A Response to the "Mea Culpa": Mathematical Analysis

Prahlad Balaji Iyengar

Disclaimer: By no means do I claim my analysis below to be complete.

1 Introduction

During Dr. Hosoi's follow-up with the students after her publication of the "Mea Culpa," I presented a sketch of a mathematical critique. Dr. Hosoi suggested that I submit it to the faculty, so I have developed it a bit further below.

For the analysis below, let $x \in \mathscr{X}$, $y \in \mathscr{Y}$ be random variables representing the true number of voters and the attendance population, respectively. Here $\mathscr{X}$ and $\mathscr{Y}$ are the sample spaces, in this case the positive integers at least equal to 30 (which I believe to be the quorum for the meeting). Then let $p_{X|Y}(x|y)$ be the posterior distribution of attendance vs. voters, and let $p_X(x)$ be the prior distribution of voters in each meeting. We are trying to extract a good guess for x based on partial data we have about y, both at the February meeting and at prior meetings.

Recall that the analysis hinges on the use of the "mean" average of the last 10 voting participation records as a reasonable estimate for the expected number of voters in this faculty meeting. From this is subtracted the actual number of "yes" votes to find that the number of "no" votes would have been insufficient to overturn the outcome. I claim that this analysis is faulty on three different levels.

2 Estimators

First, let us question the use of the mean as an estimator. In general, the mean is known to be an unbiased estimator - i.e., the following relation holds: $\mathbb{E}_{P_Y}[\hat{x}(y) - x] = 0$ where the estimator $\hat{x}(y)$ is the mean. However, it is not the case that any unbiased estimator is always an efficient estimator - in fact, the only estimator which is guaranteed to be efficient, if an efficient estimator exists, is the maximum likelihood estimator (MLE) given by:

$$\hat{x}_{ML}(y) = \arg\max_{x \in \mathscr{X}} p_Y(y;x)$$

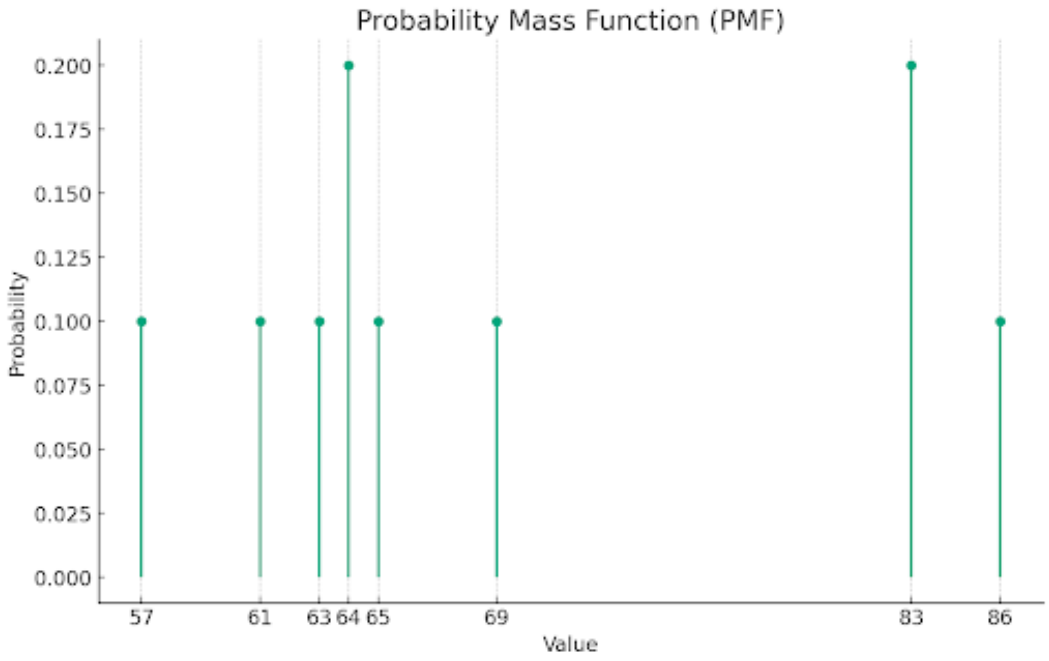
We write it in this parametrized way, rather than $p_{X|Y}(x|y)$, because we may not be in a situation which has a prior distribution over $\mathscr{X}$. The MLE estimator is in general efficient when an efficient estimator exists, but the mean is not. As a trivial example, let us consider the case where you ask a small child to pick a number from 0 through 9. Suppose the child wants to play a little game, and each time you ask, they alternate between picking either 1 or 3. In the limit, the mean of this distribution over the integers between 0 and 9 is indeed 2. However, depending on your cost function, that may not be an appropriate estimate for the next value, given that we know this distribution has a particular shape. For example, in

the binary setting where guesses are either “right” or “wrong”, the best guess is either 1 or 3. The choice of the mean as an appropriate estimator for this problem makes a particular assumption about the underlying distribution. This brings me to my next point.

3 Gaussians

Second, let us note the abuse of the Gaussian. Accompanying the analysis is a neat image of a Gaussian distribution, shading in the probability of overturning the outcome, to lend credibility. It is true that, in the case of an underlying Gaussian distribution, the mean and MLE converge. This is also true as we expand the distribution ad infinitum and take means of various samples (via the central limit theorem). But the graph provides a candidate for the raw number of votes, and suggests that this distribution is itself Gaussian. This is used to justify conducting a power-test of the hypotheses to show that the vote would have been unlikely to have been overturned. I claim this is an abuse of the Gaussian.

First, note that the actual distribution of the data provided is shown below. The graph is not obviously Gaussian, and instead looks vaguely bimodal. I will explain why I think this is a reasonable model for this data.



When the faculty vote on matters that concern them very little, I would suspect there to be low voter participation. Similarly, I would expect high voter participation on matters that concern them very greatly. Even this naive first-pass attempt at contextualizing the data would yield a bimodal distribution for voter participation - one peak for the “less interesting” votes, and one for the “more interesting” ones. As evidenced by my toy example in the previous section, using the mean as an efficient estimator for such a bimodal distribution would be erroneous.

So why do statisticians use the Gaussian so frequently? Well, they do use the Gaussian, but they use it to model statistics about the data, not the underlying distribution for the data itself. The point of the central limit theorem is to show that, if we take decently large samples from any (reasonable) distribution and compute, for each sample, some unbiased statistic (e.g., the mean), then the distribution of *those statistics* will converge to a Gaussian in the limit. This requires a bunch of data, and definitely does not mean that the underlying distribution of voters is Gaussian. Thus, this abuse of the Gaussian betrays, at best, a misunderstanding of the role of Gaussian distributions in statistics and, at worst, a deliberate manipulation of the reader (since the actual data, as plotted, is nowhere near Gaussian, and contains way too few points to be able to reliably conclude anything). I suspect that the idea was to take the mean of the data as the estimator (an erroneous step, as outlined in the previous section), and therefore the analysis invokes the the Gaussian prior $p_X(x)$. Regardless, this is a wholly unjustified assumption without fidelity to the data. But what exactly is the data representing?

4 Data

Third, let us question the data that was used. The data itself is not conducive to a rigorous analysis. If we wanted to get a distribution of voter participation, we ought to have used the percentage of voters, normalized to the attendance at each meeting at the time of the vote. The analysis rightly points out the difficulties in estimating the number of participants present at the time of the vote. But even if we assume, as written, that 95-105 people are present usually, using the raw number of voters rather than the percentage invalidates any real statistical analysis we can perform. After all, we are not drawing from a stable population! Recall that we are interested in the posterior distribution $p_{X|Y}(x|y)$ such that we can extract a guess for the number of voters given the total attendance. We cannot produce this guess without properly taking into account the attendance at each session, yet that data is nowhere to be found. To illustrate the issue, note that the two peaks in the empirical distribution occur at 64 and 83 - but $\frac{64}{105}$ and $\frac{83}{95}$ is a very different story from $\frac{64}{95}$ and $\frac{83}{105}$. So it would appear that, even if we were to accept these methodologies, deconstructed and criticized above, this analysis cannot extend to the data as provided.