

Documentation

Alvin Gavel

Contents

1	What are we trying to find out?	2
2	Simple binomial test	3
2.1	Binomial fitting	4
2.2	Finding the distribution over the difference	5
2.3	Probability of clinical significance	6
2.4	Why $P_{a, a}^x$ alone tells us all we need to know	7
2.5	How to estimate the medians \tilde{r} and \tilde{a}	9
3	What does the data look like?	10

1 What are we trying to find out?

We have given a group of participants five different tests, each one of which they had to take three different times. They were split into a test group that has been subjected to sleep deprivation and a control group that has not¹. 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, they 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. 2 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.

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

2 Simple binomial test

Our approach is intended to sidestep the question of how to interpret what the participants meant by the performance estimates they gave us, by attempting to cover all interpretations that are compatible with any reasonable definition of ‘metacognitive skill’.

Indexing participants by i and test attempts by k , we look at how likely participants who gave themselves a rating $r_{i,k}$ above the median rating \tilde{r}_k are to also have an actual performance a_i above the median performance \tilde{a}_k . We denote the probability that a participant in the control group will both rate themselves above the median rating as $P_{a,a}^t$, and the corresponding probability for a participant in the test group with $P_{a,a}^c$. If we need to refer to them we will denote the probability that a test-group participant who rated themselves above the median actually performed below the median with $P_{a,b}^t$, and analogously for the other five combinations of group, rating and performance. However, as we will see in Sect. 2.4, for each group one of the four probabilities uniquely determines the other three, so we should not need to refer to them explicitly. We also define $n_{a,a}^c$ to mean the number of participants in the control group who rated themselves above the median and actually performed above the median, and analogously for the test group and ratings and performance below the median.

The parameters $P_{a,a}^t$ and $P_{a,a}^c$ each define a binomial distribution, and by fitting them to the data we will get two probability distributions over the parameters, as described in Sect. 2.1. By convolving the two distributions we get the probability distribution over their difference $D_{a,a}$, as described in Sect. 2.2. By integrating over a segment of that distribution we can get the probability they differ enough for it to be clinically significant², as described in Sect. 2.3.

²A 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 by some amount, so that test would tell us nothing that we did not already know before we even gathered the data.

2.1 Binomial fitting

The binomial distribution describes any random process equivalent to flipping a weighted coin. That is, there are two outcomes h and t and some probability P_h of getting h and $P_t = 1 - P_h$ of getting t . If you take n samples, then the likelihood $P(n_h|n, P_h)$ of getting n_h cases of outcome h and $n_t = n - n_h$ cases of outcome t is given by:

$$P(n_h|n, P_h) = \binom{n}{n_h} P_h^{n_h} (1 - P_h)^{n - n_h} \quad (1)$$

where

$$\binom{n}{n_h} \equiv \frac{n!}{n_h! (n - n_h)!}$$

Now, say that P_h is unknown and that we want to estimate it based on these samples. By Bayes' theorem, the probability density $p(P_h|n, n_h)$ for any particular value P_h is given by:

$$p(P_h|n, n_h) = \frac{P(n_h|n, P_h) p(P_h)}{P(n_h|n)} \quad (2)$$

where $p(P_h)$ is the prior probability density over P_h and $P(n_h|n)$ is a normalisation factor. To make our life simpler we choose a flat prior. With a bit of work it can be shown that if insert this prior and (1) into (2) and require the distribution to be normalised then we end up with:

$$p(P_h|n, n_h) = \frac{P_h^{n_h} (1 - P_h)^{n - n_h}}{B(n_h + 1, n - n_h + 1)}$$

where

$$B(\alpha, \beta) \equiv \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

and Γ is the gamma function. Switching to the bulkier notation defined in the beginning of Sect. 2, we get:

$$p(P_{a,a}^x | n_{a,a}^x, n_{a,b}^x) = \frac{(P_{a,a}^t)^{n_{a,a}^x} (1 - P_{a,a}^t)^{n_{a,b}^x}}{B(n_{a,a}^x + 1, n_{a,b}^x + 1)} \quad (3)$$

where x denotes t or c .

2.2 Finding the distribution over the difference

Now, we are not interested in the probabilities $P_{a, a}^c$ and $P_{a, a}^t$ in themselves. For the purposes of our study we are only interested in the difference $D_{a, a}$ between them. It turns out that we can get the probability distribution $p(D_{a, a})$ over this difference simply by convolving the probability distributions $p(P_{a, a}^c | n_{a, a}^c, n_{a, b}^c)$ and $p(P_{a, a}^t | n_{a, a}^t, n_{a, b}^t)$.

2.3 Probability of clinical significance

Once we have the probability distribution $p(D_{a,a})$, the probability that the difference is clinically significant is given by simply by integrating between our bounds for clinical significance.

2.4 Why $P_{a, a}^x$ alone tells us all we need to know

We will only work with the probabilities $P_{a, a}^x$ 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 $P_{a, b}^x$ that a person who performed above the median gave a rating below the median is trivially given by:

$$P_{a, b}^x = 1 - P_{a, a}^x. \quad (4)$$

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

$$P_{a, b}^x = \frac{n_{a, b}^x}{n_{a, a}^x + n_{a, b}^x}. \quad (5)$$

Analogously, the probability $P_{b, a}^x$ that a participant who performed below median rated themselves above median is given by:

$$P_{b, a}^x = \frac{n_{b, a}^x}{n_{b, a}^x + n_{b, b}^x}. \quad (6)$$

Analogously, the probability $P_{b, b}^x$ that a participant who performed below median rated themselves below median is given by:

$$P_{b, b}^x = \frac{n_{b, b}^x}{n_{b, a}^x + n_{b, b}^x}. \quad (7)$$

Analogously, the probability $P_{a, a}^x$ that a participant who performed above median rated themselves above median is given by:

$$P_{a, a}^x = \frac{n_{a, a}^x}{n_{a, a}^x + n_{a, b}^x}. \quad (8)$$

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

$$n_{a, a}^x + n_{a, b}^x = n_{b, a}^x + n_{b, b}^x \quad (9)$$

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

$$n_{a, a}^x + n_{b, a}^x = n_{a, b}^x + n_{b, b}^x \quad (10)$$

Adding both sides in (9) and (10) together and simplifying gives us:

$$n_{a, a}^x = n_{b, b}^x \quad (11)$$

Analogously, subtracting both sides in (9) and (10) and simplifying gives us:

$$n_{a, b}^x = n_{b, a}^x \quad (12)$$

Inserting (11) and (12) into (7) we get:

$$P_{b, a}^x = \frac{n_{a, b}^x}{n_{a, b}^x + n_{a, a}^x}$$

Which, by (5), means that:

$$= P_{a, b}^x. \quad (13)$$

Analogously, inserting (11) and (12) into (7) gives us:

$$P_{b, b}^x = \frac{n_{a, a}^x}{n_{a, b}^x + n_{a, a}^x} \quad (14)$$

Which, by (6), means that:

$$= P_{a, a}^x. \quad (15)$$

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 $P_{a, a}^x = P_{b, b}^x$ because they are both 1, while $P_{a, b}^x = P_{b, a}^x$ 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 $P_{a, a}^x$ will equal the change in $P_{b, b}^x$ and the change in $P_{a, b}^x$ will equal the change in $P_{b, a}^x$. Completely analogous reasoning can be applied to shifting the actual performance instead of the rating. As a result, the equalities $P_{a, a}^x = P_{b, b}^x$ and $P_{a, b}^x = P_{b, a}^x$ will still hold no matter how we shift people around.

2.5 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.

3 What does the data look like?

There are six csv-files in the directory `Data`³:

- `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

The files with data for each test each need to be combined with `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`: `_(\^)_/`
- `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`.

³It 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`: `_("ʘ")_/`
- `old_time`: `_("ʘ")_/`
- `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.