

Урок 9

Множественная проверка гипотез

9.1. В чем проблема

Этот урок посвящен проблеме множественной проверки гипотез. Для того чтобы понять, в чем заключается эта проблема, можно рассмотреть несколько примеров.

9.1.1. Поиск экстрасенсов

Первый пример связан с исследованиями Джозефа Райна. Это американский ученый 50-х годов, который занимался исследованиями возможностей экстрасенсорного восприятия. Первый этап таких исследований — это поиск экстрасенсов. Джозеф Райн придумал для этого следующий эксперимент. Испытуемому предлагалось угадать цвета десяти карт, лежащих рубашкой вверх (рисунок 9.1).

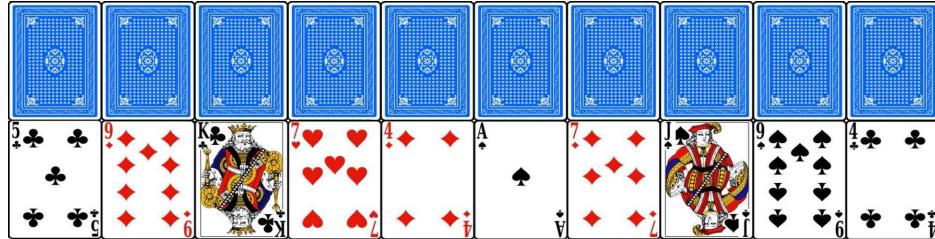


Рис. 9.1: Эксперимент по угадыванию карт

Проверялась нулевая гипотеза H_0 : испытуемый выбирает ответ наугад. Альтернативная гипотеза H_1 : испытуемый может предсказывать цвета карт. Статистика t — число карт, цвета которых угаданы, — при справедливости нулевой гипотезы имеет биномиальное распределение с параметрами $n = 10, p = 0.5$, поскольку цвета только два, и они называются наугад. Вероятность правильно назвать цвета 9 и более карт:

$$P(t \geq 9 | H_0) = 11 \cdot \frac{1}{2}^{10} = 0.0107421875.$$

То есть, если испытуемый угадывает 9 карт, получается достигаемый уровень значимости $p \approx 0.01$, и нулевую гипотезу можно с чистой совестью отклонить в пользу односторонней альтернативы.

В экспериментах Джозефа Райна процедуру отбора прошли 1000 человек. Девять из них угадали цвета 9 из 10 карт, еще двое угадали все 10 карт. Ни один из этих испытуемых в последующих экспериментах не подтвердил своих способностей, из чего Джозеф Райн сделал вывод, что экстрасенсам нельзя говорить о том, что они экстрасенсы, потому что от этого их способности сразу пропадают. Однако очевидно, что проблема кроется в чем-то другом.

Если принять гипотезу о том, что экстрасенсов не существует, то вероятность того, что из тысячи человек хотя бы один случайно угадает цвета 9 или 10 из 10 карт:

$$1 - \left(1 - 11 \cdot \frac{1}{2}^{10}\right)^{1000} \approx 0.9999796.$$

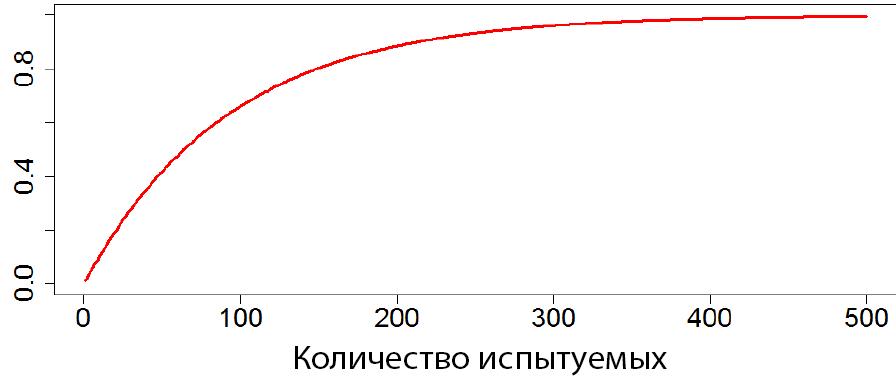


Рис. 9.2: Зависимость вероятности того, что хотя бы один испытуемый случайно угадает цвета 9 или 10 карт из 10, от количества испытуемых

На рисунке 9.2 показано, как описанная выше вероятность ведет себя в зависимости от количества испытуемых. Из графика видно, что она растет очень быстро. Уже при количестве испытуемых $N = 100$ вероятность найти хотя бы одного экстрасенса превышает $1/2$. При $N = 500$ такая вероятность уже примерно равна единице.

Тот факт, что с помощью этой статистической процедуры находятся экстрасенсы, является прекрасным примером влияния эффекта множественной проверки гипотез. При одновременной проверке большого количества гипотез вероятность совершить хотя бы одну ошибку первого рода (то есть должно отвергнуть верную нулевую гипотезу) становится очень большой.

9.1.2. Нейронаука

Еще один яркий пример действия эффекта множественной проверки гипотез можно найти в нейронауке.

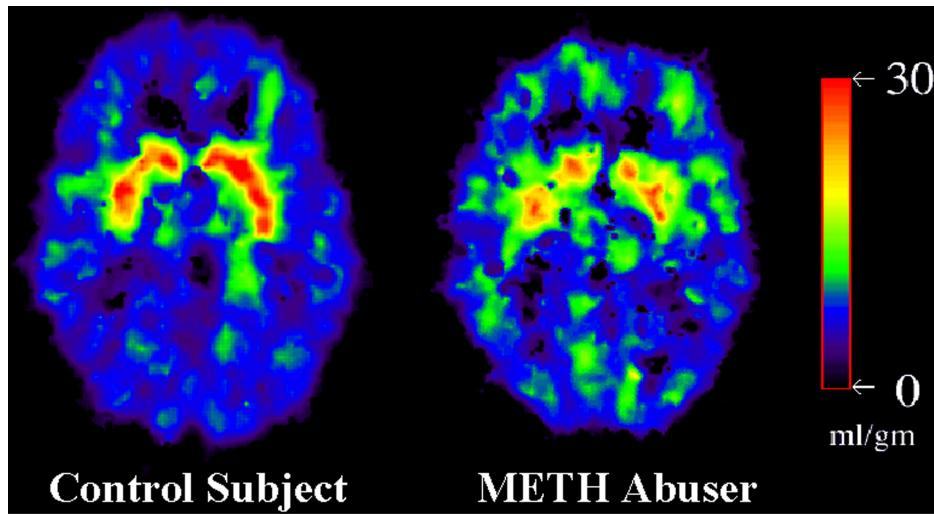


Рис. 9.3: Данные позитронно-эмиссионной томографии

Анализируются данные позитронно-эмиссионной томографии или функциональной магнитно-резонансной томографии (рисунок 9.3). Типичный дизайн такого эксперимента следующий: берётся контрольная группа испытуемых, с которыми ничего не происходит, и измеряется активность их мозга. Затем те же измерения производят с другой группой испытуемых, состояние которых каким-то образом изменили. Далее эти две выборки сравнивают, пытаясь выяснить, на какие области мозга подействовало различие между двумя экспериментальными условиями.

Решение такой задачи связано с проверкой очень большого количества гипотез. Фактически для двумерного изображения мозга гипотеза проверяется в каждой точке, для трехмерного изображения мозга, которое возникает при магнитно-резонансной томографии, гипотеза проверяется в каждом voxelе (то есть в каждом трехмерном пикселе трехмерного изображения мозга). Пикселей могут быть тысячи, voxelей могут быть миллионы. Таким образом, требуется проверить очень много гипотез. И если ничего не делать, эффект множественной проверки гипотез будет проявляться очень ярко.

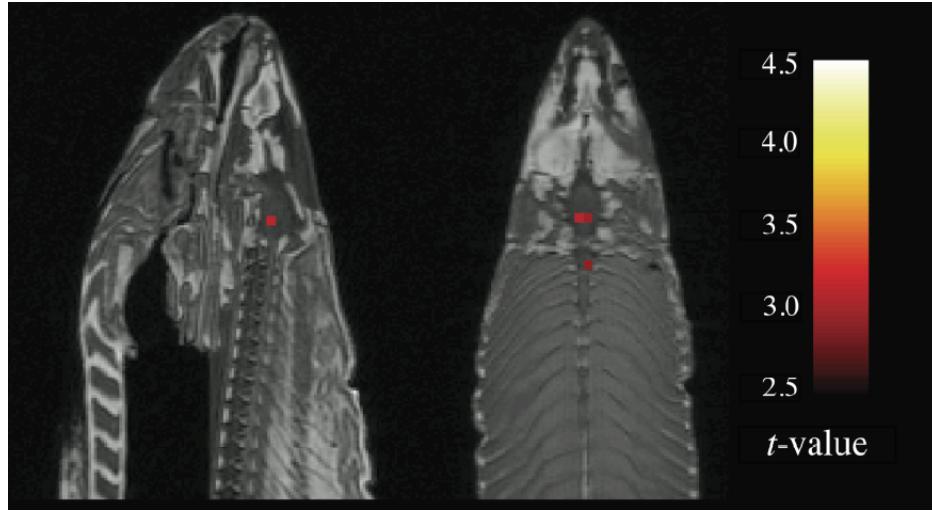


Рис. 9.4: Области мозга мёртвого лосося, в которых активность значимо изменилась при просмотре изображений

Лучше всего это демонстрирует следующий пример. Команда исследователей воспроизвела один из типичных дизайнов нейронаучных экспериментов, в котором испытуемому последовательно и много раз демонстрируются похожие стимулы, затем активность его мозга в ответ на эти стимулы сравнивается с активностью его мозга в состоянии покоя. Роль испытуемого в этом эксперименте играл мертвый лосось. В качестве стимула ему показывали картинки с изображениями людей в различных социальных ситуациях. Как видно из рисунка 9.4, задача поиска областей, которые реагируют на этот стимул, была решена успешно. В мозге лосося были выявлены области, в которых активность значимо изменилась, они на рисунке обозначены красным.

За последнее десятилетие методы анализа данных в нейронауке, в том числе методы поправки на множественную проверку гипотез, существенно улучшились. О таких методах и пойдёт речь далее.

9.2. Постановка

В этой части будет дана математическая постановка задачи множественной проверки гипотез. Для этого полезно вспомнить, как ставится задача однократной проверки гипотез.

9.2.1. Задача однократной проверки гипотез

выборка:	$X^n = (X_1, \dots, X_n)$, $X \sim \mathbf{P}$;
нулевая гипотеза:	$H_0: \mathbf{P} \in \omega$;
альтернатива:	$H_1: \mathbf{P} \notin \omega$;
статистика:	$T(X^n)$;
нулевое распределение:	$F(x)$;

Таблица 9.1: Задача однократной проверки гипотез

Задача однократной проверки гипотез ставится следующим образом. Имеется некоторая выборка X объема n из неизвестного распределения P . Проверяется нулевая гипотеза H_0 о распределении P против общей альтернативы H_1 . Это делается с помощью статистики T , которая является функцией от выборки. Для

этой статистики известно нулевое распределение $F(x)$, то есть распределение при справедливости нулевой гипотезы (рисунок 9.5).

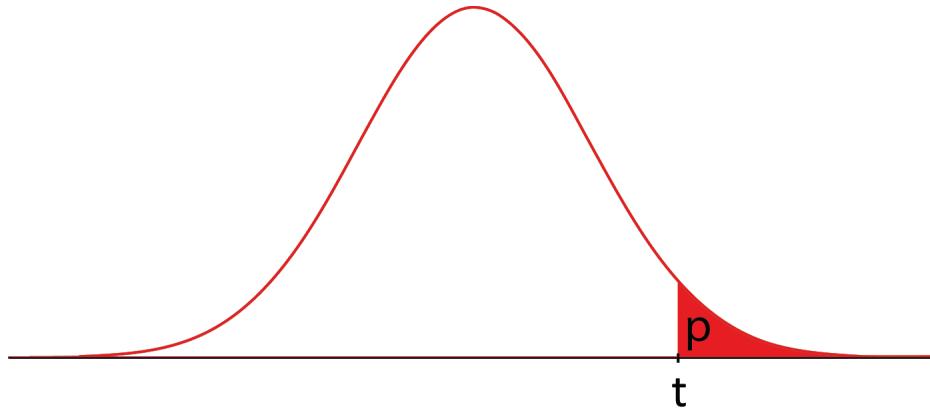


Рис. 9.5: Нулевое распределение статистики

По этому нулевому распределению, по его хвостам (разным, в зависимости от типа альтернативы) вычисляется достигаемый уровень значимости, то есть вероятность получить такое же значение статистики, какое было получено в эксперименте, или ещё более экстремальное:

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

Достигаемый уровень значимости сравнивается с порогом α — уровнем значимости (типичное значение 0.05). Если достигаемый уровень значимости меньше, чем α , то нулевая гипотеза отвергается в пользу альтернативы.

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

Таблица 9.2: Типы ошибок при проверке гипотезы

При однократной проверке гипотезы всегда есть вероятность совершить ошибку первого или второго рода (таблица 9.2). Механизм проверки гипотез построен так, что вероятность ошибки первого рода, то есть вероятность ложно отвергнуть верную нулевую гипотезу, сверху ограничена достигаемым уровнем значимости α :

$$\mathbf{P}(\text{ошибка I рода}) = \mathbf{P}(T \geq t | H_0) \leq \alpha.$$

9.2.2. Задача множественной проверки гипотез

данные:	$\mathbf{X} = \{X_1^{n_1}, \dots, X_m^{n_m}\}, \quad X_i \sim \mathbf{P}_i;$
нулевые гипотезы:	$H_i: \mathbf{P}_i \in \omega_i;$
альтернативы:	$H'_i: \mathbf{P}_i \notin \omega_i;$
статистики:	$T_i = T(X_i^{n_i});$

Таблица 9.3: Задача множественной проверки гипотез

Теперь можно разобраться с постановкой задачи множественной проверки гипотез (таблица 9.3). Пусть имеется m выборок, каждая своего размера, и из своего распределения. Каждой выборке соответствует своя гипотеза нулевая гипотеза H_i и альтернатива H'_i . Каждая из гипотез проверяется своей статистикой T_i . Для

каждой из статистик известно свое нулевое распределение. Таким образом, можно вычислить достигаемые уровни значимости всех гипотез:

$$p_i, \quad i = 1, \dots, m.$$

Для этого вводятся следующие обозначения. Пусть \mathbf{M} — это множество индексов:

$$\mathbf{M} = \{1, 2, \dots, m\};$$

\mathbf{M}_0 — это множество индексов верных нулевых гипотез, пусть его мощность равна m_0 :

$$\mathbf{M}_0 = \{i : H_i \text{ верна}\}, \quad |\mathbf{M}_0| = m_0.$$

Естественно, это множество неизвестно, потому что иначе не было бы смысла проверять гипотезы. Пусть \mathbf{R} — это множество индексов отвергаемых гипотез, а его мощность равна R :

$$\mathbf{R} = \{i : H_i \text{ отвергнута}\}, \quad |\mathbf{R}| = R$$

Тогда пересечение множеств \mathbf{R} и \mathbf{M}_0 состоит из неверно отвергнутых гипотез. Мощность этого множества обозначается V , это есть число ошибок первого рода:

$$V = |\mathbf{M}_0 \cap \mathbf{R}|.$$

	# верных H_i	# неверных H_i	Σ
# принятых H_i	U	T	$m - R$
# отвергнутых H_i	V	S	R
Σ	m_0	$m - m_0$	m

Таблица 9.4: Информация о верных и неверных, принятых и отвергнутых гипотезах для случая множественной проверки гипотез

По аналогии с задачей однократной проверки гипотез можно составить таблицу 2×2 , в которой будет стоят количество верных и неверных, принятых и отвергнутых гипотез (таблица 9.4). Из всех величин, записанных в таблице, известна только m — общее число гипотез. А единственный параметр, которым можно управлять, — это R , количество отвергаемых гипотез. При этом самая пугающая величина — это V , количество ошибок первого рода. Хочется совершать мало ошибок первого рода, но при этом единственное, что можно делать, — это перераспределять по этой таблице гипотезы из второй строки в первую. То есть, чтобы совершать мало ошибок первого рода, нужно отвергать меньше гипотез.

9.3. FWER. Поправка Бонферрони

Задача множественной проверки гипотез поставлена, теперь нужно её решить. Интерес представляет некоторая статистическая процедура, которая дает гарантии на значение V (таблица 9.2), оно не должно быть слишком большим.

9.3.1. Групповая вероятность ошибки первого рода (FWER)

Напрямую с V работать не очень удобно, поэтому, как правило, берут некоторые меры, определенные над V , и работают с ними. Одна из самых распространенных таких мер — это групповая вероятность ошибки первого рода (familywise error rate). По определению это вероятность совершить хотя бы одну ошибку первого рода:

$$\text{FWER} = P(V > 0).$$

Эту величину хочется контролировать на уровне α :

$$\text{FWER} = P(V > 0) \leq \alpha.$$

То есть, хочется построить такую статистическую процедуру, что вероятность совершить хотя бы одну ошибку первого рода будет не больше, чем α . Возникает вопрос, как этого добиться.

Единственный имеющийся в распоряжении инструмент — это уровни значимости $\alpha_1, \dots, \alpha_m$, на которых проверяются гипотезы H_1, \dots, H_m . Никаких других параметров в проверке гипотез нет. Ставится задача выбрать эти уровни так, чтобы обеспечить ограничение $\text{FWER} \leq \alpha$.

9.3.2. Поправка Бонферрони

Самый простой способ решить поставленную выше задачу — это использовать поправку Бонферрони. В методе Бонферрони достигаемые уровни значимости всех гипотез сравниваются с величиной $\frac{\alpha}{m}$:

$$\alpha_1 = \dots = \alpha_m = \frac{\alpha}{m}.$$

Альтернативный способ — преобразовать все достигаемые уровни значимости (p-value):

$$\tilde{p}_i = \min(1, mp_i).$$

Эти модифицированные достигаемые уровни значимости и будут сравниваться с исходным порогом α : H_i отвергается при $\tilde{p}_i \leq \alpha$. При такой процедуре точно так же контролируется величина FWER, как и при изменении порога.

Легко показать, что метод Бонферрони контролирует групповую вероятность ошибки первого рода на уровне α . Это будет единственная теорема в данном курсе.

Теорема Если все гипотезы $H_i, i = 1, \dots, m$ отвергаются при $p_i \leq \alpha/m$, то $\text{FWER} \leq \alpha$.

Доказательство По определению Familywise error rate (FWER) — это вероятность совершил хотя бы одну ошибку первого рода:

$$\text{FWER} = \mathbf{P}(V > 0) = \mathbf{P}\left(\bigcup_{i=1}^{m_0} \left\{p_i \leq \frac{\alpha}{m}\right\}\right).$$

Вероятность объединения событий можно оценить сверху через сумму вероятностей этих событий по неравенству Буля:

$$\mathbf{P}\left(\bigcup_{i=1}^{m_0} \left\{p_i \leq \frac{\alpha}{m}\right\}\right) \leq \sum_{i=1}^{m_0} \mathbf{P}\left(p_i \leq \frac{\alpha}{m}\right).$$

Далее можно воспользоваться свойством достигаемого уровня значимости:

$$\sum_{i=1}^{m_0} \mathbf{P}\left(p_i \leq \frac{\alpha}{m}\right) \leq \sum_{i=1}^{m_0} \frac{\alpha}{m} = \frac{m_0}{m} \alpha.$$

$m_0 < m$, следовательно

$$\frac{m_0}{m} \alpha \leq \alpha.$$

Исходное утверждение доказано.

9.3.3. Недостаток использования поправки Бонферрони

Те, кто когда-нибудь сталкивались с неравенством Буля, знают, что оценка вероятности объединения событий через сумму вероятностей этих событий очень завышенная. Действительно, чтобы получить в этом месте доказательства точное равенство, нужно вычесть вероятности всех возможных пересечений. Цепочка неравенств в доказательстве теоремы показывает, что при использовании метода Бонферрони FWER не просто меньше, чем α , а намного меньше, чем α . В идеале хочется, чтобы вероятность совершил хотя бы одну ошибку первого рода была в точности равна α . При использовании метода Бонферрони эта вероятность ограничивается гораздо более низкой величиной, чем α . Это плохо, потому что перестраховываясь в отношении ошибки первого рода, мы неизбежно совершаляем больше ошибок второго рода, то есть мощность такой статистической процедуры снижается.

9.3.4. Модельный эксперимент

Возьмем 50 выборок из нормального распределения $N(1, 1)$ и еще 150 — из стандартного нормального распределения $N(0, 1)$. Объем всех выборок $n = 20$.

На каждой из этих выборок проверяется гипотеза о равенстве среднего 0:

$$H_i: \mathbb{E}X_i = 0,$$

против двусторонней альтернативы

$$H'_i: \mathbb{E}X_i \neq 0$$

с помощью критерия Стьюдента.

	# верных H_i	# неверных H_i	Σ
# принятых H_i	142	0	142
# отвергнутых H_i	8	50	58
Σ	150	50	200

Таблица 9.5: Результаты эксперимента без поправки на множественную проверку

Если не делать никакой поправки на множественную проверку, в результате получится таблица 9.5. Видно, что отвергнуты все 50 неверных гипотез, но, к сожалению, вместе с ними отвергнуты еще и 8 верных, то есть совершилось 8 ошибок первого рода.

	# верных H_i	# неверных H_i	Σ
# принятых H_i	150	27	177
# отвергнутых H_i	0	23	23
Σ	150	50	200

Таблица 9.6: Результаты эксперимента при использовании поправки Бонферрони

Если делать поправку методом Бонферрони, гипотезы из второй строчки предыдущей таблицы перераспределяются в первую, результат — таблица 9.6. В этом случае ни одна верная нулевая гипотеза не отвергается, то есть нет ни одной ошибки первого рода. Но, к сожалению, вместе с этим исчезла возможность отвергнуть больше половины неверных нулевых гипотез: из 50 удалось отвергнуть только 23. То есть за гарантии в отношении ошибки первого рода пришлось отплатить тем, что найдено меньше неверных нулевых гипотез.

9.4. FWER. Метод Холма

9.4.1. Нисходящие методы

В методе Бонферрони уровни значимости для всех гипотез выбираются одинаковыми:

$$\alpha_1 = \dots = \alpha_m = \frac{\alpha}{m}.$$

Оказывается, если значения a_i брать не одинаковыми, а разными, можно достичь лучшего результата. Для того, чтобы это сделать, необходимо использовать нисходящую процедуру множественной проверки гипотез. В общем виде она выглядит так. Из достигаемых уровней значимости составляется вариационный ряд:

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)},$$

а все гипотезы переобозначаются так, чтобы их номера соответствовали номерам достигаемых уровней значимости в этом вариационном ряду:

$$H_{(1)}, H_{(2)}, \dots, H_{(m)}.$$

Дальше нужно самый маленький достигаемый уровень значимости $p_{(1)}$ сравнить с уровнем значимости α_1 . Если $p_{(1)} \geq \alpha_1$, то принимаются все нулевые гипотезы $H_{(1)}, H_{(2)}, \dots, H_{(m)}$, и процесс останавливается. $p_{(1)} < \alpha_1$, то отклоняется гипотеза $H_{(1)}$, и процедура продолжается. На втором шаге сравниваются $p_{(2)}$ и α_2 . Если $p_{(2)} \geq \alpha_2$, то принимаются все оставшиеся гипотезы $H_{(2)}, H_{(3)}, \dots, H_{(m)}$, и процедура завершается. Если нет — $H_{(2)}$ отвергается, процедура продолжается, и т.д.

Так в общем виде выглядит нисходящая процедура множественной проверки гипотез. Процедура называется нисходящей, несмотря на то что нулевые гипотезы перебираются по возрастанию. Это немного странно и может смущать, но идея заключается в том, что нулевые гипотезы отвергаются последовательно, начиная с наиболее значимых, то есть движение происходит по убыванию значимости.

9.4.2. Метод Холма

Метод Холма — это нисходящая процедура множественной проверки гипотез со следующими уровнями значимости:

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \dots, \alpha_i = \frac{\alpha}{m-i+1}, \dots, \alpha_m = \alpha.$$

Этот метод обеспечивает безусловный контроль над FWER. Это показать немного сложнее, чем для метода Бонферрони, поэтому доказательство здесь приведено не будет.

Вместо того, чтобы сравнивать исходные достигаемые уровни значимости с модифицированными α_i , можно их модифицировать и сравнивать с исходным порогом α . Так выглядит формула для модифицированных достигаемых уровней значимости метода Холма:

$$\tilde{p}_{(i)} = \min (1, \max ((m-i+1)p_{(i)}, \tilde{p}_{(i-1)}))$$

Метод Холма всегда мощнее, чем метод Бонферрони, то есть, он всегда отвергает не меньше гипотез, чем метод Бонферрони, потому что его уровни значимости всегда не меньше, чем из метода Бонферрони.

9.4.3. Модельный эксперимент

Для демонстрации работы метода Холма можно провести такой же модельный эксперимент с 200 гипотезами, как в предыдущем разделе.

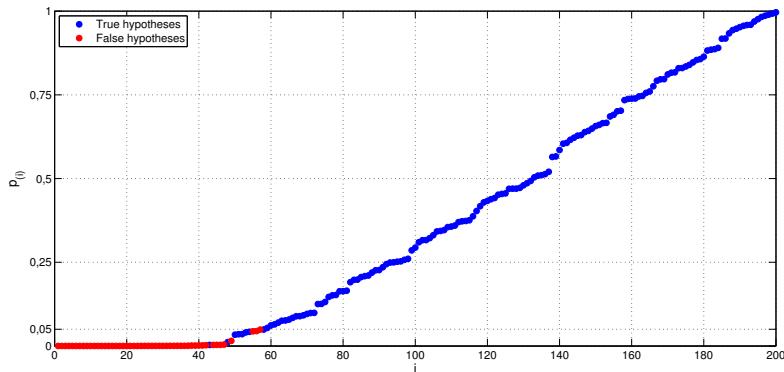
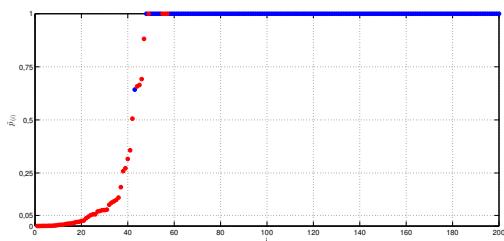
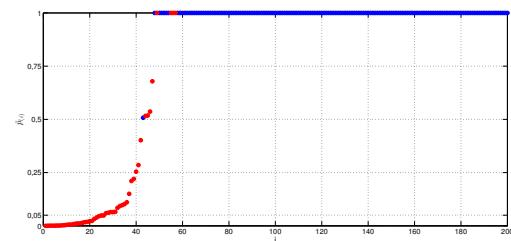


Рис. 9.6: Достигаемые уровни значимости гипотез

На графике (рисунок 9.6) показаны отсортированные достигаемые уровни значимости. По горизонтальной оси отложен номер в вариационном ряду, а по вертикальной оси — значения соответствующего достигаемого уровня значимости. Красные точки на графике — это неверные гипотезы, а синие точки — верные. Это типичный вид такого графика, на нём изображена как будто бы смесь двух треугольников: большого синего, соответствующего верным нулевым гипотезам, и маленького красного, соответствующего неверным. В месте соединения этих двух треугольников верные и неверные гипотезы смешиваются, и задача — где-то в этом месте правильно поставить порог, чтобы обеспечить теоретические гарантии на число ошибок первого рода.



(a) Модифицированные достигаемые уровни значимости гипотез при использовании поправки Бонферрони



(b) Модифицированные достигаемые уровни значимости гипотез при использовании метода Холма

Рис. 9.7

На рисунке 9.7 слева показаны модифицированные достигаемые уровни значимости, отсортированные по неубыванию, при использовании поправки Бонферрони. Результаты были приведены в таблице 9.6. Отвергаются 23 неверные нулевые гипотезы из 50 и ни одной верной.

	# верных H_i	# неверных H_i	Σ
# принятых H_i	150	24	174
# отвергнутых H_i	0	26	26
Σ	150	50	200

Таблица 9.7: Результаты проверки гипотез при использовании метода Холма

Модифицированные достигаемые уровни значимости метода Холма показаны на рисунке 9.7 справа. Этот метод позволил отвергнуть 26 из 50 гипотез и всё ещё не совершил ни одной ошибки первого рода при этом (таблица 9.7).

С одной стороны, разница между методами Холма и Бонферрони. Метод Холма не помог случиться чуду и не отверг все неверные нулевые гипотезы. С другой стороны, этот метод позволил, не делая никаких дополнительных предположений, совершить ещё три научных открытия. Ещё три гипотезы удалось отвергнуть абсолютно бесплатно. А это уже достаточный повод, чтобы пользоваться этим методом.

9.5. FDR. Метод Бенджамини-Хохберга

9.5.1. False discovery rate

В описанных ранее поправках при множественном проверке гипотез контролировалась величина групповой вероятности ошибки, то есть ограничивалась вероятность совершить хотя бы одну ошибку первого рода:

$$\text{FWER} = P(V > 0).$$

В некоторых ситуациях, например, когда проверяются десятки тысяч или миллионы гипотез, можно допустить какое-то количество ошибок первого рода ради того, чтобы увеличить мощность процедуры и отвергнуть больше неверных гипотез, то есть совершить меньше ошибок второго рода. В таких ситуациях выгоднее использовать другую меру: не familywise error rate, а false discovery rate, ожидаемую долю ложных отклонений:

$$\text{FDR} = \mathbb{E} \left(\frac{V}{\max(R, 1)} \right).$$

Для любой фиксированной процедуры множественной проверки гипотез $\text{FDR} \leq \text{FWER}$. За счет этого, если контролировать FDR, а не FWER, получается более мощная процедура, поскольку она позволяет отвергать больше гипотез.

9.5.2. Восходящие методы

Методы, которые контролируют FDR, как правило, восходящие. В каком-то смысле это противоположность нисходящих методов (таких как метод Холма), которые рассматривались до этого.

Восходящие методы работают с тем же самым вариационным рядом достигаемых уровней значимости, что и нисходящие:

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}.$$

Отличие заключается в том, что процедура начинается с другого конца этого ряда. На первом шаге самый большой p-value p_m сравнивается с соответствующей ему константой α_m . Если $p_{(m)} \leq \alpha_m$, то все нулевые гипотезы $H_{(1)}, H_{(2)}, \dots, H_{(m)}$ отвергаются, и процедура останавливается. Иначе гипотеза $H_{(m)}$ принимается, и процедура продолжается. На следующем шаге сравниваются $p_{(m-1)}$ и α_{m-1} . Если $p_{(m-1)} \leq \alpha_{m-1}$, то все нулевые гипотезы $H_{(1)}, H_{(2)}, \dots, H_{(m-1)}$ отвергаются, и процедура останавливается. Иначе принимается гипотеза $H_{(m-1)}$, процедура продолжается. И так далее.

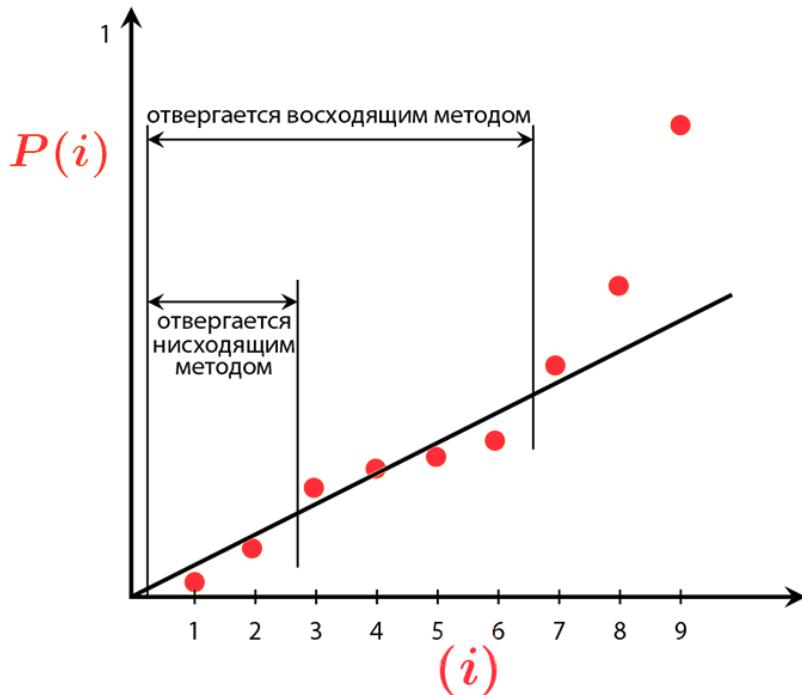


Рис. 9.8: Гипотезы, отвергаемые восходящими и нисходящими методами. Каждая точка — это гипотеза, по вертикальной оси показан соответствующий ей достигаемый уровень значимости, по горизонтальной — её номер в вариационном ряду. Линия — порог, с которым сравнивается достигаемый уровень значимости гипотез.

Если для одних и тех же α_i построить восходящую и нисходящую процедуру, то восходящая процедура всегда будет отвергать не меньше гипотез, чем нисходящая. На рисунке 9.8 показано, какие гипотезы отвергаются восходящими и нисходящими методами. При использовании восходящей процедуры движение происходит от самого большого p-value к самому маленькому. В таком случае принимаются гипотезы H_9, H_8, H_7 . Если используется нисходящая процедура, то наоборот, всё начинается с самого маленького p-value, и первые две гипотезы отвергаются, а оставшиеся семь — принимаются. Таким образом, в этом примере восходящая процедура отвергла в три раза больше гипотез, чем нисходящая. Это может происходить из-за того, что линия, соединяющая отсортированные достигаемые уровни значимости, может несколько раз пересекать прямую, задающую критические значения α .

9.5.3. Метод Бенджамини-Хохберга

Для контроля над FDR чаще всего используется метод Бенджамини-Хохберга. Это восходящая процедура с уровнями значимости

$$\alpha_1 = \frac{\alpha}{m}, \dots, \alpha_i = \frac{\alpha i}{m}, \dots, \alpha_m = \alpha.$$

Крайние уровни значимости точно также, как и в методе Холма, а вот между ними — абсолютно другие. В методе Бенджамини-Хохберга уровни значимости между α_1 и α_m меняются линейно, в то время как в методе Холма — по гиперболе.

Модифицированные достигаемые уровни значимости для метода Бенджамини-Хохберга выглядят следующим образом:

$$\tilde{\alpha}_{(i)} = \min \left(1, \frac{mp_{(i)}}{i}, \tilde{\alpha}_{(i+1)} \right).$$

Процедура восходящая, и каждый следующий p-value в ней не должен стать больше, чем предыдущий, поэтому берётся минимум из $\frac{mp_{(i)}}{i}$ и $\tilde{\alpha}_{(i+1)}$ (а также 1, поскольку это вероятность).

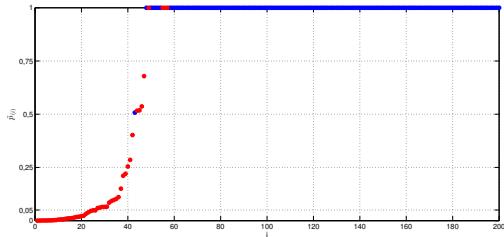
Метод Бенджамини-Хохберга обеспечивает контроль над FDR на уровне α только при условии независимости статистик, которые проверяют гипотезы. Это требование достаточно сильное. Иногда его можно ослабить, и в некоторых задачах выполняется ослабленное требование. Тем не менее, важно подчеркнуть, что процедура Бенджамини-Хохберга не является универсальной и она не применима безусловно, в отличие от метода Холма.

	# верных H_i	# неверных H_i	Σ
# принятых H_i	148	4	152
# отвергнутых H_i	2	46	48
Σ	150	50	200

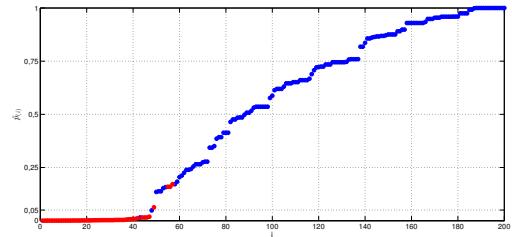
Таблица 9.8: Результаты проверки гипотез при использовании метода Бенджамини-Хохберга

9.5.4. Модельный эксперимент

Давайте протестируем метод Бенджамини-Хохберга на модельных данных из предыдущего раздела.



(a) Модифицированные уровни значимости при использовании метода Холма



(b) Модифицированные уровни значимости при использовании метода Бенджамини-Хохберга

Рис. 9.9

На рисунке 9.9 показаны отсортированные модифицированные достижимые уровни значимости, полученные методами Холма и Бенджамини-Хохберга. Они отличаются довольно сильно. Метод Холма отвергает 26 неверных гипотез, при этом не совершается ни одной ошибки первого рода (таблица ??). Метод Бенджамини-Хохберга отвергает 46 неверных гипотез (таблица 9.8), при этом совершается две ошибки первого рода ($\frac{2}{48} \approx 0.04 < 0.05$). Метод Бенджамини-Хохберга применим, так как в модельном эксперименте выборки генерировались независимо.

Итак, если контролировать FDR вместо FWER, допускается больше ошибок первого рода, но за счет этого можно критически увеличить количество отвергаемых неверных нулевых гипотез. Метод Бенджамини-Хохберга используется повсеместно, несмотря на то, что он работает далеко не всегда. Очень часто его применяют без проверки необходимого условия корректности, так делать не стоит.

9.6. Анализ подгрупп

Эффект множественной проверки гипотез ярко проявляется при анализе подгрупп. Для примера можно рассмотреть следующее исследование, в котором принимают участие 1073 пациента с ишемической болезнью сердца¹. Их делят на 2 подгруппы (в зависимости от типа лечения) и исследуют взаимосвязь между выживаемостью и типом лечения. Требуется понять, какой их двух типов лечения лучше.

Важные факторы, которые влияют на выживаемость при ишемической болезни сердца, — это число пораженных артерий, (может быть 1, 2 или 3), и тип сокращений левого желудочка (нормальный и аномальный). В таких ситуациях исследователи часто хотят посмотреть на сравнительную эффективность типов лечения отдельно во всех подгруппах по уровням важных факторов. В данном случае два фактора порождают 6 подгрупп, в каждой из них сравнивается выживаемость пациентов по двум типам лечения.

¹Ссылка на указанное исследование: Lee K.L., McNeer J.F., Starmer C.F., Harris P.J., Rosati R.A. (1980). Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation, 61(3), 508–515.

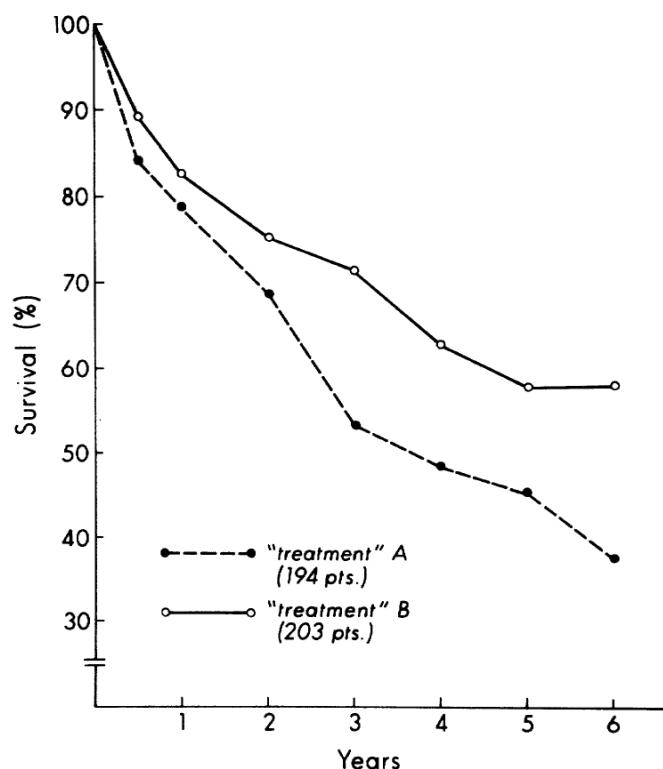


Рис. 9.10: Выживаемость людей в одной из подгрупп при разных типах лечения

В одной из 6 подгрупп были обнаружены значимые различия в выживаемости пациентов при лечении первого типа и второго. На рисунке 9.10 показаны кривые выживаемости пациентов этих подгрупп. По ним видно, что в группе лечения А после 6 лет наблюдений выжило меньше 40% пациентов, а в группе, соответствующей лечению В, — в районе 60%. Это различие статистически значимо. Кажется, что для пациентов с таким числом пораженных артерий и с таким типом сокращения левого желудочка лечение В действительно существенно эффективнее.

На самом деле эти два лечения отличаются только названием, и две группы пациентов лечились абсолютно одинаково. Эта статья была написана с целью показать необходимость поправки на множественную проверку гипотез при анализе подгрупп. Действительно, при сравнении кривых выживаемости во всех 6 подгруппах, проверяются 6 абсолютно независимых гипотез, и возникает эффект множественной проверки. Если подгрупп достаточно много, в результате будут всегда получаться какие-то значимые отклонения. Есть и исследование, в котором такая ошибка в анализе подгрупп была совершена на самом деле, — это исследование 2008 года, в котором изучалась связь между употреблением кофеина и риском возникновения рака груди². В этой статье всего было около 50 разных подгрупп, по самым разным уровням различных факторов. В частности, было показано, что употребление более чем четырех чашек кофе в день связано с увеличением риска злокачественного рака груди (достигаемый уровень значимости $p = 0.08$). Это больше, чем стандартный уровень значимости 0.05, но меньше чем либеральный уровень значимости 0.1. Кроме того, потребление кофеина связано с увеличением риска возникновения эстроген- и прогестерон- независимых опухолей, а также опухолей размером больше 2 сантиметров (достигаемый уровень значимости $p = 0.02$). Еще одно открытие: потребление кофе без кофеина связано со снижением риска возникновения рака груди у женщин в постменопаузе, принимающих гормоны (достигаемый уровень значимости $p = 0.02$).

Ясно, что за счет большого количества рассматриваемых подгрупп можно всегда получить какие-то значимые отклонения, если не делать поправку на множественную проверку. Какие-то из этих открытий с большой вероятностью окажутся ложными. В каком-то смысле это напоминает переобучение: оценивается эффективность лечения в разных подгруппах в зависимости от каких-то признаков пациента, и если эти признаки слишком сложные и их слишком большое количество, то происходит переобучение под анализируемую выборку. В качестве экстремального примера такого переобучения можно вспомнить цитату из Галена, II века

²Ссылка на указанное исследование: Ishitani K., Lin J. (2008). Caffeine consumption and the risk of breast cancer in a large prospective cohort of women. Archives of Internal Medicine, 168(18), 2022–2031.

до нашей эры: «*Все больные, принявшие это средство, вскоре выздоровели, за исключением тех, кому оно не помогло — они умерли. Отсюда очевидно, что средство помогает во всех случаях, кроме безнадежных*».

	Контроль (100)	Больные (100)	p
Мутация	1 из 100	8 из 100	0.0349
Фамилия начинается с гласной	36 из 100	40 из 100	0.6622

Таблица 9.9: Данные гипотетического эксперимента

В заключение обсуждения эффекта множественной проверки гипотез давайте рассмотрим ещё один гипотетический пример. Пусть есть 100 больных людей и 100 здоровых, и хочется понять, есть ли связь между болезнью и какой-то мутацией. В контрольной выборке из 100 человек мутация есть у одного, а в выборке больных — у 8 (таблица 9.9). По всей видимости, эта мутация достаточно редкая. Если сравнить доли людей с мутацией в выборках больных и здоровых, получится достигаемый уровень значимости $p = 0.03$, и гипотеза об отсутствии связи между мутацией и болезнью отвергается.

Пусть теперь выдвигается еще одна гипотеза: наличие заболевания связано с тем, с гласной или согласной буквы у пациентов начинаются фамилии. В контрольной выборке здоровых людей у 36 человек фамилия начинается с гласной буквы, а в выборке больных — у 40 из 100 (таблица 9.9). При сравнении этих долей биномиальным критерием получается достигаемый уровень значимости 0.66, гипотеза отклонена не будет. Проблема, однако, заключается в том, что теперь в исследовании проверяются две гипотезы, и необходимо делать поправку на множественность этой проверки.

Какой бы при этом ни использовался метод поправки, будь то метод Бонферрони, Холма или Бенджамина Хохберга, самый маленький достигаемый уровень значимости во всех них сравнивается с $\frac{\alpha}{m}$. Таким образом, если требуется обеспечить контроль над какой-то мерой числа ошибок первого рода на уровне 0.05, нужно сравнивать самое маленькое значение достигаемого уровня значимости с 0.025. Самый маленький достижимый уровень значимости в этом исследовании $p = 0.03$. Получается, эта нелепая гипотеза, которая была введена в исследование, замаскировала, возможно, неверную нулевую гипотезу, связанную с мутацией.

Отсюда вытекает рецепт лучшего способа борьбы с эффектом множественной проверки гипотез: проверять меньше гипотез. Необходимо до начала исследования подумать, какие из возможных гипотез на самом деле не представляют интереса, и отказаться от их рассмотрения. За счет этого появится возможность сделать более либеральную поправку на множественность и отвергнуть больше действительно неверных гипотез, совершив больше действительно интересных открытий. Важно, что такая фильтрация гипотез должна осуществляться именно до сбора данных. Если выбрасывать гипотезы уже после того, как стали известны достижимые уровни значимости, возникнет эффект переобучения.