Statistical Distributions

Sangbaptist

2023-10-01

Table of contents

What is a statistical distribution?

A statistical distribution is a mathematical function that defines how outcomes of an experimental trial occur randomly in a probable way (Schumacker 2017). In other words,a statistical distributions is like a map that shows us how data or outcomes are spread out. We can imagine it has a way to see all possible results of something happening and how likely each result is. A real-world example we can look at is students' grades on a test. The grades often form a pattern like a bell shape (normal distribution), where most scores are around the average, with a few very high and low scores. A teacher may use this to understand overall performance and setting grading standards.

Compenents of a statistical distribution

A statistical distribution has several key components that define its **shape**, **central tendency** and **spread**. Below are a some some of the main components of a statistical distribution:

1. **Random Variable**: In a given Sample Space Ω, a **random variable** is any rule that associates a number with each outcome in Ω. These are the variables whose values are determined by chance. In a distribution, it represents the possible outcomes or values that the distribution can take ## Tyeps of random variables:

- Discrete Random Variable: These are variables that can take on specific countable values (e.g. number of goals score by Real Madrid in a match).
- Continuous Random Variable: On the other hand, continuous random variables can take on infinite range of values within and interval.(e.g. temperature, height of students, etc.)

2. **Probability Function**: For discrete variables, the **Probability Mass Function (PMF)** gives the probability of each specific value while in continuous variables, the **Probability Density Function (PDF)** provides the density of values at each point (though exact probabilities are calculated over intervals).
3. **Cumulative Distribution Function (CDF)**: The CDF gives the probability that the variable takes on a value less of equal to a particular point. It is a cumulative measure that helps understand the likelihood of different ranges.
4. **Parameters**: Parameters of a distribution define the characteristics of the distribution. This includes:

- **Mean** (μ): The average of expected value of the distribution, showing its center.
- **Variance** (σ^2): This is the measure of the spread around the mean where higher variance indicates that data points are more spread out.
- **Standard Deviation** (σ): The square root of the variance, indicating the average distance of values from the mean.
- Some distributions have special parameters different from the the ones mentioned above.

5. **Skewness and Kurtosis**:

- **Skewness** describes the asymmetry of the distribution. A skewed distribution has has more values concentrated on one side of the mean.
- **Kurtosis** measures the "tailedness" or the peak of a distribution. High kurtosis indicates more extreme outliers while low kurtosis implies a flatter distribution.

6. **Support**: This a set of all possible values a random variable can take. For example, the support for a coin is {Head (H), Tail (T)}, while for a normal distribution, the support is all real numbers, $\Re$.
7. **Shape**: The shape of a distribution refers the form. Common distribution shapes include **bell-shaped**, **U-shaped**, **skewed distribution**. The shape often influences how the data behaves and which summary statistics are most informative.

Types of distributions:

All probability distributions can be classified as discrete probability distributions or as continuous probability distributions, depending on whether they define probabilities associated with **discrete** or **continuous variables**. One should note that each distribution has unique properties and applications, from modeling biological data to analyzing financial returns.

Discrete Distributions

1. **Bernoulli Distribution**:

2. **Binomial Distribution**: The binomial distribution is a probability distribution that describes the likelihood of a fixed number of "successes" in a fixed number of independent "trials" (or events), each with the same probability of success. It's used widely in statistics to model situations with binary (yes/no, true/false, success/failure) outcomes.

Formular:

The **binomial distribution** is a probability distribution that describes the likelihood of a fixed number of "successes" in a fixed number of independent "trials" (or events), each with the same probability of success. It's used widely in statistics to model situations with binary (yes/no, true/false, success/failure) outcomes.

1. Formula

The probability of getting exactly **k** successes in n independent trials, with probability of success p, is given by:

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

where: - $ P(X = k)$ is the probability of observing k successes, - $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the **binomial coefficient** (also called "n choose k"), representing the number of ways to achieve k successes in n trials, - p is the probability of success in a single trial, - $(1-p)$ is the probability of failure in a single trial, - (n) is the total number of trials.

2. Parameters of the Binomial Distribution

The binomial distribution is defined by two parameters: - **(n)** (number of trials): the fixed number of independent trials or experiments. - **(p)** (probability of success): the probability of a success in each individual trial.

3. Real-World Applications of the Binomial Distribution

The binomial distribution is used in a variety of real-world scenarios where outcomes are binary. Here are some examples:

- **Quality Control**: Determining the probability of finding a certain number of defective products in a batch.
- **Survey Analysis**: Estimating the probability that a certain number of people in a survey sample will agree or disagree with a statement.
- **Medical Trials**: Assessing the likelihood that a certain number of patients will experience a successful outcome (e.g., response to a treatment) in a clinical trial.
- **Finance**: Calculating the probability of achieving a specific number of wins or losses in a series of trades, where each trade is either a success or failure.

4. Shape of a binomial distribution

The shape of the binomial distribution graph depends on the parameters (n) and (p): - **Symmetry**: When $(p = 0.5)$ and (n) is large, the distribution is symmetrical. When (p) differs from 0.5, it becomes skewed. - **Skewness**: For $(p < 0.5)$, the distribution is skewed to the right. For $(p > 0.5)$, it's skewed to the left. - **Peakedness**: As (n) increases, the distribution becomes more peaked and begins to resemble a normal distribution due to the Central Limit Theorem.

For a binomial distribution with different values of (p), the graph looks like a series of bars (discrete distribution) indicating the probability of each possible number of successes from 0 up to (n).

3. **Geometric Distribution**: Counts the number of trials until the first success.
4. **Negative Binomial Distribution**: Counts the number of trials until a specified number of successes occurs.
5. **Poisson Distribution**: For the number of events in a fixed interval, given a constant rate of occurrence.

6. **Multinomial Distribution**: Generalization of the binomial distribution for more than two outcomes.
7. **Hypergeometric Distribution**: For sampling without replacement, often used in quality control.

Continuous Distributions

1. **Normal (Gaussian) Distribution**: The "bell curve," used widely in natural and social sciences.
2. **Log-Normal Distribution**: Models a variable whose logarithm is normally distributed.
3. **Uniform Distribution**: All outcomes in a specified range are equally likely.
4. **Exponential Distribution**: Models the time between events in a Poisson process.
5. **Gamma Distribution**: Generalizes the exponential distribution, useful in queuing models.
6. **Beta Distribution**: For variables constrained between 0 and 1, often used in Bayesian statistics.
7. **Weibull Distribution**: Often used in reliability analysis and survival studies.
8. **Pareto Distribution**: For variables that follow a power-law distribution, useful in economics.
9. **Cauchy Distribution**: Has heavy tails, and mean/variance are undefined.
10. **Chi-Square Distribution**: Used in hypothesis testing, particularly in tests of independence.

Multivariate Distributions

1. **Multivariate Normal Distribution**: Generalization of the normal distribution to multiple variables.
2. **Multivariate T-Distribution**: Similar to the multivariate normal but with heavier tails.
3. **Wishart Distribution**: A distribution over covariance matrices, useful in multivariate analysis.
4. **Dirichlet Distribution**: The multivariate generalization of the Beta distribution, often used in Bayesian models.

Other Important Distributions

1. **Student's T-Distribution**: Useful when sample sizes are small and population variance is unknown.
2. **F-Distribution**: Used in analysis of variance (ANOVA).
3. **Laplace Distribution**: Used for data with sharp peaks and heavy tails.

4. **Rayleigh Distribution**: Useful in signal processing, describing the distribution of magnitudes of a vector.
5. **Logistic Distribution**: Similar to normal distribution but with heavier tails; used in logistic regression.

Mixture Distributions

1. **Gaussian Mixture Model (GMM)**: Represents a mixture of multiple normal distributions, used in clustering.
2. **Hidden Markov Models (HMMs)**: Used for sequential data where each state has its own distribution.

Survival and Reliability Distributions

1. **Gumbel Distribution**: Often used to model extreme values.
2. **Frechet Distribution**: Another distribution used in extreme value theory.
3. **Log-Logistic Distribution**: Used in survival analysis and reliability modeling.

References

1. Stat Trek
2. SkillsYouNeed

Schumacker, Randall E. 2017. *Statistical Distributions*. https://doi.org/10.4135/9781506300160.n7.