Лабораторная работа №1
(теоретическая часть)

Садиев Абдурахмон, 674 группа

24.03.2020

Задача 2.4

Одеяла с электрообогревом применяются в хирургии для восстановления температуры тела пациента после операции. Имеются два вида одеял: стандартный (b0) и экспериментальный (b1).

Для 14 пациентов известно время, за которое нормальная температура тела восстанавливается при использовании одеяла каждого из видов.

Как понять, отличаются ли экспериментальные одеяла от стандартного?

Требуется:

- Записать задачу формально в виде проверяемой гипотезы и альтернативы.
- Предложить не менее 2-х критериев и соответствующих статистик для проверки этой гипотезы и описать:
  - при каких дополнительных условиях (если они есть) стоит применять тот или иной критерий
  - в чём преимущества/недостатки того или иного критерия
- Аналитически выразить достигаемый уровень значимости каждого критерия на выборке или опишите, как его получить с помощью табличных данных.

Решение

Можно проверить выполнена ли гипотеза о том, что в половине случаев элементы одной выборкибольше другой.

Запишем сразу задачу формально:

- Выборки: $\mathbf{X}_1^{(n)}$, $\mathbf{X}_2^{(n)}$, выборки связанные, $n = 14$(объемы выборок одинаковы)
- Нулевая гипотеза H_0: $\mathbb{P}\left(\mathbf{X}_1^{(n)} > \mathbf{X}_2^{(n)}\right) = \frac{1}{2}$
- Альтернатива H_1: $\mathbb{P}\left(\mathbf{X}_1^{(n)} > \mathbf{X}_2^{(n)}\right) \neq \frac{1}{2}$

Тогда применем двухвыборочный критерий знаков:

- Статистика: $T\left(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)}\right) = \sum_{i=1}^{n} [X_{1i} > X_{2i}]$

Тогда получаем, что статистика подчинина биномиальному распределению, поскольку она является суммой Бернулевских случайных величин (индекатор есть Бернулевская случайная величина), то есть мы знаем вид нулевого распределения:

- Нулевое распределение: $T\left(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)}\right) \sim Bi(n, p = \frac{1}{2})$

Конечно нужно упамянуть об ограничениях и недостатках: это непараметрический критерий, то есть он не использует никаких данных о характере распределения, и может применяться в широком спектре ситуаций, однако при этом он может иметь меньшую мощность, чем более специализированные критерии.

Но в данной ситуации мы ничего не знаем о характере распределеня данных

Найдем достигаемый уровень значимости: $p-\text{value} = \mathbb{P}(T = k) = \frac{C_n^k}{2^n}$, где $k = T\left(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)}\right)$, то есть мы уже подставили полученные данные в эту функцию. Здесь не нужно пользоваться таблицой.

Тогда правило принятия решения выглядет следующим образом:

$$\boxed{\text{нулевая гипотеза } \mathrm{H}_0 \text{ отклоняется} \Leftrightarrow p-\text{value} \notin \left[\tfrac{\alpha}{2}, \tfrac{1-\alpha}{2}\right]},$$

где α - уровень значимости.

Тогда возникает вопрос: какие можно ввести ограничение, чтобы получить критерий мощнее. Допустим, что наши выборки распределены нормально. Тогда воспользуемся t-критерий Стьюдента для связанных выборок. Поскольку для номального распределения медиана совпадает с математическим ожиданием.

Запишем задачу формально:

- Выборки: $\mathbf{X}_1^{(n)}$, $\mathbf{X}_2^{(n)}$, выборки связанные, $n = 14$(объемы выборок одинаковы) и $X_i \sim \mathcal{N}(\mu_i, \sigma_i)$ для $i = 1, 2$
- Нулевая гипотеза H_0: $\mu_1 = \mu_2$
- Альтернатива H_1: $\mu_1 \neq \mu_2$
- Статистика $T\left(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)}\right) = \frac{\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2}{S}\sqrt{n}$, где $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2}$, $D_i = X_{1i} - X_{2i}$

Тогда нулевое распределение:

- Нулевое распределение: $T\left(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)}\right) \sim St(n-1)$

Найдем достигаемый уровень значимости: $p-\text{value} = \min\left\{\mathbb{P}\left(T < t\right), \mathbb{P}\left(T > t\right)\right\}$,где $t = T\left(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)}\right)$, то есть мы уже подставили полученные данные в эту функцию. $p-\text{value} = \min\left\{St(13)_t, St(13)_{(1-t)}\right\}$, где $St(13)_t$ - t-квантиль распределения Стьдента со степенью свободы равной 13. Данное значение можно вычислить только по таблице лишь используя ассимптотическую нормальность. Будем считать тогда, что статитстика приближенно распределена как стандартное нормальное распределение. Таблицу на стандартное номальное всегда можно найти, но оценка достаточно груба для нзначения нашей статитстики, лучше пользовать правилом, которое написано ниже, оно точнее.

Тогда правило принятия решения выглядет следующим образом:

$$\boxed{\text{нулевая гипотеза } \mathrm{H}_0 \text{ отклоняется} \Leftrightarrow p-\text{value} \notin \left[\tfrac{\alpha}{2}, \tfrac{1-\alpha}{2}\right]},$$

Или иначе мы можем записать:

$$\boxed{\text{нулевая гипотеза } \mathrm{H}_0 \text{ отклоняется} \Leftrightarrow T\left(\mathbf{X}_1^{(n)}, \mathbf{X}_2^{(n)}\right) \notin \left[St(13)_{\frac{\alpha}{2}}, St(13)_{\frac{1-\alpha}{2}}\right]},$$

причем эти значения легко можно найти в таблице для $\alpha = 0.05, 0.025, 0.95, 0.975$

Задача 4.3

Правительство города М. испытывает систему обнаружения нежелаемых лиц по камерам в метро. В качестве демонстрации работоспособности системы была поставлена цель: найти и задержать опасного рецидивиста по имени Николай Вальный, а также его соучастников. Была собрана выборка из 5000 снимков лица, для которых была проверена гипотеза о несовпадении этого снимка с лицами участников команды Н. Вального. Для 100 фотографий нулевая гипотеза была отвергнута на уровне значимости $\alpha = 0.05$.

Требуется:

- Определить, в чем недостаток данного подхода и как можно его улучшить.
- Предложить наилучший, на ваш взгляд, способ для повышения качества данного решения.
- Какую меру качества контролирует данный способ? Какие гарантии он предоставляет?
- В чём недостатки данного способа?
- Как изменилась мощность при использовании предложенного вами способа относительно изначальной процедуры проверки?
- Известно, что все 5000 фотографий были сделаны для разных людей, и правительство хочет, чтобы система ни в коем случае не упустила преступников. Ответьте на те же вопросы из пунктов 2-4.
- Как изменилась мощность при использовании предложенного вами способа относительно предыдущего способа?

Решение

Очевидно, что недлстаток предложенного подхода заключается в отстутствие поправок на множественную проверку гипотез, а тут именно множественная проверка.

Согласно лекции 4, поскольку не известен характер зависимости между статистиками, то однозначно воспользуемся методом Холма. Поскольку он не учитывает характре зависимоти статистик, и, конечно, "нельзя построить контролирующую FWER процедуру мощнее, чем метод Холма". По определению, FWER - групповая вероятность ошибки первого рода (familywise error rate).

$$FWER = \mathbb{P}\left(\text{количество ошибок первого рода} > 0\right)$$

Наш выбор на метод Холма пал, поскольку он обеспечивает контроль над FWER на уровне α. Гарантии он предоставляет, что величина FWER никогда не будет больше чем α

Проблема вознкает в том, что количество ошибок второго рода может увеличиться, в следствие чего мы имеем мощность ниже, чем она была раньше. И не могу не отметить, что мы не учитывали характер взаимосвязи статистик.

Новое условие дает нам иформацию о связи статистик, а точнее говоря, они не зависимы. И стоит учитывать желание правительства, то есть они хотят, чтобы количество ошибок второго рода стемилась к нулю, "лучше лишний раз задержать, чем упустить". Тогда в силу желания правительства и новой полученной информации, мы будем спремиться к контролю велечины FDR. FDR - ожидаемая доля ложных отклонений гипотез (false discovery rate). Поэтому воспользуемся методом Бенджамини-Хохберга. Метод обеспечивает контроль над FDR на уровне α при условии, что статистики независимы. Тогда мы получим, что мощность метода возрасла,и соответственно количество ошибок второго рода меньше. Мощность возрасла поскольку условине на количество ошибок первого рода ослабли.

Кончно, метод не идеален, учистилось количество отвержения гипотезы.

Задача 2.2(решил по ошибке)

Рассмотрим фирму, которая занимается продажей лотерейных билетов. Правила лотереи следующее: все билеты являются выигрышными с вероятностью p. Билеты продаются до тех пор, пока хоть один человек не выиграет (гарантируется, что как только билет купили и он выигрышный, больше билетов не продают). Из-за вспышки коронавируса все заводы закрылись, а с ними и знание заветного p также пропало. Фирма хочет восстановить p, имея отчет о продажах билетов за последнии 5 месяцев: 8 билетов, 12 билетов, 7 билетов, 6 билетов и 12 билетов.

Требуется: Оцените методом максимального правдоподобия параметр p_0.

Теперь фирма нашла на складе несколько ящиков из билетами, которые вы можете использовать для проверки гипотезы о том, что истинное p равняется p_0 — оценке максимального правдоподобия из предыдущего пункта. Для проверки был предложен следующий эксперимент: последовательно вскрываются $n = 100$ билетов, и проводился подсчет: сколько выигрышных билетов было из данных N штук. Данный эксперимент проводился 10 раз и были получены следующие результаты: 13, 8, 11, 10, 11, 12, 7, 9, 10, 9.

- записать задачу формально;
- предложить статистику для решения данной задачи;
- получить приближенно нулевое распределение данной статистики;
- записать явно правило принятия решения на основе статистики и нулевого распределения для обеспечения уровня значимости $\alpha = 0.05$;
- проверить гипотезу по записанному критерию, для данных из условия.

Противоречат ли они гипотезе?

Решение

Согласно условию задачи мы можем переписать задачу следующим образом:

$$\mathbf{X} = (X_1 = 8, X_2 = 12, X_3 = 7, X_4 = 6, X_5 = 12)\,.$$

Причем X_i порождены геометрическим распределением: $X \backsim Geom(p_0)$, для любого i от 1 до 5. Напомним вид геомерического распределения: $\xi \backsim Geom(p_0)$ равносильно тому, что $\mathbb{P}(\xi = k) = (1-p_0)^{k-1}p_0$. Тогда запишем правдоподобие $\mathcal{L}(\mathbf{X}|p_0)$:

$$\mathcal{L}(\mathbf{X}|p_0) = \prod_{i=1}^{5}(1-p_0)^{X_i-1}p_0 = (1-p_0)^{\left(\sum_{i=1}^{5}X_i-5\right)}p_0^5$$

Запишем уравнение правдоподобия и получим оценку максимального правдоподобия (MLE) на велечину p_0:

$$\frac{\partial}{\partial p_0}\ln\mathcal{L}(\mathbf{X}|p_0) = 0$$

$$\frac{\partial}{\partial p_0}\ln\mathcal{L}(\mathbf{X}|p_0) = \frac{\partial}{\partial p_0}\ln\left((1-p_0)^{\left(\sum_{i=1}^{5}X_i-5\right)}p_0^5\right) = \frac{5}{p_0} - \frac{\left(\sum_{i=1}^{5}X_i - 5\right)}{1-p_0} = 0$$

$$p_0 = (\hat{p}_0)_{MLE} = \frac{5}{\left(\sum_{i=1}^{5}X_i\right)} = \frac{1}{9}$$

Перейдем к следующему пункту задания: у нас новая выборка, порожденная биномиальным распределением $Bi(100, p)$.

$$\widetilde{\mathbf{X}} = \left(\widetilde{X}_1 = 13, \widetilde{X}_2 = 8, \widetilde{X}_3 = 11, \widetilde{X}_4 = 10, \widetilde{X}_5 = 11, \widetilde{X}_6 = 12, \widetilde{X}_7 = 7, \widetilde{X}_8 = 9, \widetilde{X}_9 = 10, \widetilde{X}_{10} = 9\right)$$

Запишем правдоподобие и найдем p из уравнения правдоподобия:

$$\mathcal{L}\left(\widetilde{\mathbf{X}}|p\right) = \prod_{k=1}^{10} C_N^{\widetilde{X}_k} p^{\widetilde{X}_k}(1-p)^{N-\widetilde{X}_k}$$

$$\ln\left(\mathcal{L}(\mathbf{X}|p)\right) = \sum_{k=1}^{10}\ln C_N^{\widetilde{X}_k} + \ln(p)\sum_{k=1}^{10}\widetilde{X}_k + \ln(1-p)\left(10N - \sum_{k=1}^{10}\widetilde{X}_k\right)$$

$$\frac{\partial}{\partial p}\ln\mathcal{L}\left(\widetilde{\mathbf{X}}|p\right)=\frac{1}{p}\sum_{k=1}^{10}\widetilde{X}_k-\frac{1}{(1-p)}\left(10N-\sum_{k=1}^{10}\widetilde{X}_k\right)=0$$

$$\hat{p}_{\text{MLE}}=\frac{1}{10N}\sum_{k=1}^{10}\widetilde{X}_k$$

Это было маленькой подготовкой перед решением искомой задачи, а выглядет она следующим образом:

- Выборка: $\widetilde{\mathbf{X}}=(13,8,11,10,11,12,7,9,10,9)$
- Нулевая гипотеза H_0: $p_0=p$
- Альтернатива H_1: $p_0\neq p$

Стоит вопрос, какой критерий применить. Мы воспользуемся критерием отношения правдоподобия. Таким образом статистика имеет следующий вид:

- Статистика: $LR(\widetilde{\mathbf{X}})=-2\ln\left(\frac{\mathcal{L}(\widetilde{\mathbf{X}}|p_0)}{\mathcal{L}(\widetilde{\mathbf{X}}|\hat{p}_{\text{MLE}})}\right)$

Причем согласно теомере Уилкса (Wilks' theorem) мы знаем вид нулевого распределения:

- Нулевое распределение: $LR(\widetilde{\mathbf{X}})\sim\chi^2(1)$

Тогда правило принятия решения выглядет следующим образом:

$$\boxed{\text{нулевая гипотеза H}_0\text{ отклоняется}\Leftrightarrow LR(\widetilde{\mathbf{X}})>\chi^2(1)_{(1-\alpha)}},$$

где $\chi^2(1)_{(1-\alpha)}$ - $1-\alpha$ -квантиль распределения $\chi^2(1)$, причем $\chi^2(1)_{(1-\alpha)}=3.8$

Найдем значение статистики: $LR(\widetilde{\mathbf{X}})=1.3$ (численно посчитал). Таким образом мы получаем что $LR(\widetilde{\mathbf{X}})<\chi^2(1)_{(1-\alpha)}$. То есть статистические данные гипотизе не проитиворечат.