Univariate Statistics and Methodology in R

Department of Psychology, The University of Edinburgh

Academic year 2021-2022

Contents

How Assignments Were Marked **1**

Notes about the Report **1**
Question 0: Cleaning & Describing 2
Question 1: General Checks 3
Question 2: Happiness 4
Question 3: Happiness and Health 5
Question 4 6
Question 5: Predictors of Drop-out 7

How Assignments Were Marked

We marked the assignments 'positively' (that is, we looked for examples of good work and awarded marks, rather than deducting marks for getting things 'wrong'). Here are some of the things we took into account (not always applicable to all questions):

Clear written details of the analysis conducted

- was a statistical test used? was it appropriate?
- was the analysis conducted clearly described?
- were any observations excluded from the analysis, and if so, has the reason been explained?

Results

- were the results reported clearly in appropriate detail (for instance, a test statistic, standard error and p-value, not just one of these), including uncertainty in any point estimates?

Interpretation

- were the results interpreted correctly?
- was a conclusion drawn based on the results?
- was there a clear link to which bit of the result was used to draw a conclusion on the research question?

Presentation

- (if applicable) was a visualisation provided? Was it 'publication ready'?

Notes about the Report

- every student got different data.
- most relationships in the data were similar, although (because of randomness) there were some differences between reports.

Question 0: Cleaning & Describing

> Have a look at the data. Check for impossible values and deal with these in an appropriate manner. Describe the data, either in words or using suitable graphs (or a combination). Remember to detail the decisions you have made.

The key thing here is the justification given for any actions taken. If there's a choice to exclude a whole subset from all analyses at this point, then that is fine, provided there's a reasonable justification for doing so.

Notes good reponses would:

- describe the dataset, possibly in table form
- describe any actions taken wrt the raw data and justifications for taking them
- possibly cross-tabulate reasons for missingness (seen in lectures)
- There were some impossible values added to *all* students' data:
  - `age` **Integer** shouldn't have negative values (there shouldn't have been, but just in case some of the data generation led to it) **by definition of variable**. There were a couple of values ≥ 100, which should possibly have been excluded (at least discussed).
  - `accountability` **Integer** should be $5 \leq x \leq 35$ - **by definition of variable**
  - `selfmot` **Integer** should be $5 \leq x \leq 35$ - **by definition of variable**
  - `health` **Integer** should be $0 \leq x \leq 100$ - **by definition of variable**
  - `happiness` **Integer** should be $0 \leq x \leq 100$ - **by definition of variable**
  - `season` should be one of "spring", "summer", "autumn", "winter" (there was a deliberate misspelling of "autunm" here)
  - `city` should be one of "Edinburgh", "Glasgow"
  - `week_stopped` **Integer** should be $1 \leq x \leq 9$ - **by definition of variable**

```
couchto5k <-
  couchto5k %>% mutate(
    age = ifelse(age>100, NA, age),
    selfmot = ifelse(selfmot < 0, NA, selfmot),
    season = fct_relevel(factor(ifelse(season == "autunm","autumn",season)),
                         "spring","summer","autumn","winter"),
    city = factor(city),
    week_stopped = ifelse(week_stopped > 9, NA,week_stopped)
  )
```

Question 1: General Checks

- **1a.** In an earlier nationwide survey, researchers found that 45% of participants abandoned the programme before the halfway point in week 5, and a further 10% gave up before the end of the programme. Is the data in the sample you have been given in line with data from the earlier survey? Once you have created a suitable variable to map to the information in the question, you should be able to answer this using a simple statistical test.
- **1b.** Using the same three categories (stopped before week 5, stopped after week 5, completed), examine whether the patterns of attrition rates differ by city.
- **1c.** Do the average ages of participants who commenced the programme differ by city?

Notes

- this needed a new variable for early/late/no dropout, based on the week_stopped.
  We would then most likely expect the following approaches (these are examples of appropriate methods, they are not the *only* valid approaches that may be taken):
  - 1a. = χ^2 goodness of fit
  - 1b. = χ^2 test of independence/homogeneity
  - 1c. = independent t test
  - A table perhaps, and maybe a boxplot for the t.test?
- For 1a, results with very low p-values *might* be because the levels are the wrong way around in the chisq probabilities. For 1b, we didn't code anything specifically in, so we were expecting mostly non-signif results (with 5% being signif by chance!). For 1c we expected most students to find some differences.

```
couchto5k <-
  couchto5k %>% mutate(
    dropout = ifelse(week_stopped < 5, "earlydropout",
                     ifelse(week_stopped <= 8, "latedropout","nodropout"))
  )

with(couchto5k, table(dropout))
# MAKE SURE THE PROBABILITIES ARE IN THE RIGHT ORDER FOR THE LEVELS IN THE TABLE!!
# we've explicitly labelled them so that they are (alphabetically) in the order we want: earlydropout,
chisq.test(table(couchto5k$dropout), p = c(.45, .1, .45))

#with(couchto5k, table(dropout,city))
chisq.test(with(couchto5k, table(dropout,city)))

with(couchto5k, t.test(age~city))
```

Question 2: Happiness

- **2a.** Are participants' happiness ratings affected by the season they were interviewed in? Describe the way in which season influences happiness outcomes.
- **2b.** Accounting for any effects you discovered in (2a), is happiness affected by age?
- **2c.** The models you have built above explore 'baseline' effects; that is, effects that are not of primary interest to the researchers but which might affect the outcome variable of happiness. For use in question 3, pick a specific baseline model and justify why you are using this.

Notes

- the expectation was that this question (as well as 3 and 5) could be answered using simple linear models, like the ones below.
  - 2a. = `lm(happiness ~ season)`
  - 2b. = `lm(happiness ~ season + age)`
- For most students, there was probably a relationship between happiness and season, and probably no effect of age. The important thing for 2b was appropriate use of sums of squares (see https://uoepsy.github.io/usmr/labs/zz_ss.html).
  - 2c. = just a textual argument was needed here. If anyone argued strongly that age was a theoretically important predictor of happiness, and wanted to keep it in, that was OK. Ideally season would have stayed in whatever.
  - boxplots, for example, could be an appropriate visualisation for the season bit. We'd also expect some regression tables or model comparison tables.

```
m0 <- lm(happiness ~ season, couchto5k)
m1 <- lm(happiness ~ season + age, couchto5k)
anova(m1)
```

Question 3: Happiness and Health

- **3a.** Building on your baseline model, are participants' happiness ratings affected by whether or not they completed the programme? Describe the way in which programme completion influences happiness outcomes.
- **3b.** Building on the analysis in (3a), is happiness additionally affected by the "health metric"?
- **3c.** It's been hypothesised that the effects of good health are amplified by the feeling of acting healthily, such that the happiness of participants who got further along the programme might be more affected by the health metric than that of those who stopped earlier. Building on the model in (3b), can you test this hypothesis?
- **3d.** What can we conclude about the various causes of happiness in our data? Write a brief description of the effects in the model, such as you might find in an academic paper.

Notes

- 3a and 3b were (hopefully clearly) a set of nested models, running something like:
  - 3a. = `update(baseline, .~. + I(week_stopped==9))`
  - 3b. = `update(baseline, .~. + I(week_stopped==9) + health)`
- obviously the above is very compact and includes the `update()` function which you may not have seen, but 3a can be read as 'take the baseline model, and add another predictor (inside `I()`) which is `TRUE` if `week_stopped` is 9'. We expected most of you to do this in stages, something like

```
# create a new column called 'completed'
couchto5k <- couchto5k %>% mutate(completed = week_stopped==9)

# create a new model based on the baseline model
# (here we assume that the baseline model chosen only
# includes season as a predictor)
mod.3a <- lm(happiness~season+completed,couchto5k)
```

- 3c Was a bit of a sidestep because it suggests an interaction with week number rather than with 'completion'. It was probably ambiguous enough to allow for either model, though, so we awarded credit for either. The important thing here was to include the interaction (example below):

```
### example with interaction of health and week_stopped
mod.3c <- lm(happiness ~season+health*week_stopped,couchto5k)
```

- 3d: We were looking for text which made it clear that the author understood the models they'd picked.
- In general, in the generated data, we expected happiness to be affected by season (for most reports), health, and week stopped. There should (for most reports) have been an interaction between health and week_stopped. In the output below we've assumed that the baseline model includes season (only); it might also include age (this is fine; justification for which model was assessed in 2c). There was no *requirement* to draw a graph but credit was given if a good graph was presented, as long as there was a textual description to accompany any graph or table.
- For visualisations, we expected to see some plots of predicted values of happiness by completion and by health, and then (hopefully) and interaction plot for 3c. There would also ideally have been some regression tables or model comparison tables.

Question 4

- Create a subset of the data, including only those participants who completed the programme. Create a plot of the average happiness ratings grouped by season and city, that can be used in a presentation to the funders of the project.

Notes

- we'd expect most students to create a new dataframe at this point
- plots that include some kind of indication of uncertainty were preferred (this one uses `mean_se()` implicitly)
- excellent answers would describe (e.g. in caption) the specifics of the plot (e.g. "means and standard errors of .....")

```
couchto5k %>% filter(week_stopped==9) %>%
  ggplot(.,aes(x=season, y=happiness, col=city))+
  stat_summary(geom="pointrange",position=position_dodge(width=.4))+
  stat_summary(geom="path",aes(group=city),position=position_dodge(width=.4))+
  theme_minimal()
```

Question 5: Predictors of Drop-out

- **5a.** Build a model that predicts the likelihood of dropping out (at all).
- **5b.** Briefly describe the effects in your model as you would in an academic paper.
- **5c.** Draw a graph representing the probability of quitting as a function of how self motivated participants were.

- It was hard to know what people would do here – this question was, deliberately, very exploratory, to allow a bit more room for manoeuvre after the very guided nature of questions 0-4. Some of you found the `step()` function; quite a few built a model in a similar fashion to questions 2 and 3, which is a perfectly sensible approach.
- Ideally, this really should have been a logit model; extra marks were awarded for properly explaining the mapping between log-odds and probability (reminder: $p = \frac{e^l}{1+e^l}$; $l = \ln \frac{p}{1-p}$; $e^x =$ `exp(x)`; $\ln x =$ `log(x)`)
  - 5a/b. = extra marks were awarded for a narrative which made it clear how the question was approached.
  - 5c = some authors used `predict()`, which is good. The graph should ideally be in terms of probability. (The other way of doing this is in `ggplot` via `geom_smooth()` with the right method and family.)
- it is hard to decide how to present data if there are many predictors in the model (authors would have to use, e.g., colour as a dimension, or 'fix' some predictors to sensible values). Good efforts were generously rewarded.
- **Assumptions:** there were no penalties for cursory, or non-existent, assumption-checking, since we didn't go in to assumptions for logit models on the course.

```
1 couchto5k$droppedout <- factor(1*couchto5k$week_stopped!=9)
2 mdrop = glm(droppedout ~ ., couchto5k %>% select(-week_stopped,-dropout,-pptID) %>%
3               mutate(across(age:happiness, ~scale(.)[,1]))
4             , family=binomial)
```

NB. we fully admit that we're showing off a bit with the code above! Picking it apart,

- **line 1** uses the `*1` trick to turn a logical (`couchtoweek5k$week_stopped!=9`) into an integer (1 for `TRUE`, so 1 means "stopped in a week other than 9"), and turns that into a factor.
- **line 2** `droppedout ~ .` is shorthand for "`droppedout` is predicted by everything in the dataframe that isn't `droppedout`". The "data" argument is the second argument: We select relevant columns from `couchto5k` (`select(...,-X,...)` drops column X) ...
- **line 3** ... and then `scale()` (z-score) all of the columns between `age` and `happiness` (NB., this is a bit of a fudge, because it relies on column order)
- **line 4** is the third (important) argument giving the family of the linking function for `glm()`.

Obviously R code doesn't *have* to be as compact as that, but we thought that possibly one or two of you might enjoy seeing it.