

Test 01: One variable of interest Y

Richard

2024-01-01

Load required packages

```
1 library(tidyverse)
2 # if you're using macOS: library(dplyr)
3 library(skimr)
```

One variable of interest

- Sometimes, we often analyze/care about only one variable: salary, gender, interest, returns, score, pass/fail an exam, ...
- We can classify these measures into two groups:
 - Numeric variables: which ones above are numeric?
 - Categorical variables: which ones?
- Thanks to their data structure, we need different statistical tests applied to them when we ask questions about them
- Let check one example with a categorical variable: gender

Prepare Data

Please read the intro about data at [here](#)

```
1 Hsb = read_csv("data/raw/hsb.csv")
```

Check data

```
1 head(Hsb[,1:8])
```

```
# A tibble: 6 × 8
```

	id	female	race	ses	schtyp	prog	read	write
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	70	0	4	1	1	1	57	52
2	121	1	4	2	1	3	68	59
3	86	0	4	3	1	1	44	33
4	141	0	4	3	1	3	63	44
5	172	0	4	2	1	2	47	52
6	113	0	4	2	1	2	44	52

Summary data

```
1 summary(Hsb)
```

id	female	race	ses	schtyp
Min. : 1.00	Min. :0.000	Min. :1.00	Min. :1.000	Min. :1.00
1st Qu.: 50.75	1st Qu.:0.000	1st Qu.:3.00	1st Qu.:2.000	1st Qu.:1.00
Median :100.50	Median :1.000	Median :4.00	Median :2.000	Median :1.00
Mean :100.50	Mean :0.545	Mean :3.43	Mean :2.055	Mean :1.16
3rd Qu.:150.25	3rd Qu.:1.000	3rd Qu.:4.00	3rd Qu.:3.000	3rd Qu.:1.00
Max. :200.00	Max. :1.000	Max. :4.00	Max. :3.000	Max. :2.00

prog	read	write	math
Min. :1.000	Min. :28.00	Min. :31.00	Min. :33.00
1st Qu.:2.000	1st Qu.:44.00	1st Qu.:45.75	1st Qu.:45.00
Median :2.000	Median :50.00	Median :54.00	Median :52.00
Mean :2.025	Mean :52.23	Mean :52.77	Mean :52.65
3rd Qu.:2.250	3rd Qu.:60.00	3rd Qu.:60.00	3rd Qu.:59.00
Max. :3.000	Max. :76.00	Max. :67.00	Max. :75.00

science	socst
Min. :26.00	Min. :26.00

```
1 # skim_without_charts(Hsb)
```

Convert some columns to factors

```
1 Hsb = Hsb %>%  
2   mutate(  
3     race = as.factor(race),  
4     schtyp = as.factor(schtyp),  
5     prog = as.factor(prog)  
6   )
```

Questions

- How the authors construct `female` variable in the `Hsb` dataset.
- How much is the female ratio (female/total students)?
 - We often compare this ratio to which number/ratio?

Binomial test

- Hypothesis: Does the female ratio is equal to 0.5 or 50%?
- Null hypothesis H_0 : The female ratio is 50%.
- Alternative hypothesis H_1 : The female ratio is different from 50%.

R function

- Function: `prop.test`
- Usage: `prop.test(x, n, p)`
 - Recall: how to read help documentation in R?
 - `?prop.test`

R code example

```
1 prop.test(sum(Hsb$female), length(Hsb$female), p = 0.5)
```

1-sample proportions test with continuity correction

data: sum(Hsb\$female) out of length(Hsb\$female), null probability 0.5

X-squared = 1.445, df = 1, p-value = 0.2293

alternative hypothesis: true p is not equal to 0.5

95 percent confidence interval:

0.4733037 0.6149394

sample estimates:

p

0.545

Discussion

- In the sample, the female ratio is 54.5%
- The binomial test shows that the p-value is 0.2293, which is larger than the significance level (e.g., 10%)
- So we can't reject the null hypothesis
 - we don't have enough evidence to conclude that the female ratio is different than 50%

Similar questions in our life

- We often have same questions in our daily life
 - Does wearing mask prevent us from covid virus?
 - Does wearing helmet will help motorcyclists have less serious traffic accidents?

Short quizzes in classes

I will give you a sample data, let apply the `prop.test` with that data in class.

Numeric variable

- Next, switching to a numeric variable, which can receive any continuous value:
 - e.g., salary, returns, interest rate, ...
- We often want to know the mean (centralized tendency) and the variance/standard deviation of this variable:
 - e.g.1., what is the average salary after we graduated and got the first job
 - e.g.2., what is the average **write** score of all students in the class

Questions

- Check again, what is average `write` score in our `Hbs` data
- Is it equal to 50, or different

t test

- Hypothesis: Does the write score is equal to 50?
- Null hypothesis H_0 : The write score is 50.
- Alternative hypothesis H_1 : The write score is different from 50.

R function

- Function: `t.test`
- Usage:

R code example

```
1 t.test(Hsb$write, mu = 50)
```

One Sample t-test

```
data:  Hsb$write
t = 4.1403, df = 199, p-value = 5.121e-05
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 51.45332 54.09668
sample estimates:
mean of x
 52.775
```

Discussion

- How to read the results?

Short quizzes again

- Let apply `t.test` more

Another important stat to measure central tendency: median

- Let do the same test, but ask if the median is equal to 50 or not

```
1 wilcox.test(Hsb$write, mu = 50)
```

Wilcoxon signed rank test with continuity correction

data: Hsb\$write

V = 13177, p-value = 3.702e-05

alternative hypothesis: true location is not equal to 50

Next lecture

- This lecture introduces to you three important tests:
 - `prop.test`
 - one-sample `t.test`
 - one-sample median test `wilcox.test`
- In next lecture, we will consider two variables at the same time:
 - One is our variable of interest such as `write` score
 - Another one is another factor: such as `female` and `male` students
 - So the question is more like: are the write scores the same between female and male students? Or female students write better (thanks to their gifted writing skills) so they