Anna Janicka

Mathematical Statistics 2021/2022
Lecture 2

1. Introduction to Mathematical Statistics

During this lecture, we will introduce the concepts underlying mathematical statistics – the methods of inference from data. First of all, we need to define the basic object that will be under study – the statistical model (of the results of an experiment). In order to be able to do that, we need some probabilistic foundations. The main assumption on which we will base during this course is the following: the empirical data we observe and want to explore reflect the functioning of a random mechanism. In other words, we assume that the objects we will study – the collected data – will be the realizations of some random variables, defined over some probabilistic spaces. The difference between probability calculus and statistics lies in the knowledge, however; in the latter case, we usually know less about the underlying model, but have empirical data at our disposal. In order to be able to infer something from the data with statistical tools, we will need to make some assumptions about the (usually unknown!) distributions of the random variables under study. These assumptions will likely reflect theoretical implications or results of existing studies; the correctness of these assumptions will limit the results and translate to the validity of statistical considerations.

It will be useful to illustrate the difference between the probabilistic and the statistical approaches with an example. Assume that an item is produced in a facility, and that the process of production may lead to defective output. Now, if we were to look at this experiment from the perspective of probability calculus, we would do the following. *First* of all, we would need to specify the problem. The phrasing could be as follows: assume that in a production process each produced unit may be either good or defective; the result is random. Each item may be defective with probability 10%, independently of the defectiveness of all other produced items. *Second*, we would need to specify the question that we want to answer, for example: What is the chance that in a batch of 50 items, exactly 6 will be defective? What is the average number of defective elements in a batch of 50? What is the most probable number of defective elements in a batch of 50? etc. *Third*, in order to solve the problem, we would introduce a probabilistic model. In this case, we would usually describe the situation with the use of a Bernoulli scheme, with the number of trials, n, equal to 50, and the probability of success in a single trial, p, equal to 0.1. Note that if we wanted to answer different questions, for example dealing with the order of the appearance of faulty elements (What is the probability that the first item will be good, but the next four will be defective?), we would need a different model – usually one where the probability space includes all possible outcomes (of the Bernoulli scheme) understood as series of 0s (good items) and 1s (faulty items). Therefore, even in case of this simple example, we see that depending on the different questions we may want to answer, we may have different model specifications.

If we were to look at the problem from the statistical perspective, the emphasis would be elsewhere. *First*, as far as the phrasing of the problem is concerned, we would likely see something like the following: an inspector verified a set of randomly chosen items produced in a facility, noting whether the items were defective (1) or good (0). In a batch of 50 items, he obtained the following results:

0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1

Second, typical questions to be asked would include the following: based on the results obtained, and assuming that the defectiveness of elements is independent of each other, how would we assess the (unknown) probability that an element is defective? In view of the obtained results, is it possible that the level of defectiveness is equal to 10%, as the producer declares? *Third*, in order to solve the problem, we would introduce a statistical model:

Definition 1. *A* statistical model *is a triple* $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \mathcal{P})$*, where*
$\mathcal{X}$ *is the space of values of the observed random variable(s);*

$\mathcal{F}_\mathcal{X}$ is the σ-algebra of measurable events on $\mathcal{X}$; and
$\mathcal{P}$ is a family of probability distributions over $(\mathcal{X}, \mathcal{F}_\mathcal{X})$, P_θ, indexed by a parameter $\theta \in \Theta$.

Note that this definition differs from the definition of a probabilistic model in that the probability distribution is not defined unequivocally, but rather as a family of distributions, among which we will want to find the correct one (the one closest to the data). Throughout this course we will use a simplified version of the definition, however, and skip the declaration of σ-algebras for the sample spaces, assuming always the usual case (all subsets of the sample space in discrete cases, Borel subsets in case of continuous experiments). Therefore, we will be providing the specification of the statistical models in a less formal way, namely by providing $(\mathcal{X}, P, \Theta)$:

$\mathcal{X}$ – the space of values of the observed random variable;
$\mathcal{P}$ – the family of probability distributions, P_θ; and
Θ – the range of values for the unknown parameter θ.

In most cases, the range of values for the observations $\mathcal{X}$ will be an n-dimensional space, since we will need one dimension for each observation available for analysis. In our production example, we could specify the model in the following ways:

(1) If we record the results for all items separately, as above, the sample space $\mathcal{X}$ will be equal to all possible outcomes of the observations, namely all $n = 50$ element series consisting of 0s and 1s, i.e. $\mathcal{X} = \{0, 1\}^n$. Further, we would assume that all elements are independent, behave identically and have equal unknown probability of being defective, $\theta \in [0, 1]$, so the unknown probability distribution describing the probability of observing a given outcome $(x_1, x_2, \ldots, x_n) \in \mathcal{X}$ would be specified by the joint distribution

$$P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i},$$

where in the case of the data specified above we would have $X_2 = X_{10} = X_{15} = X_{32} = X_{42} = X_{50} = 1$, and all other $X_i = 0$.

(2) On the other hand, if we only recorded the total number of defective elements X, rather than the particular outcomes, we would describe the experiment with the following model: $\mathcal{X} = \{0, 1, \ldots, n\}$ and

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

for $\theta \in [0, 1]$. In the case of our observations, we would have $n = 50$ and $X = 6$.

Regardless of which of the formulations above we use, in the production example we have one unknown parameter – θ – whose value "pinpoints" one of the distributions from the assumed family of distributions. In many practical examples, and in order to be able to answer the questions formulated, we will be interested in assessing the value of this unknown parameter. This assessment procedure is referred to as *estimation*, and may be directed either at providing a single value for θ (point estimation), or providing a range of possible values for θ (interval estimation, providing so-called confidence intervals). We will broadly discuss the different methods of estimation and their properties during this course. Furthermore, we may also be interested in verifying some additional statements about the distribution (perhaps about the precise value of θ – like, is it credible that $\theta = 0.1$)? This procedure is called *hypothesis testing*, and we will also discuss such methods later on this semester.

At this point we will just signal that both estimation and hypothesis testing will be conducted based on calculations of so-called *statistics*. By a statistic we will understand any function of the observed random variables, i.e. any function $T = T(X_1, X_2, \ldots, X_n)$ of the data $X_1, X_2, \ldots, X_n$ (as a function of random variables, T is a random variable). Note that we do not allow T to depend on the unknown distribution parameter(s) θ – we must be able to calculate a statistic based on data only. Note however, that the distribution of the random variable T depends on the (true) distribution of the random variables X_i. This means that

although T does not have θ appearing in the formula explicitly, the distribution of T depends on the value of θ. This observation will be the clou to the suggested methods of reasoning in the case of both estimation and hypothesis testing.

Reverting to our production example, in the first formulation, we could define for example the following statistics:

$$T_1 = \sum_{i=1}^{n} X_i^2, \qquad T_2 = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad T_3 = \frac{1}{n}\sum_{i=1}^{n} X_i - 0.1$$

Based on the laws of large numbers, we would expect the statistic T_2 to converge to the true (unknown) value of θ, and the statistic T_3 to be zero if the true value of θ is equal to 0.1 – hinting as to how these statistics may be used later on. In the second formulation, we could have

$$T_1 = X^2, \qquad T_2 = \frac{X}{n}, \qquad T_3 = \frac{X}{n} - 0.1$$

(with analogous properties of T_2 and T_3).

What we need to stress at this point is that any reasoning conducted after the specification of the statistical model depends on the specification (and its validity). In some cases, the specification is not a source of concern – for example, if we were to repetitively toss a single coin, we would not question the use of the Bernoulli scheme as a description of the experiment. In most situations, however, the specification needs some attention. Can we be sure that the probability that an item produced will be defective is the same for all elements? Perhaps it becomes larger as time goes by? Perhaps the defects are not independent? etc. If not all assumptions made are necessarily justified, we must be aware of the fact that this influences the validity of the results of applying all statistical techniques afterwards.

The question of assuming the right distribution is especially pronounced in the case of continuous data, where we may not even have certainty as to the type of distribution in effect (can we assume that it is normal? Or perhaps we should consider some different class of distributions?). For example, in many cases of experiments with continuous outcomes, a typical assumption is that the underlying distribution is normal (on the base of the CLT, this is not a bad choice since we can expect some statistics calculated for large samples, such as sums or means, to resemble the normal distribution). What we need to be aware of, however, is that even though in such cases we will be able to estimate the values of the parameters of the normal distribution which best describes the data (i.e. the parameters θ which give the best distribution P_θ), this will not be equivalent to proving that the data indeed comes from a normal distribution. Until we positively test the assumptions made, we will only be allowed to say that from the set of distributions P_θ, one fits the data best; this is not equivalent to saying that this distribution is the true distribution of the data.

Obviously, the art of modeling reality with mathematical tools is always the art of finding a compromise between simplicity of calculations and results, and precise reproduction of reality. Theoretically, there are no "constraints" for the family of distributions $\mathcal{P}$ with which we will describe the distribution, so we could define it very generally, for example as "any continuous distribution". However, unless our aim is to test different model specifications, this is seldom done, since in a general formulation where we have different classes of distributions it may become extremely hard to pinpoint the best distribution (there is no "natural" value of parameter θ to estimate, the formulae become horrible, etc.).

We will conclude our introduction to statistical models by providing some additional examples.

(1) Periods of market growth: assume an analyst studies the length of periods of growth on the stock market. She is interested in times of continuous growth (until the first fall), measured in days. Assume that the times of growth, $X_1, X_2, \ldots, X_n$ are independent random variables from an exponential distribution with an unknown parameter λ. For this scenario, the statistical model would be:
$\mathcal{X} = (0, \infty)^n$;

The joint probability distribution may be specified either by the CDF:

$$P_\lambda(X_1 \leqslant x_1, X_2 \leqslant x_2, \ldots, X_n \leqslant x_n) = \prod_{i=1}^{n}(1 - e^{-\lambda x_i})$$

or by the density function:

$$f_\lambda(x_1, x_2, \ldots, x_n) = \lambda^n e^{-\lambda \sum x_i}$$

for $\lambda > 0$.

(2) Measurements with error: We repeat measuring a (physical) value μ. Since our measurement machine is not perfect, assume that the measurements are independent random variables $X_1, X_2, \ldots, X_n$ from a normal distribution with unknown parameters μ and σ^2. (In this case, the unknown parameter θ is two-dimensional, $\theta = (\mu, \sigma)$.) Under such assumptions, the statistical model would be:
$\mathcal{X} = \mathbb{R}^n$;
The joint probability distribution may be specified by the density function

$$f_{\mu,\sigma}(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right),$$

for θ such that $\mu \in \mathbb{R}, \sigma > 0$.

2. The Normal Model

In the previous section, we introduced the concept of a statistical model with some examples; we also signalled that a commonly made assumption is that the modeled distributions are normal, which means that analyses are frequently conducted for this set of assumptions. We will therefore continue exploring the properties of the normal model.

Assume that $X_1, X_2, \ldots, X_n$ are independent random variables from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. In many practical applications, we will be interested in the properties of different statistics calculated for this model. The most commonly used statistics are: the *sample mean*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the *sample variance*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Note that in the denominator of the expression for the sample variance, the number of observations n is diminished by 1. The rationale behind this modification of the "standard" formula for the variance are going to become clearer after exploring the properties of this statistic (the bias of this estimator), which we will do in one of the next lectures. At this point, we will just proceed with the given formula.

From the properties of the normal distribution, it follows that the sample mean has a normal distribution with parameters μ and σ^2/n. For the variance, the situation is slightly more complicated; we can prove the following theorem:

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be independent random variables from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, and let $\bar{X}$ and S^2 denote the sample mean and variance, respectively. We have that: $\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$ and $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$, and $\bar{X}$ and S^2 are independent.*

Note that the number of degrees of freedom in the chi-square distribution is equal to $n-1$, although formally the variance is a (weighted) sum of n squares. We will not support this statement with a formal proof, but some hints. Note that

$$\sum_{i=1}^{n}\frac{(X_i - \mu)^2}{\sigma^2} = (n-1)\frac{S^2}{\sigma^2} + n\frac{(\bar{X} - \mu)^2}{\sigma^2}.$$

On the left hand side we have an expression which is part of the theoretical variance; it has a chi-square distribution with n degrees of freedom (as a sum of squares of n standardized

variables). On the right hand side we have a sum of the re-scaled sample variance and a square of a standard normal variable (i.e. having a chi-square distribution with 1 degree of freedom). Showing that the two items are independent is not an easy task (intuitively it is not clear, since the variables depend on the same data and we even have the mean in the formula for the variance!), but it can be done – and in this case, looking at the number of degrees of freedom of the chi-square distribution on both sides of the equality we have that the standardized sample variance has a chi-square distribution with $n-1$ degrees of freedom.

Another distribution which often appears in connection with the normal model is the t-Student distribution. This is the distribution of the random variable

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S},$$

which, in the normal model with a sample size of n, is defined as having a t-Student distribution with $n-1$ degrees of freedom. Note that T is not a statistic (the value of μ appears in the formula), but this random variable is used for hypothesis testing (we will substitute a value that is to be tested for μ).