# Notes on Bayesian Analysis of Linear Models

compiled by D.Gueorguiev 4/15/24

## The Parametric Inference Problem

Let be a vector of real parameters, a vector of observations, is a known design matrix with size . Then the general linear model is

(1)

where where denotes the normal distribution. Here is the precision matrix of , which has covariance matrix , and is unknown.

This is the general linear model and our objective is to provide inferences for and when observing , where is the -th observation. Inference is the procedure which extracts information about from the sample .

In Bayesian models all references are based on the posterior distribution of , which is given by Bayes theorem.

### Bayes Theorem

Suppose the prior information about is represented by a probability density function , then Bayes theorem combines this information with the information contained in the sample. The likelihood function for and is

(2)

where, as before, . The likelihood function is the inherent sample information about the parameters and and is the conditional density function of the sample random variable given and .

Bayes theorem gives the conditional density of and , given the sample .

(3)

The posterior density of the pair is and represents one’s knowledge of and after observing the sample . Recall that the information about of and , before is observed is contained in the prior density . Note that the posterior density is denoted with the same symbol as the prior but with the conditional notation for as - this is just a notational convention which is adopted in the current text.

One can rewrite the posterior density as

, (4)

where K is the normalizing constant and is given by

which is the marginal probability density of .

Next, we will find the posterior density of and .

### Prior Information

The prior information about the parameters and is given in two ways. The first is when is a normal-gamma prior density, namely,

, (5)

where

, (6)

and is a element vector and is a known positive definite matrix. Thus is the conditional density of given and it is normal with mean vector and precision matrix .

The marginal prior density of is Gamma with parameters and

, (7)

We can construct the marginal prior density of as

, (8)

which is a Student t density with degrees of freedom, location vector and precision matrix .

## Appendix

### Chi-squared distribution

The chi-squared () distribution with degrees of freedom is the distribution of the squares of independent standard normal random variables.

If are independent standard normal variables, then the sum of their squares

is distributed according to the chi-squared distribution with k degrees of freedom. This fact is usually denoted with

or

The chi-squared distribution has one parameter: a positive integer that specifies the number of degrees of freedom.

The Probability Density Function of chi-squared distribution is given with

, ; for it is zero.

The Cumulative Distribution Function of chi-squared distribution is given with

where is the lower incomplete gamma function. is the regularized gamma function.

Additional Notes on chi-squared distribution

The reason why chi-squared distribution is used in hypothesis-testing is its relation to normal distribution.

Many hypothesis tests use the -statistic in a -test. As the sample size increases the sampling distribution of the test statistic approaches normal distribution by the central limit theorem. The square of standard normal distribution is distributed according to the chi-square distribution.

An additional use case of the chi-squared distribution is to model the large sample distribution of generalized likelihood ratio tests (LRT).

### Erlang distribution

The Erlang distribution is a two-parameter family of continuous probability distributions with support . The two parameters are:

* a positive integer , the “shape”
* a positive real number , the “rate”. The “scale”, , the reciprocal of the rate, is sometimes used instead.

The Erlang distribution is the distribution of a sum of independent exponential variables with mean each. Equivalently, it is the distribution of the time until the k-th event of a Poisson process with rate . The Erlang and the Poisson distributions are *complimentary*, in that while the Poisson distribution counts the events that occur in a fixed amount of time, the Erlang distribution counts the amount of time until the occurrence of a fixed number of events. When , Erlang becomes the exponential distribution. The Erlang distribution is a special case of the gamma distribution in which the shape of the distribution is discretized.

The PDF of the Erlang distribution is:

for

The parameter is the *shape* parameter, and the parameter is the *rate* parameter.

Equivalent parametrization used the *scale* parameter , the reciprocal of the rate parameter i.e. :

for

When the scale parameter equals , the distribution simplifies to the chi-squared distribution with degrees of freedom.

The CDF of Erlang distribution is

where is the lower incomplete gamma function and is the lower regularized gamma function.

The CDF may also be expressed as

Erlang-k distribution is defined by setting k in the PDF of the Erlang distribution.

The mean of Erlang distribution is given with .

### Gamma distribution

The exponential distribution, the chi-squared distribution and the Erlang distribution are special cases of Gamma distribution. There are two equivalent parametrizations of the Gamma distribution widely used today:

1 ) with shape parameter and scale parameter .

2 ) with shape parameter and rate parameter

The gamma distribution is the maximum entropy probability distribution for a random variable for which is fixed and greater than zero, and and .

A random variable that is Gamma-distributed is denoted as

The corresponding probability density function in the shape-rate parametrization is

for

where is the gamma function. For all positive integers,

A graph of a function

Description automatically generated

The CDF of Gamma distribution is the regularized Gamma function:

where is the lower incomplete Gamma function.

If is a positive integer (i.e. the distribution is an Erlang distribution), the CDF has the following series expansion:

A graph of a number of numbers

Description automatically generated

### Student -distribution

Student’s -distribution (or simply ) is a continuous distribution which generalizes the standard normal distribution. Like the latter it is symmetric around zero and bell-shaped. However, has heavier tails and the amount of probability mass in the tails is controlled by the parameter .

The probability density function of is given with

For becomes the Cauchy’s distribution with and :

because . Notice that the Cauchy’s distribution has fat tails. When increases toward infinity converges to the normal distribution which has thin tails.

For it becomes the standard normal distribution which has very thin tails.

### Multivariate t-distribution

Let is a -dimensional random vector distributed according to where is matrix. Let is a scalar random vector, distributed according to . Let is a constant vector. Then the random variable has the density

and is said to be distributed as a multivariate t-distribution with parameters . The covariance matrix is (for ).

The constructive definition of the multivariate -distribution can be used as a sampling algorithm:

1. Generate and independently.
2. Compute

This formulation gives rise to the hierarchical representation of a multivariate t-distribution as a scale-mixture of normal: where indicates a gamma distribution with density proportional to , and conditionally follows .

Note on deriving multivariate generalization of the Student -distribution

The issue here is to define a probability density function of several variables that is the appropriate generalization of the for the univariate case. In one dimension (), with and we have the probability density function

We could replace the univariate density function with a function of variables such that is replaced by a quadratic function of all , . With , the obvious choice of multivariate density function is:

which is the standard but not the only choice.

An important special case is the standard bivariate t-distribution, :

. Note that .

The difficulty with the standard representation is revealed by the last formular, which does not factorize into the product of the marginal one-dimensional distributions.

**Problem with the obvious choice of multivariate density of the t-distribution**:

When is diagonal the standard representation can be shown to have zero correlation, but the marginal distributions are not statistically independent.

There is no simple formula for the cumulative distribution function (CDF) given with:

where

Conditional Distribution of the multivariate -distribution

Let the vector …

//TODO: finish the notes on Student t-distribution

### The Maximum entropy probability distribution

In statistics the maximum entropy probability distribution has entropy that is at least as great as that of all other members of a specified class of probability distributions. According to the principle of maximum entropy, if nothing is known about a distribution except that it belongs to a certain class defined in terms of specified properties or measures, then the distribution with the largest entropy represents the least-informative result.

**Definition**: entropy of discrete r.v.

Let the discrete r.v. has distribution for . Then the entropy of is defined as

if . If we replace the term with .

**Definition**: differential entropy of continuous r.v.

Many statistical distributions of interest are those for which the moments or other measurable quantities are constrained to be constants. The following theorem by Ludwig Boltzmann gives the form of the probability density under these constraints.

**Theorem**:

*Discrete case*:

Let is a (finite or infinite) discrete subset of , and we specify functions and numbers . We consider the class of all discrete r.v.’s which are supported on and which satisfy the moment conditions

for .

If there exists a member of the class which assigns positive probability to all members of and if there exists a maximum entropy distribution for , then this distribution has the following shape:

//TODO: finish notes on Maximum entropy probability distribution

### A Bit of theory on Hypothesis Testing: Location tests and Student t-test

**Definition** A *location test* is a statistical hypothesis test that compares the location parameter of a statistical population to a given constant, or that compares the location parameters of two statistical populations to each other. Most commonly, the location parameter (or parameters) of interest are expected values but location tests based on medians or other measures of location are also used.

**Definition** *One-sample location test* compares the location parameter of one sample to a given constant. In a one-sided test, it is stated before the analysis is carried out that it is only of interest if the location parameter is either larger than or smaller than the given constant, whereas in a two-sided test, a difference in either direction is of interest.

**Definition** *Two-sample location test* compares the location parameters of two samples to each other. A common situation is where the two populations correspond to research subjects who have been treated with two different treatments (one of them possibly being a control or placebo). In this case, the goal is to assess whether one of the treatments typically yields a better response than the other. In a one-sided test, it is stated before the analysis is carried out that it is only of interest if a particular treatment yields the better responses, whereas in a two-sided test, it is of interest whether either of the treatments is superior to the other.

**Definition** *Student -test* is a location test, used to determine whether the difference between two groups of responses is statistically significant. It is most commonly applied when the test statistic follows the normal distribution if the value of the scaling term in the test statistic is estimated from data. The -test most common application is to test if the means of two populations are statistically different.

**Definition** one-sample -test

One-sample -test is a location test of whether the mean of a population has a value specified in the null hypothesis. In testing the null hypothesis that the population mean is equal to a specified value one uses the statistic

where is the sample mean, is the sample standard deviation and is the sample size. The degrees of freedom used in this test are . Although the parent population does not need to be normally distributed, the distribution of the population of sample means is assumed normal. By the Central Limit Theorem, if the observations are independent and the second moment exists, then will be approximately .

**Definition** *Two-sample -test and Welch’s test*

A *two-sample t-test* is a location test to accept or overturn the null hypothesis that assumes the means of two populations are equal, assuming the variances of the populations are also equal. The *Welch’s t-test* is a location test in which the assumption that the variances of the two populations are equal is dropped. The Welch’s t-test is often referred to as *unpaired* or *independent sample t-test* as it is applied when the statistical units underlying the two samples being compared are non-overlapping.

Summary: Two-sample -tests for difference in means involve *independent samples* (aka *unpaired samples*) or *paired samples*.

Advantages of paired -tests:

1) Paired -tests achieve greater *binary hypothesis power* (that is, probability of avoiding false negatives) than unpaired tests when the paired units are similar with respect to noise sources (confounders) that are independent of membership in the two groups being compared.

2) Paired -tests can be used to reduce the effects of confounding factors in an observational study.

#### More info on Independent (unpaired) samples

The independent samples t-test is used when two separate sets of independent and identically distributed samples are obtained, and one variable from each of the two populations is compared.

Example: Suppose we are evaluating the effect of a treatment, and we enroll subjects into a clinical study. Then we randomly assign subjects to the treatment group and subjects to the control group. With this arrangement we have two independent samples and therefore we use the unpaired form of the -test.

#### More info on Paired samples

The paired samples t-test compares the means of two measurements taken from the same individual, object or related units. Paired sample -tests typically consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (aka “repeated measures” -test).

These “paired” measurements can represent things like:

* A measurement taken at two different times (e.g. pre-test and post-test score with an intervention administered between the two time points
* A measurement taken under two different conditions (e.g. completing a test under a “control” condition and an “experimental” condition)
* Measurements taken from two parts of a subject or experimental unit.

The paired samples t-test is used to establish if there is statistical difference between two time points, between two conditions, between two measurements, between a matched pair.

Example: A typical example of the repeated measures t-test would be where subjects are tested prior to treatment and the same subjects are tested again after treatment. By comparing the same patient’s numbers before and after treatment, we are effectively using each patient as their own control. In this setup the correct rejection of the null hypothesis (here: of no difference made by the treatment) can become much more likely, with statistical power increasing simply because the random interpatient variation has been eliminated.

Note: the increase of statistical power comes at a price: more tests are required, each subject must be tested twice. Because half of the sample now depends on the other half, the paired version of Student’s -test has only degrees of freedom. Normally, there is degrees of freedom (with being the total number of observations).

Note2: A paired samples t-test based on a “*matched-pairs sample*” results from an unpaired sample that is subsequently used to form a paired sample, by using additional variables that were measured along with the variable of interest. The matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of other measured variables. This approach is sometimes used in observational studies to reduce or eliminate the effects of confounding factors.

Note3: Paired sample *t*-test is sometimes referred as “*dependent samples t-test*”, *“paired t-test”,* *“repeated measures t-test”.*

Note4: the paired sample -test can only compare the means for two (and only two) related (paired) units on a continuous outcome that is normally distributed. The paired samples t-test is not appropriate for analyses involving 1) unpaired data 2) comparison between more than two units/groups 3) a continuous outcome that is not normally distributed 4) an ordinal/ranked outcome.

#### More info on test statistics

Most test statistics have the form , where and are functions of the data.

may be sensitive to the alternate hypothesis (i.e. its magnitude tends to be larger when the alternate hypothesis is true), whereas is a scaling parameter that allows the distribution of to be determined.

As an example, in the one-sample -test

where is the sample mean from a sample of size , is the standard error of the mean, is the estimate of the standard deviation of the population, and is the population mean.

The assumptions underlying a -test in the simplest form above are that:

* follows a normal distribution and mean and variance
* follows a distribution with degrees of freedom. This assumption is met when the observations used for estimating come from a normal distribution (and iid for each group).
* and are independent.

In the -test comparing the means of two independent samples, the following assumptions should be met:

* The means of the two populations being compared should follow normal distributions. Under weak assumptions this follows in large samples from the Central Limit Theorem, even when the distribution of observations in each group is not following the normal distribution.
* If using Student’s original definition of the -test, the two populations being compared should have the same variance (testable using *F-test*, *Levene’s test*, *Bartlett’s test* or the *Brown-Forsythe test*). If the sample sizes in the two groups being compared are equal, Student’s original t-test is highly robust to the presence of unequal variances. Welch’s -test is insensitive to equality of the variances regardless of whether the sample sizes are similar.
* The data used to carry out the test should either be sampled independently from the two populations being compared or be fully paired. This is in general not testable from the data, but if the data are known to be dependent (e.g. paired by test design), a dependent t*e*st must be applied. For partially paired data, the classical independent t-tests may give invalid results as the test statistic might not follow a t distribution, while the dependent t-test is sub-optimal as it discards the unpaired data.

Most two-sample t-tests are robust to all but large deviations from the assumptions.

For exactness, the -test and the -test require normality of the sample means, and the -test additionally requires that the sample variance follows a scaled distribution, and that the sample mean and sample variance be statistically independent. Normality of the individual data values is not required if these conditions are met. By the central limit theorem, sample means of moderately large samples are often approximated by a normal distribution even if the data are not normally distributed. For data which is not normally distributed, the distribution of the sample variance may deviate substantially from a distribution.

If the sample size is large, Slutsky’s theorem implies that the distribution of the sample variance has little effect on the distribution of the test statistic. That is, as sample size n increases:

because of the Central Limit theorem

using the law of large numbers

#### One- and two-tailed tests

These represent two alternative ways of computing the statistical significance of a parameter from a data set in terms of a test statistic. A *two-tailed test* (aka *two-sided test*) is appropriate if the estimated value is greater or less than a certain range of values, for example, whether a test taker may score above or below a specific range of scores. This method is used for null hypothesis testing and if the estimated value exists in the critical areas, the alternative hypothesis is accepted over the null hypothesis. A *one-tailed test* (aka *one-sided test*) is appropriate if the estimated value may depart from the reference in only one direction, left or right but not both. An example can be whether a machine produces more than one-percent defective products. In this example, if the estimated value exists in one of the one-sided critical areas, depending on the direction of interest (greater than or less than), the alternative hypothesis is accepted over the null hypothesis.

One-tailed tests are used for asymmetric distributions that have a single tail, such as the distribution, which are common in measuring the goodness of fit, or for one side of a distribution that has two tails, such as the normal distribution, which is common in estimating location.

#### Expressions for t-tests

Explicit expressions can be used to carry out t-tests are given below. In each case the formula for a test statistic that either exactly follows or closely approximates a t-distribution under null hypothesis is given.

//TODO: finish the Appendix section on Hypothesis testing

## References

[Bayesian Analysis of Linear Models, Lyle D. Bromeling, CRC Press, 2019 (Copyright 1985)](https://www.vitalsource.com/products/bayesian-analysis-of-linear-models-broemeling-v9781351464475)

[Chi-squared distribution, Wikipedia](https://en.wikipedia.org/wiki/Chi-squared_distribution)

[Student’s -distribution, Wikipedia](https://en.wikipedia.org/wiki/Student%27s_t-distribution)

[Gamma distribution, Wikipedia](https://en.wikipedia.org/wiki/Gamma_distribution)

[Erlang distribution, Wikipedia](https://en.wikipedia.org/wiki/Erlang_distribution)

[Maximum entropy probability distribution, Wikipedia](https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution)

[Student’s -test, Wikipedia](https://en.wikipedia.org/wiki/Student%27s_t-test)

[Location test, Wikipedia](https://en.wikipedia.org/wiki/Location_test)

[One- and two-tailed tests, Wikipedia](https://en.wikipedia.org/wiki/One-_and_two-tailed_tests)

[Paired difference test, Wikipedia](https://en.wikipedia.org/wiki/Paired_difference_test)

[Incomplete Gamma function, Wikipedia](https://en.wikipedia.org/wiki/Incomplete_gamma_function)