Лекція №6

6 Перевірка статистичних гіпотез щодо закону розподілення. Параметричний і непарметричний підходи

6.1 Критерій χ^2-Пірсона

6.2 Розподіли статистик непараметричних критеріїв згоди при простих гіпотезах

6.2.1 Критерій Колмогорова

6.2.2 Критерій Смірнова

6.2.3 Критерії ω^2

6.3. Втрата непараметричними критеріями згоди „свободи від розподілу" при складних гіпотезах

Найбільш зручною для практики була б така систематизація аналітичних моделей законів розподілів похибок, яка явно показувала б їх взаємну близькість чи віддаленість, а в ідеалі дозволяла б оцінити цю близькість чи віддаленість чисельно. [1]

Закон розподілу $f(x)$ як функція характеризується набором ознак $u_1, u_2, ..., u_k$. Вибір характеру ознак, їх сенсу, кількості та форми аналітичного представлення нічим не обмежений. Ознаки можуть бути однорідними і не однорідними, локальними (тими що відносяться до окремих точок кривих розподілів) і інтегральними (тими що виражаються через інтеграли, в підінтегральні вирази яких входить закон розподілу). Щоб ознаки характеризували тільки форму закону розподілу, вони повинні бути безрозмірними та не залежити від зміщення центру розподілу.

Після вибору сукупності ознак доцільно ввести у розгляд в загальному випадку багатомірний простір ознак $u_1, u_2, \ldots, u_k$ (по ортогональних осях відкладається значення ознак). В просторі ознак кожен конкретний закон відображається точкою з координатами $u_1, u_2, \ldots, u_k$. Якщо взяти два близьких один до одного закони розподілу, то їм будуть відповідати відображаючі точки. Бажано щоб був справедливим і зворотній перехід – близьким відображаючим точкам повинні відповідати близькі в певному розумінні форми законів розподілу.

У простоті опису перевагу мають такі способи завдання ознак, при яких кількість ознак мінімальна, а побудована систематизація достатньо повно відображає особливості форми розподілів. Обмежимо задачу розглядом лише симетричних розподілів, коли ліва половина кривої щільності є точним відображенням її правої половини.

За кількістю максимумів у кривої щільності які називаються модами, закони розподілу можна розділити на ***безмодальні*** (рівномірне, трапеціедальні), ***одномодальні***, ***двохмодальні*** і ***полімодальні***. Полімодальні закони розподілу, що мають більше двох мод, виключімо з розгляду. Будемо вважати, що якщо з експериментальних даних виходить трьохмодальний і більше розподіл, то це викликано лише випадковістю малої вибірки, розподіл генеральної сукупності випадкової величини є плавним та не має більше двох мод. *Таким чином, цю класифікацію обмежимо безмодальними, одномодальними та двохмодальними.*

Обговоримо тепер вибір ознак, що характеризують форму розподілів. При використанні другого та четвертого центральних моментів форма закону розподілу чисельно характеризується

значенням ексцесу ε. Але ексцес різних розподілів коливається у нескінченних межах $(1...\infty)$, що робить цей параметр незручним. Зробимо його нелінійне перетворення в значення контрексцесу $\kappa = \frac{1}{\sqrt{\varepsilon}}$, яке для будь-яких розподілів заключене у межах від 0 (при $\varepsilon = \infty$) до 1 ($\varepsilon = 1$). Таким чином в якості першої ознаки візьмемо значення контрексцесу κ.

Однак класифікація розподілів лише за одним контрексцесом є недостатньою. Зовсім різні закони розподілу можуть мати співпадаючі значення ексцесу та контрексцесу.

В якості другої незалежної ознаки форми закону розподілу ймовірності пропонується прийняти значення ентропійного коефіцієнта $k = \frac{\Delta e}{\sigma}$, який для будь-яких законів розподілу змінюється в межах $k \in \left[0; \sqrt{\frac{\pi e}{2}} \approx 2.066\right]$.

При використанні цих двох ознак відображаюча точка з координатами k та κ буде завжди знаходитися в межах прямокутника, обмеженого значеннями $k \in [0;2.066]$ та значеннями $\kappa \in [0;1]$. Така область на площині ознак відображена на рисунку 6.1. Розмістимо на ній всі основні розподіли.

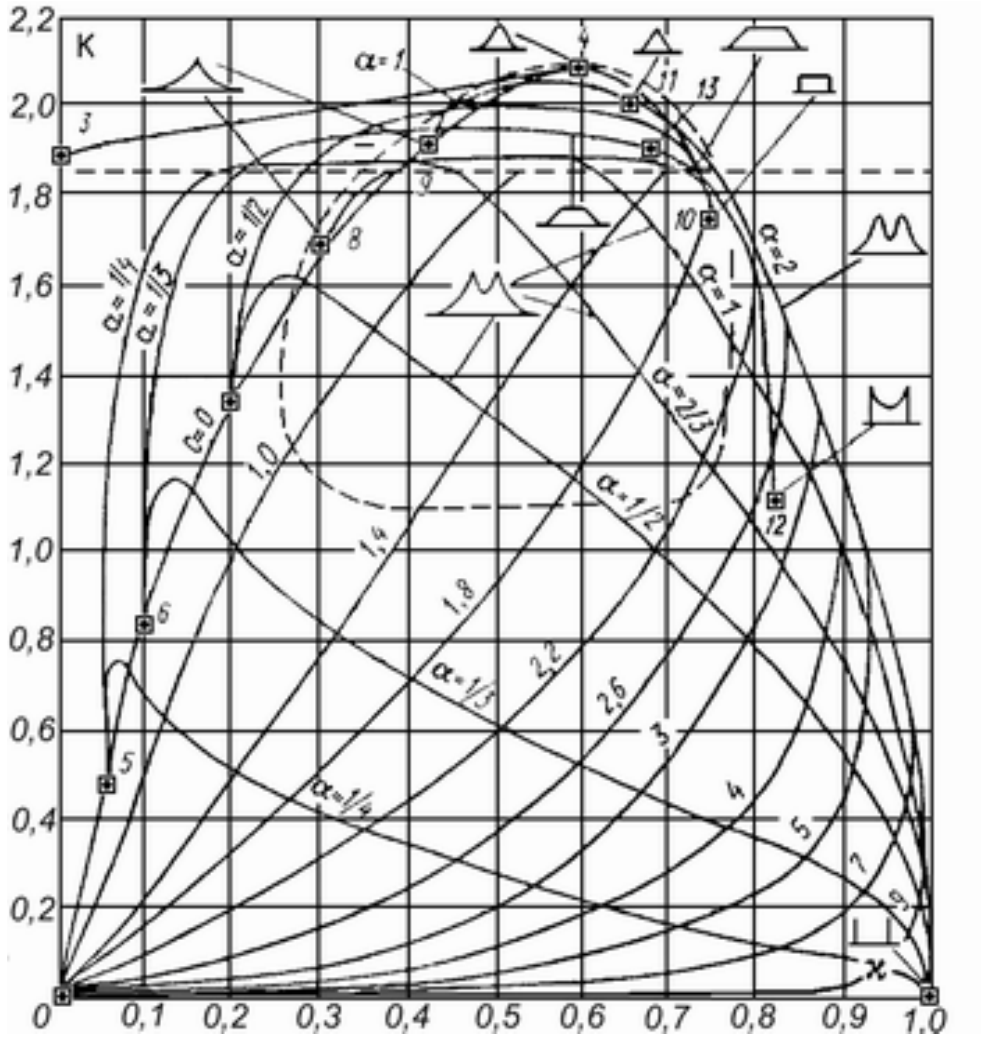

Рис. 6.1 – Схема топографічної класифікації законів розподілу випадкової величини (згідно з [1])

Чисельні оцінки форми розподілу у вигляді контрексцесу κ та ентропійного коефіцієнта k навіть при малому об'ємі вибірки експериментальних даних ($n \approx 40$) визначаються вже з достатньою точністю ($\gamma = 5 \div 10\%$). Цим можна скористатися для судження про вигляд форми кривої розподілу, що вивчається, за допомогою топографічної класифікації математичних моделей розподілів у координатах k та κ, яка показана на рисунку 6.1. Дійсно, якщо обчислити оцінки k та κ і нанести точку з цими координатами на площину $k - \kappa$, можна, окрім візуального враження від вигляду

полігона, отримати ще одну незалежну вказівку на можливу форму кривої розподілу.

Розглянемо більш детально рисунок 6.1.

Лінія, що з'єднує точки 2-3-4, є геометричним місцем точок, які відповідають сімейству розподілів Стьюдента з кількістю ступенів свободи $\nu \in [1;\infty]$.

Лінія, що з'єднує точки 2-5-6-7-8-9-4-10, є геометричним місцем точок, які відповідають класу експоненційних розподілів із показником степені $\alpha \in [0;\infty]$ Точка 2 відповідає розподілу з $\alpha \to 0$, точка 5 – з $\alpha = \frac{1}{4}$, точки 6-7-8 відповідають значенням α, які дорівнюють $\frac{1}{3}, \frac{1}{2}, \frac{2}{3}$, точка 9 відповідає розподілу Лапласа з $\alpha = 1$, точка 4 (з $\alpha = 2$) – нормальному розподілу Гауса та точка 10 (з $\alpha \to \infty$) – рівномірному розподілу.

Для опису сплощених високоентропійних розподілів (з $k > 1.87$) використані розподіли класу „шапо", тобто композиції рівномірного розподілу (точка 10) із різними експоненційними розподілами. Геометричні місця точок таких композиції знаходяться на рисунку 2.1 на лініях, що з'єднують точку 10 із точками відповідних експоненційних розподілів, починаючи від лінії 5-10 (поміченої значенням $\alpha = \frac{1}{4}$), лінії 6-10 ($\alpha = \frac{1}{3}$), ... і до лінії 4-10, що є геометричним місцем точок, які відповідають композиціям нормального (точка 4) та рівномірного (точка 10) розподілів.

Для опису низькоентропійних розподілів ($k > 1.87$) використані композиції експоненційних розподілів із дискретними двозначними розподілами виду $f(x) = 0.5 \cdot [\delta(x-a] + [\delta(x+a)]$ (точка 1 на рисунку 6.1). Геометричні місця точок таких розподілів

знаходяться на рисунку 2.1 на лініях, що з'єднують точки 5, 6, 7, 8, 9 та 4 із точкою 1 та помічені значеннями $\alpha = 2$, $\alpha = 1$, $\alpha = \frac{2}{3}$, , ... і до $\alpha = \frac{1}{4}$. Відносний вміст дискретної складової у таких розподілах зручно характеризувати відношенням $C = \frac{a}{\sigma}$, де a – напіврозмах дискретного розподілу, а σ – середнє квадратичне відхилення експоненційного. Лінії рівних C також нанесені на рисунку 6.1.

Окрім того, на рисунку 6.1 нанесені геометричні місця точок трапецеїдальних розподілів у вигляді лінії, що з'єднує точку 10, яка відповідає рівномірному розподілу, і точку 11, яка відповідає трикутному розподілу; а також композицій двох арксинусоїдальних розподілів – лінія, що з'єднує точку 12 (арксинусоїдальний розподіл) і точку 13 (композиція двох рівних арксинусоїдальних розподілів).

Для розподілів, точки яких входять до області лінії 4-10, є можливим одночасно використовувати цілого ряду рівноправних моделей у вигляді трапецій, у вигляді композицій класу „шапо" або у вигляді $f(x) = A \cdot e^{-(|x|)^{\alpha}}$, де $\alpha = 2 \div \infty$. Вибір одного з цих трьох видів аналітичного опису повинен робитися виходячи з практичної зручності подальшого використання отриманих результатів.

Характеристики, що дозволяють виконати топографічну класифікацію розподілу згідно з[1]:

Коефіцієнт форми α знаходиться з співвідношення:

$$Э = \frac{\Gamma\left(\frac{1}{\alpha}\right) \cdot \Gamma\left(\frac{5}{\alpha}\right)}{\Gamma^2\left(\frac{3}{\alpha}\right)}, \qquad (6.1)$$

де $Э = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4}$. (6.2)

Ентропійний коефіцієнт:

$$\tilde{K} = \frac{\tilde{\Delta}e}{\tilde{\tilde{\sigma}}}, \qquad (6.3)$$

де $\tilde{\tilde{\sigma}}$ - вибіркове виправлене СКВ;

$\tilde{\Delta}e = \frac{d \cdot n}{2} 10^{-\frac{1}{n}\sum_{i=1}^{m} n_j \lg n_j}$ - вибіркове ентропіне значння похибки; d – ширина интервалу вибіркового розподілу; m - число интервалів; n – объєм вибірки; n_j - частоти;

Оцінка контрексцесу (вибірковий контрексцес):

$$\tilde{\kappa} = \sqrt{\frac{\tilde{\tilde{\sigma}}^4}{\tilde{\mu}_4}}, \qquad (6.4)$$

де $\tilde{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$ - вибірковий центральний момент 4-го порядку; $\bar{x}$ - середнє арифметичне.

При перевірці узгодження емпіричного і теоретичного розподілу розрізняють *просту* та *складну* гіпотези у такій формі.

Означення 35. Гіпотеза H_0: $f(x) = f(x, \theta_0)$, де $f(.)$ - функція щільності, а θ_0 - відомий скалярний або векторний параметр теоретичного розподілу, з яким перевіряється згода, називається п*ростою*.

Означення 36. Гіпотеза H_0: $f(x) \in \{f(x, \theta), \theta \in \Theta\}$, де Θ - пространство параметрів і оцінка скалярного або векторного параметра $\tilde{\theta}$ обчислюється за тією ж самою вибіркою, за якою перевіряється гіпотеза щодо згоди, називається *складною*. Надалі будемо позначати складну гіпотезу наступним чином

$H_0 : f(x,\theta) = f(x,\tilde{\theta})$. Якщо $\tilde{\theta}$ визначається за іншою вибіркою, то гіпотеза проста.

Нижче будемо розглядати виключно складні гіпотези.

6.1. Критерій χ^2-Пірсона [2]

Перевірку за критерієм χ^2-Пірсона проводять для об'ємів вибірок $n > 100$, коли параметри математичної моделі визначені за емпіричними даними. Інтервали емпіричного та теоретичного розподілів, в яких теоретична частота $n_j' < 10$, об'єднують с сусідніми інтервалами. При цьому кількість отриманих інтервалів групування k' може бути менше або дорівнює початковій кількості інтервалів, тобто $k' \le k$. Спостережуване значення критерію обчислюється за формулою

$$\chi^2_{cn} = \sum_{j=1}^{k'} \frac{(n_j - n_j')^2}{n_j'}, ((j = 1,2,...,k'));$$

кількість ступенів вільності дорівнює

$$m = k' - r - 1,$$

де r - кількість параметрів математичної моделі.

Стосовно вибору рівня значимості доцільно притримуватися наступного правила: якщо $n < 100$, то $\alpha = 0.05$, а при $n > 100$, $\alpha = 0.01$.

Критичне значення $\chi^2_{кр}(\alpha; m)$ знаходять з таблиць критичних точок.

Якщо $\chi^2_{cn} < \chi^2_{\text{кр}}(\alpha; m)$ - H_0 стосовно передбачуваного закону відкидають.

6.2 Розподіли статистик непараметричних критеріїв згоди при простих гіпотезах [3]

6.2.1 Критерій Колмогорова

У разі простих гіпотез граничні розподіли статистик даних критеріїв згоди Колмогорова, Смірнова, ω^2 і Ω^2 Мізеса відомі *і не залежать від виду спостережуваного закону розподілу і, зокрема, від його параметрів.* Говорять, що ці критерії є “вільними від розподілу”. Це достоїнство зумовлює широке використання даних критеріїв на практиці.

Розподіл статистики

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x,\theta)|, \tag{6.5}$$

де $F_n(x)$ – емпірична функція розподілу, $F(x,\theta)$ – теоретична функція розподілу, n – об'єм вибірки, було одержано Колмогоровим в [2]. При $n \to \infty$ розподіл статистики $\sqrt{n} \cdot D_n$ сходиться рівномірно до розподілу Колмогорова

$$K(S) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2} \ . \tag{6.6}$$

Найчастіше в критерії Колмогорова (Колмогорова-Смірнова) використовується статистика виду [3]

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}}, \tag{6.7}$$

де

$$D_n = \max(D_n^+, D_n^-) \ , \tag{6.8}$$

$$D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F(x_i,\theta)\right\}, \tag{6.9}$$

$$D_n^- = \max_{1\le i\le n}\left\{F(x_i,\theta) - \frac{i-1}{n}\right\}, \tag{6.10}$$

n - об'єм вибірки, $x_1, x_2, \ldots, x_n$ - впорядковані за збільшенням вибіркові значення, $F(x,\theta)$ - функція закону розподілу, згода з яким перевіряється. Розподіл величини S_K при простій гіпотезі в межі підкоряється закону Колмогорова $K(S)$.

Якщо для обчисленого за вибіркою значення статистики S_K^* виконується нерівність

$$P\{S > S_K^*\} = 1 - K(S_K^*) > \alpha,$$

то немає підстав для відхилення гіпотези H_0.

6.2.2 Критерій Смірнова

У критерії Смірнова використовується статистика

$$D_n^+ = \sup_{|x|<\infty}\left(F_n(x) - F(x,\theta)\right) \tag{6.11}$$

або статистика

$$D_n^- = -\inf_{|x|<\infty}\left(F_n(x) - F(x,\theta)\right), \tag{6.12}$$

значення яких обчислюються за еквівалентними співвідношеннями (6.9),(6.10).

Реально в критерії звичайно використовується статистика

$$S_m = \frac{(6nD_n^+ + 1)^2}{9n}, \tag{6.13}$$

яка при простій гіпотезі в межі підкоряється розподілу χ^2 з числом ступенів свободи, рівним 2.

Гіпотеза H_0 не відкидається, якщо для обчисленого за вибіркою значення статистики S_m^*

$$P\{S_m > S_m^*\} = \int_{S_m^*}^{\infty} \frac{1}{2} e^{-x/2} dx = 1 - e^{-S_m^*/2} > \alpha .$$

6.2.3 Критерії ω^2

У критеріях типу ω^2 відстань між гіпотетичним і істинним розподілами розглядається в квадратичній метриці.

Гіпотеза H_0, що перевіряється, H_0 має вигляд

$$H_0: \int_{-\infty}^{\infty} \left\{ E\left[F_n(x)\right] - F(x) \right\}^2 \psi\left(F(x)\right) dF(x) = 0 \qquad (6.14)$$

при альтернативній гіпотезі

$$H_1: \int_{-\infty}^{\infty} \left\{ E\left[F_n(x)\right] - F(x) \right\}^2 \psi\left(F(x)\right) dF(x) > 0, \qquad (6.15)$$

де $E[\cdot]$ - оператор математичного очікування, $\psi(t)$ - задана на відрізку $0 \le t \le 1$ ненегативна функція, щодо якої передбачається, що $\psi(t)$, $t\,\psi(t)$, $t^2\psi(t)$ інтегруються на відрізку $0 \le t \le 1$. Статистика критерію виражається співвідношенням

$$\omega_n^2\left[\psi(F)\right] = \int_{-\infty}^{\infty} \left\{ E\left[F_n(x)\right] - F(x) \right\}^2 \psi\left(F(x)\right) dF(x) =$$

$$= \frac{2}{n}\sum_{i=1}^{n}\left\{g\left[F(x_i)\right] - \frac{2i-1}{2n} f\left[F(x_i)\right]\right\} + \int_0^1 (1-t)^2 \psi(t)dt\,, \qquad (6.16)$$

де

$$f(t) = \int_0^t \psi(s)ds\,, \quad g(t) = \int_0^t s\psi(s)ds\,.$$

При виборі $\psi(t) \equiv 1$ для критерію ω^2 Мізеса одержують статистику вигляду (статистику **Крамера-Мізеса-Смірнова**)

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(x_i,\theta) - \frac{2i-1}{2n}\right\}^2\,, \qquad (6.17)$$

яка при простій гіпотезі підкоряється розподілу, що має вигляд

$$a1(s) = \frac{1}{\sqrt{2s}}\sum_{j=0}^{\infty}\frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2}{16s}\right\}\times$$

$$\times\left\{I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16s}\right] - I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16s}\right]\right\}, \qquad (6.18)$$

де $I_{-\frac{1}{4}}(\cdot),\ I_{\frac{1}{4}}(\cdot)$ - модифіковані функції Бесселя,

$$I_\nu(z) = \sum_{k=0}^{\infty}\frac{\left(\frac{z}{2}\right)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)},\ |z| < \infty,\ |\arg z| < \pi\,. \qquad (6.19)$$

При виборі $\psi(t) \equiv 1/t(1-t)$ для критерію Ω^2 Мізеса статистика набуває вигляд (статистика **Андерсона-Дарлінга**)

$$S_\Omega = n\Omega_n^2 = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i,\theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1-F(x_i,\theta))\right\}$$

. (6.20)

У межі ця статистика підкоряється розподілу, що має вигляд

$$a2(s) = \frac{\sqrt{2\pi}}{s} \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)} \exp\left\{ -\frac{(4j+1)^2 \pi^2}{8s} \right\} \times$$

$$\times \int_0^{\infty} \exp\left\{ \frac{s}{8(y^2+1)} - \frac{(4j+1)^2 \pi^2 y^2}{8s} \right\} dy. \qquad (6.21)$$

Гіпотези про згоду не відкидаються, якщо виконуються нерівності

$$P\{S_\omega > S_\omega^*\} = 1 - a1(S_\omega^*) > \alpha \quad \text{і} \quad P\{S_\Omega > S_\Omega^*\} = 1 - a2(S_\Omega^*) > \alpha.$$

6.3. Втрата непараметричними критеріями згоди „свободи від розподілу” при складних гіпотезах

При перевірці складних гіпотез, коли по тій же самій вибірці оцінюються параметри спостережуваного закону розподілу ймовірності, непараметричні критерії згоди Колмогорова, Смірнова, ω^2 і Ω^2 Мізеса втрачають властивість “свободи від розподілу”. В цьому випадку граничні розподіли статистик цих критеріїв залежатимуть від закону, якому підкоряється спостережувана вибірка. Більш того, розподіли статистик непараметричних критеріїв згоди залежать до того ж і від використовуваного *методу оцінювання параметрів.* Слід також враховувати, що розподіли статистик істотно залежать від *об'єму вибірки*.

Ігнорування того, що перевіряється складна гіпотеза, неврахування фактів відмінності в складних гіпотезах приводить до некоректного застосування непараметричних критеріїв згоди на практиці і як наслідок невірним статистичним висновкам. Відмінності в граничних розподілах тих же самих статистик при

перевірці простих і складних гіпотез настільки істотні, що нехтувати цим абсолютно неприпустимо. Тому застереження проти неакуратного застосування критеріїв згоди при перевірці складних гіпотез неодноразово підіймалися на сторінках наукових видань.

У літературі наводиться низка підходів щодо використовування непараметричних критеріїв згоди в цьому випадку.

При достатньо великій вибірці її можна розбити на дві частини і по одній з них оцінювати параметри, а по іншій перевіряти згоду. У разі великих об'ємів вибірки такий підхід виправданий. Але якщо об'єм вибірки відносно невеликий, то спосіб розбиття її на дві частини відображатиметься і на оцінках параметрів, і на розподілах статистик критеріїв згоди. У цьому випадку застосовують підходи, детально наведені, наприклад, у [3].

Література

1. Ця систематизація, а на її основі й класифікація, могла бути побудована, виходячи з уявленнь добре розробленої на теперішній час теорії розпізнавання образів [Новицкий П. В., Зограф И. А. Оценка погрешностей результатов измерений. – 2-е изд., перераб. И доп. – Энергоатомиздат. Ленингр. Отделение, 1991. -304 с.].
2. Методика установления вида математической модели распределения погрешностей МИ 199-79
3. Лемешко Б. Ю., Постовалов С. Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим: Методические рекомендации. Часть II. Непараметрические критерии. – Новосибирск: Изд-во НГТУ, 1999. – С.85.