Documentation

Alvin Gavel

Contents

1 What are we trying to find out?

We have given a group of participants five different tests. We have split them into a test group that has been subjected to sleep deprivation and a control group that has not[1]. After each test we asked them to estimate their own performance. We now want to find out how sleep deprivation affects their estimated performance.

What complicates matters is that when the participants were asked to estimate how well they did, the participants were simply asked to rate their performance on a scale. This means that we have no straightforward way of checking how accurate their estimates are. If they had been asked to estimate what fraction of questions they got right, then we could straightforwardly say that, for example, someone who got 0.5 right and estimates they got 0.8 right did worse than someone who got 0.8 right and estimates they got 0.9 right. As it is, the first person might have estimated their performance as "8", and we simply do not know what – in their mind – that number was supposed to mean, which in turn means that we cannot say if it was close to their actual performance.

We can work around this problem if we assume that if people had perfectly accurate metacognition, then their performance ratings would be a monotonously increasing function of actual performance. A simple consequence of this is that people who perform above the median should also rate themselves above median, and that we can use their actual tendency to do so as a proxy for metacognitive skill. In Sect. 3 we take an approach based on this assumption. We could take the more sophisticated approach of looking at how well they are at identifying their overall ranking, seeing if, say, the person who performed seventh-best in a group of ten also gave themselves the seventh-highest rating in that group. If time permits, we may make such a test.

However, we need to point out that there is a criticism that could be made of this underlying assumption. It assumes that the participants are trying to grade themselves on a universal scale. It could be that they are only trying to rate their performance on that particular day to how they would ordinarily expect to perform – or, more likely, that they are trying to do something that is difficult to express in well-defined mathematical terms.

In Sect. 4 at a method that is arguably more sophisticated but also much harder to interpret.

[1] At least not systematically, as part of the experiment. Sometimes people just sleep badly.

2 What does the data look like?

There are six csv-files in the directory `Data`[2]:

- `kss_data.csv`
- `arithmetic_data.csv`: Data from arithmetic test
- `working_memory_data.csv`: Data from working memory test
- `episodic_memory_data.csv`: Data from Episodic memory test
- `Stroop_data.csv`: Data from Stroop test
- `simple_attention_data.csv`: Data from simple reaction time test

At the moment, we only use the file `kss_data.csv`. At time of writing I do not know what all columns are, so our best interpretation is this:

- `id`: Participant ID
- `age`: Participant age
- `woman`: Participant gender
- `sd`: Participant sleep status
- `pair`: Participant pair for social interaction tasks
- `dyad_type`: Combination of sleep status in dyad
- `X1`: ¯_(ツ)_/¯
- `date`: Date of test
- `clock`: Time of test
- `order_t`: Test order for each session
- `test_type`: Which cognitive test the participant just finished. This uses the abbreviations:
  - *M*: Arithmetic
  - *W*: Working memory
  - *ST*: Episodic memory
  - *stroop*: Stroop
  - *reactionTime*: Simple reaction time
- `rating1_type`: The type of rating in the column `rating1`
- `rating1`: Self-rated sleepiness after each test. That is, the participant's response to the question "How sleepy are you right now?"
- `rating2_type`: The type of rating in the column `rating2`.

[2]It is currently not in the repo. This may change in the future.

- `rating2`: Self-rated effort on each test. That is, the participant's response to the question "How much of an effort did you make?"
- `rating3_type`: The type of rating in column `rating1`
- `rating3`: Self-rated performance on each test. That is, the participant's response to the question "How did you perform?"
- `time_adjustment_hours`: ¯_(ツ)_/¯
- `old_time`: ¯_(ツ)_/¯
- `correct`: This is a dummy column that can be ignored.
- `time`: Order of the session in the test battery

For now, the columns of interest to us are `id`, `sd` and `rating3`. We want to know how self-rated effort depends on sleep status.

3 Simple binomial test

Instead of worrying about how to interpret what the participants' meant by the performance estimates they gave us, we simply look at how likely participants who gave themselves a rating r_i above the median $\tilde{r}$ are to also have an actual performance a_i above the median $\tilde{a}$. That is, we have two binomial distribution with some probabilities $p^{\mathrm{t}}_{\mathrm{a},\,\mathrm{a}}$ and $p^{\mathrm{c}}_{\mathrm{a},\,\mathrm{a}}$, and we want to see if they differ enough for it to be clinically significant[3], where t and c denote the test and control group, respectively.

3.1 Why $p^{\mathrm{x}}_{\mathrm{a},\,\mathrm{a}}$ alone tells us all we need to know

We will only work with the probabilities $p^{\mathrm{x}}_{\mathrm{a},\,\mathrm{a}}$ that a participant in either group who had above-median performance also gave themselves an above-median rating, since this uniquely determines any other probabilities that we might be interested in. The corresponding probabilities $p^{\mathrm{x}}_{\mathrm{a},\,\mathrm{b}}$ that a person who performed above the median gave a rating below the median is trivially given by:

$$p^{\mathrm{x}}_{\mathrm{a},\,\mathrm{b}} = 1 - p^{\mathrm{x}}_{\mathrm{a},\,\mathrm{a}}. \tag{1}$$

If we use n to denote the corresponding numbers of participants, then the probability is also definitionally given as

$$p^{\mathrm{x}}_{\mathrm{a},\,\mathrm{b}} = \frac{n^{\mathrm{x}}_{\mathrm{a},\,\mathrm{b}}}{n^{\mathrm{x}}_{\mathrm{a},\,\mathrm{a}} + n^{\mathrm{x}}_{\mathrm{a},\,\mathrm{b}}}. \tag{2}$$

Analogiously, the probability $p^{\mathrm{x}}_{1,\,\mathrm{a}}$ that a participant who performed above median rated themselves above median is given by:

$$p^{\mathrm{x}}_{\mathrm{a},\,\mathrm{a}} = \frac{n^{\mathrm{x}}_{\mathrm{a},\,\mathrm{a}}}{n^{\mathrm{x}}_{\mathrm{a},\,\mathrm{a}} + n^{\mathrm{x}}_{\mathrm{a},\,\mathrm{b}}}. \tag{3}$$

Analoguously, the probability $p^{\mathrm{x}}_{\mathrm{b},\,\mathrm{a}}$ that a participant who performed below median rated themselves above median is given by:

$$p^{\mathrm{x}}_{\mathrm{b},\,\mathrm{a}} = \frac{n^{\mathrm{x}}_{\mathrm{b},\,\mathrm{a}}}{n^{\mathrm{x}}_{\mathrm{b},\,\mathrm{a}} + n^{\mathrm{x}}_{\mathrm{b},\,\mathrm{b}}}. \tag{4}$$

Analoguously, the probability $p^{\mathrm{x}}_{\mathrm{b},\,\mathrm{b}}$ that a participant who performed below median rated themselves below median is given by:

$$p^{\mathrm{x}}_{\mathrm{b},\,\mathrm{b}} = \frac{n^{\mathrm{x}}_{\mathrm{b},\,\mathrm{b}}}{n^{\mathrm{x}}_{\mathrm{b},\,\mathrm{a}} + n^{\mathrm{x}}_{\mathrm{b},\,\mathrm{b}}}. \tag{5}$$

[3] A common mistake in this kind of study is to set up a "null hypothesis" that the two parameters are exactly the same, and then try to "falsify" that hypothesis by seeing if the difference is "statistically significant". Such a test is not scientifically meaningful. Since they are continuous parameters it is a priori given that they will differ my some amount, so that test would tell us nothing that we did not already know before we even gathered the data.

Definitionally, the number of participants who performed above and below median are equal:

$$n^{x}_{a,\,a} + n^{x}_{a,\,b} = n^{x}_{b,\,a} + n^{x}_{b,\,b} \tag{6}$$

Analoguously, the numbers of participants who rated themselves above and below median are equal:

$$n^{x}_{a,\,a} + n^{x}_{b,\,a} = n^{x}_{a,\,b} + n^{x}_{b,\,b} \tag{7}$$

Adding both sides in (6) and (7) together and simplifying gives us:

$$n^{x}_{a,\,a} = n^{x}_{b,\,b} \tag{8}$$

Analoguously, substracting both sides in (6) and (7) and simplifying gives us:

$$n^{x}_{a,\,b} = n^{x}_{b,\,a} \tag{9}$$

Inserting (8) and (9) into (4) we get:

$$p^{x}_{b,\,a} = \frac{n^{x}_{a,\,b}}{n^{x}_{a,\,b} + n^{x}_{a,\,a}}$$

Which, by (2), means that:

$$= p^{x}_{a,\,b}. \tag{10}$$

Analogiously, inserting (8) and (9) into (4) gives us:

$$p^{x}_{b,\,b} = \frac{n^{x}_{a,\,a}}{n^{x}_{a,\,b} + n^{x}_{a,\,a}} \tag{11}$$

Which, by (3), means that:

$$= p^{x}_{a,\,a}. \tag{12}$$

That said, there is a quicker way one could reach the same conclusion. We imagine starting out from an ideal state where everybody rates themselves with perfect accuracy. In this case, it is trivially true that $p^{x}_{a,\,a} = p^{x}_{b,\,b}$ because they are both 1, while $p^{x}_{a,\,b} = p^{x}_{b,\,a}$ because they are both 0. Now we imagine taking one person who performed above the median and shifting their rating from above to below the median. This moves the median rating in such a way that one person who performed below the median will now have a rating above the median. We can shift people around as much as we like, and the change to $p^{x}_{a,\,a}$ will equal the change in $p^{x}_{b,\,b}$ and the change in $p^{x}_{a,\,b}$ will equal the change in $p^{x}_{b,\,a}$. Completely analogous reasoning can be applied to shifting the actual performance instead of the rating. As a result, the equalities $p^{x}_{a,\,a} = p^{x}_{b,\,b}$ and $p^{x}_{a,\,b} = p^{x}_{b,\,a}$ will still hold no matter how we shift people around.

3.2 How to estimate the medians $\tilde{r}$ and $\tilde{a}$

The distributions of ratings r and actual performances a follow distributions that have some medians $\tilde{r}$ and $\tilde{a}$. We cannot access those numbers directly, but we can estimate them by taking the medians over the samples r_i and a_i that we actually got in the study. This leads to the question: Should we estimate separate medians for the control and test groups based on the samples for each group, or should we use a shared median for both groups?

I believe that taking separate medians for both groups is more robust to violations of the assumptions behind our analysis. There could be an overall shift between the medians of the distribution that is not due to the participants' metacognitive skill changing, but due to some difference in how they perceive the framing of the question. If we looked at a shared median for both group, we would risk confusing such an effect for a change in metacognitive skill. However, if we see that ratings are much noisier in one group than in the other, it is hard to argue that that does not reflect a genuine lowering of metacognitive skill.

4 Cumulative model

In the article *Too Tired to Know* we implement the cumulative model described in Bürkner and Vuorre (2019). The reader is primarily recommended to that article for the details, or – once it is written – the section in our article that explains the analysis. This is mostly written for myself to force myself to understand how the model works, since the root of at least half of all evil is the use of statistical recipes without first understanding them.

4.1 Basic idea

The cumulative model starts out from the assumption that the ordinal variable that we actually measure, Y, is an increasing function of some underlying unobservable variable $\tilde{Y}$. Specifically it assumes that there are K thresholds τ_k such that if $\tau_{k-1} < \tilde{Y} < \tau_k$ then we will observe $Y = k$. The underlying $\tilde{Y}$ in turn follows some probability density function f such that the probability $P(Y = k)$ is given by

$$P(Y = k) = \int_{\tau_{k-1}}^{\tau_k} f\left(\tilde{Y}\right) d\tilde{Y} \tag{13}$$

If we call the corresponding cumulative density function F, this is

$$= F(\tau_k) - F(\tau_{k-1}) \tag{14}$$

4.2 Linear model of prediction term

So far the model is very general, but to get anywhere we need to make some assumptions about what the function f actually looks like. We split $\tilde{Y}$ into a predictor term η and an error term ε, so that

$$\tilde{Y} = \eta + \varepsilon \tag{15}$$

We now assume that η is a linear function, so that

$$\eta = \sum_{n=0}^{K} b_n x_n \tag{16}$$

Inserting (16) into (15) gives us

$$\tilde{Y} = \sum_{n=0}^{K} b_n x_n + \varepsilon \tag{17}$$

Since $d\tilde{Y}/d\varepsilon = 1$ we can rewrite (13) as

$$P(Y = k) = \int_{\tau_{k-1}}^{\tau_k} f\left(\sum_{n=0}^{K} b_n x_n + \varepsilon\right) d\varepsilon \tag{18}$$

Shifting the integration boundaries by the predictor term gives us

$$= F\left(\tau_k - \sum_{n=0}^{K} b_n x_n\right) - F\left(\tau_{k-1} - \sum_{n=0}^{K} b_n x_n\right) \tag{19}$$

4.3 Normal model of error term

We now need some kind of model of what that error term looks like. To make life simpler[4], we assume that it is normally distributed. If we denote the CDF of the normal distribution with Φ, then (19) takes the form:

$$P(Y = k) = \Phi\left(\tau_k - \sum_{n=0}^{K} b_n x_n\right) - \Phi\left(\tau_{k-1} - \sum_{n=0}^{K} b_n x_n\right) \tag{20}$$

This means we have a regression problem where we need to estimate the thresholds τ_k and τ_{k-1} and the regression coefficients b_n.

[4]It is common to justify assumptions of normality with reference to the central limit theorem. I would argue that this often over-interprets the central limit theorem as being much broader than it actually is, and that the real reason why normal distributions get used so often is that they are just so very convenient.