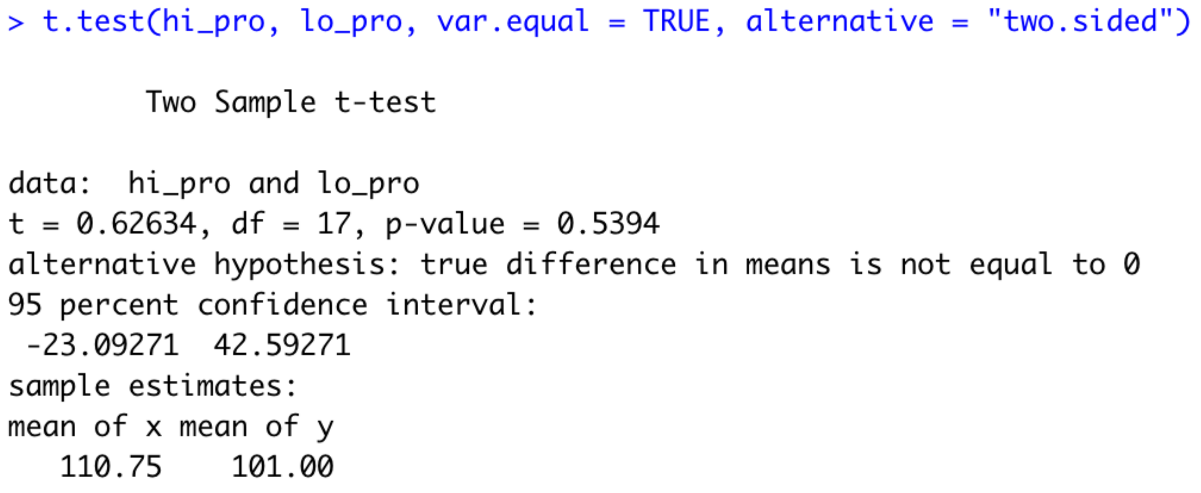
Stann-Omar Jones
Advanced Statistics for Data Science
Prof. Chong Liu

1. Consider the gain in weight of 19 female rats between 28 and 84 days after birth. 12 were fed on high protein diet and 7 on a low protein diet. Using the following data, test the hypothesis that there is no difference in weight gain between female rats raised on a high-protein diet versus those raised on a low-protein diet. Use a significance level of $\alpha = 0.05$ and assume equal variances. (“Hint: var.equal=”TRUE”)

   High protein: 134,146,104,119,124,161,107,83,113,129,97,12

   Low protein: 70,118,101,85,107,132,94

Note: This is a small sample hypothesis with known variances and so we use a two-sample t-test.



```
> t.test(hi_pro, lo_pro, var.equal = TRUE, alternative = "two.sided")

	Two Sample t-test

data:  hi_pro and lo_pro
t = 0.62634, df = 17, p-value = 0.5394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23.09271  42.59271
sample estimates:
mean of x mean of y 
   110.75    101.00 
```

Here’s how to interpret the results of the test:

data: This tells us the data that was used in the two sample t-test. In this case, we used the vectors called hi_pro and lo_pro

t: This is the t test-statistic. In this case, it is **0.62634**.

df: This is the degrees of freedom associated with the t test-statistic. In this case, it's **17**.

p-value: This is the p-value that corresponds to a t test-statistic of 0.62634 and df = 17. The p-value turns out to be 0.5394.

alternative hypothesis: This tells us the alternative hypothesis used for this particular t-test. In this case, the alternative hypothesis is that the true difference in means between the two groups is not equal to zero.

95 percent confidence interval: This tells us the 95% confidence interval for the true difference in means between the two groups. It turns out to be **[-23.09271, 42.59271]**.

sample estimates: This tells us the sample mean of each group. In this case, the sample mean of group 1 was **110.75** and the sample mean of group 2 was **101**.

The two hypotheses for this particular two sample t-test are as follows:

H_0: $\mu_1 = \mu_2$ (the two population means are equal)

H_A: $\mu_1 \neq \mu_2$ (the two population means are *not* equal)

Because the p-value of our test **(0.5394)** is greater than alpha = 0.05, we cannot reject the null hypothesis of the test. This means we do not have sufficient evidence to say that the weight gain between the two samples is different.

2. Load the "MASS" package. In the immer dataset of the "MASS" library: we have Y1 Yield in 1931, Y2 Yield in 1932. Assuming that the data in immer follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean barley yields between years 1931 and 1932 (Hint:paired t-test). Get "p-value" in a variable pvalue and "statistics" in a variable st. (Hint: ttest<-t.test(..,..,...) and then names(ttest))

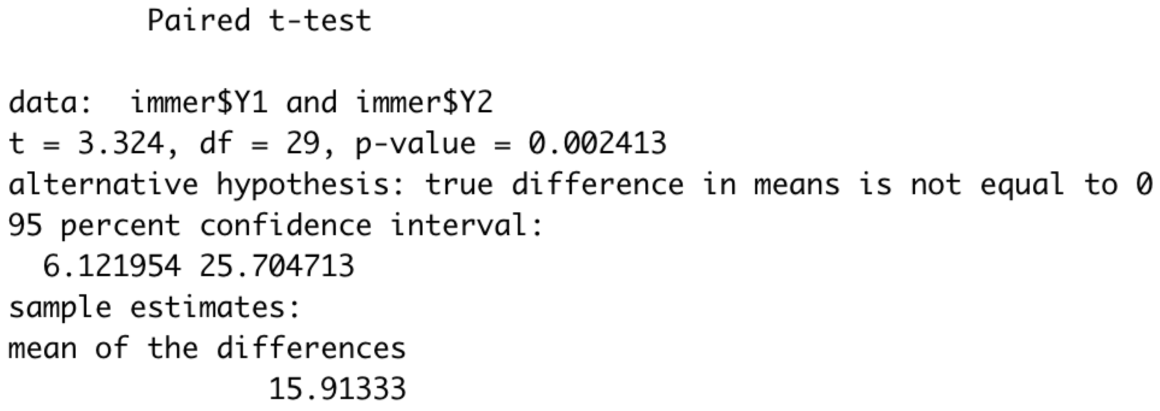
Assumptions

Assumption 1: Are the two samples paired?

Yes, since the data collected represents five varieties of barley that were grown in the same six locations in 1931 and 1932.

Assumption 2: Is this a large sample?

No, because n < 30. Since the sample size is not large enough (less than 30), we need to check whether the differences of the pairs follow a normal distribution.



```
	Paired t-test

data:  immer$Y1 and immer$Y2
t = 3.324, df = 29, p-value = 0.002413
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.121954 25.704713
sample estimates:
mean of the differences 
               15.91333 
```

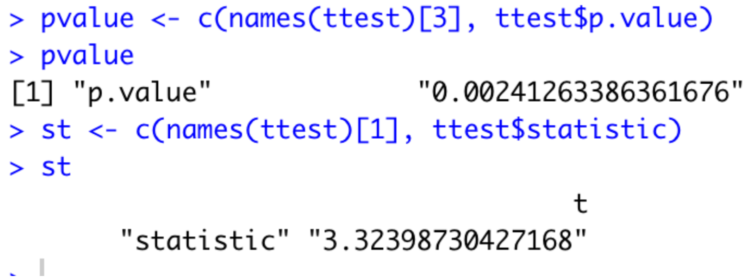
In the result above :

- **t** is the **t-test statistic** value (t = 3.324),
- **df** is the degrees of freedom (df= 29),
- **p-value** is the significance level of the **t-test** (p-value = 0.002413.
- **conf.int** is the **confidence interval** (conf.int) of the mean differences at 95% is also shown (conf.int= [6.121954, 25.704713])
- **sample estimates** is the mean differences between pairs (mean = 15.91333).

The p-value of the test is 0.002413., which is less than the significance level alpha = 0.05. We can then reject null hypothesis and conclude that the mean barley yields in years 1931 are significantly different from mean barley yield in 1932.

If the assumptions of the analysis are true, you can be 95% sure that this confidence interval contains the true difference between means.

- Get “p-value” in a variable pvalue and “statistics” in a variable st. (Hint: ttest<-t.test(..,..,…) and then names(ttest))



```
> pvalue <- c(names(ttest)[3], ttest$p.value)
> pvalue
[1] "p.value"             "0.00241263386361676"
> st <- c(names(ttest)[1], ttest$statistic)
> st
                                    t 
       "statistic" "3.32398730427168" 
```