Week 8 – Central Limit Theorem

Juan Pablo Lewinger

Last Class

Markov's inequality: for any RV $U \geq 0$:

$$P(U \geq t) \leq \frac{E[U]}{t} \qquad \text{for any } t > 0$$

Chebysheb's inequality: for any RV X, $Var[X_i] < \infty$:

$$P\left(|X - E[X]| \geq t\right) \leq \frac{Var[X]}{t^2} \qquad \text{for any } t > 0$$

Law of large numbers: $X_1, X_2, \ldots$ i.i.d with finite expectation μ:

Weak: $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$

Strong: $P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$

Basic Statistical Model

- Want to learn characteristics of a population:
  - Income of Los Angeles residents
  - Blood pressure of patients form LA county hospital
  - Vaping in among teenagers in California
  - Genotypes at a gene among individuals with Hispanic ancestry
- We model the distribution of the characteristic as a random variable X (e.g. height, blood pressure, vaping (yes vs. no), Genotypes (0,1,2)))
- To learn about X we collect a random sample: $X_1, X_2, \ldots, X_n$ from $F_X(x)$, the distribution of X
- $X_1, X_2, \ldots, X_n$ have the same probability distribution and are mutually independent.
- The distribution $F(x; \theta)$ of X_i is unknown or only partially known
- E.g. $X_i \sim N(\mu, \sigma^2)$ with known σ^2 but unknown μ
- We conceptualize the observed data $x_1, \ldots, x_n$ as realizations of the random variables $X_1, \ldots, X_n$

Sample statistics

Sample statistics are empirical summaries of the data:

- sample mean: $\bar{x}_n = \frac{x_1 + \ldots + x_n}{n}$
- sample variance: $s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$ (sometimes $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$)
- sample minimum: $x_{(1)} = \min(x_1, \ldots, x_n)$

We think of these as realizations of the corresponding random variables:

- $\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$
- $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$
- $X_{(1)} = \min(X_1, \ldots, X_n)$

Central Limit Theorem

$X_1, \ldots, X_n$ i.i.d. $E[X_i] = \mu$, $Var[X_i] = \sigma^2$ then:

Sum form: $S_n = X_1 + \ldots + X_n$

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{Var[S_n]}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} N(0,1)$$

Sample mean form: $\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$

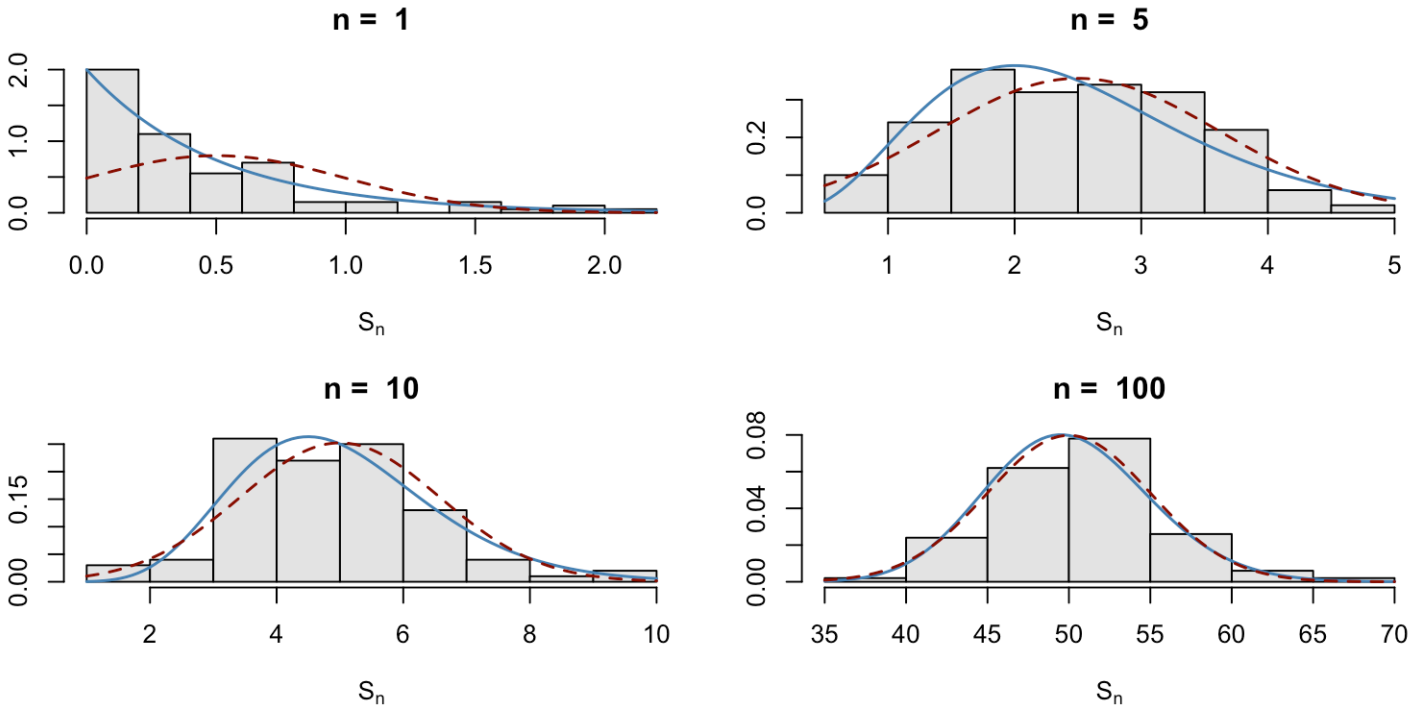
$$Z_n = \frac{\bar{X}_n - E[\bar{X}_n]}{\sqrt{Var[\bar{X}_n]}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1)$$

The symbol $\xrightarrow{D}$ denotes convergence in distribution. It means that as $n \to \infty$ the cumulative distribution function of $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ gets closer and closer to the cdf of a standard normal.

Formally, if $F_n(x) = F_{Z_n}(x)$ denotes the cdf of Z_n, then $F_n(x) \to \Phi(x)$ for every $x \in \mathbb{R}$, where $\Phi(x)$ is the cdf of a $N(0,1)$ RV.

Central Limit Theorem

Sum of n i.i.d. exponentials: $S_n = X_1 + \ldots + X_n, X_i \sim Exp(2)$



Blue line is pdf of S_n; Red line is pdf of a normal $N(\frac{n}{2}, \frac{n}{4})$

Example

A runner attempts to pace a 100m race. Her strides (steps) are independently distributed with mean $\mu = 0.97$ meters and a standard deviation of $\sigma = 0.1$. What is the (approximate) probability that her 100 strides differ from 100 meters by no more than five meters?

Estimation

- We want to learn about a distribution in a population
- e.g. proportion of democrat voters in CA, average blood pressure among covid survivors aged 70+, vaping frequency among young adults in the US, rate of patient ER night admissions in LA county hospitals
- We take a sample of size n from the population and conceptualize it as a random sample $X_1, \dots, X_n$ from a distribution $F_X(x)$ that is **totally or partially unknown to us**.
- We want to estimate specific characteristics or **parameters** of the underlying distribution:
- true mean $E[X_i] = \mu$ (e.g. X_i blood pressure)
- the true variance, $Var[X_i] = \sigma^2$ (e.g. X_i blood pressure)
- or a probability like $P(X_i = 1)$ (e.g. 1 = democrat voter; 0 = not democrat voter)
- or a rate (e.g. expected number of ER patients per hour) (X_i = number of patients within a period of 1-hour)

Estimation

Example 1: To estimate the unknown mean of a population μ (e.g. blood pressure)

- Natural to use the sample mean $\bar{X}_n$ to estimate μ because we know that for large n the sample mean will be close to the true mean.

Example 2. We can model the number of arrivals per hour at an ER unit as a $Poisson(\lambda)$. Suppose we count the arrivals during each on n (non-overlapping) 1-hour intervals to get the random sample $X_1, \ldots, X_n$. Here λ is unknown and we want to estimate it.

- A natural estimate is also the sample mean $\bar{X}_n$ because $E[X_i] = \lambda$.
- But using the sample variance S_n^2 is also reasonable because $Var[X_i] = \lambda$
- In future classes we'll learn how to choose among different options for estimating a parameter of interest

Estimator vs. estimate

- Estimate: value t that only depends on the dataset $x_1, x_2, \ldots, x_n$, i.e., t is some function of the dataset $t = h(x_1, x_2, \ldots, x_n)$. Example: $t = \bar{x}_n = \frac{x_1 + \ldots + x_n}{n}$
- *An estimate is a number* (or a vector of numbers in multi-parameter estimation problems)
- Estimator: Let $t = h(x_1, x_2, \ldots, x_n)$ be an estimate based on the dataset $x_1, x_2, \ldots, x_n$. Then t is a realization of the random variable $T = h(X_1, X_2, \ldots, X_n)$. The random variable T is called an estimator.
- *An estimator is a random variable*
- *An estimate is a realization of random variable*

Sampling distribution

Example: estimating the proportion p of LA teenagers that vape from a sample $Y_1, \dots Y_n$, $Y_i \sim Bernoulli(p)$.

- Y_i records whether the teenager vapes (yes=1 vs. no=0)
- The sample proportion $\widehat{p}_n = \frac{S_n}{n}$ of vapers is a natural estimator of p ($S_n = Y_1 + \dots + Y_n$ is the total number of vapers in the sample), because by the LLN $\widehat{p}_n \to p$
- The sampling distribution is just the standard distribution (pdf, pmf, cdf) of the random variable we call the estimator.
- For example of vapers, the sampling distribution of $\widehat{p}_n$ is the distribution of the random variable $\widehat{p}_n = \frac{S_n}{n}$

Unbiased Estimators

$X_1, \dots X_n$ a random sample from distribution F, and θ a parameter of interest about F (e.g. mean)

- An estimator T of a parameter θ is unbiased if $E[T] = \theta$
- An unbiased estimator has no systematic tendency to produce estimates that are larger than or smaller than the target parameter θ
- The difference $E[T] - \theta$ is called the bias of the estimator T
- If $bias \neq 0$ the estimator is called biased

Sample Mean and sample variance

- $X_1, \dots X_n$ a random sample with mean $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$
- The sample variance $\bar{X}_n$ is an unbiased estimator of the population mean μ.

$$E[\bar{X}_n] = \mu$$

- The sample mean $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$ is an unbiased estimator of the population variance σ^2

$$E[S_n^2] = \sigma^2$$

Unbiasedness is not preserved under general transformations

- In general, if T is an unbiased estimator of θ, $g(T)$ is not an unbiased estimate of $g(\theta)$
- Example:

  $\bar{X}_n$ is unbiased for $E[X_i] = \mu$

  But $\bar{X}_n^2$ is not unbiased for $E[X_i]^2 = \mu^2$, Because by Jensen's inequality $E[\bar{X}_n^2] > E[\bar{X}_n]^2 = \mu^2$
- Jensen's inequality: if $g(t)$ is a convex function, then $E[g(T)] \geq g(E[T])$. Equality holds only if $g(x) = at + b$
- If g linear $g(t) = at + b$, $E[g(T)] = E[aT + b] = aE[T] + b = g(E[T])$ so $g(T)$ is unbiased for $g(\theta)$.

Next week

- Read *PSD 5.6.2, 5.6.3*