Stat 411/511

The One Sample T-test

Oct 2 2015

Charlotte Wickham

stat511.cwick.co.nz

Today

Recap

We don’t know the population SD,

Z -> t

A little bit of R

Hypothesis testing & p-values

from last time

Recap

Population inference is using a sample to learn about a population.

This process relies on knowing how the sampling distribution of our statistic relates to the population distribution and our parameters of interest.

If we are interested in the population mean, assuming the sampling distribution of the sample average is Normal, leads us to a 95% confidence interval for the mean of the population,

$$\overline{X} \pm 2\frac{\sigma}{\sqrt{n}}$$

know σ, one sample Z-based CI

Standard deviation of the mean

sample mean
sample average $\overline{X}$

The standard deviation of the sampling distribution of the sample average for a sample of size n, is the population standard deviation divided by the square root of the sample size.

$$\mathrm{SD}_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

This tells us how much the sample average varies from it's mean across different possible samples.

But we usually don't know σ

Often we estimate the population standard deviation with the sample standard deviation.

I.e. we estimate σ with s.

Standard error of the mean

sample mean
sample average $\overline{X}$

If we plug in s for σ in the standard deviation of the sampling average, we called it the standard error.

The **standard error** of the sample average is an **estimate** of the **standard deviation of the sampling distribution** of the sample average.

$$\mathrm{SE}_{\overline{X}} = \frac{s}{\sqrt{n}}$$

It's an **estimate** of how much the sample average varies from it's mean across different samples.

If the **population is Normal**,

then **it is a fact** that the sampling distribution of the **t-ratio**, $\text{t-ratio} = \dfrac{\overline{X} - \mu}{\text{SE}_{\overline{X}}}$ is

a *t*-distribution with **n-1** degrees of freedom

We often write this as:

$$\text{t-ratio} \sim t_{n-1}$$

does this help us?

let's repeat our reasoning from last time,
but first some t-distribution facts...

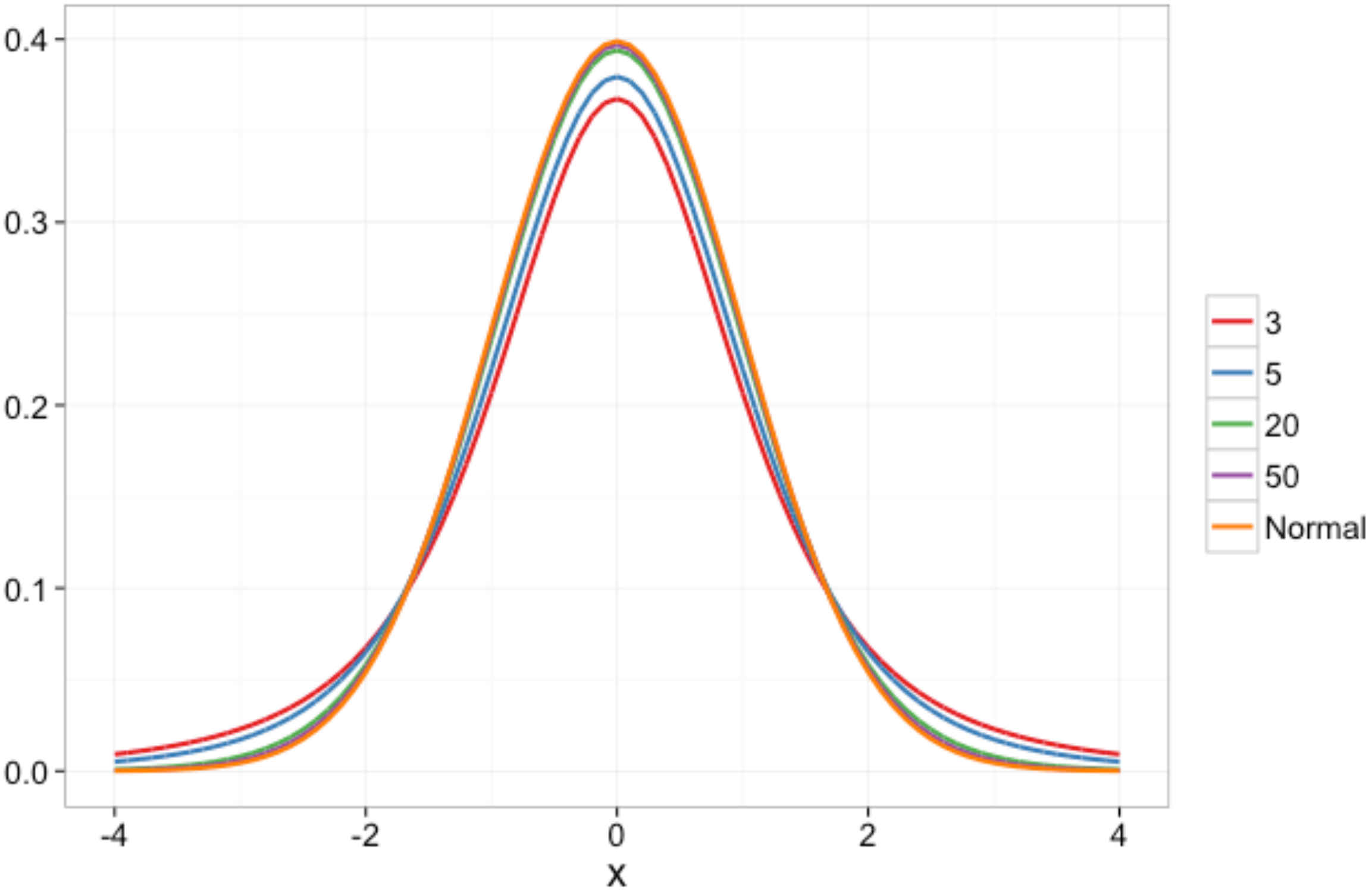
the *t*-distribution

depends on a single parameter, the degrees of freedom,

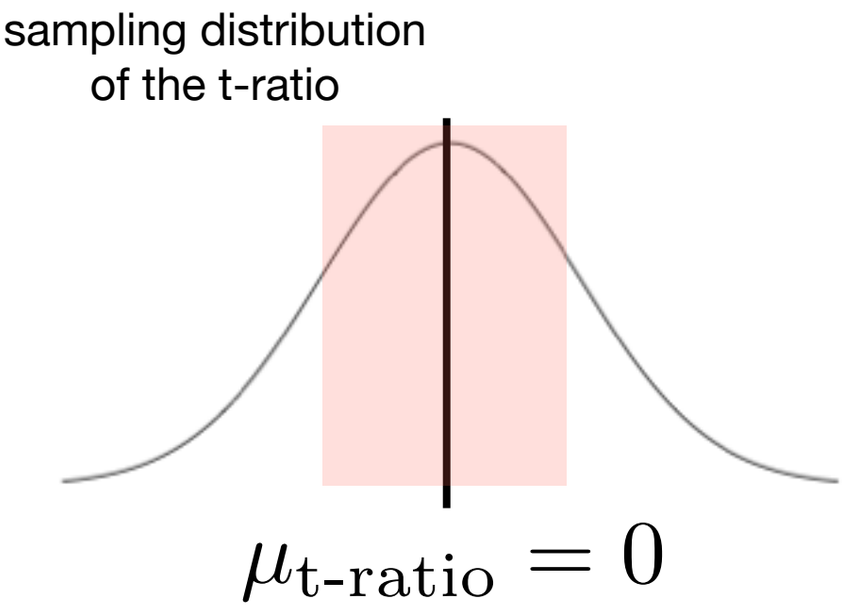
is symmetric,

has mean zero,

gets closer and closer to a Normal distribution N(0, 1) distribution as the d.f. gets larger.

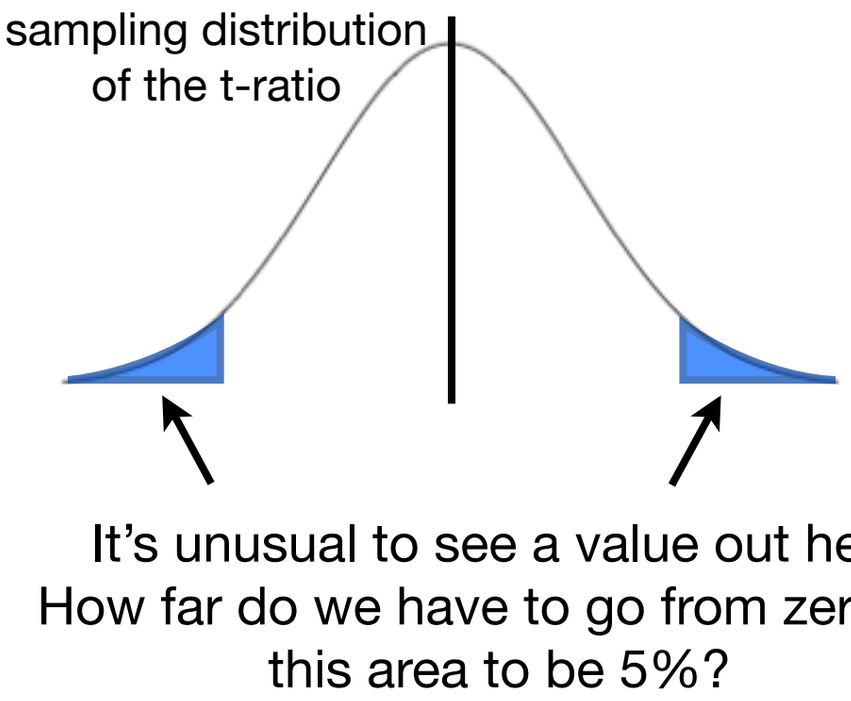

0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
x
3
5
20
50
Normal

let's repeat our reasoning...



We know it would be unusual to see a value for the t-ratio far from zero.

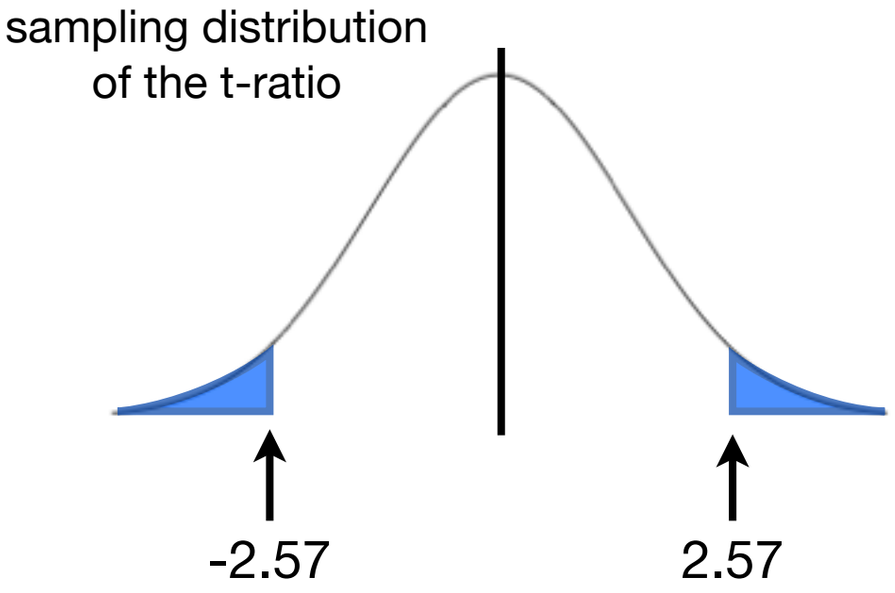
For a Normal distribution we knew about 5% of the time a value will be 2 standard deviations from the mean.



How far from the mean are the most extreme 5% of t-ratios?

look it up in table!

we won't do this now, let's call this number $t_{n-1}(0.975)$



If n = 6, so d.f. = 5, $t_5(0.975) = 2.57$

In 95% of samples, our sample t-ratio would fall **between** -2.57 and 2.57.

About 95% of the time,

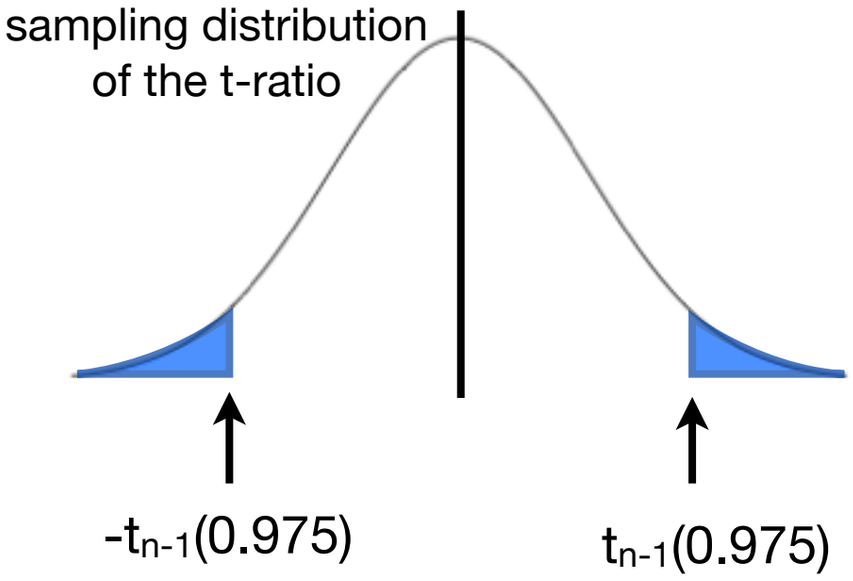
$$-2.57 < \frac{\overline{X} - \mu}{\mathrm{SE}_{\overline{X}}} < 2.57$$

Rearrange: in about 95% of possible samples,

$$\overline{X} - 2.57 \times \mathrm{SE}_{\overline{X}} < \mu < \overline{X} + 2.57 \times \mathrm{SE}_{\overline{X}}$$

think: the sample average is within 2.57 standard errors of the population mean.

A t-based confidence interval



In general,

In 95% of samples, our sample t-ratio would fall **between** $-t_{n-1}(0.975)$ and $t_{n-1}(0.975)$.

About 95% of the time,

$$-t_{n-1}(0.975) < \frac{\overline{X} - \mu}{\mathrm{SE}_{\overline{X}}} < t_{n-1}(0.975)$$

Rearrange: in about 95% of possible samples,

$$\overline{X} - t_{n-1}(0.975) \times \mathrm{SE}_{\overline{X}} < \mu < \overline{X} + t_{n-1}(0.975) \times \mathrm{SE}_{\overline{X}}$$

think: the sample average is within $t_{n-1}(0.975)$ standard errors of the population mean.

With 95% confidence the population mean is between

$\overline{X} - t_{n-1}(0.975) \times \mathrm{SE}_{\overline{X}}$ and $\overline{X} + t_{n-1}(0.975) \times \mathrm{SE}_{\overline{X}}$

A 95% **one-sample *t*** confidence interval

Recap #3:

Population inference is using a sample to learn about a population.

This process relies on knowing how the sampling distribution of our statistic relates to the population distribution and our parameters of interest.

If we are interested in the population mean, assuming the sampling distribution of the sample average is Normal, leads us to a 95% confidence interval for the mean of the population,

$$\overline{X} \pm t_{n-1}(0.975)\frac{s}{\sqrt{n}}$$

don't know σ,
one sample t-based CI

Can you check my confidence interval from slide 7 on Weds?

In R

The 0.975 quantile of a t-distribution with 5 degrees of freedom

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

the function is called qt

the first argument is called p, we are giving it the value 0.975

the second argument is called df, we are giving it the value 5

```
?qt
```

get help on the function

All the work in R is done by functions.
Functions take arguments and return some output.

To find out what arguments a function takes, ?, look at the help.

Anatomy of a function call

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

after the function name comes `(`

then the first argument, `name = value`

arguments are separated by `,`

if we've finished listing arguments we finish with a `)`

You've already seen some other functions in lab/homework 0

```
library(ggplot2)
hist(x)
qplot(x)
rnorm(100)
```

How many arguments do they have?
What are the names of the arguments?

Your turn

What's the function name?

How many arguments have I given it?

What are the arguments called?

```
mean(x = 1:10)
c(0, 1, 2, NA)
sd(x = c(0,1,2,NA), na.rm = TRUE)
data(package = "Sleuth3")
data.frame(x = 1:10)
```

My fake Example

Do textbooks cost more at the bookstore than on Amazon?

What is the mean difference in price between Amazon and the bookstore, for **all textbooks used at OSU**?

The average price difference in our sample of size, n=100, is $10 with a sample standard deviation of $5.

$$\overline{X} = 10, \quad s = 5, \quad n = 100$$

$$\text{SE}_{\overline{X}} = s/\sqrt{n} = 5/\sqrt{100} = 0.5$$

$$t_{n-1}(0.975) = t_{99}(0.975) = 1.98$$ `qt(0.975, 99)`

$$10 \pm 1.98 \times 0.5 = (9, 11)$$

With 95% confidence μ is between $9 and $11. no context

With 95% confidence we estimate that OSU textbooks on Amazon cost on average between $9 and $11 less than at the bookstore. in context

Hypothesis tests

Is this mean different from zero?

There's a correspondence between tests and confidence intervals.

If a **95%** confidence interval does not contain the hypothesized value, then the hypothesis would be rejected at the **5%** significance level.

substitute X% and 100 - X%, 95% and 5%

Hypothesis: the mean is zero, $\mu = 0$

Our 95% confidence interval: (9, 11)

Zero is not in the interval, we reject this hypothesis at the 5% significance level.

We conclude it is unlikely the mean is zero.

Quantified by a p-value, to get a p-value we need to know a little more about statistical testing.