Introduction to Statistics for the Social Sciences

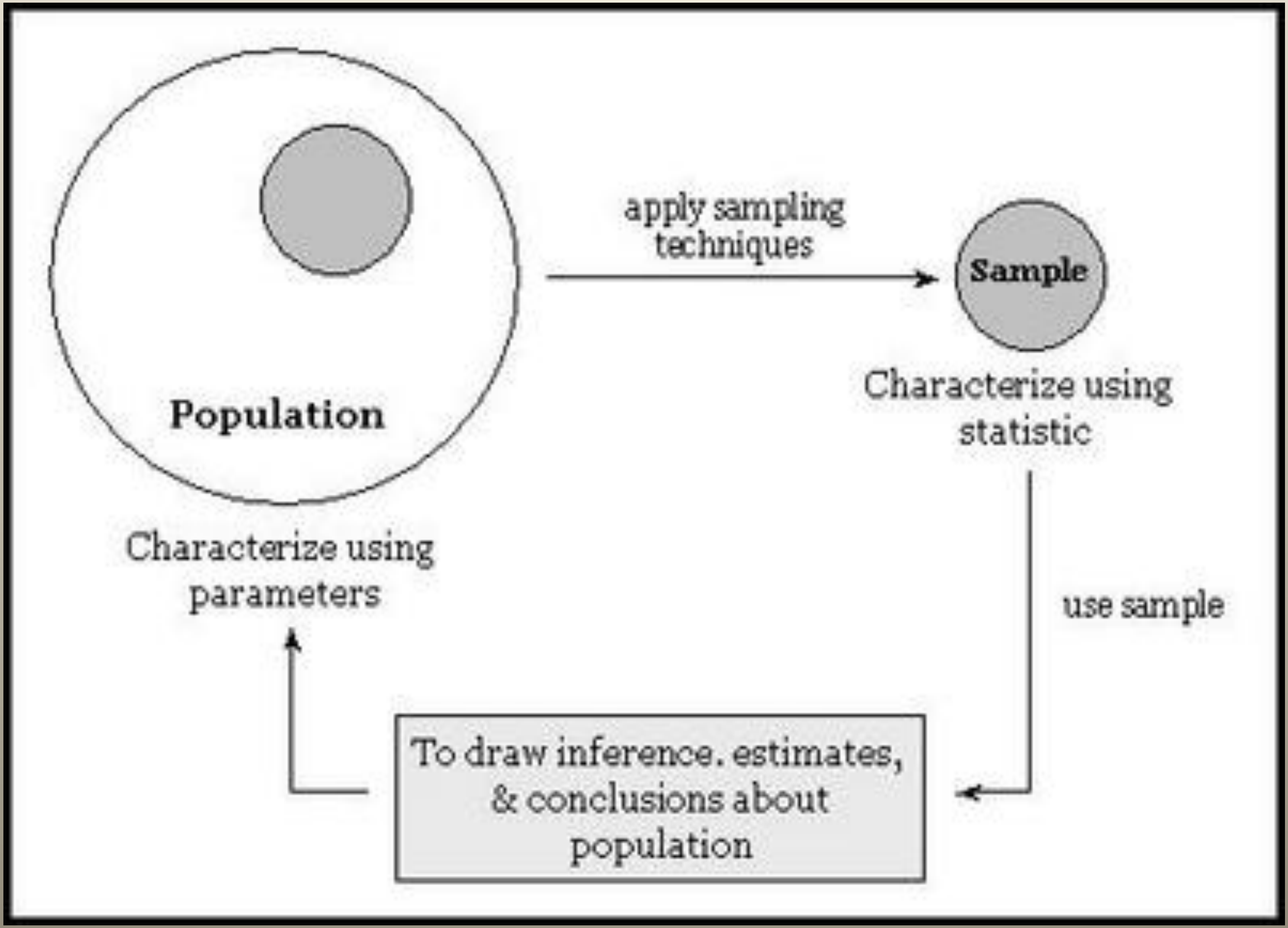
Kyle Thomas

10/13/2010

kathomas@fas.harvard.edu

Populations and Samples

- Population—The entire collection of relevant people or events
  - E.g., Everyone in the United States, everyone in a class, everyone with a disorder, etc.
- Parameters—Measures of populations (Greek)
- Sample—Some subset of an entire collection of relevant people or events
  - E.g., Some Americans, one class out of a high school, some people with a disorder, etc.
- Statistics—Measures calculated from samples (Roman—our alphabet)



Population

Characterize using parameters

apply sampling techniques

Sample

Characterize using statistic

use sample

To draw inference. estimates, & conclusions about population

Types of Variables

- Independent variable—The variable that is manipulated experimentally, or pseudo-experimentally
  - E.g., drug vs. placebo, gender, etc.
- Dependent variable—The “outcome variable,” the variable that is measured to observe the hypothesized effect
  - E.g., test score, pain tolerance, rating of something, etc.

Types of Measurement

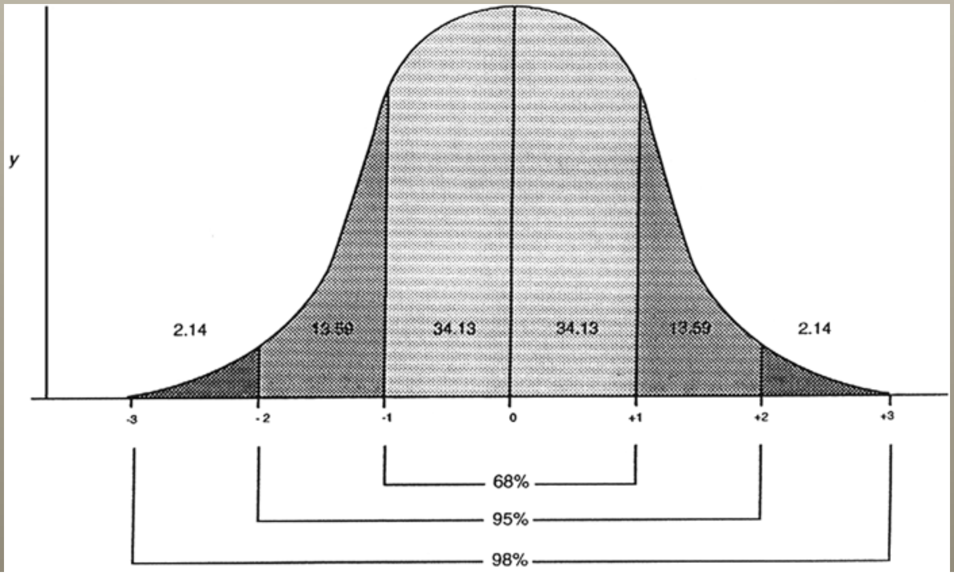
- How you measure something impacts what kinds of analysis you can do
- Nominal (categorical)—variables measured as categories
  - E.g., ethnicity, gender, location, etc.
- Ordinal (ranked)—variables can be ranked in some meaningful way, but gaps between values are uneven
  - E.g., ranking in race, class rank
- Must use *non-parametric* tests for these if they are the dependent variable

Types of Measurement

- Interval—Variables measured numerically where the distance between points is equal
  - E.g., 7-point scales, Fahrenheit temperature, etc.
- Ratio-scale—Same as interval, but with a meaningful 0-point
  - Allow statements like "subject A did twice as well as subject B"
  - E.g., GPA, Kelvin temperature, etc.
- Dependent variable needs to be measured at one of these levels for *parametric* statistics to be used!

Distributions and Parametric Statistics

- Distribution—A collection of scores
  - Parametric statistics assume a *normal distribution*
  - Normal distribution is a mathematical abstraction, never exactly exists, but stats are robust to violations of this assumption



Descriptive Statistics

- Measures that *describe* distributions
- Measures of *central tendency*
  - Mean—Average of a sample ($\overline{X}$) or a population (μ)
    - The mean of a population is typically the best guess for any given score in that population
    - Similarly, the mean of a sample is the best guess for the mean of the population it comes from
  - Median—The value that divides the distribution into 2 equal size pieces
  - Mode—The most common number in a distribution

Descriptive Statistics

- Measures of *dispersion*, or how "spread out" the data are
- Variance—A measure of dispersion, how much scores are spread out in a distribution

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

- Standard Deviation—The "average amount" that scores vary from the mean
  - Square root of variance $\sigma = \sqrt{\sigma^2}$

Inferential Statistics

- These are used when one wants to do more than just describe something
- They are used to make *inferences* about populations from samples
  - Populations can be real (e.g., all Harvard students)
  - Or hypothetical (e.g., all depressed people that take Prozac)
- Usually in research we want to know about populations, but we can only measure samples
  - Inferential statistics tell us if what we observe about samples *generalizes* to populations

Null Hypothesis Statistical Testing

- Usually in research we are testing hypotheses about how groups relate to each other
- To do this we must use NHST
- The question NHST allows us to ask is, "what is the probability that a sample was drawn from a specific population?"
- Or, more commonly, "what is the probability that multiple samples were drawn from the same population?"
  - If this probability is low we assume there is some meaningful difference between groups

NHST

- Usually in research we want to know if two groups are the same, if some manipulation has an effect, etc.
  - This is our research hypothesis-- $\mu_1 \neq \mu_2$
- These questions must be mapped onto the statistical logic of populations and samples
  - E.g., If the drug does have an effect these two groups would be two different populations (because they are different on our measure)
  - Conversely, if some drug has no effect, then people that take the drug and those that do not form one population (because they are not different on our measure)

NHST

- In order to test if our research hypothesis is true, we must compare it to a null hypothesis which is different
- Null hypothesis (statistically defined)—All samples come from one population ($\mu_1 = \mu_2$)
- Null hypothesis (conceptually defined)—The groups are not different; there is no effect; etc.
  - ***This is not the same as saying the research hypothesis is false***

NHST

- First we set up our null hypothesis:
  - Mean of population 1 = Mean of population 2
- And our research hypothesis:
  - Mean of population 1 ≠ Mean of population 2
- We can never say whether our research hypothesis is true
  - Karl Popper: It is impossible to “prove” any hypothesis, we can only argue it is our best explanation for any given data
  - Statistics: One cannot 100% say that two samples come from different populations if we do not know the population distributions
- We can only evaluate how likely the null hypothesis is to be true with inferential statistics.
  - If it is unlikely that both samples came from one population, then we *assume* our research hypothesis is a better explanation of the data

Z-Score

- Used to standardize a raw score
- Tells you how many standard deviations a score is off from the mean (z-score = number of standard deviations a score is off the mean)

$$Z = \frac{X - \mu}{\sigma}$$

Z-Test

- Used when you have a sample from a population with a known mean and standard deviation (μ, σ)
- Uses Standard Error of the Mean—the standard deviation of a sampling distribution

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$$

Single Sample t-test

- Used when [null hypothesis'] population mean is known, but not population standard deviation
- Population standard deviation *estimated* from sample's standard deviation

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}}$$

$$S_{\overline{X}} = \frac{S_X}{\sqrt{N}} = \frac{\sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}}}{\sqrt{N}}$$

Independent Samples t-test

- Used when neither [null hypothesis’] population mean or standard deviation is known
- Mean estimated from control group usually, standard deviation estimated from the standard deviation of *both* samples

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_{pooled}}$$

$$S_{pooled} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Paired Samples t-test (a.k.a. dependent samples, within-subjects)

- Single sample t-test with a difference score
- A difference score is calculated by subtracting each subject's score on one item (e.g., post-test) from that subject's score on another item (e.g., pre-test)

$$t = \frac{\overline{D} - \mu_D}{S_{\overline{D}}}$$

μ = 0 by assumption

One-way ANOVA

- Compares more than 2 groups on 1 dependent variable
  - 1 categorical Independent Variable with more than 2 levels
  - 1 Interval-level Dependent Variable
- Tests omnibus null hypothesis: $\mu_1 = \mu_2 = \mu_3 \ldots$
- Contrasts and post-hoc tests done to check for specific hypotheses: e.g., $\mu_1 > \mu_2$

One-way ANOVA

- Within-group variance is an estimate of *random* variance
- Between-group variance is an estimate of *systematic* and *random* variance
- Compares variance between the different groups to the variance within the different groups

$$F = \frac{S^2_{between}}{S^2_{within}}$$

One-way ANOVA

- Compares signal to noise ratio

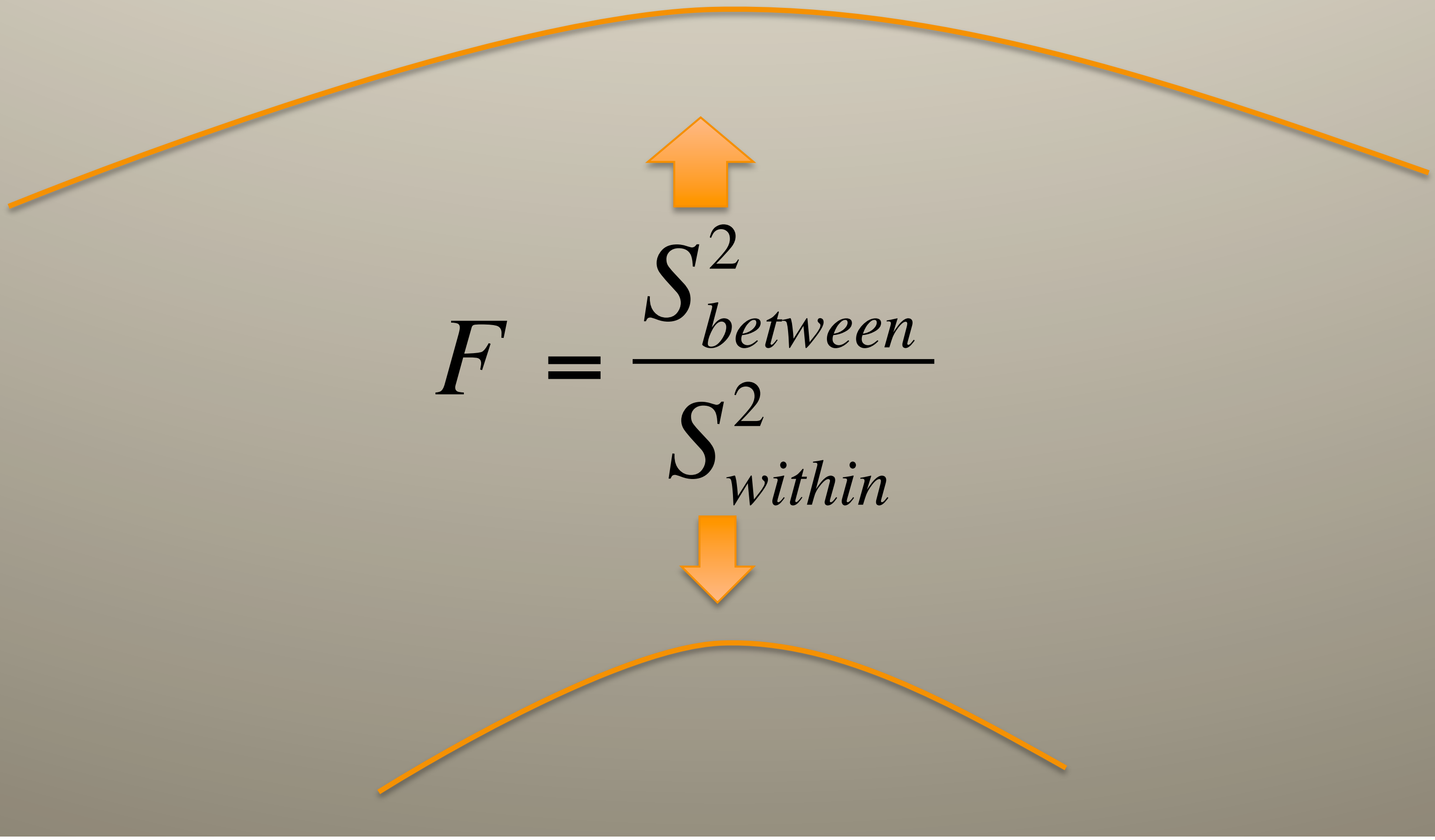
$$F = \frac{Signal + Noise}{Noise}$$

- Compares "expected" to observed

$$F = \frac{Observed}{"\,Expected"}$$

One-way ANOVA

- Compares the ratio of the populations these two measures would have likely come from



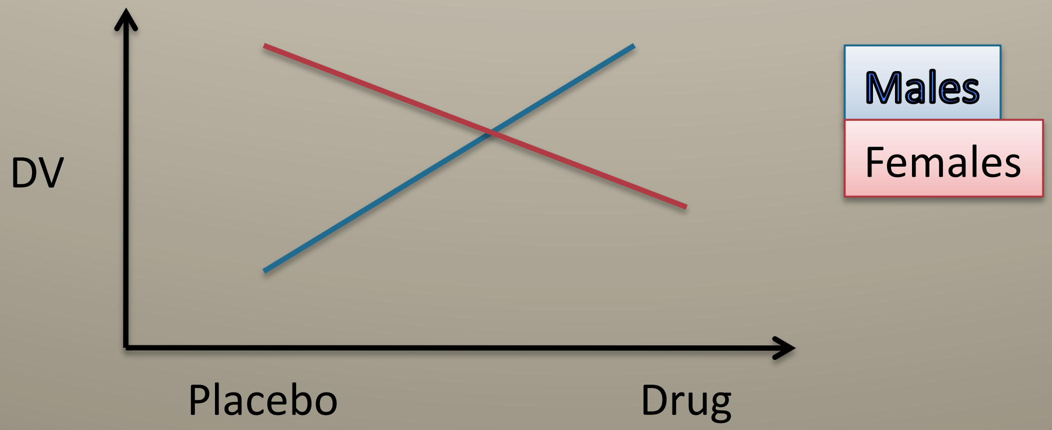
$$F = \frac{S^2_{between}}{S^2_{within}}$$

Factorial ANOVA

- Used when you have more than 1 categorical independent variable (e.g., gender and drug/placebo)
- Also tests omnibus null: $\mu_1 = \mu_2 = \mu_3 \ldots$
- But can test it for groups that vary on two or more independent variables: $\mu_{1A} = \mu_{1B} = \mu_{2A} = \mu_{2B} \ldots$
- Can also do post hoc tests and contrasts

Factorial ANOVA

- Variables must have multiple levels
- Can look at Interactions—when the effect of one variable depends upon the level of the other variable



Factorial ANOVA

- Between-Subjects—where different subjects are in different conditions
- Within-Subjects—when all subjects are in all conditions, and are compared to themselves
- Mixed-Model—Some variables are within-subjects, some are between-subjects

Non-parametric Statistics

- These must be used if your distribution is sufficiently non-normal, or if your dependent variable is not at least on an interval scale
- There are non-parametric analogs of all the previous tests, except mixed-model ANOVAs
- You can lose substantial power
- Chi-square; Mann-Whitney; Kruskal-Wallis; Sign Test; Wilcoxon; Friedman

Correlation

- Measures how two variables "track" each other, or co-vary
- Cannot infer causation
- Can be done for frequencies of nominal variables

$$r_{XY} = \frac{Cov_X Cov_Y}{\sigma_X \sigma_Y} = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{S_X S_Y (n-1)}$$

Regression

- Glorified slope intercept!
- Can take interval DVs *and* IVs

$$y = mx + b$$

$$X_{predicted} = \beta_1 X_1 + \beta_0$$

Regression coefficient

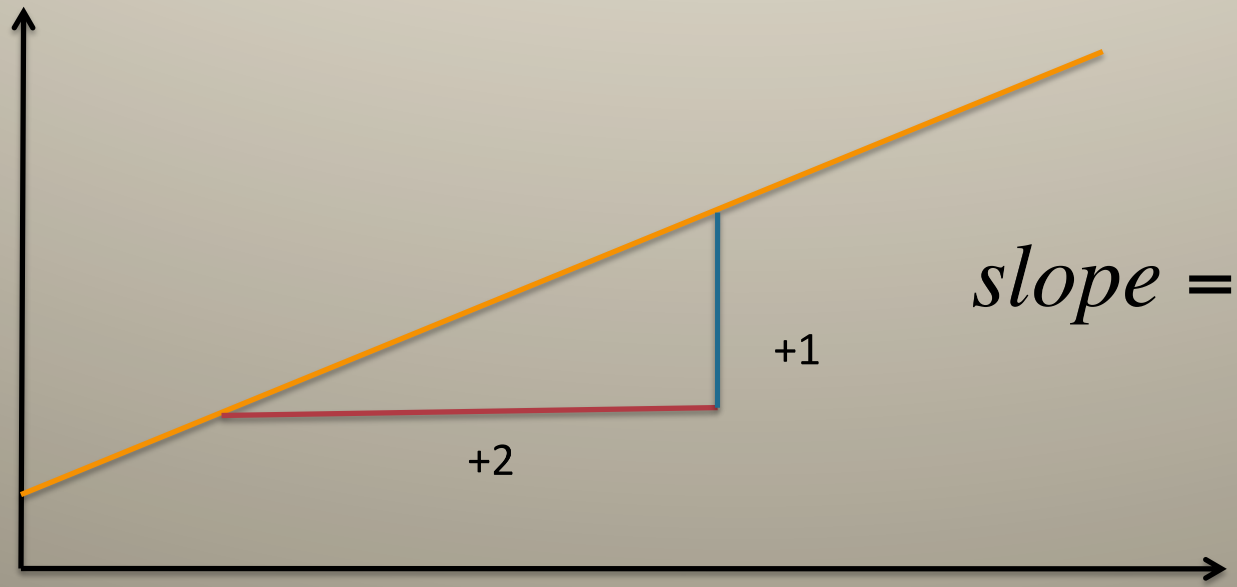
Intercept

Prediction for score
on Variable X

Score on variable 1

Regression

- Slope—How one variable changes with another (sometimes just correlation)



$$slope = \frac{rise}{run} = \frac{1}{2}$$

Regression

- Intercept—The value of our predicted variable ($X_{predicted}$) when the value of our predictor variable (X_1) is 0

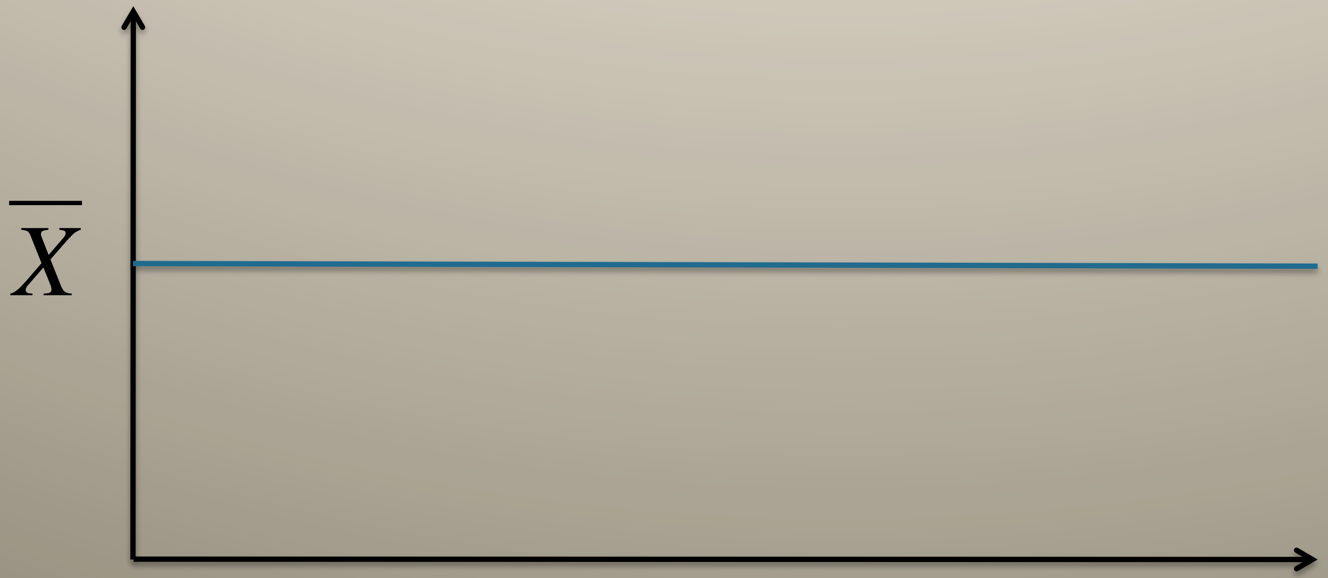
$$X_{predicted} = \beta_1 X_1 + \beta_0$$

$$X_{predicted} = \beta_1 (0) + \beta_0$$

$$X_{predicted} = \beta_0$$

Regression Logic

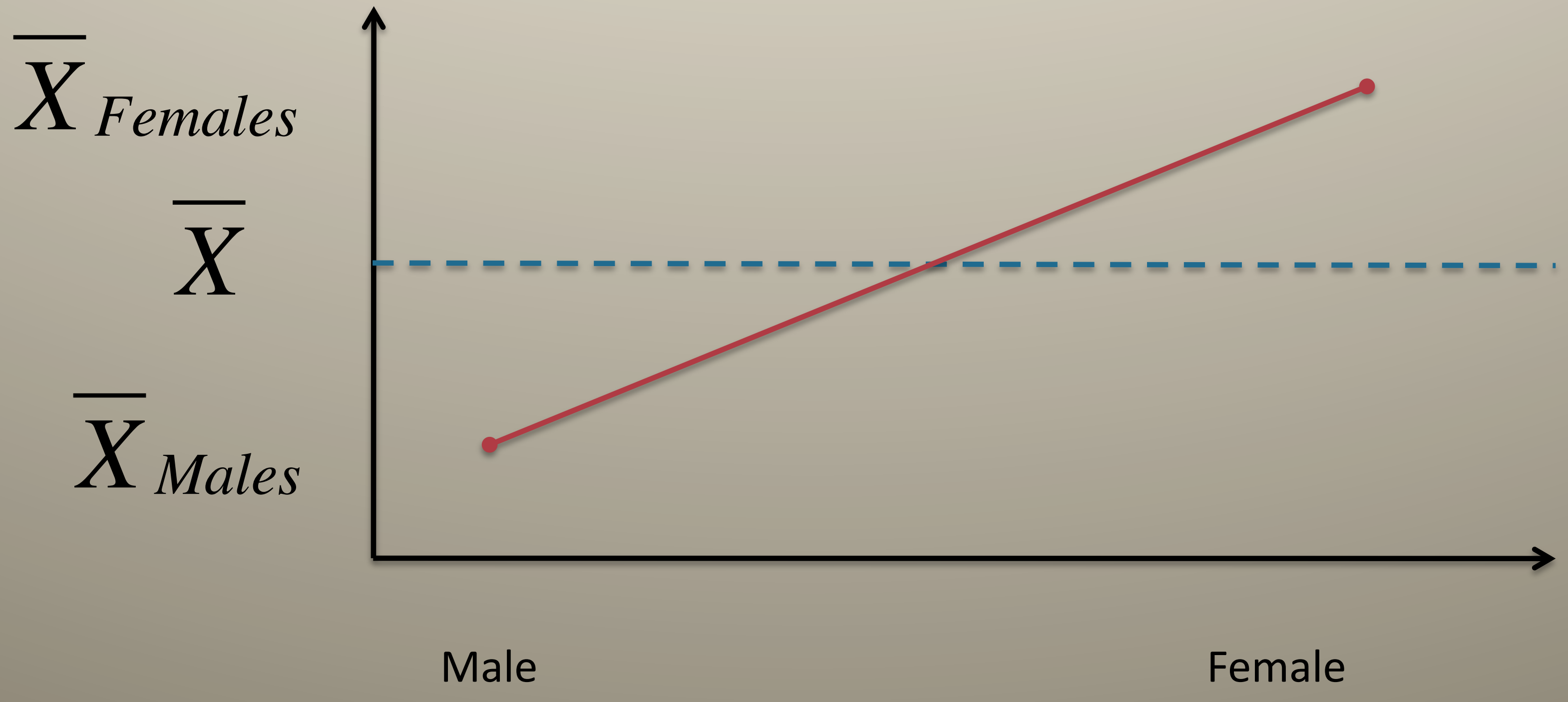
- Mean as best guess for random score, if you do not have better information



All People from Population

Regression Logic

- However, if two variables are related, you can use one to predict the value of the other one and get a better prediction than the mean



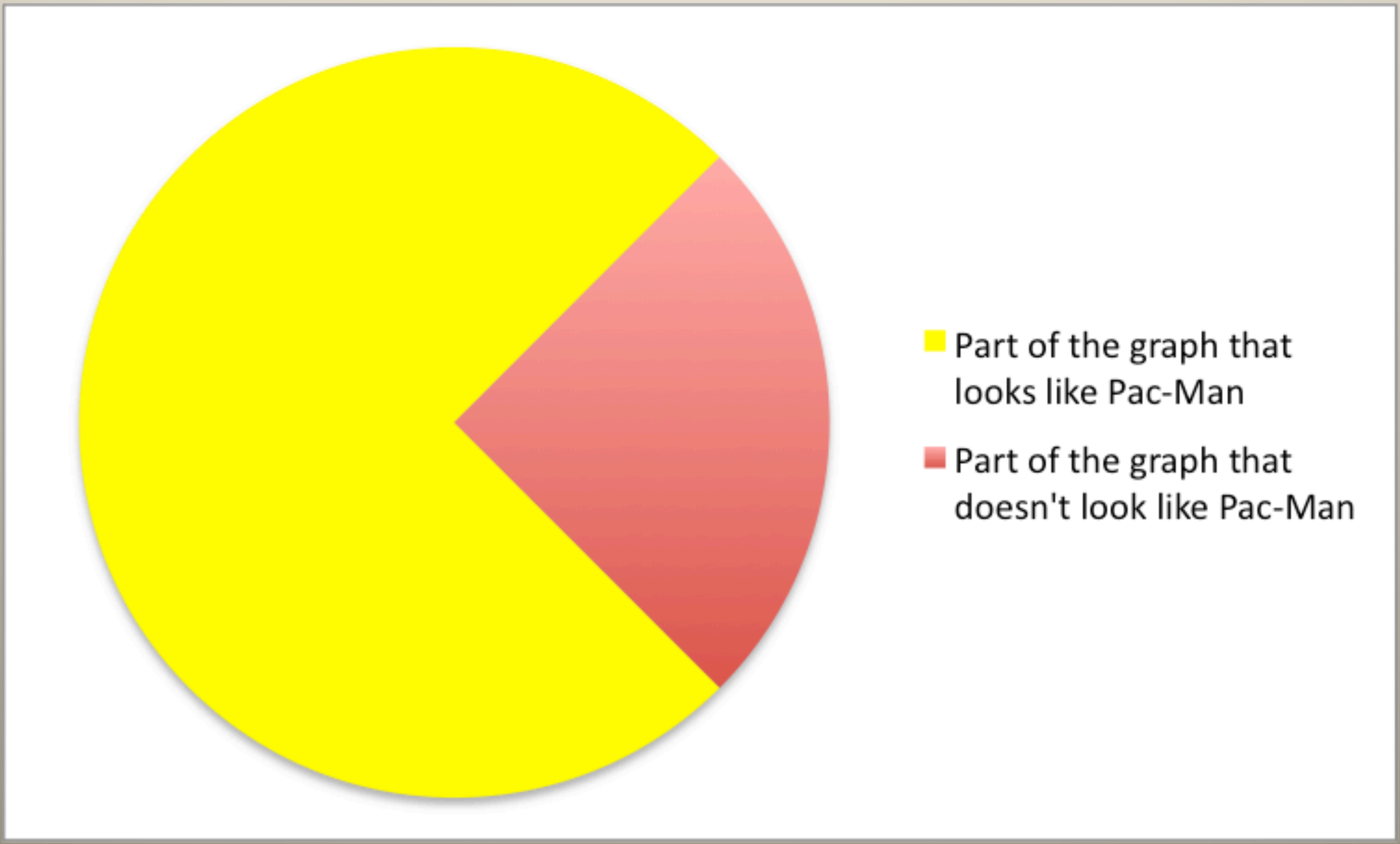
Multiple Regression

- Uses multiple independent variables to predict a dependent variable

$$X_{predicted} = \beta_1 X_1 + \beta_2 X_2 + \beta_0$$

- Mathematically, tries to capture only the *unique* contribution of each variable
  - *Caution* This cannot be done perfectly since regression works off of correlation, so more sophisticated techniques are needed to infer causality

Thanks!


Part of the graph that looks like Pac-Man
Part of the graph that doesn't look like Pac-Man