There's no 'sampling bias' for the mutual information

P.G.L. Porta Mana
<piero.mana@ntnu.no>

C. Battistin
<claudia.battistin@gmail.com>

Draft of 1 April 2019 (first drafted 31 March 2019)

This note shows that the so-called 'sampling bias' in the estimation of the mutual information for stimulus-response frequencies doesn't exist. If the estimate from the sample forecasts a high mutual information, then the response is indeed very likely to be informative. On the other hand, if we expect the response to be uninformative or if we just want to be conservative, a correct calculation of the estimate leads to a negligible mutual information. A correct frequentist analysis also shows that the estimator is unbiased. No corrections of any kind are needed.

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

1 Bias?

The mutual information between the long-run relative frequency of a signal and that of a response is a measure of how much our uncertainty about the signal is reduced by knowledge of the response. This measure is sometimes used in neuroscience, the response being some characteristic – such as the activity or the firing rate – of a neuron or of a network of neurons.

Recent works in neuroscience claim that our estimate about the mutual information for long-run frequencies from a small sample is biased; that is, in frequentist terms: the expected value of its sample estimate is different from its 'true' value. Other works have proposed various corrections to this alleged bias.

In this note we show three main results:

1. The direct estimation of the 'long-run' mutual information from a sample is reliable and doesn't need any kind of correction. On the one hand, if we are equally uncertain about the long-run response frequencies, and the estimate from a small sample suggests a large mutual information, then we should indeed expect the response to be informative. On the other hand, if we initially suspect that the long-run response frequencies should be uninformative, or if we are very conservative in our inference, then a correct calculation consistently leads to very low estimates of the mutual information.

2. For small samples, our estimate crucially depends on the pre-data probability for the long-run relative frequencies of the response conditional on the stimulus. In making such estimates we must therefore give some thought to assessing this pre-data probability, rather than correcting non-existing biases.

3. A correct frequentist analysis proves that the estimator for the long-run mutual information is unbiased.

The main result is first shown in the next section with a straightforward calculation, and explained intuitively in the subsequent section. The calculations use the example given by Panzeri et al. (2007 Fig. 1). The final section briefly discusses the third result and the importance of the pre-data probability in our estimations when the sample is small. Some ways to assess this probability are also discussed.

2 Bayes

First of all let's state what our inference is about. Given a sample of stimulus-response data we want to assess what's the most probable set of long-run[1] relative frequencies of the response conditional on each stimulus value, and from these assess what's the most probable value of the associated mutual information between stimulus and response. We assume that all stimuli appear with equal relative frequencies.

Let the stimulus s have two possible values $\{-,+\}$, and the response r ten possible values $\{1,\dots,10\}$. Let the data D be a set of n stimuli $-$ which yielded n responses (r_i^-), $i \in \{1,\dots,n\}$, and n stimuli $+$ which yielded n responses (r_i^+). The ten response state appeared with relative frequencies $\boldsymbol{q}^- := (q_r^-)$, $r \in \{1,\dots,10\}$, for the stimulus $-$, and with relative frequencies $\boldsymbol{q}^+ := (q_r^+)$ for the stimulus $+$.

We can use the concrete data summarized in fig. 1, consisting in $n = 20$ samples per stimulus.

If the long-run frequencies conditional on stimulus $-$ are $\boldsymbol{f}^- := (f_r^-)$, and conditional on $+$, $\boldsymbol{f}^+ := (f_r^+)$, then out of symmetry the probability of obtaining the data D is

$$\mathrm{p}(D \mid \boldsymbol{f}^-, \boldsymbol{f}^+, K) = \prod_r \Bigl[(f_r^-)^{n q_r^-} \; (f_r^+)^{n q_r^+} \Bigr]. \tag{1}$$

[1] 'But this *long run* is a misleading guide to current affairs. *In the long run* we are all dead.' (Keynes 2013 § 3.I, p. 65)

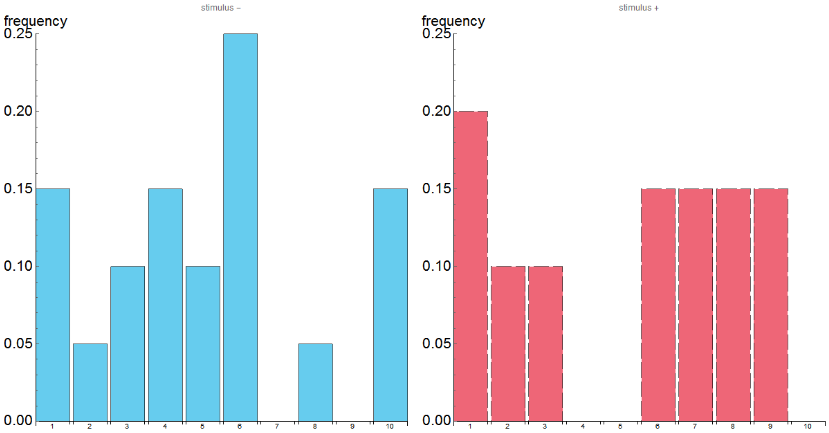

Figure 1

This is also the *likelihood* of the long-run frequencies in view of the sample. Their probability density is proportional to the likelihood, corrected by their initial probabilities $\mathrm{p}(f^-, f^+ \mid K)$

$$\begin{aligned} \mathrm{p}(f^-, f^+ \mid D, K) \propto \mathrm{p}(D \mid f^-, f^+, K)\, \mathrm{p}(f^-, f^+ \mid K) &= \\ \mathrm{p}(f^-, f^+ \mid K) \prod_r \Bigl[(f_r^-)^{nq_r^-}\ (f_r^+)^{nq_r^+}\Bigr]&. \end{aligned} \tag{2}$$

We can calculate this probability analytically when possible, or estimate it with Monte Carlo sampling. From such samples we can also estimate the probability distribution of the long-run mutual information

$$I := \sum_r \tfrac{1}{2} f_r^-\ \ln\!\left(\frac{\tfrac{1}{2} f_r^-}{\tfrac{1}{2} f_r^- + \tfrac{1}{2} f_r^+}\right) + \sum_r \tfrac{1}{2} f_r^+\ \ln\!\left(\frac{\tfrac{1}{2} f_r^+}{\tfrac{1}{2} f_r^- + \tfrac{1}{2} f_r^+}\right). \tag{3}$$

Let's consider two possible probabilities for the long-run conditional frequencies:

2.1 Uniform uncertainty about the frequencies

If we initially think that equal ranges $(\triangle f^-, \triangle f^+)$ of pairs of conditional frequencies are equally possible, then

$$\mathrm{p}(f^-, f^+ \mid K_\mathrm{u}) = 1. \tag{4}$$

Let's sample 5 000 pairs of conditional frequencies from the density (2) obtained from this probability and our example data. The resulting distribution of their associated mutual information is shown in fig. 2. It tells us that the response is likely to be informative; the most likely values of the long-range mutual information are around 0.2 bit. The next section explains intuitively why this inference is correct.

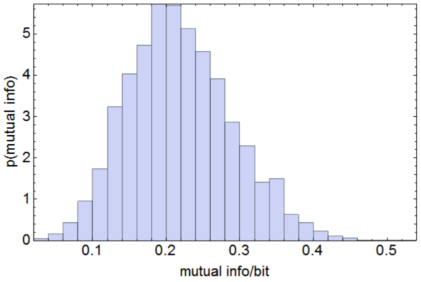

Figure 2

2.2 Conservative uncertainty about the frequencies

If we initially think that the long-run response frequencies conditional on the stimuli should be very similar, or if we simply wan to do a conservative estimate, then the initial probability will be higher for pairs with similar conditional frequencies; for example 🧩 replace with combination of Dirichlet – same effect

$$\mathrm{p}(f^-, f^+ \mid K_\mathrm{c}) \propto \exp\left[\frac{\sum_r (f_r^- - f_r^+)^2}{2\sigma^2}\right]. \tag{5}$$

This density states that the two conditional frequencies should be roughly equal, but otherwise leaves a uniform uncertainty about the values of each. Smaller values of σ represent more conservative estimates.

A sample of 5 000 pairs of conditional frequencies from the density (2) with the conservative initial density (5) 🧩 specify σ leads to the distribution of mutual information of fig. 3. The estimate now says that we should expect a negligible mutual information, with a most probable value around 100 times smaller than in the previous case.

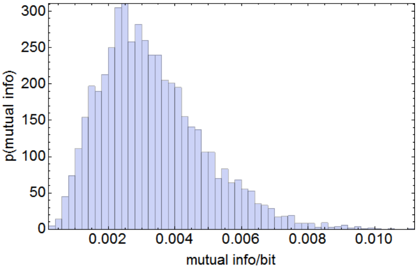

Figure 3

3 Because

Let's try to understand why the estimate of § 2.1, fig. 2, is reliable; and to understand what happens in the conservative case of § 2.2.

If we deem all pairs of conditional frequencies equally possible, let's sample 5 000 pairs uniformly. The scatter plot of the two conditional frequencies for the response value 1 is shown in fig. 4.

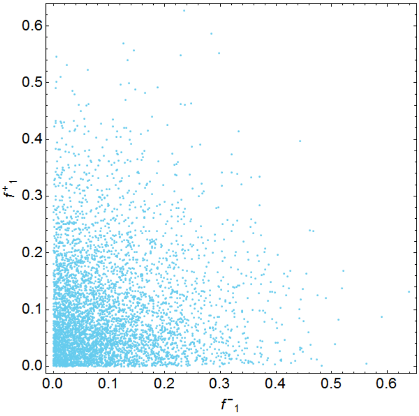

Figure 4

Let's now plot the samples showing the probability that each assigns to our observed data on the horizontal axis, and their associated mutual information on the vertical axis. We obtain the scatter plot of fig. 5. The points of all three sizes and colours are part of the plot. The pair consisting of two uniform conditional frequencies (1/10 probability

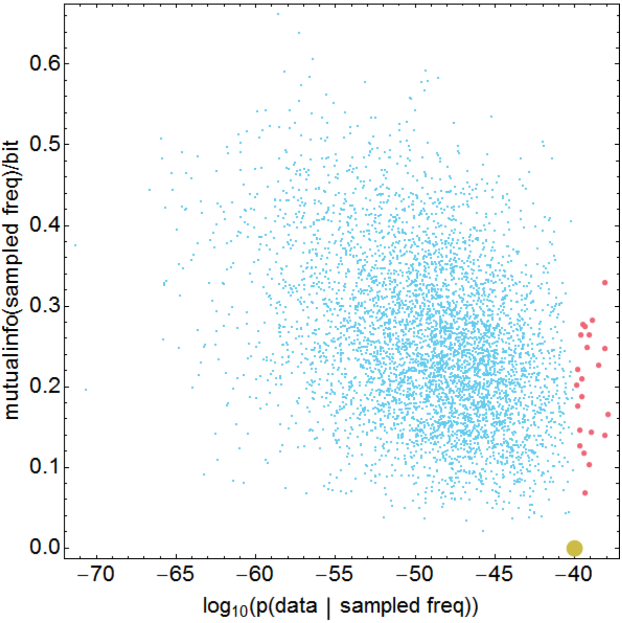

Figure 5

for each response) is the largest, yellow point. This pair of long-run conditional frequencies assigns probability 10^{-40} to the data and has zero mutual information.

It's clear from this scatter plot that the data strongly suggest a large estimated mutual information, for two reasons:

1. The pairs that assign very low probability to our data, say less than 10^{-40}, are represented by small blue points. Each one very unlikely to be the continuation of our data. But at the same time they constitute the overwhelming majority of samples, and therefore there is a non-negligible probability that the data come from one of them. Most of them have high mutual information.

2. The pairs that assign high probability to our data, 10^{-40} or more, are represented by the larger red points and the largest yellow point. The majority of them also have high mutual information. In fact, the pair of uniform conditional frequencies is an outlier: it's very unlikely that our data come from it, compared with the other possibilities.

The estimate of fig. 2 is therefore quite correct and reliable, not biased at all.

Now consider the conservative initial density (5). The scatter plot of 5 000 of the two conditional frequencies for the response value 1 is shown in fig. 6. It shows that the two frequencies are very similar to each other and close to 1/10, the probability given by the uniform distribution.

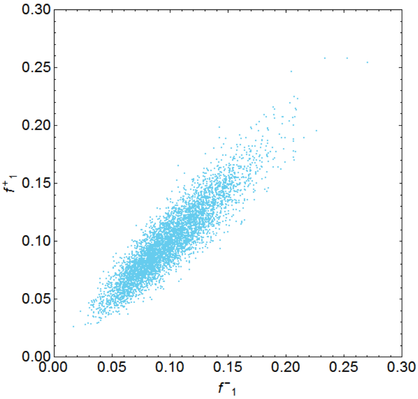

Figure 6

Figure 7 shows the samples plotted as in the previous case. There are now many pairs that assign roughly the same probability to the data as the uniform-distribution pair does. The majority of all samples a negligible mutual information. This is reflected by the estimate of fig. 3.

4 Brief

We have found that the estimates of the mutual information from a small sample have no bias whatsoever, and we've seen intuitively why they are correct. The crucial point is the specification of the initial joint probability for the long-run conditional frequencies. This specification is especially important if the data are few. It's possible to specify a probability that express a conservative guess, and the resulting estimate of the mutual information is very close to zero.

This conclusion isn't surprising: the main point is the same as, for example, in the inference of diseases or some phenotypes from genetic peculiarities and vice versa. Imagine that we know the relative frequencies of two gene variants among people who have a particular disease, and the relative frequencies among people who don't have the disease. Now we ask: given that a specific person has one of the gene variants, what's the probability that this person has the disease? The answer is that in general we don't know until we specify the incidence rate of the disease in the full population. Because even if that gene variant appears more frequently among people with the disease, the number of people having such disease may be so small that the probability still favours the person's being healthy.

Note that, even though we used Bayes's theorem in § 2, the general conclusion that there's no bias would also be reached by a serious frequentist statistician. The erroneous conclusions reached in papers claiming a sampling bias for mutual information (Panzeri et al. 2007) may be summarized thus: they confuse a *non-parametric* problem (see e.g. Wasserman 2006, a frequentist book) with a parametric one. The problem is not the estimation of a parameter that identifies a distribution in a family, but the estimation of the distribution itself: the whole distribution – or distribution*s* in our case – is the 'parameter'. The mutual

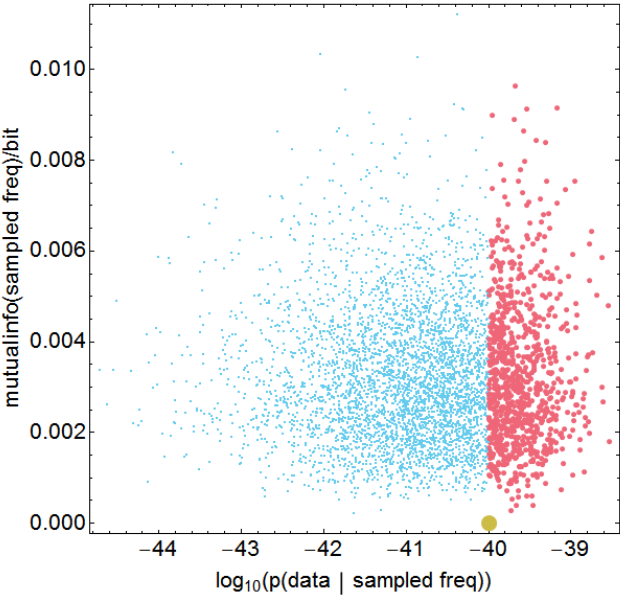

Figure 7

information is a function of this 'parameter'. The procedure of Panzeri et al. (2007 pp. 1065–1066) is therefore wrong.[2] To assess the expectation of the estimator we must sample the 'parameter'; in our case this means drawing N samples of the empirical conditional frequencies in n data and then taking their average to estimate the long-run frequency. *Then* we compute the mutual information from this estimate. If we do this with $N = 5\,000$ and $n = 20$ for a pair of uniform conditional distributions we obtain a mutual-information estimate of the order of 10^{-5}. As N increases, with $n = 20$ fixed, the estimate goes to zero. There is no bias.

The topsy-turvy point of view typical of frequentist-like analyses unfortunately leads to this kind of oversights. Our problem is not to infer data from a specific known frequency, but to infer an unknown frequency *among several possible ones* from known data – because our uncertainty about the long-run mutual information is a consequence of our uncertainty about the long-run frequency.

🧩 Paragraph on shrinkage

Baraka

PGLPM thanks Mari, Miri, & Emma for continuous encouragement and affection; Buster Keaton and Saitama for filling life with awe and inspiration; and the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

Bibliography

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

Keynes, J. M. (2013): *A Tract on Monetary Reform*, repr. of second ed. (Cambridge University Press, Cambridge). First publ. 1923.

Panzeri, S., Senatore, R., Montemurro, M. A., Petersen, R. S. (2007): *Correcting for the sampling bias problem in spike train information measures*. J. Neurophysiol. **98**^{3}, 1064–1072.

Wasserman, L. (2006): *All of Nonparametric Statistics*. (Springer, New York).

[2] Note in passing that plotting the distribution of a parameter or of a function thereof, as in Panzeri et al. (2007 Fig. 1, right columns), is anathema from a frequentist point of view: parameters do not have distributions, by divine Fisherian decree.