# Homework 6

Nora Quick

2021-11-02

## Instructions

Use the R Markdown version of this file to complete and submit your homework. Make sure you change the author in the header to your own name.

## Conceptual Questions

1. For each scenario below, **state and justify whether the data are paired or not.**

   a) Mark is tutoring European history students this summer to make extra money. He has 12 students, and so far each student has had two sessions with him. At the end of each session, they take a practice test. Mark would like to know whether there is a difference in mean score between the first and second session.

Yes, this is paired. The common variable between the two sets of data is the student he tutored and the two sets of data compaired are the first session and the second session.

   b) A chemist wishes to compare the amount of residue left behind for chemical reaction A and chemical reaction B, given a certain controlled environment for each reaction. She runs each reaction 8 times, and records the residues (in micrograms).

No, this is not paired. There is no common relationship between the two chemicals. Checking the residue left behind of each chemical is individudal to the chemical itself.

   c) A language transcriptionist translates a random sample of seven speeches. Each speech is translated from Spanish to English and from Spanish to French, and how long each transcription takes is recorded. The transciptionist would like to know if there is a difference in time it takes her to transcribe from Spanish to English compared to from Spanish to French.

Yes, this is paired. All of the languages she is test are the same and she translates them in the same order. In addition she is timing herself translating between the languages giving both similar variables and something to test between them.

2. A random sample of cars of model A and vans of model B are selected. Each is driven on flat ground over the same section of highway for one week and the gas mileage (in miles per gallon) is calculated and recorded. The measurements are stored in the vectors `mpg_cars_A` and `mpg_vans_B` as follows:

```r
mpg_cars_A <- c(
  23.35, 23.97, 28.76, 33.24, 26.66, 26.72, 28.23, 27.66, 27.12,
  25.73, 25.17, 25.80, 24.09, 26.60, 27.85, 25.07, 27.55, 31.78,
  22.84, 21.11, 28.68, 29.62, 25.21, 28.70, 26.52
)

mpg_vans_B <- c(
  28.48, 25.63, 31.91, 29.47, 26.97, 26.88, 28.81, 27.84, 30.03,
  30.03, 29.46, 29.40, 27.74, 31.45, 31.24, 29.08, 26.35, 30.56,
  29.57, 28.32, 27.78, 29.33, 29.04, 28.34, 27.53, 28.40, 28.83,
  27.38
)

t.test(x = mpg_cars_A, y = mpg_vans_B)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg_cars_A and mpg_vans_B
## t = -3.3685, df = 36.695, p-value = 0.001788
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.2981271 -0.8201872
## sample estimates:
## mean of x mean of y
##  26.72120  28.78036
```

a) Use R to test the hypothesis that the proportion of cars of model A that get at least 28 miles per gallon is different than that of vans from model B. Assume that we may use the normal approximation. **State a conclusion in the context of the problem.**

There is convincing evidence the mean difference in miles per gallon between model A and model B is not zero (paired t-test, p-value=0.0017).

b) **Give and interpret, a plausible range of values for the actual difference in proportions.**

With 95% confidence the mean difference is between 3.3 to 0.8 less for model A than model B. This means there is evidence that the average miles per hour for model A is less than the average miles per hour for model B.

c) **Why is it important that our sample sizes are “large”?*

We need large sample sizes because the there may be a lot of variability in the two models. Distance in the trips the cars take, the number of trips the cars take, etc.

3. Two teams, the Tigers and the Bears, compete in a rocket-launching competition. Each team builds one rocket, and on competition day, each rocket is launched 20 times. A laser measures the vertical distances (in feet) reached by the rockets. The winner is the team with the highest median height.

The heights recorded are as follows:

```r
tiger_heights <- c(
   11, 601, 550, 16, 100, 293, 67,  60, 474, 132,
  218,  74, 251, 492, 38, 119, 127, 106,  23, 269
)

bear_heights <- c(
  179, 86, 51, 87, 126, 82, 15, 82, 136, 55,
  171, 83, 17, 50, 142, 57, 9, 112, 32, 240
)

mood.test(tiger_heights, bear_heights, alternative = "two.sided")
```

```
##
##  Mood two-sample test of scale
##
## data:  tiger_heights and bear_heights
## Z = 1.5692, p-value = 0.1166
## alternative hypothesis: two.sided
```

And the Tiger's are declared the winners.

But the Bear's coach understands that these 20 heights are like a sample from each rockets' population of possible heights. He is curious if the rocket's have different population median heights.

**Perform a Mood's test to answer the coach's question.**

With a two-sided mood's test we find a p-value of 0.1166. This p-value indicates that the data is not significant. In the context of this problem we are assuming (NULL hypothesis) that the tigers have significant data to indicate their win. However, the p-value shows that they do not have significant data that their rocket reaches a higher height (the alternative hypothesis is true).

# R Question

Why is it important to correctly distinguish between the 2-sample t-test setup, and the paired t-test setup? These questions lead you through an example where the two procedures could lead to different conclusions.

The following code simulates three samples, `A`, `C`, and `D`:

```r
set.seed(1810)
n <- 10
A <- rnorm(n)
C <- 0.5 + (0.8 * A) + (sqrt(1 - 0.8^2) * rnorm(n))
D <- sample(C)
```

a) Using `t.test()`, conduct a **two sample t-test** of $H_0 : \mu_A - \mu_D = 0$ vs. $H_A : \mu_A - \mu_D \neq 0$, assuming unequal group variances. **Write a two sentence summary that includes an interpretation of the test result and the confidence interval**

```r
t.test(x = A, y = D)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  A and D
## t = -1.3424, df = 16.041, p-value = 0.1981
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.8272308  0.4101305
## sample estimates:
## mean of x mean of y
## 0.1772973 0.8858474
```

It is estimated the difference in A and D is greater than -1.8 and less than 0.4. So, with 95% confidence the mean weight of A is between 1.8 less and 0.4 greater than D.

b) Now using `t.test()`, conduct a **paired t-test** of $H_0 : \mu_A - \mu_D = 0$ vs. $H_A : \mu_A - \mu_D \neq 0$, assuming unequal group variances. **Write a two sentence summary that includes an interpretation of the test result and the confidence interval**

```
t.test(x = A, y = D, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  A and D
## t = -1.3375, df = 9, p-value = 0.2139
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.9069598  0.4898595
## sample estimates:
## mean of the differences
##              -0.7085501
```

It is estimated D is 0.7 larger than A. With 95% confidence the mean difference between A and D is between -1.9 and 0.5.

c) **Do the unpaired and paired tests reach the same conclusion regarding the population means of A and D?** *Be specific about where the results agree and/or disagree.*

Yes, both the unpaired and paired tests reach the same conclusion regarding the population means. They have a similar range of -1.8/-1.9 to 0.4/0.5. Additionally, the p-values of both results are similar (0.2). With only a little bit of variation within the ranges and p-values I conclude that the results agree.

d) **Repeat parts (a), (b) and (c), now comparing samples A and C.**

```
t.test(x = A, y = C)
```

```
##
##  Welch Two Sample t-test
##
## data:  A and C
## t = -1.3424, df = 16.041, p-value = 0.1981
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.8272308  0.4101305
## sample estimates:
## mean of x mean of y
## 0.1772973 0.8858474
```

```r
t.test(x = A, y = C, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  A and C
## t = -2.7164, df = 9, p-value = 0.02375
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2986192 -0.1184811
## sample estimates:
## mean of the differences
##               -0.7085501
```

It is estimated the difference in A and C is greater than -1.3 and less than 0.4. So, with 95% confidence the mean weight of A is between 1.8 less and 0.4 greater than C.

It is estimated C is 0.7 larger than A. With 95% confidence the mean difference between A and C is between -1.3 and -0.12.

No, the unpaired and paired tests do not reach the same conclusion regarding the population means. They have a differnant ranges of -1.3/-1.3 to 0.4/-0.12. Additionally, the p-values of both results are different with the unpaired p-value of 0.2 and a paired p-value of 0.02. With a variation in ranges and, more importantly, p-values I conclude that the two tests disagree.

e) You should find that the two procedures, the paired t-test and the two sample t-test, reach roughly the same conclusion when comparing samples A & D, but different conclusions when comparing samples A & C, despite the true differences in mean in both cases being 0.5. **Explain why the A & C case differs from the A & D case. In the A & C case, which procedure would be more appropriate?**

The A and C test differs from the A and D test because of the way C is calculated. I beleive that the correct test for A and C would be a one-sample test.

*You may find it helpful to either examine how the samples were generated and/or examine the following plots of the the three samples.*