Optimal Drafting in Hockey Pools

Amy E. Summers, Tim B. Swartz and Richard A. Lockhart *

Abstract

This paper considers the selection of players in a common hockey pool. The selection strategy is based on the statistical modelling of the points scored by hockey players and the number of games played. A key feature of the approach is the introduction of an optimality criterion which allows drafting to be done in real time. A simulation study suggests that the approach is superior to a number of ad-hoc strategies.

Keywords : connected graphs, limit theorems, National Hockey League, Poisson distribution, simulation.

1 INTRODUCTION

In Canada and in some of the northern states of the USA, there is a general excitement each spring, not only with the change in weather but also with the advent of the Stanley Cup Playoffs. A popular activity amongst friends and co-workers is the participation in a Stanley Cup Playoff pool.

A common Stanley Cup Playoff pool proceeds along the following lines. Amongst K drafters (participants in the pool), an order is determined from which drafter 1 selects a hockey player from any of the 16 teams in the National Hockey League (NHL) that have qualified for the Stanley Cup Playoffs. Drafter 2 then selects a player but is not allowed to select the player chosen by drafter 1. The drafting continues until the first round is complete (i.e. drafter K has made a selection). The order of drafting is then reversed for the second round, and the process continues for m rounds. At the completion of the draft, each drafter has selected a *lineup* of m players where each player accumulates points (i.e. goals plus assists) during the playoffs. The drafter whose lineup has the greatest number of total points is declared the winner. Typically, a monetary prize is given to the winner.

Now, the question arises as to how one should select hockey players. Clearly, players who are able to generate lots of points in a game have some appeal. However, this must be tempered by the strength of a player's team. For example, a weak team is likely to be eliminated early in the Stanley Cup Playoffs, and therefore, a good player on a weak team may not be as valuable as a weaker player on a stronger team. One might also consider the effect of the "eggs in one basket" syndrome. By choosing players predominantly from one team, a drafter's success is greatly influenced by the success of the team. It is fair to say that it is not obvious how to best select hockey players in a draft.

Although we have not come across any previous work concerning drafting in hockey pools, there is a considerable literature on the related problems of rating sports teams

and predicting the outcome of sporting events. For example, Berry, Reese and Larkey (1999) compare players of different eras in the sports of professional hockey, golf and baseball. Carlin (1996) uses point spreads to estimate prediction probabilities for the NCAA basketball tournament. More generally, the volume edited by Jay Bennett (1998) covers a wide range of topics related to statistical issues in sport.

This paper considers a statistical approach to the player selection problem in playoff hockey pools. More detail on all aspects of the proposed approach can be found in Summers (2005). In section 2, some statistical modelling is proposed for the number of points scored and the number of games played by hockey players. Together with Sportsbook odds, subjective probabilities, connected graphs, Newton-Raphson optimization and simulation, expectations concerning the total points by lineups are obtained. A key point is that the expectations are calculated in advance of the draft so that drafting may be done in real time. Friends and co-workers may not be entirely understanding if they need to wait long periods of time for a drafter to make a selection. In section 3, an optimality criterion is introduced for the selection of hockey players, and the optimality criterion is a simple function of the expectations derived in section 2. In section 4, we conduct a simulation study to assess the proposed selection strategy against some common ad-hoc strategies. We observe that the proposed selection strategy is arguably the best strategy. The results of an actual Stanley Cup playoff pool using our methodology are reported in section 5. We conclude with a short discussion in section 6.

2 STATISTICAL MODELLING

2.1 Distribution of Points

Consider a hockey player who is on the ice and whose team is in possession of the puck. We let p denote the probability that the team scores and the player accumulates a point during the possession. If there are n such possessions during the game, then it may be reasonable to model the number of points X obtained by the player during the game as a Binomial(n, p) random variable. Naturally, this is a simplification which assumes independence of possesions, does not take account of power plays, does not take account of teammates and opposition, etc. When n is large and p is small, we apply the standard approximation

$$X \sim \text{Poisson}(\theta)$$

where $\theta = np$. The Poisson distribution has been previously used for modelling goals scored in hockey games and has provided a good fit (Berry 2000). The Poisson is preferable to the Binomial as it is characterized by a single parameter.

A straightforward method for estimating θ is

$$\theta = \frac{\text{number of points obtained by the player in the regular season}}{\text{number of games played by the player in the regular season}}. \tag{1}$$

A drafter may choose to modify the estimator (1) in various ways such as assigning a greater weight to recent performance. From this point on, we assume that θ has been determined for each eligible player in the draft.

Consider now the number of points Y_{ki} obtained during the entire Stanley Cup Playoffs by the ith player selected by drafter k. If the player's θ value is θ_{ki} and he plays in g_{ki} playoff games, it follows that

$$Y_{ki} \mid g_{ki} \sim \text{Poisson}(g_{ki}\theta_{ki}).$$

4

We then define the total number of points obtained by lineup k

$$T_k \equiv Y_{k1} + \cdots + Y_{km}$$

and note that the unconditional distribution of T_k (i.e. unconditional with respect to $g_{k1}, \ldots, g_{km}$) is the sum of m dependent Poisson mixture distributions. Using the conditional expectation formulae and assuming that Y_{ki} and Y_{lj} are conditionally independent given g_{ki} and g_{lj} for all k, i, l, j, we obtain

$$E(T_k) = \sum_{i=1}^{m} \theta_{ki} E(g_{ki}), \tag{2}$$

$$Var(T_k) = \sum_{i=1}^{m} \theta_{ki} E(g_{ki}) + \sum_{i=1}^{m} \theta_{ki}^2 Var(g_{ki}) + 2 \sum_{i<j} \theta_{ki} \theta_{kj} Cov(g_{ki}, g_{kj}) \tag{3}$$

and

$$Cov(T_k, T_l) = \sum_{i=1}^{m} \sum_{j=1}^{m} \theta_{ki} \theta_{lj} Cov(g_{ki}, g_{lj}). \tag{4}$$

The expressions (2), (3) and (4) are used in section 3 to determine an optimal player selection strategy. The terms in the expressions are all known except for the expectations involving the number of games played, and these expectations are obtained via simulation in section 2.2. Admittedly, the conditional independence assumption may be somewhat questionable in the case of two players from the same line on the same team.

2.2   Distribution of Games

If we are able to assign probabilities for each of the playoff teams defeating each of the other playoff teams in all potential playoff games, then it is possible to simulate the entire NHL playoffs. The simulation matches teams in the first round as dictated by the
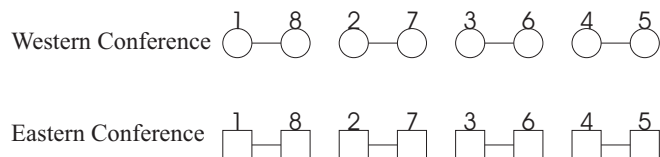
Figure 1: First round format.

actual NHL playoff matchups, and depending on which teams win the simulated series, team matchups are determined for the second round, and so on, leading to the two Stanley Cup finalists. By repeating the playoff simulation many times, the expectations appearing in the expressions (2), (3) and (4) can be estimated accurately using simple averages.

We now use Sportsbook odds, subjective probabilities, connected graphs and Newton-Raphson optimization to assign the individual game probabilities. We begin with the preliminary task of assigning the *series* probabilities $P(i,j)$, $i \neq j$ where $P(i,j)$ denotes the probability that team i defeats team j in a best of seven series. A good way to assign series probabilities is to utilize Sportsbook odds as these are known to reflect the collective wisdom of the betting public (Insley, Mok and Swartz 2004). Sportsbook odds are reported in the form $Odds(i,j) : 1$ where $Odds(i,j)$ is the payout in dollars on a winning one dollar bet on team i and $P(i,j) = 1/(Odds(i,j) + 1)$. Observe that $Odds(i,j) \cdot Odds(j,i) = 1$. A difficulty is that prior to the Stanley Cup Playoffs (when the draft takes place), Sportsbook odds are only available for actual first round matchups, and not all of the hypothetical matchups that may occur later in the playoffs. Figure 1 gives the layout for the first round of the playoffs where the numbers correspond to the seedings of the 8 teams in both the Eastern and Western conferences. Each line in the graph indicates that the probability $P(i,j)$ between the two connected teams is available via Sportsbook odds.

To complete the unspecified probabilities $P(i,j)$ where $P(j,i) = 1 - P(i,j)$, we use

the drafter's subjective hockey knowledge. The bold lines in Figure 2 represent two of the drafter's subjective probabilities. Fortunately, it is not necessary for the drafter to assign all of the $\binom{16}{2} - 8 = 112$ remaining series probabilities. By assuming that the probabilities $P(i, j)$ are "transitive", we can use the Sportbook odds and the drafter's subjective probabilities to determine the probabilities of other matchups. For example, referring to Figure 2, $P(8E, 7W)$ can be obtained by following the line from $8E$ to $7W$ whereby

$$Odds(8E, 7W) = Odds(8E, 2W) \cdot Odds(2W, 7W).$$

The goal is to create a connected graph; a graph is connected if there exists a path between each pair of vertices. In this situation, the drafter must specify a minimum of 7 subjective probabilities in order to connect the graph and thereby complete all probabilities $P(i, j)$, $i \neq j$.
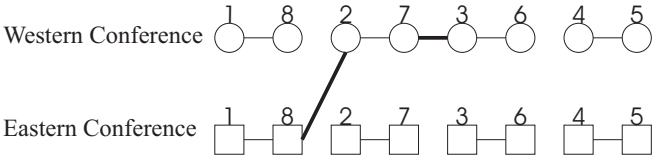
Figure 2: Drafter's probabilities in bold.

Of course, if one is going to complete the probabilities via transitivity, there should not exist different paths that lead to different probability calculations. The drafter should be "transitivity coherent" in his or her subjective probability assignments. For example, $P(1, 4) = 0.4$ and $P(4, 5) = 0.5$ imply $Odds(1, 4) = 1.5$ and $Odds(4, 5) = 1.0$ which is transitivity incoherent with respect to $P(1, 5) = 0.8$ (i.e. $Odds(1, 5) = 0.25$). Note that transitivity is a strong assumption that is not always applicable in sports, particularly when style of play and individual player matchups are paramount. However, we

believe that transitivity is fairly sensible in hockey. If a drafter is adamant that transitivity is inappropriate with respect to some teams, the drafter should simply assign the corresponding entries $P(i, j)$ rather than impute the $P(i, j)$ entries via transitivity.

Having determined the series probabilities $P(i, j)$, it is now necessary to calculate individual game probabilities. To this end, let p_{ij} denote the probability that team i defeats team j in a game on neutral ice and let ϵ denote the home ice advantage. Therefore, with respect to team i defeating team j, $p_{ij} + \epsilon$ is the probability of a home win, and $p_{ij} - \epsilon$ is the probability of an away win. We obtain an estimate of ϵ common to the league by considering the results of the regular season where

$$\epsilon = \frac{\text{(number of home team wins)} + \frac{1}{2} \text{ (number of tied games)}}{\text{total number of regular season games}}. \tag{5}$$

Keeping in mind the 2-2-1-1-1 home/away format for NHL playoff series, let team i denote the higher seeded team for whom the first two games in the series are played at home. Then

$$\begin{aligned} P(i, j) &= \text{Prob}(i \text{ wins in 4 games}) + \text{Prob}(i \text{ wins in 5 games}) \\ &+ \text{Prob}(i \text{ wins in 6 games}) + \text{Prob}(i \text{ wins in 7 games}) \end{aligned} \tag{6}$$

where

$$\text{Prob}(i \text{ wins in 4 games}) = (p_{ij} + \epsilon)^2 (p_{ij} - \epsilon)^2, \tag{7}$$

$$\begin{aligned} \text{Prob}(i \text{ wins in 5 games}) &= 2(p_{ij} + \epsilon)^2 (p_{ij} - \epsilon)^2 (1 - (p_{ij} + \epsilon)) \\ &+ 2(p_{ij} + \epsilon)^3 (p_{ij} - \epsilon)(1 - (p_{ij} - \epsilon)), \end{aligned} \tag{8}$$

$$\begin{aligned} \text{Prob}(i \text{ wins in 6 games}) &= 3(p_{ij} + \epsilon)(p_{ij} - \epsilon)^3 (1 - (p_{ij} + \epsilon))^2 \\ &+ (p_{ij} + \epsilon)^3 (p_{ij} - \epsilon)(1 - (p_{ij} - \epsilon))^2 \\ &+ 6(p_{ij} + \epsilon)^2 (p_{ij} - \epsilon)^2 (1 - (p_{ij} + \epsilon))(1 - (p_{ij} - \epsilon)), \end{aligned} \tag{9}$$

and

$$
\begin{aligned}
\text{Prob}(i \text{ wins in 7 games}) \;=\; & (p_{ij} + \epsilon)(p_{ij} - \epsilon)^3(1 - (p_{ij} - \epsilon))^3 \\
+\; & (p_{ij} + \epsilon)^4(1 - (p_{ij} - \epsilon))^3 \\
+\; & 9(p_{ij} + \epsilon)^2(p_{ij} - \epsilon)^2(1 - (p_{ij} + \epsilon))^2(1 - (p_{ij} - \epsilon)) \\
+\; & 9(p_{ij} + \epsilon)^3(p_{ij} - \epsilon)(1 - (p_{ij} + \epsilon))(1 - (p_{ij} - \epsilon))^2.
\end{aligned}
\tag{10}
$$

Substituting (7), (8), (9) and (10) in (6), we have an equation with only one unknown p_{ij}, and p_{ij} can be calculated using the Newton-Raphson algorithm. Substituting p_{ij} back into (7), (8), (9) and (10) gives us the four probabilities corresponding to team i winning the series against team j. Noting that $p_{ji} = 1 - p_{ij}$ and using formulae (7)-(10), we can directly calculate the four probabilities corresponding to team j winning the series against team i.

In summary, probability distributions with 8 outcomes are obtained for all 120 possible series. Using these distributions, the playoffs are simulated in advance of the draft to estimate the expectations appearing in the expressions (2), (3) and (4). The expressions (2), (3) and (4) can then be rapidly calculated for various lineups during the draft.

3   AN OPTIMALITY CRITERION

Without loss of generality, suppose that drafter j uses the following player selection strategy. For the first round of the draft, we suggest that drafter j select the player amongst those available whose expected number of points in the playoffs $\theta E(g)$ is greatest. Recall that θ and $E(g)$ have been determined in advance of the draft for each player, and therefore, the decision can be made in real time.

In round n of the draft, $2 \le n \le m$, we propose that drafter j choose the player whose

selection maximizes

$$Opt = \frac{1}{K-1} \sum_{i \neq j} \text{Prob}(T_j > T_i) \tag{11}$$

where $T_1, \ldots, T_K$ correspond to the current (yet incomplete) lineups in round n. We refer to Opt in (11) as the optimality criterion and interpret Opt as the average probability that lineup j accumulates more points than one of the other lineups. Alternatively, $(K-1)Opt$ can be thought of as the expected number of lineups that are beaten by lineup j. Although maximizing the function

$$\text{Prob} \left(\bigcap_{i \neq j} (T_j > T_i) \right)$$

may also seem appealing, we have a simple approximation of Opt, and we are able to maximize the approximation of Opt in real time.

In Appendix A, we suggest the approximation

$$T_j - T_k \sim \text{Normal}(\mu_{jk}, \sigma_{jk}^2) \tag{12}$$

where $\mu_{jk} = E(T_j) - E(T_k)$, $\sigma_{jk}^2 = Var(T_j) + Var(T_k) - 2Cov(T_j, T_k)$ and we refer to expressions (2), (3) and (4) which are easily modified for rounds $2 \leq n \leq m$ of the draft. Therefore Opt in (11) is approximated by

$$\frac{1}{K-1} \sum_{j \neq i} \Phi \left(\frac{\mu_{ji}}{\sigma_{ji}} \right) \tag{13}$$

where Φ is the cumulative distribution function of the standard normal distribution. A key point is that (13) can be evaluated (by computer software) for each of the unselected hockey players in round n, and in real time, we can choose that hockey player whose selection maximizes (13).

A further simplification may be possible in terms of the search over the remaining pool of hockey players. If we order the players on each of the 16 playoff hockey teams

10

in decreasing order according to their θ values, then our intuition suggests that only the highest player remaining on each of the 16 teams need to be considered in the selection process. This reduces the search to a set of only 16 players. The idea is that we never prefer a player over a teammate if the first player's ability to score points is less than the second player's ability to score points. The conjecture is addressed in Appendix B.

4 A SIMULATION STUDY

We consider an NHL playoff hockey pool with $K = 10$ drafters and $m = 10$ rounds of drafting. Our pool is based on results from the 2003-2004 NHL regular season where we obtain the player characteristics θ using (1) and we estimated the home ice advantage $\epsilon = 0.05$ as in (5). For convenience, we restrict our attention to the 10 players per playoff team having the highest θ values. This provides 160 eligible players amongst whom 100 are selected in the draft. The series win probabilities $P(i, j)$ for all $i \neq j$ are obtained using the transitivity assumption based on the subset of Sportsbook and subjective probabilities given in Table 1.

Using the series win probabilities $P(i, j)$, the corresponding single game win probabilities p_{ij} are then calculated. We then simulate the playoffs to obtain the quantities $E(g_{ki})$, $Var(g_{ki})$ and $Cov(g_{ki}, g_{lj})$ appearing in (2), (3) and (4). These quantities relate to the number of games played by each player in the playoffs and are needed to implement our optimal drafting strategy. Table 2 gives the top 10 players according to their expected number of playoff points where we recall that this quantity is a product of the player's θ and $E(g)$ values. We note that the top 10 players belong to only four different teams.

To simulate a playoff pool, we need to create "virtual" drafters. The drafters adhere to the following rules in selecting their lineups.

Table 1: Sportsbook and subjective series win probabilities $P(i, j)$ for team i defeating team j.

Team i	Team j	$P(i,j)$
Detroit	Nashville	0.84
San Jose	St. Louis	0.58
Vancouver	Calgary	0.56
Colorado	Dallas	0.55
Tampa Bay	NY Islanders	0.69
Boston	Montreal	0.60
Philadelphia	New Jersey	0.52
Toronto	Ottawa	0.48
Detroit	Vancouver	0.61
Vancouver	Colorado	0.52
Tampa Bay	Boston	0.52
Boston	Vancouver	0.53
Calgary	Ottawa	0.43
Detroit	San Jose	0.56
Calgary	New Jersey	0.45

Table 2: Top 10 hockey players according to their expected number of playoff points.

Player	Team	θ	Expected Points
Robert Lang	Detroit	1.15	17.21
Peter Forsberg	Colorado	1.41	15.35
Martin St. Louis	Tampa Bay	1.15	14.89
Pavel Datsyuk	Detroit	0.91	13.63
Cory Stillman	Tampa Bay	0.99	12.83
Brett Hull	Detroit	0.84	12.62
Brad Richards	Tampa Bay	0.96	12.52
Alex Tanguay	Colorado	1.15	12.46
Markus Naslund	Vancouver	1.08	12.15
Joe Sakic	Colorado	1.07	11.69

- Drafter 1 $\sim$ chooses players according to the optimality criterion (13).

- Drafter 2 $\sim$ chooses players with the largest θ values.

- Drafter 3 $\sim$ chooses players with the largest expected number of playoff points.

- Drafter 4 $\sim$ is an advocate of numerology. The drafter believes that the numbers 8 and 9 are lucky and chooses players with the most regular season points that are divisible by 8 or 9.

- Drafter 5 $\sim$ roots for the underdog by choosing players with the most points alternating between the lowest seeded teams in the Eastern and Western conferences.

- Drafter 6 $\sim$ roots for the favourite by choosing players with the most points alternating between the highest seeded teams in the Eastern and Western conferences.

- Drafter 7 $\sim$ chooses players with the most regular season points.

- Drafter 8 $\sim$ is a Vancouver Canucks "Superfan". The drafter chooses only Canucks and chooses Canucks according to the highest number of regular season points.

- Drafter 9 $\sim$ chooses players with the most regular season points whose first name begins with the letter S.

- Drafter 10 $\sim$ chooses players with the highest θ values from the top four seeded teams in the first four rounds of the draft. For the remaining six rounds, the drafter chooses players with the highest θ values.

Both drafter 1 (who uses the optimality criterion) and drafter 3 have an advantage. The advantage is that they have exact knowledge of an aspect of the simulation, namely the number of expected playoff points for each player. We suspect that it would be

13

unlikely for an office colleague to go to the extreme lengths of implementing the selection strategy of drafter 3. Nevertheless, we want to include a drafter who is "smart" for the sake of assessing the optimality criterion. We also remark that many of the above drafters are fairly sophisticated in their selection strategies. Our experience is that drafters are generally not this sophisticated. Therefore, our proposed approach may work better in practice than in the following test simulation.

We now consider a simulation where each drafter drafts in the position according to their number. The playoffs were simulated 10000 times and the order of finish in each simulation was recorded for each of the drafters. In Table 3, we provide the cumulative probabilities for finishing in positions $1, \ldots, 10$ in the playoff pool. We also provide $E(LD)$, the expected number of lineups that each drafter defeats. We observe that $E(LD) = 6.54$ for drafter 1 and this is the largest expectation amongst all drafters. This might be expected as the optimality criterion attempts to maximize the number of lineups defeated. However, we note that drafter 1 does not have the highest probability of finishing first in the hockey pool. This honour belongs to drafter 2, followed by drafters 8, 6 and 3. In this regard, the success of drafter 8 (the Canucks Superfan) provides some evidence that "putting all of one's eggs in one basket" has some merit. When the Canucks go deep into the playoffs, drafter 8 does very well, and conversely, when the Canucks are eliminated early in the playoffs, drafter 8 does poorly. We also note that the probability of "being in the money" (i.e. first, second or third) is the greatest for drafter 3 followed closely by drafter 1. Drafter 3 (1) is in the money 57% (56%) of the time. As many hockey pools have prizes for first, second and third place finishes, this provides further appeal for taking into account the number of games played.

We now consider the impact of draft position. We repeat the simulation where all of the drafters maintain the same order of selection except drafter 1 who is moved from

14

Table 3: Cumulative probabilities for order of finish and $E(LD)$ in a playoff pool where drafter 1 selects first.

| | Drafter | | | | | | | | | |
Finish	1	2	3	4	5	6	7	8	9	10
1st	0.11	0.21	0.15	0.02	0.01	0.16	0.07	0.19	0.00	0.06
2nd	0.32	0.32	0.36	0.10	0.03	0.29	0.20	0.24	0.01	0.15
3rd	0.56	0.40	0.57	0.16	0.05	0.36	0.34	0.26	0.02	0.27
4th	0.77	0.50	0.72	0.24	0.06	0.45	0.50	0.30	0.04	0.45
5th	0.88	0.61	0.81	0.34	0.09	0.55	0.67	0.33	0.07	0.66
6th	0.94	0.75	0.88	0.50	0.11	0.65	0.81	0.39	0.13	0.84
7th	0.98	0.87	0.94	0.72	0.14	0.73	0.92	0.47	0.30	0.94
8th	0.99	0.95	0.98	0.93	0.20	0.82	0.98	0.53	0.64	0.99
9th	1.00	1.00	1.00	0.99	0.44	0.91	1.00	0.71	0.96	1.00
10th	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$E(LD)$	6.54	5.60	6.40	3.99	1.12	4.92	5.49	3.41	2.17	5.35

the first to the last position. This means, for example, that drafter 1 has the 10th and the 11th selections in the draft. The results of the simulation are provided in Table 4. We observe that drafter 1 does about as well selecting in the last position as in the first position. For example, drafter 1 finishes second to drafter 3 in $E(LD)$ and drafter 1 is second according to being in the money. Again, we are not too disappointed by the $E(LD)$ result as drafter 3 uses specialized knowledge that would not be typically available to drafters. Our explanation for this is that drafting in the 10th position is too far away from the beginning of the draft to select one of the few really outstanding hockey players.

Now recall the discussion where it was pointed out that drafter 1 and drafter 3 were advantaged by their knowledge of the expected number of playoff points. It therefore might be asked whether there is a meaningful difference in their selection strategies. In other words, does it make sense to invoke the optimality criterion when a similar lineup might be achieved by the simpler criterion of maximizing the expected number of playoff

15

Table 4: Cumulative probabilities for order of finish and $E(LD)$ in a playoff pool where drafter 1 selects last.

	Drafter									
Finish	1	2	3	4	5	6	7	8	9	10
1st	0.13	0.13	0.18	0.02	0.01	0.22	0.07	0.18	0.00	0.05
2nd	0.31	0.25	0.46	0.08	0.03	0.32	0.19	0.23	0.01	0.14
3rd	0.50	0.36	0.64	0.15	0.05	0.39	0.34	0.27	0.01	0.29
4th	0.67	0.51	0.75	0.24	0.06	0.46	0.49	0.30	0.03	0.48
5th	0.79	0.66	0.84	0.36	0.09	0.54	0.64	0.34	0.07	0.69
6th	0.87	0.83	0.90	0.51	0.12	0.62	0.77	0.40	0.14	0.85
7th	0.93	0.94	0.95	0.72	0.15	0.70	0.88	0.47	0.33	0.94
8th	0.97	0.98	0.98	0.93	0.20	0.80	0.97	0.53	0.65	0.99
9th	0.99	1.00	1.00	0.99	0.44	0.91	1.00	0.71	0.97	1.00
10th	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$E(LD)$	6.16	5.66	6.70	3.99	1.14	4.96	5.33	3.43	2.21	5.42

points? To investigate this, we consider whom drafter 3 would have selected had drafter 3 been in the position of drafter 1. In the simulation where drafter 1 chose first, drafter 3 would have picked the same player as drafter 1 in 8 of the 10 draft positions. In the simulation where drafter 1 chose last, drafter 3 would have picked the same player as drafter 1 in only 6 of the 10 draft positions. Hence, the strategy of drafter 1 based on the optimality criterion does differ meaningfully from the simplified (but intelligent) strategy of drafter 3.

5   AN ACTUAL STANLEY CUP PLAYOFF POOL

Although a particular instance of a hockey pool does not address the long term properties of a drafting strategy, it is instructive to review our approach in a realistic setting. For the 2006 NHL playoffs, we conducted a pool with $K = 10$ drafters and $m = 10$ rounds

Table 5: Sportbook odds for the favourite in the first round series (underdog in parentheses).

Teams	Odds
Detroit (Edmonton)	0.37
Dallas (Colorado)	0.61
Calgary (Anaheim)	0.61
Nashville (San Jose)	0.92
Ottawa (Tampa Bay)	0.41
Carolina (Montreal)	0.56
New Jersey (New York)	0.54
Buffalo (Philadelphia)	0.67

of drafting. Our strategy was implemented by one of the authors (Amy), who by chance, drafted in the first position.

In Table 5, we provide the Sportsbook odds (adjusted for vigorish) for each of the favourites in the eight first round series. The odds were taken from www.vegasinsider.com two days prior to the beginning of the playoffs. The underdog is listed in parentheses. In Table 6, we provide Amy's subjective odds for seven hypothetical matchups. Some of the odds may seem extreme but this demonstrates the flexibility of the approach. All of the odds in Table 5 and Table 6 were then used to complete the graph and populate the probability matrix.

For the sake of brevity, in Table 7, we show the first 3 draft picks for each of the drafters, the draft order, playoff points accumulated by the draft picks and the total number of playoff points for the entire lineup. The 2006 playoffs were marked by upsets with Carolina defeating Edmonton in the Stanley Cup finals. However, by skill and good fortune, Amy won the hockey pool with 113 points followed by Willy with 97 points.

17

Table 6: Subjective odds for the favourite in hypothetical matchups (underdog in parentheses).

Teams	Odds
Dallas (Detroit)	0.54
Calgary (Detroit)	0.43
Calgary (Nashville)	0.25
Anaheim (New York)	0.18
Anaheim (Philadelphia)	0.82
Carolina (Ottawa)	0.92
Calgary (Carolina)	0.75

Table 7: Some summary results from the 2006 NHL playoff pool.

Drafter	Pick	Points	Drafter	Pick	Points
Amy (1)	P Elias (NJ)	16	Jen (2)	J Jagr (NY)	1
	D Briere (Buf)	19		P Schaefer (Ott)	7
	C Stillman (Car)	26		J Iginla (Cgy)	8
Total Points = 113			Total Points = 51		
Darcy (3)	D Heatley (Ott)	12	Leslie (4)	E Staal (Car)	28
	P Kariya (Nsh)	7		B Gionta (NJ)	7
	M Modano (Dal)	4		R Lang (Det)	6
Total Points = 51			Total Points = 93		
Linda (5)	J Thornton (SJ)	9	AJ (6)	B Shanahan (Det)	2
	N Lidstrom (Det)	2		J Spezza (Ott)	14
	T Holmstrom (Det)	3		R Smyth (Edm)	16
Total Points = 49			Total Points = 80		
Beth (7)	A McDonald (Ana)	9	Willy (8)	H Zetterberg (Det)	6
	A Tanguay (Col)	6		P Datsyuk (Det)	3
	A Hemsky (Edm)	17		P Marleau (SJ)	14
Total Points = 57			Total Points = 97		
Art (9)	D Alfredsson (Ott)	10	Rory (10)	J Cheechoo (SJ)	9
	T Selanne (Ana)	14		M Afinogenov (Buf)	8
	S Gomez (NJ)	9		P Forsberg (Phi)	8
Total Points = 93			Total Points = 67		

6   DISCUSSION

We have proposed a method of player selection in a common NHL hockey pool. With the aid of a computer and some preliminary simulations, player selections can be made in real time. It appears that the approach is better than a number of ad-hoc strategies.

It is clear that the approach could also be implemented for other hockey leagues (eg. junior hockey), and it might also be tweaked for pools with slightly different rules and also for different sports (eg. soccer).

7   REFERENCES

Bennett, J. (editor) (1998). *Statistics in Sport*, Arnold Publishers.

Berry, S.M. (2000). "My triple crown", In the column, A Statistician Reads the Sports Pages, *Chance*, 13 (3), 56-61.

Berry, S.M., Reese, C.S. and Larkey, P.D. (1999). "Bridging different eras in sports", *Journal of the American Statistical Association*, 94, 661-676.

Carlin, B.P. (1996). "Improved NCAA basketball tournament modeling via point spread and team strength information", *The American Statistician*, 50, 39-43.

Insley, R., Mok, L. and Swartz, T.B. (2004). "Issues related to sports gambling", *The Australian and New Zealand Journal of Statistics*, 46, 219-232.

Summers, A. (2005). "Hockey pools for profit: a simulation based player selection strategy", *MSc project, Department of Statistics and Actuarial Science, Simon Fraser University.*

8   APPENDIX A

In this section we discuss conditions under which the normal approximation to $T_j - T_k$ might be expected to be reasonable. Under our model, given the numbers of games played in the playoffs by each player on a drafter's roster, the total number of points scored T, has a Poisson distribution. When the Poisson parameter is large, and we expect this when

many players are drafted, the conditional distribution of T will be reasonably normal. However, due to the correlation in the numbers of games played by different players, it is clear that the unconditional distribution of T may be distinctly non-normal.

Consider round n of the draft and focus on one drafter. Let θ_i denote the expected goals per game for the ith player drafted and let g_i be the number of games played by this player. Put

$$G = \sum g_i \theta_i.$$

Let T be the points obtained by the players drafted through round n. Given the g_i the quantity T has a Poisson distribution with mean G. We now give conditions under which T is approximately normally distributed. Denote

$$\mu = \text{E}(G) = \sum \text{E}(g_i)\theta_i$$

and

$$\sigma^2 = \text{Var}(G).$$

Then

$$\text{E}(T) = \text{E}(G) = \sum \text{E}(g_i)\theta_i$$

and

$$\text{Var}(T) = \text{E}(G) + \text{Var}(G).$$

The basic theoretical result is this:

Theorem: *Suppose that (T_n, G_n) is a sequence of random variables such that*

- *Given G_n, the conditional distribution of T_n is Poisson with mean G_n.*

- *The mean μ_n of G_n tends to infinity with n.*

- *The variance σ_n^2 of G_n is finite.*

Assume either:

1. G_n is asymptotically normal, that is,

$$\frac{G_n - \mu_n}{\sigma_n} \Rightarrow N(0,1)$$

OR

2. Variability in G_n is asymptotically negligible in the sense that

$$\frac{\sigma_n^2}{\mu_n + \sigma_n^2} = \frac{\sigma_n^2}{\mathrm{Var}(T_n)} \to 0.$$

Then

$$\frac{T_n - \mu_n}{\sqrt{\mathrm{Var}(T_n)}} \Rightarrow N(0,1).$$

Proof: Let $\phi(\cdot\,; \lambda)$ be the characteristic function of a standardized Poisson variate with mean λ, that is, if Y has a Poisson(λ) distribution then

$$\phi(t; \lambda) = \mathrm{E}[\exp\{it(Y - \lambda)/\sqrt{\lambda}\}] = \exp\left[\lambda\{\exp(it\lambda^{-1/2}) - it\lambda^{-1/2} - 1\}\right].$$

Write

$$\tau_n^2 = \mathrm{Var}(T_n) = \mu_n + \sigma_n^2$$

and

$$Z_n \equiv \frac{T_n - \mu_n}{\tau_n} = \frac{\sqrt{G_n}\frac{T_n - G_n}{\sqrt{G_n}} + G_n - \mathrm{E}(T_n)}{\tau_n}.$$

Then

$$
\begin{aligned}
\mathrm{E}\{\exp(itZ_n)\} &= \mathrm{E}\left(\phi(t\sqrt{G_n}/\tau_n; G_n)\exp\left[it\{(G_n - \mu_n)/\tau_n\}\right]\right) \\
&= \mathrm{E}\left(\exp\{-t^2 G_n/(2\tau_n^2)\}\exp\left[it\{(G_n - \mu_n)/\tau_n\}\right]\right) \qquad (14) \\
&+ \mathrm{E}\left(\left[\phi(t\sqrt{G_n}/\tau_n; G_n) - \exp\{-t^2 G_n/(2\tau_n^2)\}\right]\exp\left[it\{(G_n - \mu_n)/\tau_n\}\right]\right).
\end{aligned}
$$

Since $\mu_n = \mathrm{E}(G_n) \le \tau_n^2$ we find $G_n/\tau_n^2 = O_P(1)$. By the central limit theorem it follows that as n tends to infinity we have

$$\left[\phi(t\sqrt{G_n}/\tau_n; G_n) - \exp\{-t^2 G_n/(2\tau_n^2)\}\right] \to 0$$

21

in probability. Moreover this quantity is bounded, and therefore the second term in (14) tends to 0. That is

$$\mathrm{E}\{\exp(itZ_n)\} = \mathrm{E}\left(\exp\left\{-t^2 G_n/(2\tau_n^2)\right\}\exp\left[it\left\{(G_n - \mu_n)/\tau_n\right\}\right]\right) + o(1). \tag{15}$$

Rewrite the leading term in (15) as

$$\mathrm{E}\left(\exp\left[\{it - t^2/(2\tau_n)\}(G_n - \mu_n)/\tau_n\right]\right)\exp(-t^2/2)\exp(t^2\sigma_n^2/(2\tau_n^2)). \tag{16}$$

Notice that $\tau_n \to \infty$ so that $t^2/\tau_n \to 0$. Under the second condition the last factor in (16) converges to 1, and

$$\frac{G_n - \mu_n}{\tau_n} = \frac{G_n - \mu_n}{\sigma_n}\frac{\sigma_n}{\tau_n} \to 0$$

in probability so that the first factor in (16) also converges to 1 leaving

$$\mathrm{E}\{\exp(itZ_n)\} \to e^{-t^2/2}$$

as desired. On the other hand if the first condition holds then

$$\mathrm{E}\left(\exp\left[\{it - t^2/(2\tau_n)\}(G_n - \mu_n)/\tau_n\right]\right) - \mathrm{E}\left(\exp\left[it(G_n - \mu_n)/\tau_n\right]\right) \to 0$$

(because $\tau_n \to \infty$). Now

$$\mathrm{E}\left(\exp\left[it(G_n - \mu_n)/\tau_n\right]\right) = \psi_n(t\sigma_n/\tau_n)$$

where ψ_n is the characteristic function of $(G_n - \mu_n)/\sigma_n$. The first condition guarantees that

$$\psi_n(t\sigma_n/\tau_n) - \exp\{-t^2(\sigma_n^2/(2\tau_n^2))\} \to 0$$

for each fixed t. It follows that

$$\mathrm{E}\{\exp(itZ_n)\}$$

22

has the same limit as

$$\exp\{-t^2(\sigma_n^2/(2\tau_n^2))\}e^{-t^2/2}\exp(t^2\sigma_n^2/(2\tau_n^2)) = e^{-t^2/2}$$

which proves the Theorem.

It remains to interpret this result in terms of our context. The condition that $\mu_n \to \infty$ means that we need the expected number of points scored in the playoffs by the first n players drafted to be large. This amounts to a requirement that the players score many points or play many games or that there be many players on a lineup.

Looking at the case where the variability in G_n is negligible amounts to considering cases where there is little randomness in the linear combination $\sum g_i\theta_i$. When players are drafted from a variety of teams we expect that the average correlations between pairs of g_i will be negative. Though an individual g_i is highly variable because a team may be eliminated early or play all the rounds the total number of games played $\sum g_i$ is clearly much less variable: the number of playoff series is fixed.

Finally we have not examined closely cases in which G itself is nearly normal. Clearly, if all players are picked from the same team, this is unlikely to be the case. If, however, players are picked from a variety of teams then we suspect that there will be enough independence amongst the g_i to produce approximate normality.

Our arguments indicate that the normal approximation should improve as rounds of drafting continue – when only 1 player has been drafted a normal approximation is unlikely to be effective.

Finally we note that the distribution of a single T_j is likely to be skewed to the right since the Poisson distributions involved are right skewed and G seems likely to be right skewed. Under these conditions it is likely that $T_j - T_k$ will be less skewed than an individual T_j, improving the resulting normal approximation to the difference.

9 APPENDIX B

We put forward the following conjecture which is intuitive and held true in all of our simulations including those of Section 4. Consider player 1 and player 2 from the same team with characteristics θ_1 and θ_2 respectively where $\theta_1 > \theta_2$. According to the optimality criterion (13), in a given round of the draft, we would never select player 2 if player 1 were available.

Unable to establish the conjecture, we prove a restricted version. Consider the n-th round of the draft where drafter k has selected a lineup with total points T_k. Drafter j considers two lineups which differ only in the n-th player selected. Let T_j be the total points corresponding to the first lineup under consideration where the n-th player selected has characteristic θ_{jn}. Let T_j' be the total points corresponding to the second lineup under consideration where the n-th player selected has characteristic $\theta_{jn} - \Delta$ where $\Delta > 0$.

Proposition: In a neighbourhood of $E(T_j) = E(T_k)$,

$$\mathrm{Prob}(T_j > T_k) > \mathrm{Prob}(T_j' > T_k)$$

where Prob refers to the normal approximation in (12).

Proof: Using (2) and the notation of Section 3, we need to establish

$$\Phi\left(\frac{E(T_j) - E(T_k)}{SD(T_j - T_k)}\right) > \Phi\left(\frac{E(T_j) - E(T_k) - \Delta E(g_{jn})}{SD(T_j' - T_k)}\right)$$

in a neighbourhood of $E(T_j) = E(T_k)$. The inequality holds if

$$f(x) = \frac{x}{SD(T_j - T_k)} - \frac{(x - \Delta E(g_{jn}))}{SD(T_j' - T_k)} > 0$$

where $x = E(T_j) - E(T_k)$. Noting that $f(0) > 0$ and $f'(x) = SD^{-1}(T_j - T_k) - SD^{-1}(T_j' - T_k)$, the proposition is proved. We remark that when the derivative is close to zero (as might be expected), the neighbourhood is large.