STAT 200

Data $\leftrightarrow$ Variable (Characteristic of Interest)

- Qualitative / Categorical: can be ordered $\rightarrow$ **ordinal** variables
- Quantitative (attach **units**): measured on a numerical scale
    - Discrete, Continuous

Displaying Data

- Categorical Variables
    - (Relative) Frequency Tables: expressed in percentages/proportions add to 100%
    - Bar charts: $Area \propto Height \propto Proportions$, with legends, labelled axes, notes
    - Pie charts: can only introduce some # of categories
    - Contingency Tables: 2 categorical variable relationship + breakdown of data
        - Simpson's Paradox: the breakdown (grouping) of data can show an opposite trend compared to all data observed altogether
- Quantitative Variables
    - Histogram: can change interval width + no gaps between intervals
    - Stem-and-leaf Display: Trailing digit is always the leaf, whereas everything before it is the stem
        - Observation = Stem + Leaf, Stem is x-axis, Leaf is y-axis
        - A stemplot is histogram rotated $90°$
    - Boxplot: 5-number summary
        - $min, Q_1, Q_2(median), Q_3, max$
        - Upper fence and Lower fence to detect outliers; When no outliers, the whiskers are the max and min

Distributions

- Conditional Distribution: **Fix** one variable, distribution for the other variable
- Marginal Distribution (Frequency taken along the margins of the table): **Collapse** one variable, distribution of the other variable
- Describing a distribution
    - Shape
        - Unimodal, Bimodal, Multimodal
    - Center

        - Mean (Arithmetic Mean)
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
$$y_i = i^{th}\ observation$$
        -
        - Median
            - Arrange data in **ascending** order
$$Odd : median = \frac{n+1}{2} th\ observation$$
$$Even : median = average\ of \frac{n}{2}\ and\ \frac{n+1}{2}\ th\ observation$$
- Spread
    - Range $= max - min$
    - IQR $= Q_3 - Q_1$
        - $y_1, y_2, y_3, y_4, y_5 \to Q_2 = y_3, Q_1 = y_2(y_1, y_2, y_3), Q_3 = y_4(y_3, y_4, y_5)$
        - $y_1, y_2, y_3, y_4, y_5, y_6 \to Q_2 = \frac{y_3+y_4}{2}, Q_1 = y_2(y_1, y_2, y_3), Q_3 = y_5(y_4, y_5, y_6)$
    - Variance and Standard Deviation (Sample: s^2, s, Theoretical: σ^2, σ)
        - Non-negative
        - More spread = Larger variance
        - SD has the same unit as the data
        - $s^2 = s = 0$, if all observations are the same
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$
- Outliers?
    - Symmetric vs. Skewed (**The skew always follows the tail**)
    - Either $> Q_3 + 1.5 \times IQR$, or $< Q_1 - 1.5 \times IQR$
- Sensitivity to outliers
    - Sensitive: Mean; Range, Variance, SD
    - Non-sensitive: Median; IQR
- What to use?
    - Range is a crude measure
    - Symmetric: Mean, Variance, SD
    - Skewed: Median, IQR
- Shifting and Scaling of observations
    - Shifting ($\pm c$ to each observation)
        - Measure of center: $\pm c$
        - Measure of spread: **Unchanged**
    - Scaling ($\times c$ to each observation)
        - Measure of center & Measure of spread: $\times c$, (variance $\to \times c^2$)
- Standardization (for 2 different scale comparison)

- z-score

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

- Normal Model
  - Bell-shaped, unimodal, perfectly symmetric about μ, SD is σ (parameters)
  - Denoted by $N(\mu, \sigma)$
  - Standardizing values from the Normal Model

$$z = \frac{observation - mean}{standard\ deviation} = \frac{y - \mu}{\sigma}$$

  - Empirical Rule
    - $|y - \mu| \leq \sigma : 68\%, |y - \mu| \leq 2\sigma : 95\%, |y - \mu| \leq 3\sigma : 99.7\%$

Scatterplots, Association & Correlation

- Scatterplot
  - 2 quantitative variables: explanatory variable, response variable
  - Each observation is (x_i, y_i)
    - Line of regression always cross $(\bar{x}, \bar{y})$, the mean-mean point
  - Direction: Positive/Negative
  - Form: Linear/Non-linear
  - How scattered: Strong/Weak/No relationship
  - Outliers?
- Correlation: Degree of linear association between 2 **quantitative variables**
  - Correlation coefficient $(-1 \leq r \leq 1)$: measure the strength of **ONLY** linear correlation
  - r is sensitive to outliers, but will not change if x&y are swapped or the observations are scaled
- Calculating r
  - Standardize x & y

$$z_x = \frac{x - \bar{x}}{s_x}$$

$$z_y = \frac{y - \bar{y}}{s_y}$$

$$r = \frac{1}{n-1} \sum z_x z_y$$

- Covariance (**NOT** standardized)

$$cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$r = \frac{cov(X, Y)}{s_x s_y}$$

- Linear Regression $(y = b_0 + b_1 x)$
  - Predict on the response given the explanatory
  - Regression line

        - $b_1 = \frac{rs_y}{s_x}$
        - $b_0 = \bar{y} - b_1\bar{x}$
        - Predicting values: $\hat{y} = b_0 + b_1x$
    - **Beware of extrapolations (outside of observed ranges)**
    - Residuals (e): $e = y - \hat{y}, \sum(y_i - \hat{y}_i) = 0$
        - minimizing $\sum(y_i - \hat{y}_i)^2$: ordinary least squares regression
        - Plots: if done properly, there will not be any pattern
    - Fitting the model
        - x & y has sufficient linearity
        - Reality check: Do the values make sense?
        - Attention to outliers - if they are on the end, they affect more (exclusion of data?)

Sampling

- Census: complete collection of individuals under study $\rightarrow$ **population**, they are costly and inefficient
- **Sample** is a subset of population
- **Statistics** $(m, s) \overrightarrow{estimate}$ **Parameters** (μ, σ)
- Randomization
    - Randomness give samples that have population characteristics to prevent non-representativeness
    - **Sample size matters**, the actual number of individuals/subjects is important
- Methods
    - Sampling frame: every subject that **SHOULD** be sampled, this defines the population clearly
    - Sampling variability: difference in characteristics from *sample to sample* (As sample size increases, the sample variability decreases)
    - Simple Random Sampling (SRS)
        - Each subject has equal chance of being selected
        - Each possible sample of size n is equally likely
    - Stratified sampling
        - First breakdown population into strata (subset with same particular characteristic)
        - Then SRS drawn within each stratum, combining to form a stratified sample
    - Cluster sampling
        - Population is naturally divided into groups or clusters
        - One-stage cluster sample: choose random clusters to form a sample
        - Two-stage cluster sample: first choose random clusters, then SRS from each selected cluster
    - Multistage sampling

- Systematic sampling
    - Selecting every k^{th} individual from the sampling frame, this works if the list has no hidden order
- Bad sampling
    - Undercoverage: completely excludes or underrepresents certain kinds of individuals
    - Convenient sampling: sample based on easy availability & accessibility
    - Voluntary response bias: individuals with strong opinions tend to respond more often $\rightarrow$ overrepresentation
    - Non-response bias
    - Response bias
        - response influenced by how questions are phrased or worded
        - misunderstanding of a question
        - unwillingness to disclose the truth

Probability and Random Variables

- Sample space (S): set of all possible outcomes of a **random phenomenon**
- Event: outcome(s) from a random phenomenon, denoted by **UPPER CASE LETTER**
- $P(A)$: probability that an event A will occur
    - $0 \leq P(A) \leq 1$, $P(A) = 0$: impossible, $P(A) = 1$: certain
    - $\sum$ Probability of all **non-overlapping** events in $S = 1$
- True positive, True negative, False positive, False negative

	SPAM	HAM
Positive (Classified as Spam)	True Positive	False Positive
Negative (Classified as Ham)	False Negative	True Negative

$$F, P = \frac{ham\ as\ spam}{ham}$$
$$F, N = \frac{spam\ as\ ham}{spam}$$

- Independence of Events
    - $P(B|A) = P(B)$ & $P(A|B) = P(A)$
        - If A and B are independent, $P(A \& B) = P(A) \times P(B)$
- Random Variables
    - Discrete RV (can come from a **countable** set of values): a *numeric* quantity of an outcome from a random phenomenon
    - set of possible values + associated theoretical probability = probability model
    - Continuous RV: **uncountable** set of values $\rightarrow$ denoted by **density curves**

- Mean: Long run average of the observed values $\rightarrow$ **weighted sum**
- Variance: how spread out the observations are
- Probability Models
  - Normal RV: $X \sim N(\mu, \sigma)$
  - Uniform RV: $X \sim U(a, b)$, a is minimum, b is maximum
    - mean: $\frac{a+b}{2}$
    - sd: $\sqrt{\frac{(b-a)^2}{12}}$
  - Binomial Model: $X \sim Bin(n, p)$
    - n identical independent trials, p is success rate, q is failure rate
    - $P(X = x) = C_n^x p^x q^{n-x}$, where $C_n^x = \frac{n!}{x!(n-x)!}$
    - Mean: np, Variance: npq, SD: $\sqrt{npq}$

Sampling Distribution

- Assumptions
  - Randomness: samples are randomly drawn
  - Independence: individual values are independent
  - Sufficiency: large sample size ($np \geq 10, nq \geq 10$)
  - Sample size $\leq 10\%$ population
- Sampling distribution of proportions
  - Distribution of sample proportions, using $\hat{p}$ to estimate p
  - $\hat{p} \sim N(p, \sqrt{\frac{pq}{n}})$, $n = 10$ in class
- Sampling distribution of means
  - Distribution of sampled mean, using $\bar{y}$ to estimate μ
  - $\bar{y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
- Central Limit Theorem (CLT)
  - $y_1, y_2, y_3, \ldots, y_n$ be independent observations (random sample), if the sample is sufficiently large, $\bar{y}$ will approximate the normal model
  - If the sample comes from a normal model, sample mean is always normal

Confidence Interval and Hypothesis Testing

- Assumptions
  - The samples are randomly chosen within the population.
  - The samples are independent of each other.
  - The sample size is large enough to be representative of the population.
  - The sample size is less than 10% of the population
  - For 2 populations
    - For independent populations, each of the populations and the samples drawn follow the above assumptions.

- For paired populations, each **pair** are random, independent, sufficient, representative, and have unknown σ

- Sample size determination for sampling proportions

$$\because ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$\therefore n = \frac{(z^*)^2 \times \hat{p}(1-\hat{p})}{ME^2}$$

Without knowing the actual proportion, choose $\hat{p} = 0.5$ for largest sample size, **round up** to the nearest integer.

- Confidence Interval - Sample size $\times 2 \rightarrow$ Standard error $\div\sqrt{2}$ (C: probability that a Confidence Interval encloses p)
  - **We are $C\%$ confident that population proportion p is in (interval)**
- Hypothesis Testing Process
  - State the **parameter of interest**
  - State whether or not the assumptions are met
  - Define relevant notations and state the null hypothesis and the alternative hypothesis
  - Define relevant variables used in the calculations
  - State the testing model used (Specify one/(independent/paired)two-sample/proportion left/right/two-tail z/t-test)
  - Calculate the test statistic and compare it with the critical value
  - State whether or not the null hypothesis is rejected
  - Formulate a one/two-sentence conclusion

	$\hat{p}$	$\bar{y}$ (σ KNOWN)	$\bar{y}$ (σ UNKNOWN)	$\bar{y_1} - \bar{y_2}$ (INDEPENDENT)	$\bar{d}$ (PAIRED)
Population Parameter	p	μ	μ	$\mu_1 - \mu_2$	μ_d
Sampling Distribution Standard Deviation	$\sqrt{\frac{p(1-p)}{n}}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{\sigma}{\sqrt{n}}$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\frac{\sigma_d}{\sqrt{n}}$
Sampling Distribution Standard Error (SE)	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\frac{s_d}{\sqrt{n}}$
Confidence Interval	$\hat{p} \pm z^* SE$	$\bar{y} \pm z^* SE$	$\bar{y} \pm t^* SE$	$(\bar{y_1} - \bar{y_2}) \pm t^* SE$	$\bar{d} \pm t^* SE$
Test Statistic	$z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$z_0 = \frac{\bar{y}-\mu_0}{SE}$	$t_0 = \frac{\bar{y}-\mu_0}{SE}$	$t_0 = \frac{(\bar{y_1}-\bar{y_2})-\Delta_0}{SE}$	$t_0 = \frac{\bar{d}-\mu_{d_0}}{SE}$
Distribution Model	$N(0,1)$	$N(0,1)$	t_{n-1}	t_{min-1}	t_{n-1}
Margin of Error	$z^* SE$	$z^* SE$	$t^* SE$	$t^* SE$	$t^* SE$
Degrees of Freedom (df)	/	/	$n-1$	$min(n_1, n_2) - 1$	$n-1$

- Hypothesis Testing for Independence between Variables
  - Null hypothesis: H_0 : not associated

    Alternative hypothesis: H_A : associated
  - Define # of rows to be r, and # of columns to be c, total # of categories $(n) = r \times c$
  - Assume H_0 is true, calculate the expected value for each category, denote the expected value for cell (i, j), row i, column j, to be $E_{i,j}$
    - Each expected value ≥ 5.
  - Test Statistic $(X^2) = \sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$
  - Degrees of freedom (df) is $df = (r-1)(c-1)$
- Rejecting H_0 or not
  - Reject H_0 if the test statistic is larger than the critical value, or the p-value is less than α (significance level)
  - Do not reject H_0 if the test statistic is less than or equal to the critical value, or the p-value is larger than or equal to α (significance level)
  - p-value: conditional probability given H_0 is true, **NOT** the probability that H_0 is true.
  - Errors
    - Type I error: rejecting H_0 when H_0 is de facto true, the probability of this happening is α
    - Type II error: failing to reject H_0 when H_0 is de facto false.
- Analysis of Variance (ANOVA) $\rightarrow$ F-test
  - Comparing 2+ independent population means, compare both the center of the groups and the spread within the groups
  - Null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

    Alternative hypothesis H_A : not all μ are equal, at least two population means are different
  - Assumptions
    - k samples drawn from k populations must be independent of each other
    - Within each sample, individual observations are randomly chosen and are independent of each other
    - Within each population, the individual observations follow the **Normal Distribution** with some μ and a **common** standard deviation σ

SOURCE OF VARIATION	df	SUM OF SQUARES	MEAN SQUARES	F-RATIO
Treatments	$k-1$	SS_T: Treatment sum of squares	$MS_T = \frac{SS_T}{k-1}$	
Error	$N-k$	SS_E: Error sum of squares	$MS_E = \frac{SS_E}{N-k}$	
Total	$N-1$	SS_{Total}		$F = \frac{MS_T}{MS_E}$

- F^* has three parameters: α (significance level), ν_1 $(k-1)$, ν_2 $(N-k)$
- Treatment sum of squares represent the variation between groups, and the Error sum of squares represent variation within a group.
- $N = \sum n$, where n is the sample size from each population; $\bar{\bar{y}} = \frac{\sum y_{i,j}}{N}$
.