HSE MDI: Mathematical Statistics

September 2022.

Class 4. Sampling Distribution. Point estimation. Unbiased estimators. Mean squared error.

Just 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, 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 $f(x)$.
- 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.

Sampling Distributions

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

- The sample mean, the arithmetic average of the values in a random sample. It is usually denoted by: $\bar{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$;
- The sample variance is the statistic defined by: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$;

- 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 $Y = T(X_1, \ldots, X_n)$ the **sampling distribution** of Y.

Statement:

Since every point estimator is a statistic, then estimators also have their sampling distributions, expectations, variances, and other characteristics 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:

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

Point Estimation

When sampling is from a population described by pdf or pmf $f(x|\theta)$, knowledge of θ *yields* knowledge of the entire population. Hence, the motivation is to seek a method of finding a good estimator of the point θ, that is, what we call a *good point estimator*.

Vague, but careful definition: A *point estimator* is any function $W(X_1, \ldots, X_n)$; so any statistic is a point estimator.

Important distinction: an *estimator* is a function from the sample, while an *estimate* is the realised value of an estimator obtained from a specific sample (when sample is actually taken).

If we are aimed at estimation of the unknown parameter θ of the population, we can denote its estimator as $\hat{\theta}$:

$$\hat{\theta} = W(X_1, \ldots, X_n).$$

Evaluating estimators

The bias of the estimator $\hat{\theta}$ is the difference between the expected value of $\hat{\theta}$ and true value of the parameter θ.

$$\mathrm{Bias}_\theta(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

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

Mean Squared Error

The Mean Squared Error of the point estimator $\hat{\theta}$ is defined by $E\left[\left(\hat{\theta}-\theta\right)^2\right]$. We also can interpret MSE of an estimator through its bias:

$$E\left[\left(\hat{\theta}-\theta\right)^2\right] = E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + E[\theta^2] =$$
$$\left(E[(\hat{\theta})^2] - E[\hat{\theta}]^2\right) + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + E[\theta^2] = \mathrm{Var}(\hat{\theta}) + \left(\mathrm{Bias}_\theta(\hat{\theta})\right)^2.$$

Thus, MSE incorporates two components: one measuring the variability of the estimator (precision), and the second measuring its bias (accuracy). An estimator with good MSE properties shall have small combined variance and bias.

For an unbiased estimator we have:

$$E\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathrm{Var}(\hat{\theta}),$$

so, if an estimator is unbiased, then its MSE equals to its variance only.

Problems

1. Random variable assumes values 0 and 1, each with probability 1/2.

   (a) Find population mean μ and variance σ^2

   (b) 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?

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.

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?