## How do we check if a variable follows the normal distribution?

1. Plot a histogram out of the sampled data. If you can fit the bell-shaped "normal" curve to the histogram, then the hypothesis that the underlying random variable follows the normal distribution can not be rejected.
2. Check Skewness and Kurtosis of the sampled data. Skewness = 0 and kurtosis = 3 are typical for a normal distribution, so the farther away they are from these values, the more non-normal the distribution.
3. Use Kolmogorov-Smirnov or/and Shapiro-Wilk tests for normality. They take into account both Skewness and Kurtosis simultaneously.
4. Check for Quantile-Quantile plot. It is a scatterplot created by plotting two sets of quantiles against one another. Normal Q-Q plot place the data points in a roughly straight line.

## What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

* The model that has high training accuracy might have low test accuracy. Without further knowledge, it is hard to know which dataset represents the population data and thus the generalizability of the algorithm is hard to measure. This should be mitigated by repeated splitting of train vs test dataset (as in cross validation).
* When there is a change in data distribution, this is called the dataset shift. If the train and test data has a different distribution, then the classifier would likely overfit to the train data.
* This issue can be overcome by using a more general learning method.
* This can occur when:
  + P(y|x) are the same but P(x) are different. (covariate shift)
  + P(y|x) are different. (concept shift)
* The causes can be:
  + Training samples are obtained in a biased way. (sample selection bias)
  + Train is different from test because of temporal, spatial changes. (non-stationary environments)
* Solution to covariate shift
  + importance weighted cv

## What is alpha- and beta-values?

Alpha is also known as the level of significance. This represents the probability of obtaining your results due to chance. The smaller this value is, the more “unusual” the results, indicating that the sample is from a different population than it’s being compared to, for example. Commonly, this value is set to .05 (or 5%), but can take on any value chosen by the research not exceeding .05.

Alpha also represents your chance of making a **Type I Error**. What’s that? The chance that you reject the null hypothesis when in reality, you should fail to reject the null hypothesis. In other words, your sample data indicates that there is a difference when in reality, there is not. Like a false positive.

The other key-value relates to the power of your study. Power refers to your study’s ability to find a difference if there is one. It logically follows that the greater the power, the more meaningful your results are. Beta = 1 – Power. Values of beta should be kept small, but do not have to be as small as alpha values. Values between .05 and .20 are acceptable.

Beta also represents the chance of making a Type II Error. As you may have guessed, this means that you came to the wrong conclusion in your study, but it’s the opposite of a Type I Error. With a Type II Error, you incorrectly fail to reject the null. In simpler terms, the data indicates that there is not a significant difference when in reality there is. Your study failed to capture a significant finding. Like a false negative.

## Bayesian and frequentist probabilities

The Bayesian interpretation uses probability to specify the degree of uncertainty that the user has about an event. It is sometimes referred to as “subjective probability” or “degree of belief”. The frequentist interpretation considers the relative frequencies of events of interest to the total number of events that occurred. The probability of an event is defined as the relative frequency of the event in the limit when one has infinite data.

## What is difference between Probability and Statistics?

Probability theory and statistics are often presented together, but they concern different aspects of uncertainty. One way of contrasting them is by the kinds of problems that are considered. Using probability, we can consider a model of some process, where the underlying uncertainty is captured by random variables, and we use the rules of probability to derive what happens. In statistics, we observe that something has happened and try to figure out the underlying process that explains the observations. In this sense, machine learning is close to statistics in its goals to construct a model that adequately represents the process that generated the data. We can use the rules of probability to obtain a “best-fitting” model for some data.

## Discrete and Continuous Probabilities

Depending on whether the target space is discrete or continuous, the natural way to refer to distributions is different. When the target space T is discrete, we can specify the probability that a random variable X takes a particular value x ∈ T , denoted as P(X = x). The expression P(X = x) for a discrete random variable X is known as the probability mass function. When the target space T is continuous, e.g., function the real line R, it is more natural to specify the probability that a random variable X is in an interval, denoted by P(a ⩽ X ⩽ b) for a < b. By convention, we specify the probability that a random variable X is less than a particular value x, denoted by P(X ⩽ x). The expression P(X ⩽ x) for cumulative a continuous random variable X is known as the cumulative distribution function

## Joint, marginal and conditional probabilities

For two random variables X and Y , the probability that X = x and Y = y is (lazily) written as p(x, y) and is called the joint probability. One can think of a probability as a function that takes state x and y and returns a real number, which is the reason we write p(x, y). The marginal probability that X takes the value x irrespective of the value of random variable Y is (lazily) written as p(x). We write X ∼ p(x) to denote that the random variable X is distributed according to p(x). If we consider only the instances where X = x, then the fraction of instances (the conditional probability) for which Y = y is written (lazily) as p(y | x).

## PDF (Probability Density Function)

Изображение выглядит как текст, Шрифт, белый, рукописный текст

Автоматически созданное описание

A function f : RD → R is called a probability density function (pdf ) if probability density function

1. ∀x ∈ R pdf D : f(x) ⩾ 0

2. Its integral exists and Z RD f(x)dx = 1.

For probability mass functions (pmf) of discrete random variables, the integral in is replaced with a sum

## PMF (Probability Mass Function)

Изображение выглядит как текст, Шрифт, белый, каллиграфия

Автоматически созданное описание

*PMF is a statistical term that describes the probability distribution of the****Discrete****random variable*

People often get confused between PDF and PMF. The PDF is applicable for continues random variable while PMF is applicable for discrete random variable For e.g, Throwing a dice (You can only select 1 to 6 numbers (countable) )

A **probability mass function** (**PMF**) is a function that gives the probability that a discrete random variable is exactly equal to some value. The PMF does not work for continuous random variables, because for a continuous random variable P(X=x)=0 for all x∈R. Instead, we can usually define the **probability density function (PDF)**. The PDF is the **density** of probability rather than the probability mass:



## CDF (Cumulative Distribution Function)

Изображение выглядит как текст, Шрифт, белый, Графика

Автоматически созданное описание

A cumulative distribution function (cdf) of a multivariate real-valued random variable X with states x ∈ RD is given by FX(x) = P(X1 ⩽ x1, . . . , XD ⩽ xD), where X = [X1, . . . , XD] ⊤, x = [x1, . . . , xD] ⊤, and the right-hand side represents the probability that random variable Xi takes the value smaller than or equal to xi . There are cdfs, which do not have corresponding pdfs. The cdf can be expressed also as the integral of the probability density function f(x)

## Bayes’ Theorem

In machine learning and Bayesian statistics, we are often interested in making inferences of unobserved (latent) random variables given that we have observed other random variables. Let us assume we have some prior knowledge p(x) about an unobserved random variable x and some relationship p(y | x) between x and a second random variable y, which we can observe. If we observe y, we can use Bayes’ theorem to draw some conclusions about x given the observed values of y.

p(x) is the prior, which encapsulates our subjective prior prior knowledge of the unobserved (latent) variable x before observing any data. We can choose any prior that makes sense to us, but it is critical to ensure that the prior has a nonzero pdf (or pmf) on all plausible x, even if they are very rare.

The likelihood p(y | x) describes how x and y are related, and in the case of discrete probability distributions, it is the probability of the data y if we were to know the latent variable x. Note that the likelihood is not a distribution in x, but only in y. We call p(y | x) either the “likelihood of x (given y)” or the “probability of y given x” but never the likelihood of y.

The posterior p(x | y) is the quantity of interest in Bayesian statistics posterior because it expresses exactly what we are interested in, i.e., what we know about x after having observed y.

p(y) := Z p(y | x)p(x)dx = EX[p(y | x)] is the marginal likelihood/evidence.

The marginal likelihood is independent of x, and it ensures that the posterior p(x | y) is normalized. The marginal likelihood can also be interpreted as the expected likelihood where we take the expectation with respect to the prior p(x).

## Explain Prior, Posterior, Likelihood

**Likelihood function** (often simply called the **likelihood**) measures how well a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) explains [observed data](https://en.wikipedia.org/wiki/Realization_(probability)) by calculating the probability of seeing that data under different [parameter](https://en.wikipedia.org/wiki/Statistical_parameter) values of the model. – p(X| θ)

A **prior probability distribution** of an uncertain quantity, often simply called the **prior**, is its assumed [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) before some evidence is taken into account – p(θ)

The **posterior probability** is a type of [conditional probability](https://en.wikipedia.org/wiki/Conditional_probability) that results from [updating](https://en.wikipedia.org/wiki/Bayesian_updating) the [prior probability](https://en.wikipedia.org/wiki/Prior_probability) with information summarized by the [likelihood](https://en.wikipedia.org/wiki/Likelihood_function) via an application of [Bayes' rule](https://en.wikipedia.org/wiki/Bayes%27_rule). – p(θ | x)

Bayes rule:

P(θ|x) = p(x| θ) \* p(θ) / p(x)

p ( x ) = ∫ p ( x | θ ) p ( θ ) d θ

## Covariance, Variance, Correlation

The covariance between two univariate random variables X, Y ∈ R is given by the expected product of their deviations from their respective means, i.e., CovX,Y [x, y] := EX,Y [(x − EX[x])(y − EY [y])]

By using the linearity of expectations, the expression in Definition 6.5 can be rewritten as the expected value of the product minus the product of the expected values, i.e., Cov[x, y] = E[xy] − E[x]E[y] .

The covariance of a variable with itself Cov[x, x] is called the variance is denoted by VX[x]. The square root of the variance is called the standard deviation and is often denoted by σ(x). The notion of covariance can be generalized to multivariate random variables.

The correlation between two random variables X, Y is given by corr[x, y] = Cov[x, y] /sqrt(V[x]V[y]) ∈ [−1, 1] . The correlation matrix is the covariance matrix of standardized random variables, x/σ(x). In other words, each random variable is divided by its standard deviation (the square root of the variance) in the correlation matrix. The covariance (and correlation) indicate how two random variables are related. Positive correlation corr[x, y] means that when x grows, then y is also expected to grow. Negative correlation means that as x increases, then y decreases.

## Gaussian Mixture

Consider a mixture of two univariate Gaussian densities p(x) = αp1(x) + (1 − α)p2(x), (6.80) where the scalar 0 < α < 1 is the mixture weight, and p1(x) and p2(x) are univariate Gaussian densities with different parameters, i.e., (µ1, σ2 1 ) ̸= (µ2, σ2 2 ). Then the mean of the mixture density p(x) is given by the weighted sum of the means of each random variable: E[x] = αµ1 + (1 − α)µ2 . The variance of the mixture density p(x) is given by V[x] = ασ2 1 + (1 − α)σ 2 2 + αµ2 1 + (1 − α)µ 2 2 − [αµ1 + (1 − α)µ2] 2 .

## Beta Distribution

We may wish to model a continuous random variable on a finite interval. The Beta distribution is a distribution over a continuous random variable µ ∈ [0, 1], which is often used to represent the probability for some binary event (e.g., the parameter governing the Bernoulli distribution). The Beta distribution Beta(α, β) (illustrated in Figure 6.11) itself is governed by two parameters α > 0, β > 0 and is defined as

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

Intuitively, α moves probability mass toward 1, whereas β moves probability mass toward 0. There are some special cases: For α = 1 = β, we obtain the uniform distribution U[0, 1]. For α, β < 1, we get a bimodal distribution with spikes at 0 and 1. For α, β > 1, the distribution is unimodal. For α, β > 1 and α = β, the distribution is unimodal, symmetric, and centered in the interval [0, 1], i.e., the mode/mean is at 1/2 .

## Conjugate Prior

According to Bayes’ theorem (6.23), the posterior is proportional to the product of the prior and the likelihood. The specification of the prior can be tricky for two reasons: First, the prior should encapsulate our knowledge about the problem before we see any data. This is often difficult to describe. Second, it is often not possible to compute the posterior distribution analytically. However, there are some priors that are computationally: conjugate priors.

A prior is conjugate for the likelihood function if the posterior is of the same form/type as the prior

Conjugacy is particularly convenient because we can algebraically calculate our posterior distribution by updating the parameters of the prior distribution.

Изображение выглядит как текст, Шрифт, снимок экрана, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

The Beta distribution is the conjugate prior for the parameter µ in both the Binomial and the Bernoulli likelihood. For a Gaussian likelihood function, we can place a conjugate Gaussian prior on the mean. The reason why the Gaussian likelihood appears twice in the table is that we need to distinguish the univariate from the multivariate case. In the univariate (scalar) case, the inverse Gamma is the conjugate prior for the variance. In the multivariate case, we use a conjugate inverse Wishart distribution as a prior on the covariance matrix. The Dirichlet distribution is the conjugate prior for the multinomial likelihood function.

## Are there any differences between the expected value and mean value?

**Expected value is used when we want to calculate the mean of a probability distribution**. This represents the average value we expect to occur before collecting any data. Mean is typically used when we want to calculate the average value of a given sample.

## How do you identify if a coin is biased?

We collect data by flipping the coin 200 times.

To perform a chi-square test (or any other statistical test), we first must establish our null hypothesis. In this example, our null hypothesis is that the coin should be equally likely to land head-up or tails-up every time. The null hypothesis allows us to state expected frequencies. For 200 tosses, we would expect 100 heads and 100 tails.

The Observed values are those we gather ourselves. The expected values are the frequencies expected, based on our null hypothesis. We total the rows and columns as indicated. It's a good idea to make sure that the row totals equal the column totals (both total to 400 in this example).

Using probability theory, statisticians have devised a way to determine if a frequency distribution differs from the expected distribution. To use this chi-square test, we first have to calculate chi-squared.

Chi-squared = (observed-expected)2/(expected)

We have two classes to consider in this example, heads and tails.

Now we have to consult a table of critical values of the chi-squared distribution.

The left-most column list the degrees of freedom (df). We determine the degrees of freedom by subtracting one from the number of classes. In this example, we have two classes (heads and tails), so our degrees of freedom is 1. Our chi-squared value is 1.28. Move across the row for 1 df until we find critical numbers that bound our value. In this case, 1.07 (corresponding to a probability of 0.30) and 1.64 (corresponding to a probability of 0.20). We can interpolate our value of 1.24 to estimate a probability of 0.27. This value means that there is a 73% chance that our coin is biased. In other words, the probability of getting 108 heads out of 200 coin tosses with a fair coin is 27%. In biological applications, a probability � 5% is usually adopted as the standard. This value means that the chances of an observed value arising by chance is only 1 in 20. Because the chi-squared value we obtained in the coin example is greater than 0.05 (0.27 to be precise), we accept the null hypothesis as true and conclude that our coin is fair.

## Exponential Family

An exponential family is a family of probability distributions, parameterized by θ ∈ RD, of the form p(x | θ) = h(x) exp (⟨θ, ϕ(x)⟩ − A(θ)) , (6.107) where ϕ(x) is the vector of sufficient statistics. In general, any inner product (Section 3.2) can be used in (6.107), and for concreteness we will use the standard dot product here (⟨θ, ϕ(x)⟩ = θ ⊤ϕ(x)). Note that the form of the exponential family is essentially a particular expression of gθ(ϕ(x)) in the Fisher-Neyman theorem  
Exponential families include many of the most common distributions. Among many others, exponential families includes the following:[[6]](https://en.wikipedia.org/wiki/Exponential_family#cite_note-7)

* [normal](https://en.wikipedia.org/wiki/Normal_distribution)
* [exponential](https://en.wikipedia.org/wiki/Exponential_distribution)
* [gamma](https://en.wikipedia.org/wiki/Gamma_distribution)
* [chi-squared](https://en.wikipedia.org/wiki/Chi-squared_distribution)
* [beta](https://en.wikipedia.org/wiki/Beta_distribution)
* [Dirichlet](https://en.wikipedia.org/wiki/Dirichlet_distribution)
* [Bernoulli](https://en.wikipedia.org/wiki/Bernoulli_distribution)
* [categorical](https://en.wikipedia.org/wiki/Categorical_distribution)
* [Poisson](https://en.wikipedia.org/wiki/Poisson_distribution)
* [Wishart](https://en.wikipedia.org/wiki/Wishart_distribution)
* [inverse Wishart](https://en.wikipedia.org/wiki/Inverse_Wishart_distribution)
* [geometric](https://en.wikipedia.org/wiki/Geometric_distribution)

## What Is a Statistical Interaction?

A statistical interaction is when two or more variables interact, and this results in a third variable being affected.

Examples. Real-world examples of interaction include: **Interaction between adding sugar to coffee and stirring the coffee**. Neither of the two individual variables has much effect on sweetness but a combination of the two does.

## What is a Skewed Distribution?

A skewed distribution occurs when one tail is longer than the other. Skewness defines the asymmetry of a distribution. Unlike the familiar normal distribution with its bell-shaped curve, these distributions are asymmetric. The two halves of the distribution are not mirror images because the data are not distributed equally on both sides of the distribution’s peak.

## Examples of Right-Skewed Distributions

Right skewed distributions are the more common form. These distributions tend to occur when there is a lower limit, and most values are relatively close to the lower bound. Values can’t be less than this bound but can fall far from the peak on the high end, causing them to skew positively.

For example, right skewed distributions can occur in the following cases:

* Time to failure cannot be less than zero, but there is no upper bound.
* Wait and response times cannot be less than zero, but there are no upper limits.
* Sales data cannot be less than zero but can have unusually large values.
* Humans have a minimum viable weight but can have large extreme values.
* Income cannot be less than zero, but there are some extremely high incomes.

## You have an 50-50 mixture of two normal distributions with the same standard deviation. How far apart do the means need to be in order for this distribution to be bimodal?

More than two standard deviations

## Given draws from a normal distribution with known parameters, how can you simulate draws from a uniform distribution?

Plug in the value to the CDF of the same random variable

## You are told that your regression model is suffering from multicollinearity. How do verify this is true and build a better model?

You should create a correlation matrix to identify and remove variables with a correlation above 75%. Keep in mind that our threshold here is subjective.

You could also calculate **VIF (variance inflation factor)** to check for the presence of multicollinearity.

You can’t just remove variables, so you should use a penalized regression model or add random noise in the correlated variables, but this approach is less ideal.

## Chi-square tests vs *t-*tests

Both [chi-square tests](https://www.scribbr.com/statistics/chi-square-tests/) and [*t*tests](https://www.scribbr.com/statistics/t-test/) can test for differences between two groups. However, a *t* test is used when you have a dependent quantitative variable and an [independent](https://www.scribbr.com/methodology/independent-and-dependent-variables/) categorical variable (with two groups). A [chi-square test of independence](https://www.scribbr.com/statistics/chi-square-test-of-independence/) is used when you have two categorical variables.

## Normal Distribution vs. t-Distribution: What’s the Difference?

A **Z-test** is a hypothesis test with a normal distribution that uses a **z-statistic**. A z-test is used when you know the population variance or if you don’t know the population variance but have a large sample size.

A **T-test** is a hypothesis test with a t-distribution that uses a **t-statistic**. You would use a t-test when you don’t know the population variance and have a small sample size.

## What is the standard error? What is the standard error of mean?

**Standard error** of a statistic is the standard deviation of its sampling distribution or an estimate of that standard deviation.

Using CLM, we can estimate the standard error of mean by using population standard deviation divided by the square root of sample size n. If the population standard deviation is unknown, we can use the sample standard deviation as an estimation.

## Standard Error of the Mean vs. Standard Deviation: What's the Difference?

The [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp) (SD) measures the amount of variability, or [dispersion](https://www.investopedia.com/terms/d/dispersion.asp), from the individual data values to the [mean](https://www.investopedia.com/terms/m/mean.asp). SD is a frequently-cited statistic in many applications from math and statistics to finance and investing.

[Standard error](https://www.investopedia.com/terms/s/standard-error.asp) of the mean (SEM) measures how far the sample mean (average) of the data is likely to be from the true population mean. The SEM is always smaller than the SD.

SEM is calculated simply by taking the standard deviation and dividing it by the square root of the sample size.

## Confidence Interval

**CI = x  +/-  t1-α/2, n-1\*(s/√n)**

where:

* **x:**sample mean
* **t:**the critical t-value, based on the significance level *α* and sample size *n*
* **s:**sample standard deviation
* **n:**sample size

## Sampling Distributions and the Standard Error of the Mean

Imagine you draw a random sample of 50 from a [population](https://statisticsbyjim.com/glossary/population/), measure a property, and calculate the mean. Now, suppose you repeat that study many times. You repeatedly draw random samples of the same size, calculate the mean for each sample, and graph all the means on a histogram. Ultimately, the histogram displays the distribution of sample means for random samples of size 50 for the characteristic you’re measuring.

[Statisticians](https://statisticsbyjim.com/glossary/statistics/) call this type of distribution a sampling distribution. And, because we’re calculating the mean, it’s the sampling distribution of the mean. There’s a different sampling distribution for each sample size.

This distribution is the sampling distribution for the above experiment. Remember that the curve describes the distribution of sample means and not individual observations. Like other distributions, sampling distributions have a central location and variability around that center.

* The center falls on the population mean because random sampling tends to converge on this value.
* The variability, or spread, describes how far sample means tend to fall from the population mean.

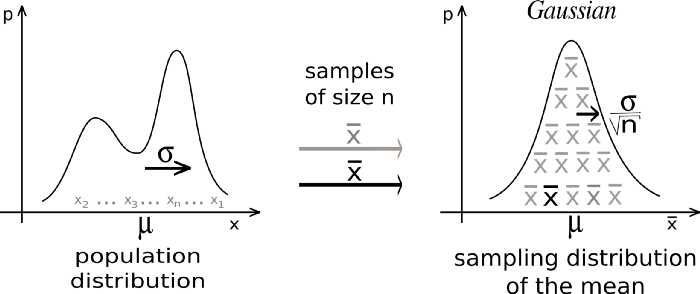
The wider the distribution, the further the sample means tend to fall from the population mean. That’s not good when you’re using sample means to estimate population means! You want narrow sampling distributions where sample means fall near the population mean.

The variability of the sampling distribution is the standard error of the mean! More specifically, the SEM is the standard deviation of the sampling distribution. For the example sampling distribution, the SEM is 3. We’ll interpret that value shortly.

The mean of the sampling distribution of sample means is mean.

## What is the Central Limit Theorem (CLM)?

The [**Central Limit Theorem**](https://en.wikipedia.org/wiki/Central_limit_theorem)states that no matter what is the population’s original distribution, when taking random samples from the population, the distribution of the means or sums from the random samples approaches a normal distribution, with mean equals to the population mean, as the random sample size gets larger



## What general conditions must be satisfied for the central limit theorem to hold?

1. The data must be sampled randomly
2. The sample values must be independent of each other
3. The sample size must be sufficiently large, generally it should be greater or equal than 30

## What does it mean if a model is heteroscedastic? what about homoscedastic?

A model is heteroscedastic when the variance in errors is **not**consistent. Conversely, a model is homoscedastic when the variances in errors is consistent.

## What does Design of Experiments mean?

**Design of experiments** also known as DOE, it is the design of any task that aims to describe and explain the variation of information under conditions that are hypothesized to reflect the variable. In essence, an experiment aims to predict an outcome based on a change in one or more inputs (independent variables).

## What does the Poisson distribution represent?

The Poisson distribution is a discrete distribution that gives the probability of the number of independent events occurring in a fixed time. An example of when you would use this is if you want to determine the likelihood of X patients coming into a hospital in a given hour.

The mean and variance are both equal to λ.

## If you had draws from a normal distribution with known parameters, how would you simulate draws from a uniform distribution?

A question like this tests your knowledge of the concepts of uniform and normal distributions.  
There’s a simple answer to this. To simulate draws from a uniform distribution, you would plug the values into the normal cumulative distribution function (CDF) for the same random variable.

This is known as the Universality of the Uniform or Probability Integral Transform.

## How do you transform a Skewed Distribution into a Normal Distribution?

To transform a **Skewed Distribution** into a **Normal Distribution** we apply some **linearized function** on it. Some common functions that achieve this goal are:

* **Logarithmic function**: We can use it to make *extremely* skewed distributions less skewed, especially for *right-skewed distributions*. The only condition is that this function is defined only for **strictly positive numbers**. $$ f(x) = ln(x) $$
* **Square root transformation n**: this one has an average effect on distribution shape: it’s weaker than *logarithmic transformation*, and it’s also used for reducing *right-skewed distributions*, but is defined only for **positive numbers**. f(x) = \sqrt{x}
* **Reciprocal transformation**: this one reverses the order among values of the same sign, so *large values* become *smaller*, but the *negative reciprocal* preserves the order among values of the same sign. The only condition is that this function is not defined for **zero values**. f(x) = 1/x
* **Exponential or Power transformation**: has a reasonable effect on distribution shape; generally, we apply power transformation (power of two usually) to reduce *left skewness*. We could also try any exponent to see which one provides better results. f(x) = x^n
* **Box-Cox Transformation**: in this transformation, we’re searching and evaluating all the other transformations and *choosing the best one*. It's defined as:

Изображение выглядит как Шрифт, белый, текст, диаграмма

Автоматически созданное описание

The exponent here is a variable called *lambda* (λ) that varies over the range of -5 to 5, and in the process of searching, we examine all values of λ. Finally, we choose the *optimal value* (resulting in the best approximation to a normal distribution) for the variable.

## What's the difference between Binomial Distribution and Geometric Distribution?

* The **Binomial distribution** describes the probability of obtaining *k* successes in *n* *Bernoulli experiments*, i.e an experiment which has only two possible outcomes, often call them *success* and *failure*. Its probability function describes the probability of getting exactly k successes in n independent *Bernoulli* trials:
* The **Geometric distribution** describes the probability of experiencing a *certain amount* of *failures* before experiencing the *first succes*s in a series of *Bernoulli experiments*. This probability is given by:

*P*(*X*=*k*)=*pk*(1−*p*)*n*−*k*

So as we can see, the key difference is that in a *binomial distribution*, there is a *fixed number of trials* meanwhile in a *geometric distribution*, we’re interested in the *number of trials* required until we obtain a *success*.

## Bayesian Inference

Focusing solely on some statistic of the posterior distribution (such as the parameter θ ∗ that maximizes the posterior) leads to loss of information, which can be critical in a system uses the prediction p(x | θ ∗ ) to make decisions. These decision-making systems typically have different objective functions than the likelihood, a squared-error loss or a mis-classification error. Therefore, having the full posterior distribution around can be extremely useful and leads to more robust decisions. Bayesian inference is about finding this posterior distribution. For a dataset X , a parameter prior p(θ), and a likelihood function, the posterior is obtained by applying Bayes’ theorem.

Parameter estimation via maximum likelihood or MAP estimation yields a consistent point estimate θ∗ of the parameters, and the key computational problem to be solved is optimization. In contrast, Bayesian inference yields a (posterior) distribution, and the key computational problem to be solved is integration. Predictions with point estimates are straightforward, whereas predictions in the Bayesian framework require solving another integration problem. However, Bayesian inference gives us a principled way to incorporate prior knowledge, account for side information, and incorporate structural knowledge, all of which is not easily done in the context of parameter estimation. Moreover, the propagation of parameter uncertainty to the prediction can be valuable in decision-making systems for risk assessment and exploration in the context of data-efficient learning.

If we do not choose a conjugate prior on the parameters, the integrals are not analytically tractable, and we cannot compute the posterior in closed form.

## Latent-Variable Models

In practice, it is sometimes useful to have additional latent variables z (besides the model parameters θ) as part of the model. These latent variables are different from the model parameters θ as they do not parametrize the model explicitly. Latent variables may describe the data-generating process, thereby contributing to the interpretability of the model. They also often simplify the structure of the model and allow us to define simpler and richer model structures. Simplification of the model structure often goes hand in hand with a smaller number of model parameters. Learning in latent-variable models (at least via maximum likelihood) can be done in a principled way using the expectation maximization (EM) algorithm.

Denoting data by x, the model parameters by θ and the latent variables by z, we obtain the conditional distribution

p(x | z, θ)

## Simpson’s paradox

**Simpson’s paradox** refers to the situations in which a trend or relationship that is observed within multiple groups disappears or reverses when the groups are combined. The quick answer to why there is Simpson’s paradox is the existence of confounding variables.