

Part I - Discussion (1 hour)**Topic (1)**

Confidence interval for the mean and the standard deviation of a normal population.

Topic (2)

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

Topic (3)

Discuss a specific kind of hypothesis test among the following ones:

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

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

The following table of data has been obtained by repeated measurements of a certain quantity x

| i | x_i | i | x_i |
|-----|-------|-----|-------|
| 1 | 0.945 | 6 | 0.859 |
| 2 | 930 | 7 | 949 |
| 3 | 881 | 8 | 911 |
| 4 | 910 | 9 | 922 |
| 5 | 901 | 10 | 908 |

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

Exercise 2

The following data are extracted from a normal population:

147 , 144 , 146 , 148 , 138 ,
144 , 145 , 158 , 145 , 143

Apply Chauvenet criterion to detect the possible outlier.

Exercise 3

An experiment has provided the table of data below:

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

All the y data are assumed to be 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 must be calculated with the least squares method. 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: $n = 10$

$$\text{Sample mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 0.9116$$

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

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

$$\bar{x} - t_{[0.975](9)} \frac{s}{\sqrt{10}} = 0.9116 - 2.262 \cdot \frac{0.027472}{\sqrt{10}} = 0.89194$$

$$\bar{x} + t_{[0.975](9)} \frac{s}{\sqrt{10}} = 0.9116 + 2.262 \cdot \frac{0.027472}{\sqrt{10}} = 0.93125$$

so that the CI writes

$$0.89194 \leq \mu \leq 0.93125$$

or, equivalently,

$$[0.9116 \pm 0.0196]$$

□ 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

The sample mean and standard deviation are given by

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 145.8 \quad s = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 5.07$$

The absolute values of the residuals with respect to the mean are listed in the table below

| x_i | $ x_i - \bar{x} $ | x_i | $ x_i - \bar{x} $ |
|-------|-------------------|-------|-------------------|
| 147 | 1.2 | 144 | 1.8 |
| 144 | 1.8 | 145 | 0.8 |
| 146 | 0.2 | 158 | 12.2 |
| 148 | 2.2 | 145 | 0.8 |
| 138 | 7.8 | 143 | 2.8 |

from which it is apparent that the datum $x_8 = 158$ turns out to be very scattered and therefore suspect. The distance of the suspect value from the mean, in units of s , holds

$$\frac{x_8 - \bar{x}}{s} = \frac{158 - 145.8}{5.1} = 2.40$$

By abruptly identifying the sample mean \bar{x} with the population mean μ and the sample variance s^2 with the corresponding population variance σ^2 , the probability that a measurement result occurs by chance at a distance larger than 2.40 standard deviations from the mean can be calculated from the Table of the cumulative distribution function of the standard normal distribution:

$$\begin{aligned} P(|x_8 - \bar{x}| \geq 2.40s) &= 1 - P(|x_8 - \bar{x}| < 2.40s) = \\ &= 1 - 2 \cdot P(\bar{x} \leq x_8 < \bar{x} + 2.40s) = \\ &= 1 - 2 \cdot 0.49180 = 0.0164 \end{aligned}$$

Out of 10 measurements we typically expect $10 \cdot 0.0164 = 0.164$ “bad” results, at a distance larger than $2.40s$ from the mean. **Since $0.164 < 1/2$, Chauvenet criterion suggests that x_8 should be rejected as an outlier.**

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:

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

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

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

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

$$y_0 = 14.3045 - 0.4995(x_0 - 23.431818) \pm t_{[0.975](20)} \cdot$$

$$\cdot \sqrt{1 + \frac{1}{22} + \frac{1}{205.9477272}(x_0 - 23.431818)^2} \sqrt{\frac{2.22324623}{20}}$$

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

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

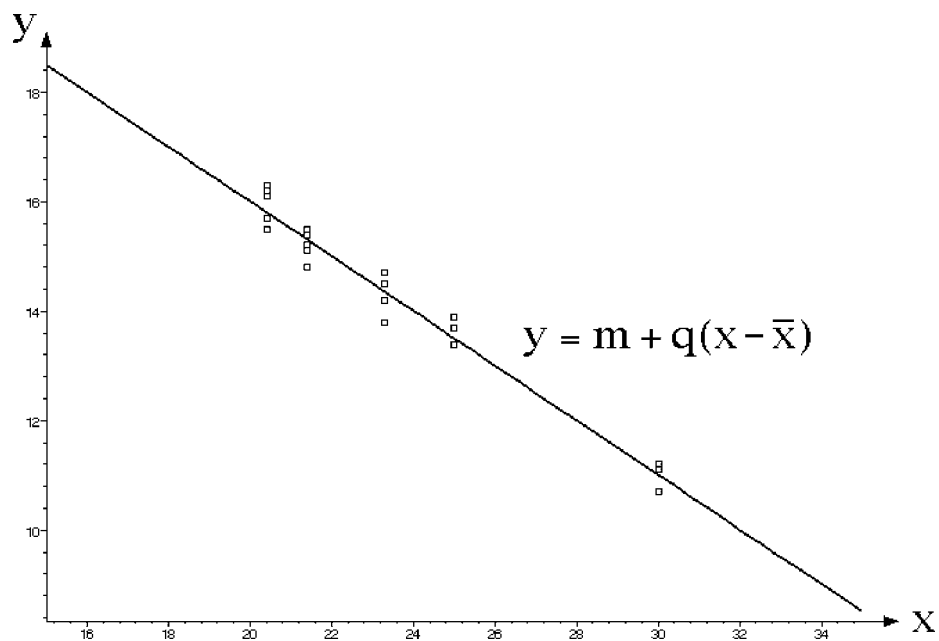
The confidence interval of y at $x = x_0 = 24.5$ is then calculated as

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

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, by enhancing the factor V of the confidence intervals for predictions:

