Exam 2 Notes

Random Variables

A **random variable** is a variable whose possible values are numerical outcomes of a random phenomenon. Random variables are classified into two types:

- **Discrete Random Variables**: These variables take on a countable number of distinct values. Example: The number of heads in a series of coin flips.
- **Continuous Random Variables**: These variables can take on any value within a given range, making the set of possible values uncountably infinite. Example: The weight of a randomly selected bag of apples.

Examples

1. **Coin Toss**: Let X denote the number of heads in a coin toss. Since the outcomes are countable (head or tail), X is a discrete random variable.
2. **Exam Scores**: Consider the scores of students in a test, ranging from 0 to 100. If Y represents a randomly selected student's score, Y is a discrete random variable because the scores, while within a range, are distinct and countable.
3. **Height of Students**: If Z represents the height of a randomly chosen student, Z is a continuous random variable because height can take on any value within a range, including decimals.

Key Properties

- **Probability Distribution**: This describes how probabilities are distributed over the values of the random variable. It's represented as a probability mass function (PMF) for discrete variables and a probability density function (PDF) for continuous variables.
- **Expected Value (Mean)**: The long-run average value of the random variable.
- **Variance and Standard Deviation**: Measures of the spread of the random variable's values around the mean.

Discrete Random Variables

Discrete random variables play a crucial role in statistical distributions, particularly when outcomes are countable. Key distributions include the Bernoulli, Binomial, and Poisson distributions, each with distinct characteristics and applications.

Bernoulli Distribution

The Bernoulli distribution models experiments with two outcomes: success (1) and failure (0), where each trial is independent.

- **Probability Mass Function (PMF)**: For a probability of success p, the PMF is defined as

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

  where $x \in \{0, 1\}$.

- **Expected Value**: $E[X] = p$ and expected value is calculated with

$$E[x] = \sum_i x_i p_i$$

  where x_i is the value of the variable and p_i is the probability of that value.

- **Variance**: $\text{Var}(X) = p(1-p)$ or with

$$\nu = \sum_i (x_i - \mu)^2 \cdot p_i = \mu[x^2] - \mu[x]^2$$

where ν is the variance, x_i is the individual observation, μ is the expected value of all observances, and p_i is the probability of a given observance x.

Binomial Distribution

This distribution extends the Bernoulli to n independent trials, each with the same success probability p.

- **PMF**: The probability of k successes in n trials is given by

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

where $\binom{n}{k}$ denotes the binomial coefficient.

- **Expected Value**: $E[X] = np$ and expected value is calculated with

$$E[x] = \sum_i x_i p_i$$

where x_i is the value of the variable and p_i is the probability of that value.

- **Variance**: $\text{Var}(X) = np(1-p)$ or with

$$\nu = \sum_i (x_i - \mu)^2 \cdot p_i = \mu[x^2] - \mu[x]^2$$

where ν is the variance, x_i is the individual observation, μ is the expected value of all observances, and p_i is the probability of a given observance x.

Poisson Distribution

The Poisson distribution models the count of events in a fixed interval, with events occurring at a constant mean rate λ, independently of the last event.

- **PMF**: For observing k events

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

where λ is the rate of the event and k is the number of events that occur.

- **Expected Value**: $E[X] = \lambda$ and expected value is calculated with

$$E[x] = \sum_i x_i p_i$$

where x_i is the value of the variable and p_i is the probability of that value.

- **Variance**: $\text{Var}(X) = \lambda$ or with

$$\nu = \sum_i (x_i - \mu)^2 \cdot p_i = \mu[x^2] - \mu[x]^2$$

where ν is the variance, x_i is the individual observation, μ is the expected value of all observances, and p_i is the probability of a given observance x.

Key Differences and Applications

- The **Bernoulli Distribution** is suitable for binary outcomes in a single trial, such as flipping a coin.
- The **Binomial Distribution** is used for counting successes in a fixed number of Bernoulli trials, like the number of heads in multiple coin flips.
- The **Poisson Distribution** efficiently models the count of events over a continuous interval, especially for rare events with a known average rate, such as emails received per hour.

These distributions are foundational for modeling discrete processes, understanding the probabilistic nature of phenomena, and form the basis for statistical inference in data science.

Continuous Random Variables

Continuous random variables take on an infinite number of values. Key distributions:

Normal (Gaussian) Distribution

A symmetric distribution centered around the mean, μ.

- **PDF**:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Properties**: Approximately 68% of the data falls within one standard deviation of the mean.

Exponential Distribution

Describes the time between events in a Poisson process.

- **PDF**:

$$f(x) = \lambda e^{-\lambda x}$$

$x \geq 0$.

- **Expected Value**:

$$E[X] = \frac{1}{\lambda}$$

- **Variance**:

$$\nu = \frac{1}{\lambda^2}$$

Standard Deviation

Standard deviation is a statistical measure that quantifies the amount of variation or dispersion in a set of data values. It indicates how much individual data points deviate from the mean (or expected value) of the data set.

Formula

The standard deviation for a **population** is denoted as σ and is calculated using the formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \sqrt{\nu}$$

where:

- N is the number of observations in the population,
- x_i represents each individual observation,
- $\bar{x}$ is the population mean,
- ν is the variance.

For a **sample** from the population, the sample standard deviation, denoted as s, uses Bessel's correction and is calculated as:

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

where:

- n is the sample size,
- x_i represents each individual sample observation,
- $\bar{x}$ is the sample mean.

Interpretation

The standard deviation is a critical tool in statistics, finance, and various fields, offering insights into the variability or risk of a dataset. In a normal distribution, approximately:

- 68% of observations fall within one standard deviation of the mean,
- 95% within two standard deviations,
- 99.7% within three standard deviations.

This empirical rule, also known as the 68-95-99.7 rule, emphasizes the role of standard deviation in evaluating the spread of a distribution.

Applications

Standard deviation is instrumental in:

- Comparing the spread between datasets with different means,
- Assessing financial risk and volatility,
- Determining the precision and reliability of statistical conclusions.

It provides a concrete measure of variability, enhancing our understanding of the distribution of data points beyond the mean.

Joint Distributions

Joint distributions are crucial in the study of probability and statistics, allowing us to analyze the behavior of two or more random variables simultaneously. This concept is pivotal for understanding the relationship between variables and for modeling complex probabilistic scenarios.

Definition

A joint distribution describes the probability distribution of two or more random variables occurring at the same time. It is represented by a joint probability mass function (PMF) for discrete variables, and a joint probability density function (PDF) for continuous variables.

Joint Probability Mass Function (PMF) for Discrete Variables

For discrete random variables X and Y, the joint PMF $p(x, y)$ gives the probability that X equals x and Y equals y simultaneously:

$$p(x, y) = P(X = x \text{ and } Y = y)$$

The sum of $p(x, y)$ over all possible values of x and y is 1.

Joint Probability Density Function (PDF) for Continuous Variables

In the case of continuous random variables X and Y, the joint PDF $f(x, y)$ describes the density of the probability that X and Y take on specific values within an infinitesimally small area. The integral of $f(x, y)$ over all possible values of x and y equals 1.

Properties and Applications

- **Marginal Distribution**: Derived from the joint distribution by summing (for PMF) or integrating (for PDF) over the range of one variable, providing the distribution of one variable irrespective of the other.
- **Conditional Distribution**: The distribution of one variable given the occurrence of another, calculated by dividing the joint distribution by the marginal distribution of the given variable.
- **Independence**: Two variables are independent if the joint distribution is the product of their marginal distributions. Independence implies that the outcome of one variable does not affect the distribution of the other.

Examples

- **Rolling Two Dice**: Analyzing the joint PMF of the sum and difference of the dice rolls to calculate probabilities for each outcome.
- **Height and Weight**: Using the joint PDF to model the relationship between height and weight in a population, enabling the calculation of probabilities for certain combinations of these variables.

Covariance, Correlation, and Independence

Understanding the relationships between two or more variables is pivotal in statistics and data science. Covariance, correlation, and independence are key concepts in this context.

Covariance

Covariance assesses the joint variability of two random variables. It is defined by the formula:

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

where E denotes the expected value. A positive covariance indicates that the variables tend to move in the same direction, while a negative covariance suggests they move in opposite directions.

Correlation

Correlation, specifically Pearson's correlation coefficient ρ, measures both the strength and direction of the linear relationship between two variables, calculated as:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Here, σ_X and σ_Y represent the standard deviations of X and Y, respectively. The correlation coefficient r ranges from -1 to 1, with -1 indicating a perfect negative linear relationship, 1 indicating a perfect positive linear relationship, and 0 indicating no linear relationship.

Independence

Two variables are considered independent if the occurrence of one does not affect the probability of occurrence of the other. Formally, X and Y are independent if and only if:

$$P(X \cap Y) = P(X)P(Y)$$

Statistically, independence implies a covariance of 0. However, a covariance of 0 does not necessarily indicate independence unless the variables are jointly normally distributed.

Relationship Between Concepts

- **Covariance** provides a measure of how two variables move together, though its scale makes it difficult to interpret the strength of their relationship.
- **Correlation** offers a standardized metric of covariance, presenting a dimensionless quantity that reflects the linear relationship's strength and direction.
- **Independence** suggests that two variables do not affect each other's outcomes, implied by a correlation of 0. However, a correlation of 0 does not ensure independence, as it only addresses linear relationships.

Sampling

Sampling is a fundamental process that involves selecting a subset of individuals or entities from a larger population to make inferences about the population's characteristics. This section covers different types of sampling, the population of interest, and the sampling frame.

Types of Sampling

Sampling methods are broadly categorized into **probability sampling** and **non-probability sampling**.

Probability Sampling

In probability sampling, every member of the population has a known, non-zero chance of being selected. It includes:

- **Simple Random Sampling**: Each subset of the population has an equal chance of being selected.
- **Stratified Sampling**: The population is divided into homogeneous subgroups, with random samples drawn from each.
- **Cluster Sampling**: The population is divided into clusters, some of which are randomly selected for sampling.
- **Systematic Sampling**: Every nth member of the population is selected, starting from a random point.

Non-Probability Sampling

Non-probability sampling does not give every member of the population a known chance of selection. Types include:

- **Convenience Sampling**: Sampling from easily accessible parts of the population.
- **Judgmental or Purposive Sampling**: Selecting members based on the researcher's judgment.
- **Quota Sampling**: Ensuring the sample reflects certain characteristics of the population.
- **Snowball Sampling**: Study subjects recruit future subjects from their acquaintances.

Population of Interest

The **population of interest** is the entire set of individuals or entities to which the study's findings aim to be generalized. Defining this population clearly is essential for the applicability and relevance of the research outcomes.

Sampling Frame

The **sampling frame** is the list from which the sample is actually drawn, ideally encompassing the entire population of interest. The presence of a *frame error*, where the frame does not perfectly match the population, can affect the sample's representativeness and the study's validity.

Multinomial Probabilities

Multinomial probabilities extend binomial probabilities to scenarios with more than two possible outcomes in each trial, crucial for analyzing experiments with multiple categories or outcomes.

Definition

The multinomial distribution generalizes the binomial distribution, modeling the probability of observing various counts among multiple categories over n trials, with each trial resulting in one of k possible outcomes.

Formula

The probability mass function (PMF) for the multinomial distribution is given by:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where:

- n is the total number of trials,
- x_i is the count of outcome i, for $i = 1, 2, \ldots, k$,
- p_i is the probability of outcome i in a single trial,
- $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} p_i = 1$.

Properties

- The sum of the probabilities of all possible outcomes is 1.
- Each variable X_i is binomially distributed, albeit not independently.
- The expected value of X_i is $n \cdot p_i$, with variance $n \cdot p_i \cdot (1 - p_i)$.

Applications

Multinomial probabilities are applied in genetics, marketing, natural language processing, and more, modeling:

- Genetic inheritance patterns of multiple alleles.
- Choice preferences among several product categories.
- Word distributions across topics in a document.

Central Limit Theorem

The Central Limit Theorem (CLT) is a cornerstone of statistics, asserting that the distribution of the sample mean of a large number of independent, identically distributed (i.i.d.) random variables approaches a normal distribution as the sample size increases, regardless of the original distribution's shape.

Statement of the Central Limit Theorem

Given i.i.d. random variables $X_1, X_2, \ldots, X_n$ with mean μ and finite variance σ^2, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ follows a normal distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$ as n approaches infinity. Formally:

$$\lim_{n\to\infty} P\left(a < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < b\right) = \Phi(b) - \Phi(a)$$

where Φ represents the cumulative distribution function of the standard normal distribution.

Implications

The CLT supports the assumption of normality in various statistical techniques, even when the population distribution is not known. It enables the approximation of probabilities associated with the sample mean, aiding in decision-making processes across numerous fields.

Applications

- In **Polling and Surveys**, it allows for the estimation of population means from sample data.
- In **Quality Control**, it assesses the distribution of means from multiple samples of product measurements.
- In **Experimental Research**, it facilitates the analysis of mean outcomes to infer effects on the population.

Standard Error and Z Value

The concepts of standard error (SE) and the z value are fundamental in statistical inference, particularly in the contexts of hypothesis testing and constructing confidence intervals.

Standard Error

The standard error measures the variability of a sample statistic (like the sample mean) relative to the actual population parameter, quantifying the uncertainty associated with the estimation. The standard error of the mean (SEM) is calculated as:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the sample size. In practice, with the population standard deviation often unknown, the SEM is estimated using the sample standard deviation (s):

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

As the sample size increases, the standard error decreases, reflecting a higher precision in estimating the population parameter.

Z Value

The z value, or z-score, indicates how many standard deviations an element is from the mean. In statistical inference, the z value is calculated to determine the distance of a sample statistic from the null hypothesis value, normalized by the standard error:

$$z = \frac{\bar{x} - \mu}{SE_{\bar{x}}}$$

where μ is the sample mean, $\bar{x}$ is the population mean under the null hypothesis, and $SE_{\bar{x}}$ is the standard error of the sample mean.

Implications and Applications

- **Hypothesis Testing**: The z value helps determine whether to reject the null hypothesis by comparing it to a critical value. If the z value falls outside the critical range, the null hypothesis is rejected.
- **Confidence Intervals**: The z value is used to calculate the margin of error when constructing confidence intervals, leveraging the normal distribution approximation for large sample sizes.

Hypothesis Testing

Hypothesis testing is a statistical method used to make decisions about a population parameter based on sample data. It involves the comparison of a null hypothesis (H_0) against an alternative hypothesis (H_1).

Key Concepts

- **Null Hypothesis (H_0)**: A statement asserting there is no effect or no difference, serving as the default assumption.
- **Alternative Hypothesis (H_1 or H_a)**: A statement contradicting the null hypothesis, asserting there is an effect or a difference.
- **Type I Error**: Occurs when the null hypothesis is incorrectly rejected (false positive).
- **Type II Error**: Occurs when the null hypothesis is not rejected when it is false (false negative).
- **Significance Level (α)**: The probability of committing a Type I error, typically set at 0.05.
- **P-value**: The probability of obtaining test results at least as extreme as the observed results, under the assumption that the null hypothesis is correct.

Steps in Hypothesis Testing

1. Formulate the null and alternative hypotheses.
2. Choose a significance level (α).
3. Collect and analyze sample data.
4. Calculate the test statistic (e.g., z, t).
5. Determine the p-value or compare the test statistic to critical values.
6. Make a decision: Reject H_0 if the evidence is against it; otherwise, do not reject H_0.

Types of Tests

Depending on the nature of the data and the hypothesis, different tests are used, including:

- **Z-test**: Used when the population variance is known and the sample size is large.
- **T-test**: Employed when the population variance is unknown and the sample size is small.
- **Chi-squared test**: Used for categorical data to assess how likely it is that an observed distribution is due to chance.
- **ANOVA**: Used to compare the means of three or more samples.

Exam 2 Cheat Sheet

Random Variables

Discrete Random Variables: These variables take on a countable number of distinct values. Example: The number of heads in a series of coin flips.

- The Bernoulli distribution models experiments with two outcomes: success (1) and failure (0), where each trial is independent: $E[X] = p, \nu = p(1-p)$

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

- The Binomial distribution extends the Bernoulli to n independent trials, each with the same success probability p: $E[X] = np, \nu = np(1-p)$

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

- The Poisson distribution models the count of events in a fixed interval, with events occurring at a constant mean rate λ, independently of the last event: $E[X] = \lambda, \nu = \lambda$

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

Continuous Random Variables: These variables can take on any value within a given range, making the set of possible values uncountably infinite. Example: The weight of a randomly selected bag of apples.

- The Normal distribution is a symmetric distribution centered around the mean, μ:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- An Exponential distribution describes the time between events in a Poisson process: $E[X] = 1/\lambda, \nu = 1/\lambda^2$

$$f(x) = \lambda e^{-\lambda x}$$

Expected Value

We calculate the expected value with

$$E[x] = \sum_i x_i p_i$$

where x_i is the value of the variable and p_i is the probability of that value.

Variance

We calculate the variance of a variable with

$$\nu = \sum_i (x_i - \mu)^2 \cdot p_i = \mu[x^2] - \mu[x]^2$$

where ν is the variance, x_i is the individual observation, μ is the expected value of all observances, and p_i is the probability of a given observance x.

Standard Deviation

Standard deviation is a statistical measure that quantifies the amount of variation or dispersion in a set of data values. It indicates how much individual data points deviate from the mean (or expected value) of the data set. It is calculated with

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \sqrt{\nu}$$

where N is the number of observations in the population, x_i represents each individual observation, $\bar{x}$ is the population mean, ν is the variance.

Covariance

Covariance assesses the joint variability of two random variables. It is defined by the formula:

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

Correlation

Correlation, specifically Pearson's correlation coefficient ρ, measures both the strength and direction of the linear relationship between two variables, calculated as:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Independence

Two variables are considered independent if the occurrence of one does not affect the probability of occurrence of the other. Formally, X and Y are independent if and only if:

$$P(X \cap Y) = P(X)P(Y)$$

Probability Sampling

In probability sampling, every member of the population has a known, non-zero chance of being selected. It includes:

- **Simple Random Sampling**: Each subset of the population has an equal chance of being selected.
- **Stratified Sampling**: The population is divided into homogeneous subgroups, with random samples drawn from each.
- **Cluster Sampling**: The population is divided into clusters, some of which are randomly selected for sampling.
- **Systematic Sampling**: Every nth member of the population is selected, starting from a random point.

Non-Probability Sampling

Non-probability sampling does not give every member of the population a known chance of selection. Types include:

- **Convenience Sampling**: Sampling from easily accessible parts of the population.
- **Judgmental or Purposive Sampling**: Selecting members based on the researcher's judgment.
- **Quota Sampling**: Ensuring the sample reflects certain characteristics of the population.
- **Snowball Sampling**: Study subjects recruit future subjects from their acquaintances.

Sampling Frame

The **sampling frame** is the list from which the sample is actually drawn, ideally encompassing the entire population of interest. The presence of a *frame error*, where the frame does not perfectly match the population, can affect the sample's representativeness and the study's validity.

Multinomial Probabilities

The multinomial distribution generalizes the binomial distribution, modeling the probability of observing various counts among multiple categories over n trials, with each trial resulting in one of k possible outcomes.

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

where n is the total number of trials, x_i is the count of outcome i, for $i = 1, 2, \ldots, k$, p_i is the probability of outcome i in a single trial, $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} p_i = 1$.

Central Limit Theorem

The Central Limit Theorem (CLT) is a cornerstone of statistics, asserting that the distribution of the sample mean of a large number of independent, identically distributed (i.i.d.) random variables approaches a normal distribution as the sample size increases, regardless of the original distribution's shape.

Standard Error

The standard error measures the variability of a sample statistic (like the sample mean) relative to the actual population parameter, quantifying the uncertainty associated with the estimation. The standard error of the mean (SEM) is calculated as:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Z Value

The z value, or z-score, indicates how many standard deviations an element is from the mean. In statistical inference, the z value is calculated to determine the distance of a sample statistic from the null hypothesis value, normalized by the standard error:

$$z = \frac{\bar{x} - \mu}{SE_{\bar{x}}}$$

where μ is the sample mean, $\bar{x}$ is the population mean under the null hypothesis, and $SE_{\bar{x}}$ is the standard error of the sample mean.

Key Concepts

Hypothesis testing is a statistical method used to make decisions about a population parameter based on sample data. It involves the comparison of a null hypothesis (H_0) against an alternative hypothesis (H_1).

- **Null Hypothesis (H_0)**: A statement asserting there is no effect or no difference, serving as the default assumption.
- **Alternative Hypothesis (H_1 or H_a)**: A statement contradicting the null hypothesis, asserting there is an effect or a difference.
- **Type I Error**: Occurs when the null hypothesis is incorrectly rejected (false positive).
- **Type II Error**: Occurs when the null hypothesis is not rejected when it is false (false negative).
- **Significance Level (α)**: The probability of committing a Type I error, typically set at 0.05.
- **P-value**: The probability of obtaining test results at least as extreme as the observed results, under the assumption that the null hypothesis is correct.