U of A

STAT151

Final EXAM

STUDY GUIDE

Lecture Notes

Section 1 / Chapter 1 in textbook - INTRODUCTION

- **Statistics** - “the science of collecting, classifying, analyzing, describing, and presenting data as well as drawing scientific conclusions about the phenomena being studied.”
  1. **Research design**
  2. **Descriptive stats**
  3. **Inferential stats**
     - **Hypothesis tests** (determine whether there is enough evidence to infer about N)
       - Compares the null hypothesis (H_0) and alternative hypothesis (H_a or H_1)
     - **Confidence intervals**
     - **Estimating about N** (based on n)

- **Purpose of statistics** - see the bigger picture, compare groups/treatments, find cause/effect relationships or associations between variables

- **Study/experimental units** - the subjects being studied

- Types of **variables** and **data**
  - **Qualitative** (categorical)
  - **Quantitative**
    - **Discrete** - countable
    - **Continuous**

Section 2 / Chapters 9-11 in textbook - GATHERING DATA AND RESEARCH DESIGN

- **Population size** - N
- **Sample size** - n
- **Sample fraction** - n/N
- **Sampling variability** - variation between samples taken from the same population. High sample size = lower variability.
- **Pilot study** - used as a trial run to:
  - Design real study
  - Test study methods, like questionnaires or recording devices

- **Parameter** vs statistic
  - **P**arameter - describes **p**opulation
    - Uses Greek letters. Population mean = μ. Population SD = σ
  - **S**tatistic - describes **s**ample
    - Sample mean = [x with line over top]. Sample SD = S

Randomness

- **Random Sampling** - an unbiased selection
  1. All individuals have an equal chance of being chosen
  2. Selection is independent (selection of one does not affect the selection of others)
- Methods:
  - **Random number table** - table with random numbers…
  - **Computer programs** - not completely random but very common
- Replacement:
  - **Sampling with replacement** - an individual can be selected more than once
  - **Sampling without replacement** - individuals can only be selected once
    - Violates the rule of independent selection (under random sampling) because not every individual will have an equal chance of being chosen. Common, esp. in social surveys.
- Types of random sampling:
  - **Simple random sampling (SRS)** - completely random and independent sampling
    - Ex. not useful: choosing random locations on a map of Alberta. Major cities are undercovered and rural communities are overcovered.
  - **Systematic** - first sample selected randomly, following are all selected sequentially
  - **Stratified** - dividing a population into homogenous subpopulations (strata) generally based on characteristics so that the strata are mutually exclusive. SRS is conducted within each strata using **proportional allocation**.
    - Ex. allocating areas of Alberta whose size is proportional to the internal population. Conduct SRS within each strata. (small section for Edmonton, larger sections in rural areas)
  - **Multistage** - selecting a sample from N, selecting a sample from the sample, selecting a sample from that sample, etc.
  - **Cluster** - select groups (clusters) and sample every individual in the clusters. Less accurate than other methods.
    - Ex. randomly select some apartment blocks then interview ALL tenants in the selected blocks
- Cluster vs Stratified:
  - Cluster → each cluster is considered one subject unit
  - Strata → elements within the strata are studied

Problems in sampling

- **Convenience sampling** (ex. observing animals next to a highway instead of deep in the wilderness; their behaviour would be different b/c they are accustomed to humans. Or surveying people in a mall.)
- **Voluntary response bias** - reduces randomness and creates bias
- **Response bias** - questions that appear to prompt or suggest a specific response.
- **Nonresponse bias** - a large number of study units fail to respond to some or any of the questions. May be caused by vague or unclear questions.
- **Incomplete sampling frame** - some members of the population are not included in the sampling frame (**sampling frame** must include all members of N)
- **Undercoverage** - portion of population given less or no representation. May be due to an incomplete sampling frame or using SRS instead of stratified.

Research design - the 5 Ws of research design

- **Study/experimental units** - on which variables are measured
- **Randomness**
- Explanatory and response variables
  - **Explanatory/predictor (independent)** - variables that are expected to affect others, but are not affected themselves (ex. age, but not height)
  - **Response (dependent)** - variables that are affected by others (ex. height, but not age)
- **Extraneous variables** - irrelevant explanatory variables; may interfere with the study. Sometimes not measured or cannot be measured (hidden variables).
- **Factors** - explanatory variables used in the study
  - Has to be **categorical**
- Aspects of design:
  - Temporal - when
  - Spatial - where
  - Purpose - why
  - Techniques - how

Observational vs Experimental

- **Observational:**
  - Called a **sample/social survey** when used for people's opinions

- Tries to estimate parameters of population
- **Random** selection of study units from target population
- No manipulation or control used, only observation
- **Population inferences** can be made
- Can observe correlation but CANNOT establish causation
- 2 types:
  - **Prospective** - subjects identified beforehand and data recorded as study proceeds
  - **Retrospective** - subjects identified and data collected after event has occurred

- **Experimental:**
  - Researcher sets up an experiment
  - **Randomness:**
    1. Study units randomly selected from population
    2. Study units are randomly assigned to treatment and control groups
  - **Manipulation of factors** (relevant explanatory variables)
    - **Treatment groups**
    - **Control groups** (involves placebos when study units are people)
  - **Extraneous variables** - controlled or made constant in all groups (constants)
  - **Response variable** - measured and recorded in all treatments and control groups
  - Both **causal** and **population** inferences can be made if selection and assignment are random
  - More accurate and definitive than observational, but may be unethical when observational are not
- **Replication** (in experimental studies)
  - Required to:
    - Confirm results
    - Apply statistical analysis
    - Estimate precision (standard deviation), give probability of accuracy
    - Number of replicates = n (number of samples / sample size)
    - Increase the "power" of the test
- **Blinding** (in experimental studies)
  - Involves those who could influence (subjects, test administrators) or evaluate (researchers, judges) the results
  - Double- vs single-blind

- Example:
  - Experimental - randomly select units, randomly allocate to Vitamin E supplement or placebo. More accurate, controlling extraneous variables, can draw causal inferences.

- Observational - many extraneous variables. Observe if there is a correlation between people taking vitamin E and having heart disease. Someone who takes vitamin E is more likely to have a healthier lifestyle, which prevents against heart disease.

Types of research design

- **Completely Randomized Single-Factor Design**
  - Test units allocated randomly to treatments/groups
  - Analyzed with **two-sample tests** if has two samples or **Single- Factor ANOVA** if there are more than 2
- **Paired Design**
  - Pairs of observations, generally each study unit is measured twice
  - **Paired Sample *t*-Test** - analyzes whether the mean difference between two (sets of) observations is zero
  - Analyzes two populations paired in space or time or by a relationship
  - Ex. before and after design
- **Randomized Block Design**
  - Uses **Randomized Block ANOVA** (Analysis of Variance)
  - Extension of the paired design
  - Experimental area is divided into blocks, each block is assumed to be homogeneous even though the blocks themselves differ
  - Requires an equal number of cells for all treatments
- **Completely Randomized Two-Factor Design**
  - Analyzed with **Two-Factor ANOVA**
  - The effects of two factors are tested at the same time
- **Multi-Way Factorial Design**
  - Analyzed with **Multi-Factor ANOVA**
  - More than two factors

Section 3 / Chapter 2 in textbook - DESCRIPTIVE STATISTICS: CATEGORICAL DATA

Grouping qualitative data

- Need to group the data before it is possible to analyze it

- **Frequency (f_i)** - number of times a value of a variable occurs
- **Frequency distribution** - a listing of all values for a variable and their frequencies. Can be either a table or a graph.
- **Relative frequency** - ratio of frequency of one value to total number of observations.
    - Class frequency / sum of all frequencies
    - $f_i / \Sigma f_i$
    - As a percent - formula x100 **(relative percent frequency)**

Other methods

- **Pie charts** - %frequency x 360° = angle
- **Bar graphs and contingency tables**
    - **Simple bar graph** - shows f of categories of one variable (same info as a pie chart)
    - **Area principle** - area under the graph must equal the value being presented
    - **Contingency tables** - gives frequencies for two qualitative variables at the same time **(bivariate data).** Also called two-way tables or cross-tabulation tables. Shows how one variable is contingent on the other.
    - **Segmented bar graph** - stacked bars. Similar to multiple bar graph.
    - **Multiple bar graph**

Table distributions

- **Joint distributions** - values in the body of a graph. Joint value of two events. Measured as %.
    - Frequency of joint event / grand total x 100
- **Marginal distributions** - total values for a variable (shown in bottom/right margins of graph). Measured as %.
    - Total frequencies of category / grand total x 100
- **Conditional distribution -** frequency distribution for one category of a variable at a time. Measured as %. Can be vertical or horizontal.
    - Frequency of specific category / total for variable x 100

Independence of variables

- Variables are independent if the distributions for the categories of one variables are all the same (ie. not dependent on the categories of the other variable)

Association of variables

- A change in one variable causes a change in another variable / one variable is dependent upon the other
- Conditional and marginal probabilities must be equal for there to be no association
- Data for a sample is a **subjective method** of deciding of there is any association. Need to use a chi-square test (inferential)
- Data for a population is an **objective method**
- A segmented bar graph can be used to assess association

Section 4 / Chapters 3-4 in textbook - DESCRIPTIVE STATISTICS: QUANTITATIVE DATA

Describing the distribution of a quantitative variable:

1. **Shape**
2. **Center**
3. **Spread**

Grouping quantitative data

- Uses **classes/bins** (used to group data)

- **Limit grouping**
  - Used more often in tables
  - **Lower class limit**
  - **Upper class limit**
  - **Class width**
  - **Class mark** - middle of the class
- **Cutpoint grouping**
  - **Used more often in graphs**
  - **Lower cutpoint**
  - **Upper cutpoint** - equivalent to the lower cutpoint of next higher class
  - **Class width** - difference between two cutpoints
  - **Class midpoint** - middle of class
- **Histograms**
  - Like a bar graph but no space between bars
  - For both discrete and continuous quantitative data
  - X-axis → classes of data
  - Y-axis → frequencies
- **Single-value frequency distribution / "grouping"** - class with one value. Used more often for discrete data.
- **Dotplots**
  - Useful for comparing two or many populations or treatments
  - Need to analyze shape, center, AND spread to reach a conclusion
- **Stemplots**
  - First 1-2 digits of the data are on the left (stem), following digits are listed on the right (leaf)
  - Like a sideways histogram but with more information
  - Can also be a **split-stem diagram**
    - Two of each number in the stem (first number takes 0-4, second takes 5-9)

        - May truncate the last digit and use the second last as the leaf
    - **Back-to-back stemplots** - use a common stem for two plots
- Comparison between histograms, dotplots, and stemplots
    - Can see the shape better in a histogram
    - A histogram can summarize large datasets, dot and stemplots are restricted to small datasets
    - Can see the details in dot and stemplots
    - Dotplots are good for visual comparison of many groups
- Other graphs for quantitative data
    - **Boxplots**
    - **Normal probability plots**
    - **Scatter plots**

Distribution shapes

- **Symmetrical** - use StDev to assess spread
    - Bell, triangle, uniform (horizontal line)
- **Skewed** - use quartiles to assess spread
    - **Left (negative)** - lower on left (leaning right)
    - **Right (positive)** - lower on right (leaning left)
    - **J-shaped** - ex. unlimited population growth (in ecology/biology)
- **Modality** - number of peaks
    - Unimodal, bimodal, multimodal

Measures of central tendency (centre)

- **Mean**
    - Very influenced by skewness (nonresistant)
    - More useful for symmetric data
    - **Population mean - μ**
        - Sum of all items in population / population size
        - $\Sigma y_i / N$
        - y = data point; i = i^{th} observation; N = population size
    - **Sample mean - ȳ**
        - Sum of all items in a sample / sample size
        - $\Sigma y_i / N$
- **Median**
    - **Resistant measure** (more resistant to outliers)
    - **Median class** - the class in which the median is found

- More useful for skewed data
- **Mode** - may be more than one (in accordance with modality)
- Comparison
  - Mean - center of gravity (if the median is the fulcrum). Skewness pulls the mean in the direction of the long tail
  - Mode - at the peak
  - Median - 50% area on one side, 50% on the other

Measures of variation (spread)

- **Range**
  - Difference between highest and lowest observations (max-min)
  - Biased by outliers
- **Sample variance - s^2**
  - Find the mean, find the distance from mean of each point, square, add and divide by n-1
- **Sample standard deviation - s**
  - Variance square root
- **Degrees of freedom (df)**
  - Number of **independent** observations
  - n - 1 in sample standard deviation
  - Explanation: http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-degrees-of-freedom-in-statistics
- **Population standard deviation - σ**
  - Denominator: N
  - σ^2 for variance
- Calculation should be rounded to one more decimal place than in the raw data

Five-number summary and boxplots

- **Percentiles** - divide the data set into 100 equal parts
- **Deciles**
- **Quartiles**
  - **First quartile** - median of first half of data set (when divided in two halves by median of entire data set)
  - **Second quartile** - median
  - **Third quartile** - median of second half of data set
  - **Interquartile range** - difference between first and third quartiles
- **Five-number summary**
  - Min, Q_1, Q_2, Q_3, Max

- Calculating **limits and outliers**
  - Lower limit → Q1 - 1.5 x interquartile range (IQR)
  - Upper limit → Q3 + 1.5 x IQR
  - Values outside of these limits are **potential outliers**
  - **Adjacent values** - the most extreme values that lie *within the limits*
    - If there are no potential outliers, the max and min are the adjacent values
- **Boxplots** (box and whisker plots)
  - Rectangles on a graph
    - Median of data = middle line (horizontal) in box
    - Q1 and Q3 = edges of box
    - Adjacent values (extreme points that lie within the limits) = whiskers
    - Potential outliers = asterisks
- When determining Q1 and Q3, don't include the median (no averaging in an odd number of items). Average all the medians in an even number of data points
- **Determining shape** using mean / median and quartiles → bottom of section 4 notes

Section 5 / Chapter 5 - THE STANDARD DEVIATION AS A RULER AND THE NORMAL MODEL

- **Density curve** - a model for a frequency distribution where the area/density under the curve represents the relative frequencies and probabilities
  - Area under curve = relative frequency = probability = percent of observations
- **Continuous probability model**
  - Smooth curve, used for continuous quantitative variables
  - Assign probabilities as the area under the density curve
  - Types of distributions:
    - Uniform
    - Normal
    - Exponential

SD as a ruler

- **Z-score** - number of SDs a data point is from the mean
  - $x = (y - \mu) / \sigma$
  - Z-score = (data point - pop. mean) / pop. SD
  - Useful for comparing grades (same grade in two classes → better relative standing in the class with the lower mean)
- **Z-distribution**
  - Same shape as original data
  - Center - mean is 0
  - Spread - SD is 1
- **Standardized normal variable**
  - Same as z-distribution
  - Creates a normal curve

Normal model

- The **normal distribution** is a specific type of continuous density curve
  - Forms a bell curve
  - Most populations are appr. normal (not completely normal)
- **Characteristics:**
  - Completely defined by its **mean** and **SD** - called the parameters (unique, like species names)
  - The notation of **N(μ, σ)** defines a normal distribution
  - **Area under curve = 1**

- Measures of center all coincide
- Extends indefinitely in either direction (only approaches the horizontal axis)
- Follows the empirical rule
- The **area under a single point is 0**

- **Empirical rule** - describes normal curves
  - The **68.26 - 95.44 - 99.74** rule:
    - 68.26% of all observations lie within 1 SD from the mean (either direction)
    - 95.44% is within 2 SDs
    - 99.74% is within 3 SDs
- **Standard Normal Table**
  - http://math.arizona.edu/~rsims/ma464/standardnormaltable.pdf
  - Represents percent of data found to the left of specific z-scores
- **Standardizing variables** using z-scores
  - Shape - no change
  - Measures of center - each point subtracts the mean, so mean becomes 0
  - Spread - each point (y) is divided by SD (lower case sigma), SD becomes 1
- **Non-linear transformations**
  - Include log, square root, etc.
  - Change the shape, center, and spread
- **Rules:**
  - Each point in a graph of continuous data has an area of 0
  - Round the answer to the same number of decimal points as in the standard normal table
- Misc.
  - Percentile notation: P_x

Assessing normality

- Knowing if the distribution is normal or appr. normal determines which kinds of tests you can use with the data
- Take a random sample and assess if it is normal or not
- **Histograms, stemplots, and dotplots**
  - Compare the distribution with a bell curve
  - Very subjective method
- **Normal probability plot**
  - Turns data into a line; if the line is straight, the data is normal (or appr. normal)
  - Also called a Q-Q plot
  - This rule should only be applied loosely to a small sample
  - Uses confidence interval lines
  - P-value (lower means not normal)
- **Chi-Square Goodness of Fit test (hypothesis test)**

    - Inferential test
    - Objective
- **Empirical rule**
    - Find z-scores and percentages and compare those with the empirical rule

Transformations of data

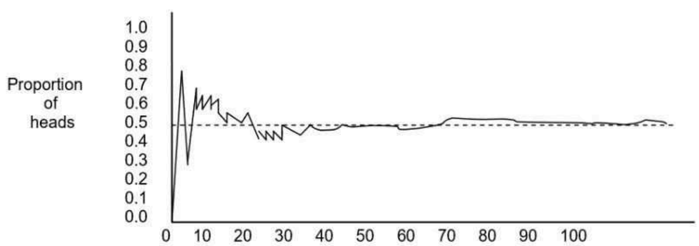
- **Linear** - adding/subtracting/multiplying/dividing by a constant for each data value
    - **Shifting data** - adding or subtracting a constant (spread and shape do not change)
    - **Rescaling data** - multiplying or dividing by a constant (shape does not change but spread does)

Section 6 / Chapters 12-13 - PROBABILITY CONCEPTS AND RULES

- **Probability theory** - the science of uncertainty
    - Mathematical basis for inferential statistics
- **Sample space (S)** - all possible outcomes for an experiment or trial
- **Outcome (O)** - a single observation of an experiment
- **Event** - any subset of the sample space (any outcome or set of outcomes)
- **Probability model** - a mathematical representation of a random phenomenon
    - Consists of sample space and a way of assigning probabilities to events
- Notation
    - **P(A)** = the probability of event A

- **Properties**
    1. Probability is between 0 and 1 or 0% and 100%, inclusive
    2. Sum of all possible outcomes or trials is 1. In other words, P(S) = 1
    3. P(impossible event) = 0
    4. P(guaranteed event) = 1

- **Interpretations of probability**
    1. **Equal-likelihood model** - prediction based on a theoretical model (ie. you will get a head 0.5 times after 1 coin flip, theoretically)
    2. **Law of large numbers (LLN)** - the probability of an event tends towards a single value the more trials there are. Example:



- **Proportion of an event** - cumulative percentage of the event

Equal-likelihood model

- **The *f*/N rule**
    - If there are N possible outcomes that are equally likely, A being an event, then:
    - P(A) = *f*/N
- Probability = relative frequency

Complementary and addition rules

- **Complement rule**, when **E** is an event that can occur and ^{c} is a negation
  - **P(E) = 1 - P(not E)**
  - **P(A^c) = 1 - P(A)**

- **Mutually exclusive events (disjoint)** - no overlap (no common outcomes)
- Events with **common outcomes** (not mutually exclusive)
  - **P(A or B)** | either A or B (or both) occur | $A \cup B$ | "A union B"
  - **P(A and B)** | both A and B at the same time | $A \cap B$ | "A intersect B"
  - **General addition rule** -
    - **P(A∪B) = P(A) + P(B) - P(A and B)**
  - **Special addition rule** - for mutually exclusive events (disjoint)
    - Same rule but without the last term because A and B can't both occur at the same time
    - You don't need to subtract the overlap because nothing overlaps
    - Mutually exclusive means that A intersect B is impossible

Conditional probabilities

- **P(B|A) "the probability of B given A"**
- **Conditional probability rule**
  - P(B|A) = (P(B intersect A)) / P(A)
- **Proof: (Joint event / total) / (total for category / total) = joint event / total for category = conditional**

Multiplication and independence rules

- **General multiplication rule** - dependent events
  - multiply both sides of conditional probability equation by the denominator
  - A x (B given A)
  - **P(B intersect A) = P(B|A) x P(A)**
- **Independence** - when the probability of one event does not affect another
  - B is independent of A if **P(B|A) = P(B)**
  - A and B are independent if **P(A and B) = P(A) x P(B)**

- **Special multiplication rule** - two or more independent events
    - P(A and B) = P(A) x P(B) (same as above)

- **Disjoint vs independent**
    - Disjoint = dependent (if one occurs, the other cannot)
    - Independent = not disjoint (can both occur, regardless of each other's occurrence)
    - Joint events (can occur together) = either dependent or independent
    - Dependent events = either joint or disjoint
- **Dependence =/= causality**
- **Tree diagrams**
    - First set of branches - unconditional probabilities of categories for one variable
    - Each node branches into categories for other variable. Number of these nodes represents number of total outcomes
- **Total probability rule**
    - P(A) = P(A and B) + P(A and B^c)

Section 7 / Chapter 14 - RANDOM VARIABLES AND PROBABILITY MODELS

- Applying mean, SD, and relative frequency distributions to probability distributions
- Random variables, two types:
  - Discrete random
  - Continuous random (ex. the normal distribution)

Probability distributions and discrete random variables

- **Random variable** - a quantitative variable whose value depends on chance (or as close to chance as possible)
- **Probability distribution** - a listing of possible outcomes with their respective probabilities (or a formula for the probabilities)
- **Discrete random variable** - a quantitative variable whose value depends on chance and can be listed (continuous data cannot). Uses a capital letter.

Formulas:

- **Sum** of the **probabilities** of a discrete random variable (ie 100%)
  - $\Sigma P(x) = 1$
- **Mean** of a discrete random variable
  - The mean is known as the **expected value (E(X))**
  - $\Sigma xP(x)$
  - Explanation of formula: https://www.thoughtco.com/formula-for-expected-value-3126269
- **Standard deviation** and **variance** of a discrete random variable
  - $\sigma = \sqrt{\Sigma(x-\mu)2\ P(x)}$
  - $Var(X) = \Sigma(x-\mu)2\ P(x)$

Interpretation of the mean of a random variable

- More observations = average of random variable X is closer to mean

Linear transformations and combinations of random variables

- **Adding/subtracting** by a constant shifts the mean/expected value but does not change the spread.
    - $E(X \pm b)$ = $\mathbf{E(X) \pm b}$
    - $Var(X \pm b)$ = $\mathbf{Var(X)}$
- **Multiplying** by a constant multiplies the mean by that constant and the variance by the square of the constant
    - $E(aX)$ = $\mathbf{aE(X)}$
    - $Var(aX)$ = $\mathbf{a^2Var(X)}$
- **Both** addition/subtraction and multiplication
    - $E(aX \pm b)$ = $\mathbf{aE(X) \pm b}$
    - $Var(aX \pm b)$ = $\mathbf{a^2Var(X)}$
    - $SD(aX \pm b)$ = $\mathbf{|a|SD(X)}$
- **Sums** of random independent variables
    - The mean of the sum is the sum of the means
        - $\mathbf{E(X + Y) = E(X) + E(Y)}$
    - The mean of the difference is the difference between the means
        - $\mathbf{E(X - Y) = E(X) - E(Y)}$
    - Variance of sum or difference is the sum of the variances
        - $\mathbf{Var(X \pm Y) = Var(X) + Var(Y)}$
    - With constants:
        - $E(aX + bY + c)$ = $\mathbf{aE(X) + bE(Y) + c}$
        - $Var(aX + bY + c)$ = $\mathbf{a^2Var(X) + b^2Var(Y)}$

Continuous probability distributions

- **Total area** under curve = 1

The Uniform Model

- Simplest continuous probability density function
- **Constant value for y** between specified values for x (a and b)
- $X \sim U(a,b)$
- **Height:**
    - Area of 1 / length
    - $1 / (b - a)$
- The random variable of X is equally likely to take values between a and b
- **Probability** between two values
    - $P(c < X < d) = (d - c) / (b - a)$
    - Distance of interest / total distance

- Expected value **(mean) = median**
    - $E(X) = \boldsymbol{(b + a) / 2}$
- **Variance**
    - $Var(X) = \boldsymbol{(b - a)^2 / 12}$
- **Interquartile range**
    - $\boldsymbol{(b - a) / 2}$

Section 8 (Chapter 15) - SAMPLING DISTRIBUTIONS

Sampling error and distributions

- **Sampling distribution of the sample mean** - distribution of the y-bar (sample mean) values for variable x and sample size n
    - Also called: **distribution of all possible sample means of a given sample size**, or **distribution of the variable y-bar**
- **Sampling error** - the error from using a sample to infer about a population, like mean or SD
- Sample size and sample error
    - When n = N, y-bar = μ

Mean and standard deviation of the sample mean

- **Mean** of sample mean
    - The mean of all possible sample means is the population mean
    - ***Mean(y-bar) =*** $\boldsymbol{\mu_{y\text{-}bar} = \mu}$
- **Standard deviation** of sample mean
    - Called the **standard error of the sample mean** because it determines the amount of sampling error to be expected when inferring to the population
    - The formula applies to **sampling with replacement from a finite population** or an **infinite population**
    - SD of y-bar is the SD of the variable divided by the square root of n
    - $\boldsymbol{\sigma_{y\text{-}bar} = \sigma / \sqrt{n}}$
    - As n increases, **SD of the sample means gets smaller** until it is 0 when n=N

Sampling distribution of the sample mean for normally distributed variables

- Involves 3 aspects:
    - **Shape** - the sampling distribution of all possible sample means are normally distributed
    - **Center** - mean(y-bar) = μ
    - **Spread** - SD(y-bar) = $\sigma/\sqrt{n}$

Standardized version of y-bar (sample mean)

- $z = (\bar{y} - \mu) / (\sigma / \sqrt{n})$
- z-score = (sample mean - population mean) / (sample mean SD)

Sampling distribution of the sample mean for any distribution type

- **Central Limit Theorem (CLT)**
    - The **y-bar (mean) variable is appr. normally distributed** for any data, esp. larger sample sizes
    - **n > 30** → generally accepted as a "large" sample size
    - See notebook
- **Shape** - normal
- **Center** - mean(y-bar) = μ
- **Spread** - SD(y-bar) = $\sigma/\sqrt{n}$

Assumptions and conditions of the sample mean distribution

- **Independence assumption** - all samples must be independently drawn from the population
- **Randomized condition** - everything must be random
- **Sample size assumption and condition** -
    - **Large enough** - sample size must be "large"
    - **10% condition** - applies to *sampling without replacement*. Sample size should be no more than 10% of the population

Sampling distribution for the difference between two means

- $\text{y-bar}_1 - \text{y-bar}_2 = \mu_{\text{y-bar1 - y-bar2}} = \boldsymbol{\mu_1 - \mu_2}$
- **Standard deviation**
    - Square root of $(\sigma^2_1/n_1 + \sigma^2_2/n_2)$

Sampling distribution of a sample proportion

- Sometimes mean and SD cannot be calculated (ex. with yes/no outcomes) - in this case, we find sample proportions
- **Population proportion (*p*)** - percent of the population with a specific attribute (a parameter)

- **Sample proportion ($\hat{p}$)** - percent of a sample from a population with a specific attribute (a statistic)
  - **y / n**
  - y = number of members with specific attribute; n = sample size
  - **Appr. normally distributed** for a large sample size (n)
    - Condition for normality - $outcome_1$ > or = 10 and $outcome_2$ > or = 10
- **Sampling distribution of sample proportion**
  - **Mean ($\hat{p}$) = p**
  - SD (standard error, or SE) = **$\sqrt{(p(1 - p) / n)}$**
- **Z-score of proportions**
  - $(\hat{p} - p)$ / SD formula

Assumptions and conditions of the sampling distribution of a sample proportion ($\hat{p}$)

- **Independence assumption**
- **Randomized condition**
- **Sample size assumption and condition**
  - **Large enough** (at least 10)
  - **10% condition**
- ^ all are the same as above

Sampling distribution for the difference between two sample proportions

- Difference → $\hat{p}_1 - \hat{p}_2 = p_1 - p_2$
- SD of difference → square root of $(((p_1(1 - p_1)) / n_1 + (p_2(1 - p_2)) / n_2)$

Section 9 (Chapters 16-18) - INFERENTIAL STATISTICS: HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

Stats in the scientific method

- Research design (sampling strategy)
- Data collection
- Descriptive and inferential stats
- Hypothesis tests
- Drawing conclusions
- Scientific writing

Inferential stats

- Two main components
  - **Hypothesis testing**
  - **Confidence intervals**

Hypotheses

- Objectives in research lead to two possible outcomes:
  - **Null hypothesis (H_0)**
    - No relationship among variables, no difference among groups
  - **Alternative hypothesis (H_a)**
    - There is a relationship between variables
- Research hypotheses are usually based on the alternative hypothesis
- Two types of objectives/hypotheses encountered in research
  - **Differences between treatments of one variable**
  - **Relationships between two or more variables** (positive or negative)
- Tailedness of hypothesis tests
  - "Tail" refers to the skewness of a distribution (left skewed = tail on left)
  - **One-tailed** → data deviates in one direction from the reference
    - Can be left-tailed or right-tailed
  - **Two-tailed** → data deviates in either direction from the reference

Various statistical test jargon

- **Test statistic** - used to decide whether or not to reject the null hypothesis
  - **Calculated value** - calculated from data
  - **Critical value** - obtained from a table showing theoretical distribution; compared with calculated value to decide about hypothesis
- **Rejection region** - set of values for the test statistic that lead to a rejection of the hypothesis
- **Nonrejection region**
- **Power of a statistical test** - probability of making a correct decision (to increase: choose most powerful test, use larger sample sizes)
- **P-value** - probability of making a type I error

Errors

- **Type I (α)** - rejecting null hypothesis when it is true
  - P (making a type I error) = α → significance level
  - Probability determined with a theoretical distribution
  - Alpha → the maximum probability of the type 1 error you will allow when rejecting H_o
- **Type II (β)** - accepting null hypothesis when it is false
  - P (making a type II error) = β
  - Generally not determined by a hypothesis test
- Relationship between alpha and beta
  - **Inversely proportional**
  - Only way to decrease both → increase sample size

Steps in hypothesis testing

1. **Choose test**
   a. ex. t-test, ANOVA, correlation
   b. Consider which hypothesis, types of variables, purpose, etc.
   c. Choose whether to test differences between one variable or association between two variables
   d. Categorical → chi-square
2. **State hypothesis**

. Also identify **significance level (α)**

a. Common α assumption = 0.05 (less than a 5% chance of making a mistake)

3. **Calculate test statistic**

. Calculated value (not critical)

4. **Decide (non)rejection of H_o and strength of evidence for or against H_o**

. Find **P-value** (probability of type I error) from theoretical distribution with appropriate n or df

a. Rules:

i. If **P-value ≤ significance level (α), H_0 is rejected**

ii. If **P-value > α, H_0 is accepted** (alternative hypothesis is rejected)

5. Conclusion (in words)

- **Critical value approach**
    - Can be used instead of step 4 to decide whether to reject H_0
    - Gives same conclusion, except for strength of evidence
    - Rules:
        - If |calculated test statistic|≥ |critical value|, H_0 is rejected
        - If |calculated test statistic|< |critical value|, H_0 is accepted

Confidence intervals

- **Point estimate** of a parameter - value of the corresponding sample statistic used to infer the parameter
    - **Of a population mean** - y-bar (sample mean used to estimate population mean)
- Confidence-interval estimate
    - **Confidence interval (CI)** - range of numbers derived from a point estimate
    - **Confidence level** - amount of confidence (as a %) that the parameter lies within the confidence interval. Confidence level = 1-alpha
    - **Confidence-interval estimate** - level and interval

(Non)parametric methods in inferential stats

- **Parametric**
    - Estimation of population parameters
    - Involve assumptions about the population:
        - Random sampling
        - Normally distributed
        - Equal variances between samples
    - Ex. t-tests, ANOVA
- **Nonparametric**
    - Fewer assumptions, only assumption is random sampling
    - Can be applied to categorical data
    - Less powerful
    - Ex. chi-square test

Section 10 (Chapters 16-19) - INFERENCES FOR ONE AND TWO POPULATION PROPORTIONS

- Two types of proportion inferences:
  - One population proportion
  - Two population proportions

Inferences for one population proportion (one-sample case)

- **Sampling distribution of the sample proportion**
  - **Mean ($\hat{p}$) = p**
  - SD (ie standard error) ($\hat{p}$) = $\sqrt{\mathbf{p(1-p)/n}}$
  - $\hat{p}$ is appr. normally distributed if:
    - Number of successes → $np \geq 10$
    - Number of failures → $n(1-p) \geq 10$

Assumptions and conditions of the sample proportion distribution

- **Independence assumption**
- **Randomized condition**
- **Sample size assumption and condition**
  - **Success/failure condition (large enough)** - number of successes (y) and failures (n - y) are both at least 10
  - **10% condition (not too large)** - when sampling w/o replacement, sample size should be no more than 10% of the population

One population proportion hypothesis testing

One-proportion z-test

1. **Check purpose and assumptions** to confirm this is an appropriate test

Purpose of test - to check for differences between a population proportion (based on a sample proportion) and a hypothesized proportion (p_0)

Assumptions:

- Simple random sample; independent sampling
- Large (at least 10)
- 10% condition

2. State the null and alternative hypotheses

a. Null hypothesis → H_0: $p = p_0$

b. Alternative hypothesis (one of the following) →

i. H_a: $p \neq p_0$ (two-tailed)

ii. H_a: $p < p_0$ (left-tailed)

iii. H_a: $p > p_0$ (right-tailed)

3. Obtain the calculated value of the test statistic

a. Proportion z-score formula

4. Decide whether to reject or accept the null hypothesis and state the strength of evidence

Difference between alpha and p-value

- Significance level (alpha) = probability of making a type 1 error
- P-value taken from a table and based on the calculated test statistic = observed probability of a type 1 error

Hypothesis test general formula

- **Test statistic = Estimate - H_0 value / SE**

Confidence intervals for one population proportion

- Confidence interval general formula
  - **Estimate ± Critical value * SE (estimate)**
- Margin of error = half of confidence interval

- Point estimate = average of two p endpoints

One proportion z-interval procedure

Purpose: find population proportion based on sample proportion

Assumptions:

1. Simple random sample, independence
2. Success/failure assumption (at least 10)
3. Sample size no more than 10% of population

1. For a given confidence level (1 - alpha), use the **z-score table to find $Z_{a/2}$** (critical value)
2. **Find confidence interval for p** from the endpoints:
   - **p-hat ± $z_{a/2}$ x √((p-hat (1- p-hat)) / n)**
3. **Interpret confidence interval**
   - If p is close to 0.5, it is more accurate and n doesn't have to be as big

Relationship between hypothesis tests and confidence intervals

- **Rejecting H_0** - if and only if the (1 - alpha) confidence interval for p does not contain the hypothesized proportion
- **Not rejecting H_0** - if the confidence interval does contain the hypothesized proportion
- **Two conditions** must be met to ensure the conclusions from a hypothesis test and confidence interval performed on the same data are the same:
  - Confidence level is the complement of the significance level applied in the hypothesis test

Determining required sample size

- **Conservative (not guessing)**
  - When ME = maximum margin of error,
  - **$n = 0.25 (z_{a/2} / ME)^2$**
- **Making an educated guess**
  - $\hat{p}_g$ = educated guess
  - **$n = (z_{a/2} / ME)^2 (\hat{p}_g (1- \hat{p}_g))$**

- Requires previous information
- Educated guess should be as close to 0.5 as possible

Plus four confidence interval for small samples

- When the success/failure condition is not met
- Sample proportion of **$\hat{p}$ = y / n** becomes **p [tilde] = (y + 2) / (n + 4)**
- Sample size **n** becomes **n [tilde] = n + 4**
- Confidence interval - same basic formula

Inferences for two population proportions

- **Distribution of the difference between two sample proportions** (large and independent samples)
  - Difference **- $\hat{p}_1 - \hat{p}_2 = p_1 - p_2$**
  - SD of the difference **- $\sqrt{((p_1(1-p_1)) / n_1) + (p_2(1-p_2)) / n_2}$**

Hypothesis test for the difference between two population proportions

Two-proportions z-test

Purpose: find difference between two population proportions based on two sample proportions

Assumptions:

1. Simple random sample, independence
2. Two independent samples
3. Large (at least 10)
4. Not too large (no more than 10% of population)

Null hypothesis: $p_1 = p_2$

Alternative hypothesis:

$p1 \neq p2$

$p1 < p2$

$p1 > p2$

1. Find the calculated value of the test statistic:
   - **$z = (\hat{p}_1 - \hat{p}_2) / \sqrt{(\hat{p}_{pooled} (1- \hat{p}_{pooled}) (1/n_1 + 1/n_2)}$**
   - H_0 value = 0

2. Decide whether to reject H0 and determine strength of evidence

Confidence interval for the difference between two population proportions

Two-proportions z-interval procedure

Purpose: find confidence interval for the difference between two population proportions based on two sample proportions

Assumptions: same as above

1. For a given confidence level (1 - alpha), **find the critical value ($Z_{a/2}$)** from a standard normal table
2. The endpoints for p are given by:
   - **$(\hat{p}_1 - \hat{p}_2) \pm z_{a/2} \times \sqrt{((p_1(1-p_1)) / n_1) + (p_2(1-p_2)) / n_2}$**
3. Interpret confidence level

Relationship between hypothesis tests and confidence intervals for two population proportion inferences (two tailed)

- **Rejecting H0** - if and only if the confidence interval does not contain 0 (endpoints are either both negative or both positive)
- **Not rejecting H0** - confidence interval contains 0 (one endpoint is positive and the other is negative)
- *If 0 is within the interval, there is no significant difference, thus the null hypothesis is not rejected*
- Conditions:
  - Confidence level is a complement of the significance level (alpha)
  - Same sidedness/tailedness

Sample size required for estimating difference between two population proportions

- Not guessing:
  - $n_1 = n_2 = 0.5 (z_{a/2} / E)^2$
- Making an educated guess using previous information

Section 11 (Chapter 23) - CHI-SQUARE TESTS

The chi-square distribution

- Applied to **categorical data**
- Chi is the Greek letter
- There is one chi-square distribution for each degree of freedom
- **Basic properties:**
    1. Total area under curve = 1
    2. The curve starts at 0 on the horizontal axis and extends infinitely to the right, never touching the axis
    3. Right skewed
    4. Higher the df → more like a normal curve
- The df is always rounded down
- Chi-square tests are right-tailed (never two-tailed)

Chi-square goodness-of-fit test

- Hypothesis test for one categorical variable

Purpose: to compare observed frequencies with expected (theoretical) frequencies

Assumptions:
- Simple random sample, independence
- Sample is no more than 10% of the population
- All expected frequencies are at least 5

Hypotheses
- H_0: no difference between observed and expected
- H_a: observed and expected frequencies are different

Calculation of expected frequencies
- E = np, where p is an expected theoretical proportion

Test statistic:

- $\square^2 = \Sigma$ (observed - expected)^2 / expected
- df = number of categories - 1

Using the chi-square table

- Calculated values are in the body of the table
- P-values at the top
- df on the side

Chi-square test for independence / association

- Hypothesis test involving two variables
- Requires a contingency table

Purpose: to test if two variables are independent or associated

Assumptions:

- Simple random sample, independent sampling
- Sample is no more than 10% of the population
- All expected frequencies are at least 5

Hypotheses

- H_0 - there is no association between the variables
- H_a - there is an association between the variables

1. **Calculate expected frequencies**
    - E = RC / n
    - E is expected frequency, R is row total, C is column total, n is sample size
2. **Calculate test statistic**
    - Same formula as above
    - df = (#rows - 1)(#columns - 1)

Contributions to the chi-square test statistic

- **Signed terms** - data points that contribute most to the test statistic (ie. most different from the expected value)
  - The (O - E)2/E terms
  - Neutral in the equation, but given signs (+/-)
- **Standardized residual**
  - Normalizes data to make it easier to compare
  - **(Observed - Expected) / √ Expected**
- **Assumptions**
  - If one or more assumptions are violated:
    - Combine rows or columns to increase expected frequencies when they are too small
    - Eliminate rows or columns where expected frequencies are too small
    - Increase sample size
- **Association and causation**
  - A chi-square test indicates association but not causation

Chi-square test for homogeneity

- Exactly the same as the test for independence/association except for the wordings of the hypotheses and conclusion

Section 12 (Chapter 20) - INFERENCES FOR ONE MEAN

- Inferences about a population are made with one sample, using two methods:
  - One proportion (section 10)
  - One mean

The *t*-distribution

- If σ is known (rare)
  - **$z = (\bar{y} - \mu_0) / (\sigma / \sqrt{n})$** (z = (y-bar - μ0) / (σ / √n))
  - Similar to proportion z-score formula
- Unknown σ:
  - Estimate σ using the sample SD and obtain:
  - **The student *t* version of the sample mean (y-bar)**
    - ***t* = (y-bar - μ) / (s / √n)**
    - df = n-1 because it uses sample SD not population SD
- The *t*-distribution is almost as statistically important as the normal distribution
- Why it is called the student t-distribution:
  - *t*-distribution inventor William Gosset originally published it under the name of "student"
- **Properties:**
  1. Total area under curve = 1
  2. Symmetrical at 0
  3. Extends infinitely in either direction, never touching the x-axis
  4. Different t-distribution for each sample size (identified by df)
  5. Higher df = *t*-curve approaches the normal curve until is is the normal curve when df = infinity

Applying the one-mean *t*-test and the one-mean *t*-interval procedure

- Guidelines for inferences:
  - **Small samples (n < 15) →** *t*-interval procedure is only used when the variable under study is normal or appr. normal
  - **Moderate samples (15 < n < 30) →** t-interval procedure is used unless the variable is very far from normal or there are outliers
  - **Large samples (n > 30) →** t-interval procedure can be used
- The *t*-test and *t*-interval procedure are fairly resistant to violations of normality but can be affected by outliers

- ***t*-table**
  - Either one-tailed, two-tailed, or both

One-mean t-test (also: one-sample)

Purpose: test for a difference between population mean and a hypothesized mean

Assumptions:
- Simple, random sample, independent
- Normal or large sample
- Sample is no more than 10% of the population

Hypotheses:
- H0: $\mu = \mu_0$
- Ha: $\mu \neq \mu_0$ or $\mu < \mu_0$ or $\mu > \mu_0$

One-mean t-test formula:
- **$t = (\bar{y} - \mu_0) / (s / \sqrt{n})$**

Confidence interval for one population mean

One mean t-interval procedure

1. For a given confidence level (1-alpha), use the t-table showing the t-test critical values to find $t_{\alpha/2}$ using the appropriate df
2. Confidence interval for μ is given by the endpoints:
   a. **$\bar{y} \pm t_{\alpha/2} \times (s / \sqrt{n})$**
3. Interpret confidence interval

Confidence intervals, margins of error, and precision

- Margin of error is half the confidence interval
- Y-bar is in the middle of the confidence interval

- **Precision:**
    - Margin of error determines the precision with which μ can be estimated
    - Increased by increasing sample size
    - Length of confidence interval is inversely proportional to precision; a shorter confidence interval is ideal