Course: Computer statistics

Lecture 4: Statistical hypotheses and criteria.

Lecturer: Oleksandr Dykhovychnyi

Лекція 4. Статистичні гіпотези і критерії.

I. Загальні поняття

У багатьох прикладних статистичних задачах виникає необхідність на підставі аналізу певних даних перевірити деяке припущення. Це припущення називають **гіпотезою**. Маючи вибірку, ми можемо висунути кілька взаємовиключних гіпотез про теоретичний розподіл і обрати одну. Завдання вибору однієї з кількох гіпотез вирішується побудовою певного правила - **статистичного критерію**. Хоча, зазвичай, за вибіркою скінченного обсягу безпомилкових висновків про розподіл зробити неможливо, тому завжди є небезпека вибрати невірну гіпотезу

Нехай $X = (x_1, x_2, ..., x_n)$ - вибірка з генеральної сукупності з розподілом F. **Гіпотезою**, позначається H, називається припущення про розподіл вибірки спостережень:

$$H = \{F = F_1\} \text{ або } H = \{F \in \Phi\},$$

де Φ - деяка підмножина множини всіх можливих розподілів. Гіпотеза H називається **простою**, якщо вона вказує на єдиний розподіл: $F = F_1$. Інакше H називається **складною**: $F \in \Phi$. Якщо є дві гіпотези, то одну з них прийнято називати **основною** - H_0, а іншу H_1 - **альтернативою.**

Нехай дана вибірка $X = (x_1, x_2, ..., x_n)$, щодо розподілу якої висунуті дві гіпотези: основна - H_0 та альтернативна - H_1. **Критерієм** називається

функція від вибірки $\pi(X)=\pi(x_1,x_2,...,x_n)$, яка приймає значення 0, у разі прийняття гіпотези H_0, та 1 - у разі прийняття гіпотези H_1. Частіше за все $\pi(X)=I\{s(X)<C\}, I(.)$ - індикатор, $s(X)$ - функція (**статистика критерію**), С - деяка стала (**поріг критерію**).

При застосуванні статистичних критерії можуть зустрічатись помилки. Помилка **першого роду** полягає у відхиленні основної гіпотези, коли вона є вірною. Імовірність такої помилки

$$\alpha = P\{\pi(X)=1/H_0\}.$$

Величину α називають **рівнем значущості** критерія.

Помилка **другого роду** полягає у відхиленні альтернативної гіпотези, коли вона є вірною. Імовірність такої помилки

$$\beta = P\{\pi(X)=0/H_1\}.$$

Величину $1-\beta$ називають потужністю критерія.

Зазвичай на множині значень вибірки X виділяють **критичну область** K наступним чином:

$$\pi(X)=\begin{cases}0, & X\in R^n\setminus K,\\ 1, & X\in K\end{cases}.$$

Зрозуміло, що є природнім намагатись зменшити обидві ймовірності помилок, але ця задача є суперечливою. Зменшення однієї з них призводить до збільшення іншої. Тому задача ставиться таким чином:

Зафіксувати рівень помилки першого роду α і вибирати той критерій, який забезпечує максимальну потужність критерію, тобто, з мінімальною ймовірністю помилки другого роду.

Є різні підходи до розв'язання цієї задачі. Розгляньмо один з можливих – підхід, який базується на **відношенні вірогідності.**

Критерій відношення вірогідності. Нехай вибірка $X=(x_1,x_2,...,x_n)$ складається з незалежних і однаково розподілених величин, про розподіл яких можливі тільки дві гіпотези: $H_0=\{F=F_0\}$ або $H_1=\{F=F_1\}$, відповідно до гіпотез розподіли мають щільності $p_0(u)$ та $p_1(u)$. Відповідні **функції вірогідності** дорівнюють:

$$L_0(X)=\Pi_{i=1}^n p_0(x_i)\ ,$$
$$L_1(X)=\Pi_{i=1}^n p_1(x_i).$$

Нехай критична множина K_c визначається наступним чином:

(*) $$K_c = \{X : L_1(X) \geq CL_0(X)\} \ ,$$

де стала C для будь-якого α забезпечує рівність $\alpha = P\{X \in K_c / H_0\}$. Тоді серед усіх критеріїв, які розрізняють гіпотези H_0 і H_1 із заданою ймовірністю помилки першого роду α, найбільш потужним є критерій з критичною областю K_c. Це твердження називають лемою **Неймана-Пірсона.**

Розгляньмо, як цей критерій працює для прийняття рішень про середнє значення нормальної генеральної сукупності.

Нехай $X = (x_1, x_2, ..., x_n)$ - вибірка з нормальної генеральної сукупності з невідомим середнім a та відомою дисперсією σ^2. Висуваємо основну гіпотезу $H_0 = \{a = a_0\}$ та альтернативну $H_1 = \{a = a_1\}$, $a_0 < a_1$. Статистикой критерия є вибіркове середнє

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} \ .$$

Якщо вибрати рівень значущості α і скласти відношення функцій вірогідності для двох нормальних сукупностей, то критична область для цього критерію виглядає так:

$$K_c = (C, \infty), \quad C = a_0 + u_{1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty \ ,$$

де $u_{1-\alpha}$ - квантіль нормального розподілу рівня $1-\alpha$.

Таким чином, гіпотеза $H_0 = \{a = a_0\}$ відхиляється, якщо $\bar{x} > C$, що є цілком логічним. При цьому потужність критерію буде наступною

$$1 - \beta = P\{\bar{x} > C / H_1\} = 1 - \Phi\left(\frac{c - a_1}{\sigma/\sqrt{n}}\right) \ ,$$

де $\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(u-a)^2}{2\sigma^2}} du, \ x \in (-\infty, \infty)$ - функція розподілу стандартного нормального розподілу.

Якщо задані рівні помилок першого й другого роду, то можна визначити мінімальний об’єм вибірки, для забезпечення цих рівнів з наступної нерівності :

$$n \geq \sigma^2 \left(\frac{u_{1-\alpha} + u_{1-\beta}}{a_1 - a_0} \right)^2 .$$

Продемонструємо все це на прикладі. Сформуємо вибірку з двох нормальних вибірок з середніми $a_0 = 1$, $a_1 = 2$ та дисперсією $\sigma^2 = 2$

Приклади:

```
#    Оптимальний критерій
     set.seed(3) #
>    alpha<-0.05 #  рівень помилки першого роду
>    sigma<-2.0 # дисперсія
>    a0<-1   # основна гіпотеза
>    a1<-2 # альтернатива
>   n0<-50 # обсяги вибірок
>   n1<-50
>   n<-n0+n1
>   x0<-rnorm(n0,a0,sigma)
>   x1<-rnorm(n1,a1,sigma)
>   x<-c(x0,x1)
>   hi<-hist(x,  probability=T,density=20, xlim=c(-3, 9),ylim=c(0,0.5),
+   breaks=40, xlab="Суміш")
>   curve(dnorm(z,a1,sigma), col="blue", add=T, xname="z")
>   curve(dnorm(z,a0,sigma), col="magenta", add=T, xname="z")
>   ualpha<-qnorm(1-alpha) # квантіль u1-alpha
>   ck<-a0+ualpha*sqrt(sigma)/sqrt(n) #критичний рівень
>   points(ck,0, col="darkred", pch=18, cex=2)
>   points(mean(x),0, col="green", pch=17, cex=2)
>   ck # порогове значення
[1] 1.232617
> mean(x) # середнє
[1] 1.522071
```

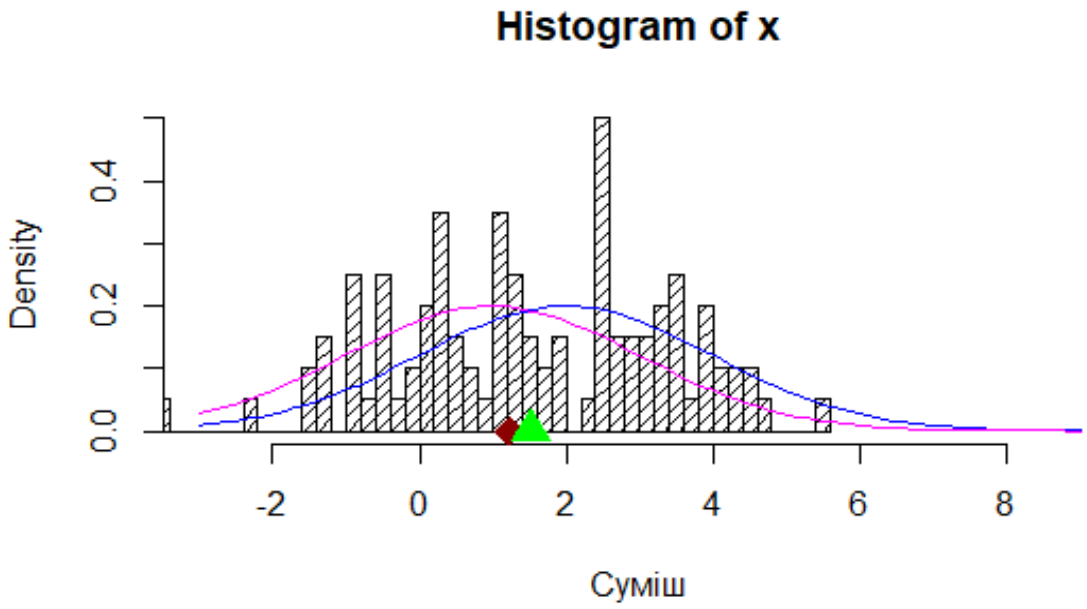

Рис. 1(Рисунок згенеровано **RStudio**)

На рисунку 1 зображені гістограма суміші обох розподілів. Червоним квадратиком відмічено C, а зеленим - трикутником вибіркове середнє $\bar{x}$. Як видно з рисунку 1 $\bar{x} > C$, отже, гіпотезу відхиллено. Обчислюємо потужність критерію, вона виявляється практично нульовою.

Приклади:

```
> #потужність критерію
> beta<-1-pnorm((ck-a1)/sqrt(sigma)*sqrt(n))
> pow<-1-beta
> pow
[1] 2.878094e-08
```

А також розраховуємо об'єм вибірки для $\beta = 0.05$

Приклади:

```
> # визначення обсягу вибірки
> betazad<-0.05
> ubetazad<-qnorm(1-betazad)
> nopt<-sigma*(ualpha+ubetazad)**2/(a0-a1)**2
> nopt
[1] 21.64435
```

Тобто, достатньо 22 значень.

II. Перевірка гіпотез про параметри нормальних генеральних сукупностей

Розглянемо перевірку ряду гіпотез, у яких порівнюють середні та дисперсії двох генеральних сукупностей.

Нехай є дві вибірки X та Y з двох нормальних генеральних сукупностей $N(a_X, {\sigma_X}^2)$ та $N(a_Y, {\sigma_Y}^2)$ об'ємами n_X, n_Y - відповідно.

1. Перед перевіркою гіпотез про рівність середніх перевіряються гіпотези про рівність дисперсій. Вважаємо середні значення вибірок невідомими. Статистика критерію:

$$S = \frac{S_X^2}{S_Y^2},$$

де $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ - вибіркова дисперсія за вибіркою X, аналогічно S_Y^2 -за вибіркою Y. Не порушуючи загальності міркувань, можна вважати, що $S_Y^2 > S_Y^2$.

Основна гіпотеза - $H_0 = \{\sigma_X^2 = \sigma_Y^2\}$, альтернативна - $H_1 = \{\sigma_X^2 > \sigma_Y^2\}$.

Тоді, при заданому рівні значущості(помилки першого роду) α, критерій працює наступним чином:

- гіпотеза H_0 приймається, якщо $S < F_{1-\frac{\alpha}{2}}(n_X - 1, n_Y - 1)$;
- гіпотеза H_0 відхиляється, якщо $S \geq F_{1-\frac{\alpha}{2}}(n_X - 1, n_Y - 1)$,

де $F_{1-\frac{\alpha}{2}}(n_X - 1, n_Y - 1)$-квантіль розподілу Фішера з $(n_X - 1, n_Y - 1)$ степенями свободи рівня $1 - \frac{\alpha}{2}$.

2. У залежності від того приймається, чи відхиляється гіпотеза про рівність дисперсій, вибирають критерій для перевірки гіпотези про рівність середніх.

Якщо гіпотезу $H_0 = \{\sigma_X^2 = \sigma_Y^2\}$ прийнято, то критерій для перевірки гіпотези про рівність середніх будується наступним чином.

Статистика критерію:

$$S = \frac{\left|\overline{X} - \overline{Y}\right|}{S_{XY}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}},$$

де $S_{XY} = \frac{(n_X - 1)S_x^2 + (n_y - 1)S_y^2}{n_Y + n_Y - 2}$.

Основна гіпотеза - $H_0 = \{a_X = a_Y\}$, альтернативна - $H_1 = \{a_X \neq a_Y\}$.

- Гіпотеза $H_0 = \{a_X = a_Y\}$ приймається, якщо $S < t_{1-\frac{\alpha}{2}}(n_X + n_Y - 2)$;
- Гіпотеза $H_1 = \{a_X \neq a_Y\}$ приймається, якщо $S \geq t_{1-\frac{\alpha}{2}}(n_X + n_Y - 2)$,

де $t_{1-\frac{\alpha}{2}}(n_X + n_Y - 2)$ - квантіль розподілу Стьюдента з $n_X + n_Y - 2$ степенями свободи рівня $1 - \frac{\alpha}{2}$.

3. Якщо гіпотеза $H_0 = \{\sigma_X^2 = \sigma_Y^2\}$ відхилена, то критерій для перевірки гіпотези про рівність середніх будується наступним чином.

Статистика критерію:

$$S = \frac{\left|\overline{X} - \overline{Y}\right|}{S_{XY}\sqrt{\frac{S^2{}_X}{n_X} + \frac{S^2{}_Y}{n_Y}}}.$$

Основна гіпотеза - $H_0 = \{a_X = a_Y\}$, альтернативна - $H_1 = \{a_X \neq a_Y\}$.

- Гіпотеза $H_0 = \{a_X = a_Y\}$ приймається, якщо

$$S < t_{1-\frac{\alpha}{2}}\left(\frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}\right);$$

- Гіпотеза $H_1 = \{a_X \neq a_Y\}$ приймається, якщо

$$S \geq t_{1-\frac{\alpha}{2}} \left(\frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y} \right)}{\frac{\left(\frac{S_X^2}{n_X} \right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y} \right)^2}{n_Y - 1}} \right).$$

Розгляньмо приклад перевірки гіпотези про рівність дисперсій двох вибірок. Згенеруємо дві вибірки з нормальним розподілом, нульовим середнім $a = 0$ і з дисперсіями $\sigma_1^2 = 1$, $\sigma_2^2 = 1.2$, відповідно, та об'ємами 50.

Приклади:

```
> #перевірка рівності дисперсій  за критерієм Фішером
> set.seed(3)
> a<-0
> n1<-50
> sigma12<-1
> sigma1<-sqrt(sigma12)
> x1<-rnorm(n1,a,sigma1) # генеруємо першу вибірку
> n2<-50
> sigma22<-1.2
> sigma2<-sqrt(sigma22)
> x2<-rnorm(n2,a,sigma2) # генеруємо другу вибірку
> var1<-var(x1) # вибіркові дисперсії
> var2<-var(x2)
> zB<-var2/var1 #статистика критерія
> zk<-qf(0.95,n1-1,n2-1) #порогове значення
> var1
[1] 0.7917457
> var2
[1] 0.812975
> zk
[1] 1.607289
> zB
[1] 1.026813
```

Як бачимо zB< zk. Тому приймається гіпотеза $H_0 = \{\sigma_1^2 = \sigma_2^2\}$.

Графічну ілюстрацію отримуємо так:

```
> plot(c(-4,4),c(0,0.5),type="n",xlab="x",ylab="density")
> lines(density(x1), col= "red", lwd = 2)
> lines(density(x2), col= "green", lwd = 2)
```

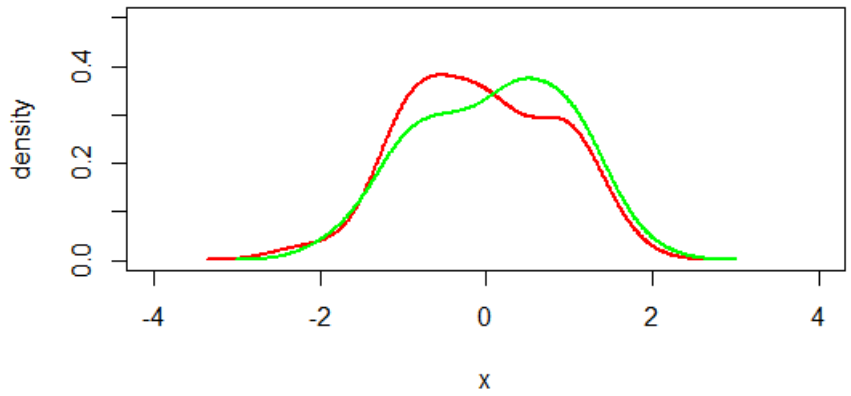

Рис. 2 (Рисунок згенеровано **RStudio**)

Разом з цим можемо перевірити туж саму гіпотезу за допомогою функції **var.test.**

Приклади:

```
> var.test(x2,x1,alternative="greater")

        F test to compare two variances

data:  x2 and x1
F = 1.0268, num df = 49, denom df = 49, p-value = 0.4633
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 0.6388477       Inf
sample estimates:
ratio of variances
          1.026813
```

Як бачимо, результати збігаються. Звернімо увагу на опцію **p-value = 0.4633**. В неї функція заносить рівень значущості критерія, він є досить високим.

Водночас можемо перевірити гіпотезу про рівність середніх.

Приклади:

```
> t.test(x2,x1,var.equal = TRUE,alternative="greater")

        Two Sample t-test

data:  x2 and x1
t = 0.88263, df = 98, p-value = 0.1898
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1393637        Inf
sample estimates:
  mean of x   mean of y
 0.09420070 -0.06392194
```

III. Критерії χ^2

Критерії χ^2 -це група статистичних критеріїв, у яких статистика критерію, за умови правдивості нульової гіпотези, має розподіл χ^2. Застосування критеріїв χ^2 є дуже поширеним для перевірки різноманітних гіпотез.

Нехай $X=(x_1,x_2,...,x_n)$ - вибірка з генеральної сукупності з розподілом F. Перевіряється гіпотеза $H_0=\{F=F_1\}$ при альтернативній гіпотезі $H_1=\{F\neq F_1\}$.

Для перевірки гіпотези за критерієм χ^2 здійснюють групування даних. Вибирають певне число інтервалів, які ділять область значень F. Після чого будують функцію розходження $\pi(X)$, як різницю теоретичних ймовірностей влучання в інтервали та теоретичних частот.

Нехай $A_1,...,A_k$ - набір інтервалів групування, що не перетинаються і на які розбита вся область значень випадкової величини, яка визначається функцією F. Позначимо як ν_j, $j=\overline{1,k}$ число елементів вибірки, що потрапили в інтервал A_j, а через p_j, $j=\overline{1,k}$ - імовірність влучення у множину A_j випадкової величини, яка має розподіл F. Очевидно, що

$$\sum_{j=1}^{k} p_j = 1 .$$

Нехай

$$\pi(X)=\sum_{j=1}^{k}\frac{\left(\nu_j-np_j\right)^2}{np_j} .$$

Тоді (теорема Пірсона), якщо правдивою є гіпотеза H_0, то при фіксованому к статистика $\pi(X) \Rightarrow \chi^2_{k-1}, n \to \infty$, тобто, наближається за розподілом до випадкової величини розподіленої за χ^2_{k-1}-розподілом з $k-1$ степенями свободи.

Критерій χ^2 застосовують для перевірки узгодження з певним розподілом. Розглянемо дискретний розподіл. Для цього згенеруємо вибірку обсягу 240, розподілену за законом Пуассона з параметром λ = 2, а потім перевіримо за критерієм χ^2 гіпотезу про узгодженість даної вибірки з розподілом Пуассона з параметром λ = 2 з рівнем значущості α = 0.05.

Приклади:

```
> #Перевірка гіпотези про параметр розподілу Пуассона
> #за допомогою критерія хі-квадрат
> set.seed(20)> x<-rpois(240,2)
# генеруємо вибірку
> statrad<-table(x) # статистичний ряд
> statrad1<-as.data.frame(statrad) # формуємо таблицю
> statrad1$Freq # частоти потрапляння в інтервал

 [1] 38 54 66 33 31 11  4  2  1

> relfr<-statrad1$Freq/240 #відносні частоти потрапляння в інтервали
> nempir<-c(38, 54, 66, 33, 31, 11,  7 )
> pth<-c(dpois(0:5,2),1-ppois(5,2))# теоретичні частоти потрапляння в
                                   #інтервал
> h<-rbind(pth[1:7],relfr[1:7]) #  об'єднана матриця частот
> h
```

```
     [,1]      [,2]      [,3]     [,4]       [,5]       [,6]       [,7]
[1,] 0.1353353 0.2706706 0.2706706 0.180447 0.09022352 0.03608941 0.01656361
[2,] 0.1583333 0.2250000 0.2750000 0.137500 0.12916667 0.04583333 0.01666667
```

```
> barplot(h,col=c(2,3),beside=T,ylab="частоти", xlab="інтервали")
          #діаграма частот
```

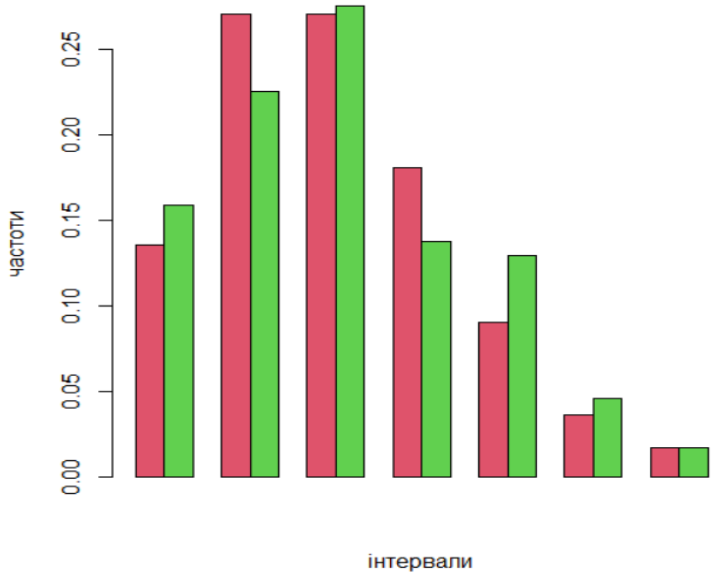

Рис. 3 (Рисунок згенеровано **RStudio**).
На рисунку 3. зображено теоретичні і емпіричні частоти.

> chisq.test(x=nempir,p=pth) # перевірка за хі квадратом

Chi-squared test for given probabilities

data: nempir
X-squared = 12.224, df = 6, p-value = 0.05715

Як видно з рисунка 1, частоти є достатньо схожими і гіпотеза приймається з рівнем значущості **p-value = 0.05715.**

IV. Критерій Колмогорова

Критерій Колмогорова застосовується як критерій перевірки згоди розподілу вибірки із заданим розподілом. Він побудований на підставі теореми Колмогорова.

Якщо $F_n^*(x) = \frac{1}{n}\sum_{x_i<x} I_{(-\infty,x)}(x_k), \quad x \in (-\infty,\infty)$ ***- емпірична функція розподілу, побудована за вибіркою*** $X = (x_1, x_2, ..., x_n)$ ***з функцією розподілу*** F ***, то***

$$P\left\{\sqrt{n}\sup_{x\in(-\infty,\infty)}\left|F_n^*(x) - F\ (x)\right| < t\right\} \to K(t), t \in (0,\infty), n \to \infty,$$

де $K(t) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2t^2}$ ***- функція Колмогорова.***

Нехай $X = (x_1, x_2, ..., x_n)$ - вибірка з генеральної сукупності з розподілом F . Перевіряється гіпотеза $H_0 = \{F = F_1\}$ при альтернативній гіпотезі $H_1 = \{F \neq F_1\}$. Статистика критерію $\pi(X) = \sqrt{n} \sup_{x \in (-\infty, \infty)} \left| F_n^*(x) - F\ (x) \right|$.

Тоді, якщо випадкова величина η має розподіл $K(x)$, а стала С є такою, що $P\{\eta > C\} = \alpha$, то критерій працює так:

- $\pi(X) < C$, то приймається гіпотеза H_0;
- $\pi(X) \geq C$, то приймається гіпотеза H_1.

Критерій Колмогорова реалізує функція **ks.test().**

Для ілюстрації згенеруємо вибірку обсягу 400, яка розподілена за нормальним розподілом з математичним сподіванням $a = 3$, і дисперсією $\sigma = 1$. За критерієм Колмогорова перевіримо гіпотезу про узгодженість розподілу даних з нормальним розподілом з тими ж самими параметрами для $a = 0.05$.

Приклади:

```
> n<-400
> x<-rnorm(n,3,1) # генеруємо вибірку
> qqnorm(x) #перевіряємо за Q-Q діаграмою
> qqline(x)
```

На рисунку 4 за допомогою **Q-Q** діаграм проілюстровано збіг згенерованого розподілу з нормальним.

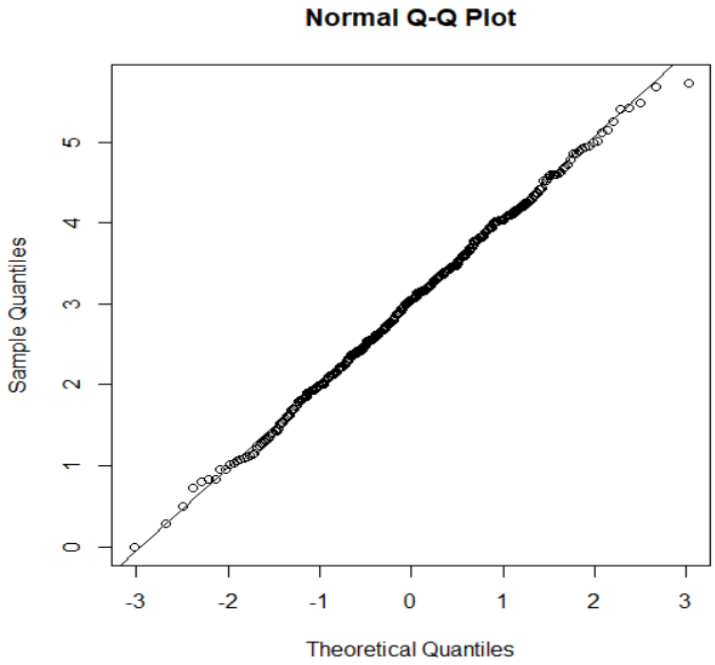

Рис. 4 (Рисунок згенеровано **RStudio**).

На рисунку 5 проілюстровано збіг емпіричної та теоретичної функцій розподілу.

Приклади:

```
> sx<-sort(x) # емпірична функція розподілу
> plot(sx,(1:n)/n,type="s",col="green")
> lines(sx,pnorm(sx,5,1),col="red") #теоретична функція розподілу
```

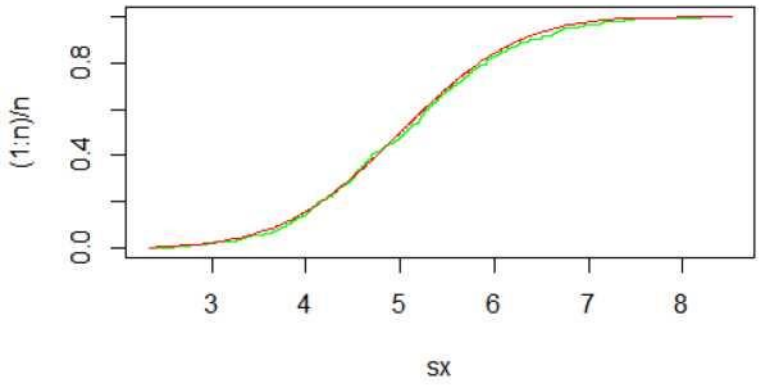

Рис. 5 (Рисунок згенеровано **RStudio**).

Далі перевіряємо збіг за тестом Колмогорова.

Приклади:

>ks.test(x,y="pnorm",mean=3,sd=1) # тест Коломогорова

data: x
D = 0.029193, p-value = 0.8849
alternative hypothesis: two-sided

Висновок: приймаємо основну гіпотезу - вибірка має нормальний розподіл, **p-value = 0.8849**
.