Прикладная статистика

Слайды к лекции 3

31 января 2023

Сергей Александрович Спирин

sspirin@hse.ru

Математическая статистика

Проверка гипотез

Проверка гипотез

- H_0 – нулевая гипотеза
- H_A (или H_1) – альтернативная гипотеза

В суде:
H_0: Человек не виновен
H_A: Человек виновен

В науке:
H_0: Эффекта нет
H_A: Эффект есть

На приеме врача:
H_0: Пациент болен
H_A: Пациент здоров

Ошибки I/II рода

- Ошибки I рода **(α)**
- Это ошибки отказа от нулевой гипотезы, когда она верна (*признать виновным невиновного при презумпции невиновности, объявить об открытии при отсутствии эффекта, принять больного за здорового*)

- Ошибки II рода **(β)**
- Это ошибки неотклонения нулевой гипотезы, когда она на самом деле ложная (*признать виновного невиновным, не заметить эффекта, начать лечить здорового*)

Ошибки I/II рода

- Ошибки I рода **(α)**
- Это ошибки отказа от нулевой гипотезы, когда она верна (*признать виновным невиновного при презумпции невиновности, объявить об открытии при отсутствии эффекта, принять больного за здорового*) – более тяжёлые по последствиям

- Ошибки II рода **(β)**
- Это ошибки неотклонения нулевой гипотезы, когда она на самом деле ложная (*признать виновного невиновным, не заметить эффекта, начать лечить здорового*) – неприятно, но не настолько опасно

Решающее правило

- Выборка – набор чисел (или несколько наборов, или набор объектов с несколькими числовыми характеристиками у каждого)
- **Решающее правило** (критерий, тест) отвергает или принимает нулевую гипотезу для каждой мыслимой выборки
- Иногда мы будем совершать ошибку первого рода
- Иногда мы будем совершать ошибку второго рода
- (Конечно, много раз мы будем правы)

P-значение (P-value)

- Если бы то, что мы предполагаем в нулевой гипотезе, было верно, то какова была бы вероятность видеть то, что мы видим в выборке (это, или еще «хуже»)?

Пример

Известно, что длина клыка глазункианского дикогрыза распределена нормально со средним 1,8 см и средним квадратичным отклонением 2 мм. Вы путешествуете по Глазункии и нашли чей-то зуб, по внешним признакам подходящий под описание клыка дикогрыза, но длиной 2,3 см. Оцените p-значение утверждения, что это зуб какого-то другого животного.

Пример

Известно, что длина клыка глазункианского дикогрыза распределена нормально со средним 1,8 см и средним квадратичным отклонением 2 мм. Вы путешествуете по Глазункии и нашли чей-то зуб, по внешним признакам подходящий под описание клыка дикогрыза, но длиной 2,3 см. Оцените p-значение утверждения, что это зуб какого-то другого животного.

- Нулевая гипотеза – число 2,3 пришло из нормального распределения со средним 1,8 и дисперсией 0,04.
- p-значение равно вероятности получить из такого распределения число, большее или равное 2,3
- считаем Z = (2,3 – 1,8) / 0,2 = 2,5

(в данном случае Z – это наша **статистика**, то есть величина, распределение которой при нулевой гипотезе нам известно)

- не глядя в таблицу, знаем, что $p < 0{,}02$
- поглядев в таблицу, оцениваем $p \approx 0{,}0062$

Пример

Известно, что длина клыка глазункианского дикогрыза распределена нормально со средним 1,8 см и средним квадратичным отклонением 2 мм. Вы путешествуете по Глазункии и нашли чей-то зуб, по внешним признакам подходящий под описание клыка дикогрыза, но длиной 2,3 см. Оцените p-значение утверждения, что это зуб какого-то другого животного.

- Нулевая гипотеза – число 2,3 пришло из нормального распределения со средним 1,8 и дисперсией 0,04.
- p-value равно вероятности получить из такого распределения число, большее или равное 2,3
- считаем $Z = (2{,}3 - 1{,}8) / 0{,}2 = 2{,}5$

(в данном случае Z – это наша **статистика**, то есть величина, распределение которой при нулевой гипотезе нам известно)

- не глядя в таблицу, знаем, что $p < 0{,}02$
- поглядев в таблицу, оцениваем $p \approx 0{,}0062$

Почему правильное p-value в этом примере вдвое больше?

Пример

Рыболову известно, что в некотором месте в определённый сезон среднее время ожидания поклёвки составляет 25 минут. Рыболов дождался сезона и пришёл в то же место, ждёт уже час, а поклёвки нет. Каково p-значение утверждения, что в этом году рыба клюёт хуже?

Пример

Рыболову известно, что в некотором месте в определённый сезон среднее время ожидания поклёвки составляет 25 минут. Рыболов дождался сезона и пришёл в то же место, ждёт уже час, а поклёвки нет. Каково p-значение утверждения, что в этом году рыба клюёт хуже?

- Нулевая гипотеза – число 60 минут пришло из экспоненциального распределения со средним λ = 25 мин.
- p-значение равно вероятности получить из такого распределения число, большее или равное 60
- считаем $p = \exp(-60/\lambda) \approx 0{,}09$

Пример

В июне прошлого года на берегу Гнилого озера было не так много комаров: в среднем один комар прилетал каждые три минуты. В июне этого года вы пришли на Гнилое озеро и уже через 10 сек. на вас сел комар. Каково p-значение утверждения, что комаров в этом году больше?

Пример

В июне прошлого года на берегу Гнилого озера было не так много комаров: в среднем один комар прилетал каждые три минуты. В июне этого года вы пришли на Гнилое озеро и уже через 10 сек. на вас сел комар. Каково p-значение утверждения, что комаров в этом году больше?

- Нулевая гипотеза – число 10 сек. пришло из экспоненциального распределения со средним $\lambda = 180$ сек.
- p-значение равно вероятности получить из такого распределения число, меньшее или равное 10
- считаем $p = 1 - \exp(-10/\lambda) \approx 0{,}054$

Решающее правило

- Решающее правило = функция от *n* чисел со значениями «принять» или «отвергнуть».
- В математической статистике выборка – *n* одинаково распределённых независимых случайных величин (поэтому принятие нулевой гипотезы становится случайным событием)
- Часто нулевая гипотеза является **простой**, то есть состоит в том, что эти с.в. распределены по некоторому заранее известному закону.
- Можно определить вероятность ошибки I рода: принимаем нулевую гипотезу, подсчитываем вероятность ответа «отвергнуть».

Решающее правило

Обычная формула решающего правила:

- Определяется **статистика**, то есть функция от наблюдений (выборки) с числовыми значениями.
- Задаётся **порог**, то есть некоторое число.
- Решающее правило:
  - отвергнуть нулевую гипотезу, если значение статистики больше порога,
  - принять в противном случае.

- В более общем случае задаётся **критическое множество** для значений статистики (например, множество всех значений, по модулю больших некоторого порога)

P-значение

- Если бы нулевая гипотеза была верна, то какова была бы вероятность видеть то, что мы видим в выборке (это, или еще «хуже»)?
- Чтобы определить p-значение, нужна только статистика (без порога).
- Вместо порога на статистику часто задаётся порог на p-значение
- Малое p-значение показывает, что вы видите что-то очень необычное с точки зрения H_0

Пример
(простой парный критерий, критерий знаков, Sign test)

Рассмотрели 769 ортологичных пар белков *E.coli* и *B.subtilis*.

В 422 парах полипептидная цепь белка из *E.coli* оказалась длиннее, а в 311 случаях — короче.

Каково p-значение утверждения, что между бактериями имеется систематическое различие в длинах полипептидных цепей ортологичных белков?

Пример
(простой парный критерий, критерий знаков, Sign test)

В 422 парах полипептидная цепь белка из *E.coli* оказалась длиннее, а в 311 случаях — короче.

Каково p-значение утверждения, что между бактериями имеется систематическое различие в длинах полипептидных цепей ортологичных белков?

- Нулевая гипотеза — большая и меньшая длины цепи равновероятны, а наблюдаемое различие случайно.
- Вероятность в 733 испытаниях с двумя равновероятными исходами получить 311 **или меньше** наблюдений одного из исходов равна примерно 0,0000235.
- Чтобы получить p-значение , надо умножить это число на 2, поскольку мы заранее не знали, какой исход получит преимущество.
- Ответ: P = 0,000047

Ошибки I/II рода

	Фактическое состояние	
	Виновен	Не виновен
Тест показывает: "виновен"	True Positive	False Positive **Type I Error**
Тест показывает: "не виновен"	False Negative **Type II Error**	True Negative

Проверка гипотез

- Р-значение зависит **от статистики и выборки**
- Уровень значимости **α** вместе со статистикой порождают **решающее правило**:
  H_0 отклоняется, если р-значение < **α**
- **α** представляет собой вероятность ошибки I рода для такого решающего правила

Проверка гипотез

- P-значение зависит **от статистики и выборки**
- Уровень значимости **α** вместе со статистикой порождают **решающее правило**:
  H_0 отклоняется, если p-значение < **α**
- **α** представляет собой вероятность ошибки I рода для такого решающего правила
  Точнее, эта вероятность не превосходит **α**

Вероятность ошибки I рода

$$\mathbf{P}(\text{ошибки I рода}) = \mathbf{P}(\text{статистика в крит. области} \mid H_0) \cdot \mathbf{P}(H_0) =$$
$$= \mathbf{P}(\text{P-value} < \boldsymbol{\alpha} \mid H_0) \cdot \mathbf{P}(H_0) \leq \mathbf{P}(\text{P-value} < \boldsymbol{\alpha} \mid H_0) = \boldsymbol{\alpha}$$

Вероятность ошибки первого рода **не больше** уровня значимости.

Она равна уровню значимости, умноженной на априорную вероятность нулевой гипотезы **P**(H_0).

Но **P**(H_0) практически никогда не известна.

Вероятность ошибки I рода

$$\mathbf{P}(\text{ошибки I рода}) = \mathbf{P}(\text{статистика в крит. области} \mid H_0) \cdot \mathbf{P}(H_0) =$$
$$= \mathbf{P}(\text{P-value} < \boldsymbol{\alpha} \mid H_0) \cdot \mathbf{P}(H_0) \leq \mathbf{P}(\text{P-value} < \boldsymbol{\alpha} \mid H_0) = \boldsymbol{\alpha}$$

Вероятность ошибки первого рода **не больше** уровня значимости.

Она равна уровню значимости, умноженной на априорную вероятность нулевой гипотезы **P**(H_0).

Но **P**(H_0) практически никогда не известна.
Хотя бывает и известна, например, для болезни с хорошо известной частотой в данной группе больных

Статистика, P-value, мощность

- Статистикой может быть **любая** функция от наблюдений
- Статистику обычно пересчитывают в p-значение по некоторой формуле.

Р-значение – это тоже статистика, но не единственная! Любую статистику можно пересчитать в p-значение, но результат для разных статистик будет разный.

- С одной и той же статистикой при разных критических множествах могут получаться разные решающие правила

*А для разных **семейств** критических множеств из одной и той же статистики будут получаться разные p-значения.*

- Каждое решающее правило имеет две основных характеристики: уровень значимости **α** и мощность.

Мощность – это вероятность отклонить H_0 при условии её несправедливости, то есть единица минус вероятность ошибки **второго** рода.

- Чем мощнее критерий при заданном **α**, тем лучше. Но численно оценить мощность в большинстве случаев невозможно.
- Хотя статистикой может быть и любая функция от наблюдений, но мощность получаемых критериев очень зависит от того, какая статистика выбрана.

Обзор задач (проверка гипотез)

- Сравнение двух выборок
  - ✓ t-тест (Стьюдента) — равны ли средние?
  - ✓ F-тест (Фишера) — равны ли дисперсии?
  - ✓ R-тест (суммы рангов Вилкоксона) = U-тест (Манна – Уитни) — есть ли систематический сдвиг?
  - ✓ Критерий Смирнова (Колмогорова – Смирнова) — совпадают ли распределения?
- Сравнение многих выборок
  - ✓ ANOVA, H-тест (Краскела — Уоллиса)
- Парные наблюдения
  - ✓ Простой парный тест (критерий знаков)
  - ✓ W-тест (Вилкоксона)
  - ✓ Сравнение средней разности с нулём: z-тест, t-тест
- Сравнение выборки с заданным распределением
  - ✓ Сравнение среднего с ожидаемым — z-тест, t-тест
  - ✓ Полиномиальный тест
  - ✓ Критерий согласия хи-квадрат (Пирсона) — приближение полиномиального теста для больших выборок
  - ✓ Критерий Колмогорова
- Таблица сопряжённости 2×2
  - ✓ Точный критерий Фишера, критерий хи-квадрат (приближение для больших выборок)
- Сравнение чисел наблюдений
  - ✓ Число успехов при заданном числе испытаний (сравнение частот)
  - ✓ Число успехов при неопределённом максимуме
- Сравнение коэффициента корреляции с нулём

Таблица стандартного нормального распределения

	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
-5	$2{,}9\cdot10^{-7}$	$4{,}8\cdot10^{-7}$	$7{,}9\cdot10^{-7}$	$1{,}3\cdot10^{-6}$	$2{,}1\cdot10^{-6}$	$3{,}4\cdot10^{-6}$	$5{,}4\cdot10^{-6}$	$8{,}5\cdot10^{-6}$	$1{,}3\cdot10^{-5}$	$2{,}1\cdot10^{-5}$
-4	$3{,}2\cdot10^{-5}$	$4{,}8\cdot10^{-5}$	$7{,}2\cdot10^{-5}$	$1{,}1\cdot10^{-4}$	$1{,}6\cdot10^{-4}$	$2{,}3\cdot10^{-4}$	$3{,}4\cdot10^{-4}$	$4{,}8\cdot10^{-4}$	$6{,}9\cdot10^{-4}$	$9{,}7\cdot10^{-4}$
-3	0,0013	0,0019	0,0026	0,0035	0,0047	0,0062	0,0082	0,011	0,014	0,018
-2	0,023	0,029	0,036	0,045	0,055	0,067	0,081	0,097	0,12	0,14
-1	0,16	0,18	0,21	0,24	0,27	0,31	0,34	0,38	0,42	0,46

$F(-1{,}96) \approx 0{,}025$

$F(-1{,}65) \approx 0{,}05$

Сравнение чисел событий

В первом эксперименте имеем n успехов из N независимых испытаний, во втором — m успехов из M независимых испытаний. Значима ли разница?

Сравнение чисел событий

В первом эксперименте имеем n успехов из N независимых испытаний, во втором — m успехов из M независимых испытаний. Значима ли разница?

Если n и $N - n$ очень большие (сотни), то можно поступить так:
• сначала оценить вероятность успеха как $p = n/N$
• затем, если $m/M < p$, то оценить вероятность получить $\leq m$ успехов при M испытаниях и вероятности успеха p
• если полученная вероятность меньше 0,05, считаем разницу значимой

Сравнение чисел событий

В первом эксперименте имеем n успехов из N независимых испытаний, во втором — m успехов из M независимых испытаний. Значима ли разница?

Пусть все числа n, m, $N - n$, $M - m$ достаточно велики
(все больше 5 и не более одного меньше 10 *).

Тогда число успехов, а значит и долю успехов можно считать распределёнными нормально.

При вероятности успеха p математическое ожидание числа успехов при **одном** испытании равно p, а дисперсия $p(1 - p)$.

Значит, матожидание числа успехов при N испытаниях равно Np, а дисперсия $Np(1 - p)$,
а для **доли** успехов матожидание p, дисперсия $p(1 - p)/N$

Оценим $p = (n + m)/(N + M)$. Тогда имеем для доли успехов:
– матожидание равно p независимо от числа испытаний,
– дисперсия при N испытаниях $D_1 = p(1 - p)/N$, при M испытаниях $D_2 = p(1 - p)/M$.

Разность двух нормальных распределений распределена тоже нормально, при этом матожидания вычитаются, а дисперсии складываются. Поэтому **разность долей** распределена нормально со средним 0 и дисперсией $D_1 + D_2$.

Статистика $Z = (n/N - m/M)/(D_1 + D_2)^{1/2}$ при нулевой гипотезе распределена как N(0,1)

* – разные есть мнения на этот счёт. Кажется, все согласны, что если хотя бы одно из чисел — 5 или меньше, нормальное приближение применять нельзя.

Сравнение чисел событий

Испытывается новая методика лечения овец, заражённых некоторым заболеванием. Из 50 овец, которых лечили старым методом, умерли 25, а из 60, леченных новым методом — 20. Как посчитать P-value утверждения, что новый метод действительно лучше?

Сравнение чисел событий

Испытывается новая методика лечения овец, заражённых некоторым заболеванием. Из 50 овец, которых лечили старым методом, умерли 25, а из 60, леченных новым методом — 20. Как посчитать P-value утверждения, что новый метод действительно лучше?

По описанной процедуре получаем Z = –1,77
Это меньше, чем –1,65
Поэтому при стандартном уровне достоверности 0,05 можно сделать вывод, что новый метод лучше.

Можно посчитать, что P-value = *F*(Z) = 0,038

Сравнение чисел событий

(двусторонняя альтернатива)

(здесь и далее фиксируем уровень значимости $\alpha = 0,05$)

Пример из книги Н.Бейли “Статистические методы в биологии”

Сравниваем всхожесть семян шпината при обработке двумя методами: А и В.

Из 80 семян, обработанных методом А, взошло 65.

Из 90 семян, обработанных методом В, взошло 80.

Вывод?

Сравнение чисел событий

(двусторонняя альтернатива)

(здесь и далее фиксируем уровень значимости $\alpha = 0{,}05$)

Пример из книги Н.Бейли "Статистические методы в биологии"

Сравниваем всхожесть семян шпината при обработке двумя методами: А и В.

Из 80 семян, обработанных методом А, взошло 65.

Из 90 семян, обработанных методом В, взошло 80.

Вывод?

Считаем Z так же, но сравниваем не с 1,65, а с 1,96
(потому что исходно методы равноправны)

Сравнение чисел событий

(двусторонняя альтернатива)

(здесь и далее фиксируем уровень значимости $\alpha = 0{,}05$)

Пример из книги Н.Бейли "Статистические методы в биологии"

Сравниваем всхожесть семян шпината при обработке двумя методами: А и В.

Из 80 семян, обработанных методом А, взошло 65.

Из 90 семян, обработанных методом В, взошло 80.

Вывод?

Решение.

Всхожесть: 0,812 для метода А и 0,889 для метода В, разность = **0,077**

$p = 145/170 \approx 0{,}853$

Дисперсия для всхожести А: $D_A = p(1-p)/80 = 0{,}00157$

Дисперсия для всхожести В: $D_B = p(1-p)/90 = 0{,}00139$

Дисперсия для разности равна сумме дисперсий:

$D = D_A + D_B = 0{,}00296 \Rightarrow \sigma \approx \mathbf{0{,}054}$

$Z \approx$ **1,43** — вывода сделать нельзя (разница незначима, нужны новые эксперименты).

Сравнение чисел событий

(при неопределённом числе испытаний)

Мальчик поймал за сезон 50 рыжих майских жуков и 33 чёрных. Можно ли утверждать, что рыжие встречаются в этом году чаще?

Сравнение чисел событий

(при неопределённом числе испытаний)

Мальчик поймал за сезон 50 рыжих майских жуков и 33 чёрных.
Можно ли утверждать, что рыжие встречаются в этом году чаще?

Решение.
При нулевой гипотезе (рыжих и чёрных поровну) среднее число пойманных жуков каждого цвета следует оценить как
$(50 + 33)/2 = 41{,}5$
Дисперсия каждого из двух наблюдений тоже равна 41,5
(почему?)
Дисперсия разности равна сумме дисперсий, то есть 83.
Поэтому $Z = (50 - 33)/\sqrt{83} \approx 1{,}87 < 1{,}96$
Опять-таки вывода сделать нельзя.

То же, при односторонней альтернативе

Мальчику сказали, что в местности, где он живёт, рыжие майские жуки встречаются чаще чёрных.
Мальчик решил это проверить, дождался мая, каждый вечер ловил майских жуков и поймал в общей сложности 50 рыжих жуков и 33 чёрных.
Можно ли утверждать, что мальчик на своём опыте подтвердил сообщённые ему сведения?

То же, при односторонней альтернативе

Мальчику сказали, что в местности, где он живёт, рыжие майские жуки встречаются чаще чёрных.
Мальчик решил это проверить, дождался мая, каждый вечер ловил майских жуков и поймал в общей сложности 50 рыжих жуков и 33 чёрных.
Можно ли утверждать, что мальчик на своём опыте подтвердил сообщённые ему сведения?

Ответ: да!
Z, как и в предыдущей задаче, получается равным 1,87
Но теперь имеется односторонняя альтернатива (мы заранее проверяли гипотезу, что рыжих больше).
Поэтому (для уровня 0,05) критическим значением является не 1,96 , а 1,65

В предыдущей задаче критическое множество: $|Z| > 1{,}96$
В этой задаче критическое множество: $Z > 1{,}65$
Вероятности попадания в эти множества одинаковы и равны 0,05

Сравнение чисел событий

При посеве на питательную среду проб воды из водоёма А выросло 49 колоний бактерий, а при посеве (такого же количества и объёма) проб из водоёма Б — 32 колонии. На основании этих данных было сделано утверждение, что в первом водоёме загрязнённость бактериями выше. Как посчитать p-значение этого утверждения?

Сравнение чисел событий

Возникло предположение, что очистные сооружения некоторого небольшого города не полностью очищают воду от бактерий. При посеве пробы воды, взятой выше города, выросло 23 колонии, а пробы, взятой ниже города — 41 колония. Как посчитать уровень значимости, на котором данные результаты подтверждают предположение?