Research Methods in English Linguistics

Inferential Statistics 1: *t* statistic

Hyunah Ahn

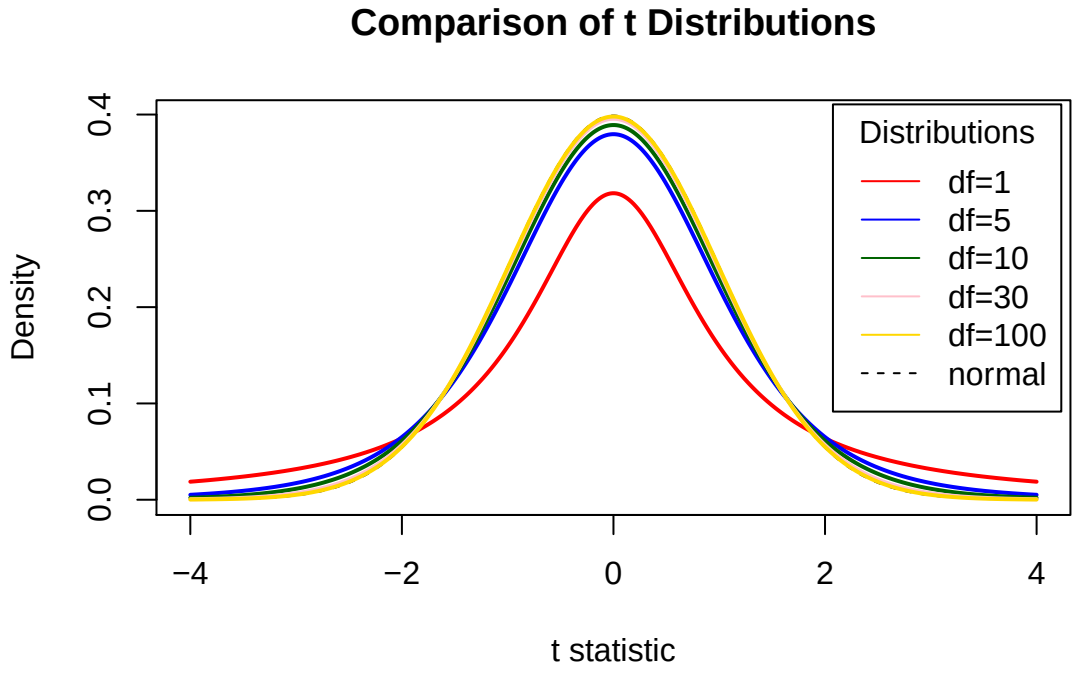
Revised: October 29, 2020

t distribution

In the previous session, we learned about z distribution or normal distribution. What we should remember is that normal distribution is about many natural phenomena in this world. But not everything is normally distributed, and even if everything IS normally distributed, we cannot measure everyone's height in a country (or even in a city) to report the national average height. Luckily, we can use t-statistic to INFER or estimate if the mean we get from a small sample is comparable to the population mean.

```
# Display the Student's t distributions with various degrees of freedom and compare them to the
    normal distribution

x ← seq(−4, 4, length=400) # Create a sequence of numbers from -4 to 4 with the length of 400
hx ← dnorm(x) # Calculate the probability density for each value in x
degf ← c(1, 5, 10, 30, 100) # Create a degree of freedom vector
colors ← c("red", "blue", "darkgreen", "pink","gold", "black") # color vector
labels ← c("df=1", "df=5","df=10", "df=30", "df=100", "normal") # label vector
plot(x, hx, type="l", lty=2, xlab="t statistic",
 ylab="Density", main="Comparison of t Distributions") # draw a normal distribution curve
for (i in 1:5){
 lines(x, dt(x,degf[i]), lwd=2, col=colors[i])
} # draw curves of varying df
legend("topright", inset=.01, title="Distributions",
 labels, lwd=1, lty=c(1, 1, 1, 1, 1, 2), col=colors) # place figure legend
```



As you can see from the figure above, when you have a smaller degree of freedom, the curve has thicker tails. This means that with a small sample size, a larger portion of data are spread out, which, in turn, requires a higher t statistic for it to reach a given percentile compared to z statistic. Remember when 95% of normally distribute data fall between the z-scores of -1.96 and 1.96 (we roughly said it was 2 but now is the time to be more precise)? Let's check that one more time.

```
pnorm(1.96) # shows the percentile ranking of z-score 1.96 in normally distributed data. It means
    that z-score 1.96 is above 97.5% of all data.
```

```
[1] 0.9750021
```

```
1-pnorm(1.96)# Since the percentile ranking ranges from 0 to 1, if we subtract pnorm(1.96) from 1,
    it gives us the amount of data that is higher than 1.96. You can see that 2.5% of data are
    above.
```

```
[1] 0.0249979
```

```
pnorm(-1.96) # If you put in the negative value, you get the proportion of data that are below two
    standard deviations from the mean. And the number is the same as 1-pnorm(1.96).
```

```
[1] 0.0249979
```

```
# In sum, a total of 5% of data are 2 sd away or farther from the mean.

(1-pnorm(1.96))+pnorm(-1.96)
```

```
[1] 0.04999579
```

```
# Since you get the same number anyway, you can simply double the first value.

2*(1-pnorm(1.96))
```

```
[1] 0.04999579
```

```
# Let's see how percentile rankings of t distribution works. Since we saw above how t distribution
    can vary depending on the degree of freedom. You need to set df in the function.

pt(1.96, df=1)
```

```
[1] 0.8498286
```

```
1-pt(1.96, df=1) # This means that about 15% of data are above the t score of 1.96
```

```
[1] 0.1501714
```

```
2*(1-pt(1.96, df=1)) # This means that about 30% of data are 2 sd away or farther from the mean.
    That's too much!
```

```
[1] 0.3003429
```

```
# Since we also learned that data with a larger degree of freedom (= data of a larger sample size)
    can have a distribution quite close to normal distribution, let's up the df value.

2*(1-pt(1.96, df=100)) # It's still slightly more than 5%.
```

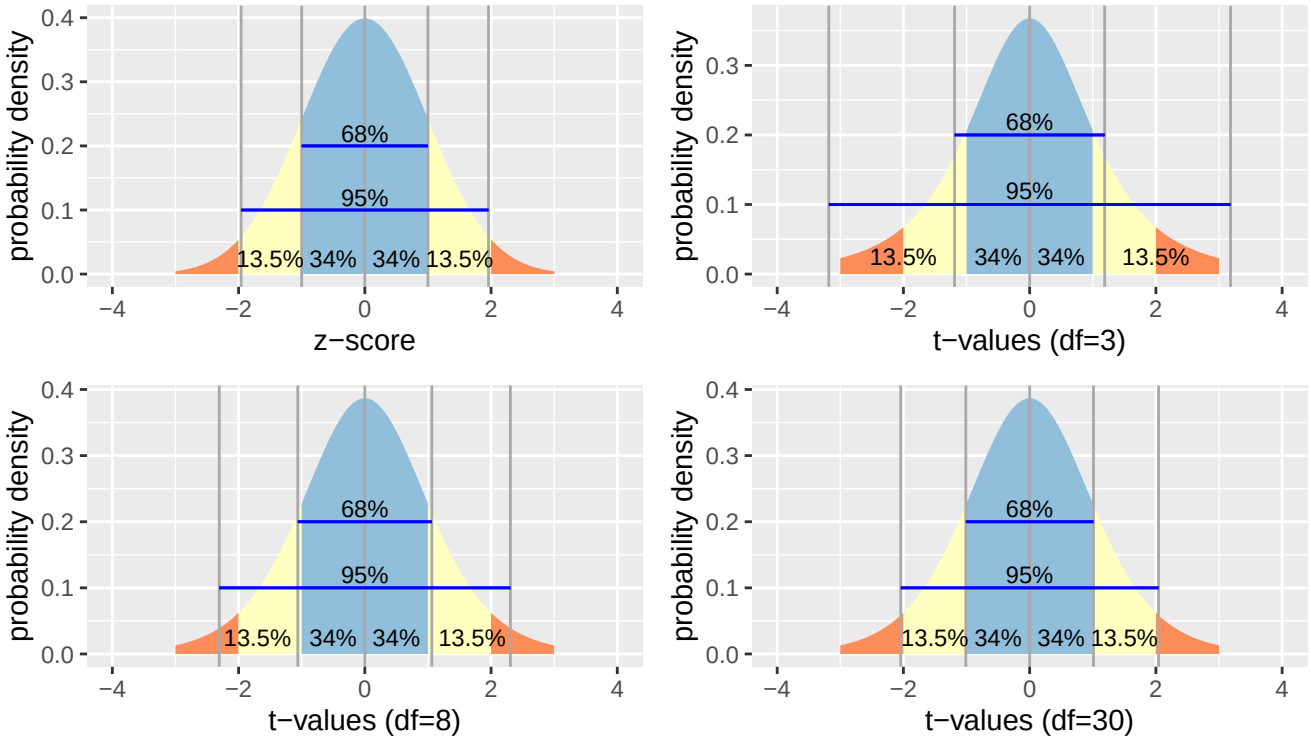
```
[1] 0.0527789
```

```
2*(1-pt(2, df=100)) # Now let's change the t value to 2. Now only 4.8% of data deviate from the
    mean 2 sd away or farther.
```

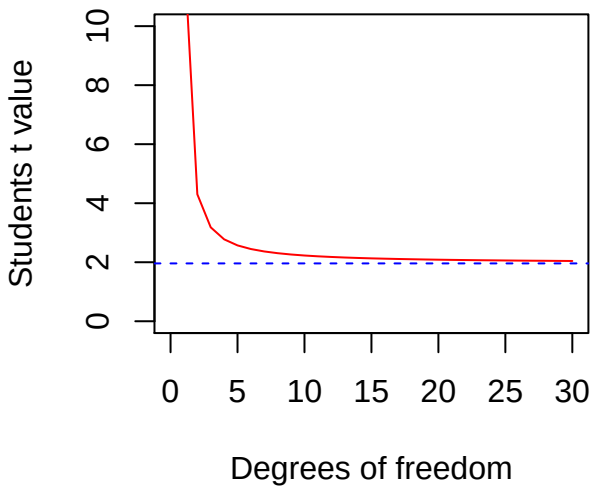
```
[1] 0.04821218
```

As a rule of thumb, you have to at least 2 in t-statistic to reduce the amount of data that deviates 2 sd away or farther.

Comparing z distribution and t distributions



As you can see from the plots above, z-distribution does not change by the number of observations (hence, degree of freedom) while t-distribution changes how many observations (or data points) are available.



As you can see from the figure above, when df is smaller than 5, you have to have a very large t statistic for it to reach 97.5%.

One sample t-test

We often hear that the average height in a certain country is reported to be a certain figure in centimeters or in inches. Let's say that some research company reported that the average height of Korean male high school students is 170cm. Now, you want to know if male students in your class (n=35) have the average height or if the male students in your class are statistically significantly taller or shorter than the national average? Then, how can you INFER that their average height is the same or not the same as the reported national average? You first have to measure the students' height, average them out, and get the t-statistic of the mean value and see if the t value reaches the 95% threshold!

Step 1: Take measurements

We have a theory or a hypothesis that the average height of male high school students in Korea is 170cm. Now, you measure the height of your students in class and find the mean of the data. Let's do it virtually in R.

```
set.seed(170)
myclass←rnorm(35, 172, 5) # Random-sample 35 numbers from a normally distributed data with the
    mean of 173 and sd of 5.

myclass # You can now see the height measurements of your 35 students.
```

```
 [1] 182.7437 175.6684 172.7914 164.3754 172.9696 179.1066 178.4673 172.3283
 [9] 179.5938 171.0750 169.3769 173.0531 174.6683 179.0022 166.1142 173.8747
[17] 175.5703 175.6251 174.3014 174.4740 165.9923 177.1603 163.9914 175.4443
[25] 176.9510 165.6848 185.1992 167.0122 164.2948 176.2005 166.4560 173.7268
[33] 172.5091 173.1423 173.5129
```

Step 2: Calculate the t statistic

Now you have a small data set, let's calculate the t-statistic. You first need its mean and sd. Then, t-statistic is calculated by dividing the difference between the sample mean ($\bar{x}$) and the true mean (μ=170cm) ($\bar{x} - \mu$) by the standard error or the standard deviation divided by the square root of the sample size ($\frac{\sigma}{\sqrt{n}}$).

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

```
myclass.mean←mean(myclass)

myclass.sd← sd(myclass)

myclass.size←length(myclass)

tvalue ←(myclass.mean−170)/(myclass.sd/sqrt(myclass.size))

tvalue
```

```
[1] 3.606674
```

Step 3: Find its p value

Now you have the t statistic, find its percentile ranking from the t distribution table. As we did before with the z distribution table, you can do it on a piece of paper or you can use the r function `pt()`.

```
pt(tvalue, df=34)
```

```
[1] 0.9995082
```

```
2*(1−pt(tvalue, df=34)) # So, what does this mean?
```

```
[1] 0.0009835857
```

```
percentage←100*2*(1−pt(tvalue, df=34))

percentage
```

```
[1] 0.09835857
```

The 'percentage' value you got above should be interpreted as the probability of obtaining the sample mean value you got when your sample mean and the true mean are the same. We call this the null hypothesis (H_0), the hypothesis that assumes that the sample mean and the true mean are equal. You can also hypothesize the other case in which the sample mean and the true mean are not equal and we call that the alternative hypothesis (H_1)

$H_0 : \bar{x} = \mu$

$H_1 : \bar{x} \neq \mu$

But should we calculate the t statistic in this complicated manner every time? No worries! R has functions written for it already!

```
t.test(myclass, mu=170)
```

```
        One Sample t-test

data:  myclass
t = 3.6067, df = 34, p-value = 0.0009836
alternative hypothesis: true mean is not equal to 170
95 percent confidence interval:
 171.4026 175.0235
sample estimates:
mean of x
 173.2131
```

The output shows the t statistic, degree of freedom, and the p-value on the first line. The alternative hypothesis (that the sample mean and the true mean are not equal) on the second line, 95% confidence intervals (both below and above), and the sample mean of the data. The p-value is the probability of obtaining the sample mean when the null hypothesis is true. When the p-value is smaller than .05, it is interpreted that the sample mean is statistically significantly different from the population mean.

As explained earlier, t statistic is a quantile value that indicates the distance between two means (in this case, the sample mean and the population mean (or hypothesized mean)). Degree of freedom is the number of values that is free to vary (n-1), and the p-value indicates the probability that the obtained difference will be observed when the null hypothesis is true (when the two means are equal). A confidence interval is a range of values that we are quite sure that a true mean (or true mean difference) can be found with a given alpha level (in this case 95%). A confidence interval in the one sample t-test is calculated as below.

You first need to find out the t-statistic at 2.5% and 97.5%. 2.5% at each end will make up for 5%. In the 95% confidence interval, you exclude the extreme 5%.

```
qt(0.025, df=34) # t value at the lowest 2.5%
```

```
[1] -2.032245
```

```
qt(0.975, df=34) # t value at the highest 2.5%
```

```
[1] 2.032245
```

lower CI = mean - standard error * t statistic at extreme 2.5%
upper CI = mean + standard error * t statistic at extreme 2.5%

```
# lower CI
mean(myclass)-(sd(myclass)/sqrt(length(myclass))*2.032245)
```

```
[1] 171.4026
```

```
# upper CI

mean(myclass)+(sd(myclass)/sqrt(length(myclass))*2.032245)
```

```
[1] 175.0235
```

Remember! You got a negative value for the lowest 2.5%, but instead of using the negative value for the calculation of the lower CI, its absolute value was used. If you are going to use the negative value as is, the formula for the lower CI should be mean + standard error * t statistic at lowest extreme 2.5%

Two sample t-test

So, that was a useful discussion! But we can do more fun stuff with t-test. Instead of comparing a sample mean to a known population mean, you can compare two different sample means. So, we had one class above and now your colleague wants to measure the height of his/her students and see if the two classes have significantly different average heights.

Step 1: Take measurements

```
set.seed(170)
theirclass<-rnorm(37, 176, 10)

theirclass
```

```
 [1] 197.4874 183.3367 177.5828 160.7508 177.9392 190.2131 188.9345 176.6567
 [9] 191.1876 174.1499 170.7538 178.1061 181.3366 190.0044 164.2285 179.7494
[17] 183.1407 183.2501 180.6029 180.9479 163.9846 186.3207 159.9828 182.8885
[25] 185.9019 163.3696 202.3984 166.0245 160.5897 184.4011 164.9120 179.4535
[33] 177.0181 178.2846 179.0259 176.5019 164.7329
```

Step 2: Calculate the t-statistic for two samples comparison

T-statistic for two samples comparison is calculated differently from that for one sample t-test. Instead of calculating the difference between the sample mean and the population mean, this time, you have to calculate the difference between sample A (myclass) amd sample B (theirclass). Then, you have to divide it by standard error of the difference. OK. First thing first, let's calculate the difference between the two sample means.

$$\bar{x}_A - \bar{x}_B$$

Now, calculate the standard error of the difference by square-rooting the standard error of the difference between the two means. When sample sizes are small (¡60 and especially ¡30), the calculation of the standard error of the difference is a bit more complicated, and it uses the concept of pooled variance. Here, we can use the simpler method below.

$$SE_{diff} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Then, two sample t-statistics are calculated as below.

First thing first, the difference of the two means ($\bar{x}_A - \bar{x}_B$)!

```
meandiff<-mean(myclass) - mean(theirclass)
meandiff
```

```
[1] -4.790977
```

Now, the standard error...

```
sediff<-sqrt(var(myclass)/length(myclass)+var(theirclass)/length(theirclass))
sediff
```

```
[1] 1.941237
```

$$t = \frac{\bar{x}_A - \bar{x}_B}{SE_{diff}}$$

```
tstat = meandiff/sediff
tstat
```

```
[1] -2.468003
```

Now, find the percentile ranking for the tstat.

```
pt(tstat, df=70)
```

```
[1] 0.008016785
```

```
2*pt(tstat, df=70) # Since we got a negative t value, we can simply double it up to get the summed
    percentile on both ends.
```

```
[1] 0.01603357
```

Again, we don't calculate t statistic manually every time we need it. Check out the cool and easy function below.

```
t.test(myclass, theirclass)
```

```
        Welch Two Sample t-test

data:  myclass and theirclass
t = -2.468, df = 53.723, p-value = 0.0168
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.6833819 -0.8985718
sample estimates:
mean of x mean of y
 173.2131  178.0040
```

Their class has a higher mean value and because a larger mean was subtracted from the smaller mean, you get a negative value for a t statistic. Let's change the order of the vectors to see if the t statistic value changes.

```
t.test(theirclass, myclass)
```

```
        Welch Two Sample t-test

data:  theirclass and myclass
t = 2.468, df = 53.723, p-value = 0.0168
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8985718 8.6833819
sample estimates:
mean of x mean of y
 178.0040  173.2131
```

It did indeed! Now you understand that it is the absolute value of t statistic (and z statistic as well) that matters to check the statistical significance.

Now, let's go over the report for each parameter. Take a look at the first table. The t-statistic is -2.468, and the degree of freedom is 53.723. The fractional degree of freedom might seem strange, but in a two sample t-test where the variance is not assumed to be the same between the two groups, calculating the degree of freedom is different from that in one sample t-test (I'll show you how to calculate one shortly below!). And the p-value indicates the probability that you will observe the two mean values obtained here when they are equal. The confidence interval is also calculated using the same formula as in the one-sample t-test.

Now, let's get the degree of freedom first!

$$df = (\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B})/(\frac{s_A^4}{n_A^2(n_A-1)} + \frac{s_B^4}{n_B^2(n_B-1)})$$

Well, don't cry! No one will ask you to manually calculate the degree of freedom in a statistic test where variances are expected to vary between groups.But let's just do this for the sake of learning.

```
df <-
(
  (
    (var(myclass)/length(myclass)) +
    (var(theirclass)/length(theirclass))
  )^2
  /
  (
    (var(myclass)^2 / (length(myclass)^2*(length(myclass)-1))) +
    (var(theirclass)^2 / (length(theirclass)^2*(length(theirclass)-1)))
  )
)

df
```

```
[1] 53.72274
```

The parentheses should be used carefully!
Let's try and calculate the confidence intervals as well.

lower CI = mean difference - standard error (of mean difference) * t statistic at extreme 2.5%

upper CI = mean difference + standard error (of mean difference) * t statistic at extreme 2.5%

Since we already know both the mean difference and the standard error of mean difference, we only need to calculate the t statistic at extreme 2.5%. That's why we needed to calculate the degree of freedom as well.

```
qt(c(0.025, 0.975), df)
```

```
[1] -2.005116  2.005116
```

Now, the lower CI

```
low<-meandiff - (sediff * (qt(0.975, df)))
low
```

```
[1] -8.683382
```

```
upp<-meandiff + (sediff * (qt(0.975, df)))
upp
```

```
[1] -0.8985718
```

You can see that the calculated CI values are the same as the ones reported in the table. When you interpret CI values, check if both values are positive or negative. In analyses of continuous variables, CI values crossing over zero means the difference estimate is not very accurate, or the statistical power is rather weak (NOTE: In statistical tests using log odds, the criterion is to see whether CI values cross 1 (more on this later)).

Paired t-test

In the example above, we compared two samples means from independent groups: my class and their class. However, what if we're comparing the two sample means from the same group. Let's say you gave midterm and final exams to your students and you want to test if their performances in midterm and finals are statistically significantly different.

Smaller standard deviation

Let's create midterm data

```
set.seed(65)
midterm<-rnorm(30, 75, 2)
midterm
```

```
 [1] 72.60636 73.09667 75.55727 72.11012 76.64330 77.49404 74.21129 72.50319
 [9] 73.69574 72.65853 75.72090 77.46054 75.84641 71.30765 74.50865 77.50766
[17] 78.12668 72.47107 70.31265 73.44788 70.67654 75.11402 74.50266 74.84338
[25] 75.59542 72.63698 78.94134 73.04705 77.74904 74.51497
```

```
mean(midterm)
```

```
[1] 74.49693
```

```
sd(midterm)
```

```
[1] 2.300255
```

And finalterm data

```
final<-rnorm(30, 78, 3)
final
```

```
 [1] 76.88293 78.87081 80.43136 82.79692 79.34112 78.63515 78.64910 75.46227
 [9] 81.79797 74.43899 81.09068 76.29334 70.63368 71.09844 80.76745 81.85636
[17] 80.94792 70.56439 73.21560 75.33373 77.36916 80.37454 83.36378 75.18807
[25] 73.34161 78.60667 77.24322 76.19854 81.26545 81.06022
```

```
mean(final)
```

```
[1] 77.77065
```

```
sd(final)
```

```
[1] 3.643675
```

Now, compare the two means using the paired t-test

```
t.test(midterm, final, paired = TRUE)
```

```

        Paired t-test

data:  midterm and final
t = -5.0008, df = 29, p-value = 2.531e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.612603 -1.934828
sample estimates:
mean of the differences
              -3.273715
```

Now, paired t-tests are a bit different from one sample t-tests (student's t) or unpaired two sample t-tests (welch's t). To make it easy for you to understand, let's combine the two vectors and create a data frame.

```
data<-as.data.frame(cbind(midterm, final))
data
```

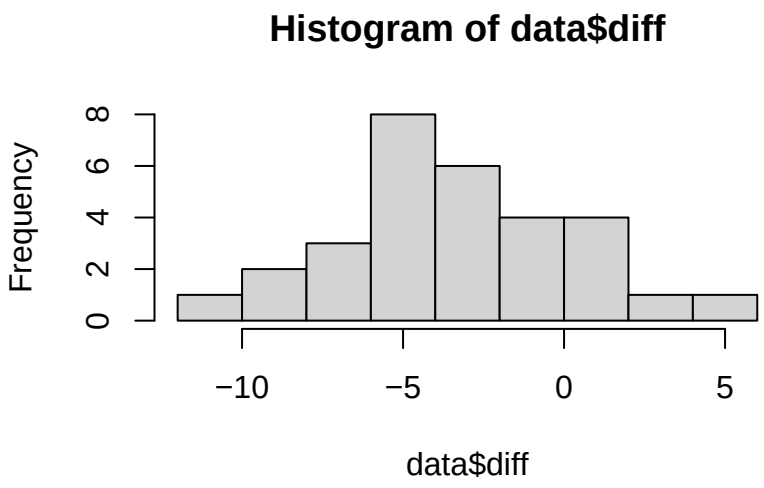
```
    midterm    final
1  72.60636 76.88293
2  73.09667 78.87081
3  75.55727 80.43136
4  72.11012 82.79692
5  76.64330 79.34112
6  77.49404 78.63515
7  74.21129 78.64910
8  72.50319 75.46227
9  73.69574 81.79797
10 72.65853 74.43899
11 75.72090 81.09068
12 77.46054 76.29334
13 75.84641 70.63368
14 71.30765 71.09844
15 74.50865 80.76745
16 77.50766 81.85636
17 78.12668 80.94792
18 72.47107 70.56439
19 70.31265 73.21560
20 73.44788 75.33373
21 70.67654 77.36916
22 75.11402 80.37454
23 74.50266 83.36378
24 74.84338 75.18807
25 75.59542 73.34161
26 72.63698 78.60667
27 78.94134 77.24322
28 73.04705 76.19854
29 77.74904 81.26545
30 74.51497 81.06022
```

You can see that students from 1 to 30 have two scores each: midterm score and final score. Now, let's create a vector that shows the differences between the two exams.

```
data$diff<-data$midterm-data$final
head(data)
```

```
   midterm    final       diff
1 72.60636 76.88293  -4.276567
2 73.09667 78.87081  -5.774147
3 75.55727 80.43136  -4.874092
4 72.11012 82.79692 -10.686806
5 76.64330 79.34112  -2.697812
6 77.49404 78.63515  -1.141108
```

```
hist(data$diff)
```



To calculate the t-statistic for paired samples, we use a different formula.

$$t_{n-1} = \frac{\bar{X}}{s/\sqrt{n}}$$

Here, $\bar{X}$ is the mean of differences. You see the third column in the data file?, $\bar{X}$ is the mean of the third column.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Now, let's take a look at what s is in the numerator.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

$X_i - \bar{X}$ is basically the difference between the individual difference from the mean difference. Let's create a vector for that and call it resid(ual).

```
data$resid<-data$diff - mean(data$diff)
data$sq.resid<-data$resid^2
head(data)
```

```
   midterm    final       diff      resid   sq.resid
1 72.60636 76.88293  -4.276567 -1.0028519  1.0057119
2 73.09667 78.87081  -5.774147 -2.5004317  6.2521589
3 75.55727 80.43136  -4.874092 -1.6003765  2.5612049
4 72.11012 82.79692 -10.686806 -7.4130902 54.9539064
5 76.64330 79.34112  -2.697812  0.5759031  0.3316643
6 77.49404 78.63515  -1.141108  2.1326069  4.5480124
```

When the residuals are squared, summed up, and divided by n-1, you get the variance of the mean differences. The square root of the variance is the standard deviation of the mean differences.

When s is...,

```
s←sqrt(
  sum(data$sq.resid)
  /
  nrow(data)-1
)

s
```

```
[1] 3.380535
```

t_{n-1} is...

```
t.paired = mean(data$diff)/(s/sqrt(nrow(data)))
t.paired
```

```
[1] -5.304154
```

You can see that the calculated t-value is the same as the one we saw in the paired t.test function output. Since there is only one group of people (with two repeated measures), the degree of freedom is 30-1. Confidence intervals are calculated as...

$\bar{X} \pm$ standard error * t statistic at extreme 2.5%

```
mean(data$diff) - ((s/sqrt(30))*(qt(0.975, 29)))
```

```
[1] -4.536028
```

```
mean(data$diff) + ((s/sqrt(30))*(qt(0.975, 29)))
```

```
[1] -2.011403
```

As you can see, the CI values do not cross over the zero, which means that the obtained t statistic has enough power for the given number of data points.

Larger standard deviation

Now, let's try and see if changing the variance will change the results.

```
set.seed(65)
midterm2←rnorm(30, 75, 15)
midterm2
```

```
 [1]  57.04769  60.72499  79.17955  53.32588  87.32478  93.70528  69.08465
 [8]  56.27395  65.21807  57.43898  80.40675  93.45406  81.34804  47.30741
[15]  71.31488  93.80745  98.45013  56.03301  39.84486  63.35910  42.57403
[22]  75.85516  71.26992  73.82538  79.46569  57.27732 104.56002  60.35289
[29]  95.61783  71.36228
```

```
mean(midterm2)
```

```
[1] 71.227
```

```
sd(midterm2)
```

```
[1] 17.25191
```

This was another set of midterm scores
And here we have final scores.

```
final2←rnorm(30, 78, 20)
final2
```

```
 [1]  70.55284  83.80541  94.20910 109.97949  86.94077  82.23431  82.32734
 [8]  61.08177 103.31978  54.25993  98.60456  66.62227  28.89120  31.98958
[15]  96.44966 103.70904  97.65279  28.42927  46.10403  60.22488  73.79441
[22]  93.83027 113.75856  59.25380  46.94406  82.04446  72.95477  65.99029
[29]  99.76965  98.40145
```

```
mean(final2)
```

```
[1] 76.47099
```

```
sd(final2)
```

```
[1] 24.29116
```

Now compare the two exam means

```
t.test(midterm2, final2, paired = TRUE)
```

```
        Paired t-test

data:  midterm2 and final2
t = -1.1705, df = 29, p-value = 0.2513
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.406839   3.918856
sample estimates:
mean of the differences
              -5.243991
```

As you can see from the results, when variances get larger, the true mean difference becomes smaller and gets harder to reach the t statistic of ± 2.

How to interpret Confidence Intervals

A confidence interval is a range of values that we are quite sure that a true mean (or true mean difference) can be found.

- From the one-sample t-test, the test output gave us two numbers (171.4026, 175.0235). The two numbers are the lower and upper boundaries of the range we expect the true mean of the population.
- From the two-sample t-test, the output gave us the lower and upper boundaries of the mean 'difference' of the two samples. In this case, you should pay attention to the plus/minus sign of the values. In confidence intervals of continuous values, the lower and upper CIs should not cross zero for the obtained statistic to be of significance.
- If you compare two different sets of paired t-tests at the end, you will see that the t-test of midterm and finalterm gave us both negative values (thus, not crossing zero) while that of midterm2 and finalterm2 gave one negative and one positive values, which means the confidence interval crossed zero. And the obtained statistic is not significant.

Questions

- What is the meaning of each p value in the two t-tests above?