Statistics for Data Science

MSc Data Science WiSe 2019/20

Prof. Dr. Dirk Ostwald

Frequentist inference

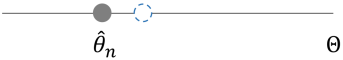
(10) Hypothesis testing

World

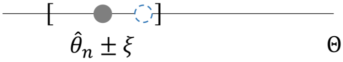
True, but unknown, parameter value

Frequentist inference

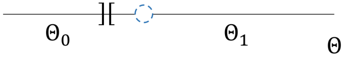
Point estimate $\theta \leftarrow \hat{\theta}_n$



Confidence interval estimate $\mathbb{P}\big(\theta \in [\hat{\theta}_n \pm \xi]\big)$



Hypothesis testing estimate $\theta \in \Theta_0$ vs. $\theta \in \Theta_1$



Bibliographic remarks

The presented material follows Ostwald et al. (2019, Supplementary Material, Section 2) for the majority of test-theoretical concepts. The discussion of the Wald test follows Wasserman (2004, Section 10.1), the development of the duality of confidence intervals and hypotheses tests is based on Czado and Schmidt (2011, Section 5.3). Finally, the introduction to p-values follows Casella and Berger (2002, Section 8.3.4). For an excellent overview on contemporary views of hypothesis testing and the use of p-values, see "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$" (2019) The American Statistician.

Hypothesis testing

- Foundations
- Test construction and examples
  - The T test
  - The Wald test
- Confidence intervals and hypotheses tests
- P-values

Hypothesis testing

- **Foundations**
- Test construction and examples
  - The T test
  - The Wald test
- Confidence intervals and hypotheses tests
- P-values

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_θ, let $\mathcal{X}$ denote the outcome space of the data such that $x \in \mathcal{X}$, and let Θ denote the parameter space of the model. Further, let Θ_0 and Θ_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 Θ_0 and Θ_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.

Definition (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,

$$\phi(X) : \mathcal{X} \to \{0, 1\}, x \mapsto \phi(X)(x), \tag{1}$$

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 a random variable, $\phi(X)$ is also a random variable.

Definition (Standard test)

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

$$\gamma(X) : \mathcal{X} \to \mathbb{R} \tag{2}$$

and a *decision rule*

$$\delta(\gamma(X)) : \mathbb{R} \to \{0, 1\} \tag{3}$$

A standard test can be written as

$$\phi(X) = \delta(\gamma(X)) : \mathcal{X} \to \{0, 1\} \tag{4}$$

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,

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

Remarks

- The events $\phi(X) = 1$ and $\gamma(X) \in R$ are equivalent.
- The events $\phi(X) = 1$ and $\gamma(X) \in R$ have the same probability.

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

$$\phi(X) : \mathcal{X} \to \{0,1\}, x \mapsto \phi(X)(x) := 1_{\{\gamma(X)(x) \geq c\}} = \begin{cases} 1 & \gamma(X)(x) \geq c \\ 0 & \gamma(X)(x) < c \end{cases} \quad (6)$$

- A *two-sided* critical value-based test takes the form

$$\phi(X) : \mathcal{X} \to \{0,1\}, x \mapsto \phi(X)(x) := 1_{\{|\gamma(X)(x)| \geq c\}} = \begin{cases} 1 & |\gamma(X)(x)| \geq c \\ 0 & |\gamma(X)(x)| < c \end{cases} \quad (7)$$

Remark

- *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 ($\phi(X) = 1$), 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 ($\phi(X) = 0$), when the null hypothesis is in fact false ($\theta \in \Theta_1$), is referred to as a *Type II error*.

Remark

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

Remarks

- The $\mathbb{P}$ subscripts Θ_0 and Θ_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 β, but this is inconsistent with the definition of the power function.

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

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

$$\alpha \leq \alpha'. \tag{8}$$

- 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 α'.

Definition (Test quality and power function)

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

$$q : \Theta \to [0,1], \theta \mapsto q(\theta) := \mathbb{E}_{\mathbb{P}_\theta}(\phi(X)). \tag{9}$$

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

$$\beta : \Theta_1 \to [0,1], \theta \mapsto \beta(\theta) := \mathbb{P}_{\Theta_1}(\phi(X) = 1). \tag{10}$$

Remarks

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

$$\mathbb{E}_{\mathbb{P}_{\Theta_0}}(\phi(X)) = 0 \cdot \mathbb{P}_{\Theta_0}(\phi(X) = 0) + 1 \cdot \mathbb{P}_{\Theta_0}(\phi(X) = 1) = \mathbb{P}_{\Theta_0}(\phi(X) = 1) = \alpha \tag{11}$$

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

$$\mathbb{E}_{\mathbb{P}_{\Theta_1}}(\phi(X)) = 0 \cdot \mathbb{P}_{\Theta_1}(\phi(X) = 0) + 1 \cdot \mathbb{P}_{\Theta_1}(\phi(X) = 1) = \mathbb{P}_{\Theta_1}(\phi(X) = 1) = \beta \tag{12}$$

Hypothesis testing

- Foundations
- **Test construction and examples**
  - The T test
  - The Wald test
- Confidence intervals and hypotheses tests
- P-values

Test construction

- Because the Type I error rate of a test is considered more important that 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.
- 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.

Test construction

The construction of a test thus typically 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 steps 1. to 4. of the above for

- The T test
- The Wald test

Hypothesis testing

- Foundations
- **Test construction and examples**
  - **The T test**
  - The Wald test
- Confidence intervals and hypotheses tests
- P-values

Example (T test)

1. Parametric statistical model

   Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ denote a random sample from a parametric statistical model with unknown expectation parameter μ 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

   $$\mu \in \Theta_0 := \{\mu_0\} \text{ and } \mu \in \Theta_1 := \mathbb{R} \setminus \{\mu_0\} \tag{13}$$

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

   $$T(X) := \frac{\sqrt{n}}{S_n}\left(\bar{X}_n - \mu_0\right) \tag{14}$$

   and the test

   $$\phi(X) := 1_{\{|T(X)| \geq c\}}. \tag{15}$$

Example (T test)

3. Test statistic distribution

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

4. Type I error rate control

   Given the form of the current test, the symmetry of $\mathrm{T}(n-1)$, and the invertibility of the CDF $\psi(t)$, $\phi(X)$ can be rendered an exact test of significance level α' by choosing the critical value

$$c_{\alpha'} := \psi^{-1}\left(1 - \frac{\alpha'}{2}; n-1\right) \tag{16}$$

   For example, for $n = 10$ and $\alpha' = 0.05$, $c_{0.05} = \psi^{-1}(0.975; 9) = 2.26$. Rejecting the null hypothesis for an observed absolute T statistic equal to or larger than $c_{0.05} = 2.26$ thus ensures a test size of $\alpha = 0.05$.

Hypothesis testing

- Foundations
- **Test construction and examples**
  - The T test
  - **The Wald test**
- Confidence intervals and hypotheses tests
- P-values

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 θ, for example $\hat{\theta}_n^{ML}$.

2. Test hypotheses, test statistic, and test

   We consider the test hypotheses

$$\theta \in \Theta_0 := \theta_0 \text{ and } \theta \in \Theta_1 := \Theta \setminus \theta_0 \tag{17}$$

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

$$W(X) := \sqrt{J_n\left(\hat{\theta}_n^{ML}\right)}\left(\hat{\theta}_n^{ML} - \theta\right) \tag{18}$$

   and the test

$$\phi(X) := 1_{\{|W(X)| \geq c\}}. \tag{19}$$

Example (Wald test)

3. Test statistic distribution

   We have seen previously, that $W \overset{a}{\sim} N(0, 1)$, i.e, its asymptotic distribution for $\theta \in \Theta_0$ is given in terms of the PDF $N(w; 0, 1)$.

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 α' by choosing the critical value

$$c_{\alpha'} := \Phi^{-1}\left(1 - \frac{\alpha'}{2}\right). \tag{20}$$

Hypothesis testing

- Foundations
- Test construction and examples
  - The T test
  - The Wald test
- **Confidence intervals and hypotheses tests**
- P-values

Theorem (Duality of confidence intervals and hypotheses tests)

Let $\mathcal{P}$ denote parametric statistical model governing the distribution of a random sample $X = (X_1, ..., X_n)$ with outcome space $\mathcal{X}$ and with PMF or PDF p_θ for $\theta \in \Theta \subseteq \mathbb{R}$.

(1) Let $[b_l(X), b_u(X)]$ be a δ-confidence interval for θ. Then the test defined by

$$\phi_\theta(X) : \mathcal{X} \to \{0, 1\}, x \mapsto \phi_\theta(X)(x) := \begin{cases} 0, & \theta \in [b_l(x), b_u(x)] \\ 1, & \theta \notin [b_l(x), b_u(x)] \end{cases} \tag{21}$$

is a test of significance level $\alpha' = 1 - \delta$ for the hypotheses

$$\Theta_0 := \{\theta\} \text{ and } \Theta_1 := \Theta \setminus \{\theta\}. \tag{22}$$

(2) Conversely, let

$$\Phi := \{\phi_\theta(X) | \theta \in \Theta\} \tag{23}$$

be a family of tests, such that $\phi_\theta(X)$ is an exact test of significance level α' for the hypotheses

$$\Theta_0 := \{\theta\} \text{ and } \Theta_1 := \Theta \setminus \{\theta\}. \tag{24}$$

Assume further, that the set

$$C := \{\theta \in \Theta | \phi_\theta(X) = 0\} \tag{25}$$

can be written as $C = [b_l(X), b_u(X)]$ for appropriately determined $b_l(X)$ and $b_u(X)$. Then C is a $\delta := 1 - \alpha'$ confidence interval for θ.

Proof

The significance level $\alpha' = 1 - \delta$ of the test defined in (1) follows from

$$\alpha' \geq \alpha = \mathbb{P}_\theta(\phi(X) = 1) = \mathbb{P}_\theta(\theta \notin [b_l(x), b_u(x)]) = 1 - \mathbb{P}_\theta(\theta \in [b_l(x), b_u(x)]) = \delta. \quad (26)$$

The confidence level $\delta = 1 - \alpha'$ of the confidence level defined in (2) follows from

$$\delta = \mathbb{P}_\theta(\theta \in C) = \mathbb{P}_\theta(\phi_\theta(X) = 0) = 1 - \mathbb{P}_\theta(\phi_\theta(X) = 1) = 1 - \alpha' \quad (27)$$

because $\phi_\theta(X)$ is assumed to be exact.

$\square$

Remarks

- δ-confidence intervals can be used to construct tests of significance level $1 - \delta$.
- Tests of significance level α' can be used to construct a $1 - \alpha'$ confidence intervals.
- In this sense, confidence intervals and hypotheses tests are “equivalent”.

Example (Constructing a hypothesis test from a confidence interval)

We have previously shown that for a random sample $X_1, ..., X_n \sim N(\mu, \sigma^2)$ with unknown expectation parameter μ and unknown variance parameter μ, $\delta \in]0,1[$, and $t_\delta := \psi^{-1}\left(\frac{1+\delta}{2}; n-1\right)$,

$$C_n := \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta\right]. \tag{28}$$

is a δ-confidence interval. With the duality of confidence intervals and hypotheses tests, we may thus define the test

$$\phi_\theta(X) : \mathcal{X} \to \{0,1\}, x \mapsto \phi_\theta(X)(x) := \begin{cases} 0, & \mu \in \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta\right] \\ 1, & \mu \notin \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta\right] \end{cases} \tag{29}$$

for the hypotheses $\Theta_0 = \mu$ and $\Theta_1 = \mathbb{R} \setminus \mu$. Then

$$\begin{aligned} \mathbb{P}_\mu\left(\phi_\mu(X) = 1\right) &= 1 - \mathbb{P}_\mu\left(\phi_\mu(X) = 0\right) \\ &= 1 - \mathbb{P}_\mu\left(\mu \in \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta\right]\right) \\ &= 1 - \delta. \end{aligned} \tag{30}$$

We thus confirmed that ϕ_θ is a test with significance level $\alpha' = 1 - \delta$.

Hypothesis testing

- Foundations
- Test construction and examples
  - The T test
  - The Wald test
- Confidence intervals and hypotheses tests
- **P-values**

Definition (p-value, valid p-value, p-value-based test)

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 $p_\theta, \theta \in \Theta$ and let $\mathcal{X}$ denote the outcome space of the random sample. A *p-value* is a statistic $\mathrm{p}(X)$ with

$$0 \le \mathrm{p}(x) \le 1 \text{ for all } x \in \mathcal{X}. \tag{31}$$

A *valid p-value* is a p-value for which

$$\mathbb{P}_\theta(\mathrm{p}(X) \le \alpha) \le \alpha \text{ for all } \theta \in \Theta \text{ and } 0 \le \alpha \le 1. \tag{32}$$

A test of the form

$$\phi(X) : \mathcal{X} \to \{0, 1\}, x \mapsto \phi(X)(x) := \begin{cases} 0, & \mathrm{p}(x) > \alpha' \\ 1, & \mathrm{p}(x) \le \alpha' \end{cases} \tag{33}$$

is called a *p-value-based test* and is of significance level α'.

Remarks

- Reporting p-values allows readers to choose their own significance levels.
- Smaller p-values imply stronger evidence for rejecting the null hypothesis.
- p-values are information-richer than the mere test outcome $\phi(X)(x)$.

Theorem (Test statistic-based p-values)

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 $p_\theta, \theta \in \Theta$ and let $\mathcal{X}$ denote the outcome space of the random sample. Further assume that

$$\gamma(X) : \mathcal{X} \to \mathbb{R} \tag{34}$$

denotes the test statistic of a critical value-based standard test, such that large value of $\gamma(X)$ provides evidence against the null hypothesis Θ_0. Then the p-value defined by

$$\mathrm{p} : \mathcal{X} \to [0,1], x \mapsto \mathrm{p}(x) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\left(\gamma(X) \geq \gamma(X)(x)\right) \tag{35}$$

is a valid p-value.

Proof

- Casella and Berger (2002, pp. 397 - 398).

Example (P-value for the two-sided t-test)

Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$. We have seen previously, that the test statistic

$$T(X) := \frac{\sqrt{n}}{S_n}\left(\bar{X}_n - \mu\right) \tag{36}$$

of the test

$$\phi(X) := 1_{\{|T(X)(x)| \geq c\}}. \tag{37}$$

is distributed according to $\mathrm{t}(n-1)$ for all values of $\theta = (\mu, \sigma^2)$. Thus the test statistic-based p-value for this test evaluates to

$$\begin{aligned}
\mathrm{p} : \mathcal{X} \to [0,1], x \mapsto \mathrm{p}(x) &:= \sup_{(\mu,\sigma^2)\in\Theta_0} 2\mathbb{P}_{\mu,\sigma^2}\left(T(X) \geq |T(X)(x)|\right) \\
&= 2\mathbb{P}_{\mu,\sigma^2}\left(T(X) \geq |T(X)(x)|\right) \\
&= 2\mathbb{P}_{\mu,\sigma^2}\left(T(X) \geq |\frac{\sqrt{n}}{S_n}\left(\bar{X}_n - \mu\right)|\right).
\end{aligned} \tag{38}$$

Intuitively, the p-value is thus the probability to observe a test statistic "as or more extreme" than the actually observed one.

References

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Czado, C. and Schmidt, T. (2011). *Mathematische Statistik*. Springer-Verlag.

Ostwald, D., Schneider, S., Bruckner, R., and Horvath, L. (2019). Power, positive predictive value, and sample size calculations for random field theory-based fmri inference. *bioRxiv*.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer.