Research Methods

Robert E Simpson

Singapore University of Technology & Design

robert_simpson@sutd.edu.sg

October 31, 2018

You should be able to use python to:

- Test whether two vectors of sample means are statistically different using the Hotelling T^2
- Test whether a vector of sample means is statistically different from a specific vector
- Perform basic vector manipulation using the numpy linalg and matrix libraries
- Compute correlation coefficients and covariances
- Write functions to perform ANOVA
- Use the built-in scipy single factor ANOVA function

ANOVA health warning

The numpy.var function computes the *population* variance by default. Therefore:

$$SSW = \sum_{i=1}^{k} n_i s_i^2$$

$$SS_{\mathrm{Total}} = Ns^2$$

However, the Excel `Var` function computes the *sample* variance– beware!

$$SSW = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

$$SS_{\mathrm{Total}} = (N - 1)s^2$$

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Multivariate methods

Multivariate analysis, AKA multifactor analysis deals with the statistical differences between samples of multiple different parameters.

- By considering multiple factors (weight, sex, age), we are able to gain more confidence on whether samples are from a similar population.
- By computing the co-variance terms we can gain even more data to gain an even higher confidence in our hypothesis.

FROM WEEK 5: Pooled variances

For small sample sizes and the variances of both populations are similar, we have the option use a pooled variance, s_p, which is defined as:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{1}$$

with degrees of freedom (d.f$= n_1 + n_2 - 2$)

- The pooled variance is simply the weighted average of the two variances
- The use of pooled variances results in tight confidence intervals (hence its appeal)

now

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t_c \sigma(\bar{x_1}, \bar{x_2}) \tag{2}$$

where $\sigma(\bar{x_1}, \bar{x_2}) = \left[s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{1/2}$

Multivariate methods

In contrast to t-tests and single factor ANOVA, multivariate methods allow multiple measures to be analysed simultaneously. Inferences are made by considering the whole system of data and this results in more accurate inferences.

Useful python functions

numpy.linalg.inv(X) Compute the inverse matrix of X

numpy.matrix.transpose(X) Compute the transpose matrix of X

X.dot(Y) Dot product of X and Y

scipy.stats.t.pdf Calculates the student t PDF

scipy.stats.t.cdf Calculates the cumulative student t-distribution

scipy.stats.t.ppf Calculates the probability for a particular combination t-value and degrees of freedom

scipy.stats.f.ppf Calculates the probability for a particular combination of f-value and degrees of freedom

scipy.pearsonr(X,Y) Calculates the Pearson correlation coefficient between X and Y

scipy.cov(X,Y) Calculates the covariance between X and Y

Case Problem 1: t-distribution

a) Plot the student t-distribution, with d.f.=1,3,5,10, & 20 over the range $-5 < t < 5$

b) Plot the cumulative student t-distribution, with d.f.=1,3,5,10, & 20 over the range $-5 < t < 5$

c) Use the plots to find the critical t-value, t_c, for one tail and two tail tests at a significance level of $\alpha = 0.05$.

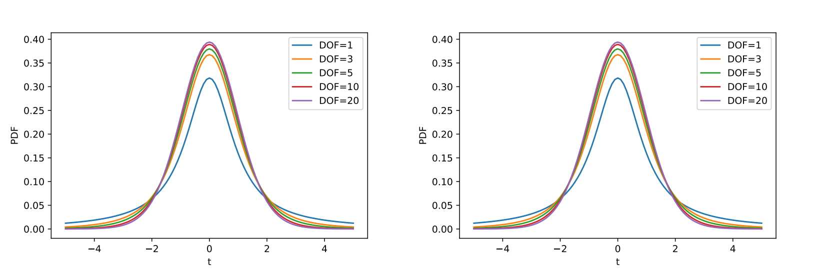
Case Problem 1 solution

```
import numpy as  np
import scipy.stats as ss
import matplotlib.pyplot as plt
from scipy.stats import f
t=np.linspace(-5,5,51)

plt.figure()
for dof in [1, 3, 5, 10, 20]:
    tPDF=ss.t.pdf(t, dof)
    plt.plot(t,tPDF)
plt.xlabel('t')
plt.ylabel('PDF')

plt.figure()
for dof in [1, 3, 5, 10, 20]:
    tCDF=ss.t.cdf(t, dof)
    plt.plot(t,tCDF)
plt.xlabel('t')
plt.ylabel('CDF')
```

Case Problem 1 solution



c) $t_c(\alpha = 0.05) = \pm 2.0871$, for two tail test

c) $t_c(\alpha = 0.05) = \pm 1.7267$, for one tail test

Case Problem 2: Paper bags

A paper manufacturer needs to develop bags for a reputable chain of grocery stores. The product development team are under the impression that the strength of the the paper increases with the weight percentage of hardwood that is included in the pulp. The table below shows measurements of the paper tensile strength at different levels of hard wood concentration in the pulp.

1. **Write a function to perform an ANOVA** to check whether the hardwood concentration influences the strength of the paper.
2. Check the output of your ANOVA function against:
   `F, p = ss.f_oneway(H5, H10, H15, H20)`.

Hardwood content (%)	Observations					
	1	2	3	4	5	6
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

Case Problem 2 solution

Identity the factor Percentage weight of hardwood, and there are four different levels

H_0: The tensile strength, τ, is equal independent of hardwood concentration, i.e. $\tau_5 = \tau_{10} = \tau_{15} = \tau_{20}$.

H_A: $\tau_i \neq \tau_j$ for at least one i, j pair

Test statistic : $f_0 = \frac{MST}{MSE}$

Significance: Reject H_0 if the P-value is less than 0.05.

Computation: Perform a single factor ANOVA

Case Problem 2 Solution

```
H5=[7,8,15,11,9,10]
H10=[12,17,13,18,19,15]
H15=[14,18,19,17,16,18]
H20=[19,25,22,23,18,20]

def SSW(*arg):
    n=len(arg[0])  #n is the number of samples in each level
    levels=len(arg)
    ssw=0
    for i in range(levels):
        var=np.var(arg[i], ddof=0)
        ssw=ssw+var
    df=(n*levels)-levels
    return (n*ssw), df

def SSB(*arg):
    grandmean=0
    levels=len(arg)
    for i in range(len(arg)):
        grandmean=grandmean+np.mean(arg[i])
    grandmean=grandmean/len(arg)
    ssb=0
    for i in range(len(arg)):
        temp=np.mean(arg[i])
        ssb=ssb+(temp-grandmean)**2
    ssb=len(arg[0])*ssb
    df=levels-1
    return ssb, df
```

Case Problem 2 Solution

```
ssw, sswdf=SSW(H5, H10, H15, H20)
ssb, ssbdf=SSB(H5, H10, H15, H20)
sst1=np.var(H5+H10+H15+H20)*len(H5+H10+H15+H20)
sst2=ssw+ssb
msw=ssw/sswdf
msb=ssb/ssbdf
F=msb/msw

Fc=f.ppf(0.95,3,20)  # use 0.95 because we are looking at the area of the RH
      tail
print "SSW=",ssw
print "MSW=", msw
print "SSB=", ssb
print "MSB=", msb
print "SSTotal=", sst1
print "F=", F
print "Critical␣F=", Fc
if F>Fc:
    print "F>Fc,␣Reject␣H0"
else:
    print "F<Fc,␣Accept␣H0"

####Using Python's built in function:
F1, p1 = ss.f_oneway(H5, H10, H15, H20)
```

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

CP2 ANOVA result

```
Grand Mean= 15.9583333333
SSE= 130.166666667
MSE= 6.50833333333
SSB= 382.791666667
F= 19.6052069996
Critical F= 3.09839121214
F>Fc, Reject H0
```

Case Problem 3: pairwise vs Hotelling T^2

Compare the mean values of two samples, X1 and X2, by a pairwise student t-test and by Hotelling T^2 methods.

Each sample is a 1-D matrix of sample means for 5 different parameters ($p = 5$). X1 contains the means of 21 observations and X2 contains the means of 28. The mean and covariance matrices of both samples have been calculated and are given below:

$$\bar{\mathbf{X}}_1 = \begin{bmatrix} 157.381 \\ 241.000 \\ 31.433 \\ 18.500 \\ 20.810 \end{bmatrix} \qquad \bar{\mathbf{X}}_2 = \begin{bmatrix} 158.429 \\ 241.571 \\ 31.479 \\ 18.446 \\ 20.839 \end{bmatrix}$$

$$\mathbf{C}_1 = \begin{bmatrix} 11.048 & 9.100 & 1.557 & 0.870 & 1.286 \\ 9.100 & 17.500 & 1.910 & 1.310 & 0.880 \\ 1.557 & 1.910 & 0.531 & 0.189 & 0.240 \\ 0.870 & 1.310 & 0.189 & 0.176 & 0.133 \\ 1.286 & 0.880 & 0.240 & 0.133 & 0.575 \end{bmatrix} \qquad \mathbf{C}_2 = \begin{bmatrix} 15.069 & 17.190 & 2.243 & 1.746 & 2.931 \\ 17.190 & 32.550 & 3.398 & 2.950 & 4.066 \\ 2.243 & 3.398 & 0.728 & 0.470 & 0.559 \\ 1.746 & 2.950 & 0.470 & 0.434 & 0.506 \\ 2.931 & 4.066 & 0.559 & 0.506 & 1.321 \end{bmatrix}$$

Case Problem 3 solution: pairwise

Each parameter is compared individually. The first parameter would be computed using pooled variance as:

$$s_1^2 = \frac{(21-1)(11.048) + (28-1)(15.069)}{21+28-2} = 13.36$$

The t-statistic would be:

$$t = \frac{(157.381 - 158.429)}{\sqrt{13.36(\frac{1}{21} + \frac{1}{28})}} = -0.99$$

For a two-tail test with df=60, the tables show that $t_c = 1.296$ at a significance level of $\alpha = 0.2$. Thus for this example, with df=47 and t=-0.99 the difference is not statistically relevant at a 20% significance level. Therefore we accept that null hypothesis and conclude that the two samples come from the same population.

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Case Problem 3 solution: pairwise

Parameter	First data set		Second data set		t-value (47 d.f.)	p-value
	Mean	Variance	Mean	Variance		
1	157.38	11.05	158.43	15.07	−0.99	0.327
2	241.00	17.50	241.57	32.55	−0.39	0.698
3	31.43	0.53	31.48	0.73	−0.20	0.842
4	18.50	0.18	18.45	0.43	0.33	0.743
5	20.81	0.58	20.84	1.32	−0.10	0.921

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Case Problem 3 solution: Hotelling T^2

$$\mathbf{C} = \left(\frac{20C_1 + 27C_2}{47}\right)$$

$$= \begin{bmatrix} 13.358 & 13.748 & 1.951 & 1.373 & 2.231 \\ 13.748 & 26.146 & 2.765 & 2.252 & 2.710 \\ 1.951 & 2.765 & 0.645 & 0.350 & 0.423 \\ 1.373 & 2.252 & 0.350 & 0.324 & 0.347 \\ 2.231 & 2.710 & 0.423 & 0.347 & 1.004 \end{bmatrix}$$

$$\mathbf{C}^{-1} = \begin{bmatrix} 0.2061 & -0.0694 & -0.2395 & 0.0785 & -0.1969 \\ -0.0694 & 0.1234 & -0.0376 & -0.5517 & 0.0277 \\ -0.2395 & -0.0376 & 4.2219 & -3.2624 & -0.0181 \\ 0.0785 & -0.5517 & -3.2624 & 11.4610 & -1.2720 \\ -0.1969 & 0.0277 & -0.0181 & -1.2720 & 1.8068 \end{bmatrix}$$

Case Problem 3 solution: Conclusion

```
duplicate C1 myC
myC=((20*C1)+(27*C2))/47
duplicate myC invC
Matrixop /O invC=Inv(myC)
variable myf
matrixop /O t2=(28*21*(X1-X2)t̂ x invC x (X1-X2))/49
myf=(28+21-5-1)*2.84167/((28+21-2)*5)
```

$\boxed{\text{This gives t2= 2.84167 and F=0.52}}$

Case Problem 3 solution

```
x1=[157.381, 241.0, 31.433, 18.500, 20.81]
x2=[158.429, 241.571, 31.479, 18.446, 20.839]
c1=[[11.048, 9.100, 1.557, 0.870, 1.286],[9.100, 17.500, 1.910, 1.310,
     0.880],[1.557, 1.910, 0.531, 0.189, 0.240],[0.870, 1.310, 0.189, 0.176,
     0.133],[1.286, 0.880, 0.240, 0.133, 0.575]]
c2=[[15.069, 17.19, 2.243, 1.746, 2.931],[17.19, 32.550, 3.398, 2.950,
     4.066],[2.243, 3.398, 0.728, 0.470, 0.559],[1.746, 2.950, 0.470, 0.434,
     0.506],[2.931, 4.066, 0.559, 0.506, 1.321]]

n1=21
n2=28
dof1=n1-1
dof2=n2-1

#pairwise Tukey
print "==================="
print "Pairwise␣Tukey"
print "==================="
tc=ss.t.ppf(0.025, dof1+dof2) #signifficance level of 0.05 is 0.025 for 2 tails
print "The␣critical␣t-value␣is", tc
for i,line1 in enumerate(x1):
    var1=c1[i][i]
    var2=c2[i][i]
    #print "variance 1=", var1, "variance 2=", var2
    s2=((dof1*var1)+(dof2*var2))/(dof1+dof2)
    t=(x1[i]-x2[i])/pow((s2*((1./dof1)+(1./dof2))),0.5)
    if abs(t)>abs(tc):
        print "Reject␣H0", x1[i], var1, x2[i], var2, s2, "t=", t
    if abs(t)<abs(tc):
        print "Accept␣H0", x1[i], var1, x2[i], var2, s2,"t=", t
```

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Case Problem 3 solution

```
#Hotelling T^2
print "==================="
print "Hotelling␣T-squard"
print "==================="

Cov=(dof1*np.array(c1) + dof2*np.array(c2))/(dof1+dof2)
invCov=np.linalg.inv(Cov)
x1minx2=np.matrix.transpose(np.array(x1)-np.array(x2))


Tsqd=((n1*n2*x1minx2.dot(invCov)).dot(np.array(x1)-np.array(x2))  )/(n1+n2)
F=(n1+n2-5-1)*Tsqd/((dof1+dof2)*5)
Fc=ss.f.ppf(0.95, 5,43)  # One right hand tail test signifficance is 0.05, and
      95% confidence
print "T-squared=", Tsqd
print "F-value", F
print "Critical␣F=", Fc  # One right hand tail test signifficance is 0.05, and
      95% confidence
if F>Fc:
    print "F>Fc,␣Reject␣H0"
else:
    print "F<Fc,␣Accept␣H0"
```

Case Problem 3 solution: Conclusion

```
===================
Pairwise t-test
===================
The critical t-value is -2.01174051048
Accept H0 157.381 11.048 158.429 15.069 13.3579361702 t= -0.971940214503
Accept H0 241.0 17.5 241.571 32.55 26.1457446809 t= -0.378515397967
Accept H0 31.433 0.531 31.479 0.728 0.644170212766 t= -0.194269876152
Accept H0 18.5 0.176 18.446 0.434 0.324212765957 t= 0.321459788858
Accept H0 20.81 0.575 20.839 1.321 1.00355319149 t= -0.0981241145141
===================
Hotelling T-squard
===================
T-squared= 2.84166792191
F-value 0.519964768689
Critical F= 2.43223647186
F<Fc, Accept H0
```

There is no evidence to support that the two groups are statistically different. This conclusion is seen from the pairwise pooled t-tests on the variances and the Hotelling T^2 test. However, there are situations when the two tests do not agree. The Hotelling T^2 method gives us a more reliable result than just looking for differences in the means of each parameter. That is, the probability of Type-1 errors decreases for the Hotelling T^2 test because we are considering more variables (covariance terms) in the test.