#UpSkillWithKalpesh

Day 07

Data Science Unlocked

From Zero to Data Hero

Advanced Statistics for Data Science

Kalpesh Pathade
@DataSimplified

Advanced Statistics

I. Correlation and Causality

Causality vs. Correlation

- **Correlation**: Refers to a statistical association between two variables. If two variables are correlated, it means that changes in one variable are associated with changes in the other variable.
    - **Example**: There is a correlation between ice cream sales and the temperature outside.
- **Causality**: Implies that one variable directly influences the other. In other words, a change in one variable causes a change in the other.
    - **Example**: Smoking causes lung cancer.
- **Key Differences**:
    - **Correlation** does not imply causation. Two variables can be correlated without one causing the other.
    - **Causality** implies a directional influence, and typically, more rigorous experimental or statistical methods are needed to establish causality (e.g., randomized controlled trials or statistical models like Granger causality).

Pearson's Correlation Coefficient

- **Definition**: Pearson's Correlation Coefficient (denoted as r) is a statistical measure that calculates the strength and direction of the linear relationship between two variables.
- Formula

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

- n = number of data points
- x = values of the first variable
- y = values of the second variable
- **Range**: The value of r ranges from -1 to 1:
  - r=11: Perfect positive correlation
  - r=−: Perfect negative correlation
  - r=0: No linear correlation
- **Assumptions**:
  - Both variables should be continuous and normally distributed.
  - The relationship should be linear.
- **Use**: Pearson's correlation is used to quantify the degree to which two variables are related in a linear fashion.

Spearman's Rank Correlation

- **Definition**: Spearman's Rank Correlation (denoted as) is a non-parametric measure that assesses how well the relationship between two variables can be described using a monotonic function.

  ρ\rho
- **Formula**:

$$\rho = \rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Where:

    - d is the difference between the ranks of the corresponding variables.
    - n is the number of data points.
- **Assumptions**:
    - This method does not require the data to be normally distributed.
    - It is used when the relationship between variables is monotonic (either increasing or decreasing).
- **Use**: It is useful when the relationship between variables is not linear, as it can capture both linear and non-linear associations.

Granger Causality Test

- **Definition**: The Granger Causality Test is a statistical hypothesis test used to determine whether one time series can predict another time series. It assesses whether past values of one time series provide information about future values of another time series.
- **Assumptions**:
    - Time series data should be stationary (i.e., mean and variance are constant over time).
    - The test involves analyzing lagged values of the time series.
- **Methodology**:
    - The null hypothesis for the test is that variable X does not Granger cause Y.
    - If the p-value is small (typically less than 0.05), we reject the null hypothesis and conclude that X Granger causes Y.
- **Use**: Granger causality is useful for understanding temporal relationships in time series data, such as stock prices, economic indicators, or sales figures.

II. Bayesian Statistics

1. Introduction to Bayesian Inference

- **Bayesian Inference** is a method of statistical inference in which Bayes' Theorem is used to update the probability estimate for a hypothesis as more evidence or information becomes available.
- Unlike frequentist methods, which rely on fixed parameters, Bayesian inference treats unknown parameters as random variables and incorporates prior knowledge into the analysis.
- The central concept is that we combine our prior beliefs with observed data to update our understanding of a hypothesis or parameter.
- **Key Concepts**:
  - **Prior Knowledge**: Information available before observing the data.
  - **Likelihood**: The probability of observing the data given a certain hypothesis.
  - **Posterior**: The updated probability of the hypothesis after observing the data.

---

2. Bayes' Theorem and its Applications

- **Bayes' Theorem** provides a way to calculate the posterior probability of an event, given the prior knowledge and likelihood of the event.
- **Formula**:

$$P(H \mid D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Where:

- P(H|D) is the posterior probability (the probability of hypothesis H after observing data D).

    - P(D|H) is the likelihood (the probability of data D given the hypothesis H).
    - P(H) is the prior probability (the initial belief about H).
    - P(D) is the marginal likelihood (the total probability of observing data D).
- **Applications of Bayes' Theorem**:
    - **Medical Diagnosis**: Updating the probability of a disease given test results.
    - **Spam Filtering**: Classifying emails as spam or not spam based on prior probabilities and observed features.
    - **Machine Learning**: In algorithms like Naive Bayes classifiers.
    - **A/B Testing**: Updating beliefs about the effectiveness of different versions of a website or product.

3. Prior, Likelihood, Posterior Distributions

- **Prior Distribution**: Represents what is known about a parameter before any data is observed. The prior can be based on previous studies, expert knowledge, or even subjective belief.
    - **Types of Priors**:
        - **Informative Prior**: Contains strong prior knowledge about the parameter.
        - **Non-informative (or flat) Prior**: Assumes little or no prior knowledge.
- **Likelihood**: Represents the likelihood of the data given a particular value of the parameter. It reflects how well the parameter fits the observed data.
- **Posterior Distribution**: After observing the data, the posterior distribution represents the updated beliefs about the parameter. The posterior combines both the prior and the likelihood through Bayes' Theorem.
    - **Formula**:

$$P(\theta \mid D) = (\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

Where:

- P(θ|D) is the posterior distribution of the parameter θ.
- P(D|θ) is the likelihood.
- P(θ) is the prior.
- P(D) is the marginal likelihood or evidence.

- **Conjugate Priors**: In some cases, the prior and posterior distributions belong to the same family of distributions, which simplifies the computation. These are known as conjugate priors.
  - **Example**: In Bayesian updating for a binomial likelihood (e.g., coin flipping), using a Beta prior results in a Beta posterior distribution.

4. Markov Chain Monte Carlo (MCMC)

- **Markov Chain Monte Carlo (MCMC)** is a class of algorithms used to sample from complex probability distributions. It is particularly useful in Bayesian statistics when the posterior distribution is difficult to compute directly.
- MCMC methods generate a sequence of samples from the posterior distribution by constructing a Markov chain whose stationary distribution is the desired posterior.
- **Key Concepts**:
  - **Markov Chain**: A sequence of random variables where the future state depends only on the current state and not on the past states.
  - **Monte Carlo**: Refers to the use of random sampling to estimate complex probabilities or expectations.
- **Popular MCMC Algorithms**:
  - **Metropolis-Hastings Algorithm**: A random walk method that generates new samples based on a proposal distribution.
  - **Gibbs Sampling**: A specific MCMC method that samples from the conditional distribution of each parameter, iteratively updating them.
- **Applications of MCMC**:

  - **Bayesian Inference**: MCMC allows sampling from posterior distributions when direct calculation is infeasible.
  - **Computational Statistics**: Used in a wide range of fields, including physics, biology, economics, and machine learning.
- **Advantages**:
  - Provides samples from the posterior distribution, which can be used for estimation and uncertainty quantification.
  - MCMC can be applied to very complex models where exact analytical solutions are not possible.

III. Non-parametric Tests

1. Wilcoxon Rank-Sum Test

- **Definition**: The Wilcoxon Rank-Sum Test (also known as the Mann-Whitney U Test) is a non-parametric test used to compare two independent groups to assess whether they come from the same distribution. It is used when the data does not follow a normal distribution.
- **Use Case**: It is typically used as an alternative to the independent t-test when the assumption of normality is violated.
- **Test Statistic**: The test ranks all the values from both groups together, then compares the sum of ranks for each group. The U statistic is calculated, and based on its distribution, the significance of the difference is assessed.
- **Hypothesis**:
  - **Null Hypothesis**: The two groups have the same distribution.
  - **Alternative Hypothesis**: The two groups have different distributions.
- **Assumptions**:
  - The two groups are independent.
  - The dependent variable is ordinal or continuous, and the data does not need to be normally distributed.

2. Kruskal-Wallis Test

- **Definition**: The Kruskal-Wallis Test is a non-parametric test used to compare three or more independent groups to determine if they come from the same distribution.
- **Use Case**: It is often used as an alternative to one-way ANOVA when the assumptions of normality or equal variances are not met.
- **Test Statistic**: The Kruskal-Wallis test ranks all the values from all groups together, then compares the average rank between the groups. The test statistic follows a chi-square distribution.
- **Hypothesis**:
  - **Null Hypothesis**: All groups have the same distribution.
  - **Alternative Hypothesis**: At least one group has a different distribution.
- **Assumptions**:
  - The groups are independent.
  - The dependent variable is ordinal or continuous, and the data does not need to follow a normal distribution.
  - The distributions of the groups have the same shape.

---

3. Friedman Test

- **Definition**: The Friedman Test is a non-parametric test used to compare three or more related (paired) groups. It is an alternative to the repeated measures ANOVA when the assumptions of normality are not met.
- **Use Case**: It is used when the data is from repeated measures or matched subjects, where you are comparing multiple conditions within the same group.
- **Test Statistic**: The Friedman test ranks the data within each block (group), and then compares the sum of ranks across the different conditions. The test statistic follows a chi-square distribution.
- **Hypothesis**:
  - **Null Hypothesis**: The distributions of all groups are the same.

- **Alternative Hypothesis**: At least one of the groups has a different distribution.

- **Assumptions**:
  - The data is ordinal or continuous.
  - The observations are paired or matched.
  - The distributions of the groups are not required to be normal.

4. Kolmogorov-Smirnov Test

- **Definition**: The Kolmogorov-Smirnov Test is a non-parametric test used to compare a sample with a reference probability distribution, or to compare two samples to determine if they come from the same distribution.
- **Use Case**: It is used to test the goodness-of-fit of a sample to a known distribution or to compare two independent samples.
- **Test Statistic**: The test compares the empirical cumulative distribution function (ECDF) of the sample to the cumulative distribution function (CDF) of the reference distribution (in the one-sample case), or compares the ECDFs of two samples.
- **Hypothesis**:
  - **Null Hypothesis**: The sample comes from the specified distribution (for one-sample) or the two samples are from the same distribution (for two-sample).
  - **Alternative Hypothesis**: The sample does not come from the specified distribution (for one-sample) or the two samples are from different distributions (for two-sample).
- **Assumptions**:
  - The data is continuous.
  - For the two-sample case, the samples are independent.
  - The Kolmogorov-Smirnov test does not require the assumption of normality.

IV. Advanced Probability Theory

1. Markov Chains

- **Definition**: A **Markov Chain** is a sequence of random variables where the future state depends only on the current state, and not on the sequence of events that preceded it. This property is known as the **Markov Property**.
- **Key Concepts**:
  - **State Space**: The set of all possible states of the system.
  - **Transition Matrix**: A matrix that defines the probabilities of moving from one state to another. Each entry p_{ij} in the matrix represents the

    probability of transitioning from state i to state j.
  - **Stationary Distribution**: A distribution over states that remains unchanged under the transition probabilities of the chain. If π is the stationary distribution, then:

    $\pi = \pi P$
  - **Absorbing Markov Chain**: A Markov Chain where some states, once entered, cannot be left.
- **Applications**:
  - **Queueing Theory**: Modeling customers arriving at a service point.
  - **PageRank Algorithm**: Used by Google to rank web pages.
  - **Weather Prediction**: Predicting future weather conditions based on current state.
- **Types of Markov Chains**:
  - **Discrete-time Markov Chain (DTMC)**: The process evolves in discrete time steps.

    - **Continuous-time Markov Chain (CTMC)**: The process evolves continuously, often used to model systems with continuous time.

2. Stochastic Processes

- **Definition**: A **Stochastic Process** is a collection of random variables indexed by time or space. It represents a system that evolves randomly over time or space.
- **Types**:
    - **Discrete-time Process**: The random variables are indexed by discrete time points.
    - **Continuous-time Process**: The random variables are indexed by continuous time.
- **Key Concepts**:
    - **State Space**: The set of all possible values of the random variables.
    - **Sample Path**: The realization of the stochastic process over time or space.
    - **Stationarity**: A stochastic process is stationary if its statistical properties do not change over time.
- **Applications**:
    - **Finance**: Modeling stock prices (e.g., Brownian motion).
    - **Queuing Theory**: Describing systems like server queues.
    - **Signal Processing**: Modeling noise in communication systems.

3. Random Walks

- **Definition**: A **Random Walk** is a type of stochastic process where an entity (such as a particle or a person) moves step by step in random directions. The direction of movement is determined by a random process, often taking steps in either a positive or negative direction.
- The mathematical formulation of a **simple random walk** can be represented as:

$$Xn = X_{n-1} + \epsilon_n$$

Where:

- Xn is the position of the entity after n steps.
- Xn−1 is the position of the entity after the previous step.
- ϵn is a random variable representing the step taken at the nth move. Typically, ϵn is either +1 or −1 with equal probability (i.e., a Bernoulli distribution).
- **Key Concepts**:
  - **Symmetric Random Walk**: Each step has an equal probability of moving in either direction.
  - **Drunkard's Walk**: A famous random walk where a person is imagined to take steps in random directions, commonly used to model diffusion processes.
  - **Absorbing Random Walk**: A random walk where the process is "absorbed" at a certain state (e.g., reaching a boundary).
- **Applications**:
  - **Physics**: Modeling diffusion and particle movement.
  - **Economics**: Describing random fluctuations in stock prices.
  - **Biology**: Modeling the movement of animals or microorganisms.

---

4. Poisson Processes

- **Definition**: A **Poisson Process** is a type of stochastic process that models the occurrence of events randomly in time or space, where the events occur independently, and the rate of occurrence is constant over time.
- **Key Properties**:
  - **Independent Increments**: The number of events occurring in non-overlapping intervals is independent.

- **Stationary Increments**: The probability of a given number of events occurring in a fixed time interval depends only on the length of the interval and not on its position in time.
- **Exponential Inter-Event Times**: The time between events follows an exponential distribution.

- **Mathematical Formulation**: The number of events that occur in a time interval of length follows a Poisson distribution:

$$P\,(N(t)\;=\;k)\;=\;\frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

Where:

- λlambda is the rate of events per unit time.
- t is the time interval.
- k is the number of events.

- **Applications**:
  - **Queueing Theory**: Modeling arrival of customers in a system.
  - **Telecommunications**: Modeling the arrival of phone calls or network packets.
  - **Reliability Engineering**: Modeling the failure of components in a system.
- **Types of Poisson Processes**:
  - **Homogeneous Poisson Process**: The rate is constant over time.

    λ
  - **Non-homogeneous Poisson Process**: The rate varies over time, often used to model events that have a varying intensity.

    λ