Statistics, MDI, Fall 2021

Gleb Karpov

Statistics

Statistics is a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty.

- Statistics is concerned with data analysis: using data to make inferences. It is concerned with questions like 'what is this data telling me?' and 'what does this data suggest it is reasonable to believe?'
- In Probability Theory we go from the assumption of the model to the probability of the specific outcome, *i.e.* from general to particular.
- The Statistics problem goes almost completely the other way around. In statistics, a sample from a given population is observed, and the goal is to learn something about that population based on the sample.
- So while the two things—probability and statistics—are closely related, there is clearly a sharp difference.

Basic definitions

We need to introduce some important words .

- **Population** – the entire collection of individuals or objects about which information is desired to be obtained.
- **Sample** is a subset of the population, selected for study in some prescribed manner.
- **Census** – study of every unit, everyone or everything, in a population. Obtaining complete information from an entire population.
- **Descriptive statistics** – the branch of statistics that includes methods for organising and summarising data.
- **Inferential statistics** – the branch of statistics that involves generalizing from a sample to the population from which it was selected, way of making inferences about populations based on samples.

Example 1

Suppose we wish to estimate the proportion p of students in HSE who attend none of the lectures since the beginning of the module. - Suppose time is limited and we can only interview 20 students at the campus. - Is it important that our survey is 'random'? How can we ensure this? - Suppose we find that 5 students have not attended any lecture. We might estimate θ by $\hat{\theta} = 5/20 = 0.25$. But how large an error might we expect $\hat{\theta}$ to have?

Example 2

Suppose the population of registered voters in Florida is divided into two groups: those who will vote democrat in the upcoming election, and those that will vote republican. To each individual in the population is associated a number, either 0 or 1, depending on whether he/she votes republican or democrat. If a sample of n individuals is taken completely at random, then the number X of democrat voters is a binomial random variable, written $X \sim Bin(n, \theta)$, where $\theta \in \Theta = [0, 1]$ is the unknown proportion of democrat voters. The statistician wants to use the data $X = x$ to learn about θ.

Yet another part with definitions

Before we formulate statistics problems in mathematical language, we need to introduce basic concepts.

- The random variables $X_1, \ldots, X_n$ are called a random sample of size n from the common distribution (population) $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables, and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, \ldots, X_n$ are called Independent and Identically Distributed (i i d) random variables with pdf or pmf $f(x)$.
- From the latter we can conclude that the joint pdf or pmf of $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = f(x_1) \cdot \ldots \cdot f(x_n) = \prod_{i=1}^{n} f(x_i)$$

- If our distribution is a part of a parametric family, then we define its pdf as $f(x|\theta)$, and the joint pdf is

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

  where the same parameter θ is used for every term in the product. The typical problem we will encounter will begin with something like "Suppose $X_1, \ldots, X_n$ is an independent sample from a distribution with PDF $f(x|\theta)$."

- Let $X_1, \ldots, X_n$ be a random sample. Let $Y = T(X_1, \ldots, X_n)$ be a function of the sample that does not depend on θ. Then Y is called a **statistic**.
- For example, statistics may give the smallest or the largest value in the sample, the average sample value, or a measure of variability in the sample observations.

The goals of Statistics

Many probability distributions depend on a small number of parameters; for example, the Poisson family depends on a single parameter λ and the Normal family on two parameters μ and σ: $\mathcal{N}(\mu, \sigma^2)$. Unless the values of the parameters are known in advance, they must be estimated from the data. Throughout we will refer to θ as a general variable for parameter.

Two kinds of inference problems

Point estimation

- Suppose $X_1, \ldots, X_n$ are iid with PDF/PMF $f(x|\theta)$.
- The point estimation problem seeks to find a quantity $\hat{\theta}$, called an estimator, depending on the values of $X_1, \ldots, X_n$, which is a "good" guess, or estimate, of the unknown true θ.
- Since $\hat{\theta}$ depend on a sample, we can formally say that $\hat{\theta} = T(X_1, \ldots, X_n)$, and so statistic T is a **point estimator** of θ.

Hypothesis testing

- Unlike the point estimation problem, the hypothesis testing problem might start with a specific question like "is θ equal to θ_0?," where θ_0 is some specified value.
- The main idea is to construct specific decision rule based on the sample $X_1, \ldots, X_n$ by which one can say whether θ belongs to the given set of parameter values or not.

Back to statistic: sampling distribution

Let us introduce some statistics that are often used and provide good summaries of the sample.

- The sample mean:

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

- The sample variance:

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

- The sample standard deviation is

$$S = \sqrt{S^2}.$$

Each statistic given a new random sample $(X_1^{(2)}, \ldots, X_n^{(2)})$ may take new value, so we can treat statistics as **random variables** themselves!

As random variables, they have their own distributions. We call the probability distribution of a statistic Y the **sampling distribution** of Y.

Statement 1:

Since every point estimator is a statistic, then estimators also have their sampling distributions, expectations, variances, and other charasteristics that random variable have.

Let $X_1, \ldots, X_n$ be a random sample of population with mean μ and variance $\sigma^2 < \infty$. Consider some characteristics of random variables sample mean $\bar{X}$, and sample variance S^2:

- $$E[\bar{X}] = \mu,$$
- $$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n},$$
- $$E[S^2] = \sigma^2.$$

Statement 2:

We say that estimator is **unbiased** if $E[\hat{\theta}] = \theta$.

Problems

Problem 1

Random variable assumes values 0 and 1, each with probability 1/2. 1. Find population mean μ and variance σ^2 2. You have 9 observations of X: $X_1, \ldots, X_9$. Consider the following estimators of the population mean μ: (i) $\hat{\mu}_1 = 0.45$, (ii) $\hat{\mu}_2 = X_1$, (iii) $\hat{\mu}_3 = \bar{X}$, (iv) $\hat{\mu}_4 = X_1 + \frac{1}{3}X_2$, (v) $\hat{\mu}_5 = \frac{2}{3}X_1 + \frac{2}{3}X_2 - \frac{1}{3}X_3$. Which of these estimators are unbiased? Calculate bias for each estimator. Which estimator is the most efficient in terms of MSE?

Problem 2

Let X_1, X_2, X_3 be a random sample from a population with mean μ and variance σ^2. Consider the following estimator of variance σ^2:

$$\hat{\sigma}^2 = c(X_1 - X_2)^2.$$

Find constant c such that $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

Problem 3

Based on a random sample of two observations, consider two competing estimators of the population mean μ:

$$\bar{X} = (X_1 + X_2)/2$$

and

$$Y = \frac{1}{3}X_1 + \frac{2}{3}X_2.$$

- Are they unbiased?
- Which estimator is more efficient in terms of MSE?