Task1: Types of data distribution

1. **Normal Distribution (Gaussian Distribution):** Also known as the bell curve, this distribution is characterized by a symmetrical, bell-shaped curve. In a normal distribution, the mean, median, and mode are all equal, and data points are evenly distributed around this central value.

2. **Uniform Distribution:** In a uniform distribution, all data points have equal probabilities of occurring. It forms a rectangle-shaped distribution, where every value within a certain range is equally likely.

3. **Exponential Distribution:** This distribution is often used to model the time between events in a Poisson process (events that occur randomly over time). It is characterized by a rapidly decreasing probability density function and is skewed to the right.

4. **Log-Normal Distribution:** The log-normal distribution is skewed to the right and is used to describe data that is not symmetric in its original form but becomes more symmetric when taking the natural logarithm of the data.

5. **Poisson Distribution:** The Poisson distribution is used to model the number of events occurring within a fixed interval of time or space. It is often applied to rare events and is characterized by a single parameter, λ (lambda), which represents the average rate of event occurrences.

6. **Binomial Distribution:** The binomial distribution models the number of successful outcomes in a fixed number of trials, where each trial has only two possible outcomes (success or failure). It is characterized by two parameters: the probability of success (p) and the number of trials (n).

7. **Bernoulli Distribution:** A special case of the binomial distribution where there is only one trial (n=1). It represents the probability of success (p) in a single trial.

8. **Geometric Distribution:** This distribution models the number of trials needed to achieve the first success in a sequence of Bernoulli trials, where each trial has a probability of success (p).

9. **Multinomial Distribution:** This distribution generalizes the binomial distribution to more than two categories or outcomes. It models the probability of observing each category in a series of independent trials.

10. **Negative Binomial Distribution:** Like the geometric distribution, the negative binomial distribution models the number of trials needed to achieve a fixed number of successes (r) in a sequence of Bernoulli trials with a probability of success (p).

11. **Chi-Square Distribution:** The chi-square distribution is commonly used in hypothesis testing and confidence interval calculations. It arises in the context of testing the goodness of fit, independence of variables, and estimating population variances.

12. **Student's t-Distribution:** The t-distribution is used for hypothesis testing when the sample size is small, and the population standard deviation is unknown. It is like the normal distribution but has heavier tails.

13. **F-Distribution:** The F-distribution is used for comparing variances or testing the equality of means from two or more populations. It commonly arises in analysis of variance (ANOVA) and regression analysis.

Task 2: Central limit theory

The Central Limit Theorem (CLT) is a fundamental concept in statistics that describes the distribution of sample means. It states that, regardless of the shape of the population distribution, the distribution of the sample means will tend to follow a normal distribution as the sample size increases, assuming certain conditions are met. The Central Limit Theorem has several key points:

1. **Population Distribution:** The CLT doesn't make any assumptions about the shape of the population distribution. It can be normal, non-normal, or any other distribution.
2. **Random Sampling:** The samples must be drawn randomly from the population. This means that each observation in the population has an equal chance of being selected.
3. **Independence:** The observations in the sample must be independent of each other. In other words, the value of one observation should not influence the value of another.
4. **Sample Size:** As the sample size (n) increases, the distribution of sample means approaches a normal distribution, even if the population distribution is not normal.
5. **Mean and Variance:** The mean of the sample means will be approximately equal to the mean of the population, and the variance of the sample means will be the population variance divided by the sample size.

Task3: Distribution and density

Probability Distribution:

A probability distribution describes how the values of a random variable are distributed. It specifies the possible values that the random variable can take on and assigns probabilities to each of those values. In essence, it provides a complete description of the probabilities associated with each outcome.

Example: Consider the roll of a fair six-sided die. The probability distribution for this random variable is:

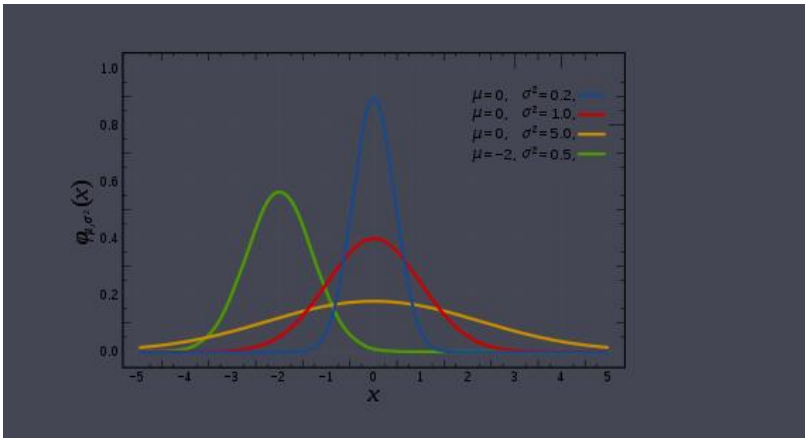
- P(X = 1) = 1/6
- P(X = 2) = 1/6
- P(X = 3) = 1/6
- P(X = 4) = 1/6
- P(X = 5) = 1/6
- P(X = 6) = 1/6

This probability distribution tells us that each number from 1 to 6 has a 1/6 probability of occurring when we roll the die.

Probability Density Function (PDF):

A probability density function, on the other hand, is typically used when dealing with continuous random variables. It describes the likelihood of a continuous random variable falling within a particular range of values. Unlike a discrete probability distribution, a PDF doesn't assign probabilities to individual values but rather provides a relative likelihood of the variable taking on a range of values.

Example: Consider the height of adult humans, which is a continuous random variable. A normal distribution (bell curve) is often used to model human heights. The probability density function for height might look something like this:



In this case, the PDF doesn't tell you the probability of a person being exactly a certain height (e.g., 175 cm), but it tells you the probability density or likelihood of a person's height falling within a range of values (e.g., between 170 cm and 180 cm)

Task4: P-value

A p-value, short for "probability value," is a statistical measure that helps us determine the strength of evidence against a null hypothesis in a hypothesis test. It quantifies the probability of obtaining observed results (or more extreme results) when the null hypothesis is true. In other words, it tells us how likely it is to see the observed data if there is no real effect or difference in the population.

Here's an example to illustrate the concept of a p-value:

Example: Hypothesis Testing for a Coin Toss

Suppose you have a coin, and you want to test whether it's a fair coin (i.e., it has an equal probability of landing heads or tails) or if it's biased in some way. You set up the following hypotheses:

- **Null Hypothesis (H0):** The coin is fair, and the probability of getting heads (H) is 0.5, i.e., P(H) = 0.5.
- **Alternative Hypothesis (H1):** The coin is biased, and the probability of getting heads is not 0.5, i.e., P(H) ≠ 0.5.

Next, you conduct an experiment in which you flip the coin 100 times, and you observe that it lands heads 60 times and tails 40 times.

Now, you want to use a statistical test (such as a two-sample proportion test or a chi-square test for goodness of fit) to determine if the results are significantly different from what you'd expect with a fair coin.

After performing the test, you calculate a p-value. Let's say you find that the p-value is 0.03.

Interpretation of the p-value:

- If the p-value is small (typically less than a pre-defined significance level, often denoted as α, e.g., α = 0.05), it suggests that the observed data is unlikely to have occurred by random

chance alone when the null hypothesis is true. In this case, with a p-value of 0.03, you might conclude that the evidence against the null hypothesis is strong, and you reject the null hypothesis in Favor of the alternative. This means you believe the coin is likely biased.

- If the p-value is large (greater than α), it suggests that the observed data is consistent with what you'd expect under the null hypothesis. In this case, you wouldn't have enough evidence to reject the null hypothesis, and you would not conclude that the coin is biased.

So, in the example above, a p-value of 0.03 suggests that the observed coin flips are unlikely to be due to chance alone if the coin were fair, leading you to conclude that there is evidence to suggest the coin is biased.

In hypothesis testing, the choice of the significance level (α) is important. It determines the threshold for considering a p-value as statistically significant. Common values for α include 0.05 and 0.01, but the choice depends on the specific context and the level of confidence required for the test.

Task5: Statistical test

Statistical tests are used in hypothesis testing to make decisions about population parameters based on sample data. They help determine if there is enough evidence to support or reject a specific hypothesis. Here are some common statistical tests and an example for each:

t-Test:

Example: Suppose you want to test if there is a significant difference in the mean test scores of two groups (Group A and Group B). You collect test scores from both groups and use a t-test to compare the means. The null hypothesis (H0) might be that there is no difference ($\mu A = \mu B$), and the alternative hypothesis (H1) is that there is a difference ($\mu A \neq \mu B$). You perform the t-test and obtain a p-value. If the p-value is less than your chosen significance level (e.g., 0.05), you may conclude that there is a statistically significant difference between the two groups.

Chi-Square Test:

Example: You want to determine if there is an association between two categorical variables, such as gender (male/female) and smoking status (smoker/non-smoker). You collect data on a sample of individuals and create a contingency table. Then, you perform a chi-square test of independence. The null hypothesis (H0) might be that there is no association between gender and smoking status, while the alternative hypothesis (H1) is that there is an association. If the chi-square test yields a p-value less than your chosen significance level, you may conclude that there is a significant association.

ANOVA (Analysis of Variance):

Example: You have data on the exam scores of students who took three different types of prep courses (A, B, and C). You want to determine if there is a statistically significant difference in mean exam scores among these groups. You perform a one-way ANOVA test. The null hypothesis (H0) is that there is no difference in mean scores among the groups ($\mu A = \mu B = \mu C$), while the alternative

hypothesis (H1) is that there is a difference. If the ANOVA test yields a p-value less than your chosen significance level, you may conclude that at least one of the groups has a different mean score.

Regression Analysis:

Example: You want to understand the relationship between a dependent variable (e.g., sales) and one or more independent variables (e.g., advertising spending and seasonality). You perform linear regression analysis. The null hypothesis (H0) might be that there is no significant linear relationship between the variables, while the alternative hypothesis (H1) is that a significant relationship exists. If the regression analysis indicates a statistically significant relationship (e.g., p-value < 0.05), you may conclude that changes in the independent variables are associated with changes in the dependent variable.

Wilcoxon Rank-Sum Test (Mann-Whitney U Test):

Example: You want to compare two independent groups' median values for a non-normally distributed variable, such as income. You collect data on income for two groups (e.g., employees of Company A and Company B) and perform a Wilcoxon rank-sum test. The null hypothesis (H0) is that there is no difference in the median income between the two groups, while the alternative hypothesis (H1) is that there is a difference. If the test results in a p-value less than your chosen significance level, you may conclude that there is a significant difference in median income between the two groups.

Task 6: Difference between Naïve Bayes and conditional probability?

Naive Bayes and conditional probability are related concepts used in probability theory and machine learning, but they are distinct in their applications and assumptions. Here's an explanation of the key differences between them:

Naive Bayes:

Nature: Naive Bayes is a classification algorithm used in machine learning and statistics.

Assumption: It assumes that the features (variables) used in the classification are conditionally independent given the class label. This is a simplifying and often unrealistic assumption, which is why it's called "naive." Features may be correlated, but Naive Bayes treats them as if they are independent.

Use: Naive Bayes is commonly used for text classification tasks, such as spam detection or sentiment analysis. It's also used in various other classification problems where the independence assumption is approximately valid or when simplicity and speed are priorities.

Bayesian Framework: Naive Bayes is based on Bayes' theorem and uses conditional probabilities to calculate the probability of a data point belonging to a particular class given its features.

Conditional Probability:

Nature: Conditional probability is a fundamental concept in probability theory and statistics.

Assumption: It does not assume independence between variables. Instead, it calculates the probability of an event occurring given that another event has already occurred.

Use: Conditional probability is used to model dependencies between events or variables. It is widely used in various statistical analyses, including hypothesis testing, Bayesian statistics, and Markov chains.

Example: If you want to calculate the probability of a student passing an exam given that they attended a study session, you are calculating a conditional probability. It answers the question, "What is the probability of event A happening given that event B has occurred?"