How long before we can label a goalie a ‘bust’?

Sample size estimation for evaluating goalie performance

By Nicholas Romanchuk

One of the challenges we face when evaluating a goalie’s current performance and even more importantly, when predicting a goalie’s future performance, is how much save percentage can vary over the course of a season. We all know that average NHL players can perform at extremely high or low levels for multiple games in a row. This can make it hard when predicting if a current player’s performance will continue, or even when assessing if a given level of performance should be viewed as extraordinary. For example, when Columbus Blue Jackets goalie Elvis Merzlikins started the 2019-20 season saving just 88.4% of shots over his first nine games, an optimistic fan might have stated:

“It’s still early… nine games isn’t enough to judge Merzlikins. Let’s see how the rest of the season plays out before we jump to any conclusions.”

Despite Merzlikins disappointing performance, is nine games a large enough sample size to start making conclusions about a goalie? I think even the most pessimistic fans would say no. But if nine games isn’t enough, than how many are? In other words, if Merzlikins really is an above average goaltender, how many games before we can detect it? Well I don’t know either. Yet…

The good news is that statistics can be used to answer this question, or rather, statistics can be used to reasonably estimate the answer. In statistics there is a concept called ‘power’, which is the probability of identifying an effect (ex. the difference between a great goalie and a mediocre one) when the effect is actually there. The reason we should care about power, is because it can be used to estimate the required sample size when identifying an effect. As mentioned, a goalie’s performance can vary significantly game-to-game, this means that when we select a random subset of games from an exceptional goalie, there is a chance that the subset games will appear as ‘mediocre’. As you could probably guess, the better the goalie is (assuming the same variation) the less overlap there will be with a mediocre one, thus lowering the chance for selecting a sample that appears as ‘mediocre’ (increased power). Similarly, the larger the sample of games, the greater the probability that the subset will approximate the exceptionally goalie’s true save percentage (increased power). This relationship can be seen in Figure 1.

**Figure 1.** The estimated number of games needed to achieve the specified power when identifying goalies who’s true save percentages differs from the league average.

The above figure was created by identifying goalies who played >40 min in >26 games for one team during the 2019-2020 season. In total, 42 goalies were identified with an average save percentage of 91.1% and an average standard deviation of 5.3%. I then held power at a desired level (ex. 80%), and used the identified average save percentage and standard deviation to estimate the required number of games when detecting a goalie who’s true save percentage differs from the league average. As you can see in Figure 1, the better the goalie is (i.e. the higher their true save percentage) the less games are needed to achieve a given power. Similarly, increasing the number of games played by a given goalie will also increase power. So, what does all this mean for the casual hockey fan? How many games before we know if our new goalie is the next perennial Vezina winner? Well like most things in life, it depends.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.** Estimated number of games needed to achieve 80% power | | | | | | | |
| Save Percentage | 93.0% | **93.5%** | 94.0% | 94.5% | 95.0% | 95.5% | 96.0% | |
| # of Games | 63.0 | **40.2** | 28.2 | 21.1 | 16.5 | 13.4 | 11.3 | |

We can start by narrowing this analysis down a little. In research we typically aim for 80% power, so let’s focus on just the minimum number of games needed to achieve 80% power (Table 1). I think we can all agree that Carey Price has been one of the best goaltenders over the last decade, posting a career high in save percentage of 93.3% in the 2014-15 season. So, if we wanted to know if our team’s new goaltender is the second coming of prime Carey Price? According to Table 1. we should wait until **~40 games** have been played before making any conclusions. In other words, if our new goaltender really is the next Carey Price, and we run a statistical comparison using a random sample of 40 games from that goalie, there is an 80% probability we will correctly conclude that our new goalie is better than average. And if we fail to identify a difference between our goalie and league average after 40 games? Well there’s still a 20% probability that we could miss an effect, or perhaps our new goalie isn’t quite as good as Carey Price is, in which case more games are needed. So next time your new goaltender starts the season saving just 88.4% of shots over his first nine game, just sit back and relax, because we need at least 40 games before we’ll know anything meaningful.

P.S. Elvis Merzlikins finished the season the with a 92.3% save percentage and was voted to the 2020 NHL All-Rookie Team.

All data scraped from [www.hockey-reference.com](http://www.hockey-reference.com)

Python code used to perform analysis: <https://github.com/NickRomanchuk>

last year for one team The graph shows the is an assessment of estimate of our ability to identify an effect. So can we use power to determine if seven games enough to detect whether Demko is an above average goaltender? Yes. But it depends on two things, i) what do we consider to be adequate power? And ii) how much better than average is Demko?

it seven games enough Power can be influenced by a number of factors; however, for this analysis we will focus on only two:

*i)* the size of the effect

*ii)* the size of the sample

If we hold power at a desired level (ex. 80%), we can manipulate the size of the effect (i.e. simulate a goalie playing at a different save percentages) and record the corresponding change in sample size. In other words, we can use a formula to calculate the required games played, at a given save percentage and variability, to achieve a desired power.

The reason power matters in this situation, is because we can use it to estimate required sample sizes. To put it simple, If Demko’s true long-run save percentage is 93.3% us our ability to detect an effect, We need to make sure we have sufficient power (general rule of thumb is at least 80%) before conducting a comparison-based experiment, since lower power means we have a greater chance of missing an effect.

In the above graph we can see how sample size changes as a function of save percentage.

To perform this analyses I first needed to know what an average save percentage and average variation of save percentage is for an NHL goalie. I began by finding every goalie that played in a game in the 2019-20 NHL season. I then removed any games where the goalie didn’t face any shots. Since some goalies played for multiple teams, and I consider teams to be a confounding factor, I only used the data from the team where they played the most games. Finally, since I needed multiple games for each goalie to get an accurate representation of the variation in save percentage, I found the median number of games played (i.e. 26) and only included goalies above the median. This left 42 goalies included in the final analysis with a mean save percentage of 91.1% and an average save percentage standard deviation of 5.3%. From there I calculated the required sample size to achieve 80%, 85%, 90%, 95%, and 99% for save percentages 93-100%. So what’s the answer? How many games are need before we can start making conclusions about a goalie’s performance? Well, let’s take a look at the results.

Here comes the big limitation of my analysis, since I want these findings which to be used as a general guide when evaluating future goalies, I need to make some assumptions about how these future goalies are going to play.

What this graph is showing us is how sample size changes for different power levels and save percentages. As you might expect the better the goalie is playing (i.e. the higher the save percentage) the less games are needed to achieve a given power. Similarly, increasing the number of games played at a given save percentage will also increase power (i.e. increase our ability to identify an effect). We can summarize this information a more succinctly in the table below. This table should be used as a rough guide when discussing sample size and goalie performance. Since 80% power is typical used when calculating the required sample size for scientific studies, this is the column we will focus on. How many games are need before we can start making conclusions about a goalie’s performance? The answer is right there in the table. A goalie needs to save 93% of shots over a 40.6 game span before we have sufficient sample size to start drawing conclusions about his level of play. This may seem like a lot, and it certainly is at just shy of half a season, but that is exactly the point. We want a larger sample size at a high save percentage to ensure that we are confident in the conclusions we make. In other words, if we ran a statistical comparison between a goalie who played 40.6 games at a hypothetical 93±6.75% save percentage vs. the league average of 90.3%, we would have an estimated 80% probability to correctly conclude that the goalie is playing at an above average level. As seen in the table, as save percentage increase sample size also decreases.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimated Number of Games to achieve specified power | | | | |
| Save Percentage (%) | 80%  (# of games) | 85%  (# of games) | 90%  (# of games) | 95%  (# of games) | 99%  (# of games) |
| 93.0 | 40.6 | 47.0 | 55.7 | 70.1 | 101.5 |
| 94.0 | 22.2 | 25.6 | 30.2 | 37.8 | 54.5 |
| 95.0 | 14.3 | 16.4 | 19.3 | 24.0 | 34.2 |
| 96.0 | 10.2 | 11.64 | 13.6 | 16.8 | 23.7 |
| 97.0 | 7.86 | 8.88 | 10.3 | 12.6 | 17.6 |
| 98.0 | 6.37 | 7.13 | 8.18 | 9.9 | 13.7 |
| 99.0 | 5.38 | 5.97 | 6.79 | 8.13 | 11.1 |
| 100.0 | 4.68 | 5.16 | 5.81 | 6.89 | 9.26 |

It is my hope that hockey fans can use this table next time they curious about a goalies play and wondering….

All data scraped from [www.hockey-reference.com](http://www.hockey-reference.com)

Python code used to perform analysis can be found at: GitHub.com