Зміст

Собственно, матстат...

1. Описова статистика.

1.1. Основні поняття математичної статистики.

- **Математична статистика** – це розділ математики, в якому вивчаються методи збору, систематизації та обробки інформації з метою виявлення існуючих закономірностей.

У математичній статистиці набір даних розглядається як реализація або спостереження деякої випадкової величини (в.в.) ξ, яка визначена на деякому ймовірнісному просторі $(\Sigma, \mathcal{F}, P)$, пов'язаний із стохастичним експериментом.

- **Генеральна сукупність** (population) – це (як правило, невідомий) ймовірнісний розподіл $\mathcal{F}$ в.в. ξ, що спостерігається (ймовірнісна міра P)
- **Вибірка** (sample) – це набір незалежних в.в. $\xi_1, \xi_2, ..., \xi_n$, кожна з яких має розподіл $\mathcal{F}$. При цьому n називається об'ємом вибірки.
- **Реалізація вибірки** – це значення $x_1, x_2, ..., x_n$ або $\overrightarrow{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$, які прийняли в.в. $\xi_1, ..., \xi_n$ в результаті конкретного стохастичного експерименту. При цьому x_k називається **варіантою**. Тобто:

$$\begin{matrix} (\Sigma, \mathcal{F}, P) \\ \omega_o \in \Sigma \end{matrix} \qquad x_1 = \xi_1(\omega_0)\ ,\ \cdots\ ,\ x_n = \xi_n(\omega_0) \qquad \begin{matrix} \text{Вибірка: } (\xi_1, ..., \xi_n) \\ \text{Реал. вибірки: } (x_1, ..., x_n) \end{matrix}$$

Основою будь-яких висновків щодо властивостей г.с. $\mathcal{F}$ є **вибірковий метод**, суть якого полягає в тому, що властивості в.в. ξ визначаються шляхом вивчення цих властивостей на випадковій вибірці. Множина всіх реалізацій S вибірки $x_1, \ldots, x_n$ називається **вибірковим простором**.

- Пара $(S, \mathcal{F})$ називається **статистичною моделлю** опису серії спостережень, які породжують вибірку.

- Якщо розподіл $\mathcal{F}_\xi$ відомий з точністю до невідомого вектора параметрів $\overrightarrow{\theta} = \begin{bmatrix} \theta_1 & \cdots & \theta_q \end{bmatrix}$ з множиною значень $\Theta(\overrightarrow{\theta} \in \Theta)$, тоді статистичну модель називають **параметричною моделлю**, а множину Θ - параметричною множиною.

- **Статистикою** (або вибірковою характеристикою) називається довільна борелева функція $g(\overrightarrow{\xi}) = g(\xi_1, \ldots, \xi_n)$ від елементів вибірки. Розподіл цієї в.в. називається **вибірковим розподілом**, а значення $g(\overrightarrow{x})$ статистики за реалізацією $\overrightarrow{x}$ - **вибірковим значенням**.

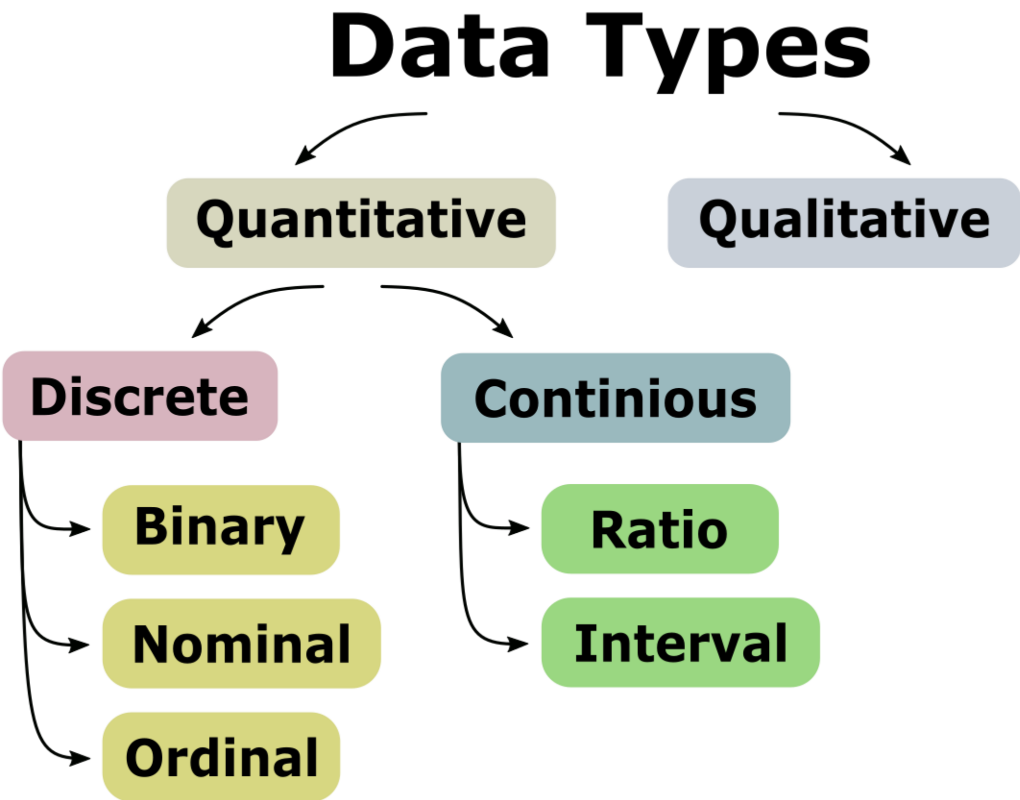
Статистичну модель називають неперервною або дискретною, якщо розподіл г.с. $\mathcal{F}_\xi$ є неперервним або дискретним.

- **Описова статистика** (або дескриптивна статистика, англ. descriptive statistics) – це розділ статистики, що займається обробкою емпіричних даних, їх систематизацією та наочним представленням у вигляді графіків та таблиць.

- **Статистичні виводи** (англ. inferential statistics) – це розділ статистики, що займається вивченням зв'язків між вибірковими даними та г.с., з яких вона отримана.

Зокрема, статистичні виводи займаються відповіддю на питання, які висновки можна зробити щодо г.с. (функція розподілу, щільність або закон розподілу, числові характеристики такі як математичне сподівання, дисперсія, моменти в.в., тощо) за вибіркою.

1.2. Типи даних.



- **Кількісні дані** використовують чисельні значення для опису об'єкта, що нас цікавить.
  - **Дискретні дані** відповідають вибірці з дискретного розподілу г.с. Дискретні дані зазвичай є цілими числами, хоча можуть задаватись і десятковими дробими. Їх особливістю є те, що вони мають "пробіли" ("діри") у своїх можливих значеннях.
  - **Неперервні дані** відповідають вибірці з неперервного розподілу г.с. Неперервні дані приймають будь-які значення з певного діапазону і не мають стрибків
- **Якісні дані** використовують описові вирази для вимірювання або класифікації об'єкта, що нас цікавить.

1.3. Первинна обробка інформації.

- **Варіаційним рядом** $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ називаються елементи вибірки, які впорядковані за зростанням. При цьому:

$$x_{(1)} = \min_{1 \le k \le n} x_k \qquad x_{(n)} = \max_{1 \le k \le n} x_k$$

  x_k називається **порядковою статистикою порядку k**.

Процес впорядкування вибірки називається **"ранжуванням"**.

Розмах вибірки R — різниця між найбільшим та найменшим елементами, тобто:

$$R = x_{(n)} - x_{(1)}$$

Нехай розподіл г.с. $\mathcal{F}$ є дискретною. Тоді нехай $x_1^*, \ldots, x_m^*$ – елементи вибірки, впорядковані за зростанням, причому кожне значення вказується лише один раз, n_k – число разів появи x_k^* в реалізації вибірки. n_k називається **частотою** появи x_k^*.

Зауважимо, що $n_1 + \cdots + n_m = n$.

Сума частот елементів $\sum\limits_{i=1}^{\mathbf{k}} x_i^*$ називається **кумулятивною частотою** n_k^*:

$$n_k^* = n_1 + ... + n_k$$

Величина $\nu_k = \dfrac{n_k}{n}$ називається **відносною частотою**.

Сума відносних частот елементів $\sum\limits_{i=1}^{k} \nu_i$ називається **кумулятивною відносною частотою** елементу x_k^*.

1.3.1. Статистичний ряд.

На основі вказаних вище характеристик можна побудувати **статистичний ряд**:

Значення (x_k^*)	Частоти (n_k)	Кумулятивні частоти (n_k^*)	Відносні частоти (ν_k)	Кумулятивні відносні частоти (ν_k^*)
x_1^*	n_1	$n_1^* = n_1$	$\nu_1 = \frac{n_1}{n}$	ν_1^*
x_2^*	n_2	$n_2^* = n_1 + n_2$	$\nu_2 = \frac{n_2}{n}$	ν_2^*
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m^*	n_m	$n_m^* = \sum_{i=1}^{m} n_i = n$	$\nu_m = \frac{n_m}{n}$	$\nu_m^* = \sum_{i=1}^{m} \nu_i = 1$

Якщо розподіл г.с. є неперервною або дискретною з великою кількістю значень, тоді використовують **інтервальний статистичний ряд**.

Область, в якому лежать всі значення реалізації вибірки розбивають на **m класів**:

$$\Delta_1 = [y_0, y_1), \cdots, \Delta_m = [y_{m-1}, y_m]$$

Кількість класів зазвичай варіюється в межах 5 - 20. Іноді, для визначення кількості класів використовують **формулою Стреджеса**:

$$m - 1 = \log_2 n$$

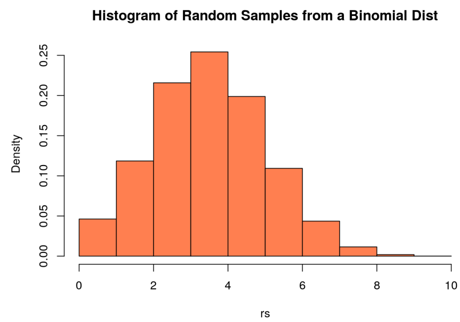
Зауваження. Реалізація вибірки може містити так звані **викиди**. У випадку наявності викідів або великої порожніх класів на початку або в кінці вибірки, доцільно використовувати при розбиті класи з відкритими кінцями (клас, що має лише нижню або верхню межу).

Зауваження. В деяких випадках спостереження можуть надаватись не як індивідуальні значення, а вже розподілені по інтервалах тобто у вигляді інтервального статистичного розподілу. В цьому випадку дані називаються **групованими**.

1.4. Графічні методи представлення інформації.

1.4.1. Гістограма.

Якщо розподіл г.с. є неперервним, гістограма будується на основі інтервального статистичного розподілу г.с. наступним чином: будують систему координат таку, що на осі абсцис будуть відображатися інтервали $\Delta_1, ..., \Delta_m$, а на осі ординат відповідні частоти.



Далі будують у вказаній системі координат прямокутники з основами $\Delta_k, k \in \overline{1, m}$, та відповідними висотами n_k.

Також використовують гістограму відносних частот, в якій висота стовпців визначається за формулою:

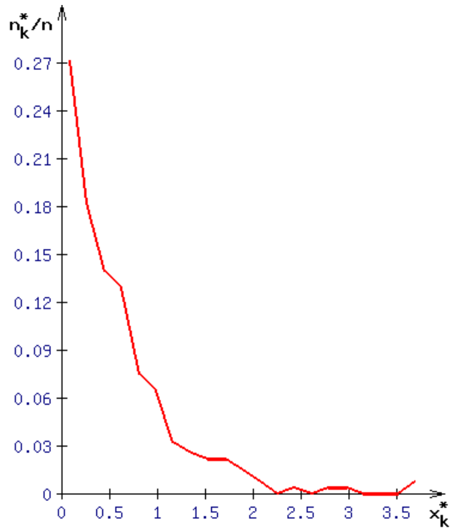
$$h_k = \frac{\nu_k}{l(\Delta_k)}$$

Зауважимо, що в цьому випадку площа отриманої фігури буде дорівнювати 1.

1.4.2. Полігон частот.

Якщо розподіл г.с. є дискретним, тоді полігон частот будується на основі статистичного розподілу г.с. наступним чином: будують систему координат таку, що на осі абсцис будуть відображатися елементи вибірки $x_1^*, \ldots, x_m^*$, а на осі ординат відповідні частоти.

Далі у вказаній системі координат будують точки $M_k(x_k^*, n_k), k = \overline{1, m}$, які з'єднують між собою у ламану $M_1 M_2 \ldots M_m$.



2. Дескриптивні міри.

Графічні методи є чудовим способом для отримання швидкого огляду вибіркових даних, але вони не є точними та не призводять самі по собі до наступних досліджень. Для цього нам потрібно введення числових параметрів таких як, наприклад, середнє.

Існують різні шляхи, за допомогою яких ми можемо спробувати описати розподіл. Розглянемо деякі з них, які корисні при описі гістограми або полігона частот.

Надалі, нехай $x_1, \ldots, x_n$ реалізація вибірки з г.с. $\mathcal{F}$.

2.1. Міри центральної тенденції (measures of central tendency).

2.1.1. Вибіркове середнє.

Вибіркове середнє $\overline{x}$ є одним із найбільш відомих параметрів центральної тенденції і знаходиться за формулою:

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

Зауваження. Вказана формула може бути використана лише у випадку, коли всі індивідуальні значення x_j є відомими.

Якщо ж відносно вибірки є відомим лише розбиття спостережень на класи (Δ_i) та частота попадань елементів вибірки у кожен з цих інтервалів (n_i). В цьому випадку групованих даних можна використовувати наступну формулу:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{m} n_i x_i^* = \sum_{i=1}^{m} \nu_i x_i^*$$

2.1.2. Вибіркова медіана.

Однією із вад вибіркового середнього воно дає нерепрезентативні результати, оскільки є чутливим до викидів у вибірці та симетрії. Тому іноді є більш привабливим використання інших параметрів. Наприклад, **вибіркової медіани**.

Нехай у вибірці відомі всі індивідуальні спостереження. Будуємо варіаційних ряд:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

Знаходимо медіану як середній елемент цього варіаційного ряду M_e^*:

$$M_e^* = \begin{cases} x_{(k+1)}, & n = 2k+1; \\ \dfrac{x_{(k)} + x_{(k+1)}}{2}, & n = 2k; \end{cases}$$

Якщо вибірка представлена групованими даними, тоді вибіркова медіана розраховується наступним чином:

1) Знаходимо клас (позначимо його номер me), що буде містити елемент, який відповідає медіані, тобто:

$$n_{me-1}^* < \frac{n}{2} \qquad n_{me}^* \geq \frac{n}{2}$$

2) Елемент знайденого класу, що буде відповідати медіані, обчислюємо за наступною формулою:

$$M_e^* = y_{me-1} + (y_{me} - y_{me-1}) \frac{\frac{n}{2} - n_{me}^* - 1}{n_{me}}$$

Зазначимо, що дріб у правій частині останньої рівності вказує нам, яку частину інтервалу ми маємо пройти, щоб дійти до медіани.

2.1.3. Вибіркова мода.

Вибіркова мода M_o^* — це елемент вибірки, який зустрічається частіше за все.

У випадку групованих даних визначення більш складне. Клас (позначимо його номер mo), який має найбільшу частоту, називається **модальним**, причому при розбитті на інтервали, вони мають бути однакової ширини.

Значення моди знаходиться за формулою:

$$M_o^* y_{mo-1} + (y_{mo} - y_{mo-1}) \frac{n_{mo} - n_{mo-1}}{(n_{mo} - n_{mo-1}) + (n_{mo} - n_{mo+1})}$$

Зауваження. Вказані три параметри центральної тенденції дають різну інформацію. Але якщо розподіл симетричний, то вони будуть давати приблизно однакові значення.

2.2. Міри розсіювання (Measure of dispersion)

При розгляді двох різних розподілів, вони можуть мати однакові або дуже близькі середні, але при цьому суттєво відрізнятись. Наприклад, при порівнянні рівня добробуту двох країн: у них може бути дуже близькими середній добробут по країні, але в одній країні всі отримують приблизно однаковий рівень достатку, а в інший можуть екстремальні значення великого достатку та злиденного. Параметри розсіювання як раз і потрібні для відокремлення таких випадків.

2.2.1. Розмах вибірки (range).

Розмах вибірки є найпростішою мірою розсіювання, який обчислюється як різниця між найбільшим та найменшим спостереженням даної вибірки:

$$R = x_{(n)} - x_{(1)}$$

2.2.2. Середнє абсолютне відхилення (mean absolute error).

Якщо відомо математичне сподівання розподілу $\mathcal{F}, \mathbb{E}\xi_j = \mu$, тоді середнє абсолютне відхилення розраховується наступним чином:

- Якщо дані не груповані, то:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|x_j - \mu|$$

- Якщо дані груповані, то:

$$MAE = \frac{1}{n}\sum_{j=1}^{m} n_i \left|x_j^* - \mu\right| = \sum_{j=1}^{m} \nu_j \left|x_j^* - \mu\right|$$

Якщо ж математичне сподівання не відомо, то μ у формулах змінюється на $\overline{x}$.

2.2.3. Вибіркова дисперсія (variance).

Найбільш корисним параметром розсіювання є **вибіркова дисперсія**. Вона є середнім квадратів відхилень елементів вибірки від середнього значення.

Якщо $\mathbb{E}\xi_j = \mu$, тоді вибіркова дисперсія розраховується наступним чином:

1) Якщо дані не груповані, то:

$$\sigma_*^2 = \mathbb{D}_\xi^* = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)^2$$

2) Якщо дані груповані, то:

$$\sigma_*^2 = \mathbb{D}_\xi^* = \frac{1}{n}\sum_{i=1}^{m} n_i(x_i^* - \mu)^2$$

Якщо ж $\mathbb{E}\xi_j$ не відомо, тоді розглядають або **вибіркову дисперсію**.

1) Якщо дані не груповані, то:

$$s^2 = \mathbb{D}_\xi^{**} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2$$

2) Якщо дані груповані, то:

$$s^2 = \mathbb{D}_\xi^{**} = \frac{1}{n}\sum_{i=1}^{m} n_i(x_i^* - \overline{x})^2$$

Або **виправлену (незміщену) вибіркову дисперсію**.

1) Якщо дані не груповані, то:

$$s_0^2 = \mathbb{D}_\xi^{***} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$$

2) Якщо дані груповані, то:

$$s_0^2 = \mathbb{D}_\xi^{***} = \frac{1}{n-1}\sum_{i=1}^{m} n_i(x_i^* - \overline{x})^2$$

2.2.4. Стандартне відхилення (standard deviation).

Оскільки при знаходженні дисперсії ми підносимо відхилення до квадрату, то не зовсім зрозумілим є одиниці вимірювання дисперсії. Тому природньо взяти від цього виразу корінь квадратний.

Таким чином, ми отримуємо величину, рівну корню квадратному від вибіркової дисперсії, яка називається **стандартним відхиленням** і задається виразом:

$$\sigma_{\xi}^{*,**,***} = \sqrt{\mathbb{D}_{\xi}^{*,**,***}}$$

2.2.5. Коефіцієнт варіації (coefficient of variation).

Коефіцієнт варіації (відносне стандартне відхилення) визначається, як процентне відношення стандартного відхилення до вибіркового середнього.

$$C_V = \frac{\sigma_{\xi}^{***}}{\overline{x}} \cdot 100\%$$

Зауваження. Він є відносним показником розсіювання(мінливості) ознаки(вибірки) по відношенню до середнього показника вибірки. (Може використовуватися лише для відносного рівня вимірювання). Використання коефіцієнта варіації є суттєвим лише при вивченні ознаки, яка має **додатне значення**.

2.3. Міри позиції.

Міри позиції є індикаторами, як окрема величина співвідноситься з іншими.

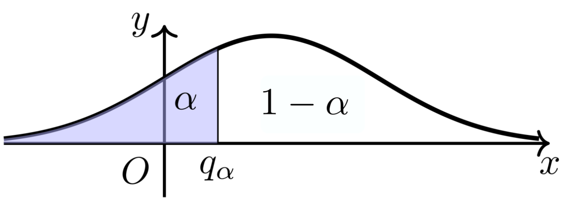
2.3.1. Квантиль.

Нехай $\alpha \in (0, 1)$. **α-квантилем** в.в. ξ називається таке число $q_\alpha \in \mathbb{R}$, що:

$$\mathbb{P}\{\xi \le q_\alpha\} \ge \alpha \qquad \mathbb{P}\{\xi \ge q_\alpha\} \ge 1 - \alpha$$

Зауважимо, що для неперервних в.в. α-квантиль однозначно знаходиться з рівняння:

$$F(q_\alpha) = \alpha$$



де F - функція розподілу в.в. ξ.
Знайдемо, як слід шукати вибіркові значення квантилів. Якщо у вибірці відомі всі індивідуальні спостереження. Побудуємо варіаційних ряд:

$$x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}$$

Елемент, що буде відповідати α-квантилю цього варіаційного ряду:

$$q_\alpha = \begin{cases} x_{(\lfloor K \rfloor + 1)}, & K = n \cdot \alpha \notin \mathbb{Z}; \\ \dfrac{x_{(K)} + x_{(K+1)}}{2}, & K = n \cdot \alpha \in \mathbb{Z}; \end{cases}$$

Для групованих даних α-квантиль q_α визначається наступним чином:

1) Знаходимо клас (позначимо його номер K_α), що буде містити α-квантиль, тобто:

$$n^*_{K_\alpha - 1} < n \cdot \alpha \qquad n^*_{K_\alpha} \ge n \cdot \alpha$$

   де n^*_i — кумулятивна частота i-го класу;

2) Обчислюємо значення α-квантиля за наступною формулою:

$$q_\alpha = y_{K_\alpha - 1} + (y_{K_\alpha} - y_{K_\alpha - 1}) \frac{\alpha \cdot n - n^*_{K_\alpha - 1}}{n_{K_\alpha}}$$

де $y_{K_{\alpha-1}}, y_{K_\alpha}$ — відповідно нижня та верхня межі класу, який містить квантиль, n_{K_α} — частота класу, який містить квантиль.

2.3.2. Процентиль.

p-процентилем (percentile) називається величина P_p, яка співпадає з квантилем порядку $\alpha = \dfrac{p}{100}$ тобто:

$$P_p = q_{\frac{p}{100}}$$

Квартилями (quartile) називають процентилі кратні 25-ти:

- $Q_1 = P_{25}$ — нижній квартиль;
- $Q_1 = P_{50} = M_e$ — медіана;
- $Q_1 = P_{50} = M_e$ — медіана;

Децилями (decile) називають процентилі кратні 10-ти:

$$D_1 = P_{10}, D_2 = P_{20}, \dots, D_9 = P_{90}$$

Міжквартильним діапазоном (interquartile range) називають величину:

$$IQR = Q_3 - Q_1$$

Помітимо, що міжквартильний діапазон є мірою розсіювання. Його недоліком є те, що він вимірює розкид в середині даних, а не по всій множині.

2.4. Міри форми.

Міри форми – є індикаторами симетричності та “згладженості” даних.

2.4.1. Вибірковий момент.

Вибірковим моментом k-го порядку називається величина:

$$\overline{x}^k = \frac{1}{n}\sum_{j=1}^{n} x_j^k$$

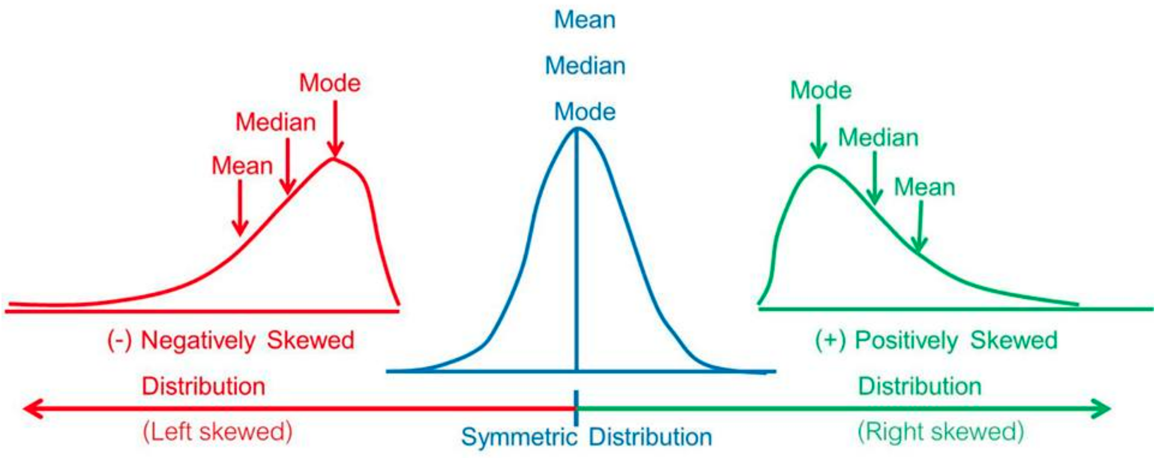
Вибірковим центральним моментом k-го порядку називається величина:

$$\overline{\mu}_k = \frac{1}{n}\sum_{j=1}^{n} (x_j - \overline{x})^k$$

2.4.2. Коефіцієнт асиметрії (skewness).

Коефіцієнтом асиметрії називається число:

$$A_S = \frac{\overline{\mu}_3}{s_0^3}$$



Коефіцієнтом асиметрії характеризує **міру скошеності розподілу** даних щодо вибіркового середнього:

- $As = 0$ – дані симетричні щодо середнього;
- $As > 0$ – дані мають правосторонню асиметрію;
- $As < 0$ – дані мають лівосторонню асиметрію.

2.5. Коефіцієнт асиметрії Пірсона.

Коефіцієнтом асиметрії Пірсона називається число:

$$Sk = \frac{\overline{x} - M_0}{s_0}$$

Недоліком коефіцієнта асиметрії Пірсона полягає в тому, що він враховує лише центральну частину розподілу.

2.5.1. Коефіцієнт ексцесу (kurtosis).

Коефіцієнтом ексцесу називається число:

$$Ek = \frac{\overline{\mu}_4}{s_0^4} - 3$$

яке характеризує міру "гостровершинності" розподілу даних у порівнянні із нормальним розподілом:

- $Ek = 0$ – розподіл даних близький до нормального;
- $Ek > 0$ – розподіл даних більш "гостровершинний" ніж нормальний розподіл;
- $Ek < 0$ – розподіл даних більш "пласковершинний" ніж нормальний розподіл.

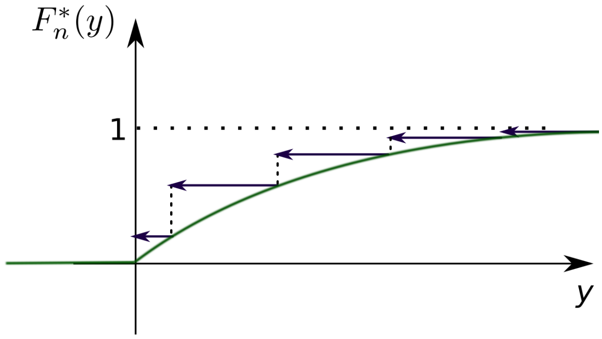
3. Властивості вибіркових характеристик.

3.1. Емпірична функція розподілу. Властивості.

Емпіричною функцією розподілу, побудованою за вибіркою $\xi_1, \ldots, \xi_n$ об'єму n, називається випадкова функція:

$$F_n^* : \mathbb{R} \times \Sigma \to [0, 1]$$

при кожному $y \in \mathbb{R}$ рівна:



$$F_n^*(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{\xi_i < y\right\}$$

Із збільшенням кількості спостережень, емпірична функція розподілу наближається до теоретичної функції розподілу г.с. Зазначимо її основні властивості.

Теорема 3.1 (Консистентність)**.** Нехай $\xi_1, \ldots, \xi_n$ – вибірка з розподілу $\mathcal{F}$ з ф.р. F та нехай F_n^* – емпірична ф.р., яку побудовано по цій вибірці. Тоді, $\forall y \in \mathbb{R}$:

$$F_n^*(y) \xrightarrow[n\to\infty]{\text{м.н.}} F(y)$$

Доведення. За означенням:

$$F_n^*(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\xi_i < y) \xrightarrow[n\to\infty]{\text{м.н.}} \underbrace{\mathbb{E}\mathbb{I}\left\{\xi_i < y\right\}}_{=F_\xi(y)}$$

Оскільки індикатори подій $\{\xi_1 < y\}, \{\xi_1 < y\}, \ldots, \{\xi_n < y\}$ є незалежними й однаково розподіленими, їх математичне сподівання скінчено:

$$\mathbb{E}\mathbb{I}\left\{\xi_1 < y\right\} = 1 \cdot \mathbb{P}\left\{\xi_1 < y\right\} + 0 \cdot \mathbb{P}\left\{\xi_1 \geq y\right\} = \mathbb{P}\left\{\xi_1 < y\right\} = F(y)$$

Застосовуючи посилений ЗВЧ : $F_n^*(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{\xi_i < y\right\} \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\mathbb{I}\left\{\xi_1 < y\right\} = F_\xi(y)$

■

Теорема 3.2 (Глівенко-Кантеллі)**.** В умовах теореми 3.1:

$$\sup_{y\in\mathbb{R}} |F_n^*(y) - F(y)| \xrightarrow[n\to\infty]{\text{м.н.}} 0$$

Без доведення $\blacksquare$.

Теорема 3.3 (Колмогорова)**.** Нехай $\xi_1, ..., \xi_n$ – вибірка з неперервною ф.р. $F \in C^1(\mathbb{R})$, а F_n^* – емпірична ф.р. **Тоді:**

$$\sqrt{n} \sup_{y\in\mathbb{R}} |F_n^*(y) - F(y)| \xrightarrow[n\to\infty]{} \eta,$$

де в.в. η має розподіл Колмогорова з неперервною ф.р.:

$$F_K(y) = \begin{cases} \sum\limits_{j=-\infty}^{\infty} (-1)^j e^{-2j^2y^2}, & y \geq 0; \\ 0, & y < 0. \end{cases}$$

Без доведення $\blacksquare$.

Нехай $\xi_1, \ldots, \xi_n$ — вибірка з розподілу $\mathcal{F}$ з функцією розподілу F, а $F_n^*(y)$ – емпірична функція розподілу. Тоді для довільного $y \in \mathbb{R}$:

1) $\mathbb{E}F_n^*(y) = F(y)$ *(незміщеність оцінки)*;
2) $\mathbb{D}F_n^*(y) = \dfrac{F(y)(1 - F(y))}{n}$;
3) $\sqrt{n}\,(F_n^*(y) - F(y)) \implies N(0, F(y)(1 - F(y)))$ при $F(y) \neq 0, 1$;
4) величина $nF_n^*(y)$ має біноміальний розподіл $B(n.F(y))$.

Доведення. Помітимо, що $\mathbb{I}\{\xi_1 < y\}$ має розподіл Бернуллі $B(F(y))$, а тому:

$$\mathbb{E}\mathbb{I}\{\xi_1 < y\} = F(y) \qquad \mathbb{D}\mathbb{I}\{\xi_1 < y\} = F(y)(1 - F(y))$$

Оскільки, крім того $\mathbb{I}\{\xi_1 < y\}, \mathbb{I}\{\xi_2 < y\}, \ldots$ є незалежними, то:

$$nF_n^*(y) = \sum_{i=1}^{n} \mathbb{I}\{\xi_1\} \ \sim \ B(n, F(y)) \implies \textit{(властивість 4)}$$

Властивості 1, 2 випливають з 4-ї. Для доведення 3-ї використаємо ЦГТ:

$$\sqrt{n}(F_n^* - F(y)) = \frac{\sum\limits_{i=1}^{n} \mathbb{I}\{\xi_i < y\} - nF(y)}{\sqrt{n}} = \frac{\sum\limits_{i=1}^{n} \mathbb{I}\{\xi_i < y\} - n\mathbb{E}\mathbb{I}\{\xi_1 < y\}}{\sqrt{n}} \implies$$

$$\implies N(0, \mathbb{D}\mathbb{I}\{\xi_1 < y\}) = N(0, F(y)(1 - F(y))) \ \ n \to \infty.$$

■

3.2. Властивості вибіркових моментів.

Нехай $\xi_1, \ldots, \xi_n$ — вибірка з розподілу $\mathcal{F}$. Тоді:

1) Якщо $\mathbb{E}|\xi_1| < \infty$, то $\mathbb{E}\overline{\xi} = \mathbb{E}\xi_1 = a$ – незміщеність $\overline{\xi}$;

2) Якщо $\mathbb{E}|\xi_1| < \infty$, то $\overline{\xi} \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\xi_1 = a$ – консистентність $\overline{\xi}$;

3) Якщо $\mathbb{D}\xi_1 < \infty, \mathbb{D}\xi_1 \neq 0$, то $\sqrt{n}(\overline{\xi} - \mathbb{E}\xi_1) \xrightarrow[n\to\infty]{} N(0, \mathbb{D}\xi_1)$

Доведення.

- Властивість 1 випливає із властивостей математичного сподівання.

- Доведення 2 та 3 випливає безпосередньо із застосування ЗВЧ Хінчина та ЦГТ, відповідно.

■

Вибірковий k-й момент $\overline{\xi}^k$ є незміщенною, консистентною та асимптотично нормальною для теоретичного k-го момента. Нехай $\xi_1, \ldots, \xi_n$ – вибірка з розподілу $\mathcal{F}$. Тоді:

1) Якщо $\mathbb{E}|\xi_1|^k < \infty, \mathbb{E}\overline{\xi}^k = \mathbb{E}\xi_1^k = m_k$;

2) Якщо $\mathbb{E}|\xi_1|^k < \infty$, то $\overline{\xi}^k \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\xi_1^k$;

3) Якщо $\mathbb{D}\xi_1^k < \infty, \mathbb{D}\xi_1^k \neq 0$, то $\sqrt{n}(\overline{\xi}^k - \mathbb{E}\xi_1^k) \xrightarrow[n\to\infty]{} N(0, \mathbb{D}\xi_1^k)$

Вибіркові дисперсії мають наступні властивості:

Нехай $\xi_1, \ldots, \xi_n$ — вибірка з розподілу $\mathcal{F}$ та $\mathbb{D}\xi_1 < \infty$. Тоді:

1) $\mathbb{E}s^2 = \frac{n-1}{n}\mathbb{D}\xi_1 = \frac{n-1}{n}\sigma^2 \neq \sigma^2, \mathbb{E}s_0^2 = \mathbb{D}\xi_1 = \sigma^2$;

2) $s^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{D}\xi_1 = \sigma^2, s_0^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{D}\xi_1 = \sigma^2$;

4. Точкові оцінки параметрів Г.С.

Нехай є генеральна сукупність $\mathcal{F}$ випадкової величини ξ з відомим розподілом, але невідомим вектором параметрів $\overrightarrow{\theta} = \begin{bmatrix} \theta_1 & \cdots & \theta_n \end{bmatrix}$.

Оцінка θ^* параметру θ — деяка статистика, значення якої "близькі" до θ:

$$\theta_n^* = \varphi(\xi_1, \ldots, \xi_n) \qquad \theta \approx \theta^*(\xi_1, ..., \xi_n)$$

4.1. Методи побудови точкових оцінок.

4.1.1. Метод моментів.

Нехай є генеральна сукупність $\mathcal{F}$ випадкової величини ξ, яка має характеристики:

Теоретичні $(\mathbb{E}\xi, \mathbb{D}\xi, \mathbb{E}\xi^k, \ldots)$ та **Вибіркові** $(\overline{\xi}, \mathbb{D}_\xi^{*,**,***})$

Ідея методу моментів – прийняти вибіркові значення характеристик за теоретичні.

4.1.2. Метод максимальної вірогідності (MLE).

Нехай $\mathcal{F}$ – дискретна генеральна сукупність. Маємо: $\xi_1, \ldots, \xi_n \xrightarrow{\text{реалізація}} x_1, \ldots, x_n$.

$$\mathcal{L}(x_1, \ldots, x_n, \theta) = \mathbb{P}\left\{\xi_1 = x_1, \ldots, \xi_n = x_n\right\} \ - \ \textbf{Likelihood function.}$$

$$\mathcal{L}(x_1, \ldots, x_n, \theta) = |\perp\!\!\!\perp| = \prod_{i=1}^{n} \mathbb{P}\left\{\xi_i = x_i\right\} = \prod_{i=1}^{\infty} \mathbb{P}_\theta\left\{\xi = x_i\right\} \xrightarrow{\theta} \max$$

Надалі максимізуємо вираз, застосувавши властивість монотонності логарифма:

$$\ln \mathcal{L}(x_1, \ldots, x_n, \theta) = \sum_{i=1}^{\infty} \ln \mathbb{P}_\theta\left\{\xi = x_i\right\} \xrightarrow{\theta} \max$$

$$\frac{\partial \ln \mathcal{L}(x_1, \ldots, x_n, \theta)}{\partial \theta} = \sum_{i=1}^{\infty} \frac{\partial}{\partial \theta} \ln \mathbb{P}_\theta\left\{\xi = x_i\right\} = 0 \implies \theta^*$$

Нехай $\mathcal{F}$ – неперервна генеральна сукупність. Маємо щільність розподілу вибірки:

$$\mathcal{L}(x_1, \ldots, x_n, \theta) = f_{\xi_1, \ldots, \xi_n}(x_1, x_2, \ldots, x_n) = |\perp\!\!\!\perp| = \prod_{i=1}^{n} f_{\xi_i}(x_i) = \prod_{i=1}^{n} f_\xi(x_i)$$

Скористалися однаковою розподіленістю величин $\xi_1, \ldots, \xi_n$ з щільністю $f_\xi(x)$.

Надалі пошук оцінки θ^* аналогічно до дискретного випадку.

4.2. Властивості оцінок.

4.2.1. Незміщеність(unbiasedness).

Означення. θ^* — **незміщена** оцінка параметру θ, якщо $\mathbb{E}\theta^* = \theta$.

Означення. θ^*_n — **асимптотично незміщена** оцінка параметру θ, якщо:

$$\mathbb{E}\theta^* \xrightarrow[n\to\infty]{} \theta$$

4.2.2. Консистентність.

Означення. θ^*_n називається **консистентною** оцінкою параметра θ, якщо:

$$\theta^*_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta \quad \text{— **слабка**} \qquad \theta^*_n \xrightarrow[n\to\infty]{\text{м.н.}} \theta \quad \text{— **сильна**}$$

Як перевіряти консистентність?

1) За (посиленим, якщо у сенсі *м.н.*) законом великих чисел.

2) За означенням збіжності ($\mathbb{P}$, *м.н.*).

3) **Лема.** Для θ^*_n: $\begin{cases} \text{(Асимптотично) незміщена.} \\ \mathbb{D}\theta^*_n \xrightarrow[n\to\infty]{} 0 \end{cases} \Longrightarrow \theta^*_n$ – слабко консистентна.

Доведення.

$$\begin{cases} \mathbb{E}\theta^*_n \xrightarrow[n\to\infty]{} 0; \\ \mathbb{D}\theta^*_n \xrightarrow[n\to\infty]{} 0. \end{cases} \Longrightarrow \left| \begin{array}{c} \text{За критерієм} \\ \mathbb{L}_2\text{-збіжності до } const \end{array} \right| \Longrightarrow \theta^*_n \xrightarrow[n\to\infty]{\mathbb{L}_2} \theta \Longrightarrow \theta^*_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta$$

■

4.3. Порівняння точкових оцінок. Ефективні оцінки.

Надалі обмежимося випадком **незміщених** оцінок параметру θ_n^*.

Нехай знайдено 2 незміщених оцінки параметру θ – θ_1^*, θ_2^*. Намагаємося визначити, яка з оцінок "краща". Абсолютно природнім критерієм порівняння може бути вираз:

$$\mathbb{E}(\theta_1^* - \theta)^2 = \theta_1^* \qquad\qquad \mathbb{E}(\theta_2^* - \theta)^2 = \mathbb{D}\theta_2^*$$

$$\mathbb{E}(\theta_1^* - \theta)^2 = \Big| \text{ незміщеність } \Big| = \mathbb{E}(\theta_1^* - \mathbb{E}\theta_1^*)^2 = \mathbb{D}\theta_2^*$$

Таким чином, сутність запропонованого критерію полягає у порівнянні дисперсій.

Означення. Оцінка θ_1^* **негірше** θ_2^*, якщо $\mathbb{D}_\theta\theta_1^* \le \mathbb{D}_\theta\theta_2^* \ \ \forall\theta$.

Приклад. $U(0,\theta), \theta_1^* = 2\overline{\xi}, \theta_2^* = 2\xi_n$. Для цих параметрів:

1) $\mathbb{D}\theta_1^* = 4\dfrac{\mathbb{D}\xi_1}{n} = \dfrac{\theta}{3n}$;

2) $\mathbb{D}\theta_2^* = \mathbb{D}(2\xi_n) = 4 \cdot \dfrac{\theta}{12} = \dfrac{\theta}{3}$.

Означення. Незміщена оцінка θ^* параметру θ називається **ефективною** в класі незміщених оцінок, якщо для Θ – множини усіх різних незміщених оцінок θ:

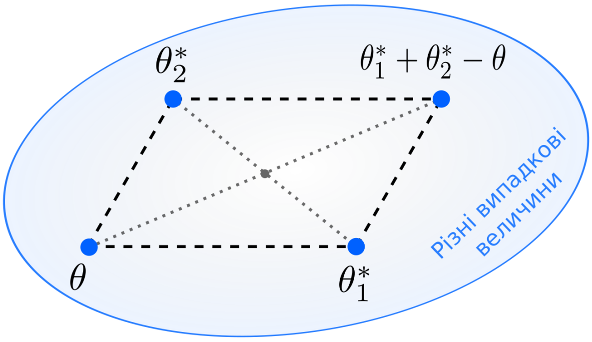
$$\mathbb{D}\theta^* = \min_{\hat{\theta}\in\Theta} \mathbb{D}\hat{\theta}$$

Теорема 4.1. В класі незміщених оцінок Θ є не більше однієї ефективної.

Доведення. Введемо функцію відстані між двома в.в.: $d(\eta_1,\eta_2) = \sqrt{\mathbb{E}(\eta_2-\eta_1)^2}$. Тоді:

$$\mathbb{E}\theta_1^* = \theta \qquad \mathbb{E}\theta_2^* = \theta \qquad \text{Припустимо: } D = \mathbb{D}\theta_1^* = \mathbb{D}\theta_2^* \le \mathbb{D}\hat{\theta}.$$

Розглянемо $\theta_3^* = \frac{\theta_1^*+\theta_2^*}{2}$, що є також незміщеною: $\mathbb{E}\theta_3^* = \mathbb{E}\frac{\theta_1^*+\theta_2^*}{2} = \frac{1}{2}(\mathbb{E}\theta_1^* + \mathbb{E}\theta_2^*) = \theta$.



$$2d^2(\theta,\theta_2^*) + 2d^2(\theta,\theta_1^*) =$$
$$= d^2(\theta, \theta_1^* + \theta_2^* - \theta) + d^2(\theta_1^*, \theta_2^*)$$
$$2\,\underbrace{\mathbb{E}(\theta_2^* - \theta)^2}_{\mathbb{D}\theta_2^* = D} + 2\,\underbrace{\mathbb{E}(\theta_1^* - \theta)^2}_{\mathbb{D}\theta_1^* = D} =$$
$$= \mathbb{E}(\theta_1^* + \theta_2^* - 2\theta)^2 + \mathbb{E}(\theta_2^* - \theta_1^*)^2$$
$$4D = 4\,\underbrace{\mathbb{E}(\frac{\theta_1^* + \theta_2^*}{2} - 2\theta)^2}_{\mathbb{D}\theta_3^*} + \mathbb{E}(\theta_2^* - \theta_1^*)^2$$

$$4D = 4\mathbb{D}\theta_3^* + \mathbb{E}(\theta_2^* - \theta_1^*)^2 \implies \mathbb{D}\theta_3^* = D - \frac{1}{4}\underbrace{\mathbb{E}(\theta_2^* - \theta_1^*)^2}_{>0} \xRightarrow{\theta_1^* \neq \theta_2^*} \mathbb{D}\theta_3^* < D$$

Отримали суперечність, адже за побудовою $\mathbb{D}\theta_3^* > D$. ■

4.3.1. Нерівність Рао-Крамера.

Задана вибірка $\overrightarrow{\xi}$ з Г.С. $\mathcal{F}$ з невідомим параметром розподілу θ.

$$\mathcal{L} = (\theta, x_1, \ldots, x_n) = \begin{cases} \prod\limits_{i=1}^{n} \mathbb{P}\{\xi = x_i\}, & \text{ДВВ;} \\ \prod\limits_{i=1}^{n} f_\xi(x_i), & \text{АНВВ.} \end{cases}$$

$$\mathbb{E}\left(\frac{\partial \ln \mathcal{L}(\theta, \overline{\xi})}{\partial \theta}\right)^2 = I(\theta) \text{ — } \textbf{кількість інформації за Фішером.}$$

Теорема 4.2 (Нерівність Рао-Крамера)**.** θ^* – незміщена оцінка параметра θ.

Якщо функція вірогідності L є “досить гладкою” за θ, то виконується нерівність:

$$\mathbb{D}\theta^* \geq \frac{1}{I(\theta)}$$

Доведення. Розглянемо випадок, $\mathcal{F}$ – абсолютно неперервна:

$$\underset{\mathbb{R}^n}{\int\cdots\int}\left[\theta^*(\overrightarrow{x})\cdot f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\right]\mathrm{d}\overrightarrow{x}=\mathbb{E}\theta^*(\overrightarrow{\xi})=\mathbb{E}\theta^*=\theta\cdot\underbrace{\underset{\mathbb{R}^n}{\int\cdots\int}f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\mathrm{d}\overrightarrow{x}}_{=1}$$

$$\theta\cdot\underset{\mathbb{R}^n}{\int\cdots\int}f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\mathrm{d}\overrightarrow{x}-\underset{\mathbb{R}^n}{\int\cdots\int}\left[\theta^*(\overrightarrow{x})\cdot f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\right]\mathrm{d}\overrightarrow{x}=0$$

$$\mathbf{J}=\underset{\mathbb{R}^n}{\int\cdots\int}\left[(\theta^*(\overrightarrow{x})-\theta)\cdot f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\right]\mathrm{d}\overrightarrow{x}=0$$

$$\frac{\partial\mathbf{J}}{\partial\theta}=\underset{\mathbb{R}^n}{\int\cdots\int}\left[\frac{\partial}{\partial\theta}\left(\theta^*(\overrightarrow{x})-\theta\right)\cdot f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)+(\theta^*(\overrightarrow{x})-\theta)\cdot\frac{\partial}{\partial\theta}f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\right]\mathrm{d}\overrightarrow{x}=$$

$$=\underset{\mathbb{R}^n}{\int\cdots\int}\left[-f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)+(\theta^*(\overrightarrow{x})-\theta)\cdot\frac{\partial f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)}{\partial\theta}\right]\mathrm{d}\overrightarrow{x}$$

$$\underset{\mathbb{R}^n}{\int\cdots\int}\left[(\theta^*(\overrightarrow{x})-\theta)\cdot\frac{\partial f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)}{\partial\theta}\right]\mathrm{d}\overrightarrow{x}=\underset{\mathbb{R}^n}{\int\cdots\int}f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\mathrm{d}\overrightarrow{x}=1$$

Із зазначених вище міркувань, отримали наступний факт:

$$\underset{\mathbb{R}^n}{\int\cdots\int}(\theta^*(\overrightarrow{x})-\theta)\cdot\underbrace{\frac{\partial f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\Big/\partial\theta}{f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)}}_{=\frac{\partial}{\partial\theta}\ln f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)}\cdot f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\mathrm{d}\overrightarrow{x}=1$$

$$1=\mathbb{E}\left[(\theta^*(\overrightarrow{x})-\theta)\cdot\frac{\partial}{\partial\theta}\ln f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\Big|_{\overrightarrow{x}=\overrightarrow{\xi}}\right]=|1|\leq$$

$$\leq\sqrt{\mathbb{E}\left(\theta^*(\overrightarrow{x})-\theta\right)^2\cdot\mathbb{E}\left(\frac{\partial}{\partial\theta}\ln f_{\overrightarrow{\xi}}(\overrightarrow{x},\theta)\Big|_{\overrightarrow{x}=\overrightarrow{\xi}}\right)^2}$$

Остаточно: $1\leq\mathbb{E}\left(\theta^*(\overrightarrow{x})-\theta\right)^2\cdot\mathbb{E}\left(\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta,\overrightarrow{\xi})\right)^2\Rightarrow\boxed{\mathbb{D}\theta_3^*\geq\frac{1}{\mathbb{E}\left(\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta,\overrightarrow{\xi})\right)^2}=\frac{1}{I(\theta)}}$ ■

Наслідок (практичний спосіб перевірки ефективності)**.** Незміщена оцінка θ^* є ефективною $\Longleftrightarrow$ в нерівності Коші-Буняковського досягається рівність $\Longleftrightarrow$ множники пропорційні $\Longleftrightarrow$ $\boxed{\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta,\overrightarrow{\xi})=C\cdot\left[\theta^*(\overrightarrow{x})-\theta\right],\text{де } C=C(\theta,n)}$.

5. Інтервальні оцінки параметрів.

Для того, щоб перейти до поняття інтервальної оцінки, нам знадобиться деякий теоретичний мінімум щодо розподілу числових характеристик гаусівських вибірок.

5.1. Розподіли числових характеристик гаусівських вибірок.

Нехай є Г.С.: $\xi \sim N(a, \sigma^2)$, що породжує вибірку $\xi_1, \ldots, \xi_n$ – незалежні, гаусівські.

- Маємо вибіркове середнє $\overline{\xi} = \frac{\xi_1 + \cdots + \xi_n}{n}$.
- Якщо правжнє математичне сподівання a відоме: $\mathbb{D}^*_\xi = \frac{1}{n}\sum\limits_{i=1}^{n}(\xi_i - a)^2$;
- Інакше можемо знайти $\mathbb{D}^{**}_\xi = \frac{1}{n}\sum\limits_{i=1}^{n}(\xi_i - \overline{\xi})^2$ або $\mathbb{D}^{***}_\xi = \frac{1}{n-1}\sum\limits_{i=1}^{n}(\xi_i - \overline{\xi})^2$.

Як розподілені ці характеристики?

5.1.1. Вибіркове середнє.

За властивостями гаусівських величин: $\boxed{\overline{\xi} \sim N\left(a, \dfrac{\sigma^2}{2}\right)}$.

5.1.2. Вибіркова дисперсія.

Нехай *відоме* $\mathbb{E}\xi = a \Longrightarrow \mathbb{D}^*_\xi$: $\dfrac{n\mathbb{D}^*_\xi}{\sigma^2} = \dfrac{n\frac{1}{n}\sum\limits_{i=1}^{n}(\xi_i - a)^2}{\sigma^2} = \sum\limits_{i=1}^{n}\left(\dfrac{\xi_i - a}{\sigma}\right)^2 = \sum\limits_{i=1}^{n}\tilde{\xi}_i^2 \sim \chi^2_n.$

$$\forall i : \xi_i \sim N(a, \sigma^2) \quad \Longrightarrow \quad \tilde{\xi}_i = \frac{\xi_i - a}{\sigma} \sim N(0,1) \quad \Longrightarrow \quad \boxed{\frac{n\mathbb{D}^*_\xi}{\sigma^2} \sim \chi^2_n}$$

Нехай $\mathbb{E}\xi$ – *невідоме*. Тоді:

$$\mathbb{D}^{**}_{\xi} = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2 = \frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right)^2 = \overline{\xi^2} - (\overline{\xi})^2$$

Спершу розглянемо величину коваріації між $\overline{\xi}$ та $\xi_i - \overline{\xi}$:

$$\operatorname{cov}(\overline{\xi}, \xi_i - \overline{\xi}) = \mathbb{E}\overline{\xi}\left(\xi_i - \overline{\xi}\right) - \mathbb{E}\overline{\xi}\cdot\underbrace{\mathbb{E}\left(\xi_i - \overline{\xi}\right)}_{=a-a=0} = \mathbb{E}\frac{\xi_1 + \dots + \xi_n}{n}\left(\xi_i - \frac{\xi_1 + \dots + \xi_n}{n}\right) =$$

$$= \mathbb{E}\left(\frac{\xi_1 + \dots + \xi_n}{n}\cdot\xi_i\right) - \mathbb{E}\left(\frac{\xi_1 + \dots + \xi_n}{n}\right)^2 = \mathbb{E}\frac{\xi_i^2}{n} + \sum_{j\neq i}\frac{\xi_i\cdot\xi_j}{n} -$$

$$-\frac{1}{n^2}\left(\mathbb{E}\xi_1^2 + \dots + \mathbb{E}\xi_n^2 + 2\mathbb{E}\xi_1\xi_2 + \dots + 2\mathbb{E}\xi_{n-1}\xi_n\right) = \left|\begin{array}{c}\mathbb{E}\xi_i^2 - (\mathbb{E}\xi_i)^2 = \mathbb{D}\xi_i\\ \mathbb{E}\xi_i^2 = a^2 + \sigma^2\end{array}\right| =$$

$$= \frac{1}{n}(a^2+\sigma^2)\frac{n-1}{n}\cdot a^2 + \frac{1}{n^2}(na^2 + n\sigma^2 - 2\frac{n(n-1)}{2}a^2) =$$

$$= \frac{a^2}{n} + \frac{\sigma^2}{n} + a^2 - \frac{a^2}{n} - \frac{a^2}{n} - \frac{\sigma^2}{n} - a^2 + \frac{a^2}{n} = \mathbf{0}$$

Остаточно, отримали, що $\overline{\xi}$ не корелює з $\xi_i - \overline{\xi}$ $\forall i$. Тобто:

$$\operatorname{corr}(\overline{\xi}, \xi_i - \overline{\xi}) = 0 \quad \forall i$$

$$\Downarrow$$

$$\operatorname{corr}\left(\overline{\xi}, \begin{bmatrix}\xi_1 - \overline{\xi}\\ \vdots \\ \xi_n - \overline{\xi}\end{bmatrix}\right) = 0$$

$$\text{За власт.} \Downarrow N(a, \sigma^2)$$

$$\overline{\xi} \perp\!\!\!\perp \begin{bmatrix}\xi_1 - \overline{\xi}\\ \vdots \\ \xi_n - \overline{\xi}\end{bmatrix}$$

$$\Downarrow$$

$$\overline{\xi} \perp\!\!\!\perp \sum_{i=1}^{n}(\xi_i - \overline{\xi})^2 = \mathbb{D}^{**}_{\xi}$$

Для гаусівської Г.С. вибіркове середнє та вибіркова дисперсія **є незалежними**.

Повернемося до визначення розподілу вибіркової дисперсії:

$$\mathbb{D}_\xi^{**} = \frac{1}{n}\sum_{i=1}^n \xi_i^2 - \left(\frac{1}{n}\sum_{i=1}^n \xi_i\right)^2 \quad \Longrightarrow \quad \frac{1}{n}\sum_{i=1}^n \xi_i^2 = \mathbb{D}_\xi^{**} + \left(\frac{1}{n}\sum_{i=1}^n \xi_i\right)^2$$

$$\frac{\xi_i - a}{\sigma} = \tilde{\xi}_i \longrightarrow \xi_i \qquad \chi_n^2 \sim \sum_{i=1}^n \tilde{\xi}_i^2 = \underbrace{n\mathbb{D}_{\tilde{\xi}}^{**}}_{\chi_{n-1}^2} + \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\xi}_i\right)^2}_{N(0,1)^2 \sim \chi_1^2}$$

В останньому міркуванні скористалися фактом незалежності, отриманим раніше:

$$\overline{\xi} \perp\!\!\!\perp \mathbb{D}_\xi^{**} \quad \Longrightarrow \quad \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\xi}_i\right)^2 \perp\!\!\!\perp n\mathbb{D}_{\tilde{\xi}}^{**}$$

Таким чином, довели, що $n\mathbb{D}_{\tilde{\xi}}^{**} \sim \chi_{n-1}^2$. Розпишемо:

$$n\mathbb{D}_{\tilde{\xi}}^{**} = \sum_{i=1}^n \left(\tilde{\xi}_i - \frac{\tilde{\xi}_1 + \dots + \tilde{\xi}_n}{n}\right)^2 = \sum_{i=1}^n \left(\frac{\xi_i - a}{\sigma} - \frac{(\xi_1 - a) + \dots + (\xi_n - a)}{n\sigma}\right)^2 =$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^n \left(\xi_i - a - \frac{\xi_1 + \dots + \xi_n}{n} + \frac{an}{n}\right)^2 = \frac{n}{\sigma^2} \cdot \frac{1}{n}\sum_{i=1}^n \left(\xi_i - \overline{\xi}\right)^2 = \boxed{\frac{n}{\sigma^2} \cdot \mathbb{D}_\xi^{**}}$$

Теорема 5.1. Остаточно, довели такі властивості:

1) $\overline{\xi} \sim N\left(a, \frac{\sigma^2}{n}\right)$

2) $\frac{n \cdot \mathbb{D}_\xi^{**}}{\sigma^2} \sim \chi_n^2$

3) $\frac{n \cdot \mathbb{D}_\xi^{**}}{\sigma^2} \sim \chi_{n-1}^2$ та $\frac{(n-1) \cdot \mathbb{D}_\xi^{***}}{\sigma^2} \sim \chi_{n-1}^2$

4) $\overline{\xi} \perp\!\!\!\perp \mathbb{D}_\xi^{**}$

Наслідок. Розглянемо деякі наслідки з попередньої теореми:

1) $\sqrt{n}\frac{\overline{\xi} - a}{\sigma} \sim N(0,1)$

2) $\frac{n \cdot \mathbb{D}_\xi^{*}}{\sigma^2} \sim \chi_n^2$

3) $\boxed{\sqrt{n-1}\frac{\overline{\xi} - a}{\sqrt{\mathbb{D}_\xi^{**}}} \sim St_{n-1}}$

4) $\frac{n\mathbb{D}_\xi^{**}}{\sigma^2} \sim \chi_{n-1}^2$

5.2. Інтервальні оцінки параметрів генеральної сукупності.

Мотивація: існує суттєвий недолік точкових оцінок — **невідома їх точність.**

Означення. Інтервал (θ_1^*, θ_2^*) називають **довірчим інтервалом** для параметру θ з довірчою імовірністю (надійністю) γ, якщо $\mathbb{P}\{(\theta_1^*, \theta_2^*) \ni \theta\} = \gamma$.

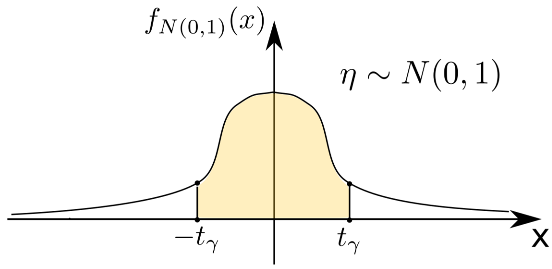
Зауваження (1). θ_1^*, θ_2^* – випадкові величини. θ – невідоме конкретне число. Таким чином, розглядаємо ймовірність потрапляння у інтервал як функцію від $\theta_1^*(\overrightarrow{\xi}), \theta_2^*(\overrightarrow{\xi})$.

Зауваження (2). Зазвичай $\gamma = 0.9, 0.95, 0.99$. $(\theta_1^*, \theta_2^*) \xrightarrow[\gamma \to 1]{} (-\infty, +\infty)$

5.2.1. Знаходження точних довірчих інтервалів для гаусівської Г.С.

1) Інтервал для $\mathbb{E}\xi = a$ при відомій дисперсії:

$$\sqrt{n}\frac{\overline{\xi} - a}{\sigma} \sim N(0,1) \quad \Longrightarrow$$



$$\gamma = \mathbb{P}\{\eta \in (-t_\gamma, t_\gamma)\} = \Phi(t_\gamma) - \Phi(-t_\gamma) = 2\Phi(t_\gamma) \Longrightarrow \boxed{t_\gamma = \Phi^{-1}\left(\frac{\gamma}{2}\right)}$$

$$\gamma = \mathbb{P}\left\{\sqrt{n}\frac{\overline{\xi} - a}{\sigma} \in (-t_\gamma, t_\gamma)\right\} \quad \Longrightarrow \quad \boxed{\overline{\xi} - \frac{\sigma t_\gamma}{\sqrt{n}} < a < \overline{\xi} + \frac{\sigma t_\gamma}{\sqrt{n}}}$$

2) Інтервал для дисперсії при відомому $\mathbb{E}\xi = a$:

$$\frac{n \cdot \mathbb{D}_\xi^*}{\sigma^2} \sim \chi_n^2 \quad \Longrightarrow$$

Аналогічно, визначаємо інтервал на PDF:

$$\eta \sim \chi_n^2 : \mathbb{P}\left\{t_{\frac{1+\sigma}{2}} < \eta < t_{\frac{1-\sigma}{2}}\right\}$$

$$t_{\frac{1+\sigma}{2}} < \frac{n\cdot\mathbb{D}^*_\xi}{\sigma^2} < t_{\frac{1-\sigma}{2}} \quad\Longrightarrow\quad \boxed{\frac{n\cdot\mathbb{D}^*_\xi}{t_{\frac{1-\sigma}{2}}} < \sigma^2 < \frac{n\cdot\mathbb{D}^*_\xi}{t_{\frac{1+\sigma}{2}}}} \qquad t \text{ – для розподілу } \chi^2_n$$

3) Інтервал для дисперсії при невідомому $\mathbb{E}\xi$:

$$\frac{n\mathbb{D}^{**}_\xi}{\sigma^2} \sim \chi^2_{n-1} \quad\Longrightarrow\quad \boxed{\frac{n\cdot\mathbb{D}^*_\xi}{t^{(n-1)}_{\frac{1-\sigma}{2}}} < \sigma^2 < \frac{n\cdot\mathbb{D}^*_\xi}{t^{(n-1)}_{\frac{1+\sigma}{2}}}} \qquad t^{(n-1)} \text{ – для розподілу } \chi^2_{n-1}$$

4) Інтервал для $\mathbb{E}\xi = a$ при невідомій дисперсії:

$$\sqrt{n-1}\frac{\overline{\xi}-a}{\sqrt{\mathbb{D}^{**}_\xi}} \sim St_{n-1} \quad\Longleftarrow$$

Знов користуємося наслідком теореми 5.1 та визначаємо інтервал на $f_{St_{n-1}}(x)$

$$-t_\gamma < \frac{\sqrt{n-1}\cdot(\overline{\xi}-a)}{\sqrt{\mathbb{D}^{**}_\xi}} < t_\gamma \quad\Longrightarrow\quad \boxed{\overline{\xi} - \frac{t_\gamma\sqrt{\mathbb{D}^{**}_\xi}}{\sqrt{n-1}} < a < \overline{\xi} + \frac{t_\gamma\sqrt{\mathbb{D}^{**}_\xi}}{\sqrt{n-1}}}$$

де t_γ – для розподілу Стьюдента з $n-1$ ступенями вільності.

5.2.2. Точні довірчі інтервали для будь-яких розподілів.

Загальний алгоритм пошуку точних довірчих інтервалів.

1) Знайти таку функцію $H(\overrightarrow{x}, \theta)$, що:
   - $H(\overrightarrow{\xi}, \theta)$ – розподіл цієї величини не залежить від θ.
   - $\exists H^{-1} : \theta = H^{-1}(\overrightarrow{\xi})$.
2) $H(\overrightarrow{\xi}, \theta)$ – розподіл не залежить від $\theta \implies \gamma = \mathbb{P}\left\{t_1 < H(\overrightarrow{\xi}, \theta) < t_2\right\}$.
3) Розв'язуємо відносно θ. Отримаємо: $\theta^*_1(\overline{\xi}) < \theta < \theta^*_2(\overline{\xi})$.

5.2.3. Асимптотичны довірчі інтервали.

Означення. $(\theta^*_{1,n}, \theta^*_{2,n})$ називається **асимптотичним довірчим інтервалом**, якщо:

$$\lim_{n\to\infty} \mathbb{P}\left\{\theta^*_{1,n} < \theta < \theta^*_{2,n}\right\} = \gamma$$

Алгоритм пошуку асимптотичного довірчого інтервалу.

1) θ^* – точкова оцінка, яка є асимптотично нормальною (наприклад, містить суми незалежних однаково розподілених випадкових величин).

2) $\dfrac{\theta^* - \mathbb{E}\theta^*}{\sqrt{\mathbb{D}\theta^*}} \approx N(0,1)$ при великих n. Шукаємо симетричний довірчий інтервал:

$$\Phi(t_\gamma) = \Phi(t_\gamma) - \Phi(0) = \frac{\gamma}{2} \quad \Longrightarrow \quad t_\gamma : \Phi(t_\gamma) = \frac{\gamma}{2}$$

$$\mathbb{P}\left\{-t_\gamma < N(0,1) < t_\gamma\right\} = \gamma \quad \Longrightarrow \quad \boxed{\mathbb{P}\left\{-t_\gamma < \frac{\theta^* - \mathbb{E}\theta^*}{\sqrt{\mathbb{D}\theta^*}} < t_\gamma\right\} \approx \gamma}$$

3) Розв'язуючи цю нерівність відносно θ, отримаємо асимптотичний довірчий интервал для θ.

6. Статистичні гіпотези. Перевірка гіпотез.

Означення. **Статистична гіпотеза** — це довільне припущення про розподіл Г.С. або його параметри, яке можна перевірити на основі вибіркових даних.

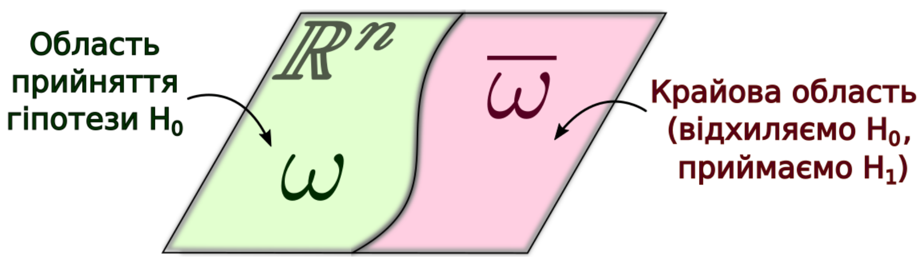
- **Непараметрична гіпотеза** — це твердження, про розподіл Г.С. Приклад: H_0 – гаусівська.
- **Параметрична гіпотеза** — це твердження про параметри Г.С. Приклад:$H_0 = Pois(\lambda)$.
- **Проста гіпотеза** — це гіпотеза, що повністю визначає розподіл Г.С. або його параметри.
- **Складна гіпотеза** — стверджує про належність Г.С. до певного сімейства розподілів.

Розглянемо H_0 – *основна гіпотеза.* Якщо ж H_0 не виправдовується, то ми приймаємо *альтернативну гіпотезу* – H_1 і відхиляємо H_0.

Означення. **Статистичний критерій** — процедура, яка на основі вибіркових даних дозволяє або прийняти H_0 , або відхилити H_0 і прийняти H_1.

Приклад. Нехай є вибірка $\xi_1, \ldots, \xi_n$, H_0 – основна гіпотеза, H_1 – альтернативна.

Вибираємо статистику $T(\xi_1, \ldots, \xi_n)$ та простір змагань $\mathbb{R}^n(\xi_1, \ldots, \xi_n)$:



Головне питання буде полягати у виборі зазначених на малюнку областей. Вони будуть визначатися задопомогою обраної статистики.

Припустимо: за умови виконання $H_0 : T(\xi_1, \ldots, \xi_n)$ з великою імовірністю потрапляє в деяку область. Якщо статистика T потрапляє в область $\omega \Rightarrow$ приймаємо H_0.

Розглянемо більш конкретний приклад. Маємо: Г.С. $\xi \sim N(a, \sigma^2)$, σ – відома.

$$H_0 = \{a = a_0\} \qquad H_1 = \{a = a_1\} \qquad a_0 < a_1$$

$$T(\xi_1, \ldots, \xi_n) = \frac{\xi_1 + \cdots + \xi_n}{n} = \overline{\xi}$$

Таким чином, оберемо $x \in (a_0, a_1)$ так, що:

- Якщо $\overline{\xi} \leq x$: приймаємо $H_0 \Rightarrow \omega = \{\overline{y} \in \mathbb{R}^n : \overline{y} \leq x\}$.
- Якщо $\overline{\xi} > x$: приймаємо $H_1 \Rightarrow \overline{\omega} = \{\overline{y} \in \mathbb{R}^n : \overline{y} > x\}$.

6.1. Помилки I-го та II-го роду.

		Істинна гіпотеза	
		H_0	H_1
Результат застосування критерію	H_0	H_0 правильно прийнята	H_0 неправильно прийнята (Похибка *II* роду)
	H_1	H_0 неправильно знехтувана (Похибка *I* роду)	H_0 правильно знехтувана

- Якщо *істинна* гіпотеза помилково відкидається, то ця помилка називається **помилкою першого роду** (англ. type I errors α errors, false positives).
- Якщо помилково приймається *хибна* гіпотеза - це **помилка другого роду** (англ. type II errors β errors, false negatives).

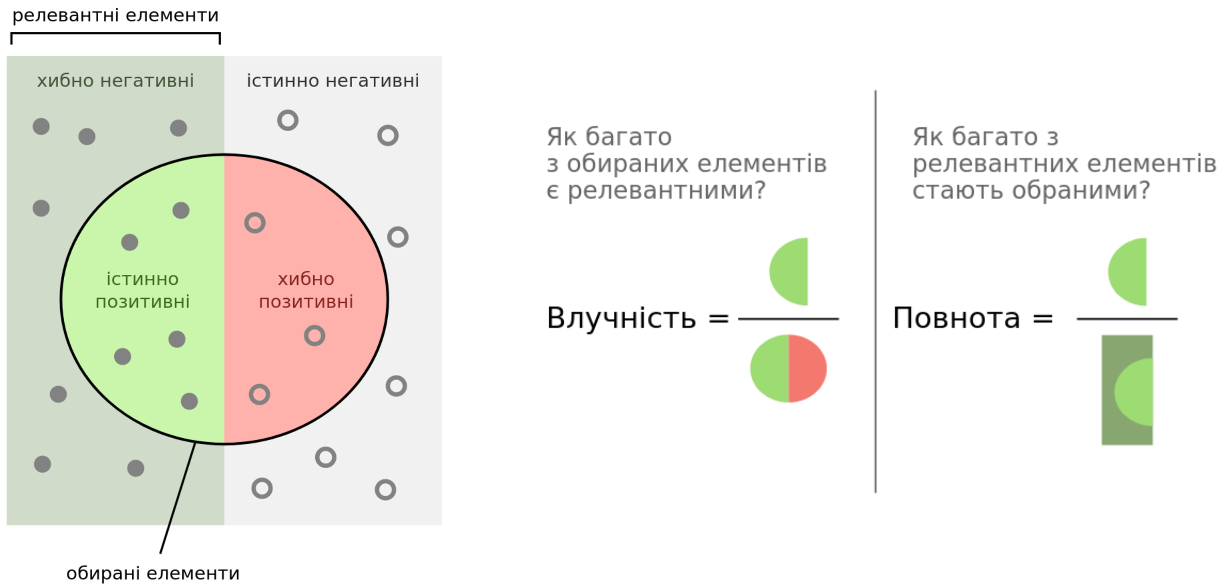
6.1.1. Влучність та повнота (Precision and recall).

В розпізнаванні образів, інформаційному пошуку та класифікації широко застосовуються дві метрики, які відображають наявність помилок першого і другого роду.

Означення. **Влучність** (англ. precision, яку також називають прогностичною значущістю позитивного результату) є часткою релевантних зразків серед знайдених.

Означення. **Повнота** (англ. recall, відома також як чутливість) є часткою загального числа позитивних зразків, яку було дійсно знайдено.

Влучність не слід плутати з **точністю** (англ. accuracy), яка є часткою правильно спрогнозованих результатів, як позитивних, так і негативних.



6.1.2. Рівень значущості та потужність критерію.

Рівень значущості критерію — це ймовірність помилки I-го роду:

$$\mathbb{P}\left\{\text{Приймаємо } H_1 \mid H_0 - \text{вірна}\right\} = \mathbb{P}\left\{\overrightarrow{\xi} \in \overline{\omega} \mid H_0 - \text{вірна}\right\} = \alpha$$

Потужність критерію дорівнює $(1-\beta)$, де β – ймовірність помилки II-го роду:

$$\mathbb{P}\left\{\text{Приймаємо } H_0 \mid H_1 - \text{вірна}\right\} = \mathbb{P}\left\{\overrightarrow{\xi} \in \omega \mid H_1 - \text{вірна}\right\} = \beta$$

Повернемося до прикладу. Саме від значущості та потужності критерію буде залежати положення "межі" x. Найчастіше вибирають значення: $\alpha = 0.01, 0.05, 0.1$.

Зафіксуємо рівень значущості $\alpha = 0.05$ та побудуємо критерій (знайдемо x).

$$\alpha = \mathbb{P}\left\{\overrightarrow{\xi} \in \overline{\omega} \;\middle|\; H_0\right\} = \mathbb{P}\left\{\overline{\xi} > x \;\middle|\; a = a_0\right\} =$$

$$= \left| \begin{array}{c} a = a_0 \Longrightarrow \xi_1, \ldots, \xi_n \sim N(a_0, \sigma^2) \\ \Downarrow \\ \overline{\xi} = \dfrac{\xi_1, \ldots, \xi_n}{n} \sim N(a_0, \sigma^2/n) \end{array} \right| = \frac{1}{2} - \Phi\left(\sqrt{n}\frac{x - a_0}{\sigma}\right) = \frac{1}{2} - \alpha$$

$$\sqrt{n}\frac{x - a_0}{\sigma} = t_\alpha \quad \Longrightarrow \quad \boxed{x = a_0 + \frac{\sigma t_\alpha}{\sqrt{n}}}$$

$$1 - \beta = 1 - \mathbb{P}\left\{\overrightarrow{\xi} \in \omega \;\middle|\; H_1\right\} = 1 - \mathbb{P}\left\{\overrightarrow{\xi} \in \omega \;\middle|\; a = a_1\right\} = \frac{1}{2} + \Phi\left(\sqrt{n} \cdot \frac{a_1 - x}{\sigma}\right)$$

Бачимо, що зменшуючи α – збільшуємо β: $\boxed{\alpha \downarrow \Rightarrow t_\alpha \to \infty \Rightarrow \beta \uparrow \Rightarrow (1 - \beta) \downarrow}$.

Знайдемо рівень значущості та потужність критерію для загального випадку:

$$\alpha = \mathbb{P}\left\{\overrightarrow{\xi} \in \overline{\omega} \;\middle|\; H_0\right\} = \underset{\overline{\omega}}{\int \cdots \int} f_{\overline{\xi}}(x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n =$$

$$= \underset{\overline{\omega}}{\int \cdots \int} \underbrace{f_\xi(x_1) \cdot \ldots \cdot f_\xi(x_n)}_{\mathcal{L}_{H_0}(\overrightarrow{x})} \mathrm{d}x_1 \ldots \mathrm{d}x_n = \underset{\overline{\omega}}{\int \cdots \int} \mathcal{L}_{H_0}(\overrightarrow{x}) \mathrm{d}x_1 \ldots \mathrm{d}x_n$$

Аналогічно, для потужності критерію: $\beta = \underset{\omega}{\int \cdots \int} \mathcal{L}_{H_1}(\overrightarrow{x}) \mathrm{d}x_1 \ldots \mathrm{d}x_n$

Зробимо висновок щодо прийняття чи відхилення гіпотези H_0:

- Більш категоричний факт – **відхилення гіпотези** H_0: на рівні значущості α, гіпотеза H_0 *суперечить* вибірковим даним.

- **Прийняття** H_0: -//- H_0 *не суперечить* вибірковим даним.

6.2. Критерій χ^2 (критерій Пірсона).

Припустимо, є Г.С. та вибірка з неї: $\overrightarrow{\xi} = (\xi_1, \ldots, \xi_n)$.

Нехай F – теоретична функція розподілу ξ. Висуваємо гіпотезу:

$$H_0 = \{F(x) = F_0(x)\} \qquad H_1 = \{F(x) \neq F_0(x)\}$$

Розіб'ємо область значень Г.С. $\mathcal{E}_\xi$ на інтервали $\Delta_1, \Delta_2, \ldots, \Delta_m$: $\begin{cases} \Delta_i \cap \Delta_j = \emptyset & \forall i \neq j \\ \bigcup\limits_i \Delta_i = \mathcal{E}_\xi \end{cases}$

Складаємо таблицю. Наведемо опис застосованих характеристик:

1) Кількість елементів, що потрапили в $\Delta_k = n_k$;

2) Середня кількість елементів, що потрапили в Δ_k, якщо розподіл вгаданий правильно, тобто: $\mathbb{E}$ (кількість елементів вибірки в Δ_k) $= \mathbb{E}\mathrm{Bin}(n, \underbrace{\mathbb{P}\left\{\xi \in \Delta_k \;\middle|\; H_0\right\}}_{=p_k})$

Характеристики	Δ_1	Δ_2	$\cdots$	Δ_k	$\cdots$	Δ_m
1. $n_k = \mathrm{card}\left\{x_i \;\middle|\; x_i \in \Delta_k\right\}$	n_1	n_2	$\cdots$	n_k	$\cdots$	n_m
2. $\mathbb{E}n_{k,H_0} = \mathbb{E}\mathrm{Bin}(n, p_k) = np_k$	np_1	np_2	$\cdots$	np_k	$\cdots$	np_m

Головне завдання може бути сформульовано наступним чином: наскільки близькими є *реальні спостереження* (I рядок) до отриманих за припущенням (II рядок). Можемо використовувати для порівняння: $\sum\limits_{k=1}^{m} \frac{(n_k - np_k)^2}{np_k} = \chi^2(n)$

Теорема 6.1 (Пірсона про критерій χ^2)**.** Якщо гіпотетичний розподіл вгадано правильно (тобто, H_0 справдовується), то:

$$\chi^2(n) \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_{m-1}$$

Вправа: Довести.

6.2.1. Застосування критерію χ^2 для перевірки простих гіпотез.

Якщо H_0 виконана, то за теоремою при великих $n\ :\ \chi^2(n)$ має розподіл приблизно χ^2_{m-1}.

Припустимо, що $\chi^2 > \chi^2_{m-1,\alpha}$. Розглянемо можливі причини:

1) Відбулася рідкісна подія ймовірності α; при цьому гіпотеза H_0 виконується, але є помилково відхиленною $\Longrightarrow$ виникає помилка I-го роду.

2) Гіпотеза H_0 не виконується $\Longrightarrow H_0$ відхиляємо, приймаємо H_1.

Запишемо вимоги до n, p_k. Дані умови є суто практичними та можуть відрізнятися:

$$\begin{matrix} \forall k \in \overline{1,m} \\ np_k \geq 5 \end{matrix} \qquad \begin{cases} m \geq 20 \\ np_k \geq 5 \end{cases} \qquad \begin{cases} m \leq 20 \\ np_k \geq 10 \end{cases}$$

6.2.2. Застосування критерію χ^2 для перевірки складних гіпотез.

Приклад складної гіпотези: $H_0 = \{$ Г.С. розподілена за законом Пуассона $\}$.

H_0 – складна гіпотеза про розподіл Г.С., s – кількість параметрів розподілу, тоді:

$$\chi^2(n,\theta^*) = \sum_{i=1}^{m} \frac{(n_i - np_i(\theta^*))^2}{np_i(\theta^*)} \qquad \widehat{\theta} = \underset{\theta^*\in\Theta}{\arg\min}\ \chi^2(n,\theta^*) \qquad p_i = \mathbb{P}\{\underbrace{\xi}_{\sim P(\theta^*)} < \Delta_i\}$$

Теорема 6.2 (Теорема Фішера)**.** Якщо виконана гіпотеза H_0, то:

$$\chi^2(n) \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_{m-s-1} \qquad \chi^2(n) = \chi^2(n,\widehat{\theta})$$

Якщо $\chi^2(n,\widehat{\theta}_{MLE}) < \chi^2_{m-s-1,\alpha}$, тоді: $\chi^2(n,\widehat{\theta}) < \chi^2(n,\widehat{\theta}_{MLE}) < \chi^2_{m-s-1,\alpha}$.

Тобто, якщо гіпотеза не суперечить даним при перевірці одною з оцінок параметрів, то гіпотезу можна прийняти на цьому рівні значущості.

6.2.3. Застосування критерію χ^2 для перевірки незалежності.

Нехай є Г.С.: $\begin{bmatrix} \xi \\ \eta \end{bmatrix}$. Та вибірка: $\begin{bmatrix} \xi_1 \\ \eta_1 \end{bmatrix}, \ldots, \begin{bmatrix} \xi_n \\ \eta_n \end{bmatrix}$. Гіпотеза: $H_0 = \{\xi \perp\!\!\!\perp \eta\}$.

- Вибираємо інтервали для ξ та η: $\mathcal{E}_\xi = \Delta_1 \cup \cdots \cup \Delta_k,\ \mathcal{E}_\eta = \nabla_1 \cup \cdots \cup \nabla_l$
- Складаємо таблицю, де n_{ij} – кількість пар $\begin{bmatrix} \xi_p \\ \eta_p \end{bmatrix} \in \Delta_i \times \nabla_j$:

$\overrightarrow{\xi} \setminus \overrightarrow{\eta}$	∇_1	$\cdots$	∇_j	$\cdots$	∇_l
Δ_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Δ_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Δ_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}

- За припущенням: $\eta \perp\!\!\!\perp \xi \Longrightarrow \mathbb{P}\left\{ \begin{bmatrix} \xi \\ \eta \end{bmatrix} \in \Delta_i \times \nabla_j \right\} = \mathbb{P}\{\xi \in \Delta_i\} \cdot \mathbb{P}\{\eta \in \nabla_j\}$

За ЗВЧ: $\begin{cases} \mathbb{P}\{\xi \in \Delta_i\} \xleftarrow[n\to\infty]{\mathbb{P}} \dfrac{n_i^*}{n} = \dfrac{\sum\limits_{j=1}^{l} n_{ij}}{n}; \\ \mathbb{P}\{\eta \in \nabla_j\} \xleftarrow[n\to\infty]{\mathbb{P}} \dfrac{n_j^{**}}{n} = \dfrac{\sum\limits_{i=1}^{k} n_{ij}}{n}. \end{cases}$ $\qquad \mathbb{P}\left\{ \begin{bmatrix} \xi \\ \eta \end{bmatrix} \in \Delta_i \times \nabla_j \right\} \xleftarrow[n\to\infty]{\mathbb{P}} \dfrac{n_{ij}}{n}$

Таким чином, якщо виконується H_0, то має бути $\dfrac{n_{ij}}{n} \approx \dfrac{n_i^*}{n} \cdot \dfrac{n_j^{**}}{n}$.

Теорема 6.3. Якщо виконано $H_0 = \{\xi \perp\!\!\!\perp \eta\}$, то:

$$\chi^2(n) \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_{(k-1)(l-1)}(n) \qquad \chi^2(n) = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_i^* \cdot n_j^{**}}{n}\right)^2}{\frac{n_i^* \cdot n_j^{**}}{n}}$$

6.3. Ранговий критерій однорідності Манна-Уітні.

- Дві незалежні Г.С.: ξ та η.
- Відповідно, дві вибірки: $\overrightarrow{\xi} = (\xi_1, \ldots, \xi_m)$ та $\overrightarrow{\eta} = (\eta_1, \ldots, \eta_n)$.
- Гіпотеза: $H_0 = \{\mathcal{F}_\xi = \mathcal{F}_\eta\}$.
- Умова: обидві Г.С. мають неперервну функцію розподілу.

Спочатку побудуємо **спільний** варіаційний ряд для реалізацій:

$$\overrightarrow{\xi} \mapsto (x_1, \ldots, x_m) \qquad \overrightarrow{\eta} \mapsto (y_1, \ldots, y_n)$$

Функція статистичного критерію: $\mathcal{W}$ — сума рангів всіх елементів вибірки $\overrightarrow{\eta}$.

$$W = \frac{1}{2}n(n+1) + \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{I}\left\{\xi_i < \eta_j\right\}$$

Знайдемо характеристики функції критерію $\mathcal{W}$ та віднормуємо:

1) $\mathbb{E}\mathcal{W} = \dfrac{n(n+1)}{2} + \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}\mathbb{P}\left\{\xi_i < \eta_j\right\} = \dfrac{n(n+1)}{2} + mn \cdot \mathbb{P}\left\{\xi < \eta\right\} = \dfrac{n(n+1)}{2} + \dfrac{mn}{2}$

2) $\mathbb{D}\mathcal{W} = \dfrac{mn(m+n+1)}{12}$

3) За ЦГТ: $\boxed{\dfrac{\mathcal{W} - \frac{n(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \xrightarrow[m,n\to\infty]{\mathrm{d}} N(0,1)}$ – за умови H_0.

6.4. Перевірка параметричних гіпотез.

6.4.1. Перевірка гіпотези про математичне сподівання для гаусівської Г.С.

Нехай задано Г.С.: $\mathcal{N}(a, \sigma)$ та гіпотезу $H_0 = \{a = a_0\}$; $H_1 = \{a = a_1\}$.

Надалі проведемо перевірку гіпотези про a в гаусівській Г.С. при відомій дисперсії.

$$\text{За гіпотези } H_0: \ \sqrt{n} \cdot \frac{\overline{\xi} - a_0}{\sigma} \sim \mathcal{N}(0, 1)$$

Побудуємо критерій. Альтернативною гіпотезою можемо взяти:

Лівостороння	Двостороння	Правостороння
$\{a < a_0\}$	$\{a \neq a_0\}$	$\{a > a_0\}$

- Для *двосторонньої* альтернативної гіпотези: $T = \sqrt{n}\dfrac{\overline{x} - a}{\sigma} \Rightarrow t : \Phi(t) = \dfrac{1-\alpha}{2}$.
- Для *правосторонньої* альтернативної гіпотези: $T \Longrightarrow t : \Phi(t) = \frac{1}{2} - \alpha$.
- Для *лівосторонньої* альтернативної гіпотези: $T \Longrightarrow t : \Phi(|t|) = \frac{1}{2} - \alpha$.

6.4.2. Перевірка гіпотези про дисперсію для гаусівської Г.С.

При невідомому математичному сподіванні можемо перевірити гіпотезу для σ^2:

$$T = \frac{n \cdot \mathbb{D}^{**}}{\sigma_0^2} \sim \chi^2_{n-1}$$

Аналогічно будується критерій для таких альтернативних гіпотез можемо взяти:

Лівостороння	Двостороння	Правостороння
$\{\sigma^2 < \sigma_0^2\}$	$\{\sigma^2 \neq \sigma_0^2\}$	$\{\sigma^2 > \sigma_0^2\}$

6.4.3. Гіпотези про співвідношення параметрів 2-х гаусівських Г.С.

Маємо дві гаусівські Г.С.: $\xi \sim \mathcal{N}(a_1, \sigma_1^2)$; $\eta \sim \mathcal{N}(a_2, \sigma_2^2) \mapsto (\xi_1, \dots, \xi_{n_1})$; $(\eta_1, \dots, \eta_{n_2})$.

1) σ_1, σ_2 **– відомі;** $H_0 = \{a_1 = a_2\}$ $H_2 = \{a_1 \neq (>, <) a_2\}$.

   Розглядаємо $\overline{\xi} - \overline{\eta} \sim \mathcal{N}(a_1 - a_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$:

$$\frac{(\overline{\xi} - \overline{\eta}) - (a_1 - a_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1) \xRightarrow{H_0 \text{ виконується}} \frac{(\overline{\xi} - \overline{\eta})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

$$T = \frac{(\overline{x} - \overline{y})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1) \qquad H_0 = \{a_1 = a_2\} \qquad H_1 = \{a_1 < a_2\}$$

$$H_1 : a_1 - a_2 < 0 \quad \Longrightarrow \quad \alpha = \Phi(-t) - \Phi(-\infty) = \frac{1}{2} - \Phi(t) \rightsquigarrow t$$

2) Якщо дисперсії **невідомі**, але однакові. Гіпотеза: $H_0 = \{a_1 = a_2\}$. Маємо:

$$\frac{n_1 \cdot \mathbb{D}_\xi^{**}}{\sigma^2} \sim \chi^2_{n_1 - 1} \text{ та } \frac{n_2 \cdot \mathbb{D}_\eta^{**}}{\sigma^2} \sim \chi^2_{n_2 - 1} \Longrightarrow \frac{n_1 \cdot \mathbb{D}_\xi^{**} + n_2 \cdot \mathbb{D}_\eta^{**}}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 1} \ (*)$$

   Розглянемо вираз $\overline{\xi} - \overline{\eta} \sim \mathcal{N}\left(a_1 - a_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})\right)$. Віднормуємо, розділивши на дисперсію. Тоді при виконанні гіпотези H_0 маємо $(**)$: $\dfrac{\overline{\xi} - \overline{\eta}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$

   Знайдемо частку від ділення виразу $(**)$ на корінь з виразу $(*)$:

$$T = \frac{(\overline{\xi} - \overline{\eta}) \cdot \sqrt{n_1 + n_2 - 2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(n_1 \cdot \mathbb{D}_\xi^{**} + n_2 \cdot \mathbb{D}_\eta^{**}\right)}} \sim \frac{\mathcal{N}(0,1) \cdot \sqrt{n_1 + n_2 - 2}}{\sqrt{\chi^2_{n_1 + n_2 - 2}}} \sim \mathrm{St}_{n_1 + n_2 - 2}$$

3) Як перевірити рівність дисперсій? Перевіряємо гіпотезу: $H_0 = \{\sigma_1^2 = \sigma_2^2\}$.

   Розглядаючи попередньо отримані факти, приходимо до вигляду критерію:

$$T = \frac{\mathbb{D}_\xi^{***}}{\mathbb{D}_\eta^{***}} \sim \mathcal{F}_{n_1, n_2}$$

6.4.4. Гіпотези про параметри бернулівської Г.С.

Нехай проведено n випробувань з p – імовірністю успіху в одному. k – кількість успіхів, випадкова величина. Гіпотеза: $H_0 = \{p = p_0\}$ $n >> 1$.

Розглянемо оцінку ймовірності успіху:

$$p^* = \frac{k}{n} \sim \frac{\mathrm{Bin}(n,p)}{n} \overset{\text{цгт}}{\approx} \frac{\mathcal{N}(np, npq)}{n} = \mathcal{N}\left(p, \frac{pq}{n}\right)$$

$$T = |H_0 \text{ виконується}| = \frac{\sqrt{n} \cdot (p^* - p_0)}{\sqrt{p_0(1-p_0)}} \sim N(0,1)$$

Для двох бернулівських Г.С.: (n_1, p_1); (n_2, p_2); $H_0 = \{p_1 = p_2\}$. Кількість успіхів:

$$k_1 - k_2 \sim \overbrace{\mathrm{Bin}(n_1, p_1) - \mathrm{Bin}(n_2, p_2)}^{\perp\!\!\!\perp}$$

$$p_1^* - p_2^* = \frac{k_1}{n_1} - \frac{k_2}{n_2} \sim \frac{\mathrm{Bin}(n_1,p_1)}{n_1} - \frac{\mathrm{Bin}(n_2,p_2)}{n_1} \approx \mathcal{N}\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

$$H_0 = \{p_1 = p_2\} \quad \Rightarrow \quad \frac{k_1}{n_1} - \frac{k_2}{n_2} = p_1^* - p_2^* \approx \mathcal{N}\left(0, p_1(1-p_1) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

$$T = \frac{p_1^* - p_2^*}{\sqrt{\left(\frac{k_1+k_2}{n_1+n_2}\right)\left(1 - \frac{k_1+k_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

При великих $n_1, n_2 : T \approx \mathcal{N}(0,1)$.

— Happy End —

Без некоторых рисунков.