Are You Sure About This?

Dealing with Uncertainty in Football Metrics

Constantinos Chappas (Email: c.chappas@gmail.com / Twitter: @cchappas)

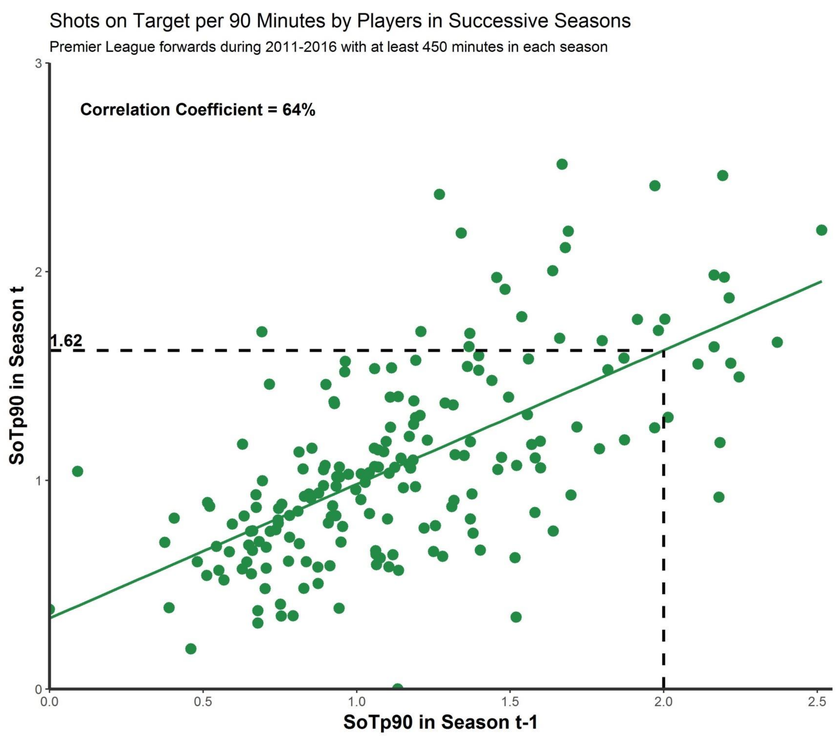
Prediction of Future Performance – Linear Regression Example

How many shots on target per 90 minutes will Harry Kane have in 2016/17?

Player Name	Season	Minutes	SoT*	SoTp90*
Harry Kane	2015/16	3368	75	2.00

***SoT**: Shots on Target

***SoTp90**: Shots on Target per 90 minutes played



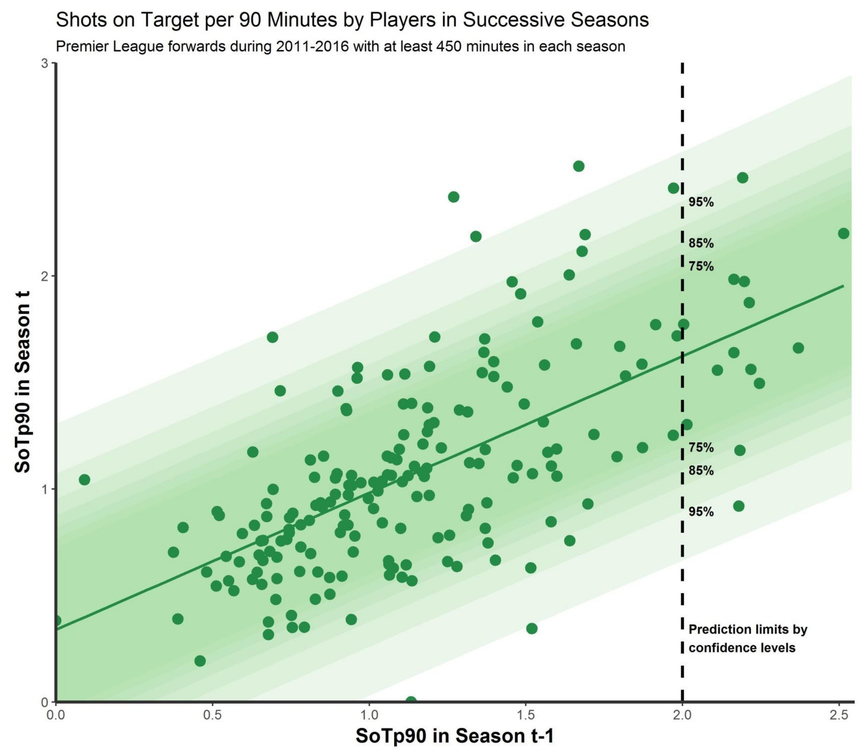
Potential Answers

1. Around 2.00. That's what he had during 2015/16, so it's reasonable to assume that he'll have similar figures.

2. Around 1.62. This is what the estimated regression line of a forward's performance against his previous season's figures suggests.

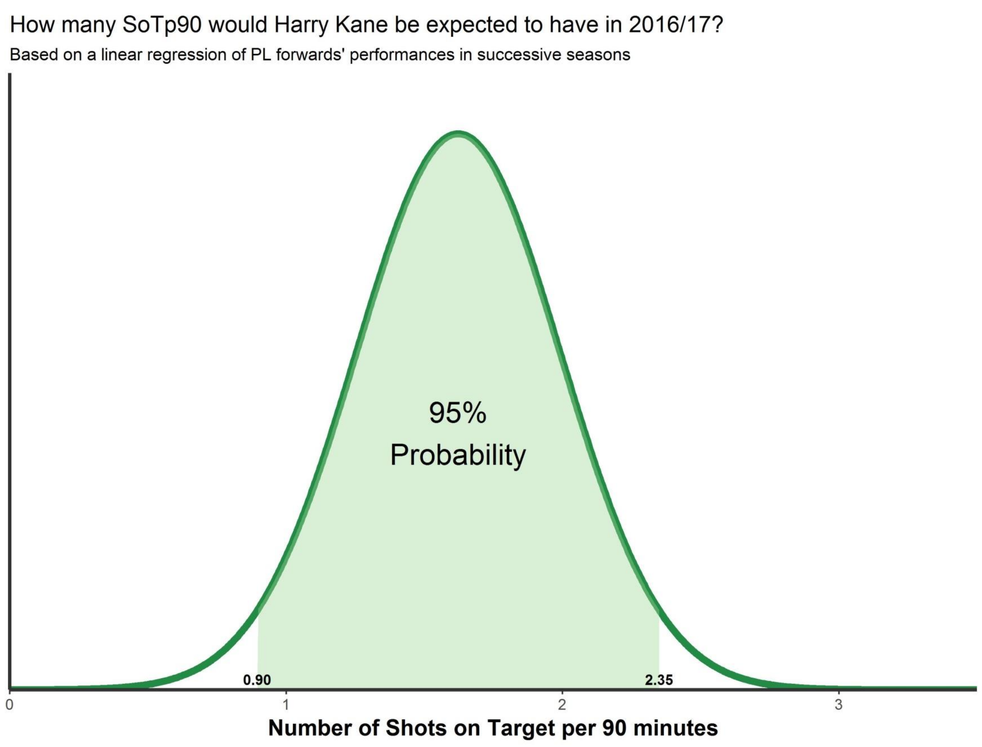
3. Uncertainty needs to be accounted for, before that question can be answered. Prediction intervals are suitable for this purpose:

A <u>prediction interval</u> is an estimate of the interval in which future observations will fall, with a certain degree of probability.



Harry Kane's **95% prediction interval** for SoTp90 is:

(0.90 – 2.35)



That means his performance is expected to be **between the 42nd (below the median) and 99th (top 1%) percentile of individual performances** within a season, at the 95% level of confidence!

The Bottom Line

This analysis is not specific to this particular metric. Any calculated statistic, especially when used to evaluate the past and indirectly forecast the future performance of a player is associated with a degree of uncertainty. That uncertainty should therefore be taken into consideration when projecting a player's future performance, no matter how simple or complicated the underlying metric is.

Why is this useful?

This analysis can be used to answer questions such as:

- How likely is it for a selected player to surpass a given level of performance?
- Comparing two players, how likely is it for one of them to produce better figures than the other? This can be done even if their past performances are based on different time frames, for example.
- How (un)likely was a player's extraordinary season in terms of particular metrics?

Further Considerations and Extensions

Other variables can be included in the analysis such as:

- Age effects
- Team / Managerial changes
- The impact of certain tactics

Expanding the model to include <u>additional variables</u> or making it more relevant by focusing on <u>certain subsets of the data</u> can alter the uncertainty in its predictions, depending on a number of factors like the usefulness of new information and the sample size.

Are You Sure About This?

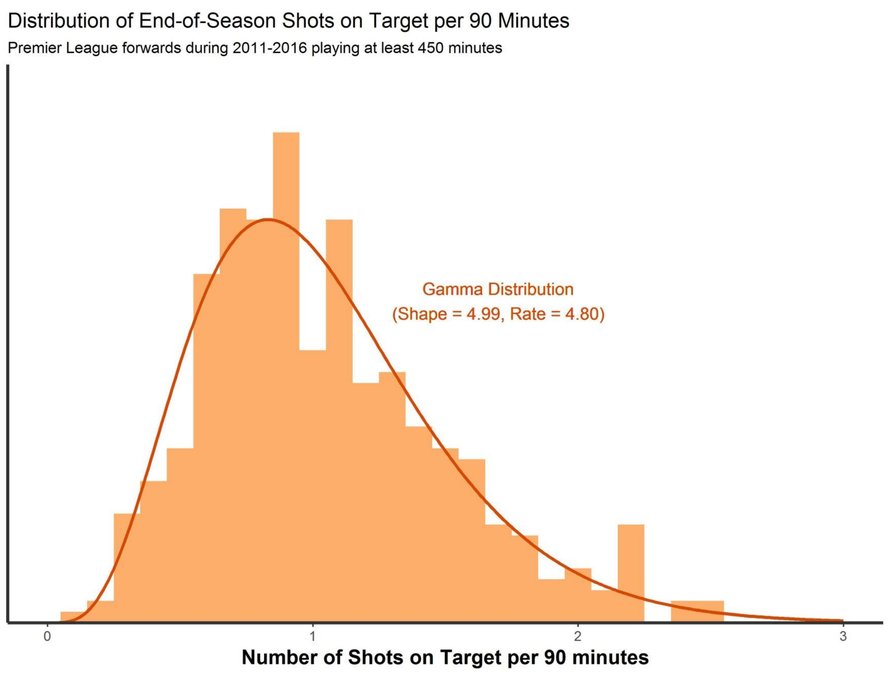
Dealing with Uncertainty in Football Metrics

Constantinos Chappas (Email: c.chappas@gmail.com / Twitter: @cchappas)

Prediction of Future Performance – The Bayesian Paradigm

How many shots on target per 90 minutes will Harry Kane have in 2016/17?

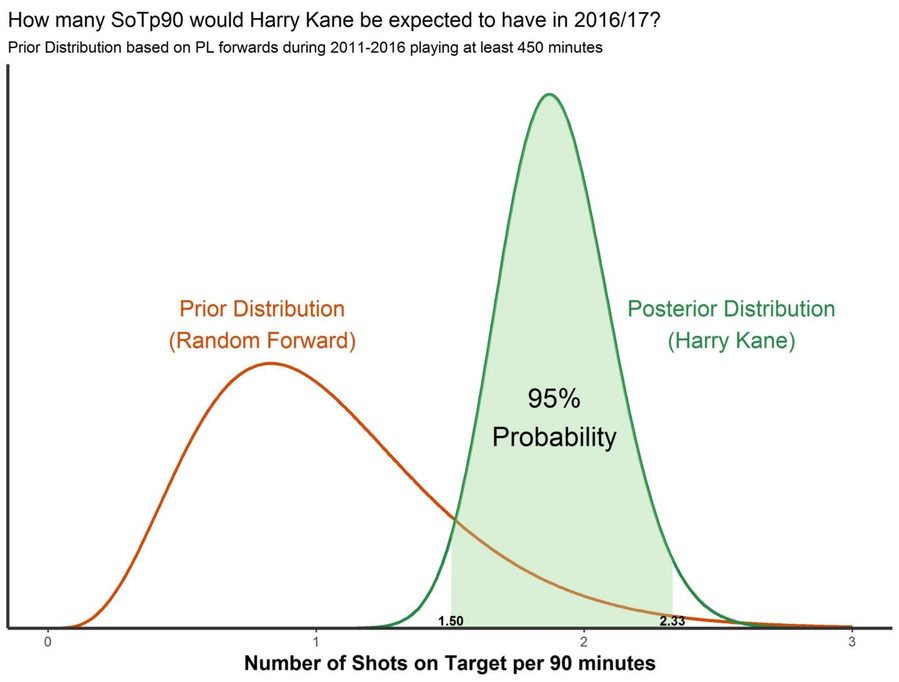
The starting point for this analysis is the distribution of the selected metric, depicted below by a histogram of SoTp90 values for the subset of players analysed.



This histogram can be used to estimate the theoretical distribution which, in this case, can be approximated by a **Gamma distribution**.

Harry Kane's **95% credible interval** for SoTp90 is:

(1.50 – 2.33)



The information contained in Harry Kane's 2015/16 performance in terms of this metric has shifted the beliefs about his expected figures upwards, compared to the prior distribution, while reducing the variance. With a 95% probability, his performance is expected to be **between the 83rd and 99th percentile of individual performances** within a season.

Final Thoughts

As in the case of the linear regression example, this approach can be extended to include significant variables that may affect the metric in question. It also provides a natural way of incorporating any new data that becomes available. But, it can also be more difficult to apply, especially when selecting appropriate prior distributions which may have strong influence on the final results.

Bayes Theorem and Inference

Using a probability theorem named Bayes Theorem, this approach allows for the analyst to combine prior beliefs about an event happening (Prior Probability) with new evidence (Likelihood), to get an updated probability (Posterior Probability). It can be written as:

$$\text{Posterior Probability} \propto \text{Likelihood} \times \text{Prior Probability}$$

The derivation of each of the posterior probabilities and therefore the posterior distribution can either be done algebraically or through simulations. From the posterior distribution, credible intervals can be calculated, which capture the uncertainty about any predictions.

In this example, the estimated Gamma distribution forms the analyst's prior beliefs. In the absence of any additional knowledge, the prior distribution is an estimate of Harry Kane's 2016/17 performance in terms of SoTp90.

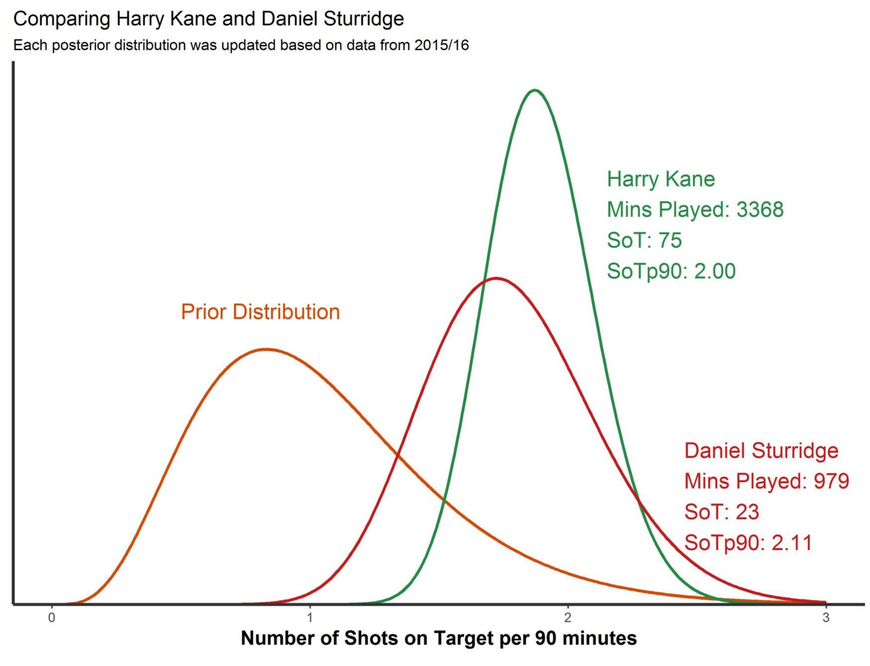
New evidence consists of the player's SoTp90 figures in the 2015/16 season:

Player Name	Season	Minutes	SoT*	SoTp90*
Harry Kane	2015/16	3368	75	2.00

***SoT**: Shots on Target
***SoTp90**: Shots on Target per 90 minutes played

Assuming that the number of shots on target in a match follows a Poisson distribution, the analyst can use the prior distribution and the available data through the Bayes Theorem to derive the posterior distribution of Harry Kane's 2016/17 SoTp90 figures.

One important aspect of the Bayesian approach is that the weight of additional information determines the amount of updating that occurs. In other words, if a player has only played a limited number of matches, the prior distribution will not be updated by as much as in the case of an ever-present player.



Even though Daniel Sturridge registered 2.11 SoTp90 during the 2015/16, the posterior distribution has not shifted by as much as Kane's because he only played for 979 minutes compared to 3368 in Kane's case. As a result, Kane is more likely to surpass Sturridge in terms of SoTp90 in 2016/17 with a probability of 62%.