CHAPTER 23. Mathematical Statistics

Sec. 23.2 Estimation of Parameters

Problem Set 23.2. Page 1108

1. **Normal distribution.** This problem is similar to Example 1 in the text, but simpler since $\mu = 0$ is given. The likelihood function remains the same, with $\mu = 0$ in the auxiliary function h. Similarly, its logarithm remains

$$\ln l = -n \ln(\sqrt{2\pi}) - n \ln \sigma - h.$$

In Example 1 you had two parameters μ and σ and two equations for determining estimates of them. You now have only one parameter to be estimated, namely, σ, and you use the second equation only. This equation looks the same as before, with $\mu = 0$. By solving it algebraically, as in Example 1, you obtain the estimate

$$\tilde{\sigma}^2 = \frac{1}{n}(x_1 + x_2^2 + \dots + x_n^2).$$

5. **Exponential distribution.** By definition the density of the exponential distribution is

$$f(x) = \theta e^{-\theta x}$$

if $x \geqq 0$ and 0 for negative x. It involves the parameter θ, for which you are supposed to find a maximum likelihood estimate. You begin with the likelihood function, which for a given sample $x_1, \dots, x_n$ is

$$l = \theta e^{-\theta x_1} \theta e^{-\theta x_2} \dots \theta e^{-\theta x_n} = \theta^n \exp[-\theta(x_1 + \dots + x_n)].$$

Next you calculate the logarithm of this, obtaining

$$\ln l = n \ln \theta - \theta(x_1 + \dots + x_n).$$

You now take the partial derivative of this with respect to θ and equate it to zero. This gives

$$\frac{n}{\theta} - (x_1 + \dots + x_n) = 0.$$

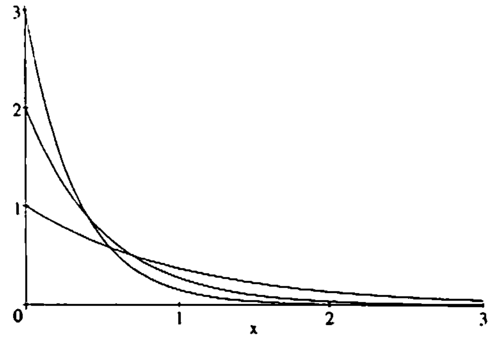
Solving algebraically for θ and denoting the solution (the desired maximum likelihood estimate) by $\hat{\theta}$ gives

$$\theta = \hat{\theta} = \frac{n}{x_1 + \dots + x_n} = \frac{1}{\bar{x}}.$$

Thus the estimate is the reciprocal of the sample mean. This looks strange, but finds its plausible explanation if you calculate the mean of the distribution, using the general definition (1b) of the mean of a continuous distribution in Sec. 22.6. You then obtain, by integrating by parts and noting that the integral-free expression obtained in this process is zero at both limits of integration

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^{\infty} x \theta e^{-\theta x}\, dx = \frac{1}{\theta}.$$

Hence the maximum likelihood estimate of $\theta = 1/\mu$ is $1/\bar{x}$. This makes good sense. The figure shows the density of the exponential distribution for three values of the parameter θ, namely, $\theta = 1, 2, 3$. You see that for large θ the mean $1/\theta$ is small. This agrees with the fact that then the curve is steep, with the area under the curve concentrated near small values of x (> 0).



Section 23.2. Problem 5. Densities for $\theta = 1,2,3$

Sec. 23.3 Confidence Intervals

Problem Set 23.3. Page 1117

3. Confidence interval for the mean when the variance is known. You must sharply distinguish between the two cases of a known σ and an unknown σ.

Case 1. When σ is known, your calculation (for a normal distribution) will involve only the normal distribution itself.

Case 2. When σ is unknown, you need another distribution, namely, the t-distribution in Table A9 in Appendix 5.

The present problem concerns Case 1 because σ is known, $\sigma = 2.5$. Hence you need no further distribution besides the normal distribution. The reason is simple. You can regard the sample mean

$$\bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$$

as an observed value of the random variable

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n),$$

and $\bar{X}$ is normal with mean μ and variance σ^2/n. Here $X_1, \dots, X_n$ are independent random variables all having the same distribution, namely, the normal distribution of the population from which the sample is taken, with mean μ and variance σ^2. This is stated in Theorem 1(b). On p. 1111 it is shown that from this theorem there follows formula (7), that is,

$$P(\bar{X} - k \leqq \mu \leqq \bar{X} + k) = \gamma. \tag{A}$$

Here the confidence level γ must be chosen. In our problem it is required that $\gamma = 0.99$. The letter k in (A) is just a short notation for $c\sigma/\sqrt{n}$. In our problem, $k = c \cdot 2.5/\sqrt{6}$. Here c depends on the choice of γ and can be calculated from (6). To make things easier, you find c (as well as all your steps) in Table 23.1, which has been obtained from (6) and (7). There you see that for $\gamma = 0.99$ the critical value c is $c = 2.576$. With this you obtain $k = 2.576 \cdot 2.5/\sqrt{6} = 2.629119$.

The next step is Step 3, in which you need the sample mean

$$\bar{x} = \frac{1}{6}(30.8 + 30.0 + 29.9 + 30.1 + 31.7 + 34.0) = 31.083\,333.$$

You can now obtain the confidence interval from (3) in the form

$$\text{CONF}_{0.99}(31.083 - 2.629 \leqq \mu \leqq 31.083 + 2.629), \tag{B}$$

that is,

$$\text{CONF}_{0.99}(28.454 \leqq \mu \leqq 33.713). \tag{C}$$

You see that the sample mean $\bar{x}$ is the midpoint of the interval, and its length is $2k$. It is rather long, but this should not surprise you because the sample size 6 is small. If you want a shorter interval, you have to take a larger sample (if you can); see Example 2 and Fig. 494. Or you can take a smaller γ, e.g. 0.95 [if you can allow a larger chance of being wrong, that is, of obtaining an interval that does not contain the unknown population mean μ (1 in about 20 cases if you choose $\gamma = 0.95$)].

Note that in (B) and (C) three digits have been dropped because it would not make too much sense to give 6D-values for the endpoints of an interval that is over 5 units long and is obtained from sample values that are rounded to 1D. Rather, you could drop even more digits if you wish.

9. **Confidence interval for the mean when the variance is unknown.** This is Case 2 (see the beginning of the previous problem), in which the normal distribution is no longer sufficient, but you need the t-distribution. Values needed for the present purpose are given in Table A9 in Appendix 5.

What is the difference compared to Case 1? We had $k = \sigma c/\sqrt{n}$. This can no longer be used because σ is unknown. Instead you have to use $k = sc/\sqrt{n}$, as shown in Table 23.2 (Sec. 23.3), with c determined from equation (9), that is, since $\gamma = 0.99$ is required,

$$F(c) = (1/2)(1+\gamma) = (1/2)(1+0.99) = 0.995.$$

The size of the given sample is $n = 20$. Hence the number of degrees of freedom of the t-distribution to be used is $n - 1 = 19$. For $F(z) = 0.995$ you find in the column for 19 degrees of freedom the value $c = 2.86$. This is your value of c in Step 4 of Table 23.2, p. 1112. The variance of a sample of size $n = 20$ is given as $s^2 = 0.09\ \text{cm}^2$. Hence the sample standard deviation is $s = 0.3$. With these values you can now calculate in Step 4 of Table 23.2 (p. 1112)

$$k = s\,c/\sqrt{n} = 0.3\cdot 2.86/\sqrt{20} = 0.191855.$$

The given sample mean is $\bar{x} = 15.50$. This is the midpoint of the confidence interval, whose endpoints are $x - k$ and $\bar{x} + k$. See (10) in Table 23.2 on p. 1112. Numerically, by inserting $\bar{x}$ and k,

$$\text{CONF}_{0.99}(15.308 \leqq \mu \leqq 15.692).$$

17. **Functions of random variables.** If X is normal with mean 40 and variance 4, then $3X$ has 3 times the mean of X, which is 120, and $3^2 = 9$ times the variance of X, which equals 36. This follows from Team Project 14(g) in Sec. 22.8, which concerns transformations of the present kind.

Similarly, $5X - 2$ has the mean $5\cdot 40 - 2 = 198$ and the variance $5^2\cdot 4 = 100$. Note that the variance is translation invariant, that is, -2 in $5X - 2$ has no effect on the variance of $5X - 2$.

Sec. 23.4 Testing of Hypotheses, Decisions

Problem Set 23.4. Page 1127

1. **Test for the mean when the variance is unknown.** The sample is

$$1,\ -1,\quad 1,\quad 3,\ -8,\quad 6,\quad 0.$$

Normality of the corresponding population is assumed. You are requested to test the hypothesis

$$\mu = \mu_0 = 0 \tag{A}$$

against the alternative

$$\mu > 0. \tag{B}$$

As in the previous section you must sharply distinguish between two cases.

Case 1. The variance σ^2 of the population is known (pp. 1122-1124).

Case 2. The variance of the population is unknown (pp.1124-1125).

The present problem belongs to Case 2 because the variance σ^2 of the population is not given. Hence you proceed as in Example 3 on p. 1124. That is, you use the t-distribution (Table A9 in Appendix 5), as follows. From the sample you calculate an observed value

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{x}{s/\sqrt{n}}. \tag{C}$$

of the random variable

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X}}{S/\sqrt{n}} \tag{D}$$

used in Example 3. Here S^2 is a random variable such that the sample variance s^2 is an observed value of S^2. This variable S^2 is given in (12), p. 1114, in the section on confidence intervals. This is not just by chance, but the determination of confidence intervals and the testing of hypotheses are based on the same theory; they are related tasks resulting by looking at situations from two different angles.

From Table A9 you calculate a critical value c, which you then compare with t in (C). From (A) and (B) you see that the test is right-sided. This is illustrated in the upper part of Fig. 497 (p. 1120). Since $\alpha = 5\%$ is required, your c will be the 95%-point of the t-distribution with $n - 1 = 7 - 1 = 6$ degrees of freedom ($n = 7$ is the sample size). In the row of Table A9 for $F(z) = 0.95$ and the column for 6 degrees of freedom you find the value

$$c = 1.94. \tag{E}$$

You now calculate t from (C) and compare it with c. If $t \leqq c$, the hypothesis is accepted. If $t > c$, the hypothesis is rejected.

The calculation will give you $\bar{x} = 0.286$ (actually, 2/7, but it would not make sense to carry along more digits), and $s = 4.309$ (recall that we use (2), p. 1106, with $n - 1 = 6$ in the denominator, which is better than n, perhaps used by your CAS, Maple, for instance). From this and (C) you obtain

$$t = \frac{0.286 - 0}{4.309/\sqrt{7}} = 0.18 < c = 1.94.$$

Hence the hypothesis is accepted.

17. **Comparison of means.** The answer on p. A49 in Appendix 2 gives you the idea of how to proceed. Case A (paired comparison) would not be appropriate because there is no indication that the automobiles used to obtain the two samples were the same, and even if they were, you would have to know the two corresponding values of the mileage, so that you could not pair. Since the samples sizes are the same ($n_1 = n_2 = 16$), you can use (12), which is obtained from (11) by obvious algebraic simplifications. Since the means and the standard deviations are given, the solution of the problem amounts to inserting these values into (12). This gives $t_0 = 3.328$.

The hypothesis is that brand B is not better than brand A. The alternative is that B is better than A. This is a right-sided test. You obtain c from Table A9 in Appendix 5 with $n_1 + n_2 - 2 = 32 - 2 = 30$ degrees of freedom, as is mentioned without proof in Example 5. No α is given. $\alpha = 5\%$ or $\alpha = 1\%$ are the usual choices. In the present application the table gives 1.70 as the 95%-point and 2.46 as the 99%-point. Hence for both choices of α the hypothesis is rejected because 3.328 is larger than both of these values. Hence on the basis of the given samples you can assert that brand B is significantly better than brand A; the higher values are not just due to chance effects.

Sec. 23.5 Quality Control

Problem Set 23.5. Page 1132

1. **Control of mean and standard deviation.** The sample mean $\bar{x}$ is an observed value of $\bar{X}$ (defined by (4) on p. 1110), which is normal with mean 1 (if the hypothesis is true) and standard deviation $\sigma/\sqrt{4} = 0.01$. The solution of the problem follows from (1) on p. 1129. You should understand that this is a two-sided test with $\alpha = 1\%$, the values -2.58 and 2.58 in (1) corresponding to the 0.5% and 99.5% points of the standardized normal distribution, and $\sigma/\sqrt{n}$ being the standard deviation of $\bar{X}$. The upper part of Fig. 501 indicates that the area under the density curve of $\bar{X}$ is divided into three parts, namely, 0.5% lies above

UCL, 0.5% below LCL, and 99% between these two points. Since $n = 4$, your simple calculation gives

$$\text{LCL} = 1 - 2.58 \cdot \frac{0.02}{\sqrt{n}} = 0.9742, \qquad \text{UCL} = 1.0258 .$$

9. **Number of defectives.** If in a production process, $p\%$ (for instance, 3%, thus $p = 0.03$) of the items are defective, then the probability of obtaining x defectives in n independent trials is given by the binomial distribution, whose probability function is (2) in Sec. 22.7. This distribution has the mean np and the variance $npq = np(1 - np)$. Now since you are required to use the three-sigma limits as LCL and UCL on a control chart for the mean, you obtain the formulas in the answer, which are

$$\text{LCL} = \mu - 3\sigma = np - 3\sqrt{npq}, \qquad \text{UCL} = \mu + 3\sigma = np + 3\sqrt{npq}.$$

This explains these formulas. Note that in the case of the normal distribution the three-sigma limits would correspond to a choice of $\alpha = 0.3\%$ (see (6c) in Sec. 22.8).

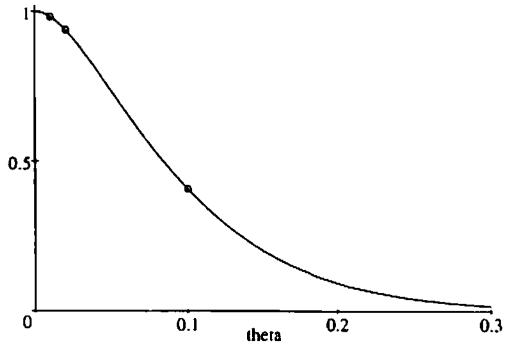
Sec. 23.6 Acceptance Sampling

Problem Set 23.6. Page 1136

1. **Sampling plan with $c = 1$.** This means that a lot is accepted if it contains one or no dull knive. Conclude from this that in (3) you have to take the sum of the first two terms (corresponding to $x = 0$ and $x = 1$). The figure shows the corresponding OC curve, given by

$$P(A, \theta) = e^{-\mu}(1 + \mu) = e^{-20\theta}(1 + 20\theta).$$

From this formula you obtain the three specific values given in the answer, namely, 0.9825 ($\theta = 1\% = 0.01$), 0.9384 ($\theta = 2\%$), and 0.4060 ($\theta = 10\%$).

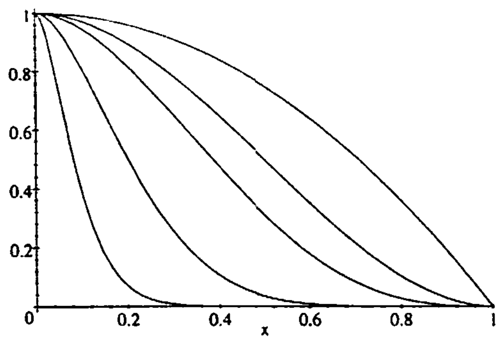


Section 23.6. Problem 1. OC curve

9. **Influence of the sample size n.** The formula in the answer is

$$P(A; \theta) = (1 - \theta)^n + n\theta(1 - \theta)^{n-1}.$$

This is the sum of the first two terms of the binomial distribution with probability of success $p = \theta$ (the probability of drawing a defective in a single trial), giving the probability of obtaining none or 1 defective in n independent trials; practically speaking, in drawing a small sample from a very large lot. The curves are the steeper the larger the sample size is.



Section 23.6. Problem 9. Sampling plans with $c = 1$ and $n = 2$ (upper curve), 3, 4, 8, 20

Sec. 23.7 Goodness of Fit. Chi-Square Test

Problem Set 23.7. Page 1140

1. **Coin flipping.** By definition, a coin is fair if heads and tails have the same probability, namely, 1/2. Hence in 100 trials you should expect 50 heads and 50 tails, except for random deviations. The problem thus amounts to deciding whether a result of 40 heads and 60 tails could still be possible if the coin is fair, that is, the result could still be due to randomness, or whether the deviation is significant, that is, whether one must assume that the coin is not fair. Accordingly, you have to test the hypothesis that the coin is fair, against the alternative that it is not fair. To perform a chi-square test, calculate χ_0^2 from (1) in Table 23.7 (Sec. 23.7). For this you need

$b_1 = 40$ Number of heads observed

$e_1 = 50$ Number of heads expected if the hypothesis is true

and similarly

$b_2 = 60$ Number of tails observed

$e_2 = 50$ Number of tails expected if the hypothesis is true.

Thus $K = 2$ in Table 23.7, and you obtain a sum of two terms, namely,

$$\chi_0^2 = \frac{(b_1 - e_1)^2}{e_1} + \frac{(b_2 - e_2)^2}{e_2} = \frac{(40 - 50)^2}{50} + \frac{(60 - 50)^2}{50} = 4.$$

(If you want to proceed literally as in Table 23.7, assign, for instance, $x = 1$ to 'head' and $x = 0$ to 'no head' (that is, 'tail') and then divide the x-axis into 2 intervals, one containing $x = 0$ and the other $x = 1$; this gives the formula for χ_0^2.) The significance level $\alpha = 5\%$ is required in the problem. You now obtain the critical c from the table of the χ^2-distribution (Table A10 in Appendix 5). For $K - 1 = 1$ degree of freedom you find that the equation $P(\chi^2 \leqq c) = 1 - \alpha = 0.95$ has the solution $c = 3.84$. Now the above $\chi_0^2 = 4$, which measures the deviation, is greater than this critical value. This means that if the hypothesis is true, the probability that an observed value of χ_0^2 falls anywhere in the interval from 3.84 to ∞ is only 5%. Since in the present case this has happened, you reject the hypothesis and assert that the coin used in obtaining that sample is not fair. 4 is not much greater than 3.84, and Table A10 shows you (in the column for 1 degree of freedom) that $\alpha = 2.5\%$ (thus 97.5%, the next value given in the table) would lead to the acceptance of the hypothesis. This is a common situation that should not surprise you.

Sec. 23.8 Nonparametric Tests

Problem Set 23.8. Page 1143

3. **Sign test.** From the results of the 11 trials drop the three that do not contribute to the decision; this is as in Example 1 of the text. This leaves you with 8 results. 7 of them (when A turned out to be better than B) you can regard as positive and the remaining 1 (when B was better than A) as negative. The hypothesis is that A is not better than B, The alternative is that A is better than B. This you can infer from the wording of the problem. Under the hypothesis, positive and negative values should have the same probability $p = q = 1/2$. Consider the random variable X = *Number of positive values among* 8 *values.* If the hypothesis is true, its possible values have the probabilities

$$f(x) = P(X = x) = \binom{8}{x}\left(\frac{1}{2}\right)^8 = 0.00391\binom{8}{x}.$$

This is a binomial distribution with $p = 1/2$; see Sec. 22.7. Conclude from this that under the hypothesis the probability of observing so many positive values, 7 or even more (namely, 8) is given by

$$f(7) + f(8) = 0.00391(8+1) = 0.0352.$$

This is less than the usual 5%, so you reject the hypothesis and assert that filters of type A are better than filters of type B. Note that in this test the hypothesis would be accepted under a significance level $\alpha = 1\% = 0.01$.

11. **Test for trend.** The sample is

$$22 \quad 19 \quad 21 \quad 20 \quad 25 \quad 18 \quad 27 \quad 30 \quad 26 \quad 24.$$

Proceed as in Example 2 of the text by listing the transpositions in the sample, beginning with the first sample value, 22, then considering the next, 19, and so on, in the given order.

22	precedes	19	21	20	18	4	transpositions
19	"	18				1	"
21	"	20	18			2	"
20	"	18				1	"
25	"	18	24			2	"
27	"	26	24			2	"
30	"	26	24			2	"
26	"	24				1	"

Hence the sample contains 15 transpositions. Its size is $n = 10$. The hypothesis is that there is no trend, that is, it makes no difference whether the animals receive different amounts of food (of course, within reasonable limits, which are not indicated in the problem). The problem requires that you test this against the alternative of positive trend.

Let T = *Number of transpositions*, as in the Example 2 of the text. You need the probability

$$P(T \leqq 15) \tag{A}$$

of f obtaining 15 or fewer transpositions in a sample of $n = 10$ values. For this, Table A12 in Appendix 5 gives the probability 0.108, almost 11%, certainly large enough for accepting the hypothesis that there is no trend. If in an experiment a somewhat unexpected result turns up, as in the present case, one should indicate precisely under what conditions the result was obtained (for instance, state the kind of food and the range of the amounts given) and one should conduct further experiments.

Sec. 23.9 Regression Analysis. Fitting Straight Lines

Problem Set 23.9. Page 1150

1. **Linear regression (straight line).** The regression line (2) is

$$y = k_0 + k_1 x.$$

Its coefficients k_0 and k_1 are obtained by solving the normal equations (10). For setting up these equations, calculate

$n = 4$	the number of sample values (4 pairs)
$\sum x_j = 30$	the sum of the x-values in the sample
$\sum x_j^2 = 306$	the sum of their squares
$\sum y_j = 119$	the sum of the y-values
$\sum x_j y_j = 1141$	the sum of the products of the x-and y-values

Hence the linear system of equations (10) in the unknowns k_0 and k_1 takes the form

$$4k_0 + 30k_1 = 119$$

$$30k_0 + 306k_1 = 1141.$$

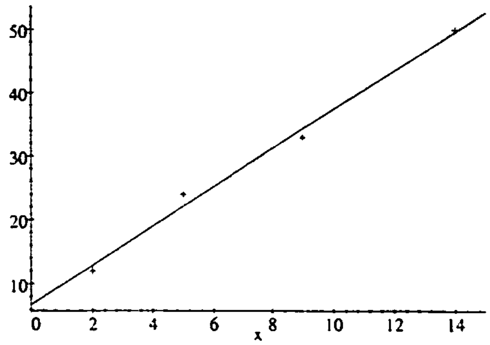
You can solve it by elimination or by Cramer's rule (p. 342), obtaining

$$k_0 = 6.74074, \qquad k_1 = 3.0679.$$

Hence the regression line is

$$y = 6.74074 + 3.0679x.$$

The figure shows this line as well as the given data as points in the xy-plane. They lie close to the line, which justifies the use of a straight line in this regression analysis. (Note that the scale on the y-axis is different from that on the x-axis.)



Section 23.9. Problem 1. Data and regression line

7. **Confidence interval for the slope κ_1 of the regression line.** So far, probability has not been involved in connection with the least squares principle, which is a *geometrical* principle. If you want to know to what extent you can "trust" the sample regression line, in particular its slope k_1, you may determine a confidence interval for the slope κ_1 of the regression line of the population, but this requires that you make assumptions involving probability, for instance, that the random variable Y for which y is an observed value, is normal for each fixed x (see Assumption (A2) on p. 1148) and that you have independence in

sampling (see Assumption (A3)). For solving the problem, proceed according to Table 23.12 on p. 1149, as follows.

1st Step. $\gamma = 95\%$ is required.

2nd Step. Determine c from (13), that is,

$$F(c) = \frac{1}{2}(1 + \gamma) = 0.975,$$

and the t-table (Table A9) in Appendix 5. Since the given sample consists of $n = 4$ pairs of values, you have to look in the column for $n - 2 = 2$ degrees of freedom, where you will find the value 4.30. This is c.

3rd Step. In the necessary calculations you can use values from Prob. 1. Calculate $3s_x^2$ from (9), using $\sum x_j = 30$ and $\sum x_j^2 = 306$. This gives (multiply by $n - 1 = 3$ on both sides)

$$3s_x^2 = 306 - \frac{1}{4} \cdot 30^2 = 81.$$

Next use $\sum y_j = 119$ and $\sum x_j y_j = 1141$ for obtaining from (8)

$$3s_{xy} = 1141 - \frac{1}{4} \cdot 30 \cdot 119 = 248.5.$$

In (14) you will need

$$\sum y_j^2 = 4309,$$

as calculated from the sample. With this you obtain from (14)

$$3s_y^2 = 4309 - \frac{1}{4} \cdot 119^2 = 768.75.$$

In (15) you will need $k_1^2 = 3.0679^2 = 9.4120$. With this, (15) gives

$$q_0 = 768.75 - 9.4120 \cdot 81 = 6.377.$$

4th Step. Now $c = 4.30$ (see before) is needed. You finally calculate

$$K = 4.30\sqrt{\frac{6.378}{281}} = 0.8531.$$

This is half the length of the confidence interval. The midpoint is the slope $k_1 = 3.0679$ of the sample regression line (see before). Hence from (16) in Table 23.12 in Sec. 23.9 you obtain the answer

$$\text{CONF}_{0.95}(3.0679 - 0.8531 \leqq \kappa_1 \leqq 3.0679 + 0.8531),$$

that is,

$$\text{CONF}_{0.95}(2.2148 \leqq \kappa_1 \leqq 3.9210).$$

It may surprise you that the points in the figure to Prob. 1 lie close to the regression line, but, nevertheless, the confidence interval for the slope of the regression line is relatively large, perhaps larger than you had expected. However, this is due to the fact that the sample used is very small ($n = 4$).