Anna Janicka

Mathematical Statistics 2020/2021
Lecture 3

1. Estimation

During the previous lecture, we signalled that the statistical model we will want to use and the usage of particular *statistics* will be determined by the questions that we want to answer on the base of the data. During this lecture, we will explore the concept of *point estimation*, i.e. the problem of how to choose, based on the data, the (single) distribution from the given family of distributions that best fits the data; in other words, how to choose the (single) best fitting value of the unknown parameter θ from the set θ of possible parameter values. In order to achieve this goal, we will introduce a special statistic, called the *estimator*:

Definition 1. *An* **estimator** *of parameter θ is any statistic $T = T(X_1, X_2, \ldots, X_n)$ with values in the set Θ (with $X_1, X_2, \ldots, X_n$ being observations for a statistical model with a family of distributions P_θ indexed by $\theta \in \Theta$).*

Note that in the definition of the estimator, we do not have the "intuitive" condition we want it to fulfill, namely that it approximates the true value of θ; an estimator is *any* function of the data, provided that it gives values from the possible range of values for θ. Obviously, we will be interested in estimators which will give us values close to what we want to obtain. We will study the different aspects of "closeness" and methods of evaluation of estimators during the next lectures. As for now, we will not define rigorously the expected property and rely on the intuitive understanding of approximating a value with a given formula, to explore the different possible approaches to estimation.

The commonly used notation for an estimator is to add a " ˆ " to the estimated value, for example $\hat{\theta}$ (if we wanted to estimate the value of parameter θ) or $\hat{g}(\theta)$ (if we wanted to estimate not the value of θ itself, but rather a function of it – for example, we could estimate σ^2, rather than σ, in the normal model).

1.1. Frequency as an estimator.

In some cases, the unknown parameter of the distribution is strictly related to the frequency of particular data outcomes. For example, this is the case in the quality control problem we introduced during the previous lecture, where we had: an observation of X, the total number of defective elements in a batch of $n = 50$ elements, with $\mathcal{X} = \{0, 1, \ldots, n\}$ and

$$P_\theta(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

for $\theta \in [0, 1]$ (we had observed $X = 6$). In this model, the parameter θ corresponds to the probability that a given element will be defective.

An obvious choice for the estimator of the unknown value of θ is, in this case, the empirical frequency. Here, we would have $\hat{\theta} = \frac{X}{n}$, and the value of the estimated parameter, based on the observed data, would be equal to $\frac{6}{50} = 0.12$. Note that if, rather than observing a single value of the number of defective elements, we were to observe sequences of 0s and 1s for the whole sample, the parameter θ could also be estimated as the empirical frequency (in this case, this would be equivalent to the average of the observed values of 0s and 1s).

Unfortunately, the choice of the best estimator is seldom as obvious. The quality control example is very simple, but it suffices to add one more (i.e., a third) possible value of the experiment outcome, and the problem gets a lot more complicated. Imagine that we are modeling the prevalence of various genotypes in a population. Assume that there are two possible versions of an allele, leading to three possible genotypes. If by θ we denote the unknown probability of a dominating allele, then the theoretical frequencies of the three genotypes would be equal to θ^2 (two times the dominating allele), $2\theta(1-\theta)$ (once the dominating version, and once the recessive version) and $(1-\theta)^2$ (for twice the recessive version of the allele). Now, assume that in a field experiment we observe N_1, N_2, N_3 individuals of the three genotypes, respectively. What function should we use as an estimator of θ? If we were to use the frequency of the first genotype only, we would take $\hat{\theta}_1 = \sqrt{\frac{N_1}{n}}$. However, we could also estimate θ on the base of the frequency of the third genotype, as $\hat{\theta}_2 = 1 - \sqrt{\frac{N_3}{n}}$. Furthermore, we could also look at $\hat{\theta}_3 = \frac{N_1}{n} + 2\frac{N_2}{n}$, and other formulae, all based on frequencies. Which of these formulas should we use? The answer is far from obvious (and we only looked at estimators based on frequency!).

1.2. **The empirical CDF as an estimator.** During the probability calculus course, we have shown (on the base of the laws of large numbers and the CLT) that the empirical cumulative distribution function constructed for a sample is a good approximation of the true CDF of the distribution (provided that we have sample sizes large enough). Therefore, the empirical CDF may also prove useful in the estimation procedure. Indeed, if we define the empirical CDF for the sample $X_1, X_2, \ldots, X_n$ as

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(X_i),$$

then, for a given value of t, this empirical CDF is a statistic, with a distribution given by the formula:

$$\mathbb{P}\left(\hat{F}_n(t) = \frac{k}{n}\right) = \binom{n}{k} (F(t))^k (1 - F(t))^{n-k},$$

for $k = 0, \ldots, n$. We may calculate the characteristics of this empirical distribution:

- The expected value of the statistic at point t is equal to $\mathbb{E}(\hat{F}_n(t)) = F(t)$ (note that we are almost dealing with a binomial distribution with parameters n and $p = F(t)$, the only difference being that instead of having the values of the random variable equal to $0, 1, \ldots, n$ we have values divided by n, namely $0, \frac{1}{n}, \frac{2}{n}, \ldots, 1$, so the expected value is the same as in the case of the binomial distribution, divided by n);
- The variance of the statistic at point t is equal to $\mathrm{Var}\hat{F}_n(t) = \frac{1}{n} F(t)(1 - F(t))$ (the variance for a binomial distribution would be $nF(t)(1 - F(t))$, and we need to divide the random variable by n so the variance is divided by n^2);
- Given the above properties, from the CLT we have that

$$\frac{\hat{F}_n(t) - F(t)}{\sqrt{F(t)(1 - F(t))}} \sqrt{n} \to_{n\to\infty} \mathcal{N}(0, 1),$$

  and what is more, from the Glivenko-Cantelli theorem we have that the convergence of the empirical distribution to the theoretical counterpart is uniform.

Therefore, it follows that we can use the the empirical CDF as an estimator of the theoretical CDF. This will be useful especially in cases where the family of probability distributions in a statistical model will have a parametrization with F, rather than a "simple" parameter θ, but not only then.

1.3. **The order statistics as estimators.** Another class of statistics, which may be used as estimators, are the order statistics. We define the i-th order statistic for a sample $X_1, X_2, \ldots, X_n$ as the i-th element of the sample when organized in ascending order. In particular, $X_{1:n}$ is the minimum value from the sample, and $X_{n:n}$ is the maximum value. For a sample of size n from a distribution with CDF equal to F, we have that the CDF of the i-th order statistic is equal to

$$F_{i:n}(t) = \mathbb{P}(X_{i:n} \leqslant t) = \sum_{k=i}^{n} \binom{n}{k} (F(t))^k (1 - F(t))^{n-k},$$

and if additionally the original distribution is continuous with density f, then the i-th order statistic is also a continuous random variable, with density equal to

$$f_{i:n}(x) = n\binom{n-1}{i-1} f(x)(F(x))^{i-1}(1 - F(x))^{n-i}.$$

1.4. **Two basic types of estimation.** From the above considerations, it follows that the empirical characteristics calculated on the base of the sample are going to be "good" estimators of their theoretical counterparts: the sample mean will be a good estimator of the expected value; the sample variance will be a good estimator of the theoretical variance; the sample median will be a good estimator of the theoretical median, and so on. These properties of the samples are the rationale underlying the two most basic methods of point estimation: the method of moments and the method of quantiles.

1.4.1. **Method of Moments Estimation.** First, we will look at a technique of estimation called the method of moments, which is based on comparisons of empirical moments with their theoretical counterparts. From the limit theorems, we know that for large samples they should be more or less equal; therefore, we will be using sample characteristics as estimators of theoretical values, which depend on unknown parameter distributions, and from that we will derive the approximated values of the parameters. If we have a k-dimensional space for the unknown probability distribution parameter θ, in the method of moments technique we will need to solve a system of k equations, such that:

- If Θ is single-dimensional, we will use one equation, usually $\mathbb{E}_\theta X = \bar{X}$;
- If Θ is two-dimensional, we will use a system of two equations, usually $\mathbb{E}_\theta X = \bar{X}$, $\mathrm{Var}_\theta X = \hat{S}^2$; etc.

We will illustrate with two simple examples.

(1) Let $X_1, X_2, \ldots, X_n$ be a sample from an exponential distribution $Exp(\lambda)$ with an unknown parameter $\lambda > 0$. We know that $\mathbb{E}_\lambda X = \frac{1}{\lambda}$, so that we will write the single equation as
$$\frac{1}{\lambda} = \bar{X},$$
and solving for λ, we get
$$\hat{\lambda}_{MM} = \frac{1}{\bar{X}}.$$

(2) Let $X_1, X_2, \ldots, X_n$ be a sample from a gamma distribution $Gamma(\alpha, \lambda)$ with unknown parameters $\alpha, \lambda > 0$. We have a two-dimensional parameter space, so we will use two equations, one for the mean and one for the variance. We know that $\mathbb{E}_{\alpha,\lambda} = \frac{\alpha}{\lambda}$ and $\mathrm{Var}_{\alpha,\lambda} = \frac{\alpha}{\lambda^2}$, so we will have
$$\frac{\alpha}{\lambda} = \bar{X}, \quad \frac{\alpha}{\lambda^2} = \hat{S}^2,$$
which gives
$$\hat{\lambda}_{MM} = \frac{\bar{X}}{\hat{S}^2}, \quad \hat{\alpha}_{MM} = \frac{\bar{X}^2}{\hat{S}^2}.$$

1.4.2. **Method of Quantiles Estimation.** The method of quantiles is identical to the method of moments, apart from the fact that instead of moments, theoretical and empirical quantiles are compared (depending on the distribution, this may be easier computationally than calculations with moments).

For the exponential model, we would calculate the estimator for the parameter λ from the equation for the median:
$$1 - e^{-\lambda Med} = \frac{1}{2},$$
from which it follows that
$$\hat{\lambda}_{MQ} = \frac{\ln 2}{Med}.$$

The method of moments and the method of quantiles are simple conceptually and (usually) computationally, but in some cases they lead to estimators which do not have some desired properties (for example, they may not behave well for small sample sizes). The third estimation technique we will introduce usually does not have these drawbacks.

1.5. **Maximum Likelihood Estimation.** A totally different way of reasoning underlies the third estimation technique we will talk about, namely: the maximum likelihood estimation. This method is based on the assumption that the value of the parameter which best fits the data, given the data, is the value of the parameter for which the probability of obtaining the given set of results is the highest (among all possible values of the parameter). Therefore, we will define the *likelihood*, as a function of the unknown parameter θ, equal to the probability (density) function of the data, treating the sample observations as given:

$$L(\theta) = f(\theta; X_1, X_2, \ldots, X_n).$$

In order to find the value of θ which best fits the data – the maximum likelihood estimator of θ – we will need to maximize the likelihood function L. In other words, $\hat{\theta}_{MLE}$ will be the maximum likelihood estimator of θ, if

$$f(\hat{\theta}_{MLE}(x_1, x_2, \ldots, x_n); x_1, x_2, \ldots, x_n) = \sup_{\theta \in \Theta} f(\theta; x_1, x_2, \ldots, x_n).$$

If we want to provide the ML estimator of a function of the parameter θ, i.e. $g(\theta)$, by convention we will provide $g(\hat{\theta}_{MLE})$.

In most practical applications, we will be looking for a ML estimator for a sample of independent observations. In this case, the likelihood function is a product of probability functions, and finding a maximum with the usual technique of taking the derivative and equaling it to zero may lead to horrible formulas. Therefore, in most cases we will take advantage of the property that a function reaches its maximum at the same point that a monotonous transformation of it – namely, the logarithm – reaches the maximum, and maximize the logarithm of the likelihood function, denoted by $l(\theta)$, instead of maximizing $L(\theta)$.

Examples:

(1) Quality control example. The class of probability distributions is given by

$$\mathbb{P}_\theta(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

so the likelihood function is equal to

$$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Instead of maximizing the likelihood, it will be easier to maximize the logarithm

$$l(\theta) = \ln \binom{n}{x} + x \ln \theta + (n-x) \ln(1-\theta),$$

which we will do by taking the derivative of $l(\theta)$ with respect to θ and equaling it to zero:

$$l'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0,$$

which leads to

$$\hat{\theta}_{ML} = \frac{x}{n}.$$

In this case, the maximum likelihood estimator is the same as the frequency estimator (and the method od moments, based on the average, estimator).

(2) Now let $X_1, X_2, \ldots, X_n$ be, again, a random sample from an exponential distribution $Exp(\lambda)$ with an unknown parameter $\lambda > 0$. The likelihood function is then equal to

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \cdot \sum_{i=1}^{n} x_i}.$$

Again, instead of maximizing the likelihood function with respect to λ, it will be much easier to maximize the logarithm of the likelihood function:

$$l(\lambda) = n \ln \lambda - \lambda \cdot \sum_{i=1}^{n} x_i,$$

so we will take the derivative of $l(\lambda)$ with respect to λ and equal it to zero:

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0,$$

which gives

$$\hat{\lambda}_{ML} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}}.$$

In this case, the maximum likelihood estimator of λ is the same as the method of moments estimator (but different from the method of quantiles estimator).

(3) Finally, let us look at a random sample of $X_1, X_2, \ldots, X_n$ from a normal model, i.e. such that $X_i \sim N(\mu, \sigma^2)$, where μ and $\sigma > 0$ are both unknown parameters. In this case, the likelihood function is equal to

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2},$$

so this time it is a function of two parameters. Again, we will maximize the log likelihood,

$$l(\mu, \sigma) = -n \ln(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

This time, the maximization procedure requires calculating two first-order conditions:

$$\frac{dl}{d\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 = 0,$$

$$\frac{dl}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n\mu}{\sigma^2} = 0.$$

Solving for μ and σ, we get

$$\hat{\mu}_{ML} = \bar{X}$$

from the second equation, and substituting into the first equation, we obtain:

$$\hat{\sigma}_{ML} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2}.$$

Note that these estimators are the same as those we would obtain with the method of moments technique, i.e. when comparing the sample average with the theoretical average and the sample variance with the theoretical variance.

Since usually we are not interested in the value of σ itself but in the variance, in such cases we would specify the maximum likelihood estimator of σ^2 as the square of the formula for σ, namely

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2.$$

Note that in the specification of the maximum likelihood method, it is not necessary that the observations are independent. If the condition of independence did not hold, we would need to specify the specific joint distribution, which would not be a product of one-dimensional marginal distributions anymore. Afterwards, the whole procedure would be performed as before.