Likelihood Ratios and Large Deviations
A Not Surprising Connection

Carlos Gonzalez

May 2, 2024

1 Introduction

Let $X_1, X_2, \ldots, X_n$ be a sequence of n iid draws from a random variable $X \sim F_X$ such that $\mathbb{E}[X] = \mu$, and $Var(X) = \sigma^2$. Let $\hat{X}_n = \frac{1}{N}\sum_i^N X_i$. Large deviations theory is usually interested in **tail probabilities**, which are objects of the following form,

$$\mathbb{P}(\hat{X}_n \geq \mu + \varepsilon) \tag{1}$$

for some $\varepsilon > 0$. More especifically, this is a tail probability on the sample average of X. Our goal is clear, we would like to understand how big it can be. An intuitive approach is then to bound this guy, leveraging the statistical properties of X itself. For example, if I know that X is such that

$$\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a \qquad \text{(Markov Inequality)} \tag{2}$$

then I might be able to say something about the tail probability of its sample average. This seems not only like a rational approach to the problem, but also the unique valid approach. At the end of the day, it looks like only assumptions on X will be able to buy us anything when bounding its sample average. Or equivalently, any assumption that we impose in the sample average must have a 1-to-1 mapping to an assumption on X.

In this piece, we explore an alternative set of assumptions and show a surprising connection between tail probabilities and Likelihood Ratio (LR) tests. Wait, what? How tests gonna help us here? They are nothing related to tail probabilities. Get some

popcorn and follow me!

2 Statistical Properties of X

One can bound objects like the one in Equation 1 in many different ways. Most of them, the ones you are probably familiar with, exploit the properties of X itself. This is the case of the Markov Inequality presented above, which creates a bound on the tail probability of X by relying only on the non-negativity of X. Stronger assumptions can get stronger bounds on the tail probabilities. Hoeffding Inequality, for instance, exploits the boundness of X_i to derive sharper bounds.

Similarly, the Cramér-Chernoff bound derives an interesting bound by assuming σ-subgaussianity of X. Remember that a random variable Z is σ-subgaussian if for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[\exp(\lambda Z)] \leq \frac{\exp(\lambda 2\sigma^2)}{2} \tag{3}$$

This property can then be used to derive a bound on the tail probability of X

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda\varepsilon) \overset{\text{MI}}{\leq} \frac{\mathbb{E}[\lambda X]}{\exp(\lambda\varepsilon)} \leq \frac{\exp(\lambda 2\sigma^2)}{\exp(\lambda\varepsilon)} \overset{\lambda=\varepsilon/\sigma^2}{=} \exp\Big(-\frac{\varepsilon^2}{2\sigma^2}\Big)$$

Some additional calculations and tricks on these results allow us to recover bounds on the tail probabilities of sample averages. This is the case of the Chebyshev Inequality (a well-known corollary to Markov) or the subgaussian sample average concentration inequality (a corollary to Cramér-Chernoff):

$$\mathbb{P}(\hat{X} \geq \mu + \varepsilon) \leq \exp\Big(-\frac{n\varepsilon^2}{2\sigma^2}\Big) \tag{4}$$

Let's stop here and make a short recap. Everything which has been done so far just makes intuitive sense. We have bounded the tail probability of an event concerning $\hat{X}$ leveraging the tail properties of X, given that $\hat{X}$ is a function of X (surprise!?). The stronger the assumptions on X, the tighter the bounds on its tail probabilities, and the tighter the bounds on the tail probabilities of its sample averages.

If you are reading this article, (i) you were probably already familiar with these sensible notions, and (ii) you still don't understand how on earth statements about the tail probabilities can be connected to Likelihood Ratio (LR) tests.

3 Likelihood Ratios

Just to make things more confusing, let's refresh what LR tests are, and how unconnected they are to tail probabilities. First of all, a LR test is a test. This means that it is a decision rule over two statements or hypothesis. In particular, it is a rule which tells you when to reject or not reject the null-hypothesis (the first statement) against the alternative (the second statement). So far, nothing to do with tail probabilities, great! Let's take a look at hypotheses tests over the population means, i.e.

$$H_0 := \mu \leq 0, \qquad H_1 := \mu > 0 \tag{5}$$

Given some data realizations of X, we define the LR test as a decision rule which rejects H_0 if and only if

$$\lambda_{LR}(X) := \frac{\sup_{\mu:\mu\leq 0} \mathcal{L}(\mu \mid X)}{\sup_{\mu} \mathcal{L}(\mu \mid X)} = \frac{\mathcal{L}(\mu_0 \mid X)}{\sup_{\mu} \mathcal{L}(\mu \mid X)} \leq c \tag{6}$$

for some $c \in [0, 1]$, where $\mathcal{L}$ is the likelihood function and the second equality serves as definition of μ_0. In words, the LR-test is a decision rule which tells you to reject the null-statement if the probability that the data X has been generated by a $\mu \in (-\infty, 0]$ is not very small compared to the probability that X has been generated by some $\mu \in \mathbb{R}$.

A simple observation now reveals that $\sup_{\mu} \mathcal{L}(\mu \mid X) = \hat{X}$. It is most likely that some data with empirical average $\hat{X}$ has been generated by a distribution with expectation $\mu = \hat{X}$. So, in fact, $\lambda_{LR}(X) = f(\mu_0, \hat{X}, X)$. If we take the data as given, we can just write $f(\mu_0, \hat{X}, X) = d(\mu_0, \hat{X})^{-1}$, where d is frequently referred to as the KL divergence between distributions $F_{\mu_0}, F_{\hat{X}}$. Intuitively, d is telling how different two distributions governed by parameters μ_0 and $\hat{X}$ are. A second observation is that by design $\mu_0 \leq \hat{X}$.

$d(\cdot, \cdot)$ has a convenient property. It is monotonic in the distance between its arguments. And, in particular, because $\mu_0 \leq 0$, the more positive $\hat{X}$ is, the larger the distance between $\hat{X}$ and μ, the higher d, and the smaller λ_{LR}. This implies that a large

distance between $\hat{X}$ and μ_0 is equivalent to a higher probability of rejecting the null using a LR-test.

4 Connection

Hey, but now we are almost there! We have made big progress. We have shown that although originally the LR-test looked like a complicated fraction of likelihoods, when dealing with statements about the population means in one-dimensional spaces, they are just "distances" between distributions governed by a candidate μ_0 and a distribution with mean $= \hat{X}$. In fact, monotonicity of d tells us that there must exist a point $a > \mu_0$ (aka $a = \mu_0 + \varepsilon$) such that if $\hat{X} \geq a$, then we reject the null-hypothesis. Wait, wait, wait... Are you saying that

$$\mathbb{P}(\text{Reject } H_0) = \mathbb{P}(\lambda_{LR} \leq c) = \mathbb{P}(d(\hat{X}, \mu) > 1/c) = \mathbb{P}(\hat{X} \geq a) = \mathbb{P}(\hat{X} \geq \mu + \varepsilon)$$

So, are you saying that the tail probability of $\hat{X}$ equals the probability of rejecting the null, aka rejecting the fact that data has been generated by μ given $\hat{X}$? Yep, that is exactly what I am saying.

Let us stop one last time to analyze why this is important, but also surprising. It is surprising, because *a priori* we have connected two unconnected notions, the probability of an event involving sample averages, and a decision rule about population means. Moreover, this result is important because distances can factor in properties about the tail probabilities of sample averages of X, without imposing further assumptions on the tail properties of X.

For instance, when analyzing Bernouilli data the following result can be derived

$$\mathbb{P}(\hat{X} \geq \mu + \varepsilon) \leq \exp(-nd(\mu + \varepsilon, \mu)) \tag{7}$$

This bound is potentially much sharper than that derived in Equation 4. This is because the notion of distance exploits information about the average of the distribution under consideration, and not just the tail probability of the underlying distribution (like Equation 4 did). For instance, in the case of Bernouilli, distance based bounds exploit the fact that when the average is close to 0 or 1, the variance is small, so bounds can be sharpened asymmetrically around 0 and 1. As you can see, this intuition is only captured at the $\hat{X}$ level, and not the X level.

What a ride this was!

References

[1] Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE, 2004.

[2] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[3] Friedrich Liese and Klaus-J Miescke. Statistical decision theory. In *Statistical Decision Theory: Estimation, Testing, and Selection*, pages 1–52. Springer, 2008.