Chapter 17: Inference of the mean when the sd is unknown

Corinne Riddell

April 3, 2020

Recap

- For the last few lectures we have assumed that the population standard deviation (σ) was known.
- We conducted the z-test and created CIs using this known σ
- Today, we generalize this framework to a setting where σ is unknown and needs to be estimated by s, the sample standard deviation

Reduced conditions for inference about a mean

- Most important: Data is a simple random sample (SRS) from a much larger population
- Some flexibility: Observations follow a Normal distribution

Estimating the standard error based on the sample

- Previously, we were told the standard deviation for the population σ and could use this to calculate the standard error of the mean as $\sigma/\sqrt{n}$
- Now, we don't know σ. So we estimate the standard error by

$$s/\sqrt{n}$$

where s is the sample standard deviation.

s **vs.** $s/\sqrt{n}$

- Remember, s is our estimate for the population standard deviation. It estimates the variation between *individuals*.
- In contrast, $s/\sqrt{n}$ is the estimate for the standard error of the mean, $\bar{x}$. $s/\sqrt{n}$ estimates the variation between *means*.

Recall the z-test!

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

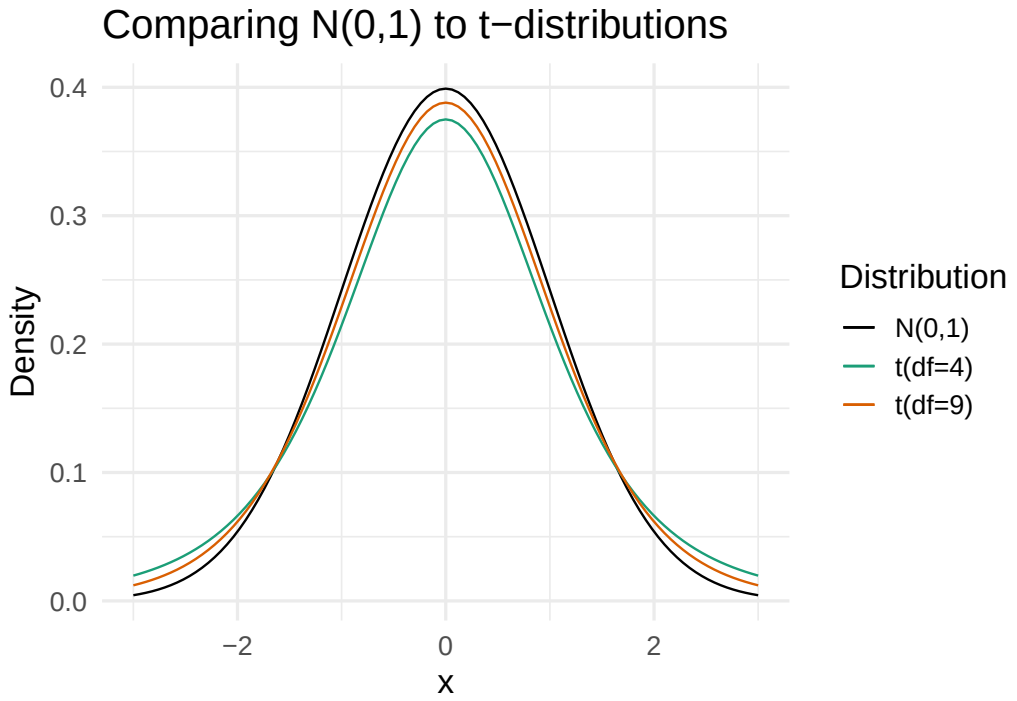
Meet the t-test!

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- What is the difference between z and t?
- The t-test statistic is more variable than the z-test statistic because we have to estimate σ using s. Because s is a statistic, it varies across samples (whereas σ is a population parameter, which implies it does not vary).
- This substitution makes the t-test not follow a Normal(0, 1) distribution. Its distribution is *more* variable than the standard Normal. Thus, we need a distribution that is like the standard Normal but a little bit wider.

Introducing the t distribution

- The t-distribution is like the standard Normal distribution, but wider.
- Its width depends on n, the sample size. This is because as n increases, our estimate s gets better and better, and approaches σ. Thus, as n increases the t-distribution approaches a Normal(0, 1) distribution.
- In the legend for this plot $df = n - 1$. We will learn more about `df` on the next slide.



Meet the t-test!

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The one-sample t statistic has a t distribution with $n - 1$ **degrees of freedom**

What are degrees of freedom? For this test, the degrees of freedom is equal to $n - 1$. The higher the degrees of freedom, the closer the shape of the t-distribution is to the Normal distribution.

Meet the t-test!

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Steps to conduct a t-test:

1. Determine whether the assumptions to conduct the t-test are met. Assumption i: Assumption ii:
2. Calculate the t-test statistic using $\bar{x}$ and s (estimated from your sample), n which is also a property of your sample, and μ_0 from the null hypothesis.
3. Compute the probability of observing this test statistic t or more extreme under the null hypothesis. This is the p-value.
4. Interpret the p-value. Is the probability very small (and shows evidence against the null distribution in favor of the alternative)? Sometimes, you will be ask to compare the p-value to a pre-defined significance level, α. Often $\alpha = 0.05$

Guess the R functions

```
pt(q = , df = , lower.tail = )
qt(p = , df = , lower.tail = )
```

Which one would we use to calculate the p-value for a hypothesis test after we calculated the t-test statistic? `pt` or `qt`?

Suppose you calculated t = -2 and you know that the sample size was 100. Write the code to calculate the p-value for a two-sided test:

```
#to fill out in class
```

Calculating a confidence interval for the t-test

Draw an SRS of size n from a large population having unknown mean μ and unknown standard deviation σ. A level C **confidence interval for** μ is:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value for the $t(n-1)$ density curve with area C between $-t*$ and $t*$.

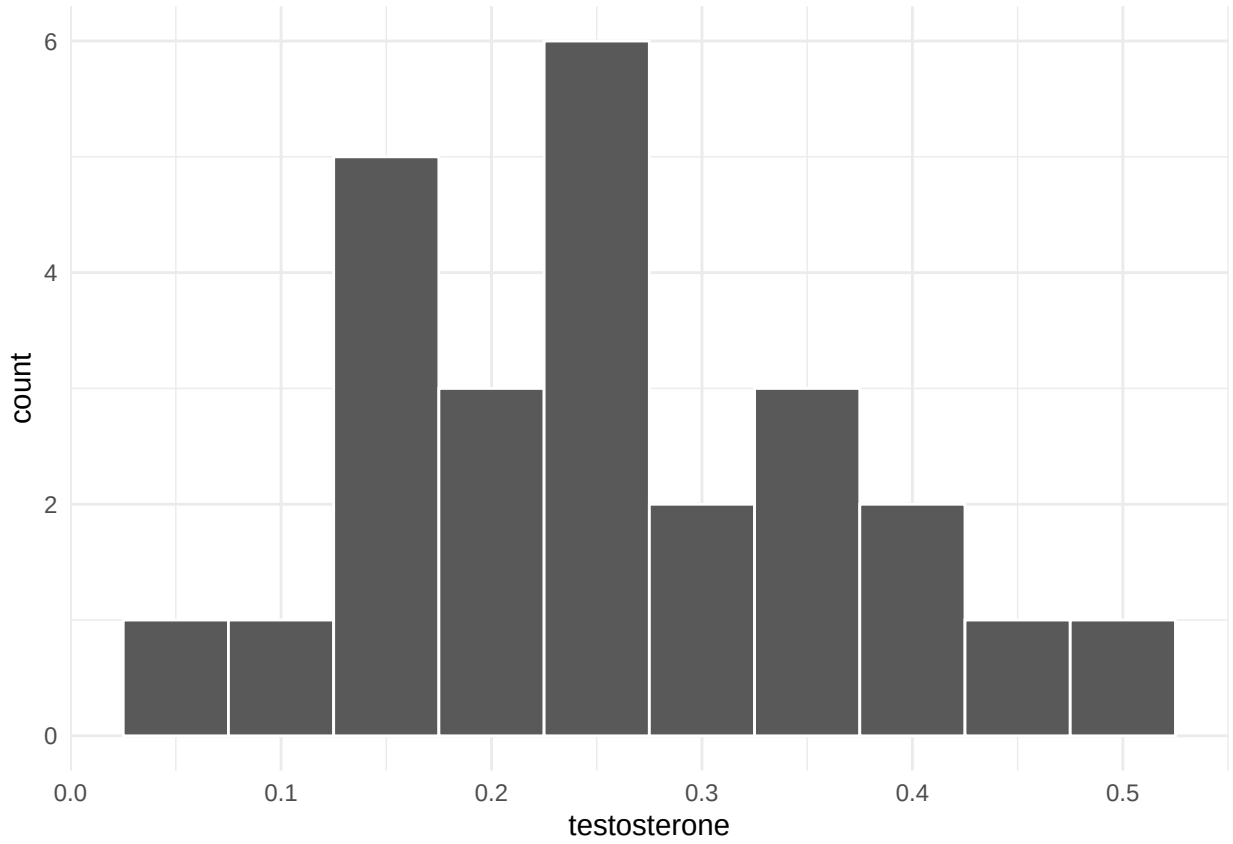
Supposing we had $n = 100$, what is t^* for a 95% confidence interval?

Example: Testosterone and obesity in adolescent males (pg 422 B&M Ed 4)

Here are the data for $n = 25$ adolescent males between the ages of 14 and 20:

```
library(tidyverse)
testosterone <- c(0.30, 0.24, 0.19, 0.17, 0.18, 0.23, 0.24, 0.06, 0.15,
                  0.17, 0.18, 0.17, 0.15, 0.12, 0.25, 0.25, 0.25, 0.32,
                  0.35, 0.37, 0.39, 0.46, 0.49, 0.42, 0.36)
dat_test <- data.frame(testosterone)

ggplot(dat_test, aes(x = testosterone)) +
  geom_histogram(binwidth = 0.05, col = "white") +
  theme_minimal()
```



Example: Testosterone and obesity in adolescent males (pg 422 B&M Ed 4)

Use R to calculate a 95% confidence interval for testosterone. We can do this using `summarize`

```
dat_test %>% summarize(sample_mean = mean(testosterone), #sample mean
                       sample_sd = sd(testosterone), #sample standard dev
                       sample_size = length(testosterone), #sample size n
                       sample_se = sample_sd/sqrt(sample_size)) #standard error of mean
```

```
##   sample_mean sample_sd sample_size  sample_se
## 1      0.2584 0.1115303          25 0.02230605
```

We still need the t^* value:

```
t_star <- qt(p = 0.975, df = 24)
t_star
```

```
## [1] 2.063899
```

Example: Testosterone and obesity in adolescent males (pg 422 B&M Ed 4)

Expand the previous code chunk to calculate the margin of error (which uses the critical t^* value), and then calculate the lower and upper CI

```
dat_test %>% summarize(sample_mean = mean(testosterone),
                       sample_sd = sd(testosterone),
                       sample_size = length(testosterone),
                       sample_se = sample_sd/sqrt(sample_size),
                       margin_of_error = sample_se*t_star,
```

```
                        lower_CI = sample_mean - margin_of_error,
                        upper_CI = sample_mean + margin_of_error)
```

```
##   sample_mean sample_sd sample_size  sample_se margin_of_error  lower_CI
## 1      0.2584 0.1115303          25 0.02230605      0.04603743 0.2123626
##    upper_CI
## 1 0.3044374
```

Interpret: The sample mean $\bar{x}$ is 0.26 and its 95% confidence interval is 0.21 to 0.30. Using this method, 95% of the confidence intervals we make will contain the true population mean μ.

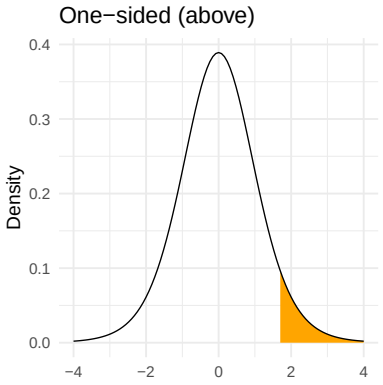
The t-test

Draw an SRS of size n from a large population having unknown mean μ and unknown standard deviation σ. To test the hypothesis $H_0 : \mu = \mu_0$, calculate the t statistic:

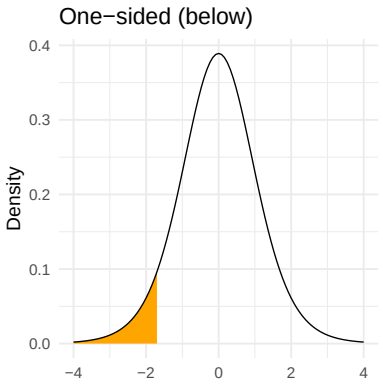
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

t comes from the t-distribution with $n - 1$ degrees of freedom. For the t we calculate from our sample, the next step is to calculate the probability that we would see this t (or a more extreme value) under the null distribution.

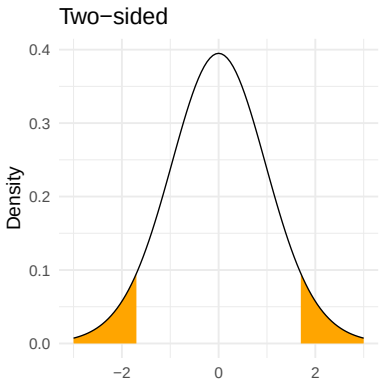
H_a: $\mu > \mu_0$ is $P(T \geq t)$, using code: `pt(q = t, df = n-1, lower.tail = F)`



H_a: $\mu < \mu_0$ is $P(T \leq t)$, using code: `pt(q = t, df = n-1)`



H_a: $\mu \neq \mu_0$ is $2 \times P(T \geq |t|)$, using code: `pt(q = t, df = n-1)*2` if your t is negative, or `pt(q = t, df = n-1, lower.tail = F) * 2` if your t is positive.



Example of a t-test (pg 426 B&M Ed 4)

Here are 18 measures of pulse wave velocity (PWV) from a sample of children diagnosed with progeria, a genetic disorder that produces rapid aging.

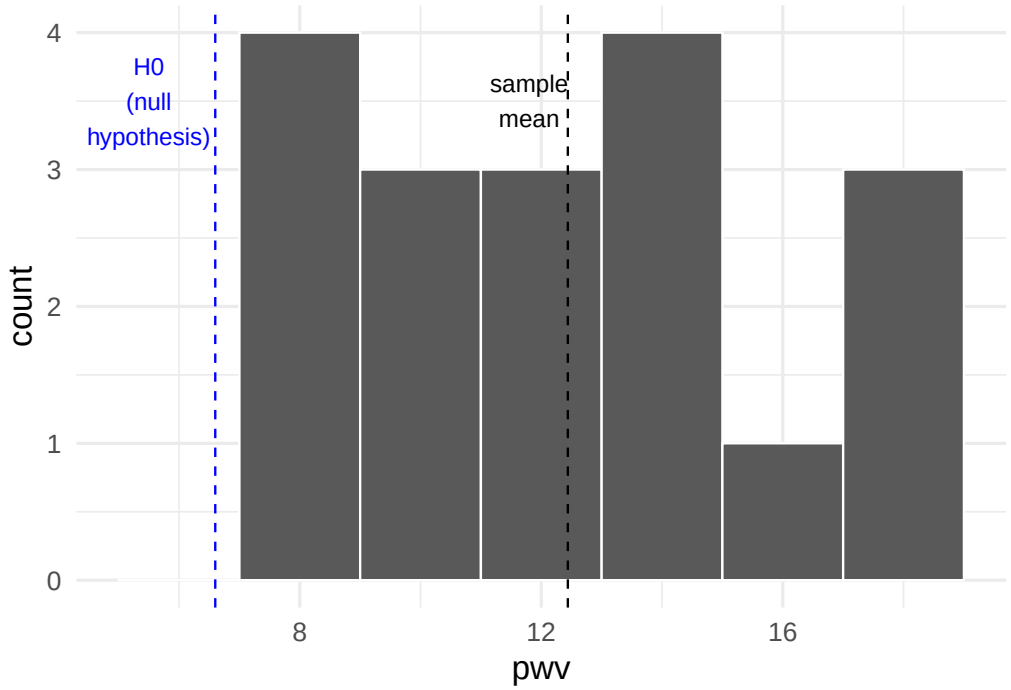
```
pwv <- c(18.8, 17.6, 17.5, 16.0, 14.8, 14.1, 13.7, 13.1, 12.9,
         12.9, 12.4, 10.1, 9.3,  9.1,  8.3,  8.3,  7.9,  7.2)

pwv_dat <- data.frame(pwv)
```

For the general population, pwv measures greater than 6.6 are considered abnormally high. We would like to test the hypothesis that the mean for this group of children is abnormally high.

That is: $H_0 : \mu = 6.6$ and $H_a : \mu > 6.6$

Look at the data and see if there is evidence against the null hypothesis



Calculations using R code

```r
pwv_dat %>%
  summarize(sample_mean = mean(pwv),
            sample_sd = sd(pwv),
            sample_size = length(pwv),
            sample_se = sample_sd/sqrt(sample_size),
            t_test = (sample_mean - 6.6)/sample_se,
            p_value = 1 - pt(t_test, df = sample_size - 1))
```

```
##   sample_mean sample_sd sample_size sample_se   t_test      p_value
## 1    12.44444  3.637747          18 0.8574252 6.816273 1.501248e-06
```

- Know also how to do these calculations by hand. For example, you could be provided with $\bar{x}$ and s for this sample and asked to compute the test statistic
- You cannot compute the p-value by hand, but should know the code required to calculate the p-value and how to interpret it.

There's a function for that...

Rather than doing the test using summarize, we could have R do it for us using `t.test`:

```r
t.test(x = pwv_dat %>% pull(pwv), alternative = "greater", mu = 6.6)
```

```
##
##  One Sample t-test
##
## data:  pwv_dat %>% pull(pwv)
## t = 6.8163, df = 17, p-value = 1.501e-06
## alternative hypothesis: true mean is greater than 6.6
## 95 percent confidence interval:
##  10.95286      Inf
## sample estimates:
## mean of x
##  12.44444
```

Matched pairs t procedures

- skip this section for now. We will come back to this next week.

Robustness of t procedures

- A confidence interval or hypothesis test is called **robust** if the confidence level or p-value does not change very much when the conditions for use of the procedure are violated.
- In particular, how robust are the procedures against non-Normality?
  - The t procedures are quite robust against non-Normality of the population except when outliers or a strong skew are present.
  - The t procedures are not robust against outliers unless the sample size is sufficiently large.

Checking assumptions

- Always plot your data first:
  - Are there any outliers?
  - Is the distribution of the data skewed?

Guidelines for using the t procedures

- The SRS condition is more important that the Normality condition
- If n < 15: Use t procedures if the data appear close to Normal (at least roughly symmetric, single peak, no outliers). If the data are skewed or there are outliers, don't use t.
- Moderate sample size > 15: The t procedures can be used except in the presence of outliers or strong skewness
- Large sample size, roughly $n \geq 40$: The t procedures can be used even for strongly skewed distributions when the sample is large, roughly $n \geq 40$

Example 17.5: Can we use t?

- Good text example. Here you are provided with four datasets and their distributions and sample sizes and are asked whether it is appropriate to use a t-test.
- Pg. 436 of edition 4.

Recap

- We use a z-test when the population sd σ is known.
- We use a t-test when the population sd has to be estimated by s.
- We compare the t-test statistic to the t-distribution with degrees of freedom equal to n-1 to calculate the probability of observing the t-value (or a more extreme value) under the null hypothesis. This is the p-value.
- When n is large, the t distribution is very close to the N(0,1) distribution. This means that we have some intuition about whether the p-value is going to be small or large when the sample large is big.