university of groningen

university of groningen

Statistiek I

Sampling

Martijn Wieling and John Nerbonne
CLCG, Rijksuniversiteit Groningen

This lecture

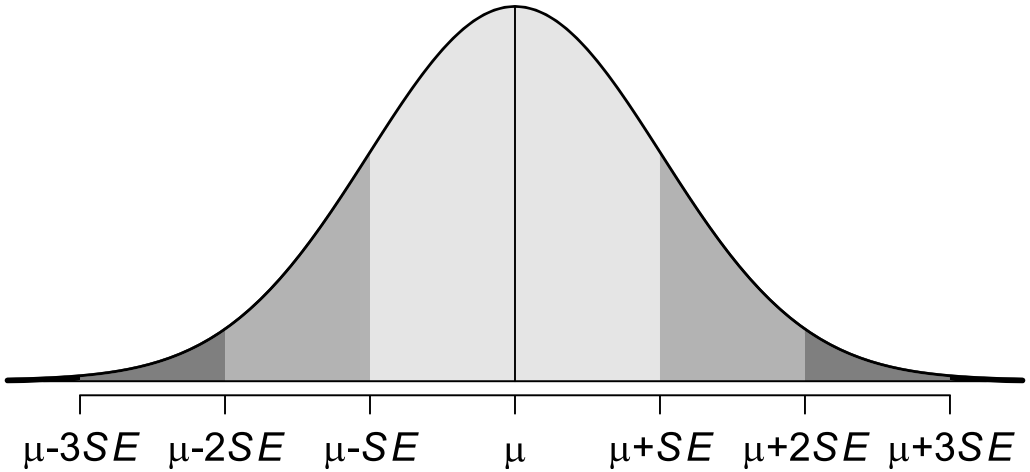
- Reasoning about the population (*populatie*) using a sample (*steekproef*)
  - Relation between population (mean) and sample (mean)
  - Confidence interval (*betrouwbaarheidsinterval*) for population mean based on sample mean
  - Testing a hypothesis (*hypothesetoets*) about the population using a sample
    - One-sided hypothesis vs. two-sided hypothesis
  - Statistical significance
  - Error types

Introduction

- Selecting a sample from a population includes an element of chance: which individuals are studied?
- Question of this lecture: **How to reason about the population using a sample?**
  - Anwered using the **Central Limit Theorem** (*centrale limietstelling*)

Central Limit Theorem

- Suppose we would gather many different samples from the population, then the distribution of the sample means will **always** be normally distributed
  - The means of these samples ($\bar{x}$) will be the population mean ($m_{\bar{x}} = \mu$)
  - The standard deviation of the sample means (standard error *SE*, *standaardfout*) is dependent on the sample size n (*steekproefgrootte*) and the population standard deviation σ (*standaardafwijking*): *SE* $= s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$



Question 1

Wat is de standaardfout van het gemiddelde?

‹ ›

15	1,5	0,15	0,015	?

d81a0057f4f80/f7f178bdfe02)

Votes: 86

Reasoning about the population (1)

- Given that the distribution of sample means is normally distributed $N(\mu, \sigma/\sqrt{n})$, having one randomly selected sample allows us to reason about the population
- Requirement: sample is **representative** (unbiased sample, *zuivere steekproef*)
  - Random selection helps avoid bias

Question 2

Welke willekeurige selectie is een zuivere steekproef om de prestaties bij dit vak te bepalen?

20 studenten aanwezig bij dit college	20 personen op de Vismarkt	20 studenten in de Harmoniekantine	20 studenten ingeschreven voor dit vak	?

ɪ81a0057f4f80/2c1453cb5efb)

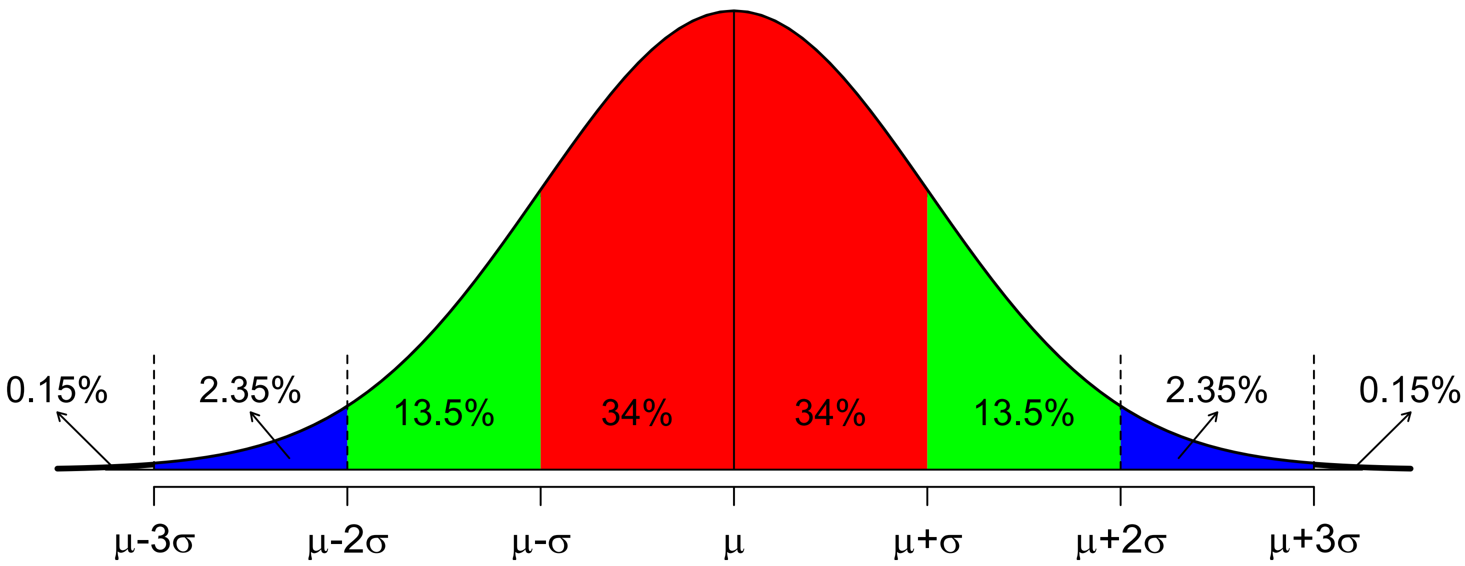
Votes: 109

Reasoning about the population (2)

- Given a representative sample:
  - We estimate the population mean to be equal to the sample mean (our best guess)
  - How certain we are of this estimate depends on the standard error: $\sigma/\sqrt{n}$
    - Increasing sample size n reduces uncertainty when reasoning about the population
      - Hard work pays off (in exactness), but it doesn't pay of quickly: $\sqrt(n)$
    - As sample means are normally distributed (CLT), we use the characteristics of the normal distribution in interpreting the sample means with respect to the population

Normal distribution

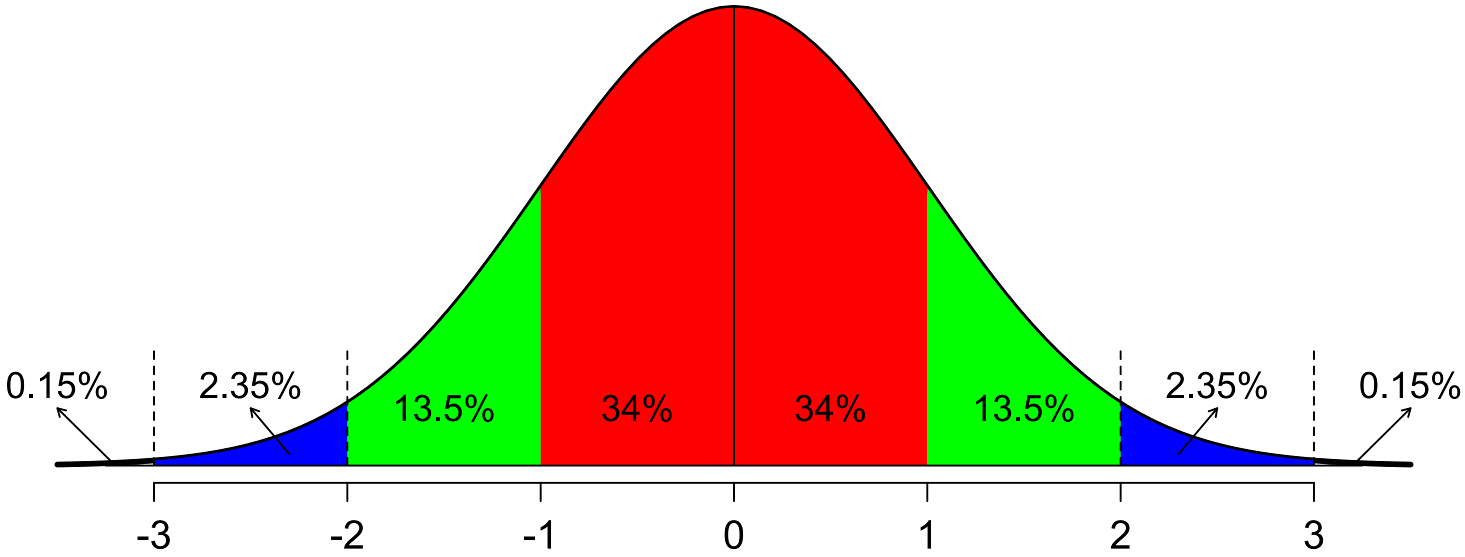
- We know the probability of an element x having a value close to the mean μ:



$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$ (34 + 34)
$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$ (34 + 34 + 13.5 + 13.5)
$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$ (34 + 34 + 13.5 + 13.5 + 2.35 + 2.35)

Normal distribution: standard z-scores

- With standardized values: $z = (x - \mu)/\sigma \Rightarrow \mu = 0$ and $\sigma = 1$



$P(-1 \leq z \leq 1) \approx 68\%$ (34 + 34)
$P(-2 \leq z \leq 2) \approx 95\%$ (34 + 34 + 13.5 + 13.5)
$P(-3 \leq z \leq 3) \approx 99.7\%$ (34 + 34 + 13.5 + 13.5 + 2.35 + 2.35)

Reasoning about the population (3)

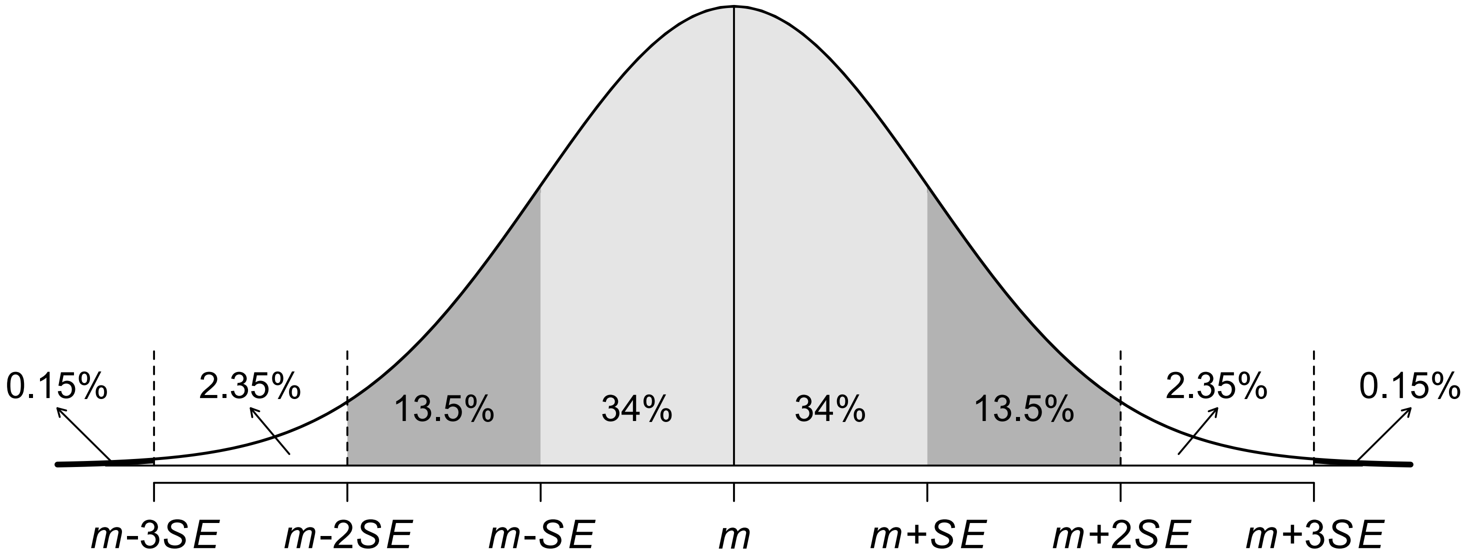
- Sample means can be interpreted in two ways:
  - Using a **confidence interval**
    - An interval which is likely to contain the true population mean
  - Using a **hypothesis test**
    - Tests if a hypothesis about the population is compatible with a sample result

Confidence interval

- **Definition**: there is an x% probability that when computing an x% confidence interval on the basis of a sample, it contains μ
  - The confidence interval gives an estimate of plausible values for the population mean
- Consider the following example:
  You want to know how many hours per week a student of the university spends earning money. The standard deviation σ for the university is 1 hr/wk.
  - You collect data from 100 randomly chosen students
  - You calculate the sample mean $m = 5$ hr/wk
  - You therefore estimate the population mean $\mu = 5$ hr/wk and *SE* $= 1/\sqrt{100} = 0.1$ hr/wk
- What is the 95% confidence interval?

Confidence interval

- According to the CLT, the sample means are normally distributed



- 95% of the sample means lie within $m \pm 2$ *SE*
  - (i.e. actually it is $m \pm 1.96$ *SE*, but we round this to $m \pm 2$ *SE*)
- With m = 5 and *SE* = 0.1, the 95% confidence interval is 5 $\pm$ 2$\times$0.1 = (4.8 hr/wk, 5.2 hr/wk)

Question 3

Wat is het 99.7%-betrouwbaarheidsinterval van het gemiddelde?

‹ ›

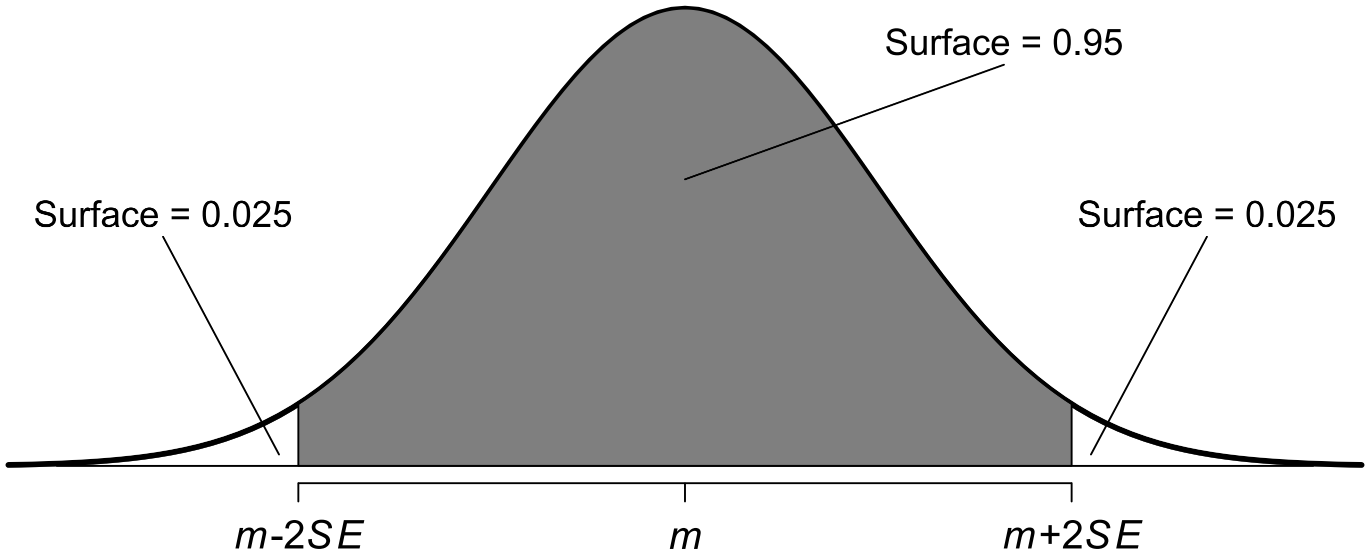
(9,11)	(9,12)	(8,12)	(7,13)	?

ɟ81a0057f4f80/776c6694d464)

Votes: 79

Confidence interval vs. significance test

- The interpretation of a confidence interval is linked to statistical significance
- A 95% confidence interval based on the sample mean m represents the values for μ for which the difference between μ and m is not significant (at the 0.05 significance threshold)
  - A value outside of the confidence interval indicates a statistically significant difference



Hypothesis

- Statistical significance is always assessed in the context of a research question formulated as a hypothesis
- Examples of hypotheses
  - *Answering online lecture questions is related to the course grade*
  - *Women and men differ in their verbal fluency*
  - *Nouns take longer to read than verbs*
- Testing these hypotheses requires **empirical** and **variable** data
  - Empirical: based on observation rather than theory alone
  - Variable: individual cases vary
- Hypotheses can be derived from theory, but also from observations if theory is incomplete

Hypothesis testing (1)

- We start from a research question:
  Is answering online lecture questions related to the course grade?
- Which we then formulate as a hypothesis (i.e. a statement):
  Answering online lecture questions is related to the course grade
- For statistics to be useful, this needs to be translated to a concrete form:
  Students answering online lecture questions score higher than those who do not

Hypothesis testing (2)

Students answering online lecture questions score higher than those who do not

- What is meant by this?
  ***All** students answering online lecture questions score higher than those who do not*?
  - Probably not, the data is variable, there are other factors:
    - Attention level of each student
    - Difficulty of the lecture
    - If the questions were answered seriously
- We need statistics to abstract away from the variability of the observations (i.e. unsystematic variation; Field, Chap. 1)
- ***On average**, students answering online lecture questions score higher than those who do not*

Testing a hypothesis using a sample

On average, students answering online lecture questions score higher than those who do not

- This hypothesis **must** be studied on the basis of a sample, i.e. a limited number of students following a course with online lecture questions
  - Of course we're interested in the population, i.e. all students who followed a course with online lecture questions
- The hypothesis concerns the population, but it is studied through a **representative sample**
  - *Students answering online lecture questions score higher than those who do not* (study based on 20 students who answered online lecture questions and 20 who did not)
  - *Women have higher verbal fluency than men* (study based on 20 men and 20 women)
  - *Nouns take longer to read than verbs* (studied on the basis of 20 people's reading of 20 nouns and verbs)

Question 4

Wat is een goed voorbeeld van een concrete, testbare hypothese?

‹ ›

Zijn vrouwen taalvaardiger dan mannen?	Vrouwen zijn taalvaardiger dan mannen.	Taalvaardigheid is gerelateerd aan geslacht.	?

ɟ81a0057f4f80/9cad7728fc15)

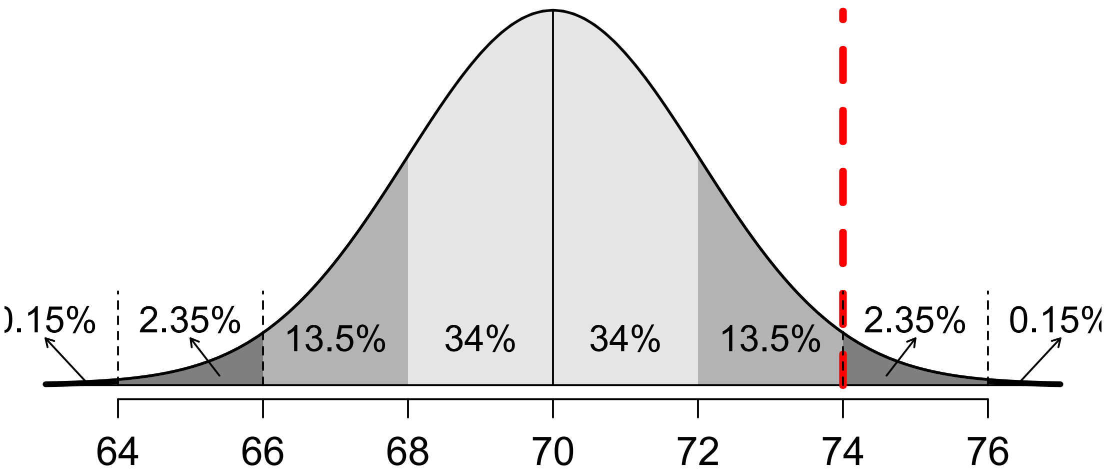
Votes: 113

Analysis: when is a difference real?

- Given a testable hypothesis:
  Students answering online lecture questions score higher than those who do not
  - You collect the final course grade for 20 randomly selected students who answered the online questions and 20 who did not
- Will any difference in average grade (in the right direction) be proof?
  - Probably not: very small differences might be due to **chance** (unsystematic variation)
- Therefore we use **statistics** to analyze the results
  - **Statistically significant** results are those unlikely to be due to chance

Our first analysis: z-test

- You think that Computer Assisted Language Learning may be effective for young kids
- You give a standard test of language proficiency (μ = 70, σ = 14) to 49 randomly chosen childen who followed a CALL program
  - You find m = 74
  - You calculate *SE* = $\sigma/\sqrt{n} = 14/\sqrt{49} = 2$
  - 74 is 2 *SE* above the population mean: at the 97.5th percentile



Conclusions of z-test

- Group with CALL scored 2 *SE* above mean (z-score of 2)
  - Chance of this is only 2.5%, so very unlikely that this is due to chance
- Conclusion: CALL programs are probably helping
  - However, it is also possible that CALL is not helping, but the effect is caused by some other factor
    - Such as the sample including lots of proficient kids
    - This is a **confounding** factor (*verstorende factor*): an influential **hidden** variable (a variable not used in a study)

Question 5

Welke factor(en) kan/kunnen verstorend zijn voor de CALL resultaten?

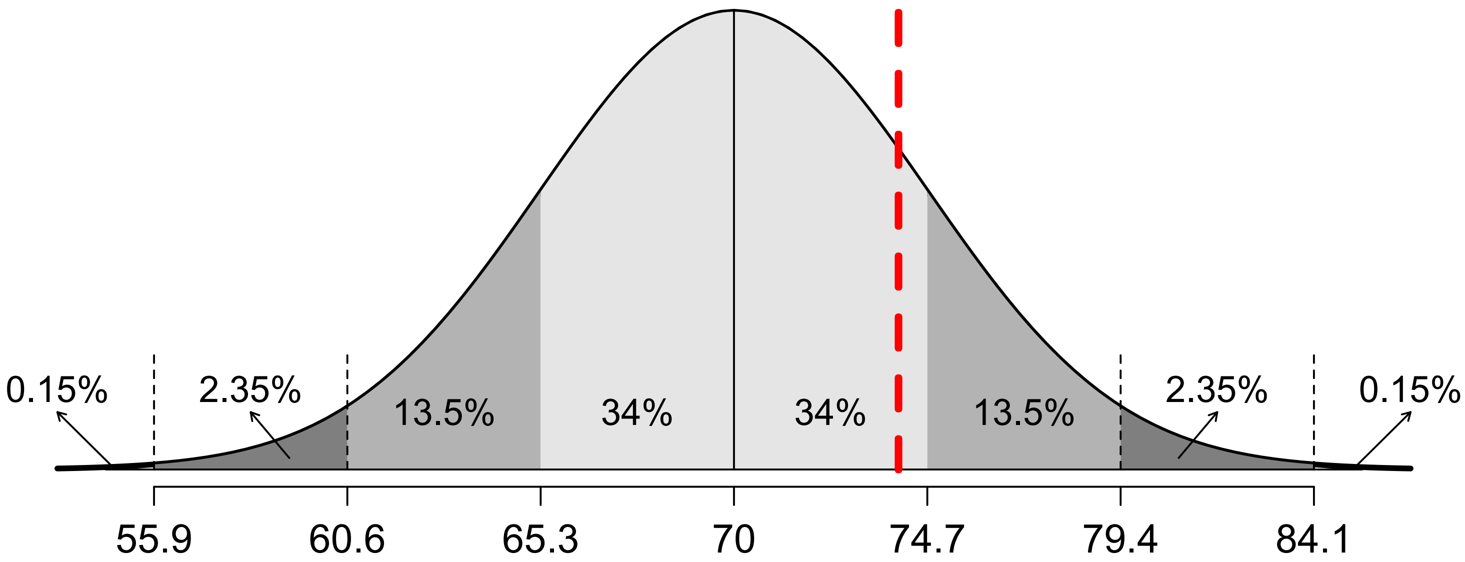
Het opleidingsniveau van de ouders	Het geslacht van de kinderen	Het weer van vandaag	Het schoolniveau van de kinderen	De steekproefgrootte	?

ɹ81a0057f4f80/410f64a5227b)

Votes: 107

Importance of sample size

- Suppose we would have used 9 children as opposed to 49, at what percentile would a sample mean of m = 74 be?
  - $SE = \sigma/\sqrt{n} = 14/\sqrt{9} \approx 4.7$
  - m = 74 is less than 1 *SE* above the mean, i.e. at less than the 84th percentile
    - Sample means of this value are found by chance more than 16% of the time (i.e. likely due to chance): not enough reason to suspect an effect of CALL



Statistical reasoning: two hypotheses

- Rather than one hypothesis, we create **two hypotheses** about the data:
  - The null hypothesis (H_0) and the alternative hypothesis (H_a)
  - The null hypothesis states that there is no relationship between two measured phenomena (e.g., CALL program and test score), while the alternative hypothesis states there is
  - For the CALL example:
    - H_0: $\mu_{CALL} = 70$ (the population mean of people using CALL is 70)
    - H_a: $\mu_{CALL} > 70$ (the population mean of people using CALL is higher than 70)
    - While m = 74, suggests that H_a is right, this might be due to chance, so we would need enough evidence (i.e. low *SE*) to accept it over the null hypothesis
    - Logically, H_0 is the inverse of H_a, and we'd expect H_0: $\mu_{CALL} \leq 70$, but we usually see '$=$' in formulations

Statistical reasoning

H_0: $\mu_{CALL} = 70$ H_a: $\mu_{CALL} > 70$

- The reasoning goes as follows:
  - Suppose H_0 is right, what is the chance p of observing a sample with m = 74?
  - To determine this, we convert 74 to a z-score: $z = (m - \mu)/$*SE* = (74 - 70) / 2 = 2
  - And look up the p-value in a table (or use a stats program): $P(z \geq 2) = 0.025$
    - The chance of observing a sample this extreme given that H_0 is true is 0.025
    - This is the p-value (measured significance level, *overschrijdingskans*)
  - If H_0 were correct and kids with CALL experience had the same language proficiency as others, then the observed sample would be expected only 2.5% of the time
    - Strong evidence **against** the null hypothesis

Statistically significant?

- We have determined H_0, H_a and the p-value
- The classical hypothesis test assesses how **unlikely** a sample must be for a test to count as significant
- We compare the p-value against this threshold significance level or α-level
- If the p-value is **lower** than the α-level (usually 0.05, but it may be lower as well), we regard the result as significant
- In sum:
    - The p-value is the chance of encountering the sample, given that the null hypothesis is true
    - The α-level is the threshold for the p-value below which we regard the result as significant
        - I.e. in that case we reject H_0 and assume H_a is true

Question 6

Wijkt de steekproef significant af van de populatie met alfa=0.05? En alfa=0.01?

‹ ›

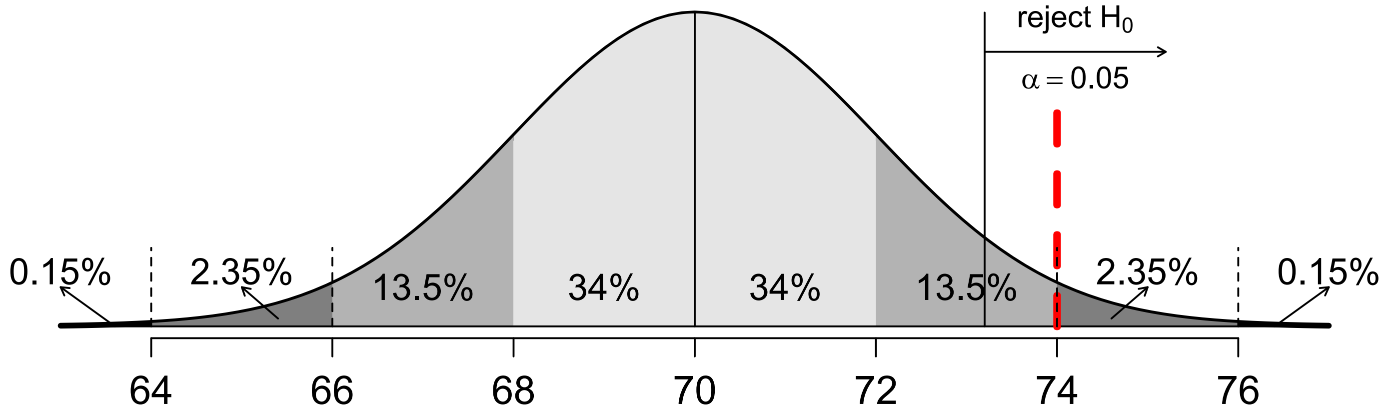
0.05: nee, 0.01: nee	0.05: nee, 0.01: ja	0.05: ja, 0.01: nee	0.05: ja, 0.01: ja	?

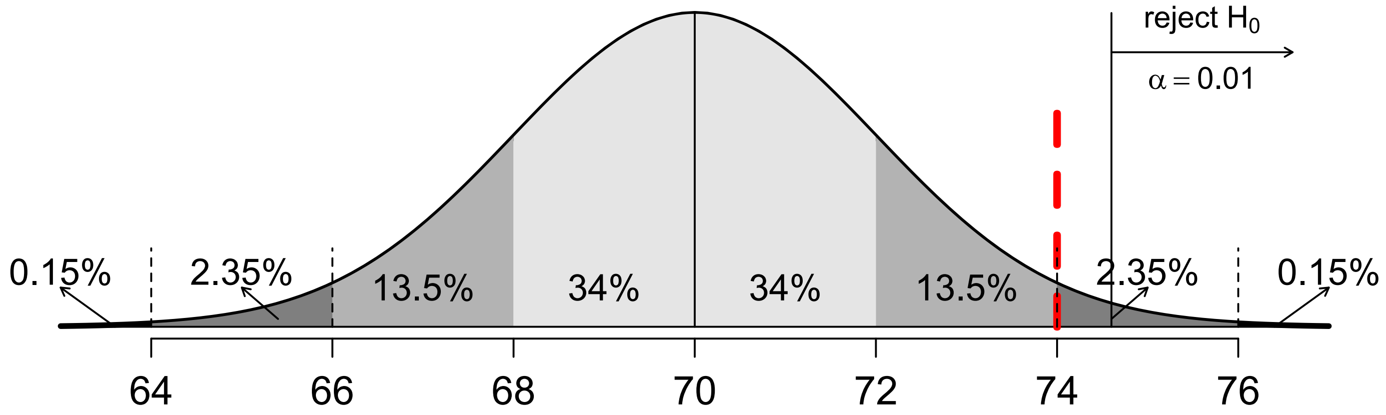
:d81a0057f4f80/1f4c29efb276)

Votes: 87

Visualizing question 6

m = 74 ($z = 2$), μ = 70, σ = 14, $n = 49$, SE = $14/\sqrt{49}$ = 2





Steps for assessing statistical significance

1. Specify H_0 and H_a
2. Specify the distribution of the sample statistic (e.g., mean) given that H_0 is true
3. Specify the α-level at which H_0 will be rejected
4. Determine the value of the statistic (e.g., mean) on the basis of a sample
5. Calculate the p-value using the distribution of the sample statistic and compare to α
   - p-value $\leq \alpha$: reject H_0 (significant result)
   - p-value $> \alpha$: do not reject H_0 (non-significant result)

Critical values

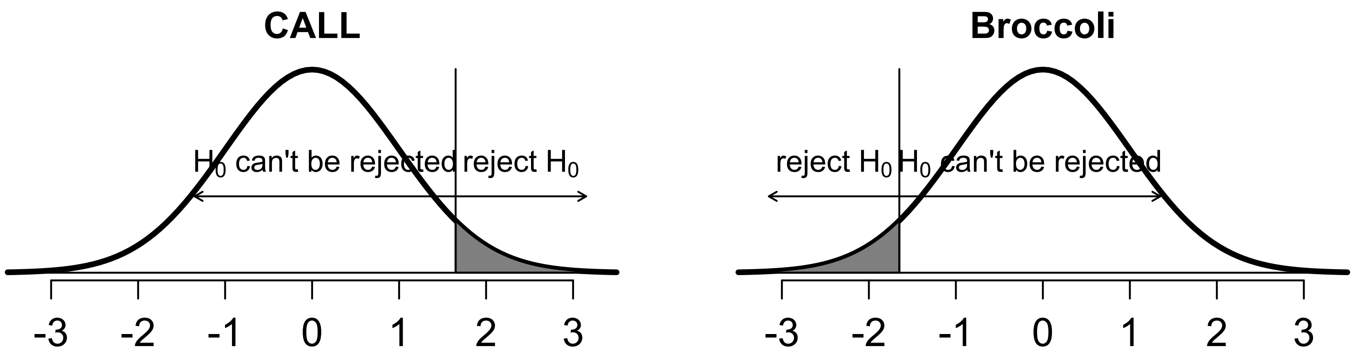
- Critical values: those values of the sample statistic which will result in a rejection of H_0
- E.g., if α is set at 0.05, the critical region is $P(z) \leq 0.05$, i.e. $z \geq 1.65$
- We can transform this to raw values using the z formula

$$z = (x - \mu)/SE$$
$$1.65 = (x - 70)/2$$
$$3.30 = x - 70$$
$$x = 73.3$$

- Thus a sample mean larger than 73.3 will result in rejection of H_0
- These critical values are automatically calculated by statistical software

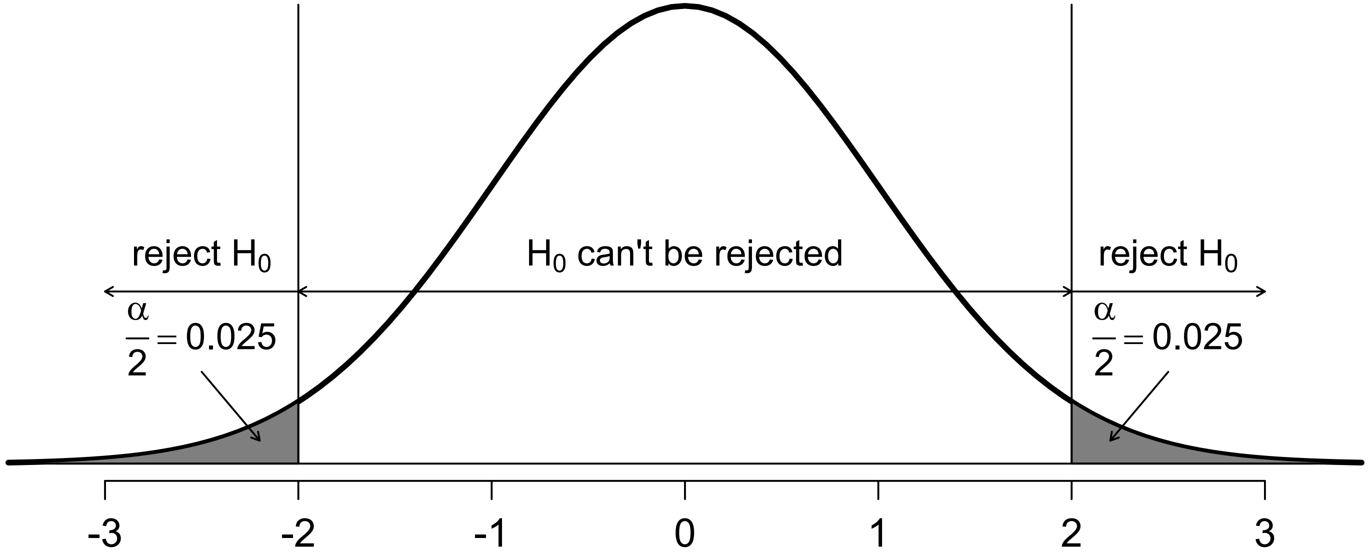
One-sided z-test

- The CALL example is a z-test, as it is based on a normal distribution with known μ and σ
- We calculate the sample mean m and the z value based on it: $z = (m - \mu)/(\sigma/\sqrt{n})$
  - We obtain the p-value linked with the z-value and compare that with the α-level
- There are different forms of z-tests:
  - H_a predicts high m: CALL improves language ability
  - H_a predicts low m: Eating broccoli lowers cholesterol levels



Two-sided z-test

- Sometimes H_a might predict not lower or higher, but just **different**
- For example, you use a statistical test for aphasia in NL developed in the UK
  - The developers claim that for non-aphasics, the distribution is $N(100, 10)$
  - You specify H_0: $\mu = 100$ and H_a: $\mu \neq 100$
  - With a significance level α of 0.05, both very high (2.5% highest) **and** very low (2.5% lowest) values give reason to reject H_0



Significance and sample size

- Recall our CALL example: H_0: $\mu_{CALL} = 70$, H_a: $\mu_{CALL} > 70$
- With a sample of 49, we have distribution $N(70, 14/\sqrt{49})$
- The sample mean m was 74 at a significance level of $p = 0.025$ (i.e. one-tailed)
  - This was significant at the α-level of 0.05, but not 0.01
- If you are certain about m = 74 and wanted significance at the 0.01 α-level, you **could** ask how large the sample would need to be

Chasing significance

- If you are certain about m = 74 and wanted significance at the 0.01 α-level, you **could** ask how large the sample would need to be
- An α-level of 0.01 (one-tailed) corresponds to $z = 2.33$ (from tables)

$$
\begin{aligned}
z &= (x-\mu)/(\sigma/\sqrt{n}) \\
2.33 &= (74-70)/(14/\sqrt{n}) \\
2.33 &= 4/(14/\sqrt{n}) \\
2.33 &= 4\sqrt{n}/14 \\
(2.33 * 14)/4 &= \sqrt{n} \\
8.2^2 &= n \\
n &\approx 67
\end{aligned}
$$

- A sample size of 67 would show significance at the α = 0.01 level, assuming m stays at 74
  - Would it make sense to collect the additional data?

Understanding significance

- Is it sensible to collect the extra data to "push" a result to significance?
  - No. At least, usually not.
- The real result (effect size, *effectgrootte*) is the difference (4 pt.), nearly 0.3σ
- "Statistically significant" implies that an effect probably is not due to chance, but the effect can be **very small**
  - If you want to know whether you should buy CALL software to learn a language, statistically significant does not tell you this
  - This is a two-edged sword, if an effect was not statistically significant, it does not mean nothing important is going on
    - You are just not sure: it could be a chance effect

Question 7

Aan welk signficant resultaat hecht je de meeste waarde?

‹ ›

p = 0.049 met n=100	p = 0.049 met n = 100000	p = 0.005 met n = 100	p = 0.005 met n = 100000	?

d81a0057f4f80/c86ed0529ac2)

Votes: 84

Misuse of significance

- **Garbage in, garbage out**: Statistics won't help an experiment with a poor design, or where data was poorly collected
- **No significance hunting**: Hypotheses should be formulated before data collection and analysis (Field, Ch. 2, "cheating")
  - Modern danger: If there are many potential variables, it is *likely* that a few turn out to be significant
    - Specific tests are necessary to correct for this
    - Exploring the data may be useful in early stages of the experiment, but only before hypothesis testing

Some remarks about hypothesis testing

- A statistical hypothesis concerns a population about which a hypothesis is made involving some statistic
  - Population: all students attending a course using online lecture questions
  - Parameter (statistic): course performance
  - Hypothesis: avg. performance of students answering online lecture questions is higher
- A hypothesis is always about a population, not a sample!
- Sample statistics include:
  - Mean
  - Frequency
  - (etc.)

Identifying hypotheses

- **Alternative hypothesis** H_a (original hypothesis) is contrasted with **null hypothesis** H_0 (hypothesis that nothing out of the ordinary is going on)
  - H_a: average performance of students answering online lecture questions higher
  - H_0: answering online lecture questions does not impact performance
- Logically H_0 should imply $\neg H_a$

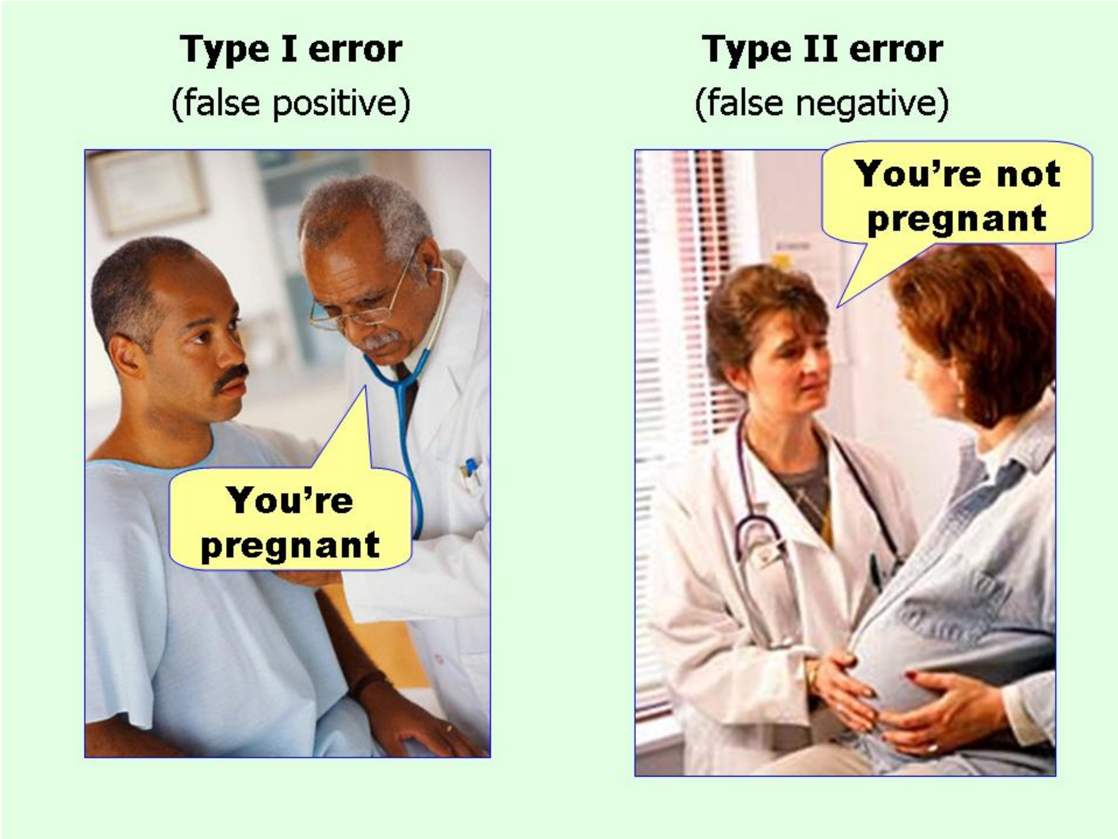
Possible errors

Of course, you could be wrong (e.g., due to an unrepresentative sample)!

H_0	**TRUE**	**FALSE**
accepted	correct	type II error
rejected	type I error	correct

- Hypothesis testing focuses on **type I errors**
  - p-value: chance of type I error
  - α-level: boundary of acceptable level of type I error
- Type II errors
  - β: chance of type II error
  - $1 - \beta$: power of statistical test
    - More sensitive tests have more power to detect an effect and are more useful

Possible errors: easier to remember



- False positive: incorrect positive (accepting H_a) result
- False negative: incorrect negative (not rejecting H_0) result

How to formulate the results?

H_0	TRUE	FALSE
accepted	correct	type II error
rejected	type I error	correct

- Results with $p = 0.06$ are not very different from $p = 0.05$, but we need a boundary
  - An α-level of 0.05 is low as the "burden of proof" is on the alternative
- If $p = 0.06$ we haven't **proven** H_0, only failed to show convincingly that it's wrong
  - This is called "retaining H_0" ("H_0 *handhaven*")

Recap

- In this lecture, we've covered
  - the difference between the population and a sample
  - how to convert a sample statistic (e.g., mean) to a z-score
  - how to calculate a confidence interval
  - how to specify a concrete testable hypothesis based on a research question
  - how to specify the null hypothesis
  - how to determine a representative sample for a given hypothesis
  - how to conduct a z-test and use the results to evaluate a hypothesis
  - what statistical significance entails
  - how to evaluate if a result is statistically signficant given a specific α-level
  - the difference between a one-tailed and a two-tailed test
  - the different error types
- **Experiment yourself**: http://eolomea.let.rug.nl/Statistiek-I/HC2 (login with s-nr)
- Next lecture: t-tests

Please evaluate this lecture

Hoe begrijpelijk vond je dit college?

‹ ›

Ik begreep alles	Ik begreep het meeste	Ik begreep ongeveer de helft	Ik begreep maar een klein deel	Ik begreep helemaal niets

d81a0057f4f80/1628acd281d9)

Votes: 101

Questions?

Thank you for your attention!

http://www.let.rug.nl/nerbonne/teach/Statistiek-I
m.b.wieling@rug.nl