Part I - Discussion (1 hour)

Topic (1)
Illustrate the general ideas of hypothesis testing. Discuss, in particular, a specific test among the following ones:

- $\mathcal{X}^2$-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;
- paired t-test for comparing the mean values of two normal populations.

Topic (2)
Determination of the confidence interval for the mean and the variance of a normal population by means of a small sample.

Part II - Open book exercises (1 hour)

Exercise 1

The following table of data has been obtained by repeated measurements of a quantity x (arbitrary units).

i	x_i	i	x_i
1	1.807	7	1.850
2	1.835	8	1.843
3	1.785	9	1.812
4	1.814	10	1.819
5	1.769	11	1.841
6	1.803	12	1.821

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 mass data (expressed in Kg):

$$194\ , \quad 193\ , \quad 208\ , \quad 198\ , \quad 188\ ,$$
$$194\ , \quad 195\ , \quad 196\ , \quad 193\ , \quad 195$$

which can be assumed to be extracted from a normal population. The datum $x_{\mathrm{sus}} = 208$ is suspect, as surprisingly far from the mean. By applying Chauvenet criterion, check whether the datum 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}
1.00	×	0.12	0.07	×	0.11
0.50	0.34	0.39	×	0.37	0.39
0.33	0.47	0.38	0.45	0.38	0.42
0.14	0.52	0.48	0.54	0.51	0.55
0.04	0.57	0.63	0.61	0.55	0.62

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

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

where $\overline{x}$ 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 $x = 0.45$ of the independent variable

Answer to Exercise 1 (Sottrarre 1 alle x)

Number of data: $n = 12$

Sample mean: $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = 2.81658$

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

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

□ The CI of the mean is

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

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

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

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

so that the CI writes

$$2.8013 \le \mu \le 2.8318$$

or, equivalently,

$$[2.8165 \pm 0.0152]$$

☐ The CI of the variance takes the form

$$\frac{1}{\mathcal{X}^2{}_{[1-\frac{\alpha}{2}](n-1)}}(n-1)s^2 \le \sigma^2 \le \frac{1}{\mathcal{X}^2{}_{[\frac{\alpha}{2}](n-1)}}(n-1)s^2$$

with $\alpha = 0.05$ and $n = 12$. Thus

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

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

and the CI becomes

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

Answer to Exercise 2 (Sottrarre 50 alle x)

The sample mean and standard deviation are given by

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

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

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

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:

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

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 (Sottrarre 1 alle x ed alle y)

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:

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

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

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

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:

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

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

In the present case the confidence intervals become therefore:

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

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

with:

$$\begin{aligned} m &= 1.4305 \\ q &= -0.4995 \\ \mathrm{SSAR} &= 0.02223246 \\ \sum_{i=1}^{22}(x_i - \overline{x})^2 &= 2.05947727 \\ t_{[0.975](20)} &= 2.086 \end{aligned}$$

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:

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

where:

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

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

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

In the present case we have $\overline{x} = 1.34318$, so that:

$$y_0 = 1.4305 - 0.4995(x_0 - 1.34318) \pm t_{[0.975](20)} \cdot$$

$$\cdot \sqrt{1 + \frac{1}{22} + \frac{1}{2.05947727}(x_0 - 1.34318)^2} \sqrt{\frac{0.02223246}{20}}$$

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

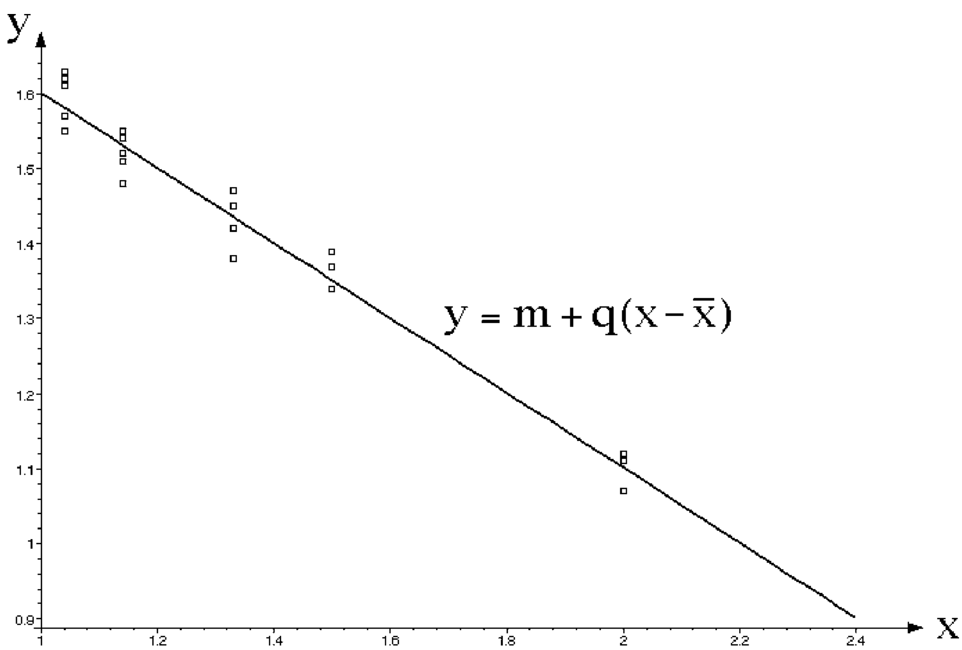
$$y_0 = 1.4305 - 0.4995(x_0 - 1.34318) \pm$$

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

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)