

**Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет информатики
и радиоэлектроники»**

**Кафедра интеллектуальных
информационных технологий**

М.Д. Степанова

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ
ДЛЯ ВЫПОЛНЕНИЯ ЛАБОРАТОРНОГО ПРАКТИКУМА**

**по дисциплине «Статистические основы индуктивного вывода,
теоретические основы машинного обучения и распознавания
в интеллектуальных системах»**

для студентов специальности I–40 03 01 “Искусственный интеллект”

**Минск
2006**

Методические указания к лабораторной работе № 1

Для выполнения лабораторной работы № 1 следует изучить методы статистической проверки гипотез.

В лабораторной работе № 1 необходимо

- 1) построить области отклонения и принятия статистической гипотезы,
- 2) проверить гипотезы о числовых значениях параметров нормального распределения.

1. Общая схема статистической проверки гипотез

Построение области принятия (отклонения) статистической гипотезы и проверка гипотезы производится в соответствии со следующей логической схемой.

1. Выдвижение гипотез H_0, H_1 .

2. Выбор **уровня значимости** α – вероятности ошибочного отклонения нулевой гипотезы. Эту величину называют также **размером критерия** (теста). Выбор величины уровня значимости α зависит от размера потерь, которые мы понесем в случае ошибочного решения. В большинстве практических задач пользуются стандартными значениями уровня значимости: $\alpha = 0,1; 0,05; 0,025; 0,01; 0,005; 0,001$.

3. Выбор **критической статистики** (критерия) $\gamma_k(x_1, x_2, \dots, x_n)$ – некоторой функции от результатов наблюдений. Эта критическая статистика γ_k сама является случайной величиной и в предположении справедливости нулевой гипотезы H_0 подчинена некоторому хорошо изученному закону распределения с плотностью $f_\gamma(u)$. Критическая статистика строится на основании **принципа отношения правдоподобия**.

4. Определение **критической области** W (множества значений критической статистики, при которых гипотеза отклоняется), исходя из следующего условия: $P(\gamma_k(x_1, x_2, \dots, x_n) \in W | H_0) = \alpha$. Из таблиц распределения $f_\gamma(u)$ находят квантили уровня $\alpha/2$ и уровня $1 - \alpha/2$, соответственно равные $\Delta(\alpha/2)$ и $\Delta(1 - \alpha/2)$. Они разделяют всю область возможных значений случайной величины γ_k на три части: 1 – **область неправдоподобно малых**

$(-\infty, \psi_{\alpha/2}]$, 2 – *правдоподобных* $(\psi_{\alpha/2}, \psi_{1-\alpha/2})$, 3 – *неправдоподобно больших* $[\psi_{1-\alpha/2}, \infty)$ значений *в условиях справедливости нулевой гипотезы H_0* (двусторонняя альтернативная гипотеза H_1).

В тех случаях, когда опасными для нашего утверждения являются только односторонние отклонения (т. е. только "слишком маленькие" или только "слишком большие" значения критической статистики γ_k), находят лишь одну квантиль. Для первого случая определяется квантиль уровня $\Delta(\alpha)$, которая будет разделять весь диапазон значений γ_k на две части: область неправдоподобно малых и область правдоподобных значений (левосторонняя альтернатива). Во втором случае вычисляется квантиль уровня $\Delta(1 - \alpha)$; она будет разделять весь диапазон значений γ_k на область правдоподобных и область неправдоподобно больших значений (правосторонняя альтернатива).

Значение границы критической области представляет собой квантиль распределения $f_\gamma(u)$ статистики критерия. Уровень квантили q устанавливается в зависимости от типа альтернативной гипотезы:

- для правосторонней альтернативы $q = 1 - \alpha$;
- для двусторонней альтернативы $q = 1 - \alpha/2$;
- для левосторонней альтернативы $q = \alpha$.

5. Определение на основе выборочных данных x_1, x_2, \dots, x_n численной величины статистики γ_k .

6. Выработка решения. Если значение статистики критерия принадлежит критической области ($\gamma_k \in W$), то гипотезу H_0 рекомендуется отклонить, в противном случае ее можно принять, так как имеющиеся данные не противоречат высказанной гипотезе.

7. Решение, принимаемое на основе статистического критерия, может оказаться ошибочным в двух случаях: когда ошибочно отклоняется гипотеза H_0 (с вероятностью α), и когда ошибочно принимается гипотеза H_0 (с вероятностью β), где вероятности ошибочных решений α и β – ошибки первого и второго рода, а величина $1 - \beta$ – мощность соответствующего критерия.

Пример 1. Найти значение границы критической области для проверки гипотезы о значении $\mu_0 = 20$ математического ожидания случайной переменной x при правосторонней альтернативе.

При решении задачи используем следующие предположения: 1) величина σ^2 известна и равна 16; 2) случайная переменная x имеет нормальное распределение; 3) объем выборки $n = 9$.

Решение. Воспользуемся приведенной выше схемой.

1. Выдвижение гипотез H_0, H_1 :

$$H_0: \mu = \mu_0 = 20, H_1: \mu > \mu_0.$$

2. Выбор уровня значимости: $\alpha = 0,05$.

3. Выбор критической статистики (критерия). В качестве статистики критерия выберем $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. При справедливости H_0 и нормальном распределении исходной случайной переменной эта статистика имеет стандартное нормальное распределение ($\mu = 0, \sigma^2 = 1$).

4. Определение границы критической области x_k . Значение границы критической области при правосторонней H_1 представляет собой квантиль уровня $p = 1 - \alpha$ распределения $f_\gamma(u)$ статистики критерия. Из таблиц стандартного нормального распределения для функции распределения $F(x; \mu, \sigma) = 1 - \alpha = 0,95$ находим значение квантили $u_{0,95}$ уровня $p = 0,95$, равное 1,64.

Следовательно, значение границы критической области x_k равно 1,64. Это означает, что при значениях статистики критерия $z \leq 1,64$ принимается гипотеза H_0 , а при значениях $z > 1,64$ – гипотеза H_1 .

Если мы хотим в качестве границы использовать среднее значение \bar{X}_k , то найдем это значение из соотношения

$$u_p = \frac{\bar{X}_k - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X}_k - 20}{4/\sqrt{9}} = 1,64.$$

Следовательно, значение границы критической области \bar{X}_k равно 22,2. Это означает, что при значениях среднего $\bar{X} \leq 22,2$ принимается гипотеза H_0 , а при значениях среднего $\bar{X} > 22,2$ – гипотеза H_1 .

2. Проверка гипотезы о значении математического ожидания нормального распределения

2.1. Случай 1: дисперсия σ^2 известна

Исходные данные. Выборка $X = (x_1, x_2, \dots, x_n)$ – независимые одинаково распределенные случайные величины.

Предположения. X имеют нормальное распределение $N(\mu, \sigma^2)$ с неизвестным математическим ожиданием μ и известной дисперсией σ^2 . Необходимо проверить гипотезу о числовом значении математического ожидания $H_0: \mu = \mu_0$.

Пример 2. На заводе выпускаются стандартные трубы с внутренним диаметром (ВД), равным 2,48 дюйма, причем дисперсия этой величины σ^2 составляет 0,0025 дюйм² ($\sigma = 0,05$). Известно, что внутренний диаметр является нормальной случайной величиной. Для выборки объема $n = 25$ произведены измерения ВД (табл.1.1). Можно ли утверждать, что выборка взята из нормальной совокупности с указанным значением математического ожидания?

Решение.

1. Выдвижение гипотез H_0, H_1 :

$H_0: \mu = \mu_0 = 2,48, H_1: \mu \neq 2,48$.

2. Выбор уровня значимости: $\alpha = 0,05$.

3. Выбор критической статистики. В качестве статистики критерия выберем $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$. При справедливости H_0 и нормальном распределении исходной случайной переменной эта статистика имеет стандартное нормальное распределение $N(0, 1)$ ($\mu = 0, \sigma^2 = 1$).

4. Определение границы критической области x_k . Определим значение квантили $u_{0,975}$ уровня $1 - \alpha/2 = 0,975$ стандартного нормального распределения: $u_{0,975} = 1,96$.

5. Определение на основе выборочных данных x_1, x_2, \dots, x_n численной величины статистики критерия.

Результаты $n = 25$ измерений внутреннего диаметра приведены в табл. 1.1.

Таблица 1.1

Значения внутреннего диаметра

X	X	X	X	X
2,408	2,468	2,511	2,561	2,649
2,430	2,474	2,529	2,565	2,654
2,438	2,478	2,541	2,576	2,659
2,458	2,494	2,543	2,589	2,674
2,467	2,501	2,545	2,610	2,687

Определим среднее значение \bar{X} по данным табл. 1.1: $\bar{X} = 2,50$. Тогда статистика критерия Z равна

$$Z = (2,50 - 2,48) / (0,05/5) = 2,00.$$

6. Выработка решения. Используем решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если } |Z| \leq \Delta(\varepsilon), \\ H_1, & \text{если } |Z| > \Delta(\varepsilon), \end{cases}$$

где $\Delta(\varepsilon)$ – квантиль уровня $\varepsilon = 1 - \alpha/2$ стандартного нормального распределения, определяющая границу критической области при двусторонней альтернативе.

В п.4 было найдено значение границы критической области $x_k = \Delta(1 - \alpha/2) = 1,96$.

При уровне значимости $\alpha = 0,05$ гипотеза о равенстве среднего значения $\mu_0 = 2,48$ отклоняется, так как $Z = 2,00 > 1,96$. Следовательно, принимается гипотеза H_1 . Это означает, математическое ожидание внутреннего диаметра трубы не равно 2,48 дюйма.

Применим другой способ проверки гипотезы, в котором используется решающее правило, основанное на P -значении (см. лекц. 8):

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если } P \geq \alpha, \\ H_1, & \text{если } P < \alpha. \end{cases}$$

Для статистики критерия, равной 2,00, вычисляется P -значение. С этой целью решаем относительно P уравнение $1 - P/2 = \Phi(2)$, где $\Phi(\cdot)$ – функция стандартного нормального распределения. Получаем $1 - P/2 = 0,977$, откуда $P = 0,046$. Так как $P = 0,046 < \alpha = 0,05$, поэтому при уровне значимости $\alpha = 0,05$ гипотеза H_0 отклоняется (принимается гипотеза H_1).

2.2. Случай 2: дисперсия σ^2 неизвестна

Исходные данные. Выборка $X = (x_1, x_2, \dots, x_n)$ – независимые одинаково распределенные случайные величины.

Предположения. X имеют нормальное распределение $N(\mu, \sigma^2)$ с неизвестным математическим ожиданием μ и неизвестной дисперсией σ^2 . Необходимо проверить гипотезу о числовом значении математического ожидания $H_0: \mu = \mu_0$.

Пример 3. На заводе выпускаются стандартные трубы с внутренним диаметром (ВД), равным 2,48 дюйма, причем дисперсия этой величины σ^2 неизвестна. Известно, что внутренний диаметр является нормальной случайной величиной. Для выборки объема $n = 25$ произведены измерения ВД (табл.1.1). Можно ли утверждать, что выборка взята из нормальной совокупности с указанным значением математического ожидания?

Решение.

1. Выдвижение гипотез H_0, H_1 :

$H_0: \mu = \mu_0 = 2,48, H_1: \mu > 2,48$.

2. Выбор уровня значимости: $\alpha = 0,05$.

3. Выбор критической статистики. В качестве статистики критерия выберем

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}},$$

где s – выборочное стандартное отклонение, $s = \sqrt{s^2}$; выборочная дисперсия вычисляется по формуле

$$s^2 = \sum_{j=1}^n (x_j - \bar{X})^2 / (n - 1).$$

При справедливости H_0 и нормальном распределении исходной случайной переменной эта статистика имеет t -распределение Стьюдента с $\nu_1 = n - 1$ степенями свободы.

4. Определение границы критической области x_k . Определим значение квантили $t_{0,95;24}$ уровня $1 - \alpha = 0,95$ t -распределения Стьюдента с $n-1=25 - 1 = 24$ степенями свободы: $t_{0,95;24} = 1,71$.

5. Определение на основе выборочных данных (табл. 1.1) численной величины статистики критерия. Для этого находим среднее значение \bar{X} внутреннего диаметра: $\bar{X} = 2,50$ и его выборочное стандартное отклонение $s = 0,05276$. Тогда статистика критерия t равна: $t = (2,50 - 2,48) / (0,05276 / \sqrt{24}) = 1,867$.

6. Выработка решения. Используем решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если } t \leq \Delta(\varepsilon), \\ H_1, & \text{если } t > \Delta(\varepsilon), \end{cases}$$

где $\Delta(\varepsilon)$ – квантиль уровня $\varepsilon = 1 - \alpha$ t -распределения Стьюдента с $n - 1$ степенями свободы, определяющая границу критической области при правосторонней альтернативе.

В п.4 было найдено значение границы критической области $x_k = \Delta(1 - \alpha) = 1,71$.

При уровне значимости $\alpha = 0,05$ гипотеза о равенстве среднего значения $\mu_0 = 2,48$ отклоняется, так как $t = 1,867 > 1,71$. Следовательно, при уровне значимости $\alpha = 0,05$ принимается гипотеза H_1 . Это означает, что математическое ожидание внутреннего диаметра трубы не равно 2,48 дюйма.

Применим другой способ проверки гипотезы, в котором используется решающее правило, основанное на P -значении:

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если } P \geq \alpha, \\ H_1, & \text{если } P < \alpha. \end{cases}$$

Находим значение функции t -распределения с $n - 1 = 24$ степенями свободы, соответствующее значению статистики $t = 1,867$: $F_t(1,867) = 0,963$. Получаем $1 - P = 0,963$, откуда $P = 0,037$. Так как $P = 0,037 < \alpha = 0,05$, поэтому при уровне значимости $\alpha = 0,05$ гипотеза H_0 отклоняется.

3. Проверка гипотезы о значении дисперсии

Исходные данные. Выборка $X = (x_1, x_2, \dots, x_n)$ – независимые одинаково распределенные случайные величины.

Предположения. X имеют нормальное распределение $N(\mu, \sigma^2)$ с неизвестными математическим ожиданием μ и дисперсией σ^2 .

Необходимо проверить гипотезу о числовом значении дисперсии

$$H_0: \sigma^2 = \sigma_0^2.$$

Пример 4. В результате длительного хронометража времени сборки узла различными сборщиками установлено, что дисперсия этого времени $\sigma_0^2 = 2$ мин². Результаты наблюдений за работой новичка представлены в табл. 1.2.

Таблица 1.2

Время сборки узла					
Время сборки одного узла в мин.	56	58	60	62	64
Частота	1	4	10	3	2

Можно ли при уровне значимости 0,05 считать, что новичок работает ритмично (в том смысле, что дисперсия затрачиваемого им времени существенно не отличается от дисперсии времени остальных сборщиков)?

Решение

1. Выдвижение гипотез H_0, H_1 :

$$H_0: \sigma^2 = \sigma_0^2, H_1: \sigma^2 \neq \sigma_0^2.$$

2. Выбор уровня значимости: $\alpha = 0,05$.

3. Выбор критической статистики. В качестве статистики критерия выберем

$$\chi^2 = (n - 1) s^2 / \sigma_0^2,$$

где s^2 – выборочная оценка дисперсии.

При справедливости H_0 и нормальном распределении исходной случайной переменной эта статистика имеет χ^2 –распределение с $\nu_2 = n - 1$ степенями свободы.

4. Определение границ критической области x_k . Определим значения квантили $\chi_{0,025;24}^2$ уровня $\alpha/2 = 0,025$ и квантили $\chi_{0,975;24}^2$ уровня $1 - \alpha/2 = 0,975$ χ^2 –распределения с $n - 1 = 25 - 1 = 24$ степенями свободы, являющиеся границами критической области при двусторонней альтернативе: $\chi_{0,025;24}^2 = 12,40$; $\chi_{0,975;24}^2 = 39,36$.

5. Определение на основе выборочных данных (табл. 1.2) численной величины статистики критерия. Для этого находим выборочную дисперсию $s^2 = 3,9894$. Тогда статистика критерия χ^2 равна: $\chi^2 = 24 \cdot 3,9894 / 2 = 47,8728$.

6. Выработка решения. Используем решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если } \Delta(\alpha/2) \leq \chi^2 \leq \Delta(1-\alpha/2), \\ H_1, & \text{если } \chi^2 < \Delta(\alpha/2) \text{ или } \chi^2 > \Delta(1-\alpha/2), \end{cases}$$

$\Delta(\varepsilon)$ – граница критической области, представляющая собой квантиль уровня $\varepsilon = 1 - \alpha/2$ χ^2 – распределения с $n - 1$ степенями свободы.

В п.4 были найдены значения границ критической области $\Delta(\alpha/2) = 12,40$ и $\Delta(1-\alpha/2) = 39,36$.

При уровне значимости $\alpha = 0,05$ гипотеза о равенстве значения дисперсии $\sigma_0^2 = 2$ отклоняется, так как $\chi^2 = 47,8728 > \chi_{0,975;24}^2 = 39,36$. Следовательно, при уровне значимости $\alpha = 0,05$ принимается гипотеза H_1 . Это означает, что дисперсия затрачиваемого новым сборщиком времени существенно отличается от дисперсии времени остальных сборщиков.

Применим другой способ проверки гипотезы, в котором используется решающее правило, основанное на P –значении:

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если } P \geq \alpha, \\ H_1, & \text{если } P < \alpha. \end{cases}$$

Для статистики критерия, равной 47,8728, вычисляется P –значение. С этой целью решаем относительно P уравнение $1 - P/2 = F(47,8728)$, где $F(\cdot)$ – функция χ^2 – распределения с $n - 1$ степенями свободы. Получаем $1 - P/2 = 0,997$, откуда $P = 0,006$. Так как $P = 0,006 < \alpha = 0,05$, поэтому при уровне значимости $\alpha = 0,05$ гипотеза H_0 отклоняется (принимается гипотеза H_1).