Did this rat learn, by watching?

Luca
<pgl@portamana.org>

Ida
<ida.v.rautio@ntnu.no>

28 July 2020; updated 3 August 2020

How to approach and answer this question? Here is a methodological analysis.

1 Some obvious but sometimes forgotten remarks

Whenever 2 moles of hydrogen and 1 mole of oxygen react, we obtain 2 moles of water. We don't obtain sometimes 2 moles, sometimes 3 moles, sometimes 1/2 mole of water. It's always 2 moles, no matter where the hydrogen and oxygen come from.

Some experiments and observations in biology and in neuroscience also have such a high level of reproducibility. For example we are always sure that any rat will become unconscious (and possibly suffer from side effects) after inhaling a specific dose of isoflurane. But the majority of experiments and observations, especially behavioural ones, are not so reproducible.

For example, after having long trained some rats for a specific task we notice that some of them yield poor results in a subsequent short performance test. The instances with poor results are caused not by a lack of learning, but by biological and environmental factors – fatigue, humanly inaudible noises, stress, peculiar mental states – which can never be fully controlled in the experimental or observational protocol, or of which we aren't even aware. It would be momentous to find a different kind of test – say, checking the activity or connectivity of some brain area – that yields perfect reproducibility instead. Hopefully Ida's research will find such a neurological test.

It is because of this lack of perfect reproducibility for a specific test that we turn to a statistical analysis; that is, analysis of collections rather than of an individual instance. Ideally we'd like to perform the test on all identically trained rats (of similar characteristics such as age, gender, and so on) that ever lived and will ever live; this is technically called a 'superpopulation'. Let's say that our test can yield scores 0 to 10, and that we perform this test on the whole superpopulation (this is a "thought

This document is designed for screen reading and two-up printing on A4 or Letter paper experiment”). Imagine that we observed 1% of all trained rats to yield score 0, and 2% to yield score 1, . . . , and 89% to yield score 10. This ideal, superpopulation statistical knowledge would be represented by a precise histogram over the scores.

If someone now presented us a trained rat and asked us to forecast what its score will be in such a test, we would reply “I have a 1% degree of belief that it’ll yield score 0, a 2% degree of belief that it’ll yield score 1, . . . , an 89% degree of belief that it’ll yield score 10”. Our beliefs would reflect the collected superpopulation statistics for all trained rats – which includes this specific rat.

The mathematical relation between superpopulation statistics, our degrees of belief in a specific instance, and the statistics of a small observed sample is given by de Finetti’s (1937) theorem.

Beliefs are all we can have and express in a specific instance. Talk of probability or “true probability” as a sort of physical property of one test is misleading, and false from a physics point of view. The outcome of every test is determined by the precise biological and neurological condition of the rat and the precise physics condition of the environment. This outcome could be perfectly predicted, given all these conditions up to the position of every single protein in every single cell and of every single air molecule, and given enough computational power. So the “physical, true probability” for each score of the test is either 0 or 1, nothing in between. See the concrete example by Diaconis et al. (2007).

The lack of reproducibility also affects our backwards inferences. If someone gives us an amount of water and we measure it to be 2 moles, and we are asked “how many moles of hydrogen and oxygen reacted to produce this amount of water?”, then we confidently reply “2 moles hydrogen and 1 mole oxygen”. If someone gives us a rat, which upon testing yields a score of 1 in the test mentioned above, and we are asked “is this a trained rat?”, we cannot be sure. Maybe it isn’t trained; or maybe it is trained but the test was one among the 2% of tests in which trained rats yield score 1.

The superpopulation statistics of a test obviously depends on the setup of the test. One test actually consisting in 100 sessions (that is, when we say “I performed this test once” we mean that 100 session were performed) could have a very sharp histogram, and the results from this test would therefore be very conclusive. The superpopulation statistics also depends on what’s measured in the test. In the examples above we spoke of a “score”, but the result of a test can consist in several quantities at once – for example the total number of successful actions, the times taken to perform each successful action, and so on. In this case the “score” isn’t just one number, but a vector of many numbers; and the corresponding histogram is thus multidimensional, possibly not even representable on paper.

In general, the more complex a test is and the more quantities are recorded during it, the sharper are its superpopulation statistics and reproducibility. Although excessive complexity can sometimes backfire, introducing too many uncontrollable factors, leading to lower reproducibility instead.

But we often don’t have the time or material resources to do complex tests. Then we must rely on unsharp statistics. It is important, however, that we keep as many variables of the test as possible for the statistics, rather than simple summaries such as the average of some quantities recorded during the test. This kind of averaged-scores often throw away much of the information that distinguishes one superpopulation from another, leading to unsharp statistics and much poorer inferences than the test actually offered (see example below).

2 The question

First of all a cursory overview of the problem. We let an untrained rat observe (possibly across several sessions) a trained rat that performs a task. This is repeated for several observing rats. We wonder whether the observers have learned the task, by watching. We try to answer this question by recording the observers’ own performances in a similar task.

The boring remarks of the previous section lead to several considerations about this general problem:

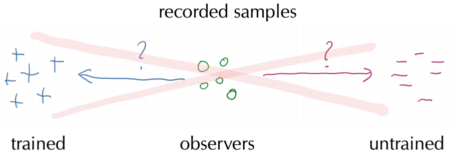
Not individuals, but superpopulations More than in the question “did this rat learn, by watching?” for one or several specific observer rats, we should be interested in the ideal superpopulation statistics of all such observers in the performance of the task. Because we’re interested in the effect of observation upon learning *for rats in general*. Therefore we really want to know how close is the *superpopulation* statistics (the histogram) of observer rats to the superpopulation statistics of trained rats, or to that of untrained rats.

Besides similarities, there could also be interesting small dissimilarities between these superpopulation statistics. For example, the statistics for the observers could be almost overlapping with that of the trained, and yet have slightly different tails.

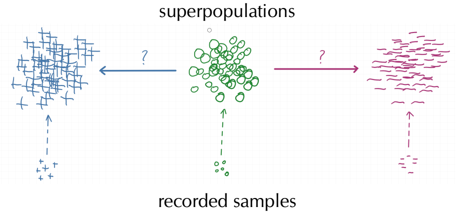
No sample comparisons We therefore should not directly compare the sample (as opposed to superpopulation) statistics of the performed tests. Because the performance of the trained, untrained, and observers could contain many outliers with respect to the superpopulation. This can especially happen if the tested rats are few. We must instead compare their superpopulations' statistics.

Our problem is that we do not know these statistics and must therefore infer them (this is again done by means of de Finetti's 1937 theorem). Here is where our recorded statistics enter: from them we infer the superpopulation statistics. Such a procedure may seem a little roundabout, but it's actually our safeguard against unwarranted inferences.

Direct sample comparison, depicted in this side picture, typically underestimates the uncertainty (statistical variability) of our inferences. The situation gets even worse if our samples contain many outliers with respect to the superpopulations: not only our result will point to the wrong direction, but it will also deceitfully appear to be quite reliable[1].

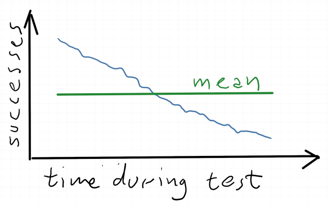


The indirect comparison through superpopulations, depicted in this side picture, takes all sources of variability and uncertainty into account instead: the generalization from our small samples to the superpopulations, and the comparison among the superpopulations. So even if our samples contain outliers and our result point to the wrong direction, the properly reckoned uncertainty will warn us that the general situation might be different. It may also happen that the data are actually good and our results are strong; then they will be even stronger because they appear despite all uncertainties taken into account.

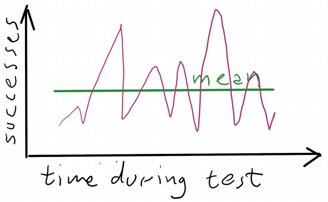


[1] this is one of the sources of failed reproducibility in science, lamented in recent times: Camerer et al. 2018; Klein et al. 2018.

Working with as many observed quantities as possible The performance of a trained rat can decrease during a test, for example owing to fatigue, as sketched in the first side picture. Or the opposite trend can occur if the rat is initially stressed for some unknown reason and the stress diminishes with time.



The casual peaks in performance of an untrained rat should instead appear unsystematic and independent on fatigue or stress, as sketched in the second side picture.



Yet both kinds of trend could have the same mean. So if we used the time-averaged performance we would be throwing away important information that distinguishes trained and untrained rats in this test. For this reason it's important to keep as many recorded quantities (ideally: all) as is computationally possible for the statistics; not just a single-number score.

The "histograms" we obtain by using multiple quantities are not the traditional ones with some score as the x axis and the population percentages as the y axis. They are two-dimensional or multi-dimensional, and therefore not directly representable as a plot – although their marginals, for example, are. But the comparison of such multidimensional superpopulation histograms proceeds, computationally, just an in the comparison of ordinary one-dimensional ones.

Comparison measures of two histograms are, for instance, their relative entropy, or their overlap, or distances such as the "earth mover's distance".

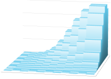
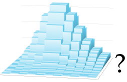
3 Summary

From the discussion above we see that the question we want to ask is slightly different:

> *how similar and different are the skill performances of observer rats (in general) to those of trained rats and of untrained rats?*

and we answer it by comparing the superpopulation statistics – histograms – of the three groups, by means of several kinds of numerical measures.

This is not really a test of alternative hypotheses, such as "have the observers learned? or not?". Yet the comparison and the data may reveal an essential similarity between observers and trained, implying that the observers did learn. And as already remarked the data can also point to interesting minor differences – which wouldn't be unreasonable, since the learning occurred in two different ways. I (Luca) believe that such an approach is more sensible and flexible than a binary- or multiple-hypothesis testing.

The hypotheses in this research appear elsewhere: in our inference of the superpopulation statistics of the three groups. We don't know these statistics and their histograms, so for each group we must ask: "is the superpopulation histogram like this:  ? Or like this:  ? Or. . . ?" (the pictures assume that a test gives us a pair of values). Each such histogram is a "hypothesis". With the probability calculus (de Finetti's theorem again) we can calculate, for each group in turn, the probability of each possible superpopulation histogram given the recorded sample.

So, for each of the three groups, we don't have one definite superpopulation histogram, but a collection of histograms with probabilities attached to them. The comparison measures between the superpopulation statistics of observers, trained, and untrained will therefore also have a probability distribution. This probability distribution tells us how uncertain we are about the degree of similarity of these statistics.

The probability formulae to be used in this approach are formally straightforward. De Finetti's theorem is used to find the probability distribution over the possible superpopulation statistics, given the sample data, for each group. The similarity measure (say, earth mover's distance) between observer rats and trained rats is computed for each possible pair of their statistics. The probability distribution over the similarity degrees is then computed by convolution. The same is done for observer rats and untrained rats.

The numerical implementation of the formulae can be more difficult and can require long computation times, with Monte Carlo sampling.

This will limit the number of quantities that we can effectively use from each test. For this reason a judicious choice of the most informative ones is necessary; focus should be especially on those that can reveal fatigue, stress, and other variability factors.

4 Hypothetical illustrative example

Here is a hypothetical example to visualize the approach just discussed. Suppose that the test yields one score in the range 0–10. From the analysis of the recorded trained, untrained, and observer rats, we obtain three probability distributions over their possible superpopulation histograms. These probability distributions are represented as 100 "representative samples" (samples drawn from that distribution) of the respective histograms in fig. 2. We see for instance that it's very probable that trained rats in general give high scores; but we are unsure whether the score with highest frequency should be 9 or 10 (some probable histograms have a peak at 9, others at 10).

Notwithstanding our uncertainty in the three superpopulation statistics, it's clear that the statistics of observers and trained are very similar, and that those of untrained rats are very dissimilar for the other two groups. This is better shown by superimposing the plots, as in fig. 1.

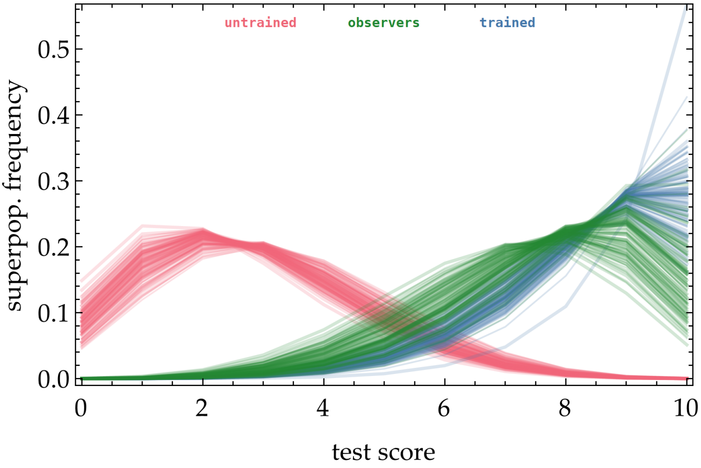


Figure 1

This plot also shows some minor differences from observers and trained, such as a probably fatter tail of the observers' statistics.

We can try to summarize the similarity of two superpopulation statistics by a single measure.

One example is the Hellinger distance[2], which roughly speaking measures the overlap of two distributions. It is equal to 0 when the distributions are exactly the same, and to 1 when the distributions have no points in common (so that with a single test we can say with certainty which population a rat belongs to).

Another example is the earth mover's (or Kantorovich-Vasershtein) distance[3], which measures how far apart the bulks of the two distributions are: from 0 when they are exactly the same, to a maximum of 10 (in our specific case) when they are both narrow and as far apart as possible.

If we knew the superpopulation histograms of the three groups we would have, for either measure, one exact value between observers and trained, and one between observers and untrained. But since we only have a probability distribution over the possible superpopulation histograms, what we obtain is a probability distribution over the possible true value of the similarity measure.

Such probability distribution in our hypothetical example is shown in fig. 3 for both distances. The medians for the Hellinger distance between observers & trained is 0.02; observers & untrained, 0.55; trained & untrained (for comparison), 0.65. For the earth mover's distance the medians are: observers & trained, 0.6; observers & untrained, 4.8; trained & untrained, 5.5. The ranges with 95% probability are reported in the figure. Clearly the statistics of observers and trained are very surely quite similar; and the statistics of observers and untrained are, again very surely, almost as dissimilar as those of trained and untrained.

The similarity measure and its probability are very useful as a simple quantitative and visual summary of the main result: observers can be said, with extreme certainty, to have learned. This measure and the probability distribution for it can be obtained and simply visualized no matter what the dimensionality of the histograms is. Even for a test yielding many recorded quantities the final plot would look like fig. 3.

[2] https://encyclopediaofmath.org/wiki/Hellinger_distance

[3] https://encyclopediaofmath.org/wiki/Wasserstein_metric

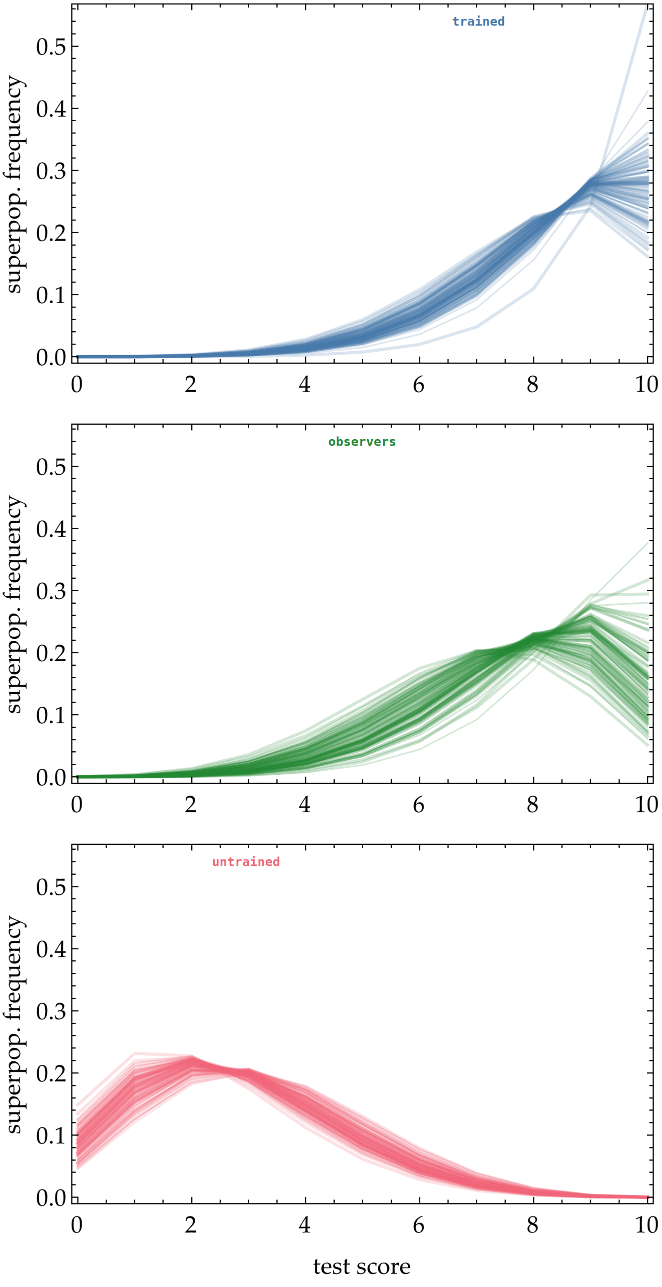

trained
superpop. frequency
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
observers
superpop. frequency
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
untrained
superpop. frequency
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
test score

Figure 2

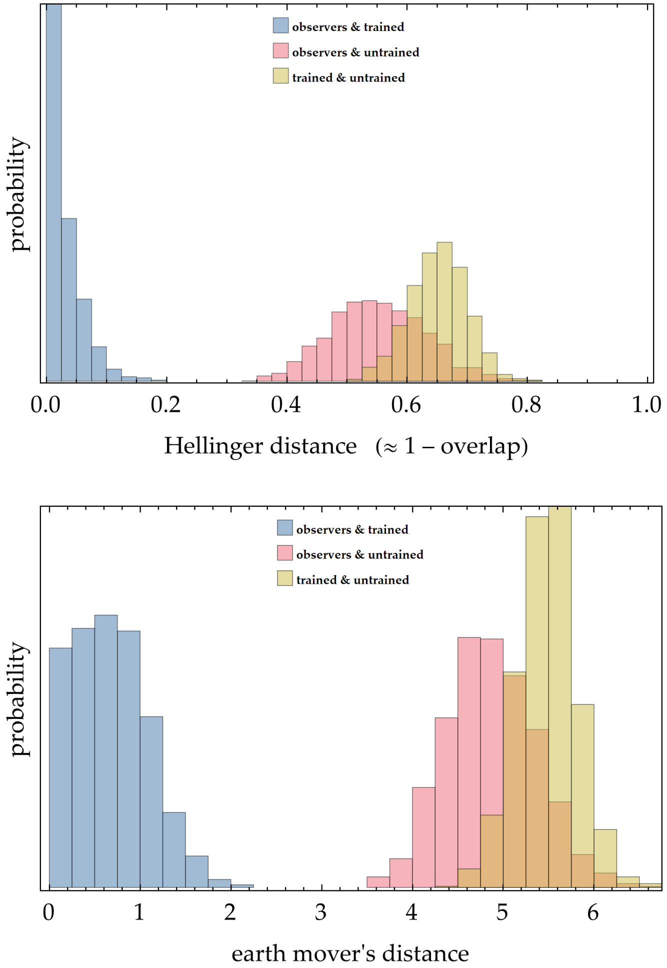


Figure 3 There's a 95% probability that the Hellinger distance between observers & trained is in the range 0–0.10; between observers & untrained in 0.40–0.71, between trained & untrained (for comparison) in 0.55–0.74. For the earth mover's distance the 95% intervals are: observers & trained, 0–1.5; observers & untrained, 3.9–5.9; trained & untrained, 4.8–6.1.

The samples of the superpopulation histograms are interesting for a deeper analysis – for example if we want to study how learning by observation may differ from learning by direct training. If the test yields more than two recorded quantities, the corresponding multi-dimensional histograms cannot be visualized. But we can always visualize, as in fig. 1, the marginal histograms for the individual quantities. This gives us a partial picture of the multi-dimensional situation (which may hide important multidimensional features, however).

Bibliography

(“de *X*” is listed under D, “van *X*” under V, and so on, regardless of national conventions.)

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., et al. (2018): *Evaluating the replicability of social science experiments in* Nature *and* Science *between 2010 and 2015*. Nat. Hum. Behav. **2**^{9}, 637–644.

de Finetti, B. (1937): *La prévision: ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**^{1}, 1–68. Transl. in Kyburg, Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

Diaconis, P., Holmes, S., Montgomery, R. (2007): *Dynamical bias in the coin toss*. SIAM Rev. **49**^{2}, 211–235. http://statweb.stanford.edu/~susan/papers/headswithJ.pdf.

Klein, R., Vianello, M., Hasselman, F., Adams, B., Adams Jr., R. B., Alper, S., Aveyard, M., Axt, J., et al. (2018): *Many Labs 2: investigating variation in replicability across sample and setting*. Adv. Meth. Pract. Psychol. Sci. **1**^{4}, 443–490.

Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.