Прикладные методы математической статистики

Домашнее задание. Вариант 4. Задача 3.

Студент: Абу Аль Лабан Н. А.
Группа: БПИ 198

Обследование потенциальных потребителей.

Задание:

В былые дни возле станции метро «Площадь Революции» проводились различные обследования (опросы и дегустации), за участие в которых можно было получить коробку конфет,шоколадку, бутылку пива или даже торт.
Представьте себя на месте добрых людей, раздающих эти продукты.
Производитель газированной воды, планирующий продвижение нового товара (вода A), заказал вам малое обследование потенциальных потребителей.

Данные содержат следующую информацию:

- `a` — общая оценка респондентом воды A по семибалльной шкале (1 — совсем не понравилось, 7 — превосходно)
- `b` — оценка респондентом воды B (предполагаемого конкурента)
- `sex` — пол респондента (0 — мужской, 1 — женский)

Требуется ответить на два вопроса. 1) Есть ли основание считать, что потенциальный потребитель предпочитает воду A воде B?
2) Связано ли отношение к воде A с полом потребителя? От этого зависит стратегия продвижения товара.

Для ответа на первый вопрос решено использовать критерий знаков, для ответа на второй — критерий ранговых сумм Уилкоксона.
Выбран уровень значимости 10%

Опираясь на данные, дайте ответ на поставленные вопросы.
В каждом случае сформулируйте основную и альтернативную гипотезы, рассчитайте статистику, приведите критическое значение (или значения) и сделайте вывод.

---

1) Есть ли основание считать, что потенциальный потребитель предпочитает воду A воде B?

Для начала необходимо составить таблицу знаков:

- Запишем наши данные.
- Посчитаем разность оценок воды A и B для каждого респондента.
- Выпишем знаки в отдельный столбец.

Таким образом получаем таблицу:

sex	a	b	difference	sign
0	4	4	0	
1	7	3	4	+
1	5	3	2	+
1	5	4	1	+
0	4	4	0	
1	3	5	-2	-
1	7	3	4	+
0	5	4	1	+
0	3	6	-3	-
1	4	6	-2	-
0	3	7	-4	-
1	7	4	3	+
0	5	5	0	
1	7	5	2	+
1	3	4	-1	-
1	7	4	3	+
0	7	4	3	+
0	3	5	-2	-
0	4	5	-1	-
0	6	7	-1	-
0	5	5	0	
1	5	4	1	+
0	1	5	-4	-
0	1	5	-4	-

sex	a	b	difference	sign
1	2	4	-2	-
0	3	5	-2	-
0	5	5	0	
0	3	6	-3	-
1	5	3	2	+
1	5	4	1	+
1	6	4	2	+
1	6	4	2	+
0	4	4	0	
1	4	5	-1	-
1	6	4	2	+
1	5	5	0	
0	4	4	0	
1	5	4	1	+
0	4	5	-1	-
0	1	4	-3	-
0	7	4	3	+
0	3	4	-1	-
0	3	4	-1	-
1	7	4	3	+
0	3	5	-2	-
1	7	3	4	+
0	3	5	-2	-
0	2	4	-2	-

total "+":	19
total "-":	21
total "=":	8
total:	48

По данным таблицы видим, что:

- В **19** случаях предпочтение отдано **воде A** (знак "+")
- В **21** случае предпочтение отдано **воде B** (знак "-")
- В **8** случаях **нет предпочтения** в выборе воды (разность равна 0)

Теперь проверим гипотезу о том, что потенциальный потребитель **предпочитает воду A**, нежели воду B.
Уровень значимости, установленный условием задания: $\alpha = 0.1$

Пусть

- X_i и Y_i - оценки воды A и воды B i-тым потребителем соответственно
- d_i - разница оценок воды A и воды B i-тым потребителем

Сформулируем основную и альтернативную гипотезы:

- $H_0 : P(X_i > Y_i) = P(X_i < Y_i)$ – Потребитель не отдает предпочтение ни одной воде
- $H_A : P(X_i > Y_i) > P(X_i < Y_i)$ – Потребитель отдает предпочтение воде A

Поскольку выборки $X = \{X_1, . . , X_40\}$ и $Y = \{Y_1, . . , Y_40\}$ - связные, мы знаем, что оценки X_i и Y_i пренадлежат одному и тому же потребителю, и можем объединить их в пару

Таким образом,
$d_i = X_i - Y_i \sim i.i.d.$

Определим числа наблюдений:

- $n^+ \sim Bi(40, 0.5)$ - число наблюдений, где $d_i > 0$, то есть предпочтение отдано воде A
- $n^- \sim Bi(40, 0.5)$ - число наблюдений, где $d_i < 0$, то есть предпочтение отдано воде B
- Наблюдения, где $d_i = 0$, то есть потребитель не отлал предпочтения ни одному из производителей воды, проигнорируем

В качестве статистики выберем $T = n^-$

В пользу H_A будут говорить значения n^- , близкие к 0.
Выберем критическую область возле 0 так, чтобы вероятность попасть туда при H_A не превосходила уровень значимости $\alpha = 0.1$
Критическая область: $T \leqslant T_{n,\alpha}$

Выпишем наблюдаемые данные:

- $n^+ = 19$ - предпочтение отдано **воде A**
- $n^- = 21$ - предпочтение отдано **воде B**
- $n = n^+ + n^- = 19 + 21 = 40$ - всего наблюдений

Таблица критерия знаков не содержит таких больших n, поэтому приблизим значения к нормальному расределению:

- $E[n^+] = np = n \cdot \frac{1}{2}$
- $D[n^+] = npq = n \cdot \frac{1}{2} \cdot \frac{1}{2} = n \cdot \frac{1}{4}$

Центрируем и нормируем:
$Z = \frac{n^+ - np}{\sqrt{npq}} = \frac{n^+ - n \cdot \frac{1}{2}}{\sqrt{n \cdot \frac{1}{4}}} = \frac{2 \cdot n^+ - n}{\sqrt{n}}$

$npq = 40 \cdot \frac{1}{2} \cdot \frac{1}{2} = 10$

Поскольку $npq = 10$, достаточно большое, значит, **интегральная теорема Муавра-Лапласа** даст хорошее приближение
Применим ее и посчитаем статистику:
$Z \sim N(1, 0)$
$Z = \frac{2 \cdot 19 - 40}{\sqrt{40}} = -0.3162$

Найдем квантиль по таблице:
$z_\alpha = z_{0.1} = -1.28$

$-0.3162 > -1.28$
$Z > z_\alpha$
Подтверждаем основную гипотезу, так как не попали в критическую облость.
Следовательно, оснований считать, что потенциальный потребитель предпочтет воду A воде B на уровне значимости 10%, **нет**.

---

2) Связано ли отношение к воде A с полом потребителя? От этого зависит стратегия продвижения товара.

Для решения задачи обратимся к критерию ранговых сумм Уилкоксона.

В качестве выборок для решения поставленной задачи обозначим:

- $X = \{X_1, .., X_m\}$ - оценки воды A мужчин
- $Y = \{Y_1, .., Y_n\}$ - оценки воды A женщин

Начнем с составления единого ранжированного ряда из обеих выборок:

1. Отсортируем элементы выборок по возрастанию оценки
2. Присвоим ранги, приписав наименьшему значению менший ранг (столбец `rg`)
3. Поскольку элементы в выборке повторяются, для одинаковым оценкам присвоим их средний ранг, такие группы называют связками (столбец `ties`)

Таким образом получаем общую таблицу рангов:

sex	a	rg	ties
0	1	1	2
0	1	2	2
0	1	3	2
1	2	4	4,5
0	2	5	4,5
1	3	6	11
0	3	7	11
0	3	8	11
1	3	9	11
0	3	10	11
0	3	11	11
0	3	12	11
0	3	13	11
0	3	14	11
0	3	15	11
0	3	16	11
0	4	17	20,5
0	4	18	20,5
1	4	19	20,5
0	4	20	20,5
0	4	21	20,5
1	4	22	20,5
0	4	23	20,5
0	4	24	20,5

sex	a	rg	ties
1	5	25	30
1	5	26	30
0	5	27	30
0	5	28	30
0	5	29	30
1	5	30	30
0	5	31	30
1	5	32	30
1	5	33	30
1	5	34	30
1	5	35	30
0	6	36	37,5
1	6	37	37,5
1	6	38	37,5
1	6	39	37,5
1	7	40	44
1	7	41	44
1	7	42	44
1	7	43	44
1	7	44	44
0	7	45	44
0	7	46	44
1	7	47	44
1	7	48	44

total M:	26
total W:	22
total:	48

Запишем ранги выборок отдельно и посчитаем суммы рангов для каждой выборки:

sex	a	RG
0	1	2
0	1	2
0	1	2
0	2	4,5
0	3	11
0	3	11
0	3	11
0	3	11
0	3	11
0	3	11
0	3	11
0	3	11
0	3	11
0	4	20,5
0	4	20,5
0	4	20,5
0	4	20,5
0	4	20,5
0	4	20,5
0	5	30
0	5	30
0	5	30
0	5	30
0	6	37,5
0	7	44
0	7	44

sex	a	RG
1	2	4,5
1	3	11
1	3	11
1	4	20,5
1	4	20,5
1	5	30
1	5	30
1	5	30
1	5	30
1	5	30
1	5	30
1	5	30
1	5	30
1	6	37,5
1	6	37,5
1	6	37,5
1	7	44
1	7	44
1	7	44
1	7	44
1	7	44
1	7	44
1	7	44

Сумма рангов (м):	478
Сумма рангов (ж):	698
Сумма рангов:	1176

Общее количество рангов:

$m = 26$ – количество элементов в первой выборке

$n = 22$ – количество элементов во второй выборке

$N = m + n = 26 + 22 = 48$

Общее количество связок (связки запишем как пару (ранг, размер связки)):

$k = |\{t_1, .., t_k\}| = |\{(2, 3); (4.5, 2); (11, 11); (20.5, 8); (30, 11); (37.5, 4); (44, 9)\}| = 7$

Сумма рангов:

$S_m = 478$ – сумма рангов элементов в первой выборке

$S_w = 698$ – сумма рангов элементов во второй выборке

Сформулируем основную и альтернативную гипотезы:

- $H_0 : H : F = G$ – Выбокрки однородны, то есть пол **не влияет** на оценку воды А
- $H_A : H : F \neq G$ – Выбокрки неоднородны, то есть пол **влияет** на оценку воды А

В качестве статистики выберем статистику Уилкоксона $W_{\text{наб}} = S_w = 698$
Уровень значимости, установленный условием задания: $\alpha = 0.1$

Поскольку мы работаем с большими выборкаими (их размеры не приведены в таблицах), прибегнем к к аппроксимации распределения W предельным распределением статистики W при $n \to \infty$ и $m \to \infty$

Для этого перейдем от величины W к $W^* = \frac{W - MW}{\sqrt{DW}}$

Здесь:

- $MW = n\frac{m+n+1}{2} = 22\frac{26+22+1}{2} = 539$
- $DW = mn\frac{m+n+1}{12}$ в обычном случае, но тк у нас есть связки,

  $DW = \frac{mn}{12}\left[(m+n+1) - \frac{\sum\limits_{i=1}^{k} t_i(t_i^2-1)}{(m+n)(m+n-1)}\right]$

In [12]:
```
ties = [3, 2, 11, 8, 11, 4, 9] # размеры связок
n = 26
m = 22

DW = 0
for ti in ties:
    DW += (ti * (ti ** 2 - 1))/((m+n)*(m+n-1))
DW = m + n + 1 - DW
DW = int(DW)
DW = m * n * DW / 12
print(DW)
```

```
2240.3333333333335
```

Таким образом,
$DW = 2240.3333$

Подставляем и считаем: $W^* = \frac{698 - 539}{\sqrt{2240.3333}} = 3.3592$

Причем
$W^* \sim N(1, 0)$

Найдем критическое значение по таблице:
$z_{\frac{\alpha}{2}} = z_{0.05} = -1.64$

Так как
$|W^*| \geqslant z_{0.05}$
$3.3592 \geqslant -1.64$

Опровергаем основную гипотезу в пользу альтернативной, так как не попали в критическую облость. Следовательно, есть основания считать, что отношение к воде А зависит от пола на уровне значимости 10%.