МОДЕЛЮВАННЯ ВПЛИВУ НЕКОНТРОЛЬОВАНИХ ФАКТОРІВ НА СТАТИСТИЧНИЙ РОЗПОДІЛ ВИПАДКОВИХ ВЕЛИЧИН, ПІДПОРЯДКОВАНИХ НОРМАЛЬНОМУ ЗАКОНУ

Хіміч Віктор Михайлович,
студент магістратури
Шаров Сергій Олегович,
студент магістратури.
Чорна Альона Віталіївна,
асистент кафедри інформатики і кібернетики
Мелітопольський державний педагогічний
університет імені Б. Хмельницького, м Мелітополь

Анотація. Розглянуто роль впливу неконтрольованих факторів при проведенні статистичного аналізу вибіркових даних, які підпорядковуються нормальному закону. Розроблено програму для визначення ступеня впливу грубих помилок на z - критерій при перевірці нульової гіпотези про рівність математичних очікувань. Показано, що в разі спотворень більше 10% випадкових величин від загальної кількості обсягу вибірки на величину трьох середньоквадратичних відхилень ймовірність помилки підвищується до 100%.

Ключові слова: дисперсія, математичне очікування, нормальний розподіл, нульова гіпотеза, випадкова величина, рівень значущості.

Вступ. Статистичні методи знайшли широке поширення у всіх областях людської діяльності. Найчастіше застосовують статистику, що використовує гіпотезу про виконання відомої функції розподілу випадкової величини, наприклад, нормального розподілу або розподілу Стьюдента [1]. У багатьох випадках вибіркові дані можуть містити елементи, пов'язані з впливом неконтрольованих факторів, що призводить до відхилення від обраного закону розподілу, в результаті чого достовірність висновків аналізу знижується, тому проведення досліджень в такому напрямку має практичний інтерес.

Об'єкт дослідження - статистична обробка експериментальних даних.

Предмет дослідження - аналіз обробки статистичних даних при наявності недоліків, що приводять до відхилення від нормального закону розподілу випадкової величини.

Мета дослідження - визначення ступеня достовірності висновків, які виходять з використанням вибіркових даних, що підпорядковуються нормальному закону при впливі неконтрольованих чинників.

Постановка проблеми. В роботі [2] запропоновано метод обробки вибіркових даних з використанням емпіричної функції розподілу випадкової величини. Подібний метод дозволяє провести оцінку адекватності висновків про виконання статистичних гіпотез при впливі неконтрольованих чинників. Представлене дослідження присвячене вивченню цієї проблеми на прикладі нормального закону розподілу. У завдання дослідження входило:

1. Створення програми *Norm* для генерування вибірки з n випадкових величин, що підпорядковуються нормальному закону при наявності збурень під впливом неконтрольованих чинників. Під збуреннями розуміється результат відносного сильної зміни деяких значень випадкових величин в кількості m, меншого в порівнянні з обсягом вибірки n.

2. Дослідження коректності висновків при перевірці нульової гіпотези про рівність математичних очікувань при наявності збурень під впливом неконтрольованих чинників.

Розробка програми для генерування випадкових величин. Генерування нормально розподілених чисел можна здійснити за допомогою методу, який заснований на використанні центральної граничної теореми (ЦГТ). Суть цього методу полягає в додаванні випадкових чисел x_i, які підпорядковуються будь-якому законом розподілу. Нехай значення x_i є числом, яке відповідає рівномірному розподілу з відрізка [0,1], і нормалізацією їх з подальшим переведенням в потрібний діапазон нормального розподілу. Згідно ЦГТ вибірка обсягу n, складена з сум виду.

$$S_k = \sum_{i=1}^{j} x_i \, , \qquad (1)$$

де $k=1,2,\ldots n$, підпорядковується нормальному закону [3, с. 399].

Перетворимо випадкову величину S_k в Z_k по формулі

$$Z_k = \left(\sum_{i=1}^{i=j} S_k - \sum_{i=1}^{i=j} m_k\right) \Big/ \sqrt{\sum_{i=1}^{j} D_i}\ , \qquad (2)$$

де m_k - математичне очікування випадкової величини S_k, D_k – її дисперсія.

Якщо S_k рівномірно розподілена на відрізку [0,1], то її математичне сподівання дорівнює m_k=0,5 [3,с.282], а дисперсія $D_k=1/12$ [3,с.310]. В цьому випадку $\sqrt{\sum_{i=1}^{j} D_i} = j/12$ і формула (2) набуває вигляду:

$$Z_k = \sqrt{12/j}\left(\sum_{i=1}^{i=j} S_k - 0{,}5j\right)\ . \qquad (3)$$

Можна показати, що випадкова величина (3) має нормалізований вигляд з математичним очікуванням $M(Z_k)$=0 і дисперсією $D(Z_k)$=1.

Формула

$$y_k = Z_k\sigma_y + M_y \qquad (4)$$

перетворює (3) в випадкову величину y_k, яка підпорядковується нормальному закону розподілу з математичним очікуванням M_y і дисперсією $\sigma_y^{\,2}$.

Нижче представлений код функції *f(double u, double v)* з програми на мові С++, яка здійснює генерування випадкових величин за формулою (4) з нормальним розподілом для математичного очікування *u* і середнє відхилення *v*:

```
double  f(double u,double v)
{
  double x=0,sum=0,y=0;
  int m=1000;
        for(int i=0;i<m;i++)
        {
          int N=0;
          N=rand();//rand() генерирует числа от 0 до 32767
      x=N;
```

```
            sum=sum+x/32767;
            y=sqrt(12.00000001/m)*(sum-0.5*m);
            y=y*v+u;
          }
   return y;
}
```

Нехай масив *w[]* заповнений за допомогою функції *f(double u, double v)* випадковими числами, які підпорядковуються нормальному закону у відповідність з формулою (4). Створення масиву *v[],* де частина елементів в кількості $u<n$ змінені в порівнянні зі значеннями елементів масиву *w[]* на величину $\delta=a\sigma_y$, може бути реалізовано за допомогою наступної функції:

```
void f_mass(int u,double v[1000],double w[1000])
{
   double x;
   x=n; x=x/100*u; u=x;
    for(int i=0;i<n;i++)
          {
                    if(i<(u+1)) v[i]= w[i]+1*1;
                    if(i>u) v[i]= w[i];
          }
}
```

Для вирішення поставленого завдання складена програма *Norm*, яка дозволяє проводити статистичний аналіз, а саме:

а) Генерувати вибіркові числові дані x_i обсягу n, що підпорядковуються нормальному закону із заданим математичним очікуванням a_x і дисперсією σ_y^2.

б) Генерувати вибіркові числові дані y_i обсягу n, що підпорядковуються нормальному закону з тим же математичним очікуванням $a_y=a_x$ і дисперсією $a_y=a_x$ при наявності $m<n$ елементів, що відрізняються від x_i в середньому на величину δ.

в) Аналізувати достовірність висновків при перевірці виконання нульової гіпотези про рівність математичних очікувань a_x і a_y в залежності від різних збурень.

Аналіз обробки статистичних даних під дією неконтрольованих чинників. Розглянемо дві вибірки з випадкових величин X і Y об'ємом n з невідомими математичними очікуваннями m_x і m_y. Нехай одна з них X, представлена масивом *w[]*, сформована за допомогою функції *f* і підпорядковується нормальному закону розподілу, а друга *Y*, представлена масивом *v[]*, створена функцією *f_mass* і містить *m* спотворених елементів. Позначимо математичні очікування дисперсій для цих вибірок через σ_x^2 і σ_y^2. Нульову гіпотезу про рівність математичних очікувань a_x і a_y представимо у вигляді:

$$H_0: a_x = a_y. \tag{5}$$

Розглянемо алгоритм перевірки гіпотези (5). Нехай результати обробки експерименту представлені у вигляді двох вибірок, кожна з об'ємом *n*. На підставі обробки цих даних отримані середньо вибіркові значення $x_{ср}$, $y_{ср}$ із дисперсіями s_x^2, s_y^2. Оскільки спотворена вибірка *Y* містить *m* чисел із зсувом до більших значень на величину δ, конкуруючу гіпотезу сформулюємо у вигляді:

$$H_1: a_y > a_x. \tag{6}$$

Для перевірки справедливості гіпотези H_0 введемо критерій [4]

$$z = \frac{\sqrt{n}(y_{ср} - x_{ср})}{\sqrt{s_x^2 + s_x^2}}\ , \tag{7}$$

Імовірність події, при якому $a_y > a_x$, дорівнює $\Phi(z_{кр})$, де $\Phi(x) = \frac{2}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$ - функція Лапласа.

Значення $z_{кр}$ для рівня значущості α визначається з умови

$$\Phi(z_{кр}) = 1\text{-}2\alpha, \tag{8}$$

При $\alpha=0,05$ функція Лапласа $\Phi(z_{кр})$=0,9 і $z_{кр}$=1,64, при α=0,25 $\Phi(z_{кр})$=0,5 і $z_{кр}$=0,67 [3].

Якщо умова $z>z_{кр}$ виконується, гіпотеза (5) приймається, в іншому випадку вона відкидається на користь альтернативної гіпотези (6).

Результати розрахунків критерію *z* з використанням програми *Norm* наведені в табл.1. Обчислення проводилися для математичних очікувань m=$m=m_x=m_y$=10 і дисперсій $\sigma^2=\sigma^2_x=\sigma^2_y=1$. Спотворення задавалися за формулою $x_i' = \delta + x_i$, де $\delta = 3\sigma$. Перший рядок у табл. 1 містить значення z, які знайдені для двох неспотворених вибірок. Як і слід було очікувати, у всіх випадках, починаючи з обсягу $n=10$, значення z виявляються менше критичної величини $z_{кр}=1,64$ для $\alpha=0,05$ і $z_{кр}=0,67$ для $\alpha=0,25$ для $\alpha=0,25$, що свідчить про виконання гіпотези (5). Підвищення числа спотворень до 40% і більше призводить до збільшення критерію z до значень вище критичної величини, що супроводжується отриманням помилкових висновків для рівня значущості $\alpha=0,05$ у всіх випадках за винятком обсягу вибірки $n=20$.

Останні три рядки табл. 1 дають інформацію про вплив факторів, що обурюють. При 10% -му спотворенні вибірки з нормальним розподілом гіпотеза (5) приймається тільки для $\alpha=0,05$ і $n<80$. При $\alpha=0,25$ вона відкидається на користь альтернативної гіпотези (6) у всіх випадках, що призводять до помилкових результатів.

Таблиця 1.

Залежність критерію z від обсягу вибірки n і кількості спотворених елементів k в процентах при $\delta = 3\sigma$ (жирним шрифтом відзначена область отримання помилкових висновків).

	n=10	20	40	80
k=0	0,63	0,20	0,22	0,18
10	1,08	1,29	1,37	**1,80**
20	1,58	**1,70**	**2,25**	**3,04**
40	**2,64**	**3,03**	**4,01**	**8,51**

Статистичний аналіз при збуреннях $\delta=12\sigma$ і більш призводить до помилкових висновків у всіх випадках.

Результати порівняння вибірок з різними математичними очікуваннями на прикладі математичних очікувань $m_x=10$, $m_y=10,5$ з однаковими дисперсіями $\sigma^2=\sigma^2_x=\sigma^2_y=1$ залежить від обсягу вибірки n, кількості збурень m і тієї вибірки, яка піддається збуренню. Якщо неконтрольовані фактори впливають на випадкові величини X, то одержувані висновки виявляються помилковими.

Якщо спотворюється вибірка *Y*, то ймовірність отримання правильних висновків підвищується.

Висновки. Розроблено програми *Norm* для генерування вибірки з *n* випадкових величин, що підпорядковуються нормальному закону при наявності збурень під впливом неконтрольованих чинників. Проведено аналіз коректності висновків при перевірці нульової гіпотези про рівність математичних очікувань при наявності збурень під впливом неконтрольованих чинників. Показано, що в разі спотворень більше 10% випадкових величин від загальної кількості обсягу вибірки на величину трьох середньоквадратичних відхилень ймовірність помилки підвищується до 100%.

Література

1. Коваленко И. Н. Теория вероятностей и математическая статистика/ И. Н. Коваленко, А. А. Филиппова – М.: Наука, 1986. – 575 с.

2. Еремеев В.С. Статистическая обработка эксперимента в случае неизвестной функции распределения / *В.С.* Еремеев, В.В. Кузьминов // Інформаційні технології в освіті. Журнал Херсонского государственного педагогического университета. – Вип. 13. – 2013. – С.44-51.

3. Жалдак М. І. Теорія ймовірностей і математична статистика: Підручник для студентів фізико-математичних спеціальностей педагог. університетів. Вид.2, перероб і доп./ М. І. Жалдак, Н. М.Кузьмина, Г. О. Михалін. – Полтава: «Довкілля-К», 2010. – 500 с.

4. Єремєєв В.С. Теорія ймовірностей та математична статистика / В.С. Єремєєв, Д. О. Сосновських, О. В. Тітова. - Мелітополь: ТОВ «Вид. будинок», 2009. – 187 с.