# Прикладная статистика, НИУ ВШЭ / Домашняя работа 2

Студент - Катерина Олейникова

7 Февраля 2022

*Примечание* Для решения большинства заданий был использован пакет stats из R.

**Задача 1** В лесу случайным образом было выбрано 7 участков одинаковой площади. На каждом участке было посчитано число взрослых сосен, росших на нём. Эти числа оказались такими: 7, 12, 9, 17, 10, 13, 15. Существенно ли варьирует число сосен?

Решение.

Сначала примем нулевую и альтернативную гипотезы:

H0: Вариативность количества деревьев на участках незначительна (распределение деревьев равномерное).
HA: Вариативность количества деревьев на участках значительна (распределение деревьев неравномерное).

Для того, чтобы определить наличие вариативности количества деревьев, используем критерий хи-квадрат (индивидуальных наблюдений здесь 7 (что больше 5) и в сумме больше 50 (не менее 20), что подтверждает применимость критерия согласия хи-квадрат.
Для нахождения значения хи-квадрата было использовано два способа: 1) с использованием функции chisq.test() и 2) по формуле:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

где O - наблюдаемое (observed);
E - ожидаемое (expected) (найдено через R и = 11,857).

$$\chi^2 = \frac{(7-E)^2 + (12-E)^2 + (9-E)^2 + (17-E)^2 + (10-E)^2 + (13-E)^2 + (15-E)^2}{E} = 6{,}145.$$

Затем находится значение степени свободы как n-1 = 7-1 = 6 (d.f.)
Принимая уровень значимости 0,05 и посмотрев по таблице распределения хи-квадрата при d.f. = 6, находим значение критической точки распределения хи-квадрата как 12,592.
Сравниваем полученное значение хи-квадрата (6,145) с табличным (12,592) - видно, что полученное значение меньше табличного, значит, здесь у нас нет оснований для отклонения нулевой гипотезы.
С помощью R было найдено соответствующее p-значение для полученного хи-квадрата: 0,407 > 0,05, следовательно, **мы не можем отвергнуть нулевую гипотезу о том, что вариативность количества деревьев на участках незначительна.**

**Задача 2** В каждом из двух прудов было поймано по 50 прудовиков. В 20 прудовиках из первого пруда и 32 прудовиках из второго были обнаружены личинки печёночных сосальщиков. На каком уровне значимости можно утверждать, что пруды различаются по заражённости прудов сосальщиком?

Решение.

Способ 1.

Сначала нужно определиться с уровнем уверенности. Пусть это будет стандартная 1/20 = 0,025.

Затем примем нулевую и альтернативную гипотезы:

H0: Пруды 1 и 2 не различаются по заражённости прудовиков.
HA: Пруд 2 отличается большей заражённостью прудовиков, чем пруд 1.

Если среднее число зараженных прудовиков печёночными сосальщиками равно μ, то реальное число распределено по Пуассону со средним μ.
Можно считать, что μ = (20 + 32)/2 = 26 (среднее между двумя наблюдениями).
При таком среднем распределение Пуассона практически не отличается от нормального распределения со средним 26 и дисперсией тоже 26 (поскольку у распределения Пуассона математическое ожидание всегда равно дисперсии и число 26 тут достаточно большое (>5-7)).
Математическое ожидание вычитается: M(X) = 26 - 26 = 0, а дисперсия складывается: D(X) = 26 + 26 = 52.
Разность двух независимо распределённых величин распределена нормально со средним 0 и дисперсией 52 (то есть с $\sigma$ = 7,211). Пересчитывая статистику в Z-score, получаем:

$$Z = \frac{\Delta x}{\sigma},$$

где $\Delta x$ - разность между двумя величинами (32 - 20 = 12);
$\sigma$ - среднеквадратическое отклонение.

$$Z = \frac{12}{7{,}211} = 1{,}664 < 1{,}96$$

1,96 является порогом на стандартно нормально распределённую случайную величину в данном случае (F(-1,96) ≈ 0,025). Также тут стоит отметить, что принята альтернатива двусторонняя, поскольку мы заранее не знаем, какой пруд заражён больше.
То есть, мы не попадаем в критическое множество, поэтому **нет основания утверждать, что пруд 2 отличается большей заражённостью прудовиков, чем пруд 1 - нулевая гипотеза не отклоняется.**

Способ 2.

Для решения по иному способу мной была рассмотрена возможность использовать как критерий хи-квадрат с поправкой Йейтса, так и точный критерий Фишера, т.к. оба критерия используются для решения таблиц 2 × 2 (что в данном случае применимо), но критерий Фишера вычисляется в случае, когда в клетках таблицы сопряжённости 2 x 2 не очень большие числа (меньше 5). Тем не менее, в качестве проверки результата точный критерий Фишера был применен в R.

Примем нулевую и альтернативную гипотезы:
H0: Зараженность прудовиков печёночными сосальщиками не зависит от пруда.
HA: Зараженность прудовиков печёночными сосальщиками зависит от пруда.
Определим различия в оценках между двумя группами с достоверностью $\alpha$ = 0,05.
Затем построим следующую таблицу сопряжённости:

| Состояние пруда | Заражён | Не заражён | Общее количество прудовиков |
|---|---|---|---|
| Пруд 1 | 20 | 30 | 50 |
| Пруд 2 | 32 | 18 | 50 |
| Общее количество | 52 | 48 | 100 |

Критерий хи-квадрат с поправкой Йейтса через R выдал следующий результат: p-value = 0,02768.
Точный критерий Фишера (двусторонний тест) показал примерно тот же результат: p-value = 0,02718.

Видно, что результаты практически одинаковы (с точностью до 3 знака).
Полученное p-значение (0,02768) меньше принятого уровня достоверности (0,05), **что свидетельствует о возможности отклонения нулевой гипотезы о том, что зараженность прудовиков печёночными сосальщиками не зависит от состояния пруда.**

**Задача 3** Геном одного из штаммов вируса SARS-CoV-2 содержит 29903 нуклеотида, которые распределены так:

| | |
|---|---|
| T | 9594 |
| A | 8954 |
| G | 5863 |
| C | 5492 |

(замечание: носителем генома является РНК, которая содержит урацил (U) вместо тимина (T), но по сложившейся традиции в базах данных используется буква T и для тимина, или урацила).

В этом геноме 2377 раз встречается слово TA. Определите, имеется ли достоверное ($\alpha = 0{,}001$) отличие частоты этого слова от ожидаемой при предположении независимого появления букв в геноме (равновероятность букв не предполагается, рассматриваем наблюдаемые частоты отдельных букв).

Решение.

Возьмем уровень достоверности $\alpha = 0{,}05$ (тогда критическое значение Z = 1,65).
Примем нулевую и альтернативную гипотезы:
H0: Отличие фактической частоты встречаемости слова TA от ожидаемой частоты недостоверно.
HA: Отличие фактической частоты встречаемости слова TA от ожидаемой достоверно.

Рассчитаем вероятность появления слова TA, учитывая появление букв A и T как двух независимых событий:

$$P(AB) = P(A) \cdot P(B)$$

$$P_2 = P(Thymines/Nucleotides) \cdot P(Adenines/Nucleotides) = \frac{9594}{29903} \cdot \frac{8954}{29903} = 0{,}09607.$$

Теоретическая (ожидаемая) частота появления TA в геноме:

$$f = n \cdot p$$

$$f_2 = Nucleotides \cdot P_2 = 29903 \cdot 0{,}09607 = 2872{,}778.$$

Собственно, по этой же формуле возможно вычислить фактическую вероятность появления TA в геноме:

$$P_1 = \frac{f_1}{Nucleotides} = \frac{2377}{29903} = 0{,}0795.$$

Обе частоты (теоретическая и фактическая) достаточно велики по значению (представим обе величины как доли числа успехов в двух независимых испытаниях N и M). Тогда и число успехов (n и m), и долю успехов можно считать распределёнными нормально.
Сперва найдем дисперсию для каждого из двух случаев по формуле:

$$D = N \cdot p \cdot (1 - p),$$

где N - число испытаний (в данном примере равно частоте f(f') появления TA в геноме).

$$D_1 = 2377 \cdot 0{,}0795 \cdot (1 - 0{,}0795) = 173{,}948.$$

$$D_2 = 2872{,}778 \cdot 0{,}09607 \cdot (1 - 0{,}09607) = 249{,}474.$$

Разность двух нормальных распределений распределена тоже нормально, при этом матожидания вычитаются, а дисперсии складываются. Поэтому разность долей распределена нормально со средним 0 и

дисперсией D1 + D2.
Далее посчитаем Z-статистику по формуле:

$$Z = \frac{n/N - m/M}{\sqrt{D1 + D2}} = \frac{p - p'}{\sqrt{D1 + D2}} = \frac{0{,}09607 - 0{,}0795}{\sqrt{173{,}948 + 249{,}474}} = 0{,}0008057 < 1{,}65.$$

Полученное Z не попало в порог критического множества, следовательно, **мы можем принять нулевую гипотезу и предположить, что отличие фактической частоты встречаемости слова ТА от ожидаемой частоты недостоверно.**

**Задача 4** Из многолетних наблюдений известно, что средняя температура воды некоторого горячего источника составляет 61,5 C°. В районе, где расположен этот источник, недавно произошло землетрясение, и геологи хотят выяснить, не повлияло ли оно на температуру источника. В файле statistics-tasks-2.4.txt находятся результаты измерений температуры источника, проведённые вскоре после землетрясения. На каком уровне значимости можно утверждать, что землетрясение повлияло на источник?

Решение.

Способ 1.