Прикладная статистика, НИУ ВШЭ / Домашняя работа 2

Студент - Катерина Олейникова

Преподаватель - Сергей Спирин

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 Δx - разность между двумя величинами (32 - 20 = 12);
σ - 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 х 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 = 0,0277.
Точный критерий Фишера (двусторонний тест) показал примерно тот же результат: p-value = 0,0272.
Видно, что результаты практически одинаковы (с точностью до 3 знака).
Полученное p-значение (0,0277) 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.

Уровень достоверности α по условию задачи равен 0,001 (тогда критическое значение Z = 3,09).
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:

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

$$P_2 = P(Thymines/Nucleotides) \cdot P(Adenines/Nucleotides) = \frac{9594}{29903} \cdot \frac{8954}{29903} = 0{,}0961.$$

Теоретическая (ожидаемая) частота появления TA в геноме:

$$f = n \cdot p$$

$$f_2 = Nucleotides \cdot P_2 = 29903 \cdot 0{,}0961 = 2872{,}778.$$

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).

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

$$D_2 = 2872{,}778 \cdot 0{,}0961 \cdot (1 - 0{,}0961) = 249{,}474.$$

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:

$$Z = \frac{n/N - m/M}{\sqrt{D1 + D2}} = \frac{p - p'}{\sqrt{D1 + D2}} = \frac{0{,}0961 - 0{,}0795}{\sqrt{173{,}948 + 249{,}474}} = 0{,}000806 < 3{,}09.$$

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.

Примем нулевую и альтернативную гипотезы:
H0: землетрясение не повлияло на источник.
HA: землетрясение повлияло на источник.
Иными словами, нулевая гипотеза опровергнет изменение средней температуры воды, а альтернативная гипотеза - подтвердит.

Задачей поставлено найти такой уровень значимости, при котором можно утверждать, что землетрясение повлияло на источник (или что то же самое - подтвердить изменение средней температуры воды), для этого необходимо будет принять альтернативную гипотезу HA.

Первый шаг - проверить выборку на нормальность распределения. Объем выборки (количество температурных значений) в текстовом файле невелик (n=19), потому применим критерий нормальности Шапиро-Уилка (данного для критерия объем выборки должен быть не меньше 3 и не больше 5000). Проверка на нормальность была реализована в R с помощью функции shapiro.test(); в результате получили значение p-value, равное 0,166 > 0,05, следовательно, *температура воды в источнике распределена равномерно.*
Затем для небольшой выборки с нормальным распределением можно применить критерий Стьюдента (t-test), реализация критерия была осуществлена в R (в результате было получено p-value = 0,0803).

Теперь подберём такой уровень значимости, при котором мы сможем отвергнуть нулевую гипотезу (то есть полученное p-значение должно быть меньше выбранного уровня значимости). Если рассматривать наиболее встречающиеся уровни значимости (10%, 5%, 1%, 0,1%), то получим следующие выражения:
0,0803 < 0,1 (что удовлетворяет условию),
0,0803 > 0,05 (не удовлетворяет условию),
0,0803 > 0,01 (не удовлетворяет условию),
0,0803 > 0,001 (не удовлетворяет условию).

То есть, **при уровне значимости 0,1 мы можем отвергнуть нулевую гипотезу в пользу альтернативной, которая свидетельствует о том, что землетрясение повлияло на источник.**

Задача 5 (Пример взят из книги: Бочаров П.П., Печинкин А.В. Теория вероятностей. Математическая статистика. 2-е изд. М.: ФИЗМАТЛИТ, 2005)

Для сравнительного анализа надежности крепёжных болтов, выпускаемых двумя заводами, были проверены на разрыв m = 24 изделия первого завода и n = 20 изделий второго. Силы натяжения ($\times 10^5$ Н), при которых произошли разрывы изделий первого и второго заводов, приведены в файле statistics-tasks-2.5.txt.
Необходимо сравнить эти две выборки по крайней мере одним (а лучше всеми) из известных методов и сделать выводы.

Решение.

Примем уровень значимости $\alpha = 0{,}05$.

Для начала можно проверить, распределены ли данные в двух выборках нормально. Применим для этого критерий Шапиро-Уилка с помощью shapiro.test() в R:
0,558 > 0,5 (для первой выборки),
0,977 > 0,5 (для второй выборки), значит, в обеих выборках данные распределены нормально.

Способ 1. Поскольку данные в обеих выборках имеют нормальное распределение, можно применять двувыборочный критерий Стьюдента (реализация осуществлена в R с помощью функции t.test()).

Для начала сформулируем нулевую и альтернативную гипотезы:
H0: обе выборки нормально распределены, и отсутствует существенная разница их средних.
HA: обе выборки нормально распределены, но имеют разные средние (при равных дисперсиях).

Среднее для первой выборки (подсчитано через функцию mean()) равно 2,619, среднее для второй выборки = 3,666.
Далее оцениваем дисперсию по формуле (подсчёт тоже реализован в R):

$$s^2 = \sum_i (X_i - \hat{X})^2/(n-1),$$

где X_i - значение для каждой из выборок;
$\hat{X}$ - среднее для каждой из выборок;
n - число наблюдений для каждой из выборок.
В итоге получаем s^2, равное 1,0825.
Рассчитаем t-статистику по формуле:

$$t = \frac{|\hat{X}_1 - \hat{X}_2|}{s/\sqrt{n}_1 + s/\sqrt{n}_2} = .. = 3{,}195.$$

Находим степень свободы как $df = n_1 + n_2 - 2 = 24 + 20 - 2 = 42$.

Сравним полученное значение t-критерия с табличным значением (для $\alpha = 0{,}05$ и $df = 42$): табличное значение t-критерия Стьюдента = 2,018 (можно найти по https://statpsy.ru/t-student/t-test-tablica/).

Рассчитанный по формуле t-критерий больше табличного (3,195 > 2,018).
В R с помощью пакета stats и функции t.test() было подсчитаны следующие значения t-критерия и p-значения: t = 3,2825, p = 0,00221.
Рассчитанное значения t-критерия больше табличного критического значения, таким образом, при уровне значимости $\alpha = 0{,}05$ наблюдаемые различия статистически значимы.
Полученное p-значение = 0,00221 меньше уровня значимости $\alpha = 0{,}05$, значит, **нулевая гипотеза отклоняется, принимается альтернативная гипотеза (то есть обе выборки распределены нормально с равными дисперсиями, но имеют разное среднее).**

Способ 2. F-тест (критерий Фишера). Применим в этом способе F-тест на равенство дисперсий.

Для начала сформулируем нулевую и альтернативную гипотезы:
H0: обе выборки распределены нормально с равными дисперсиями.
HA: обе выборки распределены нормально, но с разными дисперсиями.

Статистику посчитаем по формуле:

$$F = s_x^2/s_y^2 = \frac{\sum_i (X_i - \hat{X})^2/(k-1)}{\sum_j (Y_j - \hat{Y})^2/(l-1)},$$

где s_x^2 и s_y^2 - дисперсия для первой и второй выборки;
X_i - значение для каждого наблюдения из первой выборки;
$\hat{X}$ - среднее для первой выборки;
k - число наблюдений для первой выборки;
Y_j - значение для каждого наблюдения из второй выборки;
$\hat{Y}$ - среднее для второй выборки;
l - число наблюдений для второй выборки.

F = 0,7596 (подсчитано в R). Также F было вычислено для проверки с помощью функции var.test() (пакет stats): F = 0,7596. Значения получились идентичными.

Затем, сравнивая полученное значение F с табличным при уровне значимости $\alpha = 0{,}05$ и степенях свободы $df_1 = 20 - 1 = 19$ (для 1-ого завода) и $df_2 = 24 - 1 = 23$ (для 2-ого завода)(например, можно сравнить здесь: http://old.exponenta.ru/educat/referat/xikonkurs/student1/F-criteria.pdf), получаем, что табличное значение F = 2,123 больше полученного значения F = 0,7596.

Статистика получилась меньше критического значения, соответствующего выбранному уровню значимости ($\alpha = 0{,}05$, потому дисперсии двух рассмотренных выборок признаются одинаковыми.

Более того, при подсчете p-значения, полученного при использовании F-теста в R (var.test()), выяснено, что полученное p-value = 0,5252 > 0,05, поэтому альтернативная гипотеза отклонена, и принята нулевая гипотеза.

Можно сделать вывод, что обе выборки распределены нормально и имеют равные дисперсии.

Способ 3. Критерий Уилкоксона W (для проверки однородности выборок).

Применим такие нулевую и альтернативную гипотезы:
H0: обе выборки одинокаво распределены.
HA: обе выборки распределены неодинаково.

В качестве статистики критерия Уилкоксона используется сумма рангов из элементов выборки с меньшим количеством элементов, то есть второй выборки (n = 20). Применяя R, получаем W = 570.

Возьмем несколько уровней достоверности (0,1; 0,05; 0,01) и найдем для них нижние и верхние критическия значения статистики W (при m = 20, n = 24) из таблицы (по ссылке https://github.com/Harrix/Wilcoxon-W-Test/blob/main/-Wilcoxon-W-Test.pdf):

$\alpha = 0{,}1$ W = [379;521].
$\alpha = 0{,}05$ W = [366;534].
$\alpha = 0{,}01$ W = [341;559].

Для каждого уровня значимости рассчитанная W-статистика больше верхней критической границы (570 > 521; 570 > 534; 570 > 559), поэтому **при всех принятых уровнях значимости (α = 0,1; 0,05; 0,01) есть возможность отвергнуть нулевую гипотезу о том, что обе выборки распределены одинаково.**