II. ***Випадок обробки векторних спостережень.*** Припустимо тепер, що спостереження є векторними:

$$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n, \qquad \vec{x}_i \in \mathbb{R}^q, \ i = \overline{1,n}.$$

У векторному випадку можна скористатися такими засобами:

1) критерієм на базі F - статистики,

2) діаграмою розсіювання (у випадку $q = 2$).

1) Критерій на базі F - статистики. Задачу будемо розв'язувати шляхом перевірки гіпотези:

H_0 : *найбільш підозрілий на аномальність вимір не є викидом,* $\alpha > 0$.

1. Спочатку обчислюємо наступні величини:

$$\bar{\bar{x}}_i(n) = \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} \vec{x}_j, \quad \hat{V}_i(n) = \frac{1}{n-2}\sum_{\substack{j=1 \\ j\neq i}}^{n} \left(\vec{x}_j - \bar{\bar{x}}_i(n)\right)\left(\vec{x}_j - \bar{\bar{x}}_i(n)\right)^T, \quad i = \overline{1,n}.$$

Махаланобіса:

$$d_i^2(n) = \left(\vec{x}_i - \bar{\bar{x}}_i(n)\right)^T \hat{V}_i^{-1}(n)\left(\vec{x}_i - \bar{\bar{x}}_i(n)\right), \quad i = \overline{1,n}.$$

3. Знаходимо індекс найбільш підозрілого на аномальність виміру

$$i_0 = \arg\max_i d_i^2(n).$$

4. Визначаємо значення такої статистики:

$$F_{i_0}(n) = \frac{(n-1)(n-1-q)}{n(n-2)q} d_{i_0}^2(n).$$

З'ясувалося, що розподіл статистики $F_{i_0}(n)$ можна наблизити F –розподілом з параметрами q та $(n-1-q)$.

5. Тоді, так як *область відхилення гіпотези* H_0 - це область великих значень статистики $F_{i_0}(n)$, то гіпотезу H_0 будемо приймати, якщо справедлива наступна нерівність:

$$F_{i_0}(n) < F_\alpha(q, n-1-q),$$

де $F_\alpha(q, n-1-q)$ – 100α відсоткова точка F – розподілу з q та $n-1-q$ степенями свободи.

Якщо вектор спостережень $\vec{x}_{i_0}$ виявився аномальним, то його вилучають з вибірки, і всю процедуру повторюють починаючи з пункту 1 до того часу, доки на деякому кроці найбільш підозрілий на аномальність вимір виявиться не викидом. Після цього процедуру завершують.

2) Діаграма розсіювання. Вона описана у розділі, який присвячений розвідувальному аналізу.

Перевірка стохастичності вибірки
(Tests for Randomness)

Нехай досліджується вибірка

$$x_1, x_2, \ldots, x_n, \quad n \in \mathbb{N}.$$

Перед обробкою вибірки має сенс впевнитися, що вона є випадковою (стохастичною), а не знаходиться під впливом деякого систематичного зміщення, наприклад: монотонного або періодичного.

Розв'язувати проблему будемо шляхом перевірки гіпотези

$$H_0: \ \textit{вибірка є стохастичною}, \ \alpha > 0.$$

Для цього пропонується використовувати:

1) критерій серій на базі медіани вибірки,
2) критерій зростаючих та спадаючих серій,
3) критерій квадратів послідовних різниць (критерій Аббе),

Зупинимося детальніше на кожному з них.

для перевірки гіпотези H_0, причому альтернативна гіпотеза виглядає таким чином

H_1: *у вибірці наявне систематичне монотонне зміщення середнього.*

1. Спочатку по вибірці визначається оцінка медіани $\hat{x}_{med}$.

2. Потім під i – им членом вибірки $x_1, x_2, \ldots, x_n$, який більше $\hat{x}_{med}$ ставиться плюс, а який менше $\hat{x}_{med}$ ставиться мінус. Виміри які дорівнюють $\hat{x}_{med}$ до уваги не приймаються. Тобто символи розставляються згідно

$$\begin{cases} +, \text{якщо } x_i > \hat{x}_{med}, \\ \sqcup, \text{якщо } x_i = \hat{x}_{med}, \\ -, \text{якщо } x_i < \hat{x}_{med}. \end{cases}$$

В результаті отримаємо деяку послідовності плюсів та мінусів

$$+--+-\ldots-+$$

Означення. *Серія* - це підпослідовність підряд розташованих однакових символів у послідовності.

3. Далі обчислюємо такі дві статистики:

$\nu(n)$ - загальну кількість серій у ній,

$\tau(n)$ - кількість членів у найдовшій серії.

4. Зрозуміло, що вибірка буде мати стохастичну природу, якщо довжина найдовшої серії $\tau(n)$ не занадто довга, а загальна кількість серій $\nu(n)$ не занадто мала. Причому відомо, що розподіл статистики $\nu(n)$ можна наблизити деяким *нормальним розподілом*, а $\tau(n)$ - деяким *пуасонівським розподілом*. Тоді *область прийняття гіпотези*:

$$\begin{cases} \nu(n) > \nu_\beta(n), \\ \tau(n) < \tau_{1-\beta}(n), \end{cases}$$

де $\nu_\beta(n), \tau_\beta(n)$ – квантилі рівня β статистик $\nu(n)$ та $\tau(n)$ відповідно.

А β вибирається таким чином, щоб помилка I роду не перевищувала α. Відомо, що при фіксованому значенні β рівень значущості α буде належати інтервалу $\left[\beta, 2\beta-\beta^2\right]$. Останнє дозволяє визначити β, як розв'язок такого співвідношення $2\beta-\beta^2 \le \alpha$.

Самостійна робота. Знайти значення β, як розв'язок останньої нерівності при заданому α.

2) Критерій зростаючих та спадаючих серій. Перевірку гіпотези H_0 за допомогою цього критерію будемо здійснювати при умові, що його альтернативна гіпотеза виглядає таким чином

H_1: *у вибірці наявне систематичне монотонне або циклічне зміщення середнього.*

1. Спочатку у вибірці $x_1, x_2, \ldots, x_n$ замінюємо підряд розташовані однакові виміри одним їх представником.

$$x_1, x_2, \ldots, x_{n'}$$

ставимо символ плюс, якщо його наступний член з вибірки строго більше поточного, і ставимо символ мінус, якщо його наступний член з вибірки строго менше поточного, тобто символи під $i-$им членом розставляються згідно

$$\begin{cases} +, \text{якщо } x'_i < x'_{i+1}, \\ -, \text{якщо } x'_i > x'_{i+1}. \end{cases}$$

Отримуємо деяку послідовність довжини $(n'-1)$ плюсів та мінусів:

$$+-+-\ldots++$$

3. Далі, на базі утвореної послідовності плюсів та мінусів, визначаємо статистики $\nu(n)$ та $\tau(n)$ абсолютно аналогічно, як це робилося у попередньому критерії.

4. Потім використовується та ж сама ідея для побудови *області прийняття нашої гіпотези* про стохастичність вибірки. Вона буде мати ідентичний вигляд:

$$\begin{cases} \nu(n) > \nu_\beta(n), \\ \tau(n) < \tau_{1-\beta}(n), \end{cases}$$

де $\nu_\beta(n), \tau_\beta(n)$ – квантилі рівня β статистик $\nu(n)$ та $\tau(n)$ відповідно.

Зауваження відносно вибору β залишається у силі.

3) Критерій квадратів послідовних різниць (критерій Аббе). Даний критерій використовується при роботі *з нормальними вибірками*. *На цьому класі* вибірок він *є більш потужним* ніж попередні критерії.

Згадаємо як розуміти, що критерій K_1 є більш потужний ніж інший критерій K_2?

При перевірці гіпотези H_0 за допомогою цього критерія, у нього в якості альтернативної виступає така гіпотеза

H_1: *у вибірці наявне систематичне зміщення середнього.*

1. На основі вибірки підрахуємо таку статистику:

$$\gamma(n) = \frac{\frac{1}{2(n-1)}\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2}{\frac{1}{n-1}\left[\left(\sum_{i=1}^{n}x_i^2\right) - n\overline{x}^2(n)\right]},$$

де $\overline{x}(n) = \frac{1}{n}\sum_{i=1}^{n}x_i$.

2. Легко бачити, що до області відхилення гіпотези H_0 потрібно віднести область малих значень, тоді *область прийняття гіпотези* буде мати вигляд:

$$\gamma(n) > \gamma_\alpha(n),$$

де $\gamma_\alpha(n)$ – квантиль рівня α статистики $\gamma(n)$, який можна підрахувати таким чином

$$\gamma_\alpha(n) = \begin{cases} \text{визначається по табл. 4.9 з роботи } [*], \text{ якщо } n \le 60, \\ 1 + \dfrac{u_\alpha}{\left[n + \dfrac{1}{2}\left(1 + u_\alpha^2\right) \right]^{\frac{1}{2}}}, \text{ якщо } n > 60, \end{cases}$$

а u_α – квантиль рівня α стандартного нормального розподілу,

[*] Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.

Практика використання критеріїв перевірки вибірки на стохастичность. …

Розвідувальний аналіз (Exploratory data analysis)

Розвідувальний аналіз – це один із етапів попередньої обробки даних, який дозволяє провести візуальний експрес-аналіз даних на основі засобів їх візуалізації або перетворення, тобто представлення даних у зручному для оперативного аналізу вигляді.

Наприклад, у вигляді різноманітних графіків, діаграм, схем, таблиць і т.п. Результати цього аналізу слугуватимуть відправною точкою для планування подальшої поглибленої обробки інформації.

I. ***Випадок обробки скалярних спостережень.*** У цій ситуації у розвідувальному аналізі можна використовувати:

1) пробіт-графік (probit plot),
2) ймовірнісний графік (probability plot),
3) висячі гістобари (hanging histobars),
4) завмерлу коренеграму (suspended rootogram),
5) зображення "скринька з вусами" (box-and-whisker plot) та його модифікації
(multiple box-and-whisker plot, notched box-and-whisker plot),
6) зображення "стебло-листок" (stem-and-leaf plot), і т.д.

II. ***Випадок обробки двовимірних спостережень.*** У цій ситуації у розвідувальному аналізі можна використовувати:

1) діаграму розсіювання (scatter diagram),
2) таблиця спряженості (contingency table).

Далі буде розглянуто послідовно можливості кожного з цих засобів.

Класи розподілів типу зсув-масштабу

Для опису перших двох графічних представлень даних потрібно познайомитися з класами розподілів типу зсув-масштабу.

Означення. Клас розподілів $\mathcal{F}$ називається ***класом розподілів типу зсув-масштабу***, якщо існує така базова функція розподілу $F_0(\cdot) \in \mathcal{F}$, що для будь-якої функції розподілу $F(\cdot)$ з цього класу існують дійсні a та b $(b > 0)$ такі, що її можна представити таким чином:

$$F(x) = F_0\left(\frac{x-a}{b}\right).$$

Зауважимо, що параметр a називають ***параметром зсуву***, а b - ***параметром масштабу***.

Приклади класів розподілів типу зсув-масштабу.

1. Клас нормальних розподілів.

Дійсно довільну функцію розподілу $F(x)$ нормально розподіленої величини $\xi \sim \mathcal{N}\left(m,\sigma^2\right)$ можна представити у вигляді

$$F(x)=\Phi\left(\frac{x-m}{\sigma}\right),$$

де $\Phi(x)$ – функція розподілу нормально розподіленої величини з параметрами 0 та 1. Для цього класу розподілів: $\Phi(\cdot)$ – базова функція, $a=m,\ b=\sigma$.

2. Клас показникових (експоненціальних) розподілів.

Нехай $F(x)$ - функція показникового розподілу з параметром $\lambda\ (\lambda>0)$, тобто

$$p(x)=F'(x)=\begin{cases}\lambda e^{-\lambda x}, \text{ якщо } x\geq 0,\\ 0 \quad , \text{ якщо } x<0.\end{cases}$$

Тоді справедливо

$$F(x)=\Phi_1(\lambda x),$$

де $\Phi_1(x)$ – функція експоненціального розподілу з параметром 1. Тобто роль базової функції тут відіграє функція $\Phi_1(\cdot)$, а потрібні константи визначаються згідно з $a=0,\ b=\lambda^{-1}$.

I. Випадок обробки скалярних спостережень

Детальніше:

1) Пробіт-графік (probit plot).

Нехай $\mathcal{F}$ - деякий клас розподілів типу зсув-масштабу з базовою функцією $F_0(\cdot)$ для якої існує $F_0^{-1}(\cdot)$.

Розглянемо обробку вибірки спостережень $x_1, x_2, \ldots, x_n$ над скалярною змінною ξ з функцією розподілу $F_\xi(x)$.

Подивимося, який повинен мати вигляд побудований пробіт-графік у випадку коли функція розподілу випадкової величини ξ, яка спостерігається, належить цьому класу розподілів $\mathcal{F}$. Тоді існують a та b $(b>0)$ такі, що

$$\hat{F}_\xi(x) \approx F_\xi(x) = F_0\left(\frac{x-a}{b}\right).$$

А сам пробіт-графік буде мати такий вигляд:

$$y = F_0^{-1}(\hat{F}_\xi(x)) \approx F_0^{-1}\left(F_0\left(\frac{x-a}{b}\right)\right) = \frac{x}{b} - \frac{a}{b}.$$

Призначення. Це дозволяє використовувати цей графік для візуального розв'язку наступних задач:

1) *Перевірки гіпотези* $H_0 : F_\xi(\cdot) \in \mathcal{F}$.

   У випадку справедливості цієї гіпотези пробіт-графік буде уявляти собою приблизно деяку пряму, в протилежному випадку гіпотезу відхиляють.

2) *Виявлення наявності аномальних спостережень у вибірці.* Про присутність викидів у вибірці буде говорити наявність деяких точок графіку, які розташовані суттєво осторонь основної маси точок графіку.

2) Ймовірнісний графік (probability plot).

Побудова. Нехай $\hat{F}_\xi(x)$ – емпірична функція розподілу, яка обчислена по вибірці спостережень $x_1, x_2, \ldots, x_n$ над випадковою величиною ξ.

Ймовірнісний графік для класу розподілів $\mathcal{F}$ – це графік функції $y = \hat{F}_\xi(x)$, побудований на спеціальному *ймовірнісному папері класу розподілів* $\mathcal{F}$. Останній відрізняється від звичайного паперу зміненим масштабом по осі y. З цією метою на такому папері смугу $\{(x,y) : 0 \le y \le 1\}$ трансформують таким чином: $(x,y) \to (x, F_0^{-1}(y))$.

Призначення. Можливості та методика використання ймовірнісного графіку точно такі як і у пробіт-графіку:

1) *Перевірки гіпотези* $H_0 : F_\xi(\cdot) \in \mathcal{F}$.

У випадку справедливості цієї гіпотези пробіт-графік буде уявляти собою приблизно деяку пряму, в протилежному випадку гіпотезу відхиляють.

2) *Виявлення наявності аномальних спостережень у вибірці.* Про присутність викидів у вибірці буде говорити наявність деяких точок графіку, які розташовані суттєво осторонь основної маси точок графіку.

Якщо $\mathcal{F}$ – клас нормальних розподілів, то цей графік називають *нормальним ймовірнісним графіком*, а відповідний папір – *нормальним ймовірнісним папером*.

Наступні два засоби розвідувального аналізу торкаються візульної перевірки гіпотези нормальності.

Наступні два засоби розвідувального аналізу торкаються візульної перевірки гіпотези нормальності.

Означення. ***Нормальним розподілом найбільш узгодженим з вибіркою*** $x_1, x_2, \ldots, x_n$ називається такий нормальний закон

$$\mathcal{N}(\bar{x}(n), s^2(n)),$$

де $\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}(n)\right)^2$.

3) Висячі гістобари (hanging histobars).

Побудова. Спочатку по вибірці $x_1, x_2, \ldots, x_n$ визначають вибіркові значення математичного сподівання $\bar{x}(n)$ та дисперсії $s^2(n)$. Потім будується графік щільності нормального розподілу найбільш узгодженого з вибіркою $x_1, x_2, \ldots, x_n$, а саме $\mathcal{N}(\bar{x}(n), s^2(n))$. Далі у центрі кожного інтервалу групування даних до цієї кривої підвішують гістобару (вузенький прямокутник), висота якої пропорційна відносній частоті попадання вимірів у цей інтервал групування.

Призначення. Висячі гістобари використовують для візуальної перевірки гіпотези нормальності розподілу випадкової величини, яка спостерігається. Гіпотезу приймають, якщо основи гістобар незначно відхиляються від осі абсцис. В протилежному випадку її відхиляють.

4) Завмерла коренеграма (suspended rootogram).

Побудова. Вона представляє собою послідовність прямокутників, побудованих у центрах інтервалів групування даних вибірки $x_1, x_2, \ldots, x_n$, причому висота такого прямокутника для i-того інтервала групування даних пропорційна різниці

$$\sqrt{\nu_i^{(e)}} - \sqrt{\nu_i^{(t)}},$$

де $\nu_i^{(e)}, \nu_i^{(t)}$ – відповідно емпірична та теоретична відносні частоти попадання у i-тий інтервал групування. Остання підрахована згідно нормального розподілу найбільш узгодженого з вибіркою $x_1, x_2, \ldots, x_n$, тобто $\mathcal{N}(\bar{x}(n), s^2(n))$. Якщо $\hat{p}(x)$ - щільність нормального розподілу найбільш узгодженого з вибіркою $x_1, x_2, \ldots, x_n$, то

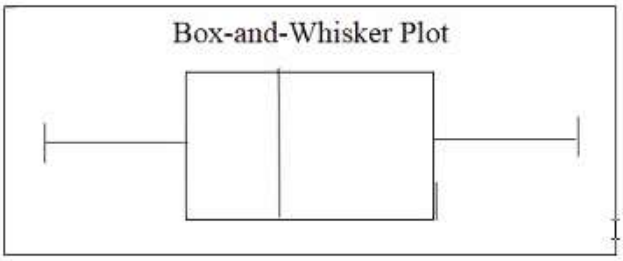
$$\nu_i^{(t)} = \int_{b_{i-1}}^{b_i} \hat{p}(x)dx,$$

де b_{i-1}, b_i - лівий та правий кінці i-того інтервалу групування.

Призначення. Це графічне представлення можна використовувати для візуальної перевірки гіпотези про нормальність розподілу випадкової величини, яка спостерігається. Остання попередньо вважається нормально розподіленою, якщо побудовані прямокутники незначно відхиляються від осі абсцис, інакше вона відхиляється.

5) Зображення "скринька з вусами" (box-and-whisker plot).

Побудова. Воно має у загальному випадку такий нижченаведений вигляд:



Модифікації: multiple box-and-whisker plot, notched box-and-whisker plot.

Призначення. Це зображення надає можливість отримати таку інформацію. Проекція середньої вертикальної лінії скриньки на вісь абсцис дає нам значення медіани, лівої границі скриньки – нижнього квартилю, правої границі скриньки – верхнього квартилю. Проекції лівого кінця лівого вуса та правого кінця правого вуса відповідно дають нам найменше найбільше значення у вибірці. При наявності у вибірці викидів (вимірів, які знаходяться від скриньки на відстані більшій ніж півтори інтерквартильної широти), на зображенні вони будуть представлені у вигляді окремих точок, відображених лівіше та правіше кінців вищевказаних ліній.

Stem-and-Leaf Plot for Variable1: unit = 100 1|2 represents 1200

	LO	18, 19, 21, 21
7	2F	455
18	2S	66666777777
30	2°	888888999999
46	3*	00000000001111111
66	3T	222222233333333333333
84	3F	444444444555555555
106	3S	66666666667777777777777
73	3°	88888888888999999999999999
47	4*	000000000011111
32	4T	22333
27	4F	444444455555555
12	4S	6666777
5	4°	89
3	5*	0
	HI	61, 73

У першому рядку вказано, що це зображення побудовано для змінної Variable1, використовуючи масштабний множник 100. Лівіше вертикальної риски вказується ведуча цифра поточного виміру з одним із фіксованих символів, а правіше вертикальної риски його наступна цифра. Врахувавши масштабний множник 100, одразу отримаємо значення поточного спостереження. Цифра, яка стоїть у першому стовпчику вказує кількість відображених спостережень у поточному рядку плюс у всіх рядках до найближчого краю зображення. У самих крайніх рядках, які починаються з абревіатур LO або HI, можуть вказуватися виміри підозрілі на аномальність.

Призначення. Зображення "стебло-листок" дозволяє візуально з'ясувати загальний вигляд розподілу даних, інтервал їх концентрації, симетричність розподілу, наявність вимірів підозрих на аномальність.

II. Випадок обробки двовимірних спостережень

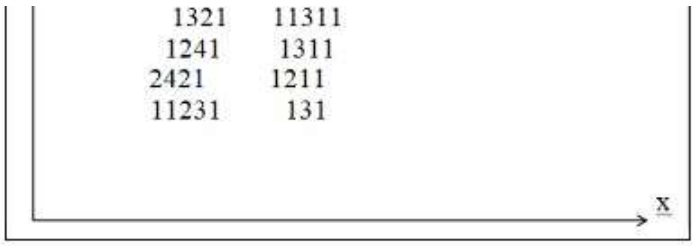
1) Діаграма розсіювання (scatter diagram).

Побудова. Нехай маємо спостереження над двома кількісними скалярними змінними

$$\xi : x_1, x_2, \ldots, x_n,$$

$$\eta : y_1, y_2, \ldots, y_n.$$

Спочатку представимо ці виміри на екрані монітора, який працює у текстовому режимі. У цій ситуації весь екран монітора розбивається на знакомісця (прямокутники). Підраховуємо скільки значень пар (x_i, y_i) з вибірки, тобто точок з координатами (x_i, y_i) попало у кожне знакомісце. А потім усі ці ненульові значення виводимо у відповідні знакомісця. Якщо ці значення з другого десятка, то можна використовувати режим «інверсії», а якщо з третього – режим «blink». У підсумку, побачимо щось на зразок такого



Якщо монітор працює у графічному режимі, то зображення на екрані формується за допомогою різнокольорових пікселів. Тут вже підраховуємо скільки точок з координатами (x_i, y_i) попало в площину кожного пікселя. А потім кожен піксель виводимо на екран тим темнішим відтінком коричневого, чим більше значень пар (x_i, y_i) з вибірки попало в площину цього пікселя.

Призначення. Діаграма розсіювання дозволяє таке:

- з`ясувати загальний вигляд залежності (класу функцій апроксимації залежності) між ξ та η,
- з`ясувати наявність аномальних спостережень.

2) Таблиця спряженості (contingency table).

Призначення. Використовується для табличного представлення спостережень над двома скалярними змінними зі скінченними множинами значень. Це можуть бути номінальні, ординальні, кількісні дискретні або кількісні неперервні змінні, спостереження над якими згруповані.

Побудова. Нехай змінна η приймає всього m_1 значень, а змінна ξ- m_2 значень. Вважаємо, що отримали n спостережень над цими двома скалярними змінними. Тоді кількість наслідків спостережень n_{ij}, коли змінна η прийняла своє $i-$те значення, а ξ своє $j-$те

значення заносимо у комірку на перетені i – того рядка та j – того стовпчика такої таблиці, яку і називають ***таблицею спряженості***:

η \ ξ	1	2	• • •	j	· · ·	m_2	$\sum$
1	n_{11}	n_{12}	• • •	n_{1j}	· · ·	n_{1m_2}	$n_{1\bullet}$
2	n_{21}	n_{22}	• • •	n_{2j}	· · ·	n_{2m_2}	$n_{2\bullet}$
⋮	⋮	⋮	• • •	⋮	· · ·	⋮	⋮
i	n_{i1}	n_{i2}	• • •	n_{ij}	· · ·	n_{im_2}	$n_{i\bullet}$
⋮	⋮	⋮	• • •	⋮	· · ·	⋮	⋮
m_1	$n_{m_1 1}$	$n_{m_1 2}$	• • •	$n_{m_1 j}$	· · ·	$n_{m_1 m_2}$	$n_{m_1\bullet}$
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$	• • •	$n_{\bullet j}$	· · ·	$n_{\bullet m_2}$	n

Значення n_{ij} називають *частотою відповідної комірки*. Значення у останньому рядку – це сума по стовпчикам, значення у останньому стовпчику – це сума по рядкам таблиці спряженості, а значення у правому нижньому кутку – це загальна кількість спостережень, тобто:

$$n_{i\bullet} = \sum_{j=1}^{m_2} n_{ij}\ ,\ n_{\bullet j} = \sum_{i=1}^{m_1} n_{ij}\ ,\ n = \sum_{i=1}^{m_1} n_{i\bullet} = \sum_{j=1}^{m_2} n_{\bullet j}\ .$$

Тут крапка у позначеннях замість індексу позначає, що було здійснено сумування по тому індексу замість якого вона фігурує.