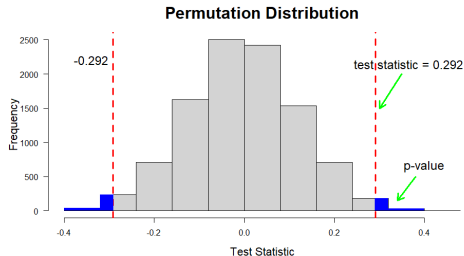
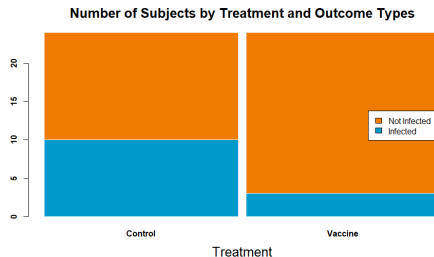


Another Case Study of Permutation Testing



Outline

- 1 The Framework of Permutation Testing
- 2 Another Case Study of Permutation Testing
- 3 Summary

Learning Objectives

In this video, we will

- Summarise the framework of permutation testing.
- Apply it to another case study.

The Framework of Permutation Testing

The Framework of Permutation Testing

- 1 State the ***Null Hypothesis*** and the ***Alternative Hypothesis***.
- 2 Pick the ***Level of Significance***.
- 3 Choose the ***Test Statistic*** and generate the permutation distribution of the test statistic, by simulating multiple permutations.
- 4 Calculate the ***p-value*** and compare it with the level of significance.
- 5 Interpret the decision.
- 6 Check the assumption.

The Difference between Permutation Testing and Classical Hypothesis Testing

- The main difference between the two approaches lies in the sampling distribution of the test statistic.
- For classical hypothesis testing:
 - ▶ We need to use the appropriate test statistic such that the theoretical sampling distribution, under the null hypothesis, is known explicitly.
- For permutation testing:
 - ▶ We can choose almost any test statistic.
 - ▶ The sampling distribution of the test statistic under the null hypothesis is generated via simulations.
 - ▶ In fact, we repeatedly generate new permutations of the sample data, and calculate the test statistic for each of them.

Another Case Study of Permutation Testing

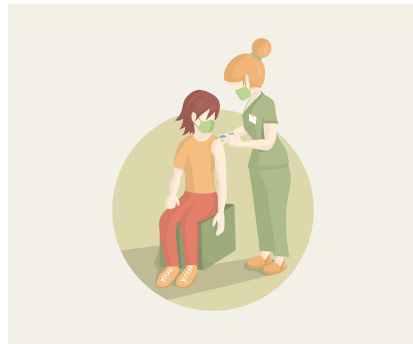
Case Study: Whether the new vaccine is effective?

Story

Suppose a researcher wished to assess if the treatment, a newly developed Covid vaccine, is effective. 48 healthy subjects were randomly assigned into two groups of equal sizes, the “Vaccine” group and the “Control” group, respectively.

Aim of the Study

To determine whether the treatment variable, receiving the vaccine or not, is associated with the outcome variable, the infection status.



Inspect the Dataset

- Let us read in the dataset and check the first few observations.

```
df2 <- read.csv('data/vaccination.csv')  
head(df2)
```

	treatment	outcome
1	Vaccine	Not Infected
2	Vaccine	Not Infected
3	Vaccine	Not Infected
4	Vaccine	Not Infected
5	Vaccine	Infected
6	Vaccine	Not Infected

- Two categorical variables:
 - ▶ “treatment” indicates whether the subject is in the vaccine group, or the control group.
 - ▶ “outcome” shows the infection status of the subject.

Inspect the Dataset

Cont'd

```
str(df2)
```

```
'data.frame':  48 obs. of  2 variables:
```

```
$ treatment: chr  "Vaccine" "Vaccine" "Vaccine" "Vaccine" ...
```

```
$ outcome  : chr  "Not Infected" "Not Infected" "Not Infected" "Not Infected" ...
```

- The two categorical variables are recorded as the “character” type, namely, each value is just a label, with no order between them.
- We also prepare the variables.

```
treatment <- df2$treatment
```

```
outcome <- df2$outcome
```

“table()” Function

- To study the association between the two categorical variables, we can use the base R function, “**table()**”, to generate the contingency table.

```
table1 <- table(df2)
table1
```

	outcome	
treatment	Infected	Not Infected
Control	10	14
Vaccine	3	21

- The test statistic can be defined as the difference of two proportions, by using the proportion of infections within the control group minus that of the vaccine group.

```
original <- table1[1,1]/sum(table1[1,]) - table1[2,1]/sum(table1[2,])
original
```

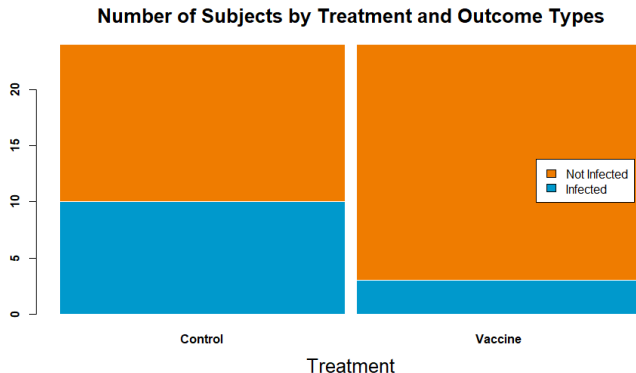
```
[1] 0.2916667
```

- “original” records the test statistic from the original sample, which is 0.292.

Barplot

- To visualise the above difference, we can use the barplot function to generate the stacked bar plot, as follows.

```
barplot(t(table1), xlab="Treatment", legend=colnames(table1),  
        main = "Number of Subjects by Treatment and Outcome Types")
```



Permutation Testing

- ① State the two hypotheses.
 - ▶ H_0 : The population proportions of infected subjects among the two groups are the same.
 - ▶ H_1 : The population proportions of infected subjects among the two groups are different.
- ② Pick the level of significance, α , as 0.05.
- ③ Define the **Test Statistic** as the difference of the two groups' proportions. Next, we will generate the permutation distribution of the test statistic, by simulating multiple permutations.

Define a Function

- We define the following function as one round of simulation.

```
permutation.test2 <- function(treatment, outcome){  
  # Generate a permutation sample  
  treatment_p <- sample(treatment, size= length(treatment), replace=  
    FALSE)  
  # Calculate the test statistic for the permutation sample  
  table_p <- table(treatment_p, outcome)  
  table_p[1,1]/sum(table_p[1,]) - table_p[2,1]/sum(table_p[2,])  
}
```

Define a Function

- We define the following function as one round of simulation.

```
permutation.test2 <- function(treatment, outcome){  
  # Generate a permutation sample  
  treatment_p <- sample(treatment, size= length(treatment), replace=  
    FALSE)  
  # Calculate the test statistic for the permutation sample  
  table_p <- table(treatment_p, outcome)  
  table_p[1,1]/sum(table_p[1,]) - table_p[2,1]/sum(table_p[2,])  
}
```

- The function includes two actions:

Define a Function

- We define the following function as one round of simulation.

```
permutation.test2 <- function(treatment, outcome){  
  # Generate a permutation sample  
  treatment_p <- sample(treatment, size= length(treatment), replace=  
    FALSE)  
  # Calculate the test statistic for the permutation sample  
  table_p <- table(treatment_p, outcome)  
  table_p[1,1]/sum(table_p[1,]) - table_p[2,1]/sum(table_p[2,])  
}
```

- The function includes two actions:
 - ① To use the sample function to generate one permutation sample by reassigning the order of the treatment variable. We will also refer to it as “***permutation by treatment***”.

Define a Function

- We define the following function as one round of simulation.

```
permutation.test2 <- function(treatment, outcome){  
  # Generate a permutation sample  
  treatment_p <- sample(treatment, size= length(treatment), replace=  
    FALSE)  
  # Calculate the test statistic for the permutation sample  
  table_p <- table(treatment_p, outcome)  
  table_p[1,1]/sum(table_p[1,]) - table_p[2,1]/sum(table_p[2,])  
}
```

- The function includes two actions:
 - 1 To use the sample function to generate one permutation sample by reassigning the order of the treatment variable. We will also refer to it as “***permutation by treatment***”.
 - 2 To calculate the test statistic. We first get the contingency table of the permutation sample, and then calculate the difference of the two proportions.

Use the “replicate()” Function to Run Multiple Simulations

- Next, we use the replicate function to generate multiple permutation samples, and collect the test statistics.

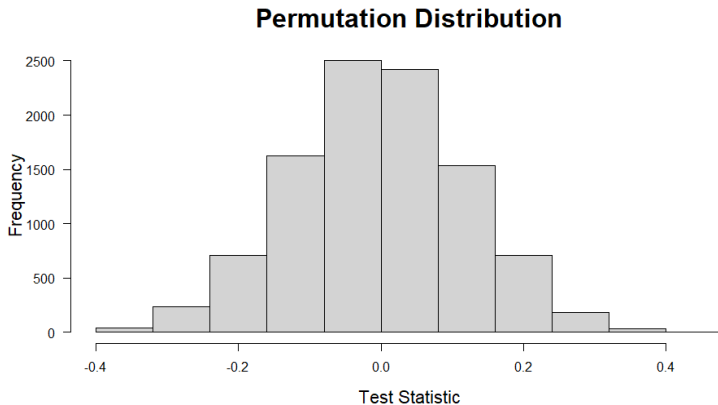
```
test2 <- replicate(10000, permutation.test2(treatment, outcome))
```

- The function, “permutation.test2”, outputs one test statistic value at each time.
- The vector, test2, will collate all the 10,000 test statistic values.

Generate the Permutation Distribution

- We use the hist function to generate the histogram of 10,000 test statistics, which approximates the permutation distribution under the null hypothesis.

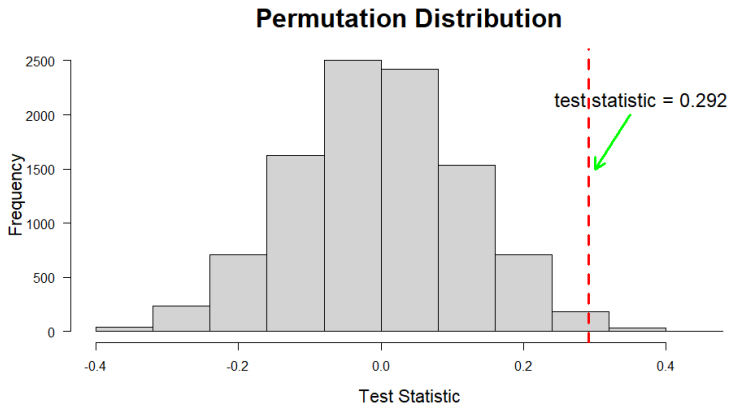
```
hist(test2, breaks = seq(-0.4, 0.5, by = 0.08),  
     main="Permutation Distribution", xlab='Test Statistic')
```



Generate the Permutation Distribution

Cont'd

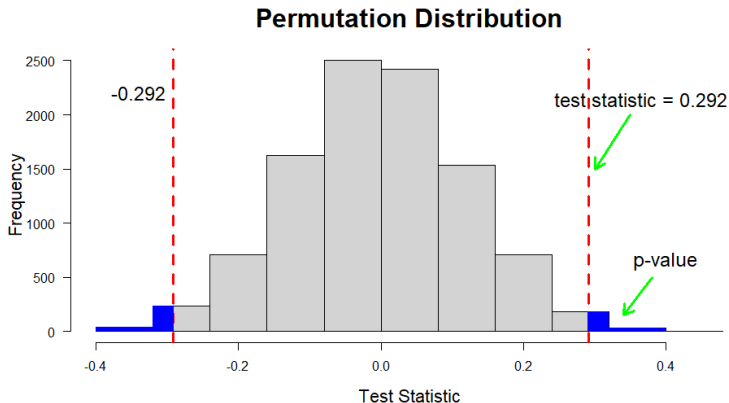
- 4 Calculate the p-value and compare it with the level of significance.
- We mark the original sample statistic, 0.292, in the chart.



Generate the Permutation Distribution

Cont'd

- 4 Calculate the p-value and compare it with the level of significance.
- We mark the original sample statistic, 0.292, in the chart.



Calculate the p-value

- We use the same R code to calculate the two-tailed p-value.

```
pvalue <- mean(abs(test2) >= abs(original))  
pvalue
```

```
[1] 0.0492
```

- The p-value implies that 492 out of the 10,000 simulated permutation samples have a difference of the proportions that is at least 29.2%.
- As $0.0492 < 0.05$, we decide to reject the null hypothesis, and conclude that there is an association between receiving the vaccine and the infection status.

Permutation Testing

Cont'd

- 5 Interpret the decision.
 - ▶ It is beneficial to receive the new vaccine, which shall lead to a lower infection risk.
 - ▶ As the p-value is extremely close to 0.05, domain expertise and other criteria should be considered before rolling out the vaccine.
 - ▶ For instance, it could be worth investigating whether the difference in proportions is consistently significant for each age-group.
- 6 Check the assumption, exchangeability.
 - ▶ It is automatically satisfied since the treatments were randomly allocated to these study participants.

Although we used the difference in proportions here, we could use some other test statistic as well.

- ▶ For example, the ratio of proportions, or the odds ratio.

Calculate the p-value in a Chi-squared Test

- We can also conduct a ***chi-squared test*** to verify the association between two categorical variables.
- It can be implemented in the following way.

```
chisq.test(df2$treatment, df2$outcome)
```

Pearson's Chi-squared test with Yates' continuity correction

data: df2\$treatment and df2\$outcome

X-squared = 3.7978, df = 1, p-value = 0.05132

- Yates' continuity correction is needed when the contingency table is 2 by 2, or when the expected frequency values of some cells are below 10.
- The p-value, 0.051, is slightly larger than the one derived from the permutation test.

Summary

Summary


We have:

- ▶ Summarised the framework of permutation testing.
 - ▶ Applied it to another example in R.
-
- One limitation of hypothesis tests is that the interpretation of a p-value is purely dichotomous.
 - In our example, for instance, the p-values from the two different approaches actually lead to opposite conclusions.

In the next video,

We will introduce an alternative approach, confidence intervals.

References

-  Mine Çetinkaya-Rundel and Johanna Hardin (2021)
Introduction to Modern Statistics