Why be interested in stats?

Vera Demberg

Saarland University

WS 2021/22

Table of Contents

How will the tutorials work?

- You should all be signed up on CMS.
- If you want to watch lectures online (live or later), do sign up for the course on TEAMS.
- You should form groups of 3 for handing in the homework sheets. For finding group members, use the Forum on CMS.
- We did the tutorial assignment yesterday. It's a large class this year, so not everybody could get their top preference (but 80% did). This was the optimal assignment.
- You all should have access to DataCamp and should have completed the assignment for last week (Intro to R; Intermediate R).

Questions

If you have a question

live: raise your hand and ask – I'll try to remember to repeat it so that remote participants know what was asked.

TEAMS: raise your virtual hand, and open your microphone when I call you up.

you can also type to the chat.

Questions are welcome!

Reading

You can read in more detail about the content of the lectures. We are following two books:

- Learning statistics with R: A tutorial for psychology students and other beginners by D. Navarro
- Statistical Methods for Psychology by David C. Howell

Today, we covered:
Navarro: chapters 1 and 2
Howell: chapter 1

Homework and Exam

- Homework: need to make a “serious attempt” at solving all tasks on each sheet.
- If you fail doing this for more than 1 sheet, you will not be allowed to take the exam.
- Copying tutorial sheet answers from previous years or peer groups is plagiarism, and leads to exclusion from the course and exam.
- Exam: Feb 22nd

Why do we use R?

R is a programming language for statistical analysis.

- it's free
- it's powerful
- there is a highly active community helping to develop R further
- there are lots of online resources for how to use R

How to learn R

- I hope that all of you already completed DataCamp's Introduction to R / Intermediate R courses.
- If not, make sure to catch up this week! (reserve enough time for it!)
- Install R on your own machine if you haven't done so yet.

Why have to learn stats?

Statistics is a very important and central part of empirical work.

Is my new

- translation system
- touch screen
- object recognition system
- ...

better than the previous system?

Table of Contents

Why not use common sense?

"You added a new feature to your new user interface. That *clearly* makes it better – why do we need any statistics?"

"Adolescents who use marijuana were found in a study to have lower IQ than their peers, who matched them in IQ at the age of 12 – seems obvious that marijuana causes a drop in IQ?"

"You have alternative methods for Bayesian hyperparameter optimization in machine learning. You compare all of them on a set of different ML problems and observe how long each method takes to converge, and what their best loss is."

Why not use common sense?

- subjective judgments cannot be trusted, we need objective criteria in science

Why not use common sense?

- subjective judgments cannot be trusted, we need objective criteria in science
- even as researchers we have biases:
  - want to believe our new system is better
  - or that our research hypotheses are true.

Why not use common sense?

- subjective judgments cannot be trusted, we need objective criteria in science
- even as researchers we have biases:
  - want to believe our new system is better
  - or that our research hypotheses are true.

Statistics can help to safeguard us from drawing incorrect conclusions.

Simpson's paradox

A real case:
Postgraduate admissions at the University of California, Berkeley in 1973:

	number of applicants	percent admitted
males	8442	44%
females	4321	35%

Simpson's paradox

A real case:
Postgraduate admissions at the University of California, Berkeley in 1973:

	number of applicants	percent admitted
males	8442	44%
females	4321	35%

Claim: females were being disadvantaged (and Berkeley got sued).

Simpson's paradox

A real case:
Postgraduate admissions at the University of California, Berkeley in 1973:

	number of applicants	percent admitted
males	8442	44%
females	4321	35%

Claim: females were being disadvantaged (and Berkeley got sued).

Reality: weak bias in favour of females!

A more detailed view of the data

The main driving factor for admission rates is the subject choice. (Some subjects are more popular; some departments can accept a larger proportion of applicants than others.)

	Male		Female	
Department	Applicants	admitted	Applicants	admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

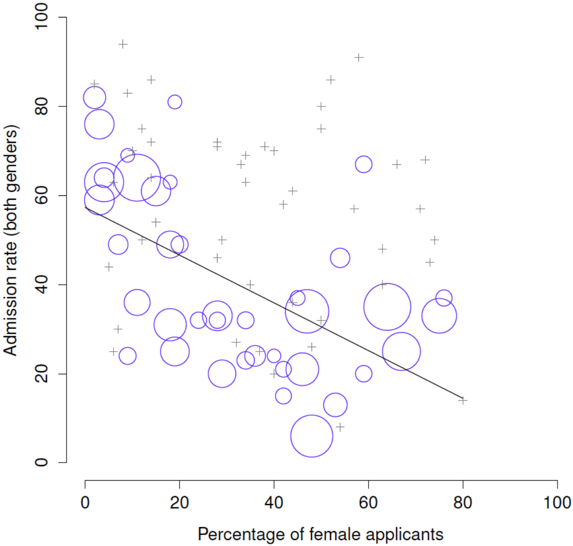
A more detailed view of the data

The main driving factor for admission rates is the subject choice. (Some subjects are more popular; some departments can accept a larger proportion of applicants than others.)

	Male		Female	
Department	Applicants	admitted	Applicants	admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Males applied to the “easier” departments, while females preferred the departments that were more difficult to get into.

Gender and Admission rate were highly correlated



What can we learn from this?

Lesson:
it's important to take into account *all* effects that have an influcence on the outcome, especially if large correlations exist!

What can we learn from this?

Lesson:
it's important to take into account *all* effects that have an influcence on the outcome, especially if large correlations exist!
(we'll get back to this)

Argument #2: We are not objective

Humans are much more likely to believe those things they want to believe.
... or the ones that they think are plausible.

Argument #2: We are not objective

Humans are much more likely to believe those things they want to believe.
... or the ones that they think are plausible.

As a researcher, that's problematic: you want your new system to be better than the old one and will find it logical when it's shows good performance, and you will look for bugs / try to understand in more detail what's going on if it shows bad performance.

Argument #2: We are not objective

Humans are much more likely to believe those things they want to believe.
... or the ones that they think are plausible.

As a researcher, that's problematic: you want your new system to be better than the old one and will find it logical when it's shows good performance, and you will look for bugs / try to understand in more detail what's going on if it shows bad performance.

Let's go through an example together.

Belief bias

Experiment where people were asked whether a conclusion was **logically valid**.

Premise 1	No cigarettes are inexpensive.
Premise 2	Some addictive things are inexpensive.
Conclusion	Therefore, some addictive things are not cigarettes.

(valid argument, believable conclusion)

Belief bias II

Premise 1	No addictive things are inexpensive.
Premise 2	Some cigarettes are inexpensive.
Conclusion	Therefore, some cigarettes are not addictive.

(valid argument, conclusion not believable)

Belief bias II

Premise 1	No addictive things are inexpensive.
Premise 2	Some cigarettes are inexpensive.
Conclusion	Therefore, some addictive things are not cigarettes.

(invalid argument, conclusion believable)

Belief bias IV

Premise 1	No cigarettes are inexpensive.
Premise 2	Some addictive things are inexpensive.
Conclusion	Therefore, some cigarettes are not addictive.

(invalid argument, conclusion not believable)

Human judgments

What we'd hope to find
(assuming people aren't biased by what they think makes sense):

	conclusion feels true	conclusion feels false
argument valid	100% say valid	100% say valid
argument invalid	0 % say valid	0% say valid

Human judgments

What we'd hope to find
(assuming people aren't biased by what they think makes sense):

	conclusion feels true	conclusion feels false
argument valid	100% say valid	100% say valid
argument invalid	0 % say valid	0% say valid

What we actually find when asking people:

	conclusion feels true	conclusion feels false
argument valid	92% say valid	
argument invalid		8% say valid

Human judgments

What we'd hope to find
(assuming people aren't biased by what they think makes sense):

	conclusion feels true	conclusion feels false
argument valid	100% say valid	100% say valid
argument invalid	0 % say valid	0% say valid

What we actually find when asking people:

	conclusion feels true	conclusion feels false
argument valid	92% say valid	46% say valid
argument invalid	92% say valid	8% say valid


WE SHOULD GO TO THE NORTH BEACH. SOMEONE SAID THE SOUTH BEACH HAS A 20% HIGHER RISK OF SHARK ATTACKS.
YEAH, BUT STATISTICALLY, TAKING THREE BEACH TRIPS INSTEAD OF TWO INCREASES OUR ODDS OF GETTING SHOT BY A SWIMMING DOG CARRYING A HANDGUN IN ITS MOUTH BY 50%!
OH NO! THIS IS OUR THIRD TRIP!
REMINDER: A 50% INCREASE IN A TINY RISK IS STILL TINY.

www.xkcd.com

Statistics...

... is a tool to help keep scientists safe from drawing incorrect conclusions.
Statistics is receiving more attention and more weight recently, due to discussions on **reproducible research**.

Table of Contents

Population

The "population" is the set of all datapoints relevant for the research question.

It's usually impractical / impossible to obtain measures for the whole population.

Therefore, we select a *random* and *representative* sample for a study

... and then calculate statistics to see whether the effect in our random sample allows us to draw conclusions about the population.

Logical steps in quantitative research

Step 1: Develop a research question

Logical steps in quantitative research

Step 1: Develop a research question
Step 2: Operationalize the research question

Logical steps in quantitative research

Step 1: Develop a research question
Step 2: Operationalize the research question
Step 3A (in ideal world): Obtain data from the whole population

Logical steps in quantitative research

Step 1: Develop a research question
Step 2: Operationalize the research question
Step 3A (in ideal world): Obtain data from the whole population
Step 3B (in reality): Collect data from a sample (e.g., corpus/experiment)

Logical steps in quantitative research

Step 1: Develop a research question
Step 2: Operationalize the research question
Step 3A (in ideal world): Obtain data from the whole population
Step 3B (in reality): Collect data from a sample (e.g., corpus/experiment)
Step 4: Describe the data: descriptive statistics

Logical steps in quantitative research

Step 1: Develop a research question
Step 2: Operationalize the research question
Step 3A (in ideal world): Obtain data from the whole population
Step 3B (in reality): Collect data from a sample (e.g., corpus/experiment)
Step 4: Describe the data: descriptive statistics
Step 5 (in ideal world, could skip this step): Make inferences

- (= inferential statistics)
- use probability theory to make an educated guess to what extent statistics from the sample applies to the population

Logical steps in quantitative research

Step 1: Develop a research question
Step 2: Operationalize the research question
Step 3A (in ideal world): Obtain data from the whole population
Step 3B (in reality): Collect data from a sample (e.g., corpus/experiment)
Step 4: Describe the data: descriptive statistics
Step 5 (in ideal world, could skip this step): Make inferences

- (= inferential statistics)
- use probability theory to make an educated guess to what extent statistics from the sample applies to the population

Step 6: Draw a conclusion

Try it out!

Please answer the following questions (take notes for yourself):

- Under what conditions would the entire student body of Saarland university be considered a "population"?
- Why would choosing names from the local phone book not produce a random sample of the residents of Saarbrücken? Who would be underrepresented and who would be overrepresented?
- Under what conditions would the entire student body of Saarland university be considered a "sample"?
- Would it be a random or nonrandom sample? Why?

Table of Contents

Terminology

Information about our sample:

we can distinguish between

- a theoretical construct (e.g., age, opinion on Trump as president)
- a measure (e.g., looking up birth certificate, asking in questionnaire, counting positive vs. negative posts on facebook, subjective ratings by participants, levels of stress, ...)
- an operationalization (is the logical connection between measure and theoretical construct, i.e. how do we derive a measure from the theoretical construct)
- a variable (the collection of actual ages / opinions in our dataset)

A variable

What can a variable look like?

- nominal scale (= categorial variable)
  no particular relationship between different values:
  e.g.: eye colour, gender, modes of transportation
  $\rightarrow$ cannot be ordered, no average

A variable

What can a variable look like?

- nominal scale (= categorial variable)
  no particular relationship between different values:
  e.g.: eye colour, gender, modes of transportation
  → cannot be ordered, no average

- ordinal scale (meaningful order exists, but not equidistant steps)
  e.g.: first, second, third prize in race, but don't know how many sec faster
  Opinions:
  1. Temperatures are rising, because of human activity
  2. Temperatures are rising, but we don't know why
  3. Temperatures are rising, but not because of humans
  4. Temperatures are not rising

  → can be ordered, but not averaged

Variables (2)

- interval scale (numeric values genuinely meaningful, but no natural zero)
    - e.g.: temperature: 3 degrees difference between 7 and 10 is same as between 30 and 33 degrees celsius; but 20 degrees is not twice as hot as 10 degrees.
    - e.g.: starting a degree in 2016 is not 1.0075 times later than starting a degree in 2001

  $\rightarrow$ no natural zero; can't divide/multiply but can add and subtract

Variables (2)

- interval scale (numeric values genuinely meaningful, but no natural zero)
  - e.g.: temperature: 3 degrees difference between 7 and 10 is same as between 30 and 33 degrees celsius; but 20 degrees is not twice as hot as 10 degrees.
  - e.g.: starting a degree in 2016 is not 1.0075 times later than starting a degree in 2001

  $\rightarrow$ no natural zero; can't divide/multiply but can add and subtract

- ratio scale (numeric values meaningful, natural zero)
  e.g.: reaction times; occurrences of two words together

  $\rightarrow$ zero really means zero; can add, subtract, divide and multiply

continuous vs. discrete

- **A continuous variable** is one in which, for any two values that you can think of, it's always logically possible to have another value in between.
- **A discrete variable** when it's sometimes the case that there can't be anything in the middle between two values.

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

Try it out! https://forms.office.com/r/RWuF33w5CA

Are the following measures
nominal / ordinal / interval / ratio?
and are they continuous / discrete?

- Life stress (number of change events in past 6 months, such as moving, new job, new partner)
- Grades in exam (German scale)
- Reaction times for pressing the button when red light comes on
- Type of accommodation students live in
- Navy ranks (commander, captain, admiral)

Try it out! https://forms.office.com/r/RWuF33w5CA

Are the following measures
nominal / ordinal / interval / ratio?
and are they continuous / discrete?

- Life stress (number of change events in past 6 months, such as moving, new job, new partner) → ordinal because a count of 6 does not necessarily mean twice as much stress as a count of 3, and difference in life stress between person with 4 vs 10 events might not be the same as between person with 10 vs. 16 events
- Grades in exam (German scale)
- Reaction times for pressing the button when red light comes on
- Type of accommodation students live in
- Navy ranks (commander, captain, admiral)

Try it out! https://forms.office.com/r/RWuF33w5CA

Are the following measures
nominal / ordinal / interval / ratio?
and are they continuous / discrete?

- Life stress (number of change events in past 6 months, such as moving, new job, new partner) → ordinal because a count of 6 does not necessarily mean twice as much stress as a count of 3, and difference in life stress between person with 4 vs 10 events might not be the same as between person with 10 vs. 16 events
- Grades in exam (German scale) → interval
- Reaction times for pressing the button when red light comes on

- Type of accommodation students live in
- Navy ranks (commander, captain, admiral)

Try it out! https://forms.office.com/r/RWuF33w5CA

Are the following measures
nominal / ordinal / interval / ratio?
and are they continuous / discrete?

- Life stress (number of change events in past 6 months, such as moving, new job, new partner) → ordinal because a count of 6 does not necessarily mean twice as much stress as a count of 3, and difference in life stress between person with 4 vs 10 events might not be the same as between person with 10 vs. 16 events
- Grades in exam (German scale) → interval
- Reaction times for pressing the button when red light comes on → ratio
- Type of accommodation students live in
- Navy ranks (commander, captain, admiral)

Try it out! https://forms.office.com/r/RWuF33w5CA

Are the following measures
nominal / ordinal / interval / ratio?
and are they continuous / discrete?

- Life stress (number of change events in past 6 months, such as moving, new job, new partner) → ordinal because a count of 6 does not necessarily mean twice as much stress as a count of 3, and difference in life stress between person with 4 vs 10 events might not be the same as between person with 10 vs. 16 events
- Grades in exam (German scale) → interval
- Reaction times for pressing the button when red light comes on → ratio
- Type of accommodation students live in → nominal
- Navy ranks (commander, captain, admiral)

Try it out! https://forms.office.com/r/RWuF33w5CA

Are the following measures
nominal / ordinal / interval / ratio?
and are they continuous / discrete?

- Life stress (number of change events in past 6 months, such as moving, new job, new partner) → ordinal because a count of 6 does not necessarily mean twice as much stress as a count of 3, and difference in life stress between person with 4 vs 10 events might not be the same as between person with 10 vs. 16 events
- Grades in exam (German scale) → interval
- Reaction times for pressing the button when red light comes on → ratio
- Type of accommodation students live in → nominal
- Navy ranks (commander, captain, admiral) → ordinal

Likert scale

- Likert scale
  (e.g.: how much do you agree that all pirates are awesome)
  1. Strongly disagree
  2. Disagree
  3. Neither agree nor disagree
  4. Agree
  5. Strongly agree

Likert scale

- Likert scale
  (e.g.: how much do you agree that all pirates are awesome)
  1. Strongly disagree
  2. Disagree
  3. Neither agree nor disagree
  4. Agree
  5. Strongly agree

→ Likert is a difficult one: quasi-interval

Roles of variables

a variable could be

- the target we're interested in
  (e.g., how much people like the new interface)
  this is the variable that should be "explained"
  aka "dependent variable"
  aka "response variable"

- other relevant information we know (e.g., which version of the system did they use; how old those people were that did the ratings)
  these variables should do the "explaining"
  aka "independent variable"
  aka "predictor"

Please let me know how much you know already about hypothesis testing. (Anonymous questionnaire)

https://forms.office.com/r/hT4c3qKYmg