Confidence Intervals

Ryan Miller

Outline

2. Confidence Intervals
    - Definition and key concepts
3. Using Central Limit theorem to find confidence interval estimates
    - Formulas for commonly used descriptive statistics

Introduction

- For the remainder of the semester our focus will be on connecting *descriptive statistics* and *probability*
  - Our fundamental question is:

Given the trends present in the sample data, what can we reliably conclude about the broader population?

Introduction

- For the remainder of the semester our focus will be on connecting *descriptive statistics* and *probability*
  - Our fundamental question is:

Given the trends present in the sample data, what can we reliably conclude about the broader population?

- For example, if we observed a sample of $n = 30$ with a mean total cholesterol level of $\bar{x} = 230$, can we confidently say the sample came from a population with *elevated cholesterol*? (defined as ≥ 220)

Point vs. interval estimates

There are two primary ways in which descriptive statistics from sample data are reported:

1) **Point estimate** - A *single value* describing what the sample data suggests is most likely true of the population (ie: $\bar{x}$ is a point estimate for μ)
2) **Interval estimate** - A *range of values* describing what the sample data suggest might plausibly be the true population parameter

Point estimates are nice, but they're *almost always wrong* (at least to some degree) due to *sampling variability*.

Interval estimates

Most interval estimates follow the format:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

- The margin of error (MOE) is intended to account for sampling variability in the data
  - To be meaningful, it must be *calibrated* to balance precision (how narrow the interval is) with reliability (how likely is the interval to contain the truth about the population)

Confidence intervals

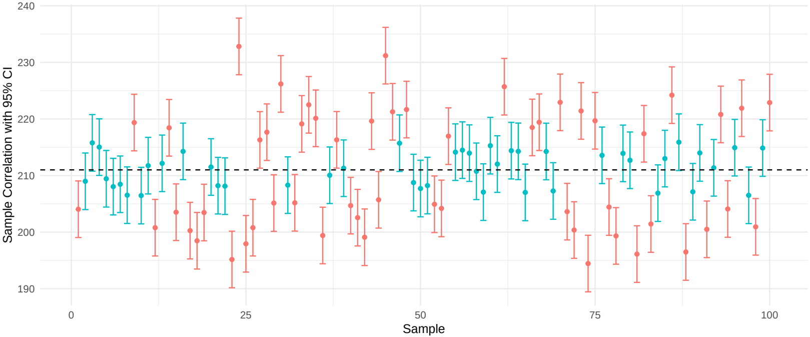
A **confidence interval** is a type of interval estimate where the margin of error is *calibrated* to achieve a *long-run* success rate known as the **confidence level**:

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

- c is a *calibration constant* needed to achieve a certain confidence level
- The *standard error*, or SE, expresses the amount of variability in the point estimate

Confidence intervals - graphical intuition

In our cholesterol example, suppose the population we sampled from actually has a mean of $\mu = 211$ and a standard deviation of $\sigma = 46$. Shown below are $\bar{x}$ ± 5 (an arbitrarily chosen MOE) for *100 different random samples*:



Remarks (graphical intuition)

- 45 of 100 random samples produced an interval estimate that contained $\mu = 211$ (the true value of the population parameter we were trying to estimate)
    - Thus, an arbitrarily chosen margin of error of 5 corresponds to a confidence level of roughly 45% (in this application)

Using Central Limit theorem

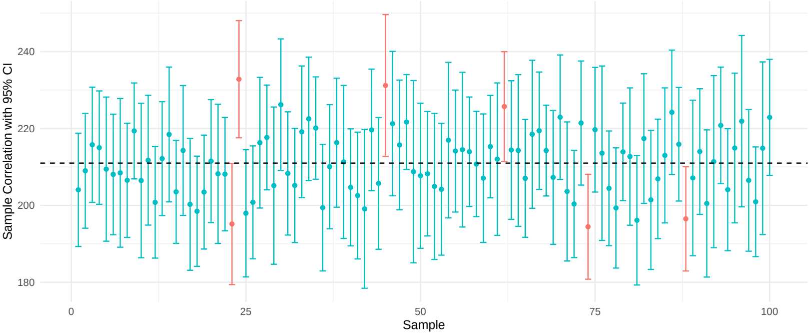
Recall that CLT suggests the following probability model for the sample average:

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

- We do not know the true mean and standard deviation of the cases within the population, but it's logical to estimate them via $\bar{x}$ and s (the sample mean and sample standard deviation)
    - We can use this probability model to determine a better margin of error

Confidence intervals - graphical intuition (part 2)

The graph below again depicts 100 different random samples, but this time the interval endpoints are the 2.5th and 97.5th percentiles from the Normal model described in the previous slide:



Notice how these intervals contain $\mu = 211$ across approximately 95% of different random samples!

Confidence intervals - general procedure

The following *generalized procedure* can be used to construct a P% confidence interval:

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

- c is a quantile that defines the middle P% of the standard Normal curve
    - For example, $c = 1.96$ is used for a 95% confidence level
- SE is estimated from the data and relies upon results from the CLT
    - For example, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is used when estimating a population proportion

Practice

A study conducted by Johns Hopkins University Hospital found that 31 of 39 babies born in their facilities at 25 weeks gestation (15 weeks early) went on to survive. Suppose the goal is to estimate the proportion of babies born under similar circumstances in similar hospitals that will survive.

1) What is the population parameter of interest? What is our *point estimate* of it?
2) Applying the CLT, what is the *SE* of our point estimate?
3) If we'd like a 98% confidence interval, what value of *c* should we use? (Hint: use `qnorm` in R)
4) Combining parts 1-3, what is the 98% CI in this example?

Practice (solution)

1) We want to estimate p, the proportion of all babies born 15 weeks early in hospital systems similar to Johns Hopkins that will survive. Our point estimate is the sample proportion, $\hat{p} = 31/39 = 0.795$
2) CLT suggests $SE = \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$ for a single proportion. Plugging in our point estimate, this is $SE = 0.065$
3) Using `qnorm()`, we find $c = 2.326$ appropriate for achieving a 98% confidence level
4) $0.795 \pm 2.326 * 0.065 = (0.644, 0.946)$, so our sample suggests, with 98% confidence, that the survival rate is anywhere between 64.4% and 94.6% at comparable hospitals

Categorical vs. quantitative outcomes

When using $\hat{p}$ to estimate p, CLT suggests:

$$SE = \sqrt{p(1-p)/n}$$

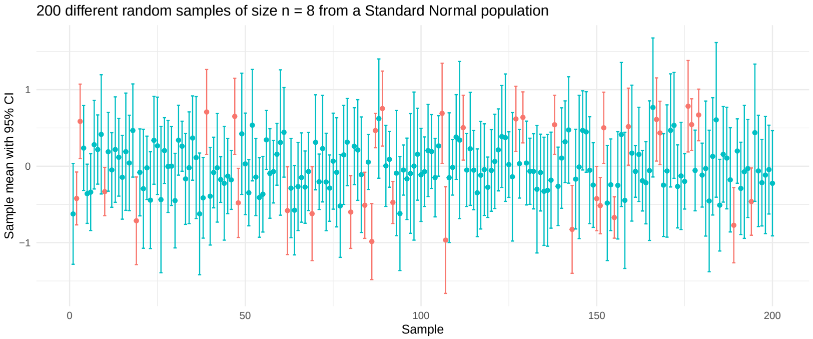
When using $\bar{x}$ to estimate μ, CLT suggests:

$$SE = \sigma/\sqrt{n}$$

- For a categorical outcome, the *SE* formula is entirely based upon the *point estimate* and the *sample size*
    - However, for a quantitative outcome, the *SE* formula involves an additional unknown parameter, σ (the standard deviation of cases *within the population*)

William Gosset and the t-distribution

For quantitative data, it seems reasonable to simply replace σ with an *estimate from the sample*, s, but this is what happens:



William Gosset and the t-distribution

- Clearly this procedure for constructing 95% CIs is *invalid*, too many random samples produced intervals that didn't contain μ
- William Gosset, an employee at Guinness Brewing, became aware of this issue in the 1890s
  - His work evaluating the yields of different barley strains often involved statistical analyses of small, Normally distributed samples

William Gosset and the t-distribution

- Clearly this procedure for constructing 95% CIs is *invalid*, too many random samples produced intervals that didn't contain μ
- William Gosset, an employee at Guinness Brewing, became aware of this issue in the 1890s
  - His work evaluating the yields of different barley strains often involved statistical analyses of small, Normally distributed samples
- 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* (sample standard deviation) interchangeably with σ (population standard deviation)

William Gosset and the t-distribution

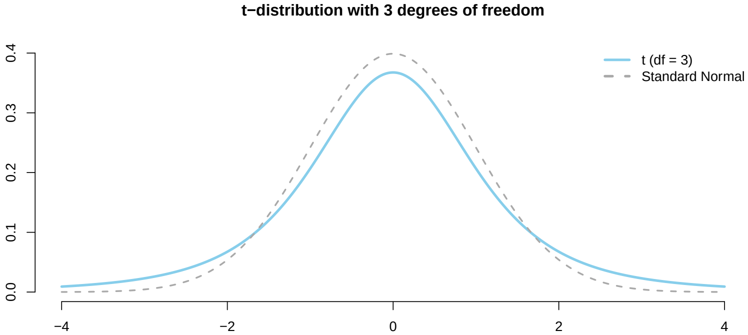
- Treating s as if it were a perfect estimate of σ results in a systematic underestimation of the total amount of variability involved in estimating μ
    - To account for the additional variability introduced by estimating σ using s, Gosset proposed a modified distribution that's slightly more spread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

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 estimating μ
    - To account for the additional variability introduced by estimating σ using s, Gosset proposed a modified distribution that's slightly more spread out than the Standard Normal curve
- 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

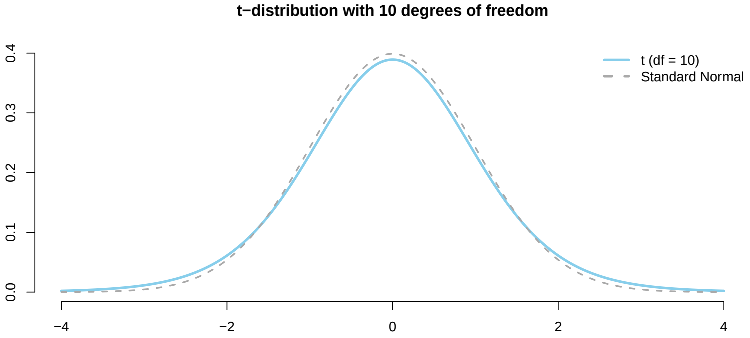
The t-distribution

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



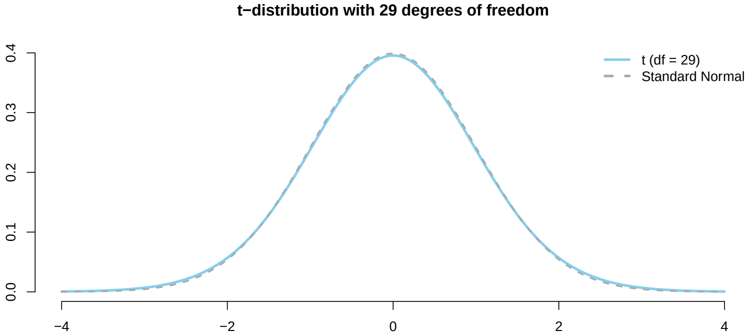
When estimating a single mean, $df = n - 1$

The t-distribution



To achieve the same level of confidence, one must go further into the tails of the *t*-distribution (as it's more spread out)

The t-distribution



As *df* increases, the *t*-distribution becomes more similar to the Normal curve (nearly indistinguishable past $n = 30$)

The t-distribution in R

The qt function is akin to qnorm, and it can be used to find *c* when calculating confidence intervals for quantitative data

```r
qt(0.975, df = 5)
```

```
## [1] 2.570582
```

```r
qt(0.975, df = 20)
```

```
## [1] 2.085963
```

```r
qt(0.975, df = 200)
```

```
## [1] 1.971896
```

```r
qnorm(0.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

The pt function is akin to pnorm, we'll use it in the future

```r
pt(2, df = 10, lower.tail = FALSE)
```

```
## [1] 0.03669402
```

```r
pnorm(2, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 0.02275013
```

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 of a Normally distributed population, answer the following:

1) What value of c should be used to properly calibrate a 95% confidence interval estimate?
2) Use these data to find a 95% CI estimate for the average delay of all flights to the part of the country where this traveler is heading.

Practice (solution)

1) For $df = 5$, the value $c = 2.571$ defines the middle 95% this t-distribution (found using `qt()`)
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: had we 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 correct confidence level (ie: it wouldn't be properly calibrated)

Confidence interval misconceptions

1) Confidence intervals are a statement about a *population parameter*, not the sample data

Confidence interval misconceptions

1) Confidence intervals are a statement about a *population parameter*, not the sample data
2) Any individual confidence interval either succeeds or fails. Consequently, the confidence level describes the *reliability* of the *procedure used* to make the interval estimate

Practice

Suppose we use a sample of $n = 30$ randomly chosen adults to calculate a 95% confidence interval for the mean cholesterol level (mg/dl) of all US adults: $203 \pm 2.045 * \frac{20}{\sqrt{30}} = (195.53, 210.47)$. Rate each of the following statements as either *true* or *false* and explain why:

1) We can be 95% confident that the sample mean from another random sample of size $n = 30$ is between 195.53 mg/dl and 210.47 mg/dl
2) It's statistically plausible that the sample mean is anywhere between 195.53 mg/dl and 210.47 mg/dl
3) We estimate that 95% of the population has cholesterol levels between 195.53 mg/dl and 210.47 mg/dl

Practice (solution)

All three statements are false, here's why:

1) The confidence interval is an estimate of a *population parameter*, it says nothing about other samples
2) Confidence intervals describe a *population parameter*, so should say that we're 95% confident that the *population's mean* is between 195.53 mg/dl and 210.47 mg/dl
3) The population parameter is the *population's mean*, so we cannot draw a conclusion about individual cases within the population

When not to use CLT

Central Limit theorem facilitates the use of relatively simple formulas to create confidence intervals from the sample data. However, these formulas will not produce valid intervals unless the following conditions are met:

- When estimating μ (a population's mean), the sample data must be approximately Normal *or* $n \geq 30$
  - Recognize that the t-distribution was created specifically for *small, Normally distributed samples*
- When estimating p (a population's proportion), at least 10 "successes" and at least 10 "failures" must be observed in the sample data (ie: $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$)

Summary

The general form of any interval estimate is given by:

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

- When using $\hat{p}$ to estimate p, CLT suggests $SE = \sqrt{\hat{p}(1-\hat{p})/n}$
  - c should be chosen from the *Normal distribution* according to the desired confidence level
- When using $\bar{x}$ to estimate μ, CLT suggests $SE = s/\sqrt{n}$
  - c should be chosen from the *t-distribution* according to the desired confidence level

We should also be careful to check that our application satisfies the conditions necessary to use the CLT