Prob Distribution

- Bernoulli
  - Single Trial
  - Success/Failure

- Binomial Distribution
  - Outcome of multiple Bernoulli
  - Probability of accumulated Bernoulli trials

- Normal Distribution
  - Middle Section: 68%
  - Wide Section: 95%
  - A Common Distribution

- Poisson
  - Prob of number of events occurring over time/space

Vocabulary

- Null Hypothesis
  - No statistical significant result (Usually in the form of difference in means)
  - Baseline

- Alternative Hypothesis
  - Has statistical significant result

- Continuous variable
  - Numeric, float
  - Variable that could technically be infinitely specific
  - Ex:
    - 0.000001
    - 12.00024
    - 5.33333332 seconds

- Discrete variable
  - A “countable” set of values
  - Numeric, usually integers, can be float
  - Basically categorical
  - Ex:
    - From 1 to 10
    - Coins, 1 or 0
    - Money, how many cents

- Categorical variable
  - Group of values
  - String
  - Ex:
    - [“dessert”, “main course”, “appetizer”]
    - [“Mac”, “windows”, “linux”]
    - [“a”, “b”, “c”, “d”, “e”]

- Independent populations
  - 2 populations that are not connected by any variable
  - Can’t be 100% sure but can make assumptions
  - Ex:
    - Sales of burgers is independent from sales of iPhones

Stat Tests

- T-tests
  - One-sample
    - Compare mean of sample to reference mean
    - Assumptions:
      - Continuous dependent (outcome) variable
      - Dependent variable is normally distributed
      - Sample is random
    - Degree of Freedom:
      - (Size of sample - 1)
  - Two-sample
    - Compare mean of sample between 2 continuous dependent populations with same categorical independent variable
    - Assumptions:
      - Continuous dependent (outcome) variable
      - Dependent variable is normally distributed
      - Random sampling
      - Populations are independent
    - Degree of Freedom:
      - (2 * Size of sample - 2)
  - Paired sample
    - Compare mean of 2 samples from the same population to measure change
    - Assumptions:
      - Continuous dependent (outcome) variable
      - Differences between pair samples are normally distributed
      - Random sampling
    - Degree of Freedom:
      - (Size of sample - 1)

- Z-tests
  - MUST HAVE MORE THAN 30 SAMPLES AND KNOW TRUE POPULATION STANDARD DEVIATION FOR ALL POPULATIONS IN USE
  - One-sample
    - Compare mean of sample to reference mean
    - Assumptions:
      - Continuous dependent (outcome) variable
      - Dependent variable is normally distributed
      - Sample is random
  - Two-sample
    - Compare mean of sample between 2 continuous dependent populations with same categorical independent variable
    - Assumptions:
      - Continuous dependent (outcome) variable
      - Dependent variable is normally distributed
      - Random sampling
      - Populations are independent
  - Paired sample
    - Compare mean of 2 samples from the same population to measure change
    - Assumptions:
      - Continuous dependent (outcome) variable
      - Differences between pair samples are normally distributed
      - Random sampling

- ANOVA tests
  - One-way
    - Categorical independent variable (with more than 2 categories/levels) and continuous dependent variable
    - Like 2-sample t-tests but with more than 2 populations
    - Assumptions:
      - Continuous dependent (outcome) variable

        - Dependent variable is normally distributed
        - Random sampling
        - Populations are independent
        - Variance across groups is same
    - Degree of Freedom:
        - (Size of sample - 1)
        - DF “between groups” and "within groups” can be referred to as “treatment” and “error”
- Two-way
    - Multiple categorical independent variables (with different categories/levels in each) and continuous dependent variable
    - Used to measure impact of different categories from different independent variables on outcome and of potential interactions between 2 categorical variables
    - Hypotheses:
        - No statistically significant difference in means across levels of variable [A, B]
        - No statistically significant interaction between variables A and B
        - Is statistically significant difference in means across levels of variable [A, B]
        - Is statistically significant interaction between variables A and B
    - Assumptions:
        - Continuous dependent (outcome) variable
        - Dependent variable is normally distributed
        - Random sampling
        - Populations are independent
        - Variance across groups is same
        - Equal sample size
    - Degree of Freedom:
        - (Size of sample - 1)
        - More frames of references for DFs like between, within, across independent variables

P-value and Confidence Interval

- p value
    - Probability of result being due to random chance
    - Usually compared to predetermined significance level
    - Ex:
        - $p > 0.05$ or $p < 0.05$
    - $p >$ predetermined level means cannot reject null hypothesis
    - $p <$ predetermined level means reject null hypothesis

- Confidence Interval
    - Range expected to contain true population mean
    - Ex:
        - (0.5, 2.6) does not cover 0
        - (-1.5, 0.8) does cover 0
    - CI covers 0 means cannot reject null hypothesis
    - CI does not cover 0 means reject null hypothesis

These metrics are all about chance and do not necessarily mean certainty.