ЧНУ ім Петра Могили

12.5.2020

Лекція №8

Перевірка статистичних гіпотез

Воробйова А.І. доцент, канд.фіз-мат наук
КАФЕДРА ІНТЕЛЕКТУАЛЬНИХ ІНФОРМАЦІЙНИХ СИСТЕМ
СЕКЦІЯ ПРИКЛАДНОЇ ТА ВИЩОЇ МАТЕМАТИКИ

Критерій узгодженості χ^2 Пірсона (перевірка гіпотез про закон розподілу):

Зміст

ТЕМА ЛЕКЦІЇ.

Схема Бернуллі.

ПЛАН ЛЕКЦІЇ

I. Поняття статистичної гіпотези.
II. Типи гіпотез.
III. Задача статистичної перевірки гіпотези.
IV. К-критерій. Критерій значущості.
V. Помилки першого та другого роду.
VI. Приклад «Перевірка гіпотези про рівність центрів розподілу двох нормальних генеральних сукупностей при відомому σ».
VII. Критерій узгодженості Пірсона (перевірка гіпотез про закон розподілу). Алгоритм дій критерія Пірсона.
VIII. Приклад. «Залежність захворюваності на бронхіт від звички до паління»
IX. Стійкість критеріїв.
X. Практичні завдання для самостійного опрацювання.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Означення та типи статистичної гіпотези.
2. В чому полягає задача статистичної перевірки гіпотези.
3. Що таке рівень значущості та довірча ймовірність?
4. Що таке К-критерій?
5. Опишіть помилки першого та другого роду.
6. Сформулюйте алгоритм дій критерія Пірсона. Та наведіть приклад.

ЗАВДАННЯ ДЛЯ ДІСТАНЦІЙНОГО ОПАНУВАННЯ

Опрацювати наданий викладачів теоретичний матеріал.

Підготувати ***опорний конспект*** (основні означення, теореми, властивості, формули).

Надати відповіді на питання самоконтролю.

І.ПОНЯТТЯ СТАТИСТИЧНОЇ ГІПОТЕЗИ.

Статистична перевірка гіпотез це розв'язання наступної задачі:

> З'ясувати чи протирічать або узгоджуються з дослідними (експериментальними) даними наші апріорні припущення про характер тих або інших подій (явищ), функцій розподілів і так далі.

Нехай ξ – досліджувана дискретна або неперервна випадкова величина.

> *Статистичною гіпотезою Н* називається припущення відносно параметрів або виду розподілу випадкової величини ξ.

Наприклад:

1.Статистичною буде гіпотеза про те, що середні розміри деталей, які виробляються на однотипних паралельно працюючих верстатах, не розрізняються між собою.

2. гіпотеза про те, що розподіл продуктивності праці робітників, що виконують однакову роботу в однакових організаційно-технічних умовах, має нормальний закон розподілу.

Перевірка статистичних гіпотез тісно пов'язана з теорією оцінювання параметрів. У природознавстві, техніці, економіці часто для з'ясування того або іншого випадкового чинника удаються до вислову гіпотез, які можна перевірити статистично, тобто спираючись на результати спостережень у випадковій вибірці.

Типи гіпотез

> Статистична гіпотеза Н називається ***простою***, якщо вона однозначно визначає розподіл випадкової величини ξ, інакше гіпотеза Н називається ***складною***.

Наприклад:

1. Простою гіпотезою є припущення про те, що випадкова величина ξ розподілена за нормальним законом $\xi \in N(0;1)$.
2. Якщо ж висловлюється припущення, що випадкова величина ξ має нормальний розподіл $\xi \in N(\theta;1)$, де $a < \theta < b$, то це буде складна гіпотеза.

Задача статистичної перевірки гіпотези

Сформулюємо задачу статистичної перевірки гіпотези в загальному вигляді.

Нехай f(x,θ) – закон розподілу випадкової величини ξ, яка залежить від одного параметра θ. Припустимо, що необхідно перевірити гіпотезу $H_0 = \{\theta = \theta_0\}$ – *нульову* гіпотезу. Гіпотезу про те, що $\theta = \theta_1$, назвемо *конкуруючою* і позначимо її через $H_1 = \{\theta = \theta_1\}$. Відмітимо, що інколи гіпотезу H_1 називають *альтернативною гіпотезою* або *альтернативою*. Оскільки розподіл випадкової величини ξ відомий і по виборці потрібно зробити припущення про параметр розподілу, то такі гіпотези називаються *параметричними*.

Таким чином, перед нами стоїть задача перевірки гіпотези H_0 відносно конкуруючої гіпотези H_1 на підставі вибірки, що складається з n незалежних спостережень $x_1, x_2, ..., x_n$ над випадковою величиною ξ.

Отже, вся можлива множина вибірок об'єму n можна розділити на дві множини (позначимо їх Q і W), які не перетинається, таких, що гіпотеза H_0, яка перевіряється, має бути відхилена, якщо спостережувана вибірка потрапляє в підмножину W, і прийнята, якщо вибірка належить Q.

W - *називають **критичною областю***;

Q - ***областю допустимих значень***.

Нульова гіпотеза є прикладом статистичного висновку: якщо *нульову гіпотезу відкинути*, то висновок полягає в тому, що у сукупності, котра розглядається є розбіжності, тобто *приймається альтернативна гіпотеза* H_1.

Ймовірність з якою може бути відхилена нульова гіпотеза, коли вона є вірною, називається ***<u>рівнем значущості</u>*** (наприклад для медико-біологічних досліджень достатнім є рівень значущості $\alpha = 0{,}05$). Рівень значущості задається заздалегідь.

Ймовірність прийняття правильності рішення (гіпотеза H_0 є вірною) називається ***<u>довірчою ймовірністю</u>*** (для медико-біологічних досліджень $p = 0{,}95$).

Перевірка гіпотез як правило зводиться до перевірки статистичних характеристик, що оцінюють параметри законів розподілу.

<u>К-критерій</u>

Правило, згідно з яким приймається або відхиляється гіпотеза H_0, називається ***критерієм К***.

<u>Критерій значущості.</u>

Критерій, заснований на використанні заздалегідь заданого рівня значущості, називають ***критерієм значущості***.

Рівень значущості α визначає "розмір" критичної області W. Положення критичної області на множині статистики Z залежить від формулювання альтернативної гіпотези H_1. Наприклад, якщо гіпотеза $H_0 = \{\theta = \theta_0\}$, що перевіряється, а альтернативна – $H_1 : (\{\theta < \theta_0\}; \text{або } \{\theta > \theta_0\})$, то критична область розміщується на правому або лівому "хвості" розподілу статистики Z. В цьому випадку критерій називається ***однобічним***.

Якщо $H_1 = \{\theta \neq \theta_0\}$, то критична область розміщується на обох "хвостах" розподілу Z, а *K* називають ***двостороннім***.

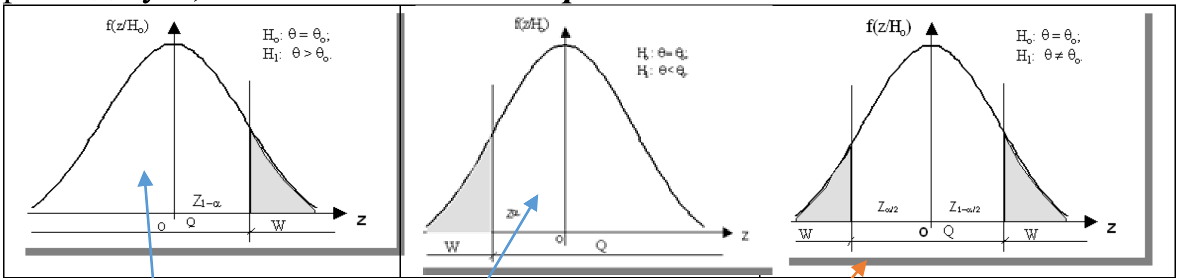


Рис. 1	Рис. 2	Рис. 3

Таким чином, перевірка параметричної статистичної гіпотези за допомогою *K*–значущості може бути розбита на наступні етапи:

1. Сформулювати гіпотезу, що перевіряється (H_0) і альтернативну (H_1);
2. Призначити рівень значущості α;
3. Вибрати статистику Z критерію для перевірки гіпотези H_0;
4. Визначити вибірковий розподіл статистики Z за умови, що вірна гіпотеза H_0;
5. Залежно від формулювання H_1 визначити критичну область W однією з нерівностей

$$Z > Z_{1-\alpha}, \; Z < Z_{\alpha}, \text{ або } \left\{Z > Z_{1-\frac{\alpha}{2}}\right\} \cup \left\{Z < Z_{\frac{\alpha}{2}}\right\};$$

6. Отримати вибірку спостережень і обчислити вибіркове значення статистики $Z_в$ критерію;

7. Прийняти статистичне рішення: якщо $Z_в \in W$, то відхилити гіпотезу H_0, як не узгоджувану з результатами спостережень. Якщо $Z_в \in Q = V \setminus W$, то прийняти гіпотезу H_0, тобто вважати, що H_0 не протирічить результатам спостережень.

Зауваження. При виборі критичної області W слід мати на увазі, що приймаючи або відхилюючи гіпотезу H_0, можна допустити помилки двох видів.

Помилки першого та другого роду

- ***Помилка першого роду*** полягає в тому, що H_0 – відхиляється, тобто приймається H_1, в той час, як насправді все ж справедлива гіпотеза H_0.
- ***Помилка другого роду*** полягає в тому, що гіпотеза H_0 приймається, в той час, як справедлива H_1. Тоді ймовірність помилки першого роду визначається рівнем значущості, оскільки

$$P\{Z \in W / H_0\} = \alpha,$$

а ймовірність помилки другого роду β можна обчислити (при простій альтернативній гіпотезі H_1)

$$P\{Z \in Q / H_1\} = \beta.$$

Отже, при перевірці гіпотез можливі помилки двох видів:

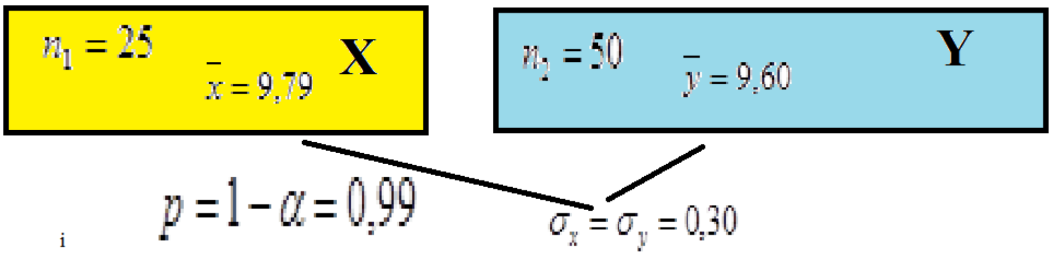
H_0 відкидається, коли вона правильна – помилка I-го роду.

H_0 приймається, коли правильна H_1 – помилка II-го роду.

Дії перевіряючого	Фактична ситуація	
	H_0 правильна	H_0 неправильна
Відкинути гіпотезу H_0	α помилка І-го роду.	1-α
Прийняти гіпотезу H_0	1-α	α помилка ІІ-го роду.

<u>Приклад «Перевірка гіпотези про рівність центрів розподілу двох нормальних генеральних сукупностей при відомому σ».</u>

В результаті двох серій вимірів з кількістю вимірів $n_1 = 25$ і $n_2 = 50$ отримані наступні середні значення досліджуваної величини: $\bar{x} = 9{,}79$; $\bar{y} = 9{,}60$ Чи можна з надійністю $p = 1 - \alpha = 0{,}99$ пояснити цю розбіжність випадковими причинами, якщо відомо, що середні квадратичні відхилення в обох серіях вимірів $\sigma_x = \sigma_y = 0{,}30$?



Розв'язання. Нехай ξ, η – незалежні випадкові величини, кожна з яких розподілена за нормальним законом $N(M(x); \sigma(x)) = N(a; \sigma)$.

Гіпотеза: $H_0 = \{M(\xi) = M(\eta)\}$,

Альтернативна: $H_1 = \{|M(\xi) - M(\eta)| > 0\}$,

$M(\xi)$, $M(\eta)$ – невідомі, а $\sigma_x = \sigma_y = 0{,}3$.

Тоді для перевірки гіпотези H_0 використовується їх найкращі оцінки $\bar{x}$ і $\bar{y}$. Відомо, що $\bar{x}$ і $\bar{y}$ мають нормальний закон розподілу з параметрами $N_1\left(M(\xi); \frac{\sigma(\xi)}{\sqrt{n_1}}\right)$ і $N_2\left(M(\eta); \frac{\sigma(\eta)}{\sqrt{n_2}}\right)$.

Вибірки – незалежні, тому $\bar{x}$ і $\bar{y}$ також незалежні, і випадкова величина, рівна різниці між $\bar{x}$, $\bar{y}$ має нормальний розподіл, причому

$$D\left(\bar{x}-\bar{y}\right)=D\left(\bar{x}\right)+D\left(\bar{y}\right)=\frac{\sigma^2(\xi)}{n_1}+\frac{\sigma^2(\eta)}{n_2}.$$

Якщо гіпотеза H_0 справедлива, то

$$M\left(\bar{x}-\bar{y}\right)=M\left(\bar{x}\right)-M\left(\bar{y}\right)=0,$$

отже, нормована різниця

$$|Z|=\frac{\left|\bar{x}-\bar{y}\right|}{\sqrt{\frac{\sigma^2(\xi)}{n_1}+\frac{\sigma^2(\eta)}{n_2}}}=\frac{|9{,}79-9{,}60|}{0{,}3\cdot\sqrt{\frac{1}{25}+\frac{1}{50}}}=2{,}59$$

підкоряється нормальному закону розподілу з математичним сподіванням, рівним нулю, і дисперсією, рівній одиниці.

По таблиці визначимо статистику Z_p, яка розділить множину Z на дві підмножини, що не перетинаються: область допустимих значень Z(Q) і критичну область Z(W). Оскільки критична область двостороння, то $\frac{1-0{,}99}{2}=0{,}005$.

Ті значення $|Z|\le Z_p=Z_{0{,}99}=2{,}58$, утворюють область допустимих значень і $W=\left\{|Z|>Z_p=Z_{0{,}99}=2{,}58\right\}$. Тепер маємо $2{,}59>2{,}58$, тому з надійністю $p=0{,}99$ можна вважати розбіжність середніх невипадковою (значимим), оскільки попадання в область значень $|Z|>2{,}58$, при нашій гіпотезі практично неможливе.

Зауваження. Проте слід зазначити, що для значень $|Z|\le 2{,}58$, ще не можна стверджувати, що гіпотеза підтвердилася: можна лише визнати допустимість гіпотези для розглянутих вибіркових спостережень до тих пір, поки ґрунтовніші дослідження не дозволять зробити протилежний висновок.

Отже, за допомогою перевірки статистичних гіпотез можна лише відхилити гіпотезу, що перевіряється, але ніколи не можна довести її справедливість.

II. КРИТЕРІЙ УЗГОДЖЕНОСТІ χ^2 ПІРСОНА (ПЕРЕВІРКА ГІПОТЕЗ ПРО ЗАКОН РОЗПОДІЛУ):

У багатьох практичних задачах закон розподілу досліджуваної випадкової величини невідомий, тобто є непараметричною гіпотезою, яка вимагає статистичної перевірки.

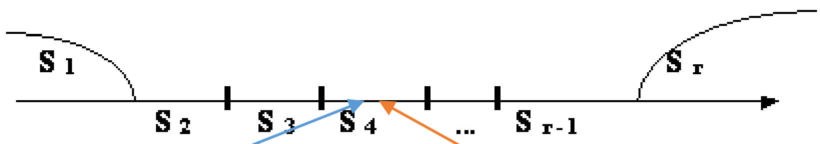
Нехай ξ – досліджувана випадкова величина. Потрібно перевірити гіпотезу H_0 – {випадкова величина ξ підкоряється закону розподілу $F(x)$}.

При побудові критерію для перевірки гіпотези H_0 використовуємо міру, введену Пірсоном, яка приводить до так званого критерію χ^2 Пірсона.

Цей критерій найчастіше вживається для перевірки гіпотези про закон розподілу. Відзначимо, що існує декілька критеріїв узгодженості: Колмогорова, Смірнова, ω^2 Мізеса та ін.

Для перевірки гіпотези H_0 зробимо вибірку, яка складається з n незалежних спостережень над випадковою величиною ξ. По вибірці побудуємо емпіричну функцію розподілу $F^*(x)$.
Порівняння емпіричного $F^*(x)$ і теоретичного розподілу відбувається за допомогою спеціально підібраної випадкової величини – критерію узгодженості.

Розіб'ємо множину значень ξ на r множин $(S_1, S_2 \ldots,)$ без спільних точок.



Обчислимо кількість елементів вибірки v_i, що попали в кожен з інтервалів S_i. Очевидно, $\sum_{i=1}^{r} v_i = n$, а $\sum_{i=1}^{r} \frac{v_i}{n} = 1$.

Через гіпотезу H_0, передбачаючи відомим закон розподілу F(x), визначимо

$$p_i = P\{\xi \in S_i\}, \quad \left(\sum_{i=1}^{r} p_i = 1\right)$$

теоретичне число значень випадкової величини ξ, що попали в інтервал S_i за формулою $n \cdot p_i$. Якщо емпіричні частоти сильно відрізняються від теоретичних, то гіпотезу H_0, що перевіряється, слід відхилити, інакше – прийняти.

Сформулюємо критерій, який би характеризував степінь розбіжності між емпіричними і теоретичними частотами.

Якщо гіпотеза H_0, що перевіряється, справедлива, то випадкова величина v_i, що характеризує кількість попадань в інтервал S_i, підкоряється біноміальному закону розподілу з математичним сподіванням $M(v_i) = n \cdot p_i$ і дисперсією $D(v_i) = n \cdot p_i \cdot q_i$. Тоді при $n \to \infty$ випадкова величина

$$y_i = \frac{v_i - n \cdot p_i}{\sqrt{n \cdot p_i \cdot q_i}},$$

розподілена нормально з $M(y_i) = 0$ і $D(y_i) = 1$.

Випадкові величини $y_1, y_2, ..., y_r$ зв'язані між собою лінійною залежності. При $n \to \infty$ статистика

$$\sum_{i=1}^{r} y_i^2 \cdot q_i = \sum_{i=1}^{r} \frac{(v_i - n \cdot p_i)^2}{n \cdot p_i}$$

має розподіл χ^2 з $k = r - 1$ степенями вільності.

Проте, якщо параметри розподілу $F(x)$ оцінюються по виборці, то при $n \to \infty$

$$\sum_{i=1}^{r} \frac{(v_i - n \cdot p_i)^2}{n \cdot p_i}$$

має χ^2 розподіл з $k = r - l - 1$ степенями вільності (l - число параметрів розподілу $F(x)$, які розраховані по виборці).

Отже, за міру розбіжності між v_i та $n \cdot p_i$ для $i = \overline{1, r}$ використовують критерій

$$\chi^2 = \sum_{i=1}^{r} \frac{(v_i - n \cdot p_i)^2}{n \cdot p_i}. \qquad (*)$$

Правило застосування критерію χ^2 зводиться до наступного.

1. Розрахувавши значення χ^2 і вибравши рівень значущості критерію α, за таблицею χ^2– розподілу визначається $\chi^2_{k\alpha}$.
2. Якщо $\chi^2 \geq \chi^2_{k\alpha}$, то гіпотеза H_0 відхиляється, якщо $\chi^2 \leq \chi^2_{k\alpha}$, то гіпотеза приймається.

Зауваження.

- При перевірці гіпотези про закон розподілу контролюється лише помилка першого роду.

Як наголошувалося раніше, статистика (*) має χ^2 розподіл лише при $n \to \infty$, тому необхідною умовою застосування критерію Пірсона є наявність в кожному S_i щонайменше 5-10 спостережень. Якщо v_i дуже малі (1-2), то має сенс об'єднати деякі S_i.

Алгоритм дій Критерій χ^2 (критерій Пірсона)

1. Скласти інтервальний статистичний ряд розподілу відносних частот

$[a_{i-1};a_i)$	$[a_0;a_1)$	$[a_1;a_2)$	$[a_2;a_3)$	$\cdots$	$[a_{n-2};a_{n-1})$	$[a_{n-1};a_n)$
$P^*_{n,i}$	$P^*_{n,1}$	$P^*_{n,2}$	$P^*_{n,3}$	$\cdots$	$P^*_{n,n-1}$	$P^*_{n,n}$

2. За складеною таблицею визначаємо число ступенів вільності $r = k - 1$, де k- кількість стовпчиків у таблиці.
3. Задати функцію f, близькою якої до функції f_n^* треба перевірить.
4. Підрахуємо ймовірності $p_i = P\left(x \in \left[a_{i-1}; a_i\right)\right) = \int\limits_{a_{i-1}}^{a_i} f(x)dx,\ i \in \overline{1,k}$.
5. Обчислимо величину

$$\chi^2_{еккс} = n\sum_{i=1}^{k}\frac{(P^*_{n,i} - p_i)^2}{p_i}$$

6. Задати рівень ***значущості*** α;
7. Визначити за спеціальною таблицею значень $\chi^2_{кр}$, що відповідає найденим параметрам α та r.
8. Якщо $\chi^2_{еккс} < \chi^2_{кр}$, гіпотеза про близькість функцій f_n^* та f не суперечить даним. В іншому випадку ця гіпотеза відкидається.

Приклад. «Залежність захворюваності на бронхіт від звички до паління»

Було проведено дослідження захворюваності на бронхіт робітників цеху залежно від звички до паління. Результати дослідження подано у таблиці:

Таблиця. Захворюваність на бронхіт робітників цеху

Рівні ознак	Наявність бронхіту (y_1)	Відсутність бронхіту (y_2)	y_1+y_2
Не палить (x_1)	5	20	25
Кинув палити (x_2)	10	40	50
Палить (x_3)	15	10	25
$x_1+x_2+x_3$	30	70	100

Формулюємо нульову гіпотезу H_0 – залежності немає.
Обираємо рівень значущості α=0,05.

За таблицею знаходимо відповідне значення критерію хі-квадрат: $\chi^2_{експ}$=5,991.

Обчислимо значення критерію $\chi^2_{кр}$: $\chi^2_{кр}$=15,950.

Оскільки $\chi^2_{експ} > \chi^2_{кр}$, нульову гіпотезу H_0 відхиляємо і приймаємо альтернативну гіпотезу H_1: залежність є.

IV.СТІЙКІСТЬ КРИТЕРІЇВ

Оскільки, існує велика кількість різних критеріїв (особливо непараметричних), це може викликати певні труднощі у спеціалістів, наведемо певні послідовність дій, притримуючись якої можна зробити правильний вибір.

Вибір методу для розв'язання задачі про порівняння параметрів розподілу вибірки

Формулювання задачі в прикладній постановці	Формулювання задачі в статистичній постановці	Додаткові умови		Метод, що застосовується
Порівняння показників контрольної та експериментальної вибірок	Перевірка гіпотези про рівність середніх (центрів розподілу) в двох незалежних вибірках	Норма-льний закон розподілу	Дисперсії вибірок рівні	t-критерій (Ст'юдента) при рівних дисперсіях
			Дисперсії вибірок не рівні	t-критерій (Ст'юдента) при нерівних дисперсіях
			Без припущення про дисперсії (але при однаковому розмірі вибірок)	t-критерій (Ст'юдента) без припущення про дисперсії
Порівняння показників вибірки до і після експерименту	Перевірка гіпотези про рівність середніх в двох залежних вибірках	**Нормальний** закон розподілу		t-критерій (Ст'юдента) для зв'язних вибірок
Чи можна вважати, що середнє значення показника дорівнює певному номінальному значенню?	Перевірка гіпотези про рівність середньої константі	**Нормальний** закон розподілу		t-критерій (Ст'юдента)
Порівняння розсіювання показників двох вибірок	Перевірка гіпотези про рівність дисперсій (про належність дисперсій до однієї генеральної сукупності)	**Нормальний** закон розподілу		F- критерій Фішера
Чи можна вважати, що в декількох вибірках має місце одне і теж значення показника?	Перевірка гіпотези про рівність середніх (про належність середніх до однієї генеральної сукупності)	**Нормальний** закон розподілу		Дисперсійний аналіз

ПРАКТИЧНІ ЗАВДАННЯ ДЛЯ САМОСТІЙНОГО ОПРАЦІВАННЯ.

Див. інд роботу.