College of Engineering & Applied Sciences

CSPB 3022

Introduction To Data Science With Probability And Statistics

Exam Notes

Taylor Larrechea

2024

Exam 1 Notes

Data Frames

A data frame in Pandas is a two-dimensional, size-mutable, and potentially heterogeneous tabular data structure with labeled axes (rows and columns). It's akin to a spreadsheet or SQL table and is one of the most commonly used Pandas data structures.

You can create a data frame from various sources, such as:

- Lists
- Dictionaries
- NumPy arrays
- CSV files
- SQL databases

The structure of data frames can be summed up by:

- **Rows**: Each row represents a single observation or record.
- **Columns**: Each column represents a variable or feature. Columns can be of different data types (integer, string, float, etc.).
- **Index**: This is the 'key' for rows, similar to an index in a database. It's an immutable array, allowing fast access to data.

Some operations that can be used with data frames are:

- **Data Manipulation**: Adding, deleting, and modifying both rows and columns.
- **Filtering**: Selecting a subset of rows or columns based on some criteria.
- **Sorting and Grouping**: Organizing data based on values in certain columns.
- **Merging and Joining**: Combining multiple data frames.
- **Handling Missing Data**: Identifying and imputing missing values.

Data Frame Operation Examples

Here are some examples of data frame manipulation in pandas:

Creation

```
import pandas as pd

# Creating a data frame from a dictionary
data = {'Name': ['Alice', 'Bob', 'Charlie'],
        'Age': [25, 30, 35],
        'City': ['New York', 'Los Angeles', 'Chicago']}
df = pd.DataFrame(data)

```

Adding A Column

```
# Adding a new column
df['Salary'] = [70000, 80000, 90000]

```

Deleting A Column

```
# Deleting a column
df.drop('Age', axis=1, inplace=True)

```

Filtering Data

```
# Filtering rows where Salary is greater than 75000
high_earners = df[df['Salary'] > 75000]

```

Sorting Data

```
# Sorting data by Salary in descending order
df_sorted = df.sort_values(by='Salary', ascending=False)

```

Merging Data Frames

```
# Creating another data frame
additional_data = pd.DataFrame({'Name': ['Alice', 'Bob'], 'Experience': [5, 10]})

# Merging data frames
merged_df = pd.merge(df, additional_data, on='Name', how='left')

```

Handling Missing Data

```
# Filling missing values with zero
df_filled = df.fillna(0)

```

Reading Data

```
# Reading data from a CSV file
df_from_csv = pd.read_csv('data.csv')

# Writing data to a CSV file
df.to_csv('output.csv', index=False)

```

Combinatorics

Combinatorics is a branch of mathematics dealing with the study of countable, discrete structures and their properties. It's particularly important in computer science, where understanding how to count and arrange objects is crucial for algorithm design, data structure optimization, and problem-solving. Here's a summary of the key concepts in combinatorics:

- **Counting Principles**
  - **The Rule of Sum**: If there are A ways to do something and B ways to do another thing, and these two things cannot happen at the same time, then there are $A + B$ ways to choose one of these actions.
  - **The Rule of Product**: If there are A ways to do something and B ways to do another thing after that, then there are $A \cdot B$ ways to perform both actions.

- **Permutations**
  - Permutations are the arrangements of objects in a specific order.
  - The number of permutations of n distinct objects is $n!$ (n factorial), which is the product of all positive integers up to n.
  - For arranging r objects out of n available objects, the formula is

$$^nP_r = \frac{n!}{(n-r)!}.$$

- **Combinations**
  - Combinations refer to the selection of objects without regard to the order.
  - The number of ways to choose r objects from n different objects is given by

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

  - Combinations are used when the order doesn't matter.
- **Binomial Theorem**
  - It provides a formula for the expansion of powers of a binomial (sum of two terms).
  - The Binomial Theorem states that:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^{n-k} \cdot b^k$$

    * This means the expansion is a sum of terms, where the exponents of a start at n and decrease to 0, while the exponents of b start at 0 and increase to n. The coefficients of each term are the corresponding binomial coefficients.
  - The coefficients of the terms in the expansion are the binomial coefficients, which can be calculated using combinations.
- **Binomial Distribution**
  - A binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials. A Bernoulli trial is an experiment with exactly two possible outcomes, typically termed "success" and "failure".
  - In the context of the binomial distribution, $P(x = k)$ denotes the probability of getting exactly k successes in n trials. The formula for this is

$$P(x = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

    * $\binom{n}{k}$ (read as "n choose k") is the binomial coefficient, representing the number of ways to choose k successes from n trials.
    * p is the probability of success on an individual trial.
    * $1 - p$ is the probability of failure (since the trials are binary, the sum of the probabilities of success and failure is 1).
    * p^k is the probability of having k successes.
    * $(1-p)^{n-k}$ is the probability of having $n - k$ failures.

groupby

The `groupby` method is used to split data into groups based on some criteria, apply a function to each group independently, and then combine the results into a data structure. The process is often summarized as split-apply-combine.

The way `groupby` works is by the following:

- **Split**: The data is divided into groups based on one or more keys. This is done by mapping a function over the index or columns of the DataFrame.
- **Apply**: A function is applied to each group independently. This function could be an aggregation (computing a summary statistic), transformation (standardizing data within a group), or filtration (removing data based on group properties).
- **Combine**: The results of the function application are combined into a new data structure.

The basic syntax for how `groupby` works is as follows:

```
df.groupby(by=None, axis=0, level=None, as_index=True, sort=True, group_keys=True, squeeze=NoDefault.
no_default, observed=False, dropna=True)

```

- **by**: Specifies the grouping criteria. Can be a function, column name, or list of column names.
- **axis**: Determines whether to group by rows (0) or columns (1).
- **level**: If the axis is a MultiIndex (hierarchical), groups by a particular level or levels.
- **as_index**: For aggregated output, returns object with group labels as the index. Setting it to False will return group labels in the columns.
- **sort**: Sorts group keys. By default, it's set to True.

groupby Example

Here is a simple example of utilizing `groupby` in Pandas:

```
import pandas as pd

# Example DataFrame
data = {
    'Date': ['2023-01-01', '2023-01-01', '2023-01-02', '2023-01-02', '2023-01-03'],
    'Product': ['A', 'B', 'A', 'A', 'B'],
    'Sales': [100, 200, 150, 100, 250]
}
df = pd.DataFrame(data)

# Group by 'Product' and sum up the sales
grouped_df = df.groupby('Product')['Sales'].sum()

print(grouped_df)

```

agg

The `agg` method in Pandas, when used with groupby, is a versatile tool for performing multiple aggregation operations on your grouped data. It allows for more flexibility than just applying a single aggregate function like `sum` or `mean` directly. With `agg`, you can apply different aggregation functions to your data simultaneously, and even specify custom functions to suit your analysis needs.

After grouping your DataFrame with the `groupby` method, you can use `agg` to specify one or more aggregation operations to apply to the grouped data. `agg` can take a variety of inputs:

- A single aggregation function as a string (e.g. `'sum'`, `'mean'`).

- A list of functions (e.g., `['sum', 'mean', 'max']`), applying each function to each column of each group.
- A dictionary where keys are column names and values are functions or lists of functions, allowing different aggregation for different columns.

agg Example

Expanding on the previous example that was used with `groupby`, the example below encapsulates how to use the `agg` function with `groupby`:

```
import pandas as pd

# Example DataFrame
data = {
    'Date': ['2023-01-01', '2023-01-01', '2023-01-02', '2023-01-02', '2023-01-03'],
    'Product': ['A', 'B', 'A', 'A', 'B'],
    'Sales': [100, 200, 150, 100, 250]
}
df = pd.DataFrame(data)

# Group by 'Product' and apply multiple aggregation functions to 'Sales'
grouped_df = df.groupby('Product')['Sales'].agg(['sum', 'mean', 'max'])

print(grouped_df)

```

This would output the total, average, and maximum sales for each product.

More advanced usage can be applied to the `agg` function, below is an example with the previous data frame in the last example of this advanced usage:

```
grouped_df = df.groupby('Product').agg({
    'Sales': ['sum', 'mean'],
    'Date': ['min', 'max']
})

```

This performs a sum and mean aggregation on the `Sales` and finds the minimum and maximum `Date` for each `Product`.

Joining DataFrames

Joining data frames is a fundamental operation in data analysis, allowing you to combine data from different sources based on common identifiers. Pandas offers several methods for joining DataFrames, akin to SQL joins, including `merge`, `join`, and `concat`. Understanding these methods and when to use each is key to effective data manipulation.

1. **merge**: The `merge` function is the most versatile method for joining two DataFrames. It allows you to perform inner, outer, left, and right joins by specifying how you want the DataFrames to be merged.
   - **Syntax**: `pd.merge(left, right, how='inner', on=None, left_on=None, right_on=None, left_index=False, right_index=False)`
   - **Parameters**:
     - `left, right`: The DataFrames you want to join.
     - `how`: Specifies the type of join `('left', 'right', 'outer', 'inner')`.
     - `on`: The column(s) to join on. Must be found in both DataFrames.
     - `left_on, right_on`: Columns from the left and right DataFrames to use as keys if they have different names.
     - `left_index, right_index`: If True, use the index (row labels) from the left or right DataFrame as its join key(s).

merge Example

Below is a simple example of how `merge` works by joining DataFrames:

```
import pandas as pd

# Example DataFrames
df1 = pd.DataFrame({'key': ['A', 'B', 'C', 'D'], 'value': range(4)})
df2 = pd.DataFrame({'key': ['B', 'D', 'D', 'E'], 'value': range(4, 8)})

# Inner join on 'key'
inner_joined = pd.merge(df1, df2, on='key', how='inner')

# Outer join on 'key'
outer_joined = pd.merge(df1, df2, on='key', how='outer')

```

2. **join**: The `join` method is a convenient method for combining DataFrames based on their indexes or on a key column. It's a simpler interface than `merge` for index-based joining.

   - **Syntax**: `DataFrame.join(other, on=None, how='left', lsuffix='', rsuffix='')`
   - **Parameters**:
     - `other`: The DataFrame to join with.
     - `on`: The column or index level names to join on in the `other` DataFrame. Must be found in both the calling DataFrame and `other`.
     - `how`: Type of join `('left', 'right', 'outer', 'inner')`.
     - `lsuffix, rsuffix`: Suffixes to apply to overlapping column names in the DataFrames.

join Example

Below is a simple example of how `join` works by joining DataFrames:

```
# Assuming df1 and df2 from the previous example
# Join df2 to df1 using the index of df1 and the 'key' column of df2
joined = df1.join(df2.set_index('key'), on='key', how='left', lsuffix='_df1', rsuffix='_df2')

```

3. **concat**: The `concat` function is used for concatenating DataFrames along a particular axis (row-wise or column-wise). It's useful for stacking DataFrames vertically or horizontally.

   - **Syntax**: `pd.concat(objs, axis=0, join='outer')`
   - **Parameters**:
     - `objs`: A sequence of DataFrames to concatenate.
     - `axis`: The axis to concatenate along (0 for rows, 1 for columns).
     - `join`: How to handle indexes on other axis `('outer', 'inner')`.

concat Example

Below is a simple example of how `concat` works by joining DataFrames:

```
# Vertical concatenation
vertical_concat = pd.concat([df1, df2])

# Horizontal concatenation, assuming same indexes
horizontal_concat = pd.concat([df1, df2], axis=1)

```

Measures Of Central Tendency

There are core ways we can measure information about data, below are some pertinent ways that we measure central tendency in data science:

1. **Mean** - The mean, often referred to as the average, is calculated by summing all the values in the data set and then dividing by the number of values. It's a measure that is sensitive to outliers, meaning that a single very high or very low value can significantly affect the mean.

   The formula for calculating the mean of a data set is:

   $$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

   where x_i represents each value in the dat set and n is the total number of values.

2. **Median** - The median is the middle value in a data set when the values are arranged in ascending or descending order. If there's an even number of observations, the median is the average of the two middle numbers. Unlike the mean, the median is not affected by outliers, making it a better measure of central tendency for skewed distributions.

   - **Finding The Median**:
     - Arrange the data in ascending order.
     - If the number of observations (n) is odd, the median is the value at position $\frac{n+1}{2}$.
     - If n is even, the median is the average of the values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

3. **Mode** - The mode is the value that appears most frequently in a data set. A data set may have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode at all if all values occur with the same frequency. The mode is particularly useful for categorical data where we want to identify the most common category.

   - **Key Characteristics**:
     - The mode is the only measure of central tendency that can be used with nominal data (data categorized without a natural order).
     - A data set may have multiple modes, making it a good measure for understanding variability in data.

Central Tendency Example

Consider the simple data set: `2,3,3,5,7,10`

Mean

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+3+3+5+7+10}{6} = \frac{30}{6} = 5.$$

Median

The data set is even, so the median is the average of the middle values (3rd and 4th values):

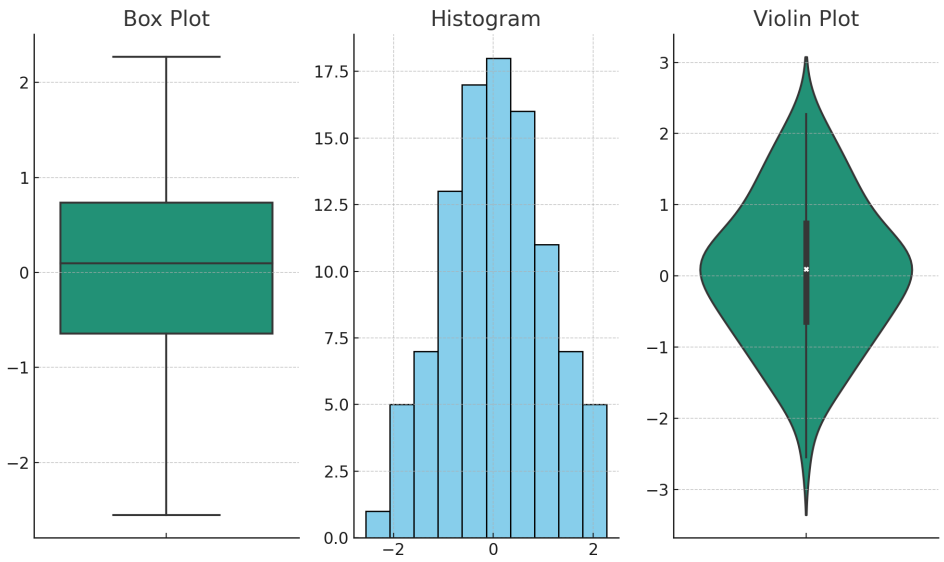
$$\text{Median} = \frac{3+5}{2} = \frac{8}{2} = 4.$$

Mode

The value `3` appears most frequently, so the mode of this data set is `3`.

Plot Interpretation

Plots are a vital concept in data science. Interpreting them is a vital skill and important for understanding data and how to draw conclusions. For this, we have created three types of plots that are common in data science. These plots can be seen below.



1. **Box Plot**: A box plot (or box-and-whisker plot) provides a visual summary of the key aspects of a distribution:
   - **Central line**: Represents the median of the data. In our box plot, the median is around 0, indicating the central tendency of the distribution.
   - **Box**: The edges of the box (the interquartile range, IQR) show the 25th percentile (Q1) and the 75th percentile (Q3) of the data. The length of the box indicates the spread of the central 50% of the data. A shorter box suggests less variability in the middle half of the data.
   - **Whiskers**: Extend from the box to show the range of the data. Typically, they extend to 1.5 * IQR beyond the quartiles, though this can vary. Points beyond the whiskers are considered outliers and are often plotted as individual points.
   - **Outliers**: Points outside the whiskers are plotted individually, indicating that they are unusual compared to the rest of the data. In our plot, there are a few outliers on both sides, indicating data points significantly higher or lower than the majority.
2. **Histogram**: A histogram displays the distribution of data by forming bins along the range of the data and then drawing bars to show the number of observations that fall in each bin:
   - **Bins**: The x-axis represents bins or intervals of values. In our histogram, data is divided into 10 bins.
   - **Height of bars**: The y-axis represents the frequency of data points within each bin. The height of a bar shows how many data points fall into that range.
   - **Distribution shape**: The histogram helps identify the distribution shape of the data, whether it's normal, skewed, or bimodal, for example. Our histogram suggests a roughly normal distribution, as the bars form a bell-shaped curve around the center.
3. **Violin Plot**: A violin plot combines elements of box plots and density plots, providing a deeper understanding of the distribution:
   - **Outer shape**: Represents the kernel density estimation of the data, showing the distribution's density. Thicker sections of the violin plot indicate a higher density of data points, while thinner sections indicate lower density.
   - **Inner box plot**: Often included within the violin, showing the median (central dot) and the interquartile range (thick bar). The violin plot in our example shows a median around 0, similar to the box plot, with the bulk of data points concentrated around the center, indicating the central tendency and variability.

- **Width**: The width of the violin at different values indicates the density of the data at those values. In our plot, the widest part is around the median, suggesting the highest density of data points is near the center of the distribution.

IQR

The Interquartile Range (IQR) is a measure of statistical dispersion, or variability, that indicates the spread of the middle 50% of a dataset. It's calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of the data. The IQR is used to build box plots, as we've seen, and is helpful for identifying outliers.

In essence, the IQR can be calculated with

$$\text{IQR} = Q3 - Q1.$$

where

- $Q1$ (the first quartile) is the median of the first half of the dataset (the 25th percentile).
- $Q3$ (the third quartile) is the median of the second half of the dataset (the 75th percentile).

IQR Strategy

The steps to calculate the IQR are as follows:

1. **Order the Data**: Arrange the data points in ascending order.
2. **Find the Median (Q2)**: This divides the dataset into two halves.
3. **Find Q1 and Q3**:
   - $Q1$ is the median of the data points to the left of the median of the dataset.
   - $Q3$ is the median of the data points to the right of the median.
4. **Calculate the IQR**: Subtract $Q1$ from $Q3$.

IQR is important because:

- **Resistant to Outliers**: Unlike range and standard deviation, the IQR focuses on the middle bulk of the data and is not affected by outliers. This makes it a more robust measure of spread for skewed distributions.
- **Identifying Outliers**: The IQR is used to define outliers. A common rule is that an outlier is any data point that lies more than 1.5 times the IQR above the third quartile (Q3) or below the first quartile (Q1).

IQR Example

Consider the sample data set `3,7,5,8,6,9`:

1. **Ordered Data Set**: `3,5,6,7,8,9`
2. **Median**: The median of this data set is $\frac{6+7}{2} = 6.5$
3. **Find Q1 and Q3**:
   - $Q1$ is the median of the lower 50%, `(3,5,6)`, and is 5
   - $Q3$ is the median of the upper 50%, `(7,8,9)`, and is 8
4. **Calculate IQR** $IQR = Q3 - Q1 = 8 - 5 = 3$

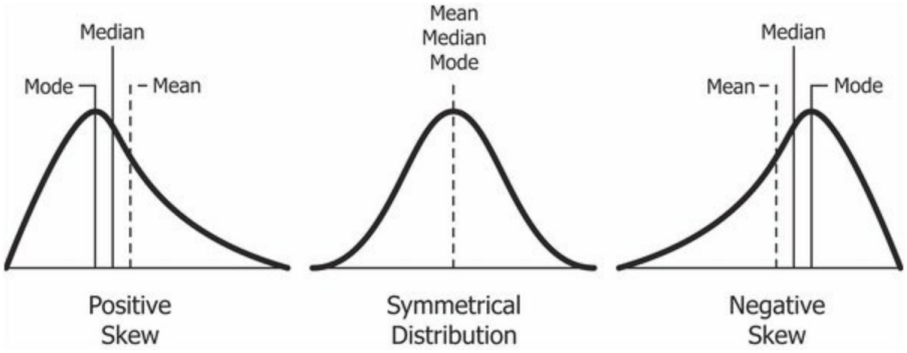
Skew

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. In simpler terms, it's a way to describe the direction and extent to which a distribution deviates from a normal distribution, where the mean, median, and mode are all the same.

The different types of skews are:

- **Positive Skew (Right-Skewed)**: The tail on the right side of the distribution is longer or fatter than the left side. In a positively skewed distribution, the mean is typically greater than the median, which is greater than the mode. This type of skew indicates that there are a number of exceptionally high values pulling the mean to the right.
- **Negative Skew (Left-Skewed)**: The tail on the left side of the distribution is longer or fatter than the right side. In a negatively skewed distribution, the mean is typically less than the median, which is less than the mode. This skew indicates that there are a number of exceptionally low values pulling the mean to the left.

When interpreting the central tendency from a skewed data set, the mean is located on the side of the skew. We can see the different central tendencies of each type of skew below.



Probability

In probability, particularly sets, there is certain terminology that is used. In the context of set notation, the following are terms that are pertinent to probability theory are

1. **Complement**: The complement of a set A, denoted as A' or A^c, represents all elements not in A but within the universal set U, which contains all possible outcomes. In probability, the complement of an event represents the probability that the event does not happen.
   - **Notation**: A^c of A'
   - **Formula**: The formula for calculating the complement of the probability of A ($P(A)$) is

$$P(A') = 1 - P(A)$$

2. **Intersection**: The intersection of two sets A and B, denoted as $A \cap B$, represents all elements that are both in A and in B. In probability, the intersection of two events represents the probability that both events occur simultaneously.
   - **Notation**: $A \cap B$
   - **Formula**: The formula for calculating $P(A \cap B)$ depends on whether events A and B are independent or dependent.
     - For **independent events**

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

- For **dependent events**, you need additional information to calculate $P(A \cap B)$ directly.

3. **Union** The union of two sets A and B, denoted as $A \cup B$, represents all elements that are in A, in B, or in both. In probability, the union of two events represents the probability that at least one of the events occurs.
   - **Notation**: $A \cup B$
   - **Formula**: The formula for calculating the union of two probabilities is

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

Event Space

Event Space, also known as Sample Space in the context of probability, is a fundamental concept that represents all possible outcomes of a probabilistic experiment. Understanding the event space is crucial for calculating probabilities, as it sets the stage for identifying specific events within the realm of all outcomes.

We define event and sample space as the following:

- **Sample Space (S)**: The set of all possible outcomes of a probabilistic experiment. It could be finite or infinite, depending on the nature of the experiment.
- **Event (E)**: Any subset of the sample space. An event is a specific outcome or a combination of outcomes that we're interested in.

The type of sample spaces are the following:

- **Finite Sample Space**: When the number of possible outcomes is countable and limited. For example, rolling a six-sided die has a finite sample space of `{1, 2, 3, 4, 5, 6}`.
- **Infinite Sample Space**: When the number of possible outcomes is uncountable or unlimited. For example, measuring the time it takes for something to happen could have an infinite number of possible outcomes.

Conditional Probability

Conditional probability is a fundamental concept in probability theory that deals with finding the probability of an event given that another event has occurred. This concept is crucial for understanding the relationship between two events and how the occurrence of one event affects the likelihood of another.

The conditional probability of an event A given that event B has occurred is denoted as $P(A|B)$ and is defined as:

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

provided that $P(B) > 0$.

The way we can interpret conditional probability is the following:

- $P(A|B)$ is read as 'the probability of A given B'.
- It quantifies how the probability of A changes when we know that B has occurred.
- This is different from the probability of A occurring in isolation, without any knowledge about B.

Key points that can be taken away from conditional probability:

- **Dependence**: Conditional probability is used when the occurrence of one event affects the probability of another, indicating a dependency between events.
- **Recalculation of Probabilities**: Knowing that B has occurred may change the sample space and, consequently, the probabilities of subsequent events.
- **Bayes' Theorem**: A fundamental result that follows from the concept of conditional probability, allowing for the update of probabilities as more evidence becomes available.

Bayes Theorem

Bayes' Theorem is a powerful tool in probability theory that allows us to update our beliefs about the probability of an event based on new evidence. It's named after Thomas Bayes (1701-1761), a British statistician and philosopher. This theorem is particularly useful in the field of data science for making predictions under uncertainty and has applications ranging from spam filtering to medical diagnosis.

Bayes' Theorem is as follows:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

where:

- $P(A|B)$ is the posterior probability of event A occurring given that B has occurred.
- $P(B|A)$ is the likelihood, the probability of event B occurring given that A has occurred.
- $P(A)$ is the prior probability of event A occurring.
- $P(B)$ is the marginal probability of event B occurring.

Independence

Independence is a key concept in probability theory that describes a situation where the occurrence of one event does not affect the probability of another event occurring. Two events, A and B, considered independent if and only if the occurrence of A has no impact on the probability of B happening, and vice versa.

The formal definition of independence is two events A and B are independent if any of the following equivalent statements holds true:

1. The probability that both A and B occur is equal to the product of the probabilities of each event occurring separately:
$$P(A \cap B) = P(A) \cdot P(B)$$
2. The conditional probability of A given B is the same as the probability of A, and vice versa:
$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

Independence intuitively means that knowing whether B shas occurred provides no information about the occurrence of A, and the reverse is also true. For instance, if you flip a fair coin twice, the outcome of the first flip does not influence the outcome of the second flip; these events are independent.

Exam 1 Cheat Sheet

The Rule of Sum

If there are A ways to do something and B ways to do another thing, and these two things cannot happen at the same time, then there are $A + B$ ways to choose one of these actions.

The Rule of Product

If there are A ways to do something and B ways to do another thing after that, then there are $A \cdot B$ ways to perform both actions.

Permutations

Permutations are the arrangements of objects in a specific order. For arranging r objects out of n available objects, the formula is

$$^nP_r = \frac{n!}{(n-r)!}$$

Combinations

Combinations refer to the selection of objects without regard to the order. The number of ways to choose r objects from n different objects is given by

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Binomial Theorem

Provides a formula for the expansion of powers of a binomial (sum of two terms).

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^{n-k} \cdot b^k$$

Binomial Distribution

A binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials

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

- $\binom{n}{k}$ (read as 'n choose k') is the binomial coefficient, representing the number of ways to choose k successes from n trials.
- p is the probability of success on an individual trial.
- $1-p$ is the probability of failure (since the trials are binary, the sum of the probabilities of success and failure is 1).
- p^k is the probability of having k successes.
- $(1-p)^{n-k}$ is the probability of having $n-k$ failures.

IQR

The spread of the middle 50% of the data set

$$\text{IQR} = Q3 - Q1$$

Skew

- **Positive Skew (Right-Skewed)**: The tail on the right side of the distribution is longer or fatter than the left side. In a positively skewed distribution, the mean is typically greater than the median, which is greater than the mode. This type of skew indicates that there are a number of exceptionally high values pulling the mean to the right.
- **Negative Skew (Left-Skewed)**: The tail on the left side of the distribution is longer or fatter than the right side. In a negatively skewed distribution, the mean is typically less than the median, which is less than the mode. This skew indicates that there are a number of exceptionally low values pulling the mean to the left.

Complement

$$P(A') = 1 - P(A)$$

Intersection

For **independent events**, the intersection is

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

Union

The union of two probabilistic events are

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 1 - P(A' \cap B') \end{aligned}$$

Conditional Probability

The conditional probability of an event A given that event B has occurred is denoted as $P(A|B)$ and is defined as

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

Individual Probability

We can calculate the individual probability based upon the conditional probability with another event and the complement of the other event, namely

$$P(B) = P(B|A)P(A) + P(B|A')P(A')$$

Contingency Table

We can use contingency tables to determine probabilities of other events that we don't information about

Prob.	$P(A)$	$P(A')$	Total
$P(B)$	$P(A \cap B)$	$P(A' \cap B)$	$P(B)$
$P(B')$	$P(A \cap B')$	$P(A' \cap B')$	$P(B')$
Total	$P(A)$	$P(A')$	1

Bayes Theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- $P(A|B)$ is the posterior probability of event A occurring given that B has occurred.
- $P(B|A)$ is the likelihood, the probability of event B occurring given that A has occurred.
- $P(A)$ is the prior probability of event A occurring.
- $P(B)$ is the marginal probability of event B occurring.

Independence

The formal definition of independence is two events A and B are independent if any of the following equivalent statements holds true:

- The probability that both A and B occur is equal to the product of the probabilities of each event occurring separately:

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

- The conditional probability of A given B is the same as the probability of A, and vice versa:

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

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

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

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

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

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

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

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:

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

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 r, 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:

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

where $\bar{x}$ 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.