Introduction and Basic Concepts

September-03-13
09:25 AM

Empirical Research

- Why do empirical research?
  - To look for patterns (regularities) in the natural world
- Regularities are **NOT** the same as constants
  - *Constants* are measure that are always the same
- Regularities usually involve a measure that has *variability*
  - Partially true variability
    - i.e. different people have different heights
  - Partially some noise and/or measurement error
- Any kind of scientific study is done through experimentation
  - Experimentation is the observation of the effects of independent variables on dependent variables
- The word variable implies that something is changing
  - Ergo… if there is no change, there is no variable
- The Independent variable is the variable manipulated by the researcher
- The Dependent variables are the variables measured by the researcher
  - This variable is *dependant* on the independent variable (in most cases)

Why do we need Stats?

- To detect empirical regularities
- To generate new knowledge
- To describe and infer properties of the natural world

What Stats can't do…

- Statistics cannot...
  - Provide explanations
  - Prove causality
  - Predict
  - Extrapolate
- Explanations and causality depend on the researcher's interpretation based on a body of data, and can never be established with full certainty
  - If you're lucky enough to get a correlation value of 1 or -1 does that mean that you've proven that there is a relationship? Or that you have proven something with complete certainty?
- Sometimes that which is 100% accurate is also 100% useless
- Stats can't tell you if your results can generalize two different groups of individuals, different, cultures, different stimuli, different treatments, etc.

Causality vs Statistics

- A causal relation is when a single event causes another if changes in the second event are due to changes in the first event
  - NOT EMPIRICAL SCIENCE
- A statistical relation is when two events are related statistically when they tend to vary together
  - EMPIRICAL SCIENCE

Statistical Relation

- Just because two events tend to occur together, doesn't mean they are *causally* related

Extrapolation vs Intrapolation

- **Intrapolation:** Statistical relation *between* variables
- **Extrapolation:** Prediction about values that have *not been observed*
  - Stats can **NEVER** do extrapolation

Statistics vs Mathematics: What's the Difference?

- Mathematics is based on premises which can't be falsified
  - e.g. if a<b, and b<c, then a<c

- Mathematics is based entirely upon *deduction*
    - If it rains, then the roads will be wet
    - Can generate explanations for events
- Statistics is based upon *induction*
    - All known life forms are carbon-based. If we find a new life form, it is probably carbon-based.
    - Can generate new knowledge, **BUT** can be false

Sampling

- A sample of participants is a portion of the target population
- The goal of sampling: Estimate certain characteristics of the target population
- Sample must be representative of the target population (needs to be large enough and possess characteristics of the target population)
    - The more a characteristic is variable, the large the sample size should be
    - Central Limit Theorem: The larger the number of randomly-drawn results that are observed, the more they approach their true distribution
- "N" is the variable representing sample size within statistics
- The larger the sample size, the better our estimate of the true mean
- The smaller the sample size, the larger the error of estimation
- Size matters bitch

Types of Measuring Scales

- Nominal: by categories
    - Compare and classify
- Ordinal: Logical Order
    - Assign a rank
- Intervals: Equal intervals
    - Compare a measurement
    - If a zero is present it is meaningless
- Ratio: Comparing proportions
    - Measurements
    - If a zero is present it means the absence of something

Making Graphs

- Nominal and Ordinal Scales: Bar Graph

- Interval and Ratio scales: histograms
    - 1. Measure the spread of the data, the large value minus the smallest value
        - Function in excel to find smallest number "=min"
        - Function in excel to find largest number "=max"
        - Remember to set the array to cover all numbers wanted (A1:A6)
        - Calculate spread in excel… =max(X0:X0) – min(X0:X0)
    - 2. Divide data into clusters
        - Clusters **MUST** be of the same **SIZE**. Not too big, not too small, just the size of Montreal
    - 3. Count the number of data point in each clusters
        - E.g. 1 to 3 = 8 (the number of data points with values ranging from one to three, is eight)
        - 4 to 6 = 4
        - 7 to 9 = 4
    - 4. Draw the Histogram
- The frequencies above are called absolute frequencies
    - These can be turned into relative frequencies (the proportion of data found within a cluster) using the following formula
    - $relative\ frequency = \frac{Absolute\ Frequency}{n}$
    - "n" **ALWAYS** means the total number of observations
- Remember to identify suspicious elements in a histogram

    - Extreme data
        - Data very far away from the distribution
        - Can shift where the mean is, in relation to the distribution
    - Asymmetry
        - When there is no clearly defined center
        - Always good to plot data before calculating the mean
        - Can shift where the mean is
    - Bimodality
        - Also called multi-modality
        - Two or more peaks, in normal distribution there is only one peak
- Distributions may take many different forms
    - Modality
        - Unimodal (one peak): I, IV, V, VI, VII
        - Bimodal (two peaks): II
        - Rectangular: III
    - Kurtosis (Curvature)
        - Mesokurtical: I, II
        - Platykurtical: V
        - Leptokurtical: IV, VI, VII

Cumulative Frequencies

- A way of creating a histogram that will always sum to 1
    - 1 in this case represents 100% of the data
    - To find cumulative frequencies you sum two relative frequencies together
    - Can smooth out graph when there is data missing

Basic Notation

- Vectoral notation
    - Group of data points… [5,1,8,2,9,4,3,1,6,8,3]
- We can let X represent the entire group of data
    - X = [5,1,8,2,9,4,3,1,6,8,3]
- Each element (or single data point) can be represented as follows…
    - $X = [x_1, x_2, \ldots. x_i]$
- So… $x_2 = 1$
- From this notation it's easy to get the sum of all data points…
    - $sum = \sum_i x_i$
- To calculate spread…
    - $spread = \max X - \min X$
- To calculate mean all fancy-like
    - $\bar{x} = \frac{\sum_i x_i}{n}$

Measures of Central Tendency and Variability

Two types of Statistics

- Descriptive Stats
    - Describe a smaple
    - E.g. mean, spread of data, etc.
- Inferential Stats
    - Goal: Infer some attributes of the population using only the data
    - E.g. t test, ANOVA

Descriptive Statistics

- Two types
    - Measures of central tendency
        - Summarize a set of observations
    - Measures of variability
        - Estimate the dispersion of the data
- **ALWAYS** look at the raw data prior to doing any calculation, **DON'T** just rely on descriptive stats
- Examples of descriptive statistics… Mean, median, mode, confidence intervals

Inferential Statistics

- Infer characteristics of the population from a sample of data

Measures of Central Tendency and Variability

September 13, 2013
2:32 PM

Two Types of Statistics

- Different stats, different conclusions
  - The curve used can alter the information that the data provides
  - i.e. Positive & Linear will mean that as X increases, Y will **always** increase
  - Or… a curve could level off meaning that as X increases, Y will increase **up to a point**
- When calculating an average it's good to know what the range of data is
  - e.g. NO ONE CARES how fast Usan Bolt was when he was 6 months old

Measures of Central Tendency

- Mean
  - The sum of Observations divided by the numbers of observations
  - A single number that describes a series of observations
  - Can reduce "noise"
  - ⭐ Remember that N represents the **total number of observations**
  - $\bar{x} = \left(\sum_{i=1}^{N} x_i\right)/N$
  - To calculate average frequency...
    - $\bar{x} = \left(\sum_{i=1}^{N} f_i x_i\right)/N$
    - ⭐ Where "f" is the frequency, you multiply the raw data by the frequency then take the sum of the multiplied data
  - To calculate weighted average
    - Where "k" represents the group
    - $\bar{x}weighted = \left(\sum_{i=1}^{k} n_i x_i\right) / \left(\sum_{i=1}^{k} n_i\right)$
  - To calculate the Harmonic Mean
    - $\tilde{x} = \frac{1}{\frac{\left(\sum_{i=1}^{N} \frac{1}{x}\right)}{N}}$
  - To Calculate the mean of clustered Data
    - Multiply the first number in column A and column B together
    - Repeat for other columns
    - Add numbers together
    - e.g. (mean = 1x2+3x4+5x6)/12
- Median
  - The value for which 50% of observations are above and 50% are below
  - ⭐ First!!! Classify data in ascending order (smallest to largest)
  - If N is odd…
    - $Median = x_{(N+1)/2}$
  - If N is even…
    - $Median = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}}}{2}$
    - E.g. In a set of 10 you would take X at position 5, and X at position 6, add the values together and then divide by 2
- Mode
  - The most frequent observation
  - Don't be stupid and complicated… you can just look at a set of data and figure this out
- Which measure of central tendency should you choose?
  - Consider vulnerability to extreme values

        - The mean **is STRONGLY** affected by extreme values
        - The median & mode are **often NOT** affected by extreme values

Measures of Variability

- Spread
    - Difference between the largest and smallest value
    - $Spread = \max(x) - \min(x)$
- Variance
    - Average amount of variability in a given set of data
        - More technically: the average of squared differences between each data point and the mean
        - Looks at the difference between each data point and the mean
        - The variance is squared because x_i - the mean will ALWAYS be zero
        - Why N - 1? The equation below underestimates the population, so we apply a correction (N - 1). This correction is determined by the degrees of freedom in the data
    - $Variance = (\sum_{i=1}^{N}(x_i - \bar{x})^2)\,/(N-1)$
        - Do the subtraction first
        - Then square
        - Then the adding
        - Then the Division
        - ⭐ Remember that it will be each individual observation (hence x_i) minus the mean, and then squared
        - ⭐ Show work on midterms
    - The larger your data set, the more likely you will encounter extreme values
    - Degrees of freedom
        - If we have 5 data points, x = [2 4 6 8 10], there are 5 degrees of freedom
        - BUT for the above equation the result can be determined exactly if we have N- 1 data points as well as the mean
            - e.g. X = [2 4 6 8 ?], mean = 6, N = 5, what is the missing number?
- Standard Deviation
    - The square root of the variance
    - How dispersed your data is around the mean
    - $Standard\ Deviation = \sqrt{Variance}$
    - The higher the standard deviation, the larger the spread around the mean of the data
    - Answers the question… On average, at what distance is a single data point from the mean?
        - Usually we expect that most data will be within 1 standard deviation of the mean
    - ⭐ Mean + S.D. and Mean - S.D. will give you two numbers, most of your data **SHOULD** be within those two numbers
    - The Mean and Variance can change independently of each other
- Standard Error
    - Standard Deviation divided by the square root of N
    - $\sigma = s/\sqrt{N}$
    - Where "s" represents the standard deviation
    - Generally… the Larger the N, the smaller the standard error
    - When the standard deviation is low, the standard error is low
    - Standard error is an indication of how the mean can generalize the target population
    - Important to report standard error
        - e.g. 6 ± 1.41
    - The standard error is the most common measure of dispersion to generate error bars in figures

Graphs

- Include titles for the axes, with units
- Include a legend if needed

- Line graphs if measuring continuous variables, bar graphs otherwise
- No unnecessary details
- Always include error bars
- Use a non-linear scale…
    - Only if the spread is large
    - Use a logarithmic axis, (log in base 10: $\log_{10}(1000) = 3$ since $10^3 = 1000$
    - Log(x) increases slowly when x increases

Inferential Statistics

September 18, 2013
4:50 PM

How to formulate a statistical Hypothesis

- Two to formulate...

H_0 The Null Hypothesis

- This is the status quo hypothesis. It is maintained until the data suggests otherwise. This is the hypothesis we are trying to reject when doing research
- E.g. no difference between psychotherapy and medication groups

H_1 The Alternative Hypothesis

- This is the hypothesis we want to demonstrate. To accept it, we must reject the null hypothesis

Elements of a Statistical Hypothesis

1. The IQ of the individual won't be higher than 120 (Null hypothesis)
2. The IQ of the individual will be higher than 120 (Alternative Hypothesis)

- Important: Assumption of a **NORMAL DISTRIBUTION**
    - Why? It is part of the definition of IQ, the bell curve
    - Some individuals have an IQ of 100, very few have a very high or low IQ
- Many variables in psychology follow a normal distribution, but not all
- Why do normal distributions emerge?
    - Reasons related to the central limit theory

Normal Distribution

- Normal distributions also assume an infinite number of possibilities
- Unimodal and symmetrical around the mean
- Mode = Median = Mean
- Asymptotic (the curve never actually touches the x axis)

Probability of Observing a Given Score

- What is the probability associated with a score of 8 when rolling the dice?
- $p = (width\ x\ height)/N$
    - Width is the width of a bar (i.e. if it goes from 7 to 8, then it's width is about 1)
    - Height is the height of the bar
- When dividing bars into more bars, width eventually approaches, and reaches zero
    - Solution: use cumulative frequencies
    - To find p(8), calculate the difference between the cumulative frequencies (CF): CF(8) - CF(7)
    - CF(8) = (1 + 2 + 3 + 4 + 5 + 6 + 7)/36
    - CF(7) = (1 + 2 + 3 + 4 + 5 + 6)/36
    - CF(8) - CF(7) = 28/36 - 21/36 = 7/36
    - Answer will be different than when using width and height
- To obtain the probability of a score x under the normal curve, we need to know the corresponding area under the curve
    - $[-\infty, x]$
    - A standard normal distribution has $\mu = 0$ and $\sigma = 1$
    - Where μ is the mean and σ is the standard deviation

What Proportion of Data is less than a specific score?

- When a standard normal distribution, everything on the axis is called a z-score
- When z=-1.75
- Find the first two digits (-1.7) in the right hand column on the z-table
- Find the third digit (0.05) on the top row
- Find where they intersect, then take that number and multiply by 100
- Only for data to the LEFT of a given score
    - For data to the right, subtract percentage of data to the left from 100%
- When trying to find data between TWO scores
    - Find the proportion of the two given scores

    - If one score is negative, you can looking up the positive equivalent
        - Subtract the positive equivalent from 100
    - Then find the different between the two proportions, AFTER converting both to percentages
- When trying to find data NOT between TWO scores
    - Follow procedure for two scores
    - Then take the difference and subtract it from 100 to find the percentage is not between two scores
- What score corresponds to a given proportion below the mean?
    - Proportion of 85%
    - Mean - 2.93, SD = 0.33
    - Look for the z-score corresponding to 85% of the data
        - Look for the z-score value that is the closest to the desired value
    1. Table(85%) = 1.04
    2. Reorder the equation to change standardized data, to the data according to the given mean and SD
        i. z = x - mean / s
        ii. sz = x - mean
        iii. x = sz + mean
            1) Mean is NORMALLY represented by "x" with a bar over the top
    3. Plug in the numbers
- What score corresponds to a given proportion above the mean
    - Prop. Of 40%
    - Mean = 2.93, SD = 0.33
    - Z table = 100% - 40% = 60%
    - $x = sz + \bar{x}$
    - Find the Z score closest to 60%
    - Calculate

Standardizing Data

- $(x - mean)/s$
- **ALWAYS** standardize the data BEFORE using the z-table

Correlation and Linear Regression

September 25, 2013
4:20 PM

Correlation and Linear Regression

- A correlation is a systematic relation between two variables
- Linear Correlation
  - Positive Correlations show that as Variable A increases so does Variable B
  - Negative Correlations show that as Variable A increases, Variable B decreases
  - When data is largely dispersed and is all over the place on the chart, there is likely little or no correlation
- Non-linear correlation shows that there is a relationship between X and Y, but that the relationship changes depending on the two points

How to Interpret Correlation

- 1.00 = Perfect Positive Correlation
- 0.00 = No Correlation
- -1.00 = Perfect Negative Correlation
- What does it mean if two variables have a perfect correlation?
  - It means we are measuring the same thing with two variables

Linear Regression

- Find the linear function that best describes a set of data
- $y = ax + b$
- Assumptions (Rules for the line to do it's job)
  1. Homogeneity of the residual error
     i. Residual Error is whatever the line cannot explain about the data
     ii. When all the residual error points are put together they should form a normal distribution
  2. Homogeneity of Variance
     i. Is the Distribution of the data the same along the regression line?

How to illustrate Correlation

- Scatter Plot
  - Can reveal non-linear relations
  - Can help identify outliers
  - Can reveal non-homogenous variance
  - Can reveal non-homogeneity of residual error
  - Can help identify points of lever
    - Outliers that shift the line from where it should go

Non-Linear Relations

- The correlation would be close to zero in the example on slide 11
  - Need to use a quadratic relation for this example
- Only focusing on linear relationships
- A correlation of zero in this case does not necessarily mean that there is no relation

Problems in interpreting a correlation

1. The Direction of the relation cannot be established
   i. You have no idea if Variable A is influencing Variable B, or Variable B is influencing Variable A
2. Non-linear relations will not show up in a correlation
3. A limited RANGE of data may hide a relation between two variables
4. Other variables may explain a relation
   i. The relation between A and B may be explained by C

Solutions to Problems

1. Directivity cannot be established
   i. Solution: Temporal Analysis
   ii. Analysis of the appearance of Variable B in relation to Variable A
2. Non-linear relationships cannot be detected when using linear correlation

i. Use non-linear regression
ii. Warning: Problem of bias versus variance. Important to analyse group data and not overfit individual data points

3. A limited range of data may hide a correlation
    i. Test a wider range of values for the variables
4. Other variables may explain the relationship
    i. Solution: Partial Correlations
    ii. Role of partial correlation
        a. Correlation between Golf and Crime = -.71
        b. Partial Correlation between golf and crime, after regressing out wealth = -.19

How to Calculate the Correlation

- Can be extremely useful to put the data into a table (see slide 25 of this lecture)

1. Obtain the mean and standard deviation of each group
2. Calculate the covariance of x and y
    i. Measures the size of the relation between x and y
    ii. $Cov_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$
3. Calculate the Correlation between x and y
    i. $r_{xy} = \frac{Cov_{xy}}{S_x S_y}$
4. Square the Correlation (Interpret the Correlation)
    i. The squared number tells us that N% of the changes in Y, can be explained by changes in X
    ii. It goes both ways, N% of the changes in X can be explained by changes in Y

Factors that influence the correlation

1. Range of the data
    i. Lower range = lower correlation
2. Using polarized groups increases the correlation
    i. Very different groups
3. With fewer data points, the correlation is underestimated

Inferences using Correlations

- H_0: The correlation in a population can be explained by chance
- H_1: The correlation in a population is superior than expected by chance
- ρ_{xy}
    - Represents the correlation at the level of the entire population
- t_{crit}
    - Examination of a t-distribution
    - Values for critical t is on the t table
    - Chance observation
- df = degrees of freedom
- $t_{(\alpha,df)} = table(value)$
    - Where α is the tolerance for being wrong (typically 0.05, so 5% of the time)
    - Where df represents the degrees of freedom
    - The value found is equal to what is considered chance expectation
- Observed t is based upon the data we observed
    - $t_{obs} = \sqrt{\frac{r_{xy}^2}{1 - r_{xy}^2} df}$
- If $t_{obs} > t_{crit}$ reject null hypothesis, and accept alternative
    - If t_{crit} is bigger, then the even could be explained by chance

Linear Regression

- We want to find the linear function that best describes the relation between two variables
- $\hat{y} = b_0 + b_1 x$
- To do so we must find two parameters: b_0 and b_1
- When b_1 is larger the line will be steeper
- How to find the b's

- $b_1 = \frac{Cov_{xy}}{s_x^2}$
- $b_0 = \bar{y} - b_1\bar{x}$

Normal Distribution

October 9, 2013
4:03 PM

How to Inter Properties of the Population When we only have Access to a Sample

- Two Scenarios
    1. We want to know if **one** sample resembles a hypothetical population
        i. e.g. Normal distribution, certain mean and SD
    2. We want to know if **two** samples are likely drawn from the same population (or two different populations)
- Based upon data and analyses we must decide to reject or maintain H_0
- A Statistical Error is any decision based on a statistical test where we draw the wrong conclusion
    - Type1 Error: Error made when H_0 is rejected, but it is true
        - We decide that the mean of a sample is significantly different from constant k, but it is actually not different
        - We decide that there is a significant difference between two means but there is no difference
    - Type 2 Error: Error made when H_0 is maintained but it is false
        - We decide that the mean of a sample is not significantly different from a constant, but it is different
        - We decide that there is no significant difference between two means, but they are actually different

Decision Threshold

- What level of risk are we willing to tolerate when making a decision to accept/reject H_0
    - Decision Threshold = Critical Value
        - e.g. Alpha = 0.05
        - This means that if we repeat the same experiment many times, a good decision will be made 95% of the time

Midterm

- 8 multiple choice
- 2 calculation, show work
- Appendix will have equations & what they mean

Term Paper

- Max 12 pages
- Individual or pairs
- Includes:
    - Title Page (titles, names, student number, data number)
    - Main Text
        - Describe the precise objectives of the study
        - Identify the Dependent and independent variables
        - Describe the null and alternative hypothesis
        - Describe the analyses that will be performed and why?
        - Start your analyses with some descriptive stats, including measures of central tendancy and dispersion. Discuss their meaning
        - Hypothesis testing: according to your H0 and H1, choose an appropriate statistical test and perform the necessary analyses
        - Global interpretation of your descriptive and inferential stats. Limitations. Alternative analyses. Conclusions
    - Figures
- 12 Point Font
- Double Spaced
- Due Nov. 29th

Evaluation

- The arguments presented are valid and related to course material as much as possible
- Text is clearly written (no fancy words)
- Graphs/tables are well presented
- Shows extensive critical thinking
- The calculations are exact (details provided)
- One point per working day will be deducted for late assignments

Significance Testing

October 9, 2013
4:09 PM

Significance Testing

- Gather all necessary information
  - Set up null and alternative hypothesis
  - Mu is the mean of the population
  - k is the constant
  - Alpha is the tolerance of error (assume 0.05 if not given)
  - Standard Deviation at the level of the population
  - x-bar is the mean of the sample
  - n is the total number of observations

1. Calculate the standard error of the mean of x (sample mean)
   a. $\sigma_x = \frac{s}{\sqrt{n}}$
2. Calculate the z score based on the data
   a. $z_{\bar{x}} = \left|\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right|$
3. Calculate the critical z value, taking a decision criterion of $\alpha = 0.05$
   a. $critical\ z = z_\alpha = 1.96$
   b. Z Score associated with a 95% chance
   c. Look for the value that is the closes to the percent
4. Compare observed z score with the calculated z score
   a. If larger than critical z, then reject the null hypothesis

Unicaudal vs. Bicaudal Decision Test

- What is the probability that the mean of the population is smaller than a given value (unicaudal decision)
- What is the probability that the mean of the population is exactly equal to a given value? (Bicaudal decision)
- Does the difference between the mean of two samples reflect the difference between two populations? (Bicaudal)
- Unicaudal alpha z = 1.65
- Bicaudal alpha = 1.96

Confidence Intervals

- A range between two values where the real score will fall between
- To find the confidence interval you need…
  - Mean of the sample
  - Population size
  - Standard Deviation
- Calculate
  - $CI_{1-\alpha} = \bar{x} - z_\alpha \sigma_{\bar{x}} \leq \mu \leq \bar{x} + z_\alpha\ \sigma_{\bar{x}}$

Difference between the means of two samples

- Calculate standard error for two samples separately
- After that, calculate the combined standard error
  - $\sigma_{\bar{x}_1 \bar{x}_2} = \sqrt{(\sigma_{\bar{x}_1})^2 + (\sigma_{\bar{x}_2})^2}$
- Then calculate the z score by subtracting the means from each other and dividing by the combined standard error

T-Test

October-24-13
02:04 PM

Why Use a T Test?

- We wish to compare 2 groups but we don't know the SD of the population
- In research we use the t test much more than the z test
    - We rarely know the population's SD
- Because we don't know the SD of the population we have to estimate the SD from the sample
- Find critical T values on a Tail Probabilities table
    - One Tail and Two Tails are Unicaudal and Bicaudal respectively
- Example
    - We want to know if a mentorship program helps students succeed at SATs. Nine participants are chosen randomly to receive a mentoring session. Will they have a score that is statistically different from a score of 1000? (General Population)
    1. Calculate the standard error of the mean of the sample
    2. Calculate the value of the observed t
    3. Calculate the critical t (bicaudal decision)
    4. Decide based upon the observed and critical t, same thing as z scores

Confidence Intervals

October-24-13
02:11 PM

Confidence Intervals

- We can calculate the confidence interval of the population mean in the following way
  - $CI_{1-\alpha} = x - t_{critical}\sigma_x \leq \mu \leq x + t_{critical}\sigma_x$
  - Where "X bar" is equal to the sample mean

Comparison between two independent groups

- Example with Independent groups: A group of primary school children (10 students); another group is formed by a class of 30 students
- Example with non-independent groups: We wish to verify the efficacy of a therapy aimed at reducing anxiety. We choose a group of participants and measure their anxiety before and after therapy
- Same Participants ≠ Independent Groups
  - Independent Groups are groups that are composed of different people
  - Dependent Groups are a group of subjects that are tested at different points in time the participants are the same
  - **THESE DEFINITIONS ARE ONLY FOR THIS CLASS**
- Two possible scenarios
  1. Two populations with the same n: $n_1 = n_2$
     i. e.g. two groups with 30 participants each
  2. The two populations don't have the same n
- In both cases 4 postulates apply
  1. Participants are assigned randomly and independently
  2. Data from both populations are normally distributed
  3. Variances are homogenous across groups
  4. There is a single dependent variable
- Both populations have the same n
  1. Test the homogeneity of variance
     i. Do both groups have comparable variance (If not t test will be of limited use)
     ii. Divide the larger variance by the smaller one
        a. $\frac{s_G^2}{s_E^2}$
     iii. The result of the above equation is called an observed F ratio
     iv. Consult F table for the critical values of F
     v. If F_{obs} is larger than F_{crit} then you can assume the variances are homogenous
  2. Calculate the standard error of the mean of the sample
  3. Calculate the Observed t
  4. Calculate the Critical t
  5. Decide based upon critical t & observed t
- Populations with different n's
  1. Calculate the variance common to the two groups
  2. Calculate the standard error of the sample mean
  3. Calculate the Observed t
  4. Calculate the critical t
  5. Decide
- Confidence intervals of the difference between two groups
  - $CI_{1-\alpha} = x - t_{critical}\sigma_x \leq \mu_1 - \mu_2 \leq x + t_{critical}\sigma_x$

Comparing Two Dependent Groups

1. Calculate the difference between pre - and post-treatment for each participant
   a. This is a third column titled "D", and D is a vector (e.g. D = [-47 -26 -27, etc]
2. Calculate the mean, variance and SD of "D"
3. Calculate the standard error of the mean of D
4. Calculate the Observed t
5. Calculate the critical t
6. Decision

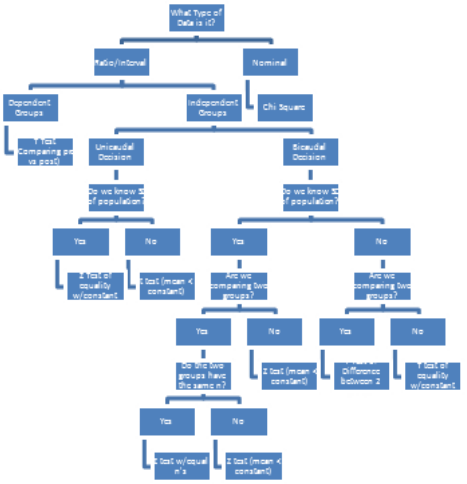
Confidence Intervals for Dependent Groups

- Same formula as the confidence interval between two groups, but replace X bar with D bar

How to report Statistical Results

- (t(4)=3.62; p<0.05)
- Where... t(df) = t_{obs};p<α
  - Use this notation for final report

What Test Should I Use?



Chi Square

November 1, 2013
3:00 PM

Parametric vs Nonparametric Tests

- A parametric test has postulates about the target population (mean, SD, distribution, etc.)
  - t-test, z-test, assumption of a normal distribution
- A Non-parametric test has no postulates about the population
  - Can be applied to any kind of problem
  - Assumes nothing
  - Chi square
- Example
  - We want to know if the latest album from an artist is preferred over three other artists
    1. Add an extra row to the bottom of the table
       i. This row will represent that if all participants did not have a preference (expected values)
    2. Compare the expected values with the observed values
    3. Set up H_0 and H_1
  - Calculations for example
    1. Calculate the observed chi-square
       i. $x^2_{obs} = \sum_{i=1}^{k} \frac{(o_i - a_i)^2}{a_i}$
          a) Where o_i is an observed value, and a_i is the expected value
          b) K = number of categories (e.g. number of artists = 4)
    2. Find the Critical Value of the Chi-Square
       i. Calculated according to the chi-square table
       ii. Required degrees of freedom (df= k -1) and a decision criterion (a = 0.05)
       iii. Crit. Value is found at the intersection of df and alpha
    3. Compare Critical Chi and Observed Chi
       i. If Crit > Obs then accept null
       ii. If Crit < Obs then accept alt

Contingency Table with Two Variables

- A contingency table allows us to see if two variables are dependent on each other
-

Gender	Physics	Engineering	Linguistics	Medicine
Men	108	345	94	17
Women	8	12	253	60

- How to solve…
  1. Calculate the total number of persons enrolled in each program and the total number of persons of each gender
     i.

Gender	Physics	Engineering	Linguistics	Medicine	Sum
Men	108	345	94	17	564
Women	8	12	253	60	333
Sum	116	357	347	77	897

  2. Calculate the expected frequencies
     i. $a_{ij} = \frac{R_i C_j}{N}$
     ii. Where *a* is the expected frequency for an element in row *i* and column *j*
     iii. R is the sum of a row, C is the total of a column, and N is the grand total
  3. Calculate the observed Chi square
  4. Calculate the critical Chi Square
     i. Degrees of Freedom
        a. $df = (n_R - 1)(n_C - 1)$

5. Decision

Concatenating Variables

Instruction	Sidewalk	Garbage	Recycling Bin	Sum
Message	41 (61.66)	477 (497.36)	385 (343.98)	903
Control	80 (59.34)	499 (478.64)	290 (331.02)	869
Total	121	976	675	1772

  - Expected frequencies in parentheses
- How to solve
  1. Calculate the expected frequencies
  2. Concatenate Variables based on the Initial question
     i. To do so, simply sum the two variables
     ii. In this case it will be Sidewalk & Garbage
  3. Calculate the Observed Chi Square
     i. NB: the "k" value changes as the number of squares in the table (categories) is reduced!!!
  4. Calculate the Critical Chi square

Measuring the relationship between variables

- To measure the degree of dependence between two variables for **nominal data** we must use Cramer's phi
  - $\phi_c = \sqrt{\frac{x^2_{obs}}{N(m-1)}}$
    - Where m is the smallest number between n_r and n_c
    - Normal correlation equation is only used for ratio and interval scale data
    - Simply square Cramer's phi to interpret it, and then multiply by 100 to get a percentage
- **WARNING**
  - Cramer's phi is influenced by
    - Sample Size: Larger samples have higher phi values
    - Uneven Totals of lines and columns: A row or column with a higher total will have more influence on the final result

Odds Ratio

- Another measure of association
- Unaffected by sample size or by uneven totals of lines and columns
- Calculating the odds ratio
  1. Calculate the proportion of the observed variables
     i. $Ratio_{GS|Recycling} = GS/Recycling$
  2. Calculate the proportion of the controlled variables
  3. Calculate the odds ratio (OR) between the two proportions
     i. $OR_{\frac{Message}{Control}} = \frac{0.74}{0.5}$
- If the odds ratio is exactly 1, then it means that there is no difference
- If the odds ratio is smaller than 1, it means that the ratio is shifted more towards the denominator
- If the odds ratio is greater than 1, it means that the ratio is shifted more towards the numerator

Statistical Power

November 6, 2013
4:35 PM

Statistical Power

- Assuming there is a difference between the groups, what is the probability that our statistical test will find it
- Statistical power occurs when we have rejected H_0 because it means that our test has found a significant difference that makes H_0 false

Statistical Inference

- Unicaudal decision based on alpha
  - Is the mean of a population equal (or inferior) to a given value
- The value of alpha is the decision criterion
  - A small value of alpha indicates that you're being very skeptical
  - As alpha decreases, beta increases
  - A large value of alpha indicates that you're being very gullible
  - As alpha increases, beta decreases
- Example
  1. Calculate the magnitude of effect
     i. Shows the size of the gap between H_0 and H_1
     ii. $\Delta_1 = \frac{|\mu - \bar{x}|}{\sigma}$
     iii. Small effect = 0.2, Medium effect = 0.5, Strong effect = 0.8
  2. Consult the Power table
     i. Find the value for statistical power
     ii. Requires: Alpha Value, Total number of observations, magnitude
        a. For the total number of subjects find the value on the table closest to your n value
        b. For the magnitude, find the value on the table closest to the delta value
     iii. Align the three values, and to the left of the n, in the power column you will find your value for statistical power
  3. Conclusion
     i. e.g. Therefore our statistical test will be able to reject H0 with 70% probability
- Factors that influence statistical power
  1. Level of Confidence (Alpha)
     i. The higher the alpha value, the higher the statistical power
  2. The magnitude (size) of effect (Delta)
     i. The higher the delta value, the higher the statistical power
  3. Variability of the population (Sigma, lower case)
     i. The higher the sigma value, the lower the statistical power
     ii. The higher the sigma, the lower the delta, and the lower the delta, the lower the statistical power
  4. The sample size (N)
     i. The larger the sample, the higher the statistical power
- Before undertaking a study, it is important to ask how many participants will be required to obtain a desired statistical power
- To find a sample size that will provide you with a given statistical power (0.9) simply consult the table
  - Try to take into account what you are measuring (e.g. babies, vs nucleotides
- How you calculate Statistical Power varies based on test

Statistical Power of a T-test with independent groups

1. Calculate Cohen's "D" (estimates magnitude of the effect)
   i. $d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 s_2^2}{2}}}$

2. Look up values in the table
    i. Intersection of Cohen's D and number of participants
3. Decision
4. In this class if there is a power of 7 or higher, it's good

- For groups that are dependent…
    - You need to calculate D
    1. Calculate the mean and SD of D
        i. Where D is the different between pre and post
    2. Calculate Cohen's D
        i. $d = |\frac{\bar{D}}{s_D}|$
    3. Decision

Statistical Power of a Correlation

1. Calculate the Correlation between two variables
    i. Don't forget to square the "r" value
    ii. Correlation also corresponds to the magnitude of the effect in this case
2. Calculate Cohen's D
    i. $d = \sqrt{\frac{r^2}{1 - r^2}}$

Limits of Statistical Tests

November 8, 2013
2:56 PM

Limits to Correlations

- A variable C may explain the correlation between A and B
  - ⭐ Solution: Partial Correlation
    - This is a statistical trick to control the effect of a third variable
    - We obtain the correlation between A and B while controlling for C
    - If the partial correlation between A and B remains significant, it can't be explained by C
    - If the partial correlation is low, we can't exclude the possibility that C explains the relation between A and B
  - Calculating Partial Correlation
    1. Calculate Correlation between A and C
    2. Take the residual values (R_A)
       i. You do this by plugging the values for C into the found equation for the line of regression
       ii. Calculate the residual error by subtracting the observed values from the predicted values and then square each difference
    3. Calculate the Correlation between B and C
    4. Take the residual values (R_B)
    5. Calculate the correlation between R_A and R_B
    6. Conclusion
- Correlation only takes care of linear dependencies between variables
- Could possibly be a three way interaction between crime and wealth
- How to do a correlation with an uneven *n* in two groups?
  - ⭐ Solution: Eliminate participants?
    - Which ones? By eliminating a participant you also lower statistical power
    - If you eliminate the outliers correlation will increase (you overestimate your correlation)
    - If you eliminate a non-outlier, correlation will decrease (you underestimate your correlation)
  - ⭐ Solution: Replace missing data with the mean of the group?
    - Problem: Reduces total variance but boosts statistical significance

Limits to T-Tests

- Correlation between data
  - T tests should only be applied to uncorrelated data within the same variable
  - NB: Certain variables may be very strongly correlated with others (e.g. Rate of Respiration is strongly correlated with Cardiac Rhythm)
    - A statistical test based on the data will overestimate true difference between H_0 and H_1
  - ⭐ Solutions
    1. Only look at data that are not correlated
    2. Apply a statistical correction to decorrelate the data
    3. ANOVA for multiple comparisons
    4. Principle components analysis

Limits to All tests of Significance

- Multiple Comparisons
  - Problem?
    - ⭐ Capitalization on Chance
    - The probability of making a type 1 error is equal to alpha
    - If we repeat a statistical test many times, the probability of a making a type 1 error increases
  - ⭐ Solution: Apply a Bonferroni correction to the alpha value

        - Take the Alpha value and divide it by the number of comparisons
        - Use the new value of alpha for all analyses that we need to perform
        - Problem with the Bonferroni correction:
            - Lower alpha, lower statistical power
            - With a very small alpha it's unlikely that you will reject H_0 when H_0 is false
- Assumption of a normal distribution
    - The t test, z test, and linear regression all assume that data is normally distributed
        - If that's not respected, then the data needs to be normalized
        - NOT THE SAME THING AS STANDARDIZING
            - Standardization: transforming data to a z distribution with a mean of 0 and SD of 1
            - Normalization: transforming data to a normal distribution where the mean and SD are the same as the original data
- Homogeneity of the Variance
    - Non-linear transformation
        - Logarithmic Transformation
            - Brings data back to a more limited interval
            - Reduces the influence of extreme data
            - Results in a more homogenous variance
            - $x_{new} = log_{10}(x + 1)$
        - DON'T USE THE LOG BUTTON!!!!!!
        - USE THE LN BUTTON
        - The Log transformation can be applied to either one, or both (two) groups)
            - When you do a log transformation the variance will decrease
            - When observed F is less than the critical F, the variances can be considered homogenous
    - You then base all observations on these new transformed values (i.e. recalculate variance)

Normalizing Data

1. Group data into clusters (like histograms
    a. e.g. 1 - 5, 6 - 10, 11 - 15, 16 - 20
2. Calculate z scores using the upper bound (higher number) of each cluster
    a. $z = \left|\frac{upper\ bound\ - mean}{SD}\right|$
    b. This is done for each cluster in the data
3. Find the area between z and the mean (consult the z table)
    a. $Area = z\ - 0.5$
    b. You subtract 0.5 because the other half of the distribution doesn't matter to you. Remember that a normal distribution has a mean of 0 and SD of 1
4. Calculate the difference of the area between each cluster
    a. To do so, start at the bottom of the chart of clustered data
    b. For the initial calculation it's the area between z and the mean minus 0.5
    c. The next calculation will be the next cluster up minus the cluster below
    d. Make sure you mark the cluster where the mean is!!!
        i. When you get to this cluster in order to calculate the difference, you add them as opposed to subtract
        ii. When you get PASSED the mean, you continue subtracting
    e. When you get near the top, and you are missing the "top" value, do 0.5 - x
    f. You can check your area by adding the numbers together in that column, they should add to 1
5. Get the new observations € according to the normal curve
    a. $E = (area)(N)$
    b. e.g. E = (area of the class)(N)

Probability I

November 13, 2013
4:46 PM

Empirical Probability

- The empirical probability is a count of the events that are observed
  - $Empirical\ Probability = \frac{Number\ of\ Events}{Total\ Number\ of\ Observations}$

Joint Events

- ∩= and
- $D \cap S$= students enrolled in psych and stats
- $D \cap S^c$= students enrolled in psych but not stats
  - Where the c stands for the compliment
- Two important concepts
  1. The complement of an event (A^c)
     i. $P(eventA) = 1 - P(eventB)$
     ii. Because… P(eventA) + P(eventB) = 1 <--- basically the probability that something will happen
  2. The intersection of two events is equal to the product of their individual probabilities
     i. $P(D \cap S) = P(D) * P(S)$
- Remember that when using probabilities to be aware of compliments, whenever it is a compliment remember to do 1 minus the other event.

Contingency Tables

- See notes on contingency tables
- For probability it works similarly…
- A = students that passed stats, B = students that passed the Psych of Personality course
-

	B	B^c	Sum
A	44	21	A= 65
A^c			
Sum:			

- Plug in values based on the combination of two variables (e.g. if A & B)

Conditional Probability

- Probability of A given B
  - Notation $P(A|B)$
  - The line isn't an absolute in this case, instead | means "given"
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- This is applicable when two events are **NOT** independent
  - e.g. the probability of one event affects the other
- How do we know if two events are independent or not?
  - They are independent if P(A|B) = P(A)
    - The equation is such because if A is independent of B, then B is not necessary and Probability would be equal to A
    - In this case, the probability of A is **NOT** affected by the probability of B
    - e.g.
      - P(A|B) = P(A)
      - $(0.71) \neq 0.65$
      - Therefore A and B are independent

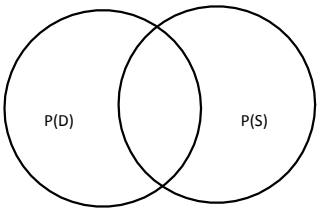
Union of Two Events

- What is the probability of getting a result greater than a given value
  - What is the probability of A or B
  - $A \cap B = A\ and\ B$ $A \cup B = A\ or\ B$
  - This question is solved via this equation
    - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
      - You subtract "A and B" because other wise you add them together

Bayes Theorem

- Allows us to calculate P(A|B) from P(B|A)
- Make sure you have the information you need to calculate using Bayes Theorem
  - You need…
    - P(A)
    - P(B)
    - P(B|A)
  - Equation
    - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Viewing Probabilities using a Venn Diagram



Probability II

November 20, 2013
4:40 PM

Calculating the total number of possible results

- Given 4 letters, we form nonsense words of three letters each
  - The letters can be used more than once
  - There are four options for the first, second and third letter…
  - Therefore
    - 4x4x4 = 64 or 4^3=64
- Essentially Take the number of possible choices for a category, and then multiply it by the number of choices for subsequent categories
  - Category 1: 4 choices
  - Category 2: 4 choices
    - There fore 4x4=16
- When the categories don't have the same number of choices
  - Do the same operation
    - Category 1: 4
    - Category 2: 3
    - Category 3: 2
      - Therefore…
      - 4x3x2=24

Permutations

- Given three letters, we want to know the total number of combinations when we do not want to repeat any letter (BAC and DAC are different)
  - Formula
    - $P_n^r = \frac{n!}{(n-r)!}$
    - Where n is the total number of choices
    - Where r is the number of elements chosen each time (in this case three)
    - ! Represents a factorial (e.g. 4! = 1x2x3x4=24
    - When showing work remember to eliminate elements that are both on the top and bottom

Combinations

- If the order of things chosen doesn't matter
- Formula
  - $C_r^n = \frac{n!}{(n-r)!\,r!}$
  - Where n is the total number of choices
  - Where r is the number of elements chosen each time

Complex Events

- What are the possible outcomes when flipping a coin
  - 2 outcomes, heads or tails
- What are the possible outcomes when throwing a die
  - 6 outcomes, #1-6
- What are the possible outcomes when flipping a coin AND throwing a die
  - Multiply the number of outcomes together
    - 2x6=12
- Example
  - In Quebec, license plates are composed of 3 letters and 3 numbers. What are the total number of plates that can be generated
    1. Calculate the number of possibilities for the letters
       i. 26 letters, 3 positions
       ii. Able to repeat letters
       iii. 26x26x26=17,576

2. Calculate the number of possibilities for the numbers
    i. 10 possible numbers (0 - 9)
    ii. 3 positions
    iii. 10x10x10 = 1,000
3. Multiply the result of steps 1 and 2
    i. 17576x1000=17576000 total car plates

- Complex evens using combinations
    - Example…
    - Emergency room with 6 nurses and 5 doctors, each time is made of 2 nurses and 3 doctors
        - Order doesn't matter
    - How many teams can we form?
        1. Figure out the number of combinations for nurses
            i. N = 6
            ii. R = 2
        2. Figure out the number of combinations for doctors
            i. N = 5
            ii. R = 3
        3. Multiply the results of 1 and 2 together