Lecture 11

Sampling Distribution & Central Limit Theorem

Readings: IntroStat Chapters 4 & 5

STAT 8010 Statistical Methods I
May 31, 2023

Whitney Huang
Clemson University

Agenda

1. Sampling Distribution
2. Central Limit Theorem (CLT)
3. Chi-Square, Student's t-, and F-Distributions

Sampling Distribution

- Independent random variables $X_1, X_2, \cdots, X_n$ with the same distribution are called a random sample

Sampling Distribution

- Independent random variables $X_1, X_2, \cdots, X_n$ with the same distribution are called a random sample

- A statistic is a function of a random sample

  Example:

Sampling Distribution

- Independent random variables $X_1, X_2, \cdots, X_n$ with the same distribution are called a random sample

- A statistic is a function of a random sample

Example:

  - Sample mean: $\bar{X}_n = \sum_{i=1}^{n} X_i / n$

Sampling Distribution

- Independent random variables $X_1, X_2, \cdots, X_n$ with the same distribution are called a random sample

- A statistic is a function of a random sample

  Example:

  - Sample mean: $\bar{X}_n = \sum_{i=1}^{n} X_i / n$

  - Sample variance: $\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 / (n-1)$

Sampling Distribution

- Independent random variables $X_1, X_2, \cdots, X_n$ with the same distribution are called a random sample

- A statistic is a function of a random sample

  Example:

  - Sample mean: $\bar{X}_n = \sum_{i=1}^{n} X_i/n$
  - Sample variance: $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2/(n-1)$
  - Sample maximum: $\max_{i=1}^{n} X_i$

Sampling Distribution

- Independent random variables $X_1, X_2, \cdots, X_n$ with the same distribution are called a random sample

- A statistic is a function of a random sample

  Example:

  - Sample mean: $\bar{X}_n = \sum_{i=1}^{n} X_i / n$

  - Sample variance: $\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 / (n-1)$

  - Sample maximum: $\max_{i=1}^{n} X_i$

- The probability distribution of a statistic is called its sampling distribution

Example

Suppose $X_1, X_2, \cdots, X_n$ is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ population, Find the sampling distribution of sample mean.

Example

Suppose $X_1, X_2, \cdots, X_n$ is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ population, Find the sampling distribution of sample mean.

$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} = \sum_{i=1}^n \frac{1}{n} X_i$. From last lecture we know that sum of normal r.v.s is still a normal r.v. Hence we only need to figure its mean and variance.

$$\mathrm{E}[\bar{X}_n] = \sum_{i=1}^n \frac{1}{n}\mu = \mu$$

$$\mathrm{Var}[\bar{X}_n] = \sum_{i=1}^n \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

Therefore, we have $\bar{X}_n \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$

Central Limit Theorem (CLT)

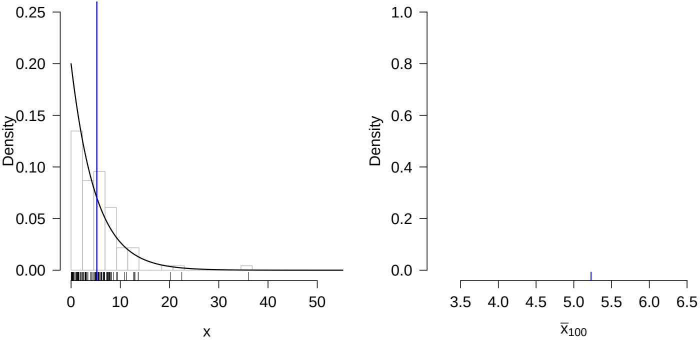
CLT

*The **sampling distribution** of the **mean** will become approximately **normally distributed** as the **sample size becomes larger**, **irrespective of the shape of the population distribution!***

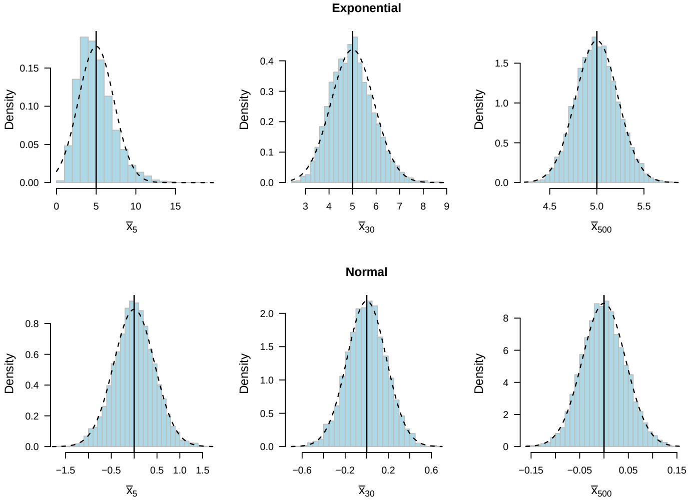
Let $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} F$ with $\mu = \mathrm{E}[X_i]$ and $\sigma^2 = \mathrm{Var}[X_i]$.
Then $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{d} \mathrm{N}(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$.

CLT In Action

1. Generate 100 (n) random numbers from an Exponential distribution (population distribution)
2. Compute the sample mean of these 100 random numbers
3. Repeat this process 120 times



CLT: Sample Size (n) and the Normal Approximation



Why CLT is Important?

- In many cases, we would like to make statistical inference about the population mean μ
  - The sample mean $\bar{X}_n$ is a sensible estimator for the population mean
  - CLT tells us the **distribution** of our estimator $\Rightarrow \bar{X}_n \approx \mathrm{N}(\mu, \frac{\sigma^2}{n})$
- Applications: Confidence Interval, Hypothesis Testing

CLT for Sample Proportions

When (binary) observations are independent and the sample size is sufficiently large, the sample proportion of success, denoted by $\hat{p}$, will tend to follow a normal distribution with the following mean and variance:

$$\mu_{\hat{p}} = p; \qquad \mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

$\hat{p} = \frac{X}{n}$, where X is a binomial random variable with parameters n and p. Then we have

$$\mathrm{E}[\hat{p}] = \mathrm{E}[X/n] = \frac{1}{n}\mathrm{E}[X] = \frac{1}{n}np = p$$

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}[X/n] = \frac{1}{n^2}\mathrm{Var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Normal distribution approximation is obtained based on normal approximation to binomial when n sufficiently large (e.g., $np \leq 5$ and $n(1-p) \geq 5$)

Chi-Square (χ^2) Distribution

If $Z_1, \cdots, Z_k$ are independent, standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom. It is usually denoted as

$$Q \sim \chi_k^2$$

Applications

Chi–squared test for assessing

- Goodness of fit
- Independence
- Homogeneity

Student's t Distribution

If $Z \sim N(0,1)$ and $V \sim \chi^2_k$ are independent, then the random variable:

$$\frac{Z}{\sqrt{V/k}}$$

follows a t-distribution with k degrees of freedom.

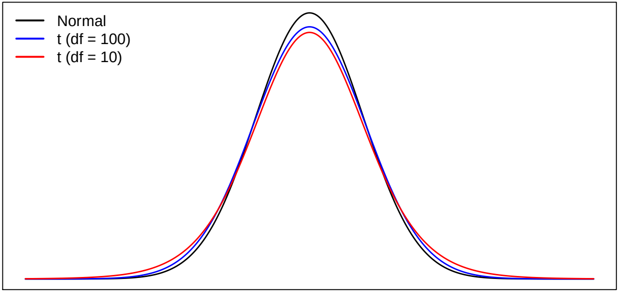
Applications:

CLT with known σ:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0,1)$$

CLT with unknown σ:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \xrightarrow{d} t_{\nu=n-1}$$



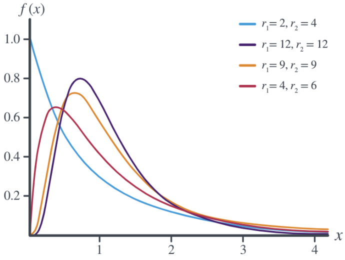
F-Distribution

If U and V are independent chi-square random variables with and degrees of freedom, k_1 and k_2, respectively, then:

$$F = \frac{U/r_1}{V/r_2}$$

follows an F-distribution with numerator degrees of freedom r_1 and denominator degrees of freedom r_2. We write

$$F \sim F_{r_1, r_2}$$



Applications:

- Testing the equality of variances of two normal populations
- Testing the equality of means of k (>2) normal populations $\Rightarrow$ ANOVA

Summary

In this lecture, we learned

- Sampling Distributions
- Central Limit Theorem (CLT)
- Chi-Squared, Student's t, and F-distributions