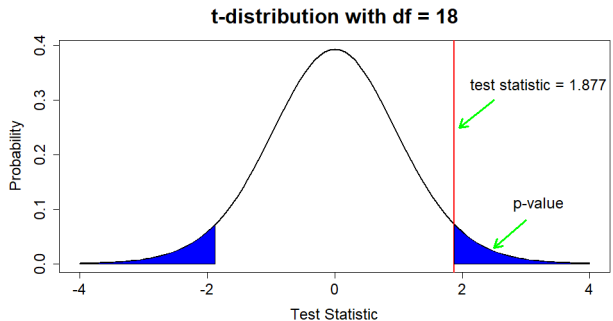
A Recap of Hypothesis Testing and its Pitfalls



Outline

Framework of Hypothesis Testing

1. State the ***Null Hypothesis*** and the ***Alternative Hypothesis***.
2. Pick the ***Level of Significance***.
3. Calculate the ***Test Statistic*** and the ***p-value***.
4. Compare the ***p-value*** with the ***Level of Significance***.
5. Interpret the decision.
6. Check the assumptions.

Learning Objectives

In this video, we will

- Review the steps of the hypothesis testing framework.
- Run the hypothesis test in ***R***.
- Summarise some pitfalls of classical hypothesis testing and discuss the solutions.

Case Study

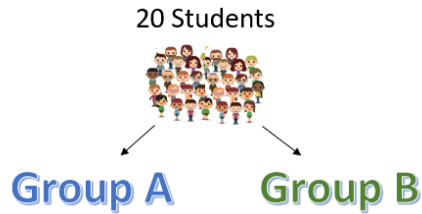
Case Study: Which paper is harder?

Story

Suppose a math professor planned to run a trial exam with two different versions of papers.

Aim of the Study

To evaluate whether the version B paper is more difficult than the version A paper.



Inspect the Dataset

- Read in the dataset and inspect the data structure.

```
df1 <- read.csv('data/math scores.csv')
str(df1)
```

```
'data.frame':  20 obs. of  2 variables:
 $ treatment: chr  "a" "a" "a" "a" ...
 $ outcome  : int  64 66 50 36 43 60 66 59 65 43 ...
```

- 20 observations.
- Two variables: "***treatment***" and "***outcome***".

Inspect the Dataset

Cont'd

- The number 1 after the comma refers to the first column of the data frame, which is also the first variable, "treatment".

```
df1[,1]
```

```
 [1] "a" "a" "a" "a" "a" "a" "a" "a" "a" "a" "b" "b" "b" "b" "b" "b" "b" "b"
[19] "b" "b"
```

- By specifying "1 : 10" before the comma, we will only quote the first 10 observations of "treatment".

```
df1[1:10,1]
```

```
[1] "a" "a" "a" "a" "a" "a" "a" "a" "a" "a"
```

```
df1[11:20,1]
```

```
[1] "b" "b" "b" "b" "b" "b" "b" "b" "b" "b"
```

Inspect the Dataset

Cont'd

- By specifying $1:10$ before the comma and number 2 after the comma, we list the scores of students from group A.

```
df1[1:10,2]
```

```
[1] 64 66 50 36 43 60 66 59 65 43
```

- We use the "mean()" function to check the mean score of each group.

```
mean(df1[1:10,2])
```

```
[1] 55.2
```

```
mean(df1[11:20,2])
```

```
[1] 64.6
```

- The mean score of Group B is 9.4 units higher than that of Group A.

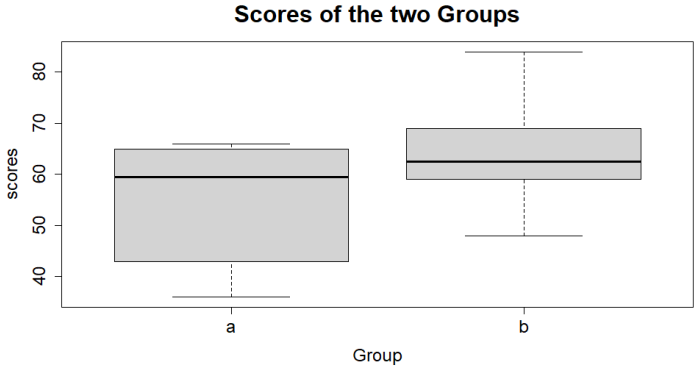
```
mean(df1[11:20,2]) - mean(df1[1:10,2])
```

```
[1] 9.4
```

Boxplot

- To visualise the distributions of the scores between the two groups, we generate the following boxplot.

```
boxplot(outcome~treatment, data = df1, ylab="scores",
        xlab="Group",main="Scores of the two Groups")
```



Hypothesis Testing

1. State the two hypotheses.
   - $\boldsymbol{H}_0$: The population mean scores of the two exam papers are the same.
   - $\boldsymbol{H}_1$: The population mean scores of the two exam papers are different.
2. Pick the level of significance, α, as 0.05.
3. Calculate the ***test statistic*** and the ***p-value***.

Test Statistic

- ***Test statistic*** is a quantity that can be calculated from the sample dataset.
- Test statistic can give us the evidence against the null hypothesis.
- It is natural to define it as the observed difference of the two samples' mean scores, which is 9.4.
- However, we also need to account for the variability in our sample.

Define a New Test Statistic

Test Statistic

$$\text{Test Statistic} = \frac{\text{Observed Difference}}{\text{SE}}$$

- Our new test statistic is the standardised difference of the two samples' mean scores.
- To calculate it, we take the observed difference, 9.4, and divide it by the standard error.
- ***Standard error***, or simply "***SE***", represents the variability of the observed statistic in the sampling distribution.
- ***Sampling distribution*** is the distribution of the test statistic.
- The test statistic follows the ***t-distribution***, under the null hypothesis and some assumptions.

t-distribution

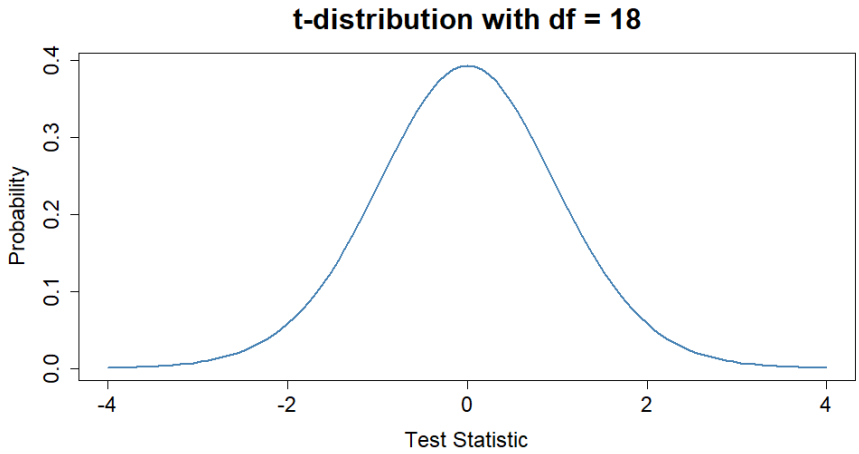
- The t-distribution is just another distribution derived by statisticians.
- The t-distribution has only one parameter, which is called "***degree of freedom***", or simply "***df***".
- In our case, df equals the total number of subjects minus the number of groups, which is $20 - 2 = 18$.
- We can use the following R code to plot the t-distribution.

```
curve(dt(x, df=18), from=-4, to=4)
```

t-distribution

Cont'd

- The t-distribution is similar to normal distribution.



Calculate p-value in R

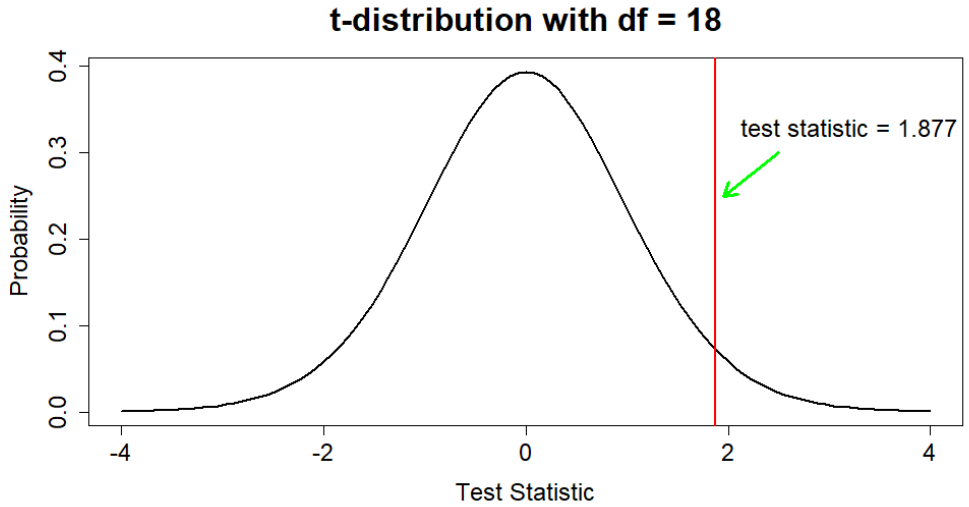
- In ***R***, we can use the function "***t.test()***" to calculate the test statistic and the p-value for two sample t-tests.

```
t.test(outcome~treatment, alternative = "two.sided", paired=F,
var.equal=T, data = df1)
```

```
Two Sample t-test
data:  outcome by treatment
t = -1.8767, df = 18, p-value = 0.07687
alternative hypothesis: true difference in means between group a and group b
is not equal to 0
95 percent confidence interval:
 -19.923268   1.123268
sample estimates:
mean in group a      mean in group b
           55.2                64.6
```

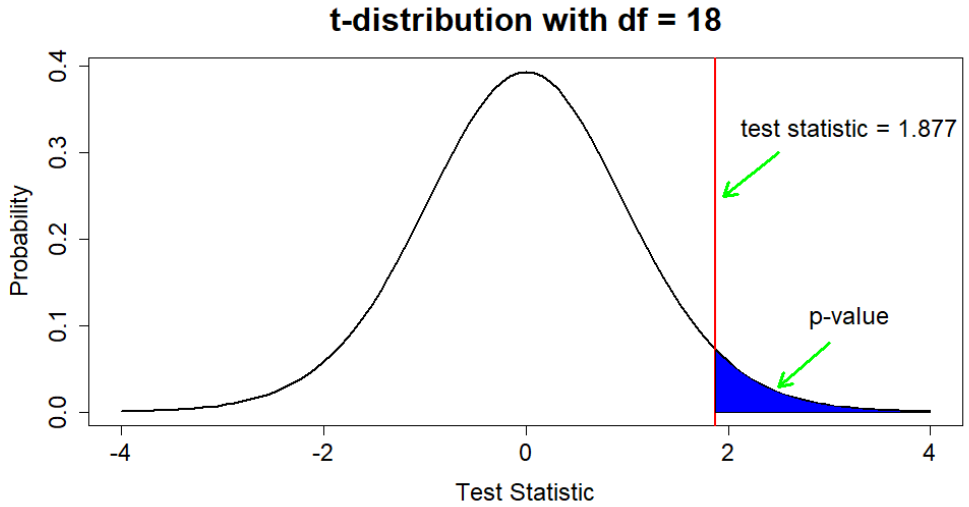
Test Statistic and p-value

- To compare group B with group A, the test statistic is 1.877.



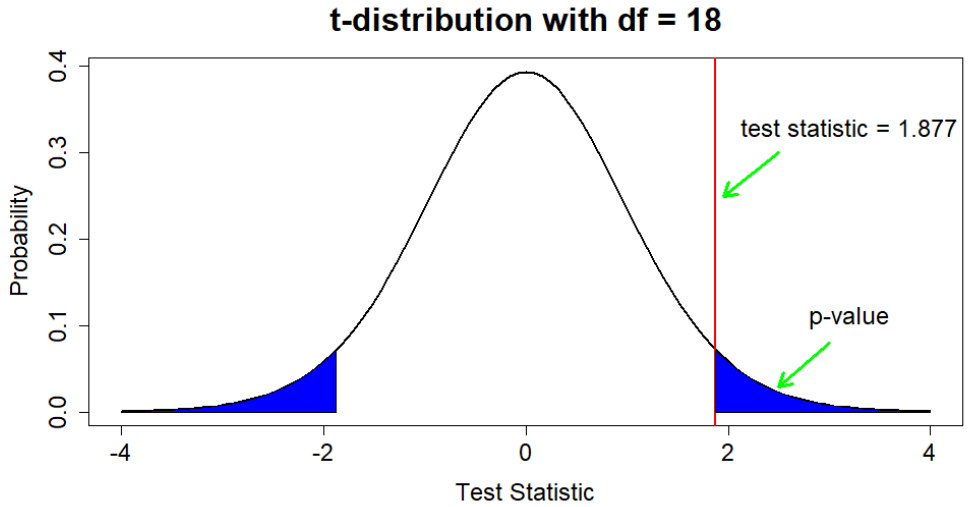
Test Statistic and p-value

- To compare group B with group A, the test statistic is 1.877.



Test Statistic and p-value

- To compare group B with group A, the test statistic is 1.877.



Hypothesis Testing

4. Compare the p-value with level of significance.
5. Interpret the decision.
   - As p-value > 0.05, we do not reject the null hypothesis and conclude that there is insufficient evidence to prove the two exam papers have different difficulty level.
6. Check the assumptions.

Check Assumptions

Assumptions of two sample t-tests

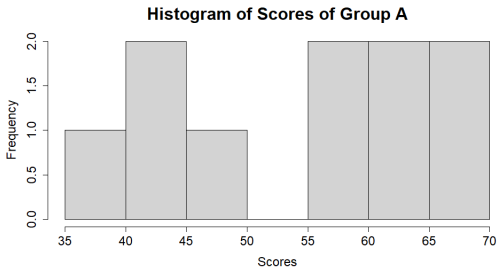
1. ***Numerical scale***: The dependent (outcome) variable is numerical.
2. ***Independence***: The observations are independent within and between the two groups.
3. ***Normality***: The population data follow normal distribution.

- The first assumption is satisfied as the outcome, math scores, is numerical.
- The second assumption is satisfied as the dataset comes from two randomly assigned groups.
- To check normality, let us take a look at the histogram and the Q-Q plot of the scores for each group.

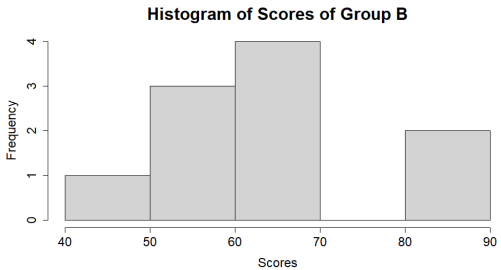
Check Histograms

- To check normality, we first look at the histogram of the scores for each group.

```
hist(df1[1:10,2], main="
   Histogram of Scores of
   Group A", xlab='Scores')
```




```
hist(df1[11:20,2], main="
   Histogram of Scores of
   Group B", xlab='Scores')
```

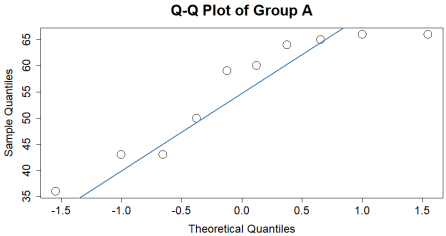


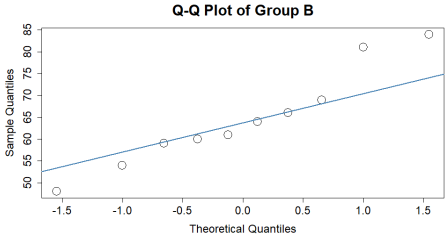
Check Q-Q Plots

- We also generate the Q-Q plots, as introduced in an earlier video.

```
qqnorm(df1[1:10,2])
qqline(df1[1:10,2], lwd =
    2)
```

```
qqnorm(df1[11:20,2])
qqline(df1[11:20,2], lwd =
    2)
```





Check Assumptions

- For histograms, the group B's plot may be approximately normal, but it is hard to tell whether the group A's plot is normal or not.
- For Q-Q plots, there are few points that are not so close to the reference line.
- There is no clear outlier.
- Based on the above plots, it is difficult to say for sure if the normality assumption is valid.
  - In such situations, we don't know how much it may affect the accuracy of the p-value.

Pitfalls of Classical Hypothesis Testing

Pitfalls of Classical Hypothesis Testing

1. It is generally a hard job to know the distribution of any test statistic.
   - In our case study, we already knew the proposed test statistic followed a t-distribution, when comparing the two samples' means.
   - In other applications, we may wish to compare median, or compute the ratio of the means.
   - How would we know what distribution the test statistic follows?

2. The assumptions may not be fully satisfied, which may affect the accuracy of p-value.
   - In general, a small sample size makes it harder to verify some assumption, say, normality.

3. The interpretation of a p-value is purely dichotomous, which may not always be appropriate.
   - For example, if the p-value is 0.0499 or 0.0501, a small variation would lead to the opposite conclusion.

4. Statistical significance doesn't mean practical significance.
   - In general, p-value should not be the sole criterion for decision making.

Some Possible Solutions

- For the 2nd issue, we can implement some other tests, which require less assumptions.
- For the 3rd and the 4th issue, we will introduce an alternative approach: Confidence interval in another video.
  - The confidence interval will provide a plausible range of values, based on which the domain expert can make some data-informed decision.

Summary

Summary

We have:

- Reviewed the classical hypothesis testing and learned how to implement it in R.
- Discussed the pitfalls of the classical hypothesis testing.

In the next video,

We would like to introduce a new approach: Permutation-based hypothesis testing.

- It requires no prior knowledge of the distribution of test statistics.

References


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


Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen (2002)
The Importance of the Normality Assumption in Large Public Health Data Sets
https://www.annualreviews.org/doi/pdf/10.1146/annurev.publhealth.23.100901.140546