PS 2010: 8. Estimation and Inference 1

Qing Chang

Department of Political Science
University of Pittsburgh

Fall 2023

- Homework 5

Agenda

- Sampling and sampling distribution
- Bias and consistency
- Confidence interval

Random Sampling

- A **population** consists of the totality of the observations with which we are concerned.
- A **sample** is a subset of a population
- A **sample statistic** is a numerical measure describing some aspect of a sample
  - Sample mean and variance from repeated draws

  A **point estimator** is a rule (formula) for calculating a sample statistic that can be used as an estimate of a population parameter.
- The probability distribution of a statistic is called a **sampling distribution**

Sample Mean and Variance

Let x_1, x_2, x_3 ... x_n are randomly draw from a distribution, then

- Sample mean: $\bar{x} = \frac{1}{n}\sum_i^n x_i$
- Sample variance: $s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$
- Sample sd: $s = \sqrt{s^2}$

Example: A comparison of coffee prices at 4 randomly selected grocery stores in San Diego showed increases from the previous month of 12, 15, 17, and 20 cents for a 1-pound bag. Find the variance of this random sample of price increases.

Sampling Distribution

- $\bar{x}$ and s^2 depends on the underling distribution, the size of the sample, and the method of choosing the samples. Therefore, they are random variables
- The probability distribution of these random variables (we call them statistics) are sampling distribution
- Again, this sampling distribution comes from repeated sampling with the same size from the same population
- This sampling distribution help us to make inference about population parameters: population mean and variance

Sampling Distribution of Sample Mean

If Let X_1, X_2, X_3 ... X_n are randomly sample of size n from a normal distribution with mean μ and variance σ^2. Then

$$\bar{X} = \frac{1}{n}\sum_i^n X_i$$

is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$

- We could write $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- It also means we could transform it to $z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Central Limit Theorem

- In the above case, sampling distribution of $\bar{X}$ is normally distributed because we draw data from normal distribution
- However, this conclusion is generally true even we draw data from unknown distribution, provided that the sample size is large ($n \geq 30$)

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $N(0, 1)$.

Example

- Central Limit Theorem (CLT) has very important application in topics such as hypothesis testing, estimation, etc. Here are some examples:
- Achievement test scores of all high school seniors in a state have mean 60 and variance 64. A random sample of n $= 100$ students from one large high school had a mean score of 58. Is there evidence to suggest that this high school is inferior?
- An important manufacturing process produces a component parts for the automotive industry. An experiment is conducted in which 100 parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation is 0.1 millimeter. What is the probability that the sample mean will be within 0.027 millimeter of the population mean.

Sampling Distribution of the Difference between Two Means

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

is approximately a standard normal variable.

Example

Given the following information on two population

Population A	Population B
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

What is the probability that a random sample of size 36 from A will have a mean that is at least 1 more than the mean of a sample size of 49 from B?

Sampling Distribution of Sample Variance

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{\sigma^2}$$

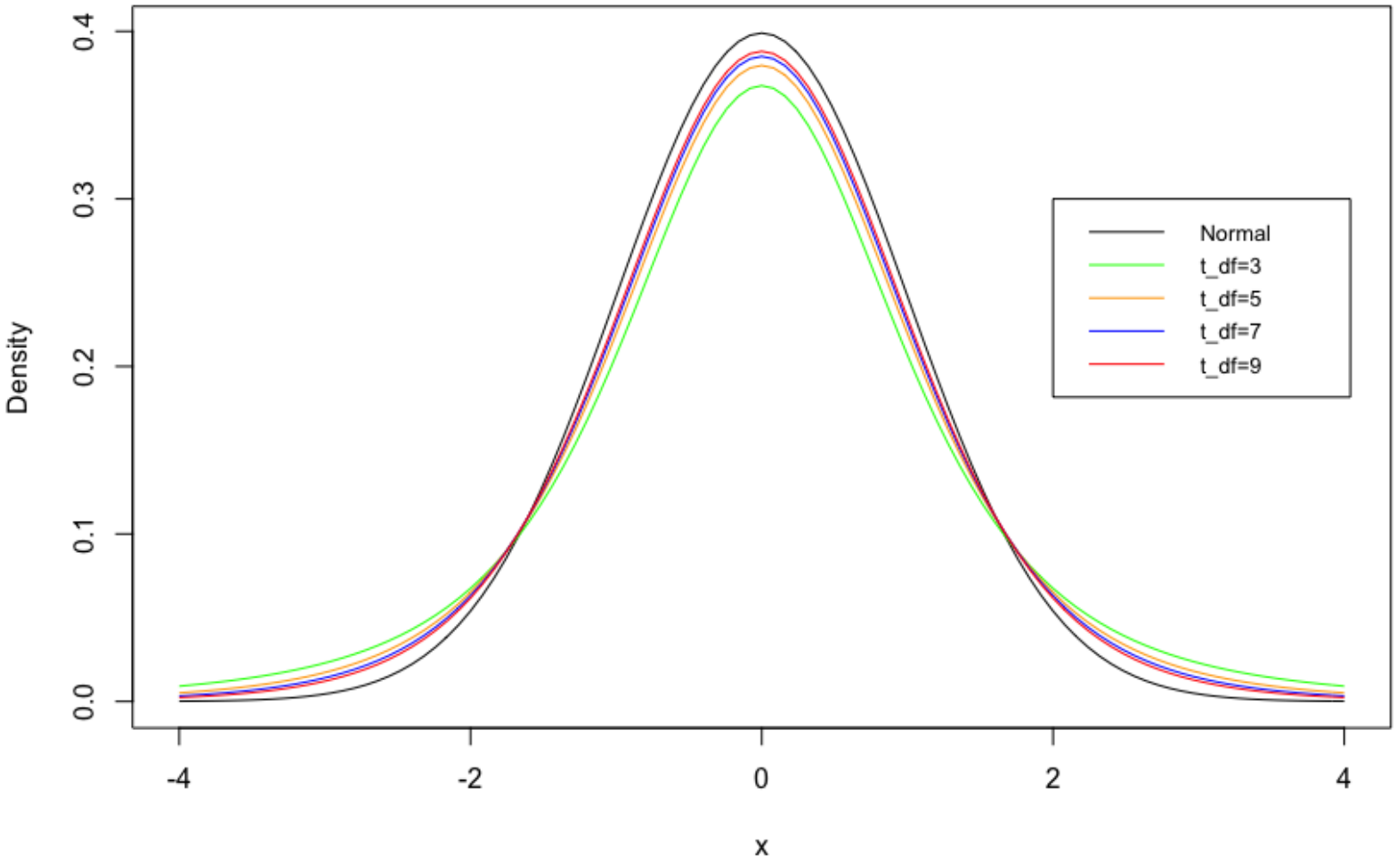
has a chi-squared distribution with $n-1$ degrees of freedom.

- Notice: $z - score = \frac{X-\mu}{\sigma}$

Student t Distribution Revisit

- In the sampling distribution of sample mean, we assume we know σ
- However, this is not the case for many applications
- We could use the sample standard deviation s to substitute σ
- If the population distribution is normal, this makes the sampling distribution of the sample mean as a student t distribution
- When sample size ≤ 30, better to use student t
- When sample size > 30, ok to use standard normal

t distribution



Example

A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he samples 25 batches each month. If the computed t-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim. What conclusion should he draw from a sample that has a mean $\bar{x} = 518$ grams per milliliter and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.

Statistical Inference

- **Statistical inference** involves drawing conclusions about features of an underlying population using relatively small samples of data from that population.
  - Frequentist
  - Bayesian
- Estimation and Hypotheses testing
- A **point estimate** of a population parameter is a single value of a statistic
  - Examples: sample mean
- However, an estimator is not expected to estimate the population parameter without error.
  - Sample size, sample value, sample method

Bias

- To measure a goodness of an estimator we can use bias measure
- An unbiased point estimator produces a sample statistic with sampling distribution that has a mean equal to the population parameter the statistic is intended to estimate.

  Let $\hat{\theta}$ be a point estimator for a parameter θ. Then $\hat{\theta}$ is an unbiased estimator if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta, \hat{\theta}$ is said to be biased.
- The bias of a point estimator $\hat{\theta}$ is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Example

If X_i are normally distributed random variables with mean μ and variance σ^2, then:

$$\bar{X} = \frac{\sum X_i}{n} \text{ and } S^2 = \frac{\sum \left(X_i - \bar{X}\right)^2}{n}$$

- Prove that $\bar{X}$ and S^2 are biased/unbiased estimator for μ and σ^2
- If it is biased estimator, what is the bias?