Test 02: Numeric Y vs categorical X

Richard

1/30/2022

Load required packages

```
library(tidyverse) # if you're using macOS, you can run: l
library(skimr)
```

Prepare Data

Please read the intro about data at here

```
Hsb <- within(
  read.csv("https://stats.idre.ucla.edu/stat/data/hsb2.csv"
    race <- as.factor(race)
    schtyp <- as.factor(schtyp)
    prog <- as.factor(prog)
})
```

Recall

- Numeric variable
- Categorical variable
- What is the key difference between them?

Some questions between numeric Y and categorical X

- We care about numeric Y for different groups in categorical X
- For example: Y is salary/score, X is gender
    - Do male employees earn more than the female co-workers
    - Do female students have higher write score than the male friends

R function

- Function: `t.test`
- Usage: ‘t.test(y $\sim$ x)
- `y` is a numeric variable
- `x` is a categorical variable with two groups
  - e.g., `female` includes only two values 0 and 1

R code example

```
t.test(Hsb$write ~ Hsb$female)
```

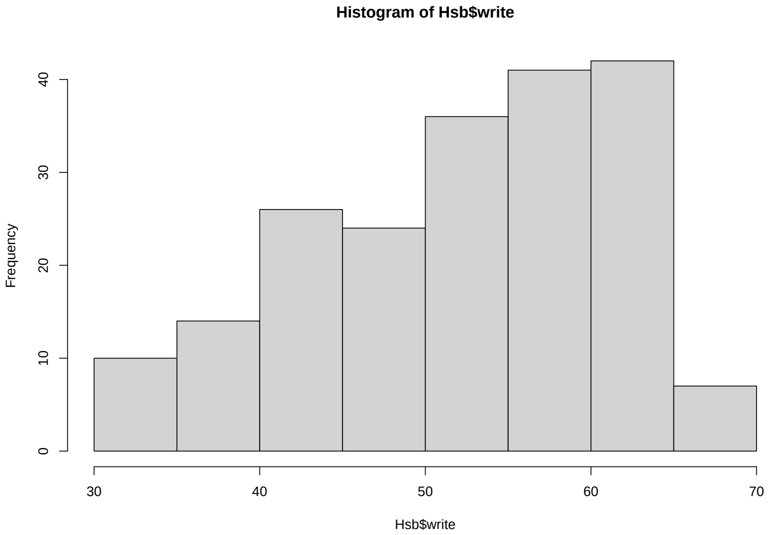
```
##
##  Welch Two Sample t-test
##
## data:  Hsb$write by Hsb$female
## t = -3.6564, df = 169.71, p-value = 0.0003409
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -7.499159 -2.240734
## sample estimates:
## mean in group 0 mean in group 1
##        50.12088        54.99083
```

Assumption of `t.test`

- It requires `y` to be normal distributed!
  - it is a very strong assumption
- If `y` is not normal distributed, we aren't confident to use t-test to answer the above research question
- So, let's check the normality assumption of `y`
  - or, we ask if `write` is normal distributed first, before we use the t-test

Check normality of `write` by histogram

```
hist(Hsb$write)
```



A normality test

```
shapiro.test(Hsb$write)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Hsb$write
## W = 0.94703, p-value = 9.867e-07
```

Discuss

- It seems that `write` does not follow normal distribution
- So we can't use t-test in this case
- Do we have an alternative test, when we don't have the normality assumption
  - Yeah! We can use The Wilcoxon-Mann-Whitney test

R function

- Function: `wilcox.test`
- Usage: `wilcox.test(y ~ x)`
- `y` is a numeric variable
- `x` is a categorical variable with two groups
  - e.g., `female` includes only two values 0 and 1

R code example

```
wilcox.test(Hsb$write ~ Hsb$female)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Hsb$write by Hsb$female
## W = 3606, p-value = 0.0008749
## alternative hypothesis: true location shift is not equal
```

Extention: more than 2 groups

- `t.test` works for 2-group X only
- If we want to check mean difference for more than 2 groups, we need to use one-way ANOVA
- For example: if the `write` score is the same for every program

```
Hsb %>% count(prog)
```

```
##   prog   n
## 1    1  45
## 2    2 105
## 3    3  50
```

R code example

```
summary(aov(Hsb$write ~ Hsb$prog))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## Hsb$prog      2   3176  1587.8   21.27 4.31e-09 ***
## Residuals   197  14703    74.6                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Discussion

- A very small p-value indicates that the write score is not the same across programs

Short quiz

- Rewrite ANOVA code above using %>% instead of `summary(aov(...))`

In sum

- In this lecture, we learn:
  - Two-sample `t.test`
  - Wilcoxon-Mann-Whitney test `wilcox.test`
  - ANOVA `aov`
- Next lecture goes to a more common case when both Y and X are numeric variables