



EBU5601

Data Design

Hypothesis Testing

Dr Chao Shu

School of Electronic Engineering and Computer Science
Queen Mary University of London
Nov. 2024

Learning Outcomes

- The main outcomes are:
 - [L07.1] Successfully set up hypothesis tests based on the questions to be investigated on the datasets
 - [L07.2] Understand the different types of errors in a hypothesis test
 - [L07.3] Determine the suitable type(s) of hypothesis test to be used for a given scenario
 - [L07.4] Perform one-sample z-tests or t-tests on collected datasets
 - [L07.5] Apply bootstrapping in one-sample tests
 - [L07.6] Perform two-sample t-tests on collected datasets
 - [L07.7] Apply bootstrapping in two-sample tests
 - [L07.8] Perform the permutation test for two-sample tests

Learning Outcomes

- The main outcomes are:

- [LO7.9] Correctly interpret p-values to draw correct conclusions

不考 [LO7.10] Understand the impact of a large sample size on hypothesis tests

[LO7.11] Differentiate statistical significance and practical significance

INTRODUCTION


Introduction


A/B Testing



Data Scientist

Monzo Bank · London, England, United Kingdom (Remote)

 Full-time · Entry level

 1,001–5,000 employees · Banking

About the job

 Location | UK Remote | London |  £65,000 - £75,000 + Benefits

Job Descriptor

- Applying your skills in quantitative analysis, data mining, and the presentation of data to see beyond the numbers and understand how our users interact with our products and how those insights can inform our product strategy
- Guide and enable product teams to measure things that matter; **initiate or help run A/B experiments** to keep improving everything we do
- Drive together with the finance team a unified company-wide understanding of the lifetime value of our users and how different product features are impacting user profitability
- Liaise with engineers to keep making sure we collect the right data to produce relevant business insights

Source: LinkedIn

Introduction

A/B Testing

- In 2013, Electronic Arts (EA) released SimCity 5
- They wanted to increase pre-orders of the game
- They used A/B testing to test different advertising scenarios
- This involves splitting users into control and treatment groups



Image Source: <https://venturebeat.com/business/electronic-arts-lays-off-350-people-the-most-under-ceo-andrew-wilson/>

Introduction

A/B Testing

- The treatment group (no ad) got 43.4% more purchases than the control group (with ad)
- The intuition that "showing an ad would increase sales" was false
- Was this result statistically significant or just by chance?



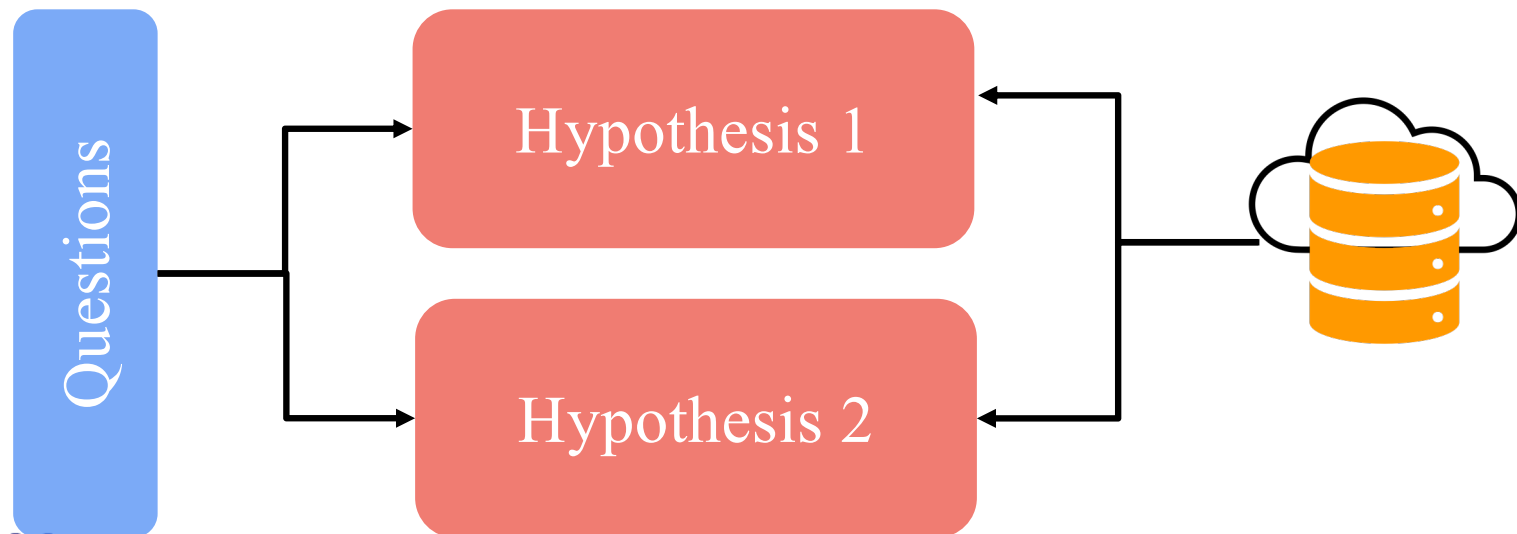
Control



Treatment

Introduction

- A statistical **hypothesis test** is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters.
 - translate a question into hypotheses.
 - collect data to justify which hypothesis is likely to be true.
- Hypothesis testing and confidence intervals allow for the **use of sample data** to draw conclusions about **an entire population**.



Introduction

Example:

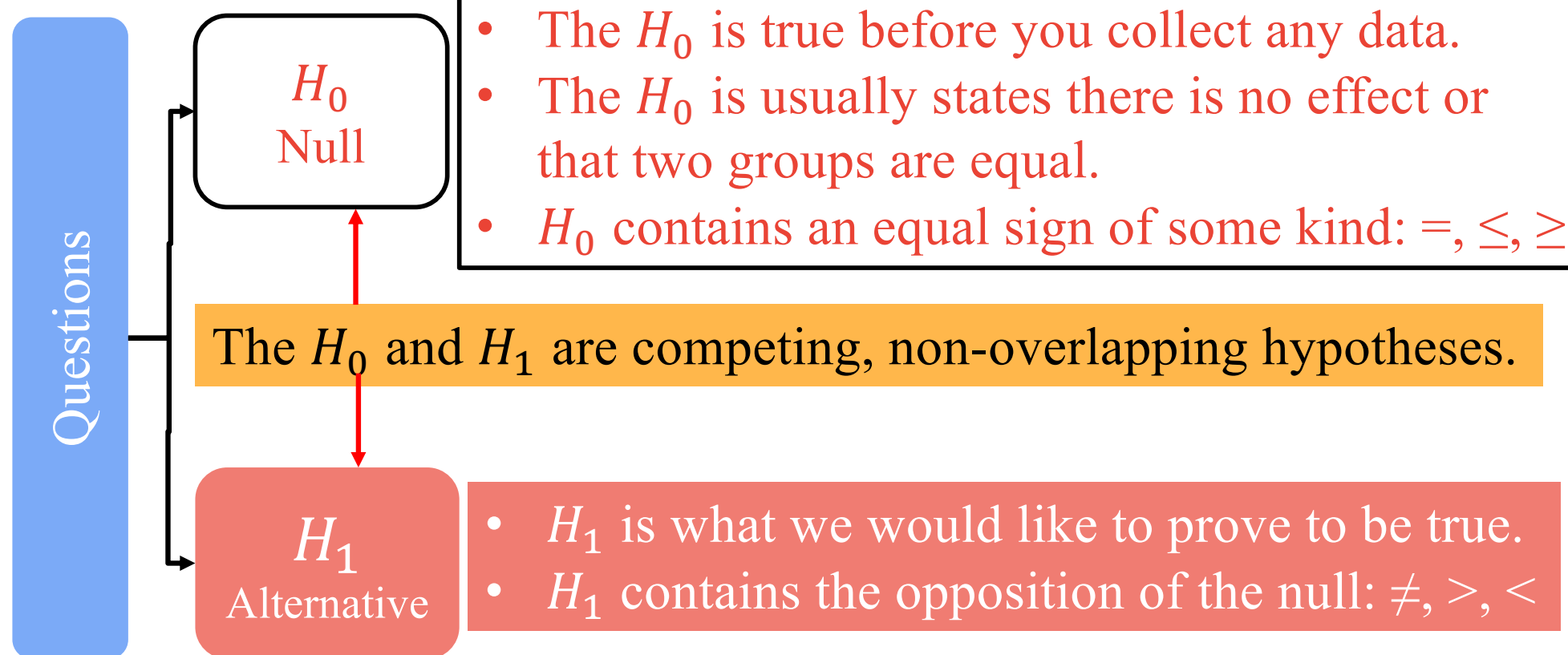
- Imagine you and your girlfriend are in a debate about what the most popular ice cream flavour is in the world.
- Hypothesis:
 - H_0 : Chocolate is the most popular flavour
 - H_1 : Vanilla is the most popular flavour
- How can we truly know this unless we talk to everyone in the world? How do we know if our conclusions are reliable?
- Use hypothesis testing to draw conclusions about a population only using sample data



SETTING UP HYPOTHESIS TESTS

Setting Up Hypothesis Tests

- A few general rules for setting up null and alternative hypotheses



Example: • Innocent until proven guilty

- H_0 : Innocent
- H_1 : Guilty

Setting Up Hypothesis Tests

Exercise:

- Imagine you create a new webpage layout and you want to know if this new page attracts more traffic than the existing page. How do you set up the null and alternative hypotheses?

Go to
www.menti.com

Enter the code

1939 7765



Or use QR code

Types of Errors

Discussion: Why the null and alternative hypotheses matter?

- Two types of errors are possible in hypothesis testing:
 - Type I Errors** and **Type II Errors**.
 - Type I Errors are usually considered the worst type of error.

Just based on this information, can you guess which truth/decision combo is for each type of error?

		Truth	
		Guilty	Innocent
Decision	Guilty		
	Innocent		

Go to
www.menti.com

Enter the code

1939 7765



Or use QR code

Types of Errors

- **Type I Error**

- The mistaken rejection of a null hypothesis that is actually true as the result of a test procedure. (Deciding the alternative (H_1) is true when actually (H_0) is true.)
- Also known as a **"false positive"**
- A type I error rate is denoted by α
- Example: an innocent person is convicted

- **Type II Error**

- The failure to reject a null hypothesis that is actually false as the result of a test procedure. (Deciding the null (H_0) is true when actually (H_1) is true.)
- Also known as a **"false negative"**
- A type II error rate is denoted by β
- Example: a guilty person is not convicted



Error Rate

- You should set up your null and alternative hypotheses so that the worst of your errors is the type I error.
- The **type I error rate** is the probability of rejecting the null hypothesis given that it is true. The test is designed to keep the type I error rate below a prespecified bound called the **significance level**, usually denoted by α and is also called the alpha level. Usually, the $\alpha = 0.05$ (5%) for research/business application, implying that it is acceptable to have a 5% probability of incorrectly rejecting the true null hypothesis (medical 1%).
- The **type II error rate** is denoted by the β and related to the power of a test, which equals $1 - \beta$
- These two types of error rates are traded off against each other: for any given sample set, the effort to reduce one type of error generally results in increasing the other type of error.

Summary

- Table of error types

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Fail to reject	Correct inference (true negative)(probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive)(probability = $1-\beta$)

Image source: https://en.wikipedia.org/wiki/Type_I_and_type_II_errors#Type_I_error

Types of Errors

Case Study: A new market campaign

- Consider we are interested in testing if the average revenue for a new marketing campaign is better than the previous marketing campaign.
- We set up the null and alternative hypotheses in the following way:
 - $H_0: \mu_n - \mu_o \leq 0$
 - $H_1: \mu_n - \mu_o > 0$
- where μ_o is the mean revenue of the old campaign, and μ_n is the mean revenue of the new campaign.
- The null hypothesis is that the mean revenue of the new campaign is less than or equal to the mean revenue of the old campaign.
- The alternative hypothesis is that the mean revenue of the new campaign is larger than the mean revenue of the old campaign.




Type I Error: Deciding the new page is better, but really the old page is better.

Type II Error: Deciding the old page is better, but really the new page is better.

Types of Errors

Case Study: Skydiving Business (An extreme case)

- Imagine you own a skydiving business. You must check parachutes to ensure they work correctly. There are two potential outcomes, either a parachute works or it doesn't. You check each parachute to make a decision.
- If **you determine** it doesn't work, you throw it out.
 - The parachute **actually** doesn't work, that is great.
 - The parachute **actually** works, you lost £50
- If **you determine** it works, you keep it for skydiver to use.
 - The parachute **actually** works, the skydiver can land safely
 - The parachute **actually** doesn't work, worst type of error possible.

		Truth	
		Works	Fails
Decision	Works		
	Fails	£-50	 £

Discussion: Is $\alpha=1\%$ enough?

Common Types of Hypothesis Tests

- Testing a population parameter
 - e.g., test whether or not the average time playing games per week for all JP students is 10 hours given survey data on the game-playing time of JP students (two-tailed test).
 - $H_0: \mu = 10$
 - $H_1: \mu \neq 10$
 - Or, test if the average time playing games per week for all JP students is more than 10 hours per week (one-tailed test).
 - $H_0: \mu \leq 10$
 - $H_1: \mu > 10$
 - One sample t-test/z-test.

Common Types of Hypothesis Tests

- Testing the difference of parameters in different population
 - test whether or not the average time playing games per week for all male JP students is the same as that for all female JP students. (two-tailed test)
 - $H_0: \mu_{boy} = \mu_{girl}$
 - $H_1: \mu_{boy} \neq \mu_{girl}$
 - Or, test if the average time playing games per week for all male JP students is longer than that for all female JP students (one-tailed test).
 - $H_0: \mu_{boy} \leq \mu_{girl}$
 - $H_1: \mu_{boy} > \mu_{girl}$
 - Two-sample t-test/z-test

Common Types of Hypothesis Tests

- Testing the difference before and after some treatment on the same individual (Paired t-test)
 - e.g, test if the average time playing games per week for male students reduces after they get a girlfriend.
 - H_0 : There is no significant difference in game-playing time before and after each male student gets a girlfriend ($\mu_d = 0$, where μ_d represents the population mean of the differences).
 - H_1 : There is a significant difference in game-playing time before and after each male student gets a girlfriend ($\mu_d \neq 0$, where μ_d represents the population mean of the differences).
 - Paired t-test

There are a lot of different hypothesis tests

Hypothesis tests are always on population parameters, never on statistics

ONE-SAMPLE TESTS

One-Sample Tests

- A **one-sample test** is a statistical hypothesis test used to determine whether the **mean of a single sample** is significantly different from a known or **hypothesized population mean**.
 - It's commonly used when you have a sample of data, and you want to assess whether that sample provides enough evidence to reject a null hypothesis about the population mean.
- The commonly used methods for one-sample tests are **one-sample z-test** and **one-sample t-test**.

One-Sample z-Tests

Example: Sleep Hours of University Students

- Suppose we've run a survey among university students collected some data about their university life.
 - 76 students, 22 types of information
- In this example, we'll focus on the amount of sleep university students get.
 - The sleep hour data of one student is missing.

ID	Gender	Classification	Height	Shoe Size	Phone Time	# of Shoes	Birth order	Pets	Happy	Funny	College	Bfast Calories	Exercise	Stat Pre	Stat Post	Phone Type	Sleep	Social Media	Impact of SocNetworking	Political	Animal	Superhero
1	male	senior	67.75	7	12	12	youngest	5	0.8	7	Natural Sciences	500	360	3		iPhone	7	180	worse	Democrat	Dog person	Batman
2	male	freshman	71	7.5	1.5	5	middle	4	0.75	8	Natural Sciences	0	200	9		Android smartphone	7	20	better	Democrat	Dog person	Batman
3	female	freshman	64	6	25	15	oldest	8	0.9	6	Natural Sciences	200	30	7	5	Android smartphone	8	60	better	Republican	Dog person	Batman
4	female	freshman	63	6.5	30	30	middle	12	0.98	9	Education	200	180	6	7	iPhone	6	60	better	Republican	Both	Superman
5	male	senior	69	6.5	23	8	oldest	4	0.75	6	Natural Sciences	0	180	4	7	iPhone	5.5	60	worse	Independent	Dog person	Superman
6	female	senior	64	8.5	13	25	oldest	1	0.95	5	Natural Sciences	250	310	7	7	iPhone	6.5	90	no impact	Democrat	Dog person	Batman
7	female	freshman	62	8.5	23	12	oldest	2	0.95	7	Nursing	200	60	7	8	Android smartphone	7	120	better	Republican	Both	Superman
8	female	freshman	64	6	50	50	youngest	10	0.9	4	Liberal Arts	200	0	5	7	iPhone	7	60	no impact	Independent	Both	Superman
9	female	freshman	66	8	10	15	youngest	25	0.9	9	Natural Sciences	0	0	8	6	iPhone	7	3	no impact	Democrat	Both	Batman
10	female	freshman	68	6.5	40	20	oldest	4	0.95		Nursing	150	240	6	10	iPhone	6	180	better	Democrat	I don't like either	Batman

Data Source: <http://sites.utexas.edu/sos/guided/inferential/numeric/claim/one-sample-t/>

One-Sample z-Tests

Example: Sleep Hours of University Students

One-sample two-tailed z-test

- **Step 1: Set up a hypothesis.**
- Suppose we are interested to know whether university students on average get as much sleep as adults get or not. We know from a report that adults get 7.25 hours of sleep on average. We can set up the hypothesis test as follows.

$$H_0: \mu = 7.25$$

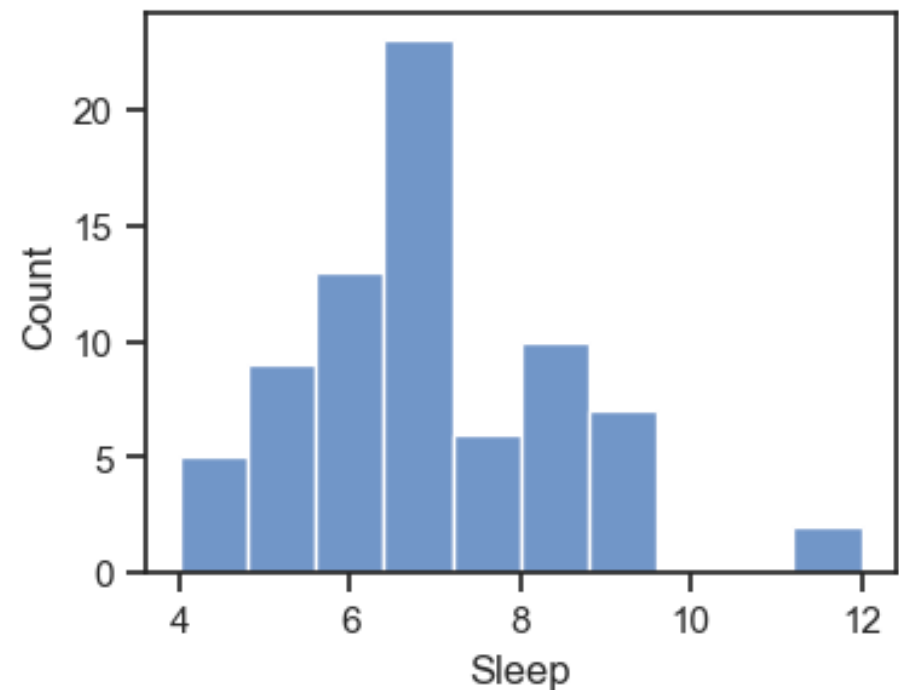
$$H_1: \mu \neq 7.25$$

One-Sample z-Tests

Example: Sleep Hours of University Students

One-sample two-tailed z-test

- **Step 2: Collect a sample and find out the summary statistics of the sample**
- Since we already have the data, let's explore the sample data.
 - Sample size = 75
 - Sample mean = 6.873
 - Sample std = 1.538
 - Sleep hours approximately follow the normal distribution



One-Sample z-Tests

Example: Sleep Hours of University Students

One-sample two-tailed z-test

- **Step 3: Define a significance level**
- Before we start to perform the hypothesis we just set up, we need to define a significance level to indicate the max Type I error we can accept. In this case, we set

$$\alpha = 0.05$$

One-Sample z-Tests

Example: Sleep Hours of University Students

One-sample two-tailed z-test

- **Step 4: Generate the sampling distribution of the statistic of interest under the null hypothesis**
- In a hypothesis test
 - Assume the null to be true.
 - Estimate how likely a sample would have a mean just like the mean of our sample when the null hypothesis is true.
 - If it is very unlikely that a sample would have a mean like the mean of our sample, we would think the null probably is not true (reject the null in favour of the alternative).
 - Otherwise, we think null would be true (fail to reject the null).
- In order to make this decision, we need to generate the sampling distribution of the sample mean under the null hypothesis

One-Sample z-Tests

Example: Sleep Hours of University Students

One-sample two-tailed z-test

- **Step 4a: Generate the null distribution (z-test)**
- Usually, we consider the sample size > 30 to be large. In this case, we could use s (sample standard deviation) as **an estimate** of the population standard deviation σ

$$SE \approx \frac{s}{\sqrt{n}}$$

- Based on the CLT, the sampling distribution of the mean under the null hypothesis (**null distribution**) will approximately be:

$$N(\mu_0, SE^2)$$

where μ_0 is the mean **under the null hypothesis**.

One-Sample z-Tests

Example: Sleep Hours of University Students

One-sample two-tailed z-test

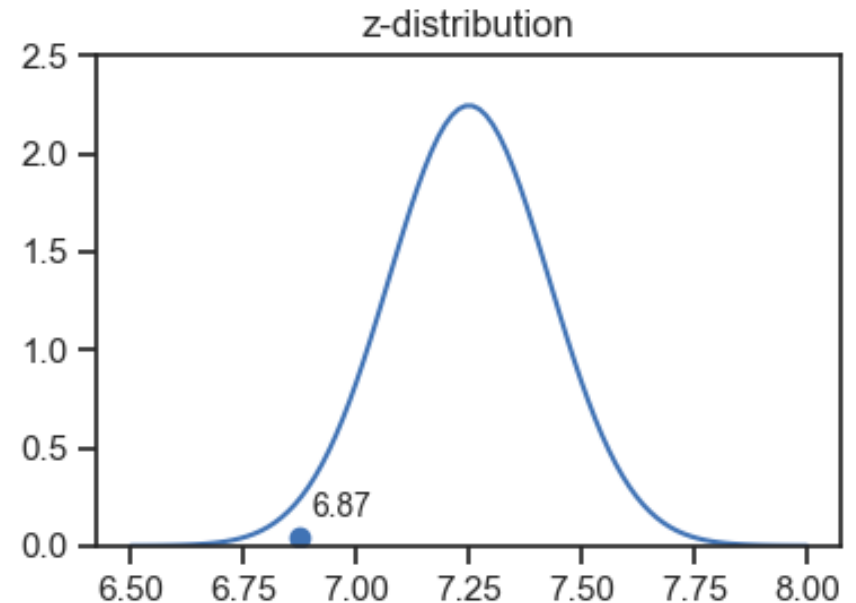
- **Step 4a: Generate the null distribution (z-test)**

- In this example

$$SE \approx \frac{s}{\sqrt{n}} = \frac{1.538}{\sqrt{75}} = 0.1776$$

- Based on the CLT, the sampling distribution of the mean under the null hypothesis (**null distribution**) will approximately be:

$$N(7.25, 0.1776^2)$$

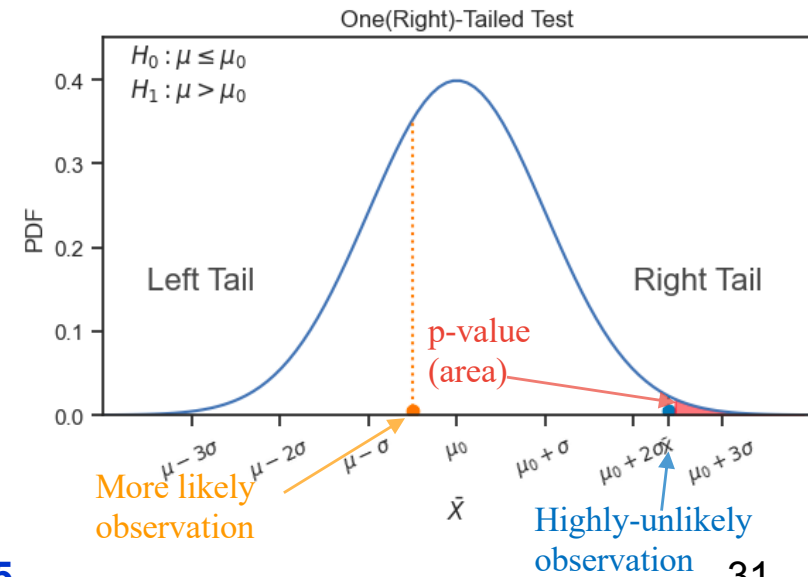
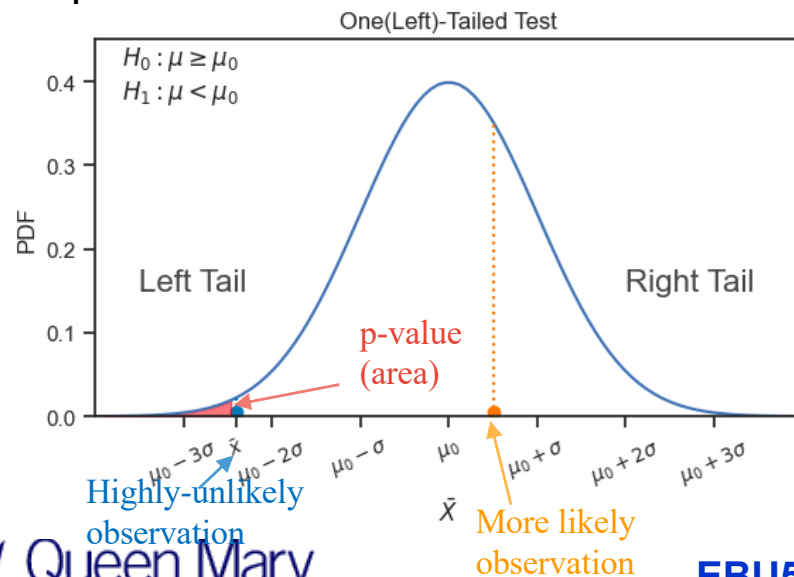
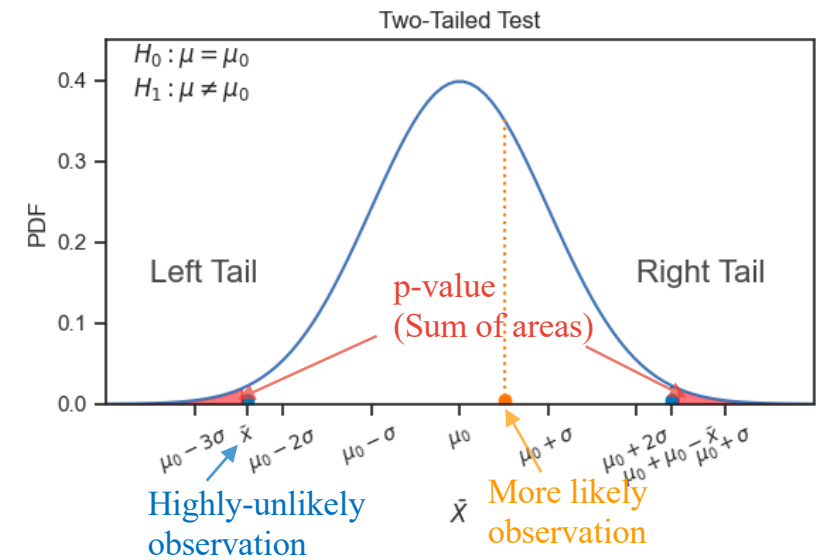


***Discussion:** How do we determine how likely a random sample would have a mean of 6.87 when null is true?*

p-Value

Definition

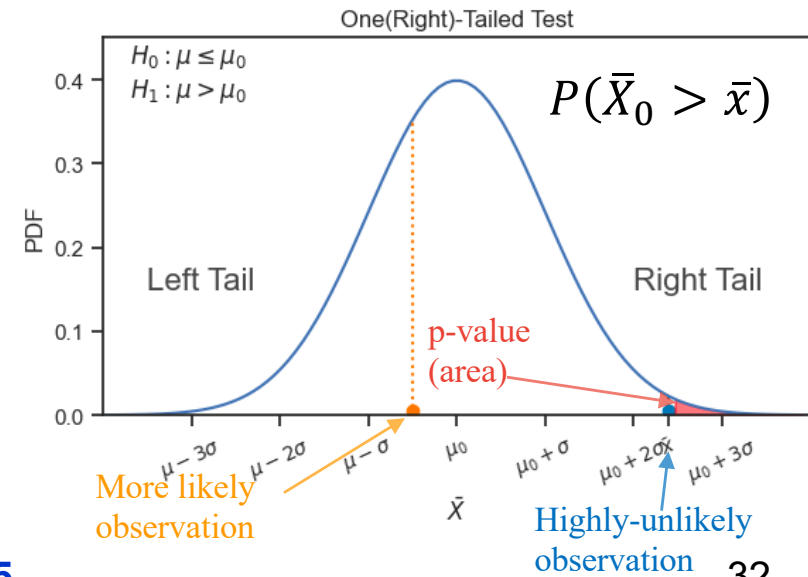
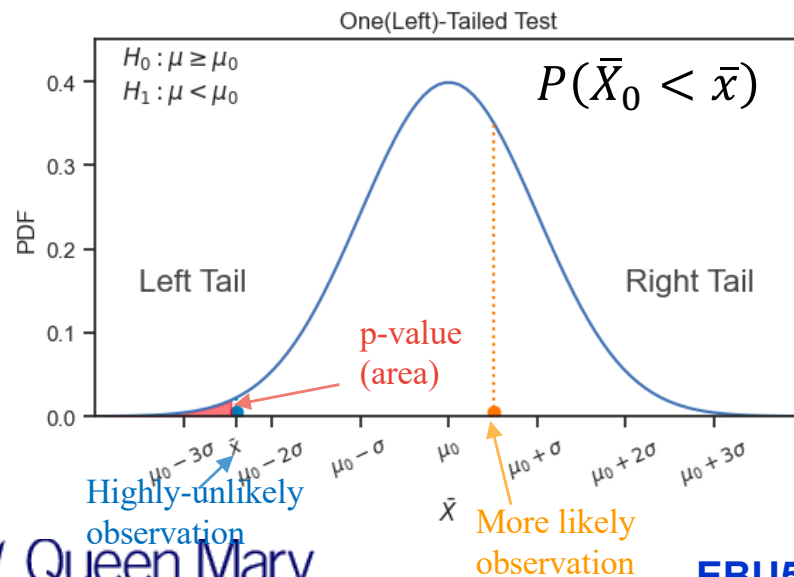
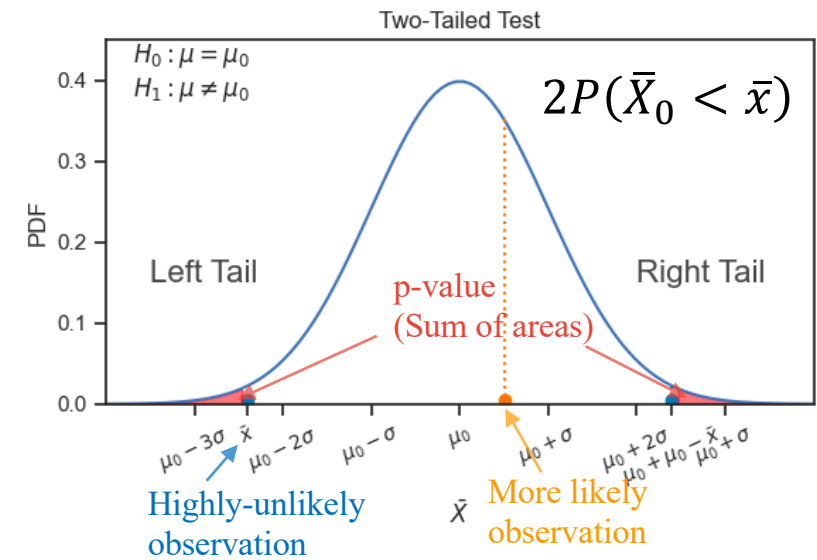
- The probability, assuming H_0 is true, that the test statistic (e.g., sample mean) would take a value as extreme or more extreme than that actually observed (i.e., more extreme in favour of the alternative hypothesis) is called the **p-value** of the test.
 - “Extreme” means “far from what we would expect if H_0 were true.”



p-Value

Calculation

- After the sampling distribution under the null (null distribution) is generated, the p-value can be calculated as shown in figures, where \bar{X}_0 is the random variable that follows the null distribution, \bar{x} is the sample mean.
 - Large p-value, large support for H_0
 - Small p-value, strong evidence against H_0



p-Value

- A small p-value suggests it is less likely to observe our statistic from the null, i.e., our statistic is more likely to have come from the alternative than the null.
- When the p-value is large, we have evidence that our statistic was likely to come from the null hypothesis. Therefore, we do not have evidence to reject the null.
- **Make decisions/Draw Conclusions**
 - if $p_value \leq \alpha$, we **reject null in favour of the alternative**
 - if $p_value \geq \alpha$, we **fail to reject null**

One-Sample z-Tests (Cont.)

Example: Sleep Hours of University Students

One-sample two-tailed z-test

- **Step 5a: Calculate the p-value (z-test)**

- Null distribution $N(7.25, 0.1776^2)$, in this two-tailed test

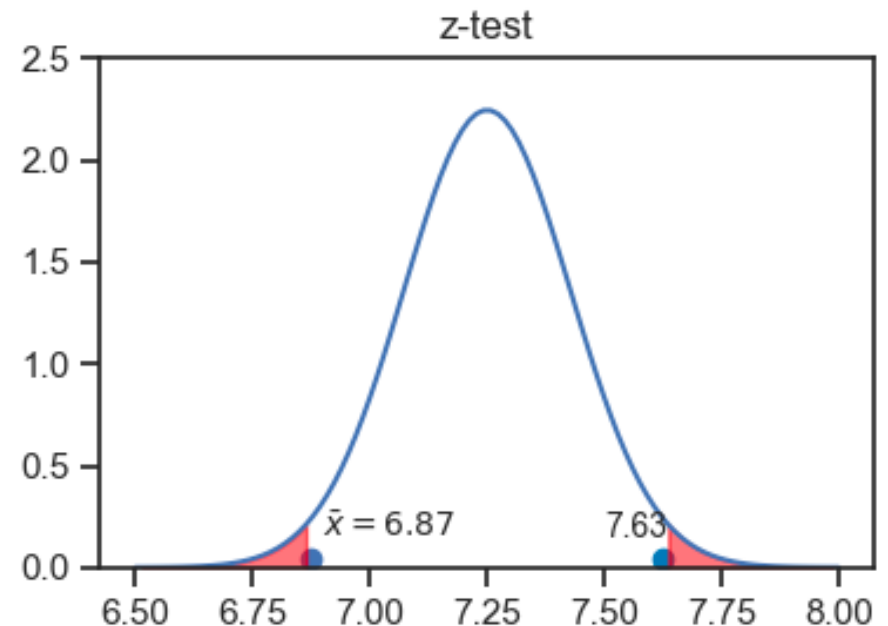
$$\text{p-value} = 2P(\bar{X} < 6.873).$$

- z-score

$$z = \frac{\bar{x} - \mu_0}{SE} = \frac{6.873 - 7.25}{0.1776} = -2.12$$

- By referring to the z-table, we can get $P(Z < -2.12) = 0.017$, so:

$$\text{p-value} = 2 * 0.017 = 0.034$$

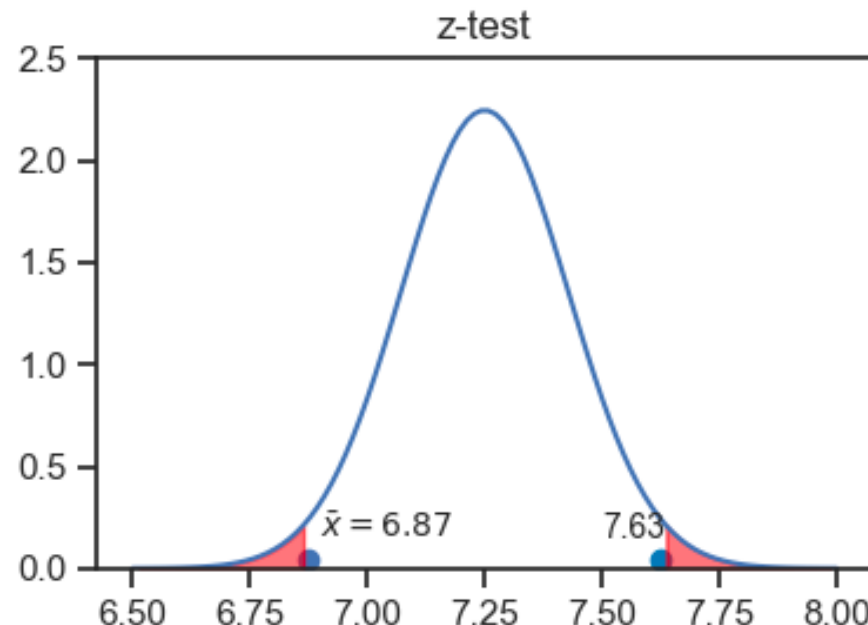


One-Sample z-Tests (Cont.)

Example: Sleep Hours of University Students

One-sample two-tailed z-test

- **Step 6: Draw conclusions**
- Because $p\text{-value} = 0.034 < \alpha (=0.05)$, we **reject the null in favour of the alternative**, i.e., there is enough evidence supporting the claim that the number of hours of sleep university students get is different from 7.25 hours.



One-Sample z-Tests (Cont.)

Example: Sleep Hours of University Students

One-sample one-tailed z-test

- **Step 1: Set up a hypothesis.**
- Suppose we now want to test if on average university students get **less** sleep than adults do. We can set up the hypothesis test as follows.

$$H_0: \mu \geq 7.25$$

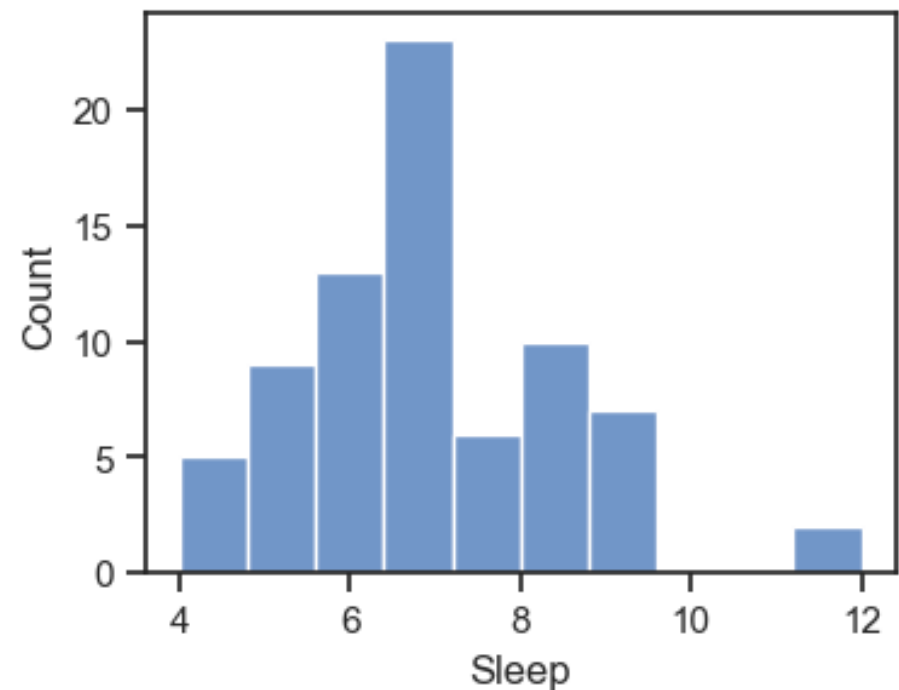
$$H_1: \mu < 7.25$$

One-Sample z-Tests (Cont.)

Example: Sleep Hours of University Students

One-sample one-tailed z-test

- **Step 2: Collect a sample and find out the summary statistics of the sample**
- We already have the data and the summary statistics.
 - Sample size = 75
 - Sample mean = 6.873
 - Sample std = 1.538
 - Sleep hours approximately follow the normal distribution



One-Sample z-Tests (Cont.)

Example: Sleep Hours of University Students

One-sample one-tailed z-test

- **Step 3: Define a significance level**
- We keep the significance level the same as the last demonstration
 $\alpha = 0.05$

One-Sample z-Tests (Cont.)

Example: Sleep Hours of University Students

One-sample one-tailed z-test

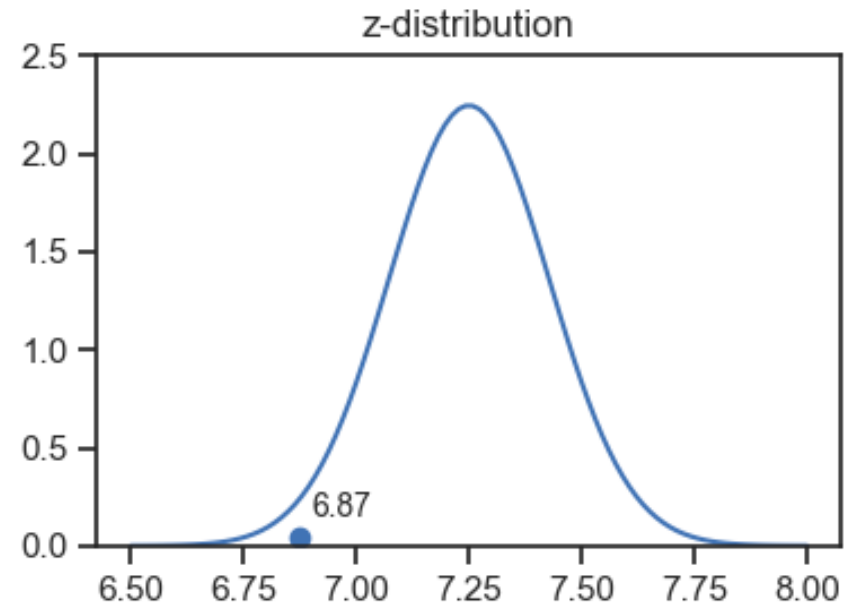
- **Step 4a: Generate the null distribution (z-test)**

- In this example

$$SE \approx \frac{s}{\sqrt{n}} = \frac{1.538}{\sqrt{75}} = 0.1776$$

- Based on the CLT, the sampling distribution of the mean under the null hypothesis (**null distribution**) will approximately be:

$$N(7.25, 0.1776^2)$$



One-Sample z-Tests (Cont.)

Example: Sleep Hours of University Students

One-sample one-tailed z-test

- **Step 5a: Calculate the p-value (z-test)**

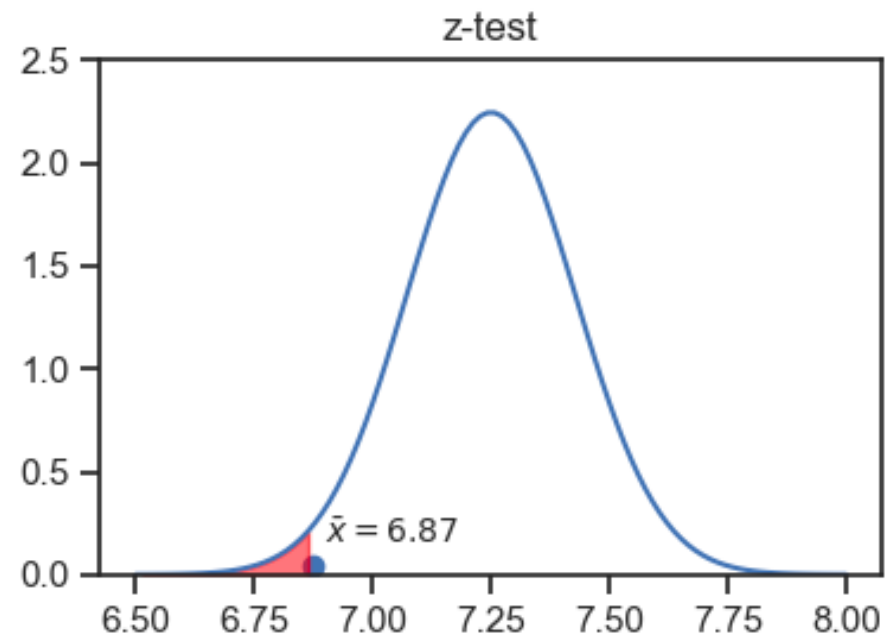
- Null distribution $N(7.25, 0.1776^2)$, in this one-tailed test

$$\text{p-value} = P(\bar{X} < 6.873).$$

- z-score

$$z = \frac{\bar{x} - \mu_0}{SE} = \frac{6.873 - 7.25}{0.1776} = -2.12$$

- By referring to the z-table, we can get $P(Z < -2.12) = 0.017$, so:
p-value=0.017



One-Sample z-Tests (Cont.)

Example: Sleep Hours of University Students

One-sample one-tailed z-test

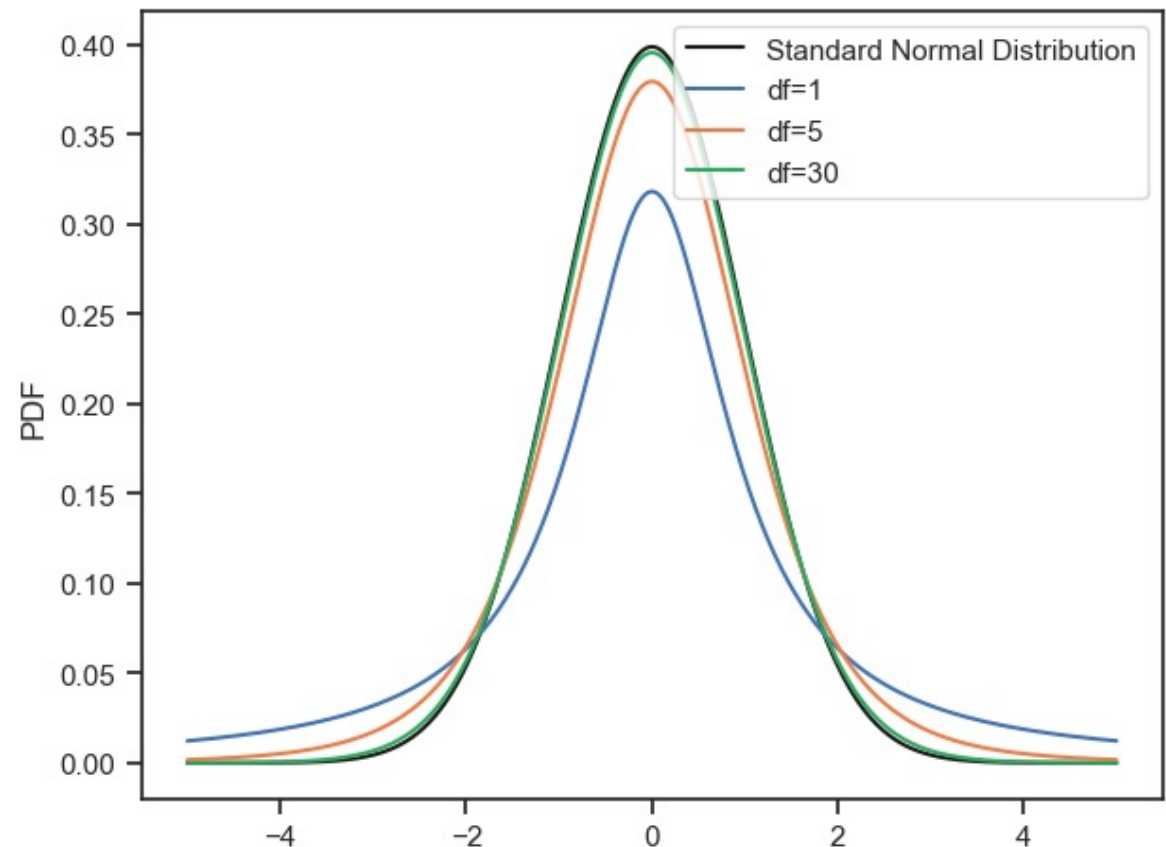
- **Step 6: Draw conclusions**
- Because $p\text{-value} = 0.017 < \alpha (=0.05)$, we reject the null in favour of the alternative, i.e., there is enough evidence supporting the claim that the number of hours of sleep university students get is less than 7.25 hours.

Recap: Student's t-Distribution

- Bell-shaped with a shorter peak and thicker tails. Designed to factor in the greater uncertainty associated with small sample sizes.
- The t-distribution describes the variability of the distances between sample means and the population mean **when the population standard deviation is unknown and the sample data approximately follow the normal distribution.**
- This distribution has only one parameter, the **degrees of freedom**, based on (but not equal to) the sample size.

Recap: Student's t-Distribution

- t-distributions with smaller degrees of freedom have thicker tails and lower peaks. At around 30 degrees of freedom, the t-distribution closely approximates the standard normal distribution (z-distribution).
- Extensively used in depicting distributions of sample statistics. You should use the t-distribution when you need to assess the mean and do not know the population standard deviation.
- It's particularly important to use it when you have a small ($n < 30$) sample size.



Recap: Student's t-Distribution

- From the CLT, we know that the sampling distribution of the sample mean follows a normal distribution

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- However, most of the time, the population standard deviation σ is unknown. If the sample size is larger enough, we consider the sample standard deviation s is a good estimate of σ

$$Z \approx \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1), \text{ if } n > 30$$

- If the sample size is small ($n < 30$), $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ will not follow the standard normal distribution, but will follow a t-distribution with a particular degree of freedom ν .

$$\bar{X} \sim \text{lst}(\mu, \tau^2, \nu)$$

where $T = \frac{\bar{X} - \mu}{\tau} \sim t_\nu$

Degree of Freedom

- In statistics, the number of **degrees of freedom** is the number of values in the final calculation of a statistic that are free to vary.
- In general, the degrees of freedom of an estimate of a parameter are equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself.

Example:

- if the variance is to be estimated from a random sample of size N then the degrees of freedom is equal to N minus the number of parameters estimated as intermediate steps (the sample mean) and is therefore equal to $N - 1$.
- For a one-sample t-test, one degree of freedom is spent estimating the mean, and the remaining $n - 1$ degrees of freedom estimate variability.

One-Sample t-Tests

Example: Sleep Hours of University Students

One-sample one-tailed t-test

- **Step 1: Set up a hypothesis.**
- Suppose we now want to test if on average university students get **less** sleep than adults do. We can set up the hypothesis test as follows.

$$H_0: \mu \geq 7.25$$

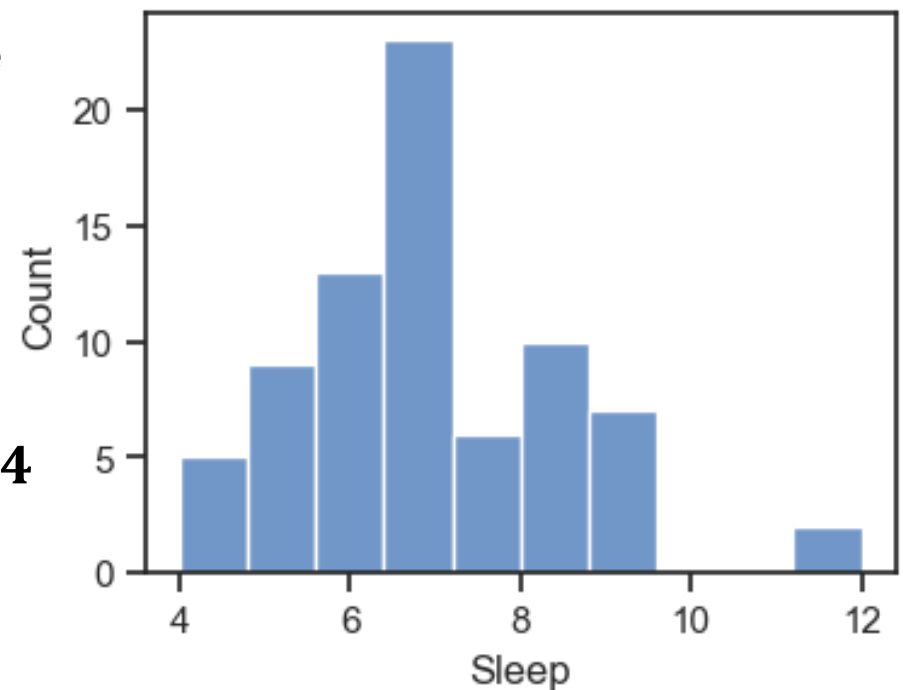
$$H_1: \mu < 7.25$$

One-Sample t-Tests

Example: Sleep Hours of University Students

One-sample one-tailed t-test

- **Step 2: Collect a sample and find out the summary statistics of the sample**
- We already have the data and the summary statistics.
 - Sample size = 75
 - Sample mean = 6.873
 - Sample std = 1.538
 - **degree of freedom (ν) = $n - 1 = 74$**
 - Sleep hours approximately follow the normal distribution



One-Sample t-Tests

Example: Sleep Hours of University Students

One-sample one-tailed t-test

- **Step 3: Define a significance level**
- We keep the significance level the same as the last demonstration
 $\alpha = 0.05$

One-Sample t-Tests

Example: Sleep Hours of University Students

One-sample one-tailed t-test

- **Step 4b: Generate the null distribution (t-test)**

- In this example

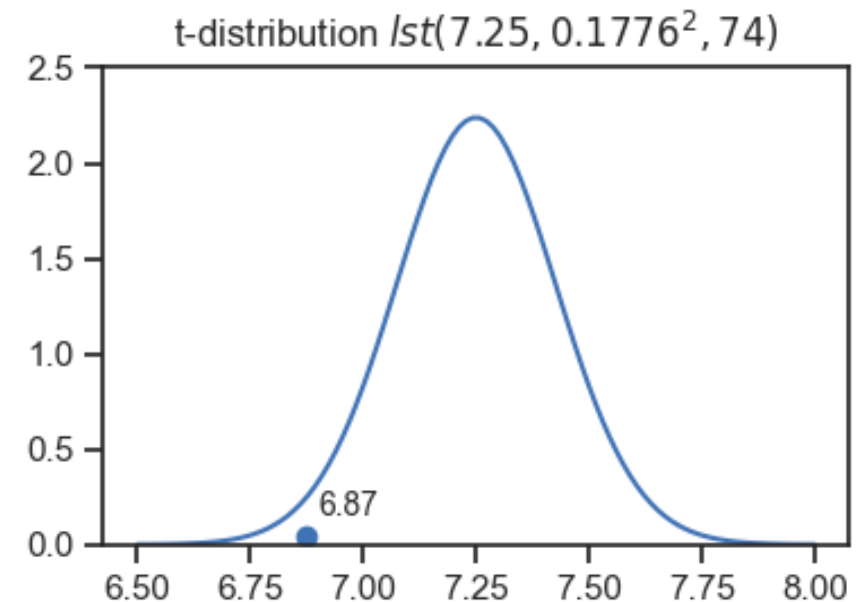
$$SE = \frac{s}{\sqrt{n}} = \frac{1.538}{\sqrt{75}} = 0.1776$$

- Based on the t-distribution definition, the sampling distribution of the mean under the null hypothesis (**null distribution**) follow a location-scale

t-distribution $lst(\mu = 7.25, \tau^2 = 0.1776^2, \nu = 74)$, i.e.,

$$T = \frac{\bar{X}_0 - 7.25}{0.1776} \sim t_{74}$$

where t_{74} is the (standardized) t-distribution with the degree of freedom of 74.



One-Sample t-Test

Example: Sleep Hours of University Students

One-sample one-tailed t-test

- **Step 5b: Calculate the p-value (t-test)**

- In this left-tailed test, we can calculate the p-value as

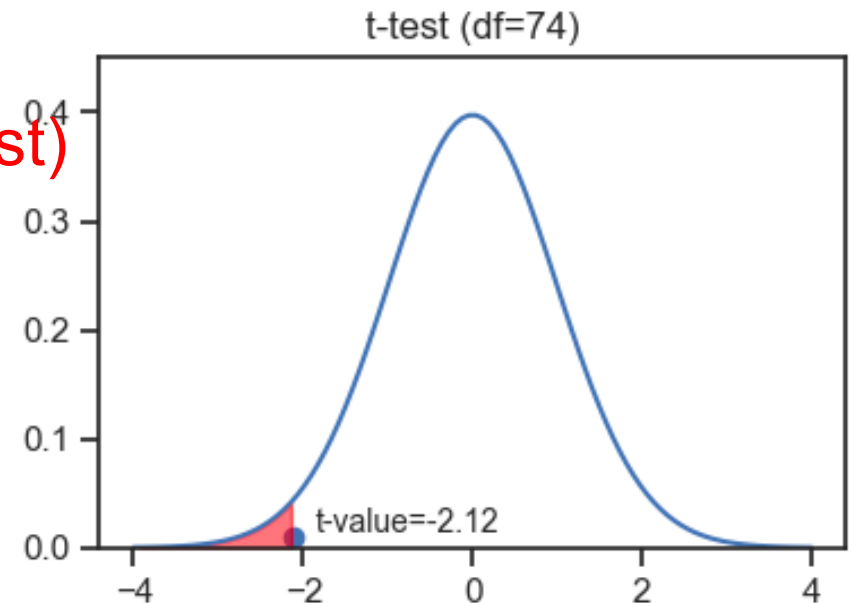
$$P(\bar{X}_0 < \bar{x})$$
$$= P\left(\frac{\bar{X}_0 - 7.25}{0.1776} < \frac{6.873 - 7.25}{0.1776}\right)$$

- t-value

$$t = \frac{\bar{x} - \mu_0}{SE} = \frac{6.873 - 7.25}{0.1776} = -2.12$$

- By using `stats.t.cdf()` in SciPy, we can get

$$p - value = P(T < -2.12) = 0.01856$$



```
1 t_value = (sample_mean - 7.25) / std_err
2 stats.t.cdf(t_value, df=74)
```

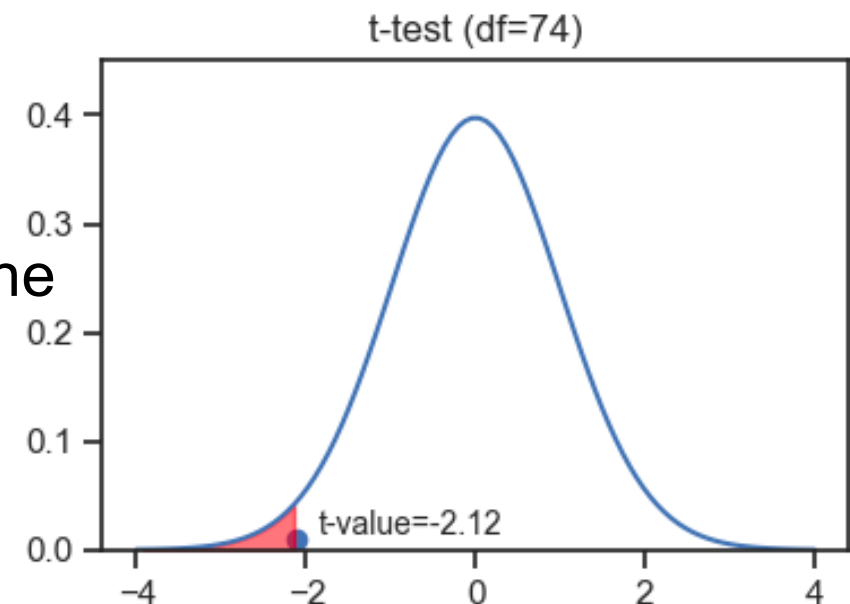
```
0.01863897743847032
```

One-Sample t-Test

Example: Sleep Hours of University Students

One-sample one-tailed t-test

- **Step 4b + 5b: One-sample t-Test using SciPy**
- By using `scipy.stats.ttest_1samp()` in SciPy, we can get
 $p - value = 0.01856$
- In the returned results, the `statistic` is the t-value, `pvalue` is the p-value, `df` is the degree of freedom.



```
1 stats.ttest_1samp(stu_survey_sleep_df['Sleep'], popmean=7.25, alternative='less')
```

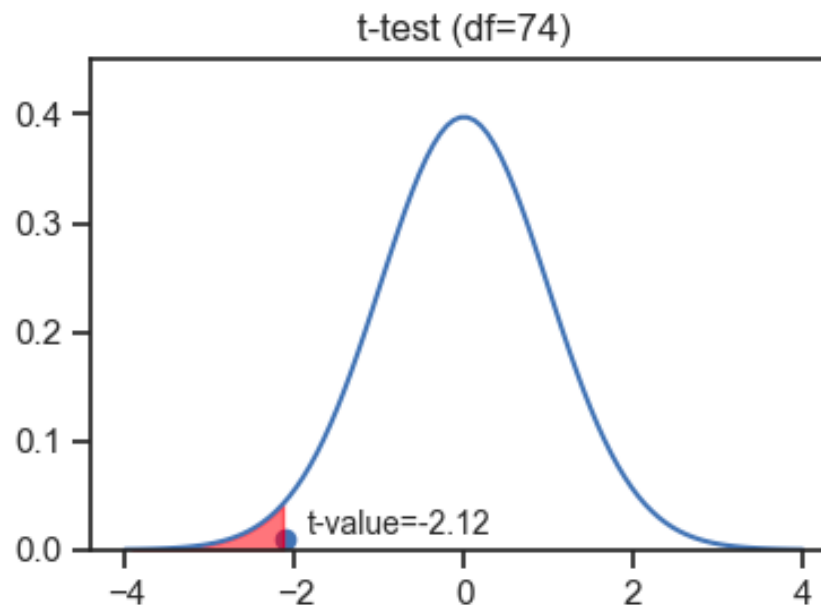
```
TtestResult(statistic=-2.120932064022875, pvalue=0.018638977438470327, df=74)
```

One-Sample t-Test

Example: Sleep Hours of University Students

One-sample one-tailed t-test

- **Step 6: Draw conclusions**
- Because $p\text{-value} = 0.01856 < \alpha (=0.05)$, we **reject the null in favour of the alternative**, i.e., there is enough evidence supporting the claim that the number of hours of sleep university students get is **less than** 7.25 hours.



One-Sample t-Test

Example: Sleep Hours of University Students

One-sample one-tailed t-test

Step 4b - 6: One-sample t-test using the t-table

- Calculate t-value as $t = (\bar{x} - \mu_0)/SE$.
 - In this example, t-value is -2.12
- Based on α and df , we can look up the **critical value** $t_{\alpha, \nu}$ (one-tailed test) or $t_{\alpha/2, \nu}$ (two-tailed test) in the t-table, which is actually the t-value when the significance level is just met.
 - In this example, we can find the critical value as 1.664

Table B

t distribution critical values

df	Tail probability <i>p</i>											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390

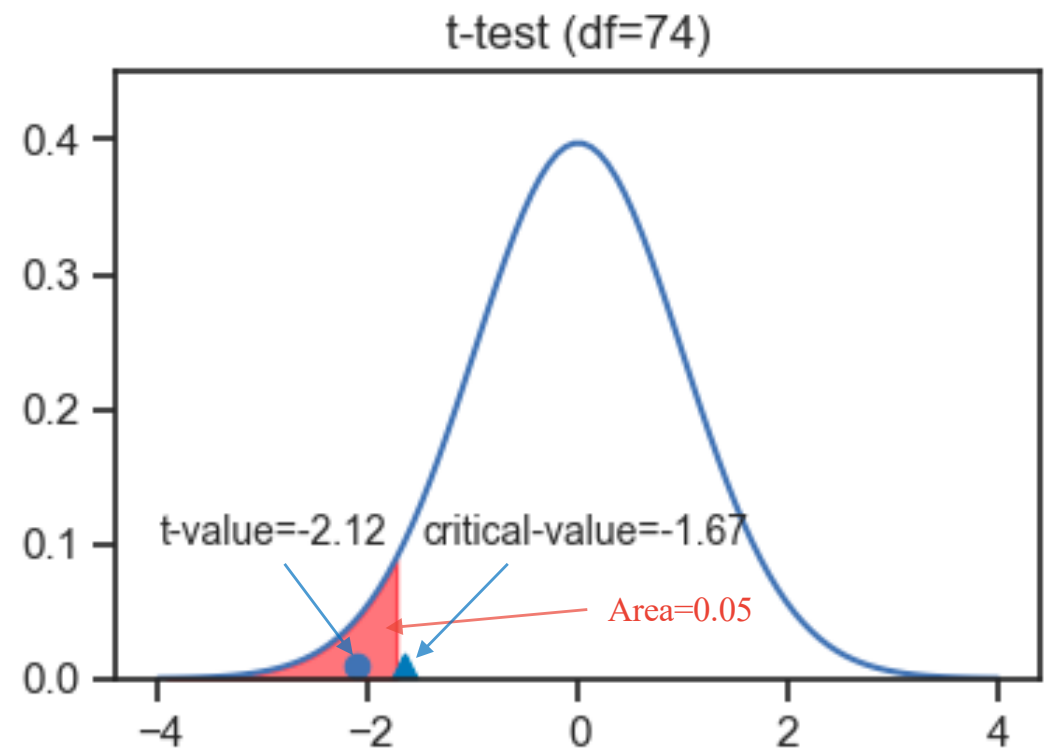
One-Sample t-Test

Example: Sleep Hours of University Students

One-sample one-tailed t-test

Step 4b - 6: One-sample t-test using the t-table

- the **critical value** in the t-table is actually the t-value when the significance level is just met
- Compare the t-value with the critical value, if $|t\text{-value}| > \text{critical value}$, reject the null. Otherwise, fail to reject the null.
 - In this example, since $|t\text{-value}| = 2.12 > 1.664$, we reject the null in favour of the alternative



Standard Error by Bootstrapping

Example: Sleep Hours of University Students

One-sample one-tailed t-test

- **Step 4: Generate the null distribution**
- z-test/t-test:
 - generate null distributions based on the estimation of the standard error s/\sqrt{n} with some assumptions, e.g., the distribution of the sample data should be approximately normal.
- Bootstrapping
 - Use bootstrapping to generate the sampling distribution of the statistic of interest
 - Obtain the SE from bootstrapped sampling distribution.
 - continue to generate null distribution in **step 4**, and continue the hypothesis test from **step 4**.

Standard Error by Bootstrapping

Demo: Sleep Hours of University Students

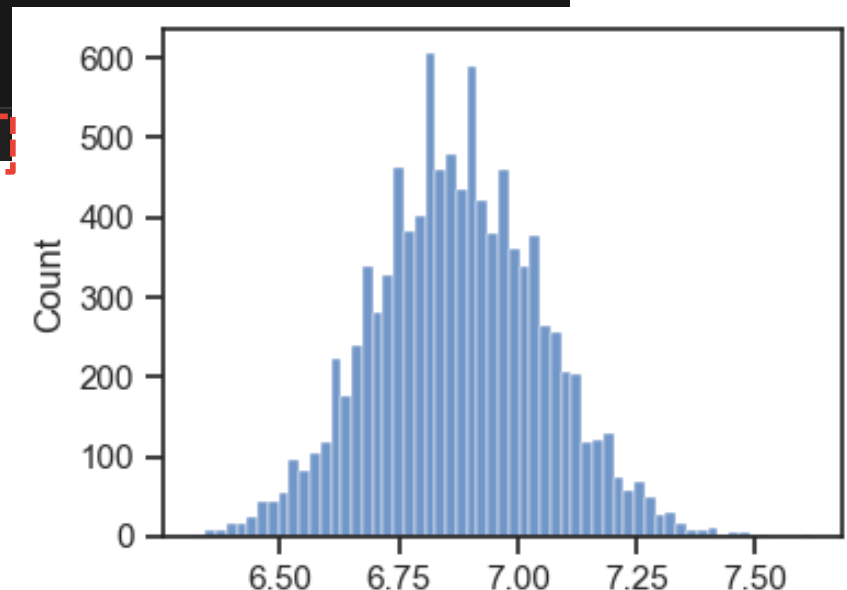
One-sample one-tailed t-test

- **Step 4: Generate the null distribution**

```
1 # Bootstrap the sample of sleep hours
2 boot_means = np.array([stu_survey_sleep_df.sample(frac=1, replace=True)['Sleep'].mean() for i in range(10000)])
3 std_err_boot = boot_means.std()
4 print("standard error of sampling distribution from bootstrapping:", std_err_boot)
5
6 fig, ax = plt.subplots(figsize=(4, 3))
7 sns.histplot(data=boot_means, ax=ax)
8
9 plt.show()
```

✓ 0.7s

standard error of sampling distribution from bootstrapping: 0.178584734496292



- the standard error (0.1786) is very close to our estimation (0.1776)

COMMON STEPS OF HYPOTHESIS TESTS

Common Steps of Hypothesis Tests

- **Step 1:** Set up a hypothesis
 - One-sample, two-sample, *paired*, *multiple*, etc.
 - One-tailed, or two-tailed.
- **Step 2:** Collect a sample.
 - Collect a random sample and find out the summary statistics of the sample e.g., sample size (n), sample mean (\bar{x}), sample standard deviation (s)
 - Check the necessary assumptions
- **Step 3:** Define a significance level.
 - Define a significance level to indicate the max Type I error you can accept (e.g., $\alpha = 0.05$)

Common Steps of Hypothesis Tests

- **Step 4:** Generate the sampling distribution of the statistic of interest under the null hypothesis (null distribution) based on the collected sample
 - **z-test:** the sampling distribution follows the normal distribution
 - when we know the population standard deviation (σ) (rarely the case)
 - $SE_z = \sigma / \sqrt{n}$
 - when the sample size (n) > 30
 - $SE_z \approx s / \sqrt{n}$

Common Steps of Hypothesis Tests

- **Step 4:** Generate the sampling distribution of the statistic of interest under the null hypothesis (null distribution) based on the collected sample
 - **t-test:** the sampling distribution follows the t-distribution
 - $SE_z = s/\sqrt{n}$
 - especially when the sample size (n) is small (≤ 30)
 - similar results to that of z-test with a large sample size
 - more general than z-test

Common Steps of Hypothesis Tests

- **Step 4:** Generate the sampling distribution of the statistic of interest under the null hypothesis (null distribution) based on the collected sample
 - Assumptions of z/t-test
 - The data are continuous.
 - The sample data have been randomly sampled from a population.
 - The sample data is approximately normally distributed.
 - SE can be obtained by bootstrapping the sample if the normality assumption is violated.

Common Steps of Hypothesis Tests

- **Step 5: Determine the p-value**

- **z-test:**

- z-order: $z = \frac{\bar{x} - \mu_0}{SE_z}$
 - p-value:
 - Two-tailed: $P(|Z| > |z|)$
 - Left-tailed: $P(Z < z)$
 - Right-tailed: $P(Z > z)$

- **t-test:**

- t-value: $t = \frac{\bar{x} - \mu_0}{SE_t}$
 - p-value:
 - Two-tailed: $P(|T| > |t|)$
 - Left-tailed: $P(T < t)$
 - Right-tailed: $P(T > t)$
 - instead of p-value, we can obtain critical value by referring to a t-table with the significance level α
 - $t_{\alpha, \nu}$ (one-tailed test)
 - or $t_{\alpha/2, \nu}$ (two-tailed test)

Common Steps of Hypothesis Tests

- **Step 6: Draw conclusions**
 - if $p - value \leq \alpha$, we reject null in favour of the alternative
 - if $p - value \geq \alpha$, we fail to reject null
 - For t-test, if we choose to use the critical value
 - if $|t\text{-value}| > \text{critical value } t_{\alpha, \nu} \text{ or } t_{\alpha/2, \nu}$, we reject null in favour of the alternative
 - otherwise, we fail to reject null.

TWO-SAMPLE TESTS

Introduction

- A **two-sample test** is a test performed on the data of two random samples, each independently obtained from a different given population. The purpose of the test is to determine whether the difference between these two populations is statistically significant.

Hypothesis Setup

- The common setups of two-sample t-tests are:
 - Two-tailed test
 - $H_0: \mu_1 = \mu_2$ or $H_0: \mu_1 - \mu_2 = 0$
 - $H_1: \mu_1 \neq \mu_2$ or $H_1: \mu_1 - \mu_2 \neq 0$
 - One-tailed test
 - $H_0: \mu_1 \leq \mu_2$ or $H_0: \mu_1 - \mu_2 \leq 0$
 - $H_1: \mu_1 > \mu_2$ or $H_1: \mu_1 - \mu_2 > 0$
- OR
- $H_0: \mu_1 \geq \mu_2$ or $H_0: \mu_1 - \mu_2 \geq 0$
 - $H_1: \mu_1 < \mu_2$ or $H_1: \mu_1 - \mu_2 < 0$

Two-Sample t-Test

- A **two-sample test** is a test performed on the data of two random samples, each independently obtained from a different given population. The purpose of the test is to determine whether the difference between these two populations is statistically significant.
- A two-sample t-test can be considered as a one-sample t-test for the **mean difference** between two populations ($\mu_{diff} = \mu_1 - \mu_2$). Under the null hypothesis, the sampling distribution of the mean difference ($\bar{X}_{diff} = \bar{X}_1 - \bar{X}_2$) has a mean of 0.

Two-Sample t-Test

t-Value:

- The t-value can be calculated as:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{diff}} = \frac{(\bar{x}_1 - \bar{x}_2)}{SE_{diff}}$$

where \bar{x}_1 is the mean of the sample drawn from population 1 with a size of n_1 and a standard deviation of s_1 , \bar{x}_2 is the mean of the sample drawn from population 2 with a size of n_2 and a standard deviation of s_2 . SE_{diff} is the standard error of the mean difference between two populations ($\bar{X}_1 - \bar{X}_2$)

Two-Sample t-Test

t-Value:

- If two populations have **equal variances**

$$SE_{diff} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the pooled variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, and s_1 and s_2 are the standard deviation of the two samples, and the degree of freedom:

$$df = n_1 + n_2 - 2$$

Two-Sample t-Test

t-Value:

- If two populations have **unequal variances**

$$SE_{diff} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

the approximate degree of freedom

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{(n_1 - 1)} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{(n_2 - 1)} \left(\frac{s_2^2}{n_2}\right)^2}$$

Two-Sample t-Tests

Example: Differences in Sleep Hours between Male and Female Students

- Same dataset as the last example.
 - 75 students' data with valid sleep hours and their genders
- In this example, we'll focus on the difference in sleep hours between male and female students.

ID	Gender	Classification	Height	Shoe Size	Phone Time	# of Shoes	Birth order	Pets	Happy	Funny	College	Bfast Calories	Exercise	Stat Pre	Stat Post	Phone Type	Sleep	Social Media	Impact of SocNetworking	Political	Animal	Superhero
1	male	senior	67.75	7	12	12	youngest	5	0.8	7	Natural Sciences	500	360	3		iPhone	7	180	worse	Democrat	Dog person	Batman
2	male	freshman	71	7.5	1.5	5	middle	4	0.75	8	Natural Sciences	0	200	9		Android smartphone	7	20	better	Democrat	Dog person	Batman
3	female	freshman	64	6	25	15	oldest	8	0.9	6	Natural Sciences	200	30	7	5	Android smartphone	8	60	better	Republican	Dog person	Batman
4	female	freshman	63	6.5	30	30	middle	12	0.98	9	Education	200	180	6	7	iPhone	6	60	better	Republican	Both	Superman
5	male	senior	69	6.5	23	8	oldest	4	0.75	6	Natural Sciences	0	180	4	7	iPhone	5.5	60	worse	Independent	Dog person	Superman
6	female	senior	64	8.5	13	25	oldest	1	0.95	5	Natural Sciences	250	310	7	7	iPhone	6.5	90	no impact	Democrat	Dog person	Batman
7	female	freshman	62	8.5	23	12	oldest	2	0.95	7	Nursing	200	60	7	8	Android smartphone	7	120	better	Republican	Both	Superman
8	female	freshman	64	6	50	50	youngest	10	0.9	4	Liberal Arts	200	0	5	7	iPhone	7	60	no impact	Independent	Both	Superman
9	female	freshman	66	8	10	15	youngest	25	0.9	9	Natural Sciences	0	0	8	6	iPhone	7	3	no impact	Democrat	Both	Batman
10	female	freshman	68	6.5	40	20	oldest	4	0.95		Nursing	150	240	6	10	iPhone	6	180	better	Democrat	I don't like either	Batman

Data Source: <http://sites.utexas.edu/sos/guided/inferential/numeric/claim/one-sample-t/>

Two-Sample t-Tests

Example: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed t-test

- **Step 1: Set up a hypothesis.**
- Suppose we are interested to know whether male university students' average sleep hours per day is significantly different from the average sleep hours per day of the female students or not. We can set up the hypothesis test as follows.

$$H_0: \mu_f - \mu_m = 0$$

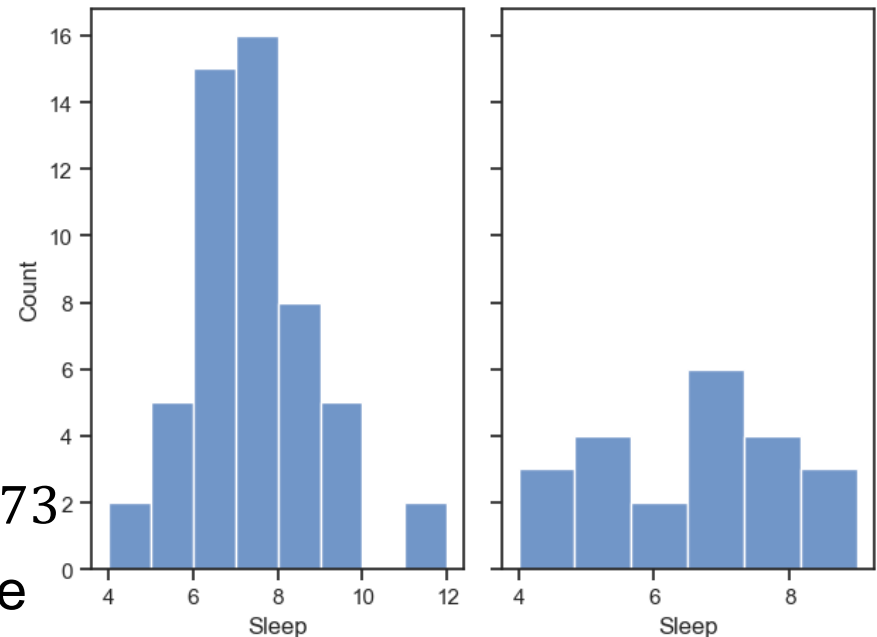
$$H_1: \mu_f - \mu_m \neq 0$$

Two-Sample t-Tests

Example: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed t-test

- **Step 2: Collect a sample and find out the summary statistics of the sample**
- We already have the data and the summary statistics.
 - Sample sizes $n_f = 53, n_m = 22$
 - Sample mean $\bar{x}_f = 7.019, \bar{x}_m = 6.523$
 - Sample std $s_f = 1.535, s_m = 1.523$
 - $s_f \approx s_m$, assume the two population distributions have approximately the same variance, so $\nu = n_f + n_m - 2 = 73$
 - the distributions of the two samples are roughly bell-shaped



Two-Sample t-Tests

Example: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed t-test

- **Step 3: Define a significance level**
- We keep the significance level the same as the last demonstration
 $\alpha = 0.05$

Two-Sample t-Tests

Example: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed t-test

- **Step 4c: Generate the null distribution (t-test)**
- Since $s_f \approx s_m$, assume the two population distributions have approximately the same variance

pooled variance:

$$s_p^2 = \frac{(n_f - 1)s_f^2 + (n_m - 1)s_m^2}{n_f + n_m - 2} = \frac{(53 - 1)1.535^2 + (22 - 1)1.523^2}{53 + 22 - 2} = 2.346$$

pooled standard deviation

$$s_p = \sqrt{s_p^2} = 1.532$$

standard error of the null distribution

$$SE_{diff} = s_p \sqrt{\frac{1}{n_f} + \frac{1}{n_m}} = 1.532 \sqrt{\frac{1}{53} + \frac{1}{22}} = 0.389$$

$$\therefore T = \frac{\bar{X}_{diff} - 0}{0.389} \sim t_{73}$$

Two-Sample t-Test

Example: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed t-test

- **Step 5c: Calculate the p-value (t-test)**
- The null distribution of the sleep hour difference

$$T = \frac{\bar{X}_{diff} - 0}{0.389} \sim t_{73}$$

- t-value

$$t = \frac{(\bar{x}_f - \bar{x}_m) - (\mu_1 - \mu_2)}{SE_{diff}} = \frac{(7.019 - 6.523) - 0}{0.389} = 1.275$$

- By using `stats.t.cdf()` in SciPy with `df=73`, we can get

$$p - value = 2P(T > 1.275) = 0.206$$

```
1 t_value = (sample_mean_female - sample_mean_male) / se_p
2 p_value = 2*(1 - stats.t.cdf(t_value, df))
3 print("t-value = {}, p-value = {}".format(t_value, p_value))
✓ 0.0s
```

```
t-value = 1.2773441806166717, p-value = 0.2055284062199625
```

Two-Sample t-Test

Example: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed t-test

- **Step 4c + 5c: Two-sample t-Test using SciPy**
- We can use `scipy.stats.ttest_ind()` in SciPy to perform two-sample t-test
 - With the assumption of equal variances between two populations, `equal_var` is set to `True`.

```
1 stats.ttest_ind(stu_survey_sleep_df[stu_survey_sleep_df['Gender'] == 'female']['Sleep'],
2                 stu_survey_sleep_df[stu_survey_sleep_df['Gender'] == 'male']['Sleep'],
3                 equal_var=True,
4                 alternative='two-sided')
```

✓ 0.0s

```
TtestResult(statistic=1.2773441806166719, pvalue=0.20552840621996235, df=73.0)
```

- In the returned results, the `statistic` is the t-value, `pvalue` is the p-value, `df` is the degree of freedom.

Two-Sample t-Test

Example: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed t-test

- **Step 4c + 5c: Two-sample t-Test using SciPy**
- We can also perform the t-test without the assumption of equal variance by setting `equal_var` is set to `False`.

```
1 stats.ttest_ind(stu_survey_sleep_df[stu_survey_sleep_df['Gender'] == 'female']['Sleep'],
2                 stu_survey_sleep_df[stu_survey_sleep_df['Gender'] == 'male']['Sleep'],
3                 equal_var=False,
4                 alternative='two-sided')
✓ 0.0s

TtestResult(statistic=1.2813008328404065, pvalue=0.2075436466844493, df=39.583166134257965)
```

- In the returned results, the `statistic` is the t-value, `pvalue` is the p-value, `df` is the degree of freedom.

Two-Sample t-Test

Example: Differences in Sleep Hours between Male and Female Students

One-sample one-tailed t-test

- **Step 6: Draw conclusions**
- Because $p\text{-value} = 0.206 > \alpha (=0.05)$, we **failed to reject the null**, i.e., there **isn't** enough evidence supporting the claim that the average sleep hours per day is different between female and male students, the difference between the average sleep hours between female and male students are not statistically significant.

Two-Sample t-Test

Exercise: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed t-test

Step 4c - 6: Two-sample t-test using the t-table

Go to
www.menti.com

Enter the code

1939 7765



Or use QR code

Table B *t* distribution critical values

df	Tail probability <i>p</i>											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390

Two-Sample z-Test by Bootstrapping

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 4: Generate the null distribution**

```
1 boot_female_sleep_means, boot_male_sleep_means, boot_diff_means = [], [], []
2
3 for i in range(10000):
4     # Bootstrapping the original sample to generate a new sample
5     boot_sample = stu_survey_sleep_df.sample(frac=1, replace=True)
6
7     # Calculate the mean sleep hours for female and male students respectively.
8     female_sleep_mean = boot_sample[boot_sample['Gender'] == 'female']['Sleep'].mean()
9     male_sleep_mean = boot_sample[boot_sample['Gender'] == 'male']['Sleep'].mean()
10
11     # Insert current mean to the list of means to generate the sampling distributions of the mean sleep hours of female and male students
12     # as well as the sampling distribution of the mean difference between female and male students.
13     boot_female_sleep_means.append(female_sleep_mean)
14     boot_male_sleep_means.append(male_sleep_mean)
15     boot_diff_means.append(female_sleep_mean - male_sleep_mean)
16
17 print("standard error of the sampling distribution of mean sleep hours of female from bootstrapping: ", np.std(boot_female_sleep_means))
18 print("standard error of the sampling distribution of mean sleep hours of male from bootstrapping:", np.std(boot_male_sleep_means))
19 print("standard error of the sampling distribution of mean difference from bootstrapping: ", np.std(boot_diff_means))
20
```

✓ 2.3s Python

```
standard error of the sampling distribution of mean sleep hours of female from bootstrapping: 0.20957152939602225
standard error of the sampling distribution of mean sleep hours of male from bootstrapping: 0.32059062692473955
standard error of the sampling distribution of mean difference from bootstrapping: 0.38286697294521543
```

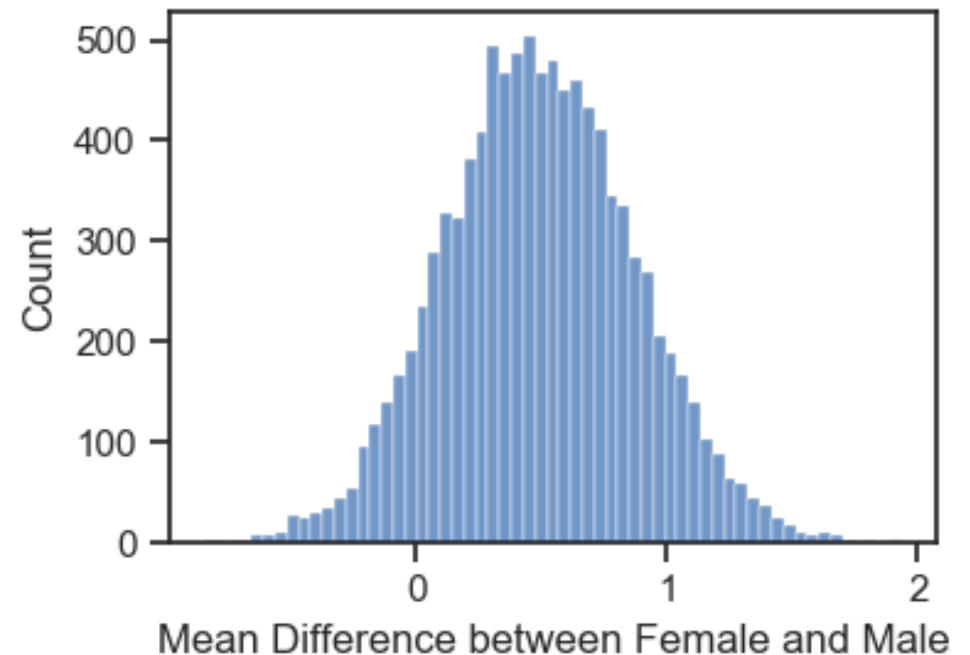
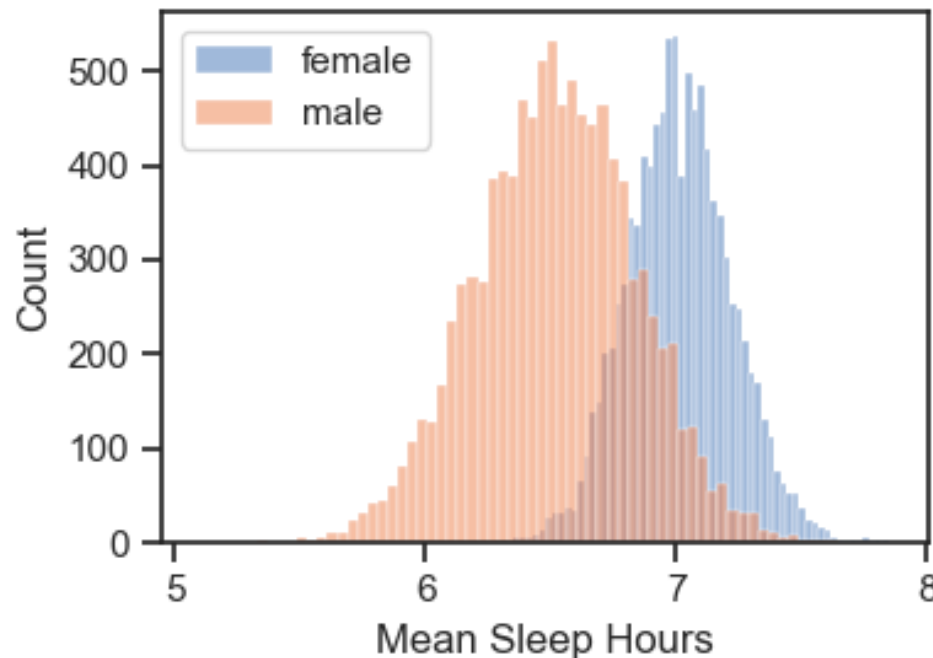
- the standard error (0.383) is very close to our estimation (0.389)

Two-Sample z-Test by Bootstrapping

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 4:** Generate the null distribution



- we can generate the null distribution of the mean difference as $N(0, 0.383^2)$

Two-Sample z-Test by Bootstrapping

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 5: Calculate the p-value (z-test)**

- z-score

$$z = \frac{(\bar{x}_f - \bar{x}_m) - (\mu_1 - \mu_2)}{SE_{diff}} = \frac{(7.019 - 6.523) - 0}{0.383} = 1.295$$

- By looking up the z-table, we can get p-value:

$$p - value = 2P(Z > 1.295) = 2 * (1 - 0.9032) = 0.1936$$

- We can get similar result using `scipy.stats.norm.cdf()`.

- **Step 6: Draw conclusions**

- Because $p\text{-value} > \alpha = 0.05$, we fail to reject the null.

PERMUTATION TESTS

Introduction

- **z/t-Tests**

- **parametric tests** where null distributions are obtained from theoretical probability distributions.
- rely on the assumption that the populations should follow the normal distribution

- **Permutation Tests**

- also known as a randomization test or re-randomization test, is a **non-parametric** statistical test used to assess whether the observed differences between groups or the observed relationship between variables in a dataset are statistically significant.
- does not rely on specific assumptions about the underlying population distribution. It is particularly useful when the assumptions of traditional parametric tests, such as the z/t-test, are not met, or when you want to perform hypothesis testing without making distributional assumptions.
- involve two or more samples

Introduction

- **The Null Distribution in Permutation Tests**

- The null hypothesis is that all samples come from the same distribution.
- Under the null hypothesis, the distribution of the test statistic is obtained by calculating all possible values of the test statistic under possible rearrangements of the observed data.
- For example, when we want to test the mean difference between two populations, the permutation test simply generates the distribution of mean differences under the assumption that the two groups are not distinct in terms of the measured variable.
- From this, one then uses the observed statistic to see to what extent this statistic is special, i.e., the likelihood of observing such a value (or larger) if the treatment labels had simply been randomized after treatment

- **Permutation Test Procedure**

- The overall steps of a permutation test are the same as a z/t-test (**step 1** to **step 6**). The differences are in **step 4** (Generate the null distribution) and **step 5** (Calculate p-value).

Permutation Procedure

The permutation procedure in **step 4** (Generate the null distribution) and **step 5** (Calculate p-value) is listed as follows:

- **Step 4d: Generate the null distribution**
 - 1. Combine the results from the different groups/samples into a single dataset.
 - 2. Shuffle the combined data and then randomly draw (**without replacement**) a resample of the same size as group A.
 - 3. From the remaining data, randomly draw (without replacement) a resample of the same size as group B.
 - 4. Do the same for groups C, D, and so on. You have now collected one set of resamples that mirror the sizes of the original samples.
 - 5. Calculate the statistic of interest for the permuted samples; this constitutes one permutation iteration.
 - 6. Repeat the previous steps R times to yield a permutation distribution of the test statistic (null distribution).

Permutation Procedure

- **Step 5d: Calculate p-value**
 - Compare the observed **test statistic** to the distribution of test statistics obtained from the permutations (step 4).
 - Calculate the proportion of permuted test statistics that are as extreme as or more extreme than the observed test statistic. This proportion is the p-value.
- **Step 6: Draw conclusions**
 - Same as other tests, If the p-value is smaller than a predetermined significance level (α), you reject the null hypothesis in favour of the alternative hypothesis.
 - Otherwise, you fail to reject the null hypothesis.

Permutation Tests

Example:

- Suppose we have two random samples drawn from two populations. The 4 values in red are drawn from one distribution, and the 5 values in blue from another. We'd like to test whether the means of the two populations are different.
 - The null hypothesis is that both groups of samples are drawn from the same distribution. $H_0: \mu_1 = \mu_2$ or $H_0: \mu_1 - \mu_2 = 0$
 - The alternative hypothesis is that the mean of the first distribution is higher than the mean of the second. $H_1: \mu_1 \neq \mu_2$ or $H_1: \mu_1 - \mu_2 \neq 0$

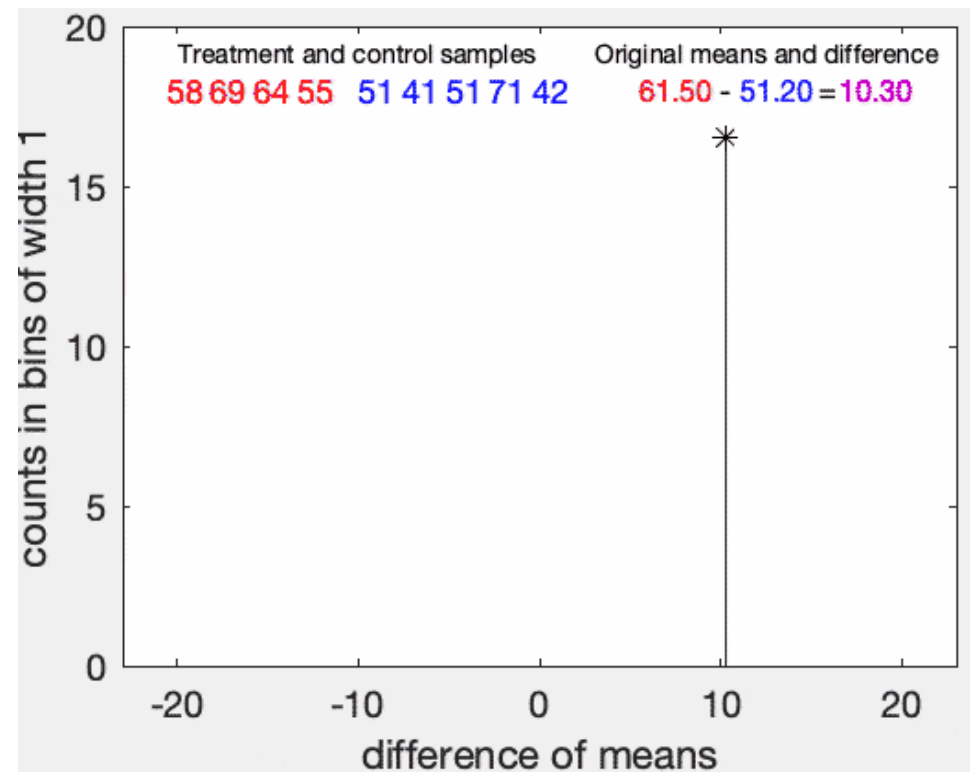
{58, 59, 64, 55}: $\bar{x}_1 = 61.5$

{51, 41, 51, 71, 42}: $\bar{x}_2 = 51.2$

Permutation Tests

Example:

- A permutation test is computed on sets of 4 and 5 random values.
 - There are 126 distinct ways to put 4 values into one group and 5 into another (9-choose-4 or 9-choose-5).
 - The p-value of the hypothesis is estimated as the proportion of permutations that give a difference as large or larger than the difference of means of the original samples.
 - In this example, we failed to reject the null at the $p = 5\%$ level.



Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 1: Set up a hypothesis.**
- Suppose we are interested to know whether male university students' average sleep hours per day is significantly different from the average sleep hours per day of the female students or not. We can set up the hypothesis test as follows.

$$H_0: \mu_f - \mu_m = 0$$

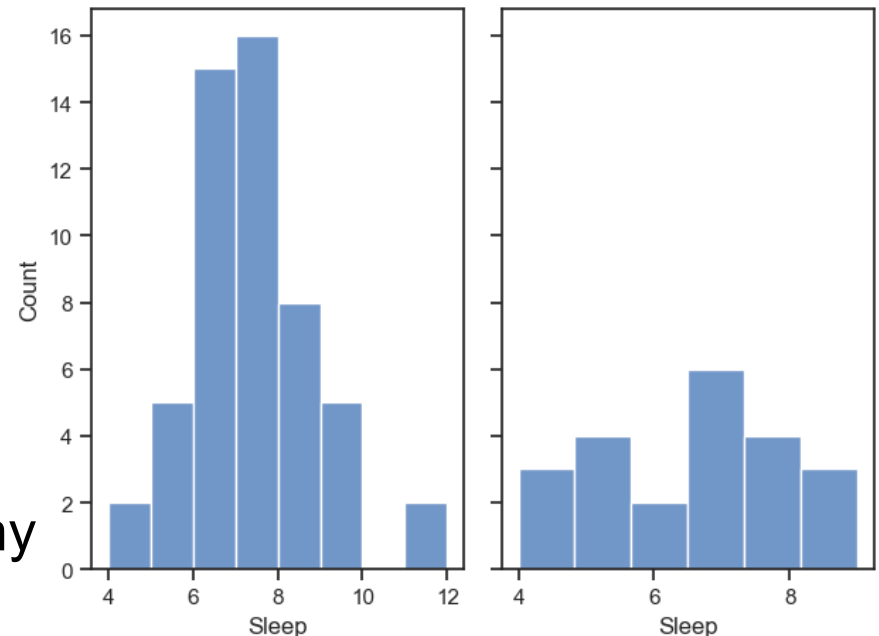
$$H_1: \mu_f - \mu_m \neq 0$$

Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 2: Collect a sample and find out the summary statistics of the sample**
- We already have the data and the summary statistics.
 - Sample sizes $n_f = 53, n_m = 22$
 - Sample mean $\bar{x}_f = 7.019, \bar{x}_m = 6.523$
 - Sample std $s_f = 1.535, s_m = 1.523$
 - Sample mean difference
 $\bar{x}_f - \bar{x}_m = 7.019 - 6.523 = 0.496$
 - the permutation test doesn't rely on any assumption of the distributions of two



Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 3: Define a significance level**
- We keep the significance level the same as the last demonstration
 $\alpha = 0.05$

Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 4: Generate the null distribution**

```
1 # Define the method to calculate the test statistic of interest
2 def diff_statistic(x, y):
3     x_mean = np.mean(x)
4     y_mean = np.mean(y)
5     mean_diff = x_mean - y_mean
6     return mean_diff
7
8 # Obtain the two original samples
9 female_sleep = stu_survey_sleep_df[stu_survey_sleep_df['Gender'] == 'female']['Sleep']
10 male_sleep = stu_survey_sleep_df[stu_survey_sleep_df['Gender'] == 'male']['Sleep']
11
12 print(female_sleep.shape, male_sleep.shape)
```

✓ 0.0s

(53,) (22,)

```
1 # Perform a permutation test for the two groups
2 pm_res = stats.permutation_test((female_sleep, male_sleep), diff_statistic, permutation_type='independent', alternative='two-sided')
3 pm_res
```

✓ 0.0s

```
PermutationTestResult(statistic=0.4961406518010296, pvalue=0.2188, null_distribution=array([-0.01843911,  0.1745283 , -0.17924528, ..., -0.661
-0.01843911,  0.68910806]))
```

statistic: the mean difference between two original samples, the null distribution is the sampling distribution of the mean difference under the null hypothesis generated by the permutation method

Python

Python

Permutation Tests

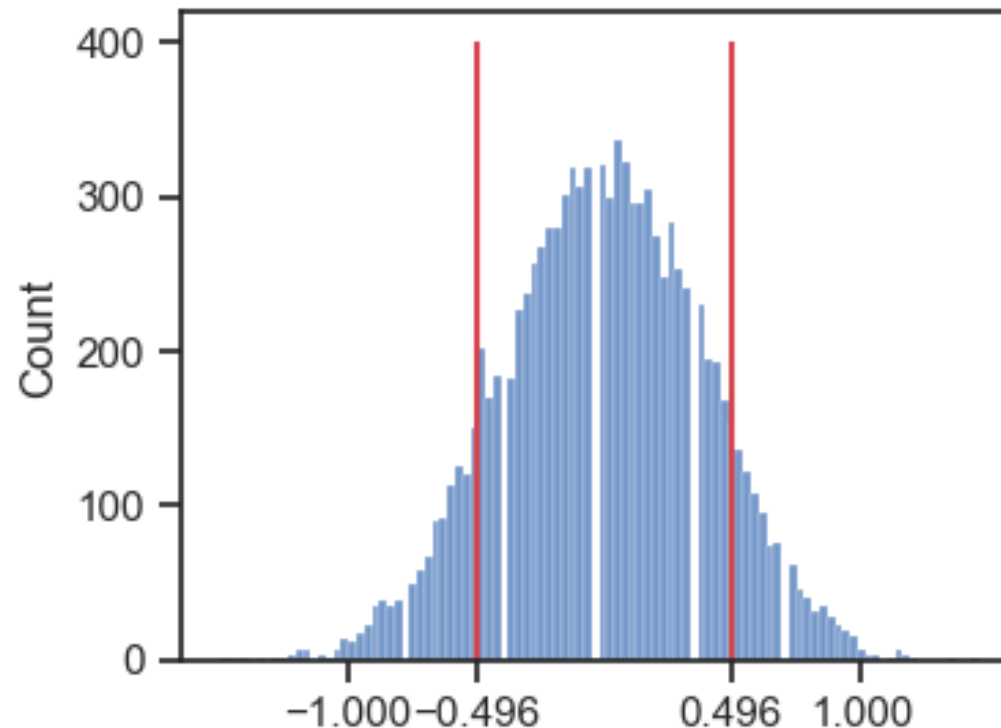
Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 4: Generate the null distribution**

```
PermutationTestResult(statistic=0.4961406518010296, pvalue=0.2188, null_distribution=array([-0.01843911, 0.1745283, -0.17924528, ..., -0.661  
-0.01843911, 0.68910806]))
```

Visualise
null_distribution



Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 5: Calculate the p-value (permutation test)**
- Read the pvalue in the returned results

```
PermutationTestResult(statistic=0.4961406518010296, pvalue=0.2188, null_distribution=array([-0.01843911, 0.1745283, -0.17924528, ..., -0.661  
-0.01843911, 0.68910806]))
```

- According to the definition.

```
1 (pm_res.null_distribution >= pm_res.statistic).mean() + (pm_res.null_distribution <= -pm_res.statistic).mean()  
✓ 0.0s  
0.2208220822082208
```

- **Step 6: Draw conclusions**
- Because $p\text{-value} > \alpha = 0.05$, we fail to reject the null.

OTHER THINGS TO CONSIDER

不考

The Impact of a Large Sample Size

- With large sample sizes, hypothesis testing leads to even the smallest of findings as **statistically significant**. However, these findings might not be **practically significant** at all.

Example: Sleep Hours of University Students

- Let's revisit the first one-sample example which test whether the average hours of sleep of university students is equal to or different from 7.25 hours.

$$H_0: \mu = 7.25$$

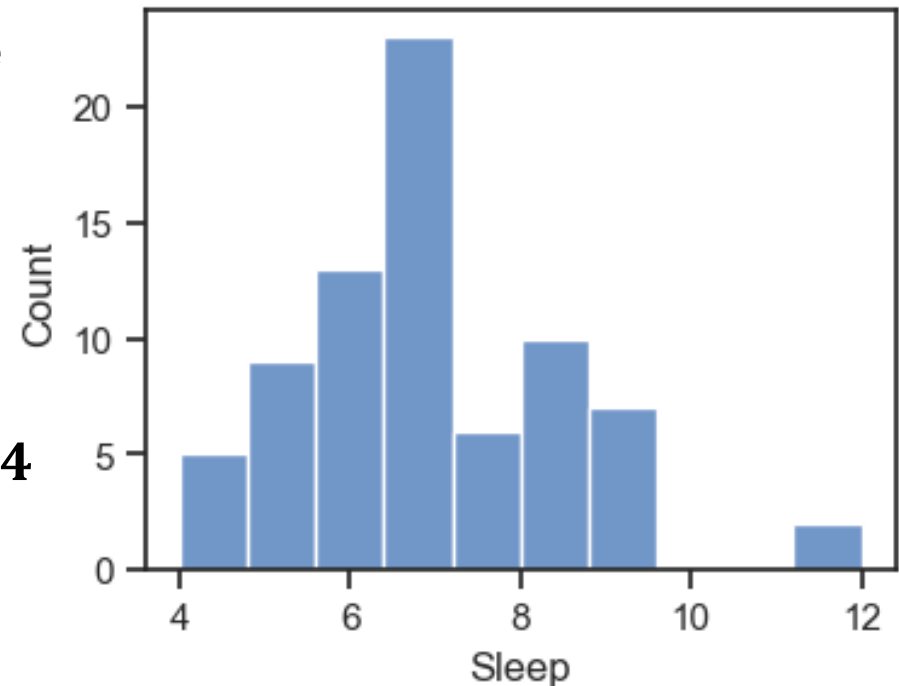
$$H_1: \mu \neq 7.25$$

The Impact of a Large Sample Size

Example: Sleep Hours of University Students

One-sample two-tailed t-test

- We already have the data and the summary statistics.
 - Sample size = 75
 - Sample mean = 6.873
 - Sample std = 1.538
 - **degree of freedom (ν) = $n - 1 = 74$**
 - Sleep hours approximately follow the normal distribution



- We keep the significance level the same as the last demonstration
 $\alpha = 0.05$

The Impact of a Large Sample Size

Example: Sleep Hours of University Students

One-sample two-tailed t-test

- If we use one-sample t-test

$$SE = \frac{s}{\sqrt{n}} = \frac{1.538}{\sqrt{75}} = 0.1776$$

- t-value

$$t = \frac{\bar{x} - \mu_0}{SE} = \frac{6.873 - 7.25}{0.1776} = -2.12$$

- By using `stats.t.cdf()` in SciPy, we can get

$$p - value = 2P(T < -2.12) = 2 * 0.01856 = 0.03712$$

- Because $p\text{-value} = 0.03712 < \alpha (=0.5)$, we **reject the null in favour of the alternative**, i.e., there is enough evidence supporting the claim that the number of hours of sleep university students get is less than 7.25 hours.

The Impact of a Large Sample Size

Exercise: Sleep Hours of University Students

One-sample two-tailed t-test

- Now suppose the sample data size is only **25**. Perform a two-tailed t-test again and draw a conclusion.
 - Sample size = **25**
 - Sample mean = 6.873
 - Sample std = 1.538
 - Sleep hours approximately follow the normal distribution
- the significance $\alpha = 0.05$

Go to
www.menti.com

Enter the code

1939 7765



Or use QR code

Statistical vs. Practical Significance

- Using confidence intervals and hypothesis testing, we are able to provide statistical significance in making decisions.
 - However, it is also important to take into consideration practical significance in making decisions.
 - Practical significance takes into consideration other factors of your situation that might not be considered directly in the results of your hypothesis test or confidence interval.
 - Constraints like space, time, or money are important in business decisions. However, they might not be accounted for directly in a statistical test.

Statistical vs. Practical Significance

Discussion: Statistical vs. Practical Significance

- If a weight loss drug helped individuals lose on average 0.2 kg over 1 years, which of the following are true statements about this result?
- This would definitely not be statistically significant
- This would definitely not be practically significant
- This would be more likely to be statistically significant if we have a small sample size
- This would be more likely to be statistically significant if we have a large sample size

Go to
www.menti.com

Enter the code

1939 7765



Or use QR code

RECAP

Hypothesis Testing

Recap

- **Setting Up Hypothesis Tests**
 - Null hypothesis
 - Alternative hypothesis
- **Types of Errors**
 - Type I error
 - type I error rate: α (significance level)
 - Type II error
 - type I error rate: β
- **Common Types of Hypothesis Tests**
 - Testing a population parameter (one-sample tests)
 - Testing the difference of parameters in different population (two-sample tests)
 - *Testing the difference before and after some treatment on the same individual (Paired t-test)*

Recap

- **Essentials for hypothesis tests**
 - Common Steps of Hypothesis Tests
 - p-value
 - t-distribution and critical value
- **One-Sample Tests**
 - One-sample z/t-tests
 - One-sample tests by bootstrapping
- **Two-Sample Tests**
 - Two-sample z/t-tests
 - Two-sample tests by bootstrapping
 - Permutation test
- **Other Things to Consider**
 - The Impact of a Large Sample Size
 - Statistical vs. Practical Significance

Questions

Use student forum on QM+
chao.shu@qmul.ac.uk