MATH 1P98 Practical Stats

1.1 Data Types

- Population - a group of individuals or objects which we intend to study
- Random Sample - of size n is a selection of n such individuals, such that each member of the population has the same chance of being included
- Parameter - a numerical piece of info about a population
- Sample Statistic - the corresponding numerical information about a random sample.
  - We use a sample stat to estimate an unknown population parameter.
- Types of Statistical Methods - Descriptive Methods - those which helps us organize and summarize data through numbers, tables & graphs.
  - Inferential Methods - take a result from a sample extending it a whole population.
- In this course, assume fixed or unchanging in time and all surveys are answered truthfully.

1.2 Percentages and Proportions.

- Proportions a ratio of the following type → $\frac{\text{Number of objects with some property}}{\text{Total number of objects.}}$ or $\frac{\text{part}}{\text{whole}}$
  - can be written as decimal
- Percentage - A percent is a proportion multiplied by 100
  - percent - $\frac{\text{part}}{\text{whole}} \times 100$

1.3 Variables

- Qualitative - a variable that allows for the classification of individuals into categories which cannot be ordered.
- Quantitative - a variable that allows for numerical measures of individuals (can be ordered).
- Discrete - denoted by integer values
- Continuous - have an infinite number of possible values
- Additional Distinctions can be made by using the NOIR Convention - N - Nominal - values that cannot be ordered
  O - Ordinal - values that can be ordered
  I - Interval - ordinal values where the differences have meaning
  R - Ratio - interval values where the ratios of the value mean something.

1.4 Frequency Distribution and Histogram

- Frequency - The number of cases a variable has a variable from a specific range
- Frequency Distribution - consists of several classes defined by conveniently chosen numerical ranges, together with the frequency of the data which falls within each class
- Numerical Range - A set of numbers between a lower number and higher number, such as "Customers that range between 20 & 30 years old"
- Classes - Lower Class Limit = smallest number that belong to each class
  - Upper Class Limit - largest number that belongs to each class
  - Class Boundaries - the numbers which separate the upper & lower limit
  - Class Width - $\frac{\text{Highest Class Upper Boundary} - \text{Lowest Class Lower Boundary}}{\#\text{ of classes}}$
  - Class Midpoints - $\frac{\text{Lower Class Limit} + \text{Upper Class Limit}}{2}$

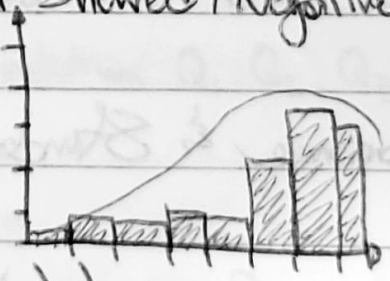
Method (Frequency Distribution)

- ① Select the number of classes (5-15)
  ② Calculate the class width
  ③ Choose the value for the lower class limit
  ④ Add the class width to the first lower class limit to get the second lower class limit and repeat until you get the number of desired classes
  ⑤ Determine corresponding upper class limits
  ⑥ List these limits in on ascending vertical column
  ⑦ Enter the total frequency for each class in the next column

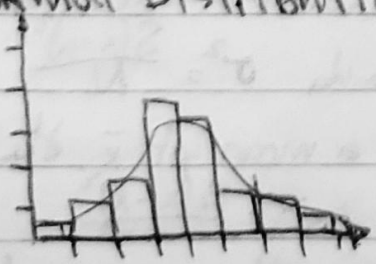
~~Shapes~~ Distributions

- Right-Skewed / Positively Skewed

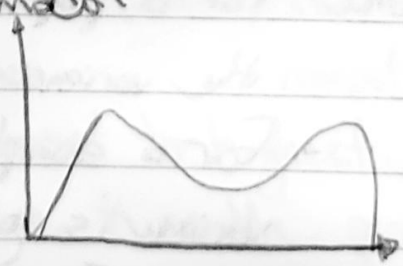
Left-Skewed / Negatively Skewed



- Normal Distribution



Bimodal

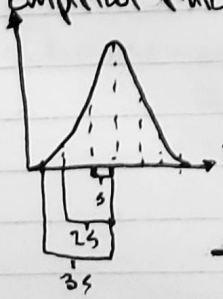


2.1 Measures of Centre

- **Mean** - The mean of a sample data set with n values is obtained by adding all values of x, written as $\sum_{i=1}^{n} x_i$ or $\sum x$ and dividing the total by n $\rightarrow$ $\bar{x} = \frac{\sum x}{n}$
  - For population, we use $\mu = \frac{\sum x}{N}$ where N is population size
- **Median** - The median is the middle value when the original data is arranged in increasing or decreasing order.
  - If there is an odd number of data values, the median is the middle number in the data set
  - If there is an even number of data values, the median is the midpoint of the 2 middle number in the data set
- **Mode** - the mode is the most frequently occuring data value
  - If the data value has no values that occurs more than once, the data set has no mode.

2.2 Variance & Standard Deviation

- **Variance** - For a population of size N with a mean of μ, the variance is given by the formula $\sigma^2 = \frac{\sum (x-\mu)^2}{N}$
  - For a sample of size n with a mean of $\bar{x}$, the variance is given by the formula $s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$
- **Standard Deviation** - For a sample of size n with a mean of $\bar{x}$; the standard deviation is $s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$
  - a measure of the typical amount data values differ from their mean
- **Empirical Rule** - Normal Distributions follow
  - ~68% of data falls within 1 standard deviation of the mean ($\bar{x} \pm s$)
  - ~95% of data falls within 2 standard deviations of the mean ($\bar{x} \pm 2s$)
  - ~99.7% of the data falls within 3 standard deviations of the mean ($\bar{x} \pm 3s$)



2.3 The z-Score

- z-Score - For any observation or data value x from a sample with mean $\bar{x}$ and standard deviation s, is given by the formula

$$z = \frac{x - \bar{x}}{s}$$

  - z-score essentially standardizes data values & scales them.

2.4 Percentiles & Quartiles.

- Percentiles - divide the data into 100 groups of about 1% of the data in each group
  - we use the notation $P_1, P_2, \ldots P_{99}$ for the boundary of each percentile.
- Quartiles - divide the data into 4 groups with about 25% of the data in each group, with the notation Q_1, Q_2, Q_3 for the boundary of each quartile
  - $P_{25} = Q_1$, $P_{50} = Q_2$, $P_{75} = Q_3$
- Percentile of x (P_x) = $\frac{\text{number of values less than } x}{\text{total number of values}} \times 100$, round numbers to the nearest integer.
- Find Percentiles - ① Sort the data from lowest to highest.
  ② Find $L = \left(\frac{k}{100}\right)n$, where k is the percentile ($Q_1 = 25 \ldots$)
  ③a If L is not an integer, round up and P_k is the L^{th} value.
  ③b If L is an integer, P_k is the L^{th} value plus the next value divided by 2.
- 5-Number Summary of a Data Set - the 5 values are ① Minimum, ② 1st Quartile (Q_1), ③ Median (Q_2), ④ 3rd Quartile (Q_3), & ⑤ Maximum

3.1 Probability

- Sample Space - the set of all possible outcomes
  - Ex. 1, 2, 3, 4, 5, & 6 outcomes on a die.
- Event - A specific collection of such outcomes.
  - Ex Getting 3 or more on a die.
- Probability - An event E occurs denoted by P(E)
  - If all potential outcomes of simple events are equally likely and if E is an event, then the probability of E occuring is $P(E) = \frac{\text{number of simple events resulting in } E}{\text{total number of simple events}}$
- Empirical Probability - A way to approximate the probability E occuring is to conduct an experiment & count the number of times that event E actually occurs
  - $P(E) = \frac{\text{number of times } E \text{ occured}}{\text{total number of trials.}}$
- Properties of Probabilities - ① The probability of any event is between 0 & 1 inclusive
  ② The probabilities of all simple events must add up to 1
  ③ The probability of an impossible event is 0.
  ④ The probability of a guaranteed event is 1.
- Complementary Event - The complementary of an event, E, denoted $\bar{E}$ is the event that occurs if and only if E does not
- Complementary Rule - The probability of an event & its complement must meet, $P(E) + P(\bar{E}) = 1$, or equivalently $P(E) = 1 - P(\bar{E})$

3.2 Probability Rules & Independent Events

- **Intersection of Two Events** - For any 2 events, if A & B occur, the probability is written as $P(A \& B)$, and the intersection of A & B is denoted as $P(A \cap B)$

- **Union of Two Events** - For any 2 events, if A & B or both occur, the probability is written as $P(A \& B)$, and the union of the events as $P(A \cup B)$

- **Conditional Probability** - For events A & B, the probability that event B occurs given that A has already occurred is $P(B|A)$ (& vice versa $P(A|B)$).
- **Independent Events** - Events are independent if one event does not affect the probability of the other.
  - $P(A) = P(A|B)$, $P(B) = P(B|A)$
- **Multiplication Rule** - For two events A & B → $P(A \& B) = P(A) \cdot P(B|A)$
  - For independent events → $P(A \& B) = P(A) \cdot P(B)$
- **Addition Rule** - for $P(A \text{ or } B)$ is → $P(A \cup B) = P(A) + P(B) - P(A \& B)$
  - For mutually exclusive events → $P(A \cup B) = P(A) + P(B)$ $(P(A \& B) = 0)$
- **Mutually Exclusive** - Events A & B have no overlap (intersection). (A) (B)

3.3 Permutations & Combinations

- **Multiplication Principle** - The total number of possible outcomes for a sequence of steps, where step 1 has n_1 possible outcomes, followed by step 2 with n_2 possible outcomes is $n_1 \times n_2$
- **Permutations** - The numbers of arrangements when r items are selected, without replacement, from set of n items is ${}_nP_r = \frac{n!}{(n-r)!}$
- **Combinations** - A set, group, list, or collection of r items is to be selected without replacement, from a set of n items. The number of possible outcomes is ${}_nC_r = \frac{n!}{r!(n-r)!}$

4.1 Probability Distributions

- **Random Variable** - A random variable returns a single numerical value as an outcome of a specific random experiment
  - Random variables are classified as discrete or continuous.
- **Probability Distribution of a Random Variable**
  - A probability distribution gives the probability for each value of the random variable (can be in the form of a graph table etc). They must satisfy
  - ① $\sum P(x) = 1$ where x takes on all possible values of the random variable.
  - ② $0 \leq P(x) \leq 1$ for every value of the random variable.
- **Mean μ**, the mean value, or expected value of X, $E(X)$

$$\mu = E(X) = \sum_{i=1}^{n} x_i P(x_i) = \sum x P(x)$$

- **Variance (σ^2)**

$$\sigma^2 = \sum (x-\mu)^2 P(x) \quad \text{or} \quad \sigma^2 = \sum x^2 P(x) - \mu^2$$

- **Standard Deviation (σ)**

$$\sigma = \sqrt{\sum (x-\mu)^2 P(x)} \quad \text{or} \quad \sigma = \sqrt{\sum x^2 P(x) - \mu^2}$$

4.2 Binomial Distribution

- **Binomial Distribution** - a probability distribution of a random variable which counts the number of successes with the following conditions
  - The procedure is repeated a fixed number of times (called trials) which are denoted by n.
  - Trials are independent of each other
  - Each trial has 2 possible outcomes: success or failure.
- Therefore, the parameters of this distribution are n & p.

- Binomial Formula - The general formula for calculating a probability of a binomial distribution is $P(x) = {}_nC_x p^x q^{n-x}$ for $x = 0, 1, \ldots n$ and x = number of successes, n = # of trails, p = probability of success in one trial, $q = 1 - p$ = probability of failure in one trial & ${}_nC_x$ is combination

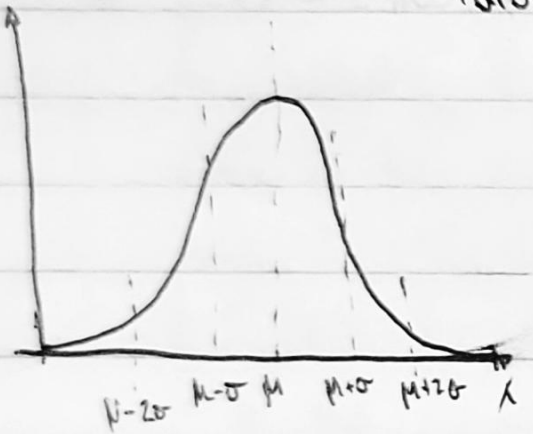
- Special formulas for Binomial Distribution
  - Mean - $\mu = np$
  - STDEV - $s = \sqrt{npq}$

4.3 Poisson Distribution

- Def. 3 - the poisson distribution is a discrete probability distribution is used to compute the probability of a given number of occurences, such as customers arrivals, or potholes on a road, in a fixed intervals. The interval can be time, distance, area, volume, etc.
  - 3 conditions for Poisson Distribution
    1. Random
    2. independent of each other
    3. equally likely to happen at any point of the fixed interval
- The general formula for calculating a probability of a value x in the Poisson distribution is $P(x) = \mu^x \frac{e^{-\mu}}{x!}$, μ = mean. $\mu = \lambda T$, where λ is the average rate of occurance per time and T is the length of the interval

4.4 Normal Distributions.

- The normal curve - is continuous probability distribution and bell-shaped
  - is symmetric, so that the mean = median = mode
  - approaches 0 probability in both tails after $\mu \pm 2\sigma$
  - total area of 1.



- Given mean, μ, & standard deviation, σ, of x to convert z to x use $x = z\sigma + \mu$
- To calculate z $\Rightarrow$ $z = \frac{x - \mu}{\sigma}$

5.1 Sampling Distribution

- Sampling Distribution - a sampling distribution of a statistic is the distribution of all possible values of the statistic when all possible samples are conducted with the same sample size n from the same population

5.2 Central Limit Theorem

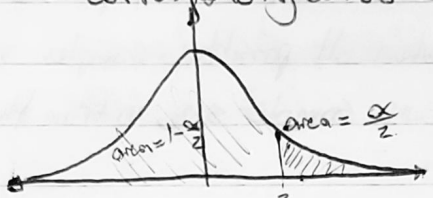
- Central Limit Theorem - the sampling distribution of $\bar{x}$, when $n \geq 30$, can always be approximation by a normal distribution with
  - $\mu_{\bar{x}} = \mu$, expected value of $\bar{x}$ is the mean
  - $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, the standard deviation of $\bar{x}$ is equal to the standard deviation of the original population divided by the root of the sample size (sometimes also called standard error of x).
- Computing z for $\bar{x}$: $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

5.3 Estimations and Confidence Intervals

- Interval Estimation - A point estimator is the statistic used to approximate the value of a population parameter.
  - For example, $\bar{x}$ is a point estimator for μ, $\hat{p}$ is used to estimate p, $\hat{s}$ is used to estimate σ.
- Confidence Interval - for a population parameter consists of a range of values and a percentage telling how confident we are that the population parameter lies in this interval. This percentage, or confidence level is the a priori probability that the true population parameter lies in this interval before the sample is selected.
  - This confidence level can be expressed (in decimal form) as $1 - \alpha$, where α is the probability of our confidence interval "missing" the value of μ.
    - confidence level = degree of confidence
    - 95% $\sim$ 0.05 = α.

Construct a Confidence Interval

- Step 1 - Choose a confidence level
- Step 2 - Find the corresponding value of z
  - $z_{\frac{\alpha}{2}}$



- Step 3 - Calculate the Margin of Error (E)
  - $E = z_{\frac{\alpha}{2}} \cdot$ standard error
  - standard error = $\bar{x} = \frac{s}{\sqrt{n}}$
    - ↳ for sample proportion $\bar{x} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Step 4 - Confidence interval for estimating μ is then given by $\bar{x} \pm E$ and for estimating p, given by $\hat{p} \pm E$.

5.4 Estimating a Population Mean.

- t-Distribution - similar to z, it is a bell-shaped, symmetric, and centered at 0 like z but with wider tails. The distribution depends of degrees of freedom (df) which equates to $n-1$. As df increases, t-distribution approaches the normal (when $n > 30$, $t \simeq$ normal).
- To find a confidence interval for μ when σ is unknown, where the population is approximately normal or $n > 30$; if one or both conditions are true: Step 1 - Calculate margin of error

  $E = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$, where $df = n-1$

  Step 2 - Calculate the confidence interval for μ

  $\bar{x} - E < \mu < \bar{x} + E$

  Note - when σ is known value σ instead of s and $z_{\frac{\alpha}{2}}$ instead of $t_{\frac{\alpha}{2}}$
- To determine sample size from σ, E & $z_{\frac{\alpha}{2}}$ use

  $n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E}\right)^2$ when σ not known use s.

5.5 Estimating a Population Proportion

- To find confidence interval for a population proportion

  Step 1 - Verify a large enough sample with $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$

  Step 2 - Find critical value for z

  Step 3 - Calculate E given by

  $$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{or} \quad z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

  n = sample size
  $\hat{q} = 1 - \hat{p}$
  $\hat{p}$ = sample portion

  Step 4 - Calculate confidence interval.

  $$\hat{p} - E < p < \hat{p} + E$$

  then state that the population portion p in this interval with $1-\alpha$ confidence.

- For a sample size n, with margin of error E

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 \hat{p}(1-\hat{p})}{E^2}$$

6.1 Hypothesis Testing - Single Mean

- Step 1 - Identify null & alternate Hypothesis
  - Null Hypothesis (H_0) - a statement that the population parameter equals some value
  - Alternate Hypothesis (H_1) - a statement that is the set of values different from the H_0 set

H_0	H_1	Test is
$\geq$	$<$	Left tailed
$\leq$	$>$	Right Tailed
$=$	$\neq$	Two-tailed

- Step 2 - Determine the appropriate ~~test~~ statistic in sampling distribution. For a test on one mean with a random sample, we use the following test statistics

Distribution	Sample Size	σ	Test Statistic
~~Approx~~ Normal	Any n	Known	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
Approx Normal	Any n	Unknown	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$, $df = n-1$
Any distribution	~~Any~~ $n \geq 30$	Known	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
Any distribution	$n > 30$	Unknown	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$, $df = n-1$

- Step 3 - Find critical values
- Step 4 - Make a decision
  - decide whether to reject H_0 or fail to reject H_0.
- Step 5 - Write a conclusion as it relates to the ~~stated~~ problem
- Type I & II Errors
  - Type I - we reject H_0 when it is true
  - Type II - we fail to reject ~~the~~ when it is false.

	If H_0 true	If H_0 false
Reject H_0	Type I Error (probability = α)	Correct decision
Fail to reject H_0	Correct decision	Type II Error (probability = β)

6.2 Hypothesis Testing - Single Proportion

- Step 1 - Identify null & alternate hypothesis.
- Step 2 - Calculate the test statistic.
  - $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$ n = number of trials, $\hat{p} = \frac{x}{n}$ sample proportion, p = population proportion with $q = 1 - p$.
- Step 3 - Find Level of significance & critical values.
- Step 4 - Decision
- Step 5 - Conclusion.

6.3 Testing Two Means

- H_0: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$, the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}}$$

df - the smaller of $n_1 - 1$ & $n_2 - 1$

6.4 Testing Two Proportions

- Independence of 2 Samples - samples from 2 populations are independent if the samples are not related in any way (no link).
  - If sample sizes are ~~independent~~ different, independent
  - If sample sizes are same may or may not be independent

	Population 1	Population 2
Population Proportion	p_1	p_2
Number of Success	x_1	x_2
Sample Size	n_1	n_2
Sample Proportion	$\hat{p}_1 = \frac{x_1}{n_1}$	$\hat{p}_2 = \frac{x_2}{n_2}$

- Test stat for independent samples

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \qquad \bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \qquad \bar{q} = 1 - \bar{p}$$

- Confidence interval For 2 proportions $(p_1 - p_2)$

$$(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E \quad \text{where } E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_1\hat{q}_1}{n_2}}$$

6.5 Dependent Data Testing.

- Test Statistic For Paired t Test

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \quad \& \quad \bar{d} - E < \mu_d < \bar{d} + E$$

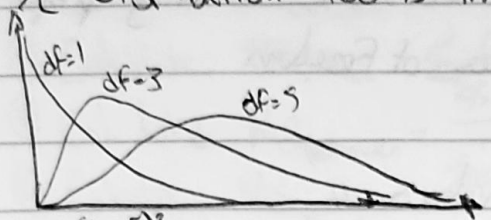
d = individual difference between 2 paired values

$$E = t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$$

$df = n - 1$

7.1 Goodness-of-Fit

- The χ^2 distribution looks like this!



  - is always right positive
  - is right skewed
  - it depends on degrees of freedom

- $\chi^2 = \sum_k \frac{(O-E)^2}{E}$

  O = observed count
  E = expected count
  k = number of categories

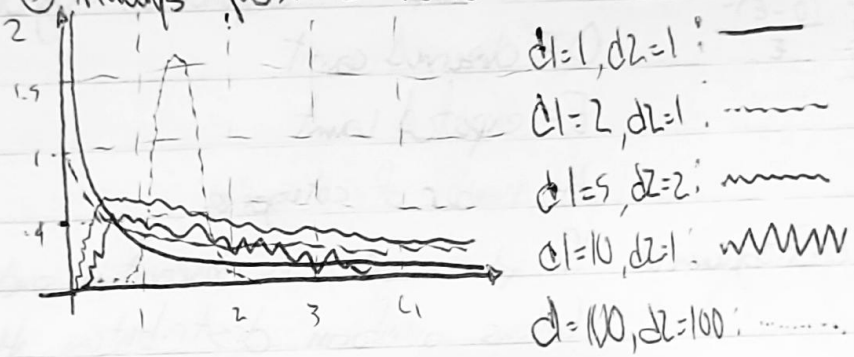
- Expected frequencies - If there are n observation and k categories which follows a uniform distribution, then the expected number in each category will be $E = \frac{n}{k}$.
  - In cases where we are given the probability for each category or any distribution, we can find the expected frequencies for the category by multiplying by the # of observations $E = np$.

7.2 Test of Independence - Contingency Tables.

- The degrees of freedom for a test of independence is given by $df = (\#\text{rows} - 1)(\#\text{colums} - 1)$
- The formula to find the expected value of any cell in row i and column j is: $E_{i,j} = \frac{(\text{row}_i \text{ total})(\text{column}_j \text{ total})}{\text{grand total}}$
  - The requirements are: - Random sampled data
    - Frequencies or actual counts (not percentages)
    - The expected values are all at least 5

73 ANOVA

- F-Distribution - The F distribution has the following properties:
  1. There are 2 degrees of freedom
  2. Skewed right
  3. Always positive values.



- The requirements of ANOVA are
  - the populations are approximately normal distributed
  - the populations have the same variances.
  - the samples are random and independent
- The null hypothesis is: H_0: $\mu_1 = \mu_2 = \dots \mu_k$. where k = number of populations sampled
- The alternate hypothesis is H_1: at least one of the means is different from the others
- F-Distribution Test Statistic - F is formed by the ratio of two variances $F = \frac{\text{Variance between samples}}{\text{variance within samples}}$
- Steps of ANOVA: let k = # of sampled populations
  n_i = size of sample i, $i = 1, \dots k$.
  $\bar{x}_i$ = means of sample i
  s_i^2 = variance of sample i
  N = total # of observation

① Find $\bar{x}_i$ & s_i^2 for each sample

② Find the mean of the sample means, $\bar{\bar{x}}$, of all data points

③ Calculate $SS_{between} = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2$

④ Calculate $SS_{within} = \sum (n_i - 1)s_i^2$

⑤ Calculate $MS_{between} = \frac{SS_{between}}{k-1} = \frac{SS_{between}}{df_1}$

⑥ Calculate $MS_{within} = \frac{SS_{within}}{N-k} = \frac{SS_{within}}{df_2}$

⑦ Calculate $F = \frac{MS_{between}}{MS_{within}}$, $df_1 = k-1$, $df_2 = N-k$

7.4 Correlation

: sample coefficient: $r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$

• $H_0: \rho = 0$ $H_1: \rho \neq 0$, $df = n-2$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

7.3 Regression

• $\hat{y} = b_0 + b_1 x$ $\quad b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$, $b_0 = \frac{\sum y}{n} - b_1\frac{\sum x}{n} = \bar{y} - b_1\bar{x}$

$y = \beta_0 + \beta_1 x$