Непараметричний підхід до перевірки гіпотез.

1 Теоретичні відомості

1.1 Тести на однорідність розподілів

1.2 Тест Колмогорова

Розглянемо кратну вибірку $X = (X_1, \ldots, X_n)$ з невідомою функцією розподілу спостержень $F(t) = \mathbf{P}(X_1 < t)$. Припустимо, що $F(t)$ є неперервною по t. Нас цікавить перевірка гіпотез про розподіл спостережень X_j.

Припустимо, з певних міркувань ми висунули гіпотетичну функцію розподілу $G(t)$ і ми хочемо перевірити чи дійсно теоретичний розподіл X_j узгоджується з теоретичним, тобто задача зводиться до перевірки статистичних гіпотез

$$\mathrm{H}_0 : F = G,\ \mathrm{H}_1 : F \neq G.$$

Якщо б мала місце основна гіпотеза, H_0, емпірична функція розподілу

$$\hat{F}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_j < t\},\ t \in \mathbb{R}$$

була б рівномірно строго консистентною оцінкою $G(t)$, тобто мала б місце така збіжність:

$$\Delta(X) = \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - G(t)| \to^{P_1} 0,\ n \to \infty. \tag{1}$$

З іншого боку, за виконання гіпотези про узгодженість розподілів, $\Delta(X)$ має швидкість збіжності до нуля порядку $1/\sqrt{n}$, бо

$$\sqrt{n}\Delta(X) \to^{W} \kappa,\ n \to \infty,$$

де κ має спеціальний розподіл Колмогорова, тобто

$$\mathbf{P}(\kappa < t) = \sum_{k\in\mathbb{Z}} (-1)^k \exp(-2k^2t^2),\ t > 0. \tag{2}$$

Результати (2) та (1) дають ідею для побудови статистичного тесту. Коли основна гіпотеза вірна, статистика $\hat{\kappa}_n = \sqrt{n}\Delta(X)$ в основному прийматиме малі значення. Підозрілими значеннями будуть великі значення $\hat{\kappa}_n$, тобто великі відхилення між вибріковим та гіпотетичним розподілами. Отже, з цих міркувань можна розглянути наступний тест для перевірки гіпотез про узгодженість розподілів:

$$\pi_K(X) = \begin{cases} 0, & \hat{\kappa}_n < k_\alpha, \\ 1, & \hat{\kappa}_n \geq k_\alpha, \end{cases} \tag{3}$$

де $k_\alpha = Q^\kappa(1-\alpha)$ – квантиль розподілу Колмогорова, α – рівень значущості тесту. Рішуче правило (3) називають тестом Колмогорова.

1.3 Тест Колмогорова-Смірнова

Розглянемо дві незалежні кратні вибірки $X = (X_1, \ldots, X_{n_X})$ та $Y = (Y_1, \ldots, Y_{n_Y})$ з невідомими функціями розподілу спостережень $F_X(t) = \mathbf{P}(X_1 < t)$ та $F_Y(t) = \mathbf{P}(Y_1 < t)$ відповідно. Припустимо, що F_X та F_Y є неперервними функціями.

Ми хочемо перевірити гіпотезу про те, чи є розподіли спостережень двох вибірок однорідним, тобто чи є розподіли F_X та F_Y однаковими. В термінах статистичних гіпотез, маємо

$$\mathrm{H}_0 : F_X = F_Y, \ \mathrm{H}_1 : F_X \neq F_Y.$$

Розглянемо емпіричні функції розподілу за кожною з вибірок

$$\hat{F}^X_{n_X}(t) = \frac{1}{n_X}\sum_{i=1}^{n_X} \mathbb{1}\{X_i < t\}, \ \hat{F}^Y_{n_Y}(t) = \frac{1}{n_Y}\sum_{j=1}^{n_Y} \mathbb{1}\{Y_j < t\}, \ t \in \mathbb{R}$$

Якщо основна гіпотеза вірна, то $\hat{F}^X_{n_X} - \hat{F}^Y_{n_X}$ рівномірно збігається до 0 при $n_X, n_Y \to \infty$. Дійсно, якщо $F_X = F_Y = F$, то з нерівності трикутника маємо

$$\begin{aligned}\Delta(X,Y) := \sup_{t\in\mathbb{R}} |\hat{F}^X_{n_X}(t) - \hat{F}^Y_{n_Y}(t)| = \sup_{t\in\mathbb{R}} |(\hat{F}^X_{n_X}(t) - F(t)) - (\hat{F}^Y_{n_Y}(t) - F(t))| \leq \\ \leq \sup_{t\in\mathbb{R}} |\hat{F}^X_{n_X}(t) - F(t)| + \sup_{t\in\mathbb{R}} |\hat{F}^Y_{n_Y}(t) - F(t)| \to^{P_1} 0, \ n_X, n_Y \to \infty.\end{aligned}$$

Розглянемо статистику $\hat{\kappa}_{n_X,n_Y} = \sqrt{\dfrac{n_X n_Y}{n_X + n_Y}} \cdot \Delta(X,Y)$. Якщо основна гіпотеза вірна, то

$$\hat{\kappa}_{n_X,n_Y} \to^W \kappa, \ n_X, n_Y \to \infty.$$

З досить схожих міркувань як у випадку тесту Колмогорова, будується тест Колмогорова-Смірнова для перевірки однорідності розподілів двох незалежних вибірок:

$$\pi_{KS}(X,Y) = \begin{cases} 0, & \hat{\kappa}_{n_X,n_Y} < k_\alpha, \\ 1, & \hat{\kappa}_{n_X,n_Y} \geq k_\alpha, \end{cases} \tag{4}$$

де $k_\alpha = Q^\kappa(1-\alpha)$ – квантиль розподілу Колмогорова, α – рівень значущості тесту. Рішуче правило (4) називають тестом Колмогорова-Смірнова.

1.4 Тест Вілкоксона

Знову розглянемо дві незалежні кратні вибірки $X = (X_1, \ldots, X_{n_X})$ та $Y = (Y_1, \ldots, Y_{n_Y})$ з невідомими функціями розподілу спостережень $F_X(t) = \mathbf{P}(X_1 < t)$ та $F_Y(t) = \mathbf{P}(Y_1 < t)$ відповідно. Припустимо, що розподіли F_X та F_Y відносяться до деякої сім'ї непрервних розподілів з зсувом:

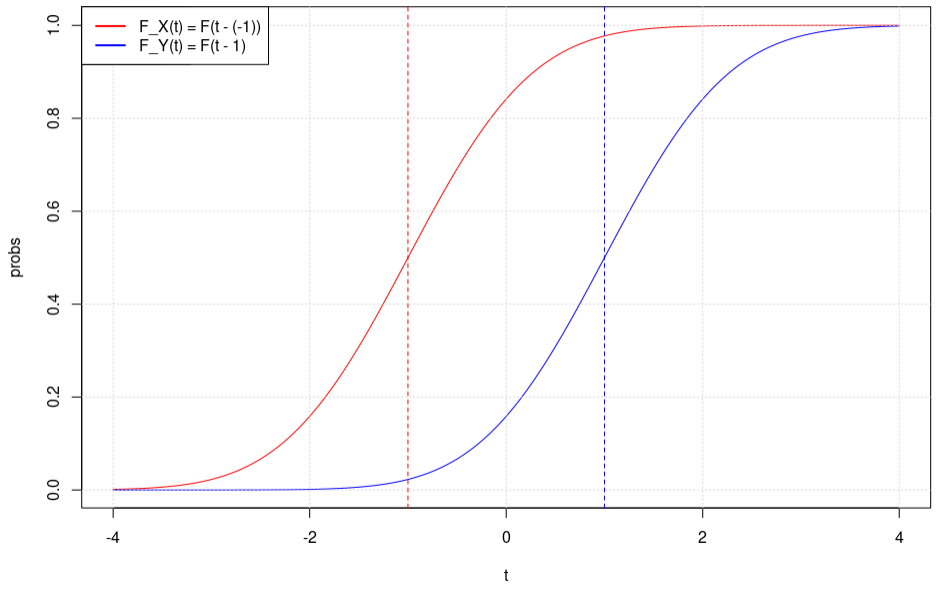
$$F_X, F_Y \in \{F(\cdot - \Delta) \mid F - \text{фіксована функція розподілу}, \Delta \in \mathbb{R} - \text{довільне}\}$$

Тобто $F_X(t) = F(t - \Delta_X)$ та $F_Y(t) = F(t - \Delta_Y)$ для деяких Δ_X та Δ_Y.

Ми хочемо перевірити гіпотезу про те, чи є розподіли спостережень двох вибірок однорідним. Якщо однорідність має місце, тоді $\Delta_X = \Delta_Y$.

В якості альтернативи до основної гіпотези вважатимемо, що Y_j переважно більші за X_j. Якщо це так, тоді $\Delta_Y > \Delta_X$.

Приклад. Нехай невідома функція розподілу F відповідає стандартного нормальному розподілу. Тоді $F(\cdot - \Delta)$ відповідає нормальному розподілу з математичним сподіванням Δ та дисперсією $\sigma^2 = 1$. Нехай розподіли X_i та Y_j беруться з сім'ї нормальних розподілів з довільним зсувом та фіксованою дисперсією, як описано вище. Конкретно, $F_X(t) = F(t-(-1))$, $F_Y(t) = F(t-1)$. Тобто значення Y_i будуть переважно більшими ніж значення X_j. Зобразимо функції розподілу F_X та F_Y нижче.



Отже, висунули наступні статистичні гіпотези

$$\mathrm{H}_0: \ F_X = F_Y, \ \mathrm{H}_1: \ F_X \neq F_Y, \ \Delta_Y > \Delta_X.$$

Для їх перевірки побудуємо тест Вілкоксона, який базується на рангових статистиках.

Нагадаємо, що ж це за звір такий, рангова статистика.

Означення. Нехай $Z = (Z_1, \ldots, Z_N)$ – деяка числова вибірка. Варіаційним рядом $\mathrm{sort}(Z)$ називають впорядковану вибірку Z за зростанням значень, тобто

$$\mathrm{sort}(Z) = (Z_{[1]}, \ldots, Z_{[N]}),\ Z_{[j]} \in Z,\ 1 \le j \le N,\ Z_{[i]} \le Z_{[j]},\ i < j.$$

Означення. Нехай $Z = (Z_1, \ldots, Z_N)$ – деяка числова вибірка та припустимо, що $Z_{[}i] < Z_{[}j]$ для всіх $i < j$. Рангом j-го спостереження ν_j у вибірці Z називають порядковий номер цього спостереження у варіаційному ряді $\mathrm{sort}(Z)$.

Якщо вибірка має повтори, тобто декілька значень мають спільне значення, то їх можна довільним чином розмістити у варіаційному ряді. Відповідні ранги для повторів називають спряженими (tied ranks). Для спряжених рангів зазвичай береться середнє арифметичне позицій.

Приклад. Розглянемо реалізацію деякої вибірки такого вигляду:

$$X = (0, -1, 2, 0, 2, 0, 1, 0)$$

Її варіаційний ряд:

$$\mathrm{sort}(X) = (-1, 0, 0, 0, 0, 1, 2, 2)$$

У вибірці лише два значення унікальні (мове іде про -1 та 1), а інші повторюються (0 та 2). Для спостережень з унікальними значеннями матимемо ранги $\nu_2 = 1$ та $\nu_7 = 6$ відповідно. Для спостережень з повторами заміщуємо ранги середнім арифметичним:

- Для $j \in \{1, 4, 6, 8\}$: спостереження займають позиції від 2 до 5, тому $\nu_j = (2+5)/2 = 3.5$,
- Для $j \in \{3, 5\}$: спостереження займають останні дві позиції, тому $\nu_j = (7+8)/2 = 7.5$.

Отже, вектор рангів має вигляд:

$$\nu_X = (\nu_1, \ldots, \nu_8) = (3.5, 1, 7.5, 3.5, 7.5, 3.5, 6, 3.5).$$

Нагадування. Так, середнє арифметичне $\{k, \ldots, k+m\}$ дорівнює $(k + (k+m))/2$:

$$k+(k+1)+\ldots+(k+m) = \sum_{j=1}^{k+m} j - \sum_{j=1}^{k-1} j = \frac{(m+1)(k+m+1)}{2} - \frac{(k-1)k}{2} = \frac{(m+1)(k+(k+m))}{2}$$

отже $(k + (k+1) + \ldots + (k+m))/(m+1) = (k + (k+m))/2$.

Тепер до побудови тесту. Тест Вілкоксона використовує статистику S_{n_X, n_Y}, яка будується для висунутих гіпотез таким чином:

1. Розглянемо розширену вибірку $Z = (Y, X) = (Y_1, \ldots, Y_{n_Y}, X_1, \ldots, X_{n_X})$,
2. Обчислюємо ранги спостережень у вибірці Z: ν_j^Z, $1 \le j \le (n_X + n_Y)$,
3. Підраховуємо суму перших n_Y рангів, тобто рангів Y в Z: $S_{n_X, n_Y} = \sum_{j=1}^{n_Y} \nu_j$.

Якщо альтернатива гіпотеза вірна, тоді у варіаційному ряді $\mathrm{sort}(Z)$ спостерження з Y будуть розміщені здебільшого в кінці, тому й матимуть дещо великі значення рангів. Значить, і сама сума S_{n_X,n_Y} може вийти великою. Тому це стане нашою 'підозрою' на користь альтернативи.

Для основної гіпотези, на противагу альтернативі, очікується 'помірні' значення тестової статистики (добре перемішані дані).

На основі попередніх міркувань, тест для висунутих гіпотез будується наступним чином:

$$\pi_W(X,Y) = \begin{cases} 0, & S_{n_X,n_Y} < w_\alpha, \\ 1, & S_{n_X,n_Y} \geq w_\alpha, \end{cases} \tag{5}$$

де w_α – деякий поріг тесту, що залежить від рівня значущості α. Рішуче правило (5) називають тестом Вілкоксона.

Поріг тесту Вілкоксона обчислюють спеціальними методами. Наприклад, якщо значення n_X та n_Y досить великі (скажімо $\min\{n_X, n_Y\} \geq 6$, $n_X + n_Y \geq 20$), можна використати асимптотичну нормальність статистики Вілкоксона:

$$\frac{S_{n_X,n_Y} - E_S}{\sqrt{V_S}} \to^W N(0,1),\ n \to \infty,$$
$$E_S = n_Y(n_X + n_Y + 1)/2, V_S = n_X n_Y (n_X + n_Y + 1)/12,$$

звідки визначається асимптотичний поріг тесту $w_\alpha = E_S + \sqrt{V_S} Q^{N(0,1)}(1-\alpha)$ для заданого рівня значущості α.

Якщо обсяги вибірок X та Y невеликі, скористаємося наступним підходом. Зауважимо, що статистика S_{n_X,n_Y} приймає цілі значення між точками $n_Y(n_Y+1)/2$ та $n_X n_Y + n_Y(n_Y+1)/2$ відповідно. Зокрема, за виконання основної гіпотези,

$$\mathbf{E}\left[S_{n_X,n_Y}\right] = \frac{n_Y(n_X + n_Y + 1)}{2} = E_S$$

це середина між заданими вище точками. Далі, розподіл S_{n_X,n_Y} симетричний відносно E_S. Для цього варто зауважити, що $S_{n_X,n_Y} - \mathbf{E}\left[S_{n_X,n_Y}\right]$ та $T_{n_X,n_Y} - \mathbf{E}\left[T_{n_X,n_Y}\right]$ мають спільний розподіл (внаслідок спільної функції розподілу спостережень), де $T_{n_X,n_Y} = \sum_{j=n_Y+1}^{n_X+n_Y} \nu_j^Z$ є сумою рангів X у Z. Крім того,

$$\begin{aligned} S_{n_X,n_Y} + T_{n_X,n_Y} &= (n_X + n_Y)(n_X + n_Y + 1)/2 = \mathbf{E}\left[S_{n_X,n_Y}\right] + \mathbf{E}\left[T_{n_X,n_Y}\right] \\ &\Rightarrow S_{n_X,n_Y} - \mathbf{E}\left[S_{n_X,n_Y}\right] = \mathbf{E}\left[T_{n_X,n_Y}\right] - T_{n_X,n_Y} = -(T_{n_X,n_Y} - \mathbf{E}\left[T_{n_X,n_Y}\right]) \end{aligned}$$

З вищенаведеного отримаємо, що

$$S_{n_X,n_Y} - \mathbf{E}\left[S_{n_X,n_Y}\right] =^d T_{n_X,n_Y} - \mathbf{E}\left[T_{n_X,n_Y}\right] =^d -(S_{n_X,n_Y} - \mathbf{E}\left[S_{n_X,n_Y}\right]),$$

що і доводить симетрію статистики Вілкоксона за виконання гіпотези H_0.

Так як, власне, обирати поріг тесту w_α? Нехай $\alpha \in (0,1)$ є деяким числом та оберемо s_α – таке найбільше число, що задовольняє нерівність $\mathbf{P}\left(S_{n_X,n_Y} \leq s_\alpha\right) \leq \alpha$ (розподіл статистики

дискретний, тому підібрати точний квантиль рівня α взагалі кажучи не вийде). Далі робимо такі махінації:

$$\begin{aligned}\mathbf{P}\left(S_{n_X,n_Y} \le s_\alpha\right) &= \mathbf{P}\left(S_{n_X,n_Y} - E_S \le s_\alpha - E_S\right) = \\ &= \mathbf{P}\left(S_{n_X,n_Y} - E_S \ge -(s_\alpha - E_S)\right) = \mathbf{P}\left(S_{n_X,n_Y} \ge E_S - (s_\alpha - E_S)\right) \le \alpha\end{aligned}$$

Таким чином нам вдалося обрати поріг тесту для висунутих вище гіпотез, тобто

$$w_\alpha = E_S - (s_\alpha - E_S) = n_Y(n_X + n_Y + 1) - s_\alpha.$$

Значення s_α можна спробувати знайти через рекурентні формули для розподілу статистики, наприклад див. [1].

Коментар.

Як адаптувати тест Вілкоксона, якщо розглядається інша одностороння альтернатива

$$\mathrm{H}_1 :\ F_X \ne F_Y,\ \Delta_Y < \Delta_X?$$

Тобто альтернатива полягає в тому, що Y_j здебільшого менші за X_i. Можемо взяти статистику Вілкоксона S_{n_X,n_Y} та вважати, що підозрілими на користь альтернативи будуть невеликі значення статистики. Тобто, тест матиме вигляд:

$$\pi_W(X,Y) = \begin{cases} 0, & S_{n_X,n_Y} > w_\alpha, \\ 1, & S_{n_X,n_Y} \le w_\alpha, \end{cases}$$

де w_α обирається або за допомогою нормальної апроксимації, або ж, для малих вибірок, $w_\alpha = s_\alpha$, де s_α визначено вище.

Ок, добре, а як тоді поводити себе у випадку двосторонньої альтернативи

$$\mathrm{H}_1 :\ F_X \ne F_Y?$$

Альтернатива в цьому разі інтерпретується дуже просто: розподіли спостережень різних вибірок різні. Можна скомбінувати попередні результати і розглянути рішуче правило вигляду:

$$\pi_W(X,Y) = \begin{cases} 0, & S_{n_X,n_Y} \in (w_\alpha^-, w_\alpha^+), \\ 1, & S_{n_X,n_Y} \notin (w_\alpha^-, w_\alpha^+), \end{cases}$$

де w_α^-, w_α^+ обираються або за допомогою нормальної апроксимації, або ж, для малих вибірок, $w_\alpha^- = s_\alpha$, $w_\alpha^+ = (n_X + n_Y + 1) - s_\alpha$. Вищенаведене правило прийняття гіпотез побудовано з міркування, що за основної гіпотези ранги повинні бути добре перемішаними та, відповідно, сума рангів має несуттєво відрізнятися від середнього E_S.

Також варто відмітити, якщо на малих вибірках критична область будується описаним вище чином, рівень значущості тесту *не перевищуватиме* α (дослідити).

Іще один коментар. Тест Вілкоксона може згодитися для перевірки гіпотези про рівність середніх, якщо має місце висунуте раніше припущення про форму розподілу даних. Це корисно, коли не можна впевнено казати про нормальність даних, відповідно й використовувати Z-тест (відома дисперсія) або T-тест (невідома дисперсія) для середнього.

Черговий коментар. Тест Вілкоксона пов'язаний з тестом Манна-Вітні (Mann-Whitney U test). У випадку другого, статистикою тесту виступає кількість разів, коли спостереження вибірки Y менші за спостереження з X:

$$U^{X>Y} = \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} \mathbf{1}\{X_i > Y_j\} = n_X n_Y + \frac{n_Y(n_Y+1)}{2} - S_{n_X,n_Y},$$

або ж кількість разів, коли спостереження з Y не менші за спостереження з X:

$$U^{X\le Y} = \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} \mathbf{1}\{X_i \le Y_j\} = n_X n_Y - U^{X>Y} = S_{n_X,n_Y} - \frac{n_Y(n_Y+1)}{2}.$$

1.5 Тести на незалежність

Мотивація. Тест Пірсона

Розглянемо дві кратні вибірки $X = (X_1, \dots, X_n)$ та $Y = (Y_1, \dots, Y_n)$. Нас цікавить, чи є спостереження з обох вибірок незалежними, тобто перевіряємо такі статистичні гіпотези

$$\mathrm{H}_0 :\ \mathrm{cor}(X_1, Y_1) = 0,\ \mathrm{H}_1 :\ \mathrm{cor}(X_1, Y_1) \neq 0,$$

де кореляція $\mathrm{cor}(X_1, Y_1)$ визначається як відношення коваріації величин та середньоквадратичних відхилень:

$$\mathrm{cor}(X_1, Y_1) = \frac{\mathrm{cov}(X_1, Y_1)}{\sqrt{\mathbf{D}\,[X_1]\,\mathbf{D}\,[Y_1]}}.$$

Для оцінки теоретичної кореляції $\mathrm{cor}(X_1, Y_1)$ розглядається коефіцієнт кореляції Пірсона:

$$r(X,Y) = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)(Y_j - \overline{Y}_n)}{\sqrt{\hat{S}_X^2 \hat{S}_Y^2}},$$

де $\overline{X}_n = \dfrac{1}{n}\sum_{j=1}^{n} X_j$, $\overline{Y}_n = \dfrac{1}{n}\sum_{j=1}^{n} Y_j$ є вибірковими середніми, а $\hat{S}_X^2 = \dfrac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$, $\hat{S}_Y^2 = \dfrac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{Y}_n)^2$ є вибірковими дисперсіями за вибірками X та Y відповідно.

Якби дані мали б гауссів розподіл, тест на кореляцію будувався б з міркувань, що статистика $(n-2)(r(X,Y))^2/(1-(r(X,Y))^2)$ має розподіл Фішера з 1 та $n-2$ ступенями вільності чисельника та знаменника відповідно (доведення методами регресійного аналізу, тест Фішера про загальну лінійну гіпотезу). Звідси тест кореляції Пірсона матиме вигляд

$$\pi_P(X) = \begin{cases} 0, & |r(X,Y)| < r_\alpha, \\ 1, & |r(X,Y)| \ge r_\alpha, \end{cases}$$

де $r_\alpha = \sqrt{(f_\alpha/(n-2))/(1+f_\alpha/(n-2))}$, $f_\alpha = Q^{F(1,n-2)}(1-\alpha)$.

Для даних не з гауссового розподілу, коефіцієнт кореляції Пірсона мало чим допоможе для перевірки гіпотези про незалежність, бо з корельованості випадкових величин необов'язково маємо незалежність. Натомість пропонується розглянути непараметричну альтернативу.

Тест Спірмена

Розглянемо вектори рангів $\nu_X = (\nu_1^X, \ldots, \nu_n^X)$, $\nu_Y = (\nu_1^Y, \ldots, \nu_n^Y)$ вибірок X, Y відповідно та введемо коефіцієнт кореляції Спірмена:

$$\tau(X,Y) = \frac{\frac{1}{n}\sum_{j=1}^n (\nu_j^X - \overline{\nu}_X)(\nu_j^Y - \overline{\nu}_Y)}{\sqrt{\hat{S}^2_{\nu_X}\hat{S}^2_{\nu_Y}}},$$

де $\overline{\nu}_X = \dfrac{1}{n}\sum_{j=1}^n \nu_j^X$, $\overline{\nu}_Y = \dfrac{1}{n}\sum_{j=1}^n \nu_j^Y$ є вибірковими середніми, а $\hat{S}^2_{\nu_X} = \dfrac{1}{n}\sum_{j=1}^n (\nu_j^X - \overline{\nu}_X)^2$, $\hat{S}^2_{\nu_Y} = \dfrac{1}{n}\sum_{j=1}^n (\nu_j^Y - \overline{\nu}_Y)^2$ є вибірковими дисперсіями за рангами вибірок X та Y відповідно.

Якщо має місце основна гіпотеза про незалежність, то

$$\sqrt{n-1}\cdot\tau(X,Y) \to^W N(0,1),\ n\to\infty.$$

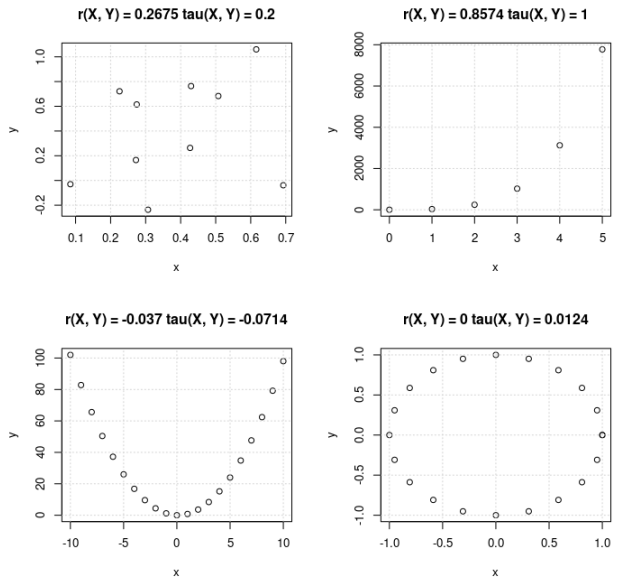
Звідси тест кореляції Спірмена матиме вигляд

$$\pi_S(X) = \begin{cases} 0, & |\tau(X,Y)| < z_\alpha/\sqrt{n-1}, \\ 1, & |\tau(X,Y)| \geq z_\alpha/\sqrt{n-1}, \end{cases}$$

де $z_\alpha = Q^{N(0,1)}(1-\alpha/2)$ для заданого рівня значущості α.

Коментар. Більш поширеним є підхід до перевірки гіпотез про некорельованість ознак. Для такої гіпотези, поріг тесту Спірмена будується дещо інакше (наприклад, для великих вибірок береться лобова підстановка у поріг тесту Пірсона, або ж інші спеціальні техніки).

Ще один коментар. Коефіцієнт кореляції Пірсона краще відстежує лінійні форми залежності, у той час як коефіцієнт кореляції Спірмена переважатиме на монотонних нелінійних залежностях. Нижче показано деякі значення коефіцієнтів кореляції для різних ситуацій.



2 Задачі

2.1 Задача 1 (Тест Колмогорова)

Наведені нижче дані являють собою час безвідмовної роботи електронного приладу. З'ясуйте, чи можна вважати, що час безвідмовної роботи цього приладу має експоненційний розподілом з інтенсивністю $\lambda = 0.01$? Час безвідмовної роботи:

$18, 21, 46, 48, 60, 63, 84, 100, 116, 117, 118, 133, 135, 163, 171, 174, 175, 213, 219, 230, 327, 328, 356, 384, 410$

Вважати, що рівень значущості $\alpha = 0.05$.

Розв'язання: За умовою задачі, маємо дані $X = (X_1, \ldots, X_n)$, розподіл $F(t) = \mathbf{P}(X_1 < t)$ є невідомим, а гіпотетичним розподілом є $G(t) = \mathbb{1}\{t > 0\} \cdot (1 - \exp(-\lambda t))$ із заданим $\lambda = 0.01$. Сформулюємо гіпотези про однорідність розподілів:

$$\mathrm{H}_0 : F = G, \ \mathrm{H}_1 : F \neq G.$$

Обчислимо статистику Колмогорова. Для обчислення найбільшого відхилення між емпіричною та гіпотетичною функкціями розподілу, достатньо знати поведінку в точках стрибку емпіричної функції розподілу

$$\Delta(X) = \max_{1 \leq k \leq n} \max\{|k/n - G(X_{(k)})|, |G(X_{(k)}) - (k-1)/n|\}.$$

Результат вище отримано на практичному занятті №2. Підрахунки відхилень наведено у таблиці, див. наступну сторінку.

Найбільше відхилення: $\Delta(X) = 0.3665$. Отже, статистика тесту $\hat{\kappa}_n \approx 1.8326$.

Поріг тесту Колмогорова при заданому рівні значущості $\alpha = 0.05$ є квантіль розподілу Колмогорова рівня 0.95: $k_\alpha = Q^\kappa(1-\alpha)$. Візьмемо $k_\alpha \approx 1.365$. Значення статистики 1.8326 більше за вказаний поріг 1.365, отже тест Колмогорова має підстави прийняти альтернативу про відмінність теоретичного розподілу від гіпотетичного.

k	$X_{(k)}$	$G(X_{(k)})$	$\delta_+ := \lvert k/n - G(X_{(k)})\rvert$	$\delta_- := \lvert G(X_{(k)}) - (k-1)/n\rvert$	$\max\{\delta_+, \delta_-\}$
1	18	0.1647	0.1647	0.1247	0.1647
2	21	0.1894	0.1494	0.1094	0.1494
3	46	0.3687	0.2887	0.2487	0.2887
4	48	0.3812	0.2612	0.2212	0.2612
5	60	0.4512	0.2912	0.2512	0.2912
6	63	0.4674	0.2674	0.2274	0.2674
7	84	0.5683	0.3283	0.2883	0.3283
8	100	0.6321	0.3521	0.3121	0.3521
9	116	0.6865	0.3665	0.3265	0.3665
10	117	0.6896	0.3296	0.2896	0.3296
11	118	0.6927	0.2927	0.2527	0.2927
12	133	0.7355	0.2955	0.2555	0.2955
13	135	0.7408	0.2608	0.2208	0.2608
14	163	0.8041	0.2841	0.2441	0.2841
15	171	0.8191	0.2591	0.2191	0.2591
16	174	0.8245	0.2245	0.1845	0.2245
17	175	0.8262	0.1862	0.1462	0.1862
18	213	0.8812	0.2012	0.1612	0.2012
19	219	0.8881	0.1681	0.1281	0.1681
20	230	0.8997	0.1397	0.0997	0.1397
21	327	0.962	0.162	0.122	0.162
22	328	0.9624	0.1224	0.0824	0.1224
23	356	0.9716	0.0916	0.0516	0.0916
24	384	0.9785	0.0585	0.0185	0.0585
25	410	0.9834	0.0234	0.0166	0.0234

Табл. 1: Обчислення для $\hat{\kappa}_n$.

2.2 Задача 2 (Тест Смірнова)

Цієї весни деякий дослідник Василь відвідав два озера. У кожному з озер він вимірював довжини деякого виду риби, випадково відловивши по 15 з кожного озера. Вимірювання такі (у см):

Довжини риб в озері №1	26	17	22	34	30	19	32	20	13	30	25	24	28	13	19
Довжини риб в озері №2	13	29	24	8	15	12	14	12	15	19	17	13	15	18	26

Василь не має жодного припущення про розподіл даних. Чи можна вважати, що розподіл довжин риби є однорідним для двох озер на рівні значущості $\alpha = 0.05$?

Розв'язання. Маємо дві незалежні вибірки $X = (X_1, \dots, X_{n_X})$ та $Y = (Y_1, \dots, Y_{n_Y})$ обсягів $n_X = n_Y = 15$. Розподіли $F_X(t)$ та $F_Y(t)$ є невідомими. Перевіримо гіпотези про однорідність розподілів двох вибірок

$$\mathrm{H}_0 : F_X = F_Y,\ \mathrm{H}_1 : F_X \neq F_Y.$$

Для обчислення статистики Колмогорова-Смірнова треба знайти найбільше відхилення між емпіричними функціями розподілу за кожною з вибірок: це достатньо перевірити в точках стрибку кожної з таких функцій.

Отже,

$$\Delta(X,Y) = \max_{Z\in\{X,Y\}} \{ \max_{1\le j\le n_Z} \max\{|\hat{F}_X(Z_j) - \hat{F}_Y(Z_j)|, |\hat{F}_X(Z_j+) - \hat{F}_Y(Z_j+)|\}\}.$$

Підрахунки відхилень наведено у таблиці, див. наступну сторінку.

Найбільше відхилення: $\Delta(X,Y) = 0.5333$. Отже, статистика тесту $\hat{\kappa}_{n_X,n_Y} \approx 1.4605$.

Поріг тесту Колмогорова при заданому рівні значущості $\alpha = 0.05$ є квантіль розподілу Колмогорова рівня 0.95: $k_\alpha = Q^\kappa(1-\alpha)$. Візьмемо $k_\alpha \approx 1.365$. Значення статистики 1.4605 більше за вказаний поріг 1.365, отже тест Колмогорова-Смірнова має підстави прийняти альтернативу про відмінність теоретичного розподілу від гіпотетичного.

Тут $Z = (X, Y)$.

Z_k	$\delta_+ := \lvert\hat{F}_X(Z_j+) - \hat{F}_Y(Z_j+)\rvert$	$\delta_- := \lvert\hat{F}_X(Z_j) - \hat{F}_Y(Z_j)\rvert$	$\max\{\delta_+, \delta_-\}$
26	0.2667	0.2667	0.2667
17	0.4667	0.4667	0.4667
22	0.4	0.3333	0.4
34	0.0667	0	0.0667
30	0.2667	0.1333	0.2667
19	0.5333	0.4667	0.5333
32	0.1333	0.0667	0.1333
20	0.4667	0.4	0.4667
13	0.2	0.2	0.2
30	0.2667	0.1333	0.2667
25	0.3333	0.2667	0.3333
24	0.3333	0.3333	0.3333
28	0.2667	0.2	0.2667
13	0.2	0.2	0.2
19	0.5333	0.4667	0.5333
13	0.2	0.2	0.2
29	0.2	0.2667	0.2667
24	0.3333	0.3333	0.3333
8	0	0.0667	0.0667
15	0.2667	0.4667	0.4667
12	0.0667	0.2	0.2
14	0.2	0.2667	0.2667
12	0.0667	0.2	0.2
15	0.2667	0.4667	0.4667
19	0.5333	0.4667	0.5333
17	0.4667	0.4667	0.4667
13	0.2	0.2	0.2
15	0.2667	0.4667	0.4667
18	0.4667	0.5333	0.5333
26	0.2667	0.2667	0.2667

Табл. 2: Обчислення для $\hat{\kappa}_{n_X,n_Y}$. Перші 15 рядків стосуються X, наступні 15 для Y.

2.3 Задача 3 (Тест Вілкоксона)

Для дослідження стійкості на стирання епоксидної пластмаси з використанням як наповнювача оксиду алюмінію у різній концентрації виготовили дві партії плиток.

Розглядалися два значення відносної концентрації наповнювача в смолі, що відповідали відношенням 1/2 до 1 і 1 до 1. Виготовили по 18 зразків плиток, при цьому дві з них були забраковані й не випробувалися. Потім кожну з плиток приводили у зворотно-поступальний рух по абразивному матеріалу (10000 циклів для кожної плитки).

Вимірювалася різниця товщини плиток до і після випробувань у дюймах:

- Концентрація 1/2 до 1: 7.0, 6.4, 7.3, 5.1, 5.7, 6.6, 5.0, 7.7, 6.8, 5.1, 4.6, 5.5, 5.8, 6.2, 4.8, 5.8
- Концентрація 1 до 1: 7.1, 5.0, 6.4, 6.9, 5.7, 6.5, 6.5, 4.0, 6.2, 6.8, 4.0, 7.5, 7.2, 5.2, 7.8, 4.8, 4.4, 6.0

Чи відрізняються плитки з епоксидної пластмаси, в яких наповнювачем був оксид алюмінію у різних концентрацях, за стійкість на стирання? Іншими словами, чи впливає концентрація наповнювача на стійкість плитки на стирання? Вважати рівень значущості $\alpha = 0.05$.

Розв'язання. Маємо дві незалежні вибірки $X = (X_1, \ldots, X_{n_X})$, $Y = (Y_1, \ldots, Y_{n_Y})$, де $n_X = 16$, $n_Y = 18$. Припустимо, що розподіли різниць X_i та Y_j можуть відрізнятися лише математичним сподіванням. Висунемо такі гіпотези для перевірки:

$$\mathrm{H}_0 : \mathbf{E}[X_1] = \mathbf{E}[Y_1], \ \mathrm{H}_1 : \mathbf{E}[X_1] \neq \mathbf{E}[Y_1]$$

Для перевірки висунутих гіпотез скористаємося двостороннім тестом Вілкоксона. Спочатку зберемо дані в єдину вибірку $Z = (Y, X)$ та підрахуємо ранги спостережень Y у наборі Z. Позиції спостережень з Y у варіаційному ряді вибірки Z наведено далі.

Вектор рангів Y у Z (зауважте, у Z наявні спряжені ранги):

$$\nu^{Y \in Z} = (29, 7.5, 20.5, 27, 13.5, 22.5, 22.5, 1.5, 18.5, 25, 1.5, 32, 30, 11, 34, 5.5, 3, 17).$$

Підраховуємо статистику Вілкоксона: $S_{n_X,n_Y} = \sum_{j=1}^{n_Y} \nu_j^{Y \in Z} = 322$.

Для визначення критичної області тесту скористаємося нормальним наближенням для статистики Вілкосона (кількість спостережень у кожній з вибірок допускає такий трюк). З того, що

$$\mathbf{P}\left(\left|\frac{S_{n_X,n_Y} - E_S}{\sqrt{V_S}}\right| < z_\alpha\right) \approx \mathbf{P}\left(|N(0,1)| < z_\alpha\right) = 1 - \alpha$$

маємо $z_\alpha = Q^{N(0,1)}(1 - \alpha/2) \approx 1.96$ та маємо довірчу область для H_0:

$$I_\alpha = (\tilde{w}_\alpha^-, \tilde{w}_\alpha^+) = (E_S - \sqrt{V_S} z_\alpha, E_S + \sqrt{V_S} z_\alpha) \approx (258.1948, 371.8052)$$

Тут $E_S = 315$ та $V_S = 840$. Якщо S_{n_X,n_Y} потрапляє в I_α, приймаємо основну гіпотезу H_0, інакше альтернативу H_1. В нашому випадку $S_{n_X,n_Y} = 322 \in I_\alpha$, що є підставою прийняти основну гіпотезу.

Звідки?	$Z_{(j)}$	j	Номер з Y
y	4	1	8
y	4	2	11
y	4.4	3	17
x	4.6	4	0
y	4.8	5	16
x	4.8	6	0
y	5	7	2
x	5	8	0
x	5.1	9	0
x	5.1	10	0
y	5.2	11	14
x	5.5	12	0
y	5.7	13	5
x	5.7	14	0
x	5.8	15	0
x	5.8	16	0
y	6	17	18
y	6.2	18	9
x	6.2	19	0
y	6.4	20	3
x	6.4	21	0
y	6.5	22	6
y	6.5	23	7
x	6.6	24	0
y	6.8	25	10
x	6.8	26	0
y	6.9	27	4
x	7	28	0
y	7.1	29	1
y	7.2	30	13
x	7.3	31	0
y	7.5	32	12
x	7.7	33	0
y	7.8	34	15

Табл. 3: Варіаційний ряд розширеної вибірки $Z = (Y, X)$ та як розміщені розмазана Y.

2.4 Задача 4 (Тест Спірмена)

Дані про смертність населення використовуються для оцінки параметрів захворюваності. Такі дані наведено для 10 регіонів у таблиці

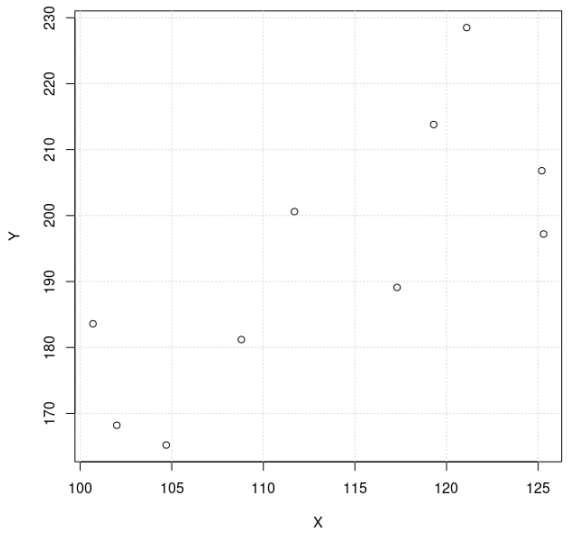
Регіон	Смертність на 10000	Захворюваність на 1000
1	125.2	206.8
2	119.3	213.8
3	125.3	197.2
4	111.7	200.6
5	117.3	189.1
6	100.7	183.6
7	108.8	181.2
8	102.0	168.2
9	104.7	165.2
10	121.1	228.5

Перевірити гіпотезу про (не)залежність смертності від захворюваності за критерієм Спірмена, вважаючи що рівень значущості $\alpha = 0.05$.

Розв'язання. Представимо значення ознак смертності та захворюваності через X та Y відповідно. Обчислюємо вектор рангів для кожної вибірки:

$$\nu_X = (9, 7, 10, 5, 6, 1, 4, 2, 3, 8),$$
$$\nu_Y = (8, 9, 6, 7, 5, 4, 3, 2, 1, 10).$$

За отриманими векторами рангів обчислюємо коефіцієнт кореляції Спірмена. Маємо таке значення: $\tau(X, Y) \approx 0.7333$. Критичний рівень беремо з нормальної апроксимації, $r_\alpha = Q^{N(0,1)}(1 - \alpha/2)/\sqrt{n-1} \approx 0.6533$. Видно, що значення коефіцієнта кореляції Пірсона за модулем перевищує заданий поріг. Отже, маємо підстави відхилити гіпотезу про некорельованість ознак.



Література

[1] H.B. Mann, D.R. Whitney. *On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other.* The Annals of Mathematical Statistics, 18(1) 50-60 March, 1947. Посилання