Overview of Client Projects

Statistical Consulting 2024-2025 – Semester 1 –

Leiden University

Version 3, Aug 27, 2024

Project 1

Name researcher: Cor de Kroon
Department: Gynaecological oncology
Faculty or Company: LUMC

What is topic of your research project / what is the research question? (max 5 lines)

In 2019 HIPEC was approved as part of the standard treatment of patients with FIGO III ovarian cancer. Until then treatment for patients with stage III and IV did not differ and consequently not much effort was put in differentiating stage III from stage IV patients.

What statistical question(s) do you want to pose to the students?

- Did the introduction of HIPEC improve survival in patients with stage III ovarian cancer?
- Did the introduction of HIPEC increase the (relative) number of patients with FIGO IV?
- Did the introduction of HIPEC improver survival in patients with ovarian cancer in general?

IKNL data of all patients with ovarian cancer diagnosed 1-4-2015 until 1-4-2023. 1000 patients, 25 variables.

Please shortly describe the data source(s) available (number of subjects, number of variables):

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Mixed models, survival.

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 2

Name researcher: Vleggeert-Lankamp
Department: Neurosurgery
Faculty or Company: Medicine

What is topic of your research project / what is the research question? (max 5 lines)

Patients that are operated on the lumbar spine for lumbar stenosis are not always satisfied with the outcome (about 2/3 is happy). We have evaluated that this is associated with a high anxiety and depression score. We would like to make a prediction model for this as we did previously in collaboration with statistics on the cervical spine

What statistical question(s) do you want to pose to the students?

Can you make a prediction model for outcome in lumbar spinal stenosis surgery. (input RDQ (functionality scale for lumbar spine) and VAS legpain and SF36 physical (anxiety and depression), output RDQ and VAS leg pain at end of follow up.

Please shortly describe the data source(s) available (number of subjects, number of variables):

160 patients. Of all patients we have baseline and outcome data

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Predictive Mixed Effects Model (predictive LMM)).

In what language is your dataset annotated (variable names and labels)?

The dataset has Dutch annotation only, this one

Project 3

Name researcher: Lodewijk Pet
Department: Clinical Epidemiology
Faculty or Company: LUMC

What is topic of your research project / what is the research question? (max 5 lines)

We are evaluating Good Research Practices (GRPs) stated in “The Netherlands Code of Conduct for Research Integrity” and “The European Code of Conduct for Research Integrity – ALLEA”. We are interested how these standards are perceived and which need to be adjusted or removed.

What statistical question(s) do you want to pose to the students?

1. What is the best way to rank these practices when using Likert scales (1 to 5)?
2. What is a better way to answer the research question instead of using Likert scales?
3. How can we compare similar GRP’s with each other? How do we take into account the fact that each individual GRP is part of a bigger concept (the entire code)?
4. How do we take into account that people change the way the answer questions during a (long) questionnaire?

Please shortly describe the data source(s) available (number of subjects, number of variables):

Questionnaire among PhD candidates who do a mandatory course.
About 300 participants
63 standards times 4 Likers scale questions per standards = 252 + 9 baseline variables

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

1. Optimal scaling
2. Tobit regression
3. Other suggestions are welcome

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 4

Name researcher: Joosje Baltussen
Department: Medical Oncology
Faculty or Company: LUMC

What is topic of your research project / what is the research question? (max 5 lines)

Despite the fact that a large proportion of patients with colorectal cancer seen in daily practice are older and live with frailty, studies investigated real world data on long-term QoL and physical functioning in this population are scarce. Therefore, we want to investigate QoL and physical functioning over time and study its association with frailty in older patients with metastatic colorectal cancer receiving anticancer treatment.

What statistical question(s) do you want to pose to the students?

We would like to give an estimate of the QoL and physical functioning over time in older patients with metastatic colorectal cancer.
Which methods are appropriate to analyse repeated measures of QoL and physical functioning?

A significant proportion of patients died over time, and of these patients we do not have information about QoL/physical functioning at several timepoints: how to take into account mortality in the longitudinal analysis?
Could a composite endpoint, taking into account mortality or a decline in QoL, be an option?
What other methods are available that consider mortality in the analysis of longitudinal PROs?
What are the advantages and disadvantages of using linear mixed models for this analysis?

Please shortly describe the data source(s) available (number of subjects, number of variables):

We use data from the ongoing, population-based Prospective Dutch Colorectal Cancer (PLCRC) cohort. This nationwide observational cohort collects longitudinal patient-reported outcomes (PROs) of patients with colorectal cancer. Dutch patients with histologically proven stage I to IV colorectal cancer can be included in the PLCRC. Participants receive various validated questionnaires to measure PROs at time of inclusion and 3, 6, 12, 18 and 24 months after inclusion. Physical functioning was assessed with the European Organization for Research and Treatment of Cancer (EORTC) QoL Questionnaire-C30 (QLQ-C30). QoL was assessed by using the EQ-5D and EQ-VAS.
For the present study, we will include 170 patients aged ≥70 years who were diagnosed with metastatic colorectal cancer

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

No: the students can choose which statistical method is most appropriate.

In what language is your dataset annotated (variable names and labels)?

The dataset has Dutch annotation only

Project 5

Name researcher: Danielle Toen (PI Louise van der Weerd)
Department: Radiology
Faculty or Company: Faculty of Medicine, Leiden University Medical Center

What is topic of your research project / what is the research question? (max 5 lines)

The focus of my research project is on iron accumulation and iron spreading in patients with Alzheimer's disease. We do not yet know where in the brain accumulation starts, and how the accumulation spreads. In this project we will develop an atlas of the brain that shows how iron accumulates as Alzheimer's disease progresses over time and compare this with the Braak scoring system for Amyloid and Tau. We do that by studying different brain regions of hundreds of brain donors with varying degrees of Alzheimer's disease.

What statistical question(s) do you want to pose to the students?

What types of statistical techniques can I apply on my data?
How do I deal with within-subject versus between-subject comparisons?
How do I deal with multiple comparisons?
What would be my statistical analysis plan?
How can I visualize my data like the iron spreading pattern?

Please shortly describe the data source(s) available (number of subjects, number of variables):

We have post-mortem tissue of 9 different brain regions from 40 subjects. On this tissue, histological stainings are performed and those stainings are scored. The number of variables in the dataset is 26 (e.g iron score per region, braak score, age at time of death, education years)

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Not applicable

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 6

Name researcher: R Berendsen
Department: Anesthesiology
Faculty or Company: LLUMC

What is topic of your research project / what is the research question? (max 5 lines)

The project's main objective is to establish a baseline inventory of the state of knowledge about altitude-related illnesses among high-altitude travelers (laypersons) using the international STrengthening Altitude Knowledge (STAK) consensus as the yardstick.

What statistical question(s) do you want to pose to the students?

At least we need to do a Psychometric test analysis on the test. Maybe a multivariate analysis: Principal component analysis (PCA) or factor analysis to identify underlying factors that explain the variance in the data, especially health variables and experience-related factors. Maybe Logistic Regression. The logistic regression can be used to identify significant predictors for achieving a pass STAK achievement score.

Please shortly describe the data source(s) available (number of subjects, number of variables):

Data is a questionnaire with 28 test questions, general question like age, profession, educational level, physical fitness, mental fitness and alpine experience questions.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation (also Dutch, French, German and Italian)

Project 7

Name researcher: H. Blom/J. Schenck/S. Hombergen/M. Munting
Department: otorhinolaryngology/ENT (KNO), physiotherapy
Faculty or Company: HagaHospital The Hague

What is topic of your research project / what is the research question? (max 5 lines)

Patients with Ménière's disease suffer from attacks of severe vertigo. This leads to hypofunction of the inner ear. Vestibular training with a physiotherapist may help patients to reverse the hypofunction. We aim to assess the outcomes of patients who underwent this training program, and study if there is a cut-off point (number of attacks) for when patients are 'trainable'.

What statistical question(s) do you want to pose to the students?

- What statistical approach should be used (eg, plan for statistical analysis)
- Is there a cut-off point in numbers of attacks, above which training is not useful?

Please shortly describe the data source(s) available (number of subjects, number of variables):

+-75 patients, per patients number of attacks in 1 year of follow-up, +-10 physiotherapeutical outcomes

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Not known yet

In what language is your dataset annotated (variable names and labels)?

The dataset has Dutch annotation only

Project 8

Name researcher: Kiki Spoelstra
Department: Cognitive Psychology (CoPAN lab)
Faculty or Company: Faculty of Social and Behavioural Sciences (FSW)

What is topic of your research project / what is the research question? (max 5 lines)

Topic: Attention towards emotionally valanced social stimuli across modalities in domestic cats
Research questions:

- Do cats respond differently to visual and auditory social stimuli of conspecifics?
- Do cats respond differently to positively and negatively valenced stimuli?

What statistical question(s) do you want to pose to the students?

- Is the statistical plan that I made adequate for answering my research questions? Improvements are welcome.
- (How) can I use (G)LMM to analyse the effect that various variables (e.g. sex, age, fearfulness) have on attention and behavioural response?

Please shortly describe the data source(s) available (number of subjects, number of variables):

- 24 cats (12 male, 12 female)
- Eight trials per cat (two pos. and two neg. categories for both visual and auditory stimuli)
- Many different variables: nine “continuous” variables (duration/frequency), over 50 behavioural variables (probably 0-1 occurrence)

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Preferably mixed models, but I have no experience and do not know if this is doable with my data.

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 9

Name researcher: Lucy Sinke
Department: CDS, LACDR
Faculty or Company: Science

What is topic of your research project / what is the research question? (max 5 lines)

In this project, we profiled gene expression in three *in vitro* liver test systems (HepG2, iPSC-HLC, and PHH) both with and without chemical exposure. In order to properly assess transcriptomic responses to these compounds, we first must characterize the baseline variability of different genes in the various test systems.

What statistical question(s) do you want to pose to the students?

What statistical methods and metrics best characterize the amount and sources of variability in baseline TempO-Seq data? How can we adjust for this baseline variability in the following analysis?

Please shortly describe the data source(s) available (number of subjects, number of variables):

PHH data: ~500 samples, ~3000 genes
HLC data: ~50 samples, ~3000 genes
HepG2 data: 125 samples, ~12000 genes

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Unsure as of yet.

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 10

Name researcher: Antoinette van Laarhoven
Department: Health, Medical and Neuropsychology
Faculty or Company: FSW, Leiden University

What is topic of your research project / what is the research question? (max 5 lines)

We investigate how patients with chronic itch and patients with chronic pain differ from healthy individuals in their itch-and pain-related perception of standardized stimuli. When applying standardized stimuli to the participants' skin, we would like to investigate whether patients with chronic itch predominantly use itch- and itch-related descriptors (e.g., itchy, tickling, prickling, …), whereas patients with chronic pain predominantly use pain- and pain-related descriptors (e.g., painful, burning, pricking) when compared to healthy individuals to describe the perceived sensations. Secondly, participants also rated their expected sensations using the same descript-tors and we are interested to see if the expectancy determined the perception. Participants also indicate the extent to which they consider each of the descriptors itch- and pain-related (irrespective of the stimuli).

What statistical question(s) do you want to pose to the students?

The main challenge of this project is that we have, potentially, 47 descriptors per stimulus, which are individually scored on itch- and pain-relatedness. We would like to hear input on how to
1) combine the scores per descriptor cluster (itch-, pain-related) to investigate whether the patients with chronic itch and pain score higher on itch- and pain-related descriptors in response to the stimuli, respectively, compared to the other two groups; 2) investigate the extent to which expectancies determine the perceptions, and how this differs per group; 3) compare the three groups on their level of differentiation in perception of the four stimuli (i.e. the number of different descriptors within the non-itch/pain-related cluster used, as opposed to itch- and pain-related descriptors); 4) irrespectively of the stimuli, investigate whether patients differ from the healthy controls in how they cluster the descriptors based on itch/pain applicability.

Please shortly describe the data source(s) available (number of subjects, number of variables):

We aim to obtain data of 60 participants (20 per group) and have currently 24 participants tested. From 4 different stimuli (with in total 13 assessment points) we ask participants the following: a) before every stimulus, for all 47 descriptors, participants report whether they expect this sensation to be perceived or not, and if so, to which extent (on a numerical rating scale with a maximum of 10); b) After each stimulus, they rate the intensity of each of the perceived descriptor(s) on a similar numerical rating scale.
We also have ratings on the extent to which they find each of the 47 descriptors itch- and pain-related (irrespective of the stimuli).

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

For question 1) we consider calculating a weighted average per individual per cluster of descriptors and compare these across the groups using mixed models, also including the 4 stimuli with multiple measurements. For 2) we were thinking of network analyses including the descriptors for both the "expected sensations" and "perceived sensations" (separately if sufficiently powered or clustered into itch-related, pain-related and other) as well as "group" as nodes. For 3) the number of used descriptors within each cluster could be calculated per participant, and using mixed models including the 4 stimuli consisting of multiple measurements, we could see if the outcome is predictive for the group the participants belong to. For 4) cluster analyses may be considered to explore how the descriptors map on itch-relatedness versus pain-relatedness and if this differs per group.

In what language is your dataset annotated (variable names and labels)?

The dataset has Dutch annotation only (note, the descriptors are in Dutch)

Project 11

Name researcher: Ep Heuvelink
Department: Horticulture and Product Physiology
Faculty or Company: Wageningen University

What is topic of your research project / what is the research question? (max 5 lines)

My research question is, how the different levels of additional Far Red light influence fruit fresh yield in a greenhouse cultivation of sweet pepper.

What statistical question(s) do you want to pose to the students?

In this research, we used split-plot design with 4 light levels as whole-plots, and 2 cultivars as sub-plots. Each combination of light treatments and cultivars has 2 replicates (2 blocks in total). The light level is quantified for each experimental unit, as cultivar also influences perceived light level due to different plant height, where variations was also found between blocks.
The question is, how to find out the quantitative relationship between light quantity and the interested variables? In other words, how do we apply regression to a split-plot design, and take block effect into account?

Please shortly describe the data source(s) available (number of subjects, number of variables):

There are 16 experimental units: 4 light levels *2 cultivars*2 replicates
There are 2 variables: 1) total fresh weight of ripe fruit, 2) total number of ripe fruits during the cultivation

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Perhaps mixed models ????

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 12

Name researcher: Nadia de Gruil
Department: Medical Oncology
Faculty or Company: LUMC

What is topic of your research project / what is the research question? (max 5 lines)

Pilot study with healthy individuals (n=9) who adhered to a fasting mimicking diet (FMD) twice in 1 month. FMD is a low-caloric diet of 4 days followed by about 28 days without dietary restrictions and then again 4 days of FMD. Blood draws were taken at baseline, day 2 and last day (~day 32). The aim is to investigate immunology changes in the lymphocytes from the blood draws using RNA seq (~700 genes), metabolic markers (4) and flow cytometry panel.

What statistical question(s) do you want to pose to the students?

6 of 9 participants have complete dataset (all 3 timepoints).
Is there an approach to analyze all 9 participants from 3 timepoints, even though 3 of them have missing data?

Please shortly describe the data source(s) available (number of subjects, number of variables):

Subjects: 9 participants, 3 timepoints with missing data
Variables: data for analysis is the RNA seq output for ~700 genes measured in log2 counts and 4 other metabolic measurements: ketone, insulin, glucose and IGF-1 level.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 13

Name researcher: Anne van Diepeningen
Department: BU Biointeractions and Plant Heath
Faculty or Company: Wageningen University and Research

What is topic of your research project / what is the research question? (max 5 lines)

The Fungus Verticillium causes wilting diseases in many crops. In Chrysanthemum resistance breeding against Verticillium is an important way to deal with this disease in future, instead of using fungicides. Currently breeding companies have their own way of doing tests and scoring ‘resistance’, or ‘susceptibility’. As bioassays are relatively expensive, they like to minimize tests. But how many plants should we test for a reliable result?

What statistical question(s) do you want to pose to the students?

- What are the best parameters to measure plant susceptibility/resistance
- How many plants should be tested for a test (big debate with breeders!)
- What are the best ranges to qualify ‘resistance/tolerance’ or ‘susceptibility’ and can we identify intermediate classes?

Please shortly describe the data source(s) available (number of subjects, number of variables):

We have done plant tests (bioassays) with different plant cultivars with a Verticillium pathogen under one standardized condition, once in winter and once in spring/summer. We scored different standard parameters (e.g. # affected plants; levels of tissue damage, but also made additional observations).

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 14

Name researcher: Prof. Dr. Matthijs Vos
Department: Theoretical and Applied Biodiversity Research
Faculty or Company: Biology, Ruhr University Bochum

What is topic of your research project / what is the research question? (max 5 lines)

We are interested in what happens to ecological communities (of f.e. animal species) when these are exposed to severe or multiple stressors over time. Do they recover? And how? (fast, slow, incomplete or complete recovery, or shifts to an alternative state instead of recovery? Do different subgroups in the community show different responses? Are there asymmetries in the response to stress and the response to stress-release?

What statistical question(s) do you want to pose to the students?

In a 42-year time series for an animal community living in a river, several disturbances happened. One of them was an insecticide spill that caused mass mortality. We would like to explain before-after differences in population densities of different species (parts of the community) as part of recovery or shift to a new state, with the difficulty that a second stressor hit the system only a few years after the first stressor, and given different species tolerances

Please shortly describe the data source(s) available (number of subjects, number of variables):

42 years of density data for ca 90 animal species in 1 river.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Appropriate tests to compare for example 10 years of density data for each of the 90 species before the insecticide event and the 10 years of data after the event; then the same for the stressor that followed within a few years, and then disentangle the effects of these subsequent events. (Some species respond to stressor 1, some to stressor 2, and some to both).

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 15

Name researcher: Group of Erik Danen (Klara Beslmuller, Michiel van Dijk,
Sabine de Winter, Erik Danen)
Department: CDS
Faculty or Company: LACDR

What is topic of your research project / what is the research question? (max 5 lines)

Cancer tumoroids migrate in the ECM. The stiffness of the ECM can greatly affect tumor metastasis. Reports have shown an increased metastasis in a stiffer microenvironment. The aim of this research is defining which stiffness in the ECM is affecting tumoroid invasion.

What statistical question(s) do you want to pose to the students?

We successfully quantified the invasion of tumoroids in different conditions. During our experiments, we quantify the Area of tumoroids for 3 independent experiments. Each independent experiment contains several technical replicates, mostly between 4 and 8. These technical replicates are different to the biological replicates (independent experiments), as the technical replicates are performed on the same day, with exactly the same cancer cells, in exactly the same collagen ECM. Technical replicates are just the different tumoroids. We also have different kinds of technical replicates: replicates in the same well (e.g. multiple image fields, multiple tumoroids per well) and replicates in the same plate but in different wells. One question would be: Can we treat all technical replicates the same? Regarding the biological replicates: those replicates are performed on a different day, with a new passage of cells and a new mixture of collagen ECM. The main statistical question thereby is whether we can treat biological and technical replicates the same in our graphs and when doing statistics on them, and, if that is not the case, how we should do statistics on these data correctly. Which test should we choose and which error bars should we show? Should we only show the SD between experiments or also the SD within the experiments?

Please shortly describe the data source(s) available (number of subjects, number of variables):

The data available are area measurements of tumoroids from three independent experiments, with a minimum of 4 technical replicates each. We have three different variables / conditions: soft ECM, stiff ECM and stiff ECM with inhibitor.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 16

Name researcher: Victorien Luppes
Department: Pediatric cardiology
Faculty or Company: LUMC

What is topic of your research project / what is the research question? (max 5 lines)

Neurodevelopmental disability is common in pediatric patients with congenital heart disease requiring surgical intervention, however, exact pathophysiologic background is not fully understood. Brain abnormalities, brain sparing and delayed head growth were discovered in prenatal ultrasounds and MRIs, suggesting that neurodevelopmental impairment might begin prenatally. Research Question: What is the impact of fetal programming on neurodevelopmental impairment in congenital heart disease? As neurodevelopment is multifactorial, and in order to estimate the impact of fetal programming on neurodevelopmental outcomes, we will examine a few demographics and perioperative parameters to correct for.

What statistical question(s) do you want to pose to the students?

This is retrospective research (largely), so we cannot influence the data. Which analysis method is most suitable to answer the research question? We have prenatal data (fetal ultrasounds), and postnatal data (follow up data in neurocognitive development; for some patients repeated measurements in time, e.g. at 3-4 years, 5-6 years, 8-9 years). We also have some demographics (e.g. educational levels of the parents) and perioperative data to correct for / take into account in the analysis. Should we analyze the data within a gradual scale, or should we analyze the data dichotomized. If dichotomized, how to decide which data are seen as normal prenatal development and which data are aberrant, especially if all prenatal ultrasounds are within normal ranges (although most tend towards aberrant values). Also, would generalized estimating equation (GEE) be an option to answer our question?

Please shortly describe the data source(s) available (number of subjects, number of variables):

There is no data source available now. However, we can provide a description of the to-be-measured variables and their properties or can provide a smaller 'example database', such that the students can simulate data.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

No, we do not know.

In what language is your dataset annotated (variable names and labels)?

I did not collect the data yet (in case of pre-registration/Statistical Analysis Plan)

Project 17

Name researcher: Dr. Elske van den Berg
Department: Psychology
Faculty or Company: Arkin Mental Health Institute, Novarum TOPGGz Center for Eating Disorders

What is topic of your research project / what is the research question? (max 5 lines)

212 clients suffering from a Binge Eating Disorder (BED) have completed a psychological treatment; digital Guided Selfhelp. Obesity is rather common amongst clients with BED. A poor body image is a core element of eating disorders but can also be found in overweight individuals. It is thought that poor body image may be associated with suboptimal treatment effect. Until now, little is know about early menarche and its impact on the development of a poor body image.

What statistical question(s) do you want to pose to the students?

(1) Is there an association between poor body image as measured with the *Body Shape Questionnaire* (BSQ,) and suboptimal treatment effect, measured with both *Eating Disorder Examination-Questionnaire* (EDE-Q global score) and number of binges at end-of-treatment ?
(2) Is there a relationship with early menarche and poor body image?
(3) Is there a relationship with childhood obesity and early menarche?

Please shortly describe the data source(s) available (number of subjects, number of variables):

Dataset of 212 participants from a RCT on efficacy of Digital Guided Selfhelp is available. We have measures at 3 timepoints (baseline, during treatment, and end of treatment).
All relevant variables are available in an excel dataset. The BSQ consists of 34 items, and the EDE-Q of 28 items. The (sub)scale scores were also constructed.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

I am open to the suggestions from the students.

In what language is your dataset annotated (variable names and labels)?

The dataset has Dutch annotation only

Project 18

Name researcher: Diego Nunez
Department: Horticulture and product physiology
Faculty or Company: Plant sciences group

What is topic of your research project / what is the research question? (max 5 lines)

Determine the effect of day lengths and dark period length altering lettuce growth morphology and carbohydrate metabolism.
We aimed to investigate how growth, morphology and carbohydrate metabolism respond to non-natural day length (so less or more than 24 h) in interaction with dark periods of 2, 4, or 6 hours

What statistical question(s) do you want to pose to the students?

How to deal with unbalanced designs and estimate marginal means
why the design is unbalanced?
Initial the experiment was planned RCBD, full factorial 5x3 with:
Five Day lengths: 20, 22 , 24 ,26 ,28 hours
Three Dark period: 2, 4, 6 hours
This would result in 15 treatments. However it turned into an unbalanced design when we only had cells available for 14 treatments then we decided to leave out 1 treatment the combination “20 h day length – 6 h dark period” that according to our criteria the less relevant to answer our research questions.

Please shortly describe the data source(s) available (number of subjects, number of variables):

14 treatments, incomplete design unbalanced, factorial 5 day lengths x 3 dark periods, variables: 6

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Mixed linear models in R package ‘lmertest’

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 19

Name researcher: Agnes Berendsen
Department: Division of Human Nutrition & Health
Faculty or Company: Wageningen University

What is topic of your research project / what is the research question? (max 5 lines)

We are studying the consequences of obesity treatment on musculoskeletal health, as both surgical and pharmacological treatment options may induce nutrient deficiencies and substantial body weight loss with consequent severe muscle mass loss and low bone mineral density.
Specifically, we would like to know whether post-obesity protein intake is related to disproportional muscle mass loss.

What statistical question(s) do you want to pose to the students?

Which method is most suitable to answer the above research question?
Currently we're using Cox proportional hazard regression models to study non-rare diseases (e.g., >10%) and logistic regression for rare diseases (e.g., outcome <10%), as otherwise log. Reg models tend to overestimate the true risk. But: would it be appropriate to use Poisson regression modelling for both situations?

Please shortly describe the data source(s) available (number of subjects, number of variables):

Currently, we are collecting the data, so no real-life data will be available. However, simulated data, or another dataset, can be used to answer the statistical question.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Survival analysis (cox proportional hazard regression) and Poisson regression.

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation
I did not collect the data yet (in case of pre-registration/Statistical Analysis Plan) > but we can make use of other existing data to simulate different models.

Project 20

Name researcher: Dr Julie Hall
Department: Health, Medical and Neuropsychology
Faculty or Company: Psychology

What is topic of your research project / what is the research question? (max 5 lines)

Can we classify subtypes of anxiety in Parkinson's disease (PD) based on their manifestations (behavioural, emotional, physiological, cognitive), and investigate the different associations between those manifestations and other PD symptoms if the manifestations may overlap? How the anxiety symptoms could vary is something I am interested in (to ultimately better diagnose anxiety in this patient group). The anxiety theory we use is described here https://www.ncbi.nlm.nih.gov/books/NBK470361/

What statistical question(s) do you want to pose to the students?

There are four different manifestations of anxiety, and we want to see the associations between these manifestations and other PD symptoms. However, the manifestations overlap, e.g. restlessness can be both behavioural and physiological. How can we investigate this if the groups are not independent from each other?

Please shortly describe the data source(s) available (number of subjects, number of variables):

A data file of 175 subject entries. We use seven different items of the Hospital and Anxiety Questionnaire to measure the four manifestations, and we have approx 12 other factors we would like to associate with these items.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

No

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 21

Name researcher: Selin Topel
Department: Institute of Psychology / Clinical Psychology Unit
Faculty or Company: Institute of Psychology - Leiden University

What is topic of your research project / what is the research question? (max 5 lines)

Do individual differences in Prospective and Inhibitory intolerance of uncertainty explain differences in information sampling behaviors and confidence in decisions?

What statistical question(s) do you want to pose to the students?

I would like to understand whether the relationship between Prospective intolerance of uncertainty (PIU) and self-reported confidence in decisions are mediated by amount of information individuals sampled prior to making a decision. PIU is measured at a higher (subject) level whereas information sampling and confidence measured at a lower (trial) level in the task. I want to do a multilevel mediation analysis and I have not done this before.

Please shortly describe the data source(s) available (number of subjects, number of variables):

N = 214; 60 trials each, information sampled, correct information, and confidence ratings for each trial as well as between-subject variables of Prospective and Inhibitory IU are the available.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

Mediation model with mixed-level data (i.e., between and within subjects trial-wise variables)

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 22

Name researcher: Rebecca Gomperts
Department:
Faculty or Company: LUMC

What is topic of your research project / what is the research question? (max 5 lines)

Safety and efficacy of mifepristone 50 mg as a weekly contraceptive

What statistical question(s) do you want to pose to the students?

- The Overall Pearl Index with weekly mifepristone 50 mg (presented in percentages with a 95% two-sided confidence interval) will be reported, as well as the Pearl Index for method failure.
- A life table analysis.
- Quality of life and sexual function changes

Please shortly describe the data source(s) available (number of subjects, number of variables):

Castor , excel export, total study will be 1000 subject
This is an interim analyses of appr 200 subjects some who started using the medication in August 2023 and some who only started in June 2024

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

In what language is your dataset annotated (variable names and labels)?

I can deliver the dataset with English annotation

Project 23

Name researcher: Neha Khandpur, Ellen Van Loo
Department: Human Nutrition and Health; Marketing and Consumer Behaviour
Faculty or Company: Wageningen University

What is topic of your research project / what is the research question? (max 5 lines)

What clusters of food products emerge based on the additives they contain?
What clusters of food products emerge based on nutrients they contain?
What clusters of food products emerge based on additives and nutrients they contain?

What statistical question(s) do you want to pose to the students?

What statistical techniques can be leveraged to identify clusters of food products, based on the two dimensions of the food product – the additives they contain and their nutritional composition?

Please shortly describe the data source(s) available (number of subjects, number of variables):

This database is a compilation of food and beverage products available in the US in May 2023. It contains the following information:

1. Name of the product
2. Serving size
3. Nutritional context per serving size
4. Ingredient list

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

We hope that the statistical consultants can give us some insights into the statistical options available, the insights the different techniques allow us to glean, and their limitations. This reference uses a continuous approach but a categorical approach to identifying ultra-processed foods would be more useful. Menichetti, G., Ravandi, B., Mozaffarian, D., & Barabási, A.-L. (2023). Machine learning prediction of the degree of food processing. Nature Communications, 14(1), 2312. doi: 10.1038/s41467-023-37457-1

In what language is your dataset annotated (variable names and labels)?

We can deliver the dataset with English annotation

Project 24

Name researcher: Sander Koenraadt
Department: Lab of Entomology
Faculty or Company: Wageningen University

What is topic of your research project / what is the research question? (max 5 lines)

Since 2020 data about nuisance due to mosquitoes are collected in the Netherlands, see https://www.naturetoday.com/intl/nl/observations/mosquito-radar. Weekly, persons are asked to quantify complaints (ordinal scale 0-3). The geographical location of the respondents is known. Many questions are of interest, but for now we want to focus on the relationship with weather and climate related variables.

What statistical question(s) do you want to pose to the students?

First step is to prepare the datafiles for mosquito nuisance and weather (from the KNMI website), and to combine them, on a weekly basis and maybe weather split regionally.
Special interest in the R-scripts to do this easily on regular basis. Next, time trends should be visualized, and relationships of complaint score with weather variables (also temperaturedays) should be studied.

Please shortly describe the data source(s) available (number of subjects, number of variables):

Data from the mosquito-radar (thousands of records, needs modification; ordinal complaint score, geographical position and other). Specific KNMI data, needs further research.

If known already, would you like to perform particular statistical techniques on the data (e.g, machine learning, mixed models, survival analysis, meta-analysis)?

We are open to your suggestions

In what language is your dataset annotated (variable names and labels)?

The dataset has Dutch annotation only (but quite irrelevant)