

First name **Family name**

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

17 April 2008

Part I - Discussion (1 hour and 1/4)

Topic (1)

The basic ideas of the linear regression technique by χ^2 -fitting: model, estimate of model parameters, goodness of fit

Topic (2)

Hypothesis testing: basic definitions and general strategy of the tests. Discuss also, in particular, two tests among those listed below:

- unpaired t -test for comparing the mean values of two normal populations;
- one-sided and two-sided t -test for the mean of a normal population;
- χ^2 -test for adapting a probability distribution to a sample;
- one-sided and two-sided F -test for the variance of a normal population;
- Kolmogorov-Smirnov test.

Part II - Open book exercises (1 hour and 1/4)**Exercise 1**

Owing to the poor epidermic transpiration, birds usually remove the excess of heat of their body by panting. In order to check if there is a relationship between the body temperature and the respiratory frequency, we consider a random sample of 15 ravens, in different environmental conditions, and for each individual we measure the body temperature T (in $^{\circ}C$) and the respiratory frequency f (in breathes per minute). The sample is assumed to be normal. The experimental results are summarized in the table below.

individual i	T_i	f_i
1	39.6	33
2	40.1	50
3	41.7	75
4	39.0	13
5	41.9	68
6	42.8	115
7	40.3	52
8	39.0	33
9	39.7	60
10	39.3	28
11	37.9	74
12	37.9	39
13	37.9	76
14	37.9	30
15	37.9	58

Test the hypothesis that the variables T and f are independent

against the alternative hypothesis that a linear correlations exists, with a significance level (a) of 5% and (b) of 1%.

Exercise 2

Assuming that the length data below (in m) belong to a normal population

694 , 690 , 692 , 696 , 676 ,
688 , 718 , 681 , 688 , 686

apply Chauvenet criterion to check whether the sample may contain an outlier.

Exercise 3

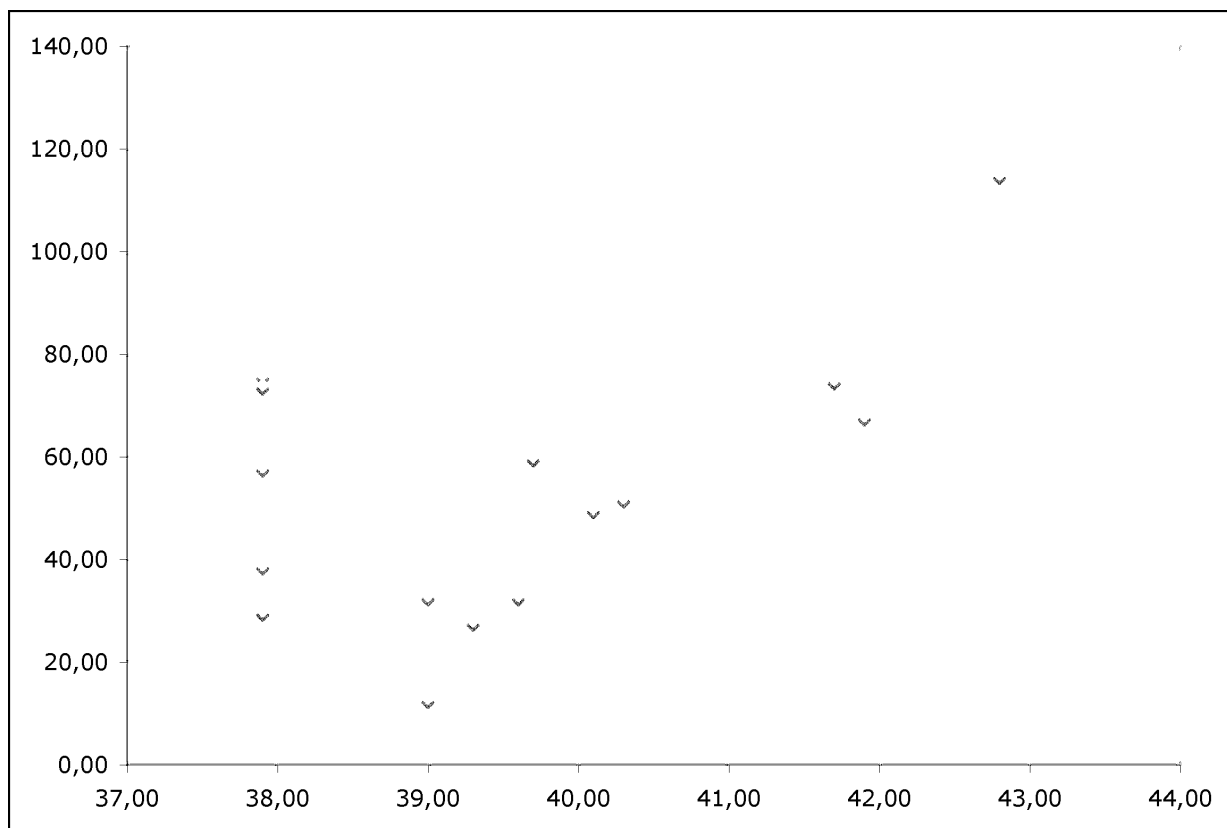
The repeated measurement of a temperature T led to the following table of data (in K):

i	T_i	i	T_i
1	306.5	9	292.1
2	310.5	10	293.3
3	297.3	11	286.5
4	278.9	12	290.1
5	295.0	13	296.9
6	289.0	14	295.3
7	285.4	15	291.7
8	280.7	16	299.5

Assuming that the population is normal, compute the confidence interval of the mean at a confidence level of 99% and the confidence interval of the standard deviation at a confidence level of 95%.

Answer to Exercise 1

Since the raven sample is completely random, both temperature T and respiratory frequency f are random variables and the problem of their stochastic independence can be tackled by analysing Pearson's linear correlation coefficient r . Notice that the variables are assumed to be normal, so that stochastic independence is equivalent to lack of correlation. The graph of the data suggests a linear relationship between T and f :



To put the analysis on a quantitative ground, we proceed to the calculation of r . We compute the sample means of the two quantities:

$$\bar{T} = \frac{1}{15} \sum_{i=1}^{15} T_i = 39.5267 \quad \bar{f} = \frac{1}{15} \sum_{i=1}^{15} f_i = 53.6000$$

and derive then the sums of residual products/squares

$$SS_{Tf} = \sum_{i=1}^{15} (T_i - \bar{T})(f_i - \bar{f}) = 301.3600$$

$$SS_{TT} = \sum_{i=1}^{15} (T_i - \bar{T})^2 = 35.8693$$

$$SS_{ff} = \sum_{i=1}^{15} (f_i - \bar{f})^2 = 9351.6000$$

so that the linear correlation coefficient becomes

$$r = \frac{SS_{Tf}}{\sqrt{SS_{TT}}\sqrt{SS_{ff}}} = \frac{301.3600}{\sqrt{35.8693}\sqrt{9351.6000}} = 0.5203$$

As T and f are normal, we can test the null hypothesis

$$H_0 : T \text{ and } f \text{ are stochastically independent}$$

against the alternative hypothesis

$$H_1 : T \text{ and } f \text{ are stochastically dependent}$$

by using the random variable

$$t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$$

which, for true H_0 , follows a Student distribution with $n - 2$ d.o.f.

In the present case we have

$$t = \sqrt{15-2} \frac{0.5203}{\sqrt{1-0.5203^2}} = 2.1969$$

The rejection region takes then the form

$$|t| > t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.975](13)} = 2.160$$

for a significance level $\alpha = 5\%$, while it becomes

$$|t| > t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.995](13)} = 3.012$$

whenever the requested significance level is $\alpha = 1\%$.

**The null hypothesis H_0 must be rejected
at a significance level of 5%!**

Therefore we conclude that a stochastic dependence between the body temperature and the respiratory frequency of birds is plausible at a 5% significance level.

In the hypothesis of normal variables, such a dependence is equivalent to a nonvanishing correlation between the same random variables.

In contrast:

**the null hypothesis H_0 cannot be rejected
at the significance level of 1%!**

The detailed calculations are listed in the table below, along with a resume of our conclusions.

x value	y value	x-mean(x)	y-mean(y)	dx ²	dy ²	dx*dy
39,60	33,00	0,0733	-20,6000	0,0054	424,3600	-1,5107
40,10	50,00	0,5733	-3,6000	0,3287	12,9600	-2,0640
41,70	75,00	2,1733	21,4000	4,7234	457,9600	46,5093
39,00	13,00	-0,5267	-40,6000	0,2774	1.648,3600	21,3827
41,90	68,00	2,3733	14,4000	5,6327	207,3600	34,1760
42,80	115,00	3,2733	61,4000	10,7147	3.769,9600	200,9827
40,30	52,00	0,7733	-1,6000	0,5980	2,5600	-1,2373
39,00	33,00	-0,5267	-20,6000	0,2774	424,3600	10,8493
39,70	60,00	0,1733	6,4000	0,0300	40,9600	1,1093
39,30	28,00	-0,2267	-25,6000	0,0514	655,3600	5,8027
37,90	74,00	-1,6267	20,4000	2,6460	416,1600	-33,1840
37,90	39,00	-1,6267	-14,6000	2,6460	213,1600	23,7493
37,90	76,00	-1,6267	22,4000	2,6460	501,7600	-36,4373
37,90	30,00	-1,6267	-23,6000	2,6460	556,9600	38,3893
37,90	58,00	-1,6267	4,4000	2,6460	19,3600	-7,1573
mean(x):	39,5267			SSxx	SSyy	SSxy
mean(y):	53,6000			35,8693	9.351,6000	301,3600

Pearson linear correlation coefficient r: 0,5203

Student's t: 2,1969

The t-value at a significance level of 5% is: 2,1600

The t-value at a significance level of 1% is instead: 3,0120

In the first case, since $t > t\text{-value}$, we conclude that the null hypothesis must be rejected at the given significance level. The variables x and y are not stochastically independent.

In the second case, $t < t\text{-value}$ and we conclude that the null hypothesis cannot be rejected at the given significance level. The variables x and y can be regarded as stochastically independent.

Answer to Exercise 2

The sample mean and standard deviation are immediately calculated as

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 690.90 \quad s = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 11.2195$$

The suspect value is 718, the farthest from the mean \bar{x} . The distance of the suspect value from the mean, in units of s , holds

$$z = \frac{x_{\text{sus}} - \bar{x}}{s} = \frac{718 - 690.9}{11.2195} = 2.4154$$

The probability that a measurement falls at a distance larger than 2.4154 standard deviations from the mean can be calculated from the Table of the cumulative distribution function of the standard normal distribution:

$$\begin{aligned} P(|x_{\text{sus}} - \bar{x}| \geq 2.4154s) &= 1 - P(|x_{\text{sus}} - \bar{x}| < 2.4154s) = \\ &= 1 - 2 \cdot P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.4154s) = \\ &= 1 - 2 \cdot 0.49214 = 0.01572 \end{aligned}$$

Notice that the probability $P = P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.4154s)$ is not directly available on the table, but can be estimated with sufficient accuracy by a linear interpolation scheme:

2.4100	0.49202
2.4154	P
2.4200	0.49224

$$\frac{2.4154 - 2.41}{2.42 - 2.41} = \frac{P - 0.49202}{0.94224 - 0.49202}$$

which provides $P = 0.49214$.

Out of 10 measurements we typically expect $10 \cdot 0.01572 = 0.1572$ “bad” results, at a distance larger than 2.4154s from the mean. **Since $0.1572 < 1/2$, Chauvenet criterion suggests that x_{sus} should be rejected as an outlier.**

The detailed calculations are summarized in the table below:

x	Dx	ABS(Dx)	Dx ²	
694,00	3,10	3,10	9,61	
690,00	-0,90	0,90	0,81	
692,00	1,10	1,10	1,21	
696,00	5,10	5,10	26,01	
676,00	-14,90	14,90	222,01	
688,00	-2,90	2,90	8,41	
718,00	27,10	27,10	734,41	
681,00	-9,90	9,90	98,01	
688,00	-2,90	2,90	8,41	
686,00	-4,90	4,90	24,01	
mean(x): 690,90		st.dev.(x): 11,2195		
Maximum of ABS(Dx): 27,1000	corresponding to the value of x:			Outlier 718,00
Distance z of the outlier from the mean in standard deviation units:				2,4154
Probability of a larger distance from the mean (see table):	z	area fro 0 to z		
	2,4100	0,49202	0,01596	
	2,4200	0,49224	0,01552	
Linearly interpolated value:	2,4154	0,49214	0,01572	
Mean number of expected events out of 10 measurements:				0,1572

The mean number of outliers is smaller than 1/2. Thus, according to Chauvenet criterion, the value 718 must be rejected as not belonging to the population.

Answer to Exercise 3

Number of data: $n = 16$

$$\text{Sample mean: } \bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = 293.0438$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (T_i - \bar{T})^2 = 70.0546$$

$$\text{Estimated standard deviation: } s = \sqrt{s^2} = 8.36986$$

□ The CI of the mean μ of T is

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

and for $\alpha = 0.01$, $n = 16$ has therefore the limits

$$\bar{T} - t_{[0.995](15)} \frac{s}{\sqrt{16}} = 293.0438 - 2.947 \cdot \frac{8.36986}{\sqrt{16}} = 286.8773$$

$$\bar{T} + t_{[0.995](15)} \frac{s}{\sqrt{16}} = 293.0438 + 2.947 \cdot \frac{8.36986}{\sqrt{16}} = 299.2103$$

so that the CI of the mean writes

$$286.8773 \leq \mu \leq 299.2103$$

or, equivalently,

$$[293.0438 \pm 6.1665]$$

□ 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 $n = 16$. Thus

$$\frac{1}{\chi^2_{[0.975](15)}} 15 s^2 = \frac{1}{27.488} \cdot 15 \cdot 70.0546 = 38.2283$$

$$\frac{1}{\chi^2_{[0.025](15)}} 15 s^2 = \frac{1}{6.262} \cdot 15 \cdot 70.0546 = 167.8089$$

and the CI becomes

$$38.2283 \leq \sigma^2 \leq 167.8089$$

The CI of the standard deviation at the confidence level of 95% is finally obtained by extracting the square root of the upper and lower bounds:

$$\sqrt{38.2283} \leq \sigma \leq \sqrt{167.8089}$$

and provides

$$6.183 \leq \sigma \leq 12.954$$

A detailed account of the calculations is given below:

T	DT	DT ²
306,5	13,45625	181,07066
310,5	17,45625	304,72066
297,3	4,25625	18,11566
278,9	-14,14375	200,04566
295,0	1,95625	3,82691
289,0	-4,04375	16,35191
285,4	-7,64375	58,42691
280,7	-12,34375	152,36816
292,1	-0,94375	0,89066
293,3	0,25625	0,06566
286,5	-6,54375	42,82066
290,1	-2,94375	8,66566
296,9	3,85625	14,87066
295,3	2,25625	5,09066
291,7	-1,34375	1,80566
299,5	6,45625	41,68316

mean:	293,04375	variance:	70,054625
		st.dev.:	8,3698641

Lower limit of the confidence interval for the mean:	286,877253
Upper limit of the confidence interval for the mean:	299,210247

Estimate of the mean:	293,04375
Estimate of the absolute error:	6,16649737

Lower limit of the confidence interval for the variance:	38,2282951
Upper limit of the confidence interval for the variance:	167,808907

Lower limit of the confidence interval for the st. dev.:	6,18290345
Upper limit of the confidence interval for the st. dev.:	12,9541077