Confidence Intervals and the Central Limit Theorem

Ryan Miller

Outline

1. The Normal distribution
2. Central Limit Theorem
3. Confidence intervals using the Central Limit theorem

Introduction

Lately we've been focused on addressing the issue of *sampling variability* using confidence intervals. The general procedure looks like:

1) Use the sample data to find a *point estimate* of the population parameter of interest

Introduction

Lately we've been focused on addressing the issue of *sampling variability* using confidence intervals. The general procedure looks like:

1) Use the sample data to find a *point estimate* of the population parameter of interest
2) Bootstrap the sample data to mimic the process of sampling from the population (ie: sampling variability)

Introduction

Lately we've been focused on addressing the issue of *sampling variability* using confidence intervals. The general procedure looks like:

1) Use the sample data to find a *point estimate* of the population parameter of interest
2) Bootstrap the sample data to mimic the process of sampling from the population (ie: sampling variability)
3) Construct a *confidence interval* by using the bootstrap distribution to estimate the point estimate's *standard error* (2-SE method) or by finding percentiles among the bootstrapped estimates (percentile method)

Generalizing the 2-SE method

- The *2-SE method* works when the bootstrap distribution is bell-shaped (due to the 68-95-99)
  - Using the **Normal Distribution**, this approach can be generalized to confidence intervals (of any confidence level):

$$\text{Point Estimate} \pm c * SE$$

- Within the interval's MOE, the multiplier, c, can be adjusted to achieve any desired confidence level

The Normal distribution

The **Normal curve**, or Normal probability function, is a mathematical function that yields a bell-shaped distribution:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

- The parameter μ is a constant that defines the *center* of the bell-curve
- The parameter σ is a constant that defines the *standard deviation* of the bell-curve (how peaked or flat it is)

The Normal distribution

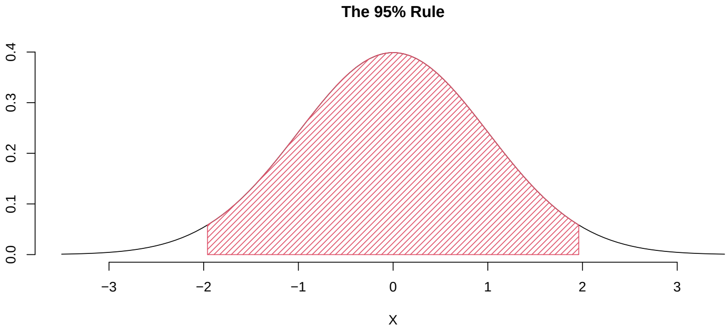
The **Normal curve**, or Normal probability function, is a mathematical function that yields a bell-shaped distribution:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

- The parameter μ is a constant that defines the *center* of the bell-curve
- The parameter σ is a constant that defines the *standard deviation* of the bell-curve (how peaked or flat it is)
- There infinitely many different Normal curves, one for each combination of μ and σ
  - We will describe them using the notation: $N(\mu, \sigma)$

The Normal distribution

If data follow a Normal distribution, the *area under the curve* describes the likelihood you see a value within a particular range:



Because the Normal probability function doesn't have a closed-form integral, we must use software to find these areas

Practice

The Theoretical Distributions section of StatKey allows us to find areas under various Normal curves:

1) Consider the **Standard Normal distribution**, or N(0,1), what values define the middle 90% of this distribution?
2) Consider a N(10,5) distribution, what proportion of this distribution is larger than 16?

Practice (solution)

1) The values of -1.645 and +1.645 define the middle 90% of the curve. This suggests we could use 1.645 as a multiplier on the SE to form a 90% CI estimate (if the sampling distribution is approximately Normal).
2) The area to the right of 16 is 0.115 on the N(10,5) curve. This suggests there's a 11.5% chance of seeing a value 16 or larger if the data follows this distribution.

Central Limit Theorem

The **Central Limit Theorem** (CLT) is a theoretical result that establishes a Normal distribution (with known SE!) for a variety of different sample estimates (provided a sufficient sample size):

$$\text{Sample Estimate} \sim N(\text{Population Parameter}, SE)$$

- The sample size needed for this theoretical result to hold differs depending on the parameter we're estimating
  - For example, $n = 30$ is generally deemed sufficient when estimating μ, a population mean
- CLT also provides a mathematical formula for an estimate's SE (see later slides)
  - The details of this formula will differ depending on the parameter we're estimating

Central Limit Theorem (one proportion)

For a *single proportion*, CLT states:

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

- This result implies that $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and a value of c from the N(0,1) curve can be used to obtain a P% CI estimate of p
- The sample size condition to use this result is $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

Practice

A 2021 study looked at the true-positive rate (sensitivity) of an Abbott Diagnostics rapid test for Covid-19 (one of the earliest such tests). Of the 84 cases with symptomatic Covid-19 who took the test, 38 had a "positive" result. Our goal is to estimate p, the overall sensitivity of this test in the target population (ie: all symptomatic Covid cases).

1) Verify that the conditions are met to use the CLT Normal approximation to construct a confidence interval estimate
2) Find the values of $\hat{p}$, its SE, and the value of c needed to construct a 99% CI estimate of p
3) Calculate and interpret the 99% CI

Practice (solution)

1) First, $\hat{p} = 38/84 = 0.452$. Then, $n\hat{p} = 84 * 0.452 = 38$ and $n(1 - \hat{p}) = 84 * (1 - 0.452 = 46$. Because both are larger than 10, the CLT Normal approximation is reasonable.
2) $\hat{p} = 38/84 = 0.452$, $SE = \sqrt{\frac{0.452(1-0.452)}{84}} = 0.054$, and $c = 2.576$ (this defines the middle 99% of a N(0,1) curve)
3) The 99% CI is $0.452 \pm 2.576 * 0.054 = (0.313, 0.591)$. Our sample suggests, with 99% confidence, that the true sensitivity of the Abbott rapid test is somewhere between 31.3% and 59.1%

Central Limit Theorem (two proportions)

For a *difference of two proportions*, CLT states:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

- Using the sample proportions: $\hat{p}_1$ and $\hat{p}_2$, as well as their denominators: n_1 and n_2 this result can be used to find the *SE* necessary to construct a confidence interval estimate of $p_1 - p_2$
- The sample size condition to use this result is $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$

Practice

The previously mentioned study also examined a test produced by Siemens. Of 72 cases with symptomatic Covid-19 that took the Siemens test, 39 had a "positive" result. Recall that 38 of 84 symptomatic cases tested positive on the Abbott test. Our goal is estimate $p_1 - p_2$, the difference in sensitivity of these two tests (at the population level)

1) Let $\hat{p}_1 = 38/84 = 0.45$ be the sample proportion for the Abbott test, and $\hat{p}_2 = 39/72 = 0.54$ be the sample proportion for the Siemens test. Find the SE for the difference in proportions, $\hat{p}_1 - \hat{p}_2$.
2) Using the CLT Normal approximation, find and interpret a 95% CI estimate for $p_1 - p_2$

Practice (solution)

1) $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.45(1-0.45)}{84} + \frac{0.54(1-0.54)}{72}} = 0.08$
2) Since $c = 1.96$, $\hat{p}_1 - \hat{p}_2 = 0.54 - 0.45 = -0.09$, and $SE = 0.08$, we calculate: $-0.09 \pm 1.96 * 0.08 =$ (-0.247, 0.067). This represent the plausible range of differences in sensitivity of these tests at the population level (estimated with 95% confidence). Because zero is included in this interval, it's plausible that the tests are actually no different.

Central Limit Theorem (one mean)

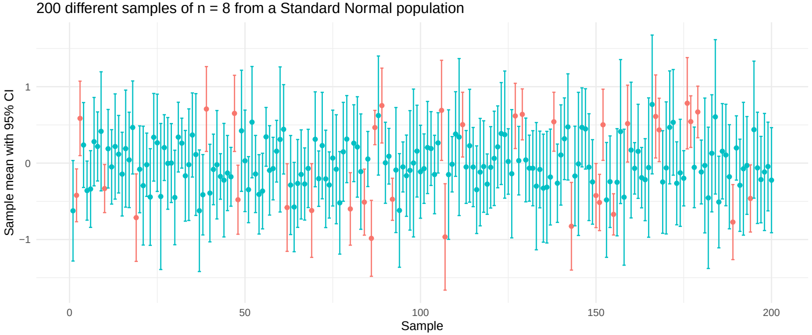
For a *single mean*, CLT states:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- σ is the standard deviation of the population

William Gosset and the t-distribution

- Different from our examples involving proportions, the previous CLT result involves a *second unknown parameter*, σ (the population's standard deviation)
  - It seems natural to simply replace this with an estimate from the sample, s, but this is what happens:



William Gosset and the t-distribution

- Clearly this 95% CI procedure is *invalid* - too many of these intervals do not contain μ
- William Gosset, a chemist working for Guinness Brewing, became aware of this issue in the 1890s
  - His work evaluating the yields of different barley strains frequently involved small sample sizes

William Gosset and the t-distribution

- Clearly this 95% CI procedure is *invalid* - too many of these intervals do not contain μ
- William Gosset, a chemist working for Guinness Brewing, became aware of this issue in the 1890s
  - His work evaluating the yields of different barley strains frequently involved small sample sizes
- In 1906, Gosset took a leave of absence from Guinness to study under Karl Pearson (developer of the correlation coefficient)
  - Gosset discovered the issue was due to using s interchangeably with σ

William Gosset and the t-distribution

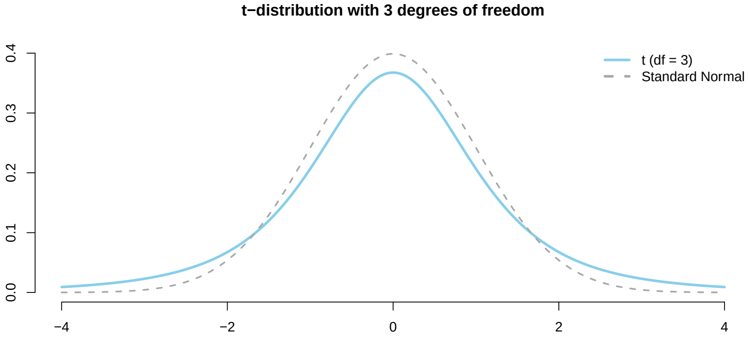
- Treating s as if it were a perfect estimate of σ results in a systematic underestimation of the total amount of variability involved in making the CI
  - To account for the additional variability introduced by estimating σ using s, a modified distribution that's slightly more spread out than the Standard Normal curve must be used

William Gosset and the t-distribution

- Treating s as if it were a perfect estimate of σ results in a systematic underestimation of the total amount of variability involved in making the CI
  - To account for the additional variability introduced by estimating σ using s, a modified distribution that's slightly more spread out than the Standard Normal curve must be used
- Typically the inventor of a new method gets to name it after themselves
  - However, Gosset was forced to publish his new distribution under the pseudonym "student" because Guinness didn't want it's competitors knowing they employed statisticians!
  - Student's t-distribution is now among the most widely used statistical results of all time

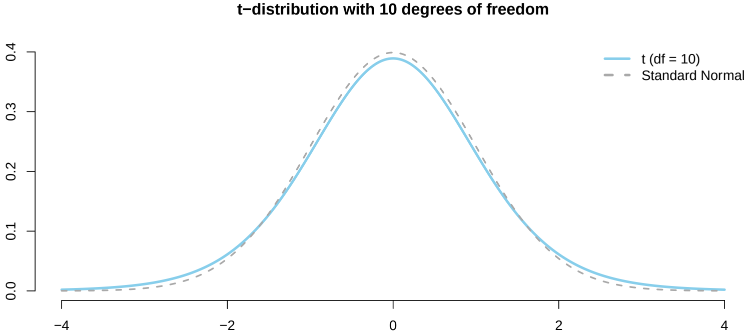
The t-distribution

The *t*-distribution accounts the additional uncertainty in small samples using a parameter known as *degrees of freedom*, or *df*:

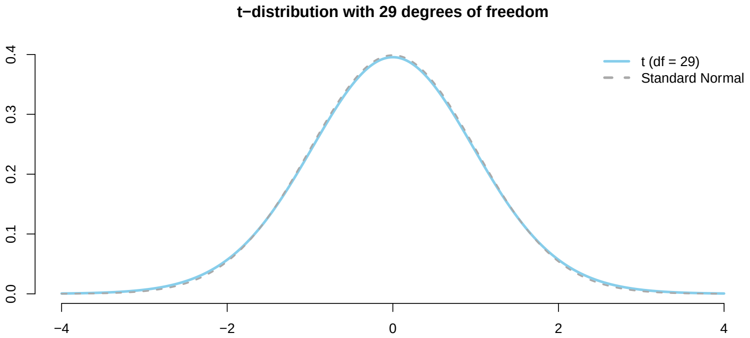


When estimating a single mean, $df = n - 1$

The t-distribution



The t-distribution



Practice

While waiting at an airport, a traveler notices 6 flights to similar a similar part of the country were delayed 6, 10, 13, 23, 45, 55 minutes. The mean delay in this sample was 25.33, with a sample standard deviation of $s = 20.2$. Assuming these data are a representative sample, answer the following:

1) How many degrees of freedom are involved when using the t-distribution to form a CI estimate? What is the value of c that should be used for 95% confidence?
2) What is the 95% CI estimate for the average delay of flights to the part of the country this traveler is heading?

Practice (solution)

1) Because $n = 6$, we'd use $df = n - 1 = 5$. For $df = 5$, $c = 2.571$ defines the middle 95% of the distribution.
2) Point Estimate $\pm$ MOE, Point estimate $= \bar{x} = 25.33$, Margin of error $= c * SE = 2.571 * \frac{20.2}{\sqrt{6}}$
   - All together, 95% CI: $25.33 \pm 2.571 * \frac{20.2}{\sqrt{6}} = (4.1, 46.5)$
   - We are 95% confident the *average* delay is somewhere between 4.1 minutes and 46.5 minutes

Note: if we'd erroneously used a Normal model (instead of the *t*-distribution), we'd get an interval that is much narrower (9.2, 41.5), but this interval wouldn't have the confidence level we are advertising (ie: it wouldn't really be a 95% CI because it would miss too often)

When to use the t-distribution

- The t-distribution was designed for small, Normally distributed samples
  - However, it can also be reliably used on large samples, regardless of their shape

	Sample data are approximately Normal	Sample data are non-Normal or skewed
Sample size is large ($n \geq 30$)	Use t-distribution	Use t-distribution
Sample size is small ($n < 30$)	Use t-distribution	*do not* use t-distribution

- For small, non-Normal samples, more robust methods (such as bootstrapping) should be used instead

Central Limit Theorem (two means)

For a *difference of two means*, CLT states:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

- Similar to applications estimating a single mean, the t-distribution should be used when s_1 and s_2 are used as estimates of σ_1 and σ_2
    - Degrees of freedom is complicated, we'll use the smaller of $n_1 - 1$ and $n_2 - 1$ as a conservative approach

Practice

To explore whether artificial light at night contributes to weight gain (in g), researchers randomly assigned 18 young mice to live in lab environments with either complete darkness or an artificial nightlight during evening hours:

Summary Statistics

Statistics	Light	Dark	Overall
Sample Size	10	8	18
Mean	6.732	4.114	5.568
Standard Deviation	2.966	1.557	2.729
Minimum	1.71	2.27	1.71
Q_1	4.99	2.68	4.00
Median	6.19	4.11	5.16
Q_3	9.17	5.28	6.94
Maximum	11.67	6.52	11.67

1) Compare the means and medians of each group as a crude assessment of whether its reasonable to assume these data came from a Normally distributed population
2) Find a 95% CI estimate for the difference in mean weight gain experienced in each group (Light - Dark)

Practice (solution)

1) Because the means and medians are reasonably close, we do not have a sufficient reason to doubt Normality
2) First, we should use $df = 7$ because $n_2 - 1$ is smaller than $n_1 - 1$. Thus, $c = 2.365$ is necessary for 95% confidence. Next, $SE = \sqrt{2.966^2/10 + 1.557^2/8} = 1.09$, therefore the 95% CI estimate is $(6.732 - 4.114) \pm 2.365 * 1.09 = (0.04, 5.20)$. With 95% confidence we can conclude that light-exposed mice exhibit a larger weight gain, with the average difference being between $+0.04$g and $+5.20$g relative to mice without exposure.

Central Limit Theorem (summary)

Estimate	Standard Error	CLT Conditions
$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$np \geq 10$ and $n(1-p) \geq 10$
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	normal population or $n \geq 30$
$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i \in \{1, 2\}$
$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	normal populations or $n_1 \geq 30$ and $n_2 \geq 30$
r	$\sqrt{\frac{1-\rho^2}{n-2}}$	normal populations or $n > 30$

Factors impacting CI width (summary)

If all other factors are held constant, the table below summarizes the impact of certain changes on the width of confidence intervals:

Change	Impact on CI width
Increasing n	decreases width (narrower CI)
Increasing confidence level	increases width (wider CI)
Increasing SE	increases width (wider CI)
Increasing number of bootstrap samples (if bootstrapping)	no impact on width