Lecture 10: Hypothesis tests

MVE055 / MSG810
Mathematical statistics and discrete mathematics

Moritz Schauer

Last updated September 28, 2021, 2021

GU & Chalmers University of Technology

Hypothesis tests

Hypothesis tests

An important problem in statistics is to test whether a theory or a *research hypothesis* is right or wrong.

Examples of such problems include:

- Does a new drug have any effect?
- Do smokers die sooner than non-smokers?
- Does the measuring device have a systematic error?

Hypothesis tests

An important problem in statistics is to test whether a theory or a *research hypothesis* is right or wrong.

Examples of such problems include:

- Does a new drug have any effect? Mean effect > 0
- Do smokers die sooner than non-smokers?
- Does the measuring device have a systematic error?

Hypothesis tests

An important problem in statistics is to test whether a theory or a *research hypothesis* is right or wrong.

Examples of such problems include:

- Does a new drug have any effect? Mean effect > 0
- Do smokers die sooner than non-smokers?
- Does the measuring device have a systematic error?

Hypothesis tests

An important problem in statistics is to test whether a theory or a *research hypothesis* is right or wrong.

Examples of such problems include:

- Does a new drug have any effect? Mean effect > 0
- Do smokers die sooner than non-smokers? Mean life time difference < 0
- Does the measuring device have a systematic error?

Hypothesis tests

An important problem in statistics is to test whether a theory or a *research hypothesis* is right or wrong.

Examples of such problems include:

- Does a new drug have any effect? Mean effect > 0
- Do smokers die sooner than non-smokers? Mean life time difference < 0
- Does the measuring device have a systematic error?

Hypothesis tests

An important problem in statistics is to test whether a theory or a *research hypothesis* is right or wrong.

Examples of such problems include:

- Does a new drug have any effect? Mean effect > 0
- Do smokers die sooner than non-smokers? Mean life time difference < 0
- Does the measuring device have a systematic error? Mean measurement error $\neq 0$

Hypothesis tests

Answers the statistical analysis could give are

1. that the research hypothesis is supported by the data (and possibly a quantification of the degree of support),
2. that the data doesn't support the hypothesis,
3. a decision rule.

Example

The length of a certain lumber from a national home building store is supposed to be 2.5 m.

Example

The length of a certain lumber from a national home building store is supposed to be 2.5 m.

A builder wants to check whether the lumber cut by the lumber mill has a mean length different smaller than 2.5 m.

Example

The length of a certain lumber from a national home building store is supposed to be 2.5 m.

A builder wants to check whether the lumber cut by the lumber mill has a mean length different smaller than 2.5 m.

A statistical formulation of this problem is that we want to test the null hypothesis

$$H_0 : \text{mean length} = 2.5\,m$$

against the alternative/research hypothesis

$$H_1 : \text{mean length} < 2.5\,m$$

Example

The length of a certain lumber from a national home building store is supposed to be 2.5 m.

A builder wants to check whether the lumber cut by the lumber mill has a mean length different smaller than 2.5 m.

A statistical formulation of this problem is that we want to test the null hypothesis

$$H_0\colon \text{mean length} = 2.5\, m$$

against the alternative/research hypothesis

$$H_1\colon \text{mean length} < 2.5\, m$$

H_1 is actionable knowledge. If H_1 is true she needs to write an angry letter.

Example

- You have new laboratory equipment to measure the chlorine content in water and want to check it. You mix water with true chlorine content 60 (you can do that very precisely), and take 6 measurements.

Example

- You have new laboratory equipment to measure the chlorine content in water and want to check it. You mix water with true chlorine content 60 (you can do that very precisely), and take 6 measurements.
- Results of the measurement are $\bar{x} = 59.62$ and $s^2 = 4.6920$.

Example

- You have new laboratory equipment to measure the chlorine content in water and want to check it. You mix water with true chlorine content 60 (you can do that very precisely), and take 6 measurements.
- Results of the measurement are $\bar{x} = 59.62$ and $s^2 = 4.6920$.
- Assume that the measurements are samples of a random variable $X \sim \mathsf{N}(\mu, \sigma^2)$.

Example

- You have new laboratory equipment to measure the chlorine content in water and want to check it. You mix water with true chlorine content 60 (you can do that very precisely), and take 6 measurements.
- Results of the measurement are $\bar{x} = 59.62$ and $s^2 = 4.6920$.
- Assume that the measurements are samples of a random variable $X \sim \mathsf{N}(\mu, \sigma^2)$.
- The question now is whether we can claim that the new equipment has systematic measurement error, $\mu \neq 60$.

Setup

A statistical formulation of this problem is that we want to test the null hypothesis

$$H_0\colon \mu = 60$$

against the alternative hypothesis or research hypothesis

$$H_1\colon \mu \neq 60.$$

If the test we perform finds that there is a systematic error, H_0 is rejected in favour of H_1.

Setup

A statistical formulation of this problem is that we want to test the null hypothesis

$$H_0\colon \mu = 60$$

against the alternative hypothesis or research hypothesis

$$H_1\colon \mu \neq 60.$$

If the test we perform finds that there is a systematic error, H_0 is rejected in favour of H_1.

Is H_1 actionable knowledge?

Setup

A statistical formulation of this problem is that we want to test the null hypothesis

$$H_0 \colon \mu = 60$$

against the alternative hypothesis or research hypothesis

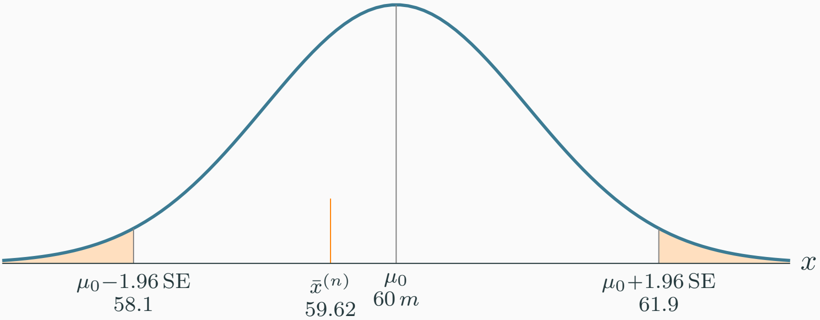
$$H_1 \colon \mu \neq 60.$$

If the test we perform finds that there is a systematic error, H_0 is rejected in favour of H_1.

Is H_1 actionable knowledge?

Choosing the alternative H_1

Choose H_1 such if someone would tell you it is true, you can do something useful with that knowledge!



$$\mathrm{SE} \approx \frac{\sqrt{4.6920}}{\sqrt{5}}$$

Decisions

The outcome of a hypothesis test can be:

- Reject H_0 (accept H_1.)
  - Action!
- Do not reject H_0
  - Could be lack of data, or H_0 being correct. The question of H_0 or H_1 is truly left open. Meh.

Decisions

The outcome of a hypothesis test can be:

- Reject H_0 (accept H_1.)
  - Action!
- Do not reject H_0
  - Could be lack of data, or H_0 being correct. The question of H_0 or H_1 is truly left open. Meh. Should still report it though.

Decision errors

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 Error
	H_1 true	Type 2 Error	✓

- A Type 1 Error is rejecting the null hypothesis when H_0 is true. We want to avoid that, control the probability for this error.
- A Type 2 Error is failing to reject the null hypothesis when H_1 is true.

Burden of proof

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$
\begin{aligned}
H_0 &: \text{ Defendant is innocent} \\
H_1 &: \text{ Defendant is guilty}
\end{aligned}
$$

Burden of proof

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_1: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty
- Declaring the defendant guilty when they are actually innocent

Burden of proof

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_1: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

Burden of proof

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_1: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

- Declaring the defendant guilty when they are actually innocent

Burden of proof

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_1: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Which error do you think is the worse error to make?

Statistical reasoning

Classical logic: If the null hypothesis is correct, then these data can not occur.
These data have occurred.
Therefore, the null hypothesis is false.

Tweak the language, so that it becomes probabilistic…

Statistical reasoning

Statistical reasoning:

If the null hypothesis is correct, then these data are highly unlikely.
These data have occurred.
Therefore, the null hypothesis is unlikely.

Statistical reasoning

Statistical reasoning:

If the null hypothesis is correct, then these data are highly unlikely.
These data have occurred.
Therefore, the null hypothesis is unlikely.

Definition

In statistical hypothesis testing, a result has statistical significance when it is very unlikely to have occurred under the null hypothesis. So significance corresponds to "statistical evidence against the null".

The significance level α is the (tolerated) probability of making a type I error:

$$\mathrm{P}(\text{reject } H_0 \mid H_0 \text{ is true }) \overset{\text{(at most)}}{=} \alpha$$

About failure to reject H_0

If you want to take a decision in the case the test fails to reject H_0, you should compute the type II error probability first. This is typically difficult.

Therefore we should avoid far reaching decisions if our tests fail to reject H_0.

Tests from confidence intervals

Data (samples from a distribution with unknown parameter μ).

Tests from confidence intervals

Data (samples from a distribution with unknown parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

Tests from confidence intervals

Data (samples from a distribution with unknown parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 \colon \mu \neq \mu_0$.

Significance level α, e.g $\alpha = 5\%$.

Decision rule: Compute a $(1 - \alpha)(= 95\%)$-confidence interval $[A, B]$ for the parameter μ. If the $\mu_0 \notin [A, B]$, reject H_0.

Tests from confidence intervals

Data (samples from a distribution with unknown parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

Significance level α, e.g $\alpha = 5\%$.

Decision rule: Compute a $(1-\alpha)(= 95\%)$-confidence interval $[A, B]$ for the parameter μ. If the $\mu_0 \notin [A, B]$, reject H_0.

Type 1 error: This rule has type 1 error of $5\,\%$, so this is a valid test for level $\alpha = 5\,\%$.

Tests from confidence intervals

Data (samples from a distribution with unknown parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

Significance level α, e.g $\alpha = 5\%$.

Decision rule: Compute a $(1 - \alpha)(= 95\%)$-confidence interval $[A, B]$ for the parameter μ. If the $\mu_0 \notin [A, B]$, reject H_0.

Type 1 error: This rule has type 1 error of $5\,\%$, so this is a valid test for level $\alpha = 5\,\%$.

Tests with test statistics

Data (samples with unknown population parameter μ).

Tests with test statistics

Data (samples with unknown population parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 : \mu \overset{\neq}{\underset{<}{>}} \mu_0$.

Tests with test statistics

Data (samples with unknown population parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 \colon \mu \overset{\neq}{\underset{<}{>}} \mu_0$.

Significance level α, e.g $\alpha = 5\%$.

Test statistic T: Typically, T comes from an estimator for our parameter

Tests with test statistics

Data (samples with unknown population parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 : \mu \overset{\neq}{\underset{<}{>}} \mu_0$.

Significance level α, e.g $\alpha = 5\%$.

Test statistic T: Typically, T comes from an estimator for our parameter with known distribution under H_0.

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \text{(example)}$$

Tests with test statistics

Data (samples with unknown population parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 : \mu \overset{\neq}{\underset{<}{>}} \mu_0$.

Significance level α, e.g $\alpha = 5\%$.

Test statistic T: Typically, T comes from an estimator for our parameter with known distribution under H_0.

$$T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \quad \text{(example)}$$

Decision rule: Reject H_0 if the p-value is less than the significance level α.
or: Reject H_0 if the T_{obs} is in the critical region/rejection region (see next slide).

Tests with test statistics

Data (samples with unknown population parameter μ).

Hypothesis about parameter. Here $H_0 : \mu = \mu_0$ and $H_1 : \mu \overset{\neq}{\underset{<}{>}} \mu_0$.

Significance level α, e.g $\alpha = 5\%$.

Test statistic T: Typically, T comes from an estimator for our parameter with known distribution under H_0.

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \text{(example)}$$

Decision rule: Reject H_0 if the p-value is less than the significance level α.
or: Reject H_0 if the T_{obs} is in the critical region/rejection region (see next slide).

Type I error: The type I error for this test is $\leqslant \alpha$.

Critical region

The critical region C_α of a test are those values of the test statistic T for which H_0 can be rejected while obeying significance level α. Typically represented by one or two critical values.

Critical region

The critical region C_α of a test are those values of the test statistic T for which H_0 can be rejected while obeying significance level α. Typically represented by one or two critical values.

We compute rejection region for the data. We reject H_0 if T_{obs} is in the rejection region.

Example: critical region for mean of normal population

We want to use a quantity T that we know the distribution of under H_0, so that we can calculate the critical region.

Example: critical region for mean of normal population

We want to use a quantity T that we know the distribution of under H_0, so that we can calculate the critical region.

In case of the normal distribution with known variance

$$(T =)Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

we know that Z under H_0 is $\mathsf{N}(0,1)$-distributed and

Example: critical region for mean of normal population

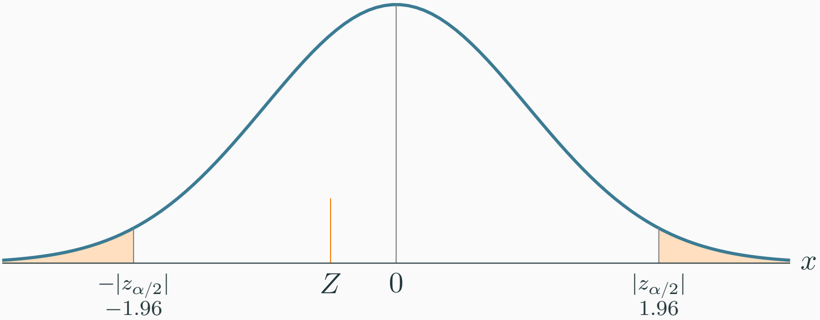
We want to use a quantity T that we know the distribution of under H_0, so that we can calculate the critical region.

In case of the normal distribution with known variance

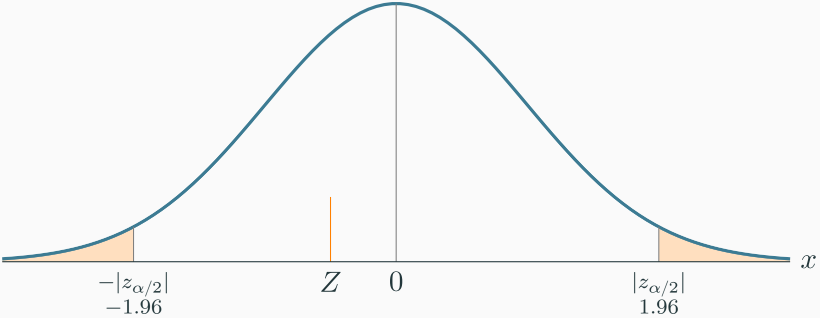
$$(T =)Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

we know that Z under H_0 is $\mathsf{N}(0, 1)$-distributed and

Reject H_0 at level α if $|Z| > z_{\alpha/2}$.



Rejection region for $\alpha = 0.05$.



Rejection region for $\alpha = 0.05$ (on the x-axis below the yellow area).

Rule: Reject H_0 (yeah) if Z is in the rejection region.

Example: p-value for mean of normal population

p-value

The p-value is the probability under the null hypothesis H_0 to obtain a test statistic T with more evidence for the alternative (more "extreme") than the one we observed, t_{obs}.

Example: p-value for normal distribution

Again we want to use a quantity T that we know the distribution of under H_0, so that we can calculate the p-value.

In case of the normal distribution with known variance

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

we know that T under H_0 is $\mathsf{N}(0, 1)$-distributed and

$$p = \mathsf{P}(|T| \geqslant |T_{obs}|) = 2 \cdot \mathsf{P}(T \geqslant |T_{obs}|) = 2(1 - \Phi(|T_{obs}|)).$$

Example: p-value for normal distribution

Again we want to use a quantity T that we know the distribution of under H_0, so that we can calculate the p-value.

In case of the normal distribution with known variance

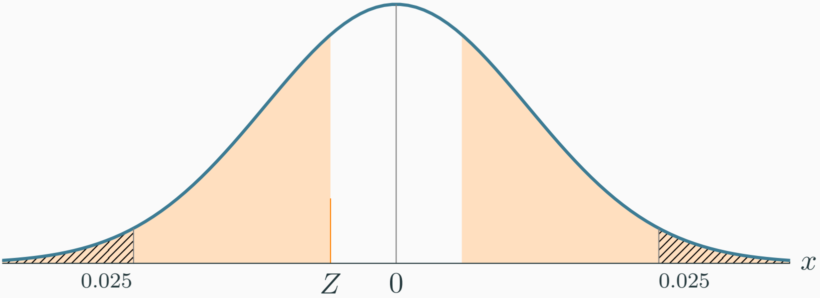
$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

we know that T under H_0 is $\mathsf{N}(0, 1)$-distributed and

$$p = \mathsf{P}(|T| \geqslant |T_{obs}|) = 2 \cdot \mathsf{P}(T \geqslant |T_{obs}|) = 2(1 - \Phi(|T_{obs}|)).$$

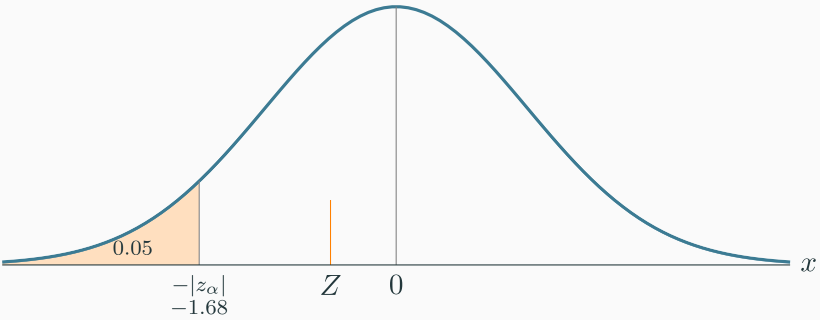
We compute p for the data. We reject H_0 if $p < \alpha$

We compute rejection region for the data. We reject H_0 if T_{obs} is in the rejection region.



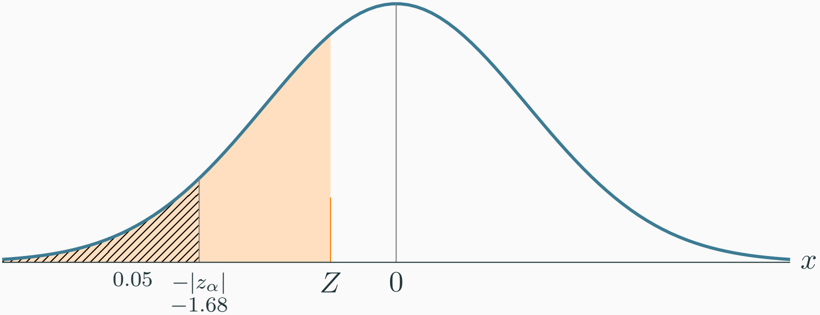
Yellow area: p-value, dashed area: $\alpha = 0.05$.

Rule: Reject H_0 if $p \leqslant \alpha$



One-sided rejection region for $\alpha = 0.05$.

Rule: Reject if Z inside the rejection region.



Yellow area: p value, dashed area: $\alpha = 0.05$.

Rule: Reject H_0 if $p < \alpha$

How many observations are needed?

A test detects a deviation of $\mu - \mu_0$ more easily if:

- If the significance level α is not very small.
- The number of observations n is large.
- The population variance relatively σ^2 is small.