BIOSTAT 702: Module 2

Estimation

Dr. Marissa Ashner

Department of Biostatistics and Bioinformatics

Fall 2025

Duke University School of Medicine

Module Goals

- Understand the basics of estimating a parameter from a sample
- Understand sampling distributions and how they can be used for interval estimation and eventual inference

Resources for this Module

Textbooks

- ST21: Chapters 7 and 10
- ADLM: Chapter 2, Section 2

Websites

- Understanding Sampling Distributions
- Reeses Pieces Sampling Simulation
- Simulating Confidence Intervals

Motivation

$$\bar{Y}, s^2 \longrightarrow \mu, \sigma^2$$

sample parameter

Why is estimation so important in statistics?

- We rarely have access to gather data on every person in the population we are interested in drawing conclusions about
- We therefore need to take a *sample* from this population, and use their data to *estimate* what we are truly interested in
  - Hopefully this would be generalizable to the whole population

Sampling

Consider this (very simple) research question: What is the average height of all Duke students?

- *Population of Interest:* All Duke Students
- *Parameter of Interest:* Average height (inches) – μ
- Assume the height of Duke Students (random variable Y) follows some *distribution* with mean (or expected value) $E(Y) = \mu$ and variance σ^2
- μ and σ^2 are *unknown*. We therefore must *estimate* them using *statistics* from a *sample* of the population.
- *Sample:* How could we draw a sample from this population? (i.e., $Y_1, \ldots, Y_n$)
  - Primary Goal: ensure the sample is representative of the population of interest

Estimator of a Population Mean

$\bar{Y}, s^2 \longrightarrow \mu, \sigma^2$

- **Estimator:** a formula used to calculate an estimate from sample data
- The best estimator of a population mean is simply the sample mean: $\bar{Y} = \sum_{i=1}^{n} Y_i$
- The best estimator of the population variance is the sample variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
  - Why do we divide by $n - 1$? unbiasedness
- *Note:* When we say "best", we mean it is unbiased and is the most efficient of all estimators of that type

best: unbiased,
most efficient

Statistics as Random Variables

- The estimators derived previously are called *sample statistics* and they themselves are random variables
- As you might imagine, different samples would yield different estimates
  - Also, any sample drawn would lead to an estimate that likely differs at least some from the unknown truth μ
- These statistics therefore have probability distributions (like all random variables) to describe the likelihood of obtaining one realization from the estimator

Sampling Distributions

$E(\bar{Y})$ 样本均值 sample mean

$Var(\bar{Y}) = \sigma^2/n$

$SE(\bar{Y}) = \sigma/\sqrt{n}$

- The sampling distribution of a statistic is the probability distribution from a sample of size n of that statistic
- This means that $\bar{Y}$ and s^2 both have their own sampling distributions

Standard Error

- The standard deviation of a sampling distribution is referred to as the *standard error*
- The SE for the sampling distribution of the mean is $\boxed{SE = \sigma/\sqrt{n}}$, estimated by $\boxed{\hat{SE} = \hat{\sigma}/\sqrt{n}}$ $\approx s/\sqrt{n}$

样本标准差

总体 $\xrightarrow{SD}$ 样本 $\xrightarrow{var(Y)}$ 统计推断 $var(\bar{Y})$

$\sigma^2 = \sqrt{\frac{(Y_i)^2}{n}}$ $s^2 = \sqrt{\frac{(Y_i)^2}{n-1}}$

SE

(描述样本均值)

standard deviation

$\left.\begin{matrix}\sigma \\ s\end{matrix}\right\}$ SD $\sqrt{var(Y)}$ $s = \sqrt{\frac{\Sigma(X_i-\bar{X})^2}{n-1}}$

描述
样本离散程度 数据描述

原始数据的波动

standard Eorror

样本均值的波动

SE $\sqrt{var(\bar{Y})}$ $SE = S/\sqrt{n}$

样本统计量 推断统计

standardized Mean difference

SMD $SMD = \frac{X_1 - X_2}{s_{pooled}}$

组间差异

$E(X)$ Expectation 长期平均

$var(X)$ variance 方差.

$var(Y) = \sigma^2$

$var(\bar{Y}) = \sigma^2/n$

$SE(\bar{Y}) = \sqrt{var(\bar{Y})} = \sigma/\sqrt{n}$

$E[Y] = \mu$

$var(\bar{Y}) = \sigma^2/n$

$var(X) = E(X^2) - [E(X)]^2$

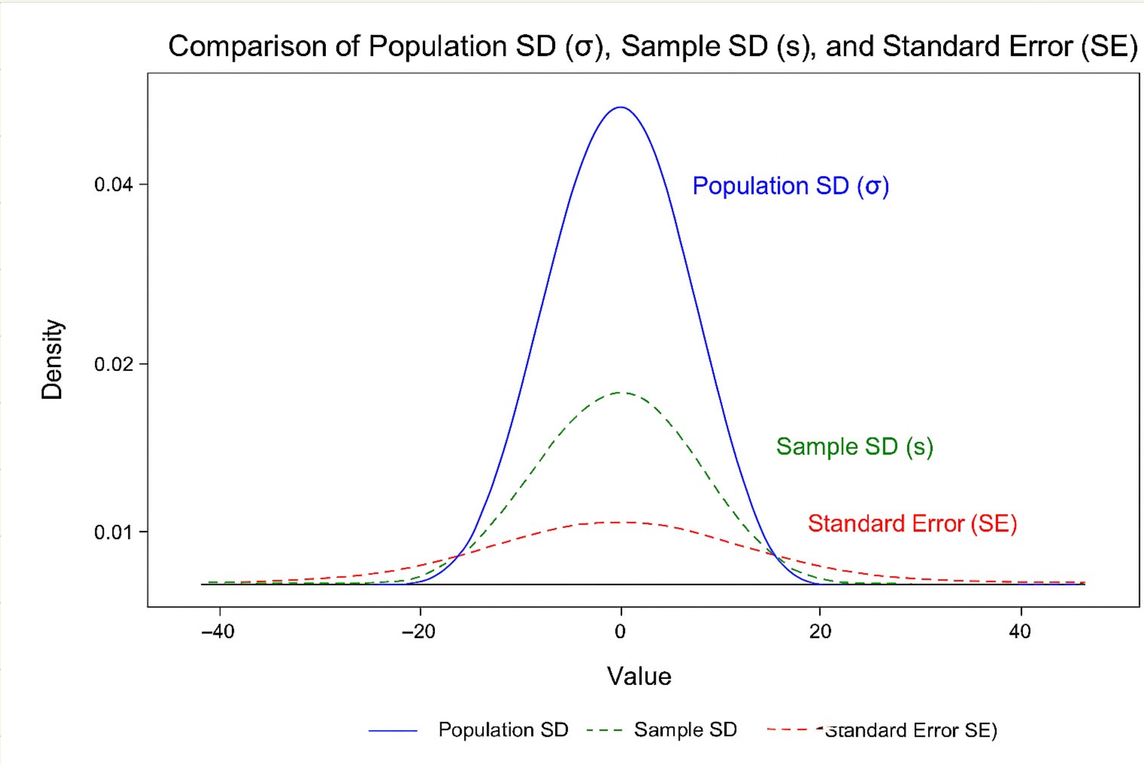
用样本推断整体

$E(X) = \sum_i x_i P(X = x_i)$

$SD = \sqrt{var(X)}$

$SE(\bar{Y}) = \sigma/\sqrt{n}$

$= s/\sqrt{}$


Comparison of Population SD (σ), Sample SD (s), and Standard Error (SE)
Population SD (σ)
Sample SD (s)
Standard Error (SE)
Density
0.04
0.02
0.01
−40
−20
0
20
40
Value
Population SD
Sample SD
Standard Error SE)

Sampling Distribution of $\bar{Y}$

- The *mean* of the sampling distribution is the unknown true parameter, μ
- The *standard deviation* is the standard error, as discussed
- But what is the actual *shape* of the sampling distribution?
  - This is important to know to make inferences about the statistics

The Central Limit Theorem

sample mean: $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

sample sum: $\sum x \sim N(n\mu, (\sqrt{n})(\sigma))$

- According to the *Central Limit Theorem* (CLT), the sampling distribution of the mean becomes normally distributed as the sample size increases, regardless of the shape of the original random variable
  - This means that the distribution of $\bar{Y}$ will become normal as n increases, even if the distribution of Y is very skewed/non-normal
  - For smaller samples, we can use the t distribution instead if the population shape is assumed to be (close to) normal
  - Even for larger samples, when we are estimating the SE (almost always), the t distribution is more accurate as it accounts for the estimation of the SE
  - For smaller samples that aren't normal-looking, may have to consider other approaches (will talk about this more later)

Point Estimates vs. Interval Estimates

- We have talked about estimating the population mean μ using the sample mean $\bar{Y}$
  - This is a *point estimate*
- We also mentioned how it is very likely for $\bar{Y} \neq \mu$, even if it's our best guess
- In order to *quantify the uncertainty* around our estimate, we can calculate an interval estimate instead
  - i.e., a Confidence Interval
  - CI = point estimate $\pm$ critical value $*$ SE
  - where the critical value depends on the assumed distribution and the confidence level

$t_{0.975, df}$

$\pm 2.5\%$

    - For a 95% CI assuming normality, the critical value is $z_{0.95} = 1.96$.

$z_{0.975}$

    - For a one-sample CI like this, we usually use the t distribution critical values instead, since we are always estimating the SE

Estimation More Broadly

- We can estimate many other things aside from the mean of a population, but we are using this as our starting point example
  - We will refer back to the ideas learned from this lecture as we estimate other quantities throughout the course
- So far, we have talked about point and interval estimation, but not inference, which is very related
  - This will come up in the next lecture!

What is the primary difference between standard deviation and standard error?

1. Standard deviation describes variability in individual data points, while standard error describes variability in sample statistics like the mean. — 82% — 27
2. Standard deviation is only used in hypothesis testing, while standard error is used to describe variability in raw data. — 3% — 1
3. Standard deviation measures how far a sample mean is from the population mean, while standard error measures how spread out the data are. — 12% — 4
4. There is no difference; the terms are interchangeable. — 3% — 1