Number of Samples Needed For Model Selection With Confidence

Sean G. Carver*

Abstract

A common measure used to quantify the similarity of two models is the Kullback-Leibler divergence, computed from a true model to an alternative model. We propose a different measure: the number of samples needed to correctly reject the alternative model with a given confidence level (e.g. 95%). Our method works as follows: (1) we simulate samples from the true model, (2) for each sample, we compute a log-likelihood ratio (3), we bootstrap and sum the log-likelihood ratios—when this sum is positive, we select the true model, (4) using simple linear regression, we determine the number of terms (i.e. number of samples) needed to make the desired quantile (e.g. 5%) fall at zero. We have tested this method on t-distributions of different degrees of freedom and have confirmed that it gives reasonably consistent results. However, we plan to apply this method to Markov chains, e.g. used for sports statistics like tennis, volleyball, and baseball. For these applications, it may be desirable to have a measure that is easier to interpret than the Kullback-Leibler divergence. How many innings are needed to falsify the model of the Yankees when simulating a model of the Orioles?

Key Words: Model selection, likelihood ratio test, Akaike information criterion, Kullback-Leibler divergence

1. Motivation

When working with two or more probabilistic models, you may want to quantify their similarity. Specifically, when observing simulations of one model, can you easily tell that the simulations do not come from another model? Or do only subtle differences between the models make this discernment difficult?

My favorite example involves baseball. What if you had score cards recording, play after play for many games, what bases had runners, and how many outs had occurred. Could you tell which teams were up at bat? In practice, there will be many uncontrolled factors, so if you require a high degree of certainty, you will only be able to distinguish between two teams if they are substantially mismatched. On the other hand, in theory, if the teams playing are *models of teams*, obeying precisely defined and known probability laws, and if the models make different, even slightly different, predictions, you can make the choice confidently. Indeed, you can have as great a chance as you want, short of being absolutely certain, of choosing the right team, provided you have enough data [2]. But how much data suffice for the confidence you demand?

Once you derive a model of each team from game records, you can ask yourself, for example: how similar is the model of the our home team, the Baltimore Orioles, to the model one of their rivals, the New York Yankees. (Baltimore hosted the Joint Statistical Meetings in 2017.) How many Baltimore-at-bat half-innings would you need to simulate to correctly reject the statement that the model of the Yankees generated the simulations, and get this answer right at least 95% of the time (or some other specified confidence level)?

*American University, 4400 Massachusetts Avenue, NW, Washington, DC, 20016

2. Baseball as a Markov Chain

Baseball is often modeled as a Markov chain [5, 1, 4], where each half-inning proceeds through a number of states. The states record which bases have runners, and how many outs have occurred. There are 8 possible combinations of runners on base (labeled: 0, 1, 2, 3, 12, 13, 23, 123), and 4 possible numbers of outs (labeled: *blank*, X, XX, XXX). Specifically, the state label 0 indicates empty bases with no outs, whereas the state label 123XX indicates loaded bases with two outs, etc. All states with three outs are combined into a single absorbing state: XXX, which signifies the end of the half-inning. There exist a total of 25 states and a total of 600 (equaling 24×25) conceivable transitions (no transitions happen from the XXX state). We note that many of these transitions remain impossible, by the rules of baseball, such as, for example, going from bases empty with two outs, to bases empty with one out, i.e. 0XX:0X. Using a recursive algorithm savvy to the rules, I counted 296 allowable transitions between states, and 304 illegal ones. Of the 296 allowable transitions, only 272 occurred at least once in the 2011 Major League season—24 never occurred (see Figure 1). For example, in 2011 no team underwent the transition 1X:1X even though it could have occurred, within the rules of the game. Had a team undergone this transition, the batter would have advanced only to first base, while the runner on first would have scored—clearly an unusual scenario, but not impossible: both 1:1 and 1XX:1XX did occur in 2011 Major League play.

We can then define a 24×25 matrix of transition probabilities, with entries $\{p_{i:j}\}$. Specifically, $p_{i:j}$ is the probability that the game will transition to state j, assuming it finds itself in state i, just prior. To specify a model of a baseball team, we must simply specify these 600 transition probabilities. However, 304 of these entries must be zero, by the rules. The rest of the probabilities will fall between 0 and 1—including perhaps dozens more zeros for any given team. Twenty-four additional constraints (sums of the entries in each row equal 1) stem from the stipulation that, from each of the 24 transient states (i.e. other than XXX), the game (or half-inning) must go on with a single new state.

To determine the values of these parameters for each team, the simplest method uses the so-called *maximum likelihood estimates* (MLEs) [3]:

$$\text{MLE of } p_{ij} = \frac{\text{number of transitions made from } i \text{ to } j}{\text{total number of transitions made from } i}$$

In calculating the MLEs for a particular team, use game records for that team. My team models come only from the game records of the team batting at home. I made this choice to make the models as different as possible. Even in the Major Leagues, ballparks differ substantially, and the home field can have a significant effect on the play. These park effects add distinctiveness to the models that also comes from the differences in the team at bat. Fixing the home field also reduces variability in the data from game to game.

Unfortunately, baseball models based on MLEs have significant drawbacks, especially when we consider the models together. Though some transitions are common, many others are not. Invariably, some of the more unusual transitions will happen for one team, in one year, but not for the other. As a consequence, MLE models will make some transitions possible for one team, but not for the other. For example, the uncommon transition 23X:3X happened once in 2011 for the Baltimore Orioles, but never in 2011 for the New York Yankees. Thus, if you ever see the 23X:3X transition in a long series of half-innings, you can immediately reject, with absolute certainty, the statement that the 2011 MLE model of the Yankees simulated the data. This automatic rejection makes the correct selection of the Orioles easy for a reason that depends more on the noise in the data than on the teams under study. For this reason, the interpretation of our results, below, will be made in the context of this rather unfortunate behavior of the MLEs for baseball.

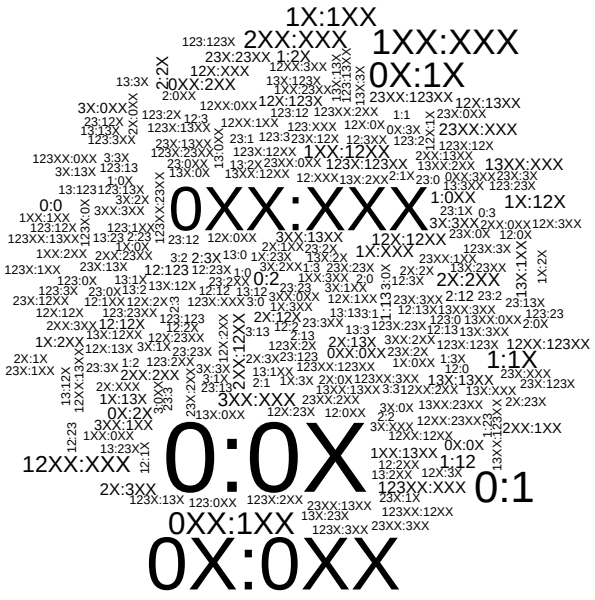

Figure 1: The 272 baseball state transitions that occurred throughout the 2011 Major League season, arranged randomly, and sized according to their frequency of occurrence. Additionally, there were 24 transitions that were possible, according to the rules of baseball, but that never occurred in 2011 Major League play (not shown).

Figure 2: Most common half-innings, as predicted from transition probabilities fit with 2011 Major League season data, arranged randomly, and sized according to predicted frequency of occurrence. Half-innings with predicted probabilities less than 0.01 are lumped together as "Rare Half Innings. With no cap on the score, there are infinitely many such rare half innings.

3. Deciding Between Models

Baseball presents one example; below, we introduce another. Our calculations work for any type of model, provided (1) you can simulate the true model to generate samples, and (2) you can calculate the likelihood of each sample, both for the true model, and for any alternative model under consideration. The relative values of these likelihoods determine which model you deem correct. Of course, with these calculations, you know in advance which model is correct, but knowing this fact allows you to calculate how many samples you would have needed to be confident, if you did not already know the answer.

Each baseball sample consists of the play-by-play records from a half-inning for the model team at bat. To create this sample with a simulation, start with bases empty and no outs (the 0 state) then successively apply the appropriate transition probabilities to "throw the dice," to see the succession of states, until reaching the end of the half-inning (the XXX state).

It is well known that in the history of Major League play, no player has ever consistently batted over .500, meaning that most plate appearances, for all batters, have led to an out. From this fact, we can guess that, throughout Major League history, the most common half-inning has been the one in which there are three outs in a row: 0:0X:0XX:XXX. Indeed, simulations with the transition probabilities (combined for all 2011 Major League teams, home and away) predict 0:0X:0XX:XXX for 31% of the half-innings. Individual teams will have different percentages for this occurrence, higher or lower, depending on the skill of their batters. True, this most common outcome does not represent a majority of all half-innings, but the next most common half-inning, 0:0X:0XX:1XX:XXX, occurs only 8% of the time.

Based on the 2011 data, there are only 9 distinct half-innings which have probabilities greater than 0.01 (predicted to occur more than 1% of the time, Figure 2). I call these *non-rare half-innings*; the rest I call *rare half-innings*. But the probabilities for the non-rare half-innings only sum to 0.56, which predicts a rare half-inning 44% of the time. As a group, rare half-innings are not particularly rare, but each one occurs less than once in 100 half-innings. How is this statement consistent with reality? Realize that, with no cap on the score, there are infinitely many possible half-innings; their probabilities sum to exactly 1 in an infinite convergent series. Necessarily, these probabilities tail off to keep the sum finite, but still, the fact that individual half-innings have such low probabilities is made up for by the fact that there are so many of them.

For a second example, we use the student t-distribution. Instead of 600 transition probabilities, t-distributions have a single parameter: the degrees of freedom. Our true model will be a t-distribution with 5 degrees of freedom, t(5), whereas our alternative model will be a t-distribution with infinite degrees of freedom, t(∞), also known as the standard Normal distribution, N(0, 1). Analogous to a single half-inning in baseball (such as 0:0X:0XX:XXX), each sample from a t-distribution is a single number on the real line, either positive, negative, or zero (such as -0.6576941). Consequently, unlike in baseball with its discrete half-innings, for the t-distributions, the samples occur on a continuum—in one dimension. As with baseball, the next step grasps the likelihoods of the different possibilities. For the t-distributions, the simplicity of the sample space allows you to plot the likelihood (on the vertical axis) against the sample (on the horizontal axis) which shows the graph of the so-called probability density function (Figure 3). This graph for N(0, 1) is just the familiar *bell curve*; the graphs for the other t-distributions appear nearly indistinguishable from the one for N(0, 1), but they are slightly warped bells, as described below.

For all t-distributions, the most likely sample is 0, which appears as the peak of the bell. Away from 0, the likelihood drops off into symmetric tails that extend to infinity

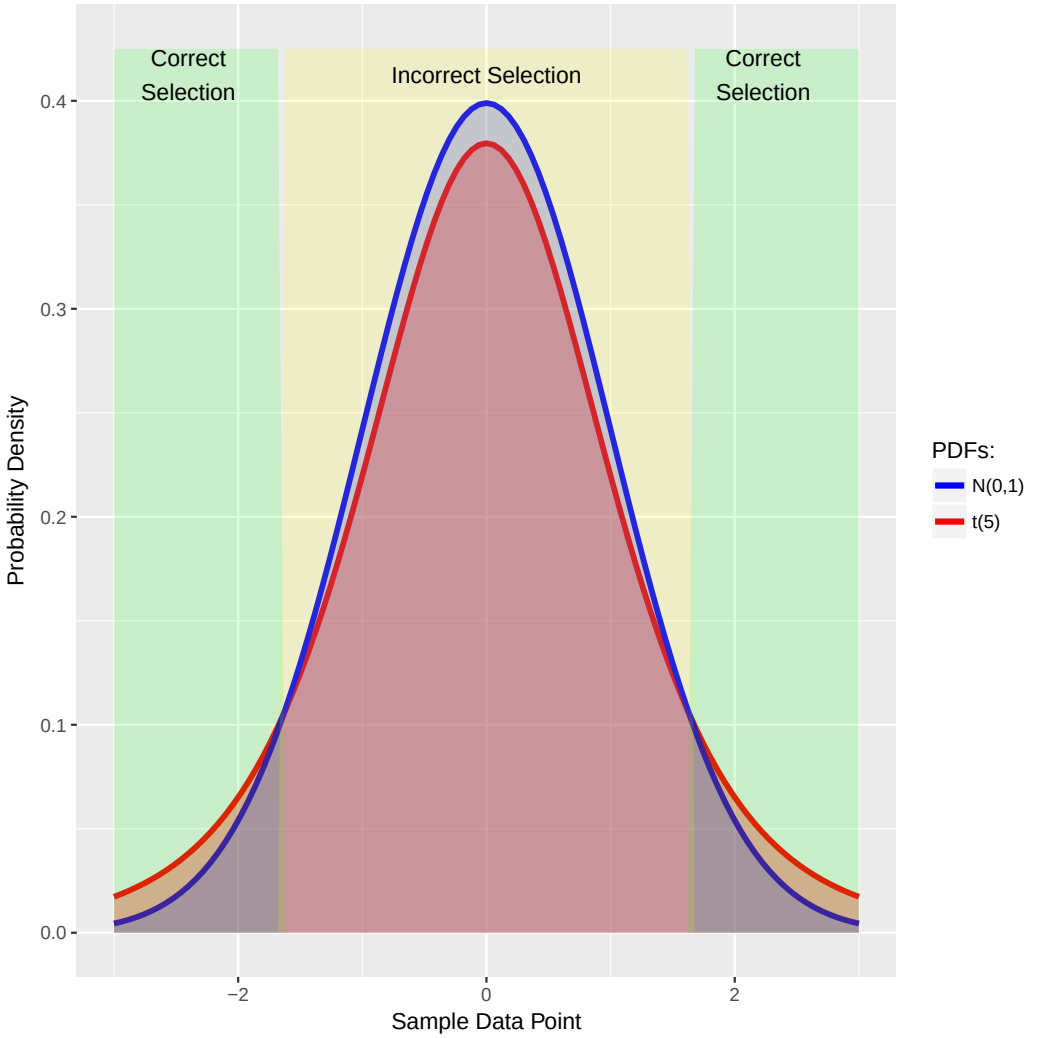

Figure 3: Selecting a model with only one data point: the probability density functions of the t(5) and N(0, 1) distributions. Assuming the t(5) distribution generates the data point, the correct selection will be made only if this data point falls in the tails of the distribution (shown with green shading), otherwise, more likely, the incorrect selection of N(0, 1) is made (shown with yellow shading). Improving the odds of a correct selection requires more data points.

in both directions. Just as in baseball, where the probabilities of the different possible half-innings sum to 1, the probability density of a t-distribution integrates to 1 (i.e. area under the bell equals 1) over the whole range of possible samples. Because the range of possible t-samples has no limit, the tails must drop off precipitously to keep the area below them finite, just as the probabilities for half-innings with higher and higher scores must similarly drop off. A key point is that this drop off occurs slower for t(5) than for N(0, 1): t-distributions have heavier tails, the smaller their degrees of freedom. In a similar fashion, in baseball, the "tail" (by which I mean: the probabilities of half-innings with higher and higher scores) drops off slower for better teams. But higher probability in the "tail(s)" must be compensated for by lower probability in the "center." The top of the bell is lower for the heavier-tailed t(5), than it is for N(0, 1). Likewise, the lowest scoring half-innings, such as 0:0X:0XX:XXX, are less likely for better teams.

To decide between models, using one sample, you make the choice based on which model has a higher likelihood, at the sample. For the t-distributions, if the sample is in one of the tails, you pick t(5); if the sample is in the center, you pick N(0, 1). The intersections of the two model's probability density functions determine the boundaries of the regions where the different models are selected. For baseball, you will tend to pick a better team if the half-inning's score is higher, and a worse team if the score is lower, although the judgement based on likelihood is often more subtle than the one just described.

In many cases, based on one sample, the answer will be wrong, most of the time. Indeed for all t-distributions, including t(5), most samples appear near the center (top of the bell), so t(5) will be rejected in favor of N(0, 1), with one sample, most of the time, even if t(5) is the true model. Likewise, good teams do often get low scoring half-innings. If this common occurence happens when choosing between teams with a single half-inning, a good team may be deemed less likely than a worse team to have produced its own data. It takes many samples, including occasional rarities in the tails, to fill in the histogram and make the correct choice, confidently.

How do we compute likelihood? For t-distributions, likelihood is the height of the bell at the sample, given by formulas (for the probability density function) easily found online. In this respect, baseball is simpler: the likelihood of a half-inning is the product of the probabilities of all its transitions. In all cases, to compute the likelihood of an ensemble of independent samples, you compute the product of the likelihoods of the individual samples. Products appear because of the independence (or, in the case of transition probabilities, the conditional independence) of the events under consideration. Note that, in working with many samples, you should, instead, work with the log-likelihoods, otherwise the numbers will get too small for a computer's floating point arithmetic to handle accurately.

We are interested in the relative likelihoods between the two models, and we select the model with the greatest likelihood. This selection goes by the name *likelihood ratio test* and in our situations, it is equivalent to making the decision based on the much championed *Akaike information criterion* or AIC. Why are the two equivalent? The formula for the AIC differs only by adding a term that adjusts for bias that inevitably occurs when the modeler makes the selection with the same data used for fitting the models [2]. We make our selection with data simulated after we fix the parameters, so in the formula for AIC, there are no fitted parameters, which implies that the two methods agree.

As suggested by the name *likelihood ratio test*, we use ratios. Specifically, after generating samples from the true model, we select models based on the ratio of the product of true-model likelihoods, over the product of alternative-model likelihoods (both using the same true-model samples). If this quantity is greater that 1, we choose the true model; otherwise we choose the alternative model. After taking logs, these quotients and products simplify to the sum of the log-likelihood ratios of the single samples. When this whole

Table 1: The *bootstrap table*: each entry is the sum of log-likelihood ratios computed from independent samples bootstrapped from a large population (size 200,000) drawn from the true model, t(5). A single log-likelihood ratio is the log of the quotient of the probability density of the true model at the sample, to the probability density of the alternative model, N(0, 1), at that same true-model sample. The sum of a number of these values, for an ensemble of independent samples, is the log-likelihood ratio of that ensemble of samples. Each entry of the k^{th} column of the table contains the sum of k of these log-likelihood ratios. A positive entry indicates a correct selection of t(5) over N(0, 1), based on the corresponding ensemble.

	One single	Sum of 2	Sum of 3	Sum of 4	Sum of 5
Repetition 1	-0.0619845	0.8031872	0.0447267	0.2339618	-0.3907877
Repetition 2	-0.0966445	0.0845291	-0.1887769	-0.2913085	5.2309346
Repetition 3	-0.0541549	-0.1628995	-0.2364334	1.6543063	1.2374765
Repetition 4	-0.0842683	0.7632357	0.4930037	-0.2023973	-0.3325500
Repetition 5	-0.0526431	-0.1469163	-0.2068822	2.0281977	0.1084606
Repetition 6	-0.0858160	1.0636336	-0.0271214	-0.2683411	-0.3006797
Repetition 7	-0.0508892	-0.1506010	-0.2433817	-0.2727284	-0.4206254

sum is greater than zero (i.e. $\log(1)$), we choose the true model; otherwise, we choose the alternative model.

The mean of the single-sample log-likelihood ratios is an unbiased estimate of the so-called *Kullback-Leibler divergence* from the true model to the alternative model [2]. While such a Monte Carlo estimate can be either positive, negative, or zero, the actual value of the Kullback-Leiber divergence (defined as the expected log-likelihood ratio) always remains nonnegative—and positive as long as the models make different predictions (known together as the *Gibbs inequality*). From this result, the central limit theorem, and a finite variance assumption, we can be sure we will, with any level of confidence short of certainty, select the true model, with enough samples, provided the two models make different predictions. In other words, our test is *consistent* [2]. For two baseball models, requiring that models make different predictions demands that at least one transition probability must be different.

4. Computing The Samples Needed

I use a brute force Monte Carlo technique. Specifically, I compute a table of sums of log-likelihood ratios (Table 1). The first column contains single log-likelihood ratios; the second contains sums of two; the third contains sums of three; and so on. The entries in the table derive from independent random samples, and independence holds across rows, as well. For large tables, the computational requirements can be prohibitive, so to economize, I compute a smaller number of samples than needed, together with their log-likelihood ratios, then use a bootstrap (random resampling with replacement) to fill in the table. This procedure introduces a bias, but one controlled by adjusting the number of samples available in the bootstrap population.

Once I compute the table, I can compute the proportion of entries in each column that are positive (each indicating a correct selection, Figure 4). The column number indicates how many samples are used for the selection, so the column number for which this proportion crosses the chosen confidence level (e.g. 0.95) *should* be my answer. With finitely many samples, however, this crossing is not well defined because the proportion varies randomly, and in a way that varies across columns. As a result, there appears a region, which

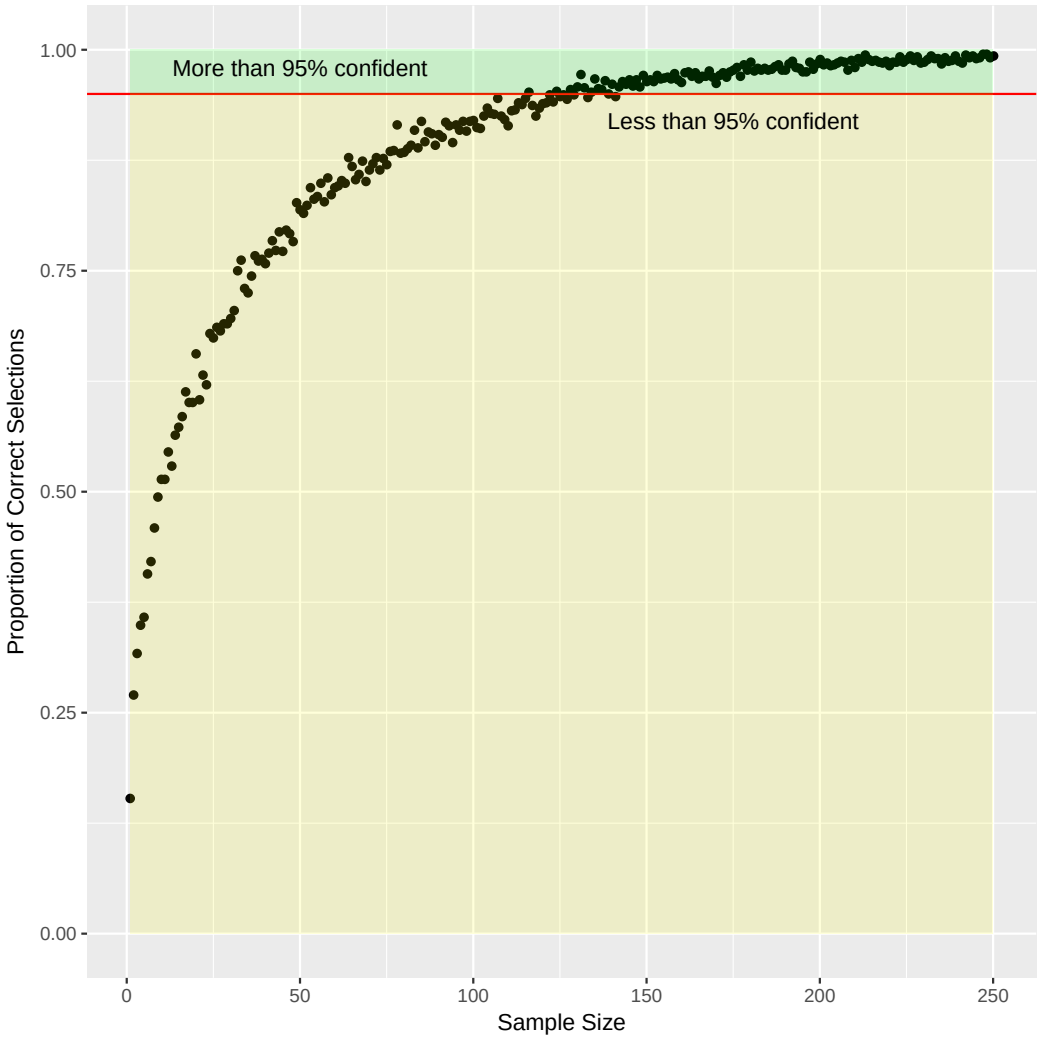

Figure 4: The proportion of correct selections (vertical axis) indicated by the proportion of positive values within each column of a "bootstrap table" (with 1000 rows and 250 columns, bootstrapped from a population of size 200,000, see Table 1). The proportion of correct selections is plotted against the sample size (horizontal axis; varying across columns of the table). Each proportion correct is necessarily a multiple of the reciprocal of the number of rows in the table. As the proportion correct crosses the value of 0.95, (boundary between yellow and green shading), we deem the confidence level acceptable.

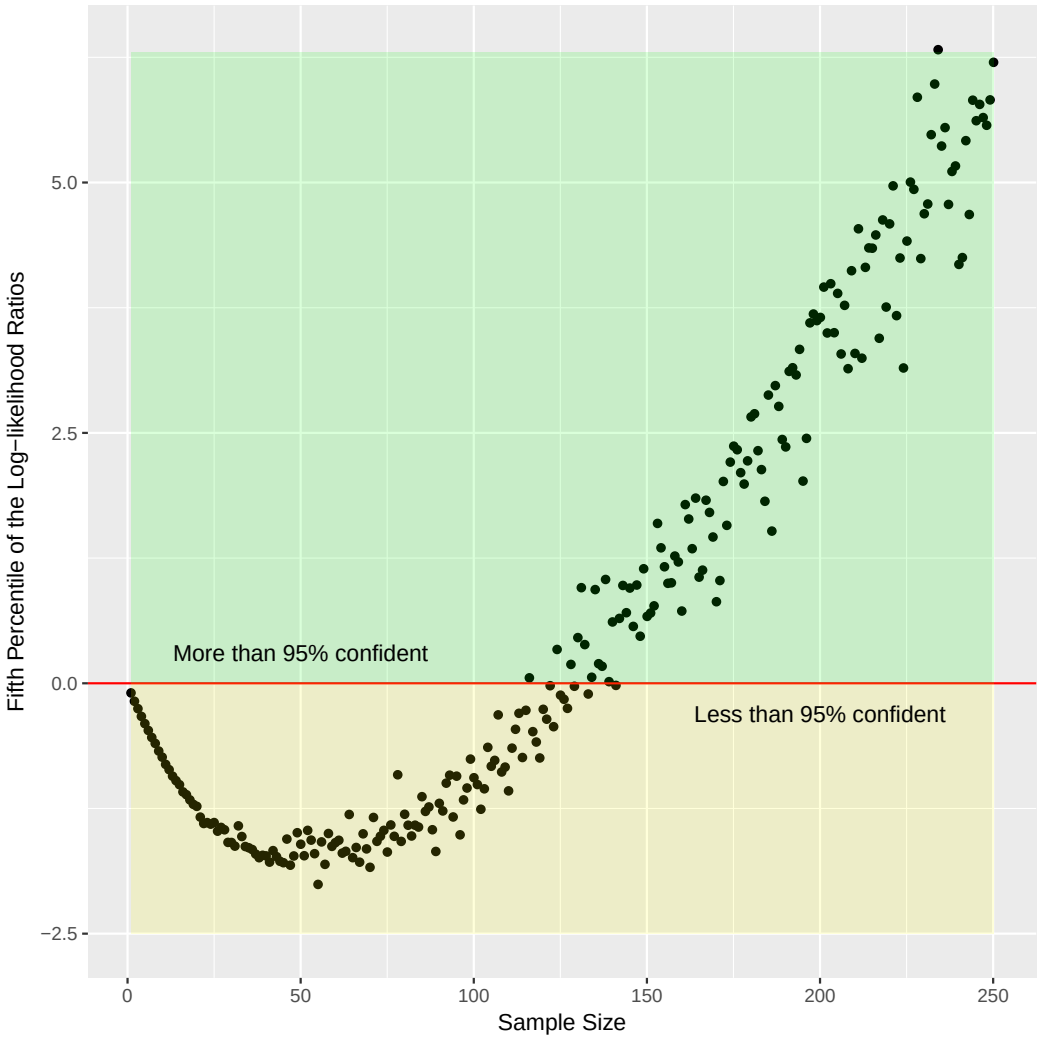

Figure 5: The fifth percentile of the log-likelihood ratios (vertical axis) within each column of a "bootstrap table" (with 1000 rows and 250 columns, bootstrapped from a population of size 200,000, see Table 1). The fifth percentile is plotted against the sample size (horizontal axis; varying across columns of the table). The fifth percentile crosses zero (the boundary between the yellow and green shading) at the same place where the proportion correct crosses 0.95, and where we deem the confidence level acceptable (green shading, see also Figure 4).

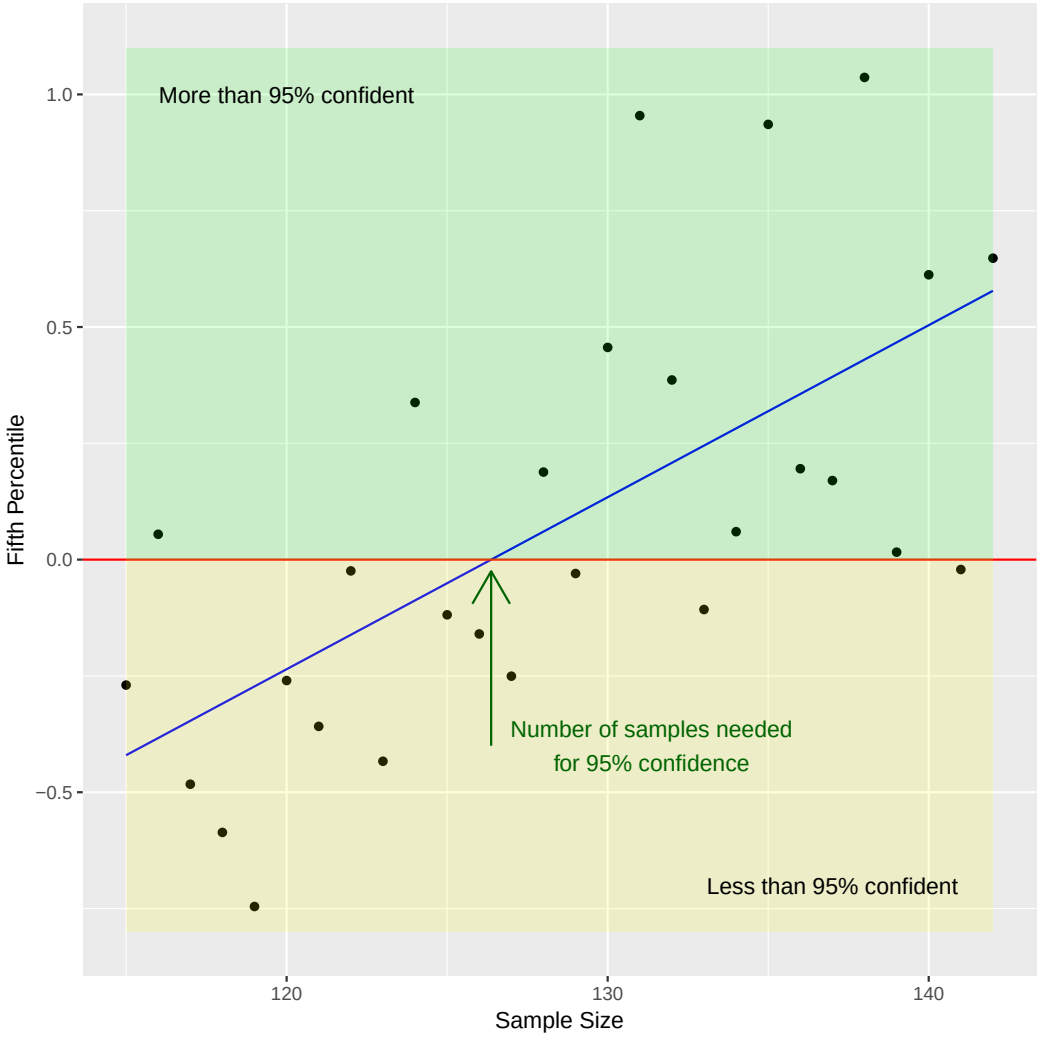

Figure 6: The region of interest (zoom-in of Figure 5). I define the region of interest to include the last sample before the first zero-crossing, the first sample after the last zero-crossing, all samples in between, but nothing else. The blue line is the linear least-squares regression line for the data in this region, the dark green arrow points to the estimate of the number of samples needed for 95% confidence.

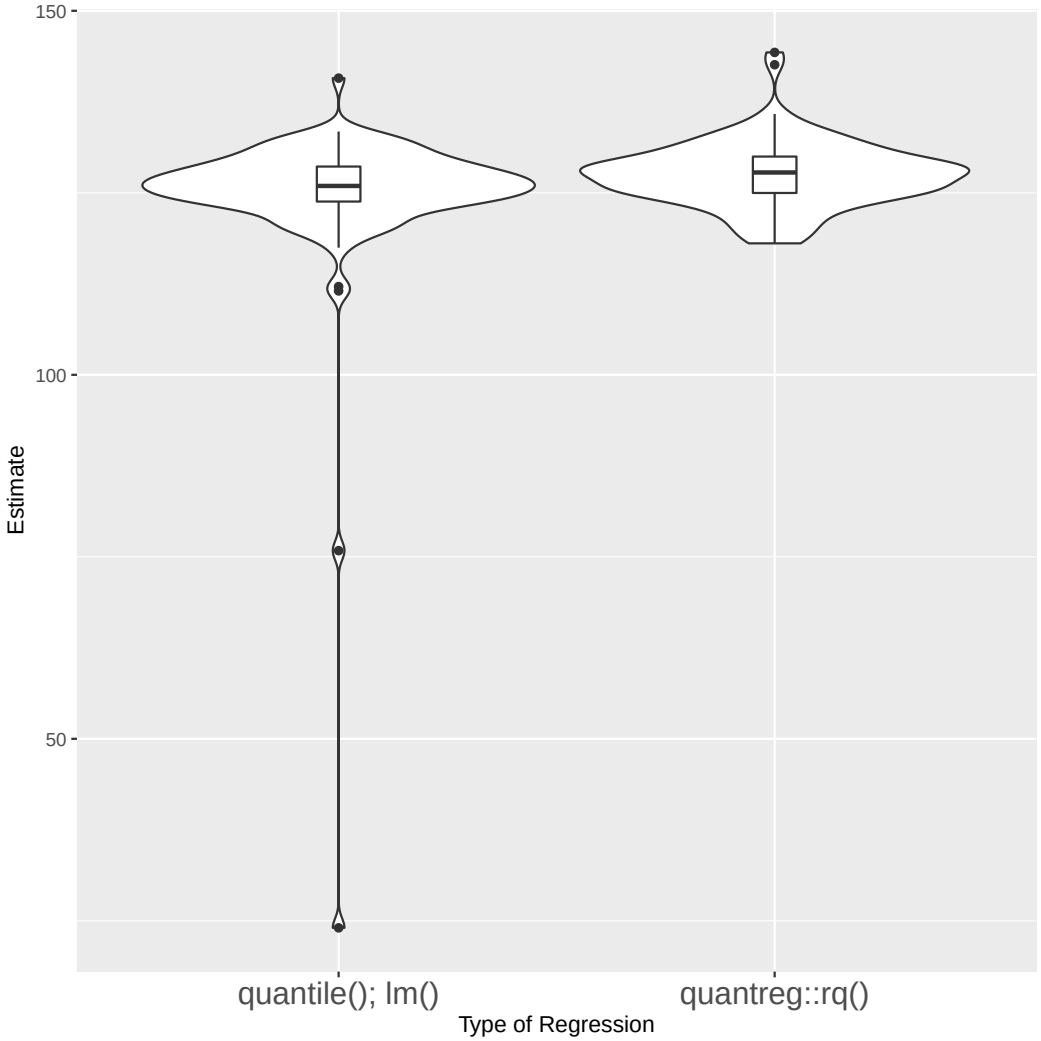

Figure 7: Distribution of the estimator of the number of samples needed to correctly reject N(0, 1) in favor of the correctly specified t(5) model. Shown are violin and box plots for two alternative methods of computing the estimate. Left: simple linear regression on the fifth percentile of the log-likelihood ratio. This procedure, shown in Figure 5, occasionally produces extreme outliers. Right: quantile regression with the R function quantreg::rq() applied to the same data.

I call the *region of interest*, where the desired crossing occurs many times.

My first solution zooms in on this region of interest and performs a linear regression on the proportion correct to pinpoint an estimate where this proportion crosses the confidence level. But this proportion must be a multiple of the reciprocal of the number of rows in the table. Unfortunately, this discretization makes for poor regression. So instead, I now work with quantiles (Figure 5). For example, the proportion correct crosses 0.95 at the same place where the 5^{th} percentile of the log-likelihood ratio crosses 0 (i.e. $\log(1)$, the threshold for the sum of the log-likelihood ratios to indicate a correct selection). As before, the quantiles are random, in a way that varies across column, so again, we must zoom in on a region of interest and perform a regression (Figure 6). But the quantiles are no longer discrete and the regression performs better.

I have tried both simple linear regression (on the quantiles, with the R command lm()) and quantile regression (directly on the entries of the bootstrapped table, with the R command quantreg::rq()) both using default arguments, where possible. A careful reading of the documentation does reveal fundamental differences between these two procedures, however, there was there was not a substantial difference in the spread of the estimates (Figure 7). On one hand, the lm() command occasionally fails, producing extreme outliers, or even nonsensical results (negative values for number of samples needed, not shown). The alternative, quantreg::rq() performed better in this regard, however the computations take almost twice as long. Indeed, using quantreg::rq() requires duplicating the intermediate step of calculating the quantiles—once to compute the region of interest, and once to perform the regression. Using lm() is more economical, in this regard.

5. Results

Using the quantreg::rq() function, I calculate that it takes 128 ± 1 samples from the t-distribution, with 5 degrees of freedom, to reject, with 95% confidence, the statement that these samples come from the standard normal distribution (mean $\pm$ standard error, averaging 100 estimates, shown in Figure 7). There was only a slight difference with the lm() function: 125 ± 2, (same number of estimates, shown in Figure 7). After eliminating the two extreme outliers from the distribution, the result changes to 126 ± 1 estimates.

On the other hand, I calculate that it takes 30 ± 1 half-innings, simulated from the MLE model derived from the 2011 Baltimore Orioles, batting at home, to reject, with 95% confidence, the statement that these half-innings were sampled from the 2011 New York Yankees MLE model.

The computations for the t-distributions, each giving a ± 1 range for the estimates, took 6 minutes and 12 minutes on a MacBook Pro, for, respectively, the lm() method, and the quantreg::rq() method. The duration of the baseball computations fell within that same range. The baseball calculation required averaging only 6 estimates, but each estimate took longer to compute—the total time elapsed was 11 minutes.

6. Discussion

One reason baseball teams are easier to distinguish than t-distributions involves the above mentioned properties of the MLE. Specifically, in a simulation of 100,000 Baltimore-at-bat half-innings, it was discovered that just under 2% had not ever occurred for the Yankees while batting in their New York home stadium in 2011. If one of these half-innings occurs in a set of Baltimore-at-bat samples, the selection is immediately made for the true model, the Orioles. At least one such half-inning is expected in approximately 45% of sequences of 30 Baltimore half-innings, making the correct selection substantially easier. A possible

solution to this problem would smooth each team's transition matrix in such a way that no transition remains impossible, for any team, that stands possible for another. After all, if a transition truly can happen, why deem it impossible, even if it never occurred for a given team one year? There exist principled ways of making this adjustment, discussed in [5], but they are beyond the scope of the present paper.

The number of samples needed for model selection with confidence offers an alternative to the Kullback-Leibler divergence to quantify the similarity of two models. This alternative may be easier to interpret for those untrained in model selection, and unfamiliar with the Kullback-Leibler divergence. However, the unlike the Kullback-Leibler divergence, the proposed measure is discrete, and substantially harder to compute.

7. Code

I provide code for my calculations, and source for this paper, using best practices for reproducible research, at: https://github.com/seancarverphd/klir. Note that the GitHub repository mentioned below in the acknowledgements contains the 2011 baseball data needed to run my code. I have not added this large data set to my own GitHub repository.

8. Acknowledgements

Rebeca Berger helped prepared the data for processing, and adapted some code for this project from the GitHub companion to [5]: https://github.com/maxtoki/baseball_R. Her adapted code is included with my own, at the GitHub address mentioned above in the Code section.

References

[1] Jim Albert. *Teaching Statistics Using Baseball*. The Mathematical Association of America, 2003.

[2] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Inofrmation-Theoretic Approach*. Springer, second edition, 2002.

[3] Morris H. DeGroot. *Probability and Statistics*. Addison-Wesley Publishing Company, second edition, 1975.

[4] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. Springer-Verlag, 1976.

[5] Max Marchi and Jim Albert. *Analyzing Baseball Data with R*. CRC Press, 2013.