СТАТИСТИЧЕСКИЙ АНАЛИЗ РАЗЛИЧНЫХ ИЗМЕРЕНИЙ ЛИЦА И ГОЛОВЫ ДЛЯ СРАВНЕНИЯ ДВУХ ГРУПП (ПРОФЕССИОНАЛЬНЫХ БОЙЦОВ И ОБЫЧНЫХ ЛЮДЕЙ)

МАТЕРИАЛЫ

В исследование использовались два набора данных содержащие информацию о различных измерениях, выполненных для двух групп людей (профессиональных бойцов смешанных единоборств и обычных людей).

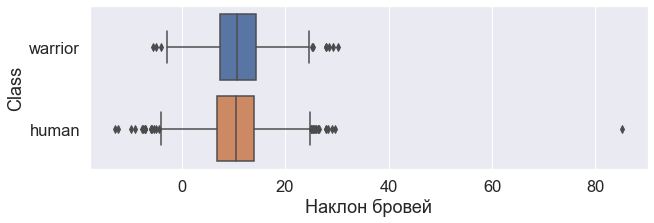
Первый набор данных имеет следующие характеристики:

* количество экземпляров – 3227;
* количество признаков – 19, все признаки являются непрерывными;
* количество классов – 2;
* количество профессиональных бойцов – 1491;
* количество обычных людей – 1736.

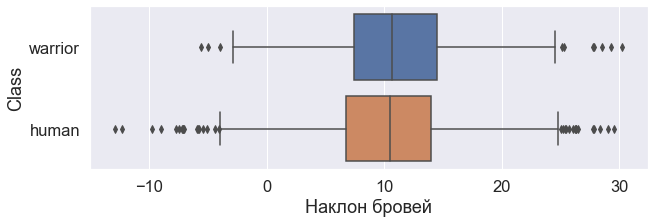
В процессе разведочного анализа обнаружены не типичные выбросы, предположительно возникшие в результате ошибки работы алгоритма распознавания черт лица.

Чтобы избавится только от ошибок измерения и удалить как можно меньше полезной информации, из набора данных удалялись объекты чьё значение какого‑либо признака отличалось от среднего арифметического сильнее чем на пять стандартных отклонений.

Таким способом было удалено 47 объектов, а ошибки измерения встретились в 14 различных признаках. На рисунках 1-14 представлены диаграммы размаха признаков до и после удаления ошибок измерения.

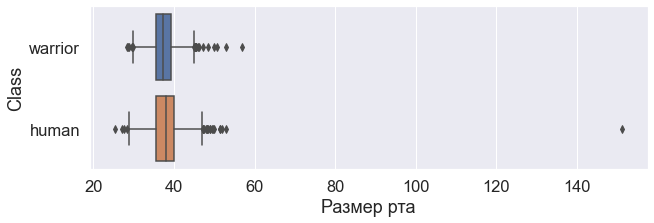


(а) — диаграмма размаха до удаления ошибок измерения

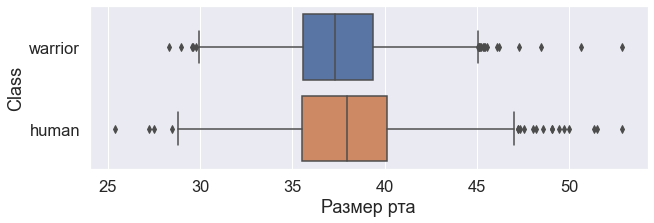


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 1 — диаграммы плотности размаха для признака ‘Наклон бровей’

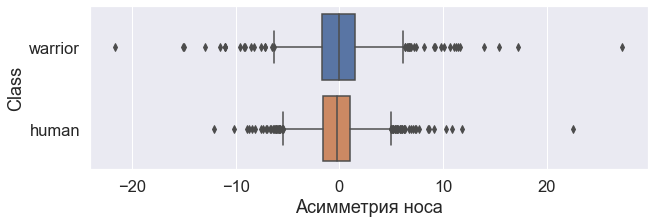


(а) — диаграмма размаха до удаления ошибок измерения

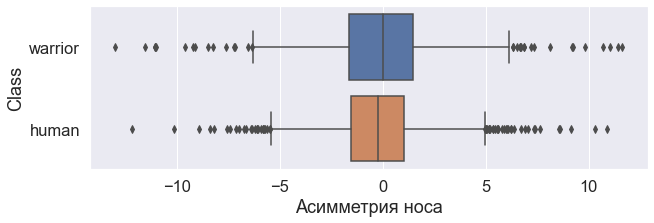


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 2 — диаграммы плотности размаха для признака ‘Размер рта’

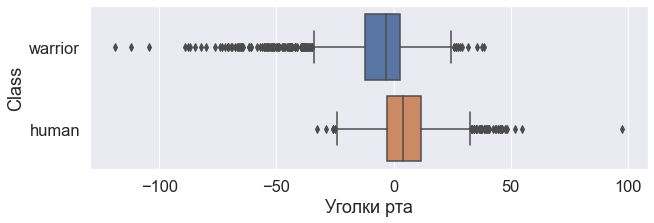


(а) — диаграмма размаха до удаления ошибок измерения

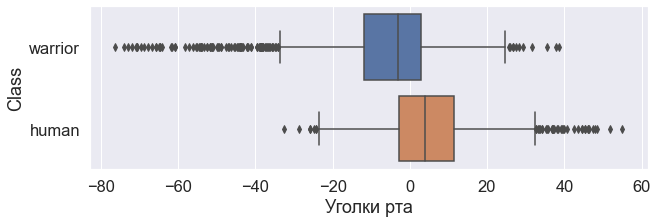


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 3 — диаграммы плотности размаха для признака ‘Асимметрия носа’

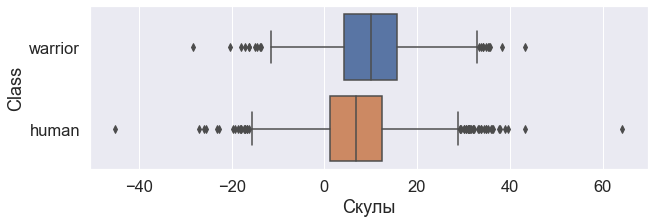


(а) — диаграмма размаха до удаления ошибок измерения

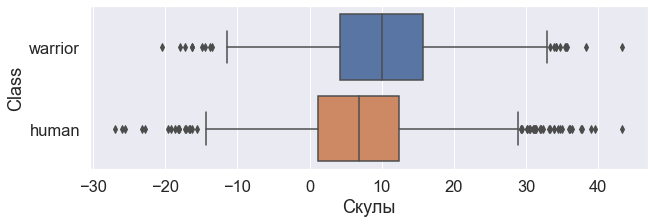


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 4 — диаграммы плотности размаха для признака ‘Уголки рта’

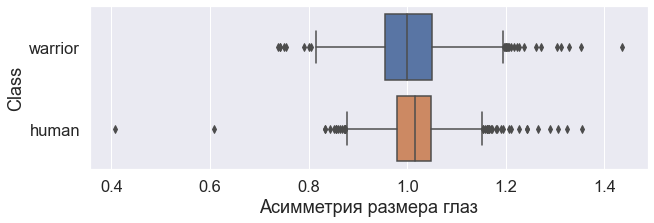


(а) — диаграмма размаха до удаления ошибок измерения

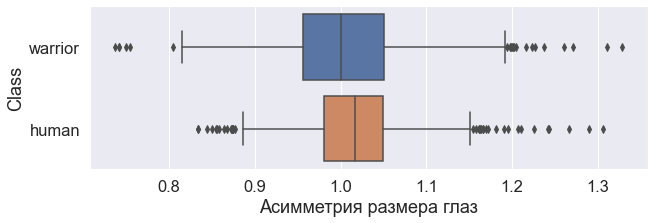


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 5 — диаграммы плотности размаха для признака ‘Скулы’

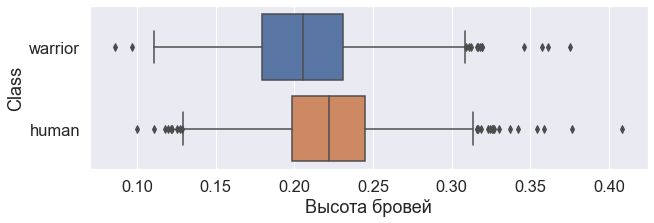


(а) — диаграмма размаха до удаления ошибок измерения

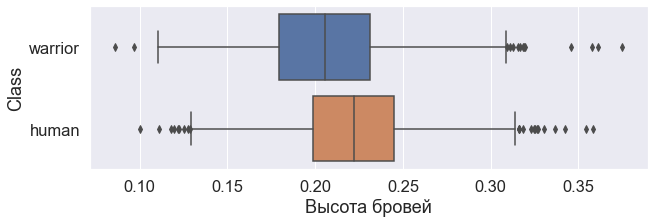


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 6 — диаграммы плотности размаха для признака ‘Асимметрия размера глаз’



(а) — диаграмма размаха до удаления ошибок измерения

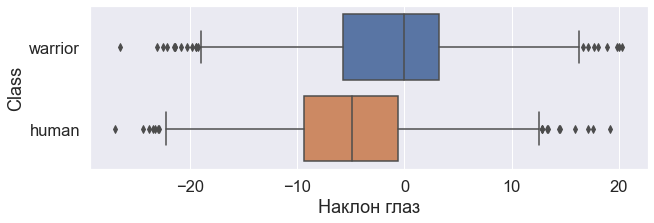


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 7 — диаграммы плотности размаха для признака ‘Высота бровей’

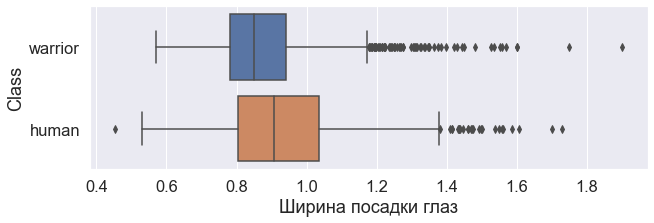


(а) — диаграмма размаха до удаления ошибок измерения

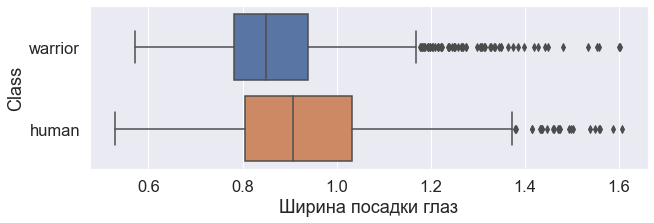


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 8 — диаграммы плотности размаха для признака ‘Наклон глаз’

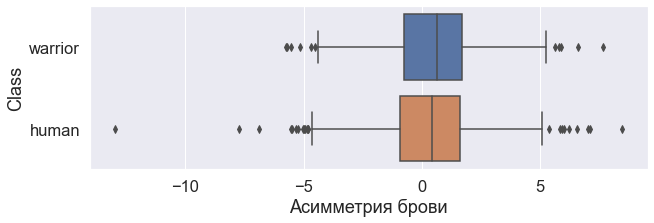


(а) — диаграмма размаха до удаления ошибок измерения

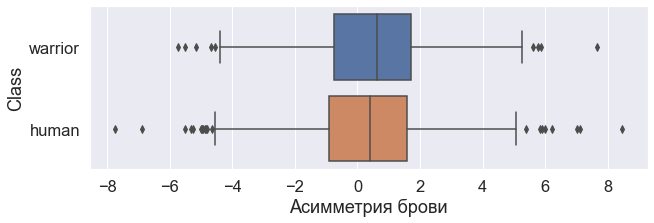


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 9 — диаграммы плотности размаха для признака ‘Наклон глаз’

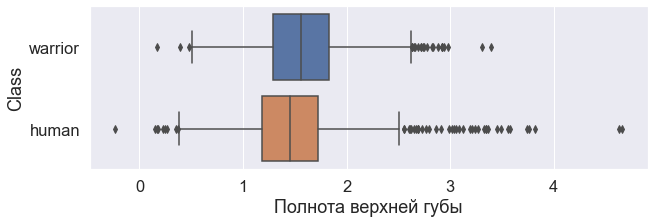


(а) — диаграмма размаха до удаления ошибок измерения

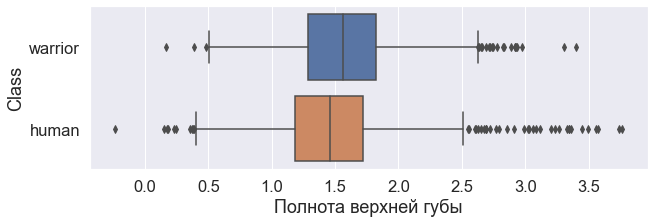


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 10 — диаграммы плотности размаха для признака ‘Наклон глаз’

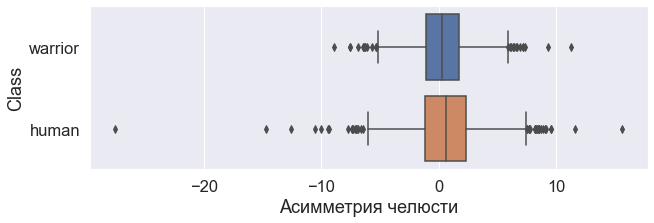


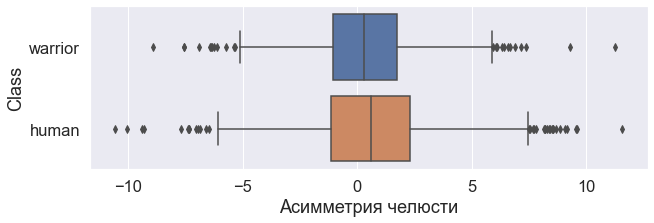
(а) — диаграмма размаха до удаления ошибок измерения



(б) — диаграмма размаха после удаления ошибок измерения

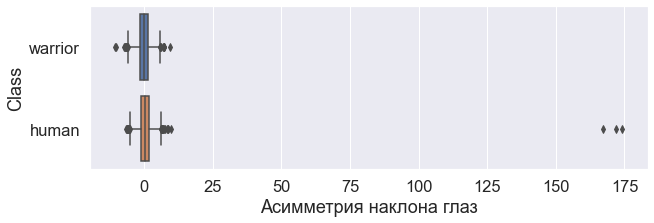
Рисунок 11 — диаграммы плотности размаха для признака ‘Наклон глаз’



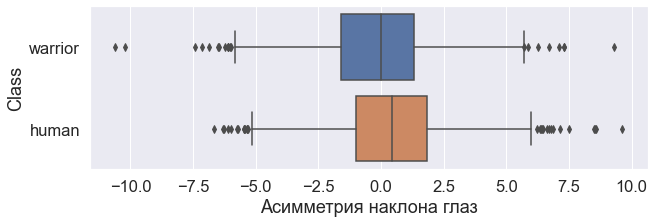
(а) — диаграмма размаха до удаления ошибок измерения 

(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 12 — диаграммы плотности размаха для признака ‘Наклон глаз’

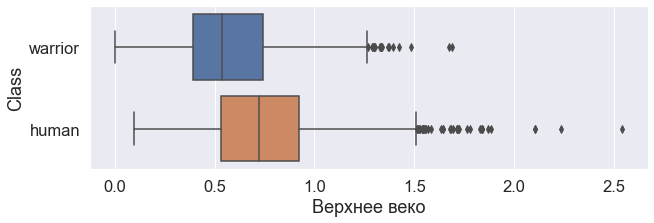


(а) — диаграмма размаха до удаления ошибок измерения

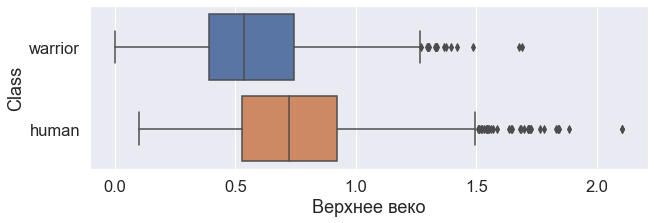


(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 13 — диаграммы плотности размаха для признака ‘Наклон глаз’



(а) — диаграмма размаха до удаления ошибок измерения



(б) — диаграмма размаха после удаления ошибок измерения

Рисунок 14 — диаграммы плотности размаха для признака ‘Наклон глаз’

После проведения предварительной обработки итоговый набор данных имеет следующие характеристики:

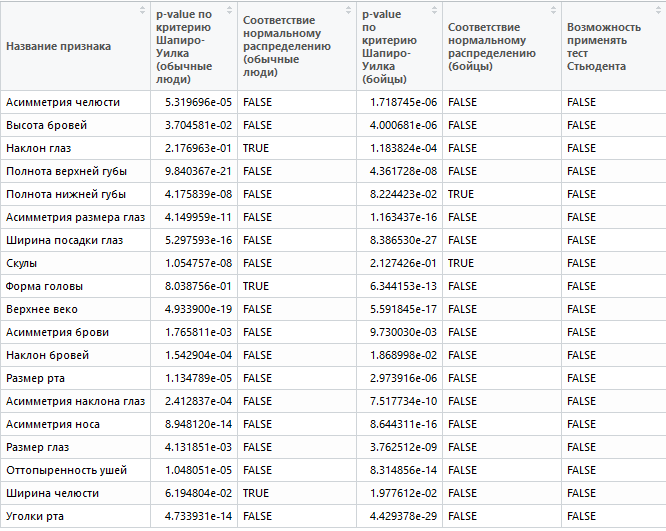
* количество экземпляров – 3188;
* количество профессиональных бойцов – 1472;
* количество обычных людей – 1716.

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Провести сравнение двух групп на предмет обнаружение статистических различий между ними можно несколькими способами. Стандартным методом для подобных исследований является использование критерия Стьюдента с нулевой гипотезой о равенстве средних в генеральных совокупностях двух сравниваемых групп. Однако данный тест имеет ряд ограничений такие как сравниваемые группы должны иметь нормальное распределение, а также их дисперсии должны быть равны. В связи с этим нам необходимо проверить данные условия.

Соответствие нормальному распределению является наиболее важным условием, так как имеются модификации теста Стьюдента для выборок с разной дисперсией. Для проверки на соответствие нормальному распределению использовался критерий Шапиро-Уилка. Нулевой гипотезой для критерия Шапиро‑Уилка является соответствие нормальному распределения, а значит если вычисленное значение p-value меньше выбранного как значимое (0.05), то необходимо заключить что распределение не соответствует нормальному. Результаты проверки представлены в таблице 1.

Таблица 1. Проверка признаков на нормальное распределение.



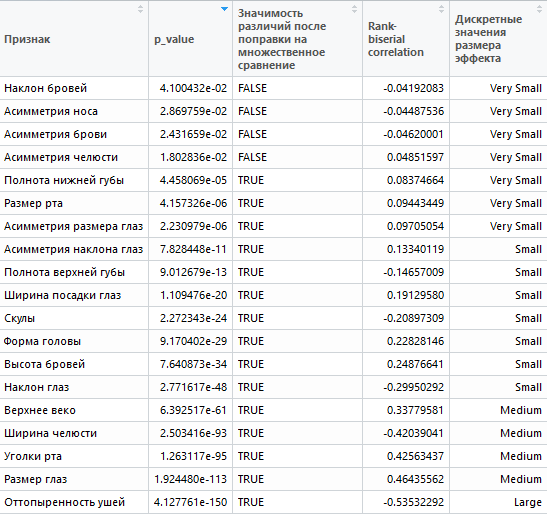
В связи с результатами проверки признаков на соответствие нормальному распределению необходимо проводить сравнение двух групп на предмет обнаружение статистических различий с помощью непараметрических тестов. В таком случае использование критерия Манна-Уитни является наиболее предпочтительных. Для вычисления размера эффекта в данном исследование использовалась рангово-бисериальная корреляция (rank biserial correlation, RBC), данный метод является стандартным для исследования размера эффекта при обнаружении статистически значимых различий с помощью критерия Манна-Уитни.

Значения рангово-бисериальная корреляция принадлежат диапазону [‑1; 1], где 0 это отсутствие эффекта, отрицательные и положительные значения сообщают об отрицательной и положительной взаимосвязи соответственно.

В качестве порога статистической значимости выбрано значение p-value равное 0.01, в качестве поправки на множественное сравнение – поправка Бонферрони (после поправки значимый p-value равен 0.0004347826).

Результаты измерения уровня статистической значимости различий и размера эффекта представлены в таблице 2.

Таблица 2. Результаты проверки статистической значимости и размера эффекта.



Как видно из таблицы признаками, не прошедшими порог статистической значимости, являются: асимметрия челюсти, асимметрия брови, наклон бровей и асимметрия носа.

Четыре признака имеют средние значения рангово-бисериальная корреляция — верхнее веко, ширина челюсти, уголки рта и размер глаз. На рисунках 15-19 представлен график плотности для каждого признака.

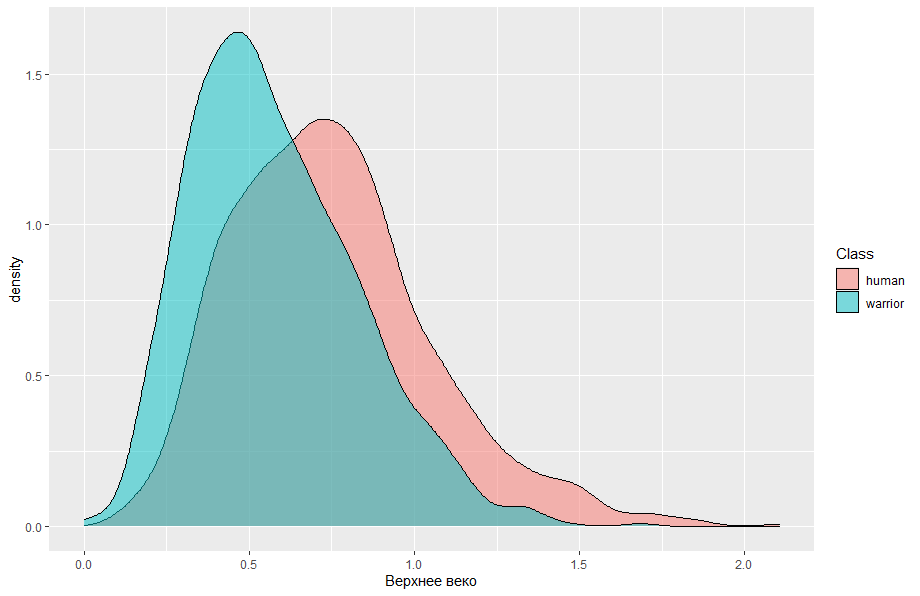


Рисунок 19 — График плотности для признака Верхнее веко

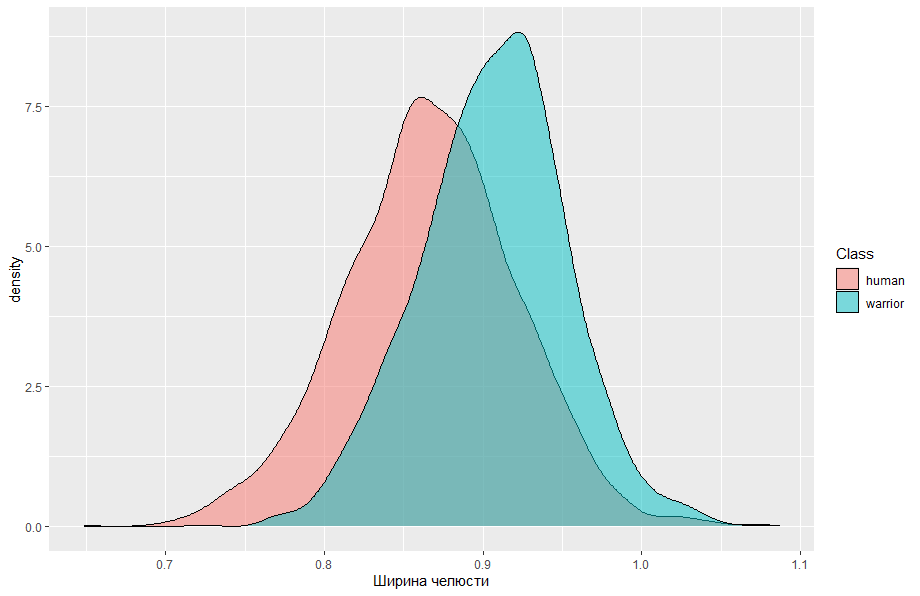


Рисунок 19 — График плотности для признака Ширина челюсти

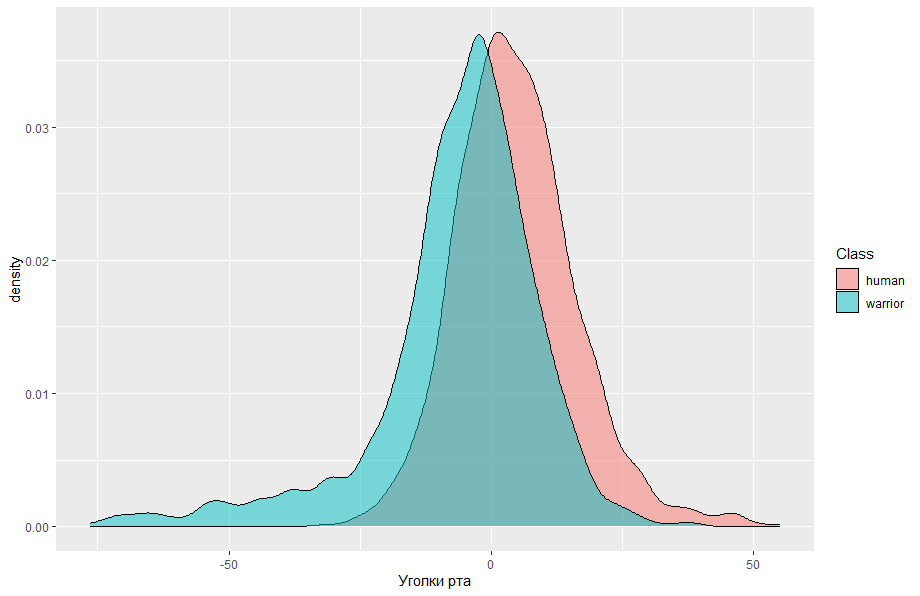


Рисунок 19 — График плотности для признака Уголки рта

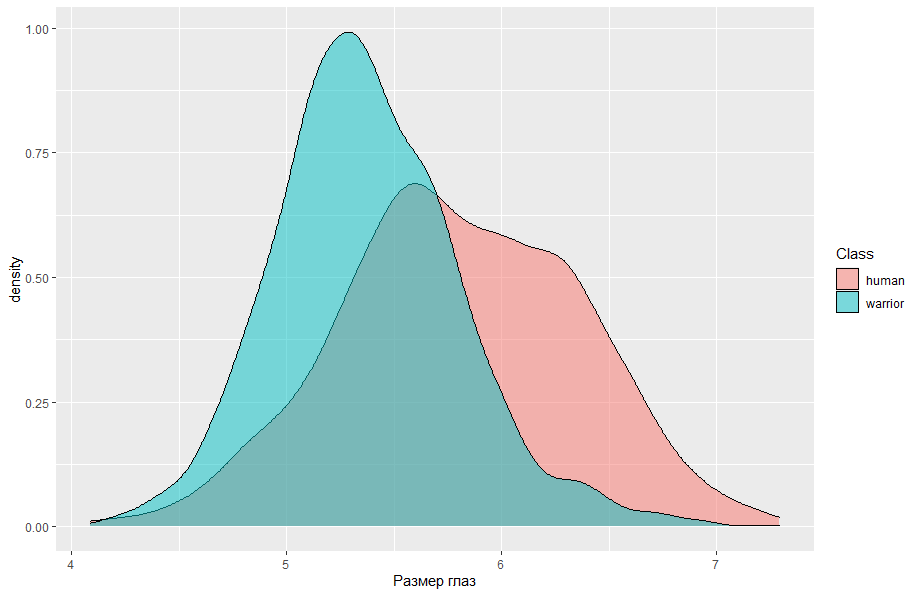


Рисунок 19 — График плотности для признака Размер глаз

Единственным признаком имеющим большой размер эффекта (больше 0.5) является оттопыренность ушей, что является вполне закономерным, так как у профессиональных борцов, а борьба один из ключевых навыков смешанных единоборств, травма уха, приводящая к его оттопыренности является профессиональной. Фактически это является **приобретенным признаком** в связи с чем и наблюдаются такие различия между двумя исследуемыми группами (рисунок 20).



Рисунок 20 — График плотности для признака Оттопыренность ушей

BOOTSTRAP

Бутстрэп (англ. bootstrap) в статистике — практический компьютерный метод исследования распределения статистик вероятностных распределений, основанный на многократной генерации выборок методом Монте-Карло на базе имеющейся выборки. Позволяет просто и быстро оценивать самые разные статистики (доверительные интервалы, дисперсию, корреляцию и так далее) для сложных моделей.

Методология его применения для построения распределения выборочной статистики представлена на рисунке 21. Из исходных данных с извлекается с возвращением несколько подвыборок (например, 5000 подвыборок), такого же размера и для каждой вычисляется необходимая статистика (например, среднее арифметическое). Затем строится распределение выбранной статистики. Теперь для полученного распределения возможно построить доверительные интервалы.

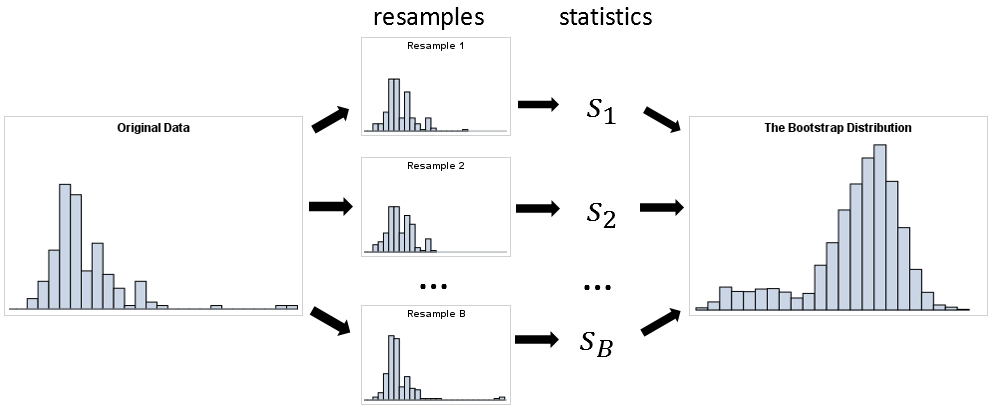


Рисунок 21 — Применение бутстрэпа для построения распределения выборочной статистики

Применение бутстрэпа для одного признака в нашем исследование заключается в следующем:

1. Выборка бойцов и обычных людей соединяются вместе.
2. Из объединенной выборки извлекаем с повторение две выборки исходного размера.
3. Считаем в каждой выборке искомую статистику и вычисляем их разность.
4. Запоминаем.
5. После повторения шагов 1-4 несколько тысяч раз мы можем построить распределение интересующей нас статистики и доверительные интервалы к нему.
6. Проверяем как соотносится наблюдаемое различие в статистиках и прогнозируемое с помощью бутстрэпа.

Таким образом мы исходим из следующей логики что если бы 2 выборки (бойцы и обычные люди) по исследуемому параметру были из одного распределения, то наблюдаемая разность исследуемой статистики находилась внутри доверительных интервалов. На рисунке 22 и 23 представлены две возможные ситуации. Пунктирные линии — это границы доверительного интервала, черная сплошная линия реально наблюдаемые различия.

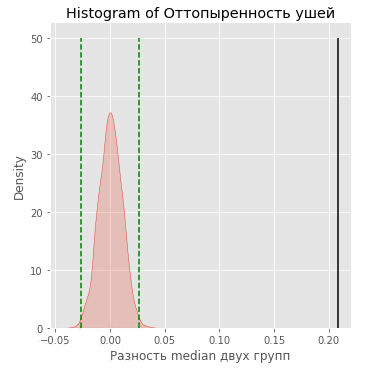


Рисунок 22 — Результаты применения бутстрэпа для признака оттопыренность ушей.

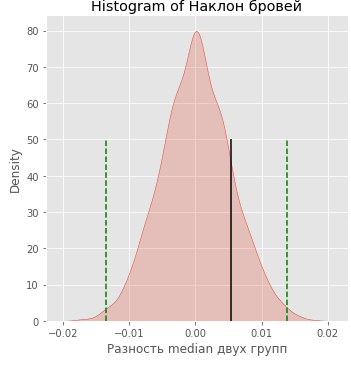


Рисунок 23 — Результаты применения бутстрэпа для признака наклон бровей.

В свою очередь мы так же можем рассчитать p-значение с помощью двустороннего статистического теста для наблюдаемых различий.

Таблица 3. Результаты проверки статистической значимости и размера эффекта.



ДОПОЛНИТЕЛЬНО

**Ящик с усами, диаграмма размаха** (англ. box-and-whiskers diagram or plot, box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.

Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

**График плотности** – это представление распределения числовой переменной. Он использует оценку плотности ядра, чтобы показать функцию плотности вероятности переменной. Это сглаженная версия гистограммы, которая используется в той же концепции.

**t-критерий Стьюдента** — типичный параметрический статистический тест, используемый для сравнения двух групп. Нулевой гипотезой является равенство средних двух генеральных совокупностей.

**U-критерий Манна**-**Уитни** — типичный непараметрический статистический тест, используемый для сравнения двух групп. Нулевой гипотезой является что две выборки взяты из одной генеральной совокупности. Фактически в статистике используется как замена t-критерий Стьюдента, когда его допущения не выполняются.

**Критерий Шапиро-Уилка** используется для проверки гипотезы H\_0: «случайная величина X распределена нормально» и является одним наиболее эффективных критериев проверки нормальности. Критерии, проверяющие нормальность выборки, являются частным случаем критериев согласия. Если выборка нормальна, можно далее применять мощные параметрические критерии, например, критерий Фишера или Стьюдента.

**Рангово-бисериальная корреляция** — это метод сообщения величины эффекта для U‑критерия Манна‑Уитни и основан на измерении ранговой корреляции. Как и другие корреляционные меры, ранг-бисериальная корреляция может варьироваться от минус единицы до плюс один, при этом значение нуля указывает на отсутствие связи.

Статья про **оттопыренность ушей** (травма ушной раковины) — <https://gol.ru/materials/1362-slomannye-usi-v-mma-eto-bol-status-kultura-ih-vosstanavlivaut-hlebnym-makisem-ili-delaut-specialno-kak>