Statistics in Data Science

overview

1. **Introduction to Statistics in Data Science**
2. **Distributions**
3. **Distribution Estimators: MoM, MLE, KDE**
4. **Point Estimates**
5. **Statistical Hypothesis Testing**
6. **Correlation**
7. **Practical Examples**

2. Distributions

- Define probability distribution.

  Probability measures how likely it is for an event to occur on a scale from 0 (the event never occurs) to 1 (the event always occurs).

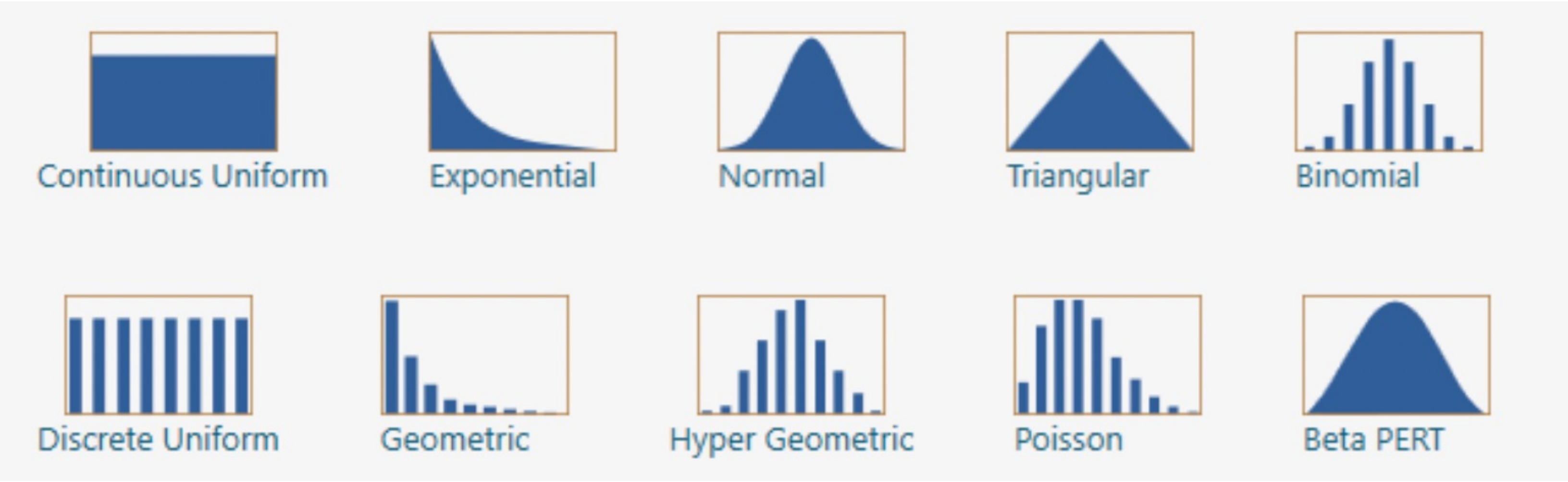
- Discrete and continuous distributions.
- common probability distributions:

  Uniform distribution; Normal distribution; Exponential distribution; Poisson distribution

- Include probability density functions (PDFs) and cumulative distribution functions (CDFs).

- A **discrete random variable** is one whose set of assumed values is countable (arises from counting).
  - values are drawn from a finite set of states.
  - Simply, this means that if I pick any two consecutive outcomes. I can't get an outcome that's in between.
  - In mathematics, we would say that the list of outcomes is countable.

- A **continuous random variable** is one whose set of assumed values is uncountable (arises from measurement.).
  - Values are drawn from a range of real-valued numerical values.



- **Continuous Probability Distributions**
- A continuous probability distribution summarizes the probability for a continuous random variable.
- The probability distribution function, or PDF, defines the probability distribution for a continuous random variable.
- Continuous probability distribution also has a cumulative distribution function, or CDF, that defines the probability of a value less than or equal to a specific numerical value from the domain.
- Distributions include:

  Normal or Gaussian distribution; Exponential distribution; Pareto distribution

- Examples:

  The probabilities of the heights of humans;The probabilities of income levels

Continous Probabilty Terminology:

- **PDF: Probability Density Function,** returns the probability of a given continuous outcome.

$$\int_a^b f(x)dx = P(a \leq X \leq b)$$

- **CDF: Cumulative Distribution Function,** returns the probability of a value less than or equal to a given outcome.

$$f(x) = P(X \leq x)$$

- **PPF: Percent-Point Function,** returns a discrete value that is less than or equal to the given probability.
    - Inverse of CDF

- **Discrete Probability Distributions**
- A discrete probability distribution summarizes the probabilities for a discrete random variable.
- The probability mass function, or PMF, defines the probability distribution for a discrete random variable.
    - It is a function that assigns a probability for specific discrete values.
- A discrete probability distribution has a *cumulative distribution function, or CDF.*
    - This is a function that assigns a probability that a discrete random variable will have a value of less than or equal to a specific discrete value.
- Distribution include:

    Bernoulli and binomial distributions;   Poisson distribution.

- Examples:

    The probabilities of dice rolls form a discrete uniform distribution.

    The probabilities of coin flips.

Discrete Probabilty Terminology:

- **PMF: Probability Mass Function**, returns the probability of a given outcome.

$$f(x) = P(X = x)$$

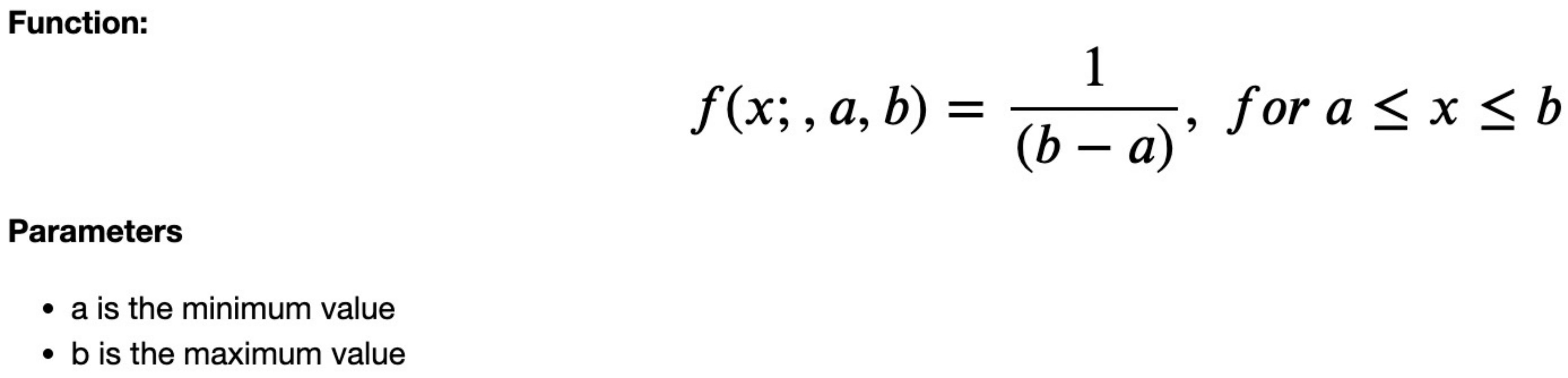
- **CDF: Cumulative Distribution Function**, returns the probability of a value less than or equal to a given outcome.

$$f(x) = P(X \leq x)$$

- **PPF: Percent-Point Function**, returns a discrete value that is less than or equal to the given probability.
    - Inverse of CDF

Uniform distribution

- Continuous
- Each value within a certain range is equally likely to occur, and values outside of the range never occur.
- Example: a die roll has six possible outcomes: 1,2,3,4,5, or 6. There is a 1/6 probability for each number being rolled.



Function:

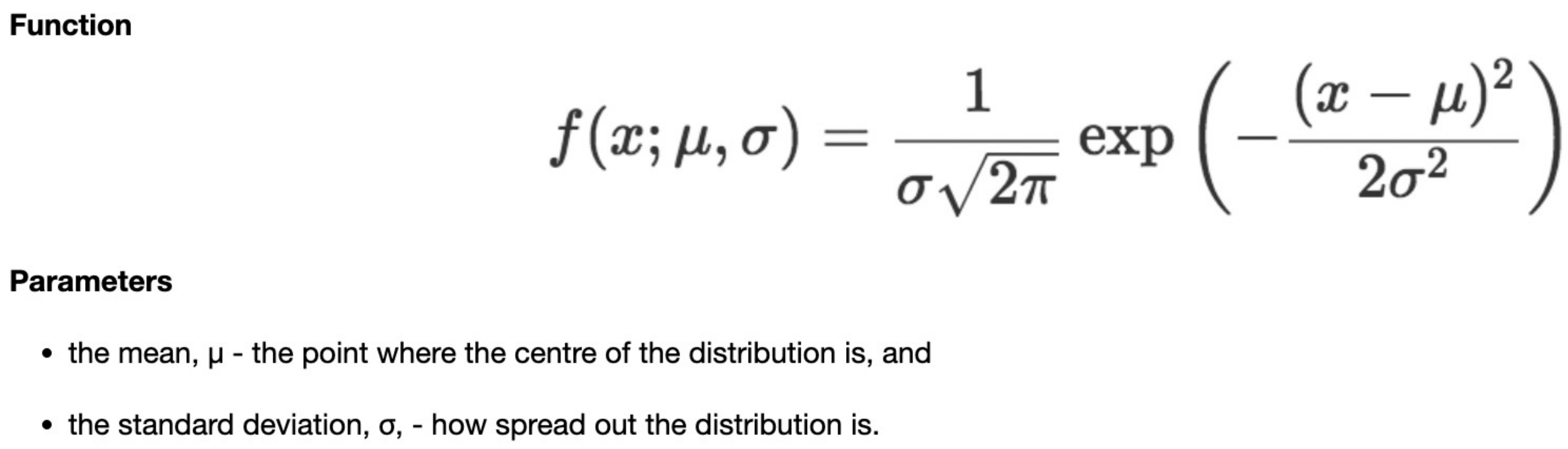
$$f(x; , a, b) = \frac{1}{(b - a)}, \ for\ a \leq x \leq b$$

Parameters

- a is the minimum value
- b is the maximum value

Normal distribution

- Continuous
- A normal distribution is defined by its center (mean) and spread (standard deviation.).
- Many common statistical tests assume distributions are normal.



Function

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Parameters

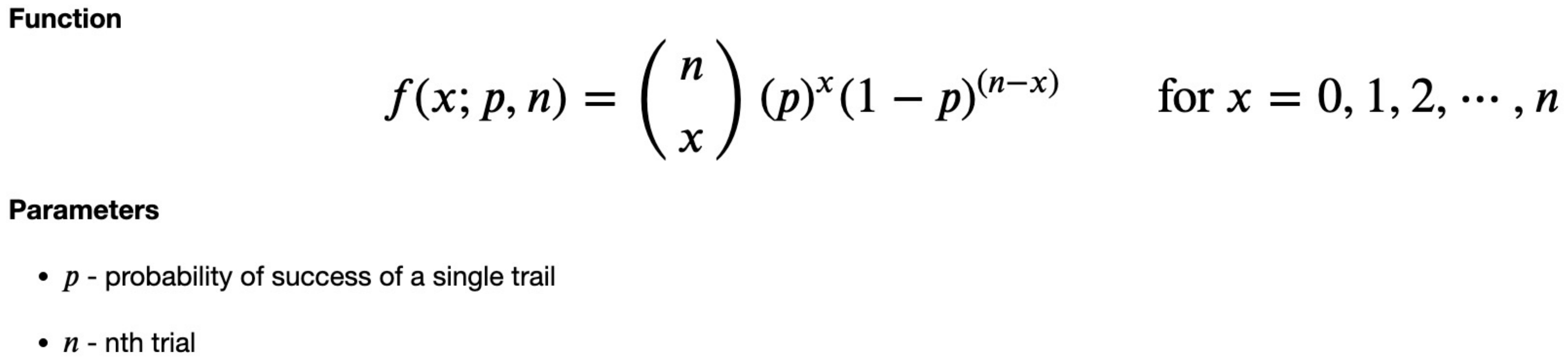
- the mean, μ - the point where the centre of the distribution is, and
- the standard deviation, σ, - how spread out the distribution is.

Binomial distribution

- The Binomial Distribution is a discrete probability distribution that models the number of successes in a fixed number of independent and identically distributed Bernoulli trials. Each Bernoulli trial has only two possible outcomes: success (usually denoted as "1") and failure (usually denoted as "0"). The Binomial Distribution is characterized by two parameters:

1. n (Number of Trials): It represents the total number of trials or experiments, each with a binary outcome (success or failure).

2. p (Probability of Success): It represents the probability of success in a single trial. It is the same for each trial and remains constant throughout the experiments.



Function

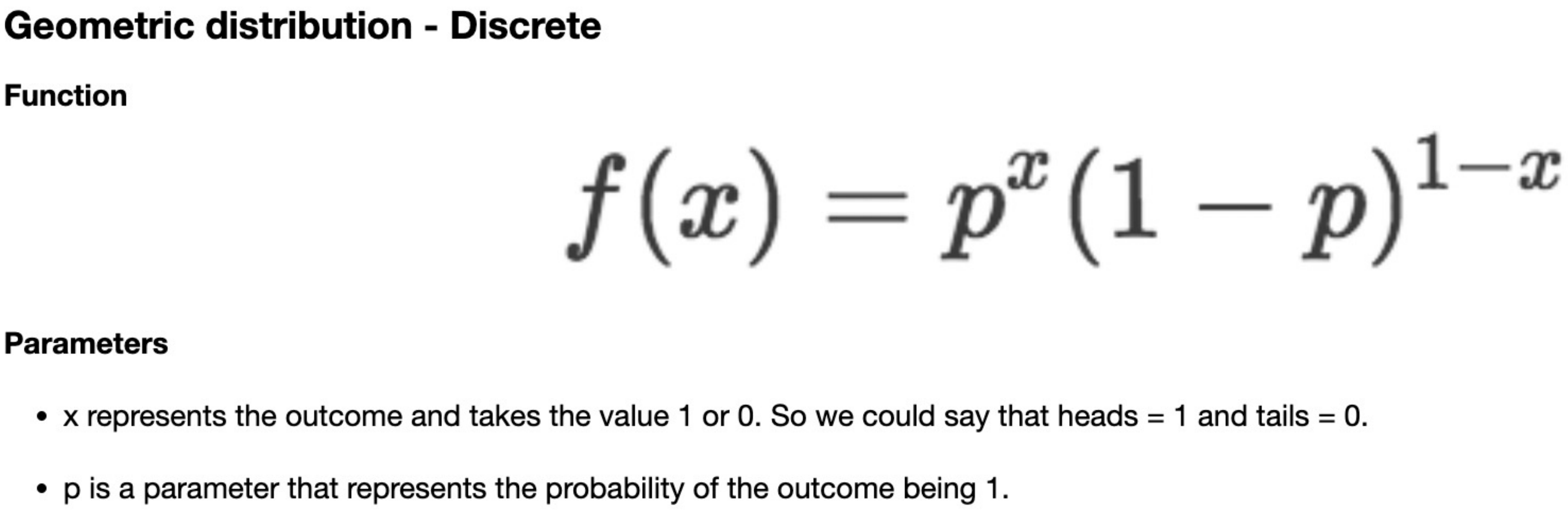
$$f(x; p, n) = \binom{n}{x} (p)^x (1-p)^{(n-x)} \qquad \text{for } x = 0, 1, 2, \cdots, n$$

Parameters

- p - probability of success of a single trail
- n - nth trial

The Geometric and Exponential Distributions

- The geometric and exponential distributions model the time it takes for an event to occur.
- The ***geometric distribution*** is discrete; Models the number of trials it takes to achieve a success in repeated experiments with a given probability of success.
- The ***exponential distribution*** is a continuous analog of the geometric distribution; Models the amount of time you have to wait before an event occurs given a certain occurrence rate.



Geometric distribution - Discrete

Function

$$f(x) = p^x(1-p)^{1-x}$$

Parameters

- x represents the outcome and takes the value 1 or 0. So we could say that heads = 1 and tails = 0.
- p is a parameter that represents the probability of the outcome being 1.

Exponential - Continuous

Function

$$f(x; \mu, \beta) = \frac{1}{\beta} e^{-(x-\mu)/\beta} \quad x \geq \mu; \beta > 0$$

Parameters

- μ is the location parameter and
- β is the scale parameter (the scale parameter is often referred to as λ which equals 1/β).
- The case where μ = 0 and β = 1 is called the standard exponential distribution.

Poisson distribution

- Discrete
- The Poisson distribution models the probability of seeing a certain number of successes within a time interval
- where the time it takes for the next success is modeled by an exponential distribution.
- The Poisson distribution can be used to model traffic, such as the number of arrivals a hospital can expect in a hour's time or the number of emails you'd expect to receive in a week.

Function

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Parameters

- $X = \{0, 1, 2, \ldots\}$
- $\lambda > 0$, where λ is both the mean and the variance of X.

$$E(X) = \mathrm{Var}(X) = \lambda$$

- e = 2.71828

Check all the figures and examples of there distributions in notebook:

https://github.com/q-tong/CS405-605-Data-Science/blob/main/Fall2023/Lecture/3.Statistics/02_Distribution.ipynb