

Università degli Studi di Trento
Doctorate School in Materials Engineering
Methods of statistical and numerical analysis. Part I
Stefano Siboni
Final test - Trento, February 17th 2006

Part I - Discussion (1 hour)

Topic (1)

Unbiased estimates for the mean and the variance of a random variable.

Topic (2)

Linear regression analysis: statistical model, sample estimates and confidence intervals of regression parameters.

Topic (3)

Discuss one of the following kinds of hypothesis test:

- χ^2 -test for adapting a probability distribution to a sample;
- t -test for the mean of a normal population;
- F -test for the variance of a normal population.

Part II - Open book exercises (1 hour)**Exercise 1**

Repeated measurements of a certain quantity q have produced the set of data below

i	q_i	i	q_i
1	2.911	6	2.901
2	2.945	7	2.949
3	2.881	8	2.930
4	2.910	9	2.922
5	2.859	10	2.908

Assuming that the population of data is normal, calculate the confidence interval of the mean and that of the variance, at the confidence level 95%.

Exercise 2

Repeated measurements of a certain quantity ξ have led to the results listed below

10.46 10.10 10.78 9.98 10.32
10.35 10.46 10.89 10.07 10.60

After an appropriate thermal treatment the measurements have been carried out again, yielding the following values

10.99 11.30 10.26 10.85 11.00 10.58
10.80 11.20 11.00 10.95 10.42 11.12

Assuming that the two sets of data are normally distributed and that the two samples have the same variance, test the hypothesis that the value of ξ has changed after the thermal treatment, at a 5% significance level.

Exercise 3

By repeating the measurement of a pair of physical quantities x and y we have obtained the following table of data:

x_i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}
30.0	11.2	—	—	10.7	11.1
25.0	13.9	13.7	—	13.9	13.4
21.4	15.4	14.8	15.2	15.1	15.5
20.4	16.1	16.3	15.7	16.2	15.5
23.3	13.8	14.7	14.5	14.2	13.8

By assuming all the y data as independent normal variables with the same standard deviation, model the data by a best-fit straight line of the form

$$y = m + q(x - \bar{x})$$

where \bar{x} denotes the sample mean of the independent variables x_i 's and the model parameters m and q are calculated by a least squares fitting. In particular, determine:

- (i) the 95%-confidence interval of the intercept m ;
- (ii) the 95%-confidence interval of the slope q ;
- (iii) the 95%-confidence interval for the prediction of y at a given value $x = 24.5$ of the independent variable

Answer to Exercise 1

Number of data: $n = 10$

$$\text{Sample mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 2.9116$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 7.5471 \cdot 10^{-4}$$

Estimated standard deviation: $s = \sqrt{s^2} = 0.027472$

□ The CI of the mean is

$$\bar{x} - t_{[1-\frac{\alpha}{2}](n-1)} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{[1-\frac{\alpha}{2}](n-1)} \frac{s}{\sqrt{n}}$$

and for $\alpha = 0.05$, $n = 10$ has therefore the limits

$$\bar{x} - t_{[0.975](9)} \frac{s}{\sqrt{10}} = 2.9116 - 2.262 \cdot \frac{0.027472}{\sqrt{10}} = 2.89194$$

$$\bar{x} + t_{[0.975](9)} \frac{s}{\sqrt{10}} = 2.9116 + 2.262 \cdot \frac{0.027472}{\sqrt{10}} = 2.93125$$

so that the CI writes

$$2.89194 \leq \mu \leq 2.93125$$

or, equivalently,

$$[2.9116 \pm 0.0196]$$

□ The CI of the variance takes the form

$$\frac{1}{\chi^2_{[1-\frac{\alpha}{2}](n-1)}}(n-1)s^2 \leq \sigma^2 \leq \frac{1}{\chi^2_{[\frac{\alpha}{2}](n-1)}}(n-1)s^2$$

with $\alpha = 0.05$ and $n = 10$. Thus

$$\frac{1}{\chi^2_{[0.975](9)}} 9 s^2 = \frac{1}{19.023} \cdot 9 \cdot 7.5471 \cdot 10^{-4} = 3.56920 \cdot 10^{-4}$$

$$\frac{1}{\chi^2_{[0.025](9)}} 9 s^2 = \frac{1}{2.700} \cdot 9 \cdot 7.5471 \cdot 10^{-4} = 25.14700 \cdot 10^{-4}$$

and the CI becomes

$$3.56920 \cdot 10^{-4} \leq \sigma^2 \leq 25.14700 \cdot 10^{-4}.$$

Notice that the latter result corresponds to a 95%-level confidence interval for the standard deviation given by

$$1.88923 \cdot 10^{-2} \leq \sigma \leq 5.01468 \cdot 10^{-2}.$$

Answer to Exercise 2

We have to apply the t -test for the comparison of the mean of two normal populations. Since the samples are not ranked in a particular way and also contain a different number of data, we use the unpaired t -test.

The first sample consists of $p = 10$ data; its mean and variance are estimated as

$$\bar{x} = \frac{1}{p} \sum_{i=1}^p x_i = 10.4010$$

$$s_x^2 = \frac{1}{p-1} \sum_{i=1}^p (x_i - \bar{x})^2 = 0.09065444.$$

As for the second sample, it contains $q = 12$ data and has mean

$$\bar{y} = \frac{1}{q} \sum_{i=1}^q y_i = 10.8725$$

while the variance can be estimated as

$$s_y^2 = \frac{1}{q-1} \sum_{i=1}^q (y_i - \bar{y})^2 = 0.09771136.$$

It is noticeable that a comparison of the variance estimates satisfactorily support the assumption that the two populations have essentially the same variance.

Therefore, we carry out the test by reckoning the t variable

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{p} + \frac{1}{q}}}$$

where the weighted mean s^2 of the variances is defined by the formula

$$s^2 = \frac{(p-1)s_x^2 + (q-1)s_y^2}{p+q-2}.$$

In the present case we get

$$s^2 = \frac{(10-1)0.09065444 + (12-1)0.09771136}{10+12-2} = 0.09453575$$

and consequently

$$t = \frac{10.4010 - 10.8725}{\sqrt{0.09453575} \sqrt{\frac{1}{10} + \frac{1}{12}}} = -3.5814834$$

t will then be used as a test variable, with a rejection region given by the union of two intervals, corresponding to large absolute values of t

$$\left\{ t \leq -t_{[1-\frac{\alpha}{2}](p+q-2)} \right\} \cup \left\{ t \geq t_{[1-\frac{\alpha}{2}](p+q-2)} \right\}$$

Since $\alpha = 0.05$ and $p + q - 2 = 20$, we read on the table of cumulative probability distributions of Student's t :

$$t_{[1-\frac{\alpha}{2}](p+q-2)} = t_{[0.975](20)} = 2.086$$

and conclude that the means of the two distributions are likely to be **different**, because $|t| = 3.5814834 > 2.086 = t_{[0.975](20)}$.

Answer to Exercise 3

The system being homoscedastic, the chi-square fitting reduces to the usual least squares fitting and the best-fit estimates of the parameters m and q can be written in the form:

$$m = \frac{1}{n} \sum_{i=1}^n y_i = 14.3045 \quad q = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = -0.4995$$

with $n = 22$ and

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i = 23.431818 \\ \sum_{i=1}^n (x_i - \bar{x}) y_i &= -102.8731813 \\ \sum_{i=1}^n (x_i - \bar{x})^2 &= 205.9477272 \end{aligned}$$

The sum of squares around regression holds then:

$$\text{SSAR} = \sum_{i=1}^n [m + q(x_i - \bar{x}) - y_i]^2 = 2.22324623$$

while $\alpha = 0.05$. At a confidence level $1 - \alpha \in (0, 1)$ the confidence interval of the intercept μ and that of the slope κ are given by:

$$\mu = m \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\frac{1}{n} \frac{\text{SSAR}}{n-2}}$$

$$\kappa = q \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{-1} \frac{\text{SSAR}}{n-2}}$$

In the present case the confidence intervals become therefore:

$$\mu = m \pm t_{[0.975](20)} \sqrt{\frac{1}{22} \frac{\text{SSAR}}{20}}$$

$$\kappa = q \pm t_{[0.975](20)} \sqrt{\left[\sum_{i=1}^{22} (x_i - \bar{x})^2 \right]^{-1} \frac{\text{SSAR}}{20}}$$

where:

$$m = 14.3045$$

$$q = -0.4995$$

$$\text{SSAR} = 2.22324623$$

$$\sum_{i=1}^{22} (x_i - \bar{x})^2 = 205.9477272$$

$$t_{[0.975](20)} = 2.086$$

the latter 0.975-value of the Student-t variable with 20 degrees of freedom being obtained on a table.

As a conclusion:

(i) the 95%-confidence interval for the intercept μ is

$$14.305 \pm 0.148 = [14.15, 14.45]$$

(ii) the 95%-confidence interval for the slope κ holds

$$-0.4995 \pm 0.0485 = [-0.548, -0.451]$$

(iii) For a homoscedastic model, the $(1 - \alpha)$ -confidence interval for the prediction of $y = y_0$ at a given abscissa $x = x_0$ is expressed by the formula:

$$\mathbb{E}(y_0) = m + q(x_0 - \bar{x}) \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{V} \sqrt{\frac{\text{SSAR}}{n-2}}$$

where:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{SSAR} = \sum_{i=1}^n [-y_i + m + q(x_i - \bar{x})]^2$$

$$V = 1 + \frac{1}{n} + \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} (x_0 - \bar{x})^2.$$

In the present case we have $\bar{x} = 23.431818$, so that:

$$y_0 = 14.3045 - 0.4995(x_0 - 23.431818) \pm t_{[0.975](20)} \cdot \sqrt{1 + \frac{1}{22} + \frac{1}{205.9477272} (x_0 - 23.431818)^2} \sqrt{\frac{2.22324623}{20}}$$

and the confidence interval for the prediction of y at $x = x_0$ reduces to:

$$y_0 = 14.3045 - 0.4995(x_0 - 23.431818) \pm 0.69549 \sqrt{1.04545 + 0.0048556(x_0 - 23.431818)^2}$$

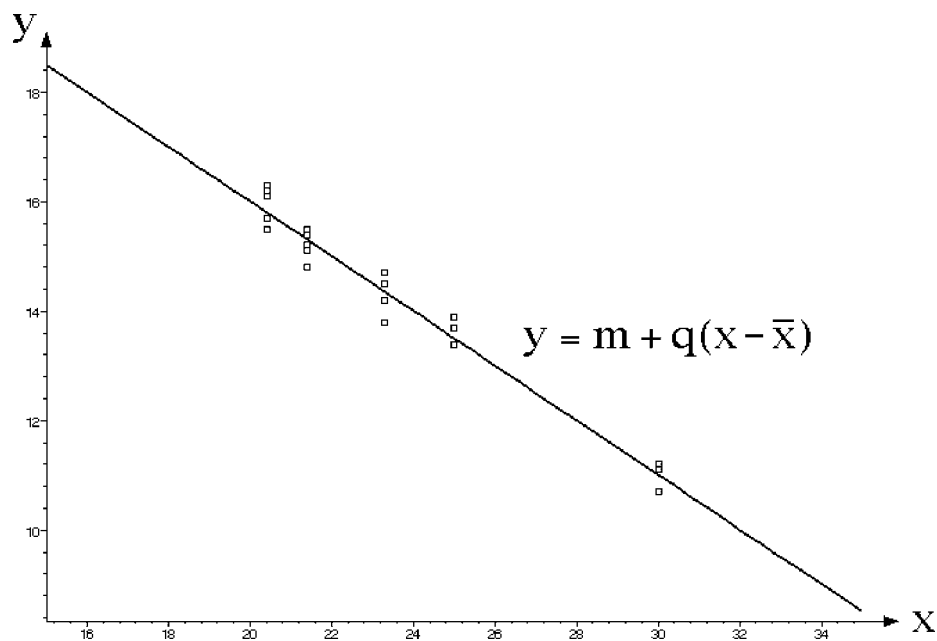
The CI of y at $x = x_0 = 24.5$ is finally calculated as

$$y_0 = 14.3045 - 0.4995(24.5 - 23.431818) \pm \\ \pm 0.69549 \sqrt{1.04545 + 0.0048556(24.5 - 23.431818)^2}$$

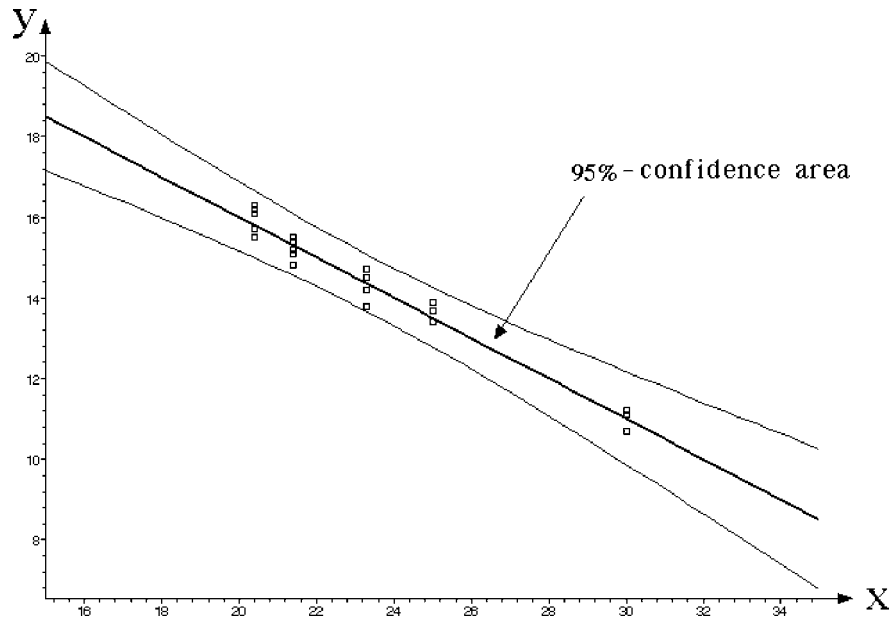
and writes therefore:

$$y_0 = [13.77 \pm 0.71] = [13.06, 14.48]$$

In the following picture the regression line is superimposed to the experimental data (represented by dots):



The confidence region for predictions (at a confidence level of 95%) is put into evidence in the figure below:



For clarity's sake, the factor V of the confidence intervals for predictions has been magnified.