

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

Provide a general illustration of the main ideas on which the hypothesis testing is based. Discuss, in particular, a specific type of 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.

Topic (2)

Small samples: determination of the confidence interval for the mean of a normal population.

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

The repeated measurement of a quantity x yielded the following table of data (arbitrary units).

i	x_i	i	x_i
1	2.850	7	2.769
2	835	8	843
3	785	9	814
4	812	10	819
5	807	11	803
6	841	12	821

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

Exercise 2

Consider the following temperature data (expressed in K):

$$247, 244, 246, 248, 238, \\ 244, 245, 258, 245, 243$$

which can be assumed to be extracted from a normal population. By applying Chauvenet criterion, check whether the odd result $x_{\text{sus}} = 258$ can be regarded as an outlier.

Exercise 3

The following (x, y) data have been obtained by an experiment:

x_i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}
1.04	1.57	1.63	1.61	1.55	1.62
1.14	1.52	1.48	1.54	1.51	1.55
1.33	1.47	1.38	1.45	1.38	1.42
1.50	1.34	1.39	×	1.37	1.39
2.00	×	1.12	1.07	×	1.11

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 $x = 1.45$ of the independent variable

Answer to Exercise 1

Number of data: $n = 12$

$$\text{Sample mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 2.81658$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 5.7645 \cdot 10^{-4}$$

Estimated standard deviation: $s = \sqrt{s^2} = 0.024009$

□ 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 $\alpha = 0.05$, $n = 12$ has therefore the limits

$$\bar{x} - t_{[0.975](11)} \frac{s}{\sqrt{12}} = 2.81658 - 2.201 \cdot \frac{0.024009}{\sqrt{12}} = 2.8013$$

$$\bar{x} + t_{[0.975](11)} \frac{s}{\sqrt{12}} = 2.81658 + 2.201 \cdot \frac{0.024009}{\sqrt{12}} = 2.8318$$

so that the CI writes

$$2.8013 \leq \mu \leq 2.8318$$

or, equivalently,

$$[2.8165 \pm 0.0152]$$

□ 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 $n = 12$. Thus

$$\frac{1}{\chi^2_{[0.975](11)}} 11 s^2 = \frac{1}{21.920} 11 \cdot 5.7645 \cdot 10^{-4} = 2.89277 \cdot 10^{-4}$$

$$\frac{1}{\chi^2_{[0.025](11)}} 11 s^2 = \frac{1}{3.816} 11 \cdot 5.7645 \cdot 10^{-4} = 16.61675 \cdot 10^{-4}$$

and the CI becomes

$$2.89277 \cdot 10^{-4} \leq \sigma^2 \leq 16.61675 \cdot 10^{-4}$$

Answer to Exercise 2

The sample mean and standard deviation are given by

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 245.8 \quad s = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 5.1$$

The distance of the suspect value from the mean, in units of s , holds

$$\frac{x_{\text{sus}} - \bar{x}}{s} = \frac{258 - 245.8}{5.1} = 2.4$$

The probability that a measurement falls at a distance larger than 2.4 standard deviations from the mean can be calculated from the Table of the cumulative distribution function of the standard normal distribution:

$$\begin{aligned} P(|x_{\text{sus}} - \bar{x}| \geq 2.4s) &= 1 - P(|x_{\text{sus}} - \bar{x}| < 2.4s) = \\ &= 1 - 2 \cdot P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.4s) = \\ &= 1 - 2 \cdot 0.49180 = 0.0164 \end{aligned}$$

Out of 10 measurements we typically expect $10 \cdot 0.0164 = 0.164$ “bad” results, at a distance larger than $2.4s$ from the mean. **Since $0.164 < 1/2$, Chauvenet criterion suggests that x_{sus} should be rejected as an outlier.**

Answer to Exercise 3

Since the standard deviations are assumed to be the same, 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:

$$m = \frac{1}{n} \sum_{i=1}^n y_i = 1.4305 \quad q = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = -0.4995$$

with $n = 22$. The sum of squares around regression holds then:

$$\text{SSAR} = \sum_{i=1}^n [m + q(x_i - \bar{x}) - y_i]^2 = 0.02223246$$

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:

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

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

with:

$$m = 1.4305$$

$$q = -0.4995$$

$$\text{SSAR} = 0.02223246$$

$$\sum_{i=1}^{22} (x_i - \bar{x})^2 = 2.05947727$$

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

As a conclusion:

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

$$1.4305 \pm 0.0148 = [1.415, 1.445]$$

(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 $\bar{x} = 1.34318$, so that:

$$y_0 = 1.4305 - 0.4995(x_0 - 1.34318) \pm t_{[0.975](20)} \cdot \sqrt{1 + \frac{1}{22} + \frac{1}{2.05947727} (x_0 - 1.34318)^2} \sqrt{\frac{0.02223246}{20}}$$

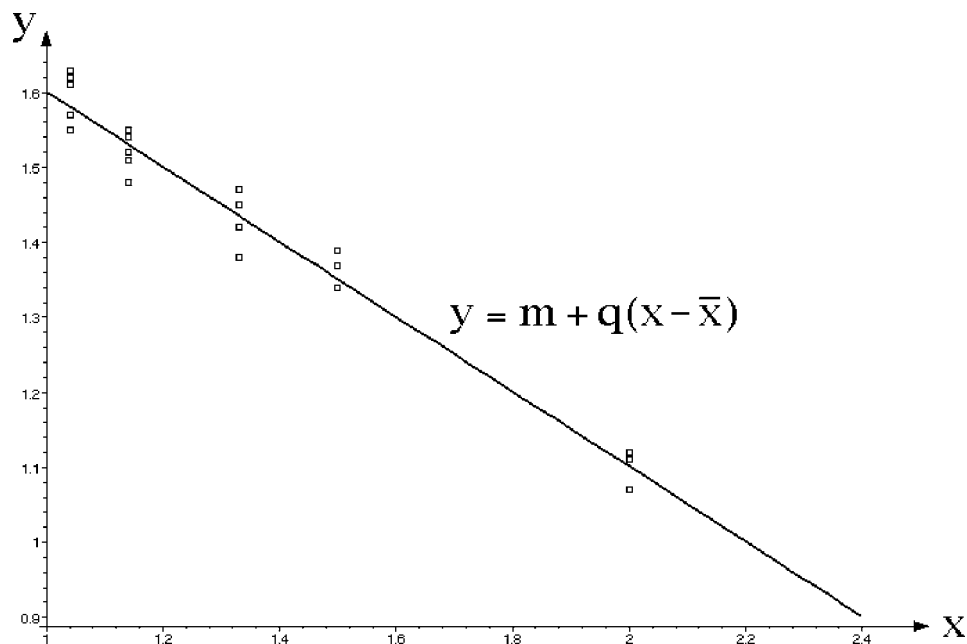
and the confidence interval for the prediction of y at $x = x_0$ reduces to:

$$y_0 = 1.4305 - 0.4995(x_0 - 1.34318) \pm 0.06955 \sqrt{1.04545 + 0.48556(x_0 - 1.34318)^2}$$

The confidence interval of y at $x = x_0 = 1.45$ writes therefore:

$$y_0 = [1.377 \pm 0.075] = [1.302, 1.452]$$

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



The confidence region for predictions (at a confidence level of 95%) is evidenced in the figure below (factor V exaggerated)

