
University of Trento
Department of Materials Engineering and Industrial Technologies

Ph. D. in Materials Engineering
Methods of statistical and numerical analysis
Final test on statistical methods

Part I - Discussion (1 hour)

Topic (1)

Sample estimates for the mean and the variance of a statistical population. Confidence intervals for the mean and the variance of a normal random variable.

Topic (2)

Linear regression of experimental data by the chi-square fitting: the standard approach to linear regression by the chi-square fitting method.

Part I - Exercises (1 hour 30')**Exercise 1**

As a result of repeated measurements of mass the following table of data has been obtained (g):

n	M_n	n	M_n
1	106.2	9	72.1
2	99.5	10	81.7
3	108.3	11	93.5
4	88.9	12	90.1
5	86.5	13	97.5
6	95.0	14	96.9
7	105.8	15	81.3
8	89.0	16	94.1

By assuming that the statistical population is normal, determine the confidence interval of the mean at the confidence level of 90%. Reckon also the confidence interval of the standard deviation at the confidence level 95%.

Exercise 2

A normal sample of temperature data (in K) is listed below:

323.2 , 324.7 , 325.8 , 324.2 , 324.6 , 326.9 ,
325.4 , 323.8 , 323.1 , 324.7 , 324.1 , 323.3 .

Test, by using Chauvenet criterion, whether the sample may contain an outlier not belonging to the statistical population.

Exercise 3

Some repeated measurements of the temperature T ($^{\circ}\text{C}$) and the surface tension γ (mN m^{-1}) of a liquid have provided the following table of data (different γ values associated to the same value of T mean that the measurement of the surface tension has been carried out many times for the same temperature):

T	γ	γ	γ	γ	γ
35.0	68.0	68.3	67.9	×	×
30.0	69.9	69.5	70.0	71.2	69.8
25.0	71.8	71.6	72.0	71.9	72.4
20.0	73.1	72.9	72.6	72.7	72.5
15.0	74.0	73.9	74.2	74.4	73.7
10.0	74.9	75.3	75.0	74.8	×

While the random error on the temperatures can be regarded as negligible, all the surface tension 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

$$\gamma = m + q(T - \bar{T}),$$

where \bar{T} denotes the sample mean of the temperatures and the model parameters m and q are calculated by the least squares method. In particular, determine:

- (i) the 95%-confidence interval of the coefficient m ;
- (ii) the 95%-confidence interval of the slope q ;
- (iii) the 95%-confidence interval for the prediction of γ at a temperature $T = 28.0$ $^{\circ}\text{C}$;
- (iv) the goodness of fit Q of the model if the common value of the standard deviation on γ were known to be $\sigma = 0.3$.

Solution of Exercise 1

The number of data of the sample is clearly $n = 16$, so that it is easy to compute the sample mean:

$$\bar{M} = \frac{1}{n} \sum_{i=1}^n M_i = 92.900$$

while the sample variance holds

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (M_i - \bar{M})^2 = 96.00266667$$

and provides the estimated standard deviation:

$$s = \sqrt{s^2} = 9.798095053$$

□ The CI of the mean is

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

In the present case the confidence level for the confidence interval of the mean must be $1 - \alpha = 90\%$, so that $\alpha = 0.10$; the number of datapoints is $n = 16$. Therefore the confidence interval has the limits

$$\begin{aligned} \bar{M} - t_{[0.95](15)} \frac{s}{\sqrt{16}} &= 92.900 - 1.753 \cdot \frac{9.798095053}{\sqrt{16}} = \\ &= 88.60598484 \end{aligned}$$

$$\begin{aligned} \bar{M} + t_{[0.95](15)} \frac{s}{\sqrt{16}} &= 92.900 + 1.753 \cdot \frac{9.798095053}{\sqrt{16}} = \\ &= 97.19401516 \end{aligned}$$

and writes

$$88.606 \leq \mu \leq 97.194$$

or, equivalently,

$$[92.900 \pm 4.294]$$

□ The CI of the variance can be written as

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

with $\alpha = 0.05$ and $n = 16$. Thus

$$\frac{1}{\chi^2_{[0.975](15)}} 15 s^2 = \frac{1}{27.488} 15 \cdot 96.00266667 = 52.38795111$$

$$\frac{1}{\chi^2_{[0.025](15)}} 15 s^2 = \frac{1}{6.262} 15 \cdot 96.00266667 = 229.9648675$$

and the CI of the variance becomes

$$52.38795111 \leq \sigma^2 \leq 229.9648675$$

Taking the square root we obtain the CI of the standard deviation

$$7.237952135 \leq \sigma \leq 15.16459256$$

which can also be put into the equivalent form

$$[11.201 \pm 3.963]$$

The main calculations involved in the previous analysis are summarized in the table below:

T	DT	DT ²	
106,2	13,300	176,890	
99,5	6,600	43,560	
108,3	15,400	237,160	
88,9	-4,000	16,000	
86,5	-6,400	40,960	
95,0	2,100	4,410	
105,8	12,900	166,410	
89,0	-3,900	15,210	
72,1	-20,800	432,640	
81,7	-11,200	125,440	
93,5	0,600	0,360	
90,1	-2,800	7,840	
97,5	4,600	21,160	
96,9	4,000	16,000	
81,3	-11,600	134,560	
94,1	1,200	1,440	
mean: 92,900		variance: 96,0026667	
		st.dev.: 9,79809505	
Lower limit of the confidence interval for the mean:		88,6059848	
Upper limit of the confidence interval for the mean:		97,1940152	
Estimate of the mean:		92,9	
Estimate of the absolute error:		4,29401516	
Lower limit of the confidence interval for the variance:		52,3879511	
Upper limit of the confidence interval for the variance:		229,964867	
Lower limit of the confidence interval for the st. dev.:		7,23795213	
Upper limit of the confidence interval for the st. dev.:		15,1645926	

Solution of Exercise 2

The sample mean and standard deviation of the $n = 12$ data are given by

$$\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = 324.48 \quad s = \sqrt{\frac{1}{11} \sum_{i=1}^{12} (x_i - \bar{x})^2} = 1.1352$$

The farthest datum from the mean \bar{x} is 326.9, and we want to check if it such an outlier actually belongs to the statistical population of the sample or not. The distance of the suspect value from the mean, in units of s , holds

$$\frac{x_{\text{sus}} - \bar{x}}{s} = \frac{326.9 - 324.48}{1.1352} = 2.1288$$

The probability that a measurement falls at a distance larger than 2.1288 standard deviations from the mean can be calculated from the Table of the cumulative distribution function $P_z(z)$ of the standard normal distribution:

$$\begin{aligned}
 P(|x_{\text{sus}} - \bar{x}| \geq 2.1288s) &= 1 - P(|x_{\text{sus}} - \bar{x}| < 2.1288s) = \\
 &= 1 - 2 \cdot P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.1288s) \\
 &= 1 - 2 \cdot P_z(2.1288) = \\
 &= 1 - 2 \cdot 0.48336 = 0.03328
 \end{aligned}$$

by using a simple linear interpolation scheme:

z	$P_z(z)$
2.1200	0.48300
2.1288	$P_z(2.1288)$
2.1300	0.48341

$$\frac{2.1288 - 2.1200}{2.1300 - 2.1200} = \frac{P_z(2.1288) - 0.48300}{0.48341 - 0.48300}$$

The typical number of results at a distance larger than 2.1288s from the mean out of 12 measurements must be $12 \cdot 0.03328 = 0.3994$. **Since $0.3994 < 1/2$, Chauvenet criterion suggests that $x_{\text{sus}} = 326.90$ should be rejected as not belonging to the statistical population.**

Let us summarize in a table the previous calculations:

x	Dx	ABS(Dx)	Dx^2		
323,20	-1,2833	1,2833	1,646944		
324,70	0,2167	0,2167	0,046944		
325,80	1,3167	1,3167	1,733611		
324,20	-0,2833	0,2833	0,080278		
324,60	0,1167	0,1167	0,013611		
326,90	2,4167	2,4167	5,840278		
325,40	0,9167	0,9167	0,840278		
323,80	-0,6833	0,6833	0,466944		
323,10	-1,3833	1,3833	1,913611		
324,70	0,2167	0,2167	0,046944		
324,10	-0,3833	0,3833	0,146944		
323,30	-1,1833	1,1833	1,400278		
mean(x):	324,48	st.dev.(x):	1,1352		
Maximum of ABS(Dx):	2,4167	corresponding to the value of x:		Outlier 326,90	
Distance z of the outlier from the mean in standard deviation units:			2,1288		
Probability of a larger distance from the mean (see table):			z	area fro 0 to z	
			2,1200	0,48300	0,03400
			2,1300	0,48341	0,03318
Linearly interpolated value:			2,1288	0,48336	0,03328
Mean number of expected events out of 12 measurements:					0,3994

The mean number of outliers is smaller than 1/2. Thus, according to Chauvenet criterion, the value 326,90 must be rejected as not belonging to the population.

Solution of Exercise 3

Since the standard deviations of the $n = 27$ data 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:

$$m = \frac{1}{n} \sum_{i=1}^n \gamma_i = 72.159259$$

$$q = \frac{\sum_{i=1}^n (T_i - \bar{T}) \gamma_i}{\sum_{i=1}^n (T_i - \bar{T})^2} = -0.265189$$

where

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = 22.03703704$$

The sum of squares around regression holds then:

$$\text{SSAR} = \sum_{i=1}^n [m + q(T_i - \bar{T}) - \gamma_i]^2 = 6.220540553$$

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:

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

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

In the present case $\alpha = 0.05$, $n = 27$ and the confidence intervals become therefore:

$$\mu = m \pm t_{[0.975](25)} \sqrt{\frac{1}{27} \frac{\text{SSAR}}{25}}$$

$$\kappa = q \pm t_{[0.975](25)} \sqrt{\left[\sum_{i=1}^{27} (T_i - \bar{T})^2 \right]^{-1} \frac{\text{SSAR}}{25}}$$

with:

$$\begin{aligned} m &= 72.15925926 \\ q &= -0.265189191 \\ \text{SSAR} &= 6.220540553 \end{aligned}$$

$$\sum_{i=1}^{27} (T_i - \bar{T})^2 = 1712.962964$$

$$t_{[0.975](25)} = 2.060$$

where the percentile can be found on the Table of Student's cumulative probability distribution or more accurately calculated by the Maple command line:

$$\text{statevalf[icdf, chisquare[25]](0.975);}$$

which provides $t_{[0.975](25)} = 2.059538553$.

As a consequence:

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

$$72.15926 \pm 0.19776 = [71.96150, 72.35702]$$

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

$$-0.26519 \pm 0.02483 = [-0.29002, -0.24036]$$

(iii) For a homoscedastic model, the $(1 - \alpha)$ -confidence interval for the prediction of $\gamma = \gamma_0$ at a given temperature $T = T_0$ is expressed by the formula:

$$\mathbb{E}(\gamma_0) = m + q(T_0 - \bar{T}) \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{V} \sqrt{\frac{\text{SSAR}}{n-2}}$$

where:

$$\begin{aligned}\bar{T} &= \frac{1}{n} \sum_{i=1}^n T_i = 22.03703704 \\ \text{SSAR} &= \sum_{i=1}^n [-\gamma_i + m + q(T_i - \bar{T})]^2 = 6.220540553 \\ V &= 1 + \frac{1}{n} + \frac{1}{n \sum_{i=1}^n (T_i - \bar{T})^2} (T_0 - \bar{T})^2 = \\ &= 1 + \frac{1}{27} + \frac{1}{1712.962964} (T_0 - 22.03703704)^2.\end{aligned}$$

so that:

$$\begin{aligned}\gamma_0 &= 72.15926 - 0.26519(T_0 - 22.03704) \pm t_{[0.975](25)} \cdot \\ &\cdot \sqrt{1 + \frac{1}{27} + \frac{1}{1712.962964} (T_0 - 22.03703704)^2} \sqrt{\frac{6.220540553}{25}}\end{aligned}$$

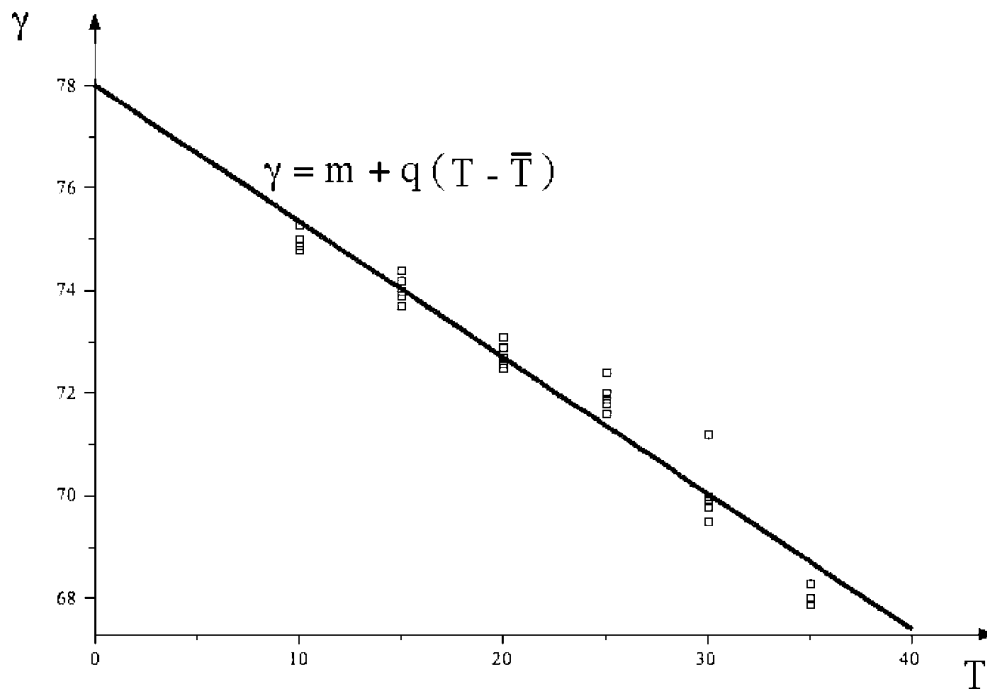
and since $t_{[0.975](25)} = 2.060$ the confidence interval for the prediction of y at $T = T_0$ reduces to:

$$\begin{aligned}\gamma_0 &= 72.15926 - 0.26519(T_0 - 22.03704) \pm \\ &\pm 1.02757 \sqrt{1.03704 + 0.0005837837834(T_0 - 22.03704)^2}\end{aligned}$$

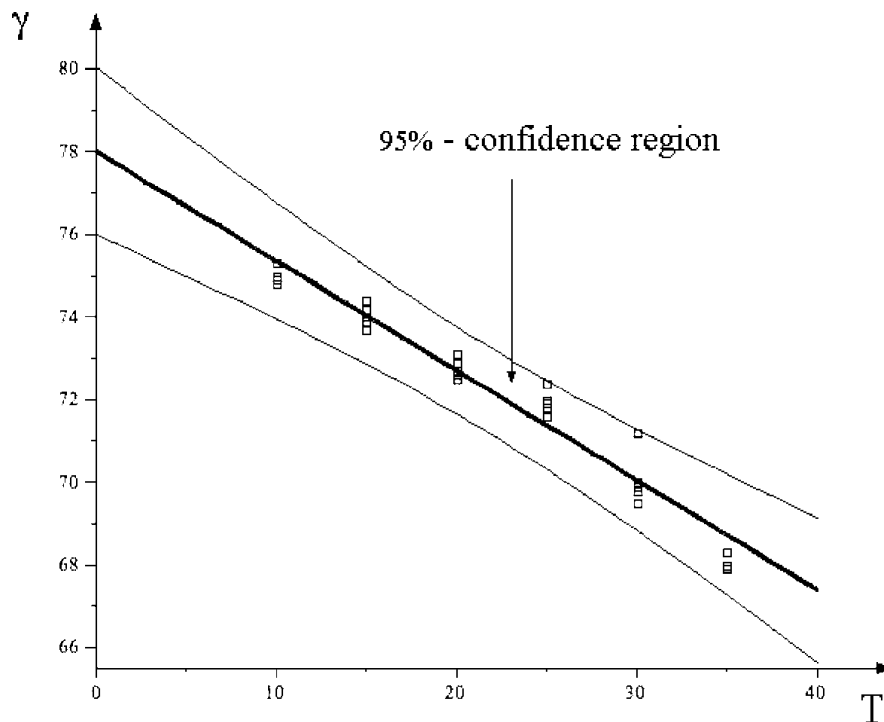
The confidence interval of γ at $T = T_0 = 28.0$ is therefore:

$$\gamma_0 = [70.578 \pm 1.057] = [69.521, 71.635]$$

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



where the factor V is magnified for clarity.

(iv) If the standard deviation of the γ data holds $\sigma = 0.3$, the normalized sum of squares around regression can be calculated as

$$\begin{aligned} \text{NSSAR} &= \frac{1}{\sigma^2} \sum_{i=1}^n [-\gamma_i + m + q(T_i - \bar{T})]^2 = \frac{1}{\sigma^2} \text{SSAR} = \\ &= \frac{1}{0.3^2} \cdot 6.220540553 = 69.11711725 \end{aligned}$$

and by definition the goodness of fit of the model is given by the integral

$$\begin{aligned} Q &= \int_{69.11711725}^{+\infty} p_{n-2}(\mathcal{X}^2) d\mathcal{X}^2 = \\ &= \int_{69.11711725}^{+\infty} p_{25}(\mathcal{X}^2) d\mathcal{X}^2 = \\ &= 1 - \int_0^{69.11711725} p_{25}(\mathcal{X}^2) d\mathcal{X}^2 = \\ &= 1 - 0.9999948002 = 5.1998 \cdot 10^{-6}. \end{aligned}$$

The integral over $[0, \text{NSSAR}]$ of the \mathcal{X}^2 probability distribution with 25 d.o.f. can be determined, for instance, by means of the Maple command line:

```
statevalf[cdf, chisquare[25]](69.11711725);
```

which computes the cumulative probability distribution of the random variable at NSSAR, or estimated by a linear extrapolation of tabulated data (although the estimate is very rough in the latter case, the accuracy is sufficient for our purposes).

The goodness of fit turns out to be very small, which, in principle, should suggest that **the regression model is probably incorrect and must be rejected.**

It should be noticed, however, that if the model were correct the NSSAR would follow a χ^2 distribution with $\nu = n - 2 = 25$ d.o.f. and the expected value of the random variable should be contained within the 99.7%-confidence interval

$$[\nu - 3\sqrt{2\nu}, \nu + 3\sqrt{2\nu}] = [3.7868, 46.2132]$$

while this is not the case, since $\text{NSSAR} = 69.11711725$.

This means that probably either the γ data are not normal or the common standard deviation σ has been underestimated (or both).

Suffice it to say that if the correct estimate of σ were 0.5, instead of 0.3, the value of NSSAR would become

$$\text{NSSAR} = \frac{1}{0.5^2} \text{SSAR} = \frac{1}{0.5^2} \cdot 6.220540553 = 24.88216221$$

so that the command

$$\text{statevalf}[\text{cdf}, \text{chisquare}[25]](24.88216221)$$

would provide a very high value of Q :

$$Q = 1 - 0.5310069109 = 0.4689930891$$

and make the regression model acceptable.