工程概率统计

Probability and Statistics for Engineering

第六章 假设检验

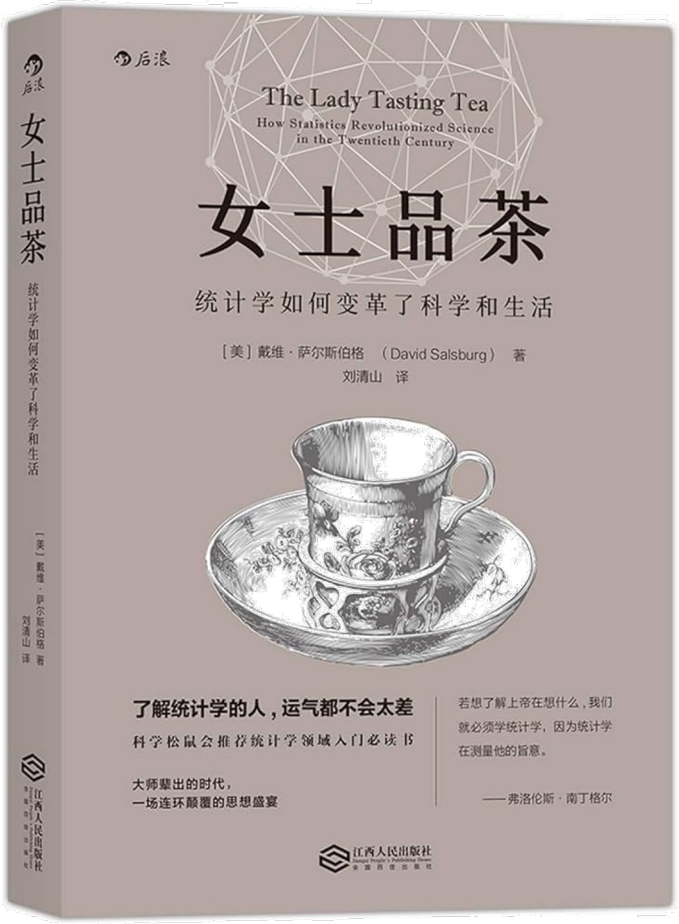
Chapter 6 Hypothesis Testing

Chapter 6 Hypothesis Testing

- 6.1 Overview and Basic Ideas
- 6.2 General Process and Examples

6.1 Overview and Basic Ideas

- Let's start this chapter with a historical story.
  - On a summer afternoon in the late 1920s in Cambridge, England, a group of university faculty members, along with their friends, were gathered around an outdoor table enjoying afternoon tea.
  - A lady insisted that the taste of tea is different depending on whether the milk is added to the tea or the tea added to the milk. However, the scientists present thought this was unlikely.
  - A man named Fisher fell into deep thought and then suggested testing the lady's claim through an experiment.
  - They brewed tea in different ways and asked her to identify how each cup of milk tea had been made.
  - The details of this experiment were recorded in Fisher's 1935 book *The Design of Experiments*.
  - Fisher did not describe the outcome of the experiment, but it is said that the lady correctly identified the method used for each cup of tea.



6.1 Overview and Basic Ideas

- Let's briefly understand the basic ideas of hypothesis testing through the example of the lady tasting tea.

Example 6.1

- There are a total of 8 cups of milk tea: 4 cups where tea is added first, and 4 cups where milk is added first. The cups are identical and are presented to the lady in a random order for tasting.
- After tasting all 8 cups, the lady indicated which ones had the tea added first, and the remaining ones are assumed to have had the milk added first.
- How can we determine from her tasting results whether the lady can truly distinguish between the two methods of making the milk tea?

6.1 Overview and Basic Ideas

Solution

6.1 Overview and Basic Ideas

Solution

6.1 Overview and Basic Ideas

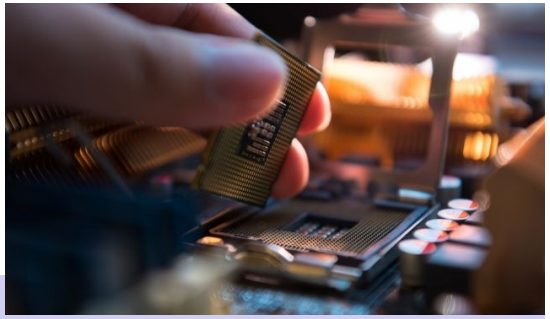
- A vital role of Statistics is in verifying a hypothesis (a statement / claim / conjecture).
- Based on a random sample (i.e., data), we would like to verify whether
  - a lady can really distinguish between the two ways of making milk tea;
  - the average internet speed is 54 Mbps, as claimed by the internet service provider;
  - the proportion of defective products is at most 3%, as promised by the manufacturer;
  - a hardware upgrade was efficient;
  - the supporting rate of a candidate in town A is higher than that in town B.
  - a new medical treatment is more effective than a standard treatment.
- Hypothesis Testing is a statistical inference method used to determine whether data supports a specific hypothesis. Hypothesis testing involves two opposing hypotheses:
  - Null hypothesis (零假设/原假设): typically the default assumption, representing no effect, no difference, or no association, denoted as H_0. It is tested through data collection to decide whether to reject it or not.
  - Alternative hypothesis (备择假设): usually the researcher's hypothesis, representing the presence of an effect, a difference, or an association, denoted as H_1.

6.1 Overview and Basic Ideas

- H_0 and H_1 are simply two mutually exclusive statements, e.g.,
  - H_0: a suspect is innocence $\leftrightarrow$ H_1: a suspect is guilty.
  - H_0: a new drug is not effective $\leftrightarrow$ H_1: a new drug is effective.
  - H_0: a hardware upgrade is not efficient $\leftrightarrow$ H_1: a hardware upgrade is effective.
- **Question:** which statement should be treated as the null hypothesis?
  - The choice of the null hypothesis should align with the context and consequences of making errors.
  - Some general guidelines are **Status Quo Assumption (现状假设)** and **Burden of Proof (举证责任)**.
  - **Status Quo Assumption:** H_0 is typically chosen as the statement that represents the status quo, no effect, no difference, or the baseline condition.
  - **Burden of Proof:** H_0 is often chosen as the statement that requires strong evidence to be overturned. The reason is that falsely rejecting H_0 can have more serious consequences than failing to reject it.
  - Therefore, hypothesis testing is like a legal process where H_0 is treated as "innocent until proven guilty" (疑罪从无). The data need to provide sufficient evidence to reject H_0.

6.1 Overview and Basic Ideas

- In the process of hypothesis testing, we test H_0 based on the sample data collected. If the sample data contradicts with H_0, then we reject H_0.



Example 6.2

- A manager evaluates effectiveness of a major hardware upgrade by running a certain process independently 50 times before the upgrade and 50 times after it.
- Based on the data collected, the average running time is 8.5 minutes (with standard deviation of 1.8 minutes) before the upgrade and 7.2 minutes (with SD of 1.6 minutes) after it.
- 1. Is the mean running time of the process before the hardware upgrade greater than 8 minutes?
- 2. Is the mean running time of the process after the hardware upgrade less than 7.5 minutes?
- 3. Does the mean running time of the process change due to the hardware upgrade?

6.1 Overview and Basic Ideas

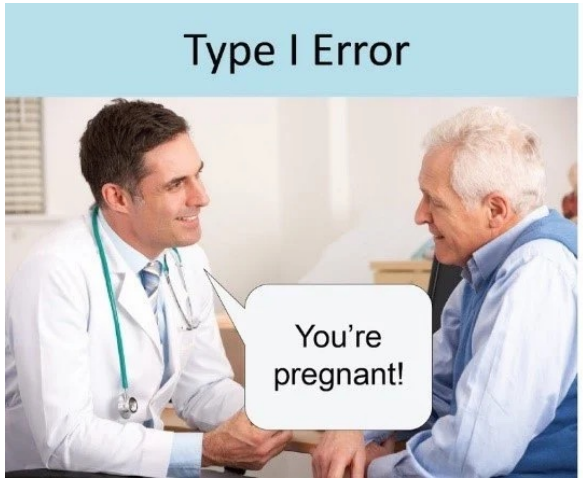
Thinking

6.1 Overview and Basic Ideas

- The set of samples $\boldsymbol{X} = (X_1, \ldots, X_n)$ such that $\bar{X} - 8 > c$, denoted by $\{\boldsymbol{X}: \bar{X} - 8 > c\}$, is called the **rejection region (RR, 拒绝域)** of the testing problem.
- Since $\bar{X}$ is a random variable, a decision made on whether to reject H_0 based on $\bar{X}$ is a random decision, while whether H_0 is true or not is deterministic (though unknown).
- This suggest that our decision might be wrong.
- Therefore, we may make two types of errors in general hypothesis testing problems:
  - **Type I error (第I类错误):** we reject H_0 when it is true.
  - **Type II error (第II类错误):** we fail to reject H_0 when it is not true.
- For a given problem, we can only make one of the two types of errors, however, we don't know which type of error we might make.

6.1 Overview and Basic Ideas

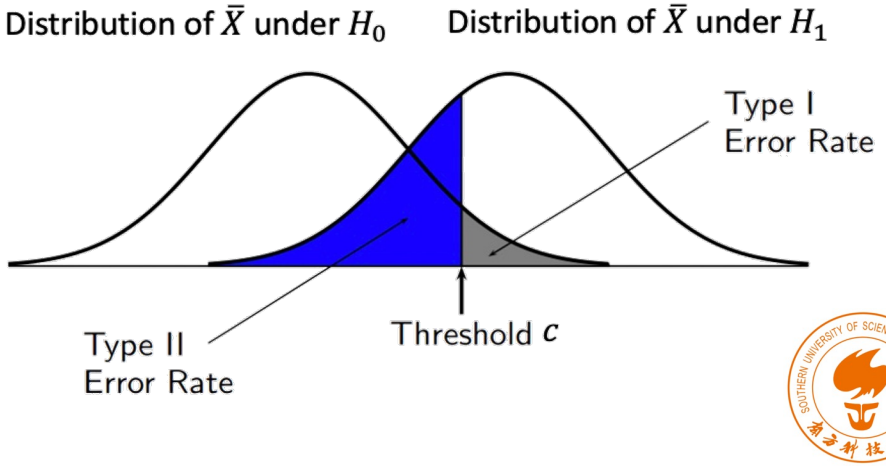
	H_0 is true	H_1 is true
Fail to reject H_0	Correct Decision	Type II Error (False Negative, 假阴性)
Reject H_0	Type I Error (False Positive, 假阳性)	Correct Decision



6.1 Overview and Basic Ideas

- Due to the randomness in sampling, errors cannot be completely avoided. However, we should try to make the probability of errors lower.
  - Type I error rate (第I类错误率): the probability of making a Type I error, $P(\boldsymbol{X} \in RR | H_0 \text{ is true})$.
  - Type II error rate (第II类错误率): the probability of making a Type II error, $P(\boldsymbol{X} \notin RR | H_1 \text{ is true})$.

- It is noted that these two error rates are traded-off against each other, i.e., they can not be reduced simultaneously. To be more specific:
  - For a given sample size and form of the RR, we cannot reduce the two error rates at the same time.
  - E.g., the RR for Example 6.2 is derived to be $\{\boldsymbol{X}: \bar{X} - 8 > c\}$, as the value of c changes, one of the two error rates becomes larger and the other becomes smaller.
  - Therefore, when determining the value of c, we need to balance between the two error rates.



6.1 Overview and Basic Ideas

- Recall that in hypothesis testing, we typically try to protect H_0 because it serves as the default or baseline assumption, ensuring that any departure from H_0 is justified by strong evidence.
- Hence, Type I error is typically considered to have more serious consequences than Type II error.
- Jerzy Neyman proposed the **significance test** (or **test of significance**, **显著性检验**) with the basic idea: try to minimize the Type II error rate provided that Type I error rate is below a given level.

Significance Test

- Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a simple random sample from the population $X \sim f(x; \theta_1, \ldots, \theta_k)$.
- For testing problem $H_0: \theta_l \in \Theta_{l0} \leftrightarrow H_1: \theta_l \in \Theta_{l1}$ and $\forall \alpha \in (0,1)$, if a test has rejection region RR such that

$$\mathrm{P}_{\boldsymbol{\theta}}(\boldsymbol{X} \in RR) \leq \alpha, \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ with } \theta_l \in \Theta_{l0},$$

α is the maximum acceptable Type I error rate

- then the test is called a **significance test of θ_l with significance level α (显著性水平为α的显著性检验)**, or simply **a test with level α (水平为α的检验)**.

- Commonly used values of α are 0.1, 0.05, 0.01.

6.1 Overview and Basic Ideas

- To minimize the Type II error rate, we set the Type I error rate at its maximum value α, then the rejection region can be determined.

Solution of Example 6.2

6.1 Overview and Basic Ideas

Solution of Example 6.2

6.1 Overview and Basic Ideas

Solution of Example 6.2

- Remarks:
  - The solution above suggest that when there are multiple parameter values in H_0 (e.g., H_0: $\mu_X \leq 8$), it can be simplified to a **simple null hypothesis (简单原假设)** involving only one parameter value which is closest to H_1 (e.g., H_0^*: $\mu_X = 8$).
  - This is why most hypothesis testing problems are formulated with a simple null hypothesis.

6.1 Overview and Basic Ideas

- Note that the decision made on whether to reject H_0 depends on the significance level α.
  - E.g., in the previous solution, if $\alpha = 0.01 \Longrightarrow z_\alpha = z_{0.01} = 2.326$, then the decision is "fail to reject H_0".
  - So, simply providing a conclusion on whether to reject H_0 may not be sufficient, we need to quantify our willingness to reject H_0. The well-known p-value (p值) serves as such a quantitative measure.
- p-value is a key concept in hypothesis testing, from Wikipedia:

> In statistical hypothesis testing, the **p-value** or **probability value** is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming the null hypothesis is correct.

  - p-value = P(observed or more extreme data|H_0 is true).
  - The idea of p-values has been used in the story of *The Lady Tasting Tea*, we obtained that "under the assumption that the lady has no ability to distinguish between the two ways, the probability of her correctly identifying the 4 cups where tea was added first is only 0.0143."

Assuming H_0 is true

Observed or more extreme data

The p-value

6.1 Overview and Basic Ideas

- A small p-value indicates that it is highly unlikely to observe the data we actually observed if H_0 is true. So, we tend to reject H_0 if p-value is small.
- The smaller the p-value, the stronger evidence we have against H_0.
- Given a significance level α, we would reject H_0 if p-value $< \alpha$, otherwise, we fail to reject H_0.
- A common misinterpretation of the p-value: p-value $= \mathrm{P}(H_0 \text{ is true} \mid \text{observed data})$. ☒

Solution of Example 6.2

6.1 Overview and Basic Ideas

- Finally, we introduce the concept of **statistical power (统计功效)** in hypothesis testing.
- Statistical power, or simply called the power, is the probability of rejecting H_0 when H_1 is true, i.e.,

$$\text{power} = \mathrm{P}(\boldsymbol{X} \in RR | H_1 \text{ is true}) = 1 - \beta,$$

- where β is the Type II error rate.
- A higher power suggests that there is a higher probability of enough evidence to claim a true positive finding. Three major factors that affect power:
  - **Significance level α**: a larger α reduces β and consequently increases the power.

    The two error rates are traded-off against each other
  - **The effect size**: the difference between the true parameter value and the values in H_0; the larger the difference, the greater the power.
  - **The sample size**: it is more difficult to detect an effect with a sample of smaller size as smaller samples are more likely to be affected by sampling error.
- We can determine the minimum sample size required to detect a given effect with a given significance level and a given power when designing an experiment or an observational study.

Chapter 6 Hypothesis Testing

- 6.1 Overview and Basic Ideas
- 6.2 General Process and Examples

6.2 General Process and Examples

- The general process of performing hypothesis testing is:

1. Formulate the hypotheses H_0 and H_1 based on the practical problem we are dealing with.

2. Choose a proper test statistic T and derive the distribution of T under H_0.

3. Determine the rejection region based on the distribution of T under H_0 and the significance level α.

4. Based on the sample observed values, compute the observed value of T and decide to reject or fail to reject H_0 at significance level α.

Or

3. Compute the observed value of T based on the sample observed values.

4. Obtain the p-value and compare it with the significance level α, consequently decide to reject or fail to reject H_0.

6.2 General Process and Examples

- The test statistic (检验统计量) T is used to discriminate between H_0 and H_1.
- When we verify a hypothesis about some parameter θ, T is usually obtained by a suitable transformation of its estimator $\hat{\theta}$ (typically a consistent estimator).
- If possible, obtain the exact distribution of T under H_0; otherwise, apply the Central Limit Theorem to derive an approximate distribution when sample size is large.
- General forms of parametric hypothesis testing: (θ is the parameter of interest, θ_0 is a value)
  - $H_0: \theta = \theta_0 \leftrightarrow H_1: \theta < \theta_0$, a one-sided, left-tail alternative hypothesis.
  - $H_0: \theta = \theta_0 \leftrightarrow H_1: \theta > \theta_0$, a one-sided, right-tail alternative hypothesis.

  单边检验
  - $H_0: \theta = \theta_0 \leftrightarrow H_1: \theta \neq \theta_0$, a two-sided alternative hypothesis. ← 双边检验
- The computation of the rejection region and the p-value would be a little bit different for different forms of tests.

6.2 General Process and Examples

Choosing a Test Statistic: A General Method

- Let $X_1, \ldots, X_n$ be a simple random sample from the population $X \sim f(x; \theta_1, \ldots, \theta_k)$. The null hypothesis is $H_0: \theta_l = \theta_{l0}$. $\hat{\theta}_l(X_1, \ldots, X_n)$ is an unbiased estimator of θ_l with variance $\sigma^2(\hat{\theta}_l) = \mathrm{Var}(\hat{\theta}_l)$.
- If $\sigma^2(\hat{\theta}_l)$ does not depend on any unknown parameter, then the test statistic is usually chosen as:

$$T = \frac{\hat{\theta}_l - \theta_{l0}}{\sigma(\hat{\theta}_l)}.$$

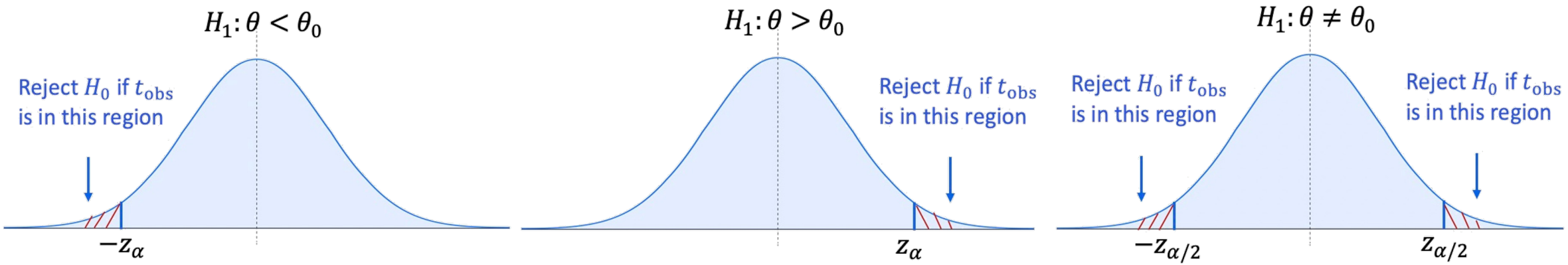
- If $\sigma^2(\hat{\theta}_l)$ depends on unknown parameters and $\hat{\sigma}^2(\hat{\theta}_l)$ is a consistent estimator of $\sigma^2(\hat{\theta}_l)$, the test statistic is usually chosen as:

$$T = \frac{\hat{\theta}_l - \theta_{l0}}{\hat{\sigma}(\hat{\theta}_l)}.$$

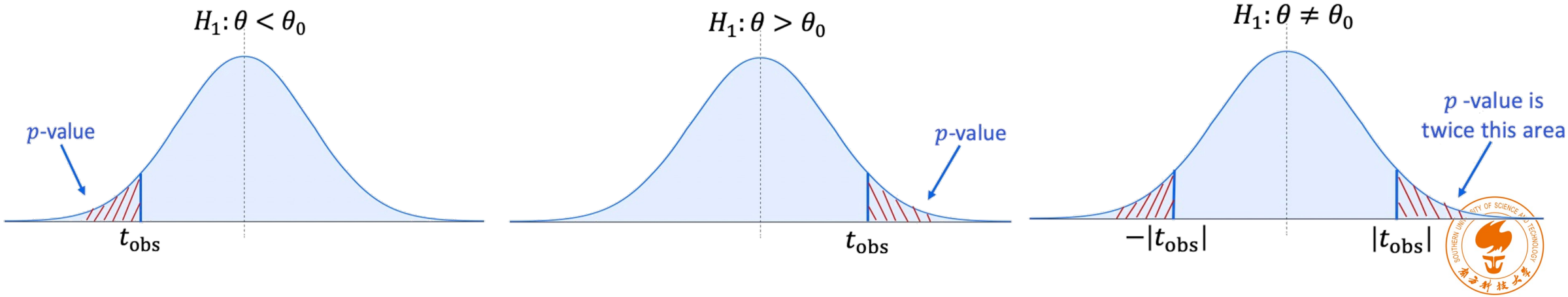
- Typically, the distribution of T under H_0 is approximately $\mathrm{N}(0, 1)$ (when the sample size is large). Under some special cases, the distribution of T is exactly $\mathrm{N}(0, 1)$.

6.2 General Process and Examples

- For the rejection region: (assuming the distribution of T under H_0 is $\mathrm{N}(0,1)$)



- For the p-value:



6.2 General Process and Examples

Solution of Example 6.2

6.2 General Process and Examples

Solution of Example 6.2

6.2 General Process and Examples

Solution of Example 6.2

6.2 General Process and Examples

Example 6.3

- A quality inspector finds 10 defective parts in a sample of 500 parts received from manufacturer A.
- Out of 400 parts from manufacturer B, she finds 12 defective ones.
- A computer-making company uses these parts in their computers and claims that the quality of parts produced by A and B is the same.
- At the 5% significance level, do we have enough evidence to disprove this claim?

Solution

6.2 General Process and Examples

Solution

6.2 General Process and Examples

- Next, let's look at an example about sample size determination.

Example 6.4

- A quality inspector is interested in testing whether the proportion of defective parts from manufacturer A is less than 2%, as promised by the manufacturer.
- He would like to randomly select n parts received from manufacturer A and perform a significance test at the 5% significance level.
- Suppose that the true defective proportion is 2.5%, find the sample size n that is necessary to achieve at least 0.90 power in the significance test.

Solution

6.2 General Process and Examples

Solution

6.2 General Process and Examples

Solution

6.2 General Process and Examples

Example 6.4 (Continued)

- How about the case if the true defective proportion is 5%? Find the sample size n that is necessary to achieve at least 0.90 power in the significance test.

Solution

6.2 General Process and Examples

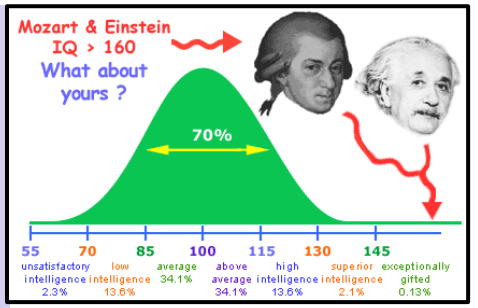
- The calculation in Example 6.4 are based on approximated distributions, so the sample size obtained may not be exact. However, the scale is correct.
- Please refer to the Monte Carlo experiments in Python.

StatisticalPower.ipynb or StatisticalPower.py

- Our experiments also verify that when the power increases with the significance level α.
- In cases when the exact distribution of the test statistic under H_1 can be obtained, the sample size calculated would be accurate. See the following example.

Example 6.5

- Suppose that the IQ test score of a randomly selected person is $X \sim \mathrm{N}(\mu, \sigma^2)$. A researcher is interested in testing whether the mean IQ test score is greater than 100.
- He would like to randomly select n people to take the IQ test and perform a significance test at the 5% significance level.
- If the true mean IQ test score is 105, and $\sigma^2 = 225$ is known, find the sample size n that is necessary to achieve at least 0.90 power in the significance test.



6.2 General Process and Examples

Solution

6.2 General Process and Examples

Solution

6.2 General Process and Examples

- Confidence intervals and significance tests are similar in that they both rely on the distribution of an estimator of the parameter of interest.
- Actually, for a two-sided testing problem $H_0: \theta = \theta_0 \longleftrightarrow H_1: \theta \neq \theta_0$, other than using the rejection region or the p-value, we can also make a decision based on the confidence interval of θ.
- The conclusion drawn from a two-tailed confidence interval is usually the same as the conclusion drawn from a two-tailed hypothesis test.
  - If the the $100(1-\alpha)\%$ confidence interval contains the hypothesized parameter value in H_0, i.e., θ_0, then a hypothesis test at the level α will almost always fail to reject H_0.
  - If the the $100(1-\alpha)\%$ confidence interval does not contain θ_0, then a hypothesis test at the level α will almost always reject H_0.
- Finally, we want to summarize the applications and deficiencies of hypothesis testing.
- Hypothesis testing is a powerful tool that provides a framework for assessing the statistical significance of hypotheses based on observational data.

6.2 General Process and Examples

- The application of hypothesis testing is very broad, encompassing fields such as scientific research, quality control, medicine and healthcare, economics and finance, psychology and social sciences, environmental studies, and market research.
- However, it is important to acknowledge that hypothesis testing have inherent deficiencies, and thus, a critical mindset should be maintained when interpreting the results of these tests.
- First, even we control the Type I error rate to be below α, it does not mean that the testing conclusion is 100% correct, as low-probability events can still occur.
- Second, hypothesis testing can only assess **statistical significance (统计显著性)**, but this does not imply **practical significance (实际显著性)**.
  - When the sample size is large enough, even minor differences can be detected, but these differences may not be significant in practical terms.
  - Practical significance should be evaluated by domain experts relevant to the research question.

6.2 General Process and Examples

睡眠与
睡在拐
时睡眠
睡眠时

原创 生

146人

专业

今天是
来越快
不足（
响全球

一项日
通过
和多
个有
体最佳
都会加
还巧妙
与睡眠

nature > s

Article | Op

Invert
and pl
study

Yanwei You
Xindong M

Scientific Re

282 Acces

图1.

研究者
为一个
龄没有
型中，
著关联

与正常
短睡眠
p=0.0
0.570
间过长
型，p
= 0.0

生物谷 >

研究人员使用两段线性回归模型计算出睡眠时间和基于对数的表型年龄之间关系的拐点为7小时(图4)，睡眠时间与表型年龄之间呈倒u型关系，这就表明：与以往研究相符（「最佳睡眠时长」竟不是"8小时"！Nature子刊：睡眠<7小时，心脏风险激增！>8小时死亡风险大幅增加！），**人体最佳的睡眠为7小时，少于或多于7个小时的睡眠都会加速表型年龄的增长。**

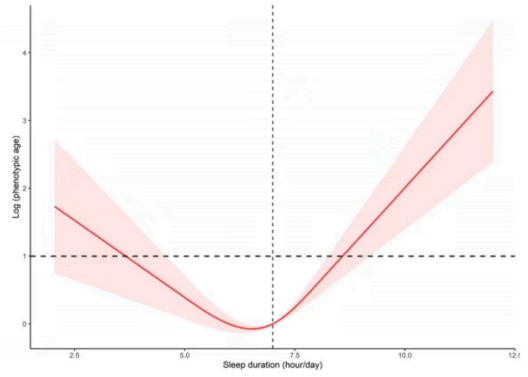


图3. 睡眠时间与LOG表型年龄的剂量-反应关系

睡懒觉是
新研究表
周末补觉
险

原创 生物谷

223人听过

专业 有趣

睡眠，是每
作，日落而
和养生哲学
大、娱乐活
「睡的晚」
成为不容忽
（CVD）、
生发展过程

近日，来
医院的研
一篇题为
catch-up
ease: Evi
and Nutri
2018"的
觉持续时
系，结果
小时的缺
超过 2小

> Sleep Health. 2024

Association
cardiovascu
Health and

Hong Zhu [1], Shouqu

Affiliations + expand

PMID: 38000943 DO

生物谷 >

研究人群的社会人口学和健康相关特征方面，在3400名成年受试者（男性1650人，女性1750人）中，333人（9.79%）被纳入CVD组。与非CVD组相比，CVD组的周末睡眠时间和周末追赶性睡眠时间均较短。然而，总的来说，CVD组比非CVD组睡眠持续时间更长。此外，年龄、性别、BMI、种族、家庭收入、吸烟、饮酒、高血压、糖尿病均与CVD相关。

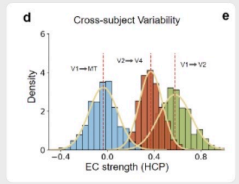
紧接着，研究人员通过单变量分析评估了周末和工作日睡眠时间与心血管疾病的关系，**结果表明，与周末追赶性睡眠时间<1小时的受试者相比，周末追赶性睡眠时间较长(> 1小时)的受试者的心血管疾病患病风险显著降低，竟达63%。**

Weekend sleep duration change	CVD (OR (95% CI))	*P*-value
No change in sleep duration	Reference	
Less sleep duration on weekends (> 1 h)	0.89 (0.43,1.85)	.74
Weekend catch-up sleep (> 1 h)	0.37 (0.23,0.58)	< .01

CVD, adults with cardiovascular disease; OR, adjusted odds ratio; 95% CI, 95% confidence interval of odds ratio.

6.2 General Process and Examples

2月1日 23:08



徐老师，请教一个很傻的统计上的问题哈

我们有批数据算出来的大脑的不同脑区的连接

这个左边蓝色的，我感觉明显就跟0比做独立样本ttest会不显著，但是因为样本量大，n=800，学生算出来就是显著的，很显著

2月1日 23:11

我算了一下，mean=-0.04, std=0.12, n=800，也确实是显著的。

但是很不符合我对大脑的理解。蓝色的这个分布看着，不应该显著呀。

统计上的显著性跟实际意义上的显著性是不一样的

那怎么来理解，统计上显著，但是直觉上看着一点儿也不显著呢?

在样本量很大的情况下，细小的差距可能在统计意义上就是显著的

没事没事，谢谢徐老师回复！这么晚，打扰您了。

匆少侠也在江湖：不好意思，刚刚看到～

这个说得应该比较清楚～

匆少侠也在江湖：

收到

这个practical significance也是会算出来一个p吗？然后我们report这个p value?

2月2日 00:17

不是哦，practical significance更像是具体问题的专家给出的判断

这个practical significance也是会算出来一个p吗?...

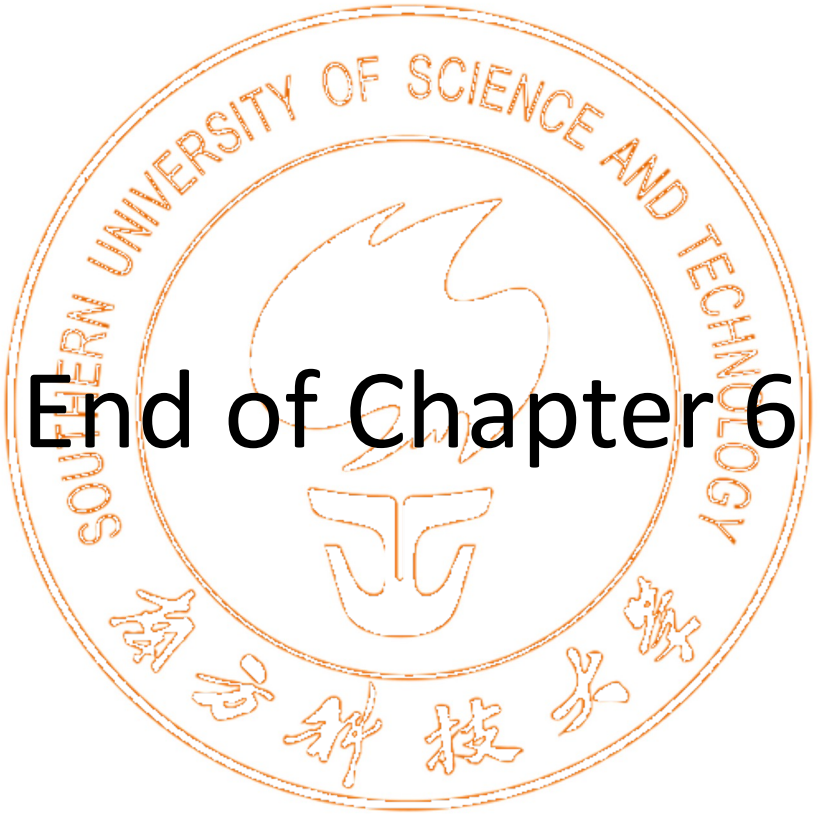
2月2日 00:18

statistical significance是从概率的角度认为观察到的差距不是由于随机抽样导致的，而是真实存在的差距

practical significance是在判断这个差距在具体的场景下是不是有实际意义

哦哦哦

practical significance也就是说，有没有实际意义，由专家说了算



End of Chapter 6