exercise_5_problems

May 9, 2022

1 Exercise Set 5

Submission by: Mahak Sadhwani, Xiongxiao Wang, Sakshi Pahujani

Due: **9:30 9 May 2022**

Discussion: **13:00 13 May 2022**

Online submission at via ILIAS in the directory Exercises / Übungen -> Submission of Exercises / Rückgabe des Übungsblätter

2 1. Parametric tests: mean [100 points]

A very common question arises when we have two sets of data (or one set of data and a model) and we ask if they differ in location. To contrast the classical and Bayesian methods for hypothesis testing, we look at the simple case of comparison of means. We deal with a Gaussian distribution, because its analytical tractability has resulted in many tests being developed for Gaussian data; and then, of course, there is the central limit theorem.

Let us suppose we have n data X_i drawn from a Gaussian of mean μ_x, and m other data Y_i, drawn from a Gaussian of **identical variance** but different mean μ_y. Call the common variance σ^2. The Bayesian method is to calculate the joint posterior distribution:

$$\mathcal{P}(\mu_x, \mu_y, \sigma) \propto \frac{1}{\sigma^{n+m+1}} \exp\left(-\frac{\Sigma_i (x_i - \mu_x)^2}{2\sigma^2}\right) \exp\left(-\frac{\Sigma_i (y_i - \mu_y)^2}{2\sigma^2}\right)$$

in which we have used the Jeffreys prior for the variance. Integrating over the 'nuisance' parameter σ, we would get the joint probability prob(μ_x, μ_y) and could use it to derive, for example, the probability that μ_x is bigger than μ_y. From this we can calculate the probability distribution of $(\mu_x - \mu_y)$. The result depends on the data via a quantity

$$t' = \frac{(\mu_x - \mu_y) - (\bar{X} - \bar{Y})}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}, \qquad \text{where} \qquad s^2 = \frac{nS_x + mS_y}{\nu}$$

with the usual mean squares $S_x = \Sigma(X_i - \bar{X})^2/n$, similarly for S_y, and $\nu = n + m - 2$. s is called pooled standard deviation. The probability for t' is

$$\mathcal{P}(t') = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t'^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}$$

We regard the data as fixed and $(\mu_x - \mu_y)$ as the variable, simply computing the probability of any particular difference in the means. We might alternatively work out the range of differences whichare, say, 90 percent probable, or we might carry the distribution of $(\mu_x - \mu_y)$ on into a later probabilistic calculation. If we instead follow the classical line of reasoning, we do not treat the μ's as random variables. Instead we guess that the difference in the averages $X - Y$ will be the statistic we need; and we calculate its distribution on the null hypothesis that $\mu_x = \mu_y$. We find that

$$t = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

follows a t-distribution with $(n + m - 2)$ degrees of freedom. This is the classical Student's t. This gives the basis of a classical hypothesis test, the t-test for means. Assuming that $(\mu_x - \mu_y) = 0$ (the null hypothesis), we calculate t. If it (or some greater value) is very unlikely, we think that the null hypothesis is ruled out.

a. Suppose we have two small sets of data, from Gaussian distributions of equal variance:

$$(-1.22, -1.17, 0.93, -0.58, -1.14) \in A,$$

and,

$$(1.03, -1.59, -0.41, 0.71, 2.10) \in B.$$

Compute the respective mean values and the pooled standard deviation s **10 Points**

[1]:
```
import numpy as np
from math import sqrt
from scipy.stats import t
import matplotlib.pyplot as plt

A = np.array((-1.22, -1.17, 0.93, -0.58, -1.14))
B = np.array((1.03, -1.59, -0.41, 0.71, 2.10))
nu = len(A)+len(B)-2
print('mean of A =',A.mean(),'mean of B =',B.mean(),' nu=',nu)
print('std of A =',A.std(),'std of B =',B.std())
s=sqrt((A.std()*len(A)+B.std()*len(B))/nu)
print('pooled std s =',s)
```

```
mean of A = -0.6359999999999999 mean of B = 0.368  nu= 8
std of A = 0.8167888343996874 std of B = 1.264442960358434
pooled std s = 1.1405129862144603
```

The mean value of set A is $\overline{A} = \frac{\sum_{i=1}^{n} a_i}{n} = -0.636$, the standard deviation is $S_a = 0.817$

The mean value of Set B is $\overline{B} = \frac{\sum_{i=1}^{m} b_i}{m} = 0.368$, the standard deviation is $S_b = 1.264$

The degree of freedom $\nu = m + n - 2 = 8$

The pooled standard deviation $s = \sqrt{\frac{n.S_a + m.S_b}{\nu}} = 1.141$

b. Compute the t statistic. Perform a two tailed test. What is the chance that these data would arise if the means were the same. What is the chance if we did a one-tailed test? **25 Points**

Two tailed test:

Null Hypothesis H_0: $\mu_x - \mu_y = 0$

Alternative Hypothesis H_a: $\mu_x - \mu_y \neq 0$

One tailed test:

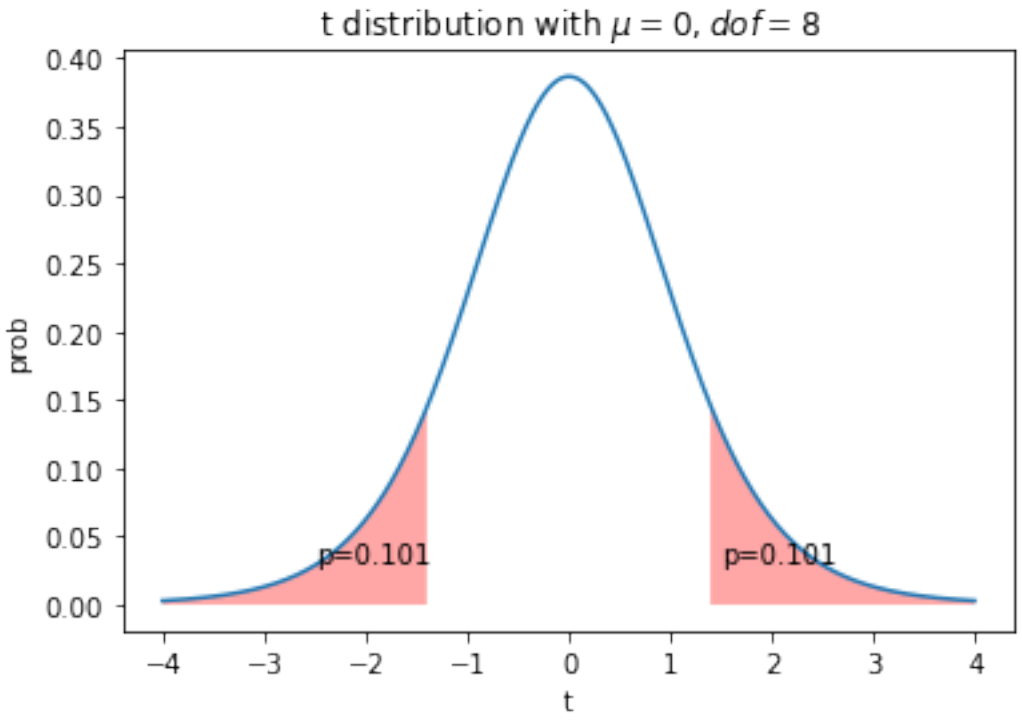
Null Hypothesis H_0: $\mu_x - \mu_y = 0$

Alternative Hypothesis H_a: $\mu_x - \mu_y \leq 0$

[2]:
```
t_value = ((A.mean()-B.mean())/(s*sqrt(1/len(A)+1/len(B))))
print('t_value=',t_value)
p_value = t.cdf(t_value,nu)
print('p_value=',p_value,'two tailed p_value=',p_value*2)
plt.fill_between(x=np.arange(-4,t_value,0.01), y1 = t.pdf(np.
 ↪arange(-4,t_value,0.01),nu),facecolor ='red',alpha =0.35)
plt.plot(np.arange(-4,4,0.01), t.pdf(np.arange(-4,4,0.01),nu))
plt.fill_between(x=np.arange(-t_value,4,0.01), y1 = t.pdf(np.
 ↪arange(-t_value,4,0.01),nu),facecolor ='red',alpha =0.35)
plt.text(x=-2.5, y=0.03, s= 'p='+str(round(p_value,3)))
plt.text(x=1.5, y=0.03, s= 'p='+str(round(p_value,3)))
plt.title(r't distribution with $\mu =0,dof = 8$')
plt.xlabel('t')
plt.ylabel(r'prob')
```

```
t_value= -1.391885409979911
p_value= 0.10071622535977316 two tailed p_value= 0.20143245071954632
```

[2]:
```
Text(0, 0.5, 'prob')
```



The t value is $t = \frac{\overline{A}-\overline{B}}{s\cdot\sqrt{\frac{1}{m}+\frac{1}{n}}} \approx -1.392$

One tailed p_value is 0.1007, two tailed p_value is 0.2014.

The probability that these data would arise if the means were the same is 0.2014

The probability for one-tailed test is 0.1007

c. Calculate the distribution of $(\mu_x - \mu_y)$ from a Bayesian point of view and plot the resulting prob$(\mu_x - \mu_y)$ as a function of $(\mu_x - \mu_y)$. What is the chance that μ_x is not smaller than μ_y? **25 Points**

$$t' = \frac{(\mu_x - \mu_y) - (\bar{X} - \bar{Y})}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}, \qquad \text{where} \qquad s^2 = \frac{nS_x + mS_y}{\nu}$$

$$\mathcal{P}(t') = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t'^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}$$

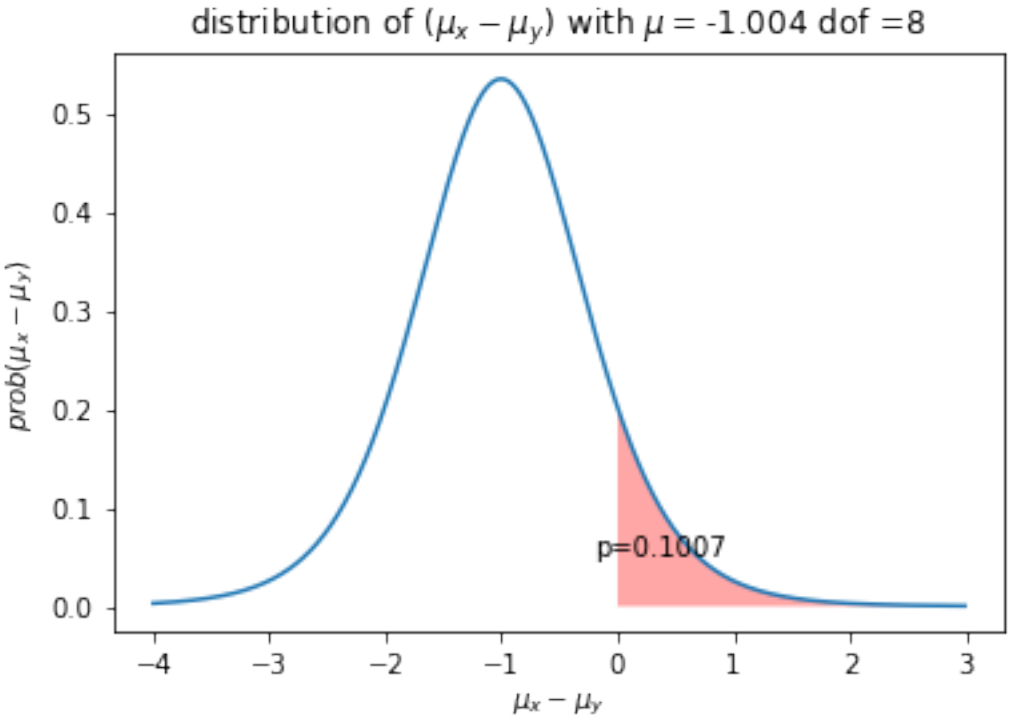
Using the two given equations, we can get the distribution of $(\mu_x - \mu_y)$ from a Bayesian point of

view:

$$\mathcal{P}(\mu_x - \mu_y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{(\frac{(\mu_x - \mu_y) - (\bar{X} - \bar{Y})}{s\sqrt{\frac{1}{m} + \frac{1}{n}}})^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}$$

[3]:
```
plt.plot(np.arange(-4,3,0.01), t.pdf(np.arange(-4,3,0.01),nu,loc=A.mean()-B.
 ↪mean(),scale = s*sqrt(1/len(A)+1/len(B))))
plt.fill_between(x=np.arange(0,3,0.01), y1 = t.pdf(np.arange(0,3,0.01),nu,loc=A.
 ↪mean()-B.mean(),scale = s*sqrt(1/len(A)+1/len(B))),facecolor ='red',alpha =0.
 ↪35)
p_red = 1-t.cdf(0,nu,loc=A.mean()-B.mean(),scale = s*sqrt(1/len(A)+1/len(B)))
plt.title('distribution of $(\mu_x-\mu_y)$ with $\mu =$'+str(A.mean()-B.
 ↪mean())+' dof ='+str(nu))
plt.xlabel('$\mu_x-\mu_y$')
plt.ylabel(r'$prob(\mu_x-\mu_y)$')
plt.text(x=-0.2,y=0.05,s='p='+str(round(p_red,4)))
print('The probability mu_x not smaller than mu_y is ',p_red)
```

```
The probability mu_x not smaller than mu_y is  0.10071622535977309
```



d. By analogous calculations we arrive at the F test for variances. Again Gaussian distributions

are assumed. The null hypothesis is $\sigma_x = \sigma_y$, the data are $X_i (i = 1, \ldots, n)$ and $Y_i (i = 1, \ldots, m)$ and the test statistic is,

$$\mathcal{F} = \frac{\Sigma_i \left(X_i - \bar{X}\right)^2 / (n-1)}{\Sigma_i \left(Y_i - \bar{Y}\right)^2 / (m-1)}$$

This follows the F ratio distribution with $(n-1)$ and $(m-1)$ degrees of freedom. The testing is the as for the Student's t. Perform a test whether the variances of the two data sets are the same. **40 Points**

Null Hypothesis H_0: $\sigma_x = \sigma_y$

Alternative Hypothesis H_a: $\sigma_x \neq \sigma_y$

We set $\alpha = 0.05$

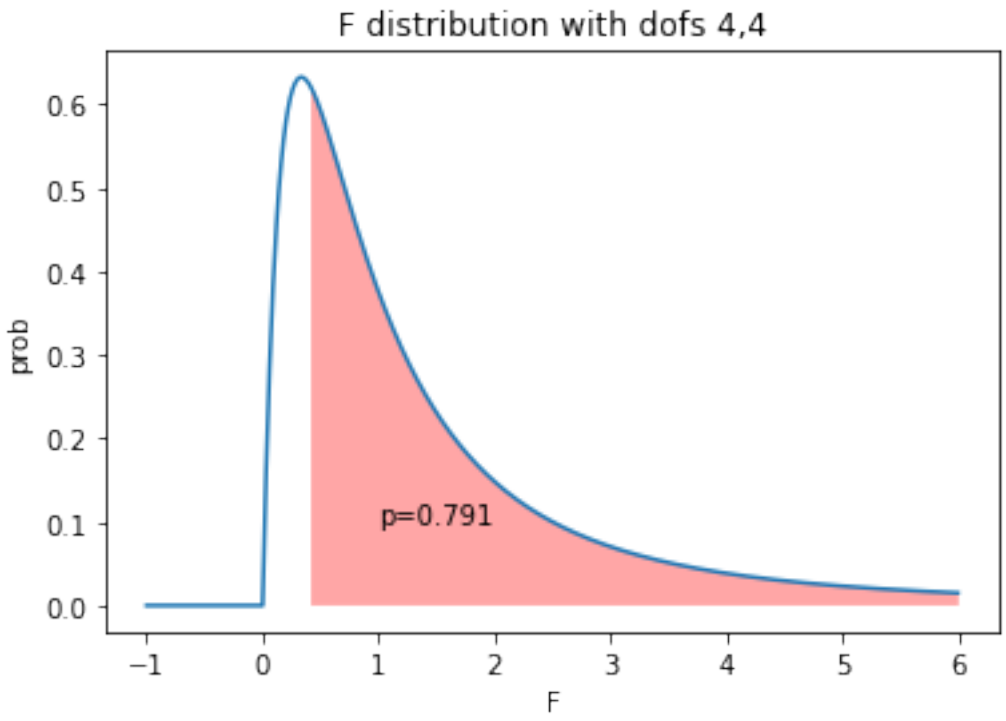
[5]:
```
from scipy.stats import f

F_num = A.var()*len(A)/(len(A)-1)
F_den = B.var()*len(B)/(len(B)-1)
F=F_num/F_den

p_value_f = 1-f.cdf(F,len(A)-1,len(B)-1)
print('The value of F is ',F,' The value of probability is ',p_value_f)
plt.plot(np.arange(-1,6,0.01),f.pdf(np.arange(-1,6,0.01),len(A)-1,len(B)-1))
plt.fill_between(x=np.arange(F,6,0.01),y1=f.pdf(np.arange(F,6,0.
 ↪01),len(A)-1,len(B)-1),facecolor ='red',alpha =0.35)
plt.title('F distribution with dofs 4,4')
plt.text(x=1,y=0.1,s='p='+str(round(p_value_f,4)))
plt.xlabel('F')
plt.ylabel('prob')
```

```
The value of F is  0.41727378259912323  The value of probability is
0.7909930575941863
```

[5]:
```
Text(0, 0.5, 'prob')
```



Decision:

Fail to reject the null hypothesis because P_value_f $> \alpha$, we still choose to belive H_0: $\sigma_x = \sigma_y$