Review

R1. Key concepts, formula and applications

Chapter 1: An Introduction to Statistics

1.1 Populations and Samples

Key Concepts:

- **Data**: Facts and figures used to draw conclusions. A **Data Set** is the collection for a specific study.
- **Variables**: Characteristics of elements, categorized as:
  - **Quantitative**: Numerical measurements (e.g., salary, age).
  - **Qualitative**: Categorical measurements, including **Nominative** (unranked, e.g., gender) and **Ordinal** (ranked, e.g., risk level).
- **Population**: All units of interest.
- **Sample**: A selected subset of the population, used when a **Census** (examining the entire population) is impractical.
- **Descriptive Statistics**: The science of describing key aspects of a data set.

1.2 Selecting a Random Sample

Key Methods:

- **Random Sample**: A sample where every population unit has an equal chance of selection.
- **Sampling Without Replacement**: The standard method where a selected unit is not returned to the population.
- **Sampling With Replacement**: A unit is returned to the population for possible reselection.
- **Systematic Sampling**: An approximate method involving the selection of every k-th unit.

Application Scenarios:

- Estimating a company's employee cell phone costs (The Cell Phone Case).
- Assessing consumer reaction to a new product design (The Marketing Research Case).
- Verifying a vehicle's fuel economy for a tax credit (The Car Mileage Case).

Chapter 2: Descriptive Statistics: Tabular and Graphical Methods

2.1 Graphically Summarizing Qualitative Data

Basic Concepts:
This section deals with summarizing non-numerical data. A **Frequency Distribution** is a table showing the count of items in non-overlapping categories. The **Relative Frequency** of a class is the proportion (frequency/total observations), often expressed as a percentage.

Key Visualizations:

- **Bar Chart:** Uses rectangles (vertical or horizontal) to represent the frequency, relative frequency, or percent frequency of each category.
- **Pie Chart:** A circle divided into slices, where each slice's size corresponds to the relative frequency or percent frequency of a category.

Application Scenario:
Used to visualize and compare categorical data, such as summarizing customer pizza preferences or analyzing purchasing patterns for different Jeep models.

2.2 Graphically Summarizing Quantitative Data

Basic Concepts:
This section focuses on summarizing numerical data by grouping it into classes to show its distribution. The main tool is the **Frequency Distribution** table, which is then visualized as a **Histogram**—a graph using rectangles to represent class frequencies.

Basic Formula:

- **Sturges' Rule (for number of classes K):** K is the smallest integer such that 2^K ≥ n (where n is the number of data points).
- **Class Length (L):** L ≈ (Largest Measurement - Smallest Measurement) / K

Application Scenario:
Used to analyze the shape of data distribution, such as examining payment times in days for an e-billing system to see if most payments are processed quickly (potentially right-skewed).

2.3 Dot Plots

Basic Concepts:
A simple graphical display where each data point is represented by a dot on a number line. It shows the individual values and the overall distribution pattern, including clusters and gaps.

Application Scenario:
Ideal for small datasets to quickly visualize the spread and concentration of data points, such as plotting the exam scores of a small class.

2.4 Stem-and-Leaf Displays

Basic Concepts:
This method separates each data value into a "stem" (leading digits) and a "leaf" (trailing digit). It provides a detailed view of the distribution while preserving the original data values, similar to a histogram turned on its side.

Application Scenario:
Used to analyze the distribution of moderately sized datasets. For example, displaying car mileage data to assess fuel economy and see if the distribution is symmetrical or skewed, which is crucial for meeting tax credit qualifications.

Chapter 3: Descriptive Statistics: Numerical Methods

3.1 Describing Central Tendency

Basic Concepts:
This section introduces measures that represent the center or middle of a dataset. A **population parameter** describes the population, while a **sample statistic** (a **point estimate**) describes a sample and estimates the population parameter. The three primary measures are:

- **Mean (μ or x̄):** The average value.
- **Median (Md):** The middle value in an ordered dataset.
- **Mode (Mo):** The most frequently occurring value.

Basic Formulas:

- Sample Mean: $\bar{x} = (\Sigma x_i) / n$
- Population Mean: $\mu = (\Sigma x_i) / N$

Application Scenario:
Used to find a typical value. For example, calculating the mean, median, and mode of invoice payment times to understand the central processing time and identify the most common (`mode`) duration.

3.2 Measures of Variation

Basic Concepts:
This section covers metrics that describe the spread or dispersion of data points around the mean.

- **Range:** The difference between the largest and smallest values.
- **Variance (σ^2 or s^2):** The average of the squared deviations from the mean.
- **Standard Deviation (σ or s):** The square root of the variance, in the original data units.
- **Empirical Rule:** For normal distributions, about 68%, 95%, and 99.7% of data fall within 1, 2, and 3 standard deviations of the mean, respectively.
- **Coefficient of Variation:** Compares relative variability between datasets with different units or means.

Basic Formulas:

- Sample Variance: $s^2 = \Sigma(x_i - \bar{x})^2 / (n - 1)$
- Sample Standard Deviation: $s = \sqrt{s^2}$
- z-score: $z = (x - \mu) / \sigma$

Application Scenario:
Used to assess consistency and risk. For instance, analyzing the standard deviation of car mileages to create tolerance intervals and ensure a vehicle model qualifies for a tax credit.

3.3 Percentiles, Quartiles, and Box-and-Whiskers Displays

Basic Concepts:
This section focuses on describing the relative standing of data points and summarizing distribution shape.

- **pth Percentile:** A value where p% of data fall at or below it.
- **Quartiles:** Q1 (25th percentile), Q2 (Median, 50th percentile), Q3 (75th percentile).
- **Interquartile Range (IQR):** $IQR = Q3 - Q1$, the range of the middle 50% of the data.
- **Five-Number Summary:** Minimum, Q1, Median, Q3, Maximum.
- **Box-and-Whiskers Plot:** A graphical display of the five-number summary used to identify skewness and **outliers**.

Application Scenario:
Used to compare distributions and identify unusual values. For example, creating a boxplot for customer satisfaction ratings to visually assess the data spread, central value, and pinpoint any outlier responses.

Chapter 4: Probability

4.1 The Concept of Probability

Basic Concepts:
An **experiment** is any process with an uncertain outcome. The **sample space** is the set of all possible outcomes. **Probability** is a numerical measure between 0 and 1 of the chance that a specific outcome or event will occur, where 0 indicates impossibility and 1 indicates certainty.

Probability Assignment Methods:

- **Classical Method:** Used when all outcomes are equally likely (e.g., probability of getting a head in a fair coin toss is 0.5).
- **Relative Frequency Method:** Based on the long-run proportion of times an outcome occurs (e.g., estimating the probability a consumer prefers a specific brand by surveying 1,000 people).
- **Subjective Method:** Based on personal judgment, experience, or expertise.

4.2 Sample Spaces and Events

Basic Concepts:
An **event** is a subset of the sample space. The probability of an event is the sum of the probabilities of the sample space outcomes that belong to that event.

Basic Formula:

- Probability of an Event A: `P(A) = (Number of outcomes in A) / (Total number of outcomes in sample space)` (Applicable when outcomes are equally likely).

4.3 Some Elementary Probability Rules

Basic Concepts and Formulas:

- **Complement:** The complement of event A ($\bar{A}$) is the event that A does not occur. `P(Ā) = 1 - P(A)`.
- **Intersection (A ∩ B):** The event that both A and B occur.
- **Union (A ∪ B):** The event that either A or B or both occur.
- **Addition Rule:**
  - For **mutually exclusive** events (A and B cannot occur together): `P(A ∪ B) = P(A) + P(B)`.
  - For any two events: `P(A ∪ B) = P(A) + P(B) - P(A ∩ B)`.

Application Scenario:
Used to calculate complex probabilities from simpler ones, such as finding the probability of drawing a Jack or a Queen from a standard deck of cards.

4.4 Conditional Probability and Independence

Basic Concepts:
Conditional Probability, denoted `P(A|B)`, is the probability of event A *given* that event B has occurred. Two events A and B are **independent** if the occurrence of one does not affect the probability of the other.

Basic Formulas:

- Conditional Probability: `P(A|B) = P(A ∩ B) / P(B)`.
- **Multiplication Rule (General):** `P(A ∩ B) = P(A) P(B|A) = P(B) P(A|B)`.
- **Multiplication Rule (Independent Events):** If A and B are independent, `P(A ∩ B) = P(A) P(B)`.

Application Scenario:
Used to analyze relationships between events, such as investigating potential gender bias in management promotions by comparing the conditional probability `P(Promotion | Male)` to `P(Promotion | Female)`.

Chapter 5: Discrete Random Variables

5.1 Two Types of Random Variables

Basic Concepts:
A **random variable** is a variable that assumes numerical values determined by the outcome of a random experiment. It quantifies uncertain outcomes. There are two main types:

- **Discrete Random Variable:** Its possible values can be counted or listed (e.g., the number of customers making a purchase, the number of heads in coin tosses).
- **Continuous Random Variable:** It can assume any numerical value in one or more intervals (e.g., waiting time, distance).

5.2 Discrete Probability Distributions

Basic Concepts:
The **probability distribution** of a discrete random variable X lists each possible value x and its associated probability $p(x)$. It can be represented in a table, graph, or formula.

Properties:

1. $0 \le p(x) \le 1$ for any value x
2. $\sum p(x) = 1$

Key Formulas:

- **Expected Value (Mean):** $\mu_X = E(X) = \sum[x \cdot p(x)]$
- **Variance:** $\sigma_X^2 = \sum[(x - \mu_X)^2 \cdot p(x)]$
- **Standard Deviation:** $\sigma_X = \sqrt{\sigma_X^2}$

Application Scenario:
Used to model and analyze countable business outcomes. For example, using historical sales data to create a probability distribution for the number of radios sold per week, calculating the expected weekly sales (μ_X), and measuring the variability (σ_X) around that expectation.

5.3 The Binomial Distribution

Basic Concepts:
The binomial distribution describes the probability of obtaining exactly x successes in n independent trials of a **binomial experiment**, which has these properties:

1. Fixed number (n) of identical trials.
2. Each trial has only two outcomes: *success* or *failure*.
3. Constant probability of success (p) for each trial.
4. Trials are independent.

Key Formula:

- **Binomial Probability:** $p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
  where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the binomial coefficient.

Mean and Variance:

- $\mu_X = np$
- $\sigma_X^2 = np(1-p)$

Application Scenario:
Ideal for situations with a fixed number of independent trials and a constant success probability. For instance, calculating the probability that exactly 2 out of 3 customers make a purchase if the historical purchase rate is 40%, or testing a drug manufacturer's claim about the rate of side effects by calculating the probability of observing a certain number of patients experiencing nausea in a sample.

Chapter 6: Continuous Random Variables

6.1 Continuous Probability Distributions

Basic Concepts:
A **continuous random variable** can assume any numerical value within intervals. Its probabilities are defined for intervals of values, not individual points, and are represented by the **area under a probability density function (PDF)**, denoted $f(x)$.

Properties of $f(x)$:

1. $f(x) \ge 0$ for all x
2. The total area under the curve of $f(x)$ is equal to 1.

Application Scenario:
Used to model inherently continuous business metrics, such as the exact waiting time for an elevator or the precise fuel efficiency (mpg) of a car model, where the probability of any single exact value is zero, but the probability for a range of values is meaningful.

6.2 The Uniform Distribution

Basic Concepts:
The **uniform distribution** describes a continuous random variable where all intervals of the same length within the distribution's range $[c, d]$ are equally likely.

Key Formulas:

- **PDF:** $f(x) = \frac{1}{d-c}$ for $c \le x \le d$
- **Probability:** $P(a \le x \le b) = \frac{b-a}{d-c}$
- **Mean:** $\mu = \frac{c+d}{2}$
- **Standard Deviation:** $\sigma = \frac{d-c}{\sqrt{12}}$

Application Scenario:
Ideal for modeling scenarios where outcomes are equally spread across a range, such as the waiting time for an elevator that arrives uniformly between 0 and 4 minutes.

6.3 The Normal Probability Distribution

Basic Concepts:
The **normal distribution** is the most important continuous distribution, characterized by its symmetrical, bell-shaped curve. It is defined by its mean (μ), which determines its center, and its standard deviation (σ), which determines its spread.

Key Formulas:

- **z-score (Standardization):** $z = \frac{x-\mu}{\sigma}$
- This converts any normal variable $X \sim N(\mu, \sigma)$ to the **standard normal distribution** $Z \sim N(0, 1)$.

Application Scenario:
Extremely widespread. Used with the **Empirical Rule** (68-95-99.7 rule) and **z-scores** to find probabilities. For example, calculating the proportion of cars with mileage between 32 and 35 mpg by converting to z-scores and using the standard normal table.

Chapter 7: Sampling Distributions

7.1 Sampling Distribution of the Sample Mean

Basic Concepts:

The **sampling distribution of the sample mean** is the probability distribution of all possible sample means obtained from all possible samples of a fixed size n from a population. It describes how the sample mean $\bar{x}$ varies from sample to sample.

Key Formulas:

- **Mean:** $\mu_{\bar{x}} = \mu$ (The mean of the sample means equals the population mean)
- **Standard Deviation (Standard Error):** $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- **Central Limit Theorem (CLT):** If n is sufficiently large (typically $n \geq 30$), the sampling distribution of $\bar{x}$ is approximately **normal**, regardless of the population's distribution shape.

Application Scenario:

Used to make probability statements about the sample mean. For example, calculating the probability that the mean amount of soda in four randomly selected bottles exceeds a certain value, or assessing the likelihood of observing a specific mean car mileage from a sample of 49 cars.

7.2 The Sampling Distribution of the Sample Proportion

Basic Concepts:

The **sampling distribution of the sample proportion** $\hat{p}$ is the probability distribution of all possible sample proportions for samples of size n. It describes how the sample proportion $\hat{p}$ varies from sample to sample.

Key Formulas:

- **Mean:** $\mu_{\hat{p}} = p$ (The mean of the sample proportions equals the population proportion)
- **Standard Deviation (Standard Error):** $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- The distribution is approximately **normal** if $np \geq 5$ and $n(1-p) \geq 5$.

Application Scenario:

Used for inference about a population proportion. For instance, a company tests a new product design on a sample; if the observed sample proportion of dissatisfied customers is very low, they use this distribution to calculate the probability of such a result and decide whether to adopt the new design based on strong evidence that the true population proportion is below a critical threshold.

Chapter 8: Confidence Intervals

8.1 z-Based Confidence Intervals for a Population Mean: σ Known

Basic Concepts:

This method constructs a confidence interval for a population mean (μ) when the population standard deviation (σ) is known. It relies on the sampling distribution of the sample mean being normal (either because the population is normal or via the Central Limit Theorem for large samples, n ≥ 30).

Basic Formulas:

- **(1-α)100% Confidence Interval:** $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- **Margin of Error (E):** $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Application Scenario:

Used to estimate a population mean with a known standard deviation. For example, estimating the mean mileage (mpg) for a car model where the historical process standard deviation is known to be 0.8 mpg.

8.2 t-Based Confidence Intervals for a Population Mean: σ Unknown

Basic Concepts:

When the population standard deviation (σ) is unknown (the usual case), the sample standard deviation (s) is used instead. The sampling distribution follows a **t-distribution** with $n-1$ degrees of freedom, which is more spread out than the normal distribution, especially for small samples.

Basic Formulas:

- **(1-α)100% Confidence Interval:** $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$

Application Scenario:

Used when estimating a population mean without knowing σ. For instance, estimating the mean debt-to-equity ratio for a bank's loan portfolio based on a sample, where the sample standard deviation s is calculated from the data.

8.3 Sample Size Determination

Basic Concepts:

This section provides a method to calculate the sample size (n) required to estimate a population mean with a specified margin of error (E) and confidence level.

Basic Formulas:

- **Sample Size for Mean (σ Known):** $n = (\frac{z_{\alpha/2}\cdot\sigma}{E})^2$

Application Scenario:

Used during the planning phase of a study. For example, determining how many car mileages need to be tested to estimate the mean mileage with a margin of error of 0.3 mpg and 95% confidence.

8.4 Confidence Intervals for a Population Proportion

Basic Concepts:

This method constructs a confidence interval for a population proportion (p), representing the percentage of units in a population that possess a certain characteristic.

Basic Formulas:

- **(1-α)100% Confidence Interval:** $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **Validity Condition:** $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$

Application Scenario:

Used to estimate a population proportion. For example, estimating the proportion of customers dissatisfied with a new product design (like a cheese spread spout) to decide if the proportion is low enough to adopt the new design.

Chapter 13: Simple Linear Regression Analysis

13.1 The Simple Linear Regression Model and the Least Squares Point Estimates

- Population Model: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Estimated Model: $\hat{y} = b_0 + b_1 x$
- Sum of Cross-Products: $SS_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$
- Sum of Squares for X: $SS_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$
- Slope Coefficient: $b_1 = \frac{SS_{xy}}{SS_{xx}}$
- Intercept: $b_0 = \bar{y} - b_1\bar{x}$

13.2 Model Assumptions and the Standard Error

- Error Sum of Squares: $SSE = \sum(y_i - \hat{y}_i)^2$
- Standard Error of Estimate: $s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}$
- Mean Square Error: $MSE = \frac{SSE}{n-2}$

13.4 Confidence and Prediction Intervals

- Confidence Interval for Mean Response: $\hat{y} \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_{xx}}}$
- Prediction Interval for Individual Response: $\hat{y} \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_{xx}}}$
- Distance Value: $\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_{xx}}$

13.5 Simple Coefficient of Determination and Correlation

- Total Sum of Squares: $SST = \sum(y_i - \bar{y})^2 = SS_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$
- Regression Sum of Squares: $SSR = \sum(\hat{y}_i - \bar{y})^2 = b_1 \cdot SS_{xy}$
- Coefficient of Determination: $r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- Correlation Coefficient: $r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$
- Relationship: $r = sign(b_1) \cdot \sqrt{r^2}$

Key Applications:

- b_1: Estimated change in Y per unit change in X
- r^2: Proportion of variance in Y explained by X
- CI: Interval estimate for mean Y at given X
- PI: Interval estimate for individual Y at given X

R2. About the final Examination

1. **Close-book**: Electronic devices such as smartphones and tablets are **forbidden**. You may bring a calculator with simple calculation functions.
2. **Time Limit**: The exam is timed and must be completed within 2 hours. Please manage your time wisely.
3. **Content**: The exam content covers Chapters 1 through 8 and Chapter 13 (excluding optional sections)..
4. **Question Types:** The exam question types are divided into basic concept questions (accounting for approximately 30% of the total score, ranging from 5 to 9 questions) and applied problems (accounting for approximately 30% of the total score, ranging from 5 to 9 questions).
5. **Difficulty:** The difficulty level of the final exam questions is similar to that of the homeworks. For more complex problems, appropriate formula hints or diagram references will be provided.