Lecture 20, Mar 3, 2023

Random Sampling – Definitions

Definition

Population: The set of all possible observations, where each observation is a realization of a random variable

Sample: a subset of the population

- e.g. if we're measuring the heights of everybody in the world, then the population would be all the heights of everyone
- Each observation is the realization of a random variable

Definition

A *random sample* with n observations, where each observation is a realization of the random varaible $X_1, \cdots, X_n$, and we assume that $f(x_1, \cdots, x_n) = f(x_1) \cdots f(x_n)$; that is, the random variables are **independently and identically distributed**

A *statistic* is a function of the random variables $X_1, \cdots, X_n$

Definition

A sample is *biased* if it always leads to under or over-estimating some statistic of interest

- The distribution of a statistic is called a *sampling distribution* (e.g. the sampling distribution of $\bar{X}$ is $f(\bar{x})$)

Properties of Normal Random Variables

- Suppose X_1, X_2 are independent and normally distributed with means μ_1, μ_2, variances σ_1^2, σ_2^2, then $X_1 + X_2$ is still normally distributed, with mean $\mu = \mu_1 + \mu_2$ and variance $\sigma = \sigma_1^2 + \sigma_2^2$
  - The mean and variance holds for any independent variable, but only when both X_1, X_2 are normal is $X_1 + X_2$ also normal
- $\frac{1}{n} X_1$ is also normal, with $\mu = \frac{\mu_1}{n}, \sigma^2 = \frac{\sigma_1^2}{n^2}$
- Suppose $X_1, \cdots, X_n$ are IID with μ, σ^2, then $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is normal with mean μ and variance $\frac{\sigma^2}{n}$
  - This follows directly from the above 2 properties
  - Notice σ^2 decreases with increasing n – that is, the more data we have in our sample, the closer we will get to the true mean
  - The standard deviation decreases on the order of $\sqrt{n}$

The Central Limit Theorem

Theorem

Central Limit Theorem: Assume a sample with $X_1, \cdots, X_n$ identically and independently distributed with mean μ and finite variance σ^2 (with no restrictions on the distribution otherwise), where the sample mean is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, let

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

then as $n \to \infty$, the distribution of Z_n converges to the standard normal

$$\lim_{n\to\infty} f(z_n) = n(z_n; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_n^2}{2}}$$

That is, the distribution of $\bar{X}_n$ itself approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$

- Z_n is never truly normally distributed but approaches a normal distribution as n gets bigger
- It doesn't matter what the actual distribution is, as we take more samples, the distribution of the average is going to look more and more like the Gaussian with smaller and smaller variance as the sample size increases
- The standard deviation of $\bar{X}_n$ is approximately $\frac{\sigma}{\sqrt{n}}$
- Example: a runner averages $\mu = 4$ minutes per mile with standard deviation $\sigma = 5$ seconds, what is the chance that the mean time of the next 20 races less than 3:58?
  - We want $P(\bar{X} < 238) = P\left(\frac{\bar{X} - 240}{\frac{5}{\sqrt{20}}} < \frac{-2}{\frac{5}{\sqrt{20}}}\right) = P(Z_{20} \leq -1.8) \approx \Phi(-1.8) = 0.036$