STAT 215
Analytic Approximations II

Colin Reimer Dawson

Oberlin College

October 5, 2017

Intervals and Tests for Means

CIs/Tests for Differences

Goals

Confidence Intervals

If we can approximate a bootstrap distribution with a theoretical model (e.g., Normal), we can construct a confidence interval.

P-values

If we can approximate a randomization distribution with a theoretical model (e.g., Normal), we can compute P-values.

The Missing Piece

We need to know what standard error to use as the standard deviation parameter.

Confidence Intervals Using a Normal Model

CI Summary

To compute a confidence interval when the bootstrap distribution is Normal, use

$$\text{Endpoint} = \text{Observed Statistic} + Z^* \cdot \text{Bootstrap SE}$$

where Z^* is the Z-score of the endpoint appropriate for the confidence level, computed from a standard normal ($\mathcal{N}(0, 1)$).

P-values Using a Normal Model

P-values from a Standard Normal

Computing P-values when the randomization distribution is Normal is the reverse process:

1. Convert the observed statistic to a z-score within the randomization distribution (i.e., using its mean and standard deviation).

$$Z_{observed} = \frac{\text{observed statistic} - \text{null parameter}}{\text{randomization SD}}$$

2. Find the relevant area beyond $Z_{observed}$ using a Standard Normal

Cases to Address

We will need standard errors to do CIs and tests for the following parameters:

1. Single Proportion
2. Single Mean
3. Difference of Proportions
4. Difference of Means
5. Mean of Differences
6. Correlation

Analytic Approximations of Sampling Distributions

Param.	Stat.	Randomization	Theory SE	Test Dist.
p	$\hat{p}$	Simulate from p_0	$\sqrt{\frac{p_0(1-p_0)}{n}}$	Normal
μ	$\bar{x}$	Bootstrap + shift	$\frac{s}{\sqrt{n}}$	t_{n-1}
$p_A - p_B$	$\hat{p}_A - \hat{p}_B$	Scramble groups	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_A} + \frac{\hat{p}(1-\hat{p})}{n_B}}$	Normal
$\mu_A - \mu_B$	$\bar{x}_A - \bar{x}_B$	Scramble groups	$\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$	$t_{\min(n_A-1, n_B-1)}$
μ_D	$\bar{x}_D$	Flip pairs*	$\frac{s_D}{\sqrt{n_D}}$	t_{n_D-1}
ρ	r	Scramble pairings	$\sqrt{\frac{1-r^2}{n-2}}$	t_{n-2}

$$CI: \quad \text{Statistic} \pm \text{Critical Value} \times \widehat{SE}$$

$$\text{Sandardized Test Statistic}: \quad \frac{\text{Statistic} - \text{Null Param.}}{\widehat{SE}}$$

Outline

Intervals and Tests for Means

CIs/Tests for Differences

Distribution of Sample Means

- Central Limit Theorem: Sampling Distribution of $\bar{x}$ is approximately Normal, for “sufficiently large” samples, or when the population distribution is Normal.
- As the sample size n goes up, the standard error goes ______________.
- What effect do you expect the *population standard deviation* to have on the standard error of the distribution of sample means? Why?

Distribution of $\bar{x}$

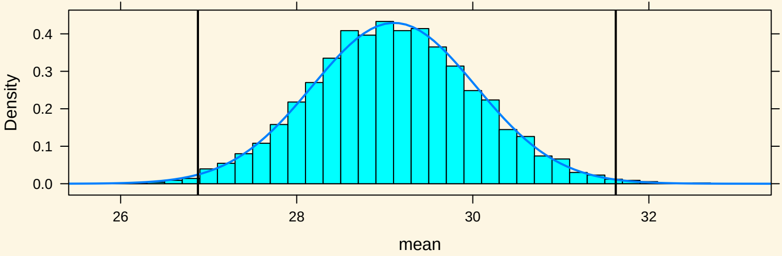
When the population mean is μ, the population standard deviation is σ, and the samples are of size n, the sampling distribution of $\bar{x}$ has mean μ and standard deviation (standard error)

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

It is also approximately Normal, when samples are large enough, OR if the population distribution is approximately Normal. The farther from Normal, the bigger the sample needs to be, but can roughly use $n \geq 27$.

Example: Mean Atlanta Commute Time, Bootstrap CI

```
require(Lock5Data); data(CommuteAtlanta)
bootstrap.means <- do(10000) * mean(~Time, data = resample(CommuteAtlanta))
CI.99.boot <- quantile(~mean, data = bootstrap.means, prob = c(0.005, 0.995))
CI.99.boot

##     0.5%    99.5%
## 26.87798 31.62404
```



Example: Mean Atlanta Commute Time CI, Normal w/ Theoretical SE

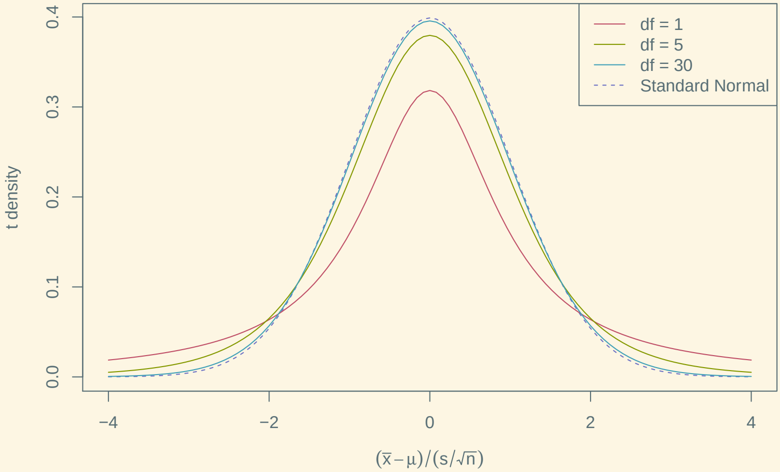
```
n <- nrow(CommuteAtlanta); xbar <- mean(~Time, data = CommuteAtlanta)
zstar.99 <- xqnorm(c(0.005, 0.995))
se.theory <- sigma / sqrt(n)
CI.99.normal.theory.se <- xbar + zstar.99 * se.theory
```

Wait, where do we get σ?

Using s instead of σ

- We only have s, the *sample* standard deviation; not σ, the *population* standard deviation.
- We can approximate SE with $\frac{s}{\sqrt{n}}$, but need to account for the fact that s itself is an estimate (and differs between samples).
- "95% of sample means are within 2SE of μ" no longer accurate: the percentage is less than this.
- How much less depends on how good an estimate s is of σ (i.e., depends on n).

A family of t distributions



CI Summary: Single Mean

To compute a confidence interval for a mean when the sampling distribution for $\bar{x}$ is approximately Normal (i.e., Normal population, or "large" n) and σ is unknown (which is almost always), use

$$\bar{x} \pm t^*_{n-1} \cdot \frac{s}{\sqrt{n}}$$

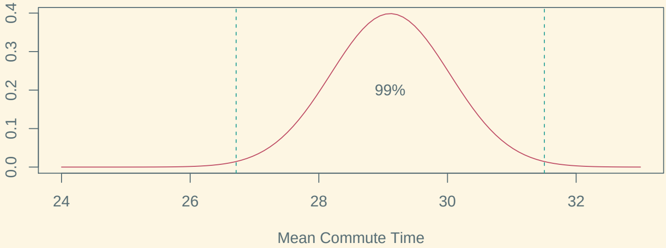
where t^*_{n-1} is the quantile appropriate for the confidence level, computed from a t-distribution with $n-1$ degrees of freedom.

Example: Atlanta Commute Time, t w/ Theoretical SE

```
n <- nrow(CommuteAtlanta); n

     [1] 500

xbar <- mean(~Time, data = CommuteAtlanta); xbar

     [1] 29.11

s <- sd(~Time, data = CommuteAtlanta); s

     [1] 20.71831

se.theory <- s / sqrt(n); se.theory

     [1] 0.926551

#df = 'degrees of freedom'
tstar.99 <- qt(c(0.005, 0.995), df = n - 1); tstar.99

     [1] -2.585718  2.585718
```

Example: Atlanta Commute Time, t w/ Theoretical SE

```
CI.99.t.theory.se <- xbar + tstar.99 * se.theory
```



P-values for a sample mean

Computing P-values when the null sampling distribution is approximately Normal (i.e., Population is normal OR sample size is "large") and σ is unknown (which is almost always) is the reverse process:

1. Convert $\bar{x}$ to a t-statistic within the theoretical distribution .

$$T_{observed} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

2. Find the relevant area beyond $T_{observed}$ using a t distribution with $n - 1$ degrees of freedom

Example: Mean Body Temperature Hypothesis Test

```
require(Lock5Data); data(BodyTemp50)
n <- nrow(BodyTemp50); xbar <- mean(~BodyTemp, data = BodyTemp50)
s <- sd(~BodyTemp, data = BodyTemp50)
se.theory <- s / sqrt(n); mu0 <- 98.6
t.observed <- (xbar - mu0) / se.theory
P.value.theory.se <- 2 * xpt(t.observed, df = n-1)
P.value.theory.se

## [1] 0.002850509
```

Analytic Approximations of Sampling Distributions

Param.	Stat.	Randomization	Theory SE	Test Dist.
p	$\hat{p}$	Simulate from p_0	$\sqrt{\frac{p_0(1-p_0)}{n}}$	Normal
μ	$\bar{x}$	Bootstrap + shift	$\frac{s}{\sqrt{n}}$	t_{n-1}
$p_A - p_B$	$\hat{p}_A - \hat{p}_B$	Scramble groups	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_A} + \frac{\hat{p}(1-\hat{p})}{n_B}}$	Normal
$\mu_A - \mu_B$	$\bar{x}_A - \bar{x}_B$	Scramble groups	$\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$	$t_{\min(n_A-1, n_B-1)}$
μ_D	$\bar{x}_D$	Flip pairs*	$\frac{s_D}{\sqrt{n_D}}$	t_{n_D-1}
ρ	r	Scramble pairings	$\sqrt{\frac{1-r^2}{n-2}}$	t_{n-2}

$$CI: \quad \text{Statistic} \pm \text{Critical Value} \times \widehat{SE}$$

$$\text{Sandardized Test Statistic}: \quad \frac{\text{Statistic} - \text{Null Param.}}{\widehat{SE}}$$

Outline

Intervals and Tests for Means

CIs/Tests for Differences

Variance and Standard Error of Differences

- If we have two *independent* samples, A and B, then quantities such as $\hat{p}_A - \hat{p}_B$ that depend on both random samples have two *independent* sources of variability.
- The result is that the difference is *more variable* than either sample statistic alone.
- In particular, the *variance* of the difference is the *sum* of the separate variances:

$$s^2_{\hat{p}_A - \hat{p}_B} = s^2_{\hat{p}_A} + s^2_{\hat{p}_B}$$

Variance and Standard Error of Proportions

- Recall: across all random samples, the *standard deviation* of the sample proportions (i.e., the standard error) is

$$s_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

  where p is the population proportion and n is the sample size.
- The *variance* of $\hat{p}$ is the square of this; i.e., the same thing without the square root.

Standard Error of Difference of Proportions

So the *variance* of the difference between two sample proportions is

$$s^2_{\hat{p}_A - \hat{p}_B} = \frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}$$

and the *standard deviation* (i.e., standard error) of the difference is

$$s_{\hat{p}_A - \hat{p}_B} = \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$$

Standard Error of Difference of Means

The exact same reasoning applies to the standard error of a difference between means of two independent samples:

$$s_{\bar{x}_A} = \frac{\sigma_A}{\sqrt{n_A}} \qquad s_{\bar{x}_B} = \frac{\sigma_B}{\sqrt{n_B}}$$

$$s^2_{\bar{x}_A} = \underline{\hspace{4cm}} \qquad s^2_{\bar{x}_B} = \underline{\hspace{4cm}}$$

$$s^2_{\bar{x}_A - \bar{x}_B} = \underline{\hspace{4cm}}$$

$$s_{\bar{x}_A - \bar{x}_B} = \underline{\hspace{4cm}}$$

Analytic Approximations of Sampling Distributions

Param.	Stat.	Randomization	Theory SE	Test Dist.
p	$\hat{p}$	Simulate from p_0	$\sqrt{\frac{p_0(1-p_0)}{n}}$	Normal
μ	$\bar{x}$	Bootstrap + shift	$\frac{s}{\sqrt{n}}$	t_{n-1}
$p_A - p_B$	$\hat{p}_A - \hat{p}_B$	Scramble groups	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_A} + \frac{\hat{p}(1-\hat{p})}{n_B}}$	Normal
$\mu_A - \mu_B$	$\bar{x}_A - \bar{x}_B$	Scramble groups	$\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$	$t_{\min(n_A-1,n_B-1)}$
μ_D	$\bar{x}_D$	Flip pairs*	$\frac{s_D}{\sqrt{n_D}}$	t_{n_D-1}
ρ	r	Scramble pairings	$\sqrt{\frac{1-r^2}{n-2}}$	t_{n-2}

$$CI: \quad \text{Statistic} \pm \text{Critical Value} \times \widehat{SE}$$

$$\text{Sandardized Test Statistic}: \quad \frac{\text{Statistic} - \text{Null Param.}}{\widehat{SE}}$$