Thursday Office Hours

203 team

10/14/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

t-tests and regressions

Here's the claim:

> If we run a t-test or a regression against the same data that has a binary (or two-category) RHS feature, we will get the same answers.
>
> When we say the "same" I mean:
>
> - The same estiamte for the difference; and,
> - The same p-value for the test.

Let's go!

```
d <- data.frame(
  id = 1:100) %>%
  mutate(
    x = sample(c('a', 'b'), size = 100, replace = TRUE),
    y = 10 + .2 * (x == 'a') + rnorm(n = 100, mean = 0, sd = 1)
  )
```

```
d %>%
  group_by(x) %>%
  summarise(
    mean_y = mean(y)
  ) # this checks out!
```

```
## # A tibble: 2 x 2
##   x     mean_y
##   <chr>  <dbl>
## 1 a      10.1
## 2 b       9.91
```

Run a t-test

```
t_test_result <- t.test(y ~ x, data = d, var.equal = TRUE)
t_test_result
```

```
##
##  Two Sample t-test
##
## data:  y by x
## t = 1.2096, df = 98, p-value = 0.2294
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1525477  0.6287847
## sample estimates:
## mean in group a mean in group b
##       10.145512        9.907393
```

```
diff_in_means <- t_test_result$estimate[1] - t_test_result$estimate[2]
(t_test_result$conf.int[1] + t_test_result$conf.int[2]) / 2  == diff_in_means
```

```
## mean in group a
##            TRUE
```

Run a linear model

```
lm_result <- lm(y ~ x, data = d)
lm_result
```

```
##
## Call:
## lm(formula = y ~ x, data = d)
##
## Coefficients:
## (Intercept)           xb
##     10.1455      -0.2381
```

```
summary(lm_result)
```

```
##
## Call:
## lm(formula = y ~ x, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.72493 -0.57766 -0.01542  0.59077  2.18824
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.1455     0.1214   83.60   <2e-16 ***
## xb           -0.2381     0.1969   -1.21    0.229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9555 on 98 degrees of freedom
## Multiple R-squared:  0.01471,    Adjusted R-squared:  0.004656
## F-statistic: 1.463 on 1 and 98 DF,  p-value: 0.2294
```

Wilcox Rank Sum Tests