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

Corinne Riddell

October 23, 2020

### Recap

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

### Reduced conditions for inference about a mean

* Data is a SRS form a much larger population (really important)
* Observations follow a Normal distribution (some leeway)

### Estimating the standard error based on the sample

* Previously, we knew the standard error of the mean to be
* Now, we don’t know , so we estimate the standard error by

where is the sample standard deviation.

### vs.

* Remember, is our estimate for the population standard deviation. It estimates the variation between *individuals*.
* In contrast, is our estimate for the standard error of the mean, . estimates how much sample *means* vary.

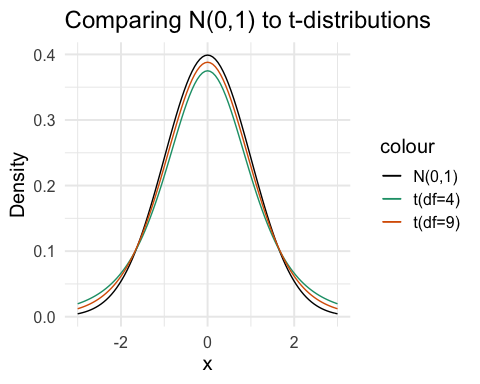
### Recall the z-test!

### Meet the t-test!

* What is the difference between and ?
* The t-test statistic is more variable than the z-test statistic because we have to estimate using . Because is a statistic, it varies across samples (whereas 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 , the sample size. This is because as increases, our estimate gets better and better, and approaches . Thus, as increases the t-distribution approaches a Normal(0, 1) distribution.



### Meet the t-test!

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

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

### Meet the t-test!

Steps to conduct a t-test:

1. Determine whether the assumptions to conduct the t-test are met.
2. Calculate the t-test statistic using and (estimated from your sample), which is also a property of your sample, and from the null hypothesis.
3. Compute the probability of observing this test statistic 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, .

### 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 from a large population having unknown mean and unknown standard deviation . A level C **confidence interval for**  is:

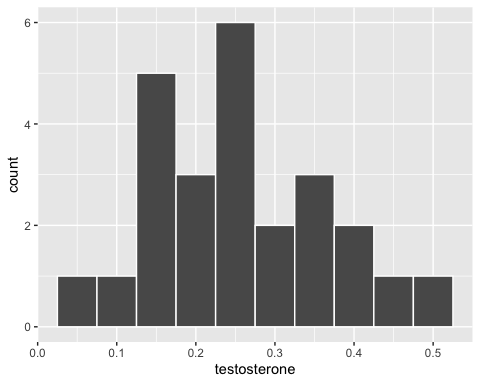
where is the critical value for the density curve with area C between and .

Supposing we had , what is for a 95% confidence interval?

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

Here are the data for 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")



### 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 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 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

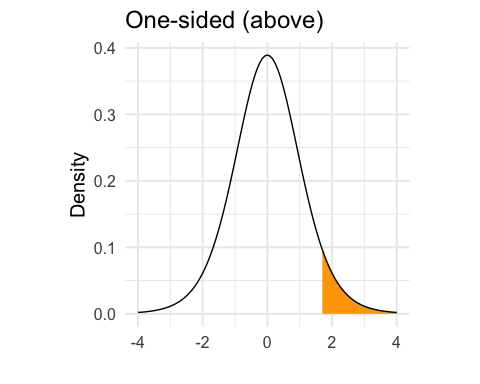
Interpret: The sample mean is 0.26 and its 95% confidence interval is 0.21 to 0.30. Using this method, 95% of the confidence intervals we make will contain the true population mean .

### The t-test

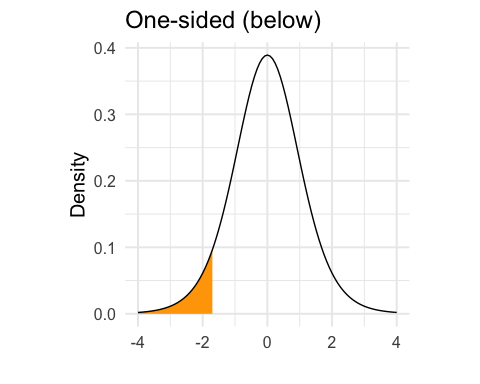
Draw an SRS of size from a large population having unknown mean and unknown standard deviation . To test the hypothesis , calculate the t statistic:

comes from the t-distribution with degrees of freedom. For the 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.

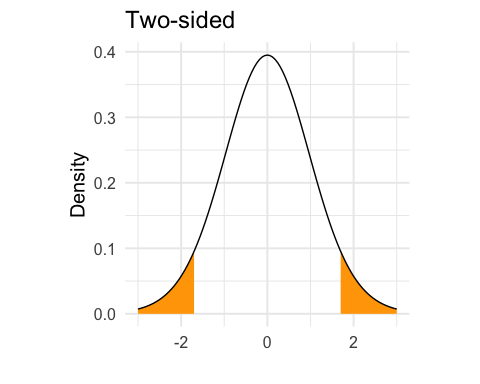
: is , using code: pt(q = t, df = n-1, lower.tail = F)



: is , using code: pt(q = t, df = n-1)



: is , 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)

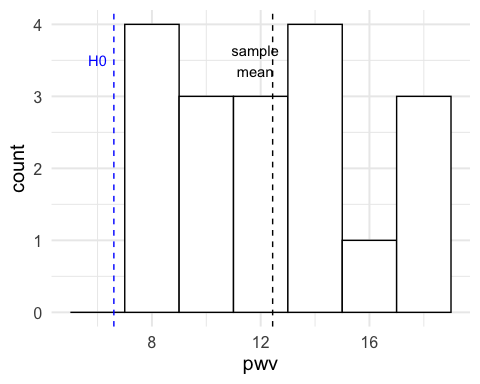
Here are 18 measures of pulse wave velocity (PWV) from a sample of children diagnosed with progeria, a genetic disorder that produces rapid aging.

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 subset of children is abnormally high.

That is: and

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



### Calculations using R code

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 and 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:

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 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 procedures are quite robust against non-Normality of the population except when outliers or strong skewness 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 procedures

* The SRS condition is more important that the Normality condition
* If n < 15: Use 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 .
* Moderate sample size > 15: The procedures can be used except in the presence of outliers or strong skewness
* Large sample size, roughly : The procedures can be used even for strongly skewed distributions when the sample is large, roughly

### Example 17.5: Can we use ?

* 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 is known
* We use a t-test when the population sd has to be estimated by
* We compare the z test statistic to a N(0,1) distribution to calculate the p-value
* We compare the t test statistics t a t distribution with degrees of freedom on n-1
* 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.