Hypothesis testing I.

Botond Szabo

Leiden University

Leiden, 26 March 2018

Outline

Example

- Suppose we want to know if exposure to asbestos is associated with lung disease.
- We take some rats and randomly divide them into two groups.
- We expose one group to asbestos and then compare the disease rate in two groups.
- We consider the following two hypotheses:
  1. The null hypothesis: the disease rate is the same in both groups.
  2. Alternative hypothesis: The disease rate is not the same in two groups.
- If the exposed group has a much higher disease rate, we will reject the null hypothesis and conclude that the evidence favours the alternative hypothesis.

Null and alternative hypotheses

- We partition the parameter space Θ into sets Θ_0 and Θ_1 and wish to test

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

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

- In our previous example θ is the difference between the disease rates in the exposed and not exposed groups, $\Theta_0 = \{0\}$, and $\Theta_1 = (0, \infty)$.

Rejection region

- Let T be a random variable and let $\mathcal{T}$ be the range of X. We test the hypothesis by finding a subset of outcomes $R \subset \mathcal{X}$, called the rejection region. If $T \in R$, we reject the null hypothesis, otherwise we retain it:

$$T \in R \Rightarrow \text{reject } H_0$$
$$T \notin R \Rightarrow \text{retain } H_0.$$

- In our previous example T is an estimate of the difference between the disease rates in the exposed and not exposed groups. We reject the null hypothesis if T is large.

Test statistic and critical value

- Usually, the rejection region R is of the form

$$R = \{x : T(x) > c\},$$

  where T is a test statistic and c is critical value.
- The problem in hypothesis testing is to find an appropriate test statistic T and an appropriate critical value c.

Warning

- There is a tendency to use hypothesis testing even when they are not appropriate.
- Point estimates and confidence intervals are at times a better option.
- Use testing only with well-defined hypotheses and in general take care of formulating them: things can go wrong when testing ill-defined or overly complex hypotheses.

Type I and type II errors

- By rejecting H_0 when H_0 is true we commit type I error.
- By retaining H_0 when H_1 is true we commit type II error.

	Retain Null	Reject Null
H_0 true	✓	type I error
H_1 true	type II error	✓

Power function

Definition

The power function of a test T with rejection region R is defined by

$$\beta(\theta) = \mathbb{P}_\theta(T \in R).$$

The size of a test is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

A test is said to be of level α, if its size is less or equal to α.

Comments

- Let $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.
- In this case $\beta(\theta_0)$ is the probability of type I error.
- $\beta(\theta_1)$ gives 1 minus the probability of type II error.
- In the ideal case we would like the probabilities of both types of errors be equal to zero. Except trivial cases, this is not possible. So we concentrate on tests of size or level α, where α is taken to be small, e.g. $\alpha = 0.05$.

Hypotheses and tests

- A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis.
- A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ (or both) is called a composite hypothesis.
- A test of the form

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

  is called a two-sided test.
- A test of the form

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

  or

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0$$

  is called a one-sided test.

Example

Example

Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$, with σ known. We want to test $H_0 : \mu \leq 0$ versus $\mu > 0$. Hence $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$. Let $T = \overline{X}_n$ and conisder the test

$$\text{reject } H_0 \text{ if } T > c.$$

The rejection region is

$$R = \{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) > c\}.$$

Example (continued)

Example

The power function is

$$\begin{aligned}\beta(\mu) &= \mathbb{P}_\mu(\overline{X}_n > c) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) \\ &= \mathbb{P}\left(Z > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right).\end{aligned}$$

The power function is increasing in μ.

Example (continued)

Example

Hence

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right).$$

For size α test we set this equal to α and solve for c to get

$$c = \frac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}.$$

We reject the null hypothesis when $\overline{X}_n > c$. Equivalently, when

$$\frac{\sqrt{n}\overline{X}_n}{\sigma} > z_\alpha = \Phi^{-1}(1-\alpha).$$

Example 2

Example

Let X be an observation from a Uniform$(0, \theta)$ distribution. Suppose one wants to test $H_0 : \theta = 2$ versus $H_1 : \theta \neq 2$. Suppose one rejects H_0 if $X \geq 1.9$. Compute the probability of type 1 error.

Most powerful test

- It would be desirable to find the test with highest power under H_1 among all size α tests.
- Such a test, if it exists, is called the most powerful test.
- Finding such tests is not easy, and often they don't even exist.
- We shall instead consider several widely used tests.

Wald test

Definition

Let θ be a scalar parameter and $\hat{\theta}$ an estimate of θ with $\hat{\text{se}}$ *its estimated standard error. Consider the test*

$$H_0 : \theta = \theta_0 \ versus \ H_1 : \theta \neq \theta_0.$$

Assume that $\hat{\theta}$ is asymptotically normal:

$$\frac{\hat{\theta} - \theta_0}{\hat{\text{se}}} \rightsquigarrow N(0,1).$$

The size α Wald test is: reject H_0 when $|W| > z_{\alpha/2}$, where

$$W = \frac{\hat{\theta} - \theta_0}{\hat{\text{se}}}.$$

Size of the Wald test

Theorem

Asymptotically, the Wald test has size α, i.e.

$$\mathbb{P}_{\theta_0}(|W| > z_{\alpha/2}) \to \alpha$$

as $n \to \infty$.

Proof.

We have

$$\begin{aligned}\mathbb{P}_{\theta_0}(|W| > z_{\alpha/2}) &= \mathbb{P}_{\theta_0}\left(\frac{|\hat{\theta} - \theta_0|}{\hat{\text{se}}} > z_{\alpha/2}\right) \\ &\to \mathbb{P}(|Z| > z_{\alpha/2}) \\ &= \alpha.\end{aligned}$$

Remark

- An alternative version of the Wald test statistic is

$$W = \frac{\hat{\theta} - \theta_0}{\text{se}_0},$$

  where se_0 is the standard error of $\hat{\theta}$ computed at θ_0.
- Both versions of the test are valid.

Power of the Wald test

Theorem

Suppose the true value of θ is $\theta^* \neq \theta_0$. The power $\beta(\theta^*)$ (the probability of correctly rejecting the null hypothesis) is approximately given by

$$1 - \Phi\left(\frac{\hat{\theta} - \theta^*}{\hat{\text{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\hat{\theta} - \theta^*}{\hat{\text{se}}} - z_{\alpha/2}\right).$$

Example

Example

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from distributions with means μ_1 and μ_2, respectively. Let us test the null hypothesis that $\mu_1 = \mu_2$. Write this as

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta \neq 0,$$

where $\delta = \mu_1 - \mu_2$. The plug-in estimate of δ is $\hat{\delta} = \overline{X}_m - \overline{Y}_n$ with estimated standard error $\hat{\text{se}} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$, where s_1^2, s_2^2 are sample variances. The size α Wald test rejects H_0 when $|W| > z_{\alpha/2}$, where

$$W = \frac{\hat{\delta} - 0}{\hat{\text{se}}} = \frac{\overline{X}_m - \overline{Y}_n}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}.$$

Example 2

Example

Let us assume that we flip a coin 100 times and 25 out of the 100 we got head. Do we have strong enough evidence to say that the coin is not fair?

Relation to confidence intervals

Theorem

The size α Wald test rejects $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$, where

$$C = (\hat{\theta} - \hat{\text{se}} z_{\alpha/2}, \hat{\theta} + \hat{\text{se}} z_{\alpha/2}).$$

In words, testing the hypothesis with Wald test is equivalent to checking whether the null value θ_0 is in the confidence interval.

Warning

- When we reject H_0, we say the result is statistically significant.
- The result may be statistically significant without the effect size being large.
- In such a case we have a result that is statistically significant, but perhaps not scientifically or practically significant.
- Any confidence interval that excludes θ_0 corresponds to rejecting θ_0. But θ_0 can be close to an endpoint of the confidence interval (not scientifically significant), or far away (scientifically significant).

Heuristics

- Consider a hypothesis test with test statistic T and the rejection region R_α at level α.
- Fix α. Upon observing the data X we evaluate the statistic $T(X)$ and ask whether the test rejects the null hypothesis at level α.
- If yes, we gradually decrease α, while asking the same question.
- Eventually we arrive at α at which the test does not reject the null hypothesis.
- This value of α is called the p-value. It measures the strength of evidence against H_0: when the p-value is small, the data offers little support in favour of H_0.

Evidence scale

- The p-value offers a 'continuous' scale of evidence against the null hypothesis and can be more informative than a simple 'accept/reject' of the hypothesis testing.
- The following evidence scale is typically used:

p-value	evidence
< 0.01	very strong evidence against H_0
$0.01 - 0.05$	strong evidence against H_0
$0.05 - 0.10$	weak evidence against H_0
> 0.1	no or little evidence against H_0

- Warning: a large p-value is not strong evidence in favour of H_0. The large p-value can occur for two reasons:
  - (i) H_0 is true, or
  - (ii) H_0 is false, but the test has low power.

p-value

Definition

Suppose for every $\alpha \in (0,1)$ *we have a size* α *test with rejection region* R_α. *Then*

$$p\text{-}value = \inf\{\alpha : T(X) \in R_\alpha\}.$$

That is, the p-value is the smallest level at which we can reject H_0.

Computation

Theorem

Suppose that the size α test is of the form

$$reject\ H_0\ if\ and\ only\ if\ T(X) \geq c_\alpha.$$

Then, with x the observed value of X,

$$p\text{-}value = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq T(x)).$$

Thus if $\Theta_0 = \{\theta_0\}$,

$$p\text{-}value = \mathbb{P}_{\theta_0}(T(X) \geq T(x)).$$

In other words, p-value is the probability under H_0 of observing a value of the test statistic the same or more extreme than what we actually observed.

Wald statistic

Theorem

Let $w = (\hat{\theta} - \theta_0)/\hat{\text{se}}$ *denote the observed value of the Wald statistic. The p-value is given by*

$$p\text{-}value = \mathbb{P}_{\theta_0}(|W| > |w|) \approx \mathbb{P}_{\theta_0}(|Z| > |w|) = 2\Phi(-|w|).$$

Distribution of the p-value

Theorem

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

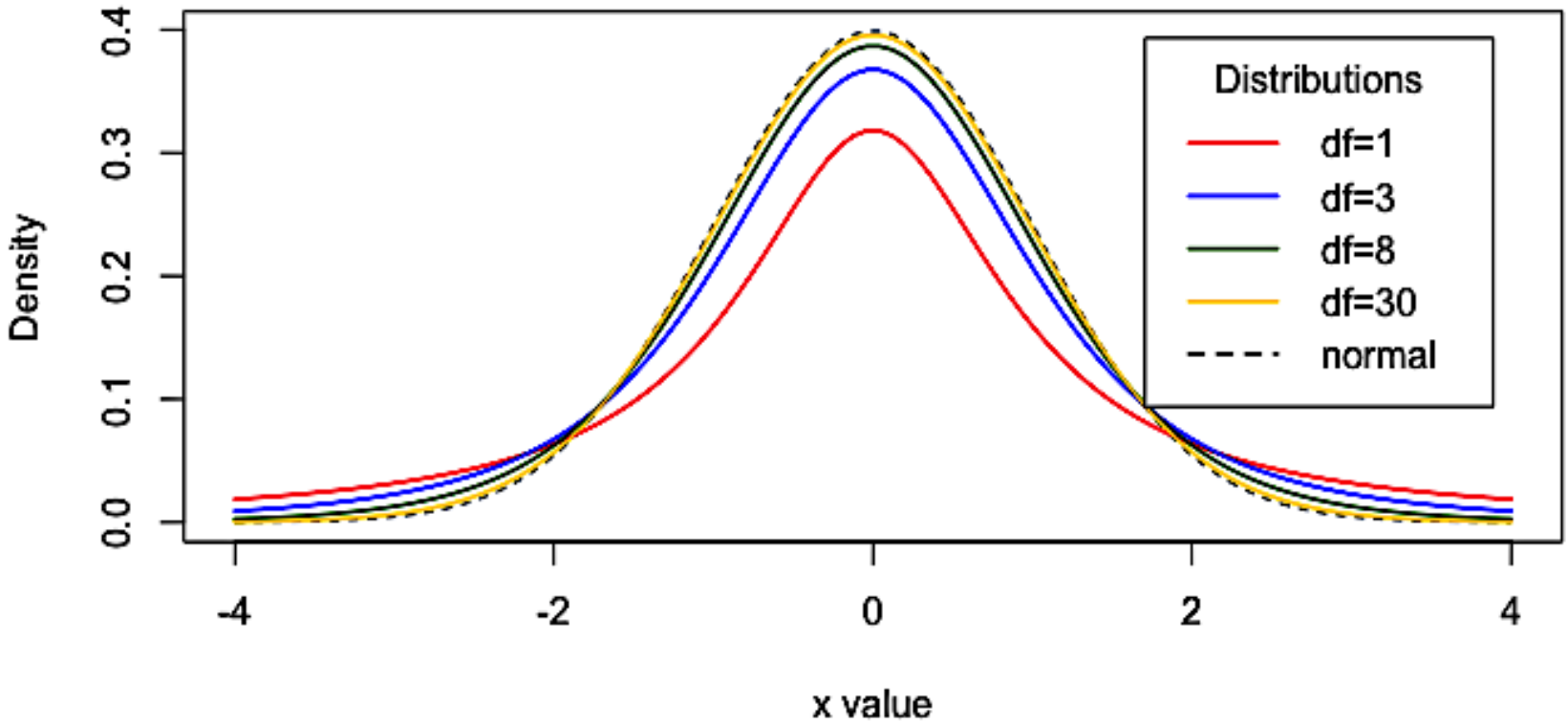
t-test

- To test $H_0 : \mu = \mu_0$ versus $\mu \neq \mu_0$, where $\mu = \mathbb{E}[X_i]$, we can use the Wald test.
- When $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ (with $\theta = (\mu, \sigma^2)$ unknown) and n is small, it is more common to use the t-test.
- A random variable has a t-distribution with k degrees of freedom, if its PDF is given by

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)\left(1+\frac{t^2}{k}\right)^{(k+1)/2}}.$$

- When $k \to \infty$, this tends to the $N(0,1)$ distribution.

t-distribution



t-test (continued)

- Let
$$T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n},$$
where S_n^2 is the sample variance.
- Under H_0, $T \sim t_{n-1}$.
- Let $t_{n-1,\alpha/2}$ denote the upper $\alpha/2$-quantile of the t-distribution.
- If we reject the null hypothesis when $|T| > t_{n-1,\alpha/2}$, we get a size α test.
- However, already for moderately large n the test is essentially identical to the Wald test.

Example

Example

The bload pressure of 6 patients were measured before taking a medicine (124, 134, 117, 150, 160, 150) and after taking the medicine (135, 136, 120, 145, 155, 126), respectively. Do we have sufficient statistical evidence that the medicine was useful in lowering the bload pressure?

Question 1

What is our aim in hypothesis testing?

Answers:

1. We check whether we can accept our theory (null hypothesis).
2. We test whether the data provides sufficient evidence to reject our theory (null hypothesis).
3. We check whether the null hypothesis or the alternative hypothesis is more likely.
4. We check whether the data is normally distributed?

Question 1

What is our aim in hypothesis testing?

Answers:

1. We check whether we can accept our theory (null hypothesis).
2. We test whether the data provides sufficient evidence to reject our theory (null hypothesis).
3. We check whether the null hypothesis or the alternative hypothesis is more likely.
4. We check whether the data is normally distributed?

Question 2

What is not true for the rejection region R?

Answers:

1. It is a random variable.
2. We reject the null hypothesis if the test statistics T is inside of R.
3. Typically is of the form $R = \{x : \ T(x) > c\}$.
4. We make decision on the null hypothesis based on the rejection region and the test statistics, which form a "pair".

Question 2

What is not true for the rejection region R?

Answers:

1. It is a random variable.
2. We reject the null hypothesis if the test statistics T is inside of R.
3. Typically is of the form $R = \{x : \ T(x) > c\}$.
4. We make decision on the null hypothesis based on the rejection region and the test statistics, which form a "pair".

Question 3

Which of the following statements is not true for the error types in hypothesis?

Answers:

1. The type I error is when we reject the null hypothesis altough it is true.
2. Type II error is when we retain the null hypothesis although it was not true.
3. We make a type II error when we do not reject the null hypothesis when the alternate is true.
4. We make a type II error if we accept the null hypothesis and it was true.

Question 3

Which of the following statements is not true for the error types in hypothesis?

Answers:

1. The type I error is when we reject the null hypothesis altough it is true.
2. Type II error is when we retain the null hypothesis although it was not true.
3. We make a type II error when we do not reject the null hypothesis when the alternate is true.
4. We make a type II error if we accept the null hypothesis and it was true.

Question 4

What is not true for the Wald test?

Answers:

1. The Wald test statistic is $W = (\hat{\theta} - \theta_0)/\hat{se}$ and the critical region $|W| > z_{\alpha/2}$.
2. $P_{\theta_0}(|W| > z_{\alpha/2}) \to \alpha$.
3. The data has to be normally distributed to apply it.
4. An alternative version of the Wald test is $W = (\hat{\theta} - \theta_0)/se_0$.

Question 4

What is not true for the Wald test?

Answers:

1. The Wald test statistic is $W = (\hat{\theta} - \theta_0)/\hat{se}$ and the critical region $|W| > z_{\alpha/2}$.
2. $P_{\theta_0}(|W| > z_{\alpha/2}) \to \alpha$.
3. The data has to be normally distributed to apply it.
4. An alternative version of the Wald test is $W = (\hat{\theta} - \theta_0)/se_0$.

Question 5

What is a p value?

Answers:

1. The probability of the null hypothesis.
2. It measures the strength of evidence against H_0.
3. It is equivalent to the first type of error.
4. It provides a strong evidence in favour of H_0.

Question 5

What is a p-value?

Answers:

1. The probability of the null hypothesis.
2. It measures the strength of evidence against H_0.
3. It is equivalent to the first type of error.
4. It provides a strong evidence in favour of H_0.

Question 6

When does one use a t-test instead of a Wald test?

Answers:

1. For modorate sample size.
2. For small sample size
3. If the data is not normally distributed.
4. If the variance of the normal distribution is known.

Question 6

When does one use a t-test instead of a Wald test?

Answers:

1. For modorate sample size.
2. For small sample size.
3. If the data is not normally distributed.
4. If the variance of the normal distribution is known.