Lecture 1
The Subject of Statistical Inference

1. Sampling from a population

The purpose in Statistical Inference is to draw relevant conclusions from given data. These conclusions may be about predicting further outcomes, evaluating risks of events, testing hypotheses, etc. In all cases, the statistician is confronted with the problem of drawing conclusion about the **population** to by using the limited information available in the data set. Typically, an experiment to collect data from the population is repeatedly performed, the replicates being independent of one another. The possible results are real numbers that form a vector of observations $\mathbf{x}=(x_1, x_2, \ldots, x_n)$. The appropriate **sample space** is R^n. There is typically a "hidden" mechanism characterizing the population that generates the data and one is looking for suitable ways to identify it. **Models** will describe this mechanism in some simplistic but hopefully useful way. For the model to be more trustworthy, continuous variables, such as time, interval measurements, etc. should be treated as such, where feasible. However, in practice, only discrete events can actually be observed. Thus, the observations will be recorded with some *unit of measurement* Δ determined by the precision of the measuring instrument. This unit of measurement is always finite in any real situation.

If empirical observations were truly continuous, then, with probability one, no two observed responses would ever be identical. This fact will sometimes be used in our theoretical derivations. On the other hand, as pointed out above, the *real life empirical observations are indeed discrete*. This fact will be utilized by us to keep some of the proofs and derivations in this course *simpler* by dealing with the discrete case only.

2. Statistical Models

If the observations were *continuous* and if we knew the density of each observation, we could calculate the **joint density** of the vector of observations:

$$L_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1).f_{X_2}(x_2)\ldots f_{X_n}(x_n) \tag{1}$$

In the case of *discrete* observations, this will be just the product of the probabilities for each of the measurements to be in a suitable interval of length Δ.

If the observations were **independent identically distributed (i.i.d.)** then all densities in (1) would be the same: $f_{X_1}(x) = f_{X_2}(x) = \ldots = f_{X_n}(x) = f(x)$. This is the most typical situation we will be discussing in our course. The need of **Statistical Inference** arises since typically, our knowledge about $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)$ is **incomplete.** Given an inference problem and having collected some data, we construct one or more set of possible **models** which may help us to understand the data generating mechanism. Basically, statistical models are working assumptions about how the data set was obtained.

For example, if our data was counts of accidents within $n = 10$ consecutive weeks on a busy crossroad, it may be reasonable to assume that a Poisson distribution with an unknown parameter λ has given rise to the data, that is, we may assume that we have 10 independent realizations of a Poisson(λ) random variable. If, on the other hand, we

measured the lengths X_i of 20 baby boys at age 4 weeks, it would be reasonable to assume

$$X_i \sim N(\mu, \sigma^2), i = 1, 2, \ldots, 20.$$

The models we use, as seen in the examples above, are usually about the shape of the density or of the cumulative distribution function of the population from which we have sampled. These models should represent, as much as possible, the available prior theoretical knowledge about the data generating mechanism. It should be noted that in most cases, *we do not* exactly know which population distribution to assume for our model. Suggesting the set of models to be validated, is a difficult matter and there is always a risk involved in this choice. The reason is that if the contemplated set of models is "too large", many of them will be similar and it will be difficult to single out the model that is best supported by the data. On the other hand, if the contemplated set of models is "too small", there exists the risk that none of them gives an adequate description of the data.

Choosing the model usually involves a close *collaboration* between the statistician and the people who formulated the inference problem. In general, we can view the statistical model as the triplet $(\mathcal{X}, \mathcal{P}, \Theta)$ where:

(random variables)

$-\mathcal{X}$ is the sample space (i.e.. the set of all possible realizations $\mathbf{X}=(X_1, X_2, \ldots, X_n)$)

$-\mathcal{P}$ is a family of model functions $P_\theta(\mathbf{X})$ that depend on the unknown parameter θ;

obtain $\hat{\theta}_n$

$-\Theta$ is the set of possible θ-values, i.e.. the parameter space indexing the models.

$\hat{\theta}_n \in \Theta$, then $P_{\hat{\theta}_n} \Rightarrow$ look for $\hat{f}$ then we get $\hat{f}$ obtain the shape of the distribution

3. The Inference problem

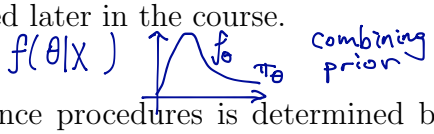
3.1. Parametric and Nonparametric procedures. Robustness. The statistical inference problem can be formulated as follows:

Once the random vector $\mathbf{X}$ *has been observed, what can be said about which members of* $\mathcal{P}$ *best describe how it was generated?*

The reason we are speaking about a *problem* here is that we do not know the exact shape of the distribution that generated the data. The reason that there exists a *possibility* of making inference rests in the fact that typically a given observation is much more probable under some distributions than under others (i.e. the observations give **information** about the distribution). This information should be combined with the *a priori* information about the distribution to do the inference. Note that we *always* have some a priori information. It could be more or less specific. When it is specific to such an extent that the shape of the distribution is known up to some *finite* number of parameters (i.e. the parameter θ is finite dimensional), we have to conduct *parametric* inference. Most of the classical statistical inference techniques are based on fairly specific assumptions regarding the population distribution and most typically the description of the population is in a parametric form. In introductory textbooks, the student just practices applying standard parametric techniques. However, to be successful in practical statistical analysis, one has to be able to deal with situations where standard parametric assumptions are not justified. A whole set of methods and techniques is available that may be classified as *nonparametric* procedures. We will be dealing with them in the second part of our course.

These procedures allow us to make inference without (or with a very limited amount of) any assumptions regarding the functional form of the underlying population distribution. Needless to say that these procedures are applicable in more general situations (which is good) but if they are applied for a situation where a particular parametric distributional shape indeed holds, the nonparametric procedures may not be as efficient as compared to a procedure specifically tailored for the corresponding parametric case (which would be bad if the specific parametric model indeed holds). The situation in practice may even be more blurred: we may know, for example, that our population is "not far away" from the (parametric) normal population but still "deviates a bit" from it. Going over in such cases directly to purely nonparametric approach would not properly address the situation since the idea about *relatively small* deviation from the baseline parametric family will be lost. The proper approach in such cases would be the *robustness* approach where we still keep the idea about the "ideal" parametric model but allow for *small* deviations from it. The aim is in such "intermediate" situations to be "close to efficient" if the parametric model holds but at the same time to be "less sensitive" to small deviations from the ideal model. These important issues will be discussed later in the course.

3.2. Bayesian inference



Another way to classify the Statistical Inference procedures is determined by the way we treat the *unknown* parameter θ. If we treat it as unknown but deterministic (fixed) then we are in a **Non-Bayesian** setting. If we consider the set of θ-values as quantities that before collecting the data, have different probabilities of occurring according to some (*a priori)* distribution, then we are speaking about *Bayesian inference.* The Bayesian approach allows us to introduce and effectively utilise any additional (prior) information about the model when such information is available. This information is entered in the model through the **prior distribution** over the set Θ of parameter values and reflects our prior belief about how likely any of the parameter values is before obtaining the information from the data. This topic will also be discussed shortly in our course.

4. Goals in Statistical Inference

Following are the most common goals in inference:

4.1. Estimation

$P_\theta \longrightarrow$ to find $\hat{\theta}_n$

We want to calculate a number (or a k-dimensional vector, or a single function) as an approximation to the numerical characteristic in question.

But let us point out immediately that there is little value in calculating an approximation to an unknown quantity without having an idea of how "good" the approximation is and how it compares with other approximations. Hence, immediately questions about **confidence interval** (or, more generally, **confidence set**) construction arise. To quote the famous English mathematician and philosopher Alfred North Whitehead, in Statistics we always "seek simplicity and distrust it".

4.2. Confidence set construction

After the observations have been made, further information to our a priori knowledge about the set Θ has been added, so it becomes plausible that the true distribution belongs to a smaller family than it was originally postulated, i.e. it becomes clear that the

unknown θ-value belongs to a *subset* of Θ. The problem of confidence set construction arises. It means, determining a (possibly small) plausible set of θ-values and clarifying the sense in which the set is plausible.

4.3. Hypotheses testing

An experimenter or a statistician sometimes has a theory which when suitably translated into mathematical language becomes a statement that the true unknown distribution belongs to a smaller family than the originally postulated one. One would like to formulate this theory in the form of a hypothesis. The data can be used then to infer whether or not his theory complies with the observations or is in such a serious disarray that would indicate that the hypothesis is false.

Finally, we will only mention here that deeper insight in all of the above goals of inference and deeper understanding of the nature of problems involved in them is given by **Statistical Decision Theory** (not to be discussed in this course). Here we define in general terms what a *statistical decision rule* is and it turns out that any of the procedures of Estimation, Confidence set construction and Hypotheses testing can be viewed as a suitably defined decision rule. Moreover, defining optimal decision rules means solving suitably formulated **mathematical optimization problems**.

Purpose of inferance:

Purpose: draw relevant conclusion from given data

conclusion about the population using limited data set

Notation:

vectors of observation: $x = (x_1, x_2, \dots, x_n)$

Sample space R^n

Unit of measurement Δ

Model is trust worthy $\Leftrightarrow$ Continuous variable

But in reality, only discrete case can be observed.

Example

$n=10$

$X_1, X_2, X_3, \dots, X_{10}$

$X_i \sim Poi(\lambda)$, $\lambda > 0$

$P(X_i = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2 \dots$

Modeling counts....

Example 2

length of babyboy at 4

$X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \dots, 20$