Small Sample Methods for Goalie Evaluation

Use all your data and quantify uncertainty

rdj

March 11, 2017

fumble.to.victory@gmail.com

What do I mean by small sample methods?

Essence of the problem

When samples are small enough, random variation can overwhelm the signal

Applying that to goalies

We don't have a particularly small sample of shots, but we have a low event rate and small sample of goals

Data herein originates from corsica.hockey unless otherwise noted (thank you!).

Measuring goalies is hard...and I don't have all the answers

Goalies are tough to evaluate

- It's tough to separate goalies from team effects, because they are always on the ice
- How do we disentangle randomness from actual performance changes?
- We have a small sample of goals

We can adjust for randomness and the information we can identify

- Goalies don't control the offense side of wins*
- Goalies don't control how many shots they face*
- Goalies don't control the quality of those shots*

*Even these are only partial truths, and very debatable

Two contrary approaches to team effects and randomness

Restrict to more comparable events

- 5v5 save percentage
- 5v5 high danger (HD) save percentage

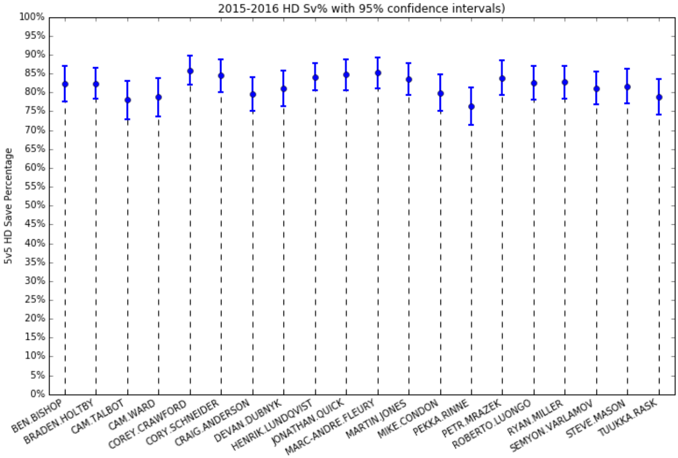
Try to account for important factors

- Adjusted save percentage (Adj.Sv%)
- Adjusted Goals Saved Above Average (adjGSAA/60)
- Expected Goals Against (xGA)

As we restrict our sample, we lose certainty

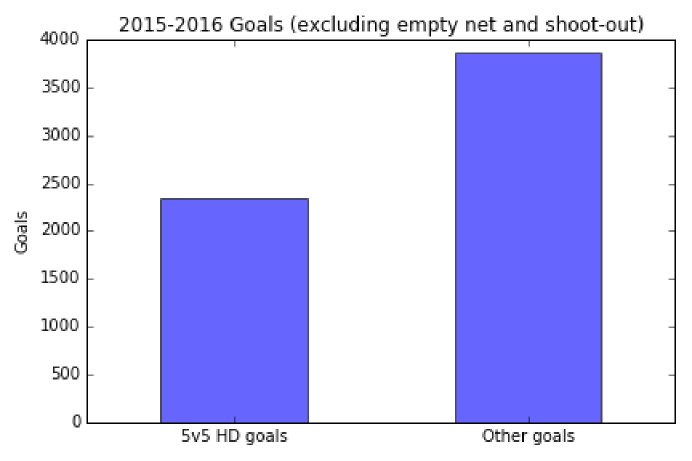
5v5 HD Sv%

Confidence intervals may be boring, but they *are* handy.



As we restrict our sample, we lose information

If differences in performance across situations vary due to the goalies themselves, then we lose that when we look at one type of shots.

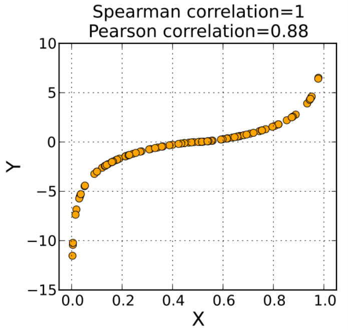


How to compare different metrics?

We can evaluate with Spearman correlation in consecutive years

- This is correlation on ranks, in [-1, 1] space
- This only makes sense if the statistic reflects goalie quality, and if better goalies are actually consistently better than worse goalies.

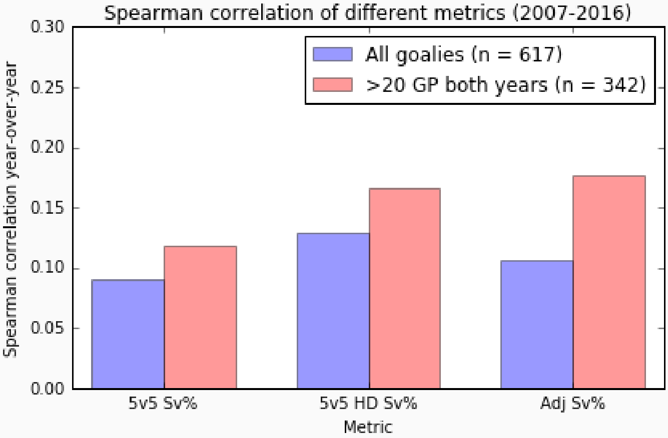
Example of Spearman correlation*



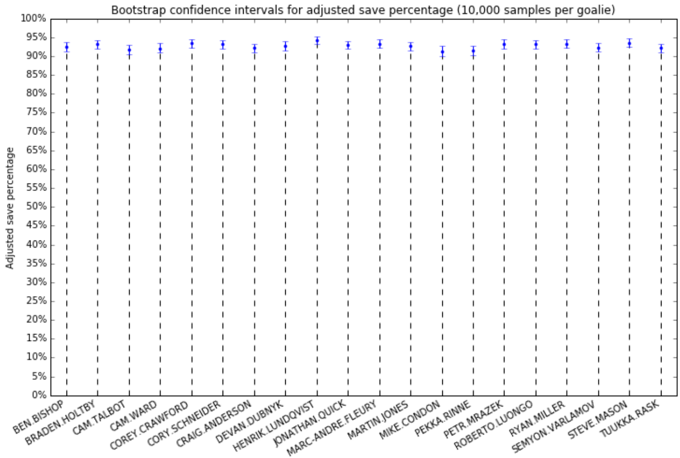
*By Skbkekas (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Adj.Sv% isn't going to end the debate

- Looking at year-over-year correlation of our ranks
- Maybe LD and MD shots are a bit too noisy for goalies with small samples



On the plus side, Adj.Sv% has tight confidence intervals



There isn't always a formula for confidence intervals

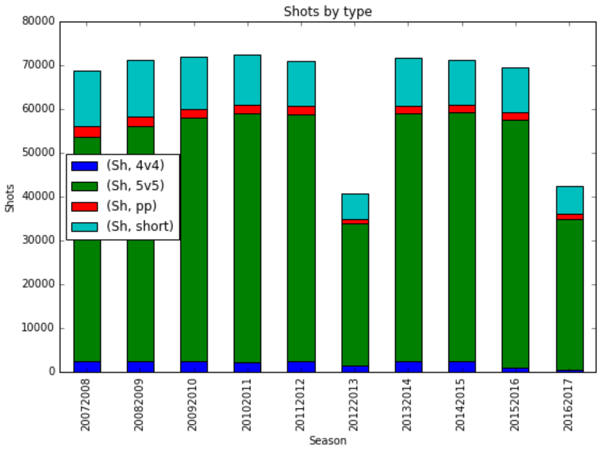
We can calculate bootstrap confidence intervals

Bootstrap sampling is simple, it just takes cycles

1. Sample randomly (with replacement) from the shots that a goalie has faced
    - Sample the same amount as the goalie faced
2. Calculate the statistic of interest (e.g. adjusted save percentage)
3. Save this value
4. Repeat steps 1-3
5. Now you have a bootstrap distribution!
    - Use percentiles to establish confidence intervals

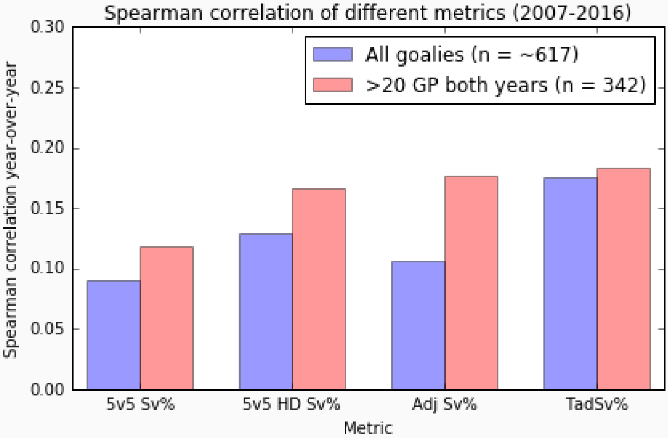
We can improve on Adj.Sv%

If we only look at 5v5 shots, we lose a lot of information
...and I don't like to give up information



Why not at least include goals when shorthanded?

- Instead of three weights, we have six (3 locations * 2 situations)
- I've been calling this Twice Adjusted Save Percentage
- It seems to perform pretty well



Why stop there?

- We could include 4v4 goals and (allowed) shorthanded goals, for example
- But now we're up to 4 * 3 = 12 weights, and not many goalies have much data for all of those
  - We already lose a few goalies with 6 weights
- Tactic (for the appendix): Apply Bayesian smoothing

My hobby: Second guessing popular opinion

Season	Vezina Winner	TadSv% Rank	TadSv% Winner
2007-08	Martin Brodeur	7	Jean-Sebastien Giguere
2008-09	Tim Thomas	1	Tim Thomas
2009-10	Ryan Miller	3	Henrik Lundqvist
2010-11	Tim Thomas	1	Tim Thomas
2011-12	Henrik Lundqvist	1	Henrik Lundqvist
2012-13	Sergei Bobrovsky	2	Jimmy Howard
2013-14	Tuukka Rask	6	Henrik Lundqvist
2014-15	Carey Price	1	Carey Price
2015-16	Braden Holtby	8	Brian Elliot

Being a top 10 goalie is very precarious

4 metrics * 5 years * 10 goalies per year

- Each stat has at least 31 unique goalies
- 47 distinct goalies represented (out of 80!)
- 11 different #1 goalies
- 5v5 HD Sv% is the most discordant: contains 10 of the 11 single-metric single-year goalies
- Most represented goalies:
  - Henrik Lundqvist: 16 of 20 lists
  - Cory Schneider: 14 of 20 lists
  - Corey Crawford: 12 of 20 lists
  - Mike Smith: 11 of 20 lists

Metrics

- 5v5 HD Sv%
- Ad.Sv%
- Tad.Sv%
- Adj.FSv%

Years

- 2011-2012 to 2015-2016

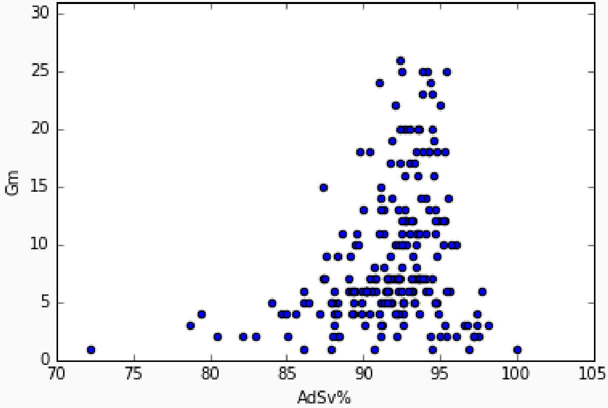
Criteria

- 1000 MP

Conditioning our statistic on sample size

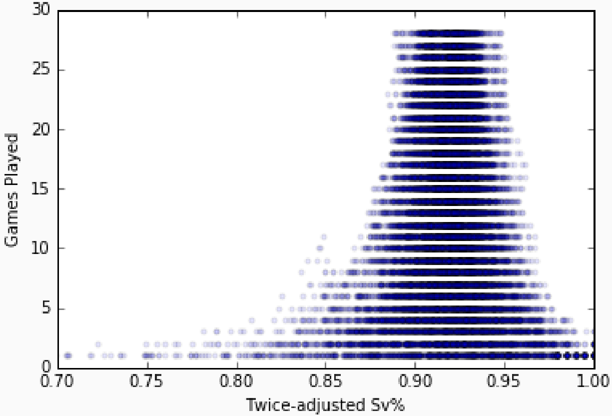
For evaluating playoff performance (or backup goalies during the regular season) randomness is an especially poignant concern

Shown here: Every playoff performance I have data for



Finding P(Result this good | games played)

- Compare N-game stat versus N-game distribution
- It's too bad we don't have a bigger sample
- ...but wait! We can fake one with regular season streaks



Best and worst playoffs

Rank	Goalie	Year	GP	TadSv%	P(Higher|GP)
1	Mike Smith	2012	16	0.948	1.90%
2	Ilya Bryzgalov	2006	12	0.952	1.90%
3	Antero Niittymaki	2011	2	0.980	3.90%
4	Tuukka Rask	2013	22	0.939	4.00%
5	Jonathan Quick	2012	20	0.938	5.40%
..	..	..	..	..	..
217	Martin Gerber	2006	7	0.869	97.6%
218	Jonathan Quick	2010	6	0.866	97.8%
219	Henrik Lundqvist	2006	3	0.835	98.6%
220	Jean-Sebastien Giguere	2006	9	0.869	99.3%
221	Marc-Andre Fleury	2010	13	0.871	100%

Data from War-on-Ice, through 2016. Zone locations differ from Corsica. Niittymaki is a bit of an odd case, with two relief appearances.

Summary

Use all your data and quantify uncertainty

- More data is better than less
- Quantify random variation through confidence intervals
- Calculate bootstrap confidence intervals if necessary
- Condition on variance to compare different sample sizes
- Good metrics should be somewhat time consistent

Thank you for your time!
Questions?

rdj
fumble.to.victory@gmail.com
oddacious.github.io
twitter.com/gentleputsch

Bayesian smoothing helps us compare goalies with fewer shots

Bayesian updating of binomial data

$$\alpha = \text{prior expectation of shots}$$

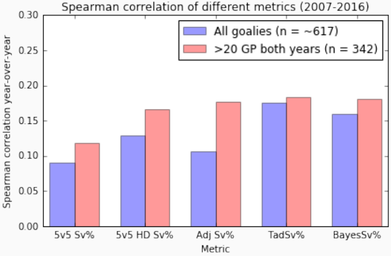
$$\beta = \text{prior expectation of goals}$$

$$\text{posterior} \sim Beta(\alpha, \beta)$$

$$\text{posterior mean} = \frac{\alpha + saves}{\alpha + \beta + shots}$$

I do this across zones and situations and use the league average as my prior, with a weight for sensitivity to the prior

Applied to goalies

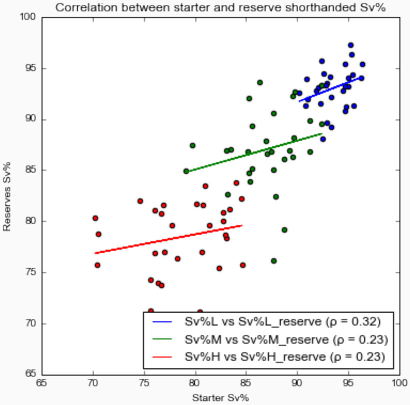


Results are not particularly better (perhaps the weights are too strong), but this technique does allow us to include all goalies.

Team effects are real

Team effects impair all metrics

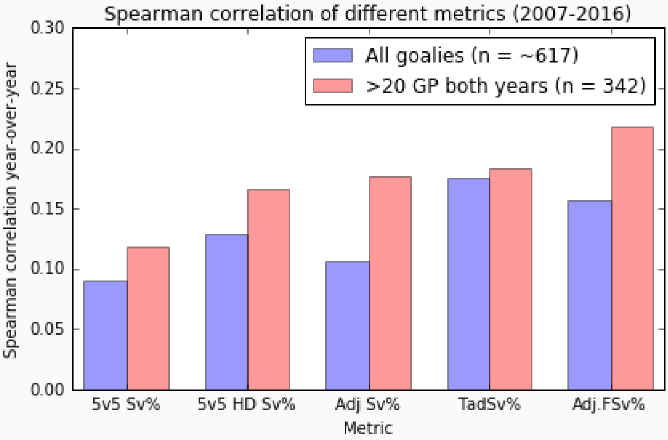
The presence of high correlation between goalies and their backups suggests that adjusting for both shot location and situation does not entirely isolate team effects*



*Alternative: The teams who find the best starters also find the best backups
Data for this chart from War-on-Ice

Adj.FSv% performing very well on large samples

- Although, this is its build sample
- Regardless, I think this is the approach we need to take



Correlation between goalies on the same team (GP > 20)

- If this is very high, our statistic is capturing a team effect