МИНОБРНАУКИ РОССИИ

Санкт-Петербургский государственный электротехнический
университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

С. В. МАЛОВ  И. Ю. МАЛОВА

# БАЗОВЫЕ МОДЕЛИ БИОСТАТИСТИКИ: АНАЛИЗ КАТЕГОРИАЛЬНЫХ ДАННЫХ

Учебное пособие

Санкт-Петербург
Издательство СПбГЭТУ «ЛЭТИ»
2021

УДК 519.2
ББК В 172я7
М хх

**С. В. Малов, И. Ю. Малова.**

М хх Базовые модели биостатистики: анализ категориальных данных: учеб. пособие. Под ред. с.

ISBN ххх-х-хххх-хх

Рассмотрено использование методов математической статистики для анализа категориальных данных биомедицинских исследований.

Предназначено для поддержки дисциплины «Биостатистика» на ФКТИ СПбГЭТУ «ЛЭТИ», будет интересно студентам и аспирантам, обучающимся по программам подготовки специалистов в области информационных технологий.

УДК 519.2
ББК В 172я7

Рецензенты: кафедра высшей математики ВШТЭ СПбГУПТД; д-р. техн. наук, Уткин Л.В. (СПбПУ им. Петра Великого)

Утверждено
редакционно-издательским советом университета
в качестве учебное пособие

ISBN ххх-х-хххх-хх

# ВВЕДЕНИЕ

В настоящее время проводится огромное количество исследовательских проектов в различных областях биологии и медицины. Изучение реальных процессов, происходящих в живых организмах, весьма затруднительно, ввиду их сложности и многогранности, поэтому одним из основных инструментов получения новых знаний представляется статистический анализ. Статистические выводы не объясняют реальных процессов, происходящих в живых организмах, но позволяют сделать определенные выводы о возможных результатах течения этих процессов и рисках нежелательных последствий, а также определить основные направления дальнейших углубленных исследований. Статистический анализ не дает возможности предсказать достоверно результат того или иного процесса. В большинстве случаев результат удается предсказать с определенной степенью точности и некоторой долей достоверности.

Проведение большого числа статистических исследований в различных областях биологии и медицины открыло серьезную проблему интерпретации результатов статистических исследований. Ввиду отсутствия абсолютной достоверности, статистические выводы иногда оказываются неверными. С точки зрения развития научного направления, более важной характеристикой результата статистического анализа является существенность статистического вывода, но далеко не все планы научных исследований позволяют исследовать эту характеристику. Например, при проведении статистического исследования об эффективности вакцинирования для предотвращения заболевания вероятность заболевания при эпидемии в случае вакцинирования может сократиться с 30% до 29%, или с 30% до 1%. В первом случае наблюдается незначительное изменение заражаемости и с медицинской точки зрения вакцину нельзя признать удачной, тогда как во втором случае заражаемость существенно снижается, что говорит об эффективности вакцины. Статистическая значимость, хоть и связана с существенностью эффекта, не дает прямого ответа на вопрос о его существенности и форме.

В основе планов большинства биостатистических экспериментов лежит нулевая гипотеза, отклонение от которой информативно с научной точки зрения. При интерпретации результатов статистического исследования следует помнить, что математическая модель эксперимента призвана лишь аппроксимировать реальное положение дел. Отклонения от нулевой гипотезы помимо интересующего эффекта могут быть обусловлены различными факторами (например, неоднородность данных), большинство из

которых невозможно эффективно контролировать, поэтому при интерпретации результатов с большим числом наблюдений значимость эффекта является вторичной, а первичными становятся его существенность и форма.

В определенный период развития биостатистических исследований решения о принятии результатов анализа основывались лишь на их статистической достоверности, а исследование существенности отклонения от нулевой гипотезы чаще всего не проводились. Успешными считались те статистические выводы, согласованность которых с нулевой гипотезой не превышала 5%. Такой подход привел к неконтролируемому росту невоспроизводимых статистических выводов. Американская статистическая ассоциация 7 марта 2016 года выпустила рекомендации о принципах корректной интерпретации статистической значимости и проведения углубленного статистического анализа, но в полной мере проблему появления невоспроизводимых выводов предложенные подходы решить не в состоянии. Дело в том, что при подготовке публикации исследователь обычно упоминает лишь успешные результаты, информативные и вносящие вклад в развитие научного мировоззрения, забывая про многие обыденные и неинформативные с научной точки зрения результаты. К последим относится согласование полученных данных и нулевой гипотезы. В рамках одного статистического исследования данную проблему можно решить путем ужесточения контроля существенности и статистической значимости успешных результатов, однако решить эту проблему для всех статистических исследований не представляется возможным.

Несмотря на выявленные недостатки, роль статистического анализа в научных исследованиях не уменьшается. Во многих областях статистический анализ остается единственным доступным инструментом научного исследования. В случае правильного использования арсенала статистических методов и корректной интерпретации результатов, статистический анализ становится эффективным инструментом пополнения объема знаний в различных областях научных исследований.

В настоящем учебном пособии упор делается на категориальный план эксперимента, в основе которого лежит простой метод накопления статистической информации, построение и анализ таблиц сопряженности признаков. Такой подход наиболее распространен при проведении биомедицинских исследований ввиду его простоты и универсальности. По сравнению с более сложными моделями, категориальная модель накопления статистической информации предъявляет минимальные требования в процессе постановки эксперимента и интерпретации результатов анализа. В

учебном пособии обсуждаются основные подходы к постановке статистического эксперимента и основы статистического анализа данных, формулируются задачи категориального анализа данных и детально обсуждаются классические и современные методы их анализа, а также рассматриваются некоторые компьютерные реализации методов категориального анализа в пакете **R**. Предполагается, что читатель знаком с основными понятиями теории вероятностей и математической статистики, линейной алгебры, математического анализа, теории меры и интеграла.

## 1. Основы статистического анализа

Статистическое исследование включает в себя сбор данных, их организацию, анализ и интерпретацию полученных результатов. Планирование статистического исследования - важнейший этап статистического анализа. На этапе планирования происходит осмысление целей предполагаемого исследования, разработка стратегии сбора статистических данных, выбор математической модели статистического эксперимента и формулировка задач статистического исследования.

Статистический эксперимент может быть активным или пассивным. При пассивном эксперименте исследователь не имеет возможности влиять на условия проведения статистического эксперимента и его роль ограничивается наблюдением за изучаемым явлением. Активный эксперимент позволяет исследователю самостоятельно формировать условия проведения эксперимента, чтобы оптимизировать возможности последующего анализа для достижения поставленных целей. При интерпретации результатов статистического анализа следует учитывать постановку статистического эксперимента, соответствие выбранной математической модели реальным условиям статистического эксперимента, возможности использованных методов статистического анализа, а также все многообразие полученных результатов, если при анализе данных использовались различные методы.

Модель статистического эксперимента включает в себя три объекта $(\mathfrak{X},\mathfrak{F},\mathcal{P})$, где $\mathfrak{X}$ – множество допустимых результатов эксперимента, $\mathfrak{F}$ – множество наблюдаемых событий или статистическая информация, получаемая в результате эксперимента, и $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ – семейство допустимых вероятностных распределений или параметризация. Следует отметить, что $(\mathfrak{X}, \mathfrak{F}, P_\theta)$ при каждом фиксированном значении $\theta$ представляет собой вероятностный эксперимент. В основе методов математической статистики лежат вероятностные принципы, но в отличие от теории вероятностей, вероятности событий зависят от параметра $\theta$.

*Накопление статистической информации.* Одним из основных принципов статистического исследования является возможность накопления информации, позволяющая делать более достоверные выводы. Наиболее часто в основе статистического эксперимента лежит выборочный принцип, основанный на повторении простого эксперимента в одинаковых условиях. Альтернативно, выборка может быть получена с использованием случайного выбора $n$ элементов из генеральной совокупности $N$ элементов, если $N >> n$. В математической статистике выборкой $X = (X_1, \ldots, X_n)$ принято называть набор независимых и одинаково распределенных случайных величин (НОРСВ) или векторов. Следует отметить, что случайный выбор из генеральной совокупности без возвращения не дает независимых случайных величин, однако если $N >> n$, то отклонение от НОРСВ несущественно. Если сама генеральная совокупность представляет собой набор НОРСВ, то случайный выбор $n$ элементов из $N$ представляет собой набор НОРСВ. Распределение выборки однозначно определяется распределением каждой ее компоненты, что позволяет параметризовать выборку распределением одной ее компоненты.

Выборочный принцип накопления статистической информации не всегда удовлетворяет потребностям статистического исследования. Например, если исследователь хочет контролировать условия проведения эксперимента, то речь по сути идет об условном распределении наблюдаемой величины при фиксированных условиях проведения эксперимента. Наблюдения могут оставаться независимыми, однако, если не делать никаких предположений о связи распределений наблюдений при различных условиях, то серьезные статистические выводы вообще говоря невозможны. Простейшим обобщением выборочной модели является модель нескольких выборок. Если существует лишь конечное число возможных условий проведения эксперимента, то в случае независимых наблюдений получаем несколько выборок по числу условий. Помимо задач, связанных с распределением каждой из выборок, появляются задачи сравнения распределений различных выборок.

При наличии многообразия условий проведения эксперимента, для обеспечения возможности накопления статистической информации требуется установить связь между распределениями в различных условиях. Наиболее распространенными моделями такого типа являются регрессионные модели. Регрессией величины $Y$ по величине $X$ называется $\mathbb{E}(Y|X) = f(X)$ – среднее значение величины $Y$ в условиях $X$. В основе регрессионной модели лежит соотношение $\mathbb{E}(Y|X) = f(X;\theta)$, $\theta \in \Theta_1$.

Следует отметить, что параметр $\theta$ не определяет распределения набора наблюдений, поэтому в регрессионной модели также может присутствовать дополнительный (мешающий) параметр.

Точные методы в статистике достаточно редки и не универсальны (требуют достаточно жестких предположений), в связи с чем важнейшую роль играют асимптотические методы. Вместо одного статистического эксперимента рассматривается последовательность статистических экспериментов $(\mathfrak{X}_n, \mathfrak{F}_n, \mathcal{P}_n)$, $\mathcal{P}_n = \{P_{\theta,n}, \theta \in \Theta\}$. Отметим, что параметрическое множество не зависит от $n$, часто это позволяет делать все более точные выводы о параметре в процессе накопления статистической информации при увеличении $n$. В качестве примера можно привести выборку $X_1, \ldots, X_n$ из распределения $P_\theta$, $\theta \in \Theta$. С ростом $n$ происходит накопление статистической информации.

*Параметризация.* Важной составляющей модели статистического эксперимента является параметризация – семейство допустимых распределений наблюдений и его формализация с помощью параметра. Параметр должен быть идентифицируемым, т.е. каждому распределению множества допустимых распределений должно соответствовать единственное значение параметра – свойство инъективности. Близким значениям параметра (если параметрическое множество плотное) должны соответствовать близкие распределения – свойство непрерывности.

Методы анализа статистических данных существенно зависят от типа параметризации. Статистические эксперименты по типу параметризации классифицируют в три группы: параметрические, семипараметрические и непараметрические. В параметрической модели параметр представляет собой вещественное число $\theta \in \Theta \subseteq \mathbb{R}$ или вектор $\theta = (\theta_1, \ldots, \theta_d) \in \Theta \subseteq \mathbb{R}^d$. Говорят, что модель семипараметрическая, если параметр состоит из вещественного числа или вектора и дополнительной (обычно функциональной) компоненты $\theta = (\theta_1, \theta_2)$ и $\theta_1 \in \mathbb{R}^d$. К непараметрическим относят все остальные статистические эксперименты.

*Подчиненные и достаточные статистики.* Для решения задач статистического анализа используются статистики – функции от наблюдений $T : \mathfrak{X} \to E$. По сути результаты статистического анализа строятся на базе значений набора статистик. В условиях статистического эксперимента $(\mathfrak{X},\mathfrak{F},\mathcal{P})$ среди статистик можно выделить два типа: подчиненные и достаточные. Подчиненной называется статистика, распределение которой не

зависит от $\theta$. Подчиненная статистика не несет никакой информации о параметре и не может быть использована для статистического анализа. Если условное распределение исходного набора наблюдений при условии $T$ не зависит от параметра модели $\theta$, то такая статистика называется достаточной. Достаточная статистика содержит ту же информацию о параметре распределения, что и исходный набор данных. Имеет смысл проводить анализ с использованием достаточных статистик. Весь исходный набор наблюдений всегда является достаточной статистикой. Достаточная статистика, обеспечивающая максимальную редукцию данных называется минимальной.

*Постановка задач математической статистики.* Цель статистического анализа данных в условиях статистического эксперимента $(\mathfrak{X}, \mathfrak{F}, \mathcal{P})$, $\{P_\theta : \theta \in \Theta\}$ — сделать те или иные выводы об исинном значении параметра $\theta$. Различают три типа задач статистического анализа: точечное оценивание параметра, доверительное оценивание параметра и проверку статистических гипотез.

Задача точечного оценивания состоит в том, чтобы по результатам наблюдений найти приближенное значение параметра. Точечная оценка – статистика $\delta : \mathfrak{X} \to \Theta$. Обычно стараются использовать оценки оптимальные в том или ином смысле.

Задача доверительного оценивания состоит в том, чтобы по результатам наблюдений выбрать область значений параметра, накрывающую теоретическое значение параметра с достаточно большой вероятностью. Доверительная оценка – статистика $\hat{\Theta} : \mathfrak{X} \to \mathcal{Y}$, где $\mathcal{Y}$ – совокупность подмножеств множества $\Theta$ определенной формы. Доверительная оценка $\hat{\Theta}$ имеет уровень доверия $1 - \alpha$, если

$$P_\theta(\theta \in \hat{\Theta}) \geq 1 - \alpha \quad \text{при любом} \quad \theta \in \Theta.$$

Если параметр распределения — вещественное число ($\Theta \subseteq \mathbb{R}$), то для доверительного оценивания используются интервалы $\hat{\Theta} = [T_1(X), T_2(x)]$. Различают односторонние ($T_1 \equiv \inf\{\Theta\}$ – правосторонний; $T_2 \equiv \sup\{\Theta\}$ – левосторонний) и двухсторонние доверительные интервалы. Чем короче длина двухстороннего доверительного интервала при фиксированном уровне доверия, тем точнее оценка. Надежность доверительной оценки определяется уровнем доверия. Точность доверительной оценки при фиксированном уровне доверия зависит от эффективности использования имеющейся статистической информации.

Наиболее распространенные доверительные множества для многомерного параметра – параллелепипеды или эллипсоиды. В случае многомерного параметра можно говорить о доверительных интервалах для отдельных параметров или функций параметров. Следует учитывать, что доверительные интервалы, построенные для нескольких параметров, можно интерпретировать для любого из параметров, но не для всех параметров в совокупности. Иногда удается построить совместные доверительные интервалы для совокупности параметров или функций параметров. Метод множественного оценивания Шеффе устанавливает связь между доверительными эллипсоидами и совместными доверительными интервалами в случае совместного нормального распределения оценок параметров или функций параметров.

Статистической гипотезой называется утверждение о теоретическом распределении, выражаемое в терминах параметра $\theta \in \Theta_H$, где $\Theta_H \subseteq \Theta$. В классической постановке выдвигают основную (или нулевую) гипотезу $H_0 : \theta \in \Theta_0$ и альтернативную гипотезы (альтернативу) $H_A : \theta \in \Theta_A$: $\Theta \cap \Theta_A = \emptyset$. Задача исследователя – по результатам наблюдений сделать выбор между основной гипотезой и альтернативой. Различные варианты истинного положения дел и решения исследователя приведены в следующей таблице.

| | Принята $H_0$ | Отвергнута $H_0$ |
|---|---|---|
| $H_0$ верна | + | Ошибка I рода |
| $H_0$ не верна | Ошибка II рода | + |

Исследователь старается минимизировать ошибки. Подход Неймана–Пирсона заключается в том, чтобы ограничить ошибку $I$ рода малым наперед заданным числом $\alpha$. Критерий – правило, согласно которому принимается или отвергается основная гипотеза. Формально, критерий – статистика $\phi : \mathfrak{X} \to [0, 1]$, определяющая вероятность отвергнуть основную гипотезу по результатам наблюдений (0 – принимаем основную гипотезу, 1 – отвергаем). Критерий называется нерандомизованным, если результаты наблюдений однозначно определяют решение $\phi(\mathfrak{X}) = \{0, 1\}$. Значение $\alpha$, ограничивающее вероятность ошибки I рода называется уровнем значимости критерия. Мощность критерия равна $1 - P_\theta$(ош. II рода). Важную роль играет среднее значение $\mathbb{E}_\theta \phi(X)$

$$b(\theta) = \mathbb{E}_\theta \phi(X) = \begin{cases} \text{вероятность ош. I рода, при } \theta \in \Theta_0 \\ \text{мощность критерия, при } \theta \in \Theta_A \end{cases}$$

Обычно в основе статистического критерия лежит статистика критерия. Статистика критерия удовлетворяет следующим условиям: (i) распределение статистики критерия не зависит от параметра при справедливости нулевой гипотезы $\theta \in \Theta_0$, (ii) это распределение изучено (существуют таблицы), (iii) распределение статистики критерия при альтернативе $\theta \in \Theta_A$ отличается от ее распределения при основной гипотезе. Пусть $T$ – статистика критерия; $P_0$ – ее распределение при основной гипотезе. Для построения статистического критерия потребуется дополнительно набор множеств $\mathcal{I}$ (доверительных), удовлетворяющих условию: для любого $\alpha$ существует $I_\alpha \in \mathcal{I}$: $P_0(T \in I_\alpha) \geq (=) 1 - \alpha$. Исходя из распределения статистики критерия, строим нерандомизованный критерий

$$\phi(X) = \begin{cases} 0, \ T \in I_\alpha \\ 1, \ T \notin I_\alpha \end{cases}$$

Рассмотренный метод построения статистического критерия позволяет определить $P$-значение — наименьшее $\alpha$, такое что $T \notin I_\alpha$. При положительном распределении $T$ наиболее часто используется $I_\alpha = [0, x_\alpha]$, где $x_\alpha$ удовлетворяет условию $P_{\theta_0}(T > x_\alpha) = 1 - F_T(x_\alpha) = \alpha$, $F_0$ функция распределения $T$ при нулевой гипотезе. В этом случае $P$-значение равно $PV = 1 - F_T(T)$ и имеет равномерное $U(0,1)$ распределение при основной гипотезе, если $F_0$ непрерывная функция.[1] Аналогичное свойство выполнено и для других тестов, построенных с использованием статистик, имеющих при основной гипотезе непрерывное распределение. Наиболее распространенным методом в медицинских исследованиях является проверка значимости отклонений от основной гипотезы. Существенным результатом является отвержение основной гипотезы – выявление значимых отклонений. Реальные значения отклонений от основной гипотезы не изучаются. Наличие богатой статистической информации позволяет выявлять даже несущественные с практической точки зрения различия. Если проверяется несколько статистических гипотез, то необходима поправка. Гипотеза согласия ставится в терминах распределения отдельных наблюдений (например, элементов выборки) о согласии с некоторым фиксированным распределением (простая) или принадлежности некоторому множеству распределений (сложная). В зависимости от имеющихся статистических данных и выбранной модели эксперимента существует четыре основных типа

[1] Используется преобразование Смирнова

статистических гипотез. Гипотеза согласия ставится в терминах распределения отдельных наблюдений (например, элементов выборки) о согласии с некоторым фиксированным распределением (простая) или принадлежности некоторому множеству распределений (сложная). При наличии двух или нескольких выборок или неодинаково распределенных результатов наблюдений, распределение которых контролируется значением сопутствующего фактора, гипотеза однородности состоит в том, что распределения в различных выборках или при различных значениях сопутствующего фактора совпадают. При наличии пары или нескольких измеряемых величин часто выдвигают гипотезу их независимости. Наконец, гипотеза случайности набора наблюдений $X_1, \ldots, X_n$ состоит в том, что он является выборкой.

*Различные типы статистических данных.* Цели статистического исследования и статистические данные должны соответствовать друг другу. Цель статистического исследования (параметр) обычно ставится в соответствие распределению некоторой случайной величины (или вектора), которую будем называть изучаемой переменной. Наблюдаемыми переменными называются случайные величины, реализации которых представляют собой статистические данные. Наиболее часто изучаемые и наблюдаемые переменные совпадают. Наблюдаемые и изучаемые переменные можно классифицировать по их типу. Количественные переменные характеризуют числовое значение изучаемой характеристики, порядковые – определяют порядок, но не величину наблюдаемой характеристики, а категориальные переменные по сути задают группировку. Числовые значения категориальных переменных не важны. Количественные переменные могут быть как непрерывными, так и дискретными. Распределение самой случайной величины и ее числовые характеристики могут представлять интерес для статистического анализа. Порядковые переменные обычно дискретные целочисленные. Интерес представляют лишь атомы соответствующего дискретного распределения с учетом порядка. Категориальные переменные могут быть как дискретными, так и вовсе нечисловыми. Интерес представляют лишь атомы соответствующего дискретного распределения.

Можно выделить несколько типов статистических данных. Простые наблюдения – случайные величины, объединяемые в векторы. В зависимости от характера наблюдаемой переменной, они могут быть количественными, порядковыми или категориальными. Временные ряды представляют собой наборы переменных, характеризующих течение изучаемого процесса во времени, который имеет смысл интерпретировать как случайный,

т.е. значения случайного процесса в выбранных точках. Обычно временной ряд представляет собой набор последовательных наблюдений одного объекта (индивида). Данные типа времени жизни включают в себя измерения времен до определенного события (например, заболевания). Часто момент интересующего события не наблюдается, что мотивирует вводить в модель цензурирование справа. В данном учебном пособии мы ограничимся рассмотрением простых наблюдений категориального или ординального типа.

## 2. Постановка задач анализа сопряженности признаков

Наблюдение категориального типа наиболее часто представляет собой дискретный случайный вектор $T = (T_1, \ldots, T_r)$ с конечным числом возможных значений для каждой компоненты (признака). Возможные значения $T_j$, необязательно числовые, называются уровнями $j$-го признака. Величины $T_1, \ldots, T_r$ могут быть качественными или ординальными. Отметим, что количественный измеряемый признак может быть трансформирован в ординальный или качественный путем группировки. Не умаляя общности считаем, что $T_j \in \{1, \ldots, d_j\}$, $j = 1, \ldots, r$. Для ординальных признаков порядок уровней сохраняется. Совместное распределение задается вероятностями $p_{i_1 \ldots i_r} = \Pr(T_1 = i_1, \ldots, T_r = i_r)$, $i_j \in \{1, \ldots, d_j\}$, $j = 1, \ldots, r$, которые удовлетворяют условиям $p_{i_1 \ldots i_r} \geq 0$, $\forall\ i_1, \ldots, i_r$ и $\sum_{i_1 \ldots i_r} p_{i_1 \ldots i_r} = 1$. Таким образом, модель статистического эксперимента является параметрической, где параметр распределения $p_{i_1,\ldots,i_r} \in \Theta$, $\Theta = \{p_{i_1,\ldots,i_r} : p_{i_1,\ldots,i_r} \geq 0, \sum_{i_1 \ldots i_r} p_{i_1 \ldots i_r} = 1\}$ имеет размерность $\dim(\Theta) = i_1 \cdot \ldots \cdot i_r - 1$.

Категориальные данные – выборка $(T_1, \ldots, T_n)$ из распределения $T$: $T_s = (T_{s1}, \ldots, T_{sr})$, $s = 1, \ldots, n$. Для вычисления функция правдоподобия $L(T; p) = \prod_{i_1 \ldots i_r} p_{i_1 \ldots i_r}^{n_{i_1 \ldots i_r}}$ достаточно иметь значения $n_{i_1 \ldots i_r}$, где $n_{i_1 \ldots i_r} = \sum_{s=1}^{n} \mathbb{1}_{\{T_{s1}=i_1,\ldots,T_{sr}=i_r\}}$ – число элементов выборки с соответствующими значениями компонент, а $\sum_{i_1,\ldots,i_r} n_{i_1 \ldots i_r} = n$ – размер выборки. Таким образом, по теореме факторизации Неймана–Фишера минимальная достаточная статистика представляет собой совокупность всех $n_{i_1 \ldots i_r}$, $i_j \in 1, \ldots, d_j$, $j = 1, \ldots, r$. Значения $n_{i_1 \ldots i_r}$ образуют массив сопряженности признаков размерности $d_1 \times \ldots \times d_r$.

## 2.1. Мультиномиальная и пуассоновская модели

Распределение дискретного случайного вектора $\nu$ называется мультиномиальным $\nu \in \mathrm{Mult}(p_1, \ldots, p_m; n)$, $m, n \in \mathbb{N}$, если

$$P(\nu_1 = n_1, \ldots, \nu_m = n_m) = \frac{n!}{n_1! \cdot \ldots \cdot n_m!} p_1^{n_1} \cdot \ldots \cdot p_m^{n_m} \mathbb{1}_{\{\sum_{i=1}^m n_i = n\}}.$$

Параметры мультиномиального распределения удовлетворяют соотношениям $p_i > 0$, $i = 1, \ldots, m$: $\sum_{i=1}^m p_i = 1$.

Сформулируем важные свойства мультиномиального распределения $\nu \in \mathrm{Mult}(p_1, \ldots, p_m; n)$.

1°. Для любой перестановки $(\sigma_1, \ldots, \sigma_m)$ индексов $(1, \ldots, m)$, $(\nu_{\sigma_1}, \ldots, \nu_{\sigma_m}) \in \mathrm{Mult}(p_{\sigma_1}, \ldots, p_{\sigma_m}; n)$ (инвариантность).

2°. При объединении уровней распределение остается мультиномиальным: $(\nu_1, \ldots, \nu_{m-2}, \nu_{m-1} + \nu_m) \in \mathrm{Mult}(p_1, \ldots, p_{m-2}, p_{m-1} + p_m; n)$. (Объединение уровней)

3°. Оценки максимального правдоподобия и выборочные оценки совпадают: $\hat{p}_i = \nu_i / n$, $i = 1, \ldots, m$.

4°. Пусть $p = (p_1, \ldots, p_m)^T$ и $\hat{p} = (p_1, \ldots, p_m)^T$. Тогда $\sqrt{n}(\hat{p} - p) \Rightarrow \mathcal{N}(0, R)$ при $n \to \infty$, с предельной матрицей ковариации $R = \|r_{ij}\|_{i,j}$, элементы которой задаются соотношениями

$$r_{ij} = \begin{cases} p_i(1 - p_i), & i = j \\ -p_i p_j, & i \neq j \end{cases}.$$

5°. Пусть $\mu_i = np_i$, $i = 1, \ldots, m$. Тогда,

$$X^2 = \sum_{i=1}^m \frac{(n_i - \mu_i)^2}{\mu_i} \Rightarrow \chi^2_{m-1}$$

при $n \to \infty$ (простой критерий $\chi^2$).

6°. Пусть $\dim(\Theta) = d$; выполнен ряд естественных условий, которым удовлетворяет совокупность $p_i = p_i(\theta)$, $\theta \in \Theta$; $n$ – известно; $\tilde{\mu}_i$ – оценка максимального правдоподобия для $np_i(\theta)$. Тогда,

$$\widehat{X}^2 = \sum_{i=1}^m \frac{(n_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \Rightarrow \chi^2_{m-d-1}$$

при $n \to \infty$ (сложный критерий $\chi^2$).

Очевидно, что массив сопряженности признаков имеет мультиномиальное распределение с фиксированным значением параметра $n$, равным размеру выборки.

Альтернативно, для построения модели категориальных данных используют распределение Пуассона. Пусть $A_1, A_2, \ldots$ – последовательность однородных событий, происходящих в случайные моменты времени $T_1, T_2, \ldots$. Поток событий называется простейшим (пуассоновским), если выполнены условия стационарности, ординарности и отсутствия последействия. Стационарность заключается в том, что вероятность появления $k$ событий в интервале $[s, s+t)$ не зависит от $s \geq 0$. Свойство ординарности состоит в том, что вероятность появления двух и более событий в малом интервале времени есть величина бесконечно малая по отношению к вероятности появления одного события в этом интервале, а отсутствие последействия означает, что числа событий, появляющихся в непересекающиеся интервалы времени, являются независимыми случайными величинами. Процесс Пуассона – точечный процесс начинающийся из нуля со скачками единичной величины в моменты появления событий простейшего потока. Интенсивность пуассоновского процесса определяется средним числом событий, появляющихся в единицу времени. Сформулируем основные свойства простейшего потока событий.

1°. Число событий $\nu(s,t)$ простейшего потока с интенсивностью $\lambda$, появляющихся в интервале времени $[s, s+t)$ имеет распределение Пуассона: $\nu(s,t) \sim \mathrm{Pois}(\lambda t)$

2°. Время ожидания $\tau(s)$ ближайшего события простейшего потока с интенсивностью $\lambda$, начиная с момента времени $s$, имеет показательное распределение: $\tau(s) \sim \mathrm{Exp}(\lambda)$

3°. Поток событий, получающийся слиянием двух (или более) независимых простейших потоков событий, будет простейшим с интенсивностью равной сумме интенсивностей исходных потоков событий

Пуассоновская модель категориального анализа основана на последовательном накоплении наблюдений. Считается, что $\nu_1, \ldots, \nu_m$ – независимые случайные величины, имеющие распределения Пуассона $\nu_i \in \mathrm{Pois}(\lambda_i)$, $\lambda_i > 0$. Иными словами, величина $\nu_i$ – число событий пуассоновского потока с интенсивностью $\lambda_i = tp_i$, накопленное к моменту времени $t > 0$. В этом случае, общее число наблюдений $n$ тоже имеет распределение Пуассона $n \in \mathrm{Pois}(t)$. Многомерное пуассоновское распределение вектора с независимыми компонентами $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_m)$ обозначим $\mathrm{Pois}(\lambda_1, \ldots, \lambda_m)$. Отметим,

что условное распределение $\boldsymbol{\nu} \in \mathrm{Pois}(\lambda_1, \ldots, \lambda_m)$ при фиксированной сумме $\sum_{i=1}^m \nu_i = n$ является мультиномиальным $\mathrm{Mult}(p_1, \ldots, p_m, n)$, где $p_i = \lambda_i / \sum_{j=1}^m \lambda_j$.

### 2.2. Многомерный эксперимент, структуризация

Рассмотрим результат статистического эксперимента $\boldsymbol{X} = (X_1, \ldots, X_r)$ : $X_i \in \{1, \ldots, d_i\}$, $i = 1, \ldots, r$, включающего $r$ признаков. Случайному вектору $\boldsymbol{X}$ можно сопоставить дискретную случайную величину $Y \in \{1, \ldots, d_1 \cdot \ldots \cdot d_r\}$ ($\boldsymbol{X} \leftrightarrow Y$) с конечным множеством значений. При выборе модели можно использовать классический или пуассоновский подход. Минимальная достаточная статистика $\{n^*\}_{j=1}^{n_{j_1 \ldots j_r}}$ по выборке из распределения $Y$ имеет мультиномиальное или пуассоновское распределение соответственно. Большинство задач касается форм зависимости компонент совместного распределения $\boldsymbol{X}$. Для постановки задач важна многомерная структура множества значений эксперимента (индекса), поэтому используют структурированный набор $\|n_{j_1 \ldots j_r}\|_{j_1 \ldots j_r}$, являющийся минимальной достаточной статистикой по выборке из распределения $\boldsymbol{X}$, но его вероятностные свойства выводятся из свойств $\{n^*\}_{j=1}^{n_{j_1 \ldots j_r}}$. В классической модели параметр представляет собой $r$-мерный массив $\|p_{j_1 \ldots j_r}\|_{j_1 \ldots j_r}$ неотрицательных чисел, обладающих свойством $\sum_{j_1 \ldots j_r} p_{j_1 \ldots j_r} = 1$. В пуассоновской модели параметр представляет собой $r$-мерный массив $\|\lambda_{j_1 \ldots j_r}\|_{j_1 \ldots j_r}$ и может быть переписан в виде упорядоченной пары $(\lambda, \|p_{j_1 \ldots j_r}\|_{j_1 \ldots j_r})$, где $\lambda_{j_1 \ldots j_r} = \lambda p_{j_1 \ldots j_r}$. Отметим, что размерность параметра в классической модели на единицу меньше, чем в пуассоновской. В пуассоновской модели параметр $\lambda$ считается мешающим и не представляет интереса для исследователя.

В заключение отметим, что многие задачи допускают формулировку в терминах условных распределений. Переход от совместных к условным распределениям востребован при наличии контроля некоторых признаков исследователем.

### 2.3. Постановка статистического эксперимента

Правильная постановка статистического исследования может существенно повысить его эффективность и сократить затраты на его организацию. Пассивный подход подразумевает простое наблюдение значений признаков. При постановке задач можно использовать как совместное, так

и условные распределения признаков. В рамках активного подхода подразумевается принудительный выбор значений одного или нескольких признаков. Обычно различают наблюдаемые и контролируемые признаки (зависимые и независимые переменные соответственно). При наличии контролируемых признаков изучение совместного распределения теряет смысл. При формировании модели активного статистического эксперимента используется условное распределение наблюдаемого признака при условии контролируемого признака.

Рассмотрим основные подходы к организации статистического эксперимента. *Разовый скрининг* (Cross sectional design (Англ.)) подразумевает измерение интересующих характеристик в фиксированный момент времени для выбранной популяции. Такой подход к постановке эксперимента по сути не требует предварительной подготовки и применим для исследования текущих характеристик исследуемых объектов. Ввиду относительно небольшой затратности, данный подход позволяет получать достаточно большие объемы данных. *Когортный план* подразумевает формирование и сопровождение когорты индивидов, разделенной на группы по значению независимой переменной. В частности, когортный дизайн используется для исследования прогрессивных изменений наблюдаемых признаков с течением времени. С точки зрения организации, когортный дизайн гораздо более затратный, чем разовый скрининг, но интерпретация результатов существенно более определенная. В условиях когортного исследования для повышения эффективности статистического анализа обычно используется активный подход. При организации когортного исследования возможно использовать ретроспективный или перспективный подходы. *Ретроспективный* подход подразумевает фиксирование значений контролируемого признака в момент получения значения наблюдаемого признака (пассивный эксперимент или контроль распределения наблюдаемого признака). При использовании *перспективного* подхода, значения контролируемого признака выбирают в начале исследования (активный эксперимент). Чтобы минимизировать влияние субъективного фактора на результат эксперимента в рамках перспективного подхода часто используют рандомизацию, заключающуюся в случайном выборе контролируюемого признака или набора контролируемых признаков. Иными словами, перед началом исследования индивиды распределяются по группам случайным образом.

При проведении когортного исследования на людях предварительное знание значений контролируемых признаков может повлиять на результат исследования (психологический фактор). В рамках слепого подхода к организации когортного исследования, значения контролируемых факторов должны быть неизвестны исследуемым индивидам. Если значения контролируемых факторов неизвестны не только исследуемым индивидам, но и исследователю, то такой подход называется двойным слепым.

## 3. Классические методы анализа категориальных данных

Методы анализа категориальных данных не ограничиваются лишь анализом таблиц сопряженности признаков, но основным приложением классических методов в задачах биостатистики являются таблицы сопряженности.

### 3.1. Анализ сопряженности двух признаков

Статистическую информацию по выборке из двумерного дискретного распределения с конечными числами уровней компонент удобно записывать в виде таблицы сопряженности (Табл. 3.1). Ставится задача изучения зависимости признаков по имеющимся статистическим данным. В зависимости от плана эксперимента можно использовать совместное или условное распределение. В классической модели совместного распределения $(Y, X)$ элементы таблицы сопряженности имеют совместное мультиномиальное распределение с параметрами $\|p_{ij}\|_{ij}$ (Табл. 3.2). При постановке активного эксперимента исследователь может контролировать значения одного из признаков (независимая переменная) и наблюдать за изменением распределения другого признака. В этом случае, правильно использовать модель условного распределения наблюдаемого признака при известном значении контролируемого признака. Считаем, что $p_{+j} > 0$, $j = 1, \ldots, d_2$ и $p_{i+} > 0$, $i = 1, \ldots, d_1$. В этом случае $(n_{1j}, \ldots, n_{d_1 j}) \in \mathrm{Mult}(p_{1|j}, \ldots, p_{d_1|j}; n_{+j})$ при всех $j = 1, \ldots, d_2$, где параметры условного распределения задаются соотношениями $p_{i|j} = p_{ij}/p_{+j}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$ (Табл. 3.3).

*Таблица 3.1.*

| $Y$ | $X$ | | | Всего |
|---|---|---|---|---|
| | 1 | ... | $d_2$ | |
| 1 | $n_{11}$ | ... | $n_{1d_2}$ | $n_{1+}$ |
| ... | ... | ... | ... | ... |
| $i$ | $n_{i1}$ | ... | $n_{id_2}$ | $n_{i+}$ |
| ... | ... | ... | ... | ... |
| $d_1$ | $n_{d_11}$ | ... | $n_{d_1d_2}$ | $n_{d_1+}$ |
| Всего | $n_{+1}$ | ... | $n_{+d_2}$ | $n$ |

*Таблица 3.2.*

| $Y$ | $X$ | | | Распр. $Y$ |
|---|---|---|---|---|
| | 1 | ... | $d_2$ | |
| 1 | $p_{11}$ | ... | $p_{1d_2}$ | $p_{1+}$ |
| ... | ... | ... | ... | ... |
| $i$ | $p_{i1}$ | ... | $p_{id_2}$ | $p_{i+}$ |
| ... | ... | ... | ... | ... |
| $d_1$ | $p_{d_11}$ | ... | $p_{d_1d_2}$ | $p_{d_1+}$ |
| Распр. $X$ | $p_{+1}$ | ... | $p_{+d_2}$ | 1 |

*Таблица 3.3.*

| $Y$ | $X$ | | |
|---|---|---|---|
| | 1 | ... | $d_2$ |
| 1 | $p_{1\|1}$ | ... | $p_{1\|d_2}$ |
| ... | ... | ... | ... |
| $i$ | $p_{i\|1}$ | ... | $p_{i\|d_2}$ |
| ... | ... | ... | ... |
| $d_1$ | $p_{d_1\|1}$ | ... | $p_{d_1\|d_2}$ |
| Сумма | 1 | 1 | 1 |

Гипотеза независимости признаков $H_0 : p_{ij} = p_{i+}p_{+j}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_1$, в модели совместного распределения соответствует гипотезе однородности $H_0^* : p_{i|j} = p_{i|1}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_1$ для условного распределения.

**Таблицы сопряженности** $2 \times 2$**.** Отдельного рассмотрения заслуживает наиболее простая и распространенная модель с бинарными признаками $(X, Y)$ $(d_1 = d_2 = 2)$. Результаты статистического эксперимента записываются в таблицу сопряженности $2 \times 2$. При работе с таблицами сопряженности $2 \times 2$ удобно использовать уровни обоих признаков $\{0, 1\}$, вместо $\{1, 2\}$. Введем специальные обозначения для условных вероятностей или пропорций $\pi_j = p_{1|j} = 1 - p_{0|j}$, $j = 0, 1$ (proportions (Англ.)). В качестве меры зависимости бинарных признаков можно использовать разность условных вероятностей $\pi_1 - \pi_0$. Отметим, что интерпретация разности условных вероятностей зависит от их значений. Разность пропроций малоинформативна, если оба значения близки к 0 (или к 1). Альтернативно, в качестве меры зависимости бинарных признаков можно использовать отношение пропорций $\pi_1/\pi_0$ (relative risks (Англ.)). Если $\pi_1$ и $\pi_0$ близки к 1, то их отношение неинформативно, и в этом случае разумнее использовать отношение $(1-\pi_1)/(1-\pi_0)$. Отношения $\text{odds}_j = \pi_j/(1-\pi_j)$, $j = 0, 1$, называются *шансами* (odds (Англ.)). Наиболее часто, в качестве меры зависимости бинарных признаков используют *отношение шансов* $\theta = \text{odds}_1/\text{odds}_0 = p_{00}p_{11}/(p_{01}p_{10})$ (odds ratio (Англ.)).

Гипотеза независимости признаков допускает представление в каждой из следующих эквивалентных форм

$$H_I : p_{ij} = p_{i+}p_{+j},\ i, j \in \{0, 1\} \quad \Leftrightarrow \quad H_I : \pi_1 = \pi_2 \quad \Leftrightarrow \quad H_I : \theta = 1.$$

Оценки максимального правдоподобия параметров модели совпадают с выборочными оценками

$$\hat{p}_{ij} = \frac{n_{ij}}{n},\ i,j \in \{0,1\}$$

При переходе к условному распределению $Y$ при условии $X$ оценки параметра $(\pi_1, \pi_2)$ задаются соотношениями $\hat{\pi}_j = \dfrac{n_{1j}}{n_{1+}},\ j \in \{0,1\}$. Соответствующая оценка отношения шансов $\theta$ имеет вид

$$\hat{\theta} = \frac{\hat{p}_{00}\hat{p}_{11}}{\hat{p}_{10}\hat{p}_{01}} = \frac{\hat{\pi}_1(1-\hat{\pi}_0)}{\hat{\pi}_0(1-\hat{\pi}_1)} = \frac{n_{00}n_{11}}{n_{10}n_{01}}.$$

При построении асимптотических доверительных интервалов для разности пропорций используются соотношения

$$\sqrt{n_i}\frac{\hat{\pi}_i - \pi_i}{\sqrt{\pi_i(1-\pi_i)}} \Rightarrow \mathcal{N}(0,1).$$

при $n_i = n_{i+} \to \infty$. В предположении $n_1 = n - n_0 = \gamma n$, $\gamma \in (0,1)$ и $n \to \infty$ с учетом условной независимости величин $\hat{\pi}_1$ и $\hat{\pi}_0$ получаем, что

$$\frac{(\hat{\pi}_1 - \hat{\pi}_0) - (\pi_1 - \pi_0)}{\sqrt{\pi_1(1-\pi_1)/n_1 + \pi_0(1-\pi_0)/n_0}} \Rightarrow \mathcal{N}(0,1).$$

Условия равномерности роста $n_1$ и $n_0$ могут быть ослаблены путем замены $\gamma$ на $\gamma_n$: $\gamma_n \in (\delta, 1-\delta)$ при всех $n$ и некотором $\delta > 0$. С использованием последнего соотношения, получаем асимптотический доверительный интервал для разности пропорций $\pi_1 - \pi_0$,

$$\left[\hat{\pi}_1 - \hat{\pi}_0 - x_\alpha\sqrt{\tfrac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \tfrac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0}}, \hat{\pi}_1 - \hat{\pi}_0 + x_\alpha\sqrt{\tfrac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \tfrac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0}}\right],$$

где $x_\alpha = \Phi^{-1}(1-\alpha/2)$ – квантиль стандартного нормального распределения. Чтобы построить доверительный интервал для отношения $\pi_1/\pi_0$ переходим к логарифмам. С использованием формулы Тейлора получаем, что

$$\sqrt{n_i}(\ln\,\hat{\pi}_i - \ln\,\pi_i) = \pi_i^{-1}\sqrt{n_i}(\hat{\pi}_i - \pi_i) + o_P(1) \Rightarrow \mathcal{N}(0, (1-\pi_i)/\pi_i).$$

Таким образом, при выполнении условия $n_{+1} = \gamma n$, $\gamma \in (0,1)$ при $n \to \infty$

$$\frac{\ln(\hat\pi_1/\hat\pi_0) - \ln(\pi_1/\pi_0)}{\sqrt{(1-\pi_1)/(n_{+1}\pi_1) + (1-\pi_0)/(n_{+0}\pi_0)}} \Rightarrow \mathcal{N}(0,1),$$

что позволяет получить асимптотический доверительный интервал для $\pi_1/\pi_0$,

$$\Big[\frac{\hat\pi_1}{\hat\pi_0} e^{-x_\alpha \hat\sigma_{rr}},\ \frac{\hat\pi_1}{\hat\pi_0} e^{x_\alpha \hat\sigma_{rr}}\Big],$$

где $\hat\sigma^2_{rr} = \dfrac{1-\hat\pi_1}{n_{+1}\hat\pi_1} + \dfrac{1-\hat\pi_0}{n_{+2}\hat\pi_0}$. Аналогичным образом, получаем асимптотическую нормальность шансов при $i = 0, 1$,

$$\sqrt{n_i}\big(\ln(\widehat{\text{odds}_i}) - \ln(\text{odds}_i)\big) = \frac{\sqrt{n_i}(\hat\pi_i - \pi_i)}{(1-\pi_i)\pi_i} + o_P(1) \Rightarrow \mathcal{N}\Big(0, \frac{1}{(1-\pi_i)\pi_i}\Big),$$

откуда выводим асимптотическую нормальность отношения шансов при $n_{+1} = \gamma n,\ \gamma \in (0,1)$

$$\frac{\ln\hat\theta - \ln\theta}{\sqrt{n_{11}^{-1} + n_{10}^{-1} + n_{01}^{-1} + n_{00}^{-1}}} \Rightarrow \mathcal{N}(0,1),$$

Таким образом, асимптотический доверительный интервал для $\theta$ равен

$$\Big[\hat\theta e^{-x_\alpha \hat\sigma_{or}},\ \hat\theta e^{x_\alpha \hat\sigma_{or}}\Big],$$

где $\hat\sigma^2_{or} = 1/n_{00} + 1/n_{10} + 1/n_{01} + 1/n_{11}$.

Далее рассмотрим задачу проверки гипотезы независимости признаков $H_I$. Наиболее известным критерием для проверки гипотезы независимости двух признаков является $\chi^2$. Разумеется, данный критерий применим и в случае бинарных признаков. Статистика критерия задаваемая соотношением $X^2 = \sum_{i,j=1}^{2} \dfrac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}$ имеет асимптотическое распределение $\chi^2_1$. Доверительная область имеет следующий вид $X^2 \le x_\alpha$, $x_\alpha : K_1(x_\alpha) = 1-\alpha$, а $P$-значение вычисляется по формуле $PV = 1 - K_1(X^2)$, где $K_1$ – функция распределения $\chi^2_1$. Еще один универсальный критерий – критерий отношения правдоподобия. Статистика критерия $G^2 = 2\sum_{i,j=1}^{2} n_{ij}\ln(n_{ij}n/(n_{i+}n_{+j}))$ асимптотически эквивалентна

статистике $X^2$, а следовательно, доверительная область определяется соотношением $X^2 \leq x_\alpha$, $x_\alpha : K_1(x_\alpha) = 1 - \alpha$, и $P$-значение вычисляется по формуле $PV = 1 - K_1(G^2)$. Отметим асимптотическую эффективность критерия отношения правдоподобия, а следовательно, и критерия $\chi^2$. Критерии $\chi^2$ и отношения правдоподобия могут давать большую ошибку при наличии малого числа наблюдений хотя бы в одной из ячеек таблицы сопряженности. Не рекомендуется использовать данные критерии, если хотя бы одна из ожидаемых частот $n_{i+}n_{+j}/n$, $i, j = 1, 2$ меньше 5.

Помимо универсальных критериев, применимых для проверки независимости двух признаков с любыми числами уровней, отметим некоторые специальные критерии для таблиц сопряженности $2 \times 2$. Асимптотические свойства отношения шансов $\hat{\theta}$, обсуждавшиеся ранее, позволяют построить $Z$-критерий (критерий типа Вальда). Статистика критерия $Z = \ln \hat{\theta}/\sigma_{or}$ при справедливости основной гипотезы $H_I : \theta = 1$ имеет асимптотическое стандартное нормальное распределение $\mathcal{N}(0, 1)$. Доверительная область двухстороннего критерия может быть записана в виде $Z \in [-x_\alpha, x_\alpha]$; $x_\alpha : \Phi(x_\alpha) = 1-\alpha/2$, а $P$-значение равно $PV = 2(1-\Phi(|Z|)) = 1-K_1(Z^2)$. Поскольку данный критерий является асимптотическим, то он также не рекомендован к применению при малом числе наблюдений в ячейках таблицы сопряженности. К достоинствам данного критерия следует причислить прозрачность альтернативы и возможность постановки задачи при односторонней альтернативе. Разумеется, для проверки гипотезы независимости при односторонней альтернативе $H_A : \theta > 1$ (или $H_A : \theta < 1$), потребуется иной подход к определению доверительного множества и $P$-значения.

Еще один специальный критерий проверки независимости – точный критерий Фишера (или чайный тест). Отличительная особенность данного критерия состоит в том, что фиксированы распределения обоих признаков. Как и в случае $Z$-критерия нулевую гипотезу удобно записывать в виде $H_I : \theta = 1$, а альтернатива может быть двухсторонней $H_A : \theta \neq 1$ или односторонней $H_A : \theta > 1$ ($H_A : \theta < 1$). Статистика критерия $n_{11}$ при справедливости основной гипотезы имеет гипергеометрическое распределение с вероятностями

$$\Pr(n_{11} = k) = \frac{C_{n_{1+}}^{k} C_{n_{0+}}^{n_{+1}-k}}{C_n^{n_{+1}}}$$

При наличии двухсторонней альтернативы доверительная область представляется в виде $n_{11} \in Q_\alpha$, где $Q_\alpha$ – наибольшее множество значений $k$: $\sum_{k':q(k')\geq q(k)} q(k') < 1-\alpha$, а $P$-значение вычисляется по формуле $PV = \sum_{k:q(k)\leq q(n_{11})} q(k)$. В случае односторонней альтернативы $H_A : \theta > 1$ доверительная область принимает вид $n_{11} \leq k_\alpha$, где $k_\alpha$ – соответствующая квантиль гипергеометрического распределения, а $P$-значение равно $PV = \sum_{k\geq n_{11}} q(k)$.

**Пример Фишера (чайный тест).** Традиционным напитком англичан является чай с молоком. Одна из коллег Фишера утверждала, что чувствует на вкус, что именно (чай или молоко) было налито первым. При этом она не утверждала, что может идентифицировать способ приготовления напитка достоверно, но все-таки чаще угадывает правильно, чем ошибается. Для проверки данного предположения было приготовлено восемь чашек чая с молоком, в четыре из которых сначала наливали молоко, а в остальные четыре — чай. Коллеге Фишера была поставлена задача выбрать четыре чашки, в которые сначала было налито молоко. В результате эксперимента три чашки были выбраны верно, а одна — ошибочно. Результат данного эксперимента можно записать в виде таблицы 2×2 (Табл. 3.4).

*Таблица 3.4.*

| Налито сначала | Предположение коллеги Фишера | | Всего |
|---|---|---|---|
| | Молоко | Чай | |
| Молоко | 3 | 1 | 4 |
| Чай | 1 | 3 | 4 |
| Всего | 4 | 4 | 8 |

Основная гипотеза состоит в том, что решение коллеги Фишера не зависит от истинного способа приготовления напитка. В терминах отношения шансов данная гипотеза может быть записана в виде $H_0 : \theta = 1$. В качестве альтернативы выбирается гипотеза $H_A : \theta > 1$. Вероятность $P_U = P(3) + P(4) = 0.243$, а следовательно, оснований для признания способности коллеги Фишера идентифицировать способ приготовления напитка на вкус на уровне значимости 0.05 (и даже на уровне значимости 0.2) нет. Предположим на минутку, что все четыре чашки были идентифицированы верно. В этом случае $P_U = 0.014$, т. е. следовало бы признать наличие способности идентифицировать способ приготовления напитка на уровне значимости 0.05.

**Таблицы сопряженности** $d_1 \times d_2$. Структура зависимости в таблицах $d_1 \times d_2$ существенно сложнее. Отметим, что размерность параметра, характеризующего зависимость, равна $(d_1 - 1)(d_2 - 1)$. В частности, для характеризации зависимости можно использовать набор отношений шансов для $(d_1 - 1)(d_2 - 1)$ миноров $2 \times 2$ таблицы сопряженности, однако интерпретировать такой параметр достаточно трудно. К сожалению, не существует простого способа характеризовать зависимость признаков. Иногда, для характеризации силы зависимости используют коэффициент неопределенности

$$U = -\frac{\sum_{i,j} p_{ij} \ln(p_{ij}/p_{i+}p_{+j})}{\sum_j p_{+j} \ln(p_{+j})}$$

Для ординальных признаков в качестве меры линейной зависимости часто используют коэффициент корреляции Пирсона. Альтернативно, для измерения силы зависимости можно использовать меры согласованности $\Pi_c = 2\sum_{ij} p_{ij}\left(\sum_{h>i}\sum_{s>j} p_{ij}\right)$ и несогласованности и $\Pi_d = 2\sum_{ij} p_{ij}\left(\sum_{h>i}\sum_{s<j} p_{ij}\right)$ соответственно. Например, коэффициент ранговой корреляции $\tau$-Кендалла равен разности $\tau = \Pi_c - \Pi_d$.

Для проверки гипотезы независимости признаков обычно используют критерий $\chi^2$ или критерий отношения правдоподобия. Статистика критерия $\chi^2$, определяемая соотношением

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n},$$

имеет при основной гипотезе асимптотическое распределение $\chi^2_{(d_1-1)(d_2-1)}$. Доверительная область определяется соотношением $X^2 \le x_\alpha$, где $x_\alpha : K_{(d_1-1)(d_2-1)}(x_\alpha) = 1 - \alpha$, а $P$-значение вычисляется по формуле $PV = 1 - K_{(d_1-1)(d_2-1)}(X^2)$. Статистика критерия отношения правдоподобия $G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \ln(n_{ij}n/(n_{i+}n_{+j}))$ асимптотически эквивалентна статистике $\chi^2$-критерия, поэтому доверительная область и формула для вычисления $P$-значения получаются аналогично.

Гипотеза некоррелированности признаков $H_L : r = 0$ формулируется с использованием коэффициента корреляции Пирсона $r$. Известно,

что $H_I \Rightarrow H_L$, но обратная импликация не верна. Таким образом, отвержение гипотезы $H_L$ автоматически влечет за собой отвержение гипотезы $H_I$. Статистика критерия проверки гипотезы $H_L$ (Trend test, (Англ.)) $M^2 = (n-1)r^2$ при справедливости гипотезы $H_L$ имеет асимптотическое распределение $\chi^2_1$. Доверительная область данного критерия имеет вид $M^2 \leq x_\alpha$, $x_\alpha : K_1(x_\alpha) = 1 - \alpha$, а $P$-значение вычисляется по формуле $PV = 1 - K_1(G^2)$.

### 3.2. Анализ сопряженности трех признаков

При наличии трех признаков $(X, Y, Z)$ структура зависимости гораздо более сложная, чем при наличии двух признаков. Полезная статистическая информация может быть записана в виде трехмерного $(d_1 \times d_2 \times d_3)$-массива сопряженности (Табл. 3.5) или, что то же самое, набор из $d_3$ двумерных $(d_1 \times d_2)$-таблиц сопряженности. В качестве модели можно использовать мультиномиальное распределение с фиксированным числом наблюдений в таблице сопряженности или распределение Пуассона предположив, что число наблюдений в таблице сопряженности имеет распределение Пуассона. Значения $X, Y, Z$ можно рассматривать как три равнозначные наблюдаемые переменные, как две наблюдаемые переменные и одна контролируемая, а также как одна наблюдаемая переменная и две контролируемые. При фиксированном числе наблюдений в таблице в случае трех наблюдаемых переменных речь идет о совместном мультиномиальном распределении, параметр которого определяется соотношениями $p_{ijs} = \Pr(X = i, Y = j, Z = s)$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$, $s = 1, \ldots, d_3$ (Табл. 3.6), а размерность параметра равна $d_1 d_2 d_3 - 1$. Если речь идет о двух наблюдаемых переменных $X, Y$ и одной контролируемой $Z$, то рассматриваются условные мультиномиальные распределения с параметром включающим в себя условные вероятности $p_{ij|s} = p_{ijs}/p_{++s}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$, $s = 1, \ldots, d_3$ (Табл. 3.7) и имеющем размерность $(d_1 d_2 - 1)d_3$. В случае контроля двух переменных $(X, Z)$ речь идет о наборе $d_2 d_3$ одномерных таблиц сопряженности, каждая из которых имеет мультиномиальное распределение, где в качестве параметров используются условные вероятности $p_{j|ik} = p_{ijk}/p_{i+k}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$, $s = 1, \ldots, d_3$ (Табл. 3.8). Изучение условных распределений и получение выводов о всей популяции носит название *стратифицированный* (stratified (Англ.)–послойный) анализ.

*Таблица 3.5.*

| $X$ | $Y$ | | | | | Всего |
|---|---|---|---|---|---|---|
| | 1 | ... | $j$ | ... | $J$ | |
| 1 | $n_{11s}$ | ... | $n_{1js}$ | ... | $n_{1d_2s}$ | $n_{1+}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $i$ | $n_{i1s}$ | ... | $n_{ijs}$ | ... | $n_{iJs}$ | $n_{i+s}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $I$ | $n_{d_11s}$ | ... | $n_{Ijs}$ | ... | $n_{d_1d_2s}$ | $n_{I+s}$ |
| Всего | $n_{+1s}$ | ... | $n_{+d_2s}$ | ... | $n_{+d_2s}$ | $n_{++s}$ |

*Таблица 3.6.*

| $X$ | $Y$ | | | | | Сумма |
|---|---|---|---|---|---|---|
| | 1 | ... | $j$ | ... | $d_2$ | |
| 1 | $p_{11s}$ | ... | $p_{1js}$ | ... | $p_{1d_2s}$ | $p_{1+s}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $i$ | $p_{i1s}$ | ... | $p_{ijs}$ | ... | $p_{iJs}$ | $p_{i+s}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $I$ | $p_{d_11s}$ | ... | $p_{d_1js}$ | ... | $p_{d_1d_2s}$ | $p_{d_1+s}$ |
| Сумма | $p_{+1s}$ | ... | $p_{+js}$ | ... | $p_{+d_2s}$ | $p_{++s}$ |

*Таблица 3.7.*

| $X$ | $Y$ | | | | | Сумма |
|---|---|---|---|---|---|---|
| | 1 | ... | $j$ | ... | $J$ | |
| 1 | $p_{11\|s}$ | ... | $p_{1j\|s}$ | ... | $p_{1J\|s}$ | $p_{1+\|s}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $i$ | $\hat{p}_{i1\|s}$ | ... | $p_{ij\|s}$ | ... | $p_{iJ\|s}$ | $p_{i+\|s}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $I$ | $p_{I1\|s}$ | ... | $p_{Ij\|s}$ | ... | $p_{IJ\|s}$ | $p_{I+\|s}$ |
| Сумма | $p_{+1\|s}$ | ... | $p_{+j\|s}$ | ... | $p_{+J\|s}$ | 1 |

*Таблица 3.8.*

| $X$ | $Y$ | | | | |
|---|---|---|---|---|---|
| | 1 | ... | $j$ | ... | $d_2$ |
| 1 | $p_{1\|1s}$ | ... | $p_{1\|js}$ | ... | $p_{1\|d_2s}$ |
| ... | ... | ... | ... | ... | ... |
| $i$ | $p_{1\|is}$ | ... | $p_{j\|is}$ | ... | $p_{d_2\|is}$ |
| ... | ... | ... | ... | ... | ... |
| $I$ | $p_{1\|d_1s}$ | ... | $p_{j\|d_1s}$ | ... | $p_{d_2\|d_1s}$ |
| Сумма | 1 | 1 | 1 | ... | 1 |

В качестве оценок параметров $p_{ijs}$ используются выборочные оценки $\hat{p}_{ijs} = n_{ijs}/n$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$, $s = 1, \ldots, d_3$, которые одновременно являются оценками максимального правдоподобия, а для оценивания условных вероятностей достаточно подставить соответствующие оценки в формулы для их вычисления. Оценки, соответствующие различным уровням (слоям) стратификации считаются независимыми. Асимптотическая нормальность оценок параметров следует непосредственно из свойств мультиномиального распределения.

Далее сформулируем наиболее востребованные гипотезы анализа таблиц сопряженности трех признаков. При наличии равнозначных признаков можно говорить об их независимости

$$H_I : p_{ijs} = p_{i++}p_{+j+}p_{++s}, \quad i = 1, \ldots, d_1, \quad i = 1, \ldots, d_2, \quad s = 1, \ldots, d_3.$$

Статистика критерия хи-квадрат для проверки данной гипотезы имеет вид

$$X^2 = \sum\nolimits_{i,j,s} \frac{(n_{ijs} - \hat{\mu}_{ijs})^2}{\hat{\mu}_{ijs}},$$

где $\hat{\mu}_{ijs} = n\hat{p}_{i++}\hat{p}_{+j+}\hat{p}_{++s} = n_{i++}n_{+j+}n_{++s}/n^2$, имеет асимптотическое распределение $\chi^2_q$, $q = d_1d_2d_3 - d_1 - d_2 - d_3 + 2$. Для вычисления $P$-значения критерия используют формулу $PV = 1 - K_q(X^2)$. Альтернативно, можно использовать критерий отношения правдоподобия, статистика которого

$G^2 = 2\sum_{ijs} n_{ijs}\ln(n_{ijs}/\hat{\mu}_{ijs})$ асимптотически эквивалентна статистике $X^2$.

Помимо независимости всех трех признаков можно рассматривать гипотезу условной независимости

$$H_{CI(Z)} : p_{ij|s} = p_{i+|s}p_{+j|s},\ i = 1,\dots,d_1,\ j = 1,\dots,d_2,\ s = 1,\dots,d_3.$$

или

$$H_{CI(Z)} : p_{i|1s} = \dots = p_{i|d_2 s},\ \ i = 1,\dots,d_1,\ \ s = 1,\dots,d_3.$$

Известно, что условная независимость не влечет независимости $X$ и $Y$. Еще одна гипотеза, проверка которой может быть интересна исследователю:

$$H_{I(Z)} : p_{ij|s} = p_{ij|1}, i = 1,\dots,d_1,\ \ j = 1,\dots,d_2,\ \ s = 2,\dots,d_3,$$

характеризует независимость вектора $(X, Y)$ от $Z$, но проверка этой гипотезы фактически не требует трехфакторного подхода, поскольку исходную задачу можно записать в виде $(d_1 d_2 \times d_3)$-таблицы. В части 2 учебного пособия структура зависимости будет наглядно представлена в терминах коэффициентов обобщенной линейной модели Пуассона.

Проверка сформулированных гипотез обычно проводится с использованием Хи-квадрат критериев для сложной гипотезы или критериев отношения правдоподобия.

**Таблицы сопряженности** $2 \times 2 \times d$. Отдельно рассмотрим случай анализа таблиц сопряженности при наличии двух бинарных наблюдаемых переменных $X, Y$ и одной контролируемой $Z$. Такая постановка не исключает возможности случайности $Z$, но для изучения совместного распределения $(X, Y, Z)$ потребуется дополнительный параметр, характеризующий распределение $Z$. Введем понятие пропорций $\pi_{j(s)} = p_{1|js}$, шансов $\mathrm{odds}_{i(s)} = \pi_{i(s)}/(1 - \pi_{i(s)})$ и отношений шансов $\theta_{XY(s)} = = \pi_{1(s)}/\pi_{2(s)}$. В данной модели разговор о совместном распределении величин $X$ и $Y$ не имеет смысла, т.е. гипотеза независимости $X$ и $Y$ не может быть сформулирована без введения дополнительных параметров $p_{++s}$, $s = 1,\dots,d$. Рассмотрим интересный пример, иллюстрирующий как появление контролируемого признака может поменять статистические выводы.

**Парадокс Симпсона.** Статистика осложнений после операции описана в следующей таблице

*Таблица 3.9*

| | Общая | | Начальная | | Запущенная | |
|---|---|---|---|---|---|---|
| Больница | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ |
| С осложнениями | 66 | 17 | 7 | 8 | 59 | 9 |
| Без осложнений | 2034 | 783 | 593 | 592 | 1441 | 191 |
| Всего | 2100 | 800 | 600 | 600 | 1500 | 200 |
| Осложнений (%) | 3.14% | 2.13% | 1.17% | 1.33% | 3.93% | 4.50% |

В левой части таблицы приведена статистика возникновения осложнений по больницам, тогда как правая часть включает в себя дополнительный фактор, характеризующий запущенность болезни к моменту проведения операции. Очевидно, что исходя из общей статистики $\theta = 1.495 > 1$, т.е. оцененная вероятность возникновения осложнений в больнице $A$ ниже, чем в больнице $B$. С другой стороны, стратификация по степени запущенности болезни дает $\theta_I = 0.874 < 1$ для начальной стадии болезни и $\theta_{II} = 0.869 < 1$ для запущенной болезни, тем самым отдавая приоритет лечению в больнице $B$, ввиду того, что в обоих случаях оцененная вероятность возникновения осложнений в больнице $B$ ниже, чем в больнице $A$. Полученная противоречивость статистических выводов объясняется тем, что процент запущенных случаев болезни в больнице B (71.4%) гораздо выше, чем в больнице A (25%), а это существенно влияет на общую статистику. Отметим также, что значения $\theta_I$ и $\theta_{II}$ очень похожи (однородность связи), но при этом существенно отличаются от единицы.

При наличии $(2 \times 2 \times d)$-таблицы сопряженности, гипотеза условной независимости может быть записана в терминах отношений шансов, стратифицированных по значениям признака $Z$

$$H_{CI(Z)} : \theta_{XY(1)} = \ldots = \theta_{XY(S)} = 1.$$

Для её проверки часто используют классический критрий Кочрана–Мантела–Хензела (CMH). В основе критерия CMH лежит асимптотическая нормальность статистики

$$\sum_s (n_{11s} - \hat{\mu}_{11s}) \Big/ \Big(\sum_s \widehat{\mathbb{D}}(n_{11s})\Big)^{1/2},$$

где

$$\hat{\mu}_{11s} = \frac{n_{1+s}n_{+1s}}{n_{++s}}, \;\; \widehat{\mathbb{D}}(n_{11s}) = \frac{n_{1+s}n_{2+s}n_{+1s}n_{+2s}}{n_{++s}^2(n_{++s}-1)}$$

— выборочные аналоги для $\mu_{11s} = \mathbb{E}_{H_{CI}}(n_{11s}) = n_{++s}p_{11s}$ и $\mathbb{D}_{H_{CI}}(n_{11s}) = n_{++s}p_{11s}(1-p_{11s})$. При выполнении основной гипотезы $H_{CI}$ данная статистика слабо сходится к стандартному нормальному распределению, а следовательно, статистика критерия CMH сходится слабо к распределению $\chi^2_1$

$$\text{CMH} = \frac{\left(\sum_s (n_{11s} - \hat{\mu}_{11s})\right)^2}{\sum_s \widehat{\mathbb{D}}(n_{11s})} \Rightarrow \chi^2_1 .$$

Для вычисления $P$-значения можно воспользоваться формулой $PV = 1 - K_1(\text{CMH})$. Отдельно отметим, что критерий CMH нечувствителен ко многим типам альтернатив, поэтому в большинстве случаев предпочтительнее использовать критерий $\chi^2$ или критерий отношения правдоподобия. Иногда рассматривают более слабую гипотезу однородности связей

$$H_{HA} : \theta_{XY(1)} = \ldots = \theta_{XY(d)}$$

Статистика классического критерия Бреслоу–Дея проверки данной гипотезы

$$\text{BD} = \sum_s \frac{(K_{ijs} - \hat{\mu}_{ijs})^2}{\hat{\mu}_{ijs}},$$

где $\hat{\mu}_{ijs}$ — оценки максимального правдоподобия в условиях $H_{HA}$, слабо сходится к распределению $\chi^2_{d-1}$. Отметим, что вычисление $\hat{\mu}_{ijs}$ достаточно затруднительно, для их вычисления используют численные методы. Кроме того, при малых объемах выборки классическая статистика Бреслова–Дея требует поправки. Для проверки гипотез $H_{CI}$ и $H_0$ при малых объемах выборок имеются точные критерии. В случае справедливости $H_{HA}$ говорят об общем отношении шансов $\theta_{XY} = \theta_{XY(1)}$, для которого используют оценку Мантела–Хензела

$$\hat{\theta}_{MH} = \frac{\sum_s (n_{11s}n_{22s}/n_{++s})}{\sum_s (n_{21s}K_{12s}/n_{++s})}.$$

# 4. Обобщенные линейные модели для анализа категориальных данных

Несмотря на то, что накопление статистической информации в таблицу сопряженности происходит с использованием выборочного принципа, обобщение модели линейной регрессии является удобным и эффективным инструментом анализа категориальных данных. Предположим, что наблюдаются бинарная переменная $Y \in \{0,1\}$ и контролируемая исследователем переменная $X \in \{0,1\}$. Данная модель соответствует условному распределению в таблице сопряженности $2 \times 2$. Распределение $Y$ при каждом фиксированном значении $X$ характеризует единственный параметр $\mathbb{E}(Y|X) = p_X = \mathbb{P}(Y = 1|X) \in [0,1]$. Таким образом, задача изучения зависимости распределения наблюдаемой переменной $Y$ от контролируемой переменной $X$ согласуется с принципом регрессии, но классическая линейная модель в данном случае неприменима. В первую очередь, это связано с необходимостью нормальности распределения наблюдаемой переменной в классической модели, но и более слабое предположение линейности регрессии[2] $\mathbb{E}_\theta(Y|X) = \alpha + \beta X$, $\theta = (\alpha, \beta)$, не кажется естественным, т.к. при различных значениях $X$ области допустимых значений параметра $\theta$ будут различными. Для устранения данного недостатка можно выбрать строго монотонную функцию $g : [0,1] \to (-\infty, \infty)$, и вместо классической линейной регрессии использовать соотношение $g(\mathbb{E}_\theta(Y|X)) = \alpha + \beta X$. В частности, можно выбрать $g(u) = \mathrm{logit}(u) = \ln(u/(1-u))$. В этом случае, наблюдаемая переменная $Y$ при фиксированном значении $X$ имеет распределение Бернулли с параметром $p_X(\theta) = \left(1 + e^{-\alpha-\beta X}\right)^{-1}$, дисперсия которого равна $p_X(\alpha,\beta)(1 - p_X(\alpha,\beta))$. При наличии более двух допустимых значений $X$ имеет смысл использовать обобщенную модель дисперсионного анализа, поскольку $X$ категориальная переменная. Таблицы сопряженности с произвольными числами уровней переменных $X$ и $Y$, а также таблицы сопряженности большей размерности можно анализировать с использованием пуассоновской обобщенной линейной модели.

## 4.1. Обобщенные линейные модели

В основе обобщенной линейной модели лежит регрессионное соотношение

$$g(\mathbb{E}_\theta(Y|\boldsymbol{X})) = \boldsymbol{X}^T\boldsymbol{\beta},$$

[2] Рассматривается простая регрессия

где строго монотонная функция $g(\mu)$ с областью значений $\mathbb{R}$ называется *функцией связи* (link) и считается известной. Область определения функции связи совпадает с множеством допустимых значений $\mathbb{E}_\theta(Y|z)$. Целесообразно выбирать функцию связи таким образом, чтобы каждому вещественному значению $g$ соответствовало некоторое значение из области определения. Подразумевается, что распределение $Y$ при условии $z$ принадлежит некоторому параметрическому семейству распределений. Для задания распределения наблюдаемой переменной $Y$ обычно используют экспоненциальные семейства с дискретными плотностями

$$f^*(y;\theta,\phi) = \exp\big((y\theta - b(\theta))/a(\phi) + c(y;\phi)\big).$$

Статистические данные представляют собой набор $(Y_1, z_1), \ldots, (Y_n, z_n)$. Наиболее часто $Y_1, \ldots Y_n$ считаются независимыми, а $z_1, \ldots, z_n$ фиксируются. К экспоненциальным семействам дискретных распределений относятся биномиальные распределения $\mathrm{Bi}(m,p)$ $(p \in (0,1))$ с дискретными плотностями вида

$$f^*(y;a,b) = \frac{m!}{y!(m-y)!}p^y(1-p)^{m-y},\ y = 0,\ldots,m;$$

семейство распределений Пуассона $\mathrm{Pois}(\lambda)$ $(\lambda > 0)$, плотности распределения которых задаются соотношениями

$$f^*(y;a,b) = \frac{\lambda^y}{y!}e^{-\lambda},\ y = 0,1,\ldots;$$

семейство отрицательных биномиальных распределений $\mathrm{Nb}(m,p)$ $(p \in (0,1))$ с дискретными плотностями

$$f^*(y;p) = \frac{(m+y-1)!}{y!(m-1)!}p^y(1-p)^m,\ y = 0,1,\ldots.$$

Отметим, что семейство мультиномиальных распределений $\mathrm{Mult}(p_1,\ldots,p_k;m)$ $(p_i \in [0,1], \sum_i p_i = 1)$ с дискретными плотностями

$$f^*(y_1,\ldots,y_k;p,b) = \frac{m!}{y_1!\cdots y_k!}p_1^{y_1}\cdots p_k^{y_k},\ y_i = 0,1,\ldots;\sum_i y_i = m.$$

является многопараметрическим экспоненциальным семейством.

## 4.2. Обобщенные линейные модели на экспоненциальных семействах

Для экспоненциальных семейств с плотностями (дискретными плотностями)

$$f(y;\theta,\phi) = \exp\big((y\theta - b(\theta))/a(\phi) + c(y;\phi)\big)$$

в предположении независимости $Y_1,\dots,Y_n$ выпишем логарифм функции правдоподобия $LL(\boldsymbol{Y};\boldsymbol{\theta},\phi) = \sum_{i=1}^{n} \ln L_i$, где

$$\ln L_i = \ln L(Y_i;\theta_i;\phi) = (Y_i\theta_i - b(\theta_i))/a(\phi) + c(Y_i;\phi).$$

Значение $\theta$ будем называть каноническим параметром, а $\phi$ — параметром дисперсии. Предположим, что $b(\theta)$ — дважды дифференцируемая функция, $a(\phi) > 0$. Исследуем

$$\ln L_i = \ln L(Y_i;\theta_i;\phi) = (Y_i\theta_i - b(\theta_i))/a(\phi) + c(Y_i;\phi).$$

Производные $\ln L_i$ по $\theta_i$ равны:

$$\frac{\partial \ln L_i}{\partial \theta_i} = (Y_i - b'(\theta_i))/a(\phi), \quad \frac{\partial^2 \ln L_i}{\partial \theta_i^2} = -b''(\theta_i)/a(\phi).$$

Исходя из условия $\mathbb{E}_\theta\Big(\frac{\partial \ln L_i}{\partial \theta_i}\Big) = 0$ заключаем, что

$$\mu_i = \mathbb{E}(Y_i) = \mathbb{E}(Y\,|\,\theta_i,\phi) = b'(\theta_i).$$

С использованием равенства

$$\mathbb{E}\Big(\frac{\partial^2 \ln L_i}{\partial \theta_i^2}\Big) = -\mathbb{E}\left(\frac{\partial \ln L_i}{\partial \theta_i}\right)^2,$$

получаем, что

$$\mathbb{D}(Y_i) = \mathbb{D}(Y|\theta_i,\phi) = b''(\theta_i)\,a(\phi).$$

Рассмотрим обобщенную линейную модель

$$g(\mu(\boldsymbol{X}_i)) = g(\mu_i) = \eta_i = \sum\nolimits_{s=1}^{m} x_{si}\beta_s$$

с матрицей регрессоров $\mathbf{X} = (\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)^T$, где $\mu(\boldsymbol{X}) = \mathbb{E}_\theta(Y|\boldsymbol{X})$. Функцию связи $g$ будем называть *канонической*, если $\eta_i = \theta_i$. Максимум

функции правдоподобия по $\boldsymbol{\beta}$ находится из системы уравнений

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_l} = \sum\nolimits_{i=1}^{n} \frac{\partial LL_i(\theta_i, \phi)}{\partial \beta_l} = \sum\nolimits_{i=1}^{n} \frac{\partial LL_i(\theta_i, \phi)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_l} = 0,$$

где $LL_i(\theta, \phi) = \ln L(Y_i; \theta, \phi)$,

$$\frac{\partial LL_i(\theta_i, \phi)}{\partial \theta_i} = \frac{Y_i - b'(\theta_i)}{a(\phi)} = \frac{Y_i - \mu_i}{a(\phi)}, \quad \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\mathbb{D}(Y_i)}{a(\phi)}, \quad \frac{\partial \eta_i}{\partial \beta_l} = x_{il}$$

и $\frac{\partial \mu_i}{\partial \eta_i} = (g^{-1}(\eta_i))'$. Получаем, что

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{il}}{\mathbb{D}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{il}}{\mathbb{D}(Y_i)} (g^{-1}(\eta_i))' = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{il}}{\mathbb{D}(Y_i) g'(\mu_i)} = 0,$$

$l = 1, \ldots, s$. Отметим, что левые части данной системы уравнений зависят от параметра $\boldsymbol{\beta}$ только через $\boldsymbol{\mu}$. В общем случае для решения системы используются численные методы. В предположении, что дисперсии величин $Y_i$ пропорциональны при различных значениях параметра $\boldsymbol{\beta}$, данную оценку можно понимать как оценку по методу наименьших квадратов с весами, обратно пропорциональными дисперсиям наблюдений.

При выполнении условий регулярности оценка максимального правдоподобия $\hat{\boldsymbol{\beta}}$ асимптотически нормальна

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \Rightarrow \mathcal{N}\big(0, \bar{\mathbb{I}}^{-1}(\boldsymbol{\beta})\big),$$

где $\bar{\mathbb{I}}(\boldsymbol{\beta}) = \lim_{n\to\infty} \mathbb{I}(\boldsymbol{\beta})/n$ и $\mathbb{I}(\boldsymbol{\beta})$ – матрица информации Фишера,

$$\mathbb{I}(\boldsymbol{\beta}) = -\Big\|\mathbb{E}\Big(\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}\Big)\Big\|_{i,j=1}^{m} = \Big\|\mathbb{E}\Big(\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_i} \frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_j}\Big)\Big\|_{i,j=1}^{m}$$

В случае экспоненциального семейства $\mathbb{I}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$, где $\mathbf{W} = \mathbf{W}(\boldsymbol{\beta}, \phi)$ — диагональная матрица с диагональными элементами $1/(\mathbb{D}(Y_r)g'(\mu_r)^2)$, $r = 1, \ldots, m$. Оценка $\hat{\mathbb{I}} = \mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$, где $\hat{\mathbf{W}} = \mathbf{W}(\hat{\boldsymbol{\beta}}, \hat{\phi})$, получается подстановкой оценки параметра в формулу для вычисления матрицы информации Фишера. В случае канонической функции связи информационная матрица не зависит от наблюдений $Y_1, \ldots, Y_n$ (неслучайна)

$$\mathbb{I}(\beta) = \Big\|\sum\nolimits_{r=1}^{n} \frac{x_{ri}x_{rj}}{a(\phi)g'(\mu_r)}\Big\|_{i,j=1}^{m} = \Big\|\sum\nolimits_{r=1}^{n} \frac{b''(\theta_r)x_{ri}x_{rj}}{a(\phi)}\Big\|_{i,j=1}^{m}$$

Перейдем к задаче доверительного оценивания. Пусть $\boldsymbol{\psi} = \mathbf{C}^T\boldsymbol{\beta}$ – функция параметра, $\mathbf{C}$ – $m \times q$ матрица ранга $q$. В качестве оценки $\boldsymbol{\psi}$ используем $\hat{\boldsymbol{\psi}} = \mathbf{C}^T\hat{\boldsymbol{\beta}}$. Асимптотическая нормальность $\hat{\boldsymbol{\beta}}$ влечет асимптотическую нормальность

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \Rightarrow \mathcal{N}(0, \Gamma_{\hat{\psi}})$$

где $\Gamma_{\hat{\psi}} = \mathbf{C}^T\bar{\mathbb{I}}^{-1}\mathbf{C}$ – предельная матрица ковариации $\hat{\boldsymbol{\psi}}$, $\hat{\Gamma}_{\hat{\psi}} = n\,\mathbf{C}^T(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{C}$ – оценка $\Gamma_{\hat{\psi}}$. Таким образом,

$$(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})^T\hat{\mathbf{B}}^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \Rightarrow \chi^2_q$$

где $\hat{\mathbf{B}} = \mathbf{C}^T(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{C}$. Получаем асимптотический доверительный эллипсоид

$$\left\{\boldsymbol{\psi} : (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})^T\hat{\mathbf{B}}^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \leq x_\alpha\right\},$$

где $x_\alpha$ – квантиль $\chi^2_q$-распределения порядка $1-\alpha$. В частности, при $q = 1$ получаем асимптотический доверительный интервал. Метод множественного оценивания Шеффе позволяет строить совместные асимптотические доверительные интервалы на базе доверительного эллипсоида.

Перейдем к рассмотрению задачи проверки статистических гипотез. Пусть $\boldsymbol{\psi} = \mathbf{C}^T\boldsymbol{\beta}$ – функция параметра, $\mathbf{C}$ – $m \times q$ матрица ранга $q$. Поставим задачу проверки значимости статистической гипотезы $H_0 : \boldsymbol{\psi} = 0$. Для проверки данной гипотезы можно использовать критерий типа Вальда, статистика которого

$$Z = \hat{\boldsymbol{\psi}}^T\hat{\mathbf{B}}^{-1}\hat{\boldsymbol{\psi}}$$

имеет асимптотическое распределение $\chi^2_q$ при основной гипотезе. Таким образом, $P$-значение данного критерия вычисляется по формуле $PV = 1 - K_q(Z)$, где $K_q$ – функция распределений $\chi^2_q$. Альтернативно, можно использовать асимптотически эквивалентный критерий отношения правдоподобия, основанный на статистике

$$G = 2LL(Y; \hat{\boldsymbol{\theta}}, \hat{\phi}) - 2LL(Y; \hat{\boldsymbol{\theta}}_H, \hat{\phi}_H),$$

где $\hat{\boldsymbol{\theta}}_H, \hat{\phi}_H$ – ОМП при ограничении $\mathbf{C}^T\boldsymbol{\beta} = 0$. Асимптотическое распределения статистики $G$ при основной гипотезе снова $\chi^2_q$, откуда получаем

формулу для вычисления $P$-значения $PV = 1 - K_q(G)$. В отличие от классической модели $Z \neq G$ в общем случае.

**Бинарное распределение наблюдаемого признака.** Формирование обобщенной линейной модели в случае бинарной наблюдаемой переменной, принимающей значения 0 и 1, обсуждалось ранее. В этом случае, наблюдаемая переменная имеет распределение Бернулли с параметром $p = p_z = \mathbb{P}(Y = 1|z)$, зависящем от ковариаты $z$. Элемент функции правдоподобия допускает запись в экспоненциальной форме $L(y;p) = p^y(1-p)^{1-y} = \exp(y\ln(p/(1-p)) - \ln(1-p))$, $y = 0, 1$, поэтому данное распределение укладывается в рамки общей теории обобщенных линейных моделей для экспоненциальных семейств с каноническим параметром $\theta = \ln(p/(1-p))$. Таким образом, $\mathbb{E}_p Y = p$; $\mathbb{D}_p Y = p(1-p)$, и каноническая функция связи будет иметь вид $g(\mu) = \mathrm{logit}(\mu) = \ln(\mu/(1-\mu))$. Компонент информационной матрицы при использовании канонической функции связи будет иметь вид $\mathbf{W} = \|w_{ij}\|_{i,j}$: $w_{ij} = p_i(1-p_i)\mathbb{1}_{\{i=j\}}$. Данная модель носит название *логистической регрессии.*

В качестве функции связи можно использовать любую строго монотонную непрерывную функцию $g : [0,1] \to \mathbb{R}$. Помимо модели логистической регрессии часто используют модель с функцией связи $g(u) = \mathrm{probit}(u) = \Phi^{-1}(u)$, $u \in [0,1]$, где $\Phi(v)$ – функция стандартного нормального распределения (Лапласа). Данная функция связи уже не будет канонической, поэтому уравнения для получения оценок и компонент ковариационной матрицы будут иметь более сложный вид.

**Распределение наблюдаемого признака с целыми неотрицательными значениями.** Наиболее часто в этом случае используют распределение Пуассона ($\mathrm{Pois}(\lambda)$). Элемент функции правдоподобия по выборке из распределения Пуассона $L(y;\lambda) = \lambda^y e^{-\lambda}/y! = \exp(y\ln(\lambda) - \lambda - \ln(y!))$, $y = 0, 1, \ldots$, зависит от параметра $\lambda$, принимающего неотрцательные значения. Очевидно, что $\theta = \ln(\lambda)$ – канонический параметр, поэтому в качестве функции связи целесообразно использовать строго монотонную непрерывную функцию на $(0,\infty)$, отображающую данный интервал в множество $\mathbb{R}$. Известное свойство распределения Пуассона – равенство математического ожидания и дисперсии: $\mathbb{E}_\lambda Y = \lambda$; $\mathbb{D}_\lambda Y = \lambda$. Каноническая функция связи для пуассоновской модели равна $g(\mu) = \ln(\mu)$. В случае использования канонической функции связи, компоненты информационной матрицы

$\mathbf{W} = \|w_{ij}\|_{i,j}$ задаются соотношениями $w_{ij} = \lambda_i \mathbb{I}_{\{i=j\}}$. Довольно часто отмеченное ранее свойство равенства математического ожидания и дисперсии нарушается, и по реальным данным наблюдается существенное превышение дисперсией математического ожидания (overdispersion (Англ.)). Чтобы предусмотреть такую возможность обычно используют обобщенное распределение Пуассона (Conway–Maxwell-Poisson). Применение данной модели обуславливается наличием избыточной или недостаточной дисперсии. Элемент функции правдоподобия для данного распределения записывается в виде $L(y;\lambda,\nu) = \frac{\lambda^y}{C(\lambda,\nu)(y!)^\nu} = \exp(y\ln\lambda - \ln C(\lambda,\nu) - \nu\ln(y!))$, где $C(\lambda,\nu) = \sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^\nu}$, $y = 0,1,\ldots$. Распределение Пуассона $\mathrm{Pois}(\lambda)$ получается, если выбрать значение параметра $\nu = 1$. Также отметим, что при $\nu\to\infty$ получается распределение Бернулли с параметром $p=\lambda/(1+\lambda)$, а при $\nu \to 0_+$ ($\lambda < 1$) получается геометрическое распределение с параметром $p = 1-\lambda$. Математическое ожидание и дисперсия обобщенного распределения Пуассона равны $\mathbb{E}_p Y = \lambda\frac{C'_\lambda(\lambda,\nu)}{C(\lambda,\nu)}$ и $\mathbb{D}_p Y = \lambda\frac{C'_\lambda(\lambda,\nu)}{C(\lambda,\nu)} + \lambda^2\Big(\frac{C''_{\lambda\lambda}(\lambda,\nu)}{C(\lambda,\nu)} - \frac{C'_\lambda(\lambda,\nu)^2}{C(\lambda,\nu)^2}\Big)$ соответственно, где $C'_\lambda(\lambda,\nu) = \sum_{j=1}^{\infty}\frac{j\lambda^{j-1}}{(j!)^\nu}$ и $C''_{\lambda\lambda}(\lambda,\nu) = \sum_{j=2}^{\infty}\frac{j(j-1)\lambda^{j-2}}{(j!)^\nu}$. В качестве функции связи обычно используют $g(\mu) = \ln\mu$, а компоненты информационной матрицы в этом случае $\mathbf{W} = \|w_{ij}\|_{i,j}$ определяются соотношениями

$$w_{ij} = \frac{\lambda C'_\lambda(\lambda,\nu)}{C(\lambda,\nu)C'_\lambda\big(\lambda,\nu) + \lambda\big(C''_{\lambda\lambda}(\lambda,\nu)C(\lambda,\nu) + C'_\lambda(\lambda,\nu)^2\big)}\mathbb{I}_{\{i=j\}}.$$

Другое классическое распределение с целыми неотрицательными значениями – геометрическое $\mathrm{Geom}(p)$, $p \in (0,1)$. Модель геометрического распределения используется довольно редко. Элемент функции правдоподобия представим в виде $L(y;\lambda) = p^y(1-p) = \exp(y\ln p + \ln(1-p))$, $y = 0,1,\ldots$. Математическое ожидание и дисперсия геометрического распределения вычисляются по формулам: $\mathbb{E}_p Y = p/(1-p)$; $\mathbb{D}_p Y = p/(1-p)^2$. Очевидно, что канонический параметр равен $\theta = \ln p \in (-\infty, 0)$, а каноническая функция связи крайне неудобна в данном случае. В качестве функции связи обычно используют $g(u) = \mathrm{logit}(u)$ как и в случае распределения Бернулли, поскольку множество значений параметра – $[0,1]$. При этом, формулы для вычисления компонент информационной матрицы $\mathbf{W} = \|w_{ij}\|_{i,j}$: $w_{ij} = p_i\mathbb{I}_{\{i=j\}}$ довольно несложные.

## 4.3. Обобщенные модели дисперсионного анализа

Термин «дисперсионный анализ» обычно используется в предположении, что ковариаты представляют собой категориальные величины, которые в дальнейшем будут называться факторами.

**Однофакторная дисперсионная модель или простая группировка.** При наличии одной ковариаты каждое наблюдение $(Y, z)$ включает в себя значение исследуемой характеристики $Y$ и фактора группировки $z$, имеющего $d$ уровней. Распределение наблюдаемой величины $Y$ определяется значением фактора группировки $z$. Обобщенная модель однофакторного анализа с функцией связи $g$ определяется соотношениями

$$g(\mathbb{E}_\theta(Y|z=i)) = \eta_j, \quad i = 1, \dots, d,$$

а наблюдаемые $Y_1, \dots, Y_d$ – независимые величины. Применяя покомпонентно функцию $g$ получаем обобщенную линейную модель для всех имеющихся наблюдений, которая в матричной форме может быть записана следующим образом:

$$g(\mathbb{E}_\theta(\boldsymbol{Y}|\boldsymbol{z})) = \mathbf{X}^T\boldsymbol{\eta}$$

где $\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 1 & 1 & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 \\ \underbrace{\phantom{00}}_{z=1} & & \underbrace{\phantom{000}}_{z=2} & & & & \underbrace{\phantom{0}}_{z=d} \end{pmatrix}$.

Поскольку речь идет о простой группировке наблюдений с одинаковыми распределениями по значениям ковариаты $z$, а порядок появления наблюдений с различными значениями $z$ не играет роли, то наблюдения иногда удобно индексировать двумя индексами: $Y_{ij}$ – $j$-е наблюдение $i$-й группы, $j = 1, \dots, n_i$, $i = 1, \dots, d$. В этом случае, получаем соотношения $g(\mathbb{E}_\theta Y_{ij}) = \eta_i$, $j = 1, \dots, n_i$, $i = 1, \dots, d$. Для построения статистических выводов дополнительно предполагается, что $Y_{ij}$ имеет распределение, принадлежащее некоторому параметрическому семейству с параметром $\theta_i$ при каждом $j = 1, \dots, n_i$, $i = 1, \dots, d$. Параметр $\eta_i$ может быть оценен только при наличии хотя бы одного наблюдения в группе ($n_i \geq 1$). При наличии пустых групп соответствующие параметры следует исключить из модели.

Сравнением $\psi$ параметров $\eta_1, \dots, \eta_d$ называется их линейная комбинация $\psi = \sum_{i=1}^d c_i\eta_i$, где $\sum_{i=1}^d c_i = 0$. В духе дисперсионного анализа

представление преобразованного среднего значения по группе

$$\eta_i = \mu + \alpha_i$$

в виде суммы взвешенного среднего $\mu = \eta_* = \sum_{i=1}^{d} v_i \eta_i$ и главных эффектов $\alpha_i = \eta_i - \eta_*$, $i = 1, \ldots, d$, по отношению к системе весов $\{v_i\}_{i=1}^{d}$: $v_i \geq 0$ и $\sum_{i=1}^{d} v_i = 1$. Очевидно, что $\alpha_* = \sum_{i=1}^{d} v_i \alpha_i = 0$. Стандартный для обобщенной модели выбор весов $v_{i_0} = 1$, $v_i = 0$ при $i \neq 1$, подразумевающий отсчет главных эффектов от выбранного базового уровня $i_0$, позволяет не накладывать дополнительных ограничений на главные эффекты, но при этом главный эффект уровня $i_0$ принимается равным нулю.

Главной задачей однофакторного дисперсионного анализа является проверка однородности групп

$$H_0 : \eta_1 = \ldots = \eta_d.$$

Данная гипотеза допускает представление в терминах главных эффектов

$$H_0 : \alpha_1 = \ldots = \alpha_{d-1} (= \alpha_d) = 0.$$

Отметим, что главные эффекты $\alpha_1, \ldots, \alpha_d$ являются сравнениями параметров группировки $\eta_1, \ldots, \eta_d$. В общем случае, гипотезу $H_0$ можно записать приравняв к нулю любые $d - 1$ линейно независимых сравнений $H_0 \Leftrightarrow H_{d-1} : \psi_1 = \ldots = \psi_{d-1} = 0$. Возможно рассмотрение более слабых гипотез

$$H_q : \psi_1 = \ldots = \psi_q = 0,$$

которые при $q < d - 1$, в отличие от $H_0$, вообще говоря будут зависеть от выбора весов. Для проверки гипотезы обычно используют критерий отношения правдоподобия, основанный на $G$-статистике

$$G = -2(LL_0(\boldsymbol{Y}; z, \theta) - LL(\boldsymbol{Y}; z, \theta))$$

где $LL_0$ – наименьшее значение логарифма функции правдоподобия при основной гипотезе, а $LL$ – наименьшее значение логарифма функции правдоподобия в общих предположениях. Еще один критерий проверки данной гипотезы использует $Z$-статистику типа Вальда,

$$Z = \hat{\boldsymbol{\psi}}^T \hat{\Gamma}_\psi \hat{\boldsymbol{\psi}},$$

где $\hat{\boldsymbol{\psi}} = (\hat{\psi}_1, \ldots, \hat{\psi}_{d-1})^T$ – набор оценок максимального правдоподобия линейно независимых сравнений параметров $\eta_1, \ldots, \eta_d$, а $\hat{\Gamma}_\psi$ – оценка предельной матрицы ковариации вектора $\boldsymbol{\psi}$. Обе статистики имеют предельные распределения $\chi^2_q$, поэтому для вычисления $P$-значений соответствующих асимптотических критериев используются формулы $PV = 1 - K_q(G)$ и $PV = 1 - K_q(Z)$ соответственно.

Для уточнения результатов проверки значимости $H_0 : \psi_1 = \ldots = \psi_{d-1} = 0$ можно использовать метод множественного оценивания Шеффе, позволяющий получить совместные асимптотические доверительные интервалы всех $\psi_i$ и их линейных комбинаций $\psi = \alpha_1\psi_1 + \ldots + \alpha_{d-1}\psi_{d-1}$

$$[\hat{\psi} - \sqrt{x_\alpha(d-1)}\hat{\sigma}_\psi, \hat{\psi} + \sqrt{x_\alpha(d-1)}\hat{\sigma}_\psi],$$

где $x_\alpha$: $K_{d-1}(x_\alpha) = \alpha$ – квантиль распределения $\chi^2_{d-1}$, а $\hat{\sigma}_\psi$ – оценка дисперсии линейной функции параметров $\psi$. Отметим, что $Z$-критерий принимает гипотезу в том, и только в том случае, если доверительный интервал для каждого $\psi$ содержит 0. Метод Шеффе позволяет выявить сравнения, ответственные за отвержение гипотезы в случае ее отвержения. В частности, с использованием метода Шеффе можно проверять односторонние гипотезы. Например, для проверки гипотезы

$$H_0 : \eta_1 < \ldots < \eta_d$$

следует построить совместные доверительные интервалы для сравнений $\psi_i = \eta_{i+1} - \eta_i$, $i = 1, \ldots, n$. Если все доверительные интервалы полностью окажутся в положительной области, то гипотезу принимают.

**Модель двухфакторного анализа.** При наличии двух факторов, влияющих на результат, каждое наблюдение можно представить в виде $(Y, \boldsymbol{z})$, где $Y$ – наблюдаемая переменная, а $\boldsymbol{z} = (z_1, z_2)$ – ковариата, где $z_1 \in \{1, \ldots, d_1\}$ и $z_2 \in \{1, \ldots, d_2\}$ – значения факторов. Статистические данные $(\boldsymbol{Y}, \boldsymbol{z})$ включают в себя значения наблюдаемых переменных $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ и набор значений соответствующих ковариат $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$. Исходная модель может быть представлена в виде простой группировки

$$g(\mathbb{E}_\theta(Y|z_1 = i, z_2 = j)) = \eta_{ij}, \quad i = 1, \ldots, d_1,\ j = 1, \ldots, d_2,$$

но чтобы разделить влияние факторов на результат используют параметризацию

$$\eta_{ij} = \mu + \alpha_i^{(1)} + \alpha_j^{(2)} + \alpha_{ij}^{(12)},$$

где $\mu = \eta_{**} = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} v_i w_j \eta_{ij}$ взвешенное среднее по отношению к наборам весов $\{v_i\}_{i=1}^{d_1} : \sum_{i=1}^{d_1} v_i = 1$, $\{w_j\}_{j=1}^{d_2} : \sum_{j=1}^{d_2} w_j = 1$; $\alpha_i^{(1)} = \eta_{i*} - \eta_{**} = \sum_{j=1}^{d_2} w_j \eta_{ij} - \eta_{**}$ и $\alpha_j^{(2)} = \eta_{*j} - \eta_{**} = \sum_{i=1}^{d_1} v_i \eta_{ij} - \eta_{**}$ – главные эффекты первого и второго факторов соответственно, и $\alpha_{ij}^{(12)} = \eta_{ij} - \eta_{i*} - \eta_{*j} + \eta_{**}$ – взаимодействия факторов. Из определения очевидно, что главные эффекты удовлетворяют соотношениям $\alpha_*^{(1)} = \sum_{i=1}^{d_1} v_i \alpha_i^{(1)} = 0$; $\alpha_*^{(2)} = \sum_{j=1}^{d_2} w_j \alpha_j^{(2)} = 0$, а взаимодействия удовлетворяют соотношениям $\alpha_{i*}^{(12)} = 0$ при всех $i = 1, \ldots, d_1$; $\alpha_{*j}^{(12)} = 0$ при всех $j = 1, \ldots, d_2$. Отметим, что главные эффекты и взаимодействия являются сравнениями параметров $\eta_{ij}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$. Как и в однофакторном анализе, наиболее удобным считается выбор весов

$$v_i = \begin{cases} 1, \ i = i_0, \\ 0, \ \text{в остальных случаях}, \end{cases} \qquad \text{и} \quad w_j = \begin{cases} 1, \ j = j_0, \\ 0, \ \text{в остальных случаях}, \end{cases}$$

при котором главные эффекты и взаимодействия отсчитываются от базовых уровней $i_0$ и $j_0$ первого и второго факторов соответственно. В этом случае главные эффекты и взаимодействия, соответствующие базовым уровням факторов равны нулю, а на остальные параметры никаких ограничений не накладываются (т.е. соответствующие сравнения параметров группировки линейно независимы).

Возможны различные варианты влияния факторов на результат (см. Рис. 1). Гипотеза отсутствия взаимодействий определяется соотношениями

$$H_{(12)} : \alpha_{ij}^{(12)} = 0, i = 1, \ldots, d_1, j = 1, \ldots, d_2$$

и не зависит от выбора весов. При выполнении $H_{(12)}$ получаем аддитивную модель

$$\eta_{ij} = \mu + \alpha_i^{(1)} + \alpha_j^{(2)}, \ i = 1, \ldots, d_1, j = 1, \ldots, d_2.$$

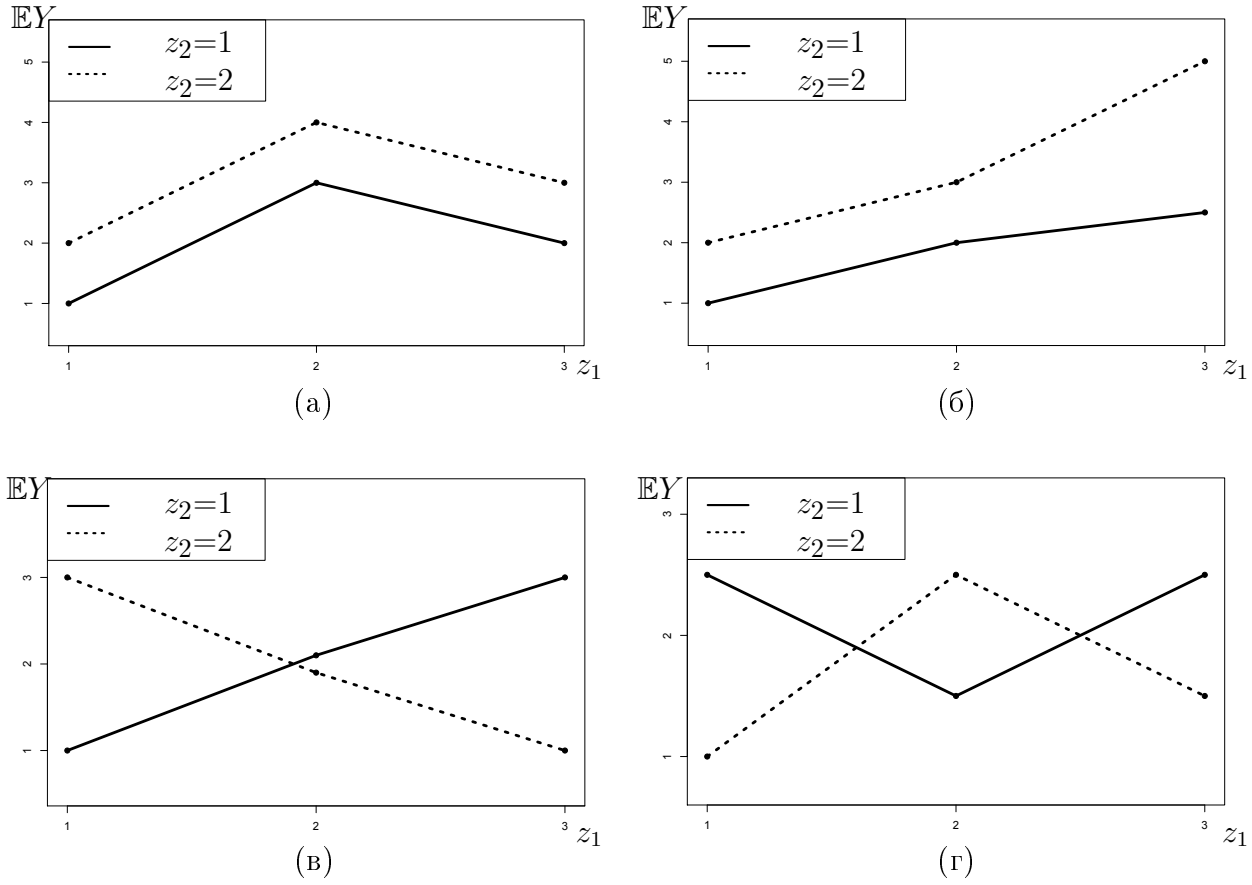


Рис. 1. *(а) Аддитивная модель; (б) Согласованное действие; (в) Эффект пересечения 1; (г) Эффект пересечения 2.*

Размерность параметра общей модели группировки равна $d_1 d_2$, тогда как размерность параметра аддитивной модели равна $d_1 + d_2 - 1$. Таким образом, размерность параметра взаимодействий равна $d_1 d_2 - (d_1 + d_2 - 1) = (d_1 - 1)(d_2 - 1)$, а гипотеза $H_{(12)}$ может быть переписана в терминах $(d_1 - 1)(d_2 - 1)$ линейно независимых взаимодействий. Поскольку взаимодействия являются сравнениями параметра $\boldsymbol{\eta}$, для проверки гипотезы отсутствия взаимодействий можно использовать $Z$-критерий типа Вальда, построенный на базе $(d_1-1)(d_2-1)$ линейно независимых взаимодействий или $G$-критерий отношения правдоподобия.

Гипотезе отсутствия влияния фактора $z_2$ на результат

$$H_{(2)} : \alpha_j^{(2)} = 0, \alpha_{ij}^{(12)} = 0, i = 1, \ldots, d_1, j = 1, \ldots, d_2$$

соответствует модель с одним фактором

$$\eta_{ij} = \mu + \alpha_i^{(1)}, \ i = 1, \ldots, d_1, j = 1, \ldots, d_2.$$

Размерность параметра данной модели равна $d_1$, поэтому для записи гипотезы $H_{(2)}$ потребуются $d_1 d_2 - d_1 = d_1(d_2 - 1)$ линейно независимых сравнений, среди которых $(d_1 - 1)(d_2 - 1)$ линейно независимых взаимодействий и $d_1 - 1$ линейно независимых главных эффекта второго фактора. Для проверки гипотезы $H_{(2)}$ снова можно использовать $Z$-критерий типа Вальда или $G$-критерий отношения правдоподобия, статистики которых при справедливости $H_{(2)}$ будут иметь предельные распределения $\chi^2_{d_1(d_2-1)}$. Аналогичным образом формулируются и проверяются гипотезы отсутствия влияния фактора $z_1$ и обоих факторов на результат. Отметим, что рассмотренные ранее гипотезы не зависят от выбора весов, поэтому их можно считать объективными. Возможна формулировка и построение критериев проверки иных гипотез, но в большинстве случаев гипотеза будет зависеть от выбора весов. Например, гипотеза

$$H_{(2*)} : \alpha_j^{(2)} = 0, j = 1, \ldots, d_2$$

при наличии ненулевых взаимодействий зависит от выбора весов и не является объективной. Отметим, что справедливость данной гипотезы при некотором выборе весов характерна при наличии эффекта пересечения факторов.

**Многофакторный анализ.** При наличии нескольких факторов, влияющих на результат, каждое наблюдение представляется в виде $(Y, \boldsymbol{z})$ как и в двухфакторном случае, где ковариата $\boldsymbol{z} = (z_1, \ldots, z_r)$ включает в себя $r$ компонент, а статистические данные представляют собой конечный набор независимых наблюдений $(\boldsymbol{Y}, \boldsymbol{z})$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ – независимые величины и $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)$ – соответствующие наборы значений ковариат. Обобщенная модель может быть представлена в виде простой группировки

$$g(\mathbb{E}_\theta(Y|z_1 = i_1, \ldots, z_r = i_r)) = \eta_{i_1,\ldots,i_r}, \quad i_s = 1, \ldots, d_s,\ s = 1, \ldots, r.$$

В целом, классификация по нескольким факторам аналогична двухфакторной. Для каждого фактора выбираются веса, исходя из которых вычисляются взвешенное среднее, главные эффекты факторов и взаимодействия,

но взаимодействия факторов имеют более сложную структуру. Помимо попарных взаимодействий $\alpha_{i,i'}^{(j,j')}$, $i = 1, \ldots, d_j$, $i' = 1, \ldots, d_{j'}$, для каждой пары факторов $(j, j')$ появляются взаимодействия всевозможных трех и большего числа факторов вплоть до взаимодействий всех факторов $\alpha_{i_1,\ldots,i_r}^{(j_1,\ldots,j_r)}$, $i_s = 1, \ldots, d_{j_s}$, $s = 1, \ldots, r$. Формулы, связывающие параметры группировки с взвешенным средним, главными эффектами и взаимодействиями строятся рекурсивно от взаимодействий меньших порядков к взаимодействиям больших порядков. В частности, взаимодействия пар факторов вычисляются аналогично двухфакторному случаю с дополнительным суммированием с весами по всем остальным индексам. Все главные эффекты и взаимодействия являются сравнениями параметров группировки. Вообще говоря, на главные эффекты и взаимодействия накладываются линейные ограничения: взвешенные суммы взаимодействий по каждому индексу при любом фиксированном наборе остальных индексов равны нулю. Линейных ограничений на параметры модели можно избежать, если отсчитывать главные эффекты и взаимодействия от базовых уровней. При этом, главные эффекты и взаимодействия, соответствующие базовым уровням, принимаются равными нулю.

При формулировке гипотез многофакторного анализа те или иные главные эффекты и взаимодействия приравнивают к нулю, начиная с взаимодействий более высоких порядков. Отметим, что гипотезы о взаимодействиях более низкого порядка при наличии взаимодействий более высоких порядков, включающих те же факторы, не являются объективными и зависят от выбора весов. Кроме того, возможность статистического исследования довольно часто ограничена тем, что оценивание параметров группировки возможно только если все комбинации уровней факторов присутствуют в исходном наборе данных. С другой стороны, отсутствие тех или иных комбинаций факторов в исходном наборе данных не исключает возможности статистического анализа при выборе более частной модели. Наконец, взаимодействия большого числа факторов плохо поддаются интерпретации. С учетом этих аргументов, довольно часто предполагают априори отсутствие взаимодействий большого числа факторов. Например, ограничиваются только взаимодействиями не более двух факторов. Для проверки выдвигаемых гипотез обычно используется $Z$-критерий типа Вальда, построенный на сравнениях параметров группировки, использующихся при формулировке гипотезы, или $G$-критерий отношения правдоподобия. Число степеней свободы предельного распределения совпадает

с числом линейно независимых сравнений, участвующих в формулировке гипотезы. Его удобно вычислять, если считать главные эффекты и взаимодействия по отношению к базовым уровням факторов.

## 4.4. Выбор оптимальной модели

При большом количестве признаков, выбор обобщенной линейной модели для проведения дальнейшего количественного исследования – одна из ключевых задач. По сути, выбирается правая часть регрессионного соотношения, тогда как функция связи изначально известна. Излишнее упрощение модели может не позволить достигнуть целей исследования, а выбор слишком подробной модели может привести к недостатку статистических данных, что не позволит делать статистические выводы, а также могут возникнуть сложности с интерпретацией результатов анализа. При наличии наиболее общей модели (saturated) вложенная модель может быть записана в виде гипотезы $H : \mathbf{C}^T\boldsymbol{\beta} = 0$, где $\boldsymbol{\beta}$ – параметр общей модели. Из совокупности допустимых моделей $H_1, \ldots, H_k$ требуется выбрать оптимальную, исходя из имеющегося набора статистических данных и целей исследования. Для выбора оптимальной модели можно использовать алгоритм последовательного исключения наименее значимых или последовательного включения наиболее значимых параметров модели. В частности, в модели многофакторного дисперсионного анализа алгоритм последовательного исключения подразумевает выдвижение и проверку объективных гипотез о взаимодействиях, начиная с гипотез отсутствия взаимодействий большего числа факторов к гипотезам об отсутствии взаимодействий меньшего числа факторов и главных эффектов. Принятие гипотезы при выбранном уровне значимости влечет исключение соответствующего параметра из модели. Процедура исключения останавливается, если все входящие в модель классы взаимодействий являются значимыми. При использовании алгоритма последовательного включения гипотезы формулируются и проверяются в обратном порядке от проверки значимости включения главных эффектов каждого фактора к взаимодействиям все более высоких порядков. В отличие от алгоритма исключения, на каждом используется своя общая модель – общей моделью считается модель с включенными дополнительными параметрами. В моделях многофакторного дисперсионного анализа часто ставится ограничение на порядок взаимодействий, поскольку взаимодействия высоких порядков не поддаются интерпретации. Оба алгоритма

основаны на изучении $p$-значений критериев проверки гипотез, являются достаточно затратными с вычислительной точки зрения, а результаты имеют большую долю субъективизма.

Для стандартизации процесса выбора оптимальной модели созданы так называемые информационные критерии. Решение принимается с использованием логарифма функции правдоподобия $LL(\boldsymbol{X},\theta) = \ln L(\boldsymbol{X},\theta)$, размерности параметра и количества статистической информации. Целевая функция информационного критерия имеет вид

$$IC_R(\boldsymbol{X};H) = -2\sup_H LL(\boldsymbol{X},\theta) + 2R(H;\boldsymbol{X}),$$

где $R(H;\boldsymbol{X})$ — неотрицательная функция пенализации, зависящая от размерности параметра в предположении $H$, от модели и количества статистической информации. Оптимальной считается модель $H$, для которой $IC_R$ принимает наименьшее значение. Наиболее часто используются критерии Акайке и Байесовский. Критерий *Акайке* (Акайке, 1973) $AIC(\boldsymbol{X};H) = IC_R(\boldsymbol{X};H)$ использует размерность параметра $\theta$ в предположении $H$ в качестве пенализации $R(H;\boldsymbol{X}) = \dim(\Theta_H)$. *Байесовский* информационный критерий (Шварц, 1978) $BIC(\boldsymbol{X};H) = IC_R(\boldsymbol{X};H)$ использует пенализацию $R(H;\boldsymbol{X}) = \dim(\Theta_H)\ln n/2$, которая определяется размерностью параметра $\theta$ в предположении $H$ и числом наблюдений. Байесовский критерий имеет более жесткую пенализацию на размерность параметра при больших $n$, чем критерий Акайке.

## 4.5. Использование обобщенных линейных моделей для анализа категориальных данных

Отметим, что задачи категориального анализа не ограничиваются только анализом таблиц сопряженности. Круг обобщенных линейных моделей, применяемых для анализа категориальных данных, гораздо шире. Ввиду особой роли анализа сопряженности признаков и необходимости связать классические методы с обобщенными линейными моделями начнем с описания именно этого типа статистического анализа. Для анализа сопряженности признаков обычно применяют модель логистической регрессии или пуассоновскую модель.

**Анализ сопряженности двух признаков.** Модель логистической регресии применима для анализа таблиц сопряженности признаков $(Y,X)$ размера $2\times d$. Использование модели логистической регрессии подразумевает изучение условного распределения наблюдаемой переменной $Y$ при

условии $X$, что, вообще говоря, не исключает использование данной модели для изучения некоторых свойств совместного распределения $(Y, X)$. Для удобства стандартизуем уровни признака $Y \in \{0, 1\}$. Поскольку речь идет об условных вероятностях $\pi_i = p_{1|j}$, $j = 1, \ldots, d$, то допускается контроль признака $X$ (независимая переменная). При фиксированном значении признака $X = i$ наблюдаемая переменная имеет распределение Бернулли с параметром $\pi_i$:

$$\text{logit}(\pi_i) = \alpha + \beta_i, \ i = 1, \ldots, d.$$

Правая часть данного соотношения представляет собой модель однофакторного дисперсионного анализа, где $\alpha$ – взвешенное среднее, $\beta_i$ – главные эффекты, удовлетворяющие соотношению $\beta_* = \sum_{i=1}^{d} v_i \beta_i = 0$, а $v_i$: $\sum_{i=1}^{d} v_i = 1$. Наиболее часто используют выбор весов $v_i = \mathbb{1}_{\{i=i_0\}}$, где $i_0$ – базовый уровень фактора $X$ (обычно наименьший или наибольший). Гипотеза однородности (эквивалентная гипотезе независимости признаков) формулируется в виде $H_0 : \beta_1 = \ldots = \beta_d = 0$. Для проверки гипотезы можно использовать критерий типа Вальда. Перепишем основную гипотезу $H_0 : \boldsymbol{\psi} = 0$, где $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{d-1})'$ – линейно независимые сравнения. Статистика критерия $Z = \hat{\boldsymbol{\psi}}' \hat{\Gamma}_{\psi}^{-1} \hat{\boldsymbol{\psi}}$, где $\hat{\Gamma}_{\psi}^{-1}$ – оценка матрицы ковариации сравнений $\boldsymbol{\psi}$, имеет предельное распределение $\chi^2_{d-1}$ при основной гипотезе, а $P$-значение вычисляется по формуле $PV = 1 - K_{d-1}(W)$, где $K_{d-1}$ – функция распределения $\chi^2_1$. Отдельно отметим, что $\beta_i - \beta_j$ – логарифм частного отношения шансов. Меры удаленности теоретического распределения от основной гипотезы можно формировать с использованием сравнений параметров $\beta_1, \ldots, \beta_d$. Критерий отношения правдоподобия, статистика которого имеет вид $G = 2(LL_S - LL_A)$ (deviance, (Англ.)), где $LL_S$ максимум логарифма функции правдоподобия в общей модели, $LL_A$ максимум логарифма функции правдоподобия в предположении $H_0$ (аддитивная модель), даже более распространен, чем критерий типа Вальда. Предельное распределение статистики $G$ при основной гипотезе – $\chi^2_{d-1}$, а $P$-значение вычисляется по формуле $PV = 1 - K_{d-1}(G)$

Аналогичным образом строятся обобщенные модели с другими непрерывными строго монотонно возрастающими функциями связи $g : (0, 1) \to \mathbb{R}$, например, $g(u) = \text{probit}(u) = \Phi^{-1}(u)$.

Лог-линейная модель применима для анализа таблиц сопряженности признаков $(Y, X)$ размера $d_1 \times d_2$ (общего вида). Предполагается, что значения в таблице сопряженности $n_{ij}$ независимы в совокупности и имеют пуассоновское распределение $\mathrm{Pois}(\lambda_{ij})$, где $\lambda_{ij} = \lambda p_{ij}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$ – параметры распределения. Обобщенная линейная модель записывается в виде

$$\ln \lambda_{ij} = \alpha + \beta_i^Y + \beta_j^X + \beta_{ij}^{YX},$$

где правая часть – модель двухфакторного дисперсионного анализа; $\alpha$ – взвешенное среднее, $\beta_i^Y$, $\beta_j^X$ – главные эффекты, удовлетворяющие соотношениям $\beta_*^Y = 0$, $\beta_*^X = 0$; $\beta^{YX}$ – взаимодействия с ограничениями $\beta_{i*}^{YX} = 0$, $\beta_{*j}^{YX} = 0$, где индекс $*$ означает суммирование с весами $\{v_i\}_{i=1}^{d_1}$ и $\{u_j\}_{j=1}^{d_2}$. Независимость факторов $X$ и $Y$ определяется отсутствием взаимодействий $H_0 : \beta_{ij}^{YX} = 0, i = 1, \ldots, d_1, j = 1, \ldots, d_2$ и не зависит от выбора весов. Соответствующая обобщенная линейная модель называется аддитивной

$$\ln \lambda_{ij} = \alpha + \beta_i^Y + \beta_j^X.$$

Покажем, что данная модель соответствует независимости компонент вектора $(X, Y)$. Действительно, в случае независимости $X$ и $Y$,

$$\mu_{ij} = tp_{ij} = tp_{i+}p_{+j}.$$

Переходя к логарифмам получаем, что

$$\ln\, \mu_{ij} = \ln(tp_{i+}) + \ln(tp_{+j}) - \ln\, t = \ln(\mu_{i+}) + \ln(\mu_{+j}) - \ln\, t.$$

Данные соотношения соответствуют аддитивной модели, если выбрать

$$\beta_i^Y = \ln(\mu_{i+}) - \sum\nolimits_i u_i \ln(\mu_{i+}), \beta_j^X = \ln(\mu_{+j}) - \sum\nolimits_j v_j \ln(\mu_{+j}),$$

$$\alpha = \sum\nolimits_i u_i \ln(\mu_{i+}) + \sum\nolimits_i v_j \ln(\mu_{+j}) - \ln\, t.$$

Обратно, исходя из аддитивной модели получаем, что

$$\ln\, \mu_{i+} = \ln\Big(\sum_{j=1}^{d_2} \exp(\alpha + \beta_i^Y + \beta_j^X)\Big) = \alpha + \beta_i^Y + \ln\Big(\sum_{j=1}^{d_2} \exp(\beta_j^X)\Big)$$

и

$$\ln\,\mu_{+j} = \ln\Big(\sum_{i=1}^{d_1}\exp(\alpha+\beta_i^Y+\beta_j^X)\Big) = \alpha+\beta_j^X+\ln\Big(\sum_{i=1}^{d_1}\exp(\beta_i^Y)\Big).$$

Отметим, что

$$t=\sum_{ij}\mu_{ij}=\sum_{ij}\exp\Big(\alpha+\beta_i^U+\beta_j^V\Big)=\exp(\alpha)\sum_{i=1}^{d_1}\exp(\beta_i^Y)\sum_{j=1}^{d_2}\exp(\beta_j^X),$$

а следовательно,

$$\alpha+\ln\Big(\textstyle\sum_{i=1}^{d_1}\exp(\beta_i^Y)\Big)+\ln\Big(\sum_{j=1}^{d_2}\exp(\beta_j^X)\Big)=\ln\,t.$$

Тогда

$$\ln\,p_{i+}p_{+j}=\ln\,\mu_{i+}+\ln\,\mu_{+j}-\ln\,t=2\alpha+\beta_i^Y+\beta_j^X+\ln\Big(\sum_{j=1}^{d_2}\exp(\beta_j^X)\Big)+$$

$$+\ln\Big(\textstyle\sum_{i=1}^{d_1}\exp(\beta_i^Y)\Big)=\alpha+\beta_i^Y+\beta_j^X-\ln\,t=\ln(\mu_{ij}/t)=\ln\,p_{ij}.$$

Таким образом, $p_{ij}=p_{i+}p_{+j}$ при всех $i=1,\ldots,d_1$, $j=1,\ldots,d_2$.

При отсчете главных эффектов и взаимодействий от базовых уровней $i_0$ и $j_0$ факторов $X$ и $Y$ соответственно, взаимодействия представляют собой логарифмы частных отношений шансов $\beta_{ij}^{YX}=\ln\big(p_{ij}p_{i_0j_0}/(p_{i_0j}p_{ij_0})\big)$, $i\neq i_0$, $j\neq j_0$. Для проверки гипотезы независимости $H_0$ можно использовать критерий типа Вальда. Для этого $H_0$ записывают с помощью $(d_1-1)(d_2-1)$ линейно-независимых сравнений $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_{(d_1-1)(d_2-1)})^T$, образующих базис линейного пространства взаимодействий. Статистика критерия типа Вальда $Z=\hat{\boldsymbol{\psi}}^T\hat{\Gamma}_{\psi}^{-1}\hat{\boldsymbol{\psi}}$ имеет предельное распределение $\chi^2_{(d_1-1)(d_2-1)}$, и ее значение не зависит от выбора сравнений $\boldsymbol{\psi}$ и от выбора весов. Для вычисления $P$-значения используют формулу $PV=1-K_{(d_1-1)(d_2-1)}(W)$. Методы множественного оценивания позволяют получать совместные доверительные интервалы для сравнений и использовать выводы для определения сравнений, ответственных за отвержение гипотезы. Наиболее часто для проверки гипотезы независимости $H_0$ используется критерий отношения правдоподобия, статистика которого $G=2(LL_S-LL_A)$ (deviance), где $LL_S$ максимум логарифма

правдоподобия в общей модели, а $LL_A$ максимум логарифма правдоподобия в аддитивной модели, имеет предельное распределение: $\chi^2_{(d_1-1)(d_2-1)}$, а $P$-значение вычисляется по формуле $PV = 1 - K_{(d_1-1)(d_2-1)}(G)$.

**Анализ сопряженности трех признаков.** При наличии трех признаков модель логистической регрессии применима для решения лишь ограниченного круга задач. В предположении, что $Y \in \{0, 1\}$ – бинарный признак, применение модели логистической регрессии ограничено условным распределением $Y$ при условии $(X, Z)$. Иными словами, можно предполагать, что $Y$ – наблюдаемый признак, а $X$ и $Z$ – контролируемые признаки. Круг задач по сути сводится к анализу зависимости $Y$ от $X$ и $Z$. Обобщенная линейная модель логистической регрессии может быть записана в виде

$$\text{logit}(\pi_{ij}) = \alpha + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ},\ i = 1, \ldots, d_1, j = 1, \ldots, d_2$$

где $\pi_{ij} = p_{1|ij} = p_{1ij}/p_{1++}$ – условные вероятности, а правая часть соотношения соответствует модели двухфакторного дисперсионного анализа, где $\alpha$ – взвешенное среднее; $\beta_i^X$ и $\beta_j^Z$ – главные эффекты, удовлетворяющие соотношениям $\beta_*^X = 0$, $\beta_*^Z = 0$ – главные эффекты факторов $X$ и $Z$ соответственно, а $\beta_{ij}$, удовлетворяющие соотношениям $\beta_{i*} = 0$, $\beta_{*j} = 0$ – взаимодействия. Гипотеза аддитивности влияния факторов $(X, Z)$ на результат $Y$

$$H_{A0} : \beta_{ij}^{XZ} = 0 \quad \forall\, i, j$$

по сути устанавливает независимость влияния факторов $X$ и $Z$ на результат $Y$, но не гарантирует условной независимости $X$ и $Z$ от $Y$. Независимость $Y$ от фактора $Z$ представляется гипотезой

$$H_Z : \beta_{ij}^{XZ} = 0, \beta_j^Z = 0 \quad \forall\, i, j$$

а гипотеза независимости признака $Y$ от $(X, Z)$ формулируется в виде

$$H_I : \beta_i^X = 0, \beta_j^Z = 0, \beta_{ij}^{XZ} = 0 \quad \forall\, i, j.$$

Более универсальной с точки зрения постановки задач является пуассоновская лог-линейная модель. Данная модель предполагает наличие совместного распределения $(X, Y, Z)$ и подразумевает использование пуассоновского подхода к накоплению статистической информации. Параметры

распределения $\lambda_{ijs} = \lambda p_{ijs}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$, $s = 1, \ldots, d_3$ – неотрицательные числа. Обобщенная линейная модель

$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X + \beta_j^Z + \beta_{ij}^{YX} + \beta_{is}^{YZ} + \beta_{js}^{XZ} + \beta_{ijs}^{YXZ},$$

где правая часть соответствует трехфакторному дисперсионному анализу с параметрами взвешенного среднего $\alpha$, главных эффектов $\beta_i^Y$, $\beta_j^X$, $\beta_s^Z$: $\beta_*^Y = 0$, $\beta_*^X = 0$, $\beta_*^Z = 0$, попарных взаимодействий $\beta_{ij}^{YX}$, $\beta_{is}^{YZ}$, $\beta_{js}^{XZ}$: $\beta_{ij*}^{YX} = 0$, $\beta_{i*s}^{YZ} = 0$, $\beta_{*js}^{XZ} = 0$, и взаимодействий 3-х факторов $\beta_{ijs}^{YXZ}$: $\beta_{ij*}^{YXZ} = \beta_{i*s}^{YXZ} = \beta_{*js}^{YXZ} = 0$. Для простоты, веса, соответствующие всем трем факторам, обычно выбирают как 0 или 1, что означает отсчет главных эффектов от заранее выбранного базового уровня соответствующего фактора, что позволяет избежать необходимости введения дополнительных ограничений на параметры модели.

Лог-линейная обобщенная линейная записывается в виде

$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X + \beta_j^Z + \beta_{ij}^{YX} + \beta_{is}^{YZ} + \beta_{js}^{XZ} + \beta_{ijs}^{YXZ}.$$

Гипотеза отсутствия взаимодействий всех трех факторов

$$H_{HA} : \beta_{ijs}^{YXZ} = 0, \quad \forall\, i, j, s$$

соответствует однородности зависимостей (при наличии двух бинарных наблюдаемых признаков $(X, Y)$ и одного контролируемого $Z$ получаем однородность отношений шансов), а соответствующая модель не содержит взаимодействий трех факторов

$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X + \beta_s^Z + \beta_{ij}^{YX} + \beta_{is}^{YZ} + \beta_{js}^{XZ}.$$

Отметим, что в общем случае свойство однородности зависимостей не зависит от выбора контролируемого фактора. Условная независимость $(Y, X)$ при условии $Z$ может быть записана в виде

$$H_{CI} : \beta_{ij}^{YX} = 0, \beta_{ijs}^{YXZ} = 0, \quad \forall\, i, j, s,$$

что соответствует лог-линейной обобщенной модели

$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X + \beta_s^Z + \beta_{is}^{YZ} + \beta_{js}^{XZ},$$

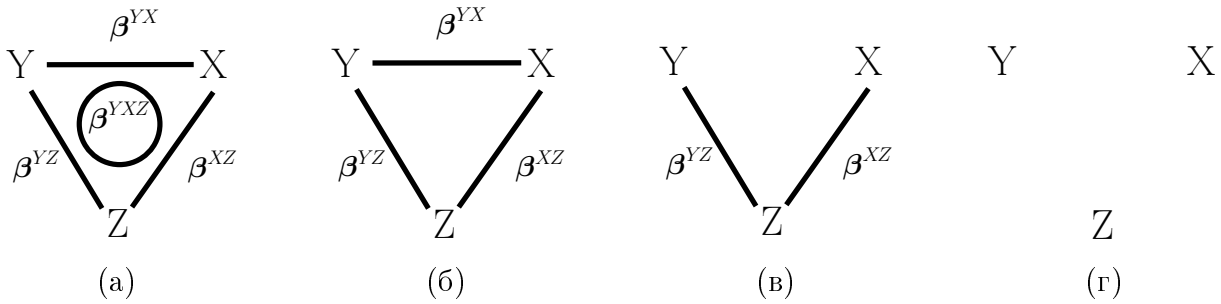


Рис. 2. *(а) Общие предположения; (б) Однородность зависимостей; (в) Условная независимость; (г) Независимость.*

а гипотезе независимости факторов $(Y, X, Z)$

$$H_I : \beta_{ij}^{YX} = 0, \beta_{is}^{YZ} = 0, \beta_{js}^{XZ} = 0, \beta_{ijs}^{YXZ} = 0, \quad \forall\, i, j, s$$

соответствует аддитивная модель

$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X + \beta_s^Z.$$

Для описания модели удобно использовать графическое представление (см. Рис. 2).

**Анализ данных сопряженности множества признаков.** Структура зависимости общего вида большого числа признаков становится настолько сложной, что для ее оценивания требуется непомерный объем статистических данных, но даже в случае возможности качественного оценивания интерпретация результатов представляется сложной задачей. Для понимания структуры зависимости удобно использовать пуассоновскую модель дисперсионного анализа, включающую всевозможные взаимодействия. При наличии множества признаков таблица сопряженности редко оказывается заполнена, поэтому роль общей модели существенно снижается. Более того, интерпретация зависимостей высокого порядка затруднительна, в связи с чем обычно либо переходят к анализу сопряженности на подмножествах признаков, либо ограничиваются лишь аддитивной моделью или попарными взаимодействиями. В этом случае, важную роль приобретают информационные критерии. Критерии Акайке и байесовский применимы в задаче выбора обобщенной модели, реализующей наилучший баланс (в том или ином смысле) соответствия модели и имеющихся статистических данных. Наиболее часто исследователь выбирает модель с использованием информационного критерия Акайке или байесовского, а затем проводит качественный анализ выбранной модели, включающий в себя интерпретацию оценок параметров, построение графиков зависимости и

доверительное оценивание построенных соотношений. Альтернативно, для выбора модели можно пользоваться методом последовательного включения параметров, или методом последовательного исключения параметров из модели, но такой подход более субъективен и менее эффективен.

**Анализ данных с категориальными наблюдаемыми переменными.** Помимо исследования сопряженности признаков к категориальным данным относят модели с категориальными наблюдаемыми переменными, тогда как контролируемые признаки могут быть произвольного вида. В большинстве случаев, количественные факторы можно подвергнуть группировке, превратив их в ординальные, но при этом часть статистической информации будет потеряна. Кроме того, возникает множество вопросов, связанных с выработкой объективных принципов группировки. Выбор распределения наблюдаемой переменной определяется ее характером. Если наблюдаемая переменная бинарная, то этот выбор ограничен распределением Бернулли, но при наличии большего или бесконечного числа уровней, выбор существенно расширяется. Иногда категориальную модель удается интерпретировать с помощью пуассоновской схемы накопления статистической информации, но довольно часто такой подход оказывается неоправданным, ввиду выявляемого различия между математическим ожиданием и дисперсией наблюдений. Тогда можно использовать отрицательное биномиальное или обобщенное пуассоновское распределение. Если число уровней наблюдаемой переменной конечно, то она заведомо имеет мультиномиальное распределение, параметры которого зависят от значений контролируемых факторов. Пуассоновский подход неприменим в этом случае, поскольку при различных значениях контролируемой переменной интенсивности пуассоновских потоков событий, отвечающих за наполнение ячеек таблицы будут различными. Недостаток модели мультиномиального распределения в том, что параметр распределения многомерный, а это диктует необходимость рассмотрения многомерной регрессионной модели, поскольку каждый из параметров распределения должен быть связан с ковариатами параметрами регрессии. Таким образом, при наличии аргументов в пользу выбора распределения Бернулли, гипергеометрического распределения или возможности аппроксимации пуассоновским, гипергеометрическим или обобщенным пуассоновским, исследователю целесообразно отказаться от использования мультиномиального распределения.

Обобщенная линейная модель может быть записана как и раньше в виде

$$g(\mathbb{E}_\theta(Y|\boldsymbol{z})) = \boldsymbol{X}^T(\boldsymbol{z})\boldsymbol{\beta},$$

но по сравнению с моделями дисперсионного анализа регрессор $\mathbf{X}$ может иметь более общий вид. При наличии одного количественного или ординального контролируемого фактора $z$ можно использовать линейную зависимость

$$g(\mathbb{E}_\theta(Y|\boldsymbol{z})) = \beta_1 + \beta_2 z.$$

Иными словами, регрессор определяется соотношением $\boldsymbol{X}^T = (1, z)$. Линейная зависимость преобразованного среднего от ковариаты далеко не всегда удовлетворяет исследователя. Тогда естественно задаться некоторым $s > 1$ и рассмотреть полиномиальную обобщенную модель

$$g(\mathbb{E}_\theta(Y|\boldsymbol{z})) = \beta_1 + \beta_2 z + \ldots + \beta_s z^{s-1},$$

для которой регрессор задается соотношением $\boldsymbol{X}^T(z) = (1, z, \ldots, z^{s-1})$. В общем случае, для задания обобщенной линейной модели возможно использование функций, отличных от полиномов.

При наличии двух и более контролируемых количественных признаков, возникает вопрос о выборе функций для описания зависимости преобразованного среднего от ковариаты. В частности, при наличии двух контролируемых признаков можно рассмотреть аддитивную линейную модель

$$g(\mathbb{E}_\theta(Y|\boldsymbol{z})) = \beta_1 + \beta_2 z_1 + \beta_3 z_2.$$

Такая модель будет соответствовать независимому вкладу контролируемых факторов в результат. Чтобы предусмотреть возможность взаимодействия факторов можно добавить в модель произведение

$$g(\mathbb{E}_\theta(Y|\boldsymbol{z})) = \beta_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 z_1 z_2,$$

или дополнить модель до полинома второго порядка от двух переменных

$$g(\mathbb{E}_\theta(Y|\boldsymbol{z})) = \beta_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 z_1^2 + \beta_5 z_2^2 + \beta_6 z_1 z_2.$$

В последнем случае, регрессор выражается через значения ковариат следующим образом $\boldsymbol{X}^T(\boldsymbol{z}) = (1, z_1, z_2, z_1^2, z_2^2, z_1 z_2)$.

В некоторых случаях часть контролируемых признаков являются количественными, а часть - качественными. Аддитивная модель влияния факторов может не устраивать исследователя. При наличии одного количественного и одного качественного фактора имеет смысл рассмотреть

совместное влияние каждого уровня качественного фактора и значения количественного фактора. Таким образом формируется модель ковариационного анализа. Гипотезы ковариационного анализа формулируются в терминах линейных функций параметров модели, а их проверка происходит согласно общей теории обобщенных линейных моделей для экспоненциальных семейств. Методы построения частных и совместных доверительных интервалов применимы также и в задачах ковариационного анализа.

## 5. Анализ категориальных данных с использованием пакета R

Пакет **R** – один из мощнейших инструментов, предназначенных для обработки статистических данных любой структуры. Простой и удобный язык позволяет не только использовать уже готовые, но и программировать собственные методы обработки статистических данных. Фактически, пакет R был создан на базе известного статистического пакета S-plus, поэтому он поддерживает большинство функций S-plus. В отличие от S-plus пакет R является свободно распространяемым программным обеспечением. Для установки пакета следует войти на сайт <http://cran.r-project.org/>, выбрать ближайшее «зеркало» (mirrow, (Англ.)) и скачать программу установки на свой компьютер. Расширение возможностей R возможно за счет установки дополнительных специализированных пакетов. Большинство функций, необходимых для анализа категориальных данных входят в базовый пакет.

**Представление данных.** Категориальные данные могут быть представлены в оригинальном виде, подразумевающем наличие определенной записи для каждого объекта исследования, в сокращенном виде, включающем всевозможные наблюдавшиеся комбинации значений факторов и числа наблюдений с соответствующими значениями факторов, а также в виде массива сопряженности. Первые два представления подразумевают запись данных в формате `data.frame`, хотя можно использовать также формат `list` или `matrix`. Третий тип записи подразумевает формат `array`, размерность которого совпадает с имеющимся числом переменных. Сокращенный вид и массив сопряженности содержат одни и те же значения, но в массиве сопряженности значения наблюдаемых факторов выносятся в названия уровней по соответствующему измерению. Данные обычно поставляются в оригинальном или сокращенном виде и вводятся с использованием команды `read.table()`. Для перевода данных из сокращенного в оригинальный вид можно использовать следующую функцию

```
tab.to.orig<-function(dct,v.names=NA,c.name="count"){
  if (!is.na(v.names)){ dct<-dct[,n%in%v.names] }
  n<-names(dct)
  v.names<-n[n!=c.name]
  d1<-dim(dct)[1]
  d2<-length(v.names)
  out.d<-list()
  for (j in 1:d2){
    out.v<-NULL
    for (i in 1:d1){
      out.v<-c(out.v,array(dct[[v.names[j]]][i],dim=dct[[c.name]][i]))
      }
    out.d[[j]]<-out.v
    }
  out.d<-as.data.frame(out.d)
  names(out.d)<-v.names
  return(out.d)
}
```

Разумеется, идентифицировать исходный порядок наблюдений по данным в сокращенном виде невозможно, поэтому используется фиксированный порядок. Необязательный параметр `v.names` позволяет выбирать переменные для анализа. По умолчанию используются все переменные, входящие в исходный массив данных. Для перевода данных из оригинального вида в сокращенный или в массив сопряженности можно использовать следующую функцию

```
orig.to.tab<-function(dt,v.names=NA,c.name="count",array=FALSE){
  if (!is.na(v.names)){ dt<-dt[,names(dt)%in%v.names] }
  z<-as.list(dt)
  v.names<-names(z)
  n.var<-length(v.names)
  tt<-table(z)
  if (array | n.var<=1){ return(tt) } else {
    tn<-dimnames(tt)
    cts<-unlist(as.list(tt))
    n.cts<-length(cts)
    out.d<-list()
    out.d[[1]]<-array(tn[[1]],dim=n.cts)
    c<-1
    lti<-length(tn[[1]])
    for (i in 2:n.var){
      c<-c*lti
      lti<-length(tn[[i]])
      out.di<-unlist(matrix(tn[[i]],nrow=c,ncol=lti,byrow=TRUE))
      out.d[[i]]<-array(out.di,dim=n.cts)
    }
```

```
    names(out.d)<-v.names
    out.d<-as.data.frame(out.d)
    out.d[[c.name]]<-cts
    return(out.d)
    }
}
```

Если параметр `array=TRUE` или если выбрана только одна переменная в исходном наборе данных, то выводится массив сопряженности, выдаваемый функцией `table()`. В остальных случаях данные выводятся в сокращенном виде. Альтернативно, для перестроения массива сопряженности в массив данных в сокращенном виде можно использовать функцию `as.data.frame.table()` пакета `data.table`.

### 5.1. Классические методы анализа категориальных данных

Классические критерии $\chi^2$ и точный критерий Фишера реализованы в виде функций `chisq.test()` и `fisher.test()` соответственно. Выводимые объекты содержат в себе всю необходимую информацию о критерии и много технической информации. Данные принимаются в оригинальном виде (два вектора) или в виде таблицы сопряженности (одна матрица). Выбор аргумента `simulate.p.value=TRUE` позволяет использовать критерий случайных перестановок на базе статистики $\chi^2$ и имеет смысл, если хотя бы одно из чисел в таблице сопряженности мало. В последнем случае, при запуске функции по умолчанию **R** выдает предупреждение. Следует помнить, что $P$-значение критерия случайных перестановок ограничено снизу параметром $1/B$, а параметр $B$ по умолчанию принимается равным 2000. Чтобы иметь возможность идентифицировать меньшие $P$-значения следует выбирать параметр $B$ побольше. Функция `fisher.test()` реализована не только для таблиц сопряженности $2 \times 2$, но и для таблиц сопряженности большего размера, для которых используется обобщенный точный критерий Фишера. В случае таблиц $2 \times 2$ вычисляется отношение шансов и 95% доверительный интервал для него. Возможна установка как односторонней, так и двухсторонней альтернативы. Реализация $G$-критерия отношения правдоподобия отсутствует в базовом пакете, но существует его реализация в виде функций `likelihood.test()` пакета `Deducer`. Синтаксис данной функции похож на `chisq.test()`. Рассмотрим, как можно построить $G$-критерий самостоятельно. Предположим, что у нас готова таблица сопряженности `tt`. Следующая программа позволяет вычислить статистику критерия `G` и соответствующее $P$-значение:

```
> d.tt<-dim(tt)
```

```
> obs.p<-tt/sum(tt)
> exp.p<-as.matrix(colSums(obs.p))%*%t(as.matrix(rowSums(obs.p)))
> G<-sum(obs.p*log(obs.p/exp.p))
> df<-(d.tt[1]-1)*(d.tt[2]-1)
> pv<-pchisq(LRS,df,lower.tail=FALSE)
```

В третьей строчке программы проводится матричное умножение двух векторов, тогда как в четвертой строчке выполняются почленные операции с матрицами.

Для решения некоторых задач анализа таблиц сопряженности размерности 3 удается адаптировать критерии $\chi^2$ и отношения правдоподобия. Для проверки условной независимости реализован критерий Кочрана–Мантела–Хензела `mantelhaen.test()`. Данные принимаются в виде массива сопряженности (массив размерности 3) или в оригинальной форме (три вектора наблюдений). В первом случае, третье измерение соответствует контролируемой переменной, а во втором случае контролируемой переменной соответствует третий вектор. Как и для критерия $\chi^2$ следует отменить поправку `correct=FALSE`, если исходные данные изначально были категориальными, а не были получены путем группировки количественных переменных. Для таблиц $2\times2\times K$ есть возможность использовать точный критерий, вместо критерия Кочрана–Мантела–Хензела. Для этого следует установить параметр `exact=TRUE`. Объект создаваемый функцией `mantelhaen.test()` содержит оценку общего отношения шансов `estimate` и доверительный интервал для неё `confint`. Легко реализовать критерий проверки условной независимости используя принцип $\chi^2$. Подготовим $dt$ – массив сопряженности размерности 3, третье измерение которого соответствует контролируемой переменной. Следующий код позволяет вычислить статистику критерия и $P$-значение критерия $\chi^2$:

```
> d<-dim(dt)
> exp.c<-array(dim=d)
> for (k in 1:d[3]){ exp.c[,,k]<-as.matrix(colSums(dt[,,k]))%*%
   t(as.matrix(rowSums(dt[,,k])))/sum(dt[,,k]) }
> X2<-sum((dt-exp.c)^2/exp.c)
> df<-(d[1]-1)*(d[2]-1)*d[3]
> pv<-pchisq(LRS,df,lower.tail=FALSE)
```

Аналогичным образом можно реализовать критерий отношения правдоподобия. Критерий Бреслова–Дея для проверки однородности зависимости в таблицах $2 \times 2 \times K$ реализован в виде функции `BreslowDayTest()` пакета `DescTools`.

## 5.2. Анализ категориальных данных с использованием обобщенных линейных моделей.

Основным инструментом обработки данных в условиях обобщенной линейной модели является функция `glm`, входящая в пакет `stats` базового набора пакетов **R**. Для запуска функции `glm` следует ввести

```
>ans<-glm(<формула>,family=<модель>,...)
```

Основными параметрами запуска `glm` являются формула `formula` и вероятностная модель `family`. Формально `family` не является обязательным аргументом функции `glm`, но поскольку по умолчанию `glm` использует `family="gaussian"`, то для решения задач категориального анализа установка пользовательского значения данного параметра является обязательной. Для решения задач категориального анализа используются биномиальное семейство распределений или семейство распределений Пуассона. Синтаксис параметра `family=<семейство распределений>(link=<функция связи>)` позволяет устанавливать пользовательское значение функции связи. Биномиальное семейство распределений `family=binomial` может быть использовано с функциями связи `"logit"`, `"probit"`, `"cachit"`, `"log"` и `"cloglog"`, среди которых по умолчанию используется логистическая регрессия `family=binomial(link="logit")`. Пуассоновское распределение `family=poisson` сочетается с функциями связи `"log"`, `"identity"`, среди которых по умолчанию выбирается логарифмическая модель `family=poisson(link="log")`.

Выбор биномиального распределения возможен только в случае если как минимум одна из категориальных переменных имеет лишь два уровня. При выборе биномиального распределения круг решаемых задач ограничен условным распределением бинарной переменной при условии других факторов. Данные для обобщенной модели с биномиальным распределением должны быть представлены в формате `data.frame`, где наблюдаемая переменная должна иметь значения из $\{0, 1\}$. Имена переменных исходного `data.frame` используются при записи формулы. Чтобы получить корректный результат следует адаптировать все переменные кроме наблюдаемой к формату «фактор». Например,

```
>dt$X<-as.factor(dt$X),
```

где `dt` – исходный массив данных; `X` – имя фактора группировки. При наличии наблюдаемой переменной $Y$ и одного фактора группировки $X$ запуск функции `glm` производится следующим образом:

```
>res<-glm(Y∼X,family="bimomial",data="dt").
```

Полученный объект `res` содержит результаты статистического анализа в условиях выбранной модели и некоторую техническую информацию. Некоторые результаты статистического анализа можно вывести на экран, применяя функцию `summary(res)`. Функция `anova(res,test="LRT")` выводит результаты последовательного включения параметров модели, где при наличии значения аргумента `test` вычисляются $P$-значения выбранного критерия значимости включения параметров, определяющих отличие соседних вложенных моделей. Для построения частных асимптотических доверительных интервалов для параметров модели можно использовать `confint(res)`. Если построить обобщенную модель логистической регрессии без учета фактора $X$, то можно получить то же $P$-значение

```
>res0<-glm(Y∼1,family="bimomial",data="dt").
```

Для дальнейшего сравнительного анализа вложенных моделей методами дисперсионного анализа используют `anova(res0,res,test="LRT")`, получая в результате $P$-значение критерия отношения правдоподобия проверки значимости параметра модели `res`, отсутствующего в модели `res0`.

По сравнению с классическими критериями, обобщенные линейные модели позволяют учитывать сопутствующие факторы, но в простых ситуациях использование классических критериев может быть предпочтительным.

**Анализ таблиц сопряженности двух признаков.** Наиболее универсальный метод анализа таблиц сопряженности – пуассоновская модель, но для таблиц сопряженности размерности 2 эффективнее использовать классические методы анализа, за исключением случаев когда необходимо делать поправку на влияние дополнительных признаков, необязательно категориальных. Данные для пуассоновской модели должны быть подготовлены в сокращенном виде. Предположим, что значения переменных $X$ и $Y$ и соответствующие частоты сформированы в массив данных `dat.pois`. Для проведения корректного анализа имеющихся данных следует адаптировать переменные $X$ и $Y$ к формату факторов `dat.pois$X<-as.factor(dat.pois$X)` и `dat.pois$Y<-as.factor(dat.pois$Y)`. Отметим, что в случае бинарных $\{0,1\}$ или нечисловых значений $X$ и $Y$ адаптация к формату фактор не критична, но в случае числовых значений $X$ и $Y$ отличных от 0 и 1 будет построена модель линейной регрессии вместо модели дисперсионного анализа. Формула для пуассоновской модели с учетом зависимости

переменных $X$ и $Y$ записывается в виде

```
>m.sat<-formula(Count~X*Y),
```

где `Count` – название переменной, содержащей числа наблюдений с соответствующими комбинациями значений факторов $X$ и $Y$. Независимости соответствует аддитивная модель

```
>m.ind<-formula(Count~X+Y).
```

Далее проводится построение моделей с учетом зависимости и в случае независимости факторов $X$ и $Y$, проводится анализ и выводится $P$-значение критерия отношения правдоподобия

```
>res.sat<-glm(m.sat,family="poisson",data="dat.pois")
>res.ind<-glm(m.ind,family="poisson",data="dat.pois")
>aov.i<-anova(res.ind,res.sat,test="LRT")
>pv<-aov.i$"Pr(>Chi)"
```

При желании использовать критерий типа Вальда следует установить параметр `test="Chisq"`. Инструмент построения асимптотических доверительных интервалов для параметров модели также реализован. По умолчанию устанавливается наименьший уровень каждого фактора в качестве базового, поэтому полученные коэффициенты представляют собой логарифмы частных отношений шансов по отношению к паре наименьших уровней. Если значения уровня фактора категориальные, то наименьшее значение вычисляется исходя из имеющейся иерархии символов. Асимптотические доверительные интервалы для параметров модели можно получить с использованием функции `confint(res.sat)`. При построении совместных доверительных интервалов для частных отношений шансов и их линейных функций требуется оценка ковариационной матрицы, которая генерируется командой `summary(res.sat)$cov.unscaled`.

При наличии хотя бы одного бинарного признака есть возможность использовать модель логистической регрессии. Тогда бинарный признак считается наблюдаемым, а другой признак – контролируемым, что не исключает возможности его случайного выбора. Для модели логистической регрессии требуется представление данных в оригинальной форме. Кроме того, уровни бинарного фактора должны быть переименованы в 0 и 1. Предположим, что данные подготовлены и записаны в таблицу данных `dat.lr`, где `dat.lr$Y` – бинарный признак с уровнями 0 и 1, а `dat.lr$X` – контролируемый признак. Рекомендуется сразу адаптировать контролируемый признак к формату фактора `dat.lr$X<-as.factor(dat.lr$X)`, если

это не было сделано ранее. Запуск модели логистической регрессии производится аналогично запуску пуассоновской модели. Снова формируем общую и гипотетическую модели

```
>m.sat<-formula(Y∼X
>m.ind<-formula(Y∼1)
```

соответственно и запускаем обобщенную линейную модель

```
>res.sat<-glm(m.sat,family="binomial",data="dat. lr")
>res.ind<-glm(m.ind,family="binomial",data="dat.lr")
>aov.i<-anova(res.ind,res.sat,test="LRT")
>pv<-aov.i$"Pr(>Chi)"
```

Чтобы использовать функцию связи, отличную от `logit` требуется скорректировать параметр `family`,например, `family=binomial(link="probit")`. Отметим, что формирование нулевой модели `res.ind` необязательно, т.к. по умолчанию функция `glm()` выдает результат проверки гипотезы по отношению к нулевой модели. Все опции, описанные при обсуждении пуассоновской модели, сохраняются.

**Анализ таблиц сопряженности трех признаков.** Технически, анализ сопряженности трех признаков реализуется таким же образом как и анализ сопряженности двух признаков. Для проведения анализа сопряженности с использованием пуассоновской модели предположим, что данные представлены в сокращенном виде и переменные `X`, `Y`, `Z`, содержащие уровни соответствующих признаков, адаптированы к типу факторов. Формируются общая модель `m.sat<-formula(Count∼X*Y*Z)`, где `Count` – название переменной, содержащей частоты, модель отсутствия взаимодействий всех трех признаков `m.ha<-formula(Count∼X*Y+X*Z+Y*Z)`, соответствующая гипотезе однородности зависимостей, а также модели условной независимости `m.ci<-formula(Count∼X*Z+Y*Z)` переменных $X$ и $Y$ при условии $Z$ и аддитивная модель независимости всех трех факторов `m.ind<-formula(Count∼X+Y+Z)`. Вычисление $P$-значений критериев проверки сформулированных гипотез осуществляет следующий код:

```
>res.sat<-glm(m.sat,family="binomial",data="dat.lr")
>res.ha<-glm(m.ha,family="binomial",data="dat.lr")
>res.ci<-glm(m.ci,family="binomial",data="dat.lr")
>res.ind<-glm(m.ind,family="binomial",data="dat.lr")
>pv.ha<-anova(res.ha,res.sat,test="LRT")$"Pr(>Chi)"
>pv.ci<-anova(res.ci,res.sat,test="LRT")$"Pr(>Chi)"
>pv.ind<-anova(res.ind,res.sat,test="LRT")$"Pr(>Chi)"
```

**Использование обобщенной модели ковариационного анализа.** Синтаксис построения модели ковариационного анализа аналогичен синтаксису модели дисперсионного анализа. Категориальные ковариаты предварительно рекомендуется адаптировать к типу факторов. Если тип соответствующей переменной числовой, то функция `glm()` будет воспринимать данную ковариату как количественную. При формировании полиномиальной модели следует сформировать новые ковариаты как степени исходной ковариаты $z$, или использовать `I(z^k)` в формуле. Например, формула с полиномом второго порядка выглядит следующим образом `f2<-formula(Y~z+I(z^2))`, тогда как формула `f1<-formula(Y~z+z^2)` соответствует линейной регрессии `f1<-formula(Y~z)`. Для формирования взаимодействий в формулах используются символы «`*`» или «`:`» (надежнее использовать «`*`»).

Для категориальных данных существует следующий набор распределений наблюдаемой переменной и функций связи:

1. `binomial` (биномиальное) – может быть использовано с функциями связи `"logit"`, `"probit"`, `"cachit"`, `"log"` и `"cloglog"`;
2. `poisson` (пуассоновское) – может быть использовано с функциями связи `"log"`, `"identity"`;
3. `quasibinomial` и `quasipoisson` отличаются от биномиального и пуассоновского введением параметра дисперсии;
4. `quasi(link=<функция связи>, variance=<функция дисперсии>)` подразумевает использование двух параметров. В качестве функций связи могут использоваться `"logit"`, `"probit"`, `"cloglog"`, `"identity"`, `"inverse"`, `"log"`, `"1/mu^2"` и `"sqrt"`, а в качестве функции дисперсии `"constant"`, `"mu(1-mu)"`, `"mu"`, `"mu^2"` и `"mu^3"`.

Обобщенная линейная модель для биномиального распределения строится с использованием `link=binomial`, где в качестве наблюдаемой переменной используются частоты $Y/m$, и устанавливается значение `weights=array(m,dim=n)`, где `Y` – наблюдаемая переменная, `m` – соответствующий параметр биномиального распределения,`n` – размер выборки. Обобщенная линейная модель для отрицательного биномиального распределения реализована функцией `glm.nb()` пакета `MASS`, где в качестве функции связи по умолчанию используется `link=log`, но дополнительно есть возможность установки `link=sqrt` или `link=identity`.

# СПИСОК ЛИТЕРАТУРЫ


1. Малов С. В. Регрессионный анализ: теоретические основы и практические рекомендации. - Спб.: Изд-во СПбГУ, 2013. - 276 стр.
2. Шеффе Г. Дисперсионный анализ. М.: Наука, 1980.
3. Коробейникров А., Малов С.В., Матвеева И.В. Основные алгоритмы численного анализа. Использование пакета R(S-plus) для анализа статистических данных. Методические указания. СПб: Изд-во СПбГЭТУ «ЛЭТИ», 2011.
4. Agresti A. An introduction to categorical data analysis. Wiley & Sons, Inc., 2006.
5. McCullagh P., Nelder J. A. Generalized linear models. 2-nd edition. London: Chapman & Hall, 1989.

## СОДЕРЖАНИЕ

Малов Сергей Васильевич,
Малова Ирина Юрьевна

**Базовые модели биостатистики: анализ категориальных данных**

Учебное пособие

Редактор Х. Х. Хххxx

---

Подписано к печати 00.00.00 Формат 60 × 84 1/16
Бумага офсетная. Печать цифровая. Печ. л. 2,0.
Гарнитура «Computer modern». Тираж 144 экз. Заказ 000.

---

Издательство СПбГЭТУ «ЛЭТИ»
197376, С.-Петербург, ул. Проф. Попова, 5