**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, возможно найденные корреляции могут быть очень полезны сами по себе.

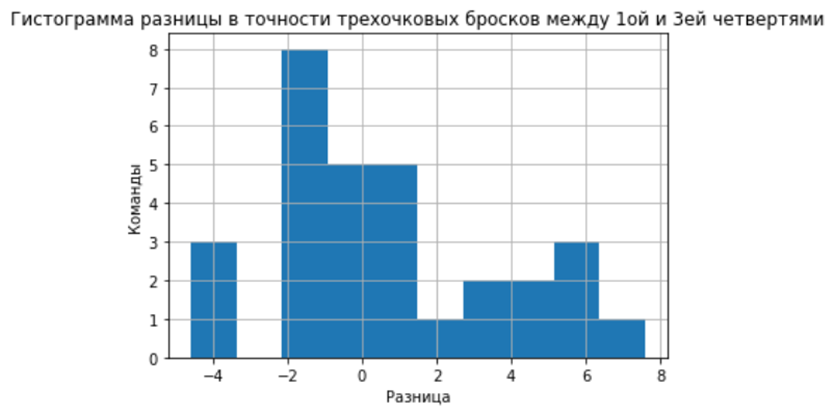
**Сформулированные вопросы**

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

**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 значимой связи. Альтернативная гипотеза полагает то, что между командными точностями 3-ех очковых бросков существует значимая связь.

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

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

μ2 -средняя 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 наблюдать те данные, которые мы наблюдаем, 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 значимая 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 после первой половины небольшое (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***

*Нулевая гипотеза H0*: между небольшим преимуществом в счете после первой половины (не более 5 очков) и итоговой победой не существует значимой связи.

*Альтернативная гипотеза Ha*: между небольшим преимуществом в счете после первой половины (не более 5 очков) и итоговой победой существует значимая связь.

χ-статистика и p-значение вычислены в Python:

χ-statistic = 10.834980000149809, p-value = 0.0009960039387090216

***Заключение***

p-значение равно 0.001, что значит это было бы очень маловерояно наблюдать те данные, которые мы наблюдаем, если бы между преимуществом в счете после первой половины (не более 5 очков) и итоговой победой не существовало связи. Основываясь на этом, мы можем опровергнуть нулевую гипотезу при уровне значимости α = 0.05, и принять альтернативную гипотезу. Мы можем заключить, что между небольшим преимуществом в счете после первой половины (не более 5 очков) и итоговой победой существует значимая связь.

Интересно, что при установленном пороге “небольшого” преимущества в 4 очка, данные не предоставляют доказательства о значимости связи.

Хи-квадрат тест, для игр с преимуществом после половины не более 4 очков: χ-statistic = 2.1520447236647535, p-value = 0.14238016058673728

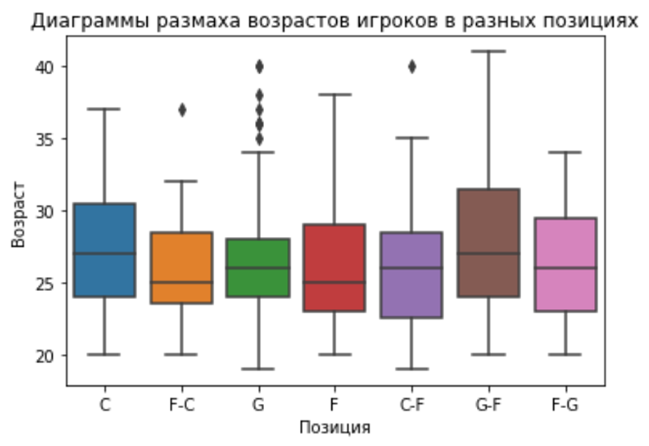
Хи-квадрат тест, для игр с преимуществом после половины ровно в 4 очка: χ-statistic = 0.768171114599685, p-value = 0.38078342795092013

Команды могут использовать эту информацию, для того чтобы быть более бдительнее при преимуществе в счете после половины менее 5 очков. Возможно, часто такое ведение в счете дает больше чувства ложной безопасности, одновременно мотивируя соперника быть лучше во второй половине, чем оно дает практического преимущества.

**3. Есть ли связь между позициями игроков и их возрастом?**

***Исследовательский анализ***

Уолтер Лемон Мл. единственный из 540 игроков кто был помечен как “PG” в наборе данных. Он не единственный игрок, рассматриваемый как “чистый” point guard в лиге, и все прочие “чистые” point guards были помечены как “G”. Потому мы будем трактовать это отдельное наблюдение как выброс, связанный с ошибкой в вводе данных, и исключим его из набора данных, используемого для анализа.

****

*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 |

Основываясь на графике и численных измерениях, мы можем сделать следующие заметки насчет имеющихся распределений:

**Центр**: график показывает что средние значения распределений предельно близки друг к другу. Максимальный средний возраст имеет позиция “C” (27.11), минимальный средний возраст имеет позиция “F-C” (25.95).

**Вариация**: судя по размаху данных, наибольшую вариацию имеют позиции “С-F” и “G” и “G-F” (21 год), наименьшую позиция “F-G” (14 лет). Если мы посмотрим на межквартильный размах, который отображает вариацию только 50 серединных процентов распределения, наибольшую вариацию имеет позиция “G-F” (7.5 лет), наименьшую позиция “G” (4 года).

**Выбросы**: позиции “F-C” и “C-F” имеют по одному верхнему выбросу (Ник Коллисон и Дирк Новитцки, соответственно). Позиция “G” имеет 5 верхних выбросов.

***Дисперсионный анализ***

В позициях “C-F” и “F-C” есть выбросы, и размер выборки в этих группах недостаточно большой для того, чтобы вступила в действие Центральная Предельная Теорема и обеспечила тот факт, что выборочное среднее имеет нормальное распределение, тем самым позволив нам гарантированно использовать дисперсионный анализ. Размер выборки для позиции “G” достаточно большой, что мы можем игнорировать выбросы в этой группе.

Нулевая гипотеза полагает то, что между позицией игрока и его возрастом не существует значимой связи. Альтернативная гипотеза полагает то, что между позицией игрока и его возрастом существует значимая связь.

*Нулевая гипотеза H0*: μ1=μ2=μ3=μ4=μ5=μ6=μ7, где

μ1, μ2,μ3,μ4,μ5,μ6,,μ7 - средний возраст игроков в позициях “C”, “C-F”, “F”, “F-C”, “F-G”, “G”, “G-F”, соответственно.

*Альтернативная гипотеза Ha*: не все μ равны.

f-статистика и p-значение вычислены в Python:

Проведение дисперсионного анализа на полном наборе данных:

f-statistic=0.955603760244476, p-value=0.45477740597358385

Проведение дисперсионного анализа на наборе данных, из которого были исключены выбросы в позициях "C-F' и 'F-C":

f-statistic=1.311318715208154, p-value=0.2501211510855592

Выбросы не повлияли на результат анализа (при уровне значимости α = 0.05).

***Заключение***

p-значение равно 0.25 с включенными выбросами и 0.45 без выбросов, что значит это не было бы маловерояно наблюдать те данные, которые мы наблюдаем, если бы между позицией игрока и его возрастом 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 между позицией игрока и его возрастом существует значимая связь.

Несмотря на популярную рекомендацию игрокам в годах подняться “вверх по позиции” (point guard -> shooting guard -> small forward -> power forward -> center) для продления карьеры, данные не предоставляют достаточно доказательства, что существенное количество игроков выбирает такой путь.

**4. Есть ли разница в национальном происхождении игроков NBA между позициями?**

***Исследовательский анализ***

Уолтер Лемон Мл. единственный из 540 игроков кто был помечен как “PG” в наборе данных. Он не единственный игрок, рассматриваемый как “чистый” point guard в лиге, и все прочие “чистые” point guards были помечены как “G”. Потому мы будем трактовать это отдельное наблюдение как выброс, связанный с ошибкой в вводе данных, и отбросим его из набора данных, используемого для анализа.

Таблица контингентности:

|  |  |  |  |
| --- | --- | --- | --- |
| Происхождение  Позиция | Интернациональный | США | Все |
| 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 |

Относительные частоты:

|  |  |  |
| --- | --- | --- |
| Происхождение  Позиция | Интернациональный | США |
| 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 |

***Критерий независимости хи-квадрат***

*Нулевая гипотеза H0*: между происхождением игроков и их позициями не существует значимой связи.

*Альтернативная гипотеза Ha*: между происхождением игроков и их позициями существует значимая связь.

χ-статистика и p-значение вычислены в Python:

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

***Заключение***

p-значение близко к 0, что значит это было бы практически невероятно наблюдать те данные, которые мы наблюдаем, если бы между между происхождением игроков и их позициями не существовало связи. Основываясь на этом, мы можем опровергнуть нулевую гипотезу при уровне значимости α = 0.05, и принять альтернативную гипотезу. Мы можем заключить, что между происхождением игроков и их позициями существует значимая связь.

Это заключение говорит нам о том, что иностранный игрок имеет наибольший шанс попасть в NBA, если он играет в позиции центрового.

Также интересна комбинация доминирования игроков из США в позиции “Guard” и отсутствие их доминирования в позиции “Center”, и как это взаимодействует с основным трендом в лиге - переходе к “smallball” (меньше традиционных больших игроков, требование новых технических навыков от больших игроков).

На образовательном уровне, NFHS и NCAA будут заинтересованы, почему для позиции центрового организации NBA часто прибегают к международным вариантам.

**5. Можем ли мы предугадать количество результативных отклонений и перехватов по размаху рук?**

Проверка на наличие линейной связи.

Для того чтобы каждое наблюдение в наборе данных было более репрезентативно по отношению к настоящему умению игрока, мы рассмотрим в этом анализе только тех игроков, кто сыграл более 15 игр в 2017-18 сезоне NBA.

**5.a.** **Можем ли мы предугадать количество результативных отклонений по размаху рук?**

***Исследовательский анализ***

Графики ниже отображают отношение между размахом рук и средним за игру количеством результативных отклонений, выполненных игроками NBA на протяжении регулярного сезона.

Диаграмма рассеяния среди всех игроков NBA:



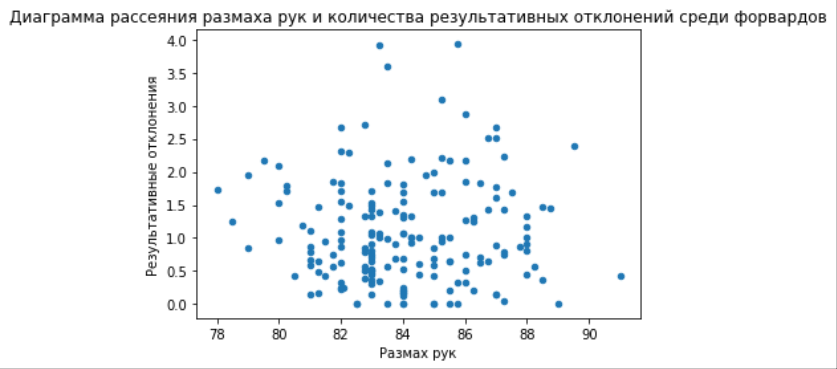
Отношение слабое отрицательное. Коэффициент корреляции r равен -0.1161291949153148.

Среди защитников:



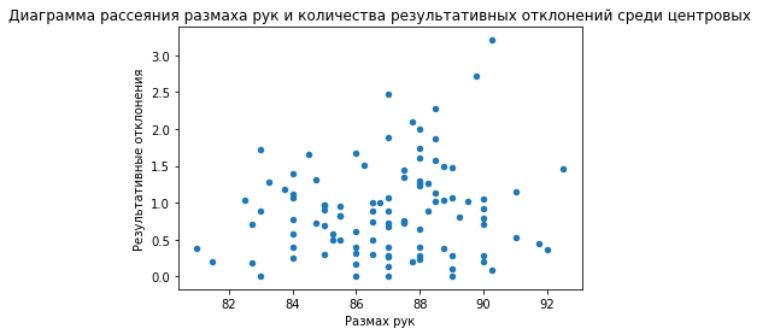
Отношение слабое положительное. Коэффициент корреляции r равен 0.1685162215044324.

Среди форвардов:



Отношение практически не существует. Коэффициент корреляции r равен rvalue=0.004339493478112684.

Среди центровых:



Отношение слабое положительное. Коэффициент корреляции r равен 0.11181464847674216

***t-критерий коэффициента линейной регрессии***

*Нулевая гипотеза H0*: между размахом рук и количеством результативных отклонений не существует линейной связи.

*Альтернативная гипотеза Ha*: между размахом рук и количеством результативных отклонений существует линейная связь.

Статистические вычисления были произведены в Python:

Среди всех игроков:

slope=-0.021622998668061597, intercept=3.0309175784659437, rvalue=-0.11612919491531477, pvalue=0.016873764176054365

Среди защитников:

slope=0.044586148284437035, intercept=-2.122655924334114, rvalue=0.1685162215044324, pvalue=0.02185052959921897

Среди форвардов:

slope=0.0013933039324277492, intercept=1.092332411440531, rvalue=0.004339493478112684, pvalue=0.9585415700906051, stderr=0.02675601850381753

Среди центровых:

slope=0.028901811699850916, intercept=-1.5653158783690861, rvalue=0.11181464847674216, pvalue=0.2886187731620592, stderr=0.02707528948056095

***Заключение***

p-значение в разрезе для защитников равно 0.022, что значит это было бы маловерояно наблюдать те данные, которые мы наблюдаем, если бы между размахом рук и количеством результативных отклонений не существовало линейной связи. Основываясь на этом мы можем опровергнуть нулевую гипотезу при уровне значимости α = 0.05, и принять альтернативную гипотезу. Мы можем заключить, что среди защитников между размахом рук и количеством результативных отклонений существует (слабая) положительная линейная связь.

Несмотря на то, что общий анализ для всех игроков также возвращает статистически значимое значение, отдельные анализы для позиций форвардов и центровых указывают на то, что имеющиеся данные не предоставляют достаточного доказательства о наличии линейной связи.

**5.b.** **Можем ли мы предугадать количество перехватов по размаху рук?**

***Исследовательский анализ***

Графики ниже отображают отношение между размахом рук и средним за игру количеством перехватов, выполненных игроками NBA на протяжении регулярного сезона.

Диаграмма рассеяния среди всех игроков NBA:



Отношение слабое отрицательное. Коэффициент корреляции r равен =-0.15365262987825057.

Среди защитников:



Отношение слабое положительное. Коэффициент корреляции r равен 0.1412903701579556.

Среди форвардов:



Отношение практически отсутствует. Коэффициент корреляции r равен -0.002638453597043652.

Среди центровых:



Отношение слабое положительное. Коэффициент корреляции r равен 0.14513319922216392.

Интересно, что в имеющемся наборе данных мы можем наблюдать парадокс Симпсона - при наличии нескольких групп данных, в каждой из которых наблюдается одинаково направленная зависимость, при объединении этих групп направление зависимости меняется на противоположное.

***t-критерий коэффициента линейной регрессии***

*Нулевая гипотеза H0*: между размахом рук и количеством перехватов не существует линейной связи.

*Альтернативная гипотеза Ha*: между размахом рук и количеством перехватов существует линейная связь.

p-значения были вычислены в Python.

Среди всех игроков:

slope=-0.015654266991884564, intercept=1.9793296177709365, rvalue=-0.15365262987825057, pvalue=0.0017152888932548994

Среди защитников:

slope=0.020907033326887838, intercept=-0.8600338179506402, rvalue=0.1412903701579556, pvalue=0.05779801365934409

Среди форвардов:

slope=-0.0004494598765432127, intercept=0.6950776427469139, rvalue=-0.002638453597043652, pvalue=0.975050662184907

Среди центровых:

slope=0.028901811699850916, intercept=-1.5653158783690861, rvalue=0.11181464847674216, pvalue=0.2886187731620592, stderr=0.02707528948056095

***Заключение***

p-значение для всех игроков и p-значения в разрезах больше уровня значимости, что значит это не было бы маловерояно наблюдать те данные, которые мы наблюдаем, если бы среди центровых между размахом рук и количеством перехватом не существовало линейной 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 среди центровых между размахом рук и количеством перехватов существует линейная связь.

С изменением в росте, мы не можем предугадать изменение количество перехватов, но в случае защитников мы можем предугадать изменение количества результативных отклонений.

*Замечание*. Все заключения в 5-ом вопросе применимы только для формирования прогнозов о выступлениях игроков, которые принимают участие более чем в 15 играх сезона.

**Подведение итогов**

В данной работе мы анализировали данные регулярного сезона NBA 2017-2018.

Мы нашли что, между позициями игроков и их возрастом нет значительной связи, однако национальное происхождение значительно различается между позициями. Было неожиданно узнать, что в командной точности трехочковых бросков нет существенной разницы между первой и третьей четвертью, и ведение в счете даже в 4 очка после половины игры не дает значительной гарантии итоговой победы. В добавок, мы обнаружили что для защитников существует положительная линейная связь между размахом рук и количеством результативных отклонений.