The ‘sampling bias’ is not a bias

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

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

It is shown that the so-called ‘sampling bias’ in the estimation of the mutual information for stimulus-response frequencies is not a bias at all. If the estimate from the sample forecasts a high mutual information, then we 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.

In either case 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 subject to a bias, a systematic error. Other works have proposed various corrections to this alleged error.

In this note I show that no such bias exists, and the estimates are reliable. 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 leads to very low estimates of the mutual information. As a consequence no corrections of any kind are needed for the estimates.

I show this first with a quick calculation in the next section, and then with a longer reasoning in §***. The final section discusses what kind of incorrect reasoning leads to the idea that there’s a bias.

The calculations use the example given by Panzeri et al.(2007 Fig. 1): the stimulus can assume two values, the response ten values.

Comment on 'sampling bias': it has a different meaning in statistics

2 Bias? No: 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 the 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 $\boldsymbol{q}^- \coloneqq (q_r^-)$, $r \in \{1,\dots,10\}$, for the stimulus $-$, and with relative frequencies $\boldsymbol{q}^+ \coloneqq (q_r^+)$ for the stimulus $+$.

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

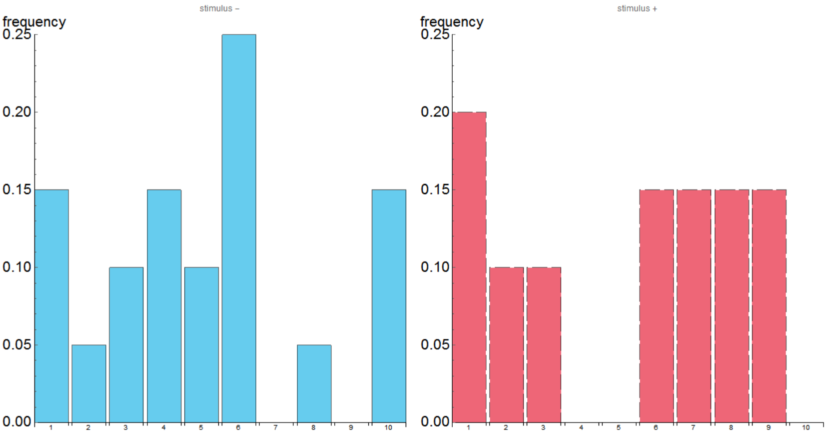

Figure 1

[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)

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

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

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 $\mathrm{p}(\boldsymbol{f}^-, \boldsymbol{f}^+ \mid K)$

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

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

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

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 $(\triangle \boldsymbol{f}^-, \triangle \boldsymbol{f}^+)$ of pairs of conditional frequencies are equally possible, then

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

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.

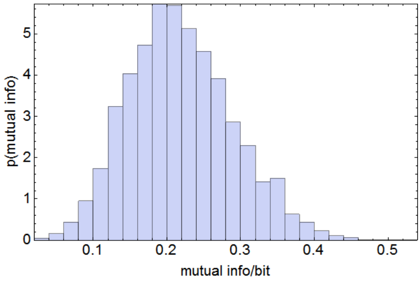

Figure 2

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

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

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

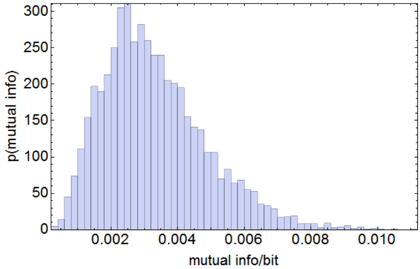

Figure 3

we should expect a negligible mutual information, with a most probable value around 100 times smaller than in the previous case.

3 Intuitive understanding

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.

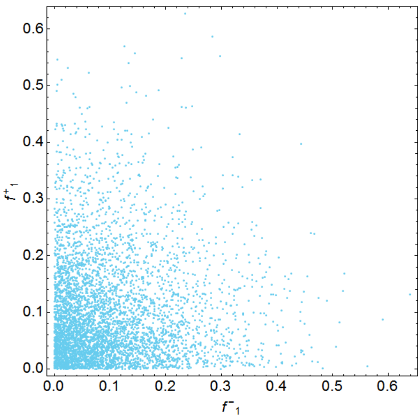

Figure 4

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 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

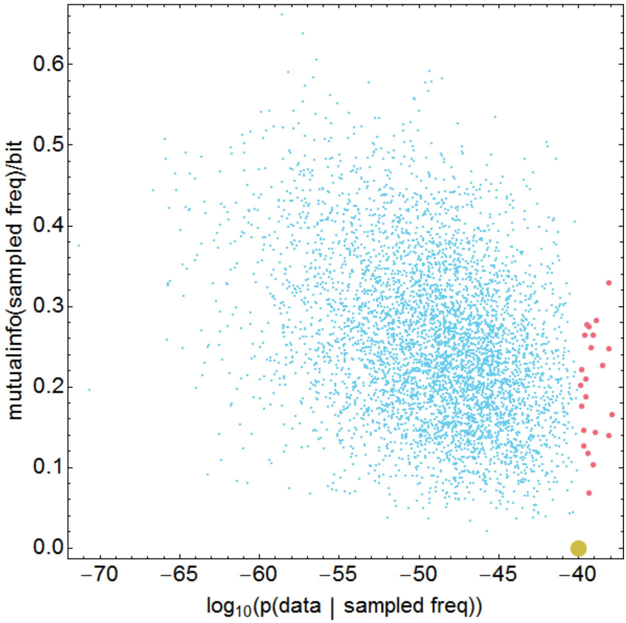

Figure 5

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.

Figure 7 shows the samples plotted as in the previous case. There are now many pairs that assign roughly the same probability to the

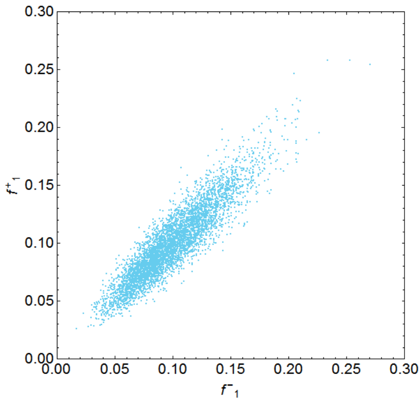

Figure 6

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.

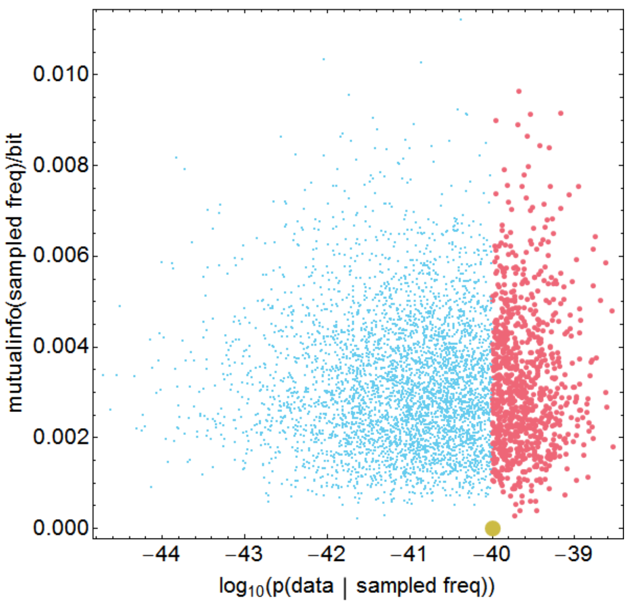

Figure 7

4 Discussion

Thanks

. . . to Mari & Miri for continuous encouragement and affection, and to Buster Keaton and Saitama for filling life with awe and inspiration. To 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. A wrong analysis that leads to wrong conclusions.