Exam 1: Chapters 1-6

Exam 2 Chapters 7-11

Chapter 7

Chapter 8

Chapter 9, 10, and 11

Post Exam 2 Material: Chapters 12-13

Objectives for Comprehensive Final Exam

The final exam will be a comprehensive two-hour assessment, covering the full scope of the course. It will predominantly focus on the material covered post-Exam 2 (approximately 60%), while also revisiting key concepts from early material (approximately 40%). This structure ensures a thorough evaluation of your understanding of the entire course content.

Exam 1: Chapters 1-6

Chapter 1: Introduction to Statistics

- Define and demonstrate knowledge of the **three branches of statistics**:
    - **Data Collection**: The process of gathering information.
    - **Descriptive Statistics**: Summarizing and organizing data.
    - **Inferential Statistics**: Drawing conclusions from data.
- Define and distinguish between a *population* and a *sample* including their respective symbols; population parameters by Greek letters, sample statistics are denoted by Latin letters.
- Determine whether a listing of objects refers to a **population or a sample**.
- Identify situations that exemplify **probability** or **inferential statistics**.

Chapter 2: Data Types and Distribution Shapes

- Identify data as **univariate, bivariate, or multivariate**.
- Recognize and classify variables as **categorical/qualitative** or **numerical/quantitative**.
- Describe the **shape of a distribution**:
    - **Peaks**: unimodal, bimodal, multimodal.

    - **Symmetry**: symmetric, right skewed, or left skewed.
    - **Outliers**: Identify and distinguish "real" outliers from the explicit points.
- **Interpret histograms** to describe shape and identification of outliers.

Chapter 3: Descriptive Statistics in R

- Given R output, identify the statistics: **mean, median, variance, standard deviation, and quartiles**.
- Understand and state the **formulas** for sample mean and sample variance:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Calculate **standard deviation from variance**:

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

- Calculate the **Interquartile Range (IQR)** and explain **quartiles** in non-mathematical terms.

$$\text{IQR} = Q_3 - Q_1$$

- Write down the **five-number summary** from R output and interpret **modified boxplots**.
- Using the five number summary identify inner and outer fences.

$$\text{IF}_\text{L} = Q_1 - 1.5 \times \text{IQR}, \quad \text{IF}_\text{H} = Q_3 + 1.5 \times \text{IQR}$$

$$\text{OF}_\text{L} = Q_1 - 3 \times \text{IQR}, \quad \text{OF}_\text{H} = Q_3 + 3 \times \text{IQR}$$

- Identify **explicit** points using the **1.5 IQR rule** and evaluate if they are **"real"**.
- Draw/complete a modified boxplot from the **five number summary** and **1.5 IQR rule**.
- Interpret the results of a modified boxplot or side-by-side boxplots.
- Decide on the appropriate **measures of location and spread** for given data.

Chapter 4: Probability

- Write down the **sample space** for experiments and determine **disjoint events**.
- Understand the **frequentist interpretation of probability**:

$$\lim_{n\to\infty} \frac{n(E)}{n} \approx P(E)$$

- State and check the axioms associated with a probability space Ω:

$$\text{For any event E} \subseteq \Omega, \quad 0 \leq P(E) \leq 1$$

$$P(\Omega) = 1$$

$$\text{For any event E} \subseteq \Omega, P(E) = \sum_{\omega \in \text{E}} P(\omega)$$

- Calculate a **probability** using:
  - The theoretical (or classical) approach (if equally likely can be assumed):

$$P(A) = \frac{\text{Number of outcomes in A}}{\text{Total number of outcomes}}$$

  - Empirical approach using a provided probability distribution table.
- Use **Venn diagrams** to visualize and calculate probabilities.
- Calculate probabilities using **probability rules**:
  - **General Addition Rule** for any events A and B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

  - **Special Addition Rule** for **disjoint events** A and B

$$P(A \cup B) = P(A) + P(B)$$

  - **Complement Rule**

$$P(A^{'}) = 1 - P(A)$$

  - **Law of Partitions**: If $B_1, \ldots, B_n$ are exhaustive and mutually exclusive events then

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

  - **Law of Total Probability**: If $B_1, \ldots, B_n$ are exhaustive and mutually exclusive then

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

  - Calculate **conditional probabilities** using **probability rules**:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

  - State and use the **general multiplication rule** to determine probabilities of intersections.
    - **General Multiplication Rule** for Two Events:

$$P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$$

    - **General Multiplication Rule** for Three Events:

$$P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid A \cap B) = P(B)P(A \mid B)P(C \mid A \cap B) = P(C)P(B \mid C)P(A \mid B \cap C) = .$$

- **Independence**
  - Two events, A and B, are independent if the occurrence of one does not affect the probability of the other: $P(A \mid B) = P(A), \;\; (B \mid A) = P(B)$
  - Special multiplication rule for independent events (Only use if you know for a fact they are independent.)

$$P(A \cap B) = P(A) \times P(B)$$

- **Bayes' Rule**:
- Baye's Rule for 2 Events:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^{'})P(A^{'})}$$

- General Baye's Rule for n Events: If $A_1, \ldots, A_n$ are exhaustive and mutually exclusive events

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)}$$

Chapter 5: Discrete Random Variables

- Recognize the properties of a **valid probability distribution** for discrete variables:
  - Each probability

$$p_X(x)$$

    satisfies

$$0 \leq p_X(x) \leq 1$$

    .
  - The sum of all probabilities $\sum p_X(x) = 1$.
- Calculate probabilities using a **probability mass function (pmf)**.
- Calculate the **mean of a discrete random variable (Expected value)**:

$$\mathrm{E}(X) = \mu_X = \sum x \cdot p(x)$$

- Calculate the **variance and standard deviation for a discrete random variable**:
  - Variance:

$$\mathrm{Var}(X) = \sigma_X^2 = \mathrm{E}[(X - \mu_X)^2] = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

  - Standard deviation:

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

Rules for Expected Value and Variance

- **LOTUS** For any real valued function $g(\cdot)$ and discrete random variable X

$$\mathrm{E}[g(X)] = \sum_x g(x) p_X(x)$$

- **Linearity of Expectation**: For any two random variables X and Y, and constants a and b,

$$\mathrm{E}(aX \pm bY) = a\mathrm{E}(X) \pm b\mathrm{E}(Y)$$

- **Variance of a Linear function**: For any random variable X and constants $a \neq 0$ and b,

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

  This shows that adding a constant b to a random variable does not change its variance, while multiplying by a scales the variance by a^2.

- **Variance of the Sum/Difference of Two Independent Random Variables**: If X and Y are independent,

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Named Distributions

- For a **Binomial distribution**, understand when it applies (BInS criteria) and how to calculate probabilities, expected values, and variances:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\mathrm{E}(X) = np, \quad \sigma_X = \sqrt{np(1-p)}$$

- For a **Poisson distribution**, recognize when it applies and how to calculate probabilities, expected values, and variances:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$\mathrm{E}(X) = \lambda, \quad \sigma_X = \sqrt{\lambda}$$

Chapter 6: Continuous Random Variables and Probability Distributions

- Determine if a function is a **legitimate density function** and calculate the **normalization constant** if necessary.
  - **Legitimate Density Functions**: A function $f(x)$ is a legitimate density function if it satisfies two conditions:
    1. $f(x) \geq 0$ for all x.
    2. The total area under the curve of $f(x)$ over its entire range equals 1, i.e., $\int_{-\infty}^{\infty} f(x)dx = 1$.
  - **Normalization Constant**: The constant required to ensure the total area under the probability density function (pdf) equals 1.
- Calculate **probabilities** for a continuous random variable using the density function:

$$P(a < X < b) = \int_a^b f(x)dx$$

- Calculate and use the **cumulative distribution function (CDF)**:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- **Percentiles and Median**:
  - **Percentile**: Solve $F(y) = p$ to find the yth percentile.
  - **Median**:

$$\int_{-\infty}^{\tilde{\mu}} f(x)dx = 0.5$$

  - **Mean (Expected Value)**:

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Named Distributions

(Note: The distributions have been written in short hand notation. You need to realize where the pdf/cdf is 0 and where the cdf is 1.)

- **Normal Distribution**:
  - Use the z-table for calculating probabilities and percentiles.
  - Normal probability plots help determine if data follow a normal distribution. Deviations suggest skewness or a non-normal distribution.
- For a **Uniform Distribution**, understand when it applies and how to calculate probabilities, percentiles, expected values, and variances:
  - Probability Density Function (pdf):

$$f(x) = \frac{1}{b-a}, \quad \text{for } a \leq x < b$$

  - CDF:

$$F(x) = \frac{x-a}{b-a}, \quad \text{for } a \leq x < b$$

  - Mean and Standard Deviation:

$$E(X) = \frac{a+b}{2}, \quad \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

- For a **Exponential Distribution**, understand when it applies and how to calculate probabilities, percentiles, expected values, and variances:
  - pdf:

$$f(x) = \lambda e^{-\lambda x}, \quad \text{for } x \geq 0$$

  - CDF:

$$F(x) = 1 - e^{-\lambda x}, \quad \text{for } x \geq 0$$

- Mean and Standard Deviation:

$$E(X) = \frac{1}{\lambda}, \quad \sigma = \frac{1}{\lambda}$$

Exam 2 Chapters 7-11

Chapter 7

- **Understanding Parameters and Statistics**: Accurately define what constitutes a parameter in the context of a population and a statistic in the context of a sample.
- **Identifying and Appreciating Sampling Distributions**:
  - Correctly identify scenarios that involve sampling distributions.
  - Explain the significance of sampling distributions in statistical inference.
  - Correctly identify the mean (μ) and standard error ($\sigma/\sqrt{n}$) of the sampling distribution of the sample mean ($\bar{X}$).
  - Compute probabilities for ranges of outcomes ($\bar{x}$) and percentiles using when given a sampling distribution, assuming the population from which samples are drawn is normally distributed.
- **Application of the Central Limit Theorem (CLT)**:
  - Assess and determine the conditions under which the distribution of sample averages from an initially unknown or non-normal population distribution can be accurately described by a normal distribution. This includes recognizing the importance of sample size and the role of the CLT in justifying the normal approximation for the distribution of sample means.
  - Clearly articulate situations when the CLT would not apply.
  - Be able to calculate probabilities and percentiles using this information.

Chapter 8

- **Understanding Observational Studies vs. Experiments**:
  - Clearly differentiate between observational studies and experiments, recognizing the unique insights each can provide.
  - Articulate the reasons behind classifying a study as observational or experimental.
  - Evaluate and identify instances of anecdotal evidence within the context of scientific research.
- **Identifying Components of Experiments**: Accurately identify the experimental units, explanatory variables, treatments or factors, levels, and response variables in various research scenarios.
- **Understanding Experimental Design Graphs**:
  - Master the ability to both draw and interpret experimental design graphs for random experiments, completely randomized, matched pairs, and block designs.

    - Justify the use of matched pair or block designs over completely randomized, understanding the specific conditions that make them preferable.
- **Evaluating Experimental Designs**: Recognize the critical factors in designing an experiment, including control, the principle of control, randomization, and replication, and assess whether an experimental design can be considered good" and the utility of single blind, and double blind experiments and when blinding is not possible.
- **Recognizing Sampling Methods**: Classify a study's sampling method and why one method may be preffered given context.
    - **Non-randomized Methods**: issues with non-random sampling and the different types.
    - **Randomized Methods**: such as simple random sampling and stratified sampling.
- **Identifying Sampling Issues**: Identify common issues in sampling such as bias, convenience sampling, self-selection, undercoverage, and nonresponse, and understand their impacts on study results.
- **Defining and Identifying Lurking Variables**:
    - Define what lurking variables are and identify potential lurking variables in given research contexts.
    - Distinguishing Between Lurking Variables: Differentiate whether a lurking variable acts as a confounding variable or is part of a common response in specific scenarios, and competently draw and interpret the corresponding diagrams.
- **Assessing Causality**: Evaluate the possibility of causality in various contexts, utilizing both statistical data and theoretical knowledge. This includes understanding when statistical inference can suggest a causal relationship and recognizing the limitations of such inferences.

Chapter 9, 10, and 11

Confidence Intervals/Bounds

- **Understanding Confidence Intervals and Confidence Bounds**:
    - Accurately define what constitutes a confidence interval and a confidence bound, including their purpose in statistical analysis.
    - Distinguish between point estimates and interval estimates in the context of statistical inference.
- **Interpreting Confidence Interval Results**:
    - Develop the skill to correctly interpret the results of a confidence interval, understanding what the interval range implies about the potential true value of the parameter being estimated.
    - Recognize the implications of different confidence levels (e.g., 95%, 99%) and how they affect the interpretation of confidence intervals.
- **Confidence Level and Random Sampling**:
    - Explain the relationship between the chosen confidence level and the principle of random sampling and the sampling distribution in the construction of confidence intervals.
    - Understand how the confidence level reflects the proportion of confidence intervals, from repeated samples, that are expected to contain the true parameter value.
- **Determining Factors That Control the Width of a Confidence Interval**:

    - Identify the factors that influence the width of a confidence interval, such as sample size, variability of the data, and the chosen confidence level.
    - Analyze the trade-off between the width of the confidence interval and the level of confidence.
- **Assumptions Required for Statistical Inference**:
    - List and understand the assumptions necessary for performing statistical inference, including normality, independence, and sample size.
    - Evaluate whether these assumptions are met in practical scenarios and understand the implications if they are not met.
- **Application of Confidence Intervals**:
    - Be proficient in interpreting computer output related to confidence intervals and know when and how to compute confidence intervals manually using the correct critical values.
    - Apply knowledge of confidence intervals, demonstrating an understanding of their practical importance in statistical inference, using provided graphs and output to support the analysis.
    - Correctly identify which computer output to use for constructing confidence intervals and understand how to apply the central limit theorem in calculating these intervals.
    - You may also need to compute the interval by hand using the correct critical value.
- **Precision of Confidence Intervals**:
    - Be able to calculate the sample size, n, needed for a particular level of precision (margin of error) for the one sample situations and determine the final answer. **ME** is the margin of error (half-width).
    - Recognize that preliminary studies or approximations may be necessary in the case where the population standard deviation is unknown.

The table below summarizes the formulas for one-sample z, one-sample t, two-sample t (independent), and two-sample t (matched pair) tests.

Summary of Confidence Interval Formulas

One-Sample Z	One-Sample t	2-Sample t (Independent)	2-Sample t (Matched Pair)
$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$	$\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2}, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\bar{d} \pm t_{\frac{\alpha}{2}, n-1} \frac{s_d}{\sqrt{n}}$

Summary of Sample Size Calculations for Confidence Intervals

One-Sample Z	One-Sample t
$n = \left(\frac{\sigma z_{\alpha/2}}{ME} \right)^2$	$n = \left(\frac{s' t_{\alpha/2, n'-1}}{ME} \right)^2$

Summary of Confidence Bound Formulas

Test Type	Lower Bound	Upper Bound
One-Sample Z	$\mu > \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}$	$\mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$
One-Sample t	$\mu > \bar{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}}$	$\mu < \bar{x} + t_{\alpha,n-1} \frac{s}{\sqrt{n}}$
2-Sample t (Independent)	$\mu > \bar{x}_1 - \bar{x}_2 - t_{\alpha,v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\mu < \bar{x}_1 - \bar{x}_2 + t_{\alpha,v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
2-Sample t (Matched Pair)	$\mu > \bar{d} - t_{\alpha,n-1} \frac{s_d}{\sqrt{n}}$	$\mu < \bar{d} + t_{\alpha,n-1} \frac{s_d}{\sqrt{n}}$

Unknwon but equal Variance

- **Pooled Variance Estimator**:
  - Be able to show that the pooled estimator is unbiased if the equal variance assumption is true.
  - Discuss the ramifications of assuming equal variances incorrectly.
  - Explore how bias in the pooled variance estimator is exacerbated when sample sizes are imbalanced, elucidating the mechanism behind this phenomenon.

Pooled Variance Estimator

$$S_p^2 = \left[\frac{n_\text{A} - 1}{n_\text{A} + n_\text{B} - 2}\right] S_\text{A}^2 + \left[\frac{n_\text{B} - 1}{n_\text{A} + n_\text{B} - 2}\right] S_\text{B}^2$$

Degrees of Freedom when Equal Variance is Assumed:

$$\textbf{df} = n_\text{A} + n_\text{B} - 2$$

Unknown and Unequal Variance and the Welch Satterthwaite Approximate Degrees of Freedom

- **Welch–Satterthwaite Approximation**:
  - Recognize the implications of incorrectly assuming equal variances between two populations.
  - Understand how the Welch t-procedure and the Satterthwaite approximation compares to the use of the pooled variance approach used in a standard two-sample t-test with the assumption of equal variances, and know when each method is appropriate.
  - Understand that the Satterthwaite Approximation is used to approximate the degrees of freedom in the two-sample t-test when the two populations do not have equal variances and the sample sizes might be unequal.
  - Answer questions regarding what happens to the formula if the sample sizes are equal, the variances are equal, or both are equal.
  - Know the limitations of the approximation, such as precision of confidence intervals and potential loss of power compared to when population variances are equal, or when alternative nonparametric tests might be more appropriate.
  - Evaluate the implications of changes in the formula for degrees of freedom on the interpretation inference results, understanding how these changes impact hypothesis

testing and statistical conclusions.

Exact Degrees of Freedom:

$$\mathbf{df} = \frac{\left(\frac{\sigma_{\mathrm{A}}^2}{n_{\mathrm{A}}} + \frac{\sigma_{\mathrm{B}}^2}{n_{\mathrm{B}}}\right)^2}{\frac{1}{n_{\mathrm{A}}-1}\left(\frac{\sigma_{\mathrm{A}}^2}{n_{\mathrm{A}}}\right)^2 + \frac{1}{n_{\mathrm{B}}-1}\left(\frac{\sigma_{\mathrm{B}}^2}{n_{\mathrm{B}}}\right)^2}$$

Welch Satterthwaite Approximate Degrees of Freedom:

$$v = \frac{\left(\frac{s_{\mathrm{A}}^2}{n_{\mathrm{A}}} + \frac{s_{\mathrm{B}}^2}{n_{\mathrm{B}}}\right)^2}{\frac{1}{n_{\mathrm{A}}-1}\left(\frac{s_{\mathrm{A}}^2}{n_{\mathrm{A}}}\right)^2 + \frac{1}{n_{\mathrm{B}}-1}\left(\frac{s_{\mathrm{B}}^2}{n_{\mathrm{B}}}\right)^2}$$

Hypothesis Testing, Error, and Power

- **Type I and Type II Errors, and Power**:
  - Comprehend the implications of Type I error (false positive rate, denoted as α), Type II error (false negative rate, denoted as β), and statistical power ($1-\beta$) in hypothesis testing.
  - Perform calculations related to Type II error and power for a specified alternative mean μ_a, given sample size n, significance level α, and null value μ_0, when the alternative hypothesis is clearly defined.
- **Sample Size, Type II Error, and Power**: Understand the direct relationship between sample size (n) and its impact on Type II error and power, including how to calculate n for a specified level of power ($1-\beta$) or Type II error (β) for a given alternative mean μ_a, when α and β or $1-\beta$ are provided.
- **Test Statistic and P-value**:
  - Demonstrate clear understanding of test statistics and p-values, including how they are obtained from hypothesis testing, and correct versus incorrect interpretations.
  - Four-Step Hypothesis Testing: Be able to systematically conduct hypothesis tests in four steps, covering parameter identification, hypothesis formulation, calculation of test statistic and p-value, and conclusion writing.

Test Statistic Formulas:

Test Type	Test Statistic Formula	Degrees of Freedom
One-Sample Z	$z_{ts} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	N/A
One-Sample t	$t_{ts} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$n-1$
2-Sample t (Independent)	$t'_{ts} = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	v
2-Sample t (Matched Pair)	$\frac{\bar{d} - \Delta_0}{s_D/\sqrt{n}}$	$n-1$

- **Hypothesis Test and Confidence Intervals/Bounds**:
  - Be able to compare the results of a confidence interval or bound with the associated hypothesis test. This includes how they are similar and how they are different.
  - Be able to determine which confidence interval or bound corresponds to which alternative hypothesis.
  - Given the results of the confidence interval/bound or hypothesis test, be able to predict the results of the other one and when this approach is valid ($C + \alpha = 1$).

Hypotheses Relationships and Code:

Hypothesis Test	Null Hypothesis	Alternative Hypothesis	Confidence Interval or Bound
Upper Tailed	$H_0 : \mu \leq \mu_0$	$H_a : \mu > \mu_0$	Lower Bound
Lower Tailed	$H_0 : \mu \geq \mu_0$	$H_a : \mu < \mu_0$	Upper Bound
Two-Tailed	$H_0 : \mu = \mu_0$	$H_a : \mu \neq \mu_0$	Interval

Hypothesis Test	p-value (z-code)	p-value (t-code)
Upper Tailed	$P(Z > z_{ts})$: **pnorm**(z_{ts}, lower.tail = **FALSE**)	$P(T > t_{ts})$: **pt**(t_{ts}, **df** = $n-1$, lower.tail = **FALSE**)
Lower Tailed	$P(Z < z_{ts})$: **pnorm**(z_{ts}, lower.tail = **TRUE**)	$P(T < t_{ts})$: **pt**(t_{ts}, **df** = $n-1$, lower.tail = **TRUE**)
Two-Tailed	$2P(Z > \lvert z_{ts}\rvert)$: 2**pnorm**(abs(z_{ts}), lower.tail = **FALSE**)	$2P(T < \lvert t_{ts}\rvert)$: 2**pt**(abs(t_{ts}), lower.tail = **FALSE**)

Statistical Significance vs. Practical Significance:

- **Understanding Statistical Significance and Practical Significance**: State and differentiate between statistical significance and practical significance, including how to determine and interpret the practical significance of inference results. Understand that practical difference and effect size are critical in evaluating the real-world importance of statistical findings.
- **Determining Practical Significance**: Given sufficient information, discern and interpret the practical significance of inference results.
- **Interpretation based on Hypothesis Test Outcome**:
  - If the decision is to fail to reject H_0, acknowledge the absence of practical significance.
  - If the decision is to reject H_0, assess the appropriate difference and effect size.
- **Difference Calculation**: Compute the difference as the absolute value of the difference between the null value and the experimental value (confidence bound or the closest limit of the confidence interval).
- **Effect Size Computation**: Calculate the effect size as the difference divided by the standard deviation.Only for one-sample, two-sample paired, or two-sample independent if equal variance is reasonable.

$$\frac{\text{difference}}{\text{standard deviation}}$$

- **Evaluating Difference and Effect Size**:
  - For the difference:
    - If the difference is small, consider the values practically the same.
    - If the difference is large, deem the values practically different.
    - Stake holders deem what differences are large or important.
    - On exams, indicate which differences are practically equivalent.
- **For the effect size**:
  - If the effect size is small, regard the values as practically indistinguishable.
  - If the effect size is large, perceive the values as practically distinct.
  - Remember practical significance is primarily determined by stake holders.
  - In the absence of stakeholder guidance, adhere to rule-of-thumb thresholds: less than 0.2 (small), 0.2 – 0.8 (moderate), greater than 0.8 (large).
- **Final Interpretation**: Recognize that practical difference only exists when both the difference and effect size indicators are large. Therefore, if one indicator is large, evaluate the other to determine the final conclusion.

Post Exam 2 Material: Chapters 12-13

Chapter 12 - ANOVA

- **Background of ANOVA**
  - Explain what is meant by One-Way ANOVA.
  - Explain why we analyze the sources of variation to determine if the means are statistically different in the population.
- **Prediction ANOVA Conclusions from Boxplots and Effects Plots**
  - Be able to make an approximate decision of whether the population means are the same or different given a side-by-side boxplot or effects plot of the data. Compare the within group variance with the between group variance and consider the respective sample sizes.
- **ANOVA Setup**
  - Be able to identify the factor, the levels of the factor, the response variable, and relate these to the populations to be compared.
- **Assumptions for ANOVA**
  - State the assumptions that are required for ANOVA to be valid and state how you would check them. If the information is available, determine if the assumptions are valid.
    - **SRS** - assumed
    - **Normality**: normal probability plot, histogram
    - **Constant standard deviation**: the ratio of maximum and minimum standard deviations.
- **One-way ANOVA Table and Hypothesis**
  - Be able to calculate or complete a One-Way ANOVA table by hand.
  - $\mathbf{SS_T = SS_A + SS_E}$
  - $\mathbf{df_T = df_A + df_E}$
  - $\mathbf{MS = \frac{SS}{df}}$

- $\mathbf{F_{TS}} = \dfrac{\mathbf{MS_A}}{\mathbf{MS_E}}$ has numerator and denominator degrees of freedom $\mathbf{df}_1 = \mathbf{df_A}$ and $\mathbf{df}_2 = \mathbf{df_E}$.

Source	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F-Statistic
Factor A (Between Groups)	$\mathbf{df_A} = k - 1$	$\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2$	$\frac{\mathbf{SS_A}}{\mathbf{df_A}}$	$\frac{\mathbf{MS_A}}{\mathbf{MS_E}}$
Error (Within Groups)	$\mathbf{df_E} = n - k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$	$\frac{\mathbf{SS_E}}{\mathbf{df_E}}$	
Total	$\mathbf{df_T} = n - 1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})^2$		

- Recognize the code for calculating ANOVA problems when the data is known using function **aov()** and the fact that you need a **summary()** function after the ANOVA calculation.
  - Perform the four-step hypothesis test for one-way ANOVA. The basic steps are similar, but not identical to the one-sample and two-sample inferences. Specifically, the alternative hypothesis in words and there are two degrees of freedom for the test statistic.
- Recognize the code to calculate the p-value.

```
pf(fts,df1,df2,lower.tail = FALSE)
```

Compare t-procedure versus F-procedure when $k = 2$ groups.

- Mathematically show that the two procedures are equivalent if $k = 2$, pooled variance approach is used, and $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$.
- Explain the advantages of using the two independent sample t-procedure over ANOVA when $k = 2$.

Multiple Comparison Procedures

- Explain what is meant by Family Wise Error Rate (FWER) and why it is important.
- Explain the differences between the different Multiple Testing Procedures: Sidak, Bonferonni, TukeyHSD, and Dunnett.
- Perform Tukey's multiple comparison analysis to determine which factors are different and construct the corresponding confidence intervals by hand given the Tukey parameter $Q_{\alpha,k,n-k}$ from code.

```
Q <- qtukey(p=C, nmeans= k, df = n - k, lower.tail = TRUE)
Q <- qtukey(p=alpha, nmeans = k, df = n - k, lower.tail = FALSE)
```

$$\bar{x}_{i.} - \bar{x}_{j.} \pm \frac{Q_{\alpha,k,n-k}}{\sqrt{2}} \sqrt{\mathbf{MS_E}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- Recognize the code and interpret the output of Tukey HSD when data is used to fit the ANOVA model.

```
fit<- aov(quantitativeVariable ~ categoricalVariable, data = dataframe)
summary(fit)
TukeyHSD(fit, ordered = TRUE, conf.level = C)
```

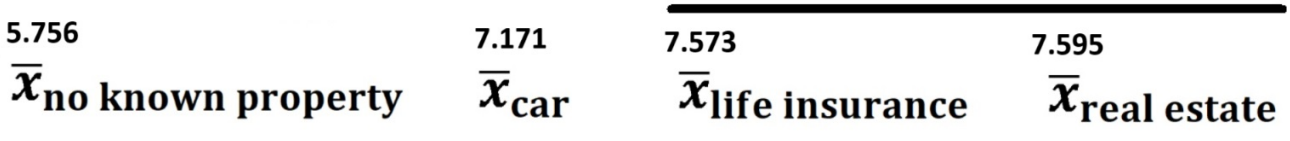
- Create a visual display by hand of the results of the TukeyHSD and explain the results in 'laymans terms'. Below is an example from CA8 Spring 2024 which includes statistical summaries, ANOVA output, TukeyHSD output, and visual diagram.

Table 1: Summary Statistics by Group

propertyMagnitude	SampleSize	Mean	StandardDeviation
car	224	7.171420	0.9155398
life_insurance	226	7.572687	0.8394031
no_known_property	221	5.756054	0.6793885
real_estate	223	7.594713	0.7950613

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## propertyMagnitude   3  500.8  166.94     253 <2e-16 ***
## Residuals         890  587.3    0.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = lamount ~ propertyMagnitude, data = credit_data_clean)
##
## $propertyMagnitude
##                                         diff        lwr        upr     p adj
## life_insurance-car                 0.40126712  0.1621434  0.6403908 0.0000012
## no_known_property-car             -1.41536579 -1.6558322 -1.1748994 0.0000000
## real_estate-car                    0.42329263  0.1833696  0.6632157 0.0000003
## no_known_property-life_insurance -1.81663291 -2.0565703 -1.5766955 0.0000000
## real_estate-life_insurance         0.02202551 -0.2173673  0.2614183 0.9917392
## real_estate-no_known_property      1.83865842  1.5979244  2.0793925 0.0000000
```



Chapter 13 - Linear Regression

- **Identifying Variables**
- Understand how to determine which variables should be treated as the explanatory variable and which should be treated as the response variable for a given situation.
- **Interpreting Scatterplots**
- Be skilled at interpreting a scatterplot in terms of pattern, direction, strength, check for validitiy of constant variance assumption, and aid in the identification of outliers, including determining if an outlier is influential or not.

The Linear Regression Model and Calculations

- Familiarize yourself with the population model for linear regression $Y = \beta_0 + \beta_1 x + \epsilon$ and clearly define all terms in the model.
- Clearly identify which parts of the population model represent the average tendencies and which parts of the model are to be considered random variables.
- Accurately state the assumptions of the simple linear regression model.
    - **(i.i.d.)** The observed pairs (x_i, y_i) for $i \in \{\mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n}\}$ are such that the y_i are considered a simple random sample (SRS) for each fixed value of $x_{\boldsymbol{i}}$.
    - **(Linearity)** The association between the explanatory variable and the response is on average linear.
    - **(Homoscedasticity)** The error terms have common variance.
    - **(Normality)** The error terms are normally distributed.
- **Check Assumptions**: Given the appropriate graphs, determine if the assumptions are met (you will need to determine which graphs are appropriate for which assumptions). You must mention ALL graphs that can be used to check each assumption.
- Describe the process in obtaining the least squares regression estimates. This could involve deriving the formulas for estimating the parameters in a simplified single parameter model for example; a case where the intercept is a known constant.
- Given summation terms know how to perform the intermediate calculations for the least squares estimates, the estimate of the common standard deviation, the sample Pearson correlation coefficient, the coefficient of determination, and the ANOVA table. Such intermediate calculations require computation of the following terms from given summations:
- $\mathbf{S_{XX}}$: The sum of squared deviations of X values from their mean.

$$\mathbf{S_{XX}} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

- $\mathbf{S_{YY}}$: The sum of squared deviations of Y values from their mean.This is also $\mathbf{SS_T}$.

$$\mathbf{S_{YY}} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

- $\mathbf{S_{XY}}$: The sum of the product of the deviations of X and Y from their respective means.

$$\mathbf{S_{XY}} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

- **Least Squares Regression Line** With these summations, we can calculate the coefficients of the least squares regression line, which are:
- **Slope** ($\hat{\beta}_1$ or b_1):

$$\hat{\beta}_1 = b_1 = \frac{\mathbf{S_{XY}}}{\mathbf{S_{XX}}}$$

- **Intercept** ($\hat{\beta}_0$ or b_0):

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}$$

- **Complete the ANOVA Table**:
  - Be able to calculate or complete a One-Way ANOVA table by hand. Useful formulas for completing the ANOVA table by hand:
    - $\mathbf{SS_T} = \mathbf{S_{YY}} = \mathbf{SS_R} + \mathbf{SS_E}$
    - $\mathbf{SS_R} = b_1\mathbf{S_{XY}}$
    - $\mathbf{df_T} = \mathbf{df_R} + \mathbf{df_E}$
    - $\mathbf{F_{TS}} = \frac{\mathbf{MS_R}}{\mathbf{MS_E}}$ has numerator and denominator degrees of freedom $\mathbf{df}_1 = \mathbf{df_R} = 1$ and $\mathbf{df}_2 = \mathbf{df_E} = n - 2$.

Source	df	SS	MS	F-Statistic
Regression	$\mathbf{df_R} = 1$	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\frac{\mathbf{SS_R}}{\mathbf{df_R}}$	$\frac{\mathbf{MS_R}}{\mathbf{MS_E}}$
Error	$\mathbf{df_E} = n - 2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\frac{\mathbf{SS_E}}{\mathbf{df_E}}$	
Total	$\mathbf{df_T} = n - 1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

- **Common Standard Deviation** ($\hat{\sigma}$ or s):

$$\hat{\sigma} = s = \sqrt{\mathbf{MS_E}}$$

- **Coefficient of Determination** (R^2): What fraction of the variation in the response is explained by the least-squares regression of y on x.

$$R^2 = \frac{\mathbf{SS_R}}{\mathbf{SS_T}}$$

- **Sample Pearson Correlation** (r): The Pearson sample correlation coefficient is a statistical measure of the strength and direction of a linear relationship between two quantitative variables.

$$r = \frac{\mathbf{S_{XY}}}{\sqrt{\mathbf{S_{XX}S_{YY}}}} = \text{sign}(b_1)\sqrt{R^2}$$

Interpretations

- Interpret R^2 including what it doesn't tell you. A high R^2 value tells us that a large proportion of the variation in the response variable is explained by the assumed linear relationship with the explanatory variable. A high R^2 value can be an indication of a good linear fit; it is not the only factor to consider when evaluating the predictive power.
    - **(Nonlinearity)** The true association may be nonlinear.
    - **(Outliers Non-Robust Measure)** The presence of outliers can have significant impact on the $\boldsymbol{R}$**^2** value.
    - **(Unexplained Variance)** The unexplained variance may still be large do to other unobservable or lurking variables that were not accounted for in the model.
    - **(Data Quality)** The presence of measurement errors that have not been accounted for in the model.
    - **(Small Data)** Small sample sizes may suggest a strong linear association when the population association is highly non-linear.
    - **(Extrapolation)** Extrapolating predictions outside the range of the fitted model may produce poor performance.
    - **(Assumption Violations)** The assumption of homogeneity of variance may be violated and the MSE would not be a reliable estimate of the variability. (Under and overestimate the errors)
- Interpret the sample pearson Correlation r. Remember that correlation is calculated assuming that the relationship is linear; therefore, it doesn't provide information regarding the form. A scatterplot is also needed to understand the relationship between x and y.
    - What happens if we swap x and y?
    - What does the sign of r indicate about the linear association?
    - What does $r = 0$ imply?
- Define the y-intercept in context. Determine if it makes sense to interpret the y-intercept by itself in a particular situation. Note that this is not simply dependent on the value of the y-intercept.
- Define the slope in context and relate it to the average rate of change.

Inference for the Least Squares Regression Line

Hypothesis Tests

Perform the hypothesis test for association (model utility test, F-test) using the four-step approach. Here's how to approach it:

- **Step 1**: Can be skipped for this procedure.
- **Step 2**: State the hypotheses:

- H_0: There is no linear association between x and Y.
- H_a: There is a linear association between x and Y.

- **Step 3**: Calculate the test statistic: State and determine the F-test statistic from the ANOVA table, the degrees of freedom, and p-value.
  - $\mathbf{F_{TS}} = \frac{\mathbf{MS_A}}{\mathbf{MS_E}}$
  - Numerator and Denominator degrees of freedom $\mathbf{df}_1 = \mathbf{df_A}$ and $\mathbf{df}_2 = \mathbf{df_E}$.
  - Determine the p-value from either the ANOVA table or from output associated with the appropriate R-code.
- **Step 4**: Decision and formal conclusion
  - Compare p-value and significance level.
  - Conclusion: Since the alternative hypothesis is that there is a linear association, this will be the data does show a linear association or the data does not show a linear association.

Perform the hypothesis test with regards to the model parameters typically regarding the slope (β_1):

- **Step 1**: This step is regarding the unknown true β the population slope of the mean response line $\mu_{Y|X=x}$.
- **Step 2**: State the hypotheses: (Null value β_{1_0} is typically 0)
  - H_0: Options –> $\beta_1 = \beta_{1_0}$, $\beta_1 \leq \beta_{1_0}$, or $\beta_1 \geq \beta_{1_0}$
  - H_a: Options –> $\beta_1 \neq \beta_{1_0}$, $\beta_1 > \beta_{1_0}$, or $\beta_1 < \beta_{1_0}$
- **Step 3**: Calculate the test statistic: State and determine the t-test statistic from the regression output or calculate it by hand, the degrees of freedom, and p-value.
  - $\mathbf{t_{TS}} = \frac{\beta_1 - \beta_{1_0}}{\sqrt{\frac{\mathbf{MS_E}}{\mathbf{S_{XX}}}}}$
  - Degrees of freedom $\mathbf{df_E} = n - 2$.
  - Determine the p-value from either the regression output or from the appropriate R-code.
- **Step 4**: Decision and formal conclusion
  - Compare p-value and significance level.
  - Conclusion: Standard conclusion template with respect to the slope.

Confidence Intervals

- **Confidence Interval for Slope**
  - Calculate the confidence interval for the slope manually using the appropriate critical value from provided R-code and output.

$$b_1 \pm t_{\alpha/2,\mathbf{df_E}} \cdot \mathbf{SE}_{\beta_1} = b_1 \pm t_{\alpha/2,\mathbf{df_E}} \cdot \sqrt{\frac{\mathbf{MS_E}}{\mathbf{S_{XX}}}}$$

  - Select the appropriate output for the confidence interval from code.

Using the Least Squares Regression Line for Prediction

- Write down the least squares line from computer output or from hand calculations ($\hat{y} = b_0 + b_1 x$).
- Make predictions using the regression line at a point x^*.

- Determine and explain when you cannot use your least squares line for prediction because of extrapolation or other reasons.
- Compute the residual $e = y - \hat{y}$ at a point using the regression line $\hat{y}$ and observation (x, y).
- Calculate and interpret the confidence interval for the mean response at $x = x^*$.

$$\hat{\mu}_{x^*} \pm t_{\alpha/2,\,\mathbf{df_E}}\mathbf{SE}_{\hat{\mu}_{x^*}}$$

  - Estimate the mean response value $\hat{\mu}_{x^*} = b_0 + b_1 x^*$.
  - Obtain the standard error for the predicted mean response at $x = x^*$ as

$$\mathbf{SE}_{\hat{\mu}_{x^*}} = \sqrt{\mathbf{MS_E}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\mathbf{S_{XX}}}\right]}$$

  - Identify the correct t-critical value from the provided R code and output using degrees of freedom $\mathbf{df_E} = n - 2$.
- Calculate and interpret the confidence interval for the response at a new point $x = x^*$.

$$\hat{Y}^* \pm t_{\alpha/2,\,\mathbf{df_E}}\mathbf{SE}_{\hat{Y}^*}$$

  - Estimate the response value $\hat{Y}^* = b_0 + b_1 x^*$.
  - Obtain the standard error for the predicted response at the new point $x = x^*$ as

$$\mathbf{SE}_{\hat{Y}^*} = \sqrt{\mathbf{MS_E}\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\mathbf{S_{XX}}}\right]}$$

  - Identify the correct t-critical value from the provided R code and output using degrees of freedom $\mathbf{df_E} = n - 2$.
- Be able to state the difference between the **confidence interval** for the **mean response** at $x = x*$ and the **prediction interval** for a particular value at $x = x*$ and when each interval would be used.