Homework 6

Rui Gao

2021-11-07

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. Paired

   b) A chemist wishes to compare the amount of residue left behind for chemical reaction A and chemical reaction B. She runs each reaction 8 times, and records the residues (in micrograms). All reactions were performed under certain controlled environment. Not paired

   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. Paired

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:

```
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
)
```

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.**

```r
x_c_2 <- c(7, 19)
n_c_2 <- c(25,25)

prop.test(x_c_2,n_c_2,correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x_c_2 out of n_c_2
## X-squared = 11.538, df = 1, p-value = 0.0006817
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.7229091 -0.2370909
## sample estimates:
## prop 1 prop 2
##   0.28   0.76
```

There is convincing evidence to say that the population proportions are different.

b) **Give and interpret, a plausible range of values for the actual difference in proportions.** The cars is 0.24 to 0.72 miles less per gallon compared with vans.

c) **Why is it important that our sample sizes are "large''?** To let the shape of samples being Normal. Because the method of tests is based on the normal distribution.If it's not normal, the accuracy level is not acceptable.

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
)
```

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.

```
heights <- c(
   11, 601, 550, 16, 100, 293, 67,  60, 474, 132,
  218,  74, 251, 492, 38, 119, 127, 106,  23, 269,
  179, 86, 51, 87, 126, 82, 15, 82, 136, 55,
  171, 83, 17, 50, 142, 57, 9, 112, 32, 240
)
median(heights)
```

```
## [1] 93.5
```

```
x_c_3 <- c(13, 7)
n_c_3 <- c(20, 20)
prop.test(x_c_3, n_c_3, correct=F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x_c_3 out of n_c_3
## X-squared = 3.6, df = 1, p-value = 0.05778
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.004376611 0.595623389
## sample estimates:
## prop 1 prop 2
##   0.65   0.35
```

There is enough evidence to say that there are different median heights.

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`:

```
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**

```
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
```

We can say we have convincing evidence that with 95% confidence level, the mean of A is 0.18 to 0.86 higher 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
```

We can say we have convincing evidence that with 95% confidence level, the mean of A is 0.19 lower to 0.49 higher than D.

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.* No, despite the p-values are greater than 0.05 and we are able to say there has difference. We have a different range and they are not at the same side of 0.Although they agree at the range of (0.18, 0.49).

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
```

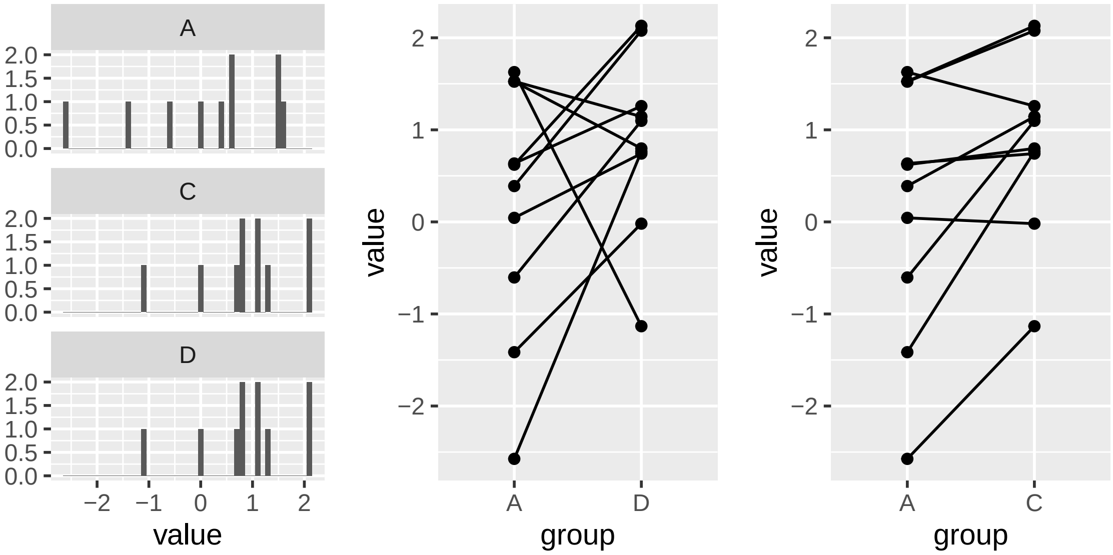
```
t.test(x = A, y = C, paired = TRUE)
```

```
##
##  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
```

They are also different.

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?** Because when we compare A and C, we are comparing the elements one by another one that is produced by the previous one.SO, the relationships are obvious. But when we compare A and D, the one and another one might not connected.

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


A
C
D
value
value
A
D
group
A
C
group