# Parametric models and the Central Limit Theory

## Inductive reasoning

* Inductive reasoning uses knowledge about a sample to draw conclusions about the underlying population.
* Extrapolating from specific cases to general rules.
* Inductive reasoning is inherently uncertain.

When using approaches to statistical inference, it is a useful methodology to fit our data to one low-dimensional parametric model, using families of probability densities where the dimension of a parameter μ is no greater than 5, 10 or 20.

## Univariate Families

The sample space X of observation x is a subset of the real line R1 (basically means that we have only one variable).

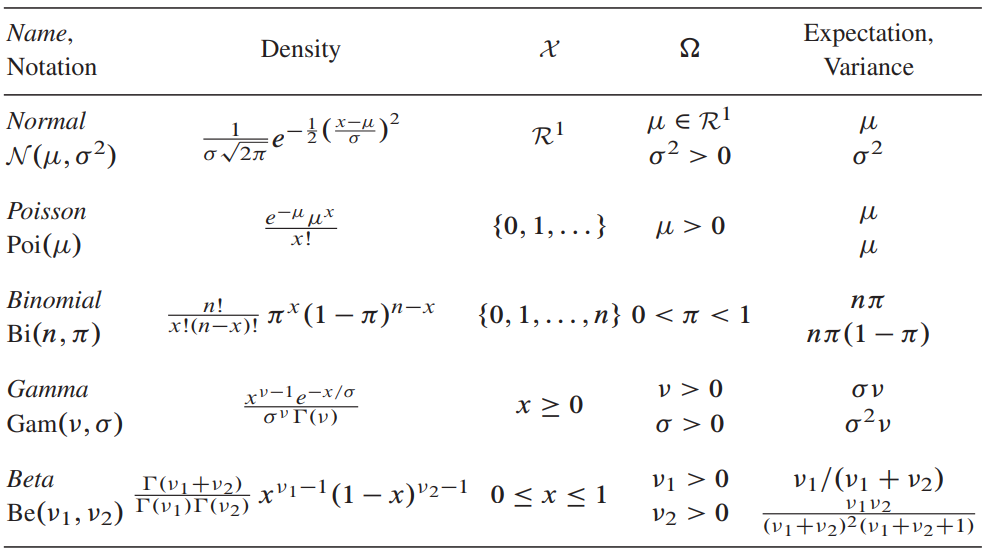


Figure 1: Five most common univariate families.

Each family have different sample spaces, making them appropriate in different situations.

## Multivariate Normal Distribution

The multivariate sample space X exists in Rp, a p-dimensional Euclidean space where p > 1.

A random vector x = (x1, x2, …, xp) has mean vector:

And p x p covariance matrix:

This basically means the expected value for an observation x is simply the expected value for each of the dimensions of the parameter space (each of the variables); and its covariance is simply the deviation or variance product of each of its dimensions with the expectancy for the dimension. In other words, how much it deviates from the expected value of the variable.

Note that the diagonal of is the variance of each variable (). On the other hand, the off-diagonal elements relate to the correlations between the coordinates of x:

## The Multinomial Distribution

Part of the multivariate distributions applies to situations in which the observations take only a finite number of discrete values *L* (could be categories or labels).

Given a number *n* of observations, yields a vector (x1, x2, …, xL) for each possible outcome (category or label). The possible outcomes are encoded in a vector of length ***L***:

-> (that is, one hot encoding of labels).

The multinomial probability model assumes that the *n* cases are independent of each other, each with probability for outcome :

Let

Indicate the vector of probabilities. The count vector ***x*** then follows the multinomial distribution,

Denoted

(for n observations, L outcomes, probability vector ).

When L = 2, the multinomial distribution reduces to the binomial with .

There is a useful relationship between the multinomial distribution and the Poisson. Calculations involving are sometimes complicated by the multinomial’s correlations, so we can approximate this using a Poisson distribution:

for large n.

## Exponential families

A p-parameter exponential family has the form

Where and y are p-vectors and A is contained in . Here is the “canonical” parameter vector and y = t(x) is the “sufficient statistic” vector. The normalizing function makes integrate to one. Note that the Poisson is a one-parameter exponential family.

The nice property is that exponential families allow to compress all inferential information into a p-dimensional statistic y.

Note: Y statistic can yield a polynomial of p dimensions for x (e.g. overfitting the data in ML models).

# Law of Large Numbers

Both the law of large numbers and central limit theorem are about many **independent** samples from **same distribution**.

The **Law of Large Numbers** tells us two things:

1. The average of many independent samples is (with high probability) close to the mean of the underlying distribution. This justifies the use of sample means as surrogates of population means.
2. 2. This density histogram of many independent samples is (with high probability) close to the graph of the density of the underlying distribution.
3. Note: Finite variance is not required for the law to apply.

The **central limit theorem** says that the sum or average of many independent copies of a random variable is approximately a normal random variable.

**LoLN**: As n grows, the probability that Xn is close to µ goes to 1.

**CLT**: As n grows, the distribution of Xn converges to the normal distribution N(µ, σ2/n).

Useful properties of CLT:

* The central limit theorem allows us to approximate a sum or average of i.i.d random variables by a normal random variable. This is extremely useful because it is usually easy to do computations with the normal distribution.
* In general, the distribution of a random variable *X* may not be familiar, or may not even be known, so you will not be able to compute the probabilities for the samples exactly. it can also happen that the exact computation is possible in theory but too computationally intensive in practice, even for a computer. The power of the CLT is that it applies whenever *X* has a mean and a variance. Note that some distributions may require larger n for the approximation to be a good one.

# Survival analysis and the EM algorithm

Survival analysis helps assess life expectancy, insurance rates and annuities by assigning a probability model to risk and life expectancy.

It uses Life Tables and Hazard Rates, where the former indicates the observed number of survivals for a given X (lifetime). Let

Be the probability of dying at age *i*, then the probability of surviving past age i-1 is:

And the *hazard rate* at age *i* is

Life tables are depicted as having each X (lifetime expectancy), number of observed survivals, number of deaths, hazard rate and survival rate. It uses current data to inference future outcomes (with all the limitations that has).

Another issue in Survival analysis comes from censored data. Censored data is data unavailable at the time of the analysis. Special cases of this is when an observed individual has survived past X years but has been lost in the data, introducing uncertainty on the outcome.

Life table curves are nonparametric because they assume no relationship between the hazard rates at each year.

## The Log-Rank Test

The log-rank test employs an extension of the life tables for the nonparametric two-sample comparison of censored survival data (something similar to Chi2 tests). Thus, it is useful to compare two life tables from different samples.

It assumes a null hypothesis where the two observed life tables come from the same distribution, and thus share equal mean and variance. We then can calculate the expected number of deaths under the null hypothesis, and calculate a Z statistic for the difference between observed and expected, for each year for the table:

## The Proportional Hazards Model

Provides a full regression analysis of censored data. It constructs the data points:

Where and are observed survival time and censoring indicator and is a known 1xp vector of covariates whose effect on survival we wish to assess. Therefore, we can run a simple linear regression that fits to each point z given c and t, and getting the coefficients of d.

## Missing Data and the EM Algorithm

Censored data is missing data. The EM algorithm is an iterative technique for solving missing-data inferential problems using only standard methods.

It works by first assigning an arbitrary value to the missing data, then computes the MLE for the parameter vector with the filled-in data (Maximizing step). Each of the missing values is replaced by its conditional expectation, given the non missing data (Expectation step). These steps are repeated until convergence.

Note, in exponential families it is guaranteed to converge to the MLE.

# Confidence Intervals

Suppose you know the underlying process, given by . The confidence intervals are obtained by finding the values of xmin and xmax such that the probability of such values is within the wanted interval. For a normal distribution is well defined, and using the CLT we can sometimes approximate different distributions to the normal and draw inference. But in some other cases, like Poisson distributions, the convergence is slow and therefore the approximation is not accurate.

## Neyman’s construction

Very close to the classical interpretation of confidence intervals.

* Repeated Sampling: Draw multiple samples and compute a confidence interval from each sample.
* Visualize Intervals: Plot all intervals—those covering the true parameter are one color (e.g., green), those missing it are another (e.g., red).
* Frequentist Interpretation: In the long run, the proportion of intervals containing the true parameter matches the confidence level (e.g., 95% of intervals contain the true value for a 95% confidence interval)

## The percentile method

Instead of obtaining the confidence intervals from the original distribution (which we typically don’t have access to):

* Generate Bootstrap samples of some statistic, and evaluate the 95% confidence interval of those. Define the bootstrap cumulative distribution function

And solve the equation:

In other words, draw bootstrap samples, and lookup the corresponding percentiles for the wanted interval.

* There often exists a transformation , such that is normal, and can be mapped to the standard error estimates.
* This method requires bootstrap sample sizes on the order of 2000.

## Bias-Corrected confidence intervals

In some cases, the percentile method should be adjusted a little downward to correct for bias. Having simulated B bootstrap replications, let be the proportion of replications less than

And define the bias-corrected value:

Where is the inverse function of the standard normal cdf. The BC level- confidence interval endpoint is defined to be:

Where is the bootstrap cdf and .