[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

2019 NHL Playoff Hockey

The data I have selected for the project is statistics from the 2019 NHL playoffs**,** if not familiar the NHL is the National Hockey League comprised of 31 teams 24 in the United States and 7 in Canada. The teams are split into 2 conferences the Western and Eastern Conference, each conference has two subconferences, in the Western Conference there is the Pacific and Central and for the Eastern there is the Atlantic and Metropolitan. In the regular season teams play 82 games, come playoff time 16 teams will move on 8 from each conference and the teams are selected from the total number of points at the end of the season. For each game a team can earn up to 2 points, and here's how the point breakdown works: 2 points for a win in regulation, if after the 3rd period the game is a tie, both teams compete in OT, each at that point receiving a point and the winner of OT or shootout receives another point.

Now to the data, I have compiled a list of stats from all the players who competed in the 2019 playoffs. Here is a breakdown of all the abbreviations on the sheet: We have age, team, position, games played, goals, assists, points (you get a point for each goal or assist), the +/- rating deals with the player getting a plus when he is on the ice when his team scores and a minus for when he is on the ice for the opposing team scoring. PIM refers to penalty minutes, then there is a breakdown of the goals the players have scored because there are 4 types: Even strength, power play, and shorthanded and game winning, it would be the same breakdown for assists except for game winning. We also have shot counts and percentages, total ice time, and average time on ice per game. Then there are minor things like the number of blocked shots, the

[redacted]

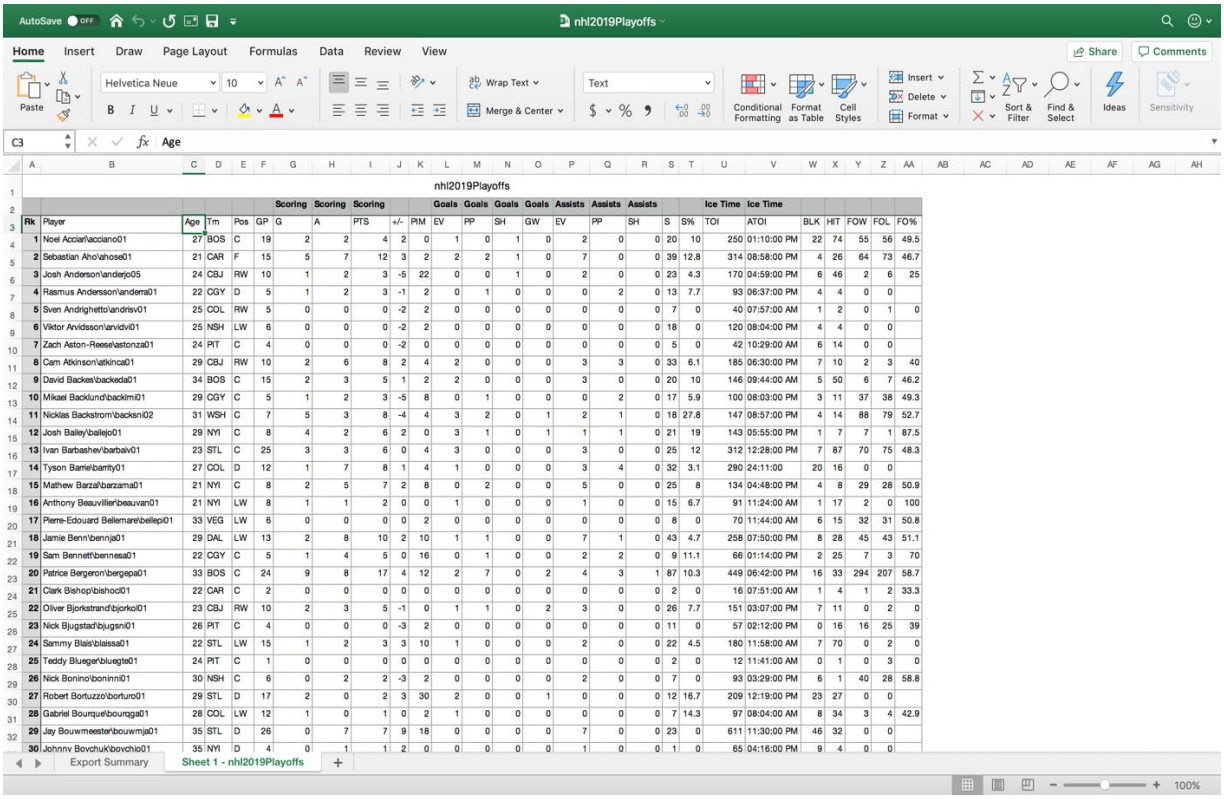
Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

number of hits, FOW which stands for faceoffs won, FOL which is faceoffs lost and finally FO% which is faceoff percentages of won by the athlete. To break down the variables in statistical terms, the categorical variables would be things such as: The position they play, the dominant hand of the player (which isn't stated on the excel sheet but could be found with a search), and the team they play for because these have no numerical value. The quantitative variables would be things like: Faceoff win percentage, shot percentage and goals and assists, as they all have a numerical value. The data I found is from a website called: Hockey-Reference: https://www.hockey-reference.com/playoffs/NHL_2019_skaters.html .

Here is a preview of the data:



nhl2019Playoffs

Rk	Player	Age	Tm	Pos	GP	G	A	PTS	+/-	PIM	Goals EV	Goals PP	Goals SH	Goals GW	Assists EV	Assists PP	Assists SH	S	S%	TOI	ATOI	BLK	HIT	FOW	FOL	FO%
1	Noel Acciari\acciano01	27	BOS	C	19	2	2	4	2	0	1	0	1	0	2	0	0	20	10	250	01:10:00 PM	22	74	55	56	49.5
2	Sebastian Aho\ahose01	21	CAR	F	15	5	7	12	3	2	2	2	1	0	7	0	0	39	12.8	314	08:58:00 PM	4	26	64	73	46.7
3	Josh Anderson\anderjo05	24	CBJ	RW	10	1	2	3	-5	22	0	0	1	0	2	0	0	23	4.3	170	04:59:00 PM	6	46	2	6	25
4	Rasmus Andersson\anderra01	22	CGY	D	5	1	2	3	-1	2	0	1	0	0	2	0	0	13	7.7	93	06:37:00 PM	4	4	0	0	
5	Sven Andrighetto\andrisv01	25	COL	RW	5	0	0	0	-2	2	0	0	0	0	0	0	0	7	0	40	07:57:00 AM	1	2	0	1	0
6	Viktor Arvidsson\arvidvi01	25	NSH	LW	6	0	0	0	-2	2	0	0	0	0	0	0	0	18	0	120	08:04:00 PM	4	4	0	0	
7	Zach Aston-Reese\astonza01	24	PIT	C	4	0	0	0	-2	0	0	0	0	0	0	0	0	5	0	42	10:29:00 AM	6	14	0	0	
8	Cam Atkinson\atkinca01	29	CBJ	RW	10	2	6	8	2	4	2	0	0	0	3	3	0	33	6.1	185	06:30:00 PM	7	10	2	3	40
9	David Backes\backeda01	34	BOS	C	15	2	3	5	1	2	2	0	0	0	3	0	0	20	10	146	09:44:00 AM	5	50	6	7	46.2
10	Mikael Backlund\backlmi01	29	CGY	C	5	1	2	3	-5	8	0	1	0	0	2	0	0	17	5.9	100	08:03:00 PM	3	11	37	38	49.3
11	Nicklas Backstrom\backsni02	31	WSH	C	7	5	3	8	-4	4	3	2	0	1	2	1	0	18	27.8	147	08:57:00 PM	4	14	88	79	52.7
12	Josh Bailey\bailejo01	29	NYI	C	8	4	2	6	2	0	3	1	0	1	1	1	0	21	19	143	05:55:00 PM	1	7	7	1	87.5
13	Ivan Barbashev\barbaiv01	23	STL	C	25	3	3	6	0	4	3	0	0	0	3	0	0	25	12	312	12:28:00 PM	7	87	70	75	48.3
14	Tyson Barrie\barrity01	27	COL	D	12	1	7	8	1	4	1	0	0	0	3	4	0	32	3.1	290	24:11:00	20	16	0	0	
15	Mathew Barzal\barzama01	21	NYI	C	8	2	5	7	2	8	0	2	0	0	5	0	0	25	8	134	04:48:00 PM	4	8	29	28	50.9
16	Anthony Beauvillier\beauvan01	21	NYI	LW	8	1	1	2	0	0	1	0	0	0	1	0	0	15	6.7	91	11:24:00 AM	1	17	2	0	100
17	Pierre-Edouard Bellemare\bellepi01	33	VEG	LW	6	0	0	0	0	2	0	0	0	0	0	0	0	8	0	70	11:44:00 AM	6	15	32	31	50.8
18	Jamie Benn\bennja01	29	DAL	LW	13	2	8	10	2	10	1	1	0	0	7	1	0	43	4.7	258	07:50:00 PM	8	28	45	43	51.1
19	Sam Bennett\bennesa01	22	CGY	C	5	1	4	5	0	16	0	1	0	0	2	2	0	9	11.1	66	01:14:00 PM	2	25	7	3	70
20	Patrice Bergeron\bergepa01	33	BOS	C	24	9	8	17	4	12	2	7	0	2	4	3	1	87	10.3	449	06:42:00 PM	16	33	294	207	58.7
21	Clark Bishop\bishocl01	22	CAR	C	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	16	07:51:00 AM	1	4	1	2	33.3
22	Oliver Bjorkstrand\bjorkoli01	23	CBJ	RW	10	2	3	5	-1	0	1	1	0	2	3	0	0	26	7.7	151	03:07:00 PM	7	11	0	2	0
23	Nick Bjugstad\bjugsni01	26	PIT	C	4	0	0	0	-3	2	0	0	0	0	0	0	0	11	0	57	02:12:00 PM	0	16	16	25	39
24	Sammy Blais\blaissa01	22	STL	LW	15	1	2	3	3	10	1	0	0	0	2	0	0	22	4.5	180	11:58:00 AM	7	70	0	2	0
25	Teddy Blueger\bluegte01	24	PIT	C	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	12	11:41:00 AM	0	1	0	3	0
26	Nick Bonino\boninni01	30	NSH	C	6	0	2	2	-3	2	0	0	0	0	2	0	0	7	0	93	03:29:00 PM	6	1	40	28	58.8
27	Robert Bortuzzo\borturo01	29	STL	D	17	2	0	2	3	30	2	0	0	1	0	0	0	12	16.7	209	12:19:00 PM	23	27	0	0	
28	Gabriel Bourque\bourqga01	28	COL	LW	12	1	0	1	0	2	1	0	0	0	0	0	0	7	14.3	97	08:04:00 AM	8	34	3	4	42.9
29	Jay Bouwmeester\bouwmja01	35	STL	D	26	0	7	7	9	18	0	0	0	0	7	0	0	23	0	611	11:30:00 PM	46	32	0	0	
30	Johnny Boychuk\boychjo01	35	NYI	D	4	0	1	1	2	0	0	0	0	0	1	0	0	1	0	65	04:16:00 PM	9	4	0	0	

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

For me I choose this topic based off the fact that I am a huge hockey particularly a fan for the Dallas Stars. A funny story my mom tells me to remind me why my interest is so high in the sport is because I was born in 1999 in April around the time the Stanley Cup was going on, reminder I'm from the Dallas area and that was the year the Dallas Stars won the cup, so she always tells me I was born to be a hockey fan. The peaking interest in the sport due to the area I am from, it is a growing hockey community with not only a professional hockey team, but 4 select hockey teams, an NAHL team, as well as the Dallas Stars Elite Hockey Club which competes in the highest level of hockey in the Dallas area for kids under the age of 18. While being in high school all my friends also played in the high school hockey league as well as competing on a select hockey team. The high school hockey association has 7 different levels 4 being varsity level and 3 junior varsity levels. I had the pleasure of watching some of my best friends play every Thursday night even watched them win the National title in 2017.

Some things I have interest in examining closer with the data would be a more in-depth look into which penalties players are taking the most and what is the least called penalty. The +/- rating is also another topic I would like to look into, mostly just the ranking of the playoff players and who had the worst and best rating. Along with that the ranking of the players with the most goals and assists. Something that isn't represented on the excel sheet but that I would also have an interest in looking into is how people thought and predicted the playoffs and series would go and who would win each series and who would win the cup.

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

From the data I took the categorical variable of position to create a frequency and relative frequency table, there are 5 positions listed on this particular set of data: C, LW, RW, F, and D. These stand for: Center, Left Wing, Right Wing, Forward, and Defenseman. In hockey its 5 vs 5 with each team also having a goalie on the ice, so there are always 3 forwards, 1 left wing, a center and a right wing and 2 defensemen on the ice at all time unless a penalty is called. The data I collected has a total of 333 players, so I would know whether or not there is a miscalculation with the frequency table. Here is what my table looked like:

Frequency	Frequency	Relative Frequency
C	122	0.366366366
LW	50	0.15015015
RW	46	0.138138138
F	2	0.006006006
D	113	0.339339339
Total	333	0.705705706

As you can see from the table I did have exactly 333 players, something unusual that I saw from the table was how close the numbers are for the center's and defensemen, as I stated above there are 2 defense men on the ice, so the number being lower than the number of centers is surprising. From more further investigation we can see a low number for both sides of the wing players, which typically means the centers are being transformed into wing players which is why the number is higher than the rest of the positions. From the table we can also determine the most popular position based off the relative frequency, this just allows us to have a view of

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

the numbers in a different format, but from that we can still determine that the centers are the most popular position. A fun fact about centers and why they are the most important player on the ice: Centers control the pace and play of the game, they are extremely dominate in both the offensive half of the ice as well as the defensive zone.

The next thing I did with the categorical variables was taking the positions again and the teams that made the playoffs to determine how many of each position was on each team. For this I took 5 different teams out of the 16 from the playoffs to create a two-way table to compare and contrast the numbers of the positions for each team. The 5 teams I decided to do were: Dallas Stars, Nashville Predators, St. Louis Blues, Pittsburg Penguins and The Washington Capitals. Here is what the table looked like:

Two Way Table:	DAL	NSH	PIT	STL	WSH
Position vs Team	DAL	NSH	PIT	STL	WSH
C	7	10	9	9	7
LW	4	2	1	4	4
RW	4	2	3	1	3
F	0	0	0	0	0
D	9	6	7	7	7

As you can see from the data out of the 5 teams Nashville had the highest number of centers followed by Pittsburg and St. Louis, then followed by Dallas and Washington. I determined from the data that the teams with a higher number of centers tended to have a lower number of wing players. You can see how Dallas and Washington have around an equal number of centers and wing players compared to the other 3 teams with around 4 wing players. Most

[redacted]

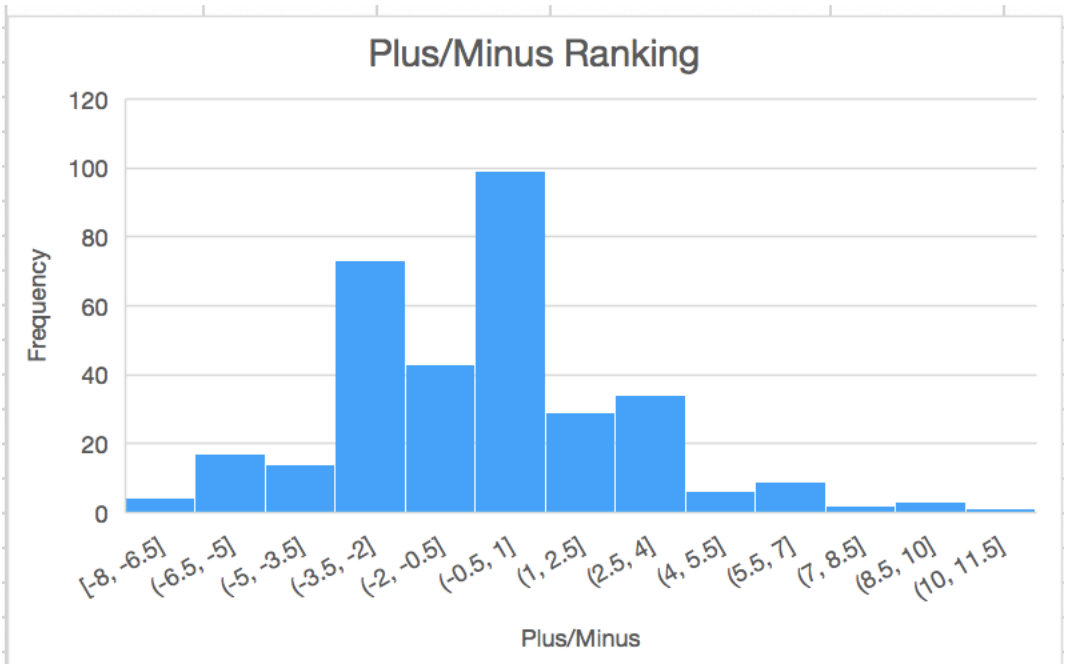
Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

teams in the playoffs and regular season carry around 7-8 defenders per game or series in case of injury or other circumstances. From the data we can see that Dallas has the highest number of defenders with 9, followed by Washington, St. Louis and Pittsburg with 7 and Nashville with 6.

For the quantitative variables I decided to work with the plus/minus ranking of all the players in the playoffs, to recap what that is, each player will get assigned a plus or a minus on a scoring play depends on which side they are on, scoring team gets a plus and the opposing team gets a minus. I took the plus/minus rankings from all players to see how it spread out from all the players in the playoffs. Here is what the data looks like for the histogram and box plot:

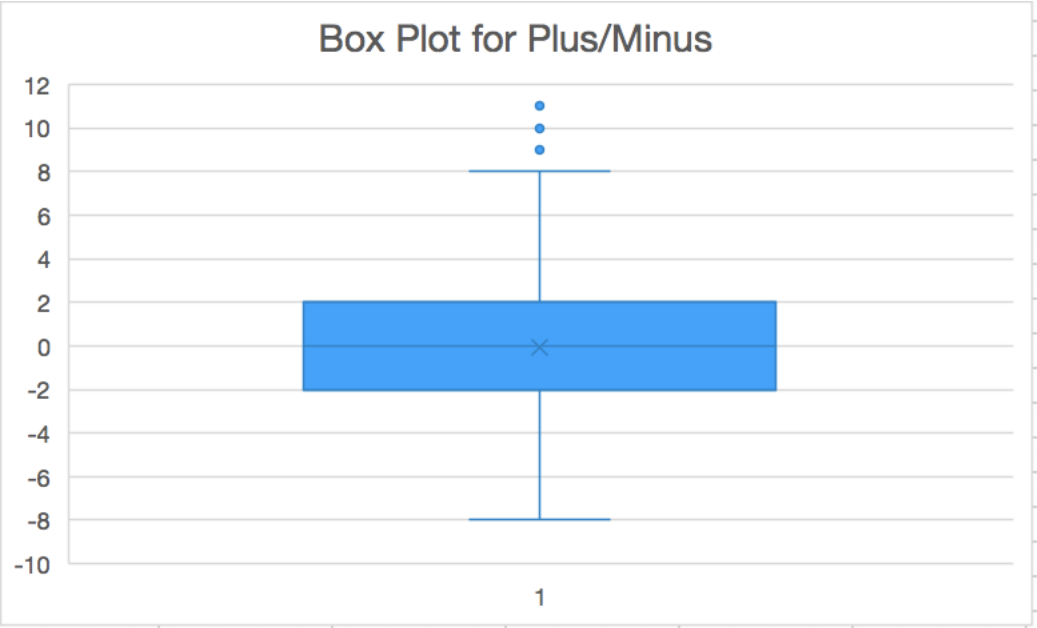


[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20



From the data we can see that there are 3 outliers on the box plots, we can also see that the range is from -8 to 8 and the median being indicated with an X at 0. From the histogram we are able to determine that many players have around a -3.5 to 1 plus/minus ranking with most players having between -0.5 to 1 since it has to tallest spike. We can also see the outliers from a different perspective on the histogram as well. For the 5 number summary of the data the median and mode were both 0, the mean was -0.0960961 and Q1 is -2 and Q3 is 2. The standard deviation from the data is 3.01648329. The data isn't skewed meaning it is symmetrical I know this mean and median are almost the same.

For the hypothesis testing potion of the project we are going to continue with the same topic from the previous test but narrow it down to focus on a singular team. I have chosen the Dallas Stars because they are my favorite team and I would like to see of their plus/minus is

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

higher or lower than the average. We will base this portion of the hypotheses on the mean and median taken from the previous section which 0 or around 0.

$$H_0: \mu Stars = 0$$

$$Ha: \mu Stars \neq 0.$$

For the categorical variables I am going to use the number of centers compared to the number of people on the ice.

$$H_0: \ pC = 1/5$$

$$Ha: : pC \neq 1/5$$

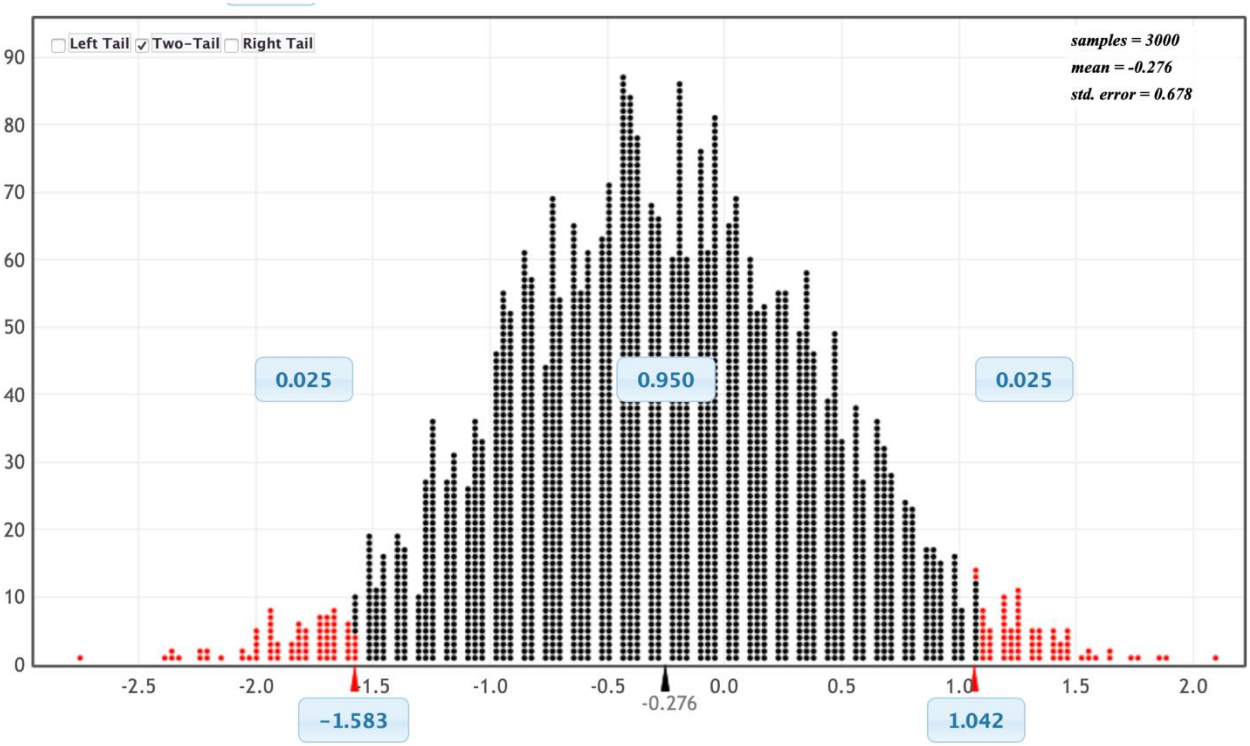
For this part of the project I am working a bootstrap for both my quantitative and categorical hypotheses, I am either going to fail to reject or reject my hypotheses created. I am using Statkey for my quantitative variable which was seeing if the Dallas Stars had a higher or lower plus/minus ranking than the average for the NHL in 2019. We calculated the average in previous steps of the project and determined the average plus/minus as 0. I used n= 3,000 for my bootstrapping sample size.

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20



This graph shows the 95% confidence interval for the plus/minus ranking as you can see the low is -1.583 and the high is 1.042. The mean for this sample was calculated at -0.276 and with a standard error of 0.678 which is all presented on the graph. As you can see the mean as stated before is -0.276 for the Stars plus/minus and the mean for the NHL was 0. This would mean that I would have to fail to reject my hypothesis because the Stars plus/minus isn't equal to the leagues plus/minus but falls within the range of the intervals. I also calculated the 85% confidence interval and found the low at -1.250 and the high at 0.708.

The next part would be to calculate my categorical which was seeing if the number of centers was equal to the number of centers that are supposed to be on the ice which is 1 out of 5

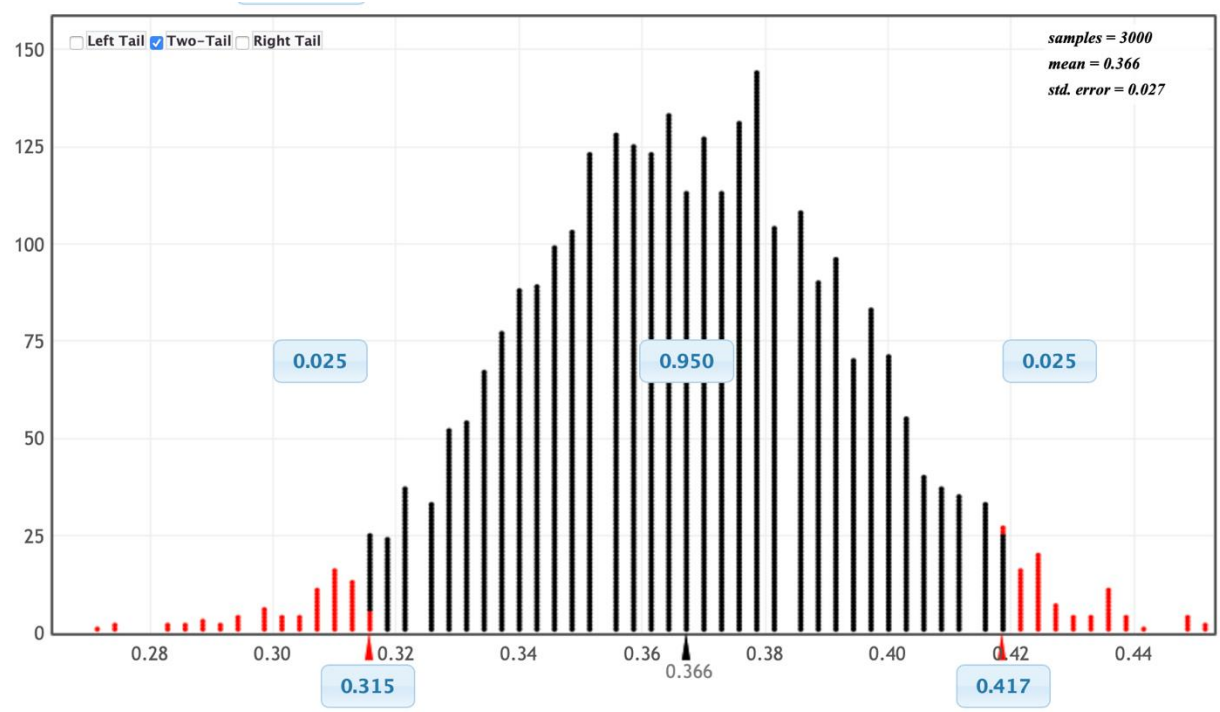
[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

players. From data in part 2 of the project we can see that there are 122 centers in the league out of 333 players. I once again used 3,000 as my sample size.



From the graph we can see the mean is 0.366 and the standard error is 0.027. We can also determine the 95% confidence interval which makes the low 0.315 and the high 0.417. I stated my hypothesis that the number of centers in the league would be equal to 1/5 of the players on the ice, that would mean 0.200 would be the number we are looking for and as I said the mean was 0.366 meaning there are more centers than the proportion meaning I would have to reject my hypothesis. I also calculated the 85% confidence interval and the low was 0.327 and the high is 0.405.

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

For this part of the project I am going to be retesting my categorical variable hypothesis. My null hypothesis was that pC= 1/5 saying that the number of centers would equal 1/5 of the players on the ice and the alternative being that pC ≠ 1/5. I used the formulas we got from the homework to compute my data. For the bootstrap I had to reject my hypothesis because the proportion didn't fall between the confidence intervals I got. When redoing the test, I still found that the proportion didn't fall between the confidence intervals I got and that is to be expected due to the fact that around 1/3 of the players who competed in the 2019 playoffs were centers, but this mean I have to reject my null hypothesis again. The formulas I used for this equation are: **Phat-P/SE** to find my z score, **SQRT(P*(1-P)/n)** this gave me my SE, for the confidence intervals =**Phat-z critical***SE** and just change the minus to a plus to get the other one and for my z critical I did **=norm.s.inv(1-0.05/2).**

Sample size	Proportion	Statistic	Alpha	confidence
333	0.2	0.36636637	0.05	
n	p	phat		95%
z*	z	SE	low 95%	high 95%
1.959963985	7.58975325	0.02191986	0.32340422	0.40932851

The next step in this project is to go back and retest my quantitative variable hypothesis, which stated that the Dallas Stars plus/minus team average either would or wouldn't equal the average of the NHL which was zero. In statistical term my hypothesis looked like this:

H_0: μStars = 0

Ha: μStars ≠ 0.

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

From my bootstrap in part 5 of the project I had to fail to reject my null hypothesis because the 0 fell in between the 95% confidence intervals I calculated. To recalculate this, I'm going to take the same group of data which is all the plus/minus rankings for each Dallas Stars player and calculate a new 95% confidence interval in Excel with formulas.

MU	X Bar	n	DF	Alpha
0	-0.291666667	24	23	0.05
SD	SE	T	T*	
3.394101875	0.692818144	-0.29166667	-1.71387153	
	SD/SQRT(N)	Xbar-MU/SE	T.INV(alpha,n-1)	
Confidence Interval 95%	high	low		
	1.066256896	-1.64959023		

For the 95% confidence interval as you can see I got my high as 1.066 and my low as -1.649, for my bootstrap in part 5 I got the high as 1.042 and my low as -1.583. These intervals I calculated are super similar my one from excel just gives me a little more range to include in. The X-bar is the mean I calculated for my samples and it is also similar to the mean from my bootstrap. I am able to fail to reject because the x-bar and the hypothesis mean fall into the range of the confidence intervals I determined.

The last step to this project is dealing with probability, I am going to go back to part 2 of the project and used the two way table I created for the positions and the number of players in those positions.

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

Frequency	Frequency	Relative Frequenc
C	122	0.366366366
LW	50	0.15015015
RW	46	0.138138138
F	2	0.006006006
D	113	0.339339339
Total	333	0.705705706

Two Way Table:	DAL	NSH	PIT	STL	WSH
Position vs Team	DAL	NSH	PIT	STL	WSH
C	7	10	9	9	7
LW	4	2	1	4	4
RW	4	2	3	1	3
F	0	0	0	0	0
D	9	6	7	7	7

For this part of the project the frequency isn't important, but we are going to determining the probability that a player on the Dallas Stars is a defenseman, in formula format this would look like:

$$P(D|Stars)=P(D\cap Stars)P(Stars)=n(LW\cap Stars)n(Stars)$$

We would end up with the fraction **9/24**, this can't be simplified any more than it already is but to put this in decimal format we get **0.375** which is a little more than a 1/3 of the players on the team. Next I want to determine the probability that a defenseman that played in the 2019 playoffs played for the Dallas Stars. In formula form that would look like:

$$P(Stars|D)=P(Stars\cap D)P(D)=n(Stars\cap D)n(D)$$

We just determined that the Stars have 9 defenseman and we also determined in part 2 that there are 333 players that participated in the 2019 playoffs. We end up with a fraction of **9/333** which is going to give us a small decimal, I got **0.02702** which is really small but it just gives you a perspective of how the number of defenseman for the Stars compares to the number of players that played in the playoffs.

I learned so much from this project before all of this I would've never been able to complete this project but I learned how to do so many new things in Excel and that can possibly

[redacted]

Introduction to Probability and Statistics

Mr. Nicholas Jacob

8/22/20

help me later on in life if needed. I’m super thankful I took this class because I came out of it with so much more knowledge than I thought I would.