APPLIED STATISTICS

Homework (in English)*

Current version October 6, 2020

Alessandra R. Brazzale

Department of Statistical Sciences

University of Padova

based on original course material

by Guido Masarotto

Master's Degrees in

Molecular Biology & Sanitary Biology

*The order of the questions was slightly changed with respect to the original Italian version.

Assignments

Below you will find a number of problems and exercises which are similar to the exam questions you will be asked to answer.

If you feel like needing more exercise, you may also try and solve the practice problems and assignment problems listed at the end of each chapter of our textbook[1]. Answers to the practice problems are provided at the end of the same.

PROBLEM 1. Briefly explain what we mean by "observed significance level" (p-value).

PROBLEM 2. Briefly explain what a confidence interval is.

PROBLEM 3. Briefly explain why we need the "central limit theorem".

PROBLEM 4. Student's two-sample t test and the Wilcoxon-Mann-Whitney test. What are they used for? In what do they differ?

PROBLEM 5. What is a type I error? And what a type II error? Why do we have to account for them? And why should we know how to calculate the probabilities with which they occur?

PROBLEM 6. What is the purpose of Bonferroni's correction? What is it for? Why do we use it?

PROBLEM 7. If we find a very large correlation coefficient between two variables, can we conclude that we discovered the existence of a causal effect?

PROBLEM 8. What is the purpose of cluster analysis? Describe, using your own words, an example of application of one of the algorithms seen.

[1] *The Analysis of Biological Data* (2nd ed) by Whitlock and Schluter, 2015

PROBLEM 9. For each of the tests seen in class, "invent/think" a situation in which they could be useful and briefly describe the experiment, the data that could be collected and the question you want to answer.

PROBLEM 10. For each of the "methods" seen in class, "invent/think" a situation in which they could be useful and briefly describe the experiment, the data that could be collected and the question you want to answer.

PROBLEM 11. A researcher has calculated a confidence interval with coverage level equal to 5% (or $\alpha = 0.95$). His motivation is that the smaller the interval is the more precise will it be.
How would you reply?

PROBLEM 12. In England, from 1200 and until coins were minted in precious metal, the honesty of the *Master of Mint* (the Royal Mint's contractor) was audited by means of a sample check, the so-called *Trial of Pyx* (*pyx* is a sacred vessel). The *Master of Mint* was subject to consequences that are far from being "pleasant" (during some centuries he was even sentenced to death) if evidence pointed to his dishonesty.

In 1799 gold guineas were checked in the following way:

- The nominal weight of a guinea was 128 grains. (360 grains make an ounce.)
- 100 guineas were randomly extracted from all those produced during the whole year. (Two royal officials went to the mint on randomly chosen days scattered throughout the year and randomly selected one or more coins from the production of that day.)
- The "extracted" guineas were gradually conserved in the *pyx*.
- At the end of the year, the vessel with the 100 guineas was weighted.
- The *Master of Mint* passed the *Trial of Pyx* if the weight of the extracted guineas was equal to the expected weight plus or minus 1/400 of the expected weight itself.

We can assume that with the technology of the time the weight of each single coin distributed as a normal variable whose mean was controlled by the *Master of Mint* and with unit variance. In addition, the true weights of the different coins can be considered independent of each other.

With these data, it was possible to calculate that:

1. the probability, say A, that a honest *Master of Mint*, who set the mean weight of gold guineas at exactly 128 grains, would survive the *Trial of Pyx*. This is
$$A = 0{,}999$$
.
2. the probability, say B, that a dishonest *Master of Mint*, who decided to steal on average 0,3 gold grains from each guinea produced, was discovered. This is
$$B = 0{,}579.$$

Questions:

(a) Implicitly what kind of statistical test did the King use to verify the honesty of his *Master of Mint*?
(a) Specifically referring to the answer you gave at the previous item, how are the two probabilities A and B called, and what do they represent?

PROBLEM 13. During a recent inspection of a hospital, the noise level (`y`, in decibels) was measured in 174 rooms and corridors of the hospital wards at random moments taken during the night. Among other things, the intent was to find out whether the hospital complies with an European directive that requires the average "night noise" to be less than 10*db*.
Some results are listed below.

```
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7716  3.8090  6.3650  7.8000 10.6200 29.0400

> t.test(y,mu=10,alternative="greater")

One Sample t-test

data:  y
t = -5.4565, df = 173, p-value = 1
alternative hypothesis: true mean is greater than 10
```

(a) Trace a boxplot which graphically summarizes the given information.
(b) Comment on the results. In particular, say whether the hospital appears to comply with the European directive.
(c) To use the one-sample t test did we have to assume that the decibels are distributed like a normal?

PROBLEM 14. In a study, increases (`y`, in grams) of lean mass was measured for 70 individuals who had been "given" a "fattening" diet.
Some results of the carried out analyses are below.

```
> shapiro.test(y)

Shapiro-Wilk normality test

data:  y
W = 0.9479, p-value = 0.005621

> t.test(y)

One Sample t-test

data:  y
t = 13.7287, df = 69, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 372.4051 499.0357
sample estimates:
mean of x
 435.7204
```

(a) Trace the histogram which is consistent with the given information.
(b) Comments on the given results. In particular: (i) explain what the Shapiro-Wilks test tells us, and (ii) translate into common language the above confidence interval.

PROBLEM 15. In the county of *Clear Waters* a survey has been conducted for years to estimate the proportion of castles inhabited by fairies. This is a sample survey: every year, 100 castles are randomly extracted and for each castle it is determined whether or not fairies live there.
This year the number of "fairy castles" is equal to 38. A young fairy who attended our Applied Statistics course brought to the attention of the Chamberlain the following analysis.

```
> prop.test(38,100,p=0.5,conf.level=0.9)

1-sample proportions test with continuity correction

data:  38 out of 100, null probability 0.5
X-squared = 5.29, df = 1, p-value = 0.02145
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.2996486 0.4670641
```



```
sample estimates:
   p
0.38
```

The poor Chamberlain, however, feels a bit lost. In particular, he wonders about the following.
(a) What does the confidence interval say?
(b) An old book written by the Duke of *Clear Waters* states that about 50% of the castles of the county are inhabited by fairies. What do the data tell us? Could this proportion still be the same?

PROBLEM 16. A team of researchers is studying a drug that could reduce the invasiveness of cancer cells. Ten cell lines were treated with the new drug (`F`) and ten were grown without treatment (`NF`). Invasiveness (`y`) was measured for each line as the percentage of cells that were able to pass a specially positioned membrane.
The results of Student's two-sample t test are below.

```
> t.test(y~g)

Welch Two Sample t-test

data:  y by g
t = -2.6695, df = 12.913, p-value = 0.01937
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.77937348 -0.08189612
sample estimates:
 mean in group F mean in group NF
       0.2581661        0.6888009
```

(a) Comment on the results.
(b) Trace the boxplots which are consistent with the given information.
(c) What would you do differently? And, why?

PROBLEM 17. A study conducted in a department for experimental medicine measured the concentration of a certain hormone in 500 adult female rabbits. Consider the results of the following data analyses. (In `R`, `y` is the variable which contains the 500 measures.)

```
> shapiro.test(y)

Shapiro-Wilk normality test
```

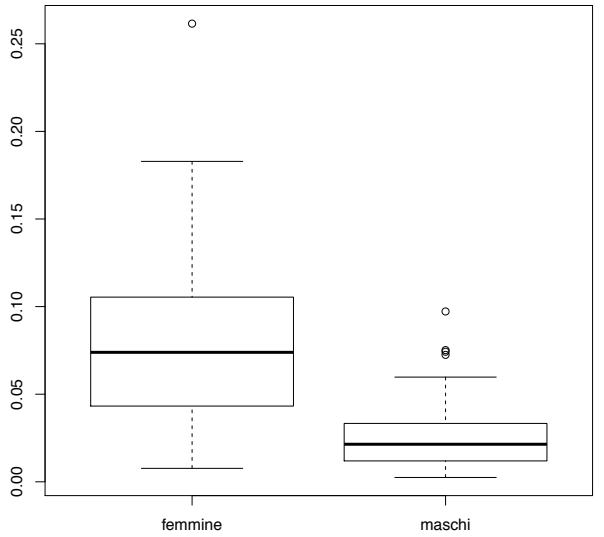
```
data:  y
W = 0.8452, p-value < 2.2e-16

> t.test(y,conf.level=0.99)

One Sample t-test

data:  y
.....................................
.....................................
99 percent confidence interval:
 114.3253 142.6109
sample estimates:
mean of x
 128.4681
```

(a) What do the results of the first function (`shapiro.test`) tell us?
(b) Explain in a concise and non-technical way what the confidence interval calculated by the second function (`t.test`) tells us.
(c) In order to use the previously calculated confidence interval, do we have to assume that the distribution of the studied hormone concentration is normal for all adult female rabbits of the breed used in the experiment?
(d) In fact, the experiment was conducted to verify the following system of hypotheses

$$H_0 : \mu = 150 \qquad \text{vs} \qquad H_1 : \mu \neq 150$$

where μ is the mean concentration of the hormone in the population of all female rabbits.
Only that... you lost the data and only hold the above code chunk of the analysis.
Besides sprinkling your head with ashes, promising that this won't happen again, and saying that, as a punishment, you will not abuse of alcohol and chocolate for the coming ten days... what else can you say?
In your opinion, should H_0 be accepted or rejected? Briefly motivate your choice.

PROBLEM 18. The concentration of a certain type of platelets was measured in blood samples of 100 men (`maschi`) and 100 women (`femmine`) aged between 30 and 60 years. The data are "shown" by the following two boxplots.



(a) Comment on the figure.
(b) Welch's two-sample t test was applied to the data. Try and guess which p-value was obtained, and explaining why.

PROBLEM 19. The alcohol level was measured for 250 male and 220 female students from a large U.S. university, when they returned on a Friday evening to the dormitories on campus. The alcohol level was classified as above/below the limit in that state for driving.
The results are shown in the following table.

	male	female
below	64	71
above	186	149

In addition, the following function call was run in `R`

```
> prop.test(c(64,71),c(250,220))
```

```
2-sample test for equality of proportions with continuity correction

data:  c(64, 71) out of c(250, 220)
X-squared = 2.2295, df = 1, p-value = 0.1354
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.15311702  0.01966247
sample estimates:
   prop 1    prop 2
0.2560000 0.3227273
```

Comment on the results.

PROBLEM 20. A new laboratory test was developed to discriminate between patients with type 1 and type 2 diabetes mellitus. To verify its reliability, a group of researchers measured the proposed parameter for

- 100 "healthy" patients (`sani`, without diabetes);
- 100 patients with type 1 diabetes (`tipo1`);
- 100 patients with type 2 diabetes (`tipo2`).

Below, you find the results of the analyses they carried out. (`y` is the response variable while `g` identifies the groups.)

```
> oneway.test(y~g)

One-way analysis of means (not assuming equal variances)

data:  y and g
F = 289.3993, num df = 2.000, denom df = 197.158, p-value < 2.2e-16

> pairwise.t.test(y,g)

Pairwise comparisons using t tests with pooled SD

data:  y and g

      sani   tipo1
tipo1 <2e-16 -
tipo2 <2e-16 0.68

P value adjustment method: holm
```


(a) Comment on the results. In particular, is it true that the newly developed laboratory test allows us to discriminate between the different types of diabetes? If not, what does it seem to allow us to do?
(b) Trace side by side the boxplots, one for each group, which are consistent with the given information.

PROBLEM 21. Although we did not discuss it in class, there are of course tests which can be used to verify the hypothesis that the variance is the same in several groups (in analogy with one-way analysis of variance and the Kruskall-Wallis test only that we now apply it to the variability of the data and not for comparing the mean/position of the groups).
A test based on the assumption of normality is Bartlett's test. A non-parametric version based on ranks is the Fligner-Killeen test. Comment on the following results knowing that (i) for both tests, the null hypothesis corresponds to equal variances, and (ii) `cholesterol` and `race` are two variables in `R` which contain cholesterol measurements and the ethnic group (white, black, Latin) for 180 students on an American campus.

```
> bartlett.test(colesterolo~razza)

Bartlett test of homogeneity of variances

data:  colesterolo by razza
Bartlett's K-squared = 2.4576, df = 2, p-value = 0.2927

> fligner.test(colesterolo~razza)

Fligner-Killeen test of homogeneity of variances

data:  colesterolo by razza
Fligner-Killeen:med chi-squared = 2.8096, df = 2, p-value = 0.2454
```

PROBLEM 22. Full moon periods have for long been described in a sinister way and, in particular, as inviting to madness. For example, W. Shakespeare has Othello say after the killing of Desdemona

> *It is the very error of the moon,*
> *She comes more near the earth than she was want,*
> *and makes men mad.*

For this reason, a psychiatric clinic in Virginia (USA) recorded the rate of hospitalization because of acute crisis for 36 consecutive "lunar periods".

The measured variables are:
– the quarter (`Quarter`) of the year to which the observation refers; it can assume values from Q1 to Q4.
– the moon phase (`Moon`); it takes the values "Before", "During" or "After" which indicate whether we are before, during or after the full moon.
– the hospitalization rate (`Admission`); the number of hospitalizations divided by the number of days of the period.

The following table shows the averages of `Admission` for the various combinations of `Quarter` and `Moon`.

```
         Quarter
Moon           Q1       Q2        Q3        Q4
  Before 11.90000 14.26667  9.766667  7.733333
  During 12.66667 17.66667 12.666667 10.666667
  After  13.16667 13.96667  9.833333  8.866667
```

A two-way analysis of variance was carried out to see whether `Quarter` and `Moon` affect admissions. The results are given bellow.

```
Response: Admission
             Sum Sq Df F value  Pr(>F)
Moon          41.51  2  1.3231 0.28504
Quarter      191.05  3  4.0593 0.01819 *
Moon:Quarter  15.84  6  0.1683 0.98286
Residuals    376.51 24
```

Comment on the results and, in particular, say whether or not we can continue to enjoy, and with what degree of peace, the full moon (which, at least to me, looks not sinister but beautiful (◡̈)).

PROBLEM 23. During a research, the growth rate (`Velocita`) of 100 virus cultures was measured. The cultures had been prepared with 5 different growth media (`Terreno`) and concerned 2 different viruses (`Virus`). In particular, 10 cultures were prepared with "medium 1" and concerned "virus A", 10 with "medium 1" and "virus B", 10 with "medium 2" and "virus A", and so on.
The following code chunk shows the results of two-way analysis of variance.

```
Response: Velocita
                 Sum Sq Df F value    Pr(>F)
Virus           0.04537  1  4.5687  0.035272 *
Terreno         2.79117  4 70.2689 < 2.2e-16 ***
Virus:Terreno   0.14175  4  3.5685  0.009504 **
Residuals 0.89373 90
```


Comment on the results. In particular, assuming that the ultimate goal of the research is to decide which growth medium to use to quickly have grow the two viruses, say whether:
(a) it is true, as someone claimed at the beginning of the research, that the medium is irrelevant, that is, that the two viruses grow with the same speed on all five mediums?
(b) it is advisable to use the same culture medium for both viruses?

PROBLEM 24. For each of the following code chunks comment on the result and draw hypothetical scatter diagrams (plot of y against x) which are consistent with the given information.

In the `R` *output,* `df` *indicates the number of observations minus two. You can of course use this information to comment though you are not asked to draw a scatter plot with hundreds or even thousands of points.*

(a)
```
> cor.test(x,y)

   Pearson's product-moment correlation

   data:  x and y
   t = 1.7328, df = 19, p-value = 0.8627
   alternative hypothesis: true correlation is not equal to 0
   95 percent confidence interval:
    -0.02484006  0.35675544
   sample estimates:
         cor
   0.1724193
```
(b)
```
> cor.test(x,y)

   Pearson's product-moment correlation

   data:  x and y
   t = -2.1866, df = 1085, p-value = 0.0001
   alternative hypothesis: true correlation is not equal to 0
   95 percent confidence interval:
    -0.075502 -0.036781
   sample estimates:
         cor
   -0.055686
```
(c)
```
> cor.test(x,y)

   Pearson's product-moment correlation
```

```
    data:  x and y
    t = -30.4461, df = 98, p-value < 2.2e-16
    alternative hypothesis: true correlation is not equal to 0
    95 percent confidence interval:
     -0.9668169 -0.9278986
    sample estimates:
           cor
    -0.9509926
```

(d)
```
> cor.test(x,y)

    Pearson's product-moment correlation

    data:  x and y
    t = 5.1665, df = 98, p-value = 1.256e-06
    alternative hypothesis: true correlation is not equal to 0
    95 percent confidence interval:
     0.2928694 0.6041839
    sample estimates:
          cor
    0.4626724
```

PROBLEM 25. A study recorded the following 3 variables fo 35 surgery patients:

Y: the depression (in percentage) of the lymphocytes;
X_1: the length (in hours) of the surgery;
X_2: a measure of the degree of invasiveness of the surgery which ranges from 0 to 4 (0 being the minimum grade and 4 the maximum grade).

A linear regression model was fitted. The results are given below.

```
> summary(lm(y~x1+x2))

Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-29.7948  -5.0823   0.2655   9.5114  22.2153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.7635     4.5223  -1.717   0.0957 .
```



```
x1            0.7985     0.5613   1.423   0.1645
x2           14.2378     1.5776   9.025 2.62e-10 ***
---

Residual standard error: 12.77 on 32 degrees of freedom
Multiple R-squared: 0.7194,Adjusted R-squared: 0.7018
F-statistic: 41.01 on 2 and 32 DF,  p-value: 1.481e-09
```

(a) Comment on the results.
(b) Is it possible to simplify the model (for instance, by reducing the number of covariates)?

PROBLEM 26. A research aimed at studying 20 different water catchment areas observed the following variables:

- `nitro`: the concentration of nitrogen (in mg/litre); it is actually an average of measurements taken over a whole year period.
- `forest`: the percentage of land around the catchment area occupied by forests and parks.
- `industrial`: the percentage of land around the catchment area occupied by industrial and commercial activities.

A linear regression model was fitted to the data using as response variable `nitro` and the remaining two as explanatory variables. The results are as follows.

```
> summary(lm(nitro~industrial+forest,data=d))

Call:
lm(formula = nitro ~ industrial + forest, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53801 -0.10484 -0.01854  0.07903  0.72398

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.096213   0.240457   8.718 1.11e-07 ***
industrial   0.187666   0.081607   2.300 0.034413 *
forest      -0.016475   0.003455  -4.769 0.000178 ***
---

Residual standard error: 0.2555 on 17 degrees of freedom
Multiple R-squared: 0.6935,Adjusted R-squared: 0.6575
F-statistic: 19.24 on 2 and 17 DF,  p-value: 4.308e-05
```

(a) What do the results tell us?
(b) On the basis of the results just seen, someone proposed that for every km^2 dedicated to industrial and commercial activities 10 be reserved for forests and woods. Does the proposal make sense to you?

PROBLEM 27. Figure 1 shows the pairs plots for the four variables recorded by a study.

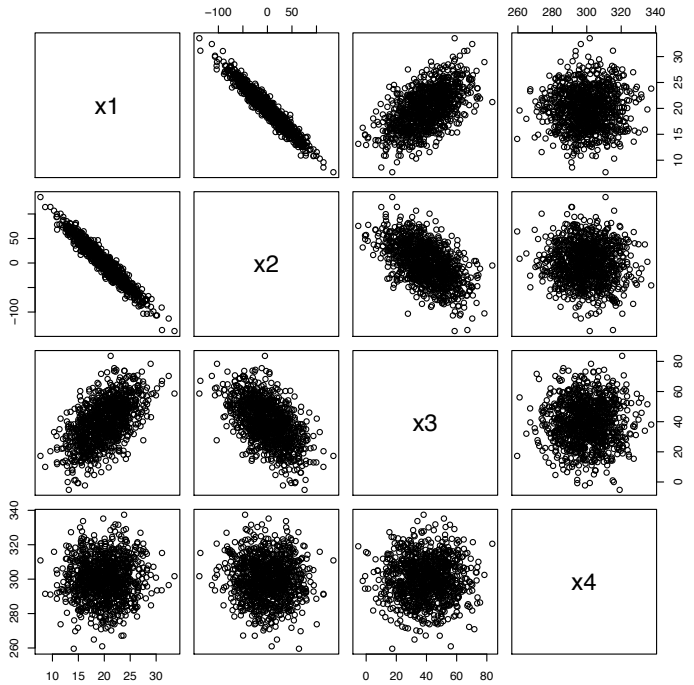


Figure 1: Pairs plots of four variables.

(a) Comment on the figure. Focus, in particular, on the relationships presents among the variables.
(b) For each pair of variables guess the value of the correlation coefficient which is consistent with the given information.

PROBLEM 28. A total of 150 insects were exposed to different dose levels of an insecticide. For every insect the following two variables were recorded:

$$Y = \begin{cases} 0 & \text{the insect is still alive after 3 hours} \\ 1 & \text{the insect is dead after 3 hours} \end{cases},$$

and the variable X which describes the dose, which for the given experiment ranges from 0 to 4 grams per litre. (That is, if $X = 3$ then the "sprayed" insecticide had been prepared by "mixing" 3 gr of "active substance" in one litre of water.)
A logistic regression model was fitted to the data and the following results were obtained.

```
> summary(glm(y~x,family="binomial"))
.....................................................
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4857     0.5824  -5.985 2.16e-09 ***
x             1.1619     0.1814   6.405 1.51e-10 ***
.....................................................
```

Comment on the results.

PROBLEM 29. A total of 35 guinea pigs were daily injected over a 10 weeks period with a substance which is potentially toxic to the liver. The animals had been divided into 5 groups, of 7 guinea pigs each, which received a different dose of the substance. (The first group received a placebo, i.e. 0 mg of the substance, the second group 0.5 mg,..., the fifth group 2.5 mg per day.) At the end of the 10 weeks period, the presence or absence of hepatic steatosis was verified for each guinea pig.
A logistic regression model was fitted to the data with the purpose to study how the probability of presenting a steatosis varied with the different dose levels of the injected substance. The results obtained in `R` are as follows.

```
> summary(glm(steatosis~dose,family="binomial"))
.....................................................
.....................................................
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1789     0.8612  -2.065   0.0389 *
dose         -0.1422     0.3799  -0.374   0.7082
.....................................................
.....................................................
```

In the above code chunk, `steatosi` is a vector of length 35 whose elements are 1 if the guinea pig showed hepatic steatosis at the end of the study and 0 otherwise; `dose` represents the different dose levels to which the guinea pigs were subjected.

(a) Comment on the achieved results.

(b) In the light of a previous research, a researcher yelled:

> "WOW !!!! It looks as if all guinea pigs, regardless of the dose level of the administered substance, have a probability of 0,5 of developing steatosis. It's like tossing a coin and observing head or cross."

Do you share this opinion? Why, or why not?

PROBLEM 30. During a research, six different numerical variables, which we identify by $V1,\ldots,V2$, were measured on 88 different objects.

The following code chunk reports the results from the principal components analysis.

```
> .PC <- princomp(~V1+V2+V3+V4+V5+V6, cor=TRUE, data=y)

> unclass(loadings(.PC))  # component loadings
       Comp.1       Comp.2      Comp.3     Comp.4      Comp.5      Comp.6
V1 -0.4591642 -0.008993128 -0.14202343  0.6316481 -0.04134766  0.60681770
V2 -0.4591643 -0.008961427  0.82563568 -0.1174850 -0.30140707 -0.05257932
V3 -0.4591641 -0.008998785 -0.46959864 -0.6326625 -0.37066412  0.17581443
V4 -0.4591642 -0.008997325 -0.27072707  0.3440183 -0.04128484 -0.77184268
V5 -0.3067786  0.663574220  0.04891266 -0.1945747  0.65115620  0.03605556
V6 -0.2501196 -0.747894471  0.04412010 -0.1753516  0.58680783  0.03249381

> .PC$sd^2  # component variances
      Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6
4.742636e+00 1.257364e+00 3.425312e-07 1.113484e-07 8.104771e-09 7.521819e-12

> summary(.PC) # proportions of variance
Importance of components:
                        Comp.1    Comp.2    Comp.3   Comp.4   Comp.5   Comp.6
Standard deviation     2.177     1.121     5.85e-04 3.33e-04 9.00e-05 2.74e-06
Proportion of Variance 0.790     0.209     5.70e-08 1.85e-08 1.35e-09 1.25e-12
Cumulative Proportion  0.790     0.999     1.000    1.000    1.000    1.000
```

Figure 2 shows the biplot of the 88 objects based on the first two principal components.

(a) Comment on the results.

(b) Which are the characteristics of object n. 29 (the one on the bottom of the figure). And which are those of object n. 86 (the one in the middle on the right)?

(c) Is there any correlation among the original variables? Or, are they all incorrelated? And the correlation, is it supposed to be large or small?

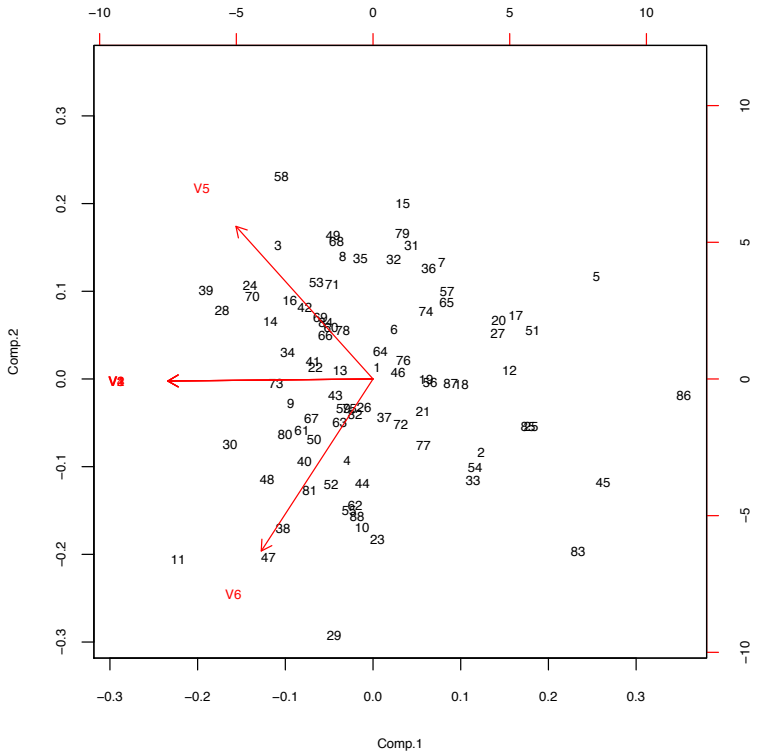


Figure 2: A biplot.

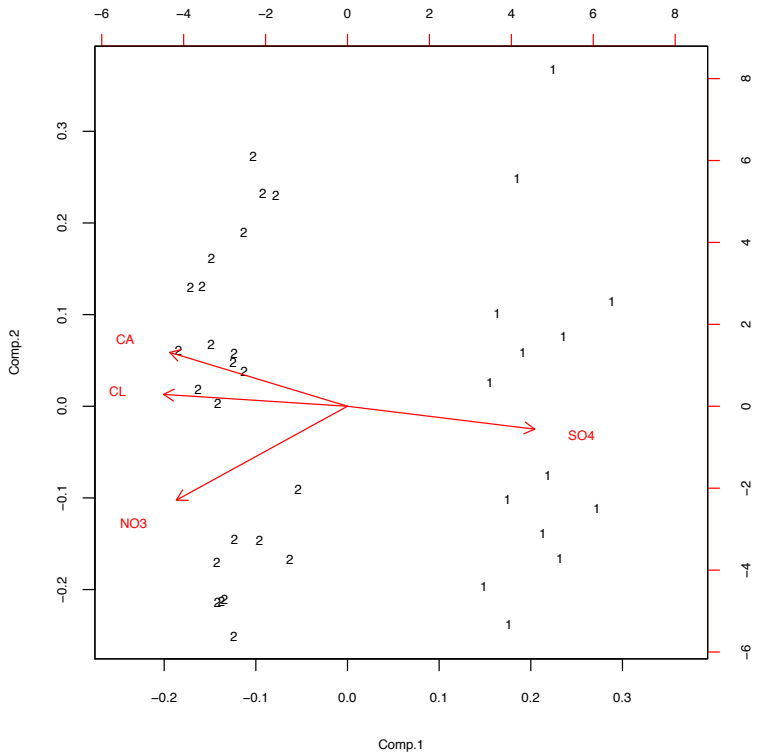


Figure 3: Biplot showing the results of KMeans applied to 37 rivers. 1/2 represent the number of the groups to which the observations are allocated by KMeans.

Problem 31. Concentrations of the following substances were measured in the water of 37 rivers or streams in a certain geographical region:

- NO_{3^-}: hereafter referred to as `NO3`,
- SO_4^{2-}: hereafter referred to as `SO4`,
- CL^-: hereafter referred to as `CL`,
- Ca^{2+}: hereafter referred to as `CA`.

The KMeans clustering algorithm was applied to the data to identify two groups. The biplot is shown in Figure 3. The averages of the 4 variables for the observations belonging to the two clusters is given below.

```
            CA         CL          NO3         SO4
Cluster 1 -1.1128571 -1.1421429 -1.0664286  1.1957143
Cluster 2  0.6791304  0.6952174  0.6486957 -0.7265217
```

Comment on the results. In particular, explain in what the two groups differ.

Problem 32. During a research, 10 numerical variables were measured on 1,000 objects. The below code chunk reports the correlation coefficients for all pairs of variables and the corresponding p-values (the raw ones and those adjusted using Holm's correction) which assess the significance of the correlation coefficients.

```
> rcorr.adjust(x[,c("V1","V2","V3","V4","V5","V6","V7","V8","V9","V10")],
               type="pearson")
       V1    V2    V3    V4    V5    V6    V7    V8    V9   V10
V1   1.00  0.01 -0.01  0.00  0.02 -0.02 -0.09 -0.01  0.00  0.01
V2   0.01  1.00  0.04  0.04 -0.01 -0.01  0.03  0.00  0.02  0.03
V3  -0.01  0.04  1.00  0.02  0.03  0.02  0.00  0.02  0.02  0.03
V4   0.00  0.04  0.02  1.00  0.02 -0.01 -0.01  0.02 -0.02  0.03
V5   0.02 -0.01  0.03  0.02  1.00  0.00 -0.03  0.05 -0.02  0.02
V6  -0.02 -0.01  0.02 -0.01  0.00  1.00  0.01 -0.02  0.00  0.01
V7  -0.09  0.03  0.00 -0.01 -0.03  0.01  1.00  0.01  0.03 -0.03
V8  -0.01  0.00  0.02  0.02  0.05 -0.02  0.01  1.00  0.01  0.01
V9   0.00  0.02  0.02 -0.02 -0.02  0.00  0.03  0.01  1.00 -0.01
V10  0.01  0.03  0.03  0.03  0.02  0.01 -0.03  0.01 -0.01  1.00

n= 1000


P
    V1     V2     V3     V4     V5     V6     V7     V8     V9     V10
V1         0.8452 0.8036 0.9324 0.5547 0.5989 0.0042 0.8420 0.9853 0.8220
V2  0.8452        0.1591 0.1686 0.7408 0.8209 0.3876 0.9874 0.6249 0.3447
V3  0.8036 0.1591        0.6296 0.2796 0.5808 0.8847 0.5958 0.5341 0.3593
V4  0.9324 0.1686 0.6296        0.4983 0.8725 0.7620 0.5337 0.4682 0.3287
```

```
V5  0.5547 0.7408 0.2796 0.4983        0.9860 0.3078 0.0928 0.4933 0.5710
V6  0.5989 0.8209 0.5808 0.8725 0.9860        0.7252 0.5023 0.9811 0.7592
V7  0.0042 0.3876 0.8847 0.7620 0.3078 0.7252        0.7382 0.4008 0.2782
V8  0.8420 0.9874 0.5958 0.5337 0.0928 0.5023 0.7382        0.6927 0.7724
V9  0.9853 0.6249 0.5341 0.4682 0.4933 0.9811 0.4008 0.6927        0.8498
V10 0.8220 0.3447 0.3593 0.3287 0.5710 0.7592 0.2782 0.7724 0.8498

 Adjusted p-values (Holm's method)
    V1     V2     V3     V4     V5     V6     V7     V8     V9     V10
V1         1.0000 1.0000 1.0000 1.0000 1.0000 0.1903 1.0000 1.0000 1.0000
V2  1.0000        1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
V3  1.0000 1.0000        1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
V4  1.0000 1.0000 1.0000        1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
V5  1.0000 1.0000 1.0000 1.0000        1.0000 1.0000 1.0000 1.0000 1.0000
V6  1.0000 1.0000 1.0000 1.0000 1.0000        1.0000 1.0000 1.0000 1.0000
V7  0.1903 1.0000 1.0000 1.0000 1.0000 1.0000        1.0000 1.0000 1.0000
V8  1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000        1.0000 1.0000
V9  1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000        1.0000
V10 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
```

Someone proposed to reduce the number of original variables using principal components? Do you think this makes sense?

PROBLEM 33. In a small pilot study, 4 clinical parameters were measured, hereafter referred to as $X1, \ldots, X4$, in blood samples of 20 individuals, 14 healthy and 6 suffering from a rather rare disease. A principal components analysis was carried out. Figure 4 shows the biplot based on the first two components (which in this case also explained more than 98% of the total variability). The points corresponding to "healthy" subjects are indicated by the letter "S" and those corresponding to "sick" subjects by the letter "M".
Comment on the figure. In particular, say which of the clinical parameters $X1, \ldots, X4$ seem, potentially, to be able to better discriminate between healthy and sick individuals.

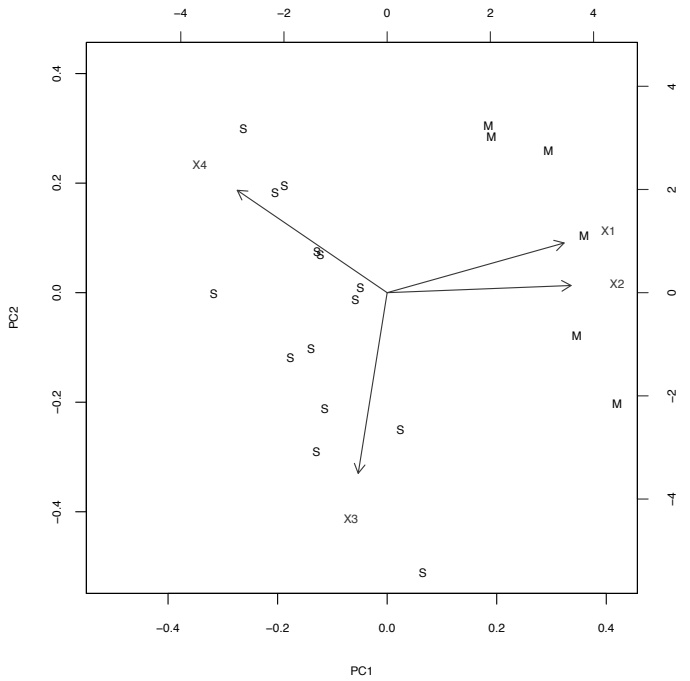


Figure 4: Biplot of the data of a small pilot study on a rare disease.

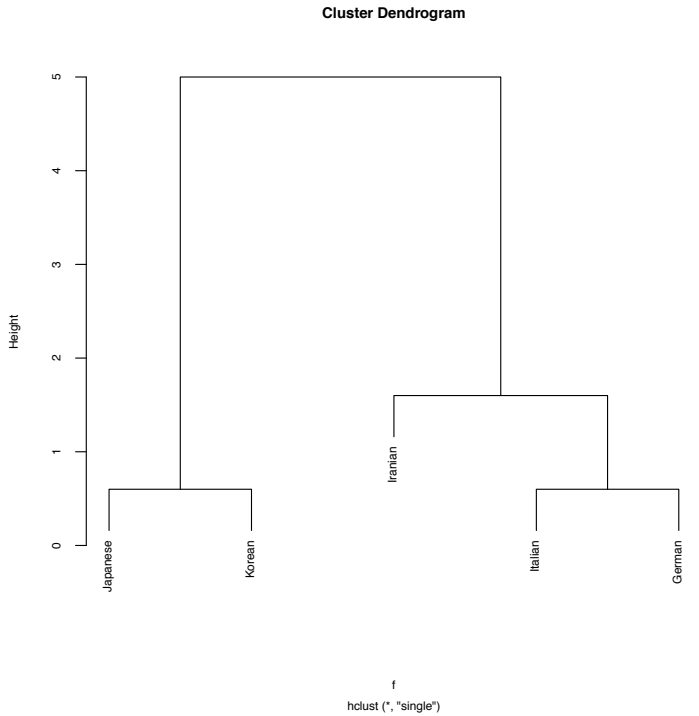


Figure 5: A dendrogram

PROBLEM 34. Figure 5 shows the dendrogram obtained by applying a hierarchical and agglomerative clustering algorithm on data that can be used to estimate the "evolutionary distance" between human beings of different races. In particular, the distance used was the one proposed by Cavalli-Sforza. For simplicity's sake, I only used five "ethnic groups". Comment on the dendrogram.