Review of the Topics for Midterm I

I. Introduction

The objective of statistics is to make inferences about a population based on information contained in a sample.

A population is the collection of all items (individuals, plants, corporations, ...) in a study. A sample is a portion of the population selected to represent the whole population.


Two Types of Statistical Analysis

a. Descriptive Statistics: Statistical methods used to develop tabular, graphical and/or numerical summaries of data

b. Inferential Statistics: The process of using data from a sample to draw conclusions about a population.

II. Descriptive Statistics

Sampling Unit, Data, Descriptive Statistics, Types Data.

III. Numerical Descriptive Measures

Sample Mean, Median, Mode, Quartiles.
Range, Sample Variance, Sample Standard Deviation, Coefficient of Variation, Interquartile Range.

The z-Score: z-score = (x-mean) / standard deviation

IV. Data Collection and Sampling
V. Probability and Discrete Probability Distributions

1. Introduction

a. What is the use of probability?
   - Probability allows us to handle variations in experimental outcomes mathematically. It helps us to deal with uncertainties.
   - Statistical inference is based on probability.

b. Experiment: Process of observing a random phenomenon.

c. Sample Space: The set of all possible distinct outcomes.

d. Event: The set of all outcomes having a certain feature.

2. Probability Rules

3. Conditional Probability and Independence

a. Let A and B be two events such that P(B) ≠ 0. The conditional probability of A given B is defined as:

\[ P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \]

b. Multiplication Rule:

   Use the definition of conditional probability:

   \[ P(A \text{ and } B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B) \]

c. Independence:

   Events A and B are independent if

   \[ P(A|B) = P(A) \text{ or } P(B|A) = P(B). \]
This leads to

\[ P(A \text{ and } B) = P(A) \cdot P(B). \]

**Example:** In a hospital 60\% of the employees are certified medical professionals and 80\% of the employees are full-time. Ninety percent of the professionals are full-time. We select an employee at random.

(a) What is the probability that the selected employee is a professional full-time employee?

(b) What is the probability that employee is a medical professional or a full-time employee?

(c) Are being a medical professional employee and being a full-time employee two independent events?

(d) What percentage of the full-time employees of this hospital is skilled?
4. Random Variables
   a. A random variable is a variable whose numerical value is determined by the outcome of a random experiment.
   b. Types of random variables:
      1. Discrete.
      2. Continuous.

5. Probability Distribution of a Discrete Random Variable
   a. The probability distribution of a discrete r.v. Y is represented by a probability function p(y) defined as: 
      \[ p(y) = P[Y = y] \]
   b. Properties of probability distribution.

6. Expected Value and Variance
   a. Let Y be a discrete random variable. The expected value (mean) of Y is defined as:
      \[ \mu = E(Y) = \sum y p(y) \]
   b. If Y is a random variable, then
      \[ E(a Y + b) = a E(Y) + b. \]
   c. Variance of a random variable Y is defined as
      \[ \sigma^2 = E(Y - \mu)^2 = \sum (y - \mu)^2 p(y) \]
   d. If Y is a random variable, then
      \[ \text{Var}(a Y + b) = a^2 \text{Var}(Y) \]
   e. If X and Y are two r.v.’s, then
      \[ E(a X + b Y + c) = a E(x) + b E(Y) + c. \]
   f. If X and Y are independent random variables, then
      \[ \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \]
      \[ \text{Var}(aX + bY + c) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) \]
VI. The Binomial Distribution

a. *The Binomial Experiment*: An experiment with the following characteristics:

1. Each trial has 2 possible outcomes, success and failure.
2. For any trial  \( P(\text{Success}) = p \) and  \( P(\text{Failure}) = 1 – p \).
3. The trials are independent of each other.

b. *Binomial Random variable*

Consider a binomial experiment. Define \( Y \) as

\[
Y = \text{Number of successes in } n \text{ trials.}
\]

\( Y \) is a random variable. The probability function of \( Y \) is:

\[
P(y) = \binom{n}{y} p^y (1-p)^{n-y} \quad y = 0, 1, \ldots, n
\]

For the binomial distribution

Mean: \( \mu = n \ p \), Variance: \( \sigma^2 = n \ p \ (1-p) \)

Example:
VII. The Normal Distribution

1. Introduction

One of the most important continuous distributions is the \textit{normal distribution}. Normal distributions are characterized by two \textit{parameters} \( \mu \) and \( \sigma \).

The empirical rule comes from the normal distribution.

2. Standard Normal Distribution

a. A normal distribution with mean 0 and standard deviation 1 is called a \textit{standard normal} distribution.

b. If \( Y \) has a normal distribution with mean \( \mu \) and standard deviation \( \sigma \), then

\[
Z = (Y - \mu) / \sigma
\]

is standard normal.

Examples:
VIII. Sampling Distributions

a. Sampling Distribution of $\bar{Y}$

- mean of the sample mean = mean of the population
- variance of the sample mean = variance of the population / sample size
- standard deviation of the sample mean = standard deviation of the population / square root of sample size

The Central Limit Theorem

For large sample sizes ($n \geq 30$), the sample mean is approximately normal with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$.

Example:

b. Sampling Distribution of a Proportion

Let $Y$ be the number of successes in $n$ independent trials. The sample proportion is defined as: 
\[ \hat{p} = \frac{Y}{n} \]

If $np > 5$ and $n(1-p) > 5$, then the sample proportion is approximately normal with mean $p$ and standard deviation $\sqrt{p(1-p)/n}$.

Example:
c. Normal Approximation to Binomial and Continuity Correction

Recall that for the binomial distribution we have:

\[
\text{Mean: } \mu = np \\
\text{Variance: } \sigma^2 = np(1-p)
\]

If \( np \geq 5 \) and \( n(1-p) \geq 5 \), we can approximate a binomial distribution with a normal distribution.

Example: