multivariate_t4

December 12, 2024

1 Multivariate Statistics Test 4

Student: Aleksandr Jan Smoliakov, VU MIF Data Science MSc year 1
Date: 2024-12-12

Note: We are assuming 0.05 significance level for all tests in this task.

[1]:
```python
import matplotlib.pyplot as plt
import pyreadstat
import pandas as pd
import pingouin as pg
import statsmodels.formula.api as smf
import statsmodels.multivariate.manova as manova
import statsmodels.stats.anova as anova
import statsmodels.stats.multicomp as multicomp
from scipy.stats import levene

pd.options.display.float_format = "{:.4f}".format
```

1.1 Task 1: Science and Math

Data: File `scmath.sav`, variables

- `group` - school's prestige (1-high, 3-low)
- `math` - mean school's Math score
- `science` - mean school's Science score

First of all, let's load the data and take a look.

[2]:
```python
df_scmath, metadata_scmath = pyreadstat.read_sav("data/scmath.sav")

df_scmath.describe()
```

[2]:

	group	math	science
count	74.0000	74.0000	74.0000
mean	2.1081	13.3784	19.8198
std	0.8204	4.4566	9.5067
min	1.0000	6.6667	3.3333
25%	1.0000	10.0000	13.3333
50%	2.0000	13.3333	20.0000

```
75%     3.0000 16.6667  26.6667
max     3.0000 26.6667  36.6667
```

1.1.1 Perform ANOVA for Science scores for all groups

We will fit a linear model with `group` as a factor and `science` as a dependent variable to test if the group has a significant effect on the science score.

- Null hypothesis: the group has no significant effect on the science score.
- Alternative hypothesis: the group has a significant effect on the science score.

[3]:
```
anova_model = smf.ols("science ~ C(group)", data=df_scmath).fit()
anova_results = anova.anova_lm(anova_model, typ=2)

anova_results
```

[3]:
```
             sum_sq      df      F  PR(>F)
C(group)   336.1923  2.0000 1.9061  0.1562
Residual  6261.4053 71.0000    NaN     NaN
```

The F-statistic of the group is 1.91, and the p-value is 0.156. Since the p-value is greater than 0.05, we fail to reject the null hypothesis.

1.1.2 Levene tests for equality of science variances in all three samples

Null hypothesis: the variances of the science scores in all three groups are equal.
Alternative hypothesis: the variances of the science scores in all three groups are not equal.

[4]:
```
levene_results = levene(
    df_scmath.loc[df_scmath["group"] == 1, "science"],
    df_scmath.loc[df_scmath["group"] == 2, "science"],
    df_scmath.loc[df_scmath["group"] == 3, "science"],
)

print("Levene test p-value:", levene_results.pvalue)
```

```
Levene test p-value: 0.11552903683721506
```

The Levene test shows that the p-value is 0.116, which means that we cannot reject the null hypothesis that the variances are equal.

We can assume that the variances of the science scores in all three groups are equal.

1.1.3 Perform ANCOVA for Science scores controlling for Math scores

Null hypothesis: the group has no significant effect on the science score after controlling for the math score.
Alternative hypothesis: the group has a significant effect on the science score after controlling for the math score.

[5]:
```
ancova_model = smf.ols("science ~ C(group) + math", data=df_scmath).fit()
ancova_results = anova.anova_lm(ancova_model, typ=2)

ancova_results
```

[5]:

	sum_sq	df	F	PR(>F)
C(group)	924.5442	2.0000	17.8991	0.0000
math	4453.5482	1.0000	172.4408	0.0000
Residual	1807.8571	70.0000	NaN	NaN

With math as a covariate:

- the p-values for the group and the math are both under 0.0001
- which means that both variables have a significant effect on the science score, and the null hypothesis is rejected

1.1.4 Post-hoc Tukey test for ANCOVA model

Sadly, Python doesn't seem to have a built-in function for Tukey's post-hoc test for ANCOVA models.

Instead, we're going to remove the effect of the math score from the science score and then perform the Tukey test on the residuals.

Null hypothesis: the means of the groups are equal.
Alternative hypothesis: the means of the groups are not equal.

[ ]:
```
tukey_results = multicomp.pairwise_tukeyhsd(
    df_scmath["science"] - ancova_model.params["math"] * df_scmath["math"],
    df_scmath["group"],
)

print(tukey_results)
```

```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower    upper  reject
-----------------------------------------------------
   1.0    2.0   -1.359 0.6413  -4.9684  2.2504  False
   1.0    3.0  -8.0716    0.0 -11.5328 -4.6104   True
   2.0    3.0  -6.7126    0.0  -10.046 -3.3793   True
-----------------------------------------------------
```

We can see the following results:

- The p-value for groups 1 (prestigious) vs 2 (very prestigious) is 0.641, which means we fail to reject the null hypothesis that the means are equal when controlling for the math score.
- The other two p-values (prestigious / v. prestigions vs not prestigious) are <0.0001, which means we reject the null hypothesis that the means are equal when controlling for the math score.

1.2 Task 2: Preferred time-spending

Data: File `Activity.sav`, variables

- `family` - preferred time-spending with family
- `social` - preferred time-spending with friends
- `work` - preferred time-spending with co-workers

First of all, let's load the data and take a look.

[6]:
```
df_activity, metadata_activity = pyreadstat.read_sav("data/Activity.sav")

df_activity.describe()
```

[6]:

	family	social	work
count	66.0000	66.0000	66.0000
mean	15.5758	15.4545	13.2424
std	4.1103	3.7670	3.6922
min	4.0000	7.0000	4.0000
25%	13.2500	13.0000	11.2500
50%	16.0000	15.5000	13.0000
75%	18.7500	18.0000	16.0000
max	25.0000	26.0000	20.0000

1.2.1 Data preparation in correct format

We will convert the data to the long format, where the columns will be transformed into separate rows.

[7]:
```
df_activity["ID"] = df_activity.index

df_activity_long = pd.melt(
    df_activity,
    id_vars=["ID"],
    value_vars=["family", "social", "work"],
)

df_activity_long
```

[7]:

	ID	variable	value
0	0	family	19.0000
1	1	family	17.0000
2	2	family	8.0000
3	3	family	13.0000
4	4	family	14.0000
..	..	…	…
193	61	work	18.0000
194	62	work	12.0000
195	63	work	16.0000

	F Value	Num DF	Den DF	Pr > F
variable	8.0916	2.0000	130.0000	0.0005

```
196  64     work 13.0000
197  65     work 10.0000

[198 rows x 3 columns]
```

1.2.2 Test Sphericity assumption

We will perform Mauchly's test of sphericity to test if the data is spherically distributed.

Null hypothesis: the data is spherically distributed.
Alternative hypothesis: the data is not spherically distributed.

[8]:
```python
mauchly_test = pg.sphericity(
    df_activity_long,
    dv="value",
    within="variable",
    subject="ID",
)

print("P-value of Mauchly's test:", mauchly_test.pval)
```

```
P-value of Mauchly's test: 0.9598403034007936
```

The Mauchly's test shows that the p-value is 0.960, which means that we cannot reject the null hypothesis that the data is spherically distributed.

We can proceed with the repeated measures ANOVA.

1.2.3 Test statistical significance

We will perform the repeated measures ANOVA to check if there are any significant differences between the three preferred time-spending types.

Null hypothesis: there are no significant differences between preference for family, social, and work time-spending types.
Alternative hypothesis: there are significant differences between preference for family, social, and work time-spending types.

[9]:
```python
anova_results = anova.AnovaRM(
    df_activity_long,
    depvar="value",
    subject="ID",
    within=["variable"],
).fit()

anova_results.summary()
```

[9]:

The p-value is 0.0005, which means that we reject the null hypothesis and conclude that there are significant differences between family, social, and work time-spending types.

1.2.4 Post hoc tests

We'll run Tukey's post hoc test to determine which pairs of variables have significantly different means.

Null hypothesis: the means of the variables are equal.
Alternative hypothesis: the means of the variables are not equal.

[10]:
```
tukey_results = multicomp.pairwise_tukeyhsd(
    df_activity_long["value"],
    df_activity_long["variable"],
)

print(tukey_results)
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
family social  -0.1212 0.9822 -1.7085  1.4661  False
family   work  -2.3333 0.0018 -3.9206 -0.7461   True
social   work  -2.2121 0.0034 -3.7994 -0.6248   True
----------------------------------------------------
```

The p-value for `family` vs `social` is 0.982, which means that we fail to reject the null hypothesis that their means are equal.

The p-value for `family` vs `work` is 0.002 and `social` vs `work` is 0.003, which means that we reject the null hypothesis that the means are equal.

1.3 Task 3: Training and Test scores

Data: File `ABk.sav`, variables

- `T` - hours trained before the test
- `school` - school location (1=small town, 2=capital, 3=rural)
- `reading` - reading test score
- `math` - math test score

First of all, let's load the data and take a look.

[11]:
```
df_abk, metadata_abk = pyreadstat.read_sav("data/ABk.sav")

df_abk.describe()
```

[11]:
```
            T  school  reading    math
count 75.0000 75.0000  75.0000 75.0000
```

```
mean    2.4133  2.0933  13.5556 10.3111
std     1.1751  0.8248   4.6848 10.3566
min     1.0000  1.0000   6.6667 -6.6667
25%     1.0000  1.0000  10.0000  3.3333
50%     2.0000  2.0000  13.3333 10.0000
75%     3.0000  3.0000  16.6667 16.6667
max     5.0000  3.0000  26.6667 46.6667
```

1.3.1 ANOVAs for reading and math

[12]:
```python
for var in ["reading", "math"]:
    anova_model = smf.ols(f"{var} ~ C(school)", data=df_abk).fit()
    anova_results = anova.anova_lm(anova_model, typ=2)
    # tukey_results = multicomp.pairwise_tukeyhsd(df_abk[var], df_abk["group"])

    print(f"ANOVA Results for {var}:")
    print(anova_results)
    print()
    # print(tukey_results)
    # print()
```

```
ANOVA Results for reading:
             sum_sq      df      F  PR(>F)
C(school)   56.9905  2.0000 1.3092  0.2764
Residual  1567.0836 72.0000    NaN     NaN

ANOVA Results for math:
             sum_sq      df      F  PR(>F)
C(school)  414.5726  2.0000 1.9840  0.1450
Residual  7522.6126 72.0000    NaN     NaN
```

The p-values for the categorical variable `school` and other variables are the following:

- `reading`: 0.276, i.e. not significant
- `math`: 0.145, i.e. not significant

1.3.2 Box test

Null hypothesis: the covariance matrices of the groups are equal.
Alternative hypothesis: at least one of the covariance matrices of the groups is different.

[19]:
```python
pg.box_m(
    df_abk,
    group="school",
    dvs=["reading", "math"],
)
```

[19]:
```
         Chi2      df    pval   equal_cov
box  10.7721 6.0000 0.0957        True
```

The p-value is 0.096, which means that we fail to reject the null hypothesis that the covariance matrices of the groups are equal.

We can assume homogeneity of covariances, and we can proceed with MANOVA.

1.3.3 Perform MANOVA with reading and math

Null hypothesis: the school location has no significant effect on the reading and math test scores. Alternative hypothesis: the school location has a significant effect on the reading and math test scores.

[13]:
```python
manova_model = manova.MANOVA.from_formula("reading + math ~ C(school)",
 ↪data=df_abk)
manova_results = manova_model.mv_test()

print(manova_results)
```

```
                  Multivariate linear model
============================================================

------------------------------------------------------------
       Intercept        Value  Num DF  Den DF F Value  Pr > F
------------------------------------------------------------
          Wilks' lambda 0.2135 2.0000 71.0000 130.7464 0.0000
         Pillai's trace 0.7865 2.0000 71.0000 130.7464 0.0000
 Hotelling-Lawley trace 3.6830 2.0000 71.0000 130.7464 0.0000
    Roy's greatest root 3.6830 2.0000 71.0000 130.7464 0.0000
------------------------------------------------------------

------------------------------------------------------------
       C(school)        Value  Num DF  Den DF  F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.6170 4.0000 142.0000  9.6944 0.0000
         Pillai's trace 0.3847 4.0000 144.0000  8.5743 0.0000
 Hotelling-Lawley trace 0.6179 4.0000  84.1709 10.9190 0.0000
    Roy's greatest root 0.6134 2.0000  72.0000 22.0822 0.0000
============================================================
```

The Wilks' Lambda test shows that the p-value is <0.0001, which means that we reject the null hypothesis and conclude that the school location has a significant effect on the reading and math test scores.

1.3.4 MANCOVA, controlling for T (hours trained)

We will incorporate the hours trained variable as a covariate in the MANOVA model, and run a MANCOVA.

Null hypothesis: the school location has no significant effect on the reading and math test scores after controlling for the hours trained.
Alternative hypothesis: the school location has a significant effect on the reading and math test scores after controlling for the hours trained.

[14]:
```
manova_model = manova.MANOVA.from_formula("reading + math ~ C(school) + T",␣
↪data=df_abk)
manova_results = manova_model.mv_test()

print(manova_results)
```

```
                  Multivariate linear model
============================================================

------------------------------------------------------------
        Intercept        Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
           Wilks' lambda 0.3748 2.0000 70.0000 58.3746 0.0000
          Pillai's trace 0.6252 2.0000 70.0000 58.3746 0.0000
  Hotelling-Lawley trace 1.6678 2.0000 70.0000 58.3746 0.0000
     Roy's greatest root 1.6678 2.0000 70.0000 58.3746 0.0000
------------------------------------------------------------

------------------------------------------------------------
       C(school)        Value  Num DF  Den DF  F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.8922 4.0000 140.0000  2.0542 0.0901
         Pillai's trace 0.1082 4.0000 142.0000  2.0299 0.0934
 Hotelling-Lawley trace 0.1204 4.0000  82.9711  2.0974 0.0884
    Roy's greatest root 0.1168 2.0000  71.0000  4.1465 0.0198
------------------------------------------------------------

------------------------------------------------------------
            T            Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
           Wilks' lambda 0.3595 2.0000 70.0000 62.3455 0.0000
          Pillai's trace 0.6405 2.0000 70.0000 62.3455 0.0000
  Hotelling-Lawley trace 1.7813 2.0000 70.0000 62.3455 0.0000
     Roy's greatest root 1.7813 2.0000 70.0000 62.3455 0.0000
============================================================
```

After controlling for the hours trained, the Wilks' Lambda test shows that the p-value is 0.090, which means that we fail to reject the null hypothesis that the school location has no significant effect on the reading and math test scores after controlling for the hours trained.