Урок 3

Проверка гипотез

3.1. Проверка гипотез: начало

Проверка статистических гипотез — это важнейший инструмент, которым необходимо владеть в совершенстве, чтобы заниматься анализом данных. В этом уроке будут разобраны все компоненты, из которых состоит этот инструмент.

3.1.1. Предсказание будущего

Можно представить себе человека, который утверждает, что он может предсказывать будущее. Не так важно, как он это делает: он может использовать гадание на кофейной гуще или делать свои предсказания на основании исторической информации, применяя обучение с учителем с хорошо измеренными признаками. Чтобы проверить его утверждение о способности предсказывать будущее, нужно провести эксперимент: записать все предсказания, сгенерировать соответствующие им события или подождать, сбудутся они или нет, а затем проверить правильность предсказаний.

Этот эксперимент порождает выборку $X^n = (X_1, \ldots, X_n)$, которая может состоять, например, из 0 и 1: $X = 1$ соответствует сбывшемуся предсказанию, а $X = 0$ — несбывшемуся. Также она может состоять из точностей предсказания, то есть разностей между фактом и прогнозом.

Предсказатель полезен, если он предсказывает лучше, чем генератор случайных чисел. Можно рассмотреть гипотезу, что предсказатель — это и есть генератор случайных чисел. Для этого нужно посмотреть на данные и подумать, свидетельствуют ли они против такого предположения. Примерно так и используется проверка гипотез.

3.1.2. Проверка гипотез: формальное определение

Теперь введём все необходимые компоненты механизма проверки гипотез формально (таблица 3.1).

выборка:	$X^n = (X_1, \ldots, X_n),\ X \sim \mathbf{P}$;
нулевая гипотеза:	$H_0\colon\ \mathbf{P} \in \omega$;
альтернатива:	$H_1\colon\ \mathbf{P} \notin \omega$;
статистика:	$T(X^n),\ \ T(X^n) \sim F(x)$ при H_0;
	$T(X^n) \not\sim F(x)$ при H_1.

Таблица 3.1: Проверка гипотез

Итак, имеется некоторая выборка из случайной величины X, которая имеет неизвестное распределение $\mathbf{P}$. Кроме того, выдвинута нулевая гипотеза об этом распределении (например, "$\mathbf{P}$ принадлежит некоторому семейству распределений ω") и альтернативная гипотеза.

Требуется проверить, глядя на имеющиеся данные, какая из двух гипотез, нулевая или альтернативная, более вероятна. Для этого используется некоторая статистика T, которая обладает очень важным свойством: если нулевая гипотеза справедлива, то точно известно, какое у статистики распределение, а если справедлива

альтернатива, то распределение статистики — какое-то другое. Распределение $F(x)$ называется нулевым распределением статистики, а пара, состоящая из статистики и нулевого распределения, образует статистический критерий для проверки нулевой гипотезы против альтернативы.

3.1.3. Нулевое распределение

Итак, пусть выборка собрана и подсчитано значение статистики на этой выборке:

$$T(X) = t.$$

Осталось понять, какова вероятность получить именно такое значение статистики при условии справедливости нулевой гипотезы. Вообще говоря, если распределение нулевой статистики непрерывно, то каждому конкретному значению соответствует нулевая вероятность, поэтому такая постановка задачи некорректна. Чтобы её переформулировать, нужно понять, какие значения статистики соответствуют альтернативной гипотезе.

Пусть, например, при справедливости альтернативы более вероятны большие значения статистики. Теперь возникает вопрос, с какой вероятностью можно получить значение статистики $T(X) \geq t$ при справедливости нулевой гипотезы. Эта вероятность является ключевым компонентом механизма проверки гипотез и называется достигаемым уровнем значимости, или p-value.

Достигаемый уровень значимости — это вероятность получить такое же значение статистики, как в эксперименте, или еще более экстремальное, при справедливости нулевой гипотезы. То, какие значения считаются экстремальными, определяется относительно альтернативной гипотезы, то есть, с учетом того, какие значения статистики более вероятны при альтернативе.

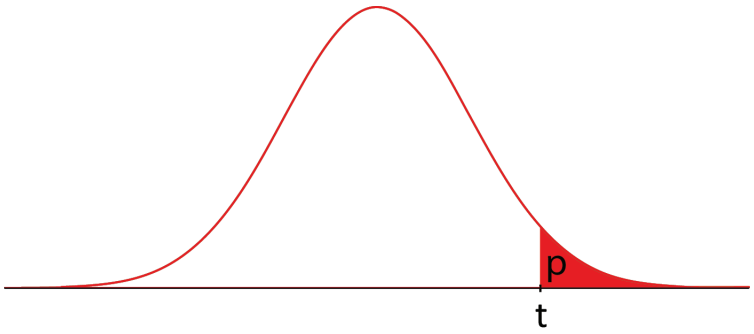

Рис. 3.1: Нулевое распределение

Зная нулевое распределение статистики и значение статистики, реализовавшееся в эксперименте, можно вычислить p-value. В случае, когда критическими, то есть, более вероятными при альтернативе, являются большие значения статистики, p-value — это интеграл от плотности нулевого распределения по правому хвосту начиная с t и до ∞ (рисунок 3.1).

Если полученное значение p-value мало, это значит, что данные свидетельствуют против нулевой гипотезы в пользу альтернативной, поскольку вероятность получить такие данные при условии, что нулевая гипотеза справедлива, мала. Обычно p-value сравнивают с порогом α, который называется уровнем значимости. Чаще всего $\alpha = 0.05$. Если $p \leq \alpha$, то нулевая гипотеза отвергается в пользу альтернативы. Если $p > \alpha$, то нулевая гипотеза не отвергается.

3.2. Ошибки I и II рода

Существенная особенность механизма проверки гипотез — его несимметричность относительно пары нулевая гипотеза – альтернатива. Эта особенность тесно связана с понятиями ошибок первого и второго рода.

Нулевая гипотеза может быть либо верна, либо неверна. В результате проверки гипотезы её можно либо принять, либо отвергнуть. Из этих соображений составлена таблица 3.2. На главной диагонали находятся верные решения: либо принимается верная нулевая гипотеза, либо отвергается неверная нулевая гипотеза. А вот на побочной диагонали располагаются ошибки. Совершить ошибку первого рода — значит отвергнуть верную нулевую гипотезу. Если же принимается неверная нулевая гипотеза, то это ошибка второго рода.

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка II рода
H_0 отвергается	Ошибка I рода	H_0 верно отвергнута

Таблица 3.2: Ошибки I и II рода

В механизме проверки гипотез ошибки первого и второго рода неравнозначны. Ошибка первого рода критичнее, вероятность отвержение нулевой гипотезы в случае, когда она верна, жестко ограничивается. Если нулевая гипотеза отвергается при значении уровня значимости $p \leq \alpha$, то вероятность ошибки первого рода получается ограниченной сверху:

$$\mathbf{P}(H_0 \text{ отвергнута} \mid H_0 \text{ верна}) = \mathbf{P}(p \leq \alpha \mid H_0) \leq \alpha.$$

Таким образом, любой корректный, хорошо построенный критерий имеет вероятность ошибки первого рода не больше, чем α.

Что касается ошибки второго рода, то она минимизируется по остаточному принципу. Понятие ошибки второго рода связано с понятием мощности статистического критерия. Мощность — это вероятность отвергнуть неверную нулевую гипотезу:

$$\text{pow} = \mathbf{P}(\text{отвергаем } H_0 \mid H_1) = 1 - \mathbf{P}(\text{принимаем } H_0 \mid H_1).$$

Чтобы найти идеальный критерий для проверки пары нулевая гипотеза – альтернатива, нужно среди всех корректных критериев выбрать критерий с максимальной мощностью.

Неравнозначность нулевой и альтернативной гипотезы видна уже на уровне терминологии. Если достигаемый уровень значимости $p \leq \alpha$, то говорят, что нулевая гипотеза отвергается в пользу альтернативы. Если достигаемый уровень значимости $p > \alpha$, то нулевая гипотеза не отвергается. Когда гипотеза не отвергается, это значит только то, что нет доказательств того что она неверна. Но отсутствие доказательств не является доказательством ее верности!

Это можно лучше понять на примере судебного процесса. Основное положение — презумпции невиновности: подсудимый по умолчанию невиновен (это нулевая гипотеза), и, если доказательств обратному нет, нельзя утверждать, что он преступник, даже если он на самом деле совершил преступление.

3.3. Достигаемый уровень значимости

Достигаемый уровень значимости — это достаточно сложная теоретическая концепция, которую часто понимают неправильно даже те, кто регулярно пользуется статистикой и проверкой гипотез.

Достигаемый уровень значимости — это вероятность при справедливости нулевой гипотезы получить такое же значение статистики, как в эксперименте, или ещё более экстремальное:

$$p = \mathbf{P}(T \geq t \mid H_0)$$

Чем ниже достигаемый уровень значимости, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Проблема определения p-value в том, что оно длинное, но из него ничего нельзя выбросить так, чтобы оно не стало неправильным. Например, часто хочется думать, что p-value — это просто вероятность справедливости нулевой гипотезы, или вероятность справедливости нулевой гипотезы при условии полученных данных, но это не так!

$$\begin{aligned} p = \mathbf{P}(T \geq t \mid H_0) &\neq \mathbf{P}(H_0) \\ &\neq \mathbf{P}(H_0 \mid T \geq t) \end{aligned}$$

Это хорошо понятно на следующем примере. В 2010 году осьминог Поль угадывал результаты матчей чемпионата мира по футболу с участием сборной Германии, выбирая из двух кормушек ту, на которой был

изображён флаг страны-победителя. Из 13 матчей, в которых он пробовал свои силы, результаты 11 ему удалось угадать. Используя эти данные как выборку, можно проверить нулевую гипотезу о том, что он выбирал кормушку наугад против альтернативы о том, у осьминога есть сверхспособности к предсказанию результатов матчей. Критерий, которым проверяется эта нулевая гипотеза, будет разобран позже. Но если его применить, получится достигаемый уровень значимости $p = 0.0112$. Это значение — не вероятность, что осьминог выбирает кормушку наугад. Вероятность того, что осьминог выбирает кормушку наугад, равна единице! $p = 0.0112$ — это именно вероятность получить такие или ещё более экстремальные данные при условии справедливости нулевой гипотезы. Эта вероятность достаточно мала, но редкие события тоже происходят. И, как правило, именно о них пишут в газетах.

3.4. Статистическая и практическая значимость

3.4.1. Размер эффекта

На самом деле эксперименты проводятся не для того, чтобы получить значение p-value. Как правило, исследователя интересует размер эффекта, то есть степень отклонения данных от нулевой гипотезы. Например, если эксперимент связан с проверкой способностей предсказателя будущего, то размер эффекта — это вероятность верного предсказания. Если проверяется эффективность лекарства, то размер эффекта — это вероятность выздоровления пациента, который это лекарство принимает, за вычетом эффекта плацебо. При запуске программы лояльности для пользователей интернет-магазина размер эффекта — это последующее увеличение среднего чека.

Размер эффекта — это величина, определенная на генеральной совокупности. Но, как правило, у исследователя есть только небольшая выборка из нее, а оценка размера эффекта по выборке — это случайная величина. Маленький достигаемый уровень значимости является показателем того, что такую оценку размера эффекта, какая получена по выборке, с маленькой вероятностью можно было получить случайно.

Достигаемый уровень значимости зависит не только от размера эффекта, но и от объема выборки, по которой оценивается эффект. Если выборка небольшая, скорее всего, нулевая гипотеза на ней не отвергается (если только она не слишком дикая). Однако с ростом объема выборки начинают проявляться все более тонкие отклонения данных от нулевой гипотезы. Велика вероятность, что на достаточно большой выборке значительная часть разумных нулевых гипотез будет отвергнута. Именно поэтому, даже если нулевая гипотеза отвергнута, это еще не значит, что полученный эффект имеет какую-то практическую значимость, её нужно оценивать отдельно. Чтобы лучше это понять, давайте рассмотрим несколько примеров.

3.4.2. Статистически значимо, практически незначимо

Первый пример связан с большим исследованием, в рамках которого на протяжении трех лет у большой выборки женщин измеряли вес, а также оценивали, насколько активно они занимаются спортом. По итогам исследования выяснилось, что женщины, которые в течение этого времени упражнялись не меньше часа в день, набрали значительно меньше веса, чем женщины, которые упражнялись менее 20 минут в день. Статистическая значимость этого результата достаточно высока: $p < 0.001$. Проблема в размере эффекта: разница в набранном весе между двумя исследуемыми группами женщин составила всего 150 граммов. 150 граммов за 3 года — это не очень много. Крайне сомнительно, что этот эффект имеет какую-то практическую значимость.

Еще один пример связан с клиническими испытаниями гормонального препарата «Премарин», который облегчает симптомы менопаузы. В 2002 году эти испытания были прерваны досрочно, поскольку было обнаружено, что прием препарата ведет к значимому увеличению риска развития рака груди (на 0.08%), инсульта (на 0.08%) и инфаркта (на 0.07%). Этот эффект статистически значим; при этом на первый взгляд кажется, что размеры эффектов ничтожны. Например, если кому-то сказать, что его любимые конфеты повышают риск возникновения инфаркта на 0.07%, вряд ли это заставит человека отказаться от этих конфет. Тем не менее, если пересчитать размеры эффектов на всю популяцию людей, которым этот препарат может быть потенциально приписан, результатом будут тысячи дополнительных смертей. Разработчики препарата не могут взять на себя эту ответственность, поэтому такой препарат немедленно запрещают и снимают с рынка.

Этот пример показывает, что практическую значимость результата нельзя определить на глаз. В идеале она должна определяться человеком, который поставил задачу и понимает предметную область.

3.4.3. Статистически незначимо, практически значимо

Еще один пример — это испытание лекарства, которое замедляет ослабление интеллекта у людей, страдающих болезнью Альцгеймера. В этом исследовании очень сложно измерить размер эффекта. В течение эксперимента одна часть испытуемых должна принимать лекарство, а другая — плацебо. Только по прошествии нескольких лет можно будет сравнить эти две группы. Поэтому такое исследование длится долгое время и дорого стоит. Если при испытании оказывается, что разница между снижением IQ в контрольной группе, где люди принимали плацебо, и тестовой группе, где люди принимали препарат, составляет 13 пунктов, это различие очень большое, и на практике этот эффект крайне значим. При этом может оказаться, что статистическая значимость не была достигнута, то есть $p > \alpha$, и формально нулевую гипотезу об отсутствии эффекта лекарства нельзя отвергнуть. Если предмет исследования очень важен, то, оказавшись в подобной ситуации, возможно, стоит продолжать исследования: набрать еще выборку, уменьшить дисперсию оценки размера эффекта и убедиться в том, что важное открытие не упущено.

3.5. Биномиальный критерий для доли

Джеймс Бонд утверждает, что он предпочитает пить мартини взболтанным, но не смешанным. Чтобы проверить это на практике, можно предложить Джеймсу Бонду пройти так называемый blind test, или слепое тестирование. Можно было бы завязать ему глаза, несколько раз предложить на выбор взболтанный и смешанный мартини, а после этого спросить, какой напиток он предпочитает. В данном случае если бы Джеймс Бонд выбирал взболтанный напиток, это считалось бы успехом, потому что его выбор соответствует его утверждению. В противном случае считалось бы, что произошла неудача, так как выбор утверждению не соответствует.

В данном случае необходимо проверить нулевую гипотезу H_0: Джеймс Бонд не различает два вида напитков и выбирает наугад, против некоторой альтернативы. Но альтернатива, вообще говоря, могла бы быть разной. С одной стороны, можно рассматривать двустороннюю альтернативу (Джеймс Бонд отличает два вида напитков, и у него есть некоторые предпочтения) или одну из односторонних (Джеймс Бонд предпочитает взболтанный мартини, так, как он утверждает, или Джеймс Бонд предпочитает смешанный). Такой эксперимент нужно провести n раз и в качестве T-статистики использовать количество единиц выборки или сумму элементов выборки. Если нулевая гипотеза справедлива, то есть Джеймс Бонд выбирает напиток наугад, то можно было бы равновероятно получить любую комбинацию из нулей и единиц. Таких комбинаций ровно 2^n, поэтому для того, чтобы получить нулевое распределение, можно было бы сгенерировать все эти наборы данных, на каждом посчитать значение статистики и таким образом получить распределение. На самом деле, в данном случае этот шаг можно пропустить, потому что исследуемая выборка состоит из нулей и единиц и взята из распределения Бернулли с вероятностью успеха p. В данном случае вероятность успеха $p = 0.5$, потому что если нулевая гипотеза справедлива, то успех и неудачи просходят равновероятно. Соответственно выборка представляет из себя сумму n независимых одинаково распределенных величин из распределения Бернулли. Значит, нулевое распределение статистики — это биномиальное распределение с параметрами n (количество экспериментов) и p (вероятность успеха).

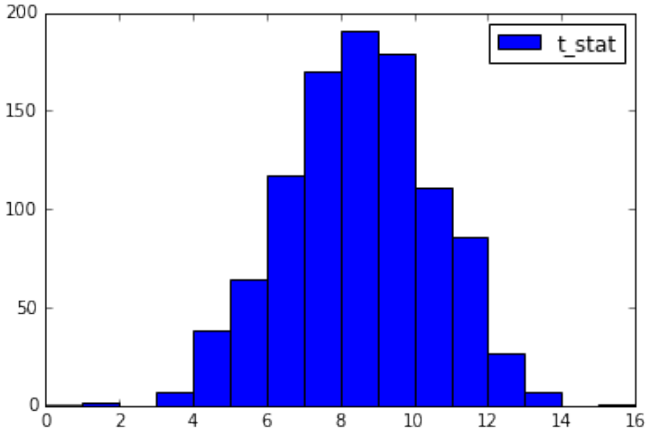

Рис. 3.2: Биномиальное распределение с параметрами $n = 16$ и $p = 0.5$

Нулевое распределение при параметрах $n = 16$ и $p = 0.5$ показано на рисунке 3.2. Оно выглядит так, как и ожидалось: пик находится в центре.

Итак, сначала можно протестировать нулевую гипотезу против односторонней альтернативы H_1: Джеймс Бонд предпочитает взболтанный мартини. При такой альтернативе более вероятно попасть в правый конец распределения, то есть получить много единиц в выборке. Пусть было проведено 16 испытаний, и при этом в 12 из них Джеймс Бонд выбрал взболтанный мартини, то есть произошел успех. Если построить соответствующее нулевое распределение, то в данном случае T-статистика была бы равна 12 и интерес представлял бы правый «хвост» распределения. В данном случае требуется просуммировать высоту столбцов, начиная со столбца, соответствующего 12, и правее, то есть правый «хвост» распределения, полученное значение — это достигаемый уровень значимости. В данном случае получается p-value 0.038, это говорит о том, что на уровне значимости 0.05 нулевая гипотеза отвергается. То есть если успех происходит 12 раз из 16, то можно сделать вывод, что Джеймс Бонд предпочитает взболтанный мартини. Если бы успехов было немного меньше, например, 11, то значение p-value стало бы больше: $p = 0.105$, то есть на уровне значимости 0.05 уже нельзя отвергнуть нулевую гипотезу.

В случае двусторонней альтернативы гипотеза H_1 переформулируется следующим образом: Джеймс Бонд предпочитает какой-то один определенный вид мартини, не требуется выбирать, какой именно. При такой альтернативе будут очень вероятны либо большие значения статистики, либо очень маленькие. При расчете достигаемого уровня значимости будут учитываться как правый, так и левый концы распределения. Если предположить, что произошло 12 успехов, то есть 12 раз Джеймс Бонд выбрал взболтанный мартини, то необходимо просуммировать тот же самый правый конец, но теперь к нему добавляется и левый. Значение достигаемого уровня значимости $p = 0.077$, это больше, чем при проверке нулевой гипотезы против односторонней альтернативы. Соответственно, в данном случае нельзя отвергнуть гипотезу на уровне значимости 0.05, однако можно отвергнуть нулевую гипотезу на уровне значимости 0.1. Можно посмотреть, достаточно ли 13 успехов, чтобы отвергнуть нулевую гипотезу на уровне 0.05. В данном случае p-value $p = 0.021$, отвергнуть нулевую гипотезу на уровне значимости 0.05 можно.

3.5.1. Критерии согласия Пирсона (хи-квадрат)

Критерий согласия Пирсона (или критерий хи-квадрат) используется для проверки того, что некоторая наблюдаемая случайная величина подчиняется тому или иному теоретическому закону распределения.

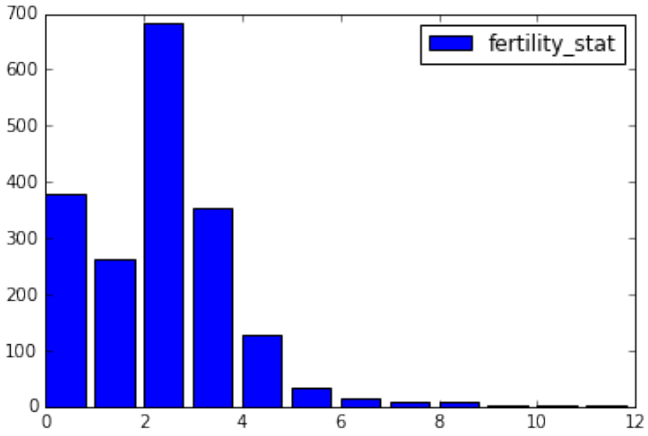

Рис. 3.3: Данные о количестве детей у женщин старше 45 лет

В качестве примера используются данные об исчерпанной рождаемости (рисунок 3.3). Этот признак связан с количеством детей, родившихся у женщины на момент окончания репродуктивного возраста (приблизительно 45 лет). Для 1878 женщин старше 45, участвующих в социологическом опросе жителей Швейцарии, известно количество детей. Этот признак – типичный счётчик, поэтому его можно попробовать оценить с помощью распределения Пуассона. В данном случае выборка — это целочисленный вектор длины n (в данном случае $n = 1878$), где каждая компонента вектора — это количество детей, рожденных у женщины. В данном случае гипотеза H_0: наблюдаемая величина имеет распределение Пуассона.

Из распределения данных (рисунок 3.3) видно, что количество детей меняется от 0 до 11. Чаще всего у женщины не более четырёх детей. Наиболее часто встречающееся количество детей — это два ребёнка.

Кажется, что такие данные должны хорошо описываться распределением Пуассона. В предыдущих курсах было показано, что лучшая оценка на параметр λ для распределения Пуассона, — это просто выборочное среднее. Если его вычислить, получается $\lambda = 1.937$.

Необходимо проверить следующую гипотезу H_0: наблюдаемая случайная величина имеет распределение Пуассона с параметром $\lambda = 2$. Это можно делать с помощью критерия согласия Пирсона. Для этого нужно подготовить данные. Первое, что представляет интерес, — это наблюдаемые частоты. Известно, сколько раз встретилось каждое количество детей, так что интересующую величину несложно получить. Тогда элемент результирующего вектора 0 говорит о том, сколько раз в нашей выборке встретилось количество детей, равное 0 (в данном случае это 379), и последний 11 элемент означает, что 11 детей встретилось всего лишь 1 раз. Это наблюдаемые частоты.

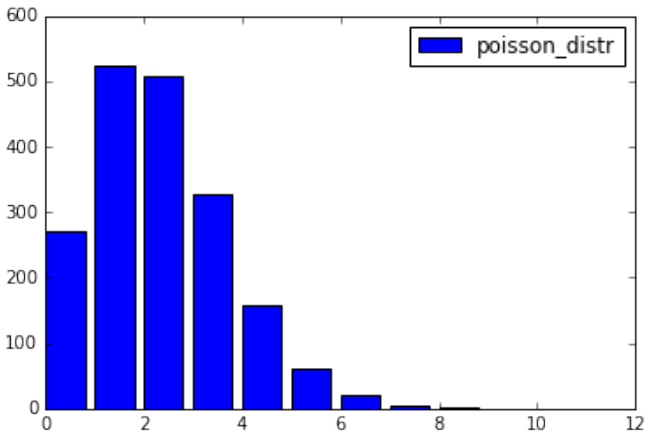

Рис. 3.4: Ожидаемые частоты для распределения Пуассона с параметром $\lambda = 2$

Теперь нужно построить так называемые ожидаемые частоты. Это те частоты, которые бы наблюдались, если бы данные имели распределение Пуассона с параметром $\lambda = 2$, и размер выборки был бы таким же. Результирующие ожидаемые частоты показаны на рисунке 3.4 Видно, что наблюдаемые частоты отличаются от ожидаемых. Строго оценить это различие можно с помощью критерия хи-квадрат. Статистика этого критерия:

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - np_i)^2}{np_i}.$$

При справедливости нулевой гипотезы статистика имеет распределение хи-квадрат с числом степеней свободы $K - 1 - m$, где m - число параметров распределения, оцененных по выборке.

Достигаемый уровень значимости, полученный с помощью критерия хи-квадрат: $p = 1.77 \times 10^(-86)$. Это значение очень близко к нулю, значит, можно смело отвергнуть нулевую гипотезу о том, что данные имеют распределение Пуассона с параметром $\lambda = 2$.

3.6. Связь между проверкой гипотез и доверительными интервалами

3.6.1. Проверка гипотез при помощи построения доверительных интервалов

Пусть требуется оценить качество предсказаний бинарного классификатора на тестовой выборке из 100 объектов. Этот классификатор верно предсказывает метку класса на 60 из 100 объектов. С одной стороны кажется, что 60 из 100 — это не очень много. С другой стороны, может быть, эта задача достаточно сложная, и предсказать лучше нельзя. Чтобы определить качество работы классификатора, его нужно сравнить с самым бесполезным классификатором — генератором случайных чисел. Если классы в задаче сбалансированы, то генератор случайных чисел в среднем будет угадывать метку у 50 объектов из 100, и вероятность угадать составит 0.5. Можно ли считать, что классификатор, который угадывает классы 60 из 100 объектов, лучше, чем генератор случайных чисел?

Чтобы ответить на этот вопрос, можно построить доверительный интервал для доли верно предсказанных меток. Результат при уровне доверия 0.95: $[0.504, 0.696]$. Соответствующая генератору случайных чисел вероятность 0.5 не содержится в этом интервале. Из этого следует, что исследуемый классификатор значимо лучше, чем генератор случайных чисел.

В общем случае, если проверяется точечная нулевая гипотеза против двусторонней альтернативы:

$$H_0 \colon \theta = \theta_0, \quad H_1 \colon \theta \neq \theta_0,$$

то производить эту проверку можно путём построения доверительного интервала, как было показано выше. Нулевая гипотеза отвергается на уровне значимости α, если доверительный интервал для θ с уровнем доверия $1-\alpha$ не содержит θ_0.

Также с помощью доверительного интервала можно вычислить достигаемый уровень значимости p. Это наибольшее значение α, при котором доверительный интервал с уровнем доверия $1-\alpha$ содержит θ_0. Таким образом, перебирая разные значения α, можно численно найти достигаемый уровень значимости.

Многие статистические критерии эквивалентны построению доверительных интервалов. Например, нормальные доверительные интервалы для доли эквивалентны z-критерию для той же самой доли (этот критерий будет описан позднее). Для таких пар нет необходимости в численном подсчёте достигаемого уровня значимости, это можно сделать аналитически.

Однако, например, для метода построения доверительных интервалов Уилсона нельзя явно записать выражение для статистики соответствующего критерия. Для подобных методов численный поиск достигаемых уровней значимости оказывается очень полезным. Например, в задаче с бинарным классификатором 95% доверительный интервал Уилсона для доли верно предсказанных меток: $[0.502, 0.691]$. Полученный интервал похож на нормальный доверительный интервал, однако в других случаях, особенно когда значение p близко к 0 или 1, доверительный интервал Уилсона может существенно отличаться (как и достигаемые уровни значимости), и лучше пользоваться именно им, поскольку он точнее. В данном случае достигаемый уровень значимости $p = 0.045$. То есть гипотеза о том, что рассматриваемый классификатор не лучше, чем генератор случайных чисел, может быть отвергнута на уровне значимости 0.05.

3.6.2. Построение доверительных интервалов с помощью критериев проверки гипотез

Выше описан метод проверки гипотез с использованием доверительных интервалов. Однако можно делать и наоборот: строить доверительные интервалы с помощью критерия, проверяющего гипотезу.

Пусть снова заданы точечная гипотеза относительно параметра θ и двусторонняя альтернатива:

$$H_0 \colon \theta = \theta_0, \quad H_1 \colon \theta \neq \theta_0.$$

В таком случае доверительный интервал с уровнем доверия $1 - \alpha$ будет состоять из всех значений θ_0, для которых такая нулевая гипотеза не отвергается на уровне значимости α против двусторонней альтернативы.

Этот метод построения доверительных интервалов не слишком конструктивный, но иногда его можно применять, если под рукой нет никакого метода получше.

3.6.3. Проверка гипотез и построение доверительных интервалов при сравнении двух классификаторов

Пусть теперь помимо описанного ранее бинарного классификатора имеется второй классификатор, который на той же самой тестовой выборке верно предсказывает метки для 75 объектов из 100. Требуется определить, какой из двух классификаторов лучше. С одной стороны, 75 больше, чем 60. Но с другой стороны, выборка из 100 объектов не очень большая, и такая разница может возникнуть и случайно.

Учесть влияние случайности можно с помощью построения доверительных интервалов. Для первого классификатора доверительный интервал Уилсона для доли верных предсказаний: $[0.502, 0.691]$. Для второго классификатора такой же доверительный интервал: $[0.657, 0.825]$. Эти доверительные интервалы пересекаются по отрезку $[0.657, 0.691]$. Но пересечение доверительных интервалов не означает, что классификаторы нельзя различить по качеству. В данном случае выдвинута точечная нулевая гипотеза относительно двух параметров, θ_1 и θ_2, и необходимо проверить её против двусторонней альтернативы:

$$H_0 \colon \theta_1 = \theta_2, \quad H_1 \colon \theta_1 \neq \theta_2.$$

Правильным решением будет построить доверительный интервал для разности параметров θ_1 и θ_2, именно она полностью соответствует выдвинутой нулевой гипотезе (если $\theta_1 = \theta_2$, значит, их разность равна нулю). 95% доверительный интервал для разности долей в данной задаче: $[0.022, 0.278]$. Этот доверительный интервал не содержит ноль, значит, можно утверждать, что второй классификатор значимо лучше.

Если инвертировать этот доверительный интервал описанным ранее способом и выбрать наибольшее значение α, при котором ноль попадает в доверительный интервал, получится уровень значимости $p = 0.022$. То есть, на уровне значимости 0.05 отвергается нулевая гипотеза о том, что два классификатора по качеству одинаковы.

I \ II	+	-	$\sum$
+	55	5	60
-	20	20	40
$\sum$	75	25	100

Таблица 3.3: Таблица сопряжённости

Ранее не было учтено, что качество классификаторов определяется на одной и той же обучающей выборке, а значит, выборки в этой задаче — связанные. В такой ситуации доверительный интервал правильнее строить другим методом. Для этого используется таблица сопряжённости 3.3, и учитывается не количество ошибок каждого классификатора отдельно, а количество объектов, на которых классификаторы дали разные ответы (20 и 5 в этой таблице). Полученный 95% доверительный интервал для разности долей в связанных выборках равен $[0.06, 0.243]$. Обратите внимание, что этот доверительный интервал уже, и его левая граница дальше отстоит от нуля, то есть, при учёте связанности увеличивается уверенность в том, что классификаторы отличаются. Для данного интервала достигаемый уровень значимости $p = 0.002$. Он почти в десять раз меньше, чем при построении доверительного интервала без учёта связанности выборок.