STATS 50: Mathematics of Sport **Spring 2019**

Week 3 – Population Variance and Statistical Significance

Lecturer: Maxime Cauchois

Warning: *these notes may contain factual errors*

1 Motivation for our study

In most sports, numbers are reported in terms of average per game performance. For instance, in the NBA, a player X will be considered a more proficient scorer than a player Y should he average more points per game, or has a better shooting percentage.

While this statistic clearly gives an indication on a player's true ability for scoring, it does not tell the whole story, because it does not even take into account players' shooting percentage.

But the truth is that those numbers may contain inherent flaws which make them unreliable, for several reasons.

First, in the case of a player who attempted only 1 FG and made it, or actually played only one game and score 30PTS, his shooting percentage will be no less than 100% in the first case, and his average number of points per game will be 30! Of course, you can feel these number are not reliable estimators of a player's true level, and generally in its rankings, the NBA solves this problem by only considering people with at least a fixed number of attempts. Our focus instead will be to give a more credible prediction, and we'll describe a method called Empirical Bayes Estimation to that extent.

Now, suppose that we have an estimation of a player's shooting percentage that we deem trustworthy and accurate. It does not tell us the whole story yet, because this value may still yield a poor prediction for the next game played. Let us consider the following (simplified) example: a player A shoots the ball with an accuracy of 100% during half of his games, and 0% in the other half, while a player B shoots at 50% during all his games. Then, in average, both players have exactly the same shooting percentage, which is 50%, but their performance is hardly comparable. This leads us to the crucial notion of consistency, which aims to answer the following question: how much variance does a player display in this performance? Can he or she be trusted to perform according to his or her average numbers? Much more than an average number, what a sports statistician is really interested in is getting a confidence interval for a given parameter. If a player shoots between 48% and 52% during 95% of his games, then this is more informative than just knowing that the player has an average shooting percentage of 50%.

2 Empirical Bayes Approach

2.1 Prior Estimation

Suppose that one wants to estimate players' batting averages in the MLB. This problem seems a bit tricky for players with very few at bat: is 4 hits out of 10 really better than 300 out 1000? In order to get a more robust estimate, here is a possible approach. First, one can try to estimate the distribution of all the batting averages across all the MLB players, in which case one obtains a histogram similar to the figure 1 (only considering players with a sufficient number of at bats).

This histogram gives us prior information on what we consider a player's real batting average to be plausible, and will be used to get more reliable estimates for players with too few at bats. We

usually model our players' batting average distribution as a $Beta(\alpha_0, \beta_0)$, where α_0 and β_0 are two "hyper-parameters" inferred from our data. This family of distributions is most commonly used for two reasons. The first one is that the batting average represents a probability and takes its values in the interval $[0, 1]$. The second one is slightly more subtle, but mostly comes down to some nice properties of the distributions, which allows easy updates to its parameters when observing new data (in this case new batting events).

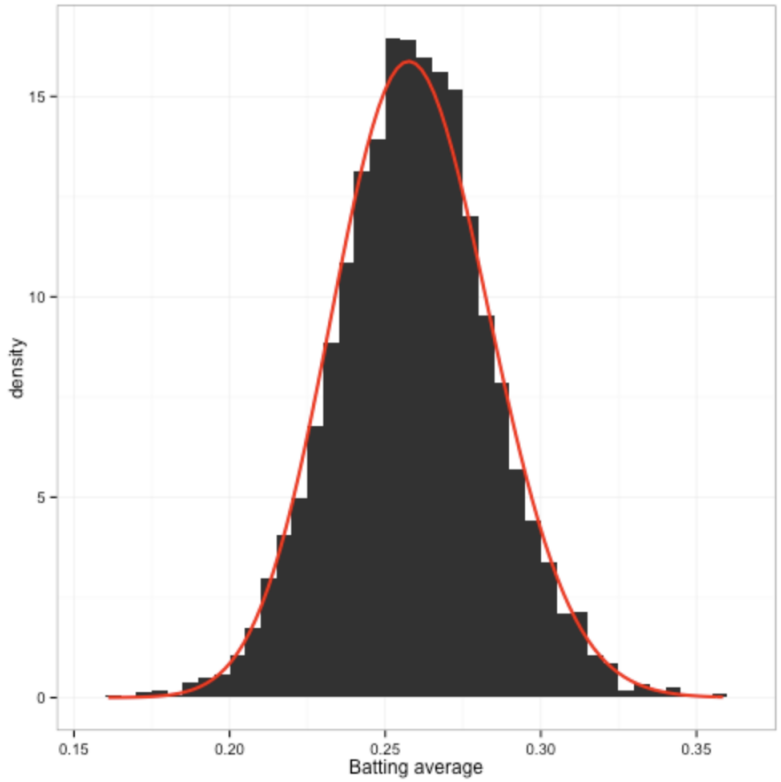

Figure 1: Batting Average and fit by a $Beta(79, 225)$ distribution

2.2 Using the prior for individual prediction

In the Bayesian framework, here are our assumptions. Each player i has a batting average $\theta_i \in (0, 1)$ that we are trying to estimate, and is independent of all the other players. However, this batting average θ_i is itself a random variable, which was drawn independently from the others, but from the same prior distribution: the same one we were trying to estimate in the previous section!

$$\theta_i \overset{i.i.d}{\sim} Beta(\alpha_0, \beta_0) \tag{1}$$

In other terms, there is a prior distribution of "raw" talent across the population, which can be approximated by the figure 1, and when a player begins his or her career, his or her batting average is drawn from this precise distribution. As a result, some are luckier than most and obtain a very high ability, and vice-versa for others.

Then, each at bat event can be seen as an independent experiment during which the player i has a probability θ_i of hitting the ball. That is, if $X_1^{(i)}$, ..., $X_n^{(i)}$ are $\{0, 1\}$-valued variables representing

whether the j at bat of player i resulted in a hit, then, we model this sequence as:

$$X_j^{(i)} \overset{i.i.d}{\sim} Bernoulli(\theta_i) \tag{2}$$

While observing the sequence of variables $X_1^{(i)}$, ..., $X_n^{(i)}$, our knowledge about θ_i gets more and more precise. Indeed, at first, without any observation, we can't say much more than (1), meaning that for each individual θ_i, we estimate it will roughly lie between 0.20 and 0.30 with high probability, However, as our number of observed events grows, our credible interval carrying our belief of where θ_i lies gets thinner and thinner. The reason is we can compute what we call the posterior distribution of θ_i given that we observed the events $X_1^{(i)}$, ..., $X_n^{(i)}$. By a nice property of the $Beta$ distribution (which is why we chose it in the first place), it turns out that:

$$\theta_i \mid X_j^{(i)} \sim Beta\left(\alpha_0 + \sum_{j=1}^{n} X_j^{(i)}, \beta_0 + n - \sum_{j=1}^{n} X_j^{(i)}\right) = Beta(\alpha_0 + \text{ Hits}, \beta_0 + \text{ Outs}) \tag{3}$$

Given that the mean of a $Beta(\alpha, \beta)$ is $\alpha/(\alpha + \beta)$, it is easy to see how any hit will tend to increase our posterior mean, whereas any out will tend to pull it down, which is somehow reassuring.

Now, for a player with 4 hits out of 10, the estimate of the batting average will be:

$$\frac{4 + \alpha_0}{10 + \alpha_0 + \beta_0} = 0.264$$

On the other hand, for a player with 300 hits out of 1000 Plate Appearances:

$$\frac{300 + \alpha_0}{1000 + \alpha_0 + \beta_0} = 0.29$$

and the estimate of the batting average will be higher.

This may seem counter-intuitive at first, but is actually very well explained. In the first scenario, our model considers it did not observe enough hits to confidently conclude that the player is well above average (that above 30%), so it will roughly output an slightly above average prediction. On the other hand, in the second scenario, such observed consistency actually means that, with high probability, the θ_i drawn for this specific player lies on the right tail of the distribution in 1, which is why the prediction is very close to the actual sample mean.

In the limit where the number of trials goes to ∞, our estimator would of course converge towards the sample mean (as well as the real θ_i in this model).

The interested reader will have a look at [2], where this model is studied in more details and made more elaborate.

3 Estimating Population Variance

A wide number of statistics based player evaluations look at a game under both a large sample assumption and a sort of "mean-field" assumption. As studied in the previous section, the large sample assumption can be slightly lessened if the model allows us to leverage prior information, and somehow pretend that the sample size is larger than it really is.

On the other hand, all our statistical estimators aim at evaluating a parameter which reflects a player or a team performance *in average*. While this is highly desirable in many scenarios, it can also fail to accurately describe a phenomenon in others. A first very natural case is the well-studied

notion of "clutchness" or ability to perform very well under pressure and at crucial moments of a game. This concept is closely related to estimating conditional probability, since it mostly asks how a player performs under specific conditions. For instance, if two basketball players shoot at an overall 50%, but one of them does it uniformly across all periods of the game (including fourth quarters) when the other has an average percentage of only 40% "under pressure", then you would probably deem that the first player is more "clutch" than the second.

In this section, we are interesting in evaluating consistency, which has to be to do with estimation of variance, or dispersion. The underlying question is the following: once I have an evaluation of a player, how confident can I be that in the next game, he/she is going to perform close to his/her average level?

In the following, suppose that we get to observe a quantity of interest X (e.g. the batting average, the shooting percentage, the average yards per pass or else a goalkeeper percentage of saved shots on goal), during N different games, where N is usually typically around a few dozens, or a couple hundred.

We therefore have a sample $(X_1, ..., X_N)$ of observations, and what we did so far was to compute the sample average $\hat{\mu}$:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

We know, from the law of large numbers, that when N grows large, $\hat{\mu}$ should be close to the real population mean $\mu = \mathbb{E}(X)$.

Several questions remain open. First, if N is not so large, we previously saw how considering prior information on μ could lead to dramatically better estimators than just the sample mean. It would thus seem natural to not only compute $\hat{\mu}$, but also to mention how many samples are sufficient to ensure that $\hat{\mu}$ is indeed a good estimator of μ. This is answered by a famous theorem called the *central limit theorem*, illustrated in figure 3. It tells us that, when $N \to \infty$ (or at least grows large), we know the distribution of $\hat{\mu}$, more precisely:

$$\hat{\mu} \dot{\sim} \mathcal{N}\left(\mu, \mathrm{Var}(X)/N\right)$$

where $\mathcal{N}(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2. Roughly, $\hat{\mu}$ and μ are close up to a $1/\sqrt{N}$ term, and the magnitude of this term is proportional to a second quantity of interest depending on the distribution of X, its variance.

In practice, one does not have access to the population variance of X (for the same reason as why μ is unknown), so one can only try to estimate it with:

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{\mu})^2 \tag{4}$$

The sample variance (4) can be used in two different settings.

First, when estimating the population mean, one can use the central limit theorem to conclude that with probability 95%, the true population mean μ lies between $\hat{\mu} - \frac{c_\alpha \hat{\sigma}}{\sqrt{N}}$ and $\hat{\mu} + \frac{c_\alpha \hat{\sigma}}{\sqrt{N}}$, where c_α depends on the amount of confidence we want (typically for 95%, $c_\alpha = 1.96$).

It turns out that $\hat{\sigma}$ can also be very informative even if the value of μ were known. Indeed, if an oracle tells us that a player has an average ability of μ in his field, how much do you really know about this player?

Let us consider a standard case of two players: one is shooting at 50% during each game, when the other alternates between 0% during half of the games and 100% during the other half. Both

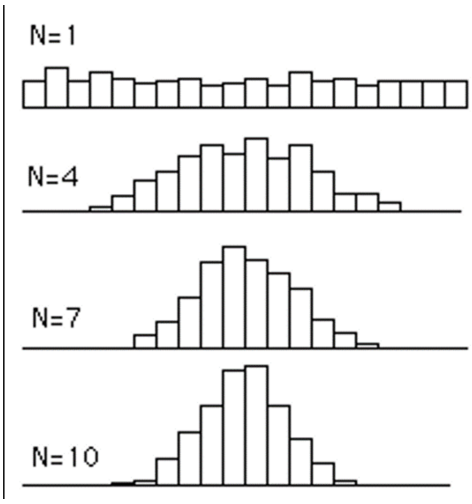

Figure 2. Convergence in Distribution: when N grows, the histogram looks more and more like a Gaussian distribution, centered around μ, and with variance σ^2/N

have the same average of 50%, and yet their performance is hardly comparable. In this case, you expect the first player to have a sample variance much lower than the second one, which might intuitively explain what the sample variance tells you about individual performance. Indeed, in both cases, you would predict the player to have a 50% shooting average during the next game, but if you expect your error to be close to 0 in the first case, it can't be less than 50% in the second! This is illustrated in figure In other terms, the sample variance gives an idea of how much variability there is in your data, and how concentrated your variable of interest is.

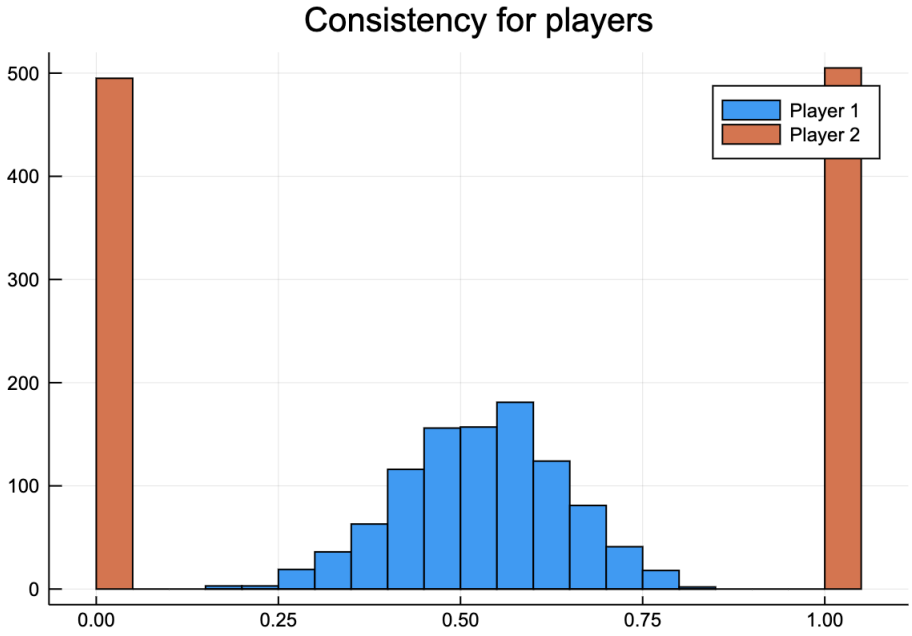

Figure 3. Histogram of players' shooting percentage during each game, assuming that both of them attempted 20 shots at each game

References

[1] Robinson, David. Understanding empirical Bayes estimation (using baseball statistics) (2015) http://varianceexplained.org/r/empirical_bayes_baseball/

[2] Jiang, Wenhua; Zhang, Cun-Hui. Empirical Bayes in-season prediction of baseball batting averages. Borrowing Strength: Theory Powering Applications A Festschrift for Lawrence D. Brown, 263–273, Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2010. doi:10.1214/10-IMSCOLL618. https://projecteuclid.org/euclid.imsc/1288099025