ASSIGNMENT 2 - Applying Statistics using Scipy

Questions To Add your Responses :

1. Explain what is the Central Limit Theorem to a non-technical audience?
2. What is Sampling and following techniques required to gather sample data?
3. Describe the Hypothesis Testing and why do we conduct it?
4. Define and briefly elaborate Central Tendency using measures with examples?
5. What's a ROC Curve? How do you differentiate Point Estimate and Confidence Interval Estimate?
6. What's the difference between Correlation and Covariance in Statistics? What are the goals of A/B Testing?
7. How would you build and test a metric to compare two user's ranked lists of movie/tv show preferences?
8. Find the probability of P(x<400) given that mean is = 1000 variance is =100
9. Provide a proper description as to when to apply Ztest, t-test, Chi-Square and annova, with examples.

Explain the questions and problems in your approach while solving them. No copy paste allowed. Description of the outcome is really important to get an understanding better.

1) Explain what is the Central Limit Theorem to a non-technical audience?

Imagine you have a giant jar filled with different colour marbles, and you want to know the average colour (let's say each colour has a number associated with it). Instead of counting all the marbles, you randomly grab a handful and calculate the average colour for that small group. If you keep doing this, taking many small groups of marbles and calculating their average colour each time, something interesting happens:

No matter what the original mix of colours in the jar is, the average colour of these small groups will start to look more and more like a normal, bell-shaped curve when you plot them on a graph. This curve will be cantered around the true average colour of all the marbles in the jar.

So, the Central Limit Theorem says that if you take enough random samples from a larger group and calculate the average each time, those averages will form a normal distribution (the bell curve) even if the original data doesn't look like a bell curve at all. This is really useful because it allows us to make predictions about the larger group even if we only look at small samples.

2) What is Sampling and following techniques required to gather sample data?

Sampling is like picking a small portion of something to understand the whole. Imagine you have a big bowl of mixed fruit, and you want to know what types of fruit are in it without having to look at every single piece. Instead, you take a small handful, which is your sample, and study it. If your sample is chosen well, it should give you a pretty good idea of what's in the entire bowl.

Sampling Techniques

There are different ways to choose a sample, and these methods are called sampling techniques. Here are some common ones:

1. **Random Sampling**:
    - What it is - This is like closing your eyes and grabbing a handful of fruit. Every piece in the bowl has an equal chance of being picked.

- Why it’s good : It’s fair and gives every piece an equal shot at being chosen, which usually leads to a sample that represents the whole bowl well.

2. **Stratified Sampling**:
    - What it is - Suppose the bowl has 50% apples, 30% oranges, and 20% grapes. To make sure your sample represents the same proportions, you divide the bowl into these groups and pick a sample from each.
    - Why it’s good- It ensures that all types of fruit (or groups in your data) are fairly represented.

3) Describe the Hypothesis Testing and why do we conduct it?

Hypothesis Testing is like playing detective in the world of data. When you have a question or a claim about something, hypothesis testing helps you figure out whether that claim is likely to be true, based on the data you collect.

Reasons of conducting hypothesis

- ✓ **To Avoid Making Assumptions**: Instead of just guessing or assuming something is true, hypothesis testing allows us to test it with data.
- ✓ **To Make Better Decisions**: Whether in medicine, business, or social sciences, hypothesis testing helps ensure that decisions are based on evidence rather than intuition or guesswork.
- ✓ **To Understand Uncertainty**: Hypothesis testing quantifies uncertainty, helping us understand how confident we can be in our conclusions.
- ✓ **To Improve Practices**: By testing different approaches or products (like new teaching methods or new drugs), we can learn what works better and make improvements.

4) Define and briefly elaborate Central Tendency using measures with examples?

Central Tendency is a statistical term that refers to the canter or typical value of a dataset. It gives us an idea of where most of the data points are located. Think of it as the "average" or the "middle point" of your data, helping you understand what a typical value might be.

There are three main measures of central tendency: **Mean, Median, and Mode**. Let's explore each one with examples:

1. **Mean (Average)**

- **What it is**: The mean is what most people think of as the average. You calculate it by adding up all the numbers in your dataset and then dividing by the number of values.
- **Example**: Imagine you have the test scores of five students: 70, 85, 90, 75, and 80.
  - **Calculate**: (70 + 85 + 90 + 75 + 80) / 5 = 400 / 5 = 80.
  - **Mean**: The mean score is 80.
- **When to use**: The mean is useful when your data is evenly distributed, without extreme values (outliers).

2. **Median**

- **What it is**: The median is the middle value when your data is arranged in order. If there's an odd number of values, it's the one in the center. If there's an even number, you take the average of the two middle numbers.
- **Example**: Using the same test scores: 70, 75, 80, 85, and 90.
  - **Arrange**: The scores in order are already 70, 75, 80, 85, 90.
  - **Median**: The middle value is 80, so the median is 80.
- **When to use**: The median is useful when your data has outliers or is skewed, as it's not affected by extremely high or low values.
- 3. **Mode**

- **What it is**: The mode is the value that appears most frequently in your data. There can be more than one mode if multiple values have the same frequency.

- **Example**: Suppose we have these test scores: 70, 75, 80, 75, and 90.
    - **Mode**: The score 75 appears twice, while all other scores appear only once, so the mode is 75.
- **When to use**: The mode is useful when you want to know the most common value in your data. It's particularly helpful with categorical data (like the most popular color or most frequent customer complaint).

5) What's a ROC Curve? How do you differentiate Point Estimate and Confidence Interval Estimate?

A **ROC Curve** (Receiver Operating Characteristic Curve) is a graphical tool used to evaluate the performance of a classification model, especially in binary classification tasks (where there are two possible outcomes, like "spam" or "not spam").

How It Works:

- **True Positive Rate (TPR)**: Also called sensitivity or recall, this is the percentage of actual positives (e.g., real spam emails) that the model correctly identifies.
- **False Positive Rate (FPR)**: This is the percentage of actual negatives (e.g., non-spam emails) that the model incorrectly identifies as positives (spam).

The ROC curve plots the True Positive Rate (y-axis) against the False Positive Rate (x-axis) at different threshold settings. Each point on the curve represents a different decision threshold for the classification model.

Why It's Useful:

- The ROC Curve helps you see the trade-off between sensitivity and specificity.

- **Area Under the Curve (AUC)**: The area under the ROC curve (AUC) is a single number that summarizes the performance of the model. An AUC of 1 indicates a perfect model, while an AUC of 0.5 suggests a model no better than random guessing.

Point Estimate vs. Confidence Interval Estimate

These are two ways of estimating an unknown population parameter (like the average height of all adults in a city) based on sample data.

Point Estimate

- **What it is**: A point estimate gives you a single value as an estimate of the population parameter. For example, if you want to estimate the average height of adults in a city, you might take a sample and find that the average height in your sample is 170 cm. That 170 cm is your point estimate of the true average height of all adults in the city.
- **Example**: Suppose you measure the heights of 100 adults and find the average height to be 170 cm. The point estimate for the average height of all adults in the city is 170 cm.
- **Pros/Cons**: While simple and straightforward, a point estimate doesn't provide any information about how certain or uncertain you are about the estimate.

Confidence Interval Estimate

- **What it is**: A confidence interval estimate provides a range of values within which the true population parameter is likely to fall, along with a level of confidence. For example, instead of just saying the average height is 170 cm, you might say, "I am 95% confident that the average height of adults in this city is between 168 cm and 172 cm."
- **Example**: After measuring the heights, you calculate that you are 95% confident that the average height of all adults in the city is between 168 cm and 172 cm. This range (168 cm to 172 cm) is your confidence interval.
- **Pros/Cons**: Confidence intervals give a clearer picture of the uncertainty associated with an estimate, but they are more complex to calculate and interpret.

6) What's the difference between Correlation and Covariance in Statistics? What are the goals of A/B Testing?

Covariance

- **What it is**: Covariance measures the direction of the relationship between two variables. It tells us whether the variables tend to increase or decrease together.
    - **Positive Covariance**: If both variables tend to increase together, the covariance is positive.
    - **Negative Covariance**: If one variable tends to increase while the other decreases, the covariance is negative.
- **Example**: Suppose you measure the number of hours studied and the test scores for a group of students. If students who study more tend to score higher, the covariance will be positive.
- **Limitation**: The magnitude of covariance is hard to interpret because it's affected by the units of the variables. For example, if one variable is in dollars and the other in meters, the covariance will be in dollar-meters, which isn't very intuitive.

Correlation

- **What it is**: Correlation also measures the relationship between two variables, but it standardizes the measure, meaning it scales the covariance to a range between -1 and 1.
    - **Correlation of 1**: Perfect positive relationship – as one variable increases, the other always increases by a fixed proportion.
    - **Correlation of -1**: Perfect negative relationship – as one variable increases, the other always decreases by a fixed proportion.
    - **Correlation of 0**: No linear relationship – the variables don't move together in any consistent way.
- **Example**: If you calculate the correlation between hours studied and test scores, a value of 0.8 would indicate a strong positive relationship, meaning as study time increases, test scores tend to increase as well.

- **Advantage**: Correlation is easier to interpret because it's unitless, making it easier to compare relationships between different pairs of variables.

Goals of A/B Testing

A/B Testing is a method used to compare two versions of something to determine which one performs better. It's commonly used in business, marketing, and product development to make data-driven decisions.

Key Goals of A/B Testing:

1. **Improve User Experience**:
    - **What It Means**: By testing different versions of a webpage, app, or product feature, you can identify which one users prefer, leading to a better overall experience.
    - **Example**: Testing two different button colors on a website to see which one gets more clicks.
2. **Increase Conversion Rates**:
    - **What It Means**: A/B testing helps identify changes that can lead to higher conversion rates, such as more sign-ups, purchases, or clicks.
    - **Example**: Testing two different headlines on a landing page to see which one leads to more sales.
3. **Make Data-Driven Decisions**:
    - **What It Means**: Instead of relying on intuition or guesswork, A/B testing allows you to use real data to make decisions.
    - **Example**: Deciding whether to implement a new feature based on whether it performs better in an A/B test.
4. **Minimize Risk**:
    - **What It Means**: By testing changes on a small scale before rolling them out to everyone, you reduce the risk of negative impacts.
    - **Example**: Testing a new pricing strategy with a subset of customers before applying it across the board.
5. **Understand User Behavior**:

- **What It Means**: A/B testing provides insights into how users interact with different elements of your product or service.
- **Example**: Testing two different navigation menus to see which one users find easier to use.

7) How would you build and test a metric to compare two user's ranked lists of movie/tv show preferences?

To compare two users' ranked lists of movie/TV show preferences, you can build a metric that measures how similar or different their rankings are. A common approach for this is to use a **distance metric** that quantifies the difference between the two lists. Here's a step-by-step guide on how you could build and test such a metric:

1. **Choose a Distance Metric**

Several distance metrics can be used to compare ranked lists. Here are a couple of popular ones:

Kendall Tau Distance:

- **What it is**: This metric counts the number of pairwise disagreements between two ranked lists. A pairwise disagreement occurs when two items are ranked in a different order by the two users.
- **How to calculate**:
  - For example, if User A ranks movies as [A, B, C] and User B ranks them as [B, A, C], Kendall Tau distance counts the number of inversions needed to convert one list into the other. Here, A and B are inverted, so the distance is 1.

Spearman's Rank Correlation Coefficient:

- **What it is**: This is a measure of how well the relationship between two ranked lists can be described using a monotonic function. It's calculated by taking the Pearson correlation coefficient between the ranks of the two lists.
- **How to calculate**:
    - Assign ranks to the items in both lists. For example, if User A ranks movies as [A, B, C] and User B ranks them as [B, C, A], assign ranks (1, 2, 3) to both lists based on their order.
    - Calculate the difference between the ranks for each item and use this to compute the Spearman correlation coefficient.

8) Find the probability of P(x<400) given that mean is = 1000 variance is =100

To find the probability P(X<400)P(X < 400)P(X<400) given that the mean μ\muμ is 1000 and the variance σ2\sigma^2σ2 is 100, we can use the properties of the normal distribution.

Steps:

1. **Identify the Parameters**:
    - Mean (μ\muμ) = 1000
    - Variance (σ2\sigma^2σ2) = 100
    - Standard deviation (σ\sigmaσ) = 100=10\sqrt{100} = 10100=10
2. **Standardize the Variable**:

- Convert the problem to the standard normal distribution (which has a mean of 0 and a standard deviation of 1) using the Z-score formula: $Z=X-\mu\sigma Z = \frac{X - \mu}{\sigma}Z=\sigma X-\mu$
- Here, $X=400X = 400X=400$, $\mu=1000\mu = 1000\mu=1000$, and $\sigma=10\sigma = 10\sigma=10$. $Z=400-100010=-60010=-60Z = \frac{400 - 1000}{10} = \frac{-600}{10} = -60Z=10400-1000=10-600=-60$

3. **Find the Probability**:
   - The probability $P(X<400)P(X < 400)P(X<400)$ is equivalent to finding $P(Z<-60)P(Z < -60)P(Z<-60)$.
   - $Z=-60Z = -60Z=-60$ is an extremely low value. For the standard normal distribution, the probability of $Z<-60Z < -60Z<-60$ is so close to 0 that it can be considered negligible.

Conclusion:

The probability $P(X<400)P(X < 400)P(X<400)$ is approximately **0** given the mean is 1000 and the variance is 100. This means it is extremely unlikely for XXX to be less than 400 in this distribution.

9 Provide a proper description as to when to apply Ztest, t-test, Chi-Square and annova, with examples.

. **Z-test**

When to Use:

Use a **Z-test** when you want to compare the means of a sample to a known population mean, and the sample size is large (typically n>30n > 30n>30).

- The population variance should be known, or the sample size should be large enough for the sample variance to approximate the population variance.

Example:

- **Scenario**: A factory claims that the average weight of a product is 500 grams. You take a sample of 50 products and find an average weight of 495 grams. You know from past records that the standard deviation of the weights is 10 grams.
- **Question**: Is the observed average weight of 495 grams significantly different from the claimed average of 500 grams?
- **Test**: Perform a Z-test to determine if the difference between the sample mean (495 grams) and the population mean (500 grams) is statistically significant.

2. **t-test**

When to Use:

- Use a **t-test** when comparing the means of one or two groups, especially when the sample size is small (n≤30n \leq 30n≤30) or when the population variance is unknown.

Types:

- **One-sample t-test**: Compare the mean of a single sample to a known value.
- **Independent two-sample t-test**: Compare the means of two independent groups.
- **Paired t-test**: Compare the means of two related groups (e.g., before and after measurements on the same subjects).

Examples:

- **One-sample t-test**: A professor claims that students in her class score an average of 75 on a test. You take a sample of 20 students and find an average score of 70. You use a one-sample t-test to see if the sample mean is significantly different from 75.
- **Independent two-sample t-test**: You want to compare the test scores of two different classes, each with 15 students. You use an independent t-test to see if there's a significant difference in average scores between the two classes.
- **Paired t-test**: You measure the blood pressure of 10 patients before and after administering a drug. You use a paired t-test to see if the drug caused a significant change in blood pressure.

3. **Chi-Square Test**

When to Use:

- Use a **Chi-Square test** to test relationships between categorical variables. It's used to determine whether there is a significant association between two categorical variables or to test if a single categorical variable fits an expected distribution.

Types:

- **Chi-Square Test for Independence**: Test whether two categorical variables are independent.
- **Chi-Square Goodness of Fit Test**: Test whether the distribution of a single categorical variable matches an expected distribution.

Examples:

- **Chi-Square Test for Independence**: You want to know if there's an association between gender (male/female) and preference for a new product (like/dislike). You collect data from a sample and use a Chi-Square test to see if gender and product preference are independent of each other.
- **Chi-Square Goodness of Fit Test**: You want to see if a dice is fair, meaning that each side should come up about the same number of times when rolled. You roll the dice 60 times, record the results, and use the Chi-Square Goodness of Fit test to see if the observed frequencies match the expected frequencies (10 for each side if the dice is fair).

4. **ANOVA (Analysis of Variance)**

When to Use:

- Use **ANOVA** when comparing the means of three or more groups. It tests whether at least one of the group means is significantly different from the others.

Types:

- **One-way ANOVA**: Compare means across one factor (e.g., three different teaching methods).
- **Two-way ANOVA**: Compare means across two factors (e.g., teaching methods and gender).

Examples:

- **One-way ANOVA**: You want to compare the effectiveness of three different diets on weight loss. You divide 60 participants into three groups, each following a different diet, and measure their weight loss after 8 weeks. You use one-way ANOVA to see if there is a significant difference in weight loss between the three diet groups.
- **Two-way ANOVA**: You want to study the effects of two different teaching methods and gender on test scores. You use two-way ANOVA to see if the teaching method and gender, or an interaction between them, significantly affect test scores.

Summary

- **Z-test**: Compare a sample mean to a known population mean (large sample size, known variance).
- **t-test**: Compare means between one or two groups (small sample size, unknown variance).
- **Chi-Square test**: Test relationships between categorical variables or fit of a categorical variable to an expected distribution.
- **ANOVA**: Compare means among three or more groups to see if at least one group differs.