Data part 2

STAT-S520

Arturo Valdivia

02-21-23

- These slides complement material from ISI Chapter 8

The sum and the average of a random sample

Let $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} \mathbb{P}$ be a random sample with $EX_i = \mu$ and $VarX_i = \sigma^2$ for $i = 1, \ldots, n$ and let's define

$$Y = \sum_{i=1}^{n} X_i$$

↳ this is also a Random Variable

as the sum of the random sample and

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

as the *sample mean* or the average of the random sample.

Expected value and variance of Y and $\bar{X}_n$

Let's use the properties of the expected value and variance to obtain:

- $EY = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} EX_i = \sum_{i=1}^{n} \mu = n\mu$
- $VarY = Var\left(\sum X_i\right) = \sum_{i=1}^{n} Var X_i = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$
- $E\bar{X}_n = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{\cancel{n}} \cancel{n}\mu = \mu$
- $Var\bar{X}_n = Var\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^{\cancel{2}}} \cancel{n}\sigma^2 = \frac{\sigma^2}{n}$

$$\boxed{Var\,\bar{X}_n = \frac{\sigma^2}{n}}$$

Simulations in R part 1

Let's work with simulations in R to show, approximately, that the expected value and variance of Y and $\bar{X}_n$ are the ones shown above.

The Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} \mathbb{P}$ with $EX_i = \mu$ and $VarX_i = \sigma^2$ for $i = 1, \ldots, n$ and

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

It can be shown that $\bar{X}_n$ converges in probability to μ, or in math notation:

$$\bar{X}_n \xrightarrow{p} \mu$$

as $n \to \infty$.

For some small $\varepsilon > 0$ (arbitrarily small)

$$P(\bar{X}_n \in (\mu - \varepsilon, \mu + \varepsilon)) \to 1$$

as $n \to \infty$

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} \mathbb{P}$ with $EX_i = \mu$ and $VarX_i = \sigma^2$ for $i = 1, \ldots, n$,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \qquad \text{and} \qquad Z \sim \text{Normal}(0, 1)$$

and we define F_n as the CDF of Z_n and Φ as the CDF of Z. The CLT states that, for any real number z,

$$F_n(z) \to \Phi(z)$$

as $n \to \infty$.

Practical use of the CLT

- The previous result tells us that when $n \to \infty$ the distribution of $\bar{X}_n$ is normal.
- However, for fairly small values of n, $\bar{X}_n$ is already approximately normally distributed
  - The rule of thumb commonly used is $n \geq 30$

Simulations in R part 2

Let's work with simulations in R that show that the WLLN and the CLT hold, approximately.