Estimating probability with regression to the mean

Jan Maly[1]

[1]Affiliation not available

August 23, 2020

Abstract

Probability is usually estimated as a quotient of successful attempts and total attempts. This article introduces my own formula, whose estimates are consistently better. The motivation for writing this paper was an effort to make the best possible predictions of the future shooting percentages of ice hockey teams, so this topic is used throughout the entire work. The topic of this paper is both relevant within a statistical context and, simultaneously, will also appeal to all involved with ice hockey in general. In the last section I explain the role of shot quality in the NHL and show the weaknesses of popular ice hockey statistics, such as the Corsi.

Keywords: probability estimation, regression to the mean, beta-binomial distribution, ice hockey

Significance of shooting percentage in ice hockey

In the beginning I want to give you a basic understanding of a topic which will be used as a motivational example throughout the entire article to create questions which will be answered in later sections, using my own new statistical techniques. All the data used in this paper was collected from game logs on www.nhl.com. They include only regular season games without overtime.

Shooting percentage is a very controversial statistic in the ice hockey world. The common definition given for it is that it is a quotient of goals and shots on goal. This definition, however, has disadvantages, which will be discussed later, and because of them, I will be working with another type of shooting percentage, which is called the *Corsi shooting percentage* . It is defined as a quotient of goals and the total number of all types of shots, including shots on goal, blocked shots, and shots missed. The total number of all these types of shots is called *Corsi* , taking its name from the former NHL goaltender coach Jim Corsi. Therefore, in this article, a *shot* refers to a Corsi shot. When I am talking about shots on goal, it will be specifically mentioned. The empirically observed sample statistic of the Corsi shooting percentage will be denoted by $shc\%$ and is defined as:

$$shc\% = \frac{goals}{corsi} \tag{1}$$

So why is this statistic controversial? Because people in the community of ice hockey statisticians keep arguing whether differences in this statistic are caused by differences in the quality of shots or by randomness and lack of data. There are a lot of people who believe that all the teams have pretty much equally efficient shots, the differences in $shc\%$ are mostly random, and it is better to leave this statistic out completely and

predict a team's future results from shot differentials, rather than goal differentials, because shots are more sustainable. This problem also has an impact in real NHL games, because there are teams whose game style is based on this idea. A perfect example is the Carolina Hurricanes, who are always among the best teams when it comes to Corsi, and yet they have not made it into the play-offs for nine consecutive years since 2009 as a result of their low shot efficiency. In this article I want to show why it is hard to get any information from this statistic and also show how to get the most out of it.

I will assume that for each team there is a probability that their shot will result in a goal, and this probability remains more or less constant over a sustained period of time. It will be denoted by $shc\%^\star$. If the Corsi is high enough, the sample statistic $shc\%$ will converge to $shc\%^\star$.

$$\lim_{corsi\to\infty} shc\% = shc\%^\star \tag{2}$$

In other words, if we observe a million shots by a certain team, the effect of randomness will be neglected and their $shc\%$ will be very close to the probability that their next shot will result in a goal. The problem is that in the 2018 NHL season the average team's Corsi in the entire season was 4,764.64. So even at the end of the season there is a big effect of randomness in $shc\%$ and hence it is significantly different from $shc\%^\star$. And at the beginning of the season, the effect of randomness is even bigger.

To show the effect of randomness, let us analyse the shooting percentages of two teams (the New York Rangers and Boston Bruins) in the 2018 NHL season. The Rangers had 4,458 shots and scored 221 goals in regular time in the regular season. The Bruins scored 261 goals from 4,858 shots. From this data, $shc\%_{\mathrm{BOS}} = 5.37\,\%$ and $shc\%_{\mathrm{NYR}} = 4.96\,\%$. It may seem that the Bruins are a better team, but if we construct 95 % confidence intervals, we can say with a confidence level[1] of 95 % that the Rangers' real shooting percentage $shc\%^\star_{\mathrm{NYR}}$ belongs to the interval $(4.32\,\%, 5.60\,\%)$. For the Bruins the confidence interval is $(4.74\,\%, 6.00\,\%)$. Therefore, it is possible that because of the lack of data, the Rangers only seem to be a worse team, because they were unlucky, and if the season was longer, their shooting percentage would be, for example, 5.2 %, while for the Bruins it would be only 5.1 %.[2]

So, as you can see, it is not easy to compare NHL teams by the quality of their shots, because we do not have enough data. As mentioned above, some people even prefer to assume in their computations that all the teams have equally successful shots and that $shc\%^\star$ is the same for all teams and can be estimated by the average of this statistic across the league.[3]

The main goals of this article are to compare these two commonly used estimators of $shc\%^\star$ (i.e. the $shc\%$ statistic and the league's average shooting percentage) and find a better estimator than these two. In the next section I will develop a formula which will be able to beat the usual two estimators and later I will compare the results of all these estimators on both artificial data and real NHL data.

[1] This says that if we construct several confidence intervals from several observations, the real shooting percentage will be within the interval in 95 % of cases.

[2] The typical shooting percentage definition uses shots on goal instead of Corsi. At this point, it should be clear why I use the Corsi shooting percentage instead of the traditional shooting percentage. For instance, the Rangers only had 2,501 shots on goal in the 2018 season. That would give a 95 % confidence interval of their shooting percentage $(7.73\,\%, 9.95\,\%)$. Both the numbers are bigger, because the types of shots that we left out did not result in a goal. Since we have omitted almost half of the shots, there is less information in the estimate. Therefore, the confidence interval is even wider and hence the estimate is less precise, and the statistic is less telling.

[3] The 2018 season's average $shc\%$ across the league was 4.93 %.

How can the estimate of probability be improved?

In this section I will introduce the formula which I will be using to estimate the real shooting percentages, so those who are not interested in the mathematical side of things can just study the formula presented at the beginning of this chapter and skip the rest.

As I have already shown, estimating an unknown probability p from sample data as:

$$\hat{p} = \frac{m}{n} \tag{3}$$

where m is the number of successful attempts and n is the number of total attempts, is inaccurate if n is not high enough. What exactly *high enough* means depends on the required precision of the estimate, but as has already been shown, when comparing the shooting percentages of ice hockey teams even several thousand observations are not enough. So how can we make the estimate more precise when we do not have unlimited numbers of observations and we need greater precision than equation (3) provides? The answer, which will be explained in more detail in this section, is that we can take N other subjects, observe their successful and total attempts $(m_1, n_1), (m_2, n_2), \ldots, (m_N, n_N)$, and then estimate p as:

$$\hat{p} = \mu + \left(\frac{s^2 - \mu(1-\mu)\nu}{\frac{n-1}{n}s^2 + \mu(1-\mu)\left(\frac{1}{n} - \nu\right)} \right) \left(\frac{m}{n} - \mu \right) \tag{4}$$

where

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \frac{m_i}{n_i} \tag{5}$$

$$\nu = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \tag{6}$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{m_i}{n_i} - \mu \right)^2 \tag{7}$$

This is my own formula and this section describes how I found it. In the following sections I will prove through both artificial and real data that it is better than estimate (3). If all the control subjects have the same number of total attempts n_c, then $\nu = \frac{1}{n_c}$. If further $n_c = n$, then the denominator can be simplified to $\frac{n-1}{n}s^2$.

Estimate (3) is commonly used to estimate probability, because it is a maximum likelihood estimate of p. The reason why this is a MLE is that during the calculation of the MLE we take advantage of Bayes' postulate, which is an assumption that the a priori probability distribution $f(p)$ of p is a uniform distribution on the interval $[0, 1]$, i.e. $f(p) = 1$ on $[0, 1]$. In other words, before we observe m and n, we assume that p can be

anything from $[0,1]$ and all the numbers from this interval are equally probable. After we observe m and n, we can construct the following probability density function of p with the parameters m and n:

$$f(p \mid m,n) = \frac{f(p)\Pr(m \mid p,n)}{\int_0^1 f(p)\Pr(m \mid p,n)\,\mathrm{dp}} = \frac{\Pr(m \mid p,n)}{\int_0^1 \Pr(m \mid p,n)\,\mathrm{dp}} = \\ = \frac{\binom{m}{n}p^m(1-p)^{n-m}}{\int_0^1 \binom{m}{n}p^m(1-p)^{n-m}\,\mathrm{dp}} = \frac{m!(n-m)!}{(n+1)!}p^m(1-p)^{n-m} \tag{8}$$

This function has a maximum at $p = \frac{m}{n}$ and therefore equation (3) gives a maximum likelihood estimate of p. So obviously the greatest weakness of estimate (3) is the assumption that p is a priori uniformly distributed on the interval $[0,1]$ and a way to improve it is to construct a more realistic a priori distribution of p by using the data of other subjects (in our case using the $shc\%$ of other NHL teams).

I will assume that the a priori probability distribution of p is a beta distribution $f(p) = \mathrm{B}(\alpha,\beta)$, with the parameters α and β, which are unknown and must be estimated from the data of other subjects. The reason why I chose beta distribution is that it is the conjugated prior of a binomial distribution. Thus, for the NHL example this will mean that each team's individual probability of scoring from a shot comes from a beta distribution $\mathrm{B}(\alpha,\beta)$, where α and β are unknown and must be estimated from the $shc\%$ statistics across the league. So how can we estimate it? There are two ways. The first approach is using the method of moments, which estimates α and β in such a way that the sample mean and variance of our data set are equal to their theoretical values. The second possibility is using the maximum likelihood estimate, which requires numerical optimization, which is impractical, because it cannot be expressed by a formula. Therefore, I will be using the method of moments.

Let us assume that we have the statistics of N control subjects, i.e. we observed their numbers of successful attempts $m_1, m_2, \ldots, m_N$ and their total numbers of attempts $n_1, n_2, \ldots, n_N$. From this data α and β can be estimated by expressing the sample mean μ and sample variance s^2 by their theoretical values:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\frac{m_i}{n_i} = \frac{\alpha}{\alpha+\beta}$$
$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{m_i}{n_i}-\mu\right)^2 = \mathrm{Var}(p) + \mathrm{E}\left[p(1-p)\frac{1}{n_i}\right] =$$
$$= \frac{\alpha+\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \mathrm{E}\left[\frac{1}{n_i}\right]\frac{\alpha}{\alpha+\beta} - \mathrm{E}\left[\frac{1}{n_i}\right]\frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{\alpha+\beta+1} =$$
$$= \frac{\alpha\beta\left(\mathrm{E}\left[\frac{1}{n_i}\right](\alpha+\beta)+1\right)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

By solving this set of equations and estimating $\mathrm{E}\left[\frac{1}{n_i}\right]$ by ν we can estimate α and β as:

where μ, ν, and s^2 are given by the formulae (7). This gives the a priori probability distribution for the subject of interest $\mathrm{B}(\hat{\alpha}, \hat{\beta})$. The a posteriori probability distribution will also be a beta distribution and will have the parameters $\hat{\alpha} + m$ and $\hat{\beta} + n - m$. Hence the a posteriori distribution is $\mathrm{B}(\hat{\alpha} + m, \hat{\beta} + n - m)$. The probability of success from a single attempt of the subject of our interest can be estimated as the mean value of this distribution:

$$\hat{p} = \frac{\hat{\alpha} + m}{\hat{\alpha} + \hat{\beta} + n}$$

(9)

By inserting $\hat{\alpha}$ and $\hat{\beta}$ from (**??**) into equation (9) we get formula (4), presented at the beginning of this section.

When studying related work, I could not find any similar probability formula which can be applied in a situation when the control subjects have different amounts of data. I only found articles about estimating the parameters of beta-binomial distribution, such as (Martinez et al., 2015), which is closely related to my problem, but a beta-binomial model assumes that all subjects have the same number of attempts and, unlike my formula, it cannot be used irrespective of how much data we have for each subject.

Performance on artificial data

In this section I will show how my formula for estimating probability performs in comparison to the trivial estimators (i.e. the quotient of successful and total attempts $\frac{m}{n}$ and the average of the control subjects' probabilities). It is not necessary to read this section if you are not interested in the mathematical part of this article. This section shows what we can expect from my formula in different situations, which can best be illustrated by means of artificial data. While constructing the formula I made a simplifying assumption that the probability remains constant throughout the entire observation, which is not fully met by the NHL data. I also assumed that the theoretical values of the probabilities are a priori distributed with beta distribution. This section will illustrate how the formula performs when these assumptions are fully met.

To compare the formulae, I ran the following simulation in MATLAB.[4] I assumed 1,000 ice hockey teams, each of which has a constant probability of scoring from a shot throughout the entire simulation. These probabilities come from a beta distribution with the parameters $\alpha = 77$ and $\beta = 1530$. These teams will be used as control subjects. First, I generated the real probabilities of scoring from a shot p_i for each team. Then I randomly generated their total numbers of shots n_i from a uniform distribution on the interval from 1,000 to 10,000 and finally I simulated these shots and counted their numbers of goals m_i. Then I generated the real probabilities of scoring for 1,000 subjects of interest. The reason why I did not use only one subject of interest is that in order to estimate the mean absolute error of our predictions, several predictions must be made. For each of those subjects I simulated 4,000 shots, counted their number of goals, and tried to reconstruct their probability of scoring (which is already known to us) in three different ways. The estimate $\hat{p_1}$ is the traditional estimate – the quotient of successful and total attempts $\frac{m}{n}$. The estimate $\hat{p_2}$ is the league average, so it is the same for every team. The estimate $\hat{p_3}$ is computed by regressing $\hat{p_1}$ to the mean $\hat{p_2}$ with formula (4). A snippet of the simulated data can be seen in Tables 1 and 2.

Since we know the real shooting percentages for the subjects of interest, I was able to compute the mean absolute error for each type of estimate. As expected, the estimator with the smallest error was $\hat{p_3}$, with a mean absolute error equal to 0.002236. The second best was the traditional estimator $\frac{m}{n}$, with a mean

[4] The code can be seen in the appendix.

team	p	m	n
1	0.0473	215	4,458
2	0.0433	437	8,912
3	0.0351	154	4,873
4	0.0519	311	5,744
5	0.0451	341	6,972
...			
1000	0.0514	256	5,213

Table 1: Data generated during the simulation for 1,000 control subjects. Each subject had between 1,000 and 10,000 shots.

team	p	m	n	$\hat{p_1}$	$\hat{p_2}$	$\hat{p_3}$
1	0.0442	168	4,000	0.0420	0.0479	0.0436
2	0.0526	228	4,000	0.0570	0.0479	0.0545
3	0.0483	211	4,000	0.0528	0.0479	0.0514
4	0.0446	160	4,000	0.0400	0.0479	0.0422
5	0.0573	239	4,000	0.0598	0.0479	0.0565
...						
1000	0.0428	182	4,000	0.0455	0.0479	0.0462
Mean absolute errors of estimators $\hat{p_1}$ - $\hat{p_3}$				0.0027	0.0042	0.0022

Table 2: Data generated during the simulation for 1,000 subjects of interest. Each subject had 4,000 shots. The mean absolute errors of the estimators are presented at the bottom.

absolute error of 0.002678. The worst was the league average $\hat{p_2}$, with a mean absolute error of 0.004172. In order to verify that this difference is statistically significant I performed a paired t-test with the null hypothesis that the mean absolute errors of $\hat{p_3}$ and $\hat{p_1}$ are equal and the alternative hypothesis that the mean absolute error of the estimator $\hat{p_3}$ is smaller.

$$H_0 : |\hat{p_3} - p| - |\hat{p_1} - p| = 0$$
$$H_A : |\hat{p_3} - p| - |\hat{p_1} - p| < 0 \quad (10)$$

The p-value of this test was $4.48e-20$. This proves that the mean absolute error of the estimator $\hat{p_3}$ is smaller than the mean absolute error of the estimator $\hat{p_1}$.

Then I lowered the number of shots of subjects of interest to 1,000 and the control subject's total numbers of shots were generated from a uniform distribution on the interval from 100 to 1,000. Once again, the best estimator was $\hat{p_3}$, with a mean absolute error of 0.003432. Only this time the estimator $\hat{p_2}$ was better than $\hat{p_1}$ as their mean absolute errors were 0.004199 and 0.005730 respectively. In general, for small numbers of attempts of the subject of interest the estimator $\hat{p_2}$ is better than $\hat{p_1}$, while for large numbers of attempts $\hat{p_1}$ is more precise than $\hat{p_2}$. But in both situations $\hat{p_3}$ is better than both these estimators. Again, I performed a paired t-test to verify the statistical significance of the difference between the estimators $\hat{p_2}$ and $\hat{p_3}$. The p-value was $9.27e-15$, so the difference is again statistically significant.

Application to NHL data

In this section, I will apply my formula to NHL data from the past few seasons. Data used in this section can be found at (Maly, 2019) and matlab codes at (Maly, 2019). First, I want to answer the question from the first section. Does shot quality exist? Or is it better to predict a team's future results from the number of shots, as many people believe these days? This question can be answered by analysing the variance of the a priori beta distribution of shooting percentages $shc\%^{\star}$ across the league. If this variance is greater than 0, then shot quality exists and the greater the variance, the more important the shot quality is. But the problem is that we can only observe the variance of $shc\%$, while we want to draw conclusions about the variance of $shc\%^{\star}$.

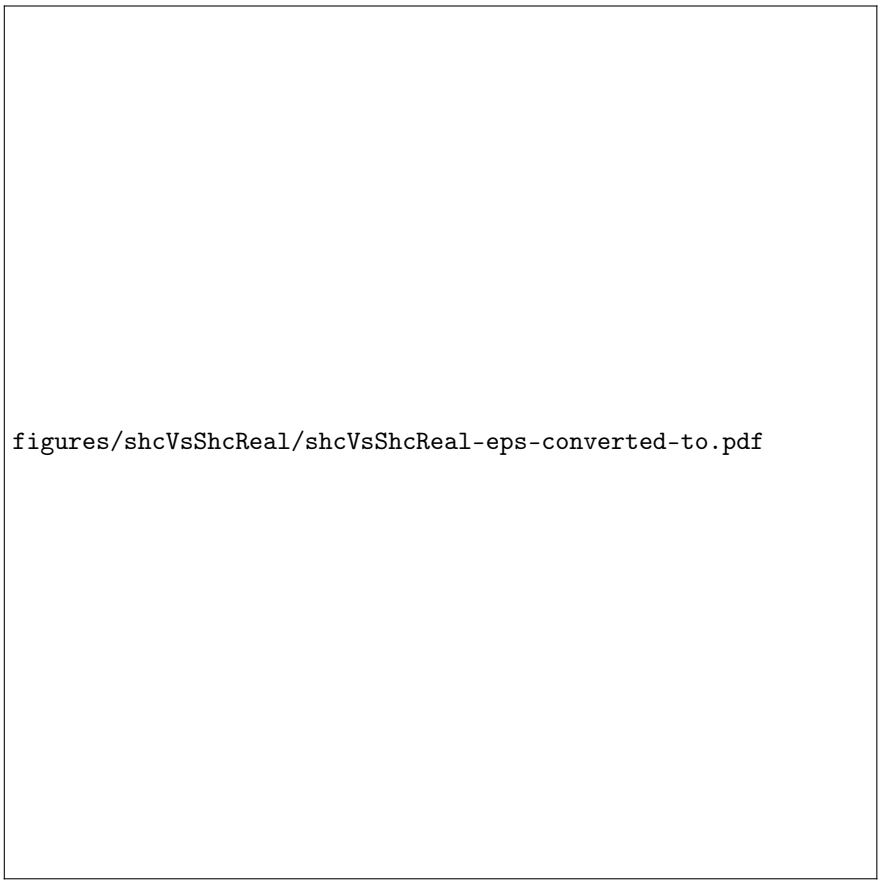

Figure 1: Probability desity functions of $shc\%^{\star}$ and $shc\%$.

As can be seen from equation (), the observed variance s^2 of the $shc\%$ statistic consists of two parts. The first element which contributes to the observed variance is the actual differences in shot quality among the teams (i.e. the variance of the a priori beta distribution). The second part of the observed variance is the variance of the binomial trials. Therefore, as we only have $shc\%$ statistics, the observed variance is increased by the randomness within this statistic and the less data we have for each subject, the greater this randomness is. For easier understanding, please see Figure 1, which depicts the difference in the probability density functions of $shc\%^{\star}$ and $shc\%$ after 1,000 shots. Using equation (), we can express the variance of

$shc\%^\star$ as:

$$\mathrm{Var}\left(shc\%^\star\right) = \frac{\mathrm{Var}\left(shc\%\right) - \mathrm{E}\left[\frac{1}{n_i}\right]\mathrm{E}\left[\frac{m_i}{n_i}\right]\left(1 - \mathrm{E}\left[\frac{m_i}{n_i}\right]\right)}{1 - \mathrm{E}\left[\frac{1}{n_i}\right]}$$

(11)

We can use this equation to estimate 95% confidence intervals of $\mathrm{Var}\left(shc\%^\star\right)$. I did it in two ways and they both provide very similar results. The first approach is to use a bootstrap[5] technique in the RStudio computer program. The second approach is to assume that $shc\%$ comes from a distribution which is close to normal, and that the observed μ and ν are close to their theoretical values, and construct the confidence interval as:

$$\mathrm{Var}\left(shc\%^\star\right) \in \left(\frac{\frac{N-1}{\chi^2_{0.975;N-1}} s^2 - \nu\mu\left(1-\mu\right)}{1-\nu}; \frac{\frac{N-1}{\chi^2_{0.025;N-1}} s^2 - \nu\mu\left(1-\mu\right)}{1-\nu} \right)$$

(12)

Now, assuming that the real Corsi shooting percentages $shc\%^\star$ are beta distributed with the parameters α and β and that this value is constant over the entire season, we can take the data for shots of all 31 NHL teams in the 2018 season[6] and from equations (7) and (11) estimate the parameters $\hat{\alpha} = 139.18$ and $\hat{\beta} = 2{,}687.54$ and the standard deviation of $shc\%^\star$ as 0.00407.

	Team	Goals	Corsi
1	Arizona Coyotes	199	4695
2	Boston Bruins	261	4870
3	Buffalo Sabres	193	4433
	...		
30	Washington Capitals	248	4496
31	Winnipeg Jets	268	4784

Table 3: Number of goals and Corsi shots of NHL teams in the 2018 season.

In order to get more data, we can use team statistics from the 2013/14 to 2017/18 seasons and consider the season's statistics of each team as one subject. In this way we get 151 subjects.[7] This data gives $\hat{\alpha} = 165.73$ and $\hat{\beta} = 3{,}287.43$. The estimated standard deviation of $shc\%^\star$ is 0.00367, with a 95% confidence interval $(0.00296; 0.00442)$ when computed by formula (12) and $(0.00302; 0.00434)$ when computed by the bootstrap method in RStudio. If the assumption that the team's shooting percentage does not change during the season seems too strong, we can take only the data for the first 1,000 shots of the season. From this data $\hat{\alpha} = 95.01$ and $\hat{\beta} = 1{,}880.82$ and the estimated standard deviation of $shc\%^\star$ is 0.00481, with a 95% confidence interval $(0.00314; 0.00647)$ when computed by formula (12) and $(0.00305; 0.00661)$ when computed by the

[5]This technique takes advantage of making several data sets out of one by randomly choosing into the new data sets from the original.

[6]As mentioned at the beginning, my data did not include overtime and play-off games.

[7]There were 30 teams in the NHL until last season, when the Las Vegas Golden Knights joined the league.

bootstrap method. The probability density functions of such an $shc\%^\star$ and its $shc\%$ after 1,000 shots are those depicted in Figure 1. On the basis of these experiments, we can conclude that the shooting percentage does exist; it is just difficult to predict it from past shooting percentages as there is a lot of randomness in it. I believe we could estimate it better by using different statistics, such as the number of passes before each shot, the number of opposing players between the goal and the player who is shooting, or the position from where the shot is taken.

Now we want to check if predicting the future Corsi shooting percentages of NHL teams using formula (4) is better in comparison to trivial estimators (i.e. the quotient of successful and total attempts and the league average). For this experiment I used the data of NHL teams from the 2013/14 to 2017/18 seasons again. I considered the season's statistics of each team as one subject, which gave 151 subjects. For each of them I took the first 2,000 shots of the season and tried to estimate the Corsi shooting percentage in the rest of the season using the remaining 150 teams as the control subjects. The first estimate $\hat{shc\%}_1$ is the quotient of successful and total attempts from the first 2,000 shots of the season of a given team. The second estimate $\hat{shc\%}_2$ is the average shooting percentage of the remaining 150 control subjects. The last estimate $\hat{shc\%}_3$ is a regression of $\hat{shc\%}_1$ to the mean $\hat{shc\%}_2$ using formula (4). For illustration, a snippet of the data is shown in Table 4. This table also shows the mean absolute errors for all three estimators. The worst is $\hat{shc\%}_1$, with a mean absolute error of 0.00561, followed by $\hat{shc\%}_2$, with an error of 0.00469, and, as expected, the best estimator was $\hat{shc\%}_3$, with an error of 0.00447.

		First 2,000 shots		Rest of the season			
	Team	Goals	Corsi	$shc\%$	$\hat{shc\%}_1$	$\hat{shc\%}_2$	$\hat{shc\%}_3$
1	Arizona 17/18	75	2000	0.0460	0.0375	0.0478	0.0436
2	Boston 17/18	98	2000	0.0568	0.0490	0.0477	0.0482
3	Buffalo 17/18	79	2000	0.0469	0.0395	0.0478	0.0444
...							
150	Washington 13/14	107	2000	0.0475	0.0535	0.0477	0.0501
151	Winnipeg 13/14	84	2000	0.0491	0.0420	0.0477	0.0454
Mean absolute error of estimators $\hat{shc\%}_1$ - $\hat{shc\%}_3$					0.00561	0.00469	0.00447

Table 4: Estimates of Corsi shooting percentages in the second part of the season from the initial 2,000 shots. The mean absolute errors of the estimators are shown at the bottom.

Shooting percentages (and probabilities in general) are not the only thing we can regress to the mean. In fact, if we know the types of a priori and a posteriori distribution, we can find formulae for regression to the mean of practically anything. The number of Corsi shots can be regressed in a similar way as shooting percentages. This technique was not discussed in the main part of the article; I will just briefly introduce it now. I assume that every team's number of shots in one game has a Poisson distribution with the parameter λ_i. I further assume that these λ_i's are constant throughout the entire season and that they are normally distributed with the parameters μ_s and σ_s^2. If we observe the total number of shots s in n games of a team of our interest and the total numbers of shots $s_1, s_2, \ldots, s_N$ of N control subjects in the same number of games, we can estimate the number of shots per game (SpG) of the team of our interest as:

$$\hat{spg} = \frac{\sum_{i=1}^{N} s_i}{nN} + \left(1 - \frac{\frac{\sum_{i=1}^{N} s_i}{N}}{\frac{1}{N-1}\sum_{i=1}^{N}\left(s_i - \frac{\sum_{j=1}^{N} s_j}{N}\right)^2}\right)\left(\frac{s}{n} - \frac{\sum_{i=1}^{N} s_i}{nN}\right) \tag{13}$$

The proof of this formula is beyond scope of this paper. Using the same data set as before, I tried to predict each team's average number of shots per game in the last 41 games of the season from the initial 41 games in three different ways. $\hat{spg}_1$ is the average number of shots per game of a given team in the first half of the season. $\hat{spg}_2$ is the average number of shots per game of all teams, except for the team of our interest. $\hat{spg}_3$ is a regression of $\hat{spg}_1$ to the mean $\hat{spg}_2$ using formula (13). The data I obtained is shown in Table 5. We can see that the estimator with the lowest mean absolute error was the one with the regression to the mean. An interesting thing to note is that shots were, on average, only regressed by less than 8 %, while shooting percentages were, on average, regressed by almost 59 %. This is why an individual team's Corsi statistics are more reliable than shooting percentages.

		First 41 games		Last 41 games			
	Team	Shots	SpG	SpG	$\hat{spg}_1$	$\hat{spg}_2$	$\hat{spg}_3$
1	Arizona 17/18	2333	56.90	57.61	56.90	55.35	56.78
2	Boston 17/18	2379	58.02	60.76	58.02	55.34	57.81
3	Buffalo 17/18	2201	53.68	54.44	53.68	55.37	53.81
...							
150	Washington 13/14	2260	55.12	52.22	55.12	55.36	55.14
151	Winnipeg 13/14	2353	57.39	55.95	57.39	55.34	57.23
Mean absolute error of estimators $\hat{spg}_1$ - $\hat{spg}_3$					2.07	3.17	2.01

Table 5: Estimates of Corsi shots per game (SpG) in the second half of the season from Corsi shots in the first half of the season. The mean absolute errors of the estimators are shown at the bottom.

The last experiment will join these two things together. I will try to estimate a team's number of goals per game (GpG) in the last 41 games of the season from their first 41 games using the same data set as before. I made four different estimates for each team. The first estimate, $\hat{g}_1$, is equal to the team's number of goals per game in the first part of the season. Estimate $\hat{g}_2$ is the average number of goals per game of all teams except the team we are making the prediction for. In the last two estimates I disassembled the goals from the first part of the season into shots and shooting percentages, applied regression to the mean to them, and finally reassembled them into goals again by multiplying these two numbers. Estimate $\hat{g}_3$ applies the regression only to shooting percentages, while $\hat{g}_4$ applies the regression to both shots and shooting percentages. The results are shown in Table 6. The first two estimators were significantly worse in comparison to the latter two. The best estimate was provided by regressing both shots and shooting percentages.

		First 41 games	Last 41 games				
	Team	GpG	GpG	$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_3$	$\hat{g}_4$
1	Arizona 17/18	2.171	2.683	2.171	2.647	2.473	2.468
2	Boston 17/18	3.195	3.171	3.195	2.640	2.969	2.958
3	Buffalo 17/18	2.171	2.537	2.171	2.647	2.393	2.398
...							
150	Washington 13/14	2.805	2.585	2.805	2.643	2.712	2.713
151	Winnipeg 13/14	2.537	2.683	2.537	2.645	2.649	2.641
Mean absolute error of estimators $\hat{g}_1$ - $\hat{g}_4$				0.311	0.285	0.261	0.258

Table 6: Estimates of goals per game (GpG) in the last 41 games of the season from the first 41 games. The mean absolute errors of the estimators are shown at the bottom.

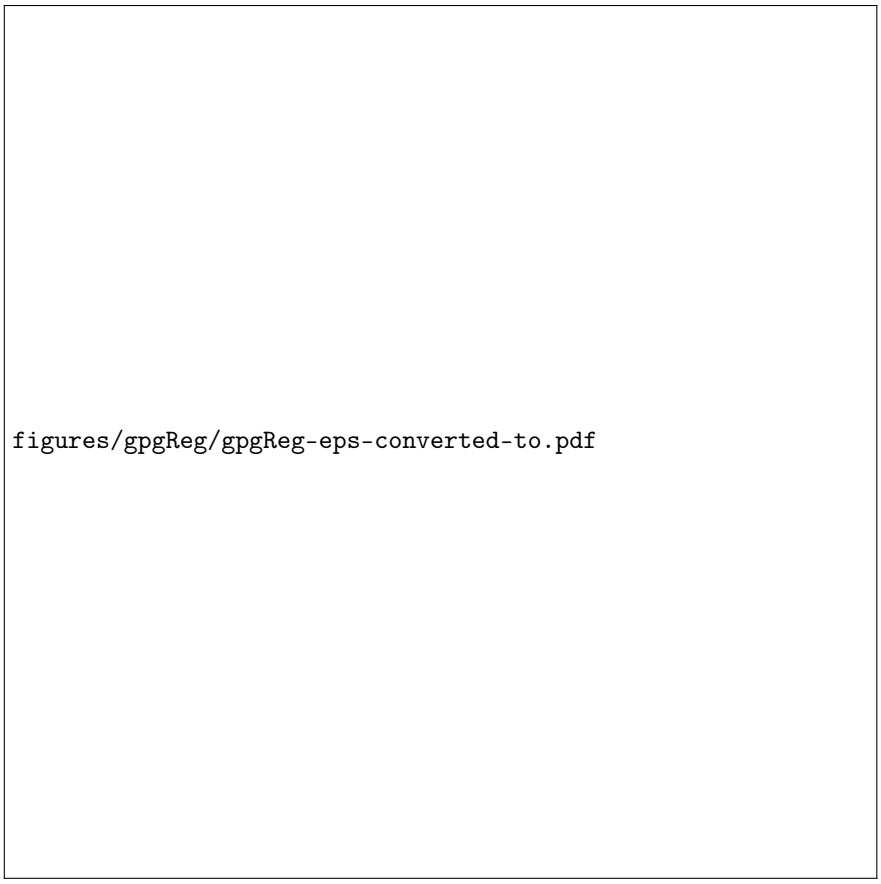

Figure 2: Linear regression between goals per game with and without regressing data from the first part of season to the mean.

Finally, I used linear regression to predict the number of goals per game in the second part of the season using data from the first part of the season. The results are shown in figure 2. The left-hand part depicts the prediction of GpG in the last 41 games from GpG in the first 41 games. The correlation coefficient of this model is 0.1317. The right-hand part describes the prediction of GpG in the last 41 games from GpG in the first 41 games with both shots and shooting percentages being regressed to the mean. The correlation coefficient of this model is 0.1682. This also confirms that the regression to the mean reduces the noise from the data and makes it more predictive. So next time you compare goaltenders on the basis of their saving percentages, you should regress their statistics to the mean using the techniques presented in this article, because a goaltender with 94 saves from 100 shots on goal is probably worse than another goaltender with 9,370 saves from 10,000 shots on goal, even though he has a higher saving percentage.

Summary

In this article I presented a formula which makes more accurate estimates of probability than a simple quotient of successful and total attempts. The idea behind this easy-to-use formula is to take advantage of the control subject's data. The key innovation is that it can be applied irrespective of how much data we have for each subject. The correctness of the formula was verified by applying it to artificial data. In the final section I applied the formula to NHL data, where I proved the importance of the shooting percentage, which

has been significantly underestimated lately. Finally, I showed how to reduce the noise from the statistics of goals and shooting percentages. This made these statistics more predictive for future observations.

Appendix - Matlab code

[language=Matlab, title=Simulated comparison of probability estimators]artificialTestBeta.m

References

Parameter estimation of the beta-binomial distribution: an application using the SAS software. (2015). *CIENCIA & NATURA*, *37*, 12. https://doi.org/10.5902/2179460X17512

NHL shots data 2013-2018 (Version V1). (2019). (Version). Harvard Dataverse. https://doi.org/10.7910/DVN/JSRN7H

NHL statistics. (2019). GitHub.