for Independence

$$k = df \text{ (degrees of freedom)} = (classes_{x1} - 1)(classes_{x2} - 1)$$

H_0 : Y and Z are independent


H_1 : Y and Z are dependent

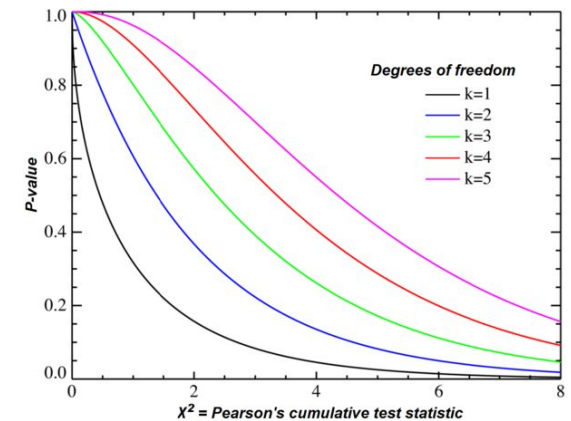
$$U = \sum_{i=0}^{classes_{x1}} \sum_{j=0}^{classes_{x2}} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

where $E_{ij} = \frac{X_{i*} \cdot X_{*j}}{n}$

Observed count: X_{ij}

Expected count: E_{ij}





	no cancer	cancer		Expected distribution	
not smoker	60	10	70	$70 \cdot 82 / 100 = 57.4$	12.6
smoker	22	8	30	24.6	5.4
	82	18	100		

Discrete Variable Comparison Metrics

- Chi-Square test for independence
- (true|false) (positive|negative) based metrics:

Discrete Variable Comparison Metrics

- Chi-Square test for independence
- (true|false) (positive|negative) based metrics:

		True condition			
		Condition positive	Condition negative		
Total population				Prevalence = $\frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	(Thank you, Wikipedia!)
Predicted condition	Predicted condition positive	True positive	False positive (Type I error)	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Test outcome positive}}$	
	Predicted condition negative	False negative (Type II error)	True negative	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Test outcome negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Test outcome negative}}$
Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$		True positive rate (TPR), Sensitivity, Recall = $\frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out = $\frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	

Discrete Variable Comparison Metrics

- Chi-Square test for independence
- **(true|false) (positive|negative) based metrics:**

		True condition		Prevalence = $\frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	
		Condition positive	Condition negative		
Predicted condition	Predicted condition positive	True positive	False positive (Type I error)	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Test outcome positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{False positive}}{\Sigma \text{Test outcome positive}}$
	Predicted condition negative	False negative (Type II error)	True negative	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Test outcome negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Test outcome negative}}$
Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$		True positive rate (TPR), Sensitivity, Recall = $\frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out = $\frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	