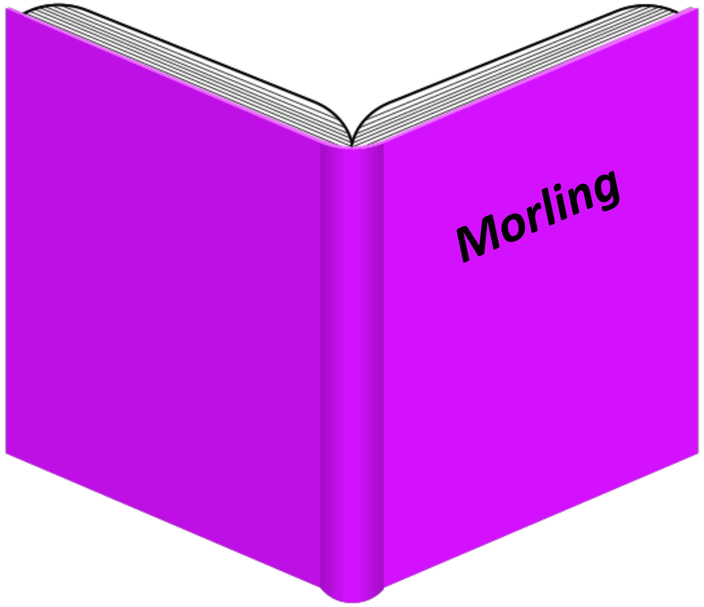


Research Methods and Statistics

Lecture 15: Hypothesis tests

Riet van Bork

Pictures source: pixabay.org

Exam is graded

- You did well! Mean: 22.8 points (exactly the same as last year, and almost same as previous exam which had mean 23)
- Wednesday we will post announcement on Canvas with some explanation on a couple of questions that were more difficult
- Now some explanation on the open question

Open question

Consider the following research example:

Benji believes there is an association between liking vegetables and doing exercises. He claims: “People who are more than others fan of eating vegetables, also exercise more!” To find evidence for this claim, he decides to ask all first-year psychology students how much they like eating vegetables and also asks them how often they exercise and computes a correlation between the scores.

Part (a): Explain whether internal validity is relevant for this example.

Open question

"People who are more than others fan of eating vegetables, also exercise more!" is an association claim!

It does not say that eating vegetables causes exercise or the other way around, it only says that these two behaviors tend to go together (an association).

It was an example of "people who.., also do.." (p.67 in book)

For more explanation, read the section "three claims" in Ch3

Internal validity is only relevant for *causal* claims. After all, only when you claim that the reason for the association is that one causes the other *then* you want to know whether there are confounders or alternative explanations.

See also lecture 7 and p. 73 of the book.

So:

External validity is not relevant for the example (1pt)

The reason is that internal validity is only relevant for causal claims, but here no causal claim is asserted (just an association). (1pt)

Open question

Part b of the question: Name the other three “big validities” and explain each of them using the example.

Construct validity: how well are the conceptual variables operationalized/measured. (0.5pt)

In this example, one could wonder whether just asking people how much they like vegetables or how often they do exercises are good measures of their actual ‘liking of vegetables’/’frequency of doing exercises’. (0.5pt)

Statistical validity: How much statistical evidence is there for the claim? Or how well is the statistical effect backed up by data? (0.5pt)

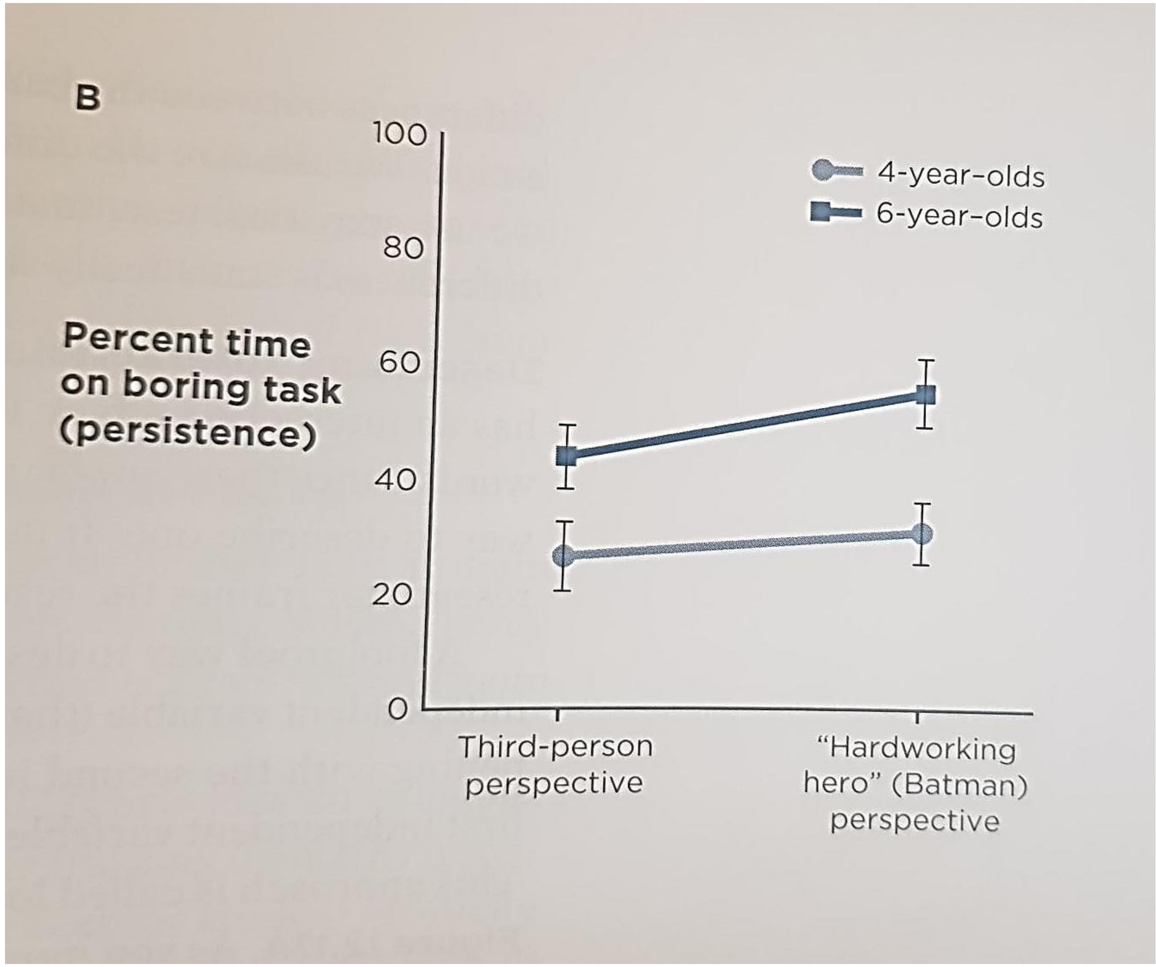
In this example, you could wonder how strong is the correlation? Or is the correlation significant? Or how precise is the estimate of the correlation? (0.5pt)

External validity: To what populations, settings and times can we generalize the findings? (0.5pt)

In this example, if there is a correlation in 1st year psychology students, does that generalize to other samples or populations? Or does this association also exist in populations with other ages? Or in other times? (0.5pt)

See Morling Ch3 “Interrogating Association Claims” pp.71-73 and table 3.6

Last week



If we would show you this graph and add "Ignore sampling variability, only consider point estimates" then the conclusion should be "there is an interaction effect, because the lines are not exactly parallel!"

However, the book concludes "There probably is no interaction (these readers reported that the interaction is not significant)" and the caption writes "the lines are nearly identical"

We use significance tests to deal with the uncertainty around the point estimates.

p.369 in Morling

Today

Hypothesis test for a proportion

Hypothesis test for a mean

Correspondence to confidence interval

Sometimes we make the wrong decision





Picture source: pixabay.com

JOURNAL ARTICLE

Maternal Diet and Other Factors Affecting Offspring Sex Ratio: A Review

Get access >

Cheryl S. Rosenfeld ✉, R. Michael Roberts

Biology of Reproduction, Volume 71, Issue 4, 1 October 2004, Pages 1063–1070, https://doi.org/10.1095/biolreprod.104.030890

Published: 01 October 2004 **Article history** ▾

Cite Permissions Share ▾

Abstract

Mammals usually produce approximately equal numbers of sons and daughters, but there are exceptions to this general rule, as has been observed in ruminant ungulate species, where the sex-allocation hypothesis of Trivers and Willard has provided a rational evolutionary underpinning to adaptive changes in sex ratio. Here, we review circumstances whereby ruminants and other mammalian species, especially rodents and primates, appear able to skew the sex ratio of their offspring. We also discuss some of the factors, both nutritional and nonnutritional, that potentially promote such skewing. Work from our laboratory, performed on mice, suggests that age of the mother and

"a diet high in saturated fats but low in carbohydrate leads to the birth of significantly more male than female offspring in mature laboratory mice, whereas when calories are supplied mainly in the form of carbohydrate rather than fat, daughters predominate"

Proportion girls

- How do we test whether a sample proportion differs from some hypothesized value

- In the population the probability of a baby girl is 0.5

- What value should we find to conclude that the high carbohydrate/low fat diet is effective in increasing the proportion girls?
  - 0.51? 0.55? 0.60?

- Posibilities:
  - Confidence interval → Chapter 8
  - Today: Hypothesis tests

Hypothesis tests

Substantive hypothesis: Statement about how the world works

→Caffeine improves concentration

→horoscope readers can predict personality based on date of birth

→Diet high on carbohydrate and low on fat will increase the probability of a baby girl

Statistical hypothesis: Statement about a population parameter

→people score higher on concentration test after having had caffeine (difference > 0)

→horoscope readers perform better than can be expected from chance (proportion correct > ¼ , when 4 personalities to choose from)

→for people who use high-carbodydrate/low-fat, proportion of girls born > 0.5

Significance test

Involves two statistical hypothesis

- Null hypothesis, H_0: Specifies a particular value for a population parameter
  - H_0: $p = p_0$
  - In this case, H_0: $p = 0.5$

- Alternative hypothesis, H_a: Specifies an alternative range of values
  - H_a: $p > p_0$ or $p < p_0$ or $p \neq p_0$
  - In this case, H_a: $p > 0.5$

Hypothesis tests

Substantive hypothesis: Statement about how the world works

→Caffeine improves concentration

→horoscope readers can predict personality based on date of birth

→Diet high on carbohydrate and low on fat will increase the probability of a baby girl

Statistical hypothesis: Statement about a population parameter

→people score higher on concentration test after having had caffeine (difference > 0)

→horoscope readers perform better than can be expected from chance (proportion correct > ¼ , when 4 personalities to choose from)

→for people who use high-carbodydrate/low-fat, proportion of girls born > 0.5

Hypothesis tests

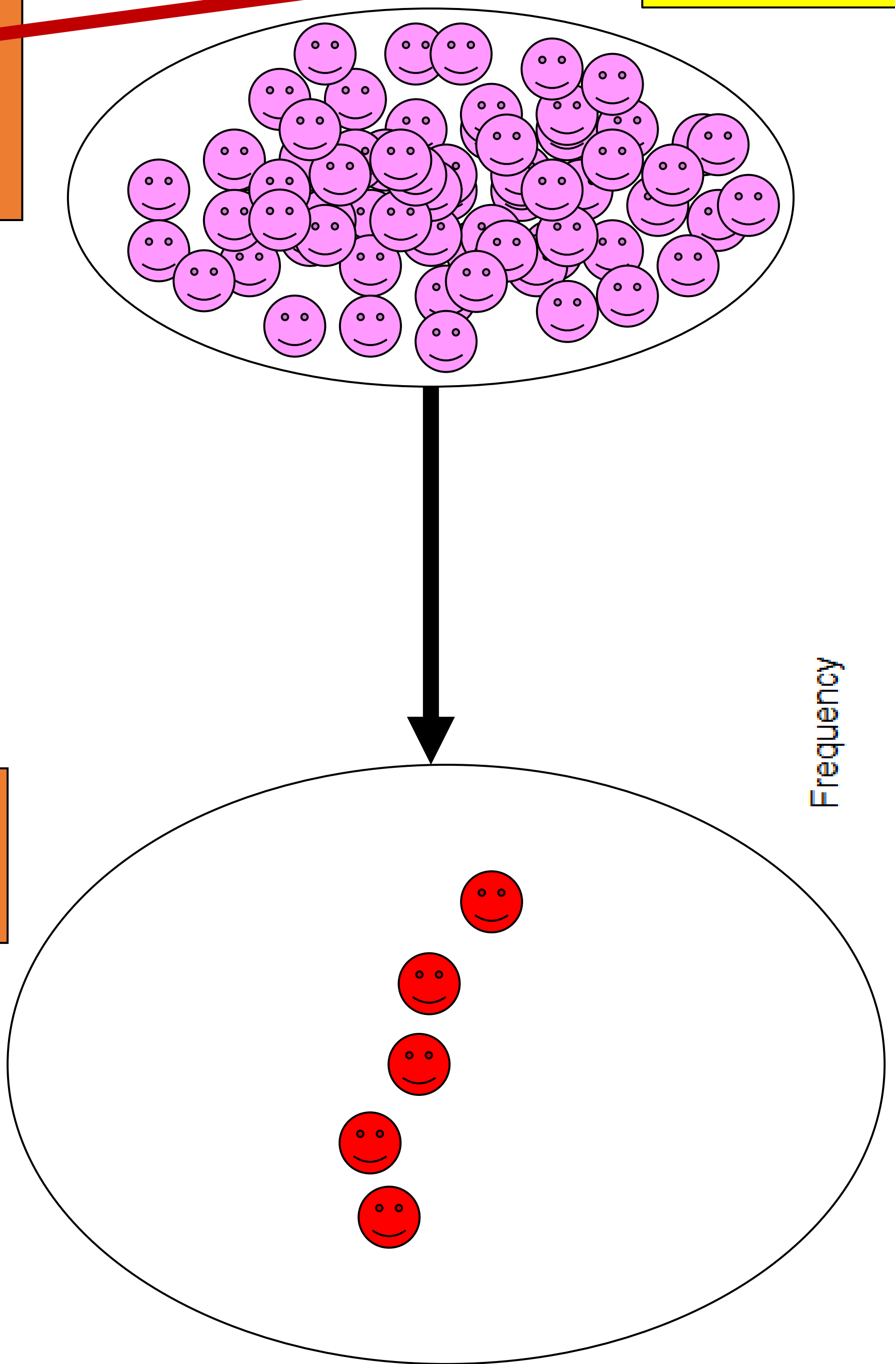
- We will test the null-hypothesis. Two conclusions possible:
  - Reject null in favor of the alternative
  - Not reject null (in that case we "retain" null, but we do not "accept" null)

- Therefore we will reason from the null: "if the null hypothesis were true, how probable would it be to find these or more extreme results?"
- So, let's assume that $p = p_0$ then what do we expect?

Population
$p_0 = 0.5$

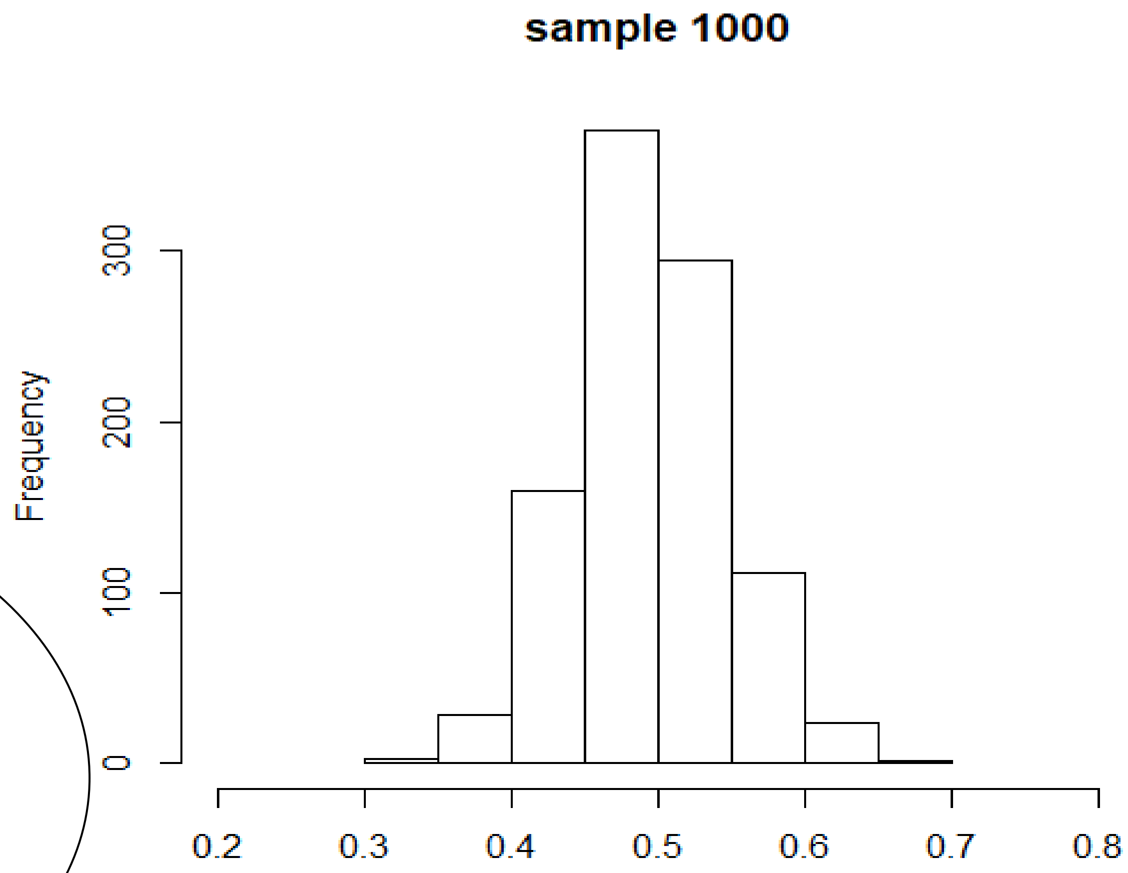
Population here refers to women using the diet

The proportion that is specified under H_0



Sample
$\hat{p} = 0.59$

For example, here suppose we take a sample of 100 women

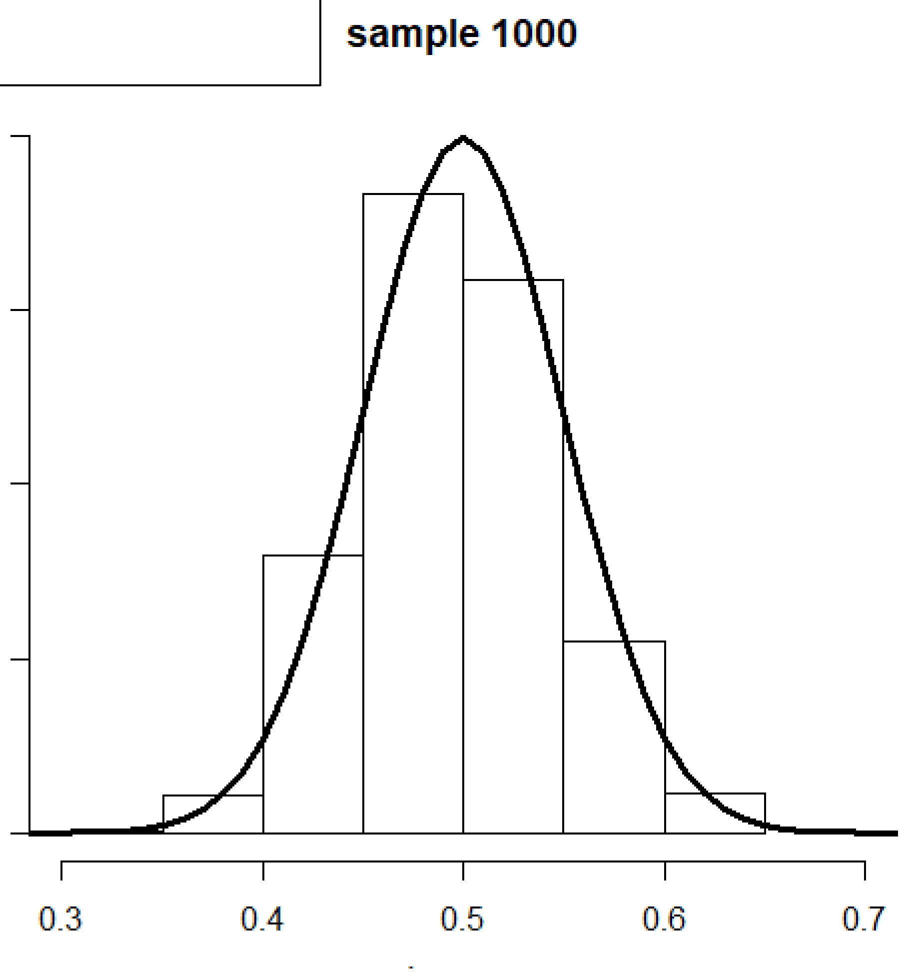


If np ≥ 15 and n(1-p) ≥ 15 then it is a normal distribution (Ch 7)

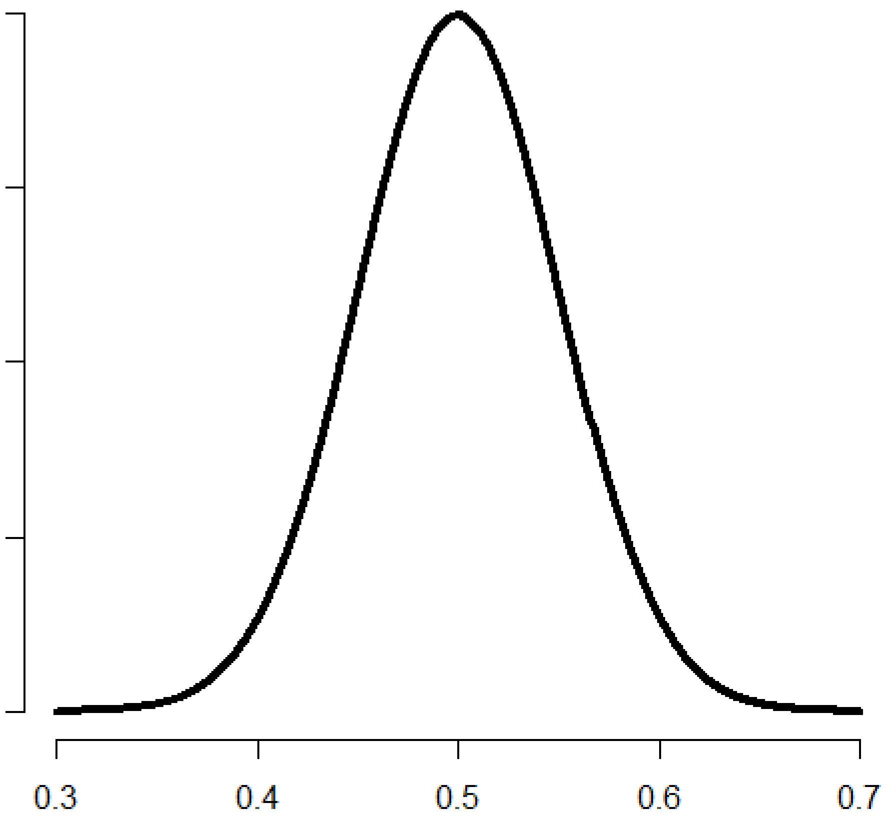
Mean: p_0

standard deviation (se_0): $\sqrt{\frac{p_0(1-p_0)}{n}}$

If H_0 is true



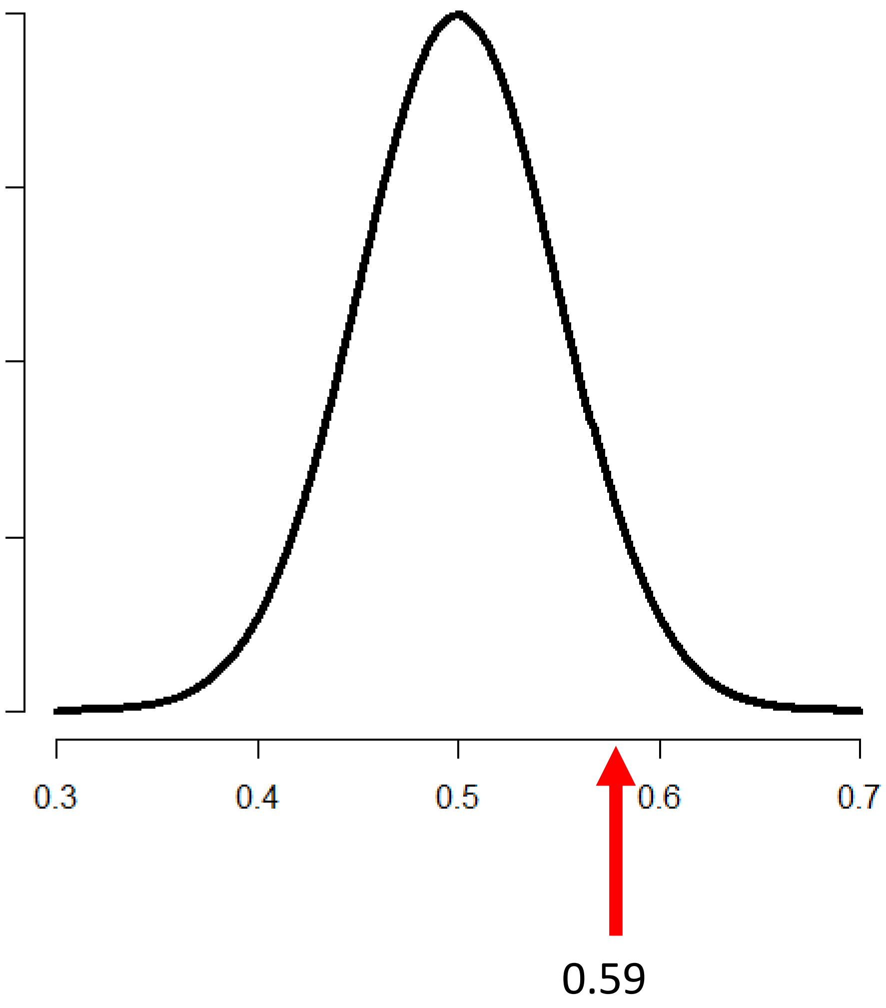
If H_0 is true



Is this sample proportion extreme if H_0 is true?

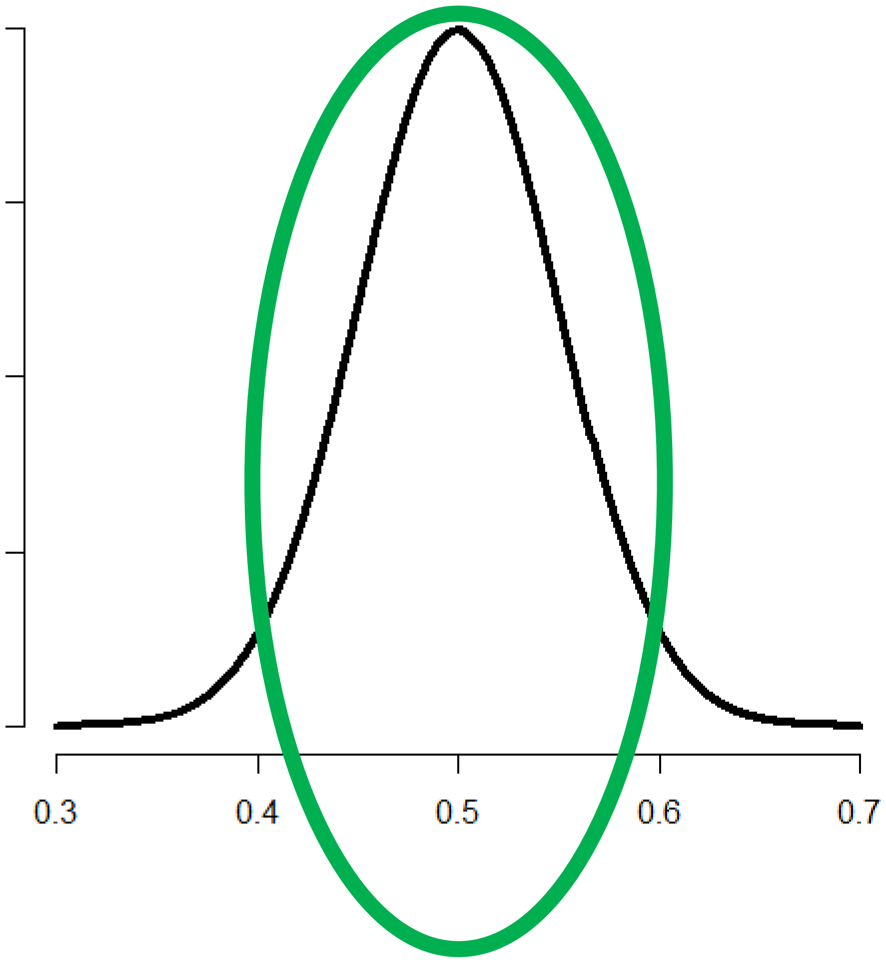
If H_0 is true

Let's suppose that the study looks at 100 women who are put on the diet, and in that sample the proportion of girls we find is 0.59 . ($\hat{p} = 0.59$)



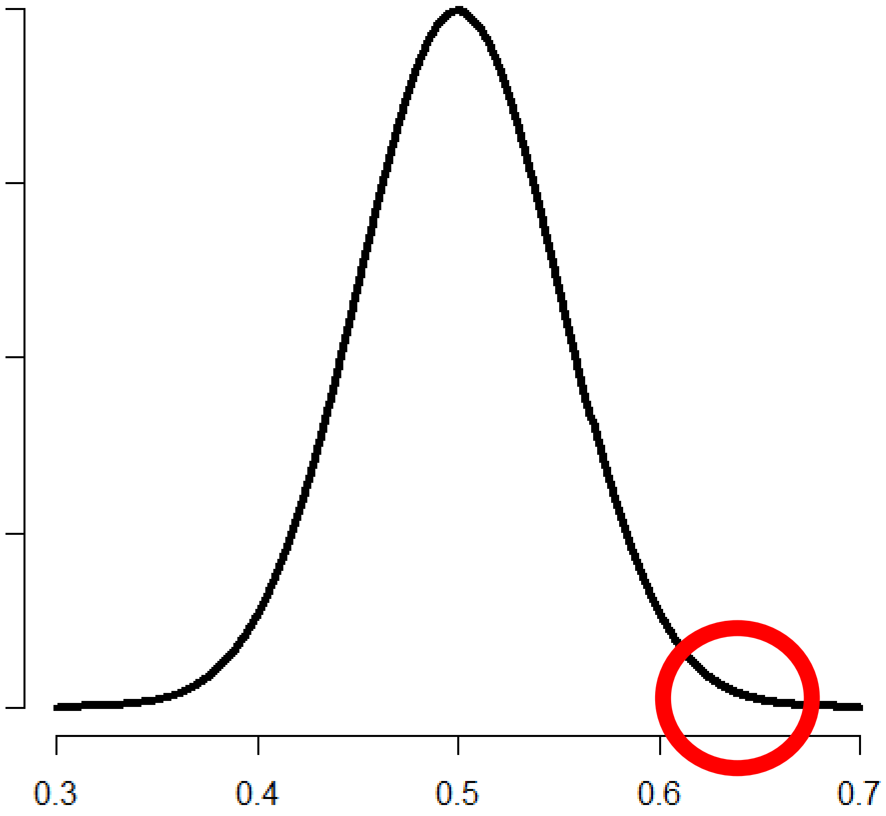
Is this sample proportion extreme if H_0 is true?

If H_0 is true



High probability to observe values in this range if H_0 is true (i.e., proportion is 0.5 in the population)

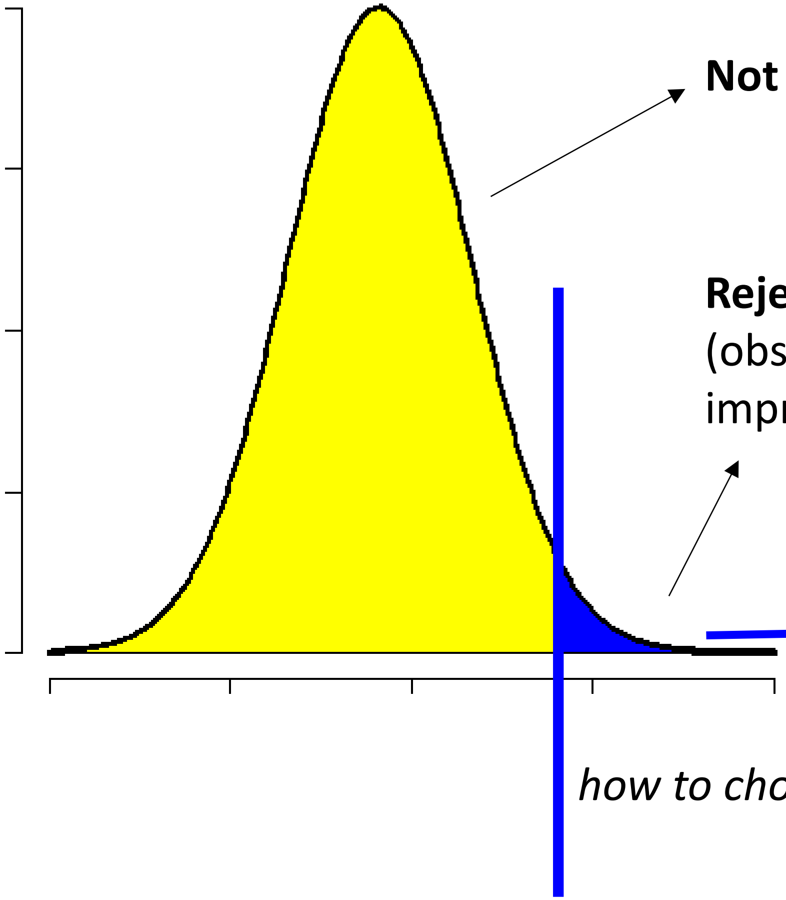
If H_0 is true



Extreme values to observe if proportion is 0.5 in the population

*Since H_a is p > 0.5 ("**larger** than .5") we **don't** consider the lower tail, only the upper tail is considered extreme*

$\alpha = 0.05$ means that if H_0 is true, then in 5% (or with probability .05) of the samples, we will incorrectly reject H_0. So, you choose an α that gives a "small" probability of an error when H_0 is true. Typically this is .05



Not reject H_0

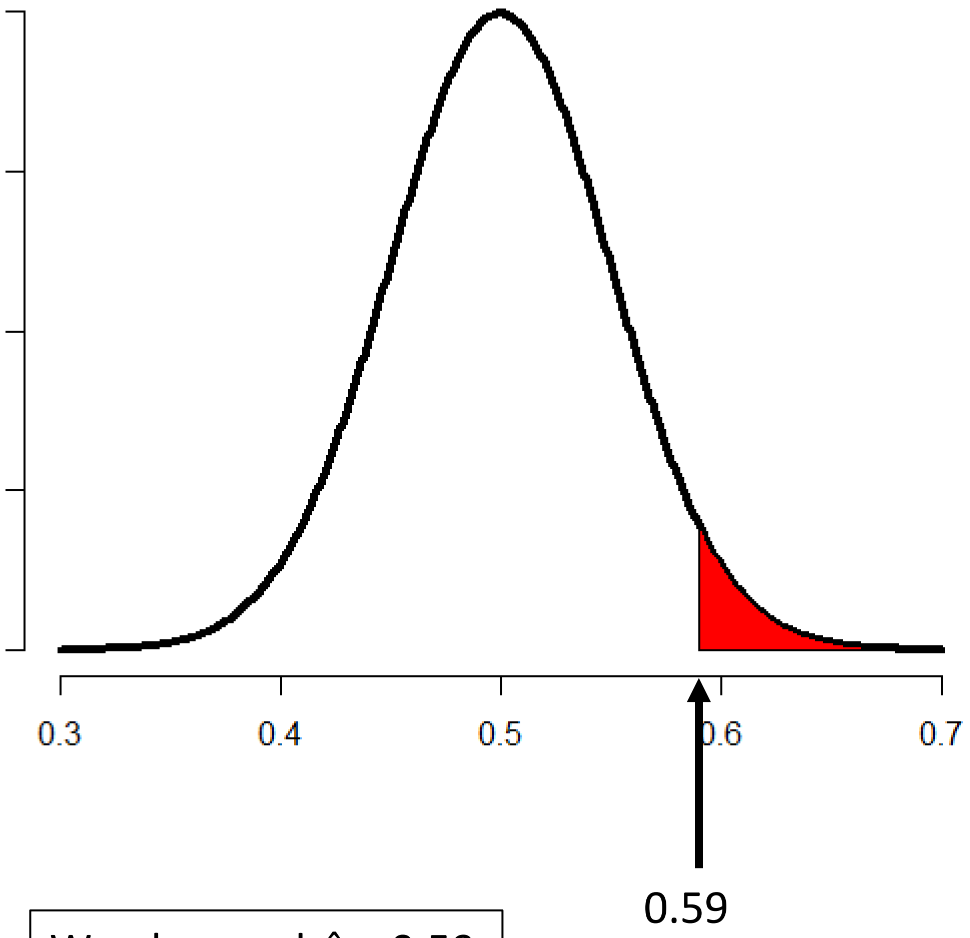
Reject H_0
(observing a statistic from this range is improbable if H_0 is true)

level of significance, α
e.g., 0.05 probability

how to choose this border??

*Since H_a is p > 0.5 ("**larger** than .5") we **don't** consider the lower tail, only the upper tail is considered extreme*

Next question: is the value you observed in this blue area or not? Calculate a p-value!
The p-value is the probability of observing the observed test statistic of more extreme if H_0 is true.
That is, the probability of the red area! Then check whether that's smaller than α.

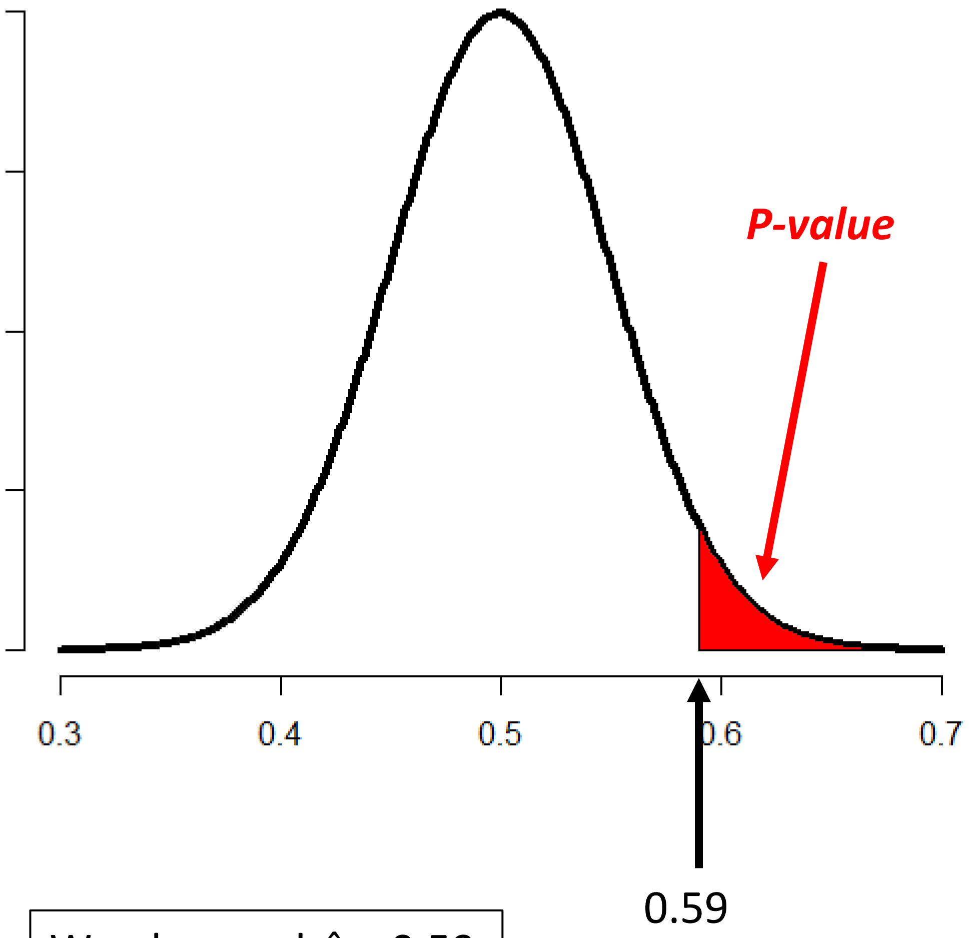


We observed $\hat{p}$ = 0.59
In a sample of n = 100

Should we reject H_0?

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}$$

Remember from lecture 9



P-value: *probability of these or extremer results if H_0 is true*

$$P(\hat{p} > 0.59)$$

$$z = \frac{\hat{p} - p_0}{se_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.59 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 1.8$$

$P(Z > 1.8) = P(\hat{p} > 0.59)$ = *p-value*

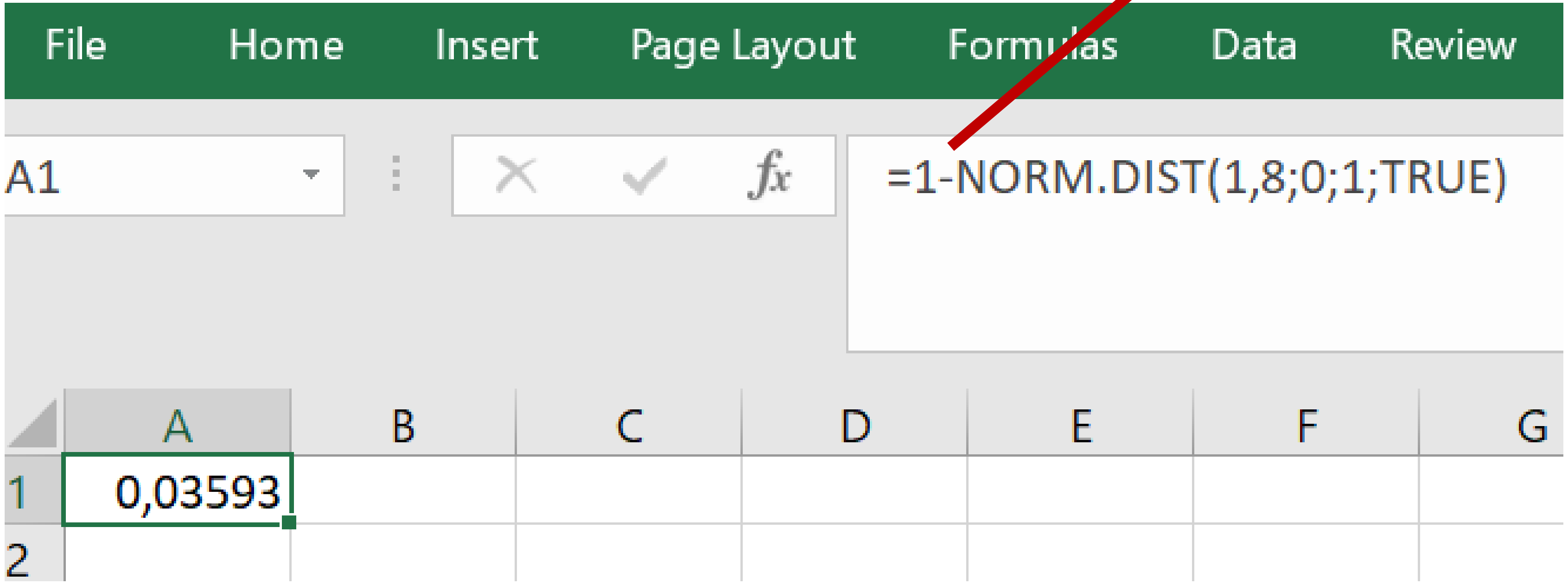
Test statistic: a measure of how far the point estimate ($\hat{p}$) falls from the parameter value in the null hypothesis, p_0

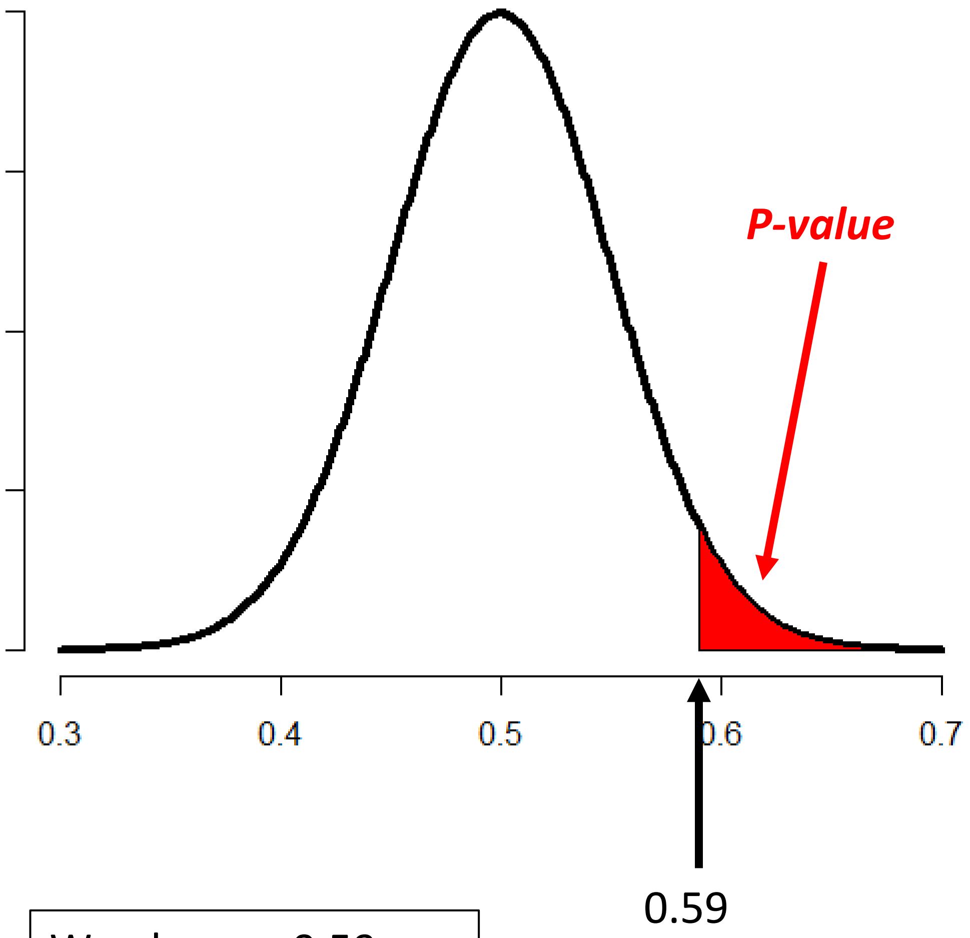
We observed $\hat{p}$ = 0.59
In a sample of n = 100

Should we reject H_0?

Calculate using excel:

We use “1 – “ to get the upper tail probability (the right side)





P-value: *probability of these or extremer results if H_0 is true*

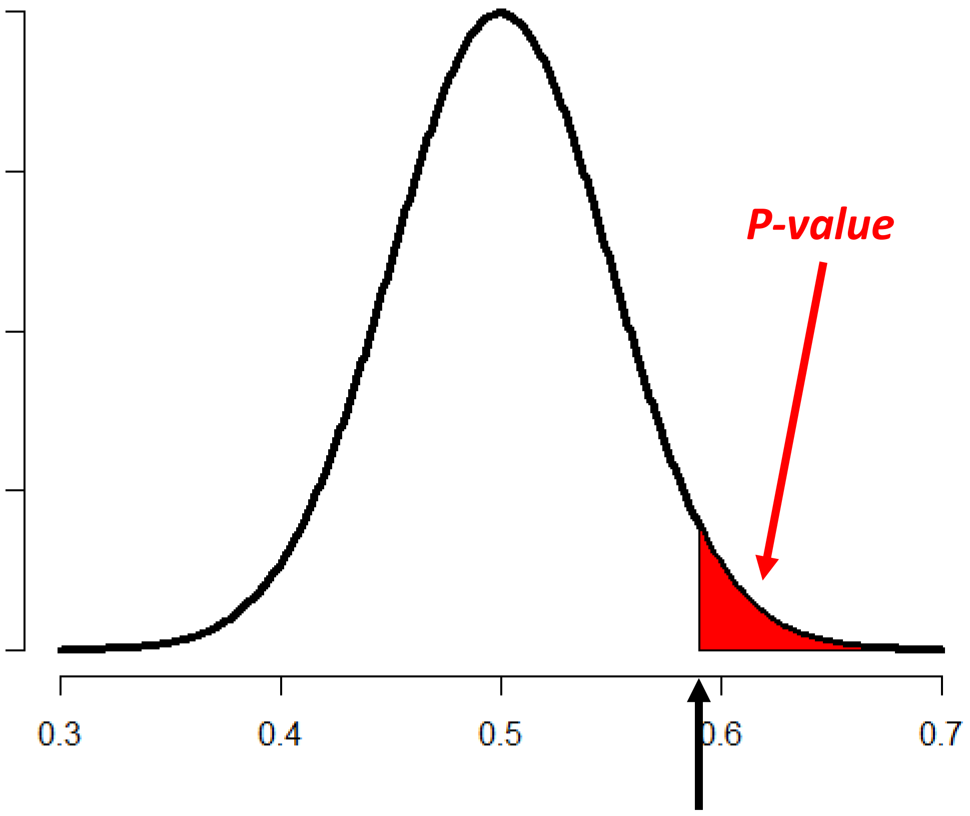
$$P(\hat{p} > 0.59)$$

$$z = \frac{\hat{p} - p_0}{se_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.59 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 1.8$$

$P(Z > 1.8) = P(\hat{p} > 0.59)$ = *p-value* = ***0.036***

We observe 0.59

Should we reject H_0?



***P-value:** probability of these or extremer results if H_0 is true*

$$P(\hat{p} > 0.59)$$

$$z = \frac{\hat{p} - p_0}{se_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.59 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 1.8$$

$P(Z > 1.8) = P(\hat{p} > 0.59)$ = *p-value* = ***0.036***

We observe 0.59

Should we reject H_0?

If P-value < α the answer is yes.

So for an α of 0.05 the answer is yes!

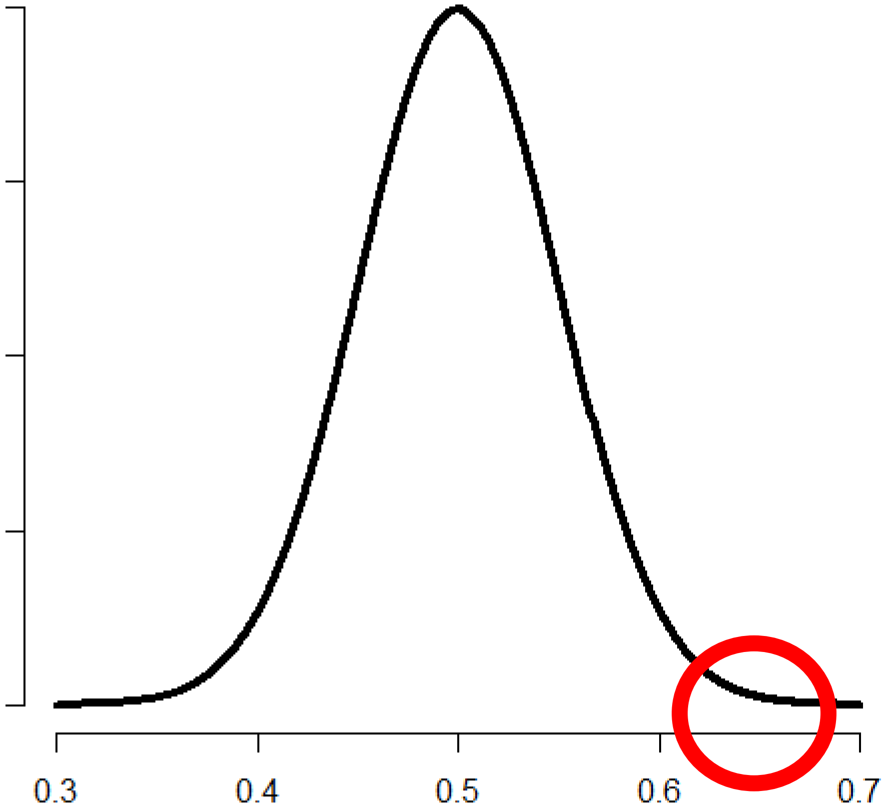
Conclusion

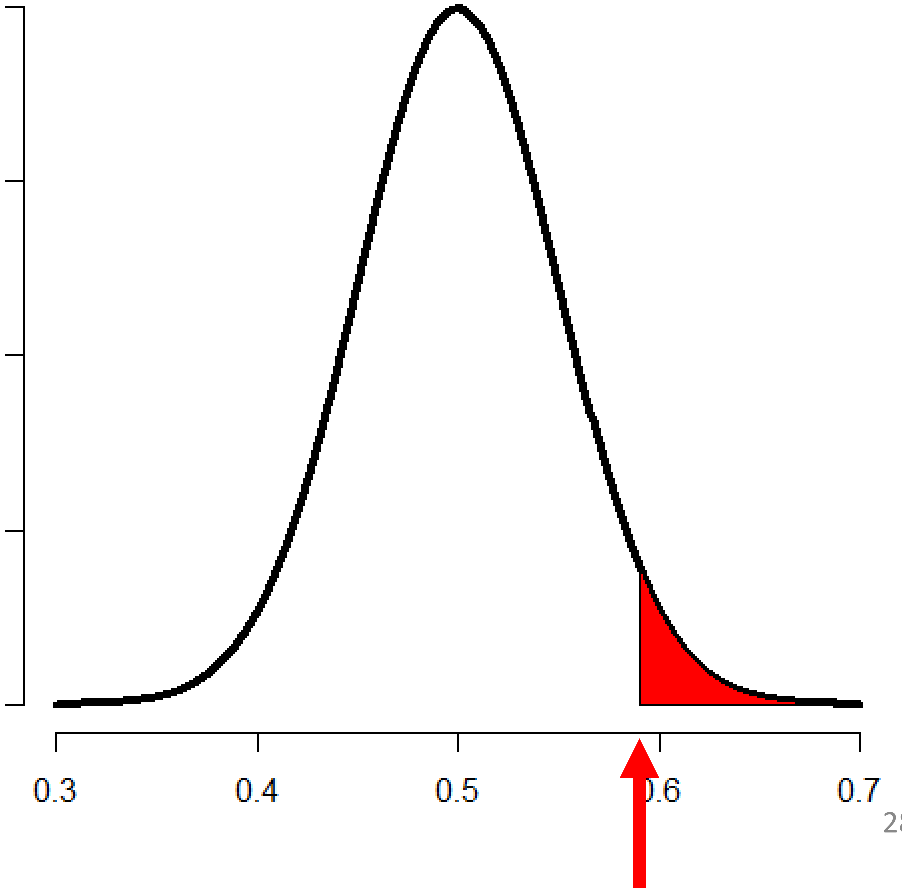
- P-value in this example equals 0.036
- Thus: If H_0 is true (i.e., birth rate girls equals 0.5) the probability of observing *these or extremer* results is 0.036
- If this probability is smaller than **the level of significance (α)**, you conclude that H_0 should be rejected in favor of H_a
- Commonly a level of significance of 0.05 is used. This means that if H_0 is true and we reject H_0 for samples that give a p-value < α, we will incorrectly reject H_0 in 5% of cases (= small error probability)
- → Thus in this case, we reject the null hypothesis and conclude that the true proportion is **significantly** larger than 0.5 (We find support for $H_a : p > 0.5$)

One sided tests

- This was a one sided test

H_0: p = 0.5 H_a: p > 0.5





One sided tests

- Example:

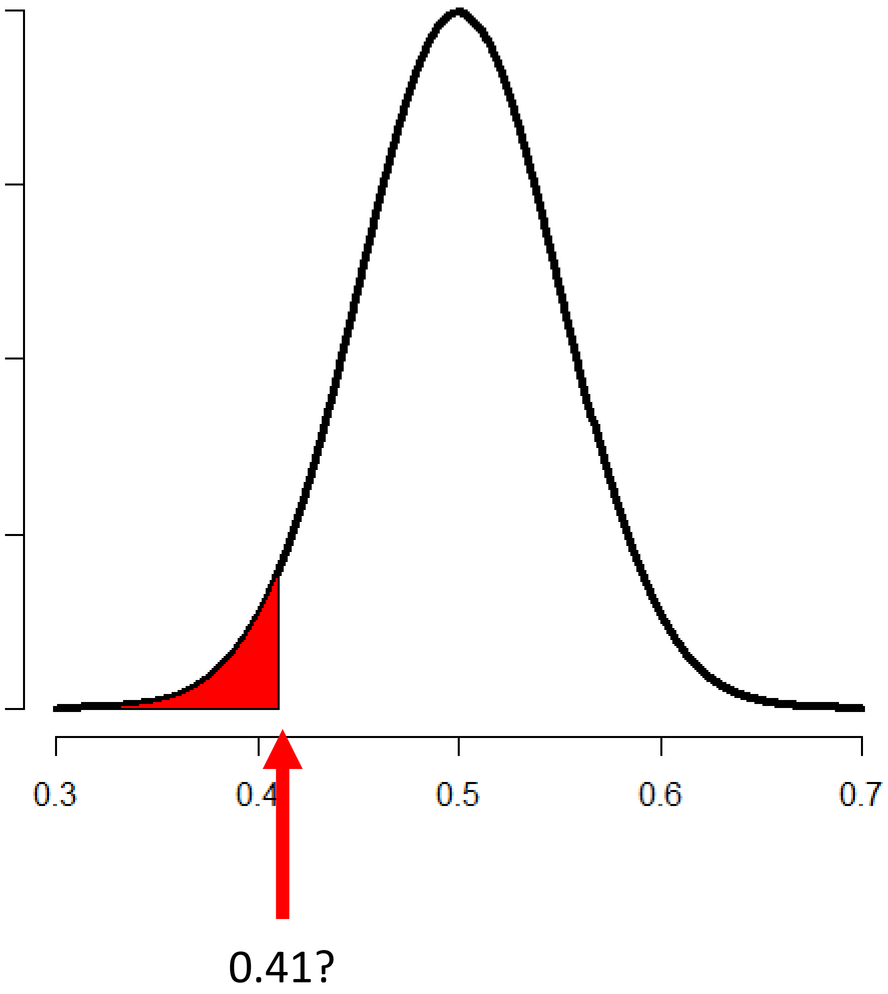
"If you have severe morning sickness during pregancy you will likely get a boy"

→Probability of girl will be *smaller* than 0.5 for people with severe morning sickness

Suppose we draw a sample and observe $\hat{p}$ = 0.41

Is this result significantly smaller than 0.5, if we use a significance level of 0.05?

One-sided tests

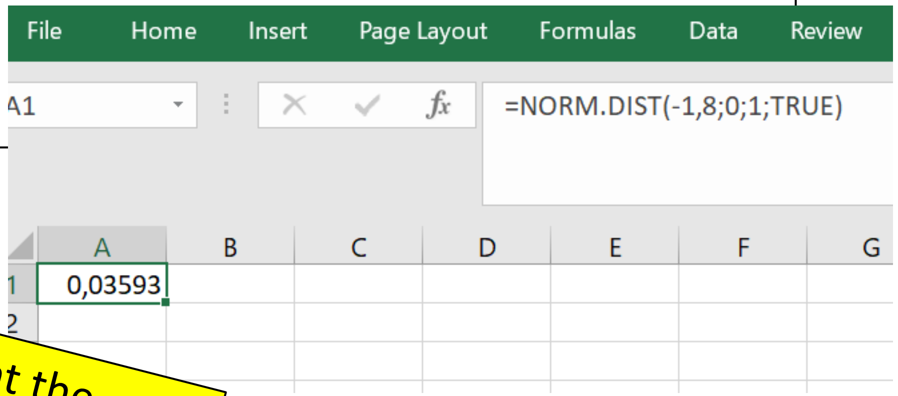


P-value: *probability of these* ***or extremer*** *results if H_0 is true*

$$P(\hat{p} < 0.41)$$

$$z = \frac{\hat{p} - p_0}{se_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \frac{0.41 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = -1.8$$

$P(z < -1.8) = 0.036$

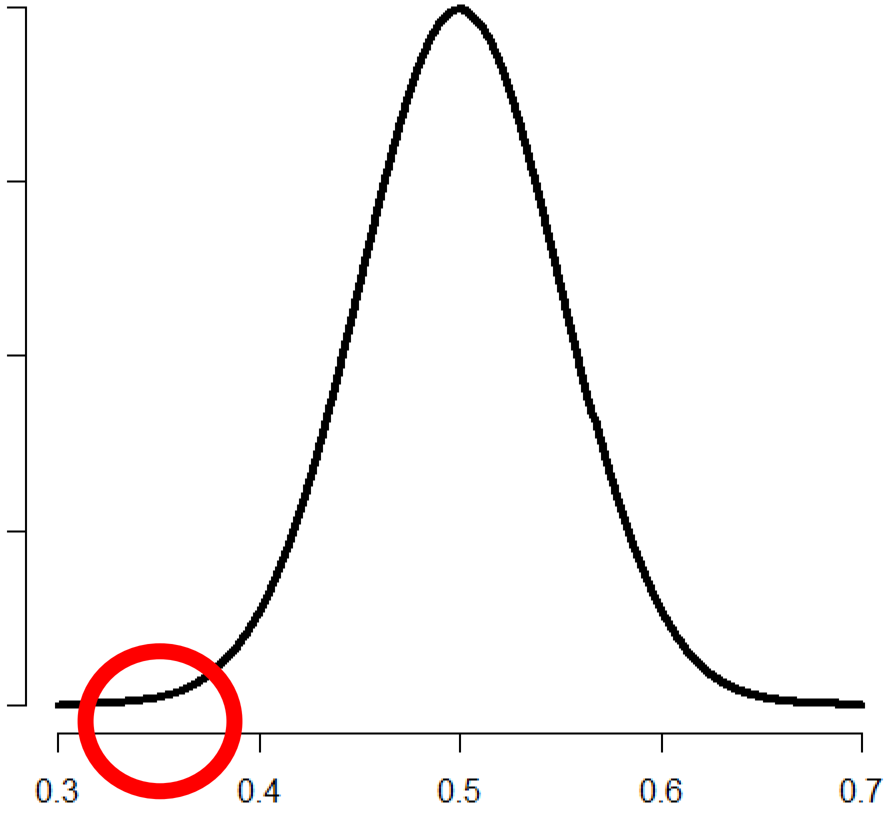


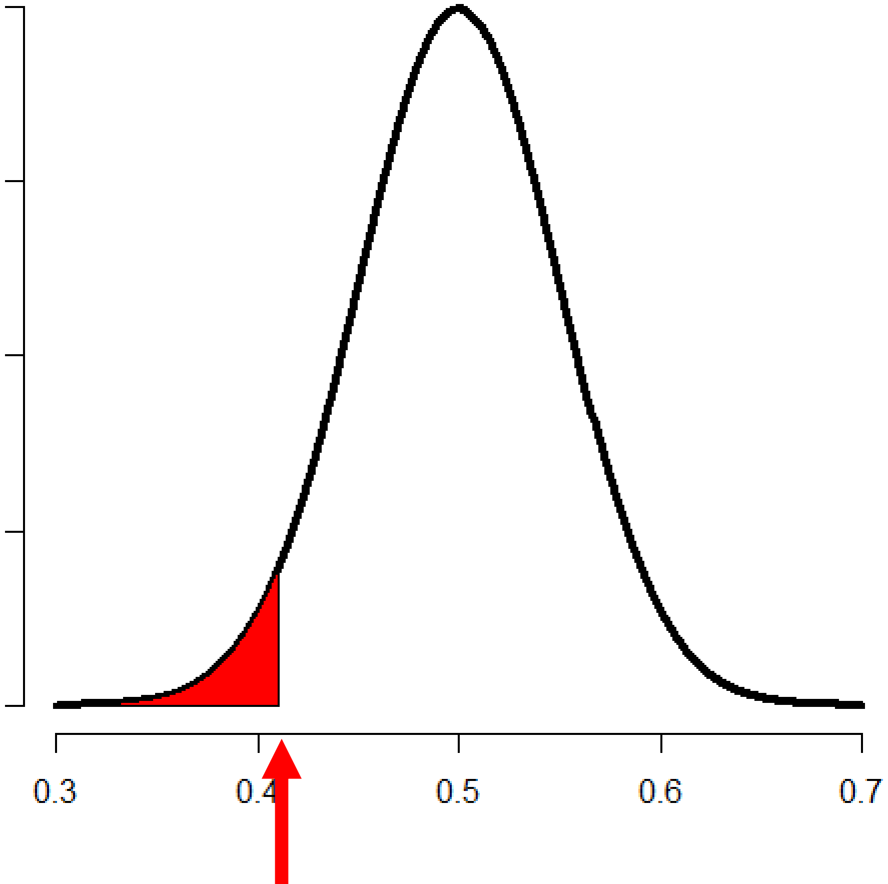
Now we want the lower tail probability!!

One sided tests

- This is also a one sided test, but different direction

H_0: p = 0.5 H_a: p **<** 0.5

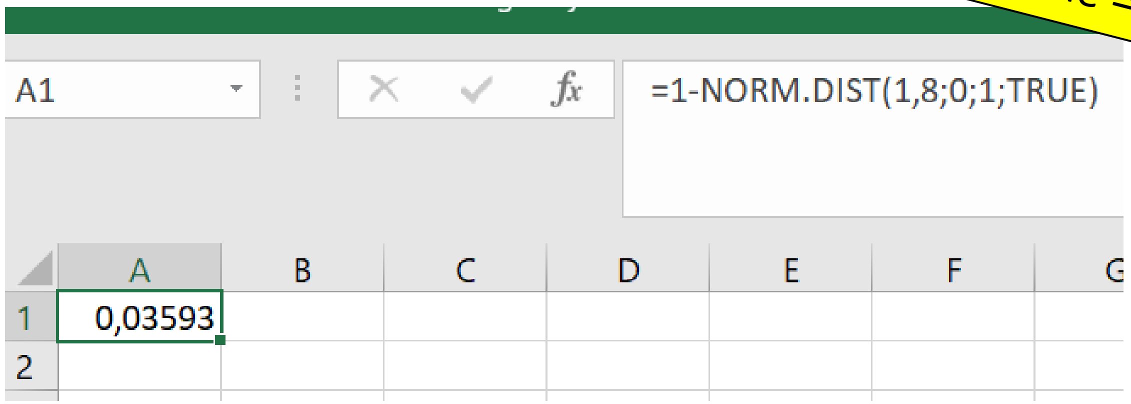


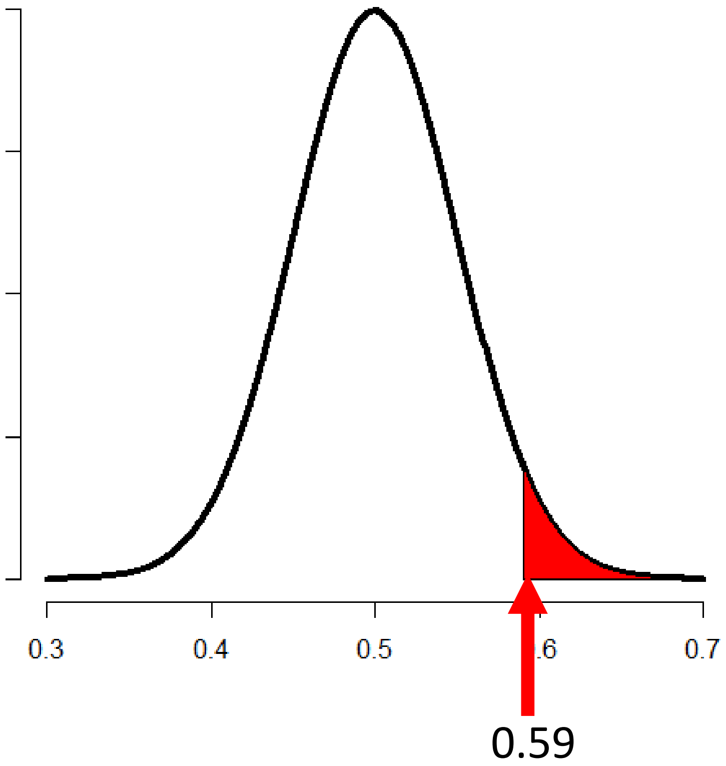


One sided tests

- Note that in a one-sided test, the choice of calculating the lower or upper tail from a observed value depends on the direction of the hypothesis!
- E.g., suppose we find that $\hat{p}$ = 0.59
- If $H_a = p > 0.5$ then:

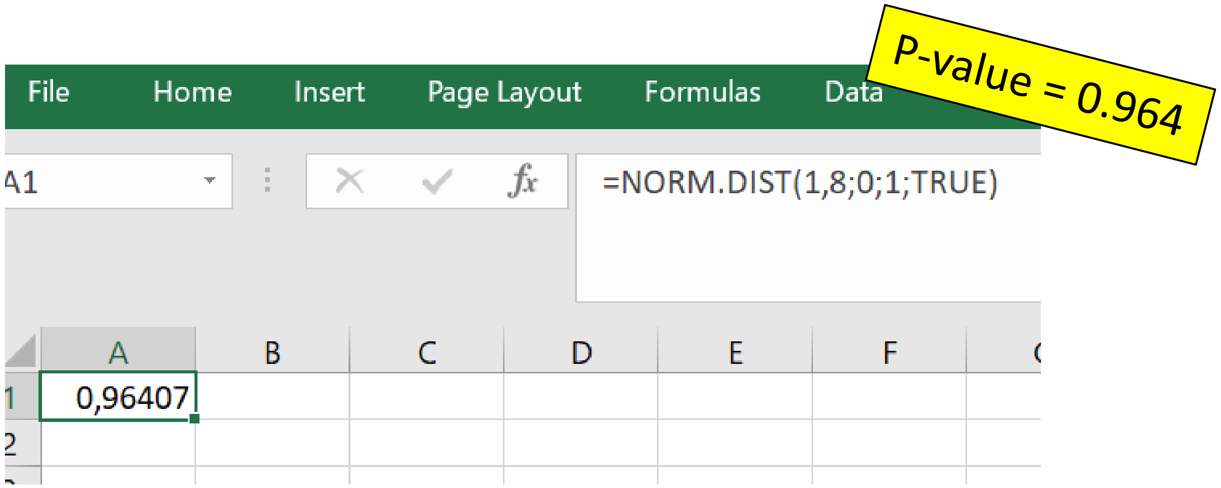
P-value = 0.036

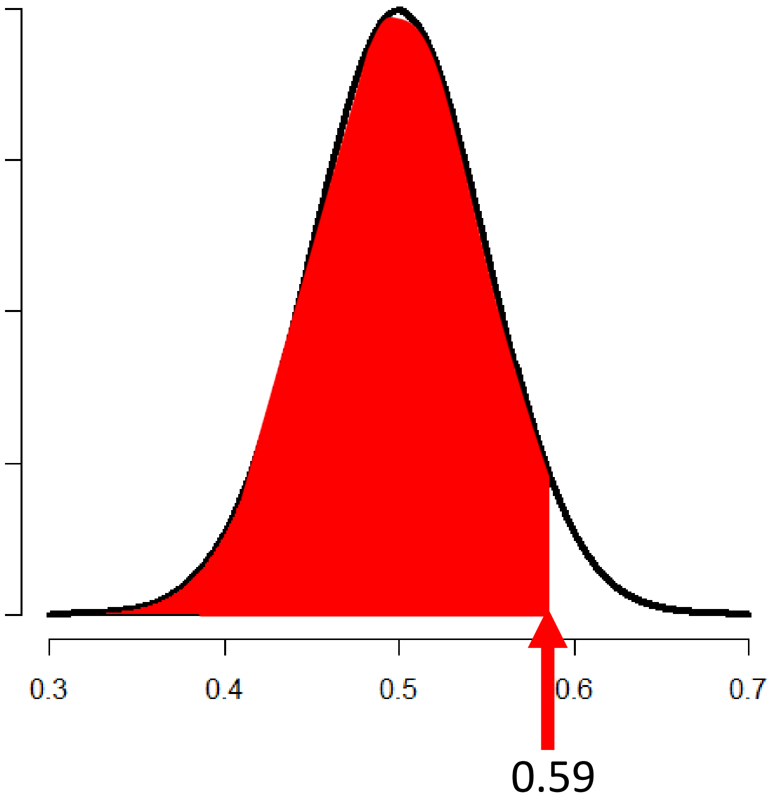




One sided tests

- Note that in a one-sided test, the choice of calculating the lower or upper tail from a observed value depends on the direction of the alternative hypothesis!
- E.g., suppose we find that $\hat{p} = 0.59$
- However, if $H_a = p < 0.5$ then:





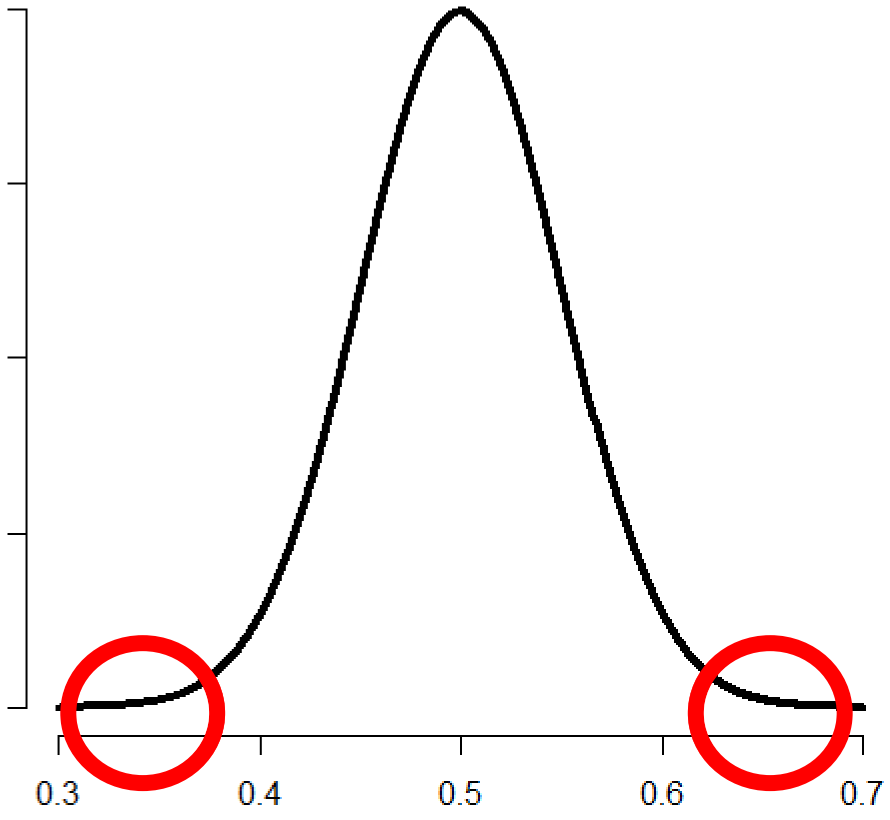
Two sided tests

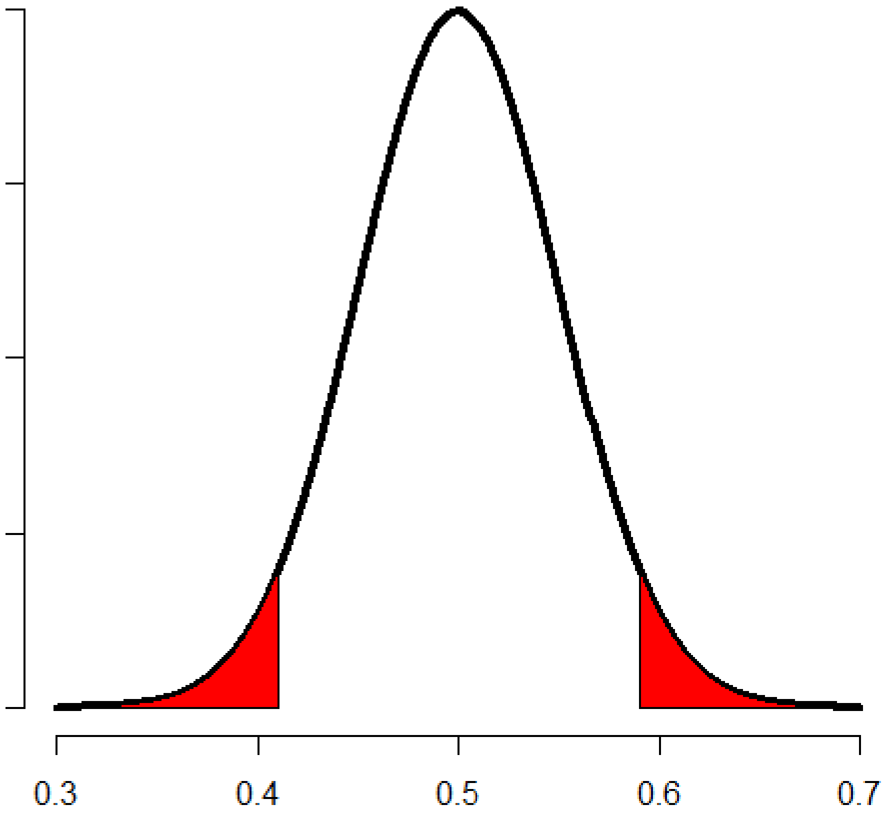
- If you don’t have a strong expectation, you need a two sided test
- E.g., is the probability on a baby girl different from 0.5?

Two sided tests

- This is a two sided test

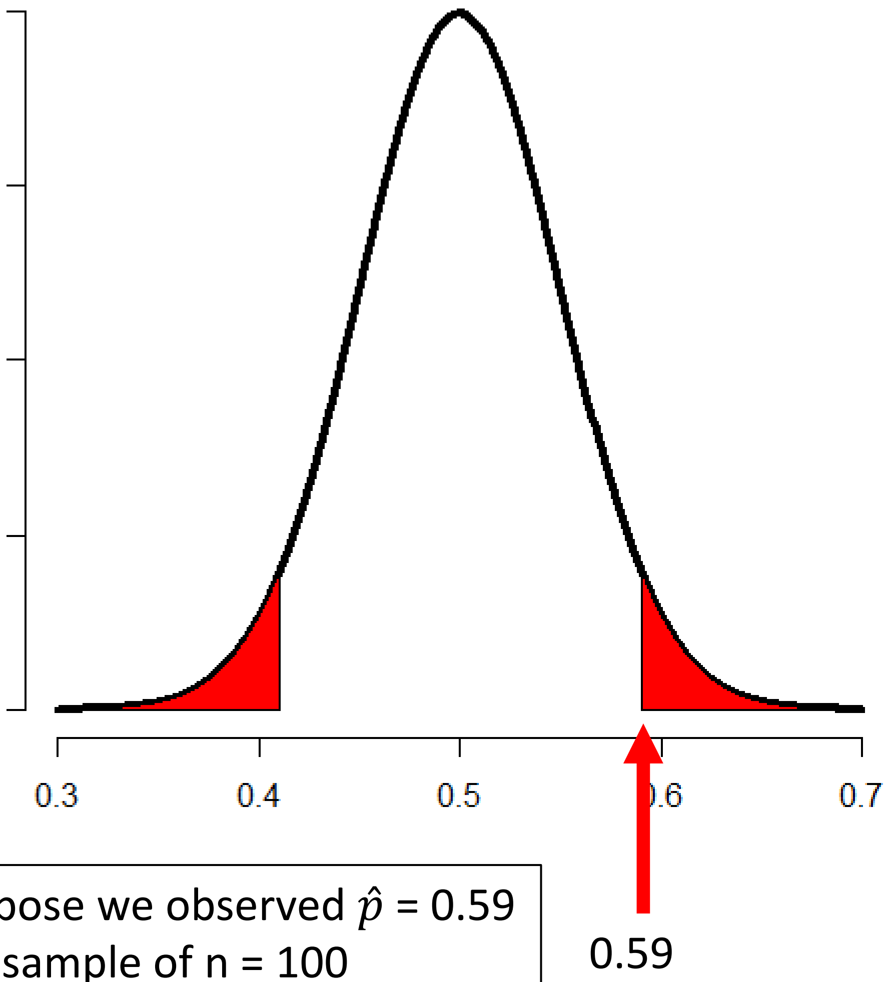
H_0: p = 0. 5 H_a: p $\neq$ 0.5





P-values in a two-sided test

Both tails are considered extreme now as we have a two-sided test!



Suppose we observed $\hat{p}$ = 0.59
In a sample of n = 100

Should we reject H_0?

P-value: *probability of these* ***or extremer*** *results if H_0 is true*

$2 \times P(\hat{p} > 0.59)$ We calculate ">" because .59 is on the left side

$$z = \frac{\hat{p} - p_0}{se_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \frac{0.59 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 1.8$$

P(z > 1.8) = 0.036 (from excel)
Thus p-value equals 2 × 0.036 = ***0.072***
For an α of 0.05, this p-value is not significant -> don't reject H_0

In a two-sided test, the alternative does not have a direction, so you calculate the smaller tail and double it

Conclusion

Based on the P-value draw a conclusion about H_0 (reject / not reject)

P-value < level of significance → reject H_0 in favor of H_a

P-value > level of significance → **not** reject H_0

reject H_0 in favor of H_a → women on diet X are more likely to get a girl

not reject H_0 → we could not reject the null hypothesis and so we did not find evidence for the effect of diet X on how likely it is to get a girl

Today

Hypothesis test for a proportion

Hypothesis test for a mean

Correspondence to confidence interval

Sometimes we make the wrong decision

Example

- A researcher is interested in whether children who bike more than 5km to school score higher on a fitness test than the general population of that age. The mean score of this age category on a certain fitness test equals 8.0.

- To test whether the fitness of these children is significantly higher than 8.0, the trainer randomly samples 31 children who bike > 5km to school and administers the fitness test

- He finds:
  - mean score: 8.43
  - standard deviation: 2.0.

- Is the mean significantly larger than 8.0?

Picture sourxe pixabay

Hypothesis test for a mean

- Specify hypotheses
    - H_0: $\mu = 8$
    - H_a: $\mu > 8$
        - Because we expect the biking to increase fitness (it will not *decrease* fitness)

Hypothesis test for a mean

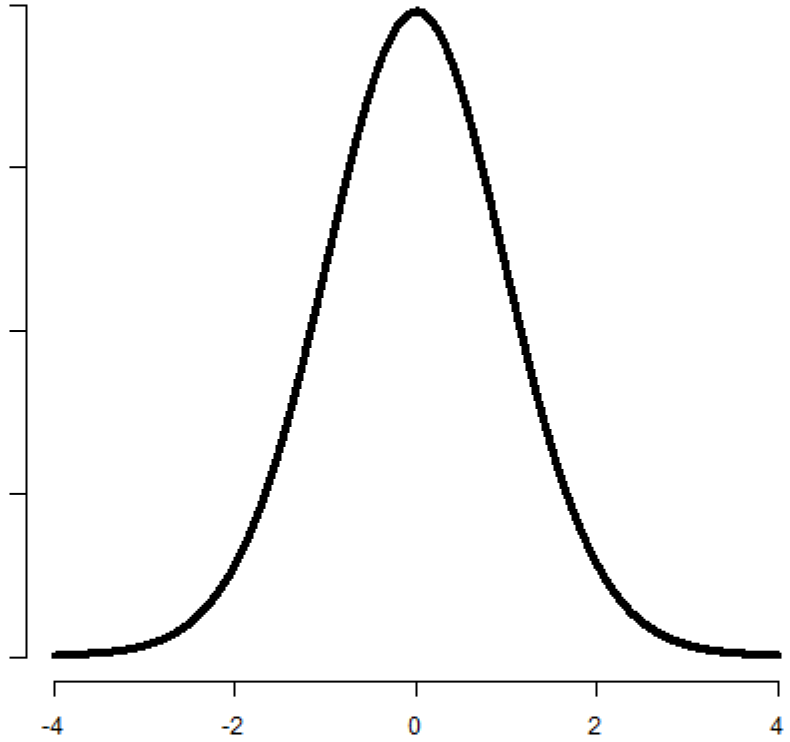
- Calculate test statistic:

  - Recall from Chapter 7: using s to estimate σ introduces additional error and therefore we use the t-distribution.

  - The test statistic will thus be a t-value:

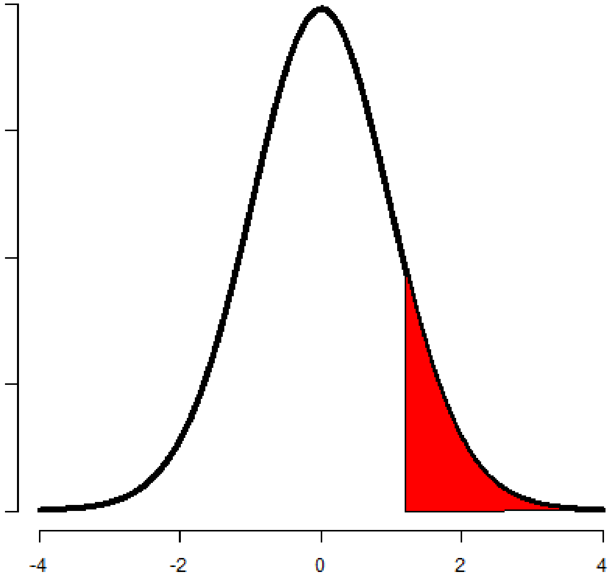
$$t = \frac{\bar{x} - \mu_0}{se} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{8.43 - 8}{2/\sqrt{31}} = 1.20$$

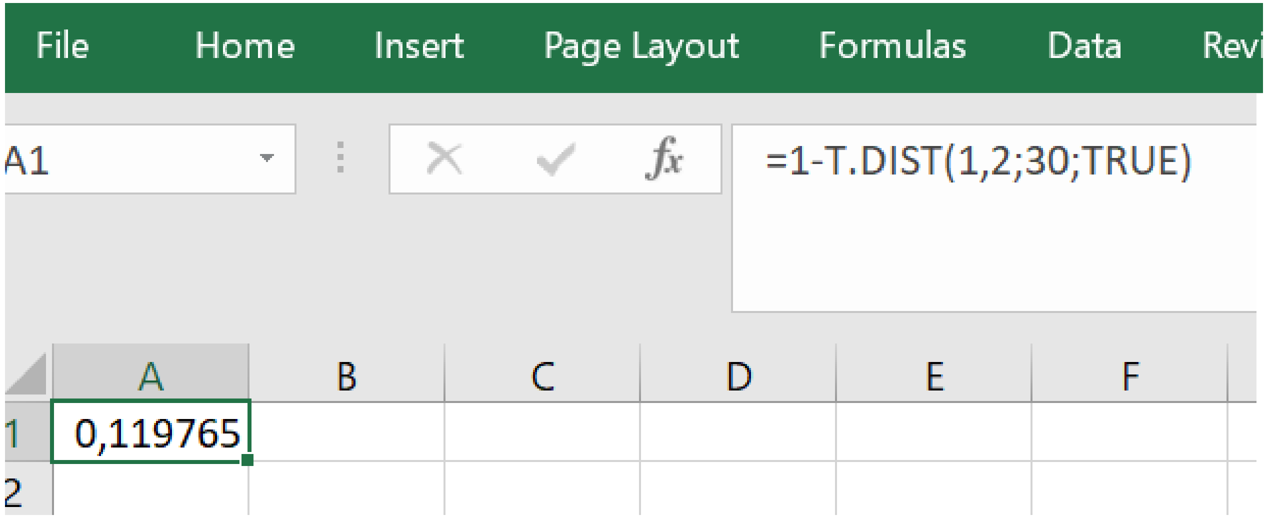
  - with *n - 1* degrees of freedom = 31 - 1 = 30

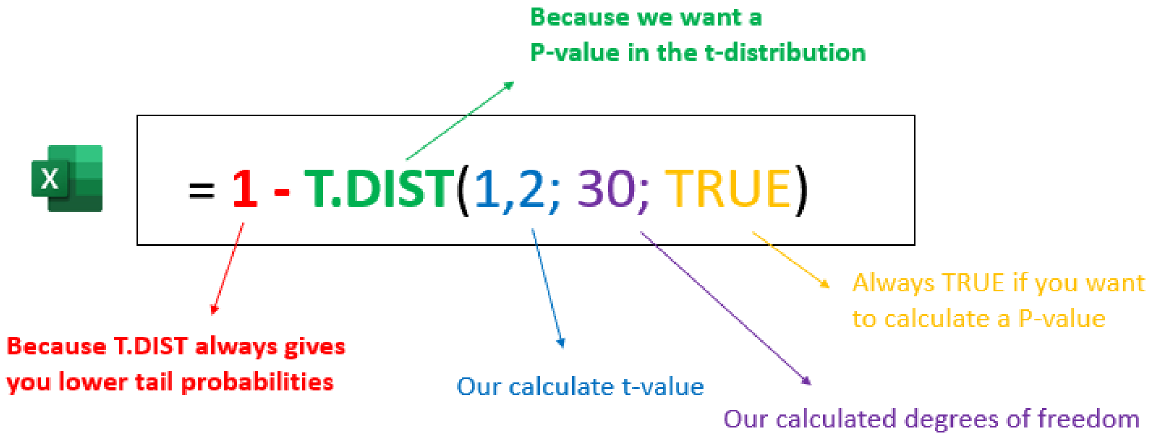


Hypothesis test for a mean

- Calculate P-value
  - $t = \frac{\bar{x}-\mu_0}{se} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{8.43-8}{2/\sqrt{31}} = 1.20$
  - with *n - 1* degrees of freedom = 31 - 1 = 30







Hypothesis test for a mean

Draw conclusion:

- *p* = 0.12, which means that the probability of observing 8.43 or larger is 0.12 if H_0 is true.

- Suppose we choose α=0.05. P-value it is not smaller than 0.05, so we do not reject H_0

- So, our observed sample mean (8.43) does **not** differ significantly from 8

- So biking did not significantly increase physical fitness scores

Today

Hypothesis test for a proportion

Hypothesis test for a mean

Correspondence to confidence interval

Sometimes we make the wrong decision

Correspondence to confidence interval

- Conclusions from a *two-sided significance test* will agree with conclusions drawn from a confidence interval

- e.g.,: a **two-sided** significance test with a significance level of 0.05 will produce the same conclusions as a 95% confidence interval

- e.g.,: a **two-sided** significance test with a significance level of 0.01 will produce the same conclusions as a 99% confidence interval

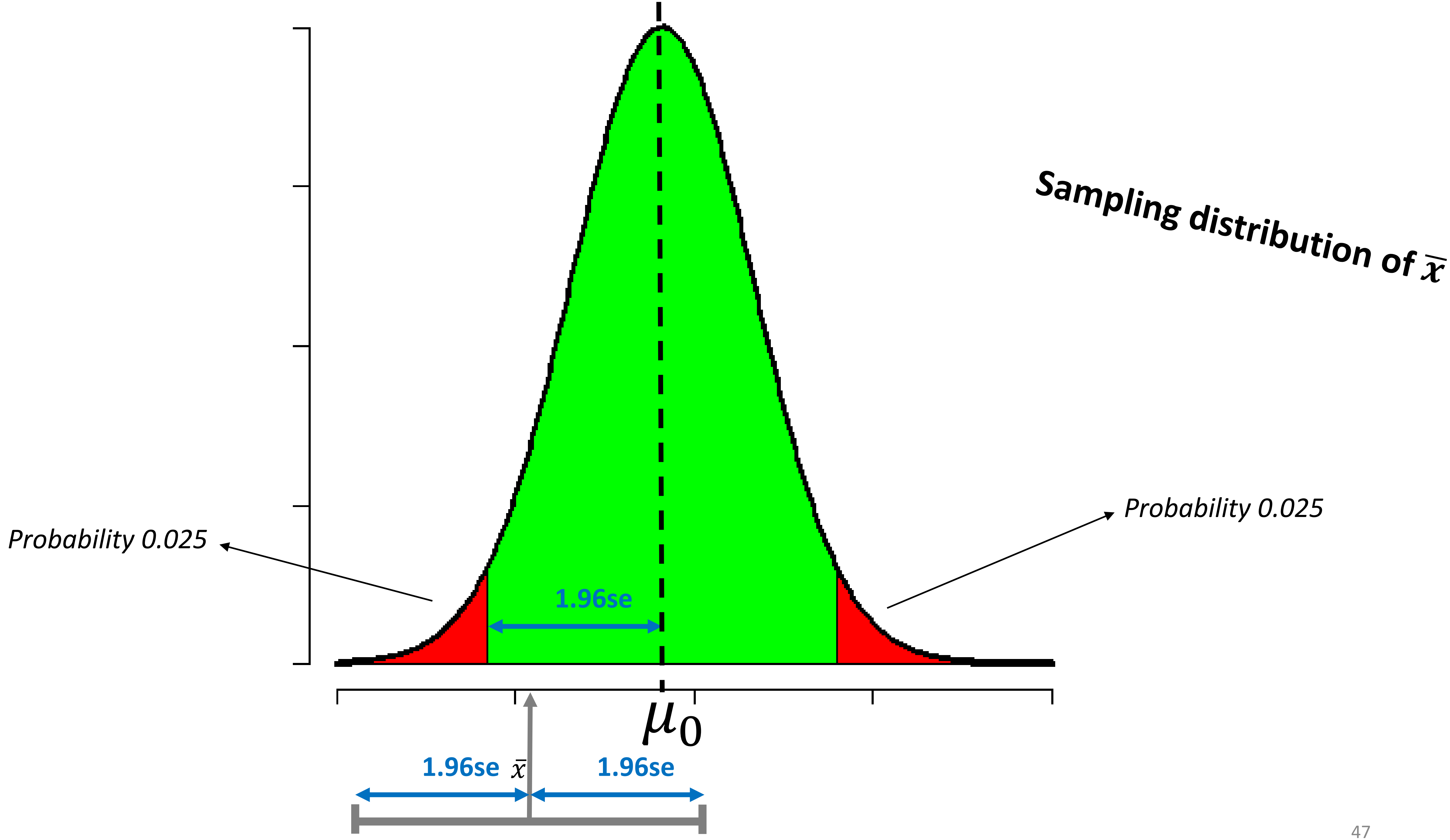
Correspondence to confidence interval

- For example, in the fitness example before:
  - mean score: 8.43
  - standard deviation: 2.0.

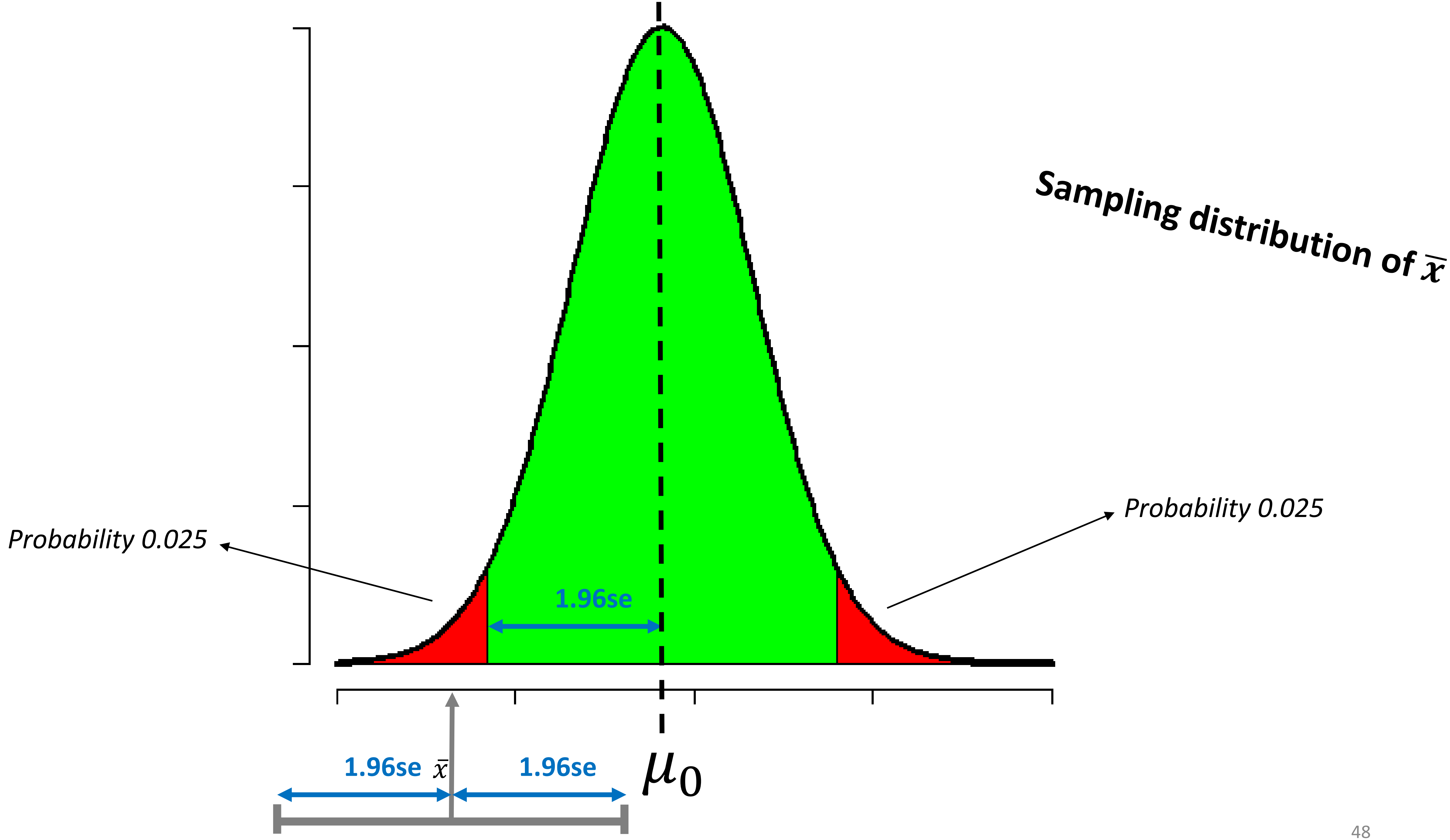
- If we did a **two-sided** significance test with significance level of 0.05 ..
  - H_0: $\mu_0 = 8$
  - H_a: $\mu_a \neq 8$
- …P-value $= 2 \times 0.12 = 0.24$, so you **won't** reject the null-hypothesis

If we calculated a 95% confidence interval, it equals (7.697; 9.16)

- The interval **does** contain 8. So, we draw the same conclusion: you **won't** reject H_0



Sampling distribution of $\overline{x}$
Probability 0.025
Probability 0.025
1.96se
μ_0
1.96se
$\bar{x}$
1.96se


Sampling distribution of $\bar{x}$
Probability 0.025
Probability 0.025
1.96se
1.96se
$\bar{x}$
1.96se
μ_0

Today

Hypothesis test for a proportion

Hypothesis test for a mean

Correspondence to confidence interval

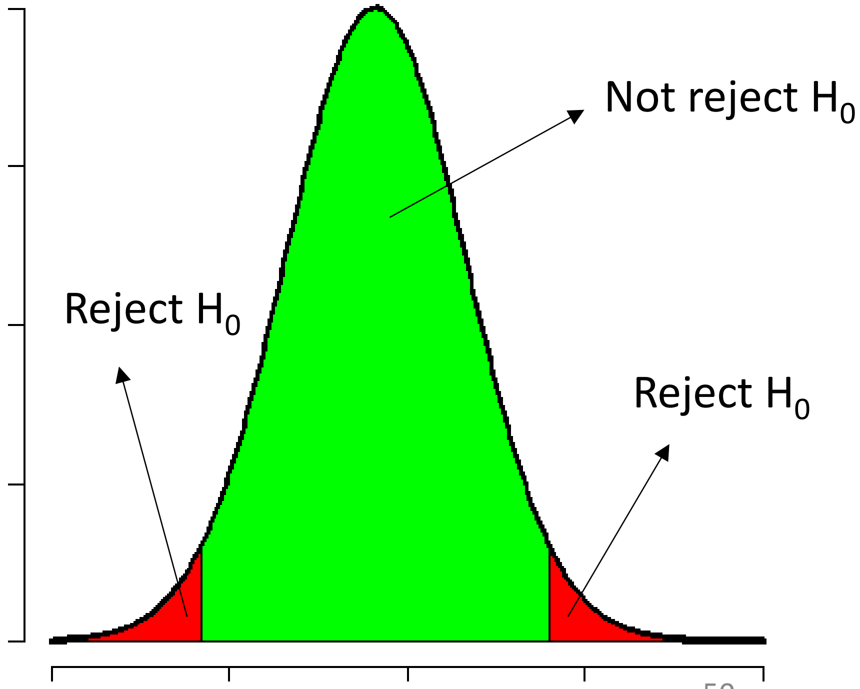
Sometimes we make the wrong decision

Sometimes we make the wrong decision

- If we find a P-value smaller than a level of significance α (for example, α = .05) we argue this:

> "If H_0 is true, it is improbable that I observe data as extreme or more extreme as what I have observed. Thus, I conclude H_0 is not true"

However there is still a chance that H_0 is true!
(i.e., 0.05 in this case)
The nice thing is, we control the probability of this error by choosing α!



Sometimes we make an error (draw wrong conclusion)

Statistical decision

True state of the world

	H_0 is not rejected	H_0 rejected
H_0 is true		
H_a is true		

Sometimes we make an error (draw wrong conclusion)

Statistical decision

True state of the world

	H_0 is not rejected	H_0 rejected
H_0 is true		*Type I Error (α)*
H_a is true		

Sometimes we make an error (draw wrong conclusion)

Statistical decision

True state of the world

	H_0 is not rejected	H_0 rejected
H_0 is true		*Type I Error (α)*
H_a is true	*Type II Error*	

Sometimes we make an error (draw wrong conclusion)

Statistical decision

True state of the world

	H_0 is not rejected	H_0 rejected
H_0 is true	***1 - P(Type I error)***	*Type I Error (α)*
H_a is true	*Type II Error*	

Sometimes we make an error (draw wrong conclusion)

Statistical decision

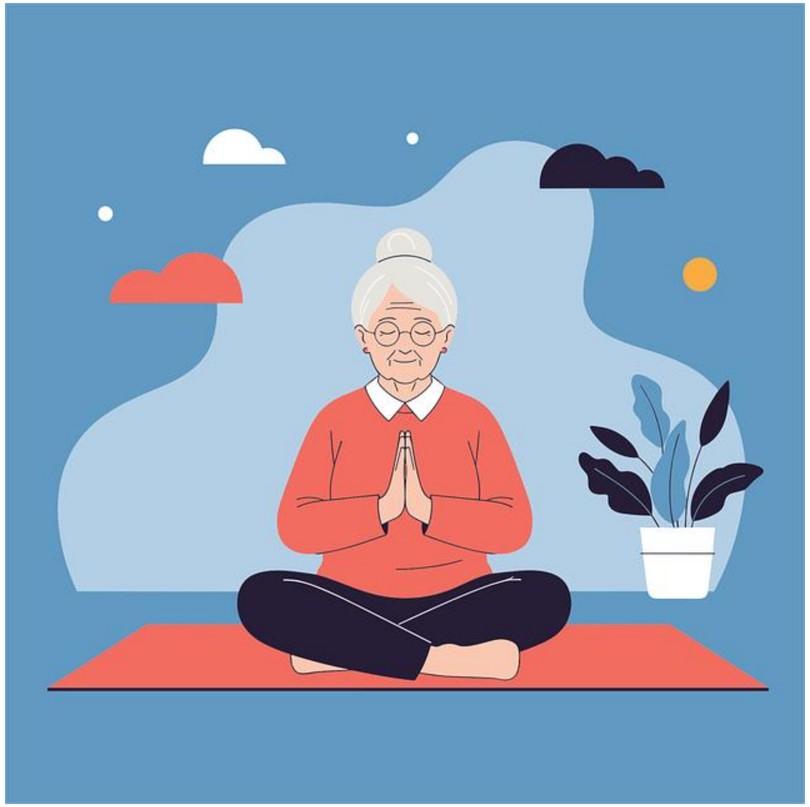
True state of the world	H_0 is not rejected	H_0 rejected
H_0 is true	***1 - P(Type I error)***	*Type I Error (α)*
H_a is true	*Type II Error*	***Power = 1 - P(Type II error)***

Example exam question

A researcher adminsters a concentration test to a group of elderly people (n = 30) who do yoga each day. It is known that elderly have a mean score of 15.1 in the population, but the researcher expects that elderly who do yoga regularly will have a better concentration (H_a: $\mu > 15.1$).

In her sample, the researcher finds a mean of 21.9 (SD = 15) . What is the p-value?

a) 0.01
b) 0.02
c) 0.05



Picture source: pixabay

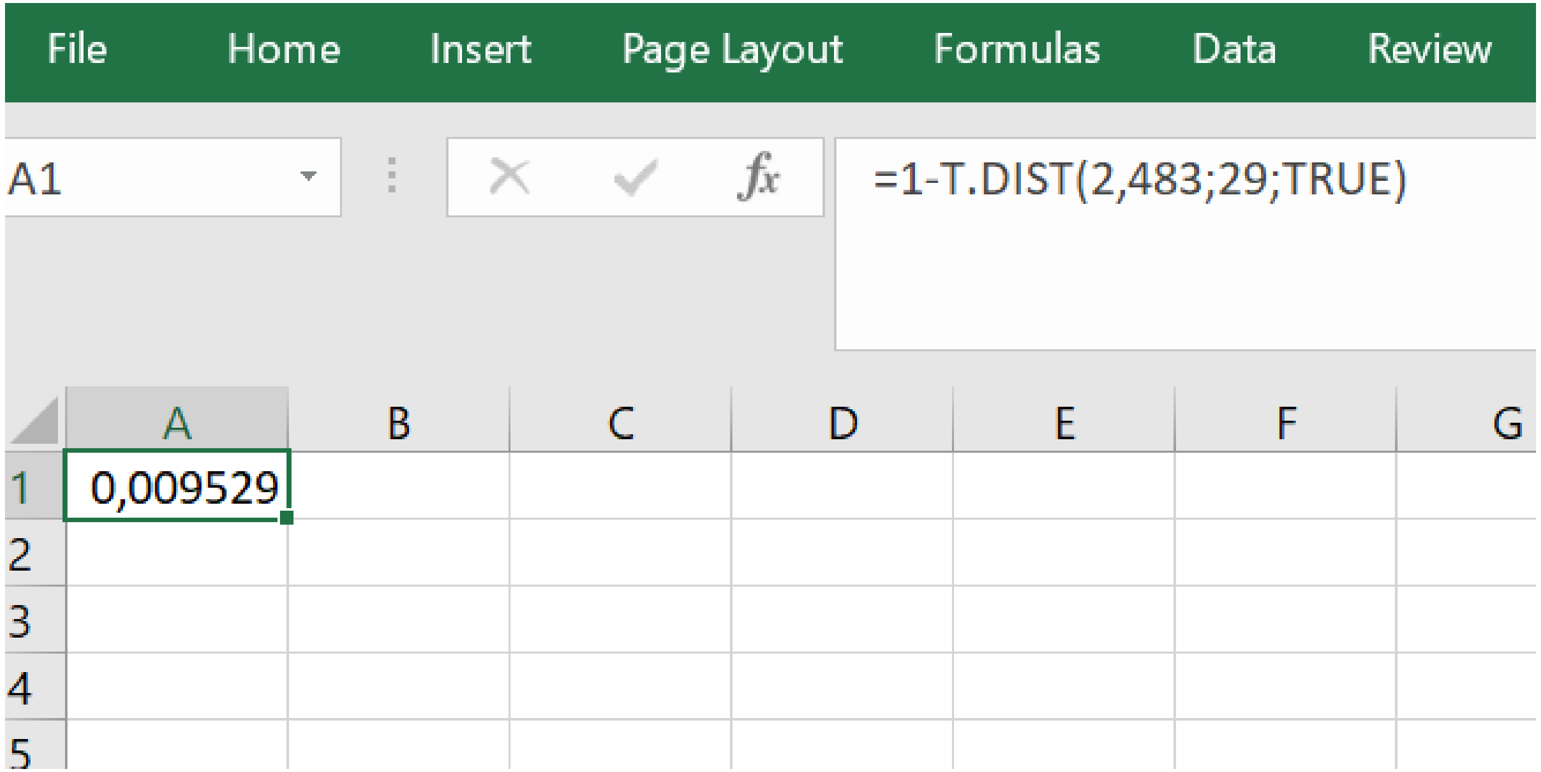
Solution example exam question

$$t = \frac{21.9 - 15.1}{15/\sqrt{30}} = 2.483$$

df = 30 – 1 = 29

P = 0.0095 (from Excel)

So, rounded this is 0.01



Example exam question

- Nina uses a level of significance of 0.05. This means that the

a) Type I error is 0.05
b) Type II error is 0.05
c) power is 0.05

Solution example exam question

- Nina uses a level of significance of 0.05. This means that the

a) Type I error is 0.05

b) Type II error is 0.05

c) power is 0.05

Example excercise from the book

p.479

9.17 Another test of astrology Examples 1, 3, and 5 referred to a study about astrology. Another part of the study used the following experiment: Professional astrologers prepared horoscopes for 83 adults. Each adult was shown three horoscopes, one of which was the one an astrologer prepared for him or her and the other two were randomly chosen from ones prepared for other subjects in the study. Each adult had to guess which of the three was his or hers. Of the 83 subjects, 28 guessed correctly.

a. Define the parameter of interest and set up the hypotheses to test that the probability of a correct prediction is $1/3$ against the astrologers' claim that it exceeds $1/3$.

b. Show that the sample proportion $= 0.337$, the standard error of the sample proportion for the test is 0.052, and the test statistic is $z = 0.08$.

APP **c.** Find the P-value. Would you conclude that people are more likely to select their horoscope than if they were randomly guessing, or are results consistent with random guessing?

Solution to 9.17

a) p = proportion of adults who guess correctly

$H_0\colon\ p = \frac{1}{3}$

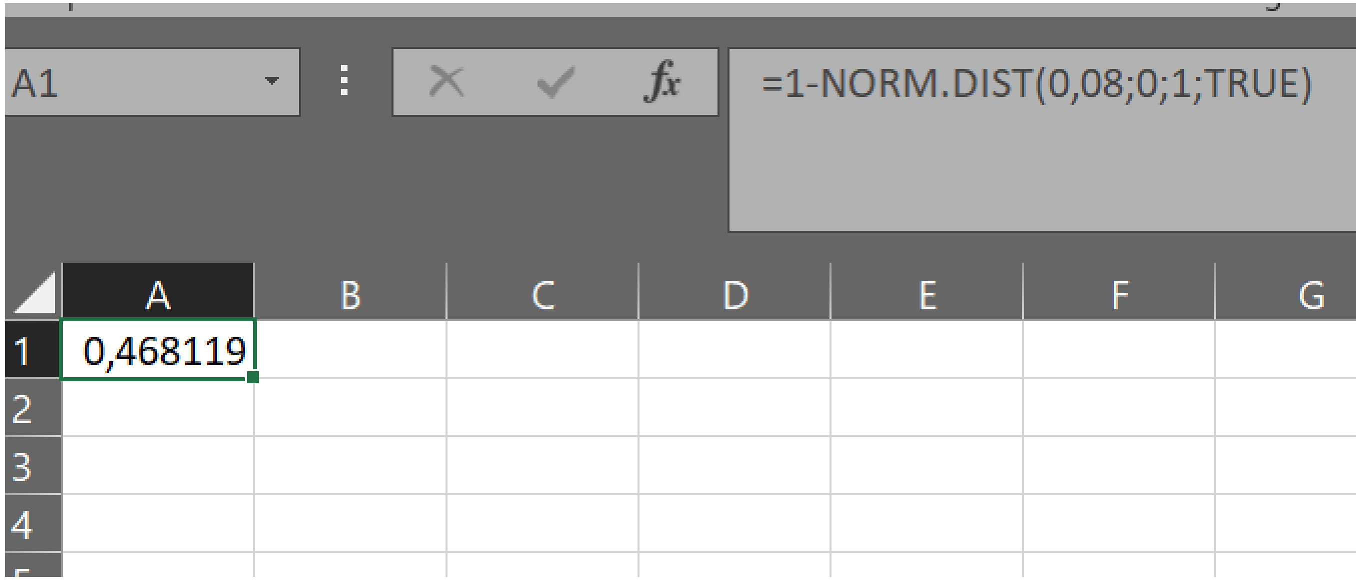
$H_a\colon p > \frac{1}{3}$

b) 28/83 = 0.337

$\sqrt{\frac{\frac{1}{3}(1-\frac{1}{3})}{83}}$=0.052

$\frac{\hat{p}\ - p_0}{se_0} = \frac{0.377 - 0.333}{0.052} = 0.08$

c) p-value = 0.47



This is much larger than an alpha of 0.05, so results are consistent with random guessing.

Some other relevant excercises in the book

- 9.2, 9.4, 9.5, 9.7, 9.10, 9.11, 9.12, 9.18, 9.23, 9.27, 9.28, 9.29, 9.30, 9.31, 9.34, 9.42, 9.43, 9.45, 9.50, 9.52

- For any questions that you like help with: ask on the discussion board on Canvas!

- The next slides summarize the five steps of a significance test

The five steps of a significance test

- Step 1: Assumptions
  - For proportion:

    Variable is categorical

    Data are obtained using randomization (random sample or random assignment)

    Sampling distribution of proportion is normal if $np \geq 15$ and $n(1-p) \geq 15$

  - For mean:
    - Variable is quantitative
    - Data are obtained using randomization (random sample or random assignment)
    - Population distribution of the variable is approximately normal

The five steps of a significance test

- Step 2: Hypothesis
  - Specify the Null and Alternative Hypothesis
  - E.g.,

H_0: $p = 0.5$
H_a: $p > 0.5$ (one sided)
or
H_a: $p < 0.5$ (one sided)
or
H_a: $p \neq 0.5$ (two sided)

The five steps of a significance test

- Step 3: Test statistic

For proportion:

$$z = \frac{\hat{p} - p_0}{se_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- For mean:

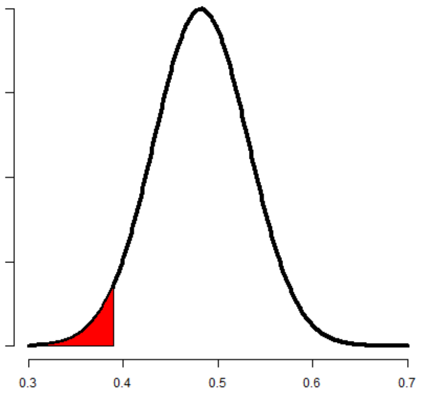
$$t = \frac{\bar{x} - \mu_0}{se} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Degrees of freedom = n - 1

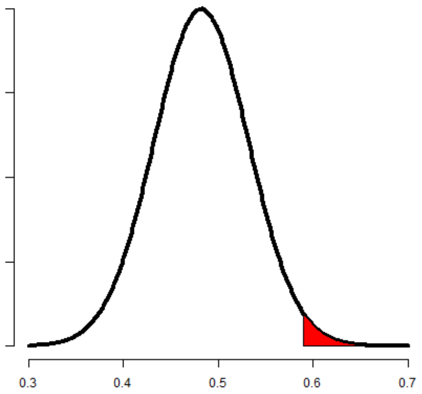
The five steps of a significance test

- Step 4: P-value
  - Calculate the relevant tail probabilities
  - E.g.,

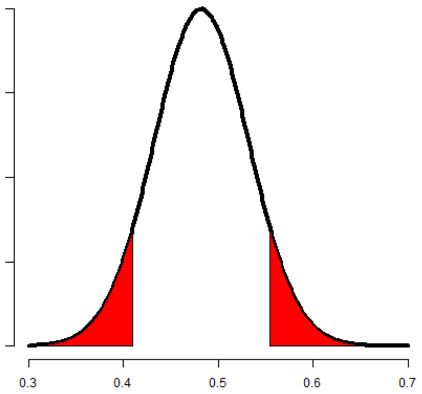
$H_a: p < p_0$



$H_a: p > p_0$



$H_a: p \neq p_0$



The five steps of a significance test

- Step 5: Conclusion

Based on the P-value draw a conclusion about H_0 (reject / not reject)

P-value < level of significance → reject H_0

P-value > level of significance → **not** reject H_0

MAKE YOUR VOICE HEARD!

Questions, suggestions, concerns?
email us:
ocstudfrac-fmg@uva.nl

Want to know more?
Scan me!

Programme Committee

- **students** and **faculty** members work together
- Advise on the **curricula**, **policy** and **teaching & exam regulations**

based on
your feedback!

fill in your
course evaluations!!