MTH209: Worksheet 6

Simulation Studies and Statistical Properties

In statistics, given data, we build estimators of various unknown parameters. This is a fundamental task in statistics. Let's take a simple example. Suppose

$$X_1, X_2, \dots, X_n \overset{iid}{\to} F_\theta\,,$$

where θ is an unknown parameter and the goal is to obtain an estimator of θ. Let T_n be the estimator of θ constructed from this sample. There are certain fundamental statistical properties that are often of interest about T_n. For instance, we would be interested in

1. Is T_n unbiased for θ? That is, is $\mathbb{E}(T_n) = \theta$?
2. What is $\text{Var}(T_n)$?
3. Is T_n consistent for θ? That is, does $T_n \overset{p}{\to} \theta$ as $n \to \infty$?
4. Is there an asymptotic distribution of T_n? That is, does there exist a $\sigma^2 > 0$ such that as $n \to \infty$

$$\sqrt{n}(T_n - \theta) \overset{d}{\to} N(0, \sigma^2)$$

A large part of statistical theory is to ensure/establish the above properties for estimators. As a reminder, recall the definitions of convergence in probability and convergence in distribution.

Convergence in Probability: We say that a sequence of random variables T_n **converges in probability** to T, written $X_n \overset{p}{\to} X$ if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|T_n - T| > \varepsilon\right) = 0.$$

Convergence in Distribution: Let $\{T_n\}_{n\geq 1}$ be a sequence of random variables and T be a random variable with distribution functions F_n and F, respectively. We say that T_n **converges in distribution** to T, written $X_n \overset{d}{\to} X$ if

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \text{for all } x \in \mathbb{R} \text{ at which } F \text{ is continuous.}$$

In this worksheet we will learn about how to "verify" these properties of estimators via simulation

Example: Sample Mean

Consider T_n to be the sample mean of the sample and θ to be the population mean. That is, we have

$$X_1, X_2, \dots, X_n \overset{iid}{\sim} F \text{ with mean } \theta \text{ and variance } \sigma^2$$

and let the estimator of θ chosen be

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

Then, we know the following already:

1. $\bar{X}_n$ is unbiased.
2. $\text{Var}(\bar{X}_n) = \sigma^2/n$
3. $\bar{X}_n$ is consistent for θ due to the law of large numbers
4. And due to the central limit theorem, as $n \to \infty$

$$\sqrt{n}(\bar{X}_n - \theta) \overset{d}{\to} N(0, \sigma^2)\,.$$

How do we "verify" this in R. We will do this in the first few problems below.

Further, given another estimator of θ, say T_n, we are also interested in knowing between T_n and $\bar{X}_n$ which one is better. Let us do these exercises in R.

Questions

1. Verify law of large numbers for the sample mean for the following F:
   a. $F = t_3$
   b. $F = t_2$
   c. $F = t_1$
2. For $F = N(0, 1)$, verify the mean and variance of the sample mean using replicated experiments (as discussed in class).
3. Verify central limit theorem for the following F for $n = 10, 100, 500$:
   a. $N(0, 1)$
   b. t_3
   c. $Gamma(a, b)$ for shape parameter $a = 100, 10, 1, .1, .01$ and rate parameter $b = 1$.
   d. t_2
4. Suppose $T_1 = \bar{X}$ and $T_2 =$ sample median are two estimators of the central parameter of a t_3 distribution ($\mu = 0$). I say an estimator is better than the other if it has smaller mean squared error:

$$\text{Mean Squared Error} = \mathbb{E}\left[(T_i - \mu)^2\right] = \text{Var}(T_i) + \left[\text{Bias}(T_i)\right]^2 \,.$$

   a. For $n = 10$, estimate the MSE for T_1 and T_2. Which one is better?
   b. Repeat for larger n. Does your answer change?
   c. Repeat for $F = N(0, 1)$.