**Описание анализированного набора данных**

Необходимые наборы данных были получены на stats.nba.com. Оттуда были загружены JSON файлы и конвертированы в CSV формат. Этот набор данных включает в себя все соответствующие данные регулярного сезона NBA 2017-2018.

Основываясь на этих данных, мы сможем экстраполировать на всю лигу, и, по причине постоянно меняющейся природе лиги, данные прошлого сезона дадут нам возможность сделать наилучший прогноз на следующие года.

Это обсервационное исследование, не эксперимент, поэтому мы не сможем установить причинные связи. Тем не менее, есть вероятность что найденные корреляции могут быть очень полезны сами по себе.

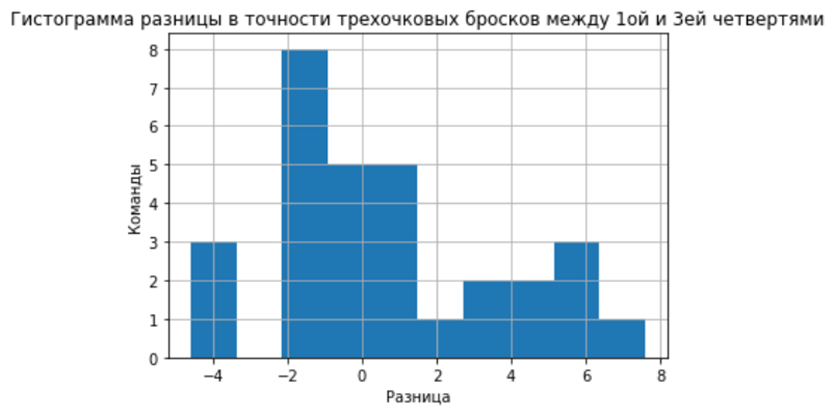
**Формулированные вопросы**

1. Есть ли разница в командной точности трехочковых бросков между 1-ой и 3-ей четвертями?
2. Есть ли связь между позициями и возрастом?
3. В играх, где преимущество в счете небольшое после первой половины (не более 5 очков), коррелирует ли ведение в счете с итоговой победой?
4. Есть ли разница в национальном происхождении игроков NBA между позициями?
5. Означает ли больший размах рук большее количество результативных отклонений и перехватов?

**1. Есть ли разница в командной точности трехочковых бросков между 1-ой и 3-ей четвертями?**

1-ая и 3-ая четверть были выбраны для сравнения

***Исследовательский анализ***



Заметьте, что этот график иллюстрирует разницу «Процент 3-х очковых в 1-ой четверти - Процент 3-х очковых в 1-ой четверти». Отрицательные значения отображают улучшение в точности трехочковых бросков в 3-ей четверти, а положительные значения отображают ухудшение в точности трехочковых бросков в 3-ей четверти.

*Описательная статистика*:

|  |  |
| --- | --- |
| счет | 30.000000 |
| среднее значение | 0.674881 |
| среднекв. отклонение | 3.145046 |
| мин | -4.581340 |
| 25% | -1.381035 |
| 50% | -0.091312 |
| 75% | 2.487729 |
| макс | 7.577884 |

Гистограмма разницы в командной точности трехочковых бросков между 1-ой и 3-ей четвертями скошена вправо (коэффициент асимметрии положителен), с одним пиком (унимодальная) и без выбросов. Распределение центрировано приблизительно на -0.1 и имеет среднеквадратическое отклонение приблизительно 3.1.

***Двухвыборочный t-критерий для зависимых выборок:***

Нулевая гипотеза полагает то, что между командными точностями 3-ех очковых бросков не существует существенной связи. Альтернативная гипотеза полагает то, что между командными точностями 3-ех очковых бросков существует существенная связь.

*Нулевая гипотеза H0*: μd = μ1−μ2 = 0, где

μ1 – точность 3-х очковых в первой четверти в процентах;

μ2 -точность 3-х очковых в третьей четверти в процентах;

μd – разница между μ1 и μ2;

*Альтернативная гипотеза Ha*: μd ≠ 0.

t-статистика и p-значение вычислены в Python:

t-statistic=1.1753325804178147, p-value=0.24942155727039717

***Заключение***

p-значение равно 0.25, что значит это не было бы маловерояно наблюсти те данные, которые мы наблюли, если бы между командными точностями трехочковых бросков в 1-ой и 3-ей четвертях не было связи. Основываясь на этом мы не можем опровергнуть нулевую гипотезу при уровне значимости α = 0.05, и принять альтернативную гипотезу. Мы не можем заключить что между командными точностями 3-ех очковых бросков существует существенная связь.

Учитывая радикальное улучшение в точности 3-ех очковых бросков (4.58%) от очень низкой в 1-ой четверти (32.86%) до выше средней в 3-ей четверти (37.44%), Миннесоте Тимбервулвс стоит рассмотреть вовлечение большего количества трехочковых бросков во время разминки. Это менее применимо к Голден Стэйт Ворриорс, так как они итак имеют хорошую точность в 1-ой четверти (38.87%), но выходят на сверхчеловеческий уровень в 3-ей четверти (43%): здесь можно видеть в количественном отображении популярное выражение «third quarter Warriors».

**2. В играх, где преимущество в счете небольшое после первой половины (не более 5 очков), коррелирует ли ведение в счете с итоговой победой?**

***Исследовательский анализ***

Таблица контингентности:

|  |  |  |  |
| --- | --- | --- | --- |
| Итоговая победа  Победа  после половины | Проигрыш | Выигрыш | Все |
| Нет | 121 | 94 | 215 |
| Да | 94 | 137 | 231 |
| Все | 215 | 231 | 446 |

Относительные частоты:

|  |  |  |
| --- | --- | --- |
| Итоговая победа  Победа  после половины | Проигрыш | Выигрыш |
| Нет | 0.562791 | 0.437209 |
| Да | 0.406926 | 0.593074 |
| Все | 0.482063 | 0.517937 |

***Критерий независимости хи-квадрат***

χ-statistic = 10.834980000149809, p-value = 0.0009960039387090216

At 5 it does significantly

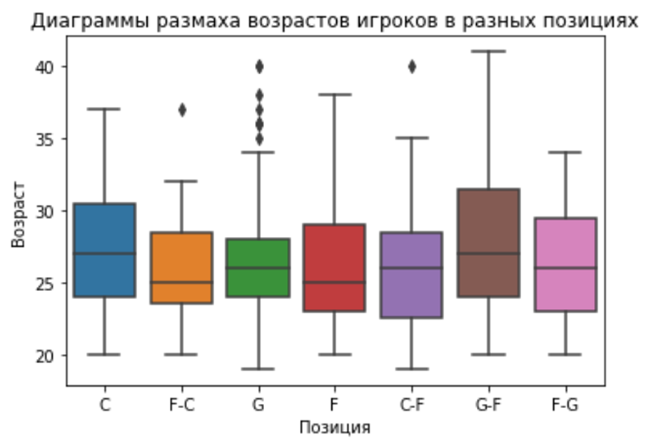
Note that at 4 doesn’t (don’t have evidence).

Exactly 4 points: χ-statistic = 0.768171114599685, p-value = 0.38078342795092013

Within 4 points: χ-statistic = 2.1520447236647535, p-value = 0.14238016058673728

Surprising that at 4 (relatively big lead) we don’t have it. Maybe it’s the complacency of winning at half-time when the game is close.

**3. Is there a relationship between position and age?**

****

Описательная статистика:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Позиция | счет | Среднее значение | Среднекв. отклонение | мин. | 25% | 50% | 75% | макс |
| C | 63.0 | 27.111111 | 4.399821 | 20.0 | 24.0 | 27.0 | 30.5 | 37.0 |
| C-F | 23.0 | 26.217391 | 5.071816 | 19.0 | 22.5 | 26.0 | 28.5 | 40.0 |
| F | 150.0 | 26.426667 | 4.045623 | 20.0 | 23.0 | 25.0 | 29.0 | 38.0 |
| F-C | 31.0 | 25.935484 | 3.881026 | 20.0 | 23.5 | 25.0 | 28.5 | 37.0 |
| F-G | 23.0 | 26.782609 | 4.155572 | 20.0 | 23.0 | 26.0 | 29.5 | 34.0 |
| G | 205.0 | 26.321951 | 4.006582 | 19.0 | 24.0 | 26.0 | 28.0 | 40.0 |
| G-F | 39.0 | 27.717949 | 5.155216 | 20.0 | 24.0 | 27.0 | 31.5 | 41.0 |

Multiple C->Q ANOVA

There are outliers in the positions “C-F” and “F-C” - Dirk Nowitzki and Nick Collison, and the sample sizes may not necessarily be enough to guarantee normal distribution of the …

Running ANOVA on the full data:

statistic=0.955603760244476, p-value=0.45477740597358385

Running ANOVA on the data without the outliers in the "C-F' and 'F-C" positions:

statistic=1.311318715208154, p-value=0.2501211510855592

The outliers did not influence the results.

Not statistically significant. Can’t reject the null hypothesis.

**4. Is the origin makeup of NBA players different across positions?**

Remove data entry outlier Walter Lemon Jr. He's not the only "pure" PG, other pure PGs were classified as guards, so we’ll consider this data entry error.

Two-way table:

|  |  |  |  |
| --- | --- | --- | --- |
| Origin /  PLAYER\_POSITION | International | USA | All |
| C | 30 | 33 | 63 |
| C-F | 8 | 15 | 23 |
| F | 27 123 | 150 |  |
| F-C | 10 | 21 | 31 |
| F-G | 4 | 19 | 23 |
| G | 28 | 177 | 205 |
| G-F | 7 | 32 | 39 |
| All | 114 | 420 | 534 |

Conditional percentages:

|  |  |  |
| --- | --- | --- |
| Origin /  PLAYER\_POSITION | International | USA |
| C | 0.476190 | 0.523810 |
| C-F | 0.347826 | 0.652174 |
| F | 0.180000 | 0.820000 |
| F-C | 0.322581 | 0.677419 |
| F-G | 0.173913 | 0.826087 |
| G | 0.136585 | 0.863415 |
| G-F | 0.179487 | 0.820513 |
| All | 0.213483 | 0.786517 |

C->C Chi-square test for independence

statistic= 39.26852697080573, p-value=6.340368194661366e-07

p-value practically zero, statistically significant

shows that international basketball more likely to be centers or big men, and the trend in US basketball towards guards (“smallball”)

**5. Does an increase in wingspan relate to increase in the number of deflections and steals?**

We’ll choose only players who had played more than 15 games.

The graphs below show the relationship between height in inches and the average number of deflections recorded by NBA players during regular season.

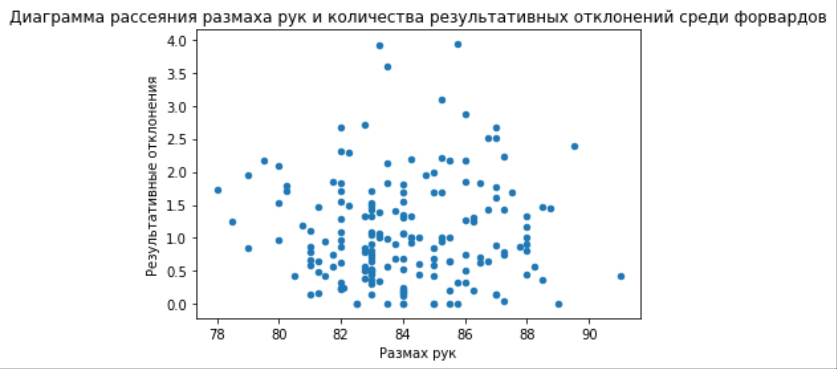
Scatterplot for all the players in the NBA:



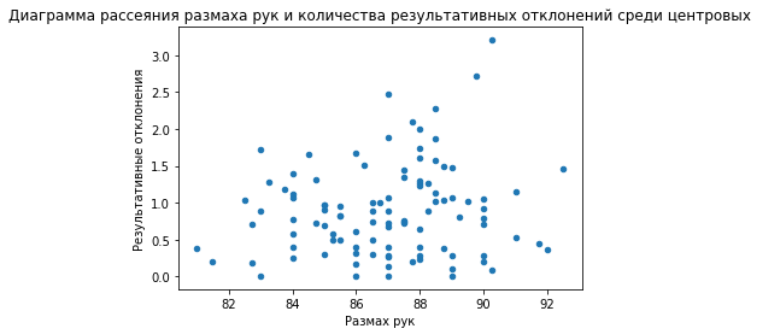
For guards:



For forwards:



For centers:

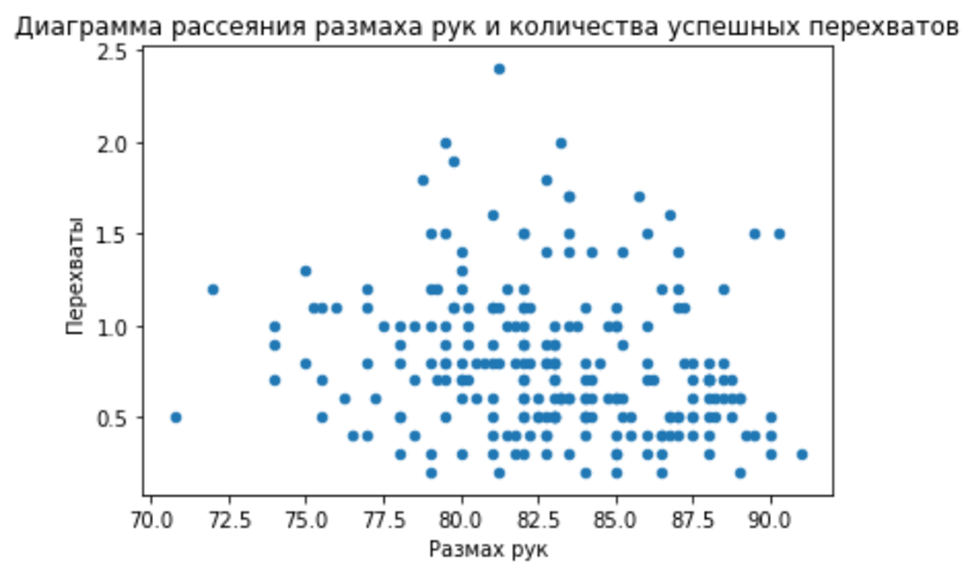


Q->Q regression t-test for the slope.

The closest we have is in the case of centers, with correlation coefficient r-value=0.14536953391373944, and test p-value=0.14288568244203492.

The graphs below show the relationship between height in inches and the average number of steals recorded by NBA players players during regular season.

Scatterplot for all the players in the NBA:



For guards:



For forwards:



For centers:



Q->Q regression t-test for the slope.

None of the relationships are significant at 0.05 significance level with the selected minimum game cutoff of 15, although for guards is close at p-value=0.05779801365934409. But with bigger cutoffs for games played (i.e. 20, 30, 40) the p-value is even bigger, so the relationship doesn’t uphold).

For centers correlation coefficient r-value=0.19791927433929551, and the test p-value=0.06612281263777153. Increasing the cutoff for minimum games played to 25 we get r-value=0.2165473177952725, and p-value=0.046525524153957186.

At the cutoff of 30 games minimum, we get r-value=0.25191047346980966, p-value=0.02242571458638422. It further upholds at the cutoff level of 40 games.

For our purposes, we can consider this statistically significant, and reject the null hypothesis. We can conclude that for centers there is, if weak, a positive linear relationship between wingspan in inches and number of steals recorded.