Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей

Кафедра электронных вычислительных машин

КОНТРОЛЬНАЯ РАБОТА

по учебной дисциплине «Системный анализ»

Выполнил:
студент гр. 300541
Строж Д. А.

Проверил:
Иванов Н. Н.

Минск 2018

СОДЕРЖАНИЕ

Задача 1. Построение линейной регрессии. Вычисление коэффициента корреляции

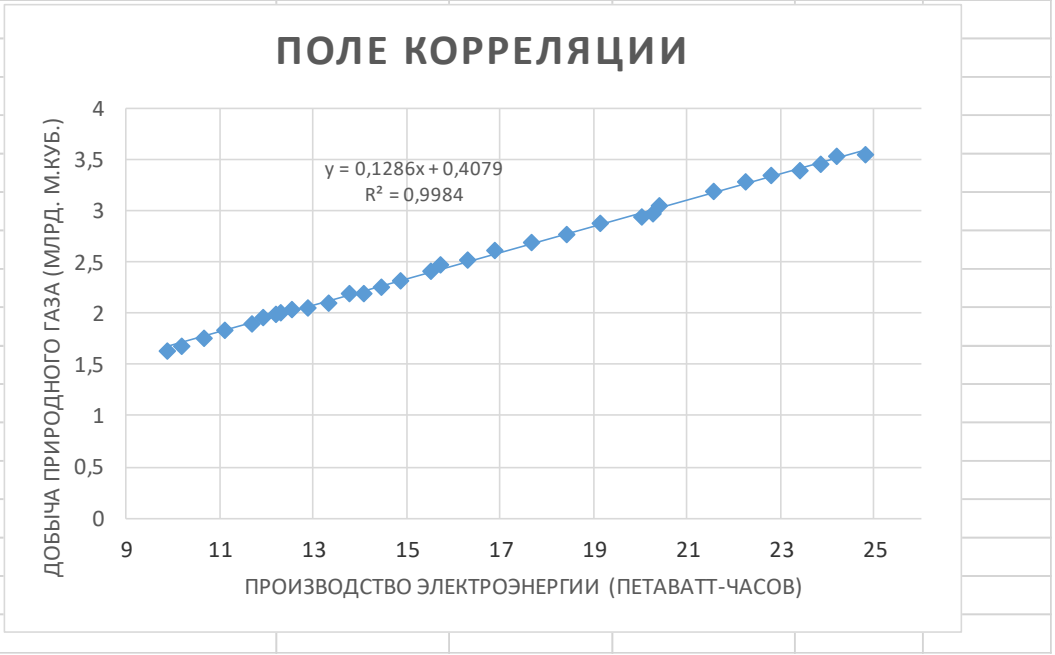
Дано: статистическая таблица соотношения добычи природного газа и производства электрической энергии с 1985 по 2016 годы.

Задача: построить линейную регрессию и вычислить коэффициент корреляции.

Решение:

Определим зависимость производства электрической энергии от добычи природного газа.

Коэффициент корреляции	0,999215837				



В прямоугольной системе координат построен график, по оси ординат определены индивидуальные значения признака зависимой переменной или переменной отклика Y (добыча природного газа) - значение, которое мы ожидаем для Y (в среднем), если мы знаем величину X, т.е. это «предсказанное значение y», а по оси абсцисс - значения признака независимой переменной или предиктора X (производство электрической энергии). Совокупность точек обоих признаков называется полем корреляции. На основании поля корреляции можно выдвинуть гипотезу (для генеральной совокупности) о том, что связь между всеми возможными значениями x и y носит линейный характер.

Линия на графике - связь между изучаемыми признаками. Она выражена уравнением прямой линии регрессии y на X: $y = ax + b$,

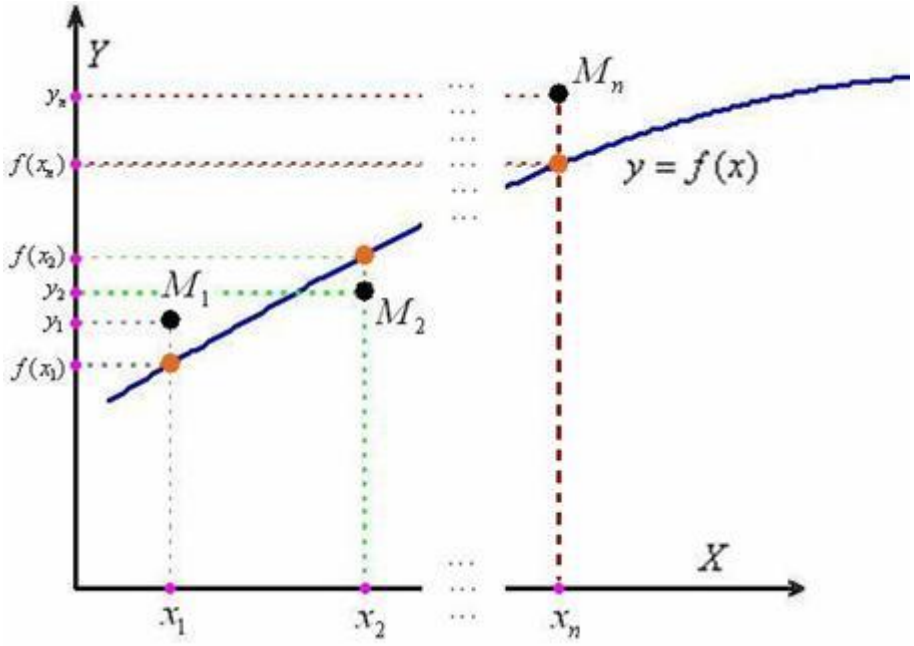
где: a – угловой коэффициент или градиент оценённой линии; она представляет собой величину, на которую Y увеличивается в среднем, если мы увеличиваем X на одну единицу.

b – свободный член (пересечение) линии оценки; это значение Y, когда X = 0.

Параметры a и b называют коэффициентами регрессии оценённой линии, хотя этот термин часто используют только для b.

Для вычисления параметров a, b используется система нормальных уравнений метода наименьших квадратов, которую можно вывести нижеприведенным способом.

Пусть некоторый график функции $y = f(x)$ приближает экспериментальные данные $M_1(x_1; y_1), M_2(x_x; y_2), \dots, M_n(x_n; y_n)$:



Чтобы оценить точность данного приближения вычислим $f(x_1), f(x_2), \dots, f(x_n)$ и разности (т.е. отклонения) $e_1 = y_1 - f(x_1), e_2 = y_2 - f(x_2), \dots, e_n = y_n - f(x_n)$. Оценим, насколько велика сумма $e_1 + e_2 + \cdots + e_n$. Так как разности могут быть отрицательными, то целесообразно суммировать их модули или, использовать распространенный метод наименьших квадратов, в котором возможные отрицательные значения ликвидируются не модулем, а возведением отклонений в квадрат: $e_1^2 + e_2^2 + \cdots + e_n^2$, после чего, если точки на графике имеют тенденцию располагаться по прямой, то следует искать уравнение

прямой $y = f(x) = ax + b$ с оптимальными значениями a и b – чтобы сумма квадратов отклонений $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$ была наименьшей.

Дифференцируем и находим производные по *x* и *y*, и после чего составим стандартную систему:

$$\begin{cases} \dfrac{\partial F}{\partial a} = 0 \\ \dfrac{\partial F}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} 2\sum_{i=1}^{n} (ax_i^2 + bx_i - x_i y_i) = 0 \\ 2\sum_{i=1}^{n} (ax_i + b - y_i) = 0 \end{cases}$$

После этого сокращаем каждое уравнение на 2 и, кроме того, разделим выражения на суммы:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0 \\ a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} b - \sum_{i=1}^{n} y_i = 0 \end{cases} \Rightarrow \begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0 \\ a\sum_{i=1}^{n} x_i + \underbrace{(b + b + \ldots + b)}_{n \text{ раз}} - \sum_{i=1}^{n} y_i = 0 \end{cases}$$

Получаем систему нормальных уравнений метода наименьших квадратов в прикладном виде:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases}$$

Ковариация - это величина, характеризующая совместное изменение значений двух параметров. Она задается как сумма произведений отклонений наблюдаемых значений X и Y от средних $\overline{X}$ и $\overline{Y}$ соответственно, т. е. $\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$, деленная на количество наблюдений.

Для выборочного коэффициента корреляции $r_{x,y}$ используется следующая формула:

$$r_{x,y} = \frac{cov(x,y)}{S_x \cdot S_y},$$

где: *cov(x,y)* – выборочная ковариация;

S_x, S_y – выборочные среднеквадратические отклонения x и y соответственно.

Чтобы понять «физический смысл» ковариации, достаточно обратить внимание на следующее свойство: если для какого-то объекта i в выборке оба значения — X_i и Y_i — окажутся высокими, то и произведение $(X_i - \overline{X})$ на $(Y_i - \overline{Y})$ будет большим и положительным. Если оба значения (по X и по Y) низки, то

произведение двух отклонений, т.е. двух отрицательных чисел, также будет положительным. Таким образом, если линейная связь X и Y положительная и велика, сумма таких произведений для всех наблюдений также будет положительна. Если связь межу X и Y обратная, то многим положительным отклонениям по X будет соответствовать отрицательные отклонения по Y, т.е. сумма отрицательных произведений отклонений будет отрицательной. При отсутствии систематической связи произведения будут иногда положительными, иногда отрицательными, а их сумма (и, следовательно, ковариация X и Y) будет, в пределе, равная нулю. Таким образом, ковариация показывает величину и направление связи, совместного изменения X и Y. Если разделить ковариацию *cov(x,y)* на среднеквадратические отклонения S_x, S_y (чтобы избавиться от влияния масштаба шкал, в которых измеряются X и Y), то мы получим искомую форму коэффициента корреляции.

Выборочная ковариация является смещенной оценкой теоретической ковариации и, чтобы размер отклонений от теоретических средних не занижался и математическое ожидание выборочной ковариации не оказалось меньше теоретической ковариации применяют несмещенную оценку теоретической ковариации, которую получают путем умножения выборочной ковариации на поправку $1/(n-1) \cdot cov(x,y)$.

Главным недостатком ковариации как показателя связи является то, что его значение зависит от единиц измерения исходных данных и не имеет критических значений, что затрудняет как сравнение различных совокупностей на предмет силы связи, так и делает невозможным установления критических значений ковариации. Этот недостаток преодолевает следующий показатель связи – коэффициент корреляции.

Коэффициент корреляции используется для определения взаимосвязи между двумя признаками. Например, можно установить зависимость между средней температурой в помещении и использованием кондиционера. Это безразмерная величина в диапазоне от -1 (при отрицательной линейной зависимости) до +1 (при положительной линейной зависимости). Близкий к нулю коэффициент указывает на отсутствие линейной зависимости между изучаемыми признаками это означает, что зависимость отсутствует, либо носит нелинейный характер. В данном случае он равен 0,999210275.

R^2 – коэффициент детерминации (в задаче имеет значение 0,9984 или 99,84%) вычисляется по формуле:

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2},$$

где s_y^2 – фактическая дисперсия зависимой переменной, $s_{\hat{y}}^2$ – дисперсия оценочных значений зависимой переменной, полученных на основании модели: $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, $s_{\hat{y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

Он означает, что расчетные параметры модели на 99,84% объясняют зависимость между изучаемыми параметрами. Чем выше коэффициент детерминации, тем качественнее модель. Хорошо – выше 0,8. Плохо – меньше 0,5. В данном случае – хорошо.

Задача 2. Проверка гипотез

Дано: статистическая таблица соотношения добычи природного газа и производства электрической энергии с 1985 по 2016 годы.

Задача: найти доверительный интервал для математического ожидания и дисперсии и проверить гипотезу о равенстве математических ожиданий двух генеральных совокупностей для известных дисперсий.

Решение:

Доверительный интервал относится к выборочной совокупности. Он показывает, на сколько параметры из выборочной совокупности могут отличаться от реальных существующих данных в генеральной совокупности. На сколько мы ошибаемся при формировании той или иной выборки мы закладываем в так называемую ошибку репрезентативности и вокруг нее строим доверительный интервал. Ширину доверительного интервала задается исследователем, варьируя точностью оценки. Доверительный интервал определяет границы, в которых будет находиться значение теоретического коэффициента регрессии с уровнем значимости α.

Уровень значимости α определяется исходя из требуемой точности. Обычно – 0.1, 0.05 или 0.01.

Доверительный интервал для математического ожидания μ:

Надежность (1-γ)	0,95	Доверительный интервал для математического ожидания	1,67737209
Уровень значимости (γ)	0,05		
Выборочное среднее ($\bar{x}$)	16,5		
Среднеквадратическое откл. (S)	4,65240837		
Объем выборки (n)	32		
		Нижняя граница	14,9
		Верхняя граница	18,2

Он определяется как

$$\left(\bar{x} - t_\gamma \frac{\sigma_в}{\sqrt{n}} \leq \mu \leq \bar{x} + t_\gamma \frac{\sigma_в}{\sqrt{n}}\right)$$

где: $\bar{x}$ - выборочное среднее арифметическое для x;

$t\frac{s}{\sqrt{n}} = \delta$ – точность оценки;

t_γ — значение аргумента функции Лапласа Φ(t), которое определяется по таблице по заданным n и γ;

$\sigma_в$ — выборочное среднеквадратическое отклонение $\sqrt{\frac{\sum n_i(x_i-\bar{x})^2}{n-1}}$;

n – объем выборки.

Суть формулы в том, что берется среднее арифметическое и далее от нее откладывается некоторое количество стандартных ошибок ($\frac{\sigma_в}{\sqrt{n}}$), умноженных на коэффициент t.

В итоге при вероятности равной 0,95 мы видим, что доверительный интервал для математического ожидания результата измерения лежит в области от 14,9 до 18,2.

Доверительный интервал для дисперсии D_X:

Число степеней свободы	31	Верхний доверительный интервал для дисперсии	15,3969463
Надежность (1-α)	0,9		
Уровень значимости (α)	0,1		
Дисперсия (S^2)	21,64490364		
Верхний квантиль	44,98534328	Нижний доверительный интервал для дисперсии	35,9240919
Нижний квантиль	19,28056856		
Выборочное среднее ($\bar{x}$)	16,5		

Он определяется как интервал между областями:

$$\frac{(n-1)s^2}{\chi^2_{\frac{1-\gamma}{2},n-1}} < D_X < \frac{(n-1)s^2}{\chi^2_{\frac{1+\gamma}{2},n-1}}$$

где: s^2 - выборочная дисперсия;

D_X - "истинная" дисперсия результата измерения, т.е. оцениваемый параметр, который нам не известен;

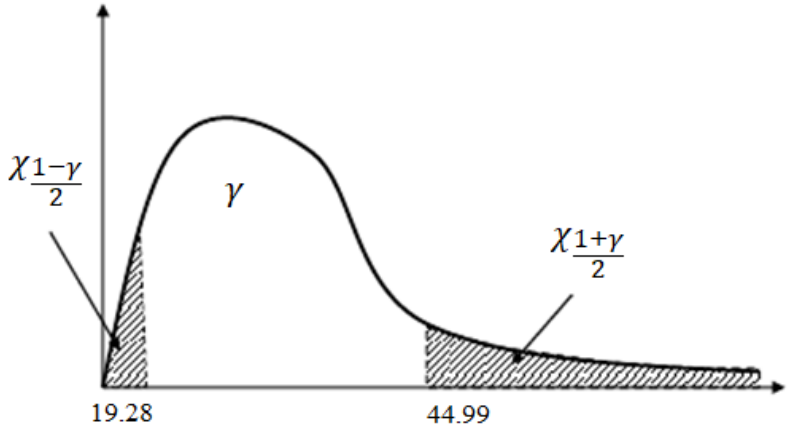
X^2 - квантиль распределения (найти его можно с помощью таблицы X^2-распределения);

γ – доверительная вероятность;

n - объем выборки.

Квантиль X^2 — это число, при котором функция распределения хи-квадрат равна заданной вероятности $(1-\gamma)/2$ либо $(1+\gamma)/2$ (для определения левого и правого X^2), и количеству степеней свободы $k = n - 1$.

На рисунке ниже показаны значения аргумента плотности распределения X^2 (19.28 и 44.99 – критические точки):



Значение выборочной дисперсии определяется формулой:

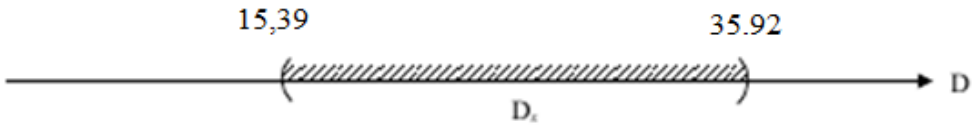
$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

где: $\bar{x}$ - выборочное среднее арифметическое для x;

$(x_i - \bar{x})$ - отклонение от средней величины для каждого значения x;

n - объем выборки.

В итоге при вероятности равной 0,9 мы видим, что доверительный интервал для "истинной" дисперсии результата измерения лежит в области от 15,39 до 35,92. Доверительный интервал для дисперсии можно изобразить следующим образом:

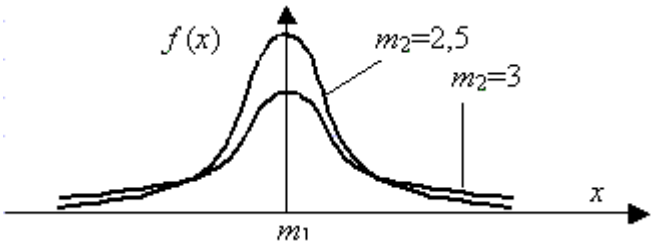


Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений. Нулевой (основной) называют выдвинутую гипотезу H_0. Конкурирующей (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

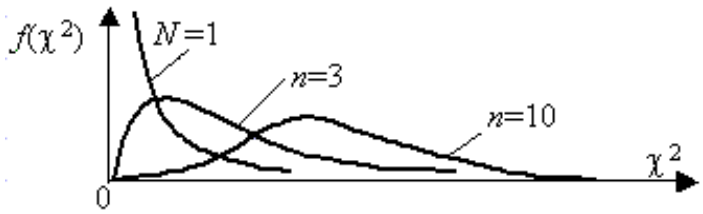
Шаги проверки статистических гипотез: формируем гипотезу; рассчитываем наблюдаемого значение критерия; находим критическую область; сравнением наблюдаемое значение критерия с критической областью и, если наблюдаемое значение критерия принадлежит критической области, то нулевую гипотезу отвергают, а если наблюдаемое значение критерия принадлежит области принятия гипотезы, то гипотезу принимают.

При проверке гипотез широкое применение находит ряд теоретических законов распределения. Важным из них является нормальное распределение. С ним связаны распределения хи-квадрат, Стьюдента, Фишера, а также интеграл вероятностей. Для указанных законов значения функций определяются по таблицам или с использованием стандартных процедур пакетов прикладных программ. Указанные таблицы содержат не значения функций распределения, а критические значения.

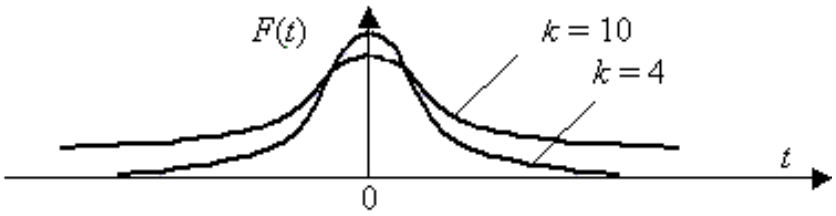
Плотность нормального распределения (m – среднее значение):



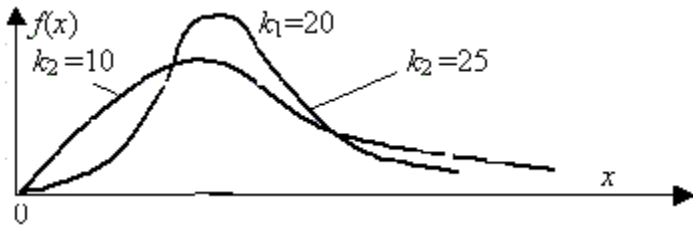
Плотность распределения хи-квадрат (n – количество случайных величин):



Плотность распределения Стьюдента (k - количество степеней свободы):



Плотность распределения Фишера (k - количество степеней свободы):



В данной задаче необходимо проверить простую гипотезу, содержащую только одно предположение используя нормальный закон распределения.

При заданном уровне значимости ($\gamma = 0{,}05$) необходимо проверить нулевую гипотезу H_0:M(X) = M(Y) о равенстве математических ожиданий двух нормальных генеральных совокупностей с известными дисперсиями при конкурирующей гипотезе H_1:M(X) ≠ M(Y).

Для этого необходимо вычислить наблюдаемое значение критерия:

$$Z_{\text{набл}} = (\bar{x} - \bar{y}) / \sqrt{\frac{D(x)}{n} + \frac{D(y)}{m}}$$

где: $\bar{x}, \bar{y}$ - выборочные средние арифметические для x и y;
$D(x), D(y)$ – генеральные дисперсии для x и y;
n, m – объемы выборок для x и y соответственно.

И по таблице Лапласа найти критическую точку $Z_{\text{кр}}$ из равенства $\Phi(Z_{\text{кр}}) = (1 - 2\gamma)/2$.

Уровень значимости (α)		0,05	
Выборочное среднее для Y		2,5	
Выборочное среднее для X		16,5	
Дисперсия Y		0,3471018	
Дисперсия X		20,9685004	
Z набл		17,1531371	
Ф (Z кр) / по таблице Лапласа		0,475	/1,96

$Z_{\text{набл}}$ = 17,15.

$\Phi(Z_{\text{кр}}) = (1 - 0{,}05)/2 = 0{,}475$.

По таблице Лапласа для равенства $\Phi(Z_{\text{кр}})$ = 0,475 получаем критическую точку, $Z_{\text{кр}}$ = 1,96.

Проверка гипотезы:

— если $|Z_{\text{набл}}| < Z_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу;

— если $|Z_{\text{набл}}| > Z_{\text{кр}}$ – нулевую гипотезу отвергают.

Получаем результат: $|Z_{\text{набл}}|$ (17,15) > $Z_{\text{кр}}$(1,96) или -1,96 > 17,15 > 1,96 – ложь, так как 17,15 не лежит в данном интервале.

Вывод: нулевую гипотезу отвергаем. Другими словами, выборочные средние двух генеральных совокупностей отличаются значительно.

Проделаем те же действия с помощью инструмента «Анализ данных» в Excel:

Двухвыборочный z-тест для средних		
	Переменная 1	*Переменная 2*
Среднее	16,53303125	2,533375
Известная дисперсия	20,9685004	0,3471018
Наблюдения	32	32
Гипотетическая разность средних	0	
z	17,15313706	
P(Z<=z) одностороннее	0	
z критическое одностороннее	1,644853627	
P(Z<=z) двухстороннее	0	
z критическое двухстороннее	1,959963985	

Видим расчетные данные с тем же результатом, т.е. - 1,64 > 17,15 > 1,64 – ложь.

Задача 3. Описательная статистика для изображений

Дано: два произвольных .JPG изображения (399x399 dpi):

cat.jpg

dog.jpg

Задача: конвертировать изображения в полутоновые (grayscale). Построить гистограммы изображений. Найти выборочное среднее, выборочное среднеквадратическое, моду и медиану для гистограмм. Найти коэффициент

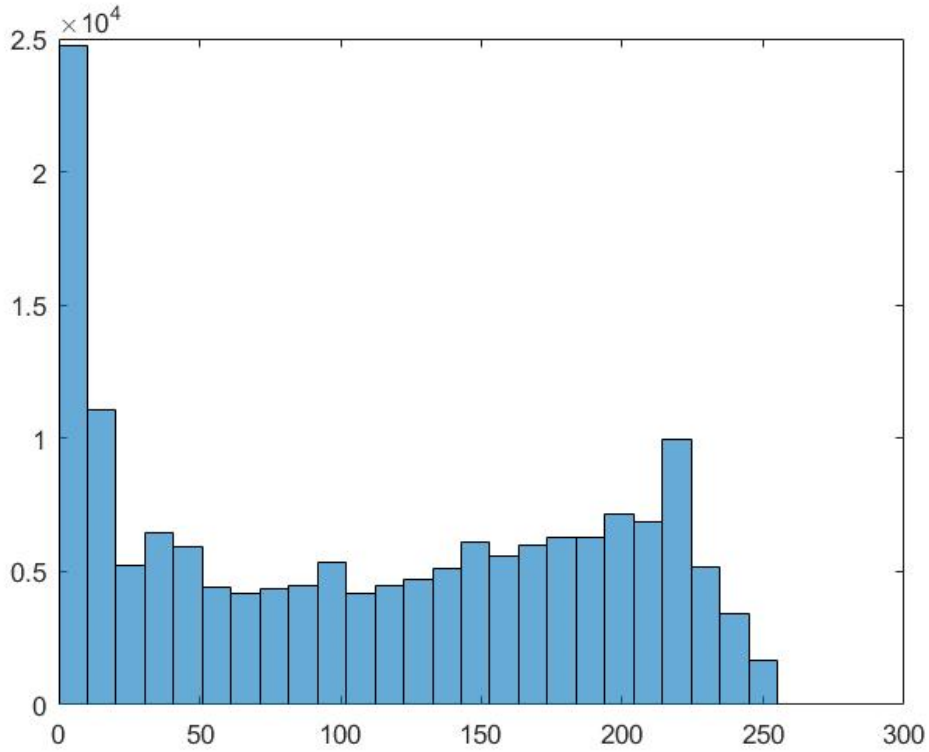
корреляции для гистограмм и изображений. Критерием Пирсона проверить гипотезы о соответствии выборочных функций распределений нормальному закону. Указание: в гистограмме использовать группировку интенсивностей, то есть разбить интервал интенсивностей [0, 255] на равные подинтервалы, напр длины 10 [0, 9), [10, 19) …, [240, 249), [250, 255) – последний интервал неполной длины, он учитывает только 6 интенсивностей.

Решение:

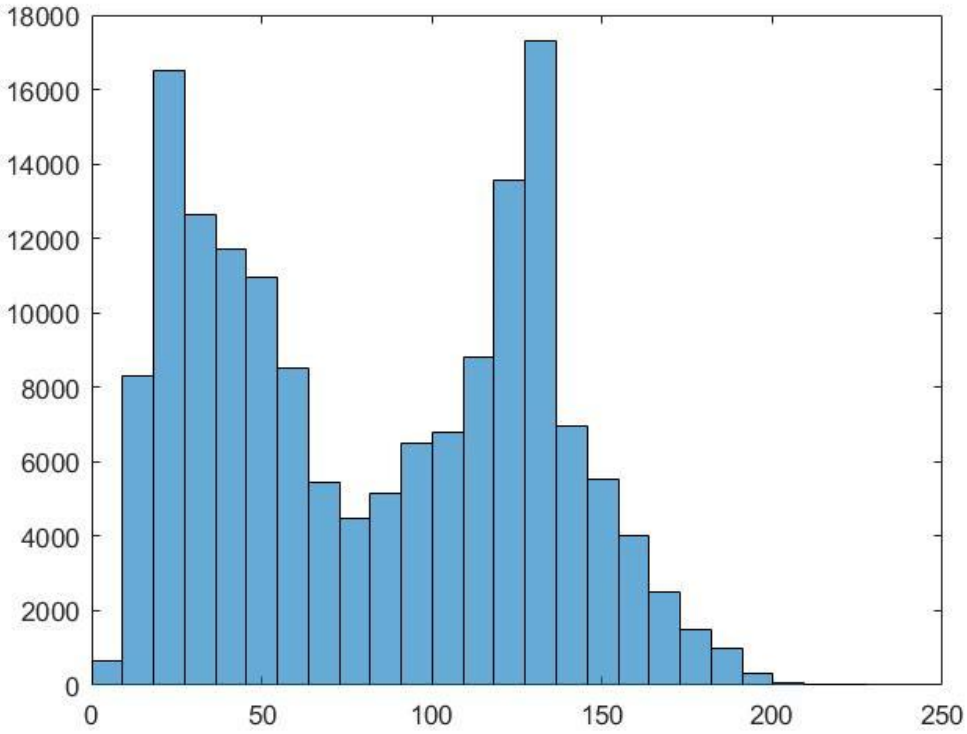
В среде MATLAB откроем два исходных изображения и сконвертируем изображения в полутоновые (grayscale) и построим гистограммы изображений этих изображений с помощью следующих команд:

```
>> CatImg=imread('cat.jpg');  % CatImg содержит матрицу изобр. cat.jpg
>> DogImg=imread('dog.jpg'); % DogImg содержит матрицу изобр. dog.jpg
>> GrayCat=rgb2gray(CatImg); % преобраз. CatImg в полутон
>> GrayDog=rgb2gray(DogImg); % преобраз. DogImg в полутон
>> histogram(GrayCat, 25); % вывод гистограммы для GrayCat
>> histogram(GrayDog, 25); % вывод гистограммы для GrayCat
>> HistCat=imhist(GrayCat, 25); % гистограмма для GrayCat
>> HistDog=imhist(GrayDog, 25); % гистограмма для GrayDog
```

Гистограмма «HistCat» изображения «cat.jpg»:



Гистограмма «HistDog» изображения «dog.jpg»:



Найдем выборочное среднее для гистограмм, т.е. оценки средних совокупностей:

```
>> Mcat=mean(HistCat); %выборочное среднее для HistCat>>
>> fprintf('Выборочное среднее для гистограммы HistCat = %f', Mcat);
Выборочное среднее для гистограммы HistCat = 6368.04
>> Mdog=mean(HistDog); %выборочное среднее для HistDog
>> fprintf('Выборочное среднее для гистограммы HistDog = %f', Mdog);
Выборочное среднее для гистограммы HistDog = 6368.04
```

Найдем выборочное среднеквадратическое для гистограмм, т.е. показатель рассеивания значений относительно математического ожидания:

```
>> Scat=std(HistCat); %среднеквадратическое для HistCat
>> fprintf('Среднеквадратическое для гистограммы HistCat = %f', Scat);
Среднеквадратическое для гистограммы HistCat = 3676.173854
>> Sdog=std(HistDog); %среднеквадратическое для HistDog
>> fprintf('Среднеквадратическое для гистограммы HistDog = %f', Sdog);
Среднеквадратическое для гистограммы HistDog = 6395.605311
```

Найдем моду для гистограмм, т.е. значение во множестве наблюдений, которое встречается наиболее часто:

```
>> MdCat=mode(HistCat); %мода для HistCat
>> fprintf('Мода для гистограммы HistCat = %f', MdCat);
Мода для гистограммы HistDog = 384.000000
>> MdDog=mode(HistDog); %мода для HistDog
>> fprintf('Мода для гистограммы HistDog = %f', MdDog);
Мода для гистограммы HistDog = 0.000000
```

Найдем медиану для гистограмм, т.е. значение срединного элемента:

```
>> MdnCat=median(HistCat); %медиана для HistCat
>> fprintf('Медиана для гистограммы HistCat = %f', MdnCat);
Медиана для гистограммы HistCat = 5604.000000
>> MdnDog=median(HistDog); %медиана для HistDog
>> fprintf('Медиана для гистограммы HistDog = %f', MdnDog);
Медиана для гистограммы HistDog = 5139.000000
```

Найдем коэффициент корреляции для гистограмм двух изображений, т.е. взаимосвязь между двумя признаками:

```
>> [R,P]=corrcoef(HistCat, HistDog);
>> fprintf('Коэффициент корреляции двух гистограмм = %f', R(1,2));
Коэффициент корреляции двух гистограмм = -0.164048
```

Найдем коэффициент корреляции для матриц двух изображений:

```
>> Rim=corr2(GrayCat,GrayDog);
>> fprintf('Коэффициент корреляции двух изображений = %f', Rim);
Коэффициент корреляции двух изображений = 0.412376
```

Критерием Пирсона проверим гипотезы о соответствии выборочных функций распределений (в виде вариант и соответствующих частот) нормальному закону. В теории для этого необходимо выполнить следующее:

1. Вычислить выборочное среднее арифметическое $\overline{x_в} = \frac{1}{n}\sum(\bar{x} - x_i)^2 n_i$ и выборочное среднеквадратическое отклонение $\sigma_в = \sqrt{\frac{\sum n_i(x_i - \bar{x})^2}{n-1}}$.
2. Выдвинуть гипотезу H_0: выборочная функция распределения HistCat либо HistDog соответствует нормальному закону с параметрами $\overline{x_в}$ и $\sigma_в$. Проверить эту гипотезу по критерию Пирсона при уровне значимости $\gamma = 0{,}05$.
3. Рассчитать теоретические частоты $n_i^0 = \frac{nh}{\sigma_в}\varphi(u_i)$, где n – объем выборки (сумма всех частот), h – шаг (разность между двумя соседними вариантами).

$$u_i = \frac{x_i - \overline{x_в}}{\sigma_в},\ \varphi(u_i) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

4. Сравнить эмпирические и теоретические частоты с помощью критерия Пирсона, а для этого составляют расчетную таблице, по которой находят наблюдаемое значение критерия $X^2_{набл} = \sum \frac{(n_i - n_i')^2}{n_i'}$. По таблице критических точек распределения X^2 , по заданному уровню значимости γ и числу степеней свободы k = s - 3 (s – число групп выборки) находят критическую точку $X^2_{кр}(\gamma, k)$ правосторонней критической области.
5. Если $X^2_{набл} < X^2_{кр}$ – нет оснований отвергнуть гипотезу о нормальном распределении. Другими словами, эмпирические и теоретические частоты различаются незначимо (случайно). Если $X^2_{набл} > X^2_{кр}$ – гипотезу отвергают. Другими словами, эмпирические и теоретические частоты различаются значимо.

Проделаем вышеописанные процедуры в среде MATLAB, используя лишь одну стандартную функцию «chi2gof», которая принимает выборку и проверяет нулевую гипотезу о соответствии выборочной функции распределения нормальному закону. Возвращаемое значение, равное 0 указывает, что chi2gof не отклоняет нулевую гипотезу на уровне значимости 5% по умолчанию. Возвращаемое значение, равное 1 указывает, что chi2gof отклоняет нулевую гипотезу на уровне значимости 5% по умолчанию.

Для гистограммы изображения cat.jpg:

```
>> hCat=chi2gof(HistCat);
>> fprintf('Возвращаемое значение hCat = %d', hCat);
Возвращаемое значение hCat = 1
```

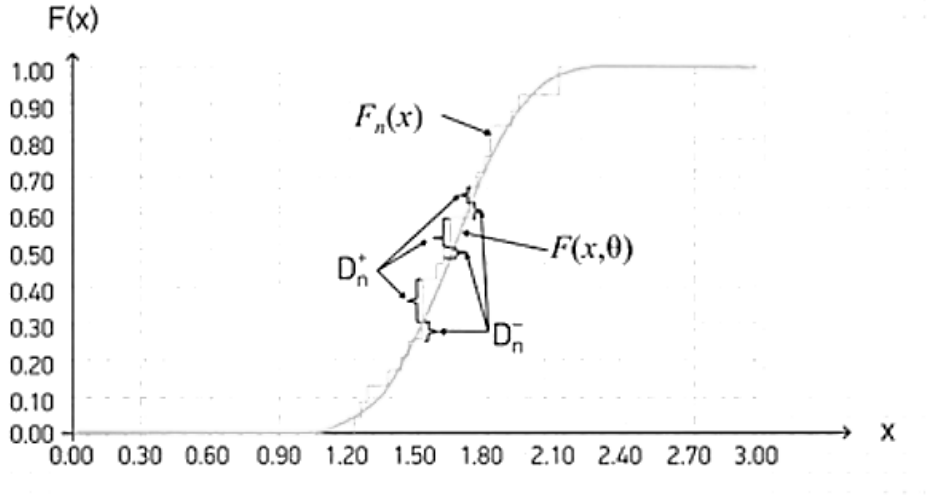
Вывод: отклоняем нулевую гипотезу на уровне значимости 5% о соответствии выборочной функции распределения нормальному закону, т.е. выборочная функция распределения HistCat не распределена по нормальному закону.

Для гистограммы изображения dog.jpg:

```
>> hDog=chi2gof(HistDog);
>> fprintf('Возвращаемое значение hDog = %d', hDog);
Возвращаемое значение hDog = 0
```

Вывод: не отклоняем нулевую гипотезу на уровне значимости 5% о соответствии выборочной функции распределения нормальному закону, т.е. выборочная функция распределения HistDog распределена по нормальному закону.

Критерий нормальности Колмогорова – Смирнова заключается в том, что он определяет, относятся ли сравниваемые два распределения к одному и тому же типу. Если сравнивать экспериментально полученное распределение с нормальным распределением, то с помощью критерия можно получить ответ о том, нормально ли наше распределение.



$F_n(x)$ – эмпирическая функция распределения.

$F(x, 0)$ – теоретическая функция распределения.
D_n^+, D_n^+ - расстояния между функциями.

Статистика критерия выглядит следующим образом:

$$D_n = sup_{|x|<\infty}|F_n(x) - F(x, 0)|,$$

где: D_n – степень различия между функциями распределениями;
x – разряды, по которым рассчитываются разницы;
$F_n(x)$ – эмпирическая функция распределения;
$F(x, 0)$ – теоретическая функция распределения;
sup – выбор максимальной по модулю величины разницы.

Критерий является правосторонним, а это значит, что если статистика, полученная по выборке, попадает в критическую область (критические значения λ_α по таблице), которая будет справа – то гипотеза отвергается, если попали в доверительную область, которая будет слева – то гипотеза не отвергается.

Задача 4. Кластерный анализ 1

Дано: 21 текст, разбитый на 3 тематические группы и 3 тематических словаря по 60 слов в каждом.

Задача: Разделить выборку, в которой на элементах измерено не менее трех параметров, методом k-средних (методом k-медианы). Методом k-средних создать кластеры текстов. Каждый текст выбран из научной литературы и состоит из 3-х страниц. Создать словари терминов для каждой науки и найти количество терминов в тексте из каждого словаря. Указания. Например, для 3-х научных направлений создать 3 словаря терминов. В каждом направлении выбрать не менее 5 текстов. Вектор признаков каждого текста будет иметь 3 компоненты, такой вектор строится для каждого текста. Из текста выделить термины 1-го словаря и записать в 1-ю компоненту вектора признаков количество таких терминов. То же выполнить для 2-го и 3-го словарей. Если количество текстов равно n, то в результате анализа выборки будет построено n векторов (n>=15). Выполнить для них кластеризацию для 3-средних.

Решение:
Для получения векторов текстов будем использовать следующий код в среде MATLAB:

```
>> f=fopen(textFile, 'r', 'n', 'Unicode'); %открываем текст для поиска
>> while feof(f)==0 %пока не конец файла
>> s=fgetl(f); %считать весь текст в одну строку
```

```
>> end
>> fclose(f); %закрываем файл
>> N1 = 0; %обнуляем первый элемент будущего вектора
>> f=fopen(firstVocabFile, 'r', 'n', 'Unicode'); %открываем словарь для поиска
>> ss = textscan(f, '%s'); %преобразуем словарь в
>> ss = ss{1,1}; %массив слов
>> fclose(f); %закрываем файл
>> %проходимся по всему тексту, находим входжения каждого слова из словаря и суммируем их количество
>> for i = 1:length(ss) N1 = N1 + length(strfind(lower(s), lower(ss(i,1)))); end;
>> N2 = 0; %обнуляем второй элемент будущего вектора
>> f=fopen(secondVocabFile, 'r', 'n', 'Unicode'); %открываем словарь для поиска
>> ss = textscan(f, '%s'); %преобразуем словарь в
>> ss = ss{1,1}; %массив слов
>> fclose(f); %закрываем файл
>> %проходимся по всему тексту, находим входжения каждого слова из словаря и суммируем их количество
>> for i = 1:length(ss) N2 = N2 + length(strfind(lower(s), lower(ss(i,1)))); end
>> N3 = 0; %обнуляем третий элемент будущего вектора
>> f=fopen(thirdVocabFile, 'r', 'n', 'Unicode'); %открываем словарь для поиска
>> ss = textscan(f, '%s'); %преобразуем словарь в
>> ss = ss{1,1}; %массив слов
>> fclose(f); %закрываем файл
>> %проходимся по всему тексту, находим входжения каждого слова из словаря и суммируем их количество
>> for i = 1:length(ss) N3 = N3 + length(strfind(lower(s), lower(ss(i,1)))); end;
>> Vector = [N1; N2; N3]; %сформировали вектор для текста
```

Для файла food1.txt получаем вектор Vfood1 = [342;34;24]. Проделаем тоже самое для оставшихся файлов и получим 21 вектор:

Vfood2 = [296;8;15]
Vfood3 = [352;10;34]
Vfood4 = [260;26;26]
Vfood5 = [476;15;73]
Vfood6 = [448;21;47]
Vfood7 = [136;7;61]

Vmashines1 = [37;390;50]
Vmashines2 = [37;498;164]
Vmashines3 = [37;162;82]

Vmashines4 = [22;357;68]
Vmashines5 = [51;201;77]
Vmashines6 = [16;261;12]
Vmashines7 = [29;250;73]

Vwater1 = [21;103;247]
Vwater2 = [126;94;254]
Vwater3 = [32;141;223]
Vwater4 = [24;73;509]
Vwater5 = [74;103;300]
Vwater6 = [185;26;406]
Vwater7 = [54;153;182]

Преобразуем 21 вектор в матрицу размерностью 21x3 и разобьём на 3 класса по методу среднего:

Матрица Matrix:

342	34	24
296	8	15
352	10	34
260	26	26
476	15	73
448	21	47
136	7	61
37	390	50
37	498	164
37	162	82
22	357	68
51	201	77
16	261	12
29	250	73
21	103	247
126	94	254
32	141	223
24	73	509
74	103	300
185	26	406
54	153	182

Выполним кластеризацию k-средних для разделения наблюдений матрицы Matrix n-на-p данных на 3 кластера и в итоге получим вектор 21x1 (cidx2), содержащий кластерные индексы каждого наблюдения и матрицу cmeans2, содержащую положение центров кластеров в матрице. Строки Matrix соответствуют точкам, а столбцы соответствуют переменным.

```
>> load fisheriris
>> [cidx2, cmeans2] = kmeans(Matrix, 3); %разделяем на 3 класса
>> ptsymb = {'bs','r^','md','go','c+'}; %условные обозначения
>> for i = 1:3 clust = find(cidx2 == i);%найдем индекс элемента массива, равного номеру кластера
>> % вывод трехмерного графика
>> plot3(Matrix(clust, 1), Matrix(clust, 2), Matrix(clust, 3), ptsymb{i}); %рисуем точки
>> hold on %держать график активным
>> scatter3(cmeans2(i,1), cmeans2(i,2), cmeans2(i,3), 100, 'ko', 'filled'); %рисуем центра в виде круга
end
>> xlabel('Food vocabulary'); ylabel('Mashines vocabulary'); zlabel('Water communic. vocabulary'); %надписи на шкалах
>> grid on %отобразим сетку
```

В итоге, после кластеризации, получаем следующую 3D размерность: