Sampling Distributions and Sampling Error

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- When data are collected, we generally assume that they come from some unknown probability distribution.
- Usually with the goal of estimating a population parameter.
- How good is the estimate?
- We need some kind of measure to determine this.

Example 1

- Assume a random variable X is uniformly distributed between 0 and 20.
- Using the theoretical formulas from a uniform distribution, we know:
  - $E[X] = (0 + 20)/2 = 10$
  - $V[X] = (20 - 0)^2/12 = 33.33$
- In R, simulate four samples with $n_1 = 5$, $n_1 = 50$, $n_1 = 100$, and $n_1 = 500$,from this distribution.
- Comment on the sample mean $(\bar{x})$ and sample variance (s^2) of each sample.

Example 2

- Use the Mean_Uniform_Histogram() function written in the example R code to do the following:
  - Generate a histogram of the means of 100 samples of $n = 10$ from a uniformly distributed variable between 0 and 20.
  - Generate a histogram of the means of 100 samples of $n = 20$ from a uniformly distributed variable between 0 and 20.
  - Generate a histogram of the means of 100 samples of $n = 100$ from a uniformly distributed variable between 0 and 20.
- What do you notice about the shape of the histograms?

Standard Error of the Mean

- The histograms in Example 2 are visualizations of the *sampling distribution of the mean*.
- We notice that the distributions become more compact as the sample sizes increase.
  - *Larger sample sizes have less sampling error.*
- **Standard Error of the Mean** $= \sigma/\sqrt{n}$
  - σ is the *known* population standard deviation.

Example 3

- We know the population standard deviation in Example 2.
- Use this and the sample sizes from Example 2 to compute the standard error of the mean for all three samples.
- Do these results support what you found visually?

Comments on Standard Error

- We will never know the *actual* population standard deviation.
- May only be able to take one sample of n observations.
- We can estimate the population standard deviation with the sample standard deviation.

Central Limit Theorem

- The Central Limit Theorem (CLT) is a very important practical result in statistics.
- The **CLT** states that if the sample size is large enough, **the sampling distribution of the mean is approximately normally distributed**, regardless of the distribution of the population.
- **The mean of the sampling distribution will be the same as that of the population.**
- **If the population is normally distributed, then the sampling distribution of the mean will also be normal for any sample size.**
- Can use probabilities from normal distributions to draw conclusions about sample means.

Normal (Gaussian) Distribution

- Continuous symmetric distribution described by a classic *bell shape*.
- One of the most widely used distributions in statistics.
- Has two parameters:
  - The mean μ (location)
  - The variance σ^2 (scale)
- $E[X] = \mu$
- $V[X] = \sigma^2$

Example 4

- Assume a random variable X follows a standard normal distribution with $\mu = 1$ and $\sigma^2 = 2$. Determine:
  - $P(X < 0.5)$
  - $P(-0.1 < X < 0.1)$
  - $P(X > 1)$

Exercise 1

- Assume X follows an exponential distribution with $\lambda = 2$.
- Calculate the theoretical expected value and variance. Use these results to repeat Example 1 and Example 2.
- *Note: You will need to make your own function to generate the histograms.*

Exercise 2

- Assume a random variable X follows a standard normal distribution with $\mu = 2$ and $\sigma^2 = 2$. Determine:
  - $P(X < 0.55)$
  - $P(-0.2 < X < 0.1)$
  - $P(X > 0.75)$

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.

- https://en.wikipedia.org/wiki/Central_limit_theorem