# Computational Methods for Statistics (VU) (706.026)

Elisabeth Lex

ISDS, TU Graz

Jan 24, 2023

# Outline

1. Introduction
2. Approach to Hypothesis Testing
3. Null and Alternative Hypotheses
4. Z-Test
5. Errors in Hypothesis Testing
6. p-values
7. Permutation Test
8. Multiple Testing

# Readings

Fahrmeir et. al, Statistik: der Weg zur Datenanalyse (Chapter 10)

Wasserman, All of Statistics (Chapter 10)

Chernick, Bootstrap Methods: A Guide for Practitioners and Researchers (Chapter 3)

# Readings

MIT OpenCourseware: Introduction to Probability and Statistics
https://ocw.mit.edu/courses/mathematics/
18-05-introduction-to-probability-and-statistics-spring-2
index.htm

MIT OpenCourseware: Statistics for Application
https://ocw.mit.edu/courses/mathematics/
18-650-statistics-for-applications-fall-2016/index.htm

# Hypothesis Testing

- Two major areas of statistical inference
  1. Parameter estimation
  2. Testing of hypotheses
- The goal is often to estimate the parameters, e.g. the user satisfaction with a new version of an app
- More often we want to use this estimate for a certain purpose
- We want to compare user satisfaction with the previous version of the app
- This is a common situation in (empirical) research

# Hypothesis Testing

### Example 14 (User Satisfaction)

We perform a user study including $n = 36$ users to estimate the user satisfaction with the new app version. The average user satisfaction in our user study is $7.1$. The current version of the app is in use for a couple of years now, and we know from numerous user studies and online rating sites that the average user satisfaction with the current version is about $6.5$. The statistical model here is: we are dealing with **two populations** of users, those using the current version and those using the new version of the app. Traditionally, to answer this type of the question we set up a **test to check if there is a substantial evidence that one mean is greater than the other mean**.

# Hypothesis Testing

### Example 14 (User Satisfaction (contd.))

We formulate the hypothesis that the new version is no better than the current version of the app. Generally, we hope that with our test we can reject this hypothesis. There is an indication that the users of the new version are more satisfied, but as previously discussed, **the sample mean is not sufficient, and we need to estimate the sample variability**.

# Hypothesis Testing

## Example 14 (User Satisfaction (contd.))

Suppose now that the sample deviation $S_n/\sqrt{n}$ in our user study with the new version of the app is $1$. Then a $95\%$ Z-score confidence interval (assuming normality of the user population of the new version of the app) is $(5.14, 9.06)$. Thus, the sample mean of $7.1$ could easily have from a population with mean smaller than $6.5$ (user satisfaction average with the current app version). Thus, we have no strong ground for rejecting the hypothesis. If, on the other hand $S_n/\sqrt{n}$ were $0.167$, then Z-score confidence intervals would be $(6.77, 7.43)$ and we could confidently reject the hypothesis and pronounce the new version of the app to be superior with respect to user satisfaction.

# Definition of hypothesis

## Definition 8 (Statistical hypothesis)

A *statistical hypothesis* $H$ is a conjecture about the distribution of one or more random variables. If the statistical hypothesis completely specifies the distribution, then it is called *simple*; otherwise it is called *composite*.

## Example 15 (Simple vs. composite hypothesis)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ with $X_1 \sim N(\theta, 25)$, where $\theta$ is an unknown population mean. The hypothesis $H : \theta = 17$ is simple because it completely specifies the distribution. On the other hand, the hypothesis $H : \theta \leq 17$ is composite because it does not completely specify the distribution.

# Definition of hypothesis test

### Definition 9 (Test of a statistical hypothesis)

A *test* of a statistical hypothesis $H$ is a rule or procedure for deciding whether to **reject** $H$.

### Example 16 (Possible test)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ with $X_1 \sim N(\theta, 25)$, where $\theta$ is an unknown population mean. Consider $H : \theta \leq 17$. One possible test is as follows: reject $H$ if and only if $\overline{X}_n > 17 + 5/\sqrt{n}$.

# Approaching Hypothesis Tests

- Basic idea: probabilistic “proof” by contradiction
  1. We assume that a hypothesis $H$ holds
  2. We compute how likely is our data if $H$ holds
  3. If the data is not “very” likely we reject $H$
- For computation in step 2: we use WLNN, CLT, etc.

# Hypothesis Test Heuristics

## Example 17 (Unfair Coin)

We toss a coin $n = 80$ times and get heads $54$ times. Can we conclude that the coin is significantly unfair?

# Hypothesis Test Heuristics

## Example 17 (Unfair Coin)

We toss a coin $n = 80$ times and get heads $54$ times. Can we conclude that the coin is significantly unfair?

## Solution.

We have $X_1, X_2, \ldots, X_n$, $n = 80$, $X_1 \sim Bernoulli(p)$.
$\overline{X}_n = 54/80 = 0.675$
$H$ : coin is fair $\implies H : p = 0.5$
By CLT we know:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} \approx N(0,1)$$

# Hypothesis Test Heuristics

## Solution (contd.)

With our data we obtain:

$$\frac{0.675 - 0.5}{\sqrt{0.5(1-0.5)}/\sqrt{80}} = 3.1305$$

$3.1305$ is not a plausible realization of a r.v. $Z \sim N(0,1)$ because:

$$P(Z > 3.1305) = 1 - \Phi\left(3.1305\right) = 0.0009$$

Conclusion: It **seems quite reasonable** to reject $H$.

## Remark 10 (p-value)

What we computed in the last step is a so-called *p-value*.

# Hypothesis Test Heuristics

## Example 18 (Unfair Coin)

We toss a coin $n = 30$ times and get heads $13$ times. Can we conclude that the coin is significantly unfair?

## Solution.

As previously:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} \approx N(0,1)$$

# Hypothesis Test Heuristics

## Solution (contd.)

With $\overline{X}_n = 13/30 \approx 0.433$:

$$\frac{0.433 - 0.5}{\sqrt{0.5(1-0.5)}/\sqrt{30}} = -0.7303$$

$-0.7303$ is a plausible realization of a r.v. $Z \sim N(0,1)$, e.g.

$$P(Z > -0.7303) = 1 - \Phi\left(-0.7303\right) = 0.7674$$

Conclusion: Our data **does not suggest** to reject $H$.

## Notebook 10 (Unfair Coin)

unfair_coin.ipynb

# Null and Alternative Hypotheses

- We go now back to the question of comparing two populations
- First user group: users of the new app version
- Second user group: users of the current app version
- Typically we will formulate **two mutually exclusive hypotheses:**
    1. **The Null Hypothesis $H_0$:** The user satisfaction is the same in the two groups.
    2. **The Alternative Hypothesis $H_1$:** The user satisfaction is not the same in the two groups.

## Remark 11 (The Null vs. The Alternative)

$H_0$ is always **cautious default**: we won't claim claim the coin is unfair unless we have a strong evidence! You can also think about $H_0$ as "nothing interesting is happening" and of $H_1$ as "something interesting is going on". For example, in a legal trial we always assume someone is innocent ($H_0$) unless the evidence strongly suggest that the person is guilty ($H_1$).

# Null and Alternative Hypotheses

## Definition 10 (Null and Alternative Hypotheses)

Let $X_1, X_2, \dots, X_n$ be a random sample of size $n$ from density $f(\cdot;\theta)$, with $\theta \in \Theta$. Let $\Theta_0$ and $\Theta_1$ be disjoint subsets of $\Theta$. We consider two hypothesis about $\theta$:

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

We call $H_0$ the *null hypothesis* and $H_1$ the *alternative hypothesis*.

## Example 18 (Unfair Coin)

In the case of coin tosses we test:

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p \neq 0.5 \end{aligned}$$

# Null and Alternative Hypotheses

**Remark 12 (Hypothesis test)**

If we believe that the true $\theta$ is either in $\Theta_0$ or in $\Theta_1$ we *test* $H_0$ *against* $H_1$. We always decide whether to *reject* $H_0$ by looking for evidence *against* $H_0$ in the data. $H_0$ and $H_1$ do not play a symmetric role: the data is only used to try to disprove $H_0$.

- **Important:** In the case of a lack of evidence, it does not mean that $H_0$ is true
- It simply means we do not have enough evidence to reject $H_0$!

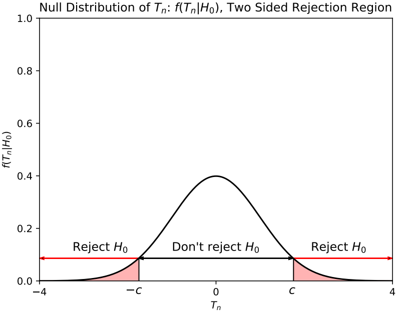
# Systematic Approach to Hypothesis Testing

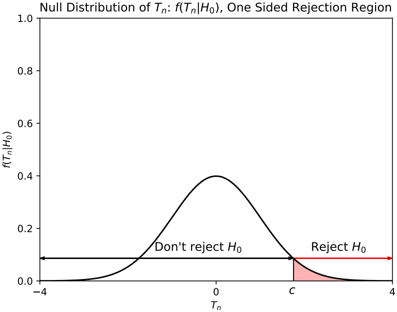
- Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from the density $f(\cdot; \theta)$
- $H_0$ and $H_1$ defines as previously
- We test $H_0$ against $H_1$ by finding an appropriate **rejection region** $R = \{T_n > c\}$:

$$\begin{aligned} t_n \in R, \text{ i.e. } T_n > c \quad &\Longrightarrow \quad \text{reject } H_0 \\ t_n \notin R, \text{ i.e. } T_n \leq c \quad &\Longrightarrow \quad \text{don't reject } H_0 \end{aligned}$$

- $T_n$ is a **test statistic** and $c$ is a **critical value**
- Problem of hypothesis testing: find an appropriate $T_n$ and $c$

# Rejection Region


Null Distribution of $T_n$: $f(T_n|H_0)$, Two Sided Rejection Region


Null Distribution of $T_n$: $f(T_n|H_0)$, One Sided Rejection Region

# Systematic Approach to Hypothesis Testing

- We assume that the null hypothesis $H_0$ holds
- We pick a test statistic $T_n$
- We decide what is **strong evidence** for us by selecting a **significance level** $\alpha$
- We compute the null sampling distribution of $T_n$, which is always conditioned on $H_0$: $f(T_n|H_0)$
- Using $f(T_n|H_0)$ and $\alpha$ we compute the critical value $c$ and determine the rejection region $R$
- We reject $H_0$ if $t_n \in R$, otherwise we don't reject $H_0$

# Z-Test

## Z-Test

Let $X_1, X_2, \dots, X_n$ be a random sample of size $n$ with $X_1 \sim N(\mu, \sigma^2)$, $\mu$ unknown, $\sigma$ known.

| | |
|---|---|
| **Hypotheses**: | $H_0 : X_1 \sim N(\mu_0, \sigma^2)$<br>$H_1 : \mu \neq \mu_0$ (two sided alternative)<br>or $H_1 : \mu > \mu_0$ (one sided alternative) |
| **Test statistic**: | $Z_n$ |
| **Null distribution**: | Assuming $H_0$: $Z_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ |
| **Significance level**: | $\alpha$ |
| **Critical value**: | $c = z_{1-\alpha/2}$ (two sided alternative)<br>$c = z_{1-\alpha}$ (one sided alternative) |
| **Rejection region**: | $R = \{\lvert Z_n \rvert > z_{1-\alpha/2}\}$ (two sided alternative)<br>$R = \{Z_n > z_{1-\alpha}\}$ (one sided alternative) |

# Rejection Region

## Example 19 (Two sided rejection region)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n = 64$ with $X_1 \sim N(\mu, 10^2)$, $\mu$ is unknown. $\overline{x}_n = 6.25$. We want to test $H_0$ against $H_1$:

$$\begin{aligned} H_0 &: \mu = 5 \\ H_1 &: \mu \neq 5 \end{aligned}$$

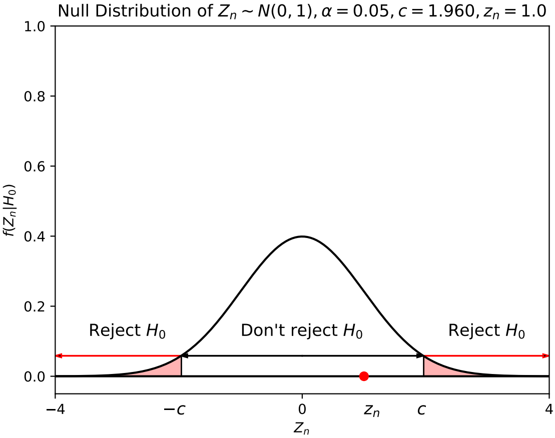
Can we reject $H_0$ at the significance level $\alpha = 0.05$?

# Rejection Region

## Solution.

We know that: $Z_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
Therefore, we compute $c = z_{1-\alpha/2}$ as two sided critical value.
We obtain $c = 1.96$, our rejection region is $R = \{|Z_n| > 1.96\}$.
Our $z_n = 1$ and it does not fall into $R$.
$\implies$ We don't reject $H_0$ at significance level $\alpha = 0.05$

# Rejection Region



# Rejection Region

### Example 20 (One sided rejection region)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n = 9$ with $X_1 \sim N(\mu, 15^2)$, $\mu$ is unknown. $\overline{x}_n = 112$. We want to test $H_0$ against $H_1$:

$$\begin{aligned} H_0 &: \mu = 100 \\ H_1 &: \mu > 100 \end{aligned}$$

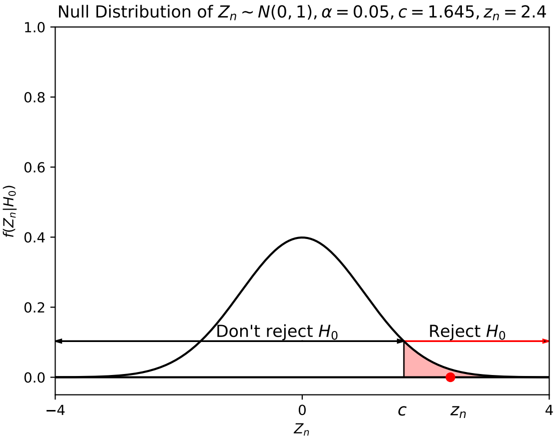
Can we reject $H_0$ at the significance level $\alpha = 0.05$?

# Rejection Region

## Solution.

We know that: $Z_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
Therefore, we compute $c = z_{1-\alpha}$ as one sided critical value.
We obtain $c = 1.645$, our rejection region is $R = \{Z_n > 1.645\}$.
Our $z_n = 2.4$ and it falls into $R$.
$\implies$ We reject $H_0$ at significance level $\alpha = 0.05$

# Rejection Region



# The Null is a Cautious Choice!

## Example 21 (Two Coins)

We have two coins:

1. $C_1$ with probability $p = 0.5$ of heads
2. $C_2$ with probability $p = 0.6$ of heads

We pick one coin at random, flip it $8$ times and get 6 heads. Answer the following three questions:

1. $H_0$: The coin is $C_1$ vs. $H_1$: The coin is $C_2$. Do we reject $H_0$ at the significance level $\alpha = 0.05$?
2. $H_0$: The coin is $C_2$ vs. $H_1$: The coin is $C_1$. Do we reject $H_0$ at the significance level $\alpha = 0.05$?
3. Are the answers to $1$ and $2$ paradoxical?

# The Null is a Cautious Choice!

## Solution.

To find the rejection regions we need the tables for $B(8, 0.5)$ and $B(8, 0.6)$:

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $p(k\|p = 0.5)$ | 0.004 | 0.031 | 0.109 | 0.219 | 0.273 | 0.219 | 0.109 | 0.031 | 0.004 |
| $p(k\|p = 0.6)$ | 0.001 | 0.008 | 0.041 | 0.124 | 0.232 | 0.279 | 0.209 | 0.090 | 0.017 |

# The Null is a Cautious Choice!

## Solution.

$H_0$: The coin is $C_1$ vs. $H_1$: The coin is $C_2$. $\alpha = 0.05$

$0.6 > 0.5 \implies$ one sided (right) rejection region

# The Null is a Cautious Choice!

## Solution.

$H_0$: The coin is $C_1$ vs. $H_1$: The coin is $C_2$. $\alpha = 0.05$

$0.6 > 0.5 \implies$ one sided (right) rejection region

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $p(k\|p = 0.5)$ | 0.004 | 0.031 | 0.109 | 0.219 | 0.273 | 0.219 | 0.109 | 0.031 | 0.004 |
| $p(k\|p = 0.6)$ | 0.001 | 0.008 | 0.041 | 0.124 | 0.232 | 0.279 | 0.209 | 0.090 | 0.017 |

Since we got $6$ heads we do not reject $H_0$.

# The Null is a Cautious Choice!

## Solution (contd.)

$H_0$: The coin is $C_2$ vs. $H_1$: The coin is $C_1$. $\alpha = 0.05$
$0.5 < 0.6 \implies$ one sided (left) rejection region

# The Null is a Cautious Choice!

## Solution (contd.)

$H_0$: The coin is $C_2$ vs. $H_1$: The coin is $C_1$. $\alpha = 0.05$
$0.5 < 0.6 \implies$ one sided (left) rejection region

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $p(k\|p = 0.5)$ | 0.004 | 0.031 | 0.109 | 0.219 | 0.273 | 0.219 | 0.109 | 0.031 | 0.004 |
| $p(k\|p = 0.6)$ | 0.001 | 0.008 | 0.041 | 0.124 | 0.232 | 0.279 | 0.209 | 0.090 | 0.017 |

Since we got 6 heads we do not reject $H_0$.
Paradox: the fact that we don't reject $C_1$ in favor of $C_2$ or $C_2$ in favor of $C_1$ reflects asymmetry of hypothesis testing. The null hypothesis is a cautious choice. We only reject $H_0$ if the data is extremely unlikely when we assume $H_0$. This is not the case for either $C_1$ or $C_2$.

# One More Example

## Example 22

Suppose $n = 3$ radar guns are set up along a stretch of road to catch people driving over the speed limit of $50$ km/h. Each radar gun is known to have a normal measurement error $N(0, 5^2)$. For a passing car, let $\overline{X}_n$ be the average of the $n$ readings. The police's default assumption is that car is not speeding (cautious choice: innocent until proven guilty!).

1. Describe the above story in the context of statistical hypothesis testing. Are the most natural null and alternative hypotheses simple or compound?
2. The police would like to set a threshold on $\overline{X}_n$ for issuing tickets so that no more than $4\%$ of the tickets are given in error. Use the statistical hypothesis testing described in $1$ to determine what threshold the police should set when using $n = 3$ radars.

# One More Example

## Solution.

Let $\mu$ be the actual speed of a given driver. Hence, $X_1 \sim N(\mu, 5^2)$ and therefore $\overline{X}_n \sim N(\mu, 5^2/3)$. The most natural hypothesis are:

# One More Example

## Solution.

Let $\mu$ be the actual speed of a given driver. Hence, $X_1 \sim N(\mu, 5^2)$ and therefore $\overline{X}_n \sim N(\mu, 5^2/3)$. The most natural hypothesis are:

$H_0$: $\mu \leq 50$, i.e., the driver is not speeding

$H_1$: $\mu > 50$, i.e., the driver is speeding

Both hypotheses are composite, however we can work with $H_0$: $\mu = 50$, which is simple.

# One More Example

## Solution.

The null distribution is $\overline{X}_n \sim N(50, 5^2/3)$ and we are looking for one sided rejection region ($H_1$: $\mu > 50$) with $\alpha = 0.04$:

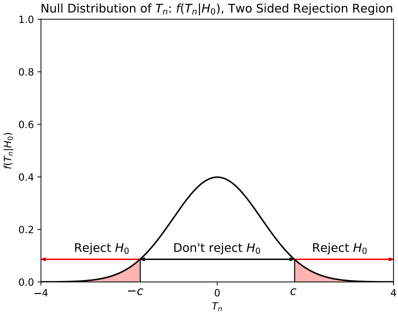
$$c = \mu + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 50 + 1.7507\frac{5}{\sqrt{3}} = 55.054 \text{ km/h}$$

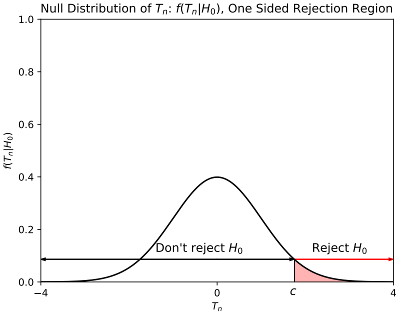
## Notebook 11 (Radar Guns)

`radar_guns_z_test.ipynb`

# What is the Significance Level?

- The probability that we will reject $H_0$ even if it is true





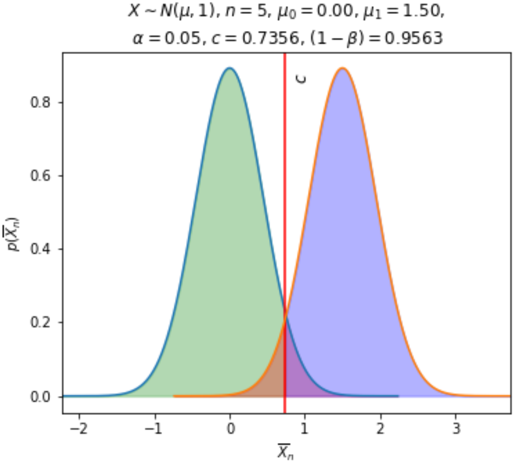
- When we reject an $H_0$ that is true we are making a **false positive**

Notebook 11 (Radar Guns)

`radar_guns_z_test.ipynb`

# Errors

- What other kind of error can we make?



- When we don't reject an $H_0$ when $H_1$ is in fact true: **false negative**

# Errors

| | $H_0$ true | $H_0$ false |
|---|---|---|
| Don't reject $H_0$ | true negative (tn)<br>prob. $= 1 - \alpha$<br>Correct inference | false negative (fn)<br>prob. $= \beta$<br>Type II error |
| Reject $H_0$ | false positive (fp)<br>prob. $= \alpha$<br>Type I error | true positive (tp)<br>prob. $= 1 - \beta$ (Power)<br>Correct inference |

# Errors

- We can not simultaneously reduce false positives and false negatives
- If we reduce $\alpha$ we reduce false positives but we move critical value
- Then more probability mass will be on the other side of the critical value if $H_1$ is true
- Hence, we will increase false negatives
- Similar behavior if we want to control for false negatives, therefore the convention is to control for false positives

Notebook 12 (FP vs FN Trade-Off)

`radar_guns_roc.ipynb`

# Scientific vs. Statistical Significance

- When we reject $H_0$ we say that the result is **statistically significant**
- A result might be statistically significant but the effect size might be small
- In such a case, we have a statistical significance but no **scientific or practical significance**:

$$\text{statistical significance} \not\Rightarrow \text{scientific significance}$$

- In such cases, confidence intervals are more informative than hypothesis tests

# Scientific vs. Statistical Significance

## Example 23 (Scientific vs. Statistical Significance)

Suppose we extend an app by adding two features and perform two separate user satisfaction studies ($n = 100$ in both studies). For the first feature we obtain an average user satisfaction of $6.6$, and for the second of $7.1$, $\sqrt{S_n/n} = 0.05$ in both studies. The old version of the app had the average user satisfaction of $6.5$. For both studies we define $H_0$: no improvement in the new versions. Can we reject $H_0$ for both features at the significance level $\alpha = 0.05$? Compare these results with 95% confidence intervals for the sample means for both features.

# Scientific vs. Statistical Significance

## Solution.

For both features:

$$\begin{aligned} H_0 &: \quad \mu = 6.5 \\ H_1 &: \quad \mu > 6.5 \end{aligned}$$

We perform one sided Z-Test by computing the critical value:

$$c = \mu + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 6.5 + 1.645 \cdot 0.05 = 6.58225$$

Thus, for both features we reject $H_0$. The 95% Z-score confidence intervals are:
Feature 1: $6.6 \pm 0.098 = (6.502, 6.698)$
Feature 2: $7.1 \pm 0.098 = (7.002, 7.198)$

# Scientific vs. Statistical Significance

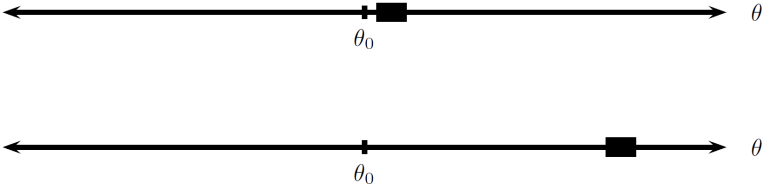


Figure: From All of Statistics by Wasserman

# p-values

- Previous example illustrated that reporting that we reject $H_0$ is not always informative
- Instead we could ask for every $\alpha$ whether the test rejects at that level
- Generally, if the test rejects at $\alpha$ it will also reject at level $\alpha' > \alpha$
- Hence, there is a smallest $\alpha$ at which the test rejects
- We call this number **p-value**
- Note that the p-value is also a r.v. and has a sampling distribution

# p-value

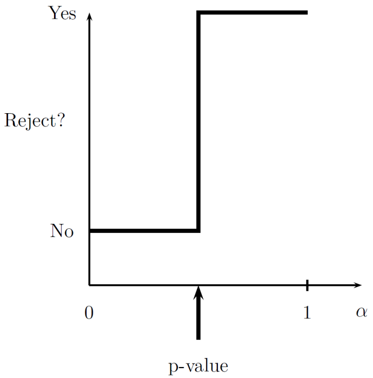


Figure: From All of Statistics by Wasserman

# p-value

## Warning!

- **Warning!** A large p-value is not strong evidence in favor of $H_0$. A large p-value can occur because $H_0$ is true or $H_0$ is false but the test has low power.
- **Warning!** p-value is not $P(H_0|Data)$, i.e. the p-value is not the probability that the null hypothesis is true
- **Correct interpretation:** The p-value is the probability under $H_0$ of observing a value of test statistic the same as or more extreme than what we actually observed!

# p-value

### Theorem 4 (Distribution of p-value)

*If the test statistic has a continuous distribution, then under $H_0$: $\theta = \theta_0$, the p-value has a uniform distribution $Unif(0,1)$. Therefore, if we reject $H_0$ when the p-value is less than $\alpha$, the probability of false positive (type I error) is $\alpha$.*

### Remark 13 (Interpretation)

If $H_0$ is true, the p-value is like a random draw from a $Unif(0,1)$ distribution. If $H_1$ is true, the distribution of the p-value will tend to concentrate closer to $0$.

### Notebook 10 (Unfair Coin)

`unfair_coin.ipynb`

# The Permutation Test

- The permutation test is a resampling method for testing whether two distributions are the same
- This test is exact: it is not based on large sample approximations
- Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$, $X_1 \sim F_X$
- Let $Y_1, Y_2, \ldots, Y_m$ be a random sample of size $m$, $Y_1 \sim F_Y$
- With permutation test we are testing:

$$H_0 : F_X = F_Y \text{ against } H_1 : F_X \neq F_Y$$

# Basic Idea of the Permutation Test

- Let $T_N = t(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ be some test statistic, where $N = n + m$
- E.g. $T_n = |\overline{X}_n - \overline{Y}_m|$
- We consider all $N!$ permutations of the data $X_1, \ldots, X_n, Y_1, \ldots, Y_m$
- For each permutation we compute the test statistics $T$
- We denote these values with $T_1^*, \ldots, T_{N!}^*$

# Basic Idea of the Permutation Test

- How likely are each of the $T_1^*, \ldots, T_{N!}^*$ under the $H_0$?

# Basic Idea of the Permutation Test

- How likely are each of the $T_1^*, \ldots, T_{N!}^*$ under the $H_0$?
- Equally likely!
- The distribution $P_0$ that puts $1/N!$ mass on each $T_j^*$ is called the **permutation distribution** of $T$
- Let $t_n$ be the observed value of the test statistic
- Assuming we reject when $T$ is large, the p-value of the permutation test:

$$p - value = P_0(T* > t_N) = \frac{1}{N!} \sum_{j=1}^{N!} I(T_j^* \geq t_N)$$

# Toy Example

## Example 24 (Toy Example: Permutation Test)

Suppose the data are: $(X_1, X_2, Y_1) = (3, 9, 1)$. Let $T(X_1, X_2, Y_1) = |\overline{X}_n - \overline{Y}_m|$, i.e. $t_N = 5$. Compute the p-value of the test statistic $T$.

## Solution.

| permutation | $T^*$ | $P_0(T^*)$ |
|---|---|---|
| $(3, 9, 1)$ | $5$ | $1/6$ |
| $(1, 3, 9)$ | $7$ | $1/6$ |
| $(1, 9, 3)$ | $2$ | $1/6$ |
| $(3, 1, 9)$ | $7$ | $1/6$ |
| $(9, 1, 3)$ | $2$ | $1/6$ |
| $(9, 3, 1)$ | $5$ | $1/6$ |

p-value: $P(T^* \geq 5) = 4/6$

# Basic Idea of the Permutation Test

- Usually, it is not practical to evaluate all $N!$ permutations
- We can approximate the p-value by simulating random permutations
- The fraction of times $T_j^* > t_N$ among these samples approximate the p-value

# Algorithm for Permutation Test

1. Compute the observed value point of the test statistic: $t_N = t(x_1, \ldots, x_n, y_1, \ldots, y_m)$
2. Randomly permute the data. Compute the statistic $T^*$ again using the permuted data
3. Repeat the previous step $b$ times to obtain $T_1^*, \ldots, T_b^*$
4. The approximate p-value:

$$\frac{1 + \sum_{j=1}^{b} I(T_j^* \geq t_N)}{b+1}$$

# Are beer drinkers more attractive to mosquitos?

- We permute labels since under $H_0$ they are meaningless

Beer

| | | | |
|---|---|---|---|
| 27 | 19 | 20 | 20 |
| 23 | 17 | 21 | 24 |
| 31 | 26 | 28 | 20 |
| 27 | 19 | 25 | 31 |
| 24 | 28 | 24 | 29 |
| 21 | 21 | 18 | 27 |
| 20 | | | |

Water

| | | | |
|---|---|---|---|
| 21 | 19 | 13 | 22 |
| 15 | 22 | 15 | 22 |
| 20 | 12 | 24 | 24 |
| 21 | 19 | 18 | 16 |
| 23 | 20 | | |

$\overline{X}_n = 23.6000$ $\quad$ $\overline{Y}_m = 19.2222$

$$\overline{X}_n - \overline{Y}_m = 4.3778$$

## Notebook 13 (Beer/Water)

`beer_water.ipynb`

# Driving Behavior

## Example 25

Suppose we have a list of cities with the car driver velocities on weekdays and weekends. Test if the driving behavior in each city is different on weekdays as compared to weekends.

## Solution.

We take $H_0$: no difference in driving behavior and test it against $H_1$: driving behavior is different at weekdays and weekends for a given city. We use permutation test for the difference in means.

## Notebook 14 (Driving Behavior)

`car_drivers.ipynb`

# Multiple Testing

- In some situations we conduct multiple tests
- In the previous example we tested for e.g. multiple cities
- Suppose each test is conducted at level $\alpha$
- For any one test, the chance of falsely rejecting the null is $\alpha$
- But the chance of at least one false rejection is much higher
- This is the **multiple testing problem**
- The problem comes up in many data mining applications where we sometimes test thousands or millions of hypotheses

# Bonferroni Correction

- Consider $m$ hypothesis tests: $H_{0i}$ vs. $H_{1i}$, $i = 1, \dots, m$
- We denote with $P_1, \dots, P_m$ the $m$ p-values for these tests:

### Bonferroni Correction

Given p-values $P_1, \dots, P_m$ reject null hypothesis $H_{0i}$ if:

$$P_i < \frac{\alpha}{m}.$$

# Bonferroni Correction

### Theorem 5 (Bonferroni Correction)

*Using the Bonferroni correction, the probability of falsely rejecting any null hypothesis is less thanb or equal to* $\alpha$

# Bonferroni Correction

## Theorem 5 (Bonferroni Correction)

*Using the Bonferroni correction, the probability of falsely rejecting any null hypothesis is less thanb or equal to* $\alpha$

## Proof.

Let $R$ be the event that at least one null hypothesis is falsely rejected. Let $R_i$ be the event that the $i^{th}$ null hypothesis is falsely rejected:

$$P(R) = P\left(\bigcup_{i=1}^{m} R_i\right) \leq \sum_{i=1}^{m} P(R_i) = \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha$$

□

## Notebook 14 (Driving Behavior)

`car_drivers.ipynb`