Exercise-book
-
Part-V
Statistics
-

Andras Vetier

2013 06 22

Contents

1 Regression in one-dimension

PROBLEMS

1. *Minimizing the the expected value of* $|X - c|$
   A certain type of electric bulbs has exponential life-time X with parameter 2.5.

   (a) What is the number c for which the expected value of $|X - c|$ is minimal?

   (b) Check your result with simulation. Take the optimal c value, call it c_{opt}, and another c value, call it c_{other}. Simulate 1000 experiments, $X_1, \ldots, X_{1000}$, and approximate the expected values of $|X - c_{\text{opt}}|$ and $|X - c_{\text{other}}|$ by the averages

   $$\frac{|X_1 - c_{\text{opt}}| + \ldots + |X_{1000} - c_{\text{opt}}|}{1000}$$

   and

   $$\frac{|X_1 - c_{\text{other}}| + \ldots + |X_{1000} - c_{\text{other}}|}{1000}$$

   respectively. You may be convinced that the first average, in most cases, is smaller than the second one.

2. *Minimizing the the expected value of* $(X - c)^2$
   A certain type of electric bulbs has exponential life-time X with parameter 2.5.

   (a) What is the number c for which the expected value of $(X - c)^2$ is minimal?

   (b) Check your result with simulation. Take the optimal c value, call it c_{opt}, and another c value, call it c_{other}. Simulate 1000 experiments, $X_1, \ldots, X_{1000}$, and approximate the expected values of $(X - c_{\text{opt}})^2$ and $(X - c_{\text{other}})^2$ by the averages

   $$\frac{(X_1 - c_{\text{opt}})^2 + \ldots + (X_{1000} - c_{\text{opt}})^2}{1000}$$

   and

   $$\frac{(X_1 - c_{\text{other}})^2 + \ldots + (X_{1000} - c_{\text{other}})^2}{1000}$$

   respectively You may be convinced that the first average, in most cases, is smaller than the second one.

2 Regression in two-dimensions

EXCEL

Here are files to study regression problems:

Demonstration file: Regression
ef-200-92-00

Demonstration file: Conditional distributions, -expected value, -median, $(RND_1 RND_2; RND_1)$
ef-200-93-00

Demonstration file: Conditional distributions, -expected value, -median, $(RND_1 RND_2; RND_1/RND_2)$
ef-200-94-00

PROBLEMS

3. *Water level in a reservoir - regression curves*
Let us assume that the water level of a water reservoir is measured in such a scale that the minimum level corresponds to 0, the maximum level corresponds to 1. Let us denote the water level on a day by X, and the water level two days later by Y. Let us assume that the two-dimensional random variable (X, Y) has the following density function:

$$f(x,y) = \frac{4}{5}(1+xy) \quad \text{if} \quad 0 < x < 1, \;\; 0 < y < 1$$

(a) Find the conditional expected value of Y on condition that $X = x$,

(b) Find the conditional median of Y on condition that $X = x$.

Solution:

(a) The density function of X is:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_0^1 \frac{4}{5}(1+xy)\, dy =$$

$$= \frac{4}{5}\left(1 + \frac{x}{2}\right) = \frac{4}{5} + \frac{2}{5}x \quad \text{if} \quad 0 < x < 1$$

The conditional density function of Y on condition that $X = x$ is:

$$f_{2|1}(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{\frac{4}{5}(1+xy)}{\frac{4}{5}\left(1+\frac{x}{2}\right)} = \frac{1+xy}{1+\frac{x}{2}}$$

The conditional expected value of Y on condition that $X = x$ is:

$$\mathbf{E}(Y|X = x) = \int_{-\infty}^{\infty} y\, f_{2|1}(y|x)\, dy = \int_0^1 y\, \frac{1 + xy}{1 + \frac{x}{2}}\, dy = \frac{\frac{1}{2} + \frac{x}{3}}{1 + \frac{x}{2}}$$

Solution
Sol-05-02-01

(b) The conditional distribution function of Y on condition that $X = x$ is:

$$F_{2|1}(y|x) = \int_{-\infty}^{y} f_{2|1}(y|x)\, dy = \int_0^y \frac{1 + xy}{1 + \frac{x}{2}}\, dy = \frac{y + x\frac{y^2}{2}}{1 + \frac{x}{2}}$$

$$= \left(\frac{2}{2 + x}\right) y + \left(\frac{x}{2 + x}\right) y^2 \quad \text{if} \quad 0 < x < 1, \quad 0 < y < 1$$

The conditional median is the solution to the equation

$$F_{2|1}(y|x) = \frac{1}{2}$$

that is,

$$\left(\frac{2}{2 + x}\right) y + \left(\frac{x}{2 + x}\right) y^2 = \frac{1}{2}$$

The solution, that is, the conditional median of Y on condition that $X = x$ is:

$$y = \frac{-2 + \sqrt{4 + 2x(2 + x)}}{2x}$$

Solution
Sol-05-02-02

4. *Checking the results of the previous problem by simulation*

(a) Take the regression curve defined by the conditional expected values of Y, and denote this function by $k_{\text{opt}}(x)$:

$$k_{\text{opt}}(x) = \frac{\frac{1}{2} + \frac{x}{3}}{1 + \frac{x}{2}}$$

Take an arbitrary other function $k_{\text{other}}(x)$, too. Simulate 1000 experiments, $X_1, \ldots, X_{1000}$, and approximate the expected values of $(Y - k_{\text{opt}}(X))^2$ and $(Y - k_{\text{other}}(X))^2$ by the averages

$$\frac{(Y_1 - k_{\text{opt}}(X_1))^2 + \ldots + (Y_{1000} - k_{\text{opt}}(X_{1000}))^2}{1000}$$

and

$$\frac{(Y_1 - k_{\text{other}}(X_1))^2 + \ldots + (Y_{1000} - k_{\text{other}}(X_{1000}))^2}{1000}$$

respectively You may be convinced that the first average, in most cases, is smaller than the second one.

(b) Take the regression curve defined by the conditional medians of Y, and denote this function by $k_{\text{opt}}(x)$:

$$k_{\text{opt}}(x) = \frac{-2 + \sqrt{4 + 2x(2 + x)}}{2x}$$

Take an arbitrary other function $k_{\text{other}}(x)$, too. Simulate 1000 experiments, $X_1, \ldots, X_{1000}$, and approximate the expected values of $(Y - k_{\text{opt}}(X))^2$ and $|Y - k_{\text{other}}(X)|$ by the averages

$$\frac{|Y_1 - k_{\text{opt}}(X_1)| + \ldots + |Y_{1000} - k_{\text{opt}}(X_{1000})|}{1000}$$

and

$$\frac{|Y_1 - k_{\text{other}}(X_1)| + \ldots + |Y_{1000} - k_{\text{other}}(X_{1000})|}{1000}$$

respectively You may be convinced that the first average, in most cases, is smaller than the second one.

3 Linear regression

PROBLEMS

5. *Water level in a reservoir - linear regression*
Let us assume that the water level of a water reservoir is measured in such a scale that the minimum level corresponds to 0, the maximum level corresponds to 1. Let us denote the water level on a day by X, and the water level two days later by Y. Let us assume that the two-dimensional random variable (X, Y) has the following density function:

$$f(x, y) = \frac{4}{5}(1 + xy) \quad \text{if} \quad 0 < x < 1, \;\; 0 < y < 1$$

(a) Calculate the expected value of X and Y.

(b) Calculate the second moments, the variances and the standard deviations of X and Y.

(c) Calculate the covariance between X and Y.

(d) Calculate the correlation coefficient between X and Y.

(e) Set up the equation of the regression line when Y is estimated from X.

6. *Checking the results of the previous problem by simulation*
Check by simulation that the regression line gives the best linear estimation. Take the regression line, and denote the linear function by $k_{\text{opt}}(x)$. Take an arbitrary other line $k_{\text{other}}(x)$, too. Simulate 1000 experiments, $X_1, \ldots, X_{1000}$, and approximate the expected values of $(Y - k_{\text{opt}}(X))^2$ and $(Y - k_{\text{other}}(X))^2$ by the averages

$$\frac{(Y_1 - k_{\text{opt}}(X_1))^2 + \ldots + (Y_{1000} - k_{\text{opt}}(X_{1000}))^2}{1000}$$

and

$$\frac{(Y_1 - k_{\text{other}}(X_1))^2 + \ldots + (Y_{1000} - k_{\text{other}}(X_{1000}))^2}{1000}$$

respectively You may be convinced that the first average, in most cases, is smaller than the second one.

4 Confidence intervals

EXCEL

Here are files to study construction of confidence intervals.

Demonstration file: Finite confidence interval for the expected value, using standard normal distribution
ef-300-23-00

Demonstration file: Infinitely long confidence interval for the expected value, using standard normal distribution
ef-300-23-50

Demonstration file: Finite confidence interval for the expected value, using t-distribution
ef-300-24-00

PROBLEMS

7. *Constructing a confidence interval*
   Assume that the weight of a randomly chosen child in a certain school is normally distributed with an unknown expected value and a standard deviation 4 kg. What is the confidence interval for the expected value if we have the following experimental results: 15.4 kg, 26.4 kg, 24.1 kg, 20.5 kg, 15.8 kg, 26.4 kg, 25.0 kg, 21.6 kg, 86.4 kg, 27.7 kg, and we want that the confidence interval contains the expected value with a probability

   (a) 0.95 ?

   (b) 0.68

8. *How long is the confidence interval*
   Assume that the height of men in a certain country is normally distributed with an unknown expected value μ and a standard deviation 20 cm. Assume that you have measured 5 men, and their heights are: 173.2 cm; 176.4 cm; 185.3 cm; 177.7 cm; 173.0 cm. Using our method associated to the probability value

   (a) $p = 0.90$,

   (b) $p = 0.95$,

   what confidence interval do you declare for μ? How long is the interval?

9. *How long is the confidence interval*
   Assume that you have measured 20 men, and their heights are:

   173.2 cm; 176.9 cm; 185.3 cm; 177.7 cm; 173.0 cm;

   183.5 cm; 166.4 cm; 169.2 cm; 187.7 cm; 192.7 cm;

   163.4 cm; 186.3 cm; 185.9 cm; 177.7 cm; 173.2 cm,

   170.0 cm; 175.4 cm; 175.3 cm; 197.1 cm; 174.3 cm.

   Using our method associated to the probability value

   (a) $p = 0.90$,

   (b) $p = 0.95$

   what confidence interval do you declare for μ? How long is the interval?

10. *U-test*
    Assume that the weight of a randomly chosen woman in a certain country is normally distributed with an unknown expected value and a standard deviation 4 kg. What is the confidence interval for the expected value, if we want that it contains the expected value with probability 0.95, and we have the following experimental results:

    (a) 65.4 kg, 76.4 kg, 74.1 kg, 70.5 kg;

    (b) 65.8 kg, 76.4 kg, 75.0 kg, 76.6 kg, 86.4 kg, 77.7 kg;

    (c) 65.4 kg, 76.4 kg, 74.1 kg, 70.5 kg, 65.8 kg, 76.4 kg, 75.0 kg, 76.6 kg, 86.4 kg, 77.7 kg ?

5 U-tests

EXCEL

In the following simulation files, the average of the experimental results is directly simulated. The experimental results themselves are not simulated.

Demonstration file: U-test 1: Case of "less than", when n is given
ef-300-25-20

Demonstration file: U-test 1: Case of "less than", when n is given, version B
ef-300-25-30

Demonstration file: U-test 1: Case of "less than", when n is given, version C
ef-300-25-40

Demonstration file: U-test 2: Case of "less than", when n is calculated
ef-300-26-00

Demonstration file: U-test 3: Case of "equality", when n is given
ef-300-27-20

Demonstration file: U-test 3: Case of "equality", when n is given, version B
ef-300-27-30

Demonstration file: U-test 3: Case of "equality", when n is given, version C
ef-300-27-40

Demonstration file: U-test 4: Case of "equality", when n is calculated
ef-300-28-00

Demonstration file: U-test 5: Case of "equality", when an interval is considered instead of the point μ_0
ef-300-29-00

Here are files to study the U-test for "two populations".

Demonstration file: U-test, two populations (Version B)
ef-300-29-40

Demonstration file: U-test, two populations (Version C)
ef-300-29-50

PROBLEMS

Use the above files to answer the following problems.

11. *U-test*
Assume that the weight of a randomly chosen man in a certain country is normally distributed with an unknown expected value μ and a standard deviation 10 kg. You are testing the hypothesis that $\mu = 82$ with a method for which it is true that when the hypothesis holds, then the probability of accepting the hypothesis is $p = 0.95$. Find the critical values (symmetrical about 82), if you use a sample of size 5. Find the critical values (symmetrical about 82), if you use a sample of size 10. Would you accept the hypothesis $\mu = 82$, if you have the following experimental results:

    (a) 85.4 kg, 76.4 kg, 84.1 kg, 80.5 kg, 87.7 kg;

    (b) 85.4 kg, 76.4 kg, 84.1 kg, 80.5 kg, 85.8 kg, 76.4 kg, 75.0 kg, 76.6 kg, 86.4 kg, 87.7 kg?

12. *U-test*
Assume that the weight of a randomly chosen man in a certain country is normally distributed with an unknown expected value μ and a standard deviation 7 kg. If you use the significance level $p = 0.95$, then would you accept the hypothesis $\mu = 82$, if we have the following experimental results:

    (a) 85.4 kg, 76.4 kg, 84.1 kg, 80.5 kg;

    (b) 85.8 kg, 76.4 kg, 75.0 kg, 76.6 kg, 86.4 kg, 87.7 kg;

    (c) 85.4 kg, 76.4 kg,84.1 kg, 80.5 kg, 85.8 kg, 76.4 kg, 75.0 kg, 76.6 kg, 86.4 kg, 87.7 kg?

6 *** T-tests

EXCEL

Here are files to study the T-test for "equality".

Demonstration file: T-test, equality, sample average, 1 experiment (version A)
ef-300-31-00

Demonstration file: T-test, equality, sample average, 1000 experiments (version A)
ef-300-32-00

Demonstration file: T-test, equality, sample average, 1 experiment (version B, no figures)
ef-300-33-00

Demonstration file: T-test, equality, sample average, 1 experiment (version B with figures)
ef-300-34-00

Here are files to study the T-test for "less-than".

Demonstration file: T-test, less-than, sample average, 1 experiment (version A)
ef-300-37-00

Demonstration file: T-test, less-than, sample average, 1000 experiments (version A)
ef-300-38-00

Demonstration file: T-test, less-than, sample average, 1 experiment (version B, no figures)
ef-300-39-00

Demonstration file: T-test, less-than, sample average, 1 experiment (version B with figures)
ef-300-40-00

PROBLEMS

Use the above files to answer the following problems.

13. *t-test*
Assume that the weight of a randomly chosen man in a certain country is normally distributed with an unknown expected value μ and a standard deviation. You are testing the hypothesis that $\mu = 82$ with a method for which it is true that when the hypothesis holds, then the probability of accepting the hypothesis is $p = 0.95$. Find the critical values (symmetrical about 82), if you use a sample of size 5. Find the critical values (symmetrical about 82), if you use a sample of size 10. Would you accept the hypothesis $\mu = 82$, if you have the following experimental results:

    (a) 85.4 kg, 76.4 kg, 84.1 kg, 80.5 kg, 87.7 kg;

    (b) 85.4 kg, 76.4 kg, 84.1 kg, 80.5 kg, 85.8 kg, 76.4 kg, 75.0 kg, 76.6 kg, 86.4 kg, 87.7 kg?

14. *U-test*
Assume that the weight of a randomly chosen man in a certain country is normally distributed with an unknown expected value μ and standard deviation. If you use the significance level $p = 0.95$, then would you accept the hypothesis $\mu = 82$, if we have the following experimental results:

(a) 85.4 kg, 76.4 kg, 84.1 kg, 80.5 kg;

(b) 85.8 kg, 76.4 kg, 75.0 kg, 76.6 kg, 86.4 kg, 87.7 kg;

(c) 85.4 kg, 76.4 kg,84.1 kg, 80.5 kg, 85.8 kg, 76.4 kg, 75.0 kg, 76.6 kg, 86.4 kg, 87.7 kg?

7 *** Chi-square-test for fitness

EXCEL

Here are files to study chi-square-test for fitness.

Demonstration file: Chi-square-test for fitness (version B, no figures)
ef-300-50-00

Demonstration file: Chi-square-test for fitness (version C, with figures)
ef-300-51-00

Chi-square-test for normality. Details will be written here later.

Demonstration file: Chi-square-test for normality
ef-300-50-10

Demonstration file: Chi-square-test for independence
ef-300-50-50

PROBLEMS

Use the above files to answer the following problems.

15. *Is the die fair?*
Somebody made 80 tosses with a die. Here are the results:

2	1	6	4	1	3	3	5	2	5
3	4	6	1	4	4	5	2	3	5
5	6	1	6	1	6	4	4	1	1
2	1	5	6	1	1	5	3	2	3
6	2	2	4	4	1	5	5	2	3
5	3	6	5	4	6	3	6	6	4
1	6	5	1	5	4	3	3	2	1
6	3	6	6	2	4	1	4	5	6

Would you accept that the die is fair?

16. *Is the die fair?*
Somebody made 100 tosses with a die. Here are the results:

4	2	6	4	4	3	4	5	5	4
6	1	4	4	4	6	4	4	4	6
5	1	3	6	1	4	1	5	5	5
6	5	6	5	6	2	3	2	4	4
2	5	4	3	3	5	5	6	5	6
6	6	1	4	5	1	4	1	6	6
2	2	5	2	6	6	5	4	4	1
4	4	4	1	6	4	4	1	6	5
6	5	3	4	5	6	6	4	2	6
6	6	3	5	6	4	6	5	5	6

Would you accept that the die is fair?

17. *Are the coins fair?*
Somebody states that he tossed 2 fair coins 50 times, and checked how many heads he got with the 2 coins. Here are the results:

0	0	1	0	1	1	0	2	0	1
0	2	1	1	1	0	2	1	0	1
2	0	1	0	1	1	0	2	2	2
2	1	1	2	1	1	0	1	1	2
0	2	1	1	2	0	1	2	2	0

Would you accept that the coins are fair?

18. *Are the coins fair?*
Somebody states that he tossed 2 fair coins 60 times, and checked how many heads he got with the 2 coins. Here are the results:

2	2	0	1	1	1	1	1	1	0
1	2	1	2	1	2	1	1	1	0
1	1	1	1	1	0	2	2	1	0
0	1	1	0	2	1	1	2	1	1
2	2	0	2	2	1	1	0	1	1
1	2	1	0	2	0	0	1	2	1

Would you accept that the coins are fair?

8 *** Chi-test for standard deviation (Chi-square-test for variance)

EXCEL

Here are files to study Chi-test for standard deviation (Chi-square-test for variance).

Demonstration file: Chi-test for standard deviation (Chi-square-test for variance), 1 experiment (version A)
ef-300-52-00

Demonstration file: Chi-test for standard deviation (Chi-square-test for variance), 1000 experiments (version A)
ef-300-53-00

Demonstration file: Chi-test for standard deviation (Chi-square-test for variance) (version B, no figures)
ef-300-54-00

Demonstration file: Chi-test for standard deviation (Chi-square-test for variance) (version C, with figures)
ef-300-55-00

PROBLEMS

Use the above files to answer the following problems.

19. *Is the standard deviation 10?*
Here are 10 experimental results for a normally distributed random variable: 85.30, 90.81, 106.55, 99.45, 93.75, 96.84, 118.09, 99.29, 92.71, 90.73. Would you accept that the standard deviation of the random variable is 10?

20. *Is the standard deviation 25?*
Here are 12 experimental results for a normally distributed random variable: 123.67, 83.08, 118.91, 119.85, 88.77, 58.75, 124.30, 117.37, 87.14, 148.49, 47.97, 57.26. Would you accept that the standard deviation of the random variable is 20?

9 *** F-test for equality of variances (of standard deviations)

EXCEL

Here are files to study F-tests for variances (standard deviation).

Demonstration file: F-test for equality of variances (of standard deviations) (version B, no figures)
ef-300-56-00

Demonstration file: F-test for equality of variances (of standard deviations) (version B with figures)
ef-300-57-00

Demonstration file: F-test for "less-than" of variances (of standard deviations) (version B, no figures)
ef-300-58-00

Demonstration file: F-test for "less-than" of variances (of standard deviations) (version B with figures)
ef-300-59-00

PROBLEMS

Use the above file to answer the following problems.

21. *Are the standard deviations equal?*
Here are 11 experimental results for a normally distributed random variable X:

123.67, 83.08, 118.91, 119.85, 88.77, 58.75, 124.30, 117.37, 87.14, 148.49, 47.97.

Here are 9 experimental results for a normally distributed random variable Y:

136.66, 112.20, 44.14, 58.42, 143.21, 155.20, 121.47, 50.81, 114.76.

Would you accept that the standard deviation of X is equal to the standard deviation of Y?

22. *Are the standard deviations equal?*
Here are 13 experimental results for a normally distributed random variable X:

57.55, 56.33, 44.28, 63.74, 52.30, 58.25, 51.55, 41.41, 45.14, 37.53, 54.12, 54.73, 56.41.

Here are 14 experimental results for a normally distributed random variable Y:

15.33, 19.64, 21.70, 24.96, 20.67, 21.58, 30.04, 17.45, 14.35, 14.21, 17.43, 23.26, 18.94, 20.77.

Would you accept that the standard deviation of X is equal to the standard deviation of Y?

10 *** Test with ANOVA (Analysis of variance)

EXCEL

This is a file to study ANOVA (Analysis of variance).

Demonstration file: ANOVA (Analysis of variance)
ef-300-49-00

PROBLEMS

Use the above file to answer the following problems.

23. *Are all the expected values equal?*
Four normally distributed random variables are considered.

We have 12 measurement results for the first:

14.74, 10.6, 13.38, 20.63, 9.34, 9.44, 16.52, 9.04, 19.48, 16.21, 21.86, 12.6,

9 measurement results for the second:

16.04, 9.18, 17.81, 8.94, 16.37, 14.77, 8.60, 13.76, 11.44,

8 measurement results for the third:

15.76, 17.88, 13.38, 8.06, 9.98, 15.85, 10.03, 13.55,

12 measurement results for the fourth:

13.12, 12.95, 13.60, 10.79, 0.26, 10.72, 7.17, 4.51, 4.98, 5.83, 11.82, 15.31.

Would you accept that all the expected values equal?

24. *Are all the expected values equal?*
Four normally distributed random variables are considered.

We have 12 measurement results for the first:

18.98, 18.73, 15.31, 6.70, 18.73, 17.00, 11.21, 18.48, 14.99, 19.09, 16.00, 14.96,

9 measurement results for the second:

8.05, 8.35, 5.61, 8.28, 10.22, 6.62, 10.93, 1.13, 13.65,

8 measurement results for the third:

27.17, 21.73, 13.04, 14.66, 17.13, 19.94, 15.84, 15.85,

12 measurement results for the fourth:

12.55, 3.34, 6.48, 10.6, 6.13, 14.08, 10.57, 10.53, 0.54, 5.12, 9.17, 22.58.

Would you accept that all the expected values equal?