knowledgehut
upGrad

Deep Dive into Probability Theory & Probability Statistics

Exploring Statistical Foundations in Data Science

By the End of this Session:

- Define Bayes' theorem and conditional probability.
- Differentiate between independent and mutually exclusive events.
- Demonstrate a comprehensive understanding of the Central Limit Theorem.
- Assess the statistical significance of distribution curves.
- Communicate probability concepts effectively through oral and written means.
- Apply probability concepts to real-world problems and scenarios.

Probability

"Probability theory is nothing but common sense reduced to calculation."

-Pierre-Simon Laplace

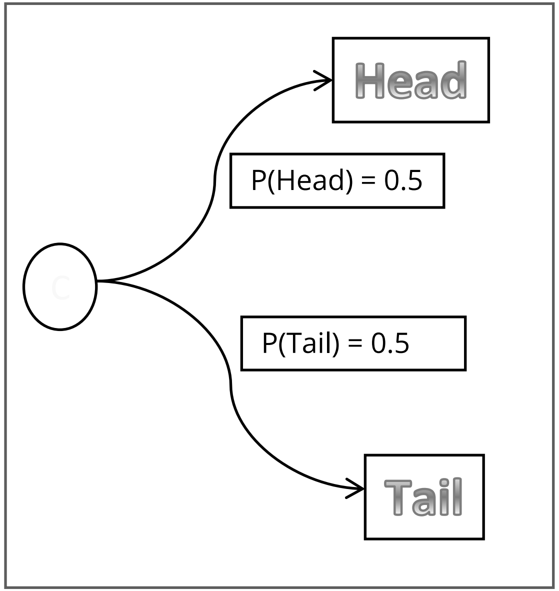
Probability is a measure of the likelihood of an event occurring.

It is a number between 0 and 1, where 0 means the event is impossible and 1 means the event is certain.

The probability of an event is calculated by dividing the number of outcomes that favor the event by the total number of possible outcomes.

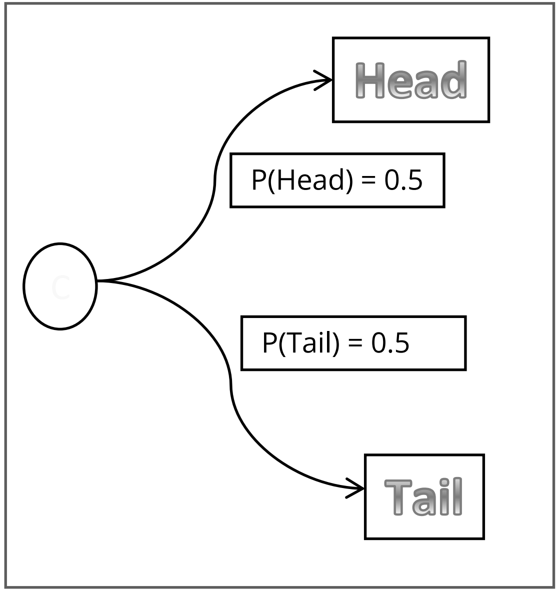
Components of Probability

- **Sample Space (S):** The sample space is the set of all possible outcomes of the experiment.
- **Event (E):** An event is the subset of Sample space.
- **Probability (P):** The probability of an Event (E) is the measure of how likely it is to occur.



Components of Probability

- For Example: Let's consider an Experiment of tossing a coin.
- If you toss a coin one time, the sample space is the set of all possible outcomes that can appear: Head and Tail.
- Number of Samples = 2
- Let's consider the Event to be the occurrence of Head for the given experiment.
- P(Head) = Number of favorable outcomes/Total number of outcomes = ½ = 0.5



Pop Quiz

What is Probability?

a. A measure of how likely an event is to occur

b. A measure of how unlikely an event is to occur

c. A measure of how certain an event is to occur

d. A measure of how impossible an event is to occur

Pop Quiz

What is Probability?

a. **A measure of how likely an event is to occur**

b. A measure of how unlikely an event is to occur

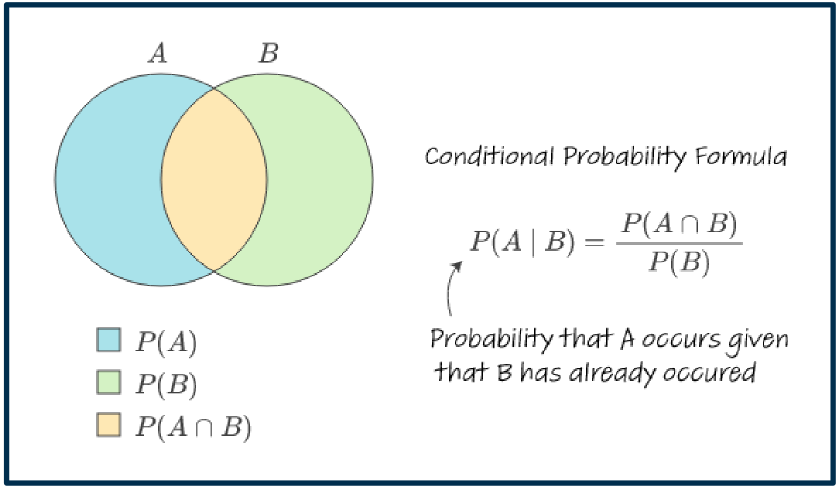
c. A measure of how certain an event is to occur

d. A measure of how impossible an event is to occur

Conditional Probability

Conditional probability is the probability of an event occurring, given that another event has already occurred.

It is denoted by P(A|B), where A is the event we are trying to find the probability of, and B is the event that has already occurred.



Independent Events

Independent events are events where the occurrence or non-occurrence of one event does not affect the probability of the other event happening.

Example:

Consider tossing a fair coin and rolling a fair six-sided dice. The outcome of the coin toss (heads or tails) does not affect the outcome of rolling the dice (getting any of the six numbers).

Example

Let’s assume we are rolling a dice, and we want to find the probability of getting a 3, given that the first roll was a 6.

Solution:
A is the event of getting a 3.
B is the event of getting a 6.

Note: Here, the occurrence of 3 on the second roll will not get affected by the first roll.

$P(A|B) = P(A \cap B) / P(B) = P(A) * P(B) / P(B) = P(A) = 1/6$

Where,
$P(A|B)$ is the probability of A given B.
$P(A \cap B)$ is the probability of A and B occurring together.
$P(B)$ is the probability of B occurring.

Pop Quiz

Q. In a deck of playing cards, there are 52 cards with 4 different suits: hearts, diamonds, clubs, and spades. If you randomly draw one card from the deck without replacement, what is the probability that the second card you draw is a heart, given that the first card drawn was a diamond?

A. 1/4

B. 1/13

C. 1/26

D. 13/51

Pop Quiz

Q. In a deck of playing cards, there are 52 cards with 4 different suits: hearts, diamonds, clubs, and spades. If you randomly draw one card from the deck without replacement, what is the probability that the second card you draw is a heart, given that the first card drawn was a diamond?

A. 1/4

B. 1/13

C. **1/26**

D. 13/51

Dependent Events

Dependent events are events where the occurrence or non-occurrence of one event affects the probability of the other event happening.

Example:

Weather Conditions and Outdoor Activities.

The probability of having a rainy day affects the probability of going for a picnic.

Example

Let's assume that we have to select 2 red balls from a bag containing 10 red and 10 blue balls. (Without replacement)

Solution:
Total Number of Balls = 10 + 10 = 20

Let A be the event of selecting 1st red ball and B be the event of selecting 2nd red ball.

P(A) = 10/20
P(B) = 9/19

$P(A \cap B) = P(A) * P(B|A) = 10/20 * 9/19 = 0.23$

Mutually Exclusive Events

Two events are said to be mutually exclusive if they cannot occur at the same time.

Example:

Flipping a coin and getting heads is mutually exclusive with flipping a coin and getting tails.

Example

What is the probability of a dice showing the number 1 or 6?

Solution:

P(1) is the probability of getting a number 3

P(6) is the probability of getting the number 5

P(1) = 1/6 and P(6) = 1/6

P(1 or 6) = P(3) + P(5)

P(1 or 6) = (1/6) + (1/6) = 2/6

P(1 or 6) = 1/3

Poll Time

What will be the probability of getting a 3 on a dice, when it is rolled one time?

a. 1/6

b. 2/6

c. 3/6

d. 4/6

Poll Time

What will be the probability of getting a 3 on a dice, when it is rolled one time?

a. **1/6**

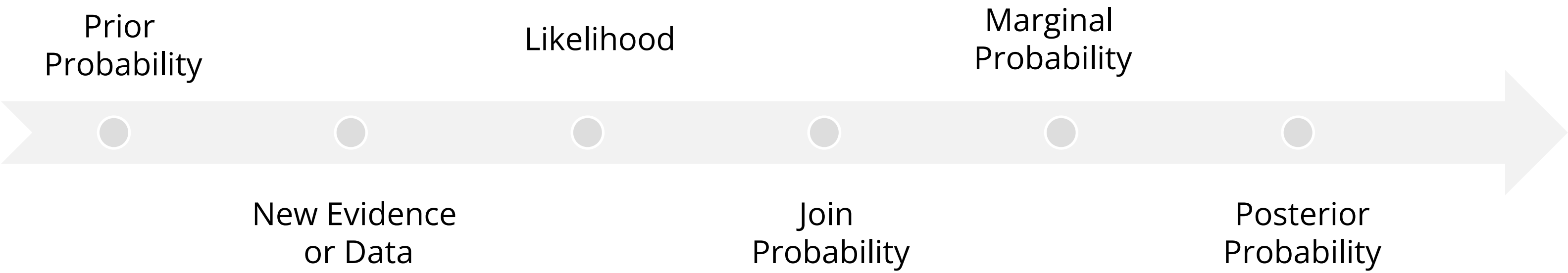
b. 2/6

c. 3/6

d. 4/6

Bayes’ Theorem

Road Map of Bayes Theorem



Bayes' Theorem

- **Prior probability:** The probability of an event occurring before any evidence is considered.

- **Likelihood:** The probability of observing the evidence given that the event has occurred.

- **Joint probability:** The probability of two events occurring together.

- **Marginal probability:** The probability of an event occurring independently of other events.

- **Posterior probability:** The probability of an event occurring after the evidence has been considered.

Let's understand all of these using examples.

Bayes' Theorem

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

P(A|B) is the posterior probability of event A occurring given event B. This is the probability that we are interested in calculating.

P(A) is the prior probability of event A occurring. This is our belief about the probability of event A occurring before we have any evidence.

P(B|A) is the likelihood of event B occurring given event A. This is the probability of observing event B if event A has already occurred.

P(B) is the probability of event B occurring. This is the probability of observing event B regardless of whether or not event A has occurred.

Bayes’ Theorem: Example

- Example: Breast Cancer Screening
- Suppose there is a mammography screening test for breast cancer that has the following characteristics:
- The test is 90% accurate when a woman has breast cancer.
- The test is 95% accurate when a woman does not have breast cancer.
- In the general population, 1% of women have breast cancer (prior probability).
- Now, let's consider a scenario where a woman receives a positive test result from the mammography screening test, and we want to determine the probability that she actually has breast cancer.

Bayes' Theorem: Example

Solution:

Prior probability of breast cancer = 0.01

Likelihood of a positive test result given breast cancer = 0.9

Likelihood of positive test result given no breast cancer = 0.05

Marginal probability of positive test result = 0.07

Posterior probability of breast cancer given positive test result = (0.01 * 0.9) / 0.07 = 0.128

Thus, the probability that the woman actually has breast cancer is 12.857%, given that she tested positive.


Q&A

Poll Time

Two events are mutually exclusive if:

a. They cannot both occur at the same time

b. They can both occur at the same time, but the probability of both occurring is very small

c. They can both occur at the same time, but the probability of both occurring is very high

d. They cannot occur at the same time, but the probability of one event occurring does not affect the probability of the other event occurring

Poll Time

Two events are mutually exclusive if:

a. **They cannot both occur at the same time.**

b. They can both occur at the same time, but the probability of both occurring is very small

c. They can both occur at the same time, but the probability of both occurring is very high

d. They cannot occur at the same time, but the probability of one event occurring does not affect the probability of the other event occurring

Central Limit Theorem

The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution.

In other words, the CLT states that if you take a large enough sample from a population, the distribution of the sample means will be approximately normal, even if the population distribution is not normal.

Assumptions:

- The samples must be independent.
- The samples must be identically distributed.
- The sample size must be large enough.

Violation of Rules:

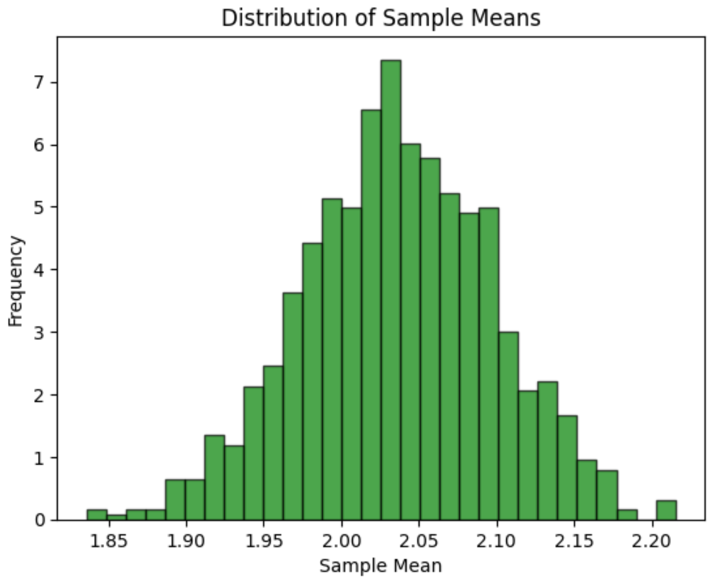
Example 1. Suppose we are sampling for the height of athletes, and we consider the height of the same athlete twice. Then we violate assumption 1.

Example 2. Suppose while sampling for heights, we consider some samples of weights. Then we violate assumption 2.

Example 3. Suppose while sampling for heights, we consider just 2 samples. Then we violate assumption 3.

Central Limit Theorem

```
# Importing Necessary Libraries
import numpy as np
import matplotlib.pyplot as plt

# Defined the dimension of population and sample.
population_size = 10000
sample_size = 1000
num_samples = 1000

# Creating random population for a given population size.
population = np.random.exponential(scale=2, size=population_size)

# Initiating variable to store the means with initial values as 0.
sample_means = np.zeros(num_samples)

# Storing Means of Created Samples
for i in range(num_samples):
    sample = np.random.choice(population, size=sample_size, replace=False)
    sample_means[i] = np.mean(sample)

# Plotting the Means
plt.hist(sample_means, bins=30, density=True, alpha=0.7, color='green',
         edgecolor='black')
plt.xlabel('Sample Mean')
plt.ylabel('Frequency')
plt.title('Distribution of Sample Means')
plt.show()
```



Pop Quiz

Q. A manufacturing company produces widgets, and the weights of the widgets follow an unknown distribution. The company takes a sample of 100 widgets and calculates the mean weight. According to the Central Limit Theorem, as the sample size increases, what happens to the sampling distribution of the mean?

A. It becomes skewed

B. It approaches a uniform distribution

C. It becomes more spread out

D. It approaches a normal distribution

Pop Quiz

Q. A manufacturing company produces widgets, and the weights of the widgets follow an unknown distribution. The company takes a sample of 100 widgets and calculates the mean weight. According to the Central Limit Theorem, as the sample size increases, what happens to the sampling distribution of the mean?

A. It becomes skewed

B. It approaches a uniform distribution

C. It becomes more spread out

D. It approaches a normal distribution

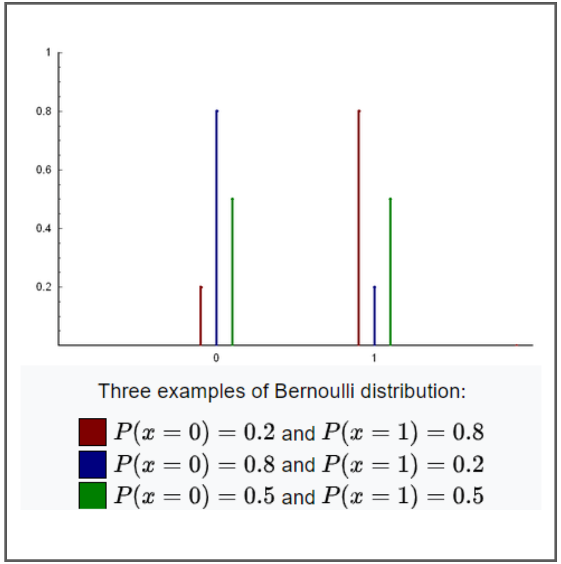
Bernoulli Distribution

The Bernoulli theorem is a fundamental concept in probability theory that deals with the probability of success or failure in a single trial of an experiment.

Example:

Consider a fair coin flip. Let's define success as getting a "heads" outcome and failure as getting a "tails" outcome.

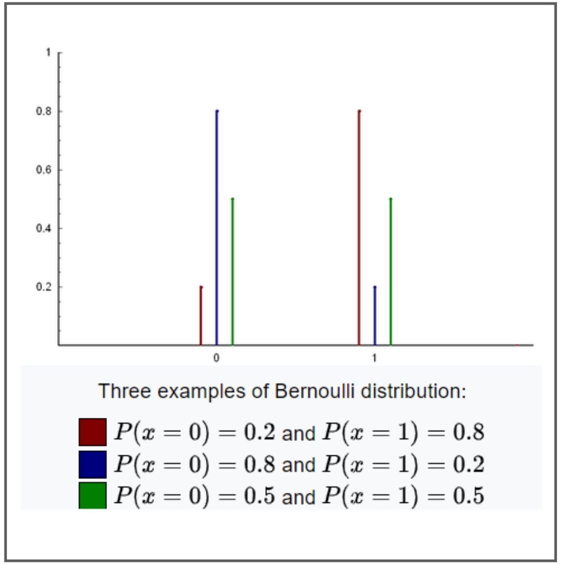
- The probability of success (getting "heads") is denoted as p.
- The probability of failure (getting "tails") is denoted as q = 1 - p.



Bernoulli Distribution

Bernoulli Distribution:

- The Bernoulli distribution represents the probability distribution of a random variable that takes only two possible outcomes, typically labeled as 1 for success and 0 for failure.
- It is characterized by a single parameter p, representing the probability of success.



Bernoulli Distribution

Example of Bernoulli Distribution:

Suppose we have a biased coin with a probability of getting heads (success) as $p = 0.6$.

- The probability of getting a head outcome (success) is 0.6.
- The probability of getting a tails outcome (failure) is 0.4 (1 - 0.6).
- The probability distribution function (PMF) of the Bernoulli distribution can be represented as: $P(X = 1) = p$ and $P(X = 0) = q$
- The expected value (mean) of a Bernoulli distribution is $E(X) = p$, and the variance is $Var(X) = p * q$.

Pop Quiz

Q. In a coin toss experiment, the variable "success" is recorded as 1 if the coin lands head up and 0 if it lands tails up. What type of probability distribution can be used to model this scenario?

A. Normal distribution

B. Poisson distribution

C. Binomial distribution

D. Bernoulli distribution

Pop Quiz

Q. In a coin toss experiment, the variable "success" is recorded as 1 if the coin lands head up and 0 if it lands tails up. What type of probability distribution can be used to model this scenario?

A. Normal distribution

B. Poisson distribution

C. Binomial distribution

D. **Bernoulli distribution**

Binomial Distribution

- The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials.
- It is widely used in statistics and probability theory to analyze and predict outcomes in various real-world scenarios.

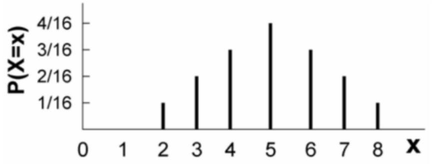
The binomial distribution is defined by two parameters:
n: The number of trials or observations.
p: The probability of success in each trial.

Probability Mass Function (PMF):

- The probability mass function of the binomial distribution gives the probability of obtaining exactly k successes in n trials.

- The PMF formula for the binomial distribution is:

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



X is the sum of outcomes got by rolling a 4 sided dice twice.

Binomial Distribution

Example of Binomial Distribution:
A fair coin is flipped 10 times. What is the probability of getting exactly 5 heads?

Solution:
Let success be flipping heads, and failure is flipping tails.
P(Success) = p = 0.5
P(Failure) = q = 0.5

The following formula gives the probability of getting exactly 5 heads in 10 trials:

P(X = 5) = nC5 * p^5 * q^5

Where, n = 10, p = 0.5, and q = 0.5

P(X = 5) = 10C5 * (0.5)^5 * (0.5)^5 = 252/1024

This is about a 25% chance of getting exactly 5 heads in 10 trials.

Poisson Distribution

The Poisson distribution is a discrete probability distribution that represents the number of events that occur in a fixed interval of time or space.

$$\boxed{P_X(k) = \frac{e^{-\lambda} * \lambda^k}{k!}}$$

- **Px(k) = P(X = k)** represents the probability that the random variable **X** takes on the value **k**.
- **λ** (lambda) is the average rate or expected value of the Poisson distribution.
- **e** is the base of the natural logarithm, approximately 2.71828.
- **k** is the number of events or occurrences for which we want to calculate the probability.
- **k!** denotes the factorial of **k**, which is the product of all positive integers less than or equal to **k**.

Poisson Distribution

- In Figure 1, X axis represent the number of Events and Y axis represents the probability of getting those events.

- The mean of the distribution is denoted by λ.

- The probability of getting a certain number of events is given by the Poisson Distribution.

- For Example, the probability of getting 2 events in a Poisson distribution with mean λ = 3 is 0.183, and so on.

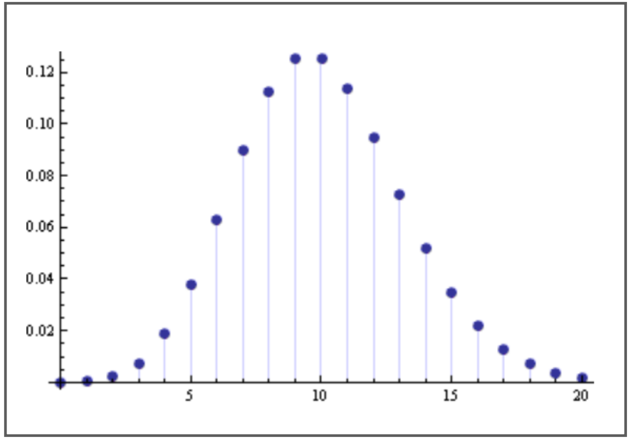


Figure 1

Poisson Distribution

- The graph shows that the probability of getting a certain number of events decreases as the number of events increases. This is because the Poisson distribution is a bell-shaped curve. The bell-shaped curve is also called the normal distribution.

- The peak of the bell-shaped curve is at the mean of the distribution, which is λ in the case of the Poisson distribution. This means that the probability of getting the mean number of events is highest.

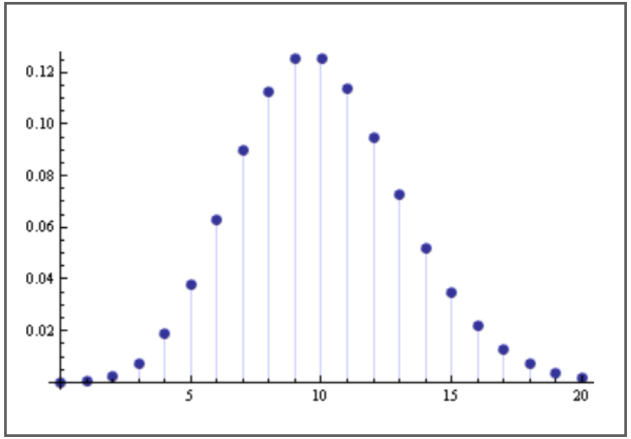


Figure 1

Poisson Distribution

Example:

A call center receives an average of 10 phone calls per hour.

The Poisson distribution can be used to calculate the probability of receiving a certain number of phone calls in an hour.

For example, the probability of receiving 5 phone calls in an hour is:

Using,
$\lambda = 10$
$K = 5$

P(5) = (10^5) * e^(-10) / 5!

This is about a 2.7% chance of receiving 5 phone calls in an hour.

Pop Quiz

Q. In a manufacturing process, the average number of defects per day is 4. What is the probability of observing exactly 3 defects on a randomly chosen day, assuming the defects follow a Poisson distribution?

A. 0.075

B. 0.091

C. 0.128

D. 0.183

Pop Quiz

Q. In a manufacturing process, the average number of defects per day is 4. What is the probability of observing exactly 3 defects on a randomly chosen day, assuming the defects follow a Poisson distribution?

A. 0.075

B. **0.091**

C. 0.128

D. 0.183


Q&A

Summary

- Defined conditional probability (P(A|B)) and explained Bayes' theorem
- Discussed key probability distributions:
  - Binomial: Modeled successes in fixed trials
  - Poisson: Modeled rare event occurrences in fixed intervals
  - Bernoulli: Represented single trials with two outcomes
- Analyzed the statistical significance of distribution curves
- Evaluated and thought critically about probability problems
- Applied probability concepts to real-world problems

Activity

In a bolt factory, three machines M_1, M_2, and M_3 manufacture 2000, 2500, and 4000 bolts every day. Of their output 3%, 4%, and 2.5% are defective bolts. One of the bolts is drawn very randomly from a day's production and is found to be defective. What is the probability that it was produced by machine M_2?

Activity

Hint to solve the previous problem:

1. **Define the variables and parameters.**

- A: The bolt was produced by machine M_2.
- B: The bolt is defective.
- P(A | B): The probability that the bolt was produced by machine M_2 given that it is defective.
- P(B | A): The probability that the bolt is defective given that it was produced by machine M_2.
- P(A): The probability that the bolt was produced by machine M_2.
- P(B): The probability that the bolt is defective.

2. **Calculate the probability of the event B happening.**

3. **Calculate the probability of the event A happening given that event B happened.**

THANK YOU!

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut upGrad

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

By the End of This Session, You Will:

- Understand probability distributions and their role in describing random variables.
- Identify common probability distributions and their characteristics.
- Apply probability distributions to model real-world data and make predictions.
- Enhance data analysis, probability assessment, and decision-making skills using probability distributions.
- Build a foundation for further exploration and application of probability distributions.

What Have You Learned So Far?

- Probability is a measure of the likelihood of an event occurring.
- Conditional probability deals with the probability of an event occurring given that another event has already occurred.
- Bayes' theorem allows us to update the probability of an event based on new information or evidence.
- Bernoulli's distribution models a binary outcome, where an event can have only two possible outcomes.
- Poisson distribution is used to model the number of events occurring in a fixed interval of time or space.
- Binomial distribution models the number of successes in a fixed number of independent Bernoulli trials.

Poll Time

Q. A retail store manager wants to estimate the number of customers who will visit the store on a particular day. Which of the following probability distributions would be the most appropriate for this task?

a. Bernoulli distribution

b. Binomial distribution

c. Poisson distribution

Problem Statement

Problem:
Akshay is interested in understanding the spread of data about athletes' performances and earnings. He wants to use probability distributions to identify outliers and other useful information in the data.

Goal:
Akshay's goal is to use probability distributions to gain a better understanding of the data and to identify any potential problems or trends.

Approach:
1. Akshay will use basic probability distributions to understand the spread of the data.
2. He will also use statistical tools to identify outliers and potential problems or trends.

1. Akshay expects to be able to identify outliers and trends in the data.
2. He also expects to gain a better understanding of the spread of the data.

Areas to Focus

Areas to Focus

- Trends in the data
- Distribution of the data
- Earnings of different athletes
- Most likely combinations of elements in the data

Pop Quiz

Q. Which of the following is NOT an area of focus for Akshay in his analysis of the data?

a. Outliers can have a significant impact on statistical analysis

b. Outliers can be used to identify potential problems or trends in the data

c. Outliers are always caused by errors in data collection or processing

d. Outliers can represent real-world phenomena not captured by the rest of the data

Pop Quiz

Q. Which of the following is NOT an area of focus for Akshay in his analysis of the data?

a. Outliers can have a significant impact on statistical analysis

b. Outliers can be used to identify potential problems or trends in the data

c. **Outliers are always caused by errors in data collection or processing**

d. Outliers can represent real-world phenomena not captured by the rest of the data

Understanding the Data

Sneak Peak into the Data

Actual Data

rcc	wcc	hc	hg	ferr	bmi	ssf	pcBfat	lbm	ht	wt	sex	sport
3.96	7.5	37.5	12.3	60	20.56	109.1	19.75	63.32	195.9	78.9	f	B_Ball
4.41	8.3	38.2	12.7	68	20.67	102.8	21.30	58.55	189.7	74.4	f	B_Ball
4.14	5.0	36.4	11.6	21	21.86	104.6	19.88	55.36	177.8	69.1	f	B_Ball
4.11	5.3	37.3	12.6	69	21.88	126.4	23.66	57.18	185.0	74.9	f	B_Ball
4.45	6.8	41.5	14.0	29	18.96	80.3	17.64	53.20	184.6	64.6	f	B_Ball
4.10	4.4	37.4	12.5	42	21.04	75.2	15.58	53.77	174.0	63.7	f	B_Ball
4.31	5.3	39.6	12.8	73	21.69	87.2	19.99	60.17	186.2	75.2	f	B_Ball
4.42	5.7	39.9	13.2	44	20.62	97.9	22.43	48.33	173.8	62.3	f	B_Ball
4.30	8.9	41.1	13.5	41	22.64	75.1	17.95	54.57	171.4	66.5	f	B_Ball
4.51	4.4	41.6	12.7	44	19.44	65.1	15.07	53.42	179.9	62.9	f	B_Ball

Variable	Description	Units
rcc	Red Blood Cell Count	10^12-1
wcc	White Blood Cell Count	10^12 per liter
hc	Hematocrit	percent
hg	Hemoglobin Concentration	g per decaliter
fe	Plasma Ferritins	ng dl-1
bmi	Body Mass Index	kg cm^-2 * 10^2
ssf	Sum of Skin Folds	-
pcbfat	Percent Body Fat	-
lbm	Lean Body Mass	kg
ht	Height	cm
wt	Weight	kg
sex	Gender	f or m
sport	Sport	B_Ball, Row, Netball, Swim, Field, T_400m, T_Sprnt, Tennis, Gym, or W_Polo

What Are You Going to Build?

We have conducted a comprehensive statistical analysis of the athlete dataset, including probability distribution fitting, outlier detection, skewness and kurtosis, descriptive statistics, and normal distribution assessment.

This analysis will provide valuable insights into the athlete dataset and help us to better understand the performance of athletes and make informed decisions about training and coaching.

The next steps could include further exploratory data analysis, hypothesis testing, predictive modeling, and interpretation and reporting.

Pop Quiz

Q. A doctor is analyzing the blood test results of a patient. The patient's red blood cell count (rcc) is significantly lower than the normal range. Which of the following statistical analysis techniques would be most helpful in identifying the cause of the low rcc?

a. Probability distribution fitting

b. Skewness and kurtosis

c. Descriptive statistics

d. Outlier detection

Pop Quiz

Q. A doctor is analyzing the blood test results of a patient. The patient's red blood cell count (rcc) is significantly lower than the normal range. Which of the following statistical analysis techniques would be most helpful in identifying the cause of the low rcc?

a. Probability distribution fitting

b. Skewness and kurtosis

c. Descriptive statistics

d. Outlier detection


Q&A

Hands-on: Case Study Questions

Pop Quiz

Q. Which of the following statistical measures would be most helpful in summarizing the height data for swimming athletes?

a. Mean

b. Median

c. Mode

d. Standard deviation

Pop Quiz

Q. Which of the following statistical measures would be most helpful in summarizing the height data for swimming athletes?

a. Mean

b. Median

c. Mode

d. Standard deviation


Q&A

Activity 1

Pre-requisites:

- Basic knowledge of probability distributions
- Python programming skills
- NumPy, pandas, and Matplotlib libraries

Scenario:

The following code loads the breast_cancer dataset from the sklearn library:

```
import numpy as np
import pandas as pd
from sklearn.datasets import load_breast_cancer

data = load_breast_cancer()
df = pd.DataFrame(data.data, columns=data.feature_names)
```

Tasks:

- What kind of distribution is the given dataset?
- Identify any outliers in the dataset.
- Determine the skewness of the dataset.

Activity 2

Pre-requisites:
- Basic knowledge of probability distributions
- Python programming skills
- NumPy, Pandas, and Matplotlib libraries

Scenario:

The following code loads the diabetes dataset from the sklearn library:

```
import numpy as np
import pandas as pd
from sklearn.datasets import load_diabetes

data = load_diabetes()
df = pd.DataFrame(data.data, columns=data.feature_names)
```

Tasks:
- What is the kurtosis of the diabetes dataset?
- Is the diabetes dataset normally distributed?
- Can you fit a Poisson distribution to the diabetes dataset?

Summary

- Explored the concept of probability distributions, which provided a mathematical framework for describing the possible values of a random variable and their likelihoods.
- Discussed common probability distributions such as normal distribution, uniform distribution, binomial distribution, Bernoulli distribution, Poisson distribution, and exponential distribution.
- Highlighted the practical applications of these distributions in modeling real-world data and making predictions about future events.
- Gained insights into data analysis, probability assessment, and informed decision-making through the understanding and utilization of probability distributions.

Session Feedback

Next Session:

Excel Case Study - II

THANK YOU

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut
upGrad