Central Limit Theorem

Ryan Miller

Outline

1. Central Limit theorem
   - assumptions, results, and implications
2. Practice using Central Limit theorem

Introduction

- Statisticians will often focus on single numbers that summarize trends within sample data
  - The sample mean, denoted $\bar{x}$, summarizes the center of numerical data

Introduction

- Statisticians will often focus on single numbers that summarize trends within sample data
  - The sample mean, denoted $\bar{x}$, summarizes the center of numerical data
- Because the act of obtaining sample data is a random process, $\bar{x}$ is an observed realization of a continuous *random variable*
  - That is, if the process used to collect our data were repeated many times, we'd expect different values of $\bar{x}$ each time (and these values would follow some probability model)

The sample average as a random variable

- With modern computing, it's relatively easy to study the behavior of $\bar{x}$ across different random samples
- The Sampling Distribution for a mean section of StatKey is a nice interactive tool for understanding the random process of acquiring sample data
- Particularly important is the role of n
  - When n is large, the distribution of sample means tends to be symmetric and bell-shaped, regardless of how the data itself is distributed (with the exception of extreme outliers)

Central Limit theorem (CLT)

- Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with a common expectation, $E(X)$, and a common standard deviation, $SD(X)$
- If $\bar{X}$ denotes the average of these random variables, then:

$$\sqrt{n}\left(\frac{\bar{X} - E(X_i)}{SD(X_i)}\right) \to N(0, 1)$$

With some abusive notation, CLT suggests:

$$\bar{x} \sim N\left(E(X), \frac{SD(X)}{\sqrt{n}}\right)$$

Partial justification

Notice $\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_n$, so using our knowledge of linear combinations of random variables:

$$E(\bar{X}) = \frac{1}{n}E(X_1) + \ldots + \frac{1}{n}E(X_n) = \frac{1}{n}(n * E(X))$$

Partial justification

Notice $\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_n$, so using our knowledge of linear combinations of random variables:

$$E(\bar{X}) = \frac{1}{n}E(X_1) + \ldots + \frac{1}{n}E(X_n) = \frac{1}{n}(n * E(X))$$

Similarly:

$$Var(\bar{X}) = \frac{1}{n^2}Var(X_1) + \ldots + \frac{1}{n^2}Var(X_n) = \frac{1}{n^2}(n * Var(X))$$

So:

$$SD(\bar{X}) = \frac{SD(X)}{\sqrt{n}}$$

Establishing Normality requires a more complicated proof that is beyond the scope of this course (but recognize we studied this empirically using StatKey)

Proportions as averages

The sample proportion is really just an average of n independent Bernoulli random variables:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \qquad \hat{p} = \frac{1+0+0+1+\ldots+1}{n}$$

Applying the Central Limit theorem, and considering what you know about Bernoulli random variables, what is an approximate distribution for $\hat{p}$?

Proportions as averages

The sample proportion is really just an average of n independent Bernoulli random variables:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \qquad \hat{p} = \frac{1+0+0+1+\ldots+1}{n}$$

Applying the Central Limit theorem, and considering what you know about Bernoulli random variables, what is an approximate distribution for $\hat{p}$?

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Practice

According to the National Center for Health Statistics, the distribution of serum cholesterol levels for 20- to 74-year-old males living in the United States has mean 211 mg/dl, and a standard deviation of 46 mg/d

1) Suppose we plan to collect a sample of 25 individuals and measure their cholesterol levels. What is the probability that the *sample average* will be above 230?
2) If we plan to a collect a sample of 50 individuals, what is the probability that the *sample average* will be above 230?

Practice (solution)

1) CLT suggests N(211, $46/\sqrt{25}$) as a model for $\bar{x}$, using `pnorm` we find $P(\bar{X} \geq 230) = 0.0174$
2) CLT suggests N(211, $46/\sqrt{50}$) as a model for $\bar{x}$, using `pnorm` we find $P(\bar{X} \geq 230) = 0.0014$

Practice (continued)

According to the National Center for Health Statistics, the distribution of serum cholesterol levels for 20- to 74-year-old males living in the United States has mean 211 mg/dl, and a standard deviation of 46 mg/d

1) Suppose we are planning to collect a sample of 25 individuals and measure their cholesterol levels. What two values would we expect the middle 95% of the sample averages to fall between?
2) If we plan to collect a sample of 50 individuals, what two values would we expect the middle 95% of the sample averages to fall between?

Practice (solution)

1) Using `qnorm`, in 95% of random samples of size $n = 25$ the mean will fall between 193.36 and 228.64
2) Using `qnorm`, in 95% of random samples of size $n = 50$ the mean will fall between 198.53 and 223.47

Remarks (sample size)

- Central Limit theorem illustrates the connection between sample size and the amount of uncertainty present in the sample data
  - Larger sample sizes will produce estimates with lower variability

Remarks (next steps)

- The power of the Central Limit theorem is that it allows us to build reliable probability models for things we have minimal data on
  - Very often, we'll use the *sample mean* and *standard deviation* to model the *sampling distribution*

Remarks (next steps)

- The power of the Central Limit theorem is that it allows us to build reliable probability models for things we have minimal data on
  - Very often, we'll use the *sample mean* and *standard deviation* to model the *sampling distribution*
- As we'll soon see, this will provide us a framework for two fundamental statistical techniques:
  - **Confidence Intervals** - a method of estimation that takes into account statistical uncertainty in the sample data
  - **Hypothesis Tests** - a method for determining whether associations seen in the sample data might be explained by *random chance*