Тема. Суттєвість різниці між вибірковими долями та її визначення за критерієм Стьюдента. Суттєвість різниці між співвідношеннями класів та її визначення за методом хі-квадрат.

t-критерій Стьюдента порівняння двох середніх значень нормальних генеральних сукупностей, дисперсії яких невідомі, але однакові (за даними малих незалежних вибірок). Нехай генеральні сукупності X і Y мають нормальні розподіли. Із цих генеральних сукупностей взято незалежні вибірки невеликих обсягів n_X і n_Y ($n_X \le 30$, $n_Y \le 30$) і для них знайдено вибіркові середні значення $\bar{x}$ і $\bar{y}$ і виправлені вибіркові дисперсії s_X^2 і s_Y^2.

Нехай, крім того, генеральні дисперсії цих сукупностей хоч і невідомі, але рівні між собою. Якщо немає підстав так вважати, то спочатку треба перевірити гіпотезу про рівність генеральних дисперсій, скориставшись F-критерієм Фішера-Снедекора.

Критерій Стьюдента дає можливість встановити, чи відмінності між двома незалежними вибірками зумовлені випадковими причинами чи ні.

Як правило, знайдені вибіркові середні значення $\bar{x}$ і $\bar{y}$ відрізняються одне від одного. Нехай

$$\bar{x} > \bar{y}$$

(якщо буде виконуватися протилежна нерівність, то можна поміняти позначення генеральних сукупностей X і Y).

Але це розходження може бути зумовлене або різними випадковими причинами, наприклад, випадковим відбором об'єктів, або тим, що самі генеральні середні значення $M(X)$ і $M(Y)$ різні. Треба встановити, значуще (суттєво) чи незначуще (несуттєво) відрізняються вибіркові середні значення $\bar{x}$ і $\bar{y}$.

Для цього треба при заданому рівні значущості α перевірити нульову гіпотезу про рівність генеральних середніх значень (математичних сподівань):

$$H_0 : M(X) = M(Y).$$

Альтернативною гіпотезою може бути або гіпотеза

$$H_1 : M(X) > M(Y) ,$$

або гіпотеза

$$H_1 : M(X) \ne M(Y) .$$

Емпіричне значення критерію Стьюдента обчислюють за формулою:

$$T_{\text{емп}} = \frac{\bar{x} - \bar{y}}{\sqrt{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}} \sqrt{\frac{n_X n_Y (n_X + n_Y - 2)}{n_X + n_Y}} .$$

Для першого випадку (конкуруюча гіпотеза має вигляд $H_1 : M(X) > M(Y)$) будують правобічну критичну область. Критичну точку при заданому рівні значущості α і кількості ступенів свободи $k = n_X + n_Y - 2$

$$t_{\text{кр одноб}}(\alpha; k) = t_{\text{кр одноб}}(\alpha; n_X + n_Y - 2)$$

знаходять за таблицею критичних точок розподілу Стьюдента (див. додаток 2, таблиця 5) для *однобічної* критичної області.

Для другого випадку (конкуруюча гіпотеза має вигляд $H_1 : M(X) \neq M(Y)$) будують двобічну критичну область. Критичну точку при заданому рівні значущості α і кількості ступенів свободи $k = n_X + n_Y - 2$

$$t_{\text{кр двоб}}(\alpha; k) = t_{\text{кр двоб}}(\alpha; n_X + n_Y - 2)$$

знаходять за таблицею критичних точок розподілу Стьюдента (для *двобічної* критичної області.

Суттєвість різниці між співвідношеннями класів та її визначення за методом хі-квадрат.

Часто необхідно знати закон розподілу ознаки у генеральній сукупності. Наприклад, є підстави вважати, що він має вигляд А. Тоді висувають гіпотезу (припущення): **генеральна сукупність розподілена за законом А.** У цій гіпотезі йде мова про вигляд невідомого закону розподілу. Іноді закон розподілу генеральної сукупності відомий, але його параметри (числові характеристики) невідомі. Тоді висувають гіпотезу: **невідомий параметр $\boldsymbol{\theta}$ дорівнює $\boldsymbol{\theta_o}$.** Ця гіпотеза вказує припущену величину параметра відомого розподілу. Можливі інші гіпотези: про рівність параметрів двох різних розподілів, про незалежність вибірок тощо.

Означення. **Статистичними** називають гіпотези про вигляд розподілу генеральної сукупності або про параметри відомих розподілів.

Наприклад, статистичними будуть гіпотези: а) генеральна сукупність розподілена за нормальним законом; б) дисперсії двох сукупностей, розподілених за законом Пуассона, рівні між собою.

Приклад нестатистичної гіпотези (оскільки не йде мова ні про вигляд закону розподілу, ні про його параметри): значна частина людей, народжених у другому півріччі, має краще розвинену праву частину мозку, яка здійснює образне мислення.

Разом із припущеною гіпотезою завжди можна розглядати протилежну їй гіпотезу, які доцільно розрізняти.

Означення. **Основною (нульовою)** називають припущену гіпотезу і позначають $\boldsymbol{H_o}$.

Означення. **Альтернативною (конкурентною)** називають гіпотезу, що суперечить основній і позначають $\boldsymbol{H_1}$.

Наприклад, якщо $\boldsymbol{H_o : M(X) = 6}$, то $\boldsymbol{H_1 : M(X) \neq 6}$.

Гіпотези можуть містити тільки одне припущення (**прості**) або більше одного припущення (**складні**). Наприклад, якщо $\boldsymbol{\lambda}$ - параметр показникового розподілу, то гіпотеза $\boldsymbol{H_o : \lambda = 6}$ - проста, а гіпотеза $\boldsymbol{H_o : \lambda > 6}$ - складна (містить нескінченну множину гіпотез).

Статистична гіпотеза, яка висунута, може бути правильною або неправильною, тому виникає необхідність її **статистичної перевірки** (перевірка за даними вибірки). При цьому за даними **випадкової вибірки** можна зробити хибний висновок.

Означення. Якщо за висновком буде відкинута правильна гіпотеза, то кажуть, що це похибка першого роду.

Означення. **Якщо за висновком буде прийнята хибна гіпотеза, то кажуть, що це похибка другого роду.**

Відмітимо, що наслідки похибок другого роду більш небезпечні, ніж наслідки похибок першого роду.

__Означення__. Імовірність здійснити похибку першого роду називають **рівнем значущості.**

Рівень значущості найчастіше позначають α і приймають рівним 0,01 або 0,05. Якщо $\alpha = 0{,}05$, то це значить, що в п'яти випадках із 100 ми ризикуємо дістати похибку першого роду (відкинути правильну гіпотезу).

Наприклад, при контролі якості продукції імовірність признати неякісними якісні вироби називають **ризиком виробника**, а імовірність признати якісними неякісні вироби називають **ризиком споживача.**

КРИТЕРІЇ УЗГОДЖЕННЯ ДЛЯ ПЕРЕВІРКИ ГІПОТЕЗ.

__Означення__. **Статистичним критерієм узгодження перевірки гіпотези (або просто критерієм)** називають випадкову величину **К** (вибіркову функцію), розподіл якої (точний або наближений) відомий і яка застосовується для перевірки основної гіпотези.

__Зауваження__. Якщо статистична характеристика (вибіркова функція) розподілена нормально, то критерій позначають не буквою **К**, а літерою **Z** (а процес перевірки **Z-тестуванням**). Якщо статистична характеристика розподілена за законом Фішера-Снедекора, то її позначають **F (F–тестування)**. У випадку розподілу статистичної характеристики за законом Стьюдента її позначають **t (t–тестування)**, а у випадку закону "хі-квадрат" - χ^2 (χ^2**-тестування**).

__Означення__. **Спостереженим значенням критерію узгодження** називають значення відповідного критерію, обчислене за даними вибірки.

__Означення__. **Критичною областю** називають множину можливих значень критерію, при яких основна гіпотеза відхиляється. Є однобічні та двобічні критичні області.

__Означення__. **Областю прийняття гіпотези (областю допустимих значень)** називають множину можливих значень критерію, при яких основна гіпотеза приймається.

Для знаходження критичних областей (та областей прийняття гіпотез) задають рівень значущості α, визначають кількості ступенів вільності (це поняття буде розглянуто далі), а потім шукають критичну точку $k_{кр}$ із умови $P(K > k_{kp}) = \alpha$ у випадку правобічної критичної області. Ця точка відокремлює критичну область від області прийняття гіпотези.

__Зауваження__. **Єдиним способом одночасного зменшення імовірностей похибок першого та другого роду є збільшення об'єму вибірки**.

КРИТЕРІЙ УЗГОДЖЕННЯ ПІРСОНА (χ^2 -КРИТЕРІЙ).

Критерій Пірсона ефективно використовують для перевірки гіпотези про розподіл генеральної сукупності (теоретичний розподіл). Критерієм перевірки основної гіпотези про вигляд теоретичного розподілу беруть випадкову величину χ^2, що визначається через порівняння емпіричних (вибіркових) та теоретичних частот. Ця ВВ не залежить від виду закону, а залежить тільки від рівня значущості α та кількості ступенів вільності k, яка визначається як різниця між зменшеною на одиницю кількістю варіант x_i (або інтервалів варіант) та кількістю параметрів розподілу. Тобто $k = n - 1 - l$, де n - кількість варіант (або інтервалів варіант), а l - кількість параметрів розподілу.

Критичне значення (критична точка) $\chi^2_{kp} = \chi^2(\alpha;k)$ знаходиться за відповідними таблицями (або за спеціальними функціями Excel).

<u>Правило Пірсона.</u> Щоб при заданому рівні значущості α перевірити основну гіпотезу H_o: генеральна сукупність розподілена за певним законом, потрібно:

1) припустити наявність певного закону розподілу, знайти його параметри та побудувати (записати) цей закон;

2) обчислити за цим законом теоретичні частоти m_i^T для кожної варіанти x_i (або інтервалу варіант);

3) обчислити спостережене значення критерію за формулою

$$\chi^2_{cn} = \sum_{i=1}^{k} \frac{\left(m_i - m_i^T\right)^2}{m_i^T};$$

4) знайти за таблицями критичну точку $\chi^2_{kp} = \chi^2(\alpha;k)$;

5) порівняти χ^2_{kp} та χ^2_{cn} і зробити висновок:

якщо $\chi^2_{cn} < \chi^2_{kp}$, то гіпотеза приймається,

якщо $\chi^2_{cn} > \chi^2_{kp}$, то гіпотеза відхиляється.

Для нашого приклада (див.аналіз вибіркових даних на попередній лекції) висунемо основну гіпотезу H_o: **платня усіх робітників фірми (генеральна сукупність) розподілена за нормальним законом з параметрами**:

математичне сподівання $a = M(X) = X_{\tilde{A}} \approx X_B = 325{,}8$ грн.

стандарт $\sigma = \sigma_{\tilde{A}} \approx \sigma_B = 22{,}67$ грн.

Щільність розподілу імовірностей (диференціальна функція):

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \approx \frac{1}{22{,}67\sqrt{2\pi}} e^{-\frac{(x-325{,}8)^2}{1027{,}72}}.$$

Функція розподілу імовірностей (інтегральна функція):

$F(x) = \frac{1}{2} + \Phi\left(\frac{x - 325{,}8}{22{,}67}\right)$, де $\Phi(t)$ - інтегральна функція Лапласа.

Усі подальші розрахунки див. на Додатку 5 (інтервали з теоретичними або емпіричними частотами, меншими 5, приєднуються до сусідніх). За таблицями (або за спеціальними функціями Excel) для рівня значущості $\alpha = 0{,}05$ при кількості ступенів вільності $k = 7 - 1 - 2 = 4$: $\chi^2_{êðèò} = \chi^2(\alpha = 0{,}05; k = 4) = 9{,}49$.

ВИСНОВОК: оскільки $\chi^2_{ðîçð} = 9{,}14 < \chi^2_{êðèò} = 9{,}49$, то гіпотеза H_o: платня усіх робітників фірми (генеральна сукупність) розподілена за нормальним законом, приймається.