Маголего «Анализ данных»

Гаврилина Александра,
магистратура «Суперкомпьютерное моделирование в науке и инженерии», группа МСКМ191
aagavrilina@edu.hse.ru
gavrilinasanya@gmail.com

1. Откройте файл experim.dta. Подберите подходящий тест для ответа на вопросы ниже. Обоснуйте свой выбор, сформулируйте гипотезы и выводы.

1.1. Кто испытывает большие опасения по отношению к статистике в момент времени 1 (fear of statistics at time 1), мужчины или женщины?

Гипотеза H0: средние значения переменной fost1 для мужчин и женщин не отличаются
Гипотеза H1: средние значения переменной fost1 для мужчин и женщин отличаются

Сравним средние значения переменной fost1 в двух группах (мужчины/женщины). Анализируемая переменная fost1 интервальная.

Проверка переменной fost1 на нормальность:

swilk fost1

```
. swilk fost1

                    Shapiro-Wilk W test for normal data

    Variable |        Obs        W         V         z       Prob>z
-------------+-------------------------------------------------------
       fost1 |         30    0.97841     0.686    -0.779    0.78189
```

Значимость p-value = 0.78189 > 0.05, значит, не значимо отличается от нормального распределения.

В выборе 30 наблюдений. Наблюдения не влияют друг на друга.

Применим t-test для независимых выборок, выполнив команду:

ttest fost1, by(sex)

Результат:

```
. ttest fost1, by(sex)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |      15        41.2    1.467749    5.684566    38.05199    44.34801
  female |      15    39.13333    1.170538    4.533473    36.62278    41.64389
---------+--------------------------------------------------------------------
combined |      30    40.16667    .9420976    5.160081    38.23986    42.09347
---------+--------------------------------------------------------------------
    diff |            2.066667     1.87735               -1.778911    5.912245
------------------------------------------------------------------------------
    diff = mean(male) - mean(female)                              t =   1.1008
Ho: diff = 0                                     degrees of freedom =       28

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.8598         Pr(|T| > |t|) = 0.2803          Pr(T > t) = 0.1402
```

P(|T| > |t|) = 0.2803 > 0.05 – не можем принять гипотезу H1.

Вывод: не принимаем гипотезу H1.

1.2. Произошли ли изменения в ощущении уверенности студентами по отношению к статистике после внешнего воздействия, произошедшего с момента времени 1 до момента времени 2 (confidence time1 (confid1) и confidence time2 (confid2))?

Гипотеза H0: изменений в ощущении уверенности не произошло
Гипотеза H1: изменения в ощущении уверенности произошли

Переменная интервальная, выборка 30.
Тест на нормальность для выборки меньше 100 (тест Шапиро-Уилка):
swilk confid1
swilk confid2

```
. swilk confid1

                    Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+-------------------------------------------------------
     confid1 |         30     0.95259       1.507     0.848    0.19824

. swilk confid2

                    Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+-------------------------------------------------------
     confid2 |         30     0.97317       0.853    -0.329    0.62900
```

В обоих случаях значимость p-value > 0.05, следовательно, распределение переменных не значимо отличается от нормального.

Используем t-test для двух связанных переменных по одной и той же выборке:

ttest confid1 == confid2

```
. ttest confid1 == confid2

Paired t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 confid1 |      30          19    .980265    5.369133    16.99513    21.00487
 confid2 |      30    21.86667    1.021306    5.593921    19.77786    23.95547
---------+--------------------------------------------------------------------
    diff |      30   -2.866667    .8679919    4.754188   -4.641909   -1.091424
------------------------------------------------------------------------------
     mean(diff) = mean(confid1 - confid2)                          t =  -3.3026
Ho: mean(diff) = 0                               degrees of freedom =       29

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0013         Pr(|T| > |t|) = 0.0025          Pr(T > t) = 0.9987
```

P(|T| > |t|) = 0.0025 < 0.05 – принимаем гипотезу H1.

Вывод: принимаем гипотезу H1 (изменения в ощущении уверенности произошли).

1.3. Изменялся ли уровень депрессии в разные моменты измерения?

Переменные depress1, depress2, depress3.

Попарно сравним уровни депрессии:

ttest depress1 == depress2

Гипотеза H0: изменений нет
Гипотеза H1: изменения есть

```
. ttest depress1 == depress2

Paired t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
depress1	30	42.53333	.8383527	4.591847	40.81871	44.24796
depress2	30	40.73333	1.007938	5.520703	38.67187	42.7948
diff	30	1.8	.4558685	2.496895	.8676442	2.732356

```
     mean(diff) = mean(depress1 - depress2)                      t =   3.9485
 Ho: mean(diff) = 0                               degrees of freedom =       29

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9998         Pr(|T| > |t|) = 0.0005          Pr(T > t) = 0.0002
```

Pr(|T|) > |t|) = 0.0005 < 0.05 – принимаем гипотезу H1 о различном уровне депрессии.

ttest depress2 == depress3

Гипотеза H0: изменений нет
Гипотеза H1: изменения есть

```
. ttest depress2 == depress3

Paired t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
depress2	30	40.73333	1.007938	5.520703	38.67187	42.7948
depress3	30	39.2	.9289717	5.088188	37.30004	41.09996
diff	30	1.533333	.3415089	1.870521	.8348691	2.231798

```
     mean(diff) = mean(depress2 - depress3)                      t =   4.4899
 Ho: mean(diff) = 0                               degrees of freedom =       29

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9999         Pr(|T| > |t|) = 0.0001          Pr(T > t) = 0.0001
```

Pr(|T|) > |t|) = 0.0001 < 0.05 – принимаем гипотезу H1 о различном уровне депрессии.

ttest depress1 == depress3

Гипотеза H0: изменений нет
Гипотеза H1: изменения есть

```
. ttest depress1 == depress3

Paired t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
depress1	30	42.53333	.8383527	4.591847	40.81871	44.24796
depress3	30	39.2	.9289717	5.088188	37.30004	41.09996
diff	30	3.333333	.4632056	2.537081	2.385972	4.280695

```
     mean(diff) = mean(depress1 - depress3)                      t =   7.1962
 Ho: mean(diff) = 0                               degrees of freedom =       29

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Pr(|T|) > |t|) = 0.0000 < 0.05 – принимаем гипотезу H1 о различном уровне депрессии.

2. Откройте файл reading.dta. Проведите однофакторный дисперсионный анализ с использованием следующих переменных:

Зависимая переменная: score
Независимая переменная: class

Гипотеза H0: скорость чтения у учеников разных классов одинаковая
Гипотеза H1: скорость чтения у учеников разных классов разная

Однофакторный дисперсионный анализ (One-Way ANOVA):
oneway score class, scheffe

```
. oneway score class, scheffe

                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups         3458.2        4       864.55       3.25     0.0125
 Within groups      78544.7167      295    266.253277
------------------------------------------------------------------------
    Total           82002.9167      299    274.257246

Bartlett's test for equal variances:  chi2(4) =   0.8657  Prob>chi2 = 0.929
```

Prob > F: 0.0125 < 0.05, значит, имеются различия. Отвергаем гипотезу H0, принимаем H1.

Проведите апостериорный тест. Сформулируйте гипотезы. Интерпретируйте результаты анализа.

Comparison of reading score by class nested in program
(Scheffe)

Row Mean- Col Mean	1	2	3	4
2	-3.41667 0.859			
3	-2.66667 0.938	.75 1.000		
4	.1 1.000	3.51667 0.845	2.76667 0.930	
5	6.23333 0.359	9.65 0.035	8.9 0.066	6.13333 0.377

Вывод: различие в скорости чтения наблюдается для 2 и 5 классов – значимость менее 0.05 (0.035).

3. Откройте файл survey.dta. Подберите подходящий статистический тест для ответа на вопросы ниже. Обоснуйте свой выбор теста, сформулируйте гипотезы и выводы.

3.1. Оцените отличаются ли значения переменной «Total of perceived stress (tpstress)» в разных возрастных группах, представленных переменной «age 5 groups (agegp5)».

Гипотеза H0: средние значения переменной tpstress не отличаются в разных возрастных группах
Гипотеза H1: средние значения переменной tpstress отличаются в разных возрастных группах

Для сравнения средних величин в трех и более группах (в данном случае 5) используем однофакторный дисперсионный анализ (One-Way ANOVA).

Выполним команду:
oneway tpstress agegp5, scheffe

Результат:

```
. oneway tpstress agegp5, scheffe

                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      500.761358      4    125.19034       3.75     0.0051
 Within groups      14271.0816    428   33.3436486
------------------------------------------------------------------------
    Total            14771.843    432   34.1940809

Bartlett's test for equal variances:  chi2(4) =   5.3065  Prob>chi2 = 0.257
```

Prob > F: 0.0051 < 0.05. Значит, есть различия: отвергаем гипотезу H0, принимаем гипотезу H1.

Апостериорный тест:

Comparison of Total perceived stress by age 5 groups
(Scheffe)

Row Mean-Col Mean	18 - 24	25 - 32	33 - 40	41 - 49
25 - 32	-2.95099 0.021			
33 - 40	-1.83386 0.357	1.11713 0.814		
41 - 49	-1.9811 0.239	.96989 0.866	-.14724 1.000	
50+	-2.8489 0.038	.102084 1.000	-1.01505 0.873	-.867806 0.916

Вывод: различия есть между группами (значимость менее 0.05):

- 18-24 и 50+ (0.038);
- 18-24 и 25-32 (0.021).

3.2. Испытывают ли курильщики больший стресс, чем некурящие люди? Переменные: smoke и Total perceived stress (tpstress).

Гипотеза H0: курильщики не испытывают больший стресс, чем некурящие люди
Гипотеза H1: курильщики испытывают больший стресс, чем некурящие люди

Первое, что пришло в голову – провести t-test аналогично заданию 1.1, для этого нужно проверить, имеются ли какие-то ограничения на проведение теста.

Переменная tpstress должна быть интервальной – это выполняется.
В выборе больше 30 наблюдений – выполняется.
Наблюдения не влияют друг на друга.
Распределение значений не должно значимо отличаться от нормального распределения, проверим это. Размер выборки больше 100. Проверим переменную tpstress на нормальность с помощью теста Колмогорова-Смирнова. Создадим вспомогательные переменные tpstress_mu и tpstress_s:

```
. egen tpstress_mu = mean(tpstress)

. egen tpstress_s = sd(tpstress)
```

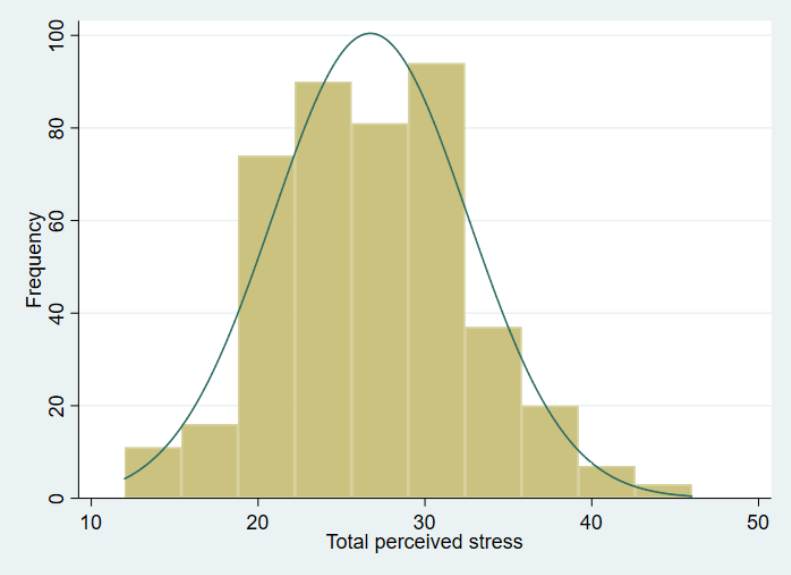
```
. ksmirnov tpstress = normprob((tpstress-tpstress_mu)/tpstress_s)

One-sample Kolmogorov-Smirnov test against theoretical distribution
           normprob((tpstress-tpstress_mu)/tpstress_s)

 Smaller group       D       P-value
 ---------------------------------------
 tpstress:           0.0691    0.016
 Cumulative:        -0.0326    0.398
 Combined K-S:       0.0691    0.032

Note: Ties exist in dataset;
      there are 34 unique values out of 433 observations.
```

Значимо не отличается от нормального. Можно также построить график и убедиться:



Однако число наблюдений в сравниваемых подгруппах (курильщики/некурящие) не является равнозначным (курильщиков в 4 раза меньше). Поэтому t-test использовать не можем. В таком случае рекомендовано использование аналогичных непараметрических тестов, например, Wilcoxon Rank sum Test.

Выполним команду:

ranksum tpstress, by(smoke) porder

Результат:

```
. ranksum tpstress, by(smoke) porder

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      smoke |       obs    rank sum    expected
------------+---------------------------------
        YES |        84       18704       18102
         NO |       346       73961       74563
------------+---------------------------------
   combined |       430       92665       92665

unadjusted variance  1043882.00
adjustment for ties    -3116.18
                     ----------
adjusted variance    1040765.82

Ho: tpstress(smoke==YES) = tpstress(smoke==NO)
             z =   0.590
    Prob > |z| =   0.5551

P{tpstress(smoke==YES) > tpstress(smoke==NO)} = 0.521
```

P(|T|) > |t|) = 0.521 > 0.05 – не принимаем гипотезу H1, принимаем H0.

Вывод:
Принимаем гипотезу H0 о том, что курильщики не испытывают больший стресс.

3.3. Наблюдаются ли различные результаты прохождения теста по самооценке в трёх разных возрастных группах? Переменные: Total self esteem (tslfest) и agegp3.

Гипотеза H0: различий результатов прохождения теста в разных возрастных группах нет
Гипотеза H1: различия результатов прохождения теста в разных возрастных группах есть

Для сравнения величин в трех и более группах применяется однофакторный дисперсионный анализ (One-Way ANOVA).

Выполним команду:
oneway tslfest agegp3, scheffe

Результат:

```
. oneway tslfest agegp3, scheffe

                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      258.075226      2   129.037613      4.51     0.0116
 Within groups      12402.4752    433   28.6431299
------------------------------------------------------------------------
    Total           12660.5505    435   29.1047137

Bartlett's test for equal variances:  chi2(2) =   0.9962  Prob>chi2 = 0.608
```

Prob > F: 0.0116 < 0.05, значит, имеются различия в группах. Отвергаем гипотезу H0, принимаем H1. Посмотрим, в каких группах имеются отличия:

```
                Comparison of Total Self esteem by age 3 groups
                                (Scheffe)
Row Mean-
Col Mean    |     18 - 29     30 - 44
------------+----------------------
 30 - 44    |     .988211
            |       0.278

      45+   |     1.90639     .918177
            |       0.012       0.350
```

<u>Вывод:</u> различие наблюдается между группами 18-29 и 45+ (значимость 0.012 < 0.05).