Институт Транспорта и Связи

Заочное отделение

ЛАБОРАТОРНАЯ РАБОТА №2

по дисциплине

“МЕТОДЫ КОМПЬЮТЕРНОЙ ОБРАБОТКИ СТАТИСТИЧЕСКИХ ДАННЫХ”

Тема: “Проверка гипотезы о виде закона распределения”

Вариант №6

Выполнил: ст. Козлов С. А.

ст. код. 34524

24 июня 2002 г.

Проверил: доцент Люмкис В. Д.

Рига 2002

1. *В некоторой переменной промоделировать 50 нормально распределенных чисел с математическим ожиданием μ = 3 и σ = 3. Построить гистограмму. Для другой переменной промоделировать 50 случайных величин, имеющих экспоненциальное распределение с параметром λ = 1/4. Построить гистограмму.*

Для генерации нормально распределенных чисел N(3,3) использовалась следующая формула:

```
VAR1=((rnd(1)+rnd(1)+rnd(1)+rnd(1)+rnd(1)+rnd(1))-3)*sqrt(2)*3+3
```

Для генерации экспоненциально распределенных чисел использовалась следующая формула:

```
VAR2=-4*log(1-rnd(1))
```

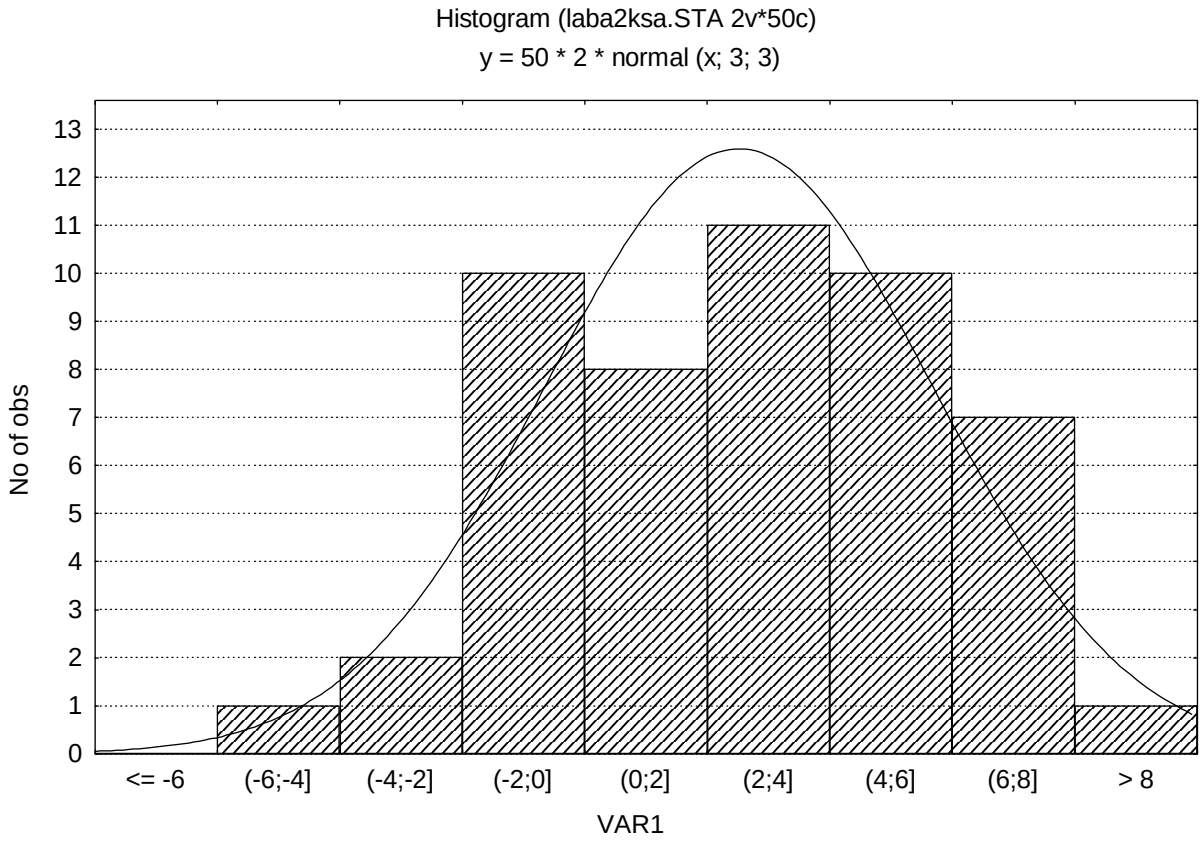
Далее в таблице приведены полученные значения(VAR1 – нормально распределенная последовательность, VAR2 - экспоненциально распределенная последовательность):

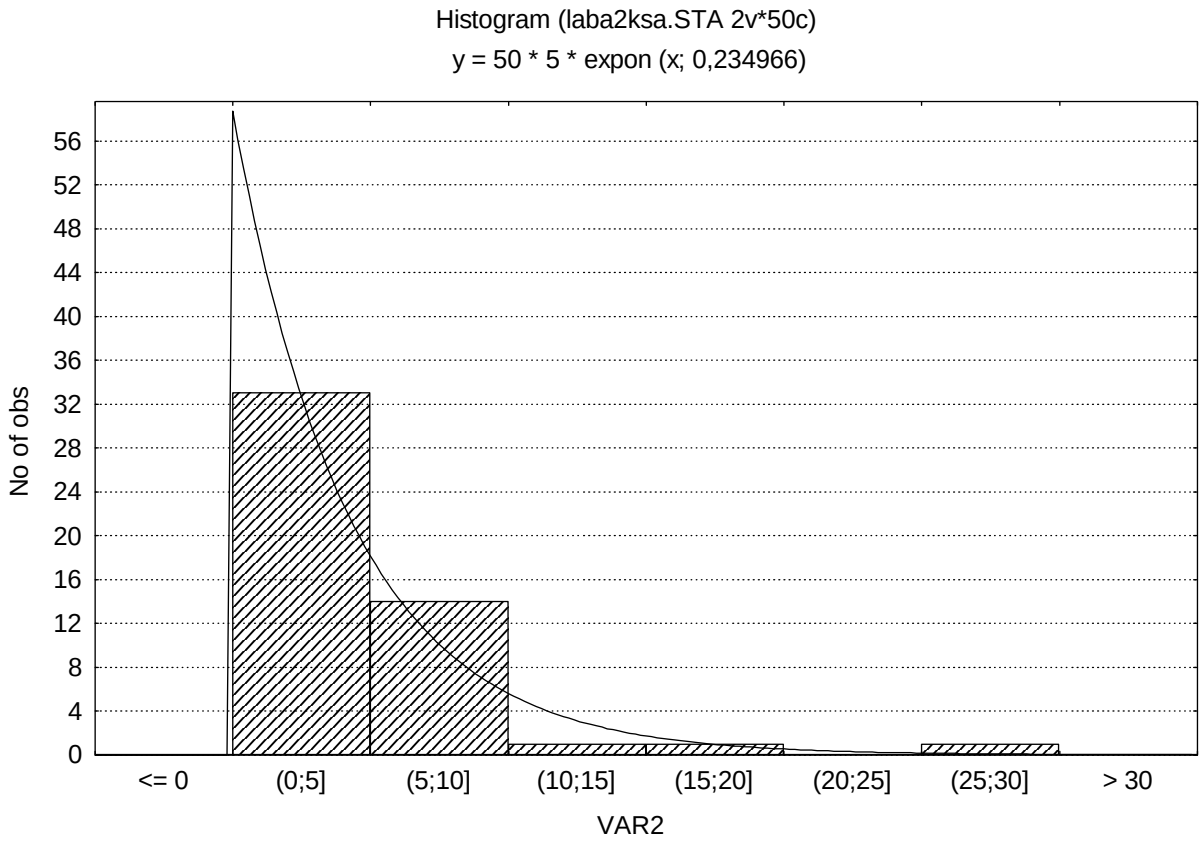
	VAR1	VAR2
1	3,387	6,358
2	3,881	7,297
3	-,987	2,855
4	,370	,309
5	4,184	,949
6	-5,042	,487
7	3,238	3,245
8	8,332	28,876
9	4,079	,932
10	3,902	11,722
11	-3,235	,848
12	5,496	9,728
13	2,995	1,870
14	4,596	4,237
15	,924	4,989
16	-,557	4,396
17	6,106	2,169
18	3,474	6,192

	VAR1	VAR2
19	-,346	9,640
20	4,771	,209
21	1,397	,042
22	-1,739	5,676
23	4,568	6,943
24	4,475	4,084
25	1,926	7,214
26	-2,839	5,853
27	-1,961	3,261
28	1,496	2,623
29	,044	,751
30	5,166	1,012
31	4,635	4,570
32	7,679	,058
33	-,940	5,568
34	7,188	1,226
35	2,380	1,151
36	-,146	,584

	VAR1	VAR2
37	3,518	,890
38	7,403	3,522
39	4,798	15,200
40	,019	6,617
41	6,309	6,843
42	6,012	,869
43	-,248	3,530
44	3,843	,562
45	-1,949	,074
46	6,770	1,235
47	-,334	8,290
48	3,569	1,226
49	2,987	,641
50	,113	5,376

Далее приведены гистограммы для каждой переменной:





■

2. *Проверить гипотезу о распределении генеральной совокупности, используя критерии Колмогорова-Смирнова и хи-квадрат. Предположить, что совокупность распределена по одному из 3 непрерывных законов (нормальному, экспоненциальному, хи-квадрат). Проверки произвести при уровне значимости 0.05.*

Критерий Колмогорова-Смирнова основан на статистике:

$$D_n = \max_{-\infty<x<+\infty} |F_n(x) - F(x)|.$$

$F_n(x)$ - эмпирическая функция распределения, построенная по данным выборки после её упорядочения(вариационный ряд);

$F(x)$ - некоторая конкретная фиксированная функция распределения.

Область принятия гипотезы $(0, K_{1-\alpha} \cdot \sqrt{\frac{1}{n}})$, при объёме выборки больше 50 значение квантиля Колмогорова может быть вычислено по формуле:

$$K_{1-\alpha} \approx \sqrt{-\frac{\ln \alpha}{2}}.$$

Согласно задания:

α=0,05;

$K_{0,95}$=1,224;

область принятия гипотезы: (0, 0.173).

Формула статистики Пирсона (хи-квадрат):

$$\chi^2 = \sum_{i=1}^{m} \frac{(a_i - np_i)^2}{np_i}$$

n - количество значений выборки;

a_i - количество значений выборки попавших в **i**-ый интервал;

m - количество интервалов;

p_i - теоретические вероятности.

По Probability Calculator:

```
ChiI(2)=5,991465     p=,050000
ChiI(3)=7,814728     p=,050000
ChiI(4)=9,487729     p=,050000
```

Соответствующие области принятия решения представлены в таблице:

Степень свободы, df	ОПР
2	0..5,991
3	0..7,815
4	0..9,488

Далее приведены результаты обработки переменных VAR1 и VAR2 пакетом Statistica на соответствие заданным законам распределения, с последующим анализом результатов:

```
STAT.     Variable VAR1 ; distribution: Normal (laba2ksa.sta)
NONPAR    Kolmogorov-Smirnov d = ,0809302, p = n.s.
STATS     Chi-Square: 5,786040, df = 3, p = ,1225198 (df adjusted)
```

Значения как критерия Колмогорова так и хи-квадрат попадают в ОПГ, т.о. принимается основная гипотеза: совокупность VAR1 распределена по нормальному закону.

```
STAT.     Variable VAR1 ; distribution: Exponential (laba2ksa.sta)
NONPAR    Kolmogorov-Smirnov d = ,2167609, p < ,05
STATS     Chi-Square: 17,16509, df = 3, p = ,0006549 (df adjusted)
```

Оба критерия попали в критическую область – принимаем альтернативную гипотезу: закон распределения совокупности VAR1 отличается от экспоненциального.

```
STAT.     Variable VAR1 ; distribution: Chi-Square (laba2ksa.sta)
NONPAR    Kolmogorov-Smirnov d = ,2117414, p < ,05
STATS     Chi-Square: 19,28883, df = 3, p = ,0002389 (df adjusted)
```

Оба критерия попали в критическую область – принимаем альтернативную гипотезу: закон распределения совокупности VAR1 отличается от хи-квадрат.

```
STAT.     Variable VAR2 ; distribution: Normal (laba2ksa.sta)
NONPAR    Kolmogorov-Smirnov d = ,1933095, p < ,05
STATS     Chi-Square: 16,58990, df = 4, p = ,0023256 (df adjusted)
```

Оба критерия попали в критическую область – принимаем альтернативную гипотезу: закон распределения совокупности VAR2 отличается от нормального.

```
STAT.     Variable VAR2 ; distribution: Exponential (laba2ksa.sta)
NONPAR    Kolmogorov-Smirnov d = ,0493185, p = n.s.
STATS     Chi-Square: 2,604781, df = 2, p = ,2718955 (df adjusted)
```

Оба критерия попали в ОПГ – принимаем основную гипотезу: совокупность VAR2 распределена по экспоненциальному закону.

```
STAT.     Variable VAR2 ; distribution: Chi-Square (laba2ksa.sta)
NONPAR    Kolmogorov-Smirnov d = ,1894332, p < ,10
STATS     Chi-Square: 13,70202, df = 3, p = ,0033444 (df adjusted)
```

Оба критерия попали в критическую область – принимаем альтернативную гипотезу: закон распределения совокупности VAR2 отличается от хи-квадрат.

■

3. *На базе выборки из нормального распределения проверить гипотезу о нормальном распределении с помощью приближенного метода.*

Сущность приближенного метода сводится к оценке двух характеристик распределения Асимметрии (Skewness) β_1 и Эксцесса(Kurtosis) β_2.

Области принятия решения определяются формулами:

$$|\beta_1| < \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot\sigma_{\beta_1};\ \left|\beta_2+\frac{6}{n+1}\right| < \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot\sigma_{_2};$$

где:

$$\sigma_{\beta_1} \approx \sqrt{\frac{6\cdot(n-2)}{(n+1)\cdot(n+3)}};\ \sigma_{\beta_2} \approx \sqrt{\frac{24n\cdot(n-2)\cdot(n-3)}{(n+1)^2\cdot(n+3)\cdot(n+5)}}.$$

Примечание: Приведенные упрощения верны в случае большой выборки $>10^3$ и в конкретном случае представляются более чем грубыми.

По Probability Calculator квантиль нормального распределения: `Z(0;1)=1,959964 p=,0500000.`

Для n=50:

$$\sigma_{\beta_1} = 0{,}326;\ \sigma_{\beta_2} = 0{,}598;\ |\beta_1| < 0{,}638948264;\left|\beta_2+\frac{6}{n+1}\right| < 1{,}172058472;$$

```
STAT.     Descriptive Statistics (laba2ksa.sta)
BASIC
STATS

Variable      Mean     Skewness   Kurtosis
    VAR1   2,514176   -,262348   -,652543
```

в таком случае: $\left|\beta_2+\frac{6}{n+1}\right| = 0{,}535$ и все полученные значения попадают в ОПР и, следовательно, принимается основная гипотеза о нормальности распределения исследуемой совокупности.

■

4. *а) Ввести данные об оборотах фирмы до и после проведения рекламной компании. Данные ввести путем моделирования нормальной выборки до проведения рекламной компании с параметром* ***μ1*** *и моделирования выборки после компании с* ***μ2*** *=* ***μ1 + d,*** *(d > 0, например,* ***μ1 = 5, d = 0.8*** *при* ***σ = 1***, ***n*** *- объем выборки и* ***n = 50****). Провести исследование эффективности рекламной компании по данным выборок. Дать пояснение смысла проверки гипотез о параметрах. Реализовать проверку гипотезы* ***H: μ1 = μ2*** *против альтернативы* ***K: μ1 ≠ μ1*** *на базе распределения Стьюдента.*

Для генерации нормально распределенных чисел использовались следующие формулы:

Выборка до рекламной компании:

```
VAR3=((Rnd(1)+Rnd(1)+Rnd(1)+Rnd(1)+Rnd(1)+Rnd(1))-3)*SQRT(2)*1+5
```

Выборка после рекламной компании:

```
VAR4=((Rnd(1)+Rnd(1)+Rnd(1)+Rnd(1)+Rnd(1)+Rnd(1))-
3)*SQRT(2)*1+5+0,8
```

Далее в таблице приведены полученные значения:

	VAR3	VAR4
1	4,027	3,382
2	5,452	5,473
3	4,814	4,973
4	7,005	4,340
5	4,695	4,651
6	5,008	8,777
7	5,694	4,354
8	4,404	6,875
9	3,969	6,303
10	4,690	5,902
11	6,204	8,462
12	4,806	4,978
13	3,258	6,486
14	5,171	6,379
15	5,258	5,937
16	4,313	4,882
17	6,196	4,758
18	4,504	6,366

	VAR3	VAR4
19	6,266	3,508
20	4,022	6,617
21	5,238	4,772
22	4,609	4,144
23	4,611	4,103
24	4,672	4,848
25	5,377	6,003
26	5,382	6,078
27	5,187	5,118
28	5,878	5,603
29	5,801	5,720
30	4,820	7,464
31	4,926	6,932
32	4,312	6,604
33	4,514	5,565
34	5,147	6,267
35	5,390	7,642
36	6,369	6,221

	VAR3	VAR4
37	3,324	7,166
38	5,165	5,891
39	4,486	5,212
40	7,262	4,305
41	3,278	6,689
42	4,256	5,581
43	2,697	7,053
44	4,390	6,198
45	6,109	5,299
46	3,926	6,729
47	3,990	6,018
48	5,426	5,881
49	4,726	5,840
50	4,031	6,808

Формула критерия Стьюдента (t-критерия) следующая:

$$\hat{t}(n_1+n_2-2)=\frac{\bar{x}_1(n_1)-\bar{x}_2(n_2)}{\tilde{s}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}};$$

где $\bar{x}_1(n_1)$ и $\bar{x}_2(n_2)$ - выборочные средние первой и второй выборки, $\tilde{s}^2$ - оценка дисперсии, составленная из оценок дисперсии для каждой группы данных:

$$\tilde{s}^2=\frac{1}{n_1+n_2-2}[(n_1-1)\tilde{s}_1^2(n_1)+(n_2-1)\tilde{s}_2^2(n_2)];$$

$$\tilde{s}_j^2(n)=\frac{1}{n_j-1}\sum_{i=1}^{n}(x_i-x_j(n))^2,\forall j=1,2.$$

Далее представлен результат обработки совокупностей VAR3 и VAR4 пакетом Statistica:

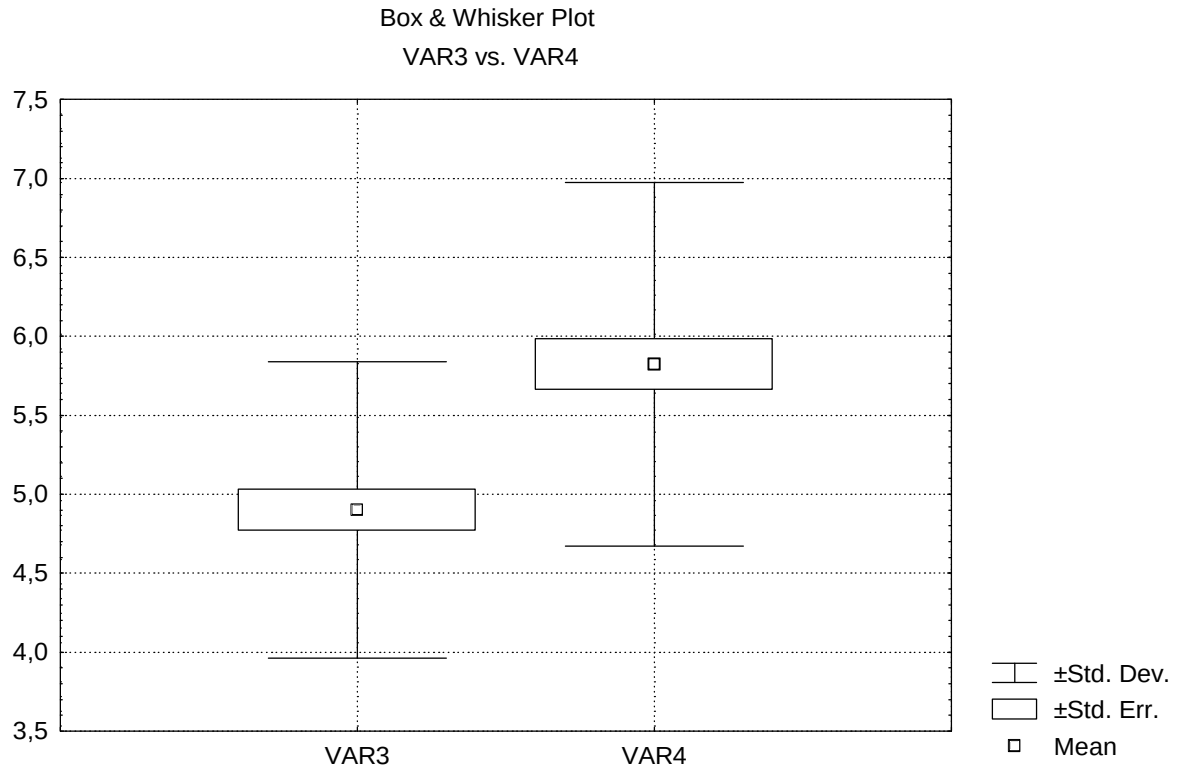
```
STAT.                    T-test for Independent Samples (laba2ksa.sta)
BASIC                    Note: Variables were treated as independent samples
STATS
                            Mean       Mean
Group 1  vs. Group 2     Group 1    Group 2    t-value      df          p
  VAR3   vs.     VAR4   4,901104*  5,823157*  -4,38551*     98*   ,000029*
```

По Probability Calculator t(98)=1,984467 p=,050000.

Т.о. значение критерия попадает в критическую область и, следовательно, принимаем альтернативную гипотезу о неравенстве матожиданий сравниваемых совокупностей. Следовательно проведенная рекламная компания дала некоторый эффект.

■

b) *В ходе проверки гипотезы реализовать визуализацию средних с помощью графиков типа BOX-WHISKERS.*



Приведённая диаграмма наглядно демонстрирует выводы сделанные в предыдущем подпункте: заметна разница в среднем значении анализируемых совокупностей, превышающая ошибку среднего.

■

24 июня 2002 г. /Козлов С. А./