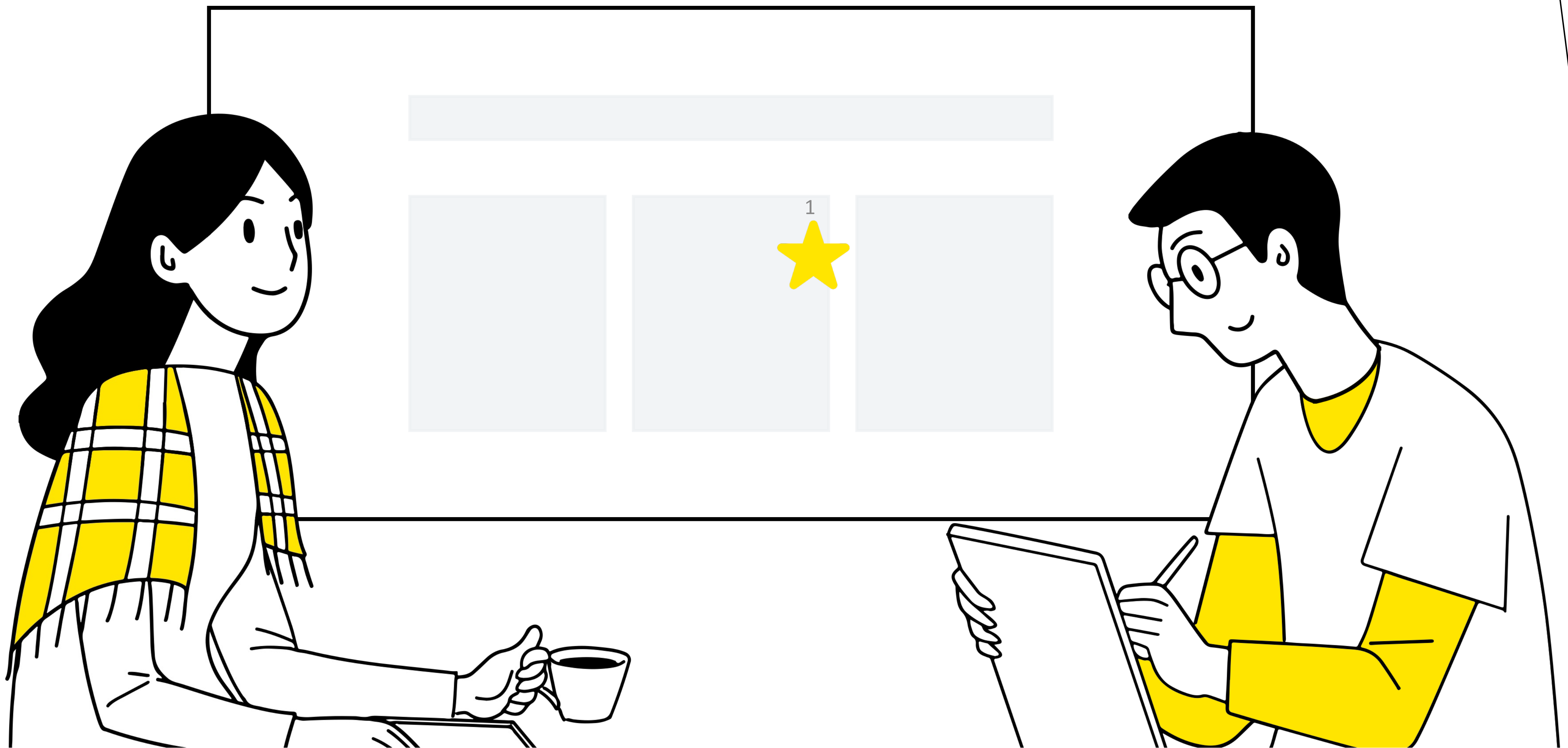
Statistics Team 5

Oct 17, 2023

Mann Whitney U Test



Members:
江沛晴
侯登耀
郭旻霏
李品嫻

Outline

1 **Intro**

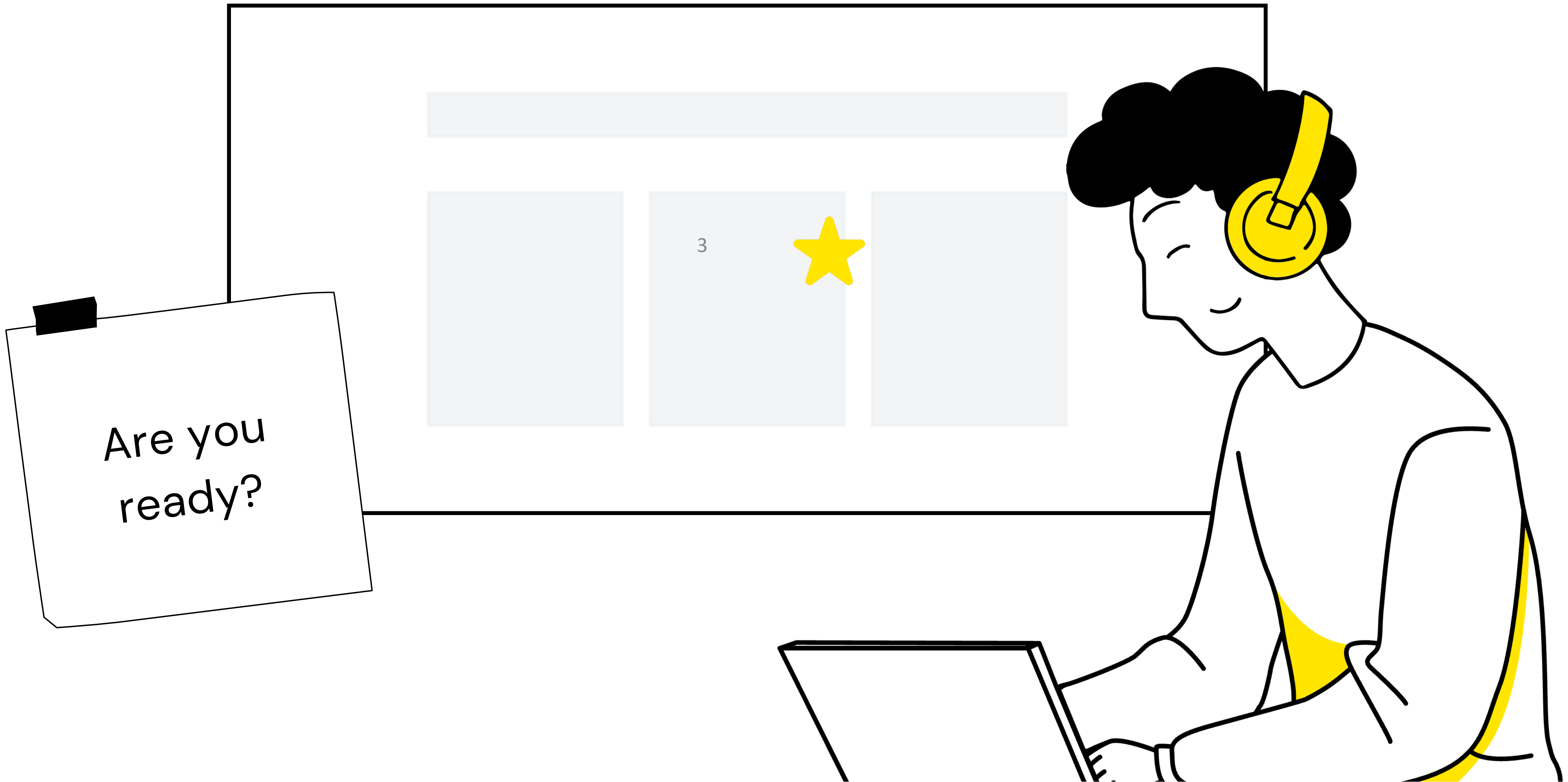
Motivation and purpose 、 Situations of the testing、Mann-Whitney Test

2 **Theory**

2

3 **Software simulation**

1 Intro



Motivations and Purpose

The procedures for testing two population parameters based on two independent samples.

For testing two population means, the t test can be used, However, alternative (nonparametric) procedures such as the median test and the Mann-Whitney test for **comparing two population medians** are considered when the data **do not meet the basic normal assumption** for the t test.

- Tests of the Difference between Two Medians
  - Median Test
  - **Mann-Whitney Test**

Situations of the testing

Nonparametric statistics

The statistic inference method that **the distribution of group is unknown** and it’s **not normal**, or **sample size is small.**

Nonparametric test

The tests **do not require the assumption of a normal distribution.**

The Mann–Whitney test is **nonparametric test of two independent ordinal samples.**

Mann–Whitney Test

Assumptions

- Two independent random samples $X_1, X_2, X_3, \dots, X_{n1}$ and $Y_1, Y_2, Y_3, \dots, Y_{n2}$ are from two populations with unknown medians M_X and M_Y, where the two populations are identical except different locations.
- The measurement is at least ordinal.
- The variable of interest is continuous.

6

Test Statistic

- Rank the combined sample from smallest to largest.
- Let the test statistic $T = S - n_1(n_1 + 1)/2$, where S is the sum of the rank of the observations from population 1.

Mann-Whitney Test

Decision Rule

- For H_0: $M_X = M_Y$ H_1: $M_X \neq M_Y$
  - Reject H_0 if the value of T is sufficiently small or large.
- For H_0: $M_X \leq M_Y$ H_1: $M_X > M_Y$
  - Reject H_0 if the value of T is sufficiently large.
- For H_0: $M_X \geq M_Y$ H_1: $M_X < M_Y$
  - Reject H_0 if the value of T is sufficiently small.

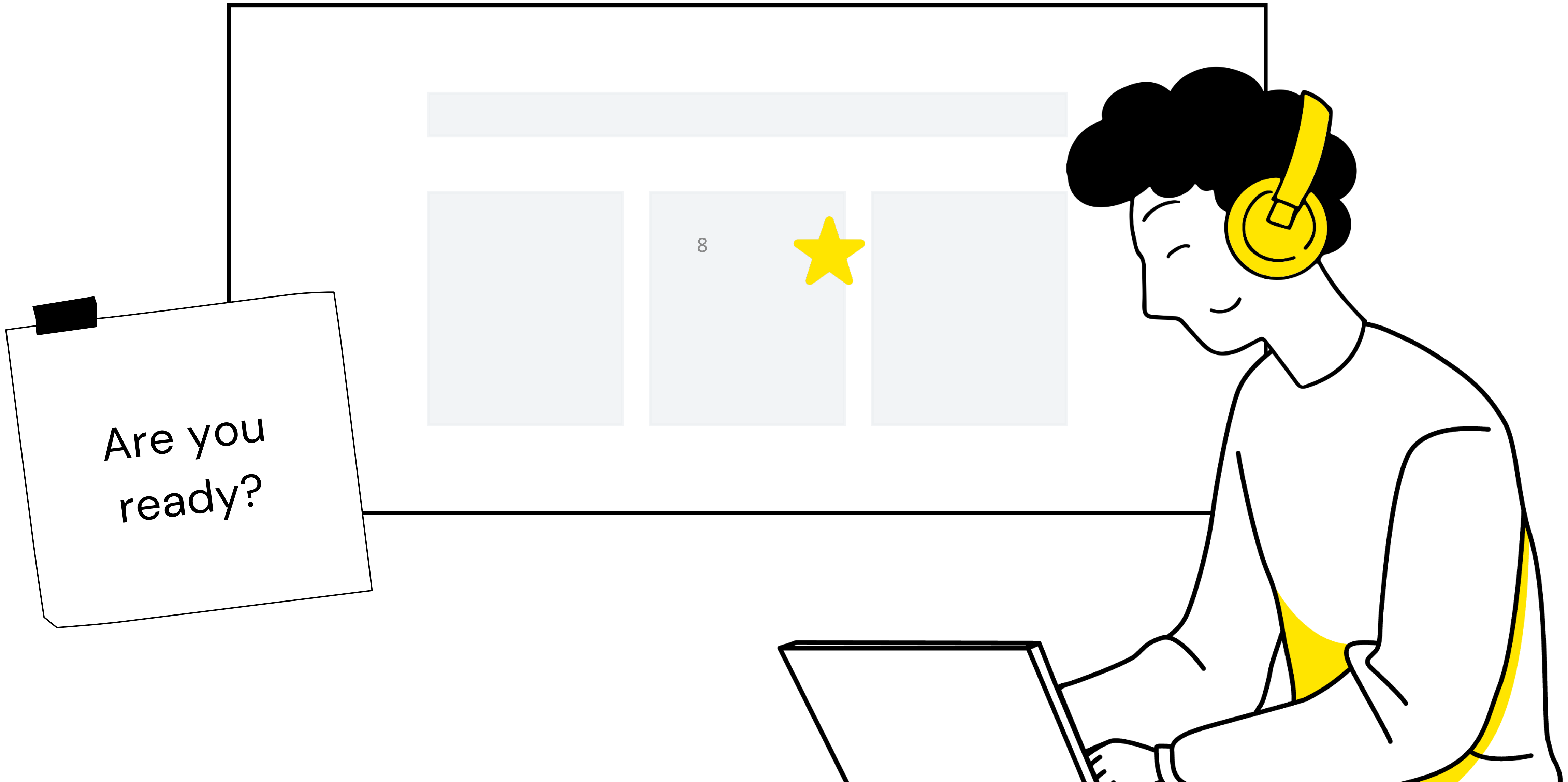
7

Large-Sample Approximation

$$Z = \frac{T - n_1 n_2/2}{\sqrt{n_1 n_2(n_1 + n_2 + 1)/12}} \approx N$$

Large sample : $n_1 + n_2 \geq 20$

2 Theory



Example_ insurance

Question："Can insurance agent who have received interpersonal relationship training courses leave a 'good' impression on customers?"

GroupX unit:score	18	15	9	10	14	16	11	13	19	17
Rank	19	16	9	10.5	15	17	12	14	20	18
GroupY unit:score	12	10	8	1	2	7	5	3	6	4
Rank	13	10.5	8	1	2	7	5	3	6	4

M_X : The median of insurance agents who have participated in training courses
M_Y :The median of insurance agents who have not participated in training courses

n_X : *GroupX sample size*
n_Y : *GroupY sample size*

H_0 : $M_X \leq M_Y$ H_1: $M_X > M_Y$

$\alpha = 0.05$

Example_ insurance

Question : "Can insurance agent who have received interpersonal relationship training courses leave a 'good' impression on customers?"

GroupX unit:score	18	15	9	10	14	16	11	13	19	17
Rank	19	16	9	10.5	15	17	12	14	20	18
GroupY unit:score	12	10	8	1	2	7	5	3	6	4
Rank	13	10.5	8	1	2	7	5	3	6	4

$S = \textit{sum of rank of score from GroupX} = 150.5$

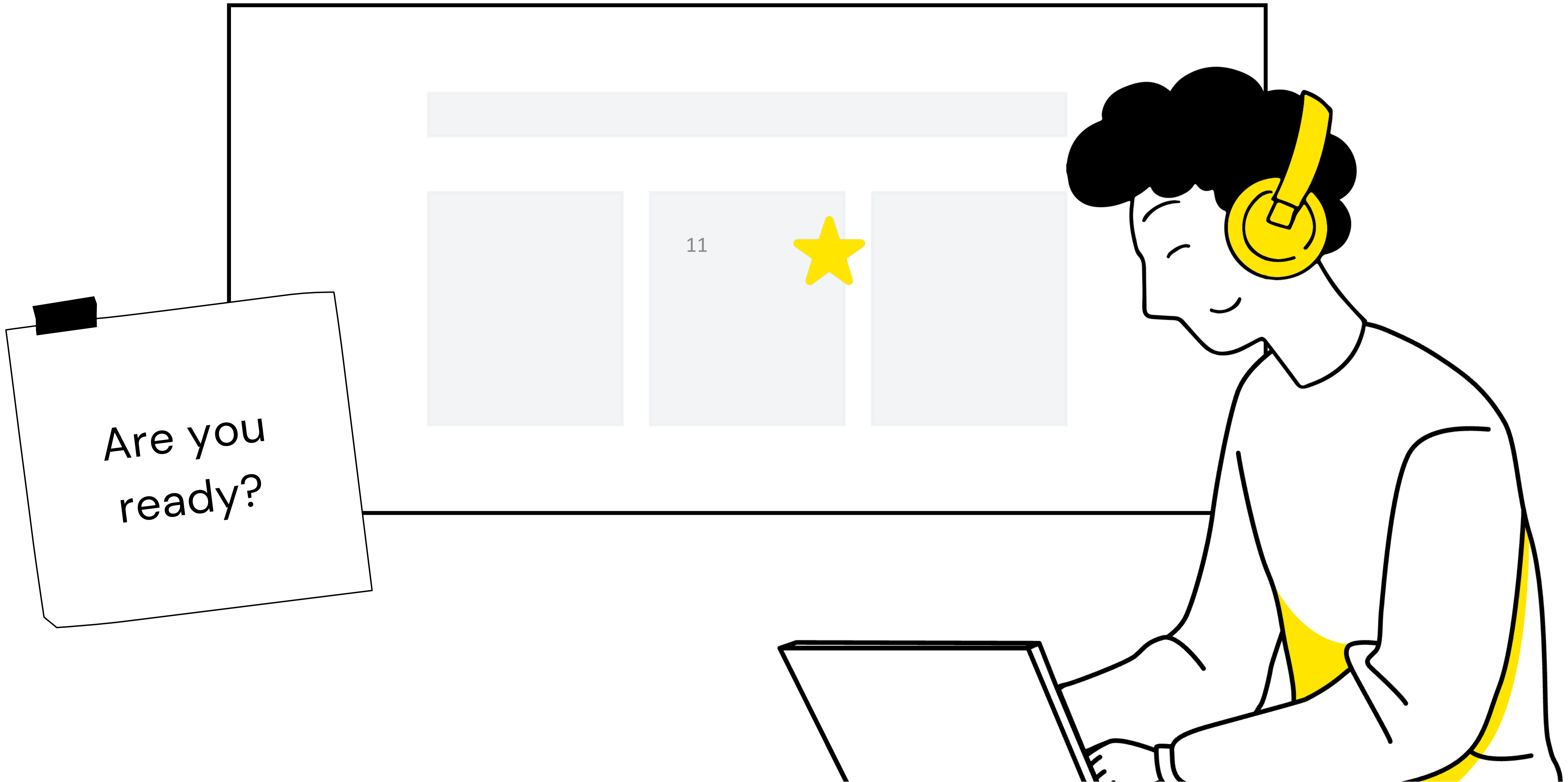
$$T = S - \frac{n_x(n_x+1)}{2} = 150.5 - \frac{10\times 11}{2} = 95.5$$

$$Z = \frac{95.5 - \frac{10\times 10}{2}}{\sqrt{10\times 10(10+10+1)/12}} = 3.439477$$

Since $Z = 3.439477 > Z_{\alpha=0.05} = 1.645$
Reject H_0 at $\alpha = 0.05$

"We have significant evidence to show that insurance agents who have received interpersonal relationship training courses are more capable of leaving a 'good' impression on customers compared to those who have not received such training."

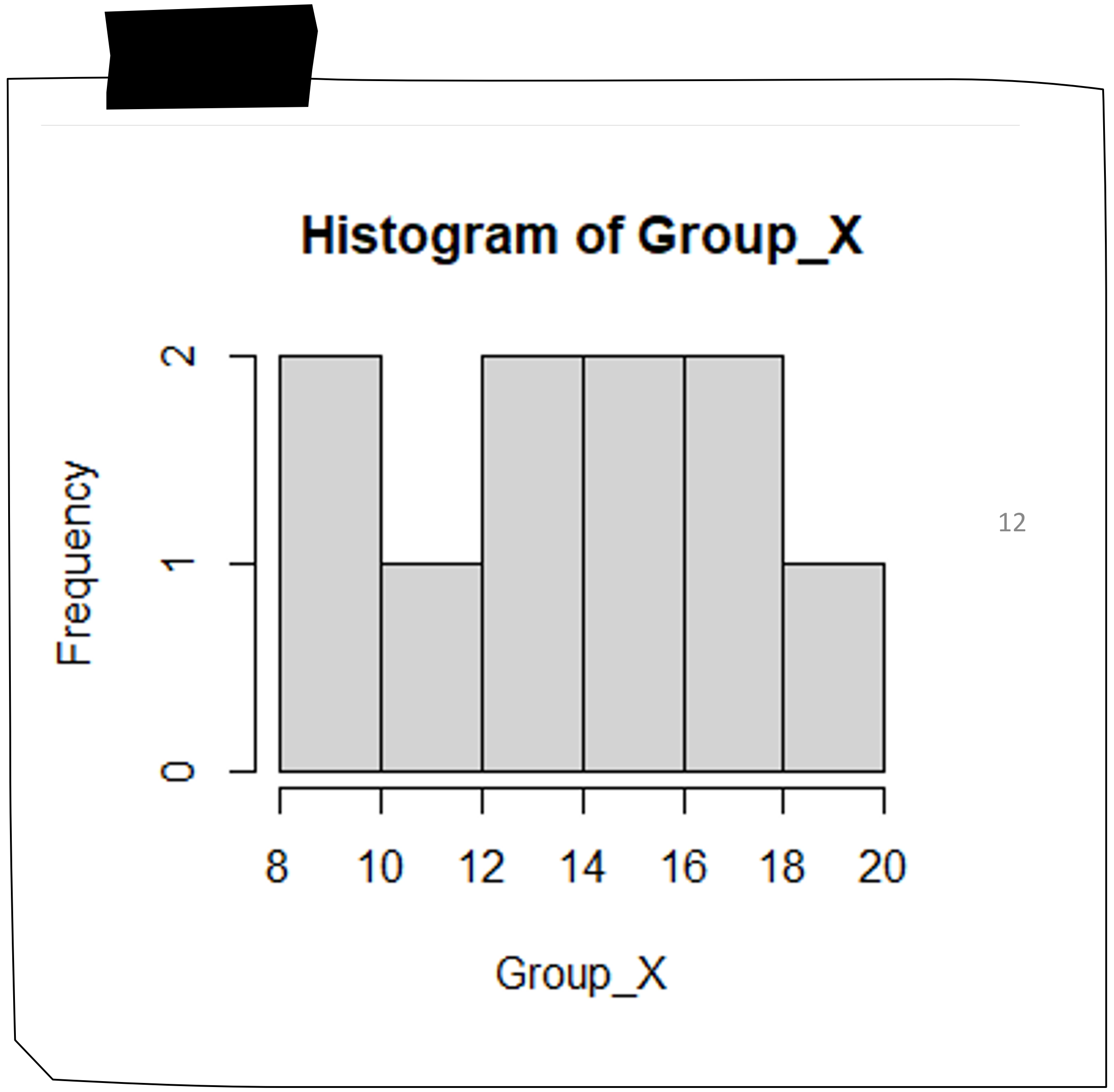
3 Software Simulation

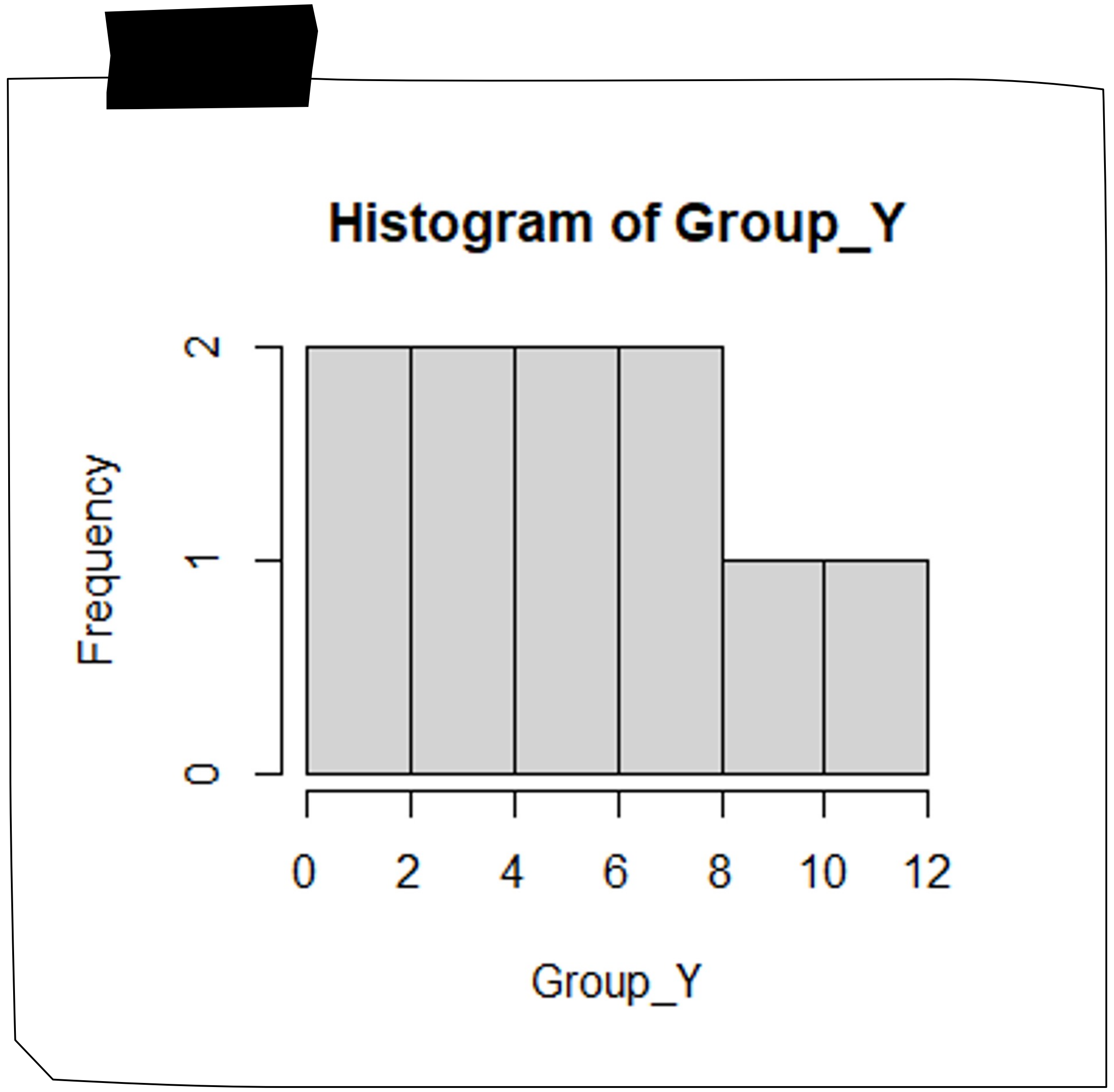


Software simulation

score	1	2	3	4	5	6	7	8	9	10
Group X	18	15	9	10	14	16	11	13	19	17
Group Y	12	10	8	1	2	7	5	3	6	4

```
Group_X <- c(18, 15, 9, 10, 14, 16, 11, 13, 19, 17); Group_Y <- c(12, 10, 8, 1, 2, 7, 5, 3, 6, 4)
hist(Group_X ); hist(Group_Y)
data <- data.frame(Group_X, Group_Y)
```





Software simulation

```r
SCORE = c(Group_X, Group_Y)
TYPE = rep(c("Group_X", "Group_Y"), each = 10)
# Now creating a dataframe
DATASET <- data.frame(TYPE, SCORE, stringsAsFactors = TRUE)
library(dplyr)
group_by(DATASET,TYPE) %>%  summarise(
        count = n(),    median = median(SCORE, na.rm = TRUE),
        IQR = IQR(SCORE, na.rm = TRUE)   )
```

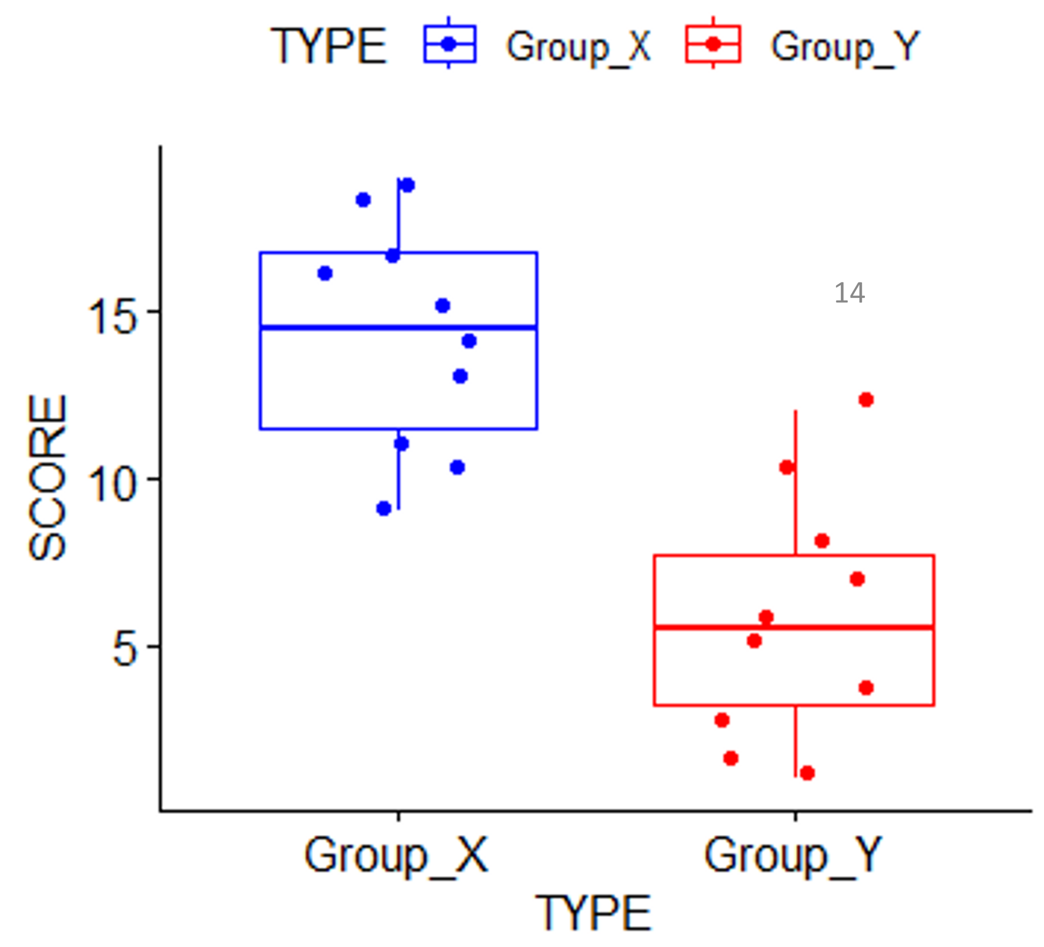
13

TYPE	count	median	IQR
Group_X	10	14.5	5.25
Group_Y	10	5.5	4.5

TYPE	SCORE
Group_X	18
Group_X	15
Group_X	9
Group_X	10
Group_X	14
Group_X	16
Group_X	11
Group_X	13
Group_X	19
Group_X	17
Group_Y	12
Group_Y	10
Group_Y	8
Group_Y	1
Group_Y	2
Group_Y	7
Group_Y	5
Group_Y	3
Group_Y	6
Group_Y	4

Software simulation

```
# loading package for boxplot
library("ggpubr")
ggboxplot(DATASET, x= "TYPE", y= "SCORE",
          color = "TYPE", palette = c("blue", "red"),
          ylab = "SCORE", xlab = "TYPE", add = "jitter")
```



TYPE	SCORE
Group_X	18
Group_X	15
Group_X	9
Group_X	10
Group_X	14
Group_X	16
Group_X	11
Group_X	13
Group_X	19
Group_X	17
Group_Y	12
Group_Y	10
Group_Y	8
Group_Y	1
Group_Y	2
Group_Y	7
Group_Y	5
Group_Y	3
Group_Y	6
Group_Y	4

Software simulation

```
res <- wilcox.test(SCORE~ TYPE, data = DATASET,
                   exact = FALSE,
                   alternative = "greater")
print(res)
```

M_X : The median of insurance agents who have participated in training courses
M_Y :The median of insurance agents who have not participated in training courses

H_0 : $M_X \leq M_Y$ H_1: $M_X > M_Y$ $\alpha = 0.05$

GroupX unit:score	18	15	9	10	14	16	11	13	19	17
Rank	19	16	9	10.5	15	17	12	14	20	18
GroupY unit:score	12	10	8	1	2	7	5	3	6	4
Rank	13	10.5	8	1	2	7	5	3	6	4

W statistic	**95.5**
P-value	**0.0003**
Conclusion 15	**Reject H0. Compared with those who have not participated in training, people who have participated in training have a better impression on customers.**

Thank you for listening