Bus 701: Advanced Statistics

Harald Schmidbauer

Chapter 11:

Hypothesis Testing

11.1 An Introductory Example

The problem.

- Suppose we know that a typical audience rating of a certain TV program in the past was $p = 10\% = 0.1$.
- Today, it was observed that 350 out of 4000 people (i.e., we have a random sample of 4000 people) were watching this program.

Is today a typical day? —

11.1 An Introductory Example

Expectation vs. randomness.

Is today a typical day? —

- IF it is a typical day, we'd expect some 400 people to be watching. . .
- So maybe today is not a typical day?!
- On the other hand: The sample is a $random$ sample.

We need some kind of decision rule to decide whether it is a typical day or not.

11.1 An Introductory Example

A stochastic model.

Is today a typical day? — To ponder this question, we need a stochastic model. The sample of 4000 is described by

$$X_i = \begin{cases} 1 & \text{if person number } i \text{ is watching the program,} \\ 0 & \text{otherwise,} \end{cases}$$

and $i = 1, \ldots, 4000$.

Then, what can we say about the distribution of

$$\sum_{i=1}^{4000} X_i \ldots ?$$

11.1 An Introductory Example

A stochastic model.

IF today is a typical day:

$$\sum_{i=1}^{4000} X_i \quad \sim \quad \mathrm{B}(4000, 0.1)$$

$$\sum_{i=1}^{4000} X_i \quad \sim \quad \mathrm{N}(400, 360) \quad \text{approximately}$$

$$\hat{p} = \frac{1}{4000} \sum_{i=1}^{4000} X_i \quad \sim \quad \mathrm{N}(0.1, 360/4000^2)$$

Our *observed* $\hat{p}$ was 350/4000=8.75%! This is less than the expected 0.1=10% on a usual day.

11.1 An Introductory Example

The prob-value.

The crucial question is now:

> **If** today is a typical day, what is the probability of observing a $\hat{p}$ which is as far, or even further, off the expected 10%, as 8.75%?

This probability is called the prob-value of the hypothesis: "Today is a typical day".

11.1 An Introductory Example

Calculating the prob-value.

The prob-value is $1 - P(0.0875 \leq \hat{p} \leq 0.1125)$.
This can be calculated easily by standardizing $\hat{p}$:

$$\begin{aligned} & 1 - P\left(\frac{0.0875-0.1}{\sqrt{\frac{0.1\cdot 0.9}{4000}}} \leq \frac{\hat{p}-0.1}{\sqrt{\frac{0.1\cdot 0.9}{4000}}} \leq \frac{0.1125-0.1}{\sqrt{\frac{0.1\cdot 0.9}{4000}}}\right) \\ = \; & 1 - P(-2.635 \leq Z \leq +2.635) = 0.0084 \end{aligned}$$

since $Z \sim \mathrm{N}(0,1)$ if today is a typical day (otherwise not!).

The prob-value is very small indeed — less than 1%!

11.1 An Introductory Example

Two explanations for what has happened.

- The question is: Is today a typical day?
- We observed: 350 in 4000 people were watching, that is: $\hat{p}$ =8.75%.
- The probability of observing a $\hat{p}$ as far off as 8.75% is very small.

We conclude from this:

- Either today is a typical day, and something very unlikely has happened.
- Or today is not a typical day!

11.1 An Introductory Example

Statistical hypothesis testing.

The theory of statistical hypothesis testing goes one step further.

- We have just tested

the null hypothesis	$H_0:$	$p = p_0 = 10\%$
against the alternative	$H_1:$	$p \neq p_0 = 10\%$

- Here, p is the true, unknown parameter; p_0 is called the hypothesized value.

- Since the prob-value of H_0 is less than $\alpha = 5\%$, we reject H_0 and decide: Today is not a typical day.

11.1 An Introductory Example

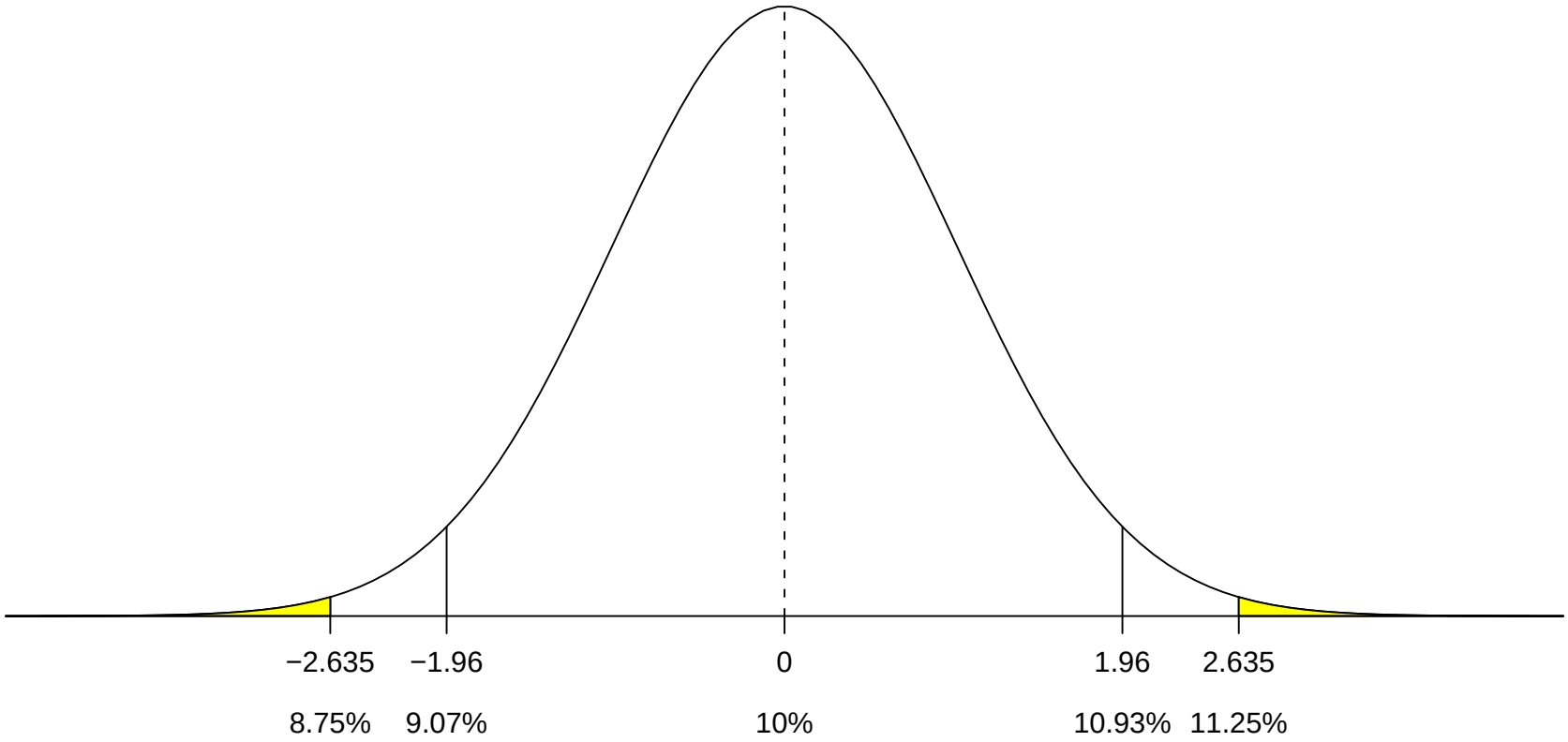
An introductory example.

- This procedure is called a significance test.
- The threshold α is called the significance level.
- Like any method in inductive statistics, it is risky: The decision may be wrong.
- α is actually an error probability: It is the probability that H_0 is rejected even though it is true.

11.1 An Introductory Example

An introductory example.

The relationship between α, the prob-value, and the observed $\hat{p}$ can be illustrated as follows:



11.1 An Introductory Example

An introductory example.

That is: H_0 will be rejected if and only if

$$\hat{p} \text{ is outside } [9.07\%, 10.93\%]$$

or, equivalently:

$$\frac{\hat{p} - 0.1}{\sqrt{\frac{0.1 \cdot 0.9}{4000}}} \text{ is outside } [-1.96, +1.96]$$

or, again equivalently:

$$\text{The prob-value is less than } \alpha = 5\%.$$

11.1 An Introductory Example

An introductory example.

There is another equivalent, very comfortable way to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$:

1. Compute a 95% confidence interval for θ.

2. Reject H_0 if and only if the hypothesized value θ_0 is **not** in this confidence interval.

11.1 An Introductory Example

Example: Audience rating.

Again, let p = true audience rating of the program.

We observed that 350 in the random sample of 4000 were watching the program. Approximate 95% confidence interval (with the hypothesized $p_0 = 0.1$ in the standard error term):

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{p_0(1-p_0)}{n}} = 0.0875 \pm 1.96 \cdot \sqrt{\frac{0.1 \cdot 0.9}{4000}};$$

the 95% confidence interval for p is [7.8%, 9.7%].

This means: $H_0 : p = 0.1$ is rejected against $H_1 : p \neq 0.1$.

We say: p was found to be significantly different from 10%.

11.2 Structure of a Significance Test

Three procedures to test a hypothesis. We assume:

- X is our variable of interest; its distribution depends on an unknown parameter θ.
- We want to test:

  $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$

  Here, θ is the true and unknown parameter;
  θ_0 is the hypothesized value.
- We have chosen a significance level α (typically, $\alpha = 0.05$).

In the following, we shall review the three procedures to test H_0.

11.2 Structure of a Significance Test

Procedure I.

1. Specify the test statistic $T = T(X_1, \ldots, X_n)$.
   - Its distribution must be (approximately) known in the case that H_0 is true.
   - T can be a (standardized) point estimator for θ.
2. Observe the data, i.e. the realizations of $X_1, \ldots, X_n$.
3. Compute the prob-value of H_0.
4. Make the decision: Reject H_0 if the prob-value is less than α; otherwise don't reject H_0.

11.2 Structure of a Significance Test

Procedure II.

1. Specify the test statistic $T = T(X_1, \ldots, X_n)$.
   - Its distribution must be (approximately) known in the case that H_0 is true.
   - T can be a (standardized) point estimator for θ.
2. Determine a critical region C such that $P_{\theta_0}(T \in C) = \alpha$.
   - "Critical" means: critical for H_0.
   - C can consist of "too small" and "too large" values for T.
3. Observe the data, i.e. the realizations of $X_1, \ldots, X_n$.
4. Make the decision: Reject H_0 if $T \in C$; otherwise don't reject H_0.

11.2 Structure of a Significance Test

Procedure III.

1. Specify a point estimator $\hat{\theta}$ for θ.
   - Its distribution must be (approximately) known in the case that H_0 is true.

2. Observe the data, i.e. the realizations of $X_1, \ldots, X_n$.

3. Compute an (approximate) $(1-\alpha)\cdot 100\%$ confidence interval $[C_1, C_2]$ for θ, assuming H_0 is true.

4. Make the decision: Reject H_0 if $\theta_0 \notin [C_1, C_2]$; otherwise don't reject H_0.

11.2 Structure of a Significance Test

Procedure III — Example 1.

The $Alpha$ company produces steel tubes.

- The steel tube process: cut-to-length operation; generates tubes that have a normally distributed length (measured in inches) with mean μ and standard deviation σ.
- From previous operations, it is known that $\sigma = 0.1$, while μ is unknown, due to a new adjustment of the process.
- The required average length is 12 inches.
- A sample of 15 tubes had lengths 11.73, 12.02, 11.99, 11.86, 12.11, 12.11, 12.02, 12.01, 11.89, 11.96, 12.12, 11.91, 11.98, 12.03, 11.95.
- Is this in line with the required average length?
- $H_0 : \mu = \mu_0 = 12$ is not rejected against $H_1 : \mu \neq \mu_0 = 12$, because $\mu = 12$ is contained in the 95% confidence interval: [11.93, 12.03]

11.2 Structure of a Significance Test

Procedure III — Example 2.

Analyzing returns on stocks.

Approximate 95% confidence intervals for the kurtosis were

Bovespa:	$[-0.47, 3.82]$
Dow-Jones:	$[1.81, 5.99]$
DAX:	$[1.79, 3.87]$

It turns out that Bovespa is different with respect to its kurtosis! —

For Dow-Jones as well as for DAX, the kurtosis was found to be significantly different from 0. Not so for Bovespa!

11.3 Errors in Significance Tests

Type I and type II errors.

An inductive conclusion is necessarily risky. Two kinds of error can happen. This can be described in a table:

		true situation	
		H_0 true	H_0 false
our decision:	reject H_0	type I error	no error
	don't reject H_0	no error	type II error

11.3 Errors in Significance Tests

Type I and type II errors.

- Significance tests are constructed such that the probability of a type I error is under control and small — it is α.
- However, the probability of a type II error is not under control.
  It can be as large as $1 - \alpha$, that is: 95%!
- This indicates a fundamental asymmetry of a significance test.
- This means: We can be confident we have found something only if H_0 is rejected.
  Not rejecting H_0 does not provide us with any new information.

11.3 Errors in Significance Tests

Type I and type II errors.

Why can the type II error probability become so large?

Example: Audience rating. Consider this situation:

$$H_0 : p = p_0 = 10\% \quad \text{against} \quad H_1 : p \neq p_0 = 10\%$$

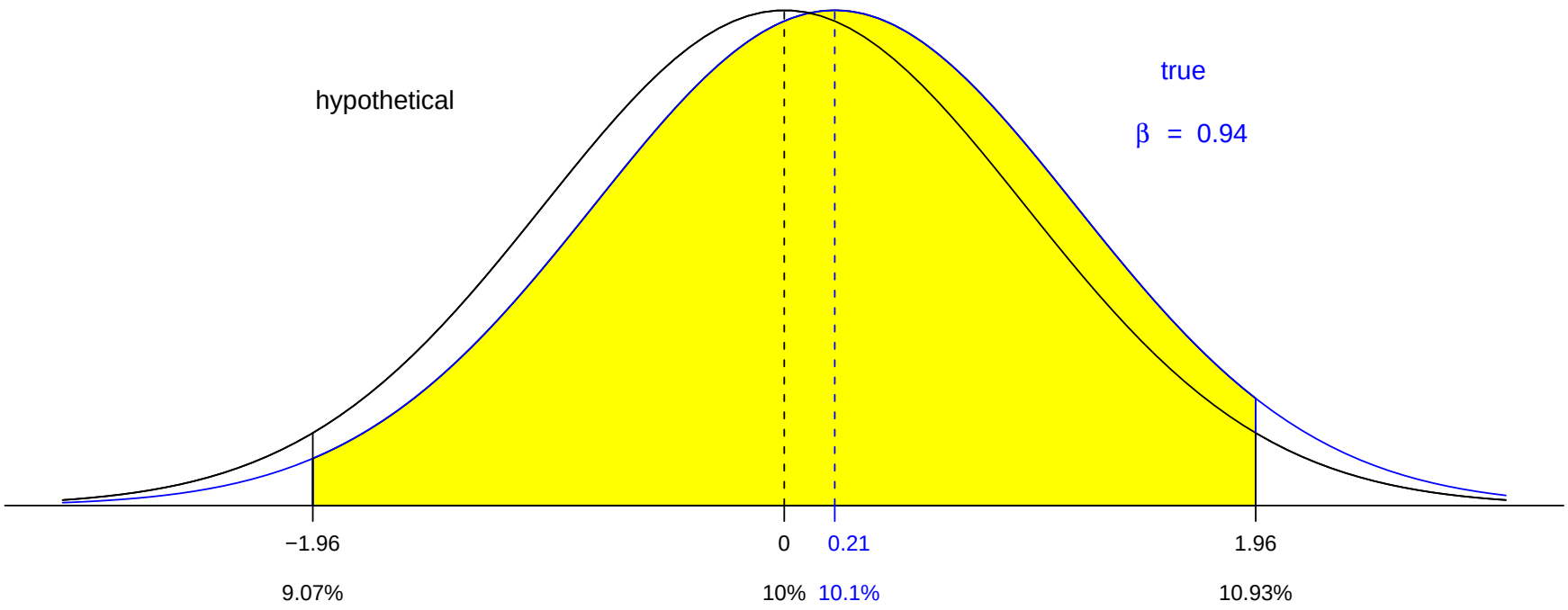
Now suppose the true p is not $p = 10\%$, but $p = 10.1\%$.

- Then H_0 is false, but there is practically no chance to detect this small difference.
- That is, the probability to reject will be nearly the same as if p was exactly 10%.
- In other words, the probability of a type II error is about 95%.

11.3 Errors in Significance Tests

Type I and type II errors.

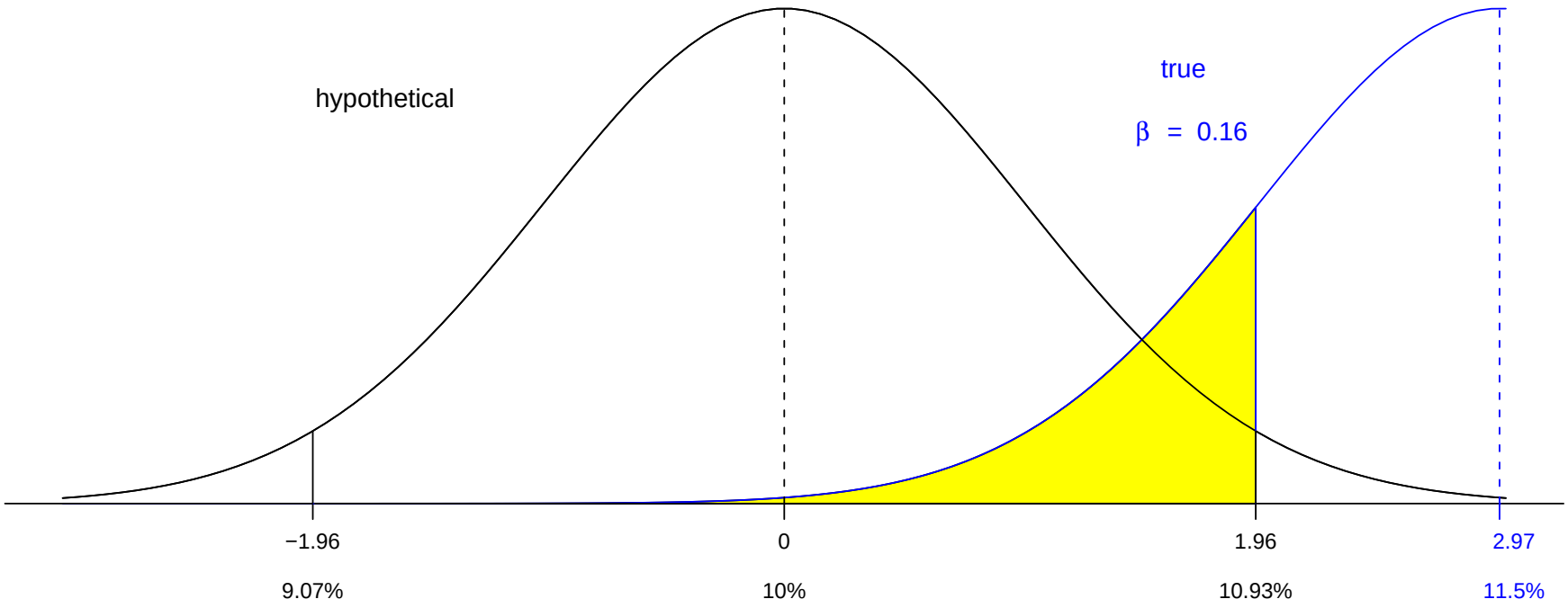
The following picture shows that the type II error probability can be very large.



11.3 Errors in Significance Tests

Type I and type II errors.

If the true parameter is further away from the hypothesized value, the type II error probability becomes smaller.



11.3 Errors in Significance Tests

Asymmetry of significance tests.

The asymmetry of significance tests has consequences for the correct formulation of H_0 and H_1.

Example: Audience rating.

Which null hypothesis H_0 should be tested against which alternative H_1? —

This depends on the research interest!

We shall see three perspectives.

11.3 Errors in Significance Tests

Example: Audience rating.

Perspective of. . .

- . . . research institute: They have no particular interest in showing that p is large or small — all they want to know is: It today's p different from the p in the past, or not?
  They will test:

$$H_0 : p = p_0 = 10\% \quad \text{against} \quad H_1 : p \neq p_0 = 10\%$$

11.3 Errors in Significance Tests

Example: Audience rating.

Perspective of. . .

- . . . TV channel's program director: She may want to show that today's audience rating is higher than in the past: "We gained market share!"
  She has to test:

$$H_0 : p \leq p_0 = 10\% \quad \text{against} \quad H_1 : p > p_0 = 10\%$$

  If H_0 is rejected, she has indeed evidence that her statement is true.

11.3 Errors in Significance Tests

Example: Audience rating.

Perspective of. . .

- . . . company having their TV commercial broadcast during that program: They will want to show that today's audience rating is less than in the past: "Broadcasting fees have to go down!"
  They have to test:

$$H_0 : p \geq p_0 = 10\% \quad \text{against} \quad H_1 : p < p_0 = 10\%$$

  If H_0 is rejected, they have indeed evidence that the audience rating has decreased.

11.3 Errors in Significance Tests

Example: Audience rating.

We conclude with a numerical example of the company perspective. To be tested:

$$H_0 : p \geq p_0 = 10\% \quad \text{against} \quad H_1 : p < p_0 = 10\%;$$

critical: small values of $\hat{p}$. ("Critical" always means: critical for H_0.)

If we observed a sample of 4000, with $\hat{p} = 8.75\%$, the prob-value is the probability that we observe a $\hat{p}$ as small as, or even smaller than, 8.75%, if the true p is $p = 10\%$.

11.3 Errors in Significance Tests

Example: Audience rating.

This probability is:

$$\text{prob-value } = P(\hat{p} \leq 0.0875) = \ldots = 0.0042$$

Since prob-value $< 5\%$, we reject H_0, and decide:

> The audience rating that day was significantly smaller than in the past. — There is evidence that the audience rating has gone down.

(Observe that this is useless for the TV channel's program director.)

11.4 The Power of a Test

Possible errors and power of a test.

- For any significance test, the type I error probability is always (at most) α.
- Without further consideration, the type II error probability is not under control.
- A "good", "powerful" test should have a small type II error probability, that is:
  A false null hypothesis should be rejected with high probability.

11.4 The Power of a Test

The power function.

Consider a test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. The function

$$\theta \mapsto P_\theta(\ H_0 \text{ is rejected }\ )$$

is called the power function of the test. — It holds that:

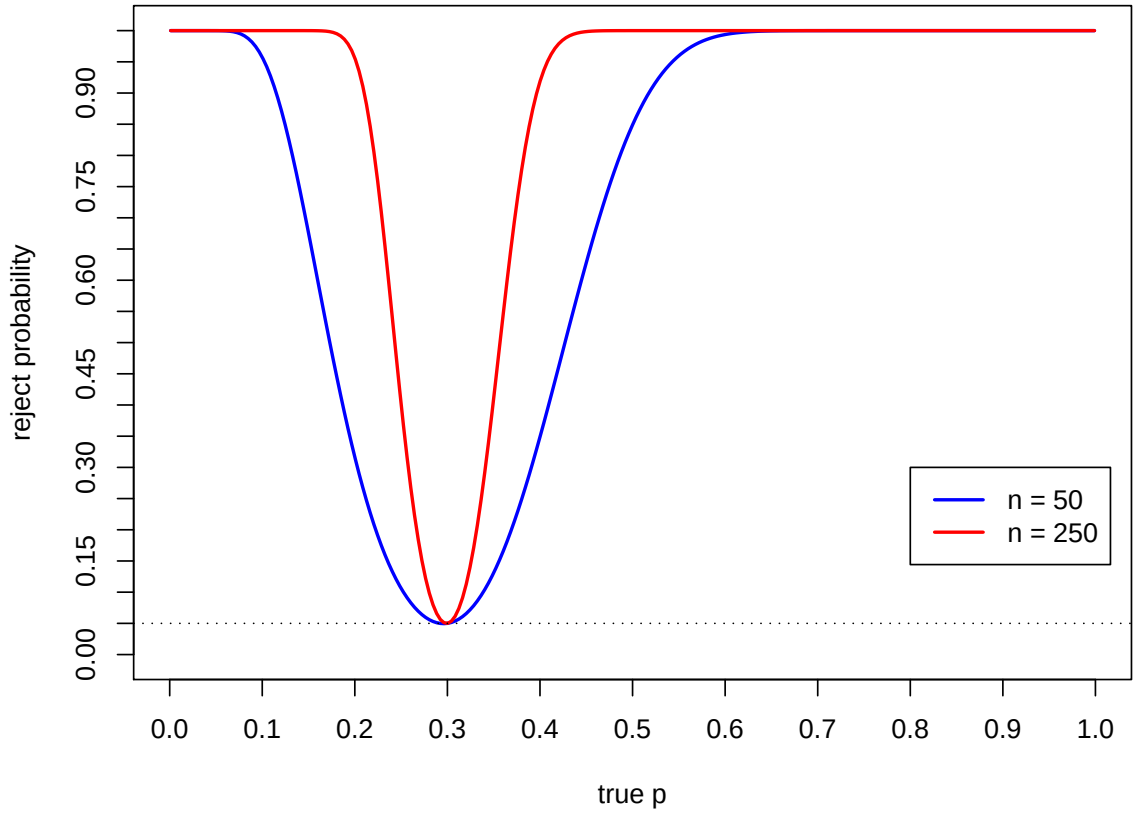
- $P_{\theta_0}(\ H_0 \text{ is rejected }\ ) = \alpha$.
- For $\theta \neq \theta_0$, $1 - P_\theta(\ H_0 \text{ is rejected }\ )$ is the probability of a type II error.

(Which changes have to be made in the case of a one-sided test?)

11.4 The Power of a Test

Plot of a power function.

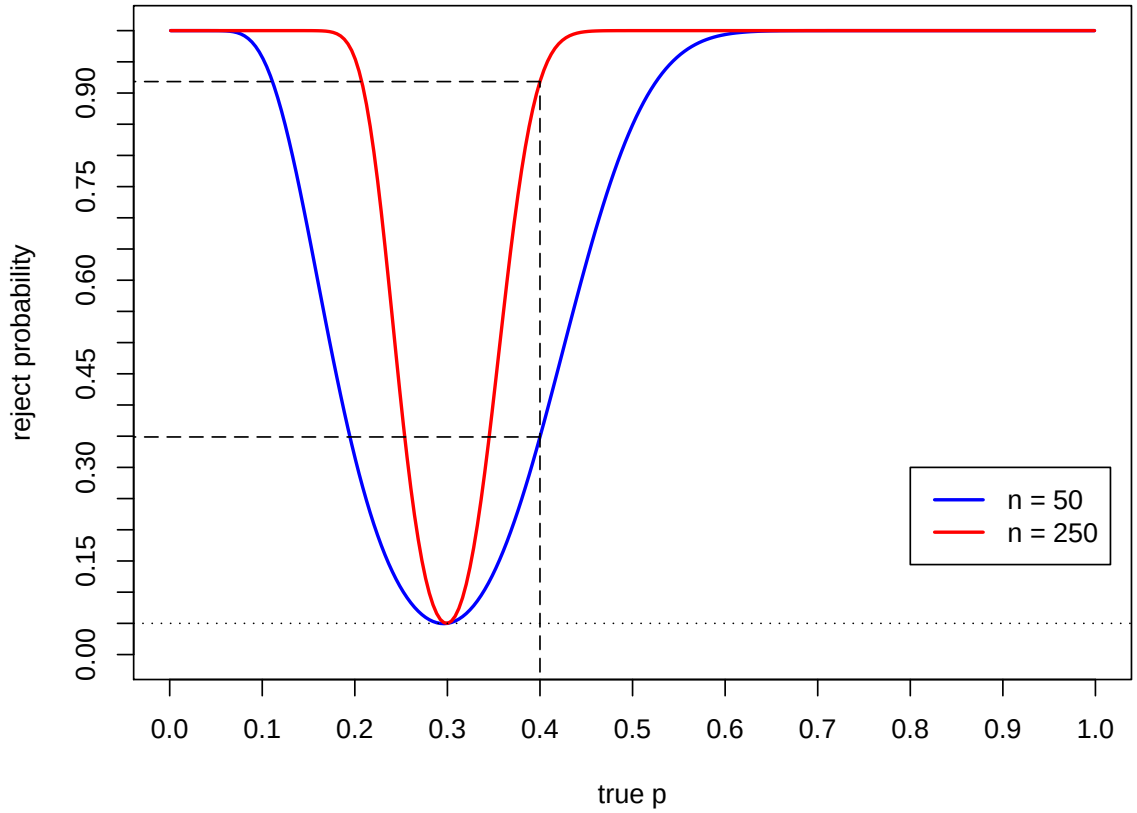
Testing $H_0 : p = 0.3$ against $H_1 : p \neq 0.3$. Here is a plot of the power function of this test for two different sample sizes:



11.4 The Power of a Test

Plot of a power function.

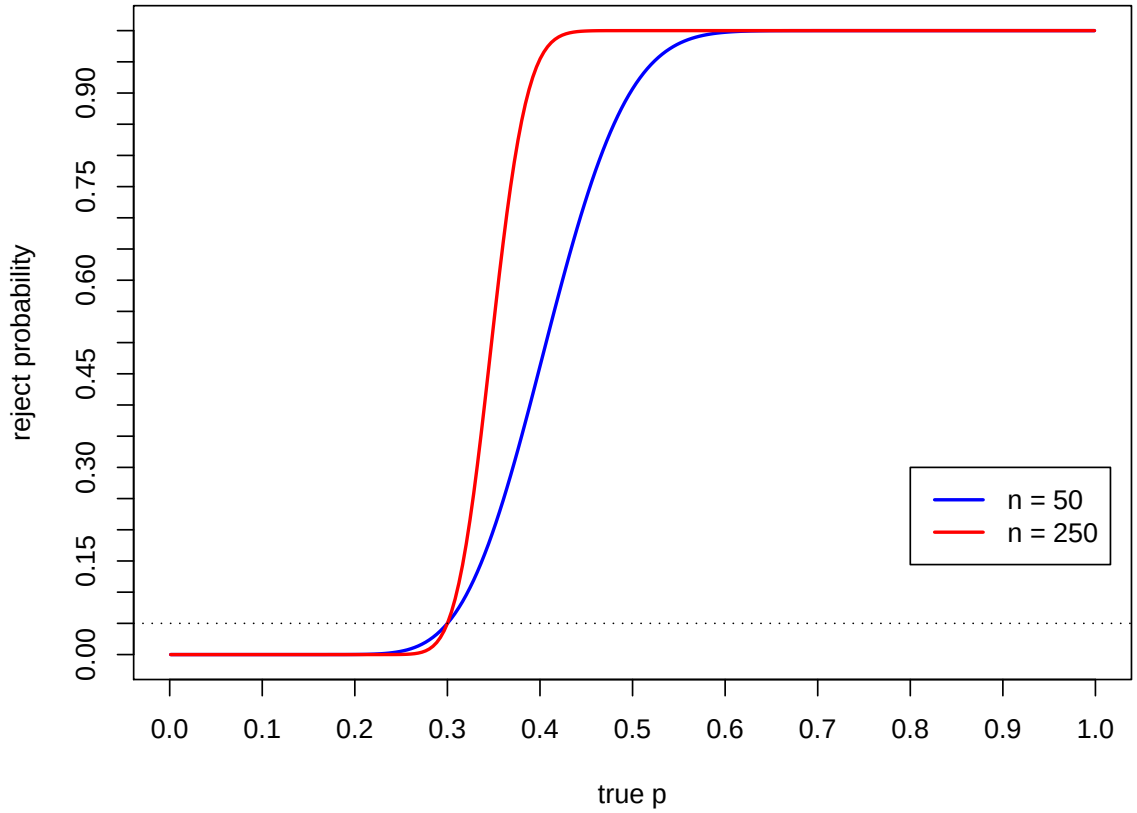
This plot shows the "power" of the test to detect the difference between hypothesized $p_0 = 0.3$ and true $p = 0.4$.



11.4 The Power of a Test

Plot of a power function — a one-sided test.

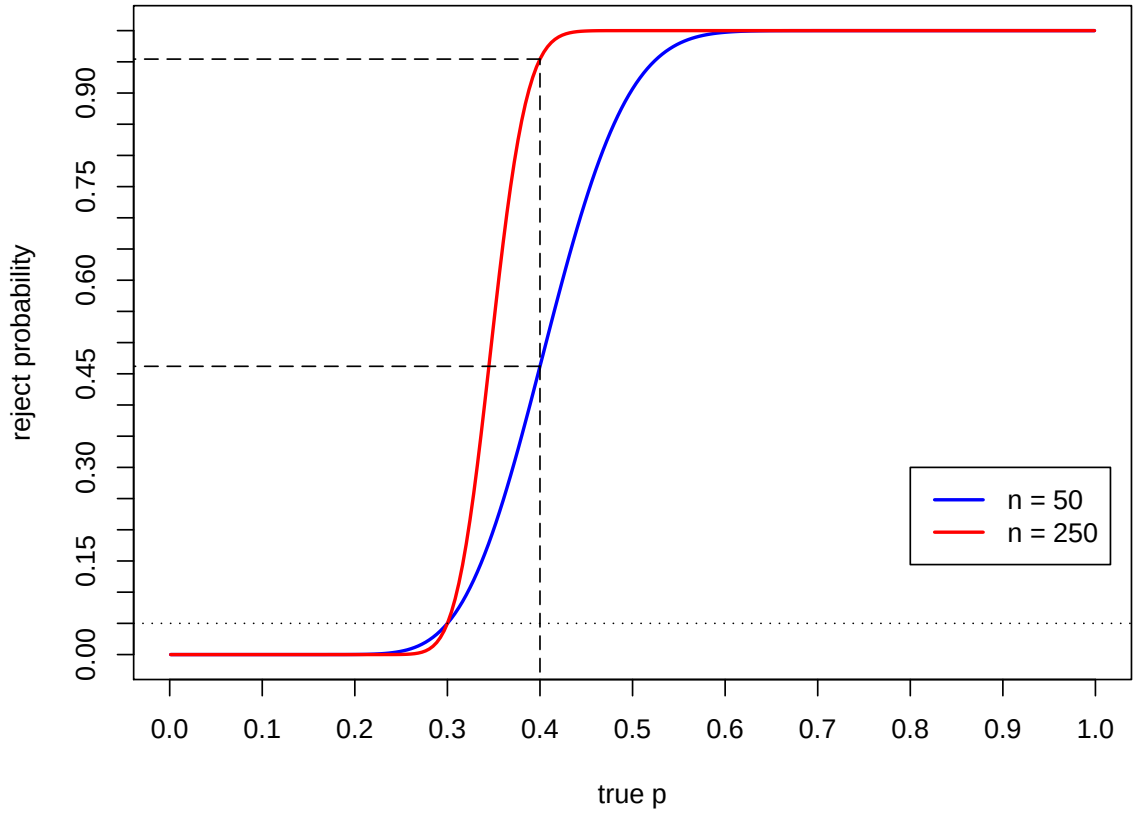
Testing $H_0 : p \leq 0.3$ against $H_1 : p > 0.3$. Here is a plot of the power function of this test for two different sample sizes:



11.4 The Power of a Test

Plot of a power function — a one-sided test.

This plot shows the "power" of the test to detect the difference between hypothesized $p_0 \leq 0.3$ and true $p = 0.4$.



11.4 The Power of a Test

An example from quality control.

- A lot of thousands of items is delivered.
- An unknown share p of the items is defective.
- We are willing to accept the lot if $p \leq 4\%$.
- We draw a random sample from the lot to decide if we accept or reject the lot.
- How large should the sample size be if we want a reject probability of at least 90% if $p = 8\%$?

11.5 Hypotheses

Where does a hypothesis come from? — How is it tested?

- How should null hypothesis and alternative hypothesis be formulated?
- We have seen: This depends on the research interest.
- IMPORTANT:
  It is not admissible to use the same dataset to derive and test a null hypothesis.

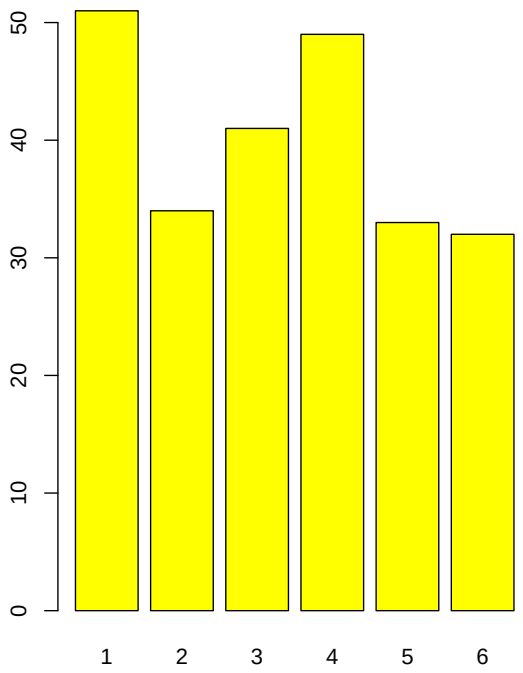
11.5 Hypotheses

A random experiment.

- A die is rolled 240 times.
- We determine the two outcomes with the highest frequencies.
- Let $p =$ combined probability of these two outcomes.
  (If the die is unbiased, $p = 1/3$.)
- Then we test, using the $same\ data$, $H_0 : p \leq 1/3$ against $H_1 : p > 1/3$.
- What is wrong with this procedure?

11.5 Hypotheses

A typical outcome of this experiment.



- Frequencies:

outcome	1	2	3	4	5	6
frequency	51	34	41	49	33	32

- Let $p = P($ die falls 1 or 4 $)$
- $H_0 : p \leq 1/3, \quad H_1 : p > 1/3$
- p-value: 0.0043
- H_0 is rejected!

11.5 Hypotheses

What is wrong with this procedure?

- The problem with the procedure in this example is that the same data are used
  - to formulate the null hypothesis
  - and to test the null hypothesis.
- The consequence is that the type I error probability is not under control anymore.
- Here, the probability of rejecting the null hypothesis $H_0 : p \leq 1/3$ against $H_1 : p > 1/3$ is more than 40%!

11.5 Hypotheses

So, in order to work correctly, where does a hypothesis come from?

A hypothesis can. . .

- . . . reflect a target value.
- . . . reflect assumed ignorance or neutrality.
- . . . be intended to confirm a theory by empirical evidence.

Important:

- A hypothesis cannot be tested reliably using the data which gave rise to that hypothesis.

11.5 Hypotheses

A famous example: The lady tasting tea.

A lady claims she can tell what was poured into the cup first: tea or milk. Is she exaggerating?

- Let $p = P($ the lady judges correctly when tasting a single cup $)$
- We test: $H_0 : p = 1/2$ (or $H_0 : p \leq 1/2$) against $H_1 : p > 1/2$.
- How many cups in a row would the lady have to judge correctly so that we can say: $Her\ success\ rate\ is\ significantly\ larger\ than\ 50\%$?
- What are the type I, type II error probabilities?