Методические указания к лабораторной работе № 2

Для выполнения лабораторной работы № 2 следует изучить методы статистической проверки гипотез.

В лабораторной работе № 2 необходимо

1) проверить гипотезу согласия с помощью критериев Пирсона и Колмогорова,

2) проверить гипотезы об однородности параметров (математических ожиданий и дисперсий) двух нормальных распределений.

1. Проверка соответствия выбранной модели распределения исходным данным (критерии согласия)

Критерии согласия предназначены для проверки гипотезы

$$H_0 : F_\xi(x) = F_0(x;\theta^{(1)},...,\theta^{(s)}) \qquad (2.1)$$

Этот тип критериев основан на использовании различных мер расстояний между анализируемой эмпирической функцией распределения, определяемой по выборке, и гипотетической модельной $F_0(x;\theta^{(1)},...,\theta^{(s)})$.

При выполнении лабораторной работы необходимо проверить гипотезу согласия с помощью критериев Пирсона и Колмогорова.

1.1. Критерий χ^2 Пирсона

Критерий χ^2 Пирсона позволяет проверить гипотезу (2.1), когда значения параметров $\theta^{(1)},...,\theta^{(s)}$ неизвестны и данные группированы. Процедура проверки гипотезы состоит из следующих шагов.

1. Диапазон значений исследуемой случайной величины ξ разбивается на k взаимно исключающих и непересекающихся интервалов $I_1,...,I_k$. Длина интервалов разбиения не обязательно одинакова.

2. На основании выборочных данных $x_1, x_2, ..., x_n$ строятся статистические оценки $\hat{\theta}^{(1)},...,\hat{\theta}^{(s)}$ неизвестных параметров $\theta^{(1)},...,\theta^{(s)}$, от которых зависит закон распределения F.

3. Подсчитывается число наблюдений n_i, попадающих в каждый интервал группирования I_i, $i = 1,\ldots, k$.

4. Вычисляются вероятности событий $\xi \in I_i$, т.е. вероятности p_i попадания случайной величины ξ в интервал I_i:

$$p_i = F_0(l_i;\ \hat{\theta}^{(1)},\ldots,\hat{\theta}^{(s)}) - F_0(l_{i-1};\ \hat{\theta}^{(1)},\ldots,\hat{\theta}^{(s)}),$$

где и l_i – левый и правый концы i-го интервала группирования.

5. Вычисляется ожидаемое число наблюдений ν_i в интервале I_i при условии справедливости гипотезы H_0: $\nu_i = n\,p_i$.

6. Вычисляется статистика

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \nu_i)^2}{\nu_i},$$

которая при верной H_0 имеет χ^2–распределение с $f = k - s - 1$ степенями свободы.

7. ***Гипотеза*** о том, что исследуемая случайная величина ξ подчиняется закону распределения F_0, ***принимается на уровне значимости α***, если

$$\Delta(\alpha/2) \le \chi^2 < \Delta(1 - \alpha/2),$$

где $\Delta(\varepsilon)$ – квантиль уровня ε имеет χ^2– распределение с $f = k - s - 1$ степенями свободы.

Если $\chi^2 \ge \Delta\ (1 - \alpha/2)$, ***гипотеза H_0 отклоняется***, так как выполнение неравенства свидетельствует о слишком большом отклонении исследуемого закона распределения от $F_0(x)$. Случай $\chi^2 < \Delta\ (\alpha/2)$ требует дополнительного исследования. Слишком малые значения статистики критерия говорят о неудачном выборе закона F (завышение числа параметров), нарушении технологии выборочного обследования и т.д.

Пример 2.1. Требуется проверить гипотезу H_0 (2.1) о том, что генеральная совокупность имеет стандартное нормальное распределение ($\mu = 0$, $\sigma^2 = 1$), т.е. H_0 : $F\ (x) = \Phi\ (x)$. Для проверки этой гипотезы используем критерий согласия χ^2 на уровне значимости $\alpha = 0{,}05$ по результатам наблюдений выборки объема $n = 100$.

Воспользуемся приведенной выше схемой проверки гипотезы согласия с помощью χ^2-критерия.

Найдем статистические оценки параметров нормального распределения. Статистические оценки параметров нормального распределения, полученные по результатам наблюдений, составляют: среднее значение $\overline{X} = 0{,}0025$, выборочная дисперсия $s^2 = 2{,}5398$, среднее квадратическое отклонение $s = 1{,}5937$.

Результаты выполнения пунктов 1, 3 – 5 представлены в табл.2.1. Структура этой таблицы такова:

столбец 1 – значения границ интервалов группирования;

столбец 2 – число наблюдений n_i, попадающих в каждый интервал группирования (наблюдаемая частота);

столбец 3 – наблюдаемое число значений случайной величины, не превышающих верхнюю границу рассматриваемого интервала (наблюдаемая накопленная частота);

столбцы 4 и 5 – значения частоты в столбцах 2 и 3, выраженные в %;

столбец 6 – ожидаемое число наблюдений ν_i в интервале I_i (ожидаемая частота);

столбец 7 – теоретическое число значений случайной величины, не превышающих верхнюю границу рассматриваемого интервала (накопленная ожидаемая частота);

столбцы 8 и 9 – величины из столбцов 6 и 7, выраженные в %: вероятности p_i попадания случайной величины ξ в интервал I_i; столбец 9 – значения теоретической функции распределения $F_0(l_i;\ \hat{\theta}^{(1)},...,\hat{\theta}^{(s)})$;

столбец 10 – разность между эмпирической и теоретической функциями распределения.

В результате вычислений получаем значение статистики $\chi^2 = 5{,}3818$ и число степеней свободы $f = 7$.

Таблица 2.1

Результаты вычислений критерия согласия

Границы интерва-лов	Наблю-даемая частота	Накопл. наблюд. частота	Наблюд. частота %	Накопл. набл.час-тота, %	Ожидае-мая частота	Накопл. ожидаем. частота	Ожид. частота, %	Ожид. накопл. част.,%	Набл.– ожид. %
1	2	3	4	5	6	7	8	9	10
<= -4,	0	0	0	0	0,596	0,596	0,596	0,596	-0,596
-3,5	1	1	1	1	0,821	1,417	0,821	1,417	0,179
-3,0	0	1	0	1	1,655	3,072	1,655	3,072	-1,655
-2,5	3	4	3	4	3,007	6,079	3,007	6,079	-0,007
-2,0	9	13	9	13	4,930	11,009	4,930	11,009	4,070
-1,5	6	19	6	19	7,293	18,302	7,293	18,302	-1,293
-1,0	9	28	9	28	9,734	28,036	9,734	28,036	-0,734
-,5	14	42	14	42	11,722	39,758	11,722	39,758	2,278
0,0	15	57	15	57	12,735	52,493	12,735	52,493	2,265
,5	9	66	9	66	12,483	64,976	12,483	64,976	-3,483
1,0	12	78	12	78	11,040	76,016	11,040	76,016	0,960
1,5	5	83	5	83	8,809	84,825	8,809	84,825	-3,809
2,0	8	91	8	91	6,342	91,167	6,342	91,167	1,658
2,5	4	95	4	95	4,119	95,286	4,119	95,286	-0,119
3,0	1	96	1	96	2,414	97,700	2,414	97,700	-1,414
3,5	1	97	1	97	1,276	98,976	1,276	98,976	-0,276
4,0	2	99	2	99	0,609	99,585	0,609	99,585	1,391
4,5	1	100	1	100	0,262	99,847	0,262	99,847	0,738
∞	0	100	0	100	0,153	100,000	0,153	100,000	-0,153

По таблицам χ^2 – распределения или с помощью программы **Probability Calculator** (Вероятностный Калькулятор) пакета STATISTICA определяем границы области принятия H_0: $\chi^2_{0.025;7} = 1{,}690$ и $\chi^2_{0.975;7} = 15{,}750$. Так как $1{,}690 < \chi^2 = 5{,}3818 < 15{,}750$, то можно сделать вывод о том, что результаты наблюдений не противоречат гипотезе H_0, т.е. выборочные данные принадлежат совокупности со стандартным нормальным распределением.

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 $\chi^2 = 5{,}3818$, вычисляется *P*–значение. Его величина $p = 0{,}613$, поэтому при уровне значимости $\alpha = 0{,}05$ нулевая гипотеза о нормальности распределения принимается.

1.2. Критерий Колмогорова–Смирнова

Когда модельное распределение известно полностью и является непрерывным, для проверки гипотезы согласия (2.1) целесообразно использовать критерий Колмогорова–Смирнова.

Определим расстояние Колмогорова между эмпирической $F_э(x)$ и теоретической $F_0(x)$ функциями распределения:

$$D = \max_x | F_э(x) - F_0(x) | .$$

Решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если} \quad n^{1/2} D < \Delta(\alpha), \\ H_1, & \text{если} \quad n^{1/2} D \geq \Delta(\alpha). \end{cases}$$

Пример 2.2. Проверить на уровне значимости α = 0,05 гипотезу H_0 (2.1) о том, что распределение анализируемой случайной величины является стандартным нормальным ($\mu = 0$, $\sigma^2 = 1$). Получена выборка объема $n = 100$. Для проверки этой гипотезы используем критерий Колмогорова. В результате обработки данных имеем $D = 0{,}0451$. Отсюда значение статистики критерия $n^{1/2} D = \sqrt{100} \cdot 0{,}0451 = 0{,}451$. По табл. 2.2 при α = 0,05 находим границу области принятия нулевой гипотезы $\Delta(\alpha) = 1{,}36$. Так как $n^{1/2} D = 0{,}451 < 1{,}36$, то результаты наблюдений не противоречат гипотезе H_0. Следовательно, полученная выборка подчиняется стандартному нормальному распределению.

Таблица 2.2

Значения квантилей распределения Колмогорова

α	0.01	0.05	0.1	0.2
Δ	1.63	1.36	1.22	1.07

Результаты проверки гипотезы согласия с помощью критериев χ^2-Пирсона и Колмогорова показаны на рис. 2.1.

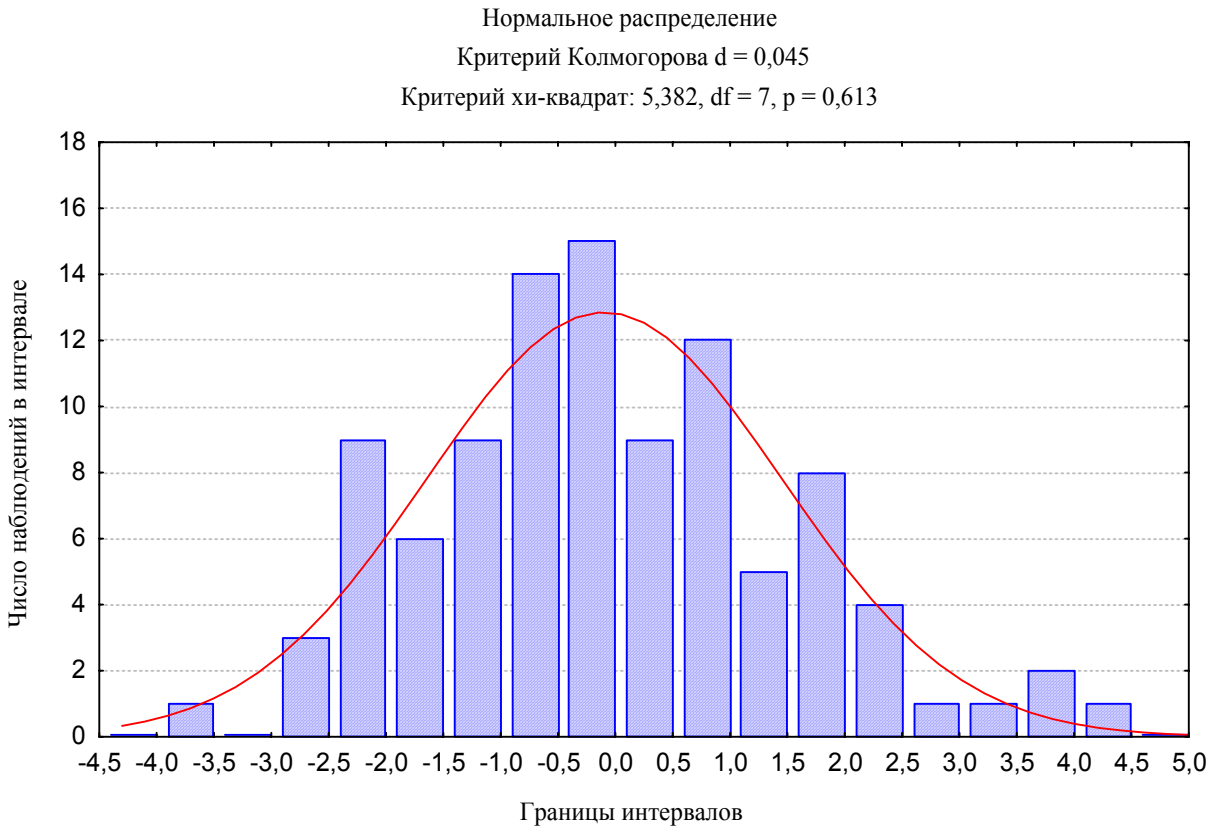

Рис. 2.1. Результаты проверки гипотезы согласия

1.3. Проверка гипотезы согласия в пакете STATISTICA

Для проверки гипотезы согласия необходимо выполнить следующую последовательность действий.

1. Из главного диалогового окна пакета вызвать программный модуль «Непараметрическая статистика» **Nonparametrics/Distrib.**

2. После появления диалогового окна **Nonparametric Statistics** задать с помощью радиокнопки режим выбора распределения **Distribution Fitting**, выбрать вид распределения из списков непрерывных и дискретных распределений и нажать кнопку **ОК**.

3. В открывшемся диалоговом окне **Fitting Continuous Distributions** с помощью кнопки **Variables** перейти к заданию анализируемой переменной.

4. После задания переменной по кнопке **ОК** получить таблицу с результатами проверки гипотезы согласия.

5. С помощью кнопки **GRAPH** построить гистограмму и графики плотности и функции распределения.

2. Проверка гипотез однородности

Гипотезы равенства (однородности) математических ожиданий H_1 и дисперсий H_2 и их двусторонние альтернативы $\overline{H}_1$ и $\overline{H}_2$ можно записать в виде

$$H_1:\ \mu_1=\mu_2,\ \overline{H}_1:\ \mu_1\neq\mu_2; \tag{2.2}$$

$$H_2:\ \sigma_1^2=\sigma_2^2,\ \overline{H}_2:\ \sigma_1^2\neq\sigma_2^2. \tag{2.3}$$

2.1. Однородность математических ожиданий

Случай равных дисперсий ($\sigma_1^2=\sigma_2^2$)

Исходные данные. Две независимые случайные выборки $X^{(1)}$ и $X^{(2)}$ объемов n_1 и n_2 соответственно: $X^{(i)}=(x_{i1},...,x_{in_i})$, $i=1, 2$.

Предположения. Выборки извлечены из нормальных распределений с равными дисперсиями ($\sigma_1^2=\sigma_2^2$).

Пример 2.3. Сравниваются по урожайности два сорта пшеницы. Сорт *A* – обычная разновидность, сорт *B* – новый гибрид. Была засеяна одинаковая площадь пшеницей каждого сорта, причем условия созревания на обоих участках были одинаковы. Средний урожай сорта *A* – 32 ц/га с дисперсией, равной 5,9. Средний урожай сорта *B* – 36,2 ц/га с дисперсией, равной 11,2. Является ли урожайность сорта *B* значительно более высокой, чем урожайность сорта *A*?

Решение

1. Выдвижение гипотез H_0, H_1:

$H_0:\ \ \mu_A=\mu_B,\ H_1:\ \mu_A<\mu_B$.

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

3. Выбор критической статистики (критерия). В качестве статистики критерия выберем двухвыборочную статистику Стьюдента:

$$t=(\overline{X}_1-\overline{X}_2)/(s^2\ (1/n_1+1/n_2))^{1/2}, \tag{2.4}$$

$$s^2=\sum_{i=1}^{2}(n_i-1)s_i^2/(\sum_{i=1}^{2}n_i-2)\text{–объединенная выборочная дисперсия.} \tag{2.5}$$

При справедливости H_0 и нормальном распределении исходной случайной переменной статистика имеет t–распределение Стьюдента с n_1+n_2-2 степенями свободы.

4. Определение границы критической области x_k. Значение границы критической области при левосторонней H_1 представляет собой квантиль уровня $p=\alpha$ распределения статистики критерия. Определим значение квантили $t_{0,05;48}$ уровня $\alpha=0,05$ t–распределения Стьюдента с $n_1+n_2-2=48$ степенями свободы: $t_{0,05;48}=-1,645$.

Следовательно, значение границы критической области x_k равно –1,645. Это означает, что при значениях статистики критерия $t \geq -1,645$ принимается гипотеза H_0, а при значениях $t < -1,645$ – гипотеза H_1.

5. Определение по формулам (2.4) и (2.5) численной величины статистики критерия: $t=-5,08$.

6. Выработка решения. Используем решающее правило:

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

где $\Delta(\varepsilon)$ – квантиль уровня α t–распределения Стьюдента с n_1+n_2-2 степенями свободы, определяющая границу критической области при левосторонней альтернативе.

В п.4 было найдено значение границы критической области $x_k=\Delta(\alpha)=-1,645$.

При уровне значимости $\alpha=0,05$ гипотеза о равенстве среднего значения $\mu_A=\mu_B$, отклоняется, так как $t=-5,08<-1,645$. Следовательно, при уровне значимости $\alpha=0,05$ принимается гипотеза H_1. Это означает, что урожайность сорта B значительно более высокая, чем урожайность сорта A.

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 $n_1+n_2-2=48$ степенями свободы, соответствующее значению статистики $t=-5,08$: $F_t(-5,08)=0,000003$. В условиях левосторонней альтернативы получаем $P=F_t(-5,08)=$

= 0,000003. Так как $P < \alpha = 0{,}05$, поэтому при уровне значимости $\alpha = 0{,}05$ гипотеза H_0 отклоняется.

Случай неравных дисперсий ($\sigma_1^2 \neq \sigma_2^2$)

При нарушении условия равенства дисперсий ($\sigma_1^2 \neq \sigma_2^2$) для проверки гипотезы (2.2) используется статистика Уэлча

$$t_1 = (\overline{X}_1 - \overline{X}_2 - \delta) / (s_1^2/n_1 + s_2^2/n_2)^{1/2},$$

имеющая t–распределения Стьюдента с ν_1 степенями свободы,

$$\nu_1 = \left(\frac{c_1^2}{n_1 - 1} + \frac{(1 - c_1^2)^2}{n_2 - 1} \right)^{-1}, \quad c_1 = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}.$$

Последовательность проверки гипотезы (2.2) в условиях неравенства дисперсий аналогична случаю равных дисперсий.

2.2. Однородность дисперсий

Исходные данные. Две независимые случайные выборки $X^{(1)}$ и $X^{(2)}$ объемов n_1 и n_2 соответственно: $X^{(i)} = (x_{i1}, ..., x_{in_i})$, $i = 1, 2$.

Предположения. Выборки извлечены из нормальных распределений.

Пример 2.4. В Примере 2.3 сравнивалась урожайность двух сортов пшеницы. Можно ли утверждать равенство дисперсий для урожайности сорта *A* и урожайности сорта *B*?

Решение

1. Выдвижение гипотез H_0, H_1:

H_0 : $\sigma_1^2 = \sigma_2^2$, H_1: $\sigma_1^2 \neq \sigma_2^2$.

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

3. Выбор критической статистики (критерия). В качестве статистики критерия выберем

$$F = \frac{s_1^2}{s_2^2} = \frac{\sum_{i=1}^{n_1} (x_{1i} - \overline{X}_1)^2 / (n_1 - 1)}{\sum_{i=1}^{n_2} (x_{2i} - \overline{X}_2)^2 / (n_2 - 1)}, \qquad (2.6)$$

При справедливости H_0 и нормальном распределении исходной случайной переменной статистика (2.6) имеет F–распределение Фишера с числами

степеней свободы числителя и знаменателя, равными соответственно $n_1 - 1$ и $n_2 - 1$.

4. Определение границы критической области x_k. Значение границ критической области при двусторонней H_1 представляет собой квантили уровней $p_1 = \alpha/2$ и $p_2 = 1 - \alpha/2$ распределения статистики критерия. Из таблиц F–распределения для значений функции распределения, равных соответственно 0,025 и 0,975, и степеней свободы $n_1 - 1 = 24$ и $n_2 - 1 = 24$ находим значение квантилей $F_{0,025;24} = 0,4407$ и $F_{0,975;24} = 2,2693$. Это означает, что при значениях статистики критерия $0,4407 \le F \le 2,2693$ принимается гипотеза H_0, а при значениях $F < 0,4407$ или $F > 2,2693$ – гипотеза H_1.

5. Определение по формуле (2.6) численной величины статистики критерия: $F = 11,2/5,9 = 1,90$.

6. Выработка решения. Используем решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_2, & \text{если } \Delta(\alpha/2) \le F \le \Delta(1-\alpha/2), \\ \overline{H}_2, & \text{если } F < \Delta(\alpha/2) \text{ или } F > \Delta(1-\alpha/2), \end{cases}$$

где $\Delta(\varepsilon)$ – порог теста, определяемый при верной H_0 (2.3) как квантиль уровня ε F–распределения Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы.

В п.4 были найдены значения границ критической области $x_{k1} = \Delta(\alpha/2) = 0,4407$ и $x_{k2} = \Delta(1 - \alpha/2) = 2,2693$.

При уровне значимости $\alpha = 0,05$ гипотеза о равенстве дисперсий $\sigma_1^2 = \sigma_2^2$ принимается, так как $0,4407 < F = 1,90 < 2,2693$. Следовательно, при уровне значимости $\alpha = 0,05$ принимается гипотеза H_0. Это означает равенство дисперсий для урожайности сорта A и урожайности сорта B.

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

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

Находим значение функции F–распределения с $n_1 - 1 = 24$ и $n_2 - 1 = 24$ степенями свободы, соответствующее значению статистики $F = 1,90$: $F_F(1,90) = 0,9385$. В условиях двусторонней альтернативы получаем $1 - P/2 = F_t(1,90) = 0,9386$. Отсюда $P = 0,1228$. Так как $P > \alpha = 0,05$, поэтому при уровне значимости $\alpha = 0,05$ гипотеза H_0 принимается.

2.3. Проверка гипотез однородности в пакете STATISTICA

Для проверки гипотез однородности необходимо выполнить следующую последовательность действий.

1. Из главного диалогового окна пакета вызвать программный модуль «Основная статистика» **Basic Statistics.**

2. После появления диалогового окна **Basic Statistics and Tables** выбрать в нем строку **T-test for independent samples** для вызова программы вычисления *t*–критерия Стьюдента и *F* –критерия и нажать кнопку **ОК**.

3. В открывшемся диалоговом окне **T-test for independent samples** с помощью кнопки **Variables** перейти к заданию анализируемых переменных: в левом списке задать группирующую переменную, в правом списке – анализируемые переменные. В полях Code for group указать код анализируемых групп (выборок). Затем выставить флажок для опции проверки гипотезы равенства средних значений в условиях неравенства дисперсий: **T-test for separate variance estimates.**

4. После задания переменных и установки необходимых режимов вычислений по кнопке **ОК** получить таблицу с результатами проверки гипотезы однородности средних и дисперсий.

5. С помощью кнопки **Categorized histograms** построить гистограммы анализируемых признаков для каждой выборки.