Probability Tutorial

Understanding Uncertainty

Tiago Paixão

September 20, 2023

Outline

Introduction

Introduction

- Probability is a measure of uncertainty.
- It plays a fundamental role in statistics and modeling.
- Numeric range: 0 (impossible) to 1 (certain).

Sample Space and Events

Sample Space and Events

- Sample Space (S): Set of all possible outcomes.
- Event (E): Subset of the sample space.
- Example: Rolling a fair six-sided die

$$S = \{1, 2, 3, 4, 5, 6\},$$

$$E = \{2, 4, 6\}.$$

Probability Rules

Probability Rules

1. Probability of an Event (P(E)):

$$P(E) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Possible Outcomes}}$$

2. Complement Rule:

$$P(\neg E) = 1 - P(E)$$

3. Addition Rule for Mutually Exclusive Events:

$$P(A \text{ or } B) = P(A) + P(B)$$

4. Multiplication Rule for Independent Events:

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Probability Rules

1. Multiplication Rule for Dependent Events:

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

2. Conditional Probability:

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

3. Bayes' Theorem:

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

4. Law of Total Probability:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

Example: DNA Sequence Mutations

Probability Rule of Multiplication for Independent Events

- In DNA sequencing, mutations can occur at specific positions along a DNA sequence.
- Mutations are categorized as transitions (e.g., A to G) or transversions (e.g., A to C).
- Let's consider two positions: Position 1 and Position 2.
- Probability of a transition mutation at each position: $P(\text{Transition}) = 0.2$.
- Probability of a transversion mutation at each position: $P(\text{Transversion}) = 0.1$.
- We want to calculate the probability of specific mutation combinations at both positions.

Example: DNA Sequence Mutations

Probability Rule of Multiplication for Independent Events

- In DNA sequencing, mutations can occur at specific positions along a DNA sequence.
- Mutations are categorized as transitions (e.g., A to G) or transversions (e.g., A to C).
- Let's consider two positions: Position 1 and Position 2.
- Probability of a transition mutation at each position: $P(\text{Transition}) = 0.2$.
- Probability of a transversion mutation at each position: $P(\text{Transversion}) = 0.1$.
- We want to calculate the probability of specific mutation combinations at both positions.

Position 1/2	Transition	Transversion
Transition	0.04	0.02

Bayes' Rule Example: Rare Disease Testing

Scenario:

- Disease: Disease X (a rare disease).
- Prevalence: Disease X is very rare, with a prevalence of 0.1

Diagnostic Test:

- Test Accuracy:
  - Sensitivity (True Positive Rate): 98%
  - Specificity (True Negative Rate): 95%

Bayes' Rule Example: Rare Disease Testing (Contd.)

Question: Given that an individual tests positive for Disease X, what is the probability that they actually have the disease?

Solution using Bayes' Rule:

- $P(D)$: Prevalence of the disease = 0.1% = 0.001.
- $P(T|D)$: Probability of testing positive if you have the disease (Sensitivity) = 98% = 0.98.
- $P(T|\neg D)$: Probability of testing positive if you don't have the disease (False Positive Rate) = 5% = 0.05.

Bayes' Rule Example: Rare Disease Testing (Contd.)

Using Bayes' Rule:

To find $P(D|T)$, the probability of having the disease given a positive test result, we use Bayes' rule:

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

To find $P(T)$, we use the law of total probability:

$$P(T) = P(T|D) \cdot P(D) + P(T|\neg D) \cdot P(\neg D)$$

Where $P(\neg D)$ is the probability of not having the disease.

Now, we can calculate $P(T)$.

Bayes’ Rule Example: Rare Disease Testing (Contd.)

Calculating $P(T)$:

Using the law of total probability:

$$P(T) = (0.98 \cdot 0.001) + (0.05 \cdot 0.999) \approx 0.05093$$

Now, we can calculate $P(D|T)$ using Bayes’ rule:

$$P(D|T) = \frac{0.98 \cdot 0.001}{0.05093} \approx 0.0192$$

So, if an individual tests positive for Disease X, the probability that they actually have the disease is approximately 1.92%.

Discrete Probability Distributions

In probability theory, discrete probability distributions describe the probabilities of individual outcomes in a discrete (countable) sample space.

Discrete Probability Distributions

In probability theory, discrete probability distributions describe the probabilities of individual outcomes in a discrete (countable) sample space.

- Bernoulli Distribution

Bernoulli Distribution models a single trial with two possible outcomes, usually labeled as ”success” and ”failure.”

Discrete Probability Distributions

In probability theory, discrete probability distributions describe the probabilities of individual outcomes in a discrete (countable) sample space.

- Bernoulli Distribution
- Binomial Distribution

Bernoulli Distribution models a single trial with two possible outcomes, usually labeled as ”success” and ”failure.”

Binomial Distribution models the number of successes in a fixed number of independent Bernoulli trials.

Discrete Probability Distributions

In probability theory, discrete probability distributions describe the probabilities of individual outcomes in a discrete (countable) sample space.

- Bernoulli Distribution
- Binomial Distribution
- Poisson Distribution

Bernoulli Distribution models a single trial with two possible outcomes, usually labeled as ”success” and ”failure.”

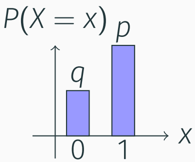
Binomial Distribution models the number of successes in a fixed number of independent Bernoulli trials.

Poisson Distribution models the number of events that occur in a fixed interval of time or space.

Bernoulli Distribution

Description:

- Models a single trial with two possible outcomes.
- Often labeled as "success" (S) and "failure" (F).
- Probability of success: p.
- Probability of failure: $q = 1 - p$.



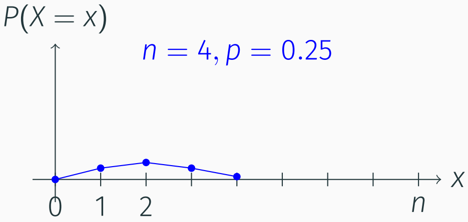
PMF:

$$P(X = x) = \begin{cases} p, & \text{if } x = 1 \\ q, & \text{if } x = 0 \end{cases}$$

Binomial Distribution

Description:

- Models the number of successes in a fixed number of independent Bernoulli trials.
- Parameters: n (number of trials) and p (probability of success in each trial).



PMF:

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

Inheritance of a Recessive Trait

Example: Inheritance of Red Hair (rr)

- Consider a population with a gene for hair color.
- There are two alleles: ”R” (dominant) and ”r” (recessive).
- Individuals inherit two alleles, one from each parent.
- Let’s find the probability distribution for the number of offspring inheriting the ”rr” genotype in a family.

Probability of Inheriting the Recessive Trait

Binomial Distribution

- Probability of inheriting the "rr" genotype from both parents: p^2.
- In a family with n children, we want to find the probability of k children inheriting "rr."

Probability Formula

$$P(X = k) = \binom{n}{k} \cdot (p^2)^k \cdot (1 - p^2)^{n-k}$$

- $\binom{n}{k}$: Number of ways to choose k children out of n.
- p^2: Probability of each child inheriting "rr."
- $(1 - p^2)$: Probability of each child not inheriting "rr."

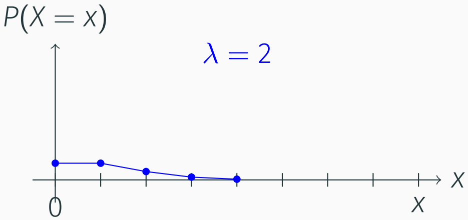
Poisson Distribution

Description:

- Models the number of events occurring in a fixed interval of time or space.
- Limit of Binomial when n is large and p is small
- Parameter: λ (average rate of events).

PMF:

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$



Continuous distribution

Continuous vs. Discrete Distributions

Discrete Distributions

Continuous Distributions

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.

Continuous Distributions

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.
- Examples: Binomial, Poisson, Bernoulli.

Continuous Distributions

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.
- Examples: Binomial, Poisson, Bernoulli.
- Probability Mass Function (PMF).

Continuous Distributions

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.
- Examples: Binomial, Poisson, Bernoulli.
- Probability Mass Function (PMF).
- Probability associated with **individual points**.

Continuous Distributions

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.
- Examples: Binomial, Poisson, Bernoulli.
- Probability Mass Function (PMF).
- Probability associated with **individual points**.

Continuous Distributions

- Deal with random variables that can take **any value** in a range.

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.
- Examples: Binomial, Poisson, Bernoulli.
- Probability Mass Function (PMF).
- Probability associated with **individual points**.

Continuous Distributions

- Deal with random variables that can take **any value** in a range.
- Examples: Normal, Uniform, Exponential.

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.
- Examples: Binomial, Poisson, Bernoulli.
- Probability Mass Function (PMF).
- Probability associated with **individual points**.

Continuous Distributions

- Deal with random variables that can take **any value** in a range.
- Examples: Normal, Uniform, Exponential.
- Probability Density Function (PDF).

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.
- Examples: Binomial, Poisson, Bernoulli.
- Probability Mass Function (PMF).
- Probability associated with **individual points**.

Continuous Distributions

- Deal with random variables that can take **any value** in a range.
- Examples: Normal, Uniform, Exponential.
- Probability Density Function (PDF).
- Probability associated with **intervals**.

Continuous vs. Discrete Distributions

Discrete Distributions

- Deal with random variables that take on **countable** values.
- Examples: Binomial, Poisson, Bernoulli.
- Probability Mass Function (PMF).
- Probability associated with **individual points**.

Continuous Distributions

- Deal with random variables that can take **any value** in a range.
- Examples: Normal, Uniform, Exponential.
- Probability Density Function (PDF).
- Probability associated with **intervals**.

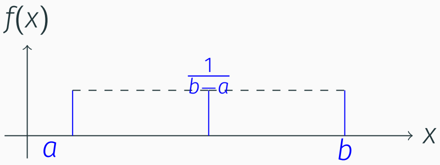
Key Difference:

- Discrete: Probability at **specific points**.
- Continuous: probability **density** , Probability over **ranges or intervals**.

Uniform Distribution

Description:

- Models continuous random variables with a constant probability density.
- Parameters: a (lower bound) and b (upper bound).



PDF (Probability Density Function):

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

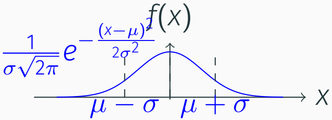
Normal Distribution (Gaussian)

Description:

- Models a continuous random variable with a bell-shaped probability density.
- Parameters: μ (mean) and σ (standard deviation).

PDF (Probability Density Function):

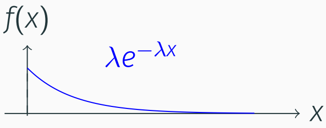
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Exponential Distribution

Description:

- Models the time between events in a Poisson process.
- Parameter: λ (rate parameter, average number of events per unit time).



PDF (Probability Density Function):

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

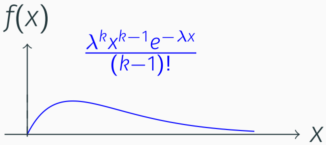
Gamma Distribution

Description:

- Models the waiting time until a Poisson process reaches a certain number of events.
- Parameters: k (shape) and λ (rate).

PDF (Probability Density Function):

$$f(x) = \begin{cases} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$



Central Limit Theorem

Central Limit Theorem (CLT)

Statement:

Conditions:

Illustration:

Central Limit Theorem (CLT)

Statement:

- The CLT states that the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the original distribution.

Conditions:

Illustration:

Central Limit Theorem (CLT)

Statement:

- The CLT states that the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the original distribution.

Conditions:

- Random variables must be i.i.d.

Illustration:

Central Limit Theorem (CLT)

Statement:

- The CLT states that the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the original distribution.

Conditions:

- Random variables must be i.i.d.
- The number of variables must be *sufficiently* large.

Illustration:

Central Limit Theorem (CLT)

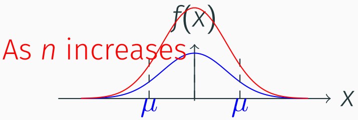
Statement:

- The CLT states that the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the original distribution.

Conditions:

- Random variables must be i.i.d.
- The number of variables must be *sufficiently* large.

Illustration:



Central Limit Theorem (CLT)

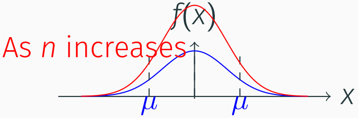
Statement:

- The CLT states that the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the original distribution.

Conditions:

- Random variables must be i.i.d.
- The number of variables must be *sufficiently* large.

Illustration:



- The CLT is a fundamental concept in statistics and helps explain why the normal distribution appears in various real-world situations.

Probability and random sampling in R

Normal Distribution Functions in R

Normal Distribution (`norm`)

- R provides several functions for working with the Normal Distribution.
- Key functions for the Normal Distribution:
    - *dnorm(x, mean = μ, sd = σ)* - Probability Density Function (PDF)
    - *pnorm(q, mean = μ, sd = σ)* - Cumulative Distribution Function (CDF)
    - *qnorm(p, mean = μ, sd = σ)* - Quantile Function (Inverse CDF)
    - *rnorm(n, mean = μ, sd = σ)* - Random Number Generation

Parameters:

- μ: Mean of the distribution.
- σ: Standard deviation of the distribution.
- x: Values to evaluate (PDF).
- q: Quantiles (CDF and quantile).
- p: Probabilities (CDF and quantile).
- n: Number of random samples (random number generation).

Probability Distributions

- R provides functions to work with various probability distributions.
- Some of the most common distributions include:
  - Binomial Distribution (*rbinom*)
  - Poisson Distribution (*rpois*)
  - Normal Distribution (*rnorm*)
  - Uniform Distribution (*runif*)
  - Exponential Distribution (*rexp*)
  - Gamma Distribution (*rgamma*)

Conclusion

Conclusion

- Probability is a fundamental concept in statistics.
- These rules help us understand uncertainty and make informed decisions.
- Explore advanced probability topics for deeper understanding.