Statistics for Data Science

MSc Data Science WiSe 2019/20

Prof. Dr. Dirk Ostwald
(10) Hypothesis testing
Bibliographic remarks

The presented material follows Ostwald et al. (2019, Supplementary Material, Section 2) for the majority of test-theoretical concepts.
Hypothesis testing

- Foundations
- Significance levels and p-values
- Examples
  - The T test
  - The Wald test
  - The likelihood ratio test
Hypothesis testing

- **Foundations**
- Significance levels and p-values
- Examples
  - The T test
  - The Wald test
  - The likelihood ratio test
Foundations

Definition (Test hypotheses)

Let \( \mathcal{P} \) denote a parametric statistical model governing the distribution of a random sample \( X = (X_1, ..., X_n) \) with a PMF or PDF \( p_\theta \), let \( \mathcal{X} \) denote the outcome space of the data such that \( x \in \mathcal{X} \), and let \( \Theta \) denote the parameter space of the model. Further, let \( \Theta_0 \) and \( \Theta_1 \) denote a partition of the parameter space, such that \( \Theta = \Theta_0 \cup \Theta_1 \) and \( \Theta_0 \cap \Theta_1 = \emptyset \). Then a test hypothesis is a statement about the parameter governing the distribution of \( X \) in relation to the parameter space subsets \( \Theta_0 \) and \( \Theta_1 \). Specifically

- \( H_0 : \theta \in \Theta_0 \) is referred to as null hypothesis, and
- \( H_1 : \theta \in \Theta_1 \) is referred to as alternative hypothesis.

Remarks

- We assume that both null and alternative hypothesis exist.
- The null hypothesis is not necessarily the hypothesis \( \Theta_0 = \{0\} \).
- The null hypothesis is the hypothesis one is willing to reject.
Definitions (Simple and composite hypotheses)

- A *simple hypothesis* refers to a subset of parameter space containing a single element, such as \( \Theta_0 := \{ \theta_0 \} \)
- A *composite hypothesis* refers to a subset of parameter space containing more than one element, such as \( \Theta_0 := \mathbb{R}_{\leq 0} \).

Remark

- The often encountered null hypothesis \( \Theta_0 = \{ 0 \} \) is an example for a simple hypothesis and is also referred to as *nil hypothesis*.
Definition (Test)

Given a test hypotheses scenario, a test is a mapping from the data outcome space to the set \( \{0, 1\} \), formally

\[
\phi(X) : \mathcal{X} \rightarrow \{0, 1\}, \; x \mapsto \phi(X)(x), \tag{1}
\]

where

- 0 represents the act of not rejecting the null hypothesis.
- 1 represents the act of rejecting the null hypothesis.

Remarks

- Rejecting the null hypothesis \( \Leftrightarrow \) Accepting the alternative hypothesis.
- Not rejecting the null hypothesis \( \Leftrightarrow \) Rejecting the alternative hypothesis.
- Accepting the null hypothesis \( \Leftrightarrow \) Rejecting the alternative hypothesis.
- Because \( X \) is random, \( \phi(X) := \phi(X = \cdot) \) is also random.
## Definition (Standard test)

A *standard test* is given by the composition of a *test statistic* 

\[
\gamma(X) : \mathcal{X} \to \mathbb{R}
\]  

(2)

and a *decision rule* 

\[
\delta(\gamma(X)) : \mathbb{R} \to \{0, 1\}
\]  

(3)

A standard test can be written as 

\[
\phi(X) = \delta(\gamma(X)) : \mathcal{X} \to \{0, 1\}
\]  

(4)

### Remarks

- Because \( X \) is random, both \( \gamma(X) := \gamma(X = \cdot) \) and \( \delta(\gamma(X)) \) are random.
Definition (Test rejection region)

The subset of the test statistic’s outcome space for which the test takes on the value 1 is referred to as the rejection region $R$ of the test. Formally,

$$R := \{\gamma(X) \in \mathbb{R} | \phi(X) = 1\} \subset \mathbb{R}.$$  \hfill (5)

Remarks

- The events $\phi(X) = 1$ and $\gamma(X) \in R$ are equivalent.
Foundations

Definition (One-sided and two-sided critical value-based tests)

A critical value-based test is a standard test with a critical value \( c \in \mathbb{R} \)-dependent decision rule.

- A one-sided critical value-based test takes the form
  \[
  \phi(X) : \mathcal{X} \rightarrow \{0, 1\}, x \mapsto \phi(X)(x) := 1\{\gamma(X)(x) \geq c\} = \begin{cases} 
  1 & \gamma(X)(x) \geq c \\
  0 & \gamma(X)(x) < c 
\end{cases} \tag{6}
  \]

- A two-sided critical value-based test takes the form
  \[
  \phi(X) : \mathcal{X} \rightarrow \{0, 1\}, x \mapsto \phi(X)(x) := 1\{|\gamma(X)(x)| \geq c\} = \begin{cases} 
  1 & |\gamma(X)(x)| \geq c \\
  0 & |\gamma(X)(x)| < c 
\end{cases} \tag{7}
  \]

Remarks

- T-tests are familiar examples of critical value-based tests: using the sample mean and sample standard deviation, a realization of the data \( X \) is first transformed into the value of the t-statistic, whose size is then compared to a critical value in order to decide for rejecting the null hypothesis or not.
Definition (Test errors)

When conducting a hypothesis test, two kinds for errors can occur:

- Rejecting the null hypothesis \( \phi(Y) = 1 \), when the null hypothesis is in fact true \( (\theta \in \Theta_0) \), is referred to as a Type I error.

- Not rejecting the null hypothesis \( \phi(Y) = 0 \), when the null hypothesis is in fact false \( (\theta \in \Theta_1) \), is referred to as a Type II error.

Remarks

- Type I errors are often considered more detrimental than Type II errors.
Definition (Test error probabilities)

- The probability of a Type I error is referred to as the **size** of a test and commonly denoted by $\alpha = [0, 1]$, $\alpha := P_{\Theta_0}(\phi(Y) = 1)$. Its complementary probability $P_{\Theta_0}(\phi(Y) = 0) = 1 - \alpha$ is referred to as the **specificity** of a test.
- The probability of a Type II error $P_{\Theta_1}(\phi(Y) = 0)$ lacks a common denomination. Its complementary probability $\beta := P_{\Theta_1}(\phi(Y) = 1)$ is referred to as the **power** of a test.

Remarks

- The $P$ subscripts $\Theta_0$ and $\Theta_1$ indicate that null/alternative hypothesis hold.
- The size of a test is also referred to as the Type I error rate.
- The probability of a Type II error is sometimes denoted by $\beta$, but this is inconsistent with the definition of the power function.
Definition (Test quality and power function)

For a test $\phi(Y)$, the test quality function is defined as

$$q : \Theta \rightarrow [0, 1], \theta \mapsto q(\theta) := \mathbb{E}_{P_{\theta}}(\phi(Y)),$$

(8)

For $\theta \in \Theta_1$, the test quality function is also referred to as the test’s power function, and is denoted by

$$\beta : \Theta_1 \rightarrow [0, 1], \theta \mapsto \beta(\theta) := \mathbb{P}_{\Theta_1}(\phi(Y) = 1).$$

(9)

Remarks

- The test quality function summarizes a test’s size and power as function of $\theta$.
- For $\theta \in \Theta_0$, the test quality function value evaluates to

$$\mathbb{E}_{P_{\Theta_0}}(\phi(Y)) = 0 \cdot \mathbb{P}_{\Theta_0}(\phi(Y) = 0) + 1 \cdot \mathbb{P}_{\Theta_0}(\phi(Y) = 1) = \mathbb{P}_{\Theta_0}(\phi(Y) = 1) = \alpha,$$

(10)

- For $\theta \in \Theta_1$, the test quality function value evaluates to

$$\mathbb{E}_{P_{\Theta_1}}(\phi(Y)) = 0 \cdot \mathbb{P}_{\Theta_1}(\phi(Y) = 0) + 1 \cdot \mathbb{P}_{\Theta_1}(\phi(Y) = 1) = \mathbb{P}_{\Theta_1}(\phi(Y) = 1) = \beta$$

(11)
A common procedure for constructing hypothesis tests

(1) Because the Type I error rate of a test is considered more important than the Type II error rate of a test, the test size is usually fixed first, e.g., by selecting a significance level such as $\alpha' = 0.05$ and an associated critical value $c_{\alpha'}$ of the test statistic.

(2) Given a desired significance level, different tests or statistical models (e.g., sample sizes) are then compared in their ability to minimize the probability of the test’s Type II error, i.e., maximize the test’s power.
Hypothesis testing

• Foundations

• **Significance levels and p-values**

• Examples

  • The T test
  • The Wald test
  • The likelihood ratio test
Significance levels and p-values

Definition (Significance level, conservative, exact, and liberal tests)

A test is said to be of significance level $\alpha' \in [0, 1]$, if its size $\alpha$ is smaller than or equal to $\alpha'$, i.e., if

$$\alpha \leq \alpha'.$$

(12)

- A test is called **conservative**, if $\alpha \leq \alpha'$.
- A test is called **exact**, if $\alpha = \alpha'$.
- A test is called **liberal**, if $\alpha > \alpha'$.

Remarks

- The size and the significance level of a test are two different things.
- A liberal test is not of significance level $\alpha'$.
Hypothesis testing

- Foundations
- Significance levels and p-values

- Examples
  - The T test
  - The Wald test
  - The likelihood ratio test
The construction of a hypothesis test commonly involves

1. The definition of a parametric statistical model.
2. The definition of the test hypotheses, test statistic, and test.
3. The assessment of the test statistic distribution.
4. The establishment of Type I error rate control.
5. The assessment of the test’s power function.

In the following, we demonstrate the above for

1. The T test
2. The Wald test
3. The likelihood ratio test
Hypothesis testing

- Foundations
- Significance levels and p-values
- **Examples**
  - The T test
  - The Wald test
  - The likelihood ratio test
Examples

Example (T test)

1. Parametric statistical model

   Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ denote a random sample from a parametric statistical model with unknown expectation parameter $\mu \in \mathbb{R}$ and unknown variance parameter $\sigma^2 > 0$.

2. Test hypotheses, test statistic, and test

   For the parameter space of the expectation parameter $\Theta := \mathbb{R}$, we consider the test hypotheses
   
   $$\mu \in \Theta_0 := \mu_0 \text{ and } \mu \in \Theta_1 := \mathbb{R} \setminus \mu$$
   
   A standard two-sided test can then be constructed by considering the test statistic
   
   $$T(X = \cdot) : \mathbb{R}^n \to \mathbb{R}, x \mapsto T(X = x) := \frac{\sqrt{n}}{S_n} (\bar{X}_n - \mu_0)$$
   
   and the test
   
   $$\phi(X = \cdot) : \mathbb{R}^n \to \{0, 1\}, x \mapsto \phi(X = x) := 1_{\{|T(X=x)| \geq c\}}.$$
Example (T test)

3. Test statistic distribution

We have seen previously that for $\mu \in \Theta_0$, $T \sim T(n - 1)$, i.e, its distribution for $\mu \in \Theta_0$ is given in terms of the PDF

$$p_{\Theta_0}(t) = \tau(t; n - 1)$$  \quad (16)

4. Type I error rate control

Given the form of the current test and the symmetry of $T(n - 1)$, $\phi(X)$ can be rendered an exact test of significance level $\alpha'$ by choosing a critical value $c_{\alpha'}$ such that

$$\mathbb{P}_{\Theta_0}(\phi(X) = 1) = \mathbb{P}_{\Theta_0}(|T(X)| \geq c_{\alpha'}) = 1 - \int_{-\infty}^{c_{\alpha'}} \tau(x; n - 1) \, dx = \alpha/2. \quad (17)$$
Example (T test)

5. Test quality function
Hypothesis testing

- Foundations
- Significance levels and p-values

**Examples**

- The T test
- **The Wald test**
- The likelihood ratio test
Example (Wald test)

1. Parametric statistical model

Let $X_1, ..., X_n \sim p_\theta$ denote a random sample from a parametric statistical model with unknown parameter $\theta \in \Theta$. Let $\hat{\theta}_n$ denote an asymptotically normally distributed estimator for $\theta$, for example $\hat{\theta}_n^{ML}$.

2. Test hypotheses, test statistic, and test

We consider the test hypotheses

$$\theta \in \Theta_0 := \theta_0 \text{ and } \theta \in \Theta_1 := \Theta \setminus \theta_0 \quad (18)$$

A standard two-sided test can then be constructed by considering the test statistic

$$W(X = \cdot) : \mathbb{R}^n \to \mathbb{R}, x \mapsto W(X = x) := \sqrt{I_n(\hat{\theta}_n^{ML})} (\hat{\theta}_n^{ML} - \theta) \quad (19)$$

and the test

$$\phi(X = \cdot) : \mathbb{R}^n \to \{0, 1\}, x \mapsto \phi(X = x) := 1\{|W(X=x)| \geq c\}. \quad (20)$$
Example (Wald test)

3. Test statistic distribution
   We have seen previously, that $W \sim N(0, 1)$, i.e, its asymptotic distribution for $\theta \in \Theta_0$ is given in terms of the PDF
   \[ p_{\Theta_0}(w) = N(w; 0, 1) \]  
   (21)

4. Type I error rate control
   Given the form of the current test and the symmetry of $N(w; 0, 1)$, $\phi(X)$ can be rendered an asymptotically exact test of significance level $\alpha'$ by choosing a critical value $c_{\alpha'}$ such that
   \[ P_{\Theta_0}(\phi(X) = 1) = P_{\Theta_0}(|W(X)| \geq c_{\alpha'}) = 1 - \int_{-\infty}^{c_{\alpha'}} N(x; 0, 1) \, dx = \alpha/2. \]  
   (22)
Hypothesis testing

• Foundations

• Significance levels and p-values

• Examples
  • The T test
  • The Wald test
  • The likelihood ratio test
Hypothesis testing

• Foundations

• Significance levels and p-values

• Examples
  • The T test
  • The Wald test
  • The likelihood ratio test

• Confidence intervals and hypothesis tests
References