Statistics Theory

Siim Erik Pugal

January 2023

1 Relation between Random Variable, Its Distribution and Probability

A random variable is a variable that can take on different values randomly. The distribution of a random variable is a description of the possible values that the random variable can take on, and the probabilities associated with those values.

For example, consider the random variable X that represents the number of heads that appears when a coin is flipped three times. The possible values of X are $0, 1, 2$, and 3, and the distribution of X is given by the probabilities associated with each of these values. If the coin is fair, then the probability of getting 0 heads is $(1/2)^3 = 1/8$, the probability of getting 1 head is $3 \cdot (1/2)^2 \cdot (1/2) = 3/8$, the probability of getting 2 heads is $3 \cdot (1/2) \cdot (1/2)^2 = 3/8$, and the probability of getting 3 heads is $(1/2)^3 = 1/8$.

The distribution of a random variable can be described using a probability mass function (for discrete variables) or a probability density function (for continuous variables). The probability mass function gives the probability for each possible value of the random variable, and the probability density function gives the probability for any value within a range.

For example, the probability mass function for the random variable X described above is given by:

$$f(x) = \begin{cases} 1/8 & \text{if } x = 0 \\ 3/8 & \text{if } x = 1 \\ 3/8 & \text{if } x = 2 \\ 1/8 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

In general, the probability of a given event is given by the sum of the probabilities of all the possible outcomes that make up that event. For example, the probability of getting at least one head when flipping the coin three times is the sum of the probabilities of getting 1 head, 2 heads, and 3 heads, which is $3/8 + 3/8 + 1/8 = 7/8$.

2 Random Vector, Covariation and Correlation

A random vector is a set of multiple random variables. It is a mathematical object that represents a set of variables that can vary randomly.

Covariation is a measure of how two random variables change together. If two random variables X and Y have a positive covariation, then they tend to increase or decrease together. If they have a negative covariation, then one tends to increase as the other decreases. The covariation between X and Y is given by the following equation:

$$Cov(X,Y) = K_x(X,Y) = E[(X - E[X])(Y - E[Y])]$$

where $E[X]$ and $E[Y]$ are the expected values of X and Y, respectively.

Correlation is a measure of the linear relationship between two random variables. It is a normalized version of covariation, with values ranging from -1 to 1. A value of 1 indicates a perfect positive linear relationship, a value of -1 indicates a perfect negative linear relationship, and a value of 0 indicates no linear relationship. The correlation between X and Y is given by the following equation:

$$Corr(X,Y) = R_x(X,Y) = \frac{Cov(X,Y)}{\sigma(X) \cdot \sigma(Y)}$$

where $\sigma(X)$ and $\sigma(Y)$ are the standard deviations of X and Y, respectively.

3 Assumptions and Statement of Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) is a fundamental statistical theorem that states that, under certain conditions, the mean of a large number of independent, identically distributed random variables will be approximately normally distributed, regardless of the distribution of the individual variables.

There are several assumptions that are made in the statement of the CLT:

1. The random variables are independent: This means that the value of one random variable does not affect the value of the others.

2. The random variables are identically distributed: This means that they all have the same probability distribution.

3. The sample size is large: The CLT is typically applied to samples with sizes of at least 30, although the precise value needed for the sample size to be considered "large" may vary depending on the specific application.

The statement of the CLT is as follows:

If $X_1, X_2, ..., X_n$ are independent and identically distributed random variables with mean μ and standard deviation σ, and if Y is the mean of these random variables (i.e. $Y = (X_1 + X_2 + ... + X_n)/n$), then the distribution of Y will be approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$ as n becomes large.

In other words, the CLT states that the distribution of the mean of a large number of independent, identically distributed random variables will be approximately normal, regardless of the distribution of the individual variables. This result is extremely useful for statistical inference, because it allows us to use normal distributions to approximate the distribution of a sample mean and make statistical conclusions about a population based on sample data.

4 Random process and its stationarity

A random process is a mathematical model that describes a sequence of events or variables that evolve over time and are governed by chance. It is a way to represent the behavior of a system that exhibits randomness.

A random process is said to be stationary if the statistical properties of the process do not change over time. In other words, the mean, variance, and other statistical properties of the process are constant over time.

There are two types of stationarity: weak stationarity and strict stationarity. A random process is said to be weakly stationary if the mean of the process is constant over time and the autocovariance of the process depends only on the time lag between observations, not on the specific time at which the observations are made. A random process is said to be strictly stationary if it is both weakly stationary and has a constant variance.

An example of a stationary process is a random walk, in which the value of the process at each time step is determined by the value at the previous time step plus a random disturbance. The mean and variance of the random walk are constant over time, so it is a stationary process.

On the other hand, a random process that exhibits a trend or a changing variance over time is not stationary. For example, a random process that represents the stock price of a company would not be stationary, because the stock price is likely to change over time due to various factors such as company performance, market conditions, and so on.

5 The essence of Markov-Chain. Kolmogorov-Chapman equation

A Markov chain is a mathematical model that represents a sequence of events in which the probability of each event depends only on the state of the previous event. It is a type of random process that is used to model systems that evolve over time and exhibit a certain degree of memorylessness.

The essence of a Markov chain is that it represents a sequence of events in which the future depends only on the present, not on the past. This means that the probability of each event in the sequence depends only on the current state of the system, and not on the sequence of events that led up to it.

The Kolmogorov-Chapman equation is a fundamental equation in the theory of Markov chains. It is used to determine the transition probabilities between states in a Markov chain. The equation is given by:

$$P(X_t = j | X_0 = i) = \sum_{k=1}^{t-1} P(X_{t-1} = k | X_0 = i) \cdot P(X_t = j | X_{t-1} = k)$$

where $P(X_t = j|X_0 = i)$ is the probability of being in state j at time t, given that the process started in state i at time 0, and $P(X_t = j|X_{t-1} = k)$ is the transition probability from state k to state j.

The Kolmogorov-Chapman equation is often used to analyze the long-term behavior of Markov chains, such as the probability of being in a particular state at a given time, or the expected number of steps needed to reach a certain state. It is a key tool in the analysis of

Markov chains and has many applications in fields such as economics, computer science, and engineering.

6 Poisson process. Relation between Poisson, exponential and gamma distribution.

A Poisson process is a type of continuous-time random process that is used to model the occurrence of events over time. It is characterized by a constant rate of occurrence, meaning that the probability of an event occurring in a given time interval is proportional to the length of the interval.

The countable random process $N(t)$ is called a Poisson process if

1. $N(0) = 0$.
2. $N(s) \geq N(t)$ if $s > t$.
3. All $N(t_2 - t_1)$ and $N(t_4 - t_3)$ are independent, $t_4 > t_3 > t_2 > t_1$.
4. Number of events in time interval t have a Poisson distribution with mean value λt. That means
$$P[N(t+s) - N(s) = n] = \frac{(\nu t)^n}{n!} \exp(-\nu t)$$
where $n = 0, 1, 2, \ldots$.

The Poisson process is related to the exponential and gamma distributions. The exponential distribution is often used to model the time between events in a Poisson process, and the gamma distribution is used to model the sum of a large number of independent, exponentially distributed random variables.

Poisson and exponential distributions are very strongly related but they're fundamentally different because the Poisson is discrete (a count variable) and the exponential is continuous (a waiting time).

How are they related? If the time between a certain type of event is exponentially distributed with rate λ, then the number of events in a given time period of length t follows a Poisson distribution with parameter λt.

For example, if shooting stars appear in the sky at a rate of λ per unit time, then the time you wait until you see your first shooting star is distributed exponentially with rate λ. If you watch the night sky for t units of time, then you could see $0, 1, 2, \ldots$ shooting stars. The number of shooting stars that you count in this time is a $Poisson(\lambda t)$ random variable.

How long must I wait before I see n shooting stars? The answer is a sum of independent exponentially distributed random variables, and it follows a $gamma(\lambda, n)$ distribution (also sometimes called an Erlang distribution, to distinguish it from the general gamma distribution where n is allowed to be a non-integer).

The probability density function (PDF) of the exponential distribution is given by:

$$f(x) = \lambda \cdot e^{-\lambda \cdot x}$$

where λ is the rate parameter, which determines the shape of the distribution. The mean and variance of the exponential distribution are both equal to $1/\lambda$.

The PDF of the gamma distribution is given by:

$$f(x) = \frac{1}{\Gamma(k)} \cdot (\lambda \cdot x)^{k-1} \cdot e^{-\lambda \cdot x}$$

where λ is the rate parameter and k is the shape parameter. The mean and variance of the gamma distribution are both equal to k/λ.

The Poisson process is often used to model the occurrence of rare events, such as radioactive decay, the arrival of customers at a store, or the occurrence of earthquakes. It is a widely used model in many fields, including physics, engineering, and computer science.

7 Markov Chain with continuous time. Birth and Death process with an example

A Markov chain with continuous time is a type of random process that evolves over time according to the rules of a Markov chain, but with the added constraint that the time between events is continuous rather than discrete. This means that the probability of an event occurring in a given time interval is proportional to the length of the interval, as in a Poisson process.

A birth and death process is a specific type of Markov chain with continuous time that models the dynamics of a system in which the state can change by either increasing or decreasing by one unit at a time. It is often used to model systems in which there is a balance between the creation and destruction of objects, such as in population dynamics or in queueing theory.

An example of a birth and death process is a queueing system in which customers arrive at a rate of λ (Pihlak uses μ to represent birth) and are served at a rate of μ (Pihlak uses ν to represent death). In this case, the state of the system is the number of customers in the queue, and the system can transition to a higher state (increase in the number of customers) when a new customer arrives, or to a lower state (decrease in the number of customers) when a customer is served.

The probability of transitioning from one state to another in a birth and death process is governed by the following transition rates:

$$P(X_t + dt = i + 1 | X_t = i) = \lambda_i \cdot dt$$

$$P(X_t + dt = i - 1 | X_t = i) = \mu_i \cdot dt$$

where λ_i is the rate at which the system transitions to a higher state from state i, and μ_i is the rate at which it transitions to a lower state. The transition rates are constant over time, so the process is a Markov chain with continuous time.

8 Brownian motion and normal distribution

Brownian motion, also known as the Brownian random walk or the Wiener process, is a mathematical model that describes the random motion of a particle due to the constant bombardment of surrounding molecules. It is named after the Scottish botanist Robert Brown, who observed the random motion of pollen particles suspended in water under a microscope in the 19th century.

The path of a particle undergoing Brownian motion is a random walk, meaning that it consists of a series of discrete steps in which the particle moves in a random direction. The steps are typically assumed to be normally distributed, with a mean of 0 and a variance that increases linearly with time.

The normal distribution is a continuous probability distribution that is often used to model data that are symmetrically distributed around a mean. It is characterized by its mean and variance, and is defined by the following probability density function:

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

where μ is the mean and σ is the standard deviation.

In the context of Brownian motion, the normal distribution is used to model the distribution of the steps that the particle takes. The mean of the distribution is 0, because the particle is equally likely to move in any direction, and the variance is proportional to the length of the time interval over which the motion is observed.

The Brownian motion model has many applications in fields such as physics, finance, and biology, and has played a key role in the development of modern probability theory and statistical physics.

9 Statistic as estimator. Mean Square Error (MSE). Maximum Likelihood Estimation (MLE). Interval Estimation

A statistic is a random variable that is composed on the basis of a sample and is used to estimate a parameter, denoted as θ_j. The estimator of the parameter, $\hat{\theta}_j$, is denoted as $T(X)$ and can have a realization that is a rational number, interval, or set. The probability of the parameter taking the value θ_j is denoted as $\alpha = P(T(X) = \theta_j)$ in the discrete case and in the continuous case, the probability of the parameter belonging to a set $\mathcal{H}\alpha$ is denoted as $\alpha = P(T(X) \in \mathcal{H}\alpha)$. This can also be denoted as $\alpha = P(\hat{\theta}_j = \theta_j)$ or $\alpha = P(\hat{\theta}_j \in \mathcal{H}\alpha)$.

The mean square error (MSE) of an estimator is a measure of the accuracy of the estimator. It is defined as the average squared difference between the estimator and the true value of the population parameter being estimated. The MSE is given by the following equation:

$$\text{MSE}(\hat{\theta}_j) = E(\hat{\theta}_j - \theta_j)^2$$

$\hat{\theta}_j$ - mean square error of point estimator; θ_j - mean square error of parameter. Another way to define is:

$$\text{MSE} = E[(\textsc{estimator} - \textsc{true value})^2]$$

where $E[...]$ denotes the expected value. The MSE is often used to compare the performance of different estimators, with a lower MSE indicating a better estimator.

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model based on the observed data. It involves finding the parameter values that maximize the likelihood function, which is a function that expresses the probability of the observed data given the model and the parameter values. In other words, the MLE is the parameter value that maximizes the likelihood function.

Interval estimation is a statistical method that is used to estimate the value of a population parameter by constructing a confidence interval around the estimate. A confidence interval is an interval of values that is calculated from the sample data and is expected to contain the true value of the population parameter with a certain level of confidence. The confidence level is typically chosen to be 95% or 99%.

For example, suppose we want to estimate the mean height of a population of adult males. We take a sample of n adult males and measure their heights. The sample mean is our estimate of the population mean, and we can use this estimate to construct a confidence interval around the mean. The confidence interval will have a certain width, which depends on the sample size, the level of confidence, and the variance of the sample. The larger the sample size and the smaller the variance, the narrower the confidence interval will be.

10 Statistical Tests. How to interpret probability p-value?

Statistical tests are procedures that are used to assess the evidence provided by a sample of data against a hypothesis about a population parameter. They involve calculating a test statistic based on the sample data and comparing it to a reference distribution to determine the likelihood of obtaining the observed data if the hypothesis were true.

The probability p-value is a measure of the statistical significance of the observed data. It is the probability of obtaining a test statistic that is at least as extreme as the one observed, given that the null hypothesis is true. The null hypothesis is a statement that is assumed to be true unless there is sufficient evidence to reject it.

If the p-value is low (typically less than 0.05), it indicates that the observed data are unlikely to have occurred by chance if the null hypothesis is true, and therefore there is evidence to reject the null hypothesis. On the other hand, if the p-value is high (greater than 0.05), it indicates that the observed data are consistent with the null hypothesis and there is not enough evidence to reject it.

It is important to note that the p-value does not tell us whether the null hypothesis is actually true or false, but only whether the observed data provide sufficient evidence to reject it. Therefore, it is important to interpret the p-value in the context of the research question and the specific characteristics of the sample data.

11 General Linear Model (GLM). Least Square Estimation (LSE)

The general linear model (GLM) is a flexible statistical model that can be used to analyze the relationship between a response variable and one or more predictor variables. It is a generalization of the linear regression model that allows for the response variable to be modeled using different types of probability distributions.

In the GLM, the response variable is modeled as a linear combination of the predictor variables and an error term, which is assumed to be independently and identically distributed according to a specified probability distribution. The parameters of the model (the coefficients of the predictor variables and the parameters of the error distribution) are estimated using a method called least squares estimation (LSE).

LSE is a method of estimating the parameters of a statistical model by minimizing the sum of the squared differences between the observed data and the model's predicted values. It is a widely used method for fitting linear models, and is often referred to as ordinary least squares (OLS) when applied to linear regression models.

The GLM is a widely used statistical model that has many applications in fields such as psychology, economics, and biology. It is particularly useful for modeling data that are not well-described by a normal distribution, such as data that are skewed or have outliers.

12 Differences between classical and non-parametric Statistics

Classical statistics and non-parametric statistics are two broad categories of statistical methods that are used to analyze data and make statistical inferences about a population.

Classical statistics is based on the assumption that the data follows a particular probability distribution, such as the normal distribution. It involves estimating the parameters of the distribution from the sample data and using these estimates to make statistical inferences about the population. Classical statistical methods include techniques such as t-tests, ANOVA, and linear regression.

Non-parametric statistics, on the other hand, does not make any assumptions about the underlying probability distribution of the data. It is based on the ranks or order of the data rather than the actual values, and is therefore more robust to departures from normality. Non-parametric statistical methods include techniques such as the Wilcoxon rank-sum test, the Kruskal-Wallis test, and the Spearman rank correlation coefficient.

One key difference between classical and non-parametric statistics is the level of assumptions that they make about the data. Classical statistics assumes that the data follows a particular probability distribution and requires that the sample size be sufficiently large in order to make reliable inferences. Non-parametric statistics, on the other hand, does not make any assumptions about the distribution of the data and is generally more robust to deviations from the assumptions of the model. However, non-parametric methods may be less powerful than classical methods and may require a larger sample size to achieve the same level of statistical significance.

13 Essence of Bayesian Inference. Prior distribution, Likelihood and Posterior distribution

Bayesian inference is a statistical method that is based on the principles of Bayesian probability. It involves updating our beliefs about the probability of an event or the value of a parameter based on new evidence or data.

In Bayesian inference, we begin by specifying a prior distribution, which represents our initial belief about the probability of an event or the value of a parameter. This prior belief can be based on previous knowledge or experience, or it can be a subjective belief.

Next, we observe some data or evidence and use it to update our belief about the probability of the event or the value of the parameter. This updating process is based on the likelihood of the data given the event or parameter, which represents the probability of observing the data if the event or parameter has a certain value.

The updated belief about the probability of the event or the value of the parameter is called the posterior distribution. It is obtained by combining the prior distribution and the likelihood of the data using Bayes' theorem. The posterior distribution represents our updated belief about the event or parameter based on the data that we have observed.

Bayesian inference is a flexible and powerful statistical method that allows us to incorporate prior knowledge and subjective beliefs into the analysis of data, and to update our beliefs as new data become available. It has many applications in fields such as machine learning, economics, and biology.