FEGYVERNEKI SÁNDOR,

PROBABILITY THEORY AND MATHEMATICAL STATISTICS

7

A Műszaki Földtudományi Alapszak tananyagainak kifejlesztése a TÁMOP 4.1.2-08/1/A-2009-0033 pályázat keretében valósult meg.

VII. Introduction to Mathematical Statistics

1. Mathematical Statistics

Definition: A random sample of size n is a sequence of independent, identically distributed random variables $X_1, X_2, \dots, X_n$.

Estimation

An estimator $\hat{\vartheta}$ of a parameter (statistical characteristic) ϑ of a random variable X is a random variable which depends upon a random sample $X_1, X_2, \dots, X_n$.

Some estimators:

sample mean: $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$,

sample variance: $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

An estimator $\hat{\vartheta}$ of ϑ is unbiased if $E(\hat{\vartheta}) = \vartheta$.

An estimator $\hat{\vartheta}$ having the property that, for each $\varepsilon > 0$,

$$\lim_{n \to +\infty} P(|\hat{\vartheta} - \vartheta| < \varepsilon) = 1,$$

is called a consistent estimator.

An unbiased estimator $\hat{\vartheta}$ is the efficient estimator of ϑ if $Var(\hat{\vartheta}) < Var(\hat{\vartheta}_1)$ when $\hat{\vartheta}_1$ is any other unbiased estimator of ϑ.

The term $E(\hat{\vartheta}) - \vartheta$ is called the bias of $\hat{\vartheta}$.

Confidence interval

A random interval, completely determined by the sample and independent of unknown characteristics, which covers the unknown scalar statistical characteristic ϑ with a given probability $1 - \alpha$ is called a confidence interval; for this characteristic corresponds to a confidence coefficient $1 - \alpha$. The quantity α is called a significance level of estimate deviation.

Theorem: Let $X_1, X_2, \dots, X_n$ be the values of a random sample from a population determined by the random variable X which has finite mean m and variance σ^2. Suppose further that either X is normally distributed or n is large enough that $\overline{X}$ can be considered normally distributed. Then, if we assume σ is known, the confidence interval is given by

$$\overline{X} \pm E,$$

where

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

z_α is defined to be the largest value of z such that $\Phi(z) = 1 - \alpha$.

Theorem: Suppose $X_1, X_2, \ldots, X_n$ are the values of a random sample from a population determined by a normally distributed random variable X with unknown mean and variance.

(i) The confidence interval for the mean of X is given by

$$\overline{X} \pm E,$$

where

$$E = t_{\alpha/2}\sqrt{\frac{S^2}{n}},$$

and $t_{\alpha/2}$ is defined by $P(T > t_{\alpha/2}) = \frac{\alpha}{2}$, where T has a Student's t -distribution with $n-1$ degrees of freedom.

(ii) The confidence interval for σ^2 is given by

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1S^2}{\chi^2_{1-\alpha/2}},$$

where the chi-square distribution has $n-1$ degrees of freedom.

Maximum likelihood estimation:

Suppose that X is a random variable, discrete or continuous, whose distribution depends upon a single parameter ϑ. Let $X_1, X_2, \ldots, X_n$ be an observed sample. If X is discrete, the probability that a random sample consists of exactly these values is given by

$$L(\vartheta) = p(X_1)p(X_2)\ldots p(X_n),$$

where p is the probability mass function of X. The function L is called the likelihood function. If X is continuous with density function f, then the likelihood function L is defined by

$$L(\vartheta) = f(X_1)f(X_2)\ldots f(X_n).$$

The maximum likelihood estimate of ϑ is the value of ϑ that maximizes the value of the likelihood function.

In many cases it is more convenient to work with $l = lnL$, the logarithmic likelihood function. Since the logarithm function is a monotonically increasing function, a maximum of L is a maximum of l and vice versa.

Hypothesis testing

Hypothesis testing is a procedure for determining, from information contained in a random sample from a population, whether to accept or reject a certain statement (hypothesis) about the random variable determining the population.

Hypothesis Test Procedure

1. upon a null hypothesis H_0 and an alternative hypothesis H_1.
2. a test statistic, that is, a formula for calculating a number based upon the random sample, say $t(X_1,\dots,X_n)$.
3. a rejection region (sometimes called a critical region) for values of the test statistics, that is, choose a set of possible test statistic values such that if H_0 is true, then the probability that the value of the statistic will fall in the rejection region is α. Here α is a preselected value, called the level of significance of the test.
4. the test statistic of a random sample from the population. If this falls in the rejection region, reject H_0 and accept H_1, otherwise accept H_0.

There are two distinct types of error which may arise from significance testing

Type I error: we accept H_1 when H_0 is true,

Type II error: we accept H_0 when it is false.

We denote its size for a particular test by β.

The "goodness" of a particular statistical test, for fixed α and H_0, is measured by the *power of the test* which is

$$1-\beta = P(rejecting\ H_0 \mid H_1\ is\ true).$$

In some cases it can be proven that a particular test has the maximum power of any test of a hypothesis H_0 at the α level of significance against an alternative hypothesis; such a test is called a *uniformly most powerful test.*

A *goodness-of-fit test* is a special hypothesis test in which the null hypothesis is that the population is determined by a particular type of a random variable and the alternative is that it is not.

Chi-Square Goodness-of-Fit Test: Each element of a given random sample $X_1,\dots,X_n$ falls into exactly one of k categories $C_1, C_2, \dots, C_k$. This test will determine at the α level of significance whether or not it is reasonable to suppose the observed distribution of the n sample values into categories is consistent with the null hypothesis that X has the given distribution.

Algorithm of Chi-Square Test:

- *Step 1* Count the number O_i of observed elements in category C_i, for $i = 1,2,\dots,k$.
- *Step 2* On the basis of the null hypothesis, calculate E_i, the expected number of elements in category C_i, for $i = 1,2,\dots,k$.
- *Step 3* Calculate the chi-square statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

- *Step 4* Calculate the number of degrees of freedom m of the underlying chi-square distribution. Set $m = k-1$. Then subtract one from m for each independent parameter that is estimated from the data to generate the E_i values in Step 2.

- *Step 5* Find the critical value χ_α^2 such that the probability a chi-square random variable with m degrees of freedom will exceed χ_α^2 is α.
- *Step 6* If $\chi \geq \chi_\alpha^2$ then reject H_0, otherwise accept H_0.

Remark: It has been found, empirically, that the chi-square test works best when all the E_i are at least 5.

PROBLEM

EXAMPLE 1 A Service receives requests for service in a Poisson pattern and wants to estimate the average arrival rate λ from the sample $k_1, k_2, \ldots, k_n$ of arrivals per one-minute interval. Find the maximum likelihood estimation for λ.

Solution:

Let $\bar{k} = \dfrac{k_1 + k_2 + \cdots + k_n}{n}$.

Thus

$$l = lnL = -ln(k_1!k_2!\cdots k_n!) - n\lambda + n\bar{k}ln\lambda,$$

therefore

$$\frac{\partial l}{\partial \lambda} = -n + \frac{n\bar{k}}{\lambda} = 0.$$

Solving for λ yields

$$\hat{\lambda} = \bar{k} = \frac{k_1 + k_2 + \cdots + k_n}{n}.$$

Thus the sample mean is the maximum likelihood estimate for λ.

PROBLEM

EXAMPLE 2 Before the introduction of a new drug, the success rate in the treatment of patients for a certain disease was 43%. However, of 88 patients treated with the drug, 51 recovered. On the basis of this evidence, is the drug significantly effective in treating the disease?

Solution:

Null Hypothesis: That the drug has no effect, i.e., the probability of a patient recovering is still 0.43.

We therefore have a binomial distribution, with $n = 88, p = 0.43$.

Since n is large we can approximate by a normal distribution with $m = np = 37.8$ and $\sigma = \sqrt{np(1-p)} = 4.64$.

Therefore $P(51$ or more recoveries $) = 1 - \Phi(\dfrac{50.5 - 37.8}{4.64}) = 0.003$ (using a continuity correction).

As $0.003 < 0.01,$ this result is significant at the 1% level. So the probability, on the basis of the null hypothesis, of obtaining the observed result is very remote, and we can say with considerable confidence that the drug is effective.

PROBLEM

EXAMPLE 3 The average number of lines of code per programmer-day for each of 30 programs produced at Huzunga Enterprises has been collected. The average number of lines of code per programmer-day, X, is normally distributed. If $\overline{X} = 75$ and $S^2 = 90$, find 95 % confidence intervals for the mean and for the variance of X.

Solution: For 29 degrees of freedom, $\chi^2_{0.975} = 16.047$ and $\chi^2_{0.025} = 45.722$. Hence, the confidence interval for variance is the interval from 57.084 to 162.647.

PROBLEM

EXAMPLE 4 Data is collected on the number of messages arriving at a message switching center during the peak period. The results are tabulated in Table 4.1. The table shows that for the 500 different minutes observed, there were no message arrivals during 4 of the minutes, exactly one message arrival during 17 different observed minutes, etc. At the 5% level of significance, does the arrival pattern appear to be Poisson?

i:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
O_i:	4	17	42	66	86	101	69	52	24	19	11	4	1	2	0	2

Table 7.1: Data on Number of Message Arrivals in One-Minute Period

---

Solution: We use the chi-square test. Let X be a random variable that counts the number of arrivals during a one-minute period. Step 1 has already been done. To determine the E_i or expected numbers we must find the probability of i arrivals in one minute, assuming a Poisson distribution. Since the parameter λ of the assumed Poisson process was not part of the null hypothesis, we must estimate it from data. λ should be the average number of arrivals per minute so we use the estimate

$$\hat{\lambda} = \sum_{i=0}^{15} \frac{iO_i}{500},$$

that is, the total number of arrivals in 500 minutes, divided by 500. This calculation yields 5.022, so we use the estimate of $\hat{\lambda} = 5$ arrivals per minute. Using this λ and the Poisson formula

$$P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = 0, 1, 2, \ldots.$$

We calculate the E_i by the formula

$$E_i = 500\, P(X = i).$$

We must modify the Table 7.1 since there are too few expected in a number of the categories and the total probabilities do not add up to 1. We make a new category of " 11 or more arrivals" and assignate it the probability $1 - P(X \leq 10)$. Then we calculate

$$\chi^2 = \sum_{i=0}^{11} \frac{(O_i - E_i)^2}{E_i} = 6.057.$$

We estimated the parameter λ to generate the E_i values, so the number of degrees of freedom m in our underlying chi-square distribution is 10. The critical value $\chi^2_{0.05}$ is 18.31. Since 6.057 is less than 18.31, we accept the hypothesis that the arrival pattern is Poisson.

2. Exercises

Solutions: ⊙ visible ○ invisible

1. The lengths, in millimetres, of eight full-grown fish of a certain species were found to be:

56, 49, 68, 58, 63, 60, 55, 59.

Obtain unbiased estimates for the mean and variance of the length of the species.

Solution: 58.5, 31.7.

2. A sample of 15 leaves was obtained from a certain plant, and the areas of these leaves, in square centimetres, are given below:

4.60, 5.85, 5.13, 3.94, 5.61, 4.08, 4.71, 4.29, 6.01, 5.46, 5.80, 6.33, 4.43, 4.90, 5.09.

Obtain unbiased estimates of the mean and variance of leaf area (correct to 3 decimal places).

Solution: 5.082, 0.551.

3. t is an unbiased estimator for a parameter θ, and $t_i, i = 1, 2, \ldots, n$ are values of t obtained from n experiments.

i. Prove that $\frac{1}{n}\sum_{i=1}^{n} t_i$ is an unbiased estimate of θ.

ii. Find the relationship between the numbers $a_i, i = 1, \ldots, n$ if $\sum a_i t_i$ is an unbiased estimate of θ.

Solution: (ii) $\sum_{i=1}^{n} a_i = 1.$

4. A firm produces jars of jam in such a way that the net mass of jam per jar is known to be normally distributed with mean 340 g and standard deviation 12 g. The jars are packed in cartons of $16.$ In what percentage of the cartons will the average mass of jam per jar be less than 336 g?

The machine breaks down. After it is repaired, it is found that the average mass of jam in the first 100 jars it fills is 342 g. Has the performance of the machine significantly altered?

Solution: $9\%,\quad P(|mass - 340| \geq 2) = 0.096.$

5. Over a number of years an average of 40 out of 100 patients who underwent a difficult operation survived. Last year new medical techniques were introduced and 73 out of 150 patients survived the operation. Explain whether or not it is statistically justified to continue using these techniques.

Solution: Significant at 5% level.

6. The masses of a certain species of rabbit are known to be normally distributed with mean 1.68 kg and standard

deviation 0.24 kg. Nine rabbits are fed on specially enriched foodstuffs, and their average mass after two moths is found to be 1.85 kg. Test at

i. the 0.05 level,

ii. the 0.01 level, the hypothesis that the foodstuff does not increase the rabbits' masses.

Solutions:
(i) Significant, (ii) not significant evidence against the hypothesis.

---

7. According to a certain hypothesis the variable X is $U(0,1)$. Determine the probability that the smallest member of a sample of n observations of X has value not less than a where $0 \le a \le 1$.

Ten observations were taken of X; the values of the smallest and largest were 0.2 and 0.7. Calculate the probability that ten observations:

i. are each not less than 0.2,

ii. are each not greater than 0.7,

iii. lie within an interval of length 0.5.

State whether your results would lead you to reject the hypothesis.

Solutions:
$(1-a)^n$, (i) 0.107, (ii) 0.028, (iii) 0.001.

---

8. It is known that the random variable X is normally distributed with standard deviation 8. The average size of a sample of 10 values of X is found to be 63. Find (i) 90%, (ii) 95%, (iii) 99% confidence limits for the mean value of X.

Solutions:
(i) 58.8, 67.2, (ii) 58.0, 68.0, (iii) 56.5, 69.5.

---

9. The melting points, in $^\circ C$, of 10 samples of a certain metal were found to be 1154, 1151, 1154, 1150, 1148, 1152, 1155, 1153, 1149, 1154. Past experience indicates that these observations will be normally distributed, with standard deviation equal to $3^\circ C$. Find (i) 95%, (ii) 99% confidence interval for the mean melting point of the metal.

Solutions:
(i) 1150.1, 1153.9, (ii) 1149.6, 1154.4.

---

10. In a random sample of 1000 housewives from a large population, 300 stated that they used a certain detergent. Show that 95% confidence limits for the proportion of the population using this detergent are (approximately) 0.272 to 0.328. (Use a normal approximation to the binomial distribution, and an estimated value of σ).

Following an advertising campaign a second sample of 800 housewives was taken, and of these 260 stated that they used the detergent. Is this evidence of the success of the campaign?

Subsequently it was decided to give a free gift with each packet of the detergent, and later in a random sample of 600 housewives 216 stated that they used the detergent. Does this indicate the success of the free gift scheme?

Solutions:
No; yes.

---

11. It is known that an examination paper is marked in such a way that the standard deviation of the marks is 15.1. In a certain school, 80 candidates take the examination, and they have an average mark of 57.4. Find (i) 95%, (ii) 99% confidence interval for the mean mark in the examination.

Solutions:
(i) 54.1, 60.7, (ii) 53.05, 61.75.

---

12. A factory manufacturing ammeters tests them for zero errors in their calibration. From past routine tests, it is known that the standard deviation of these errors is 0.3.

A batch of 9 ammeters taken from one worker's production has zero errors of 1.0, -0.1, -0.3, 1.6, 0.5, 0.4, 0.5, 0.2, -0.2. Test whether there is evidence of bias in the ammeters produced by this worker, and establish a 95% confidence interval for the mean zero error of this ammeters.

Solution:
Highly significant that there is bias, 0.2 to 0.6.

---

13. We wish to test a die by casting it many times. Suppose that, in an actual experiment consisting of 600 casts, the following distribution of occurrences resulted: $(100,100,115,110,90,85)$. Should we consider this to be evidence of a fair die (significance level 0.10) ?

Solution: $\chi^2_{0.10,5} = 9.24 > 6.50$. Significant.

---

14. A milk distributor claims that the milk sold by his company contains on the average 0.0110 pounds of butterfat per quart. A research laboratory checks this claim by taking random samples. It is known that the standard deviation of butterfat for this distributor is 0.0012 pounds.

i. The laboratory takes a random sample of 16 quarts and finds an arithmetical average of 0.0108 pound of butterfat. Would you conclude that the distributor's claim was false? (Use a significance level 0.1.)

ii. What is the error of type II if the true mean equals 0.0105 ?

iii. How big a sample is needed to make the answer in (ii) equal 0.01 ?

Solutions:
(i) No, (ii) $\beta = 0.7378$.

---

15. It is required to compare the effect of two dyes on cotton fibres. A random sample of 10 pieces of yarn were chosen; 5 pieces were treated with dye A, and 5 with dye B. The results were

Dye	A	4	5	8	8	10
Dye	B	6	2	9	4	5

Solution:
(i) Not significant, (ii) $n = 5$, (iv) $1.5 \le m \le 4.5$.

i. Test the significance of the difference between the two dyes. (Assume the normality, common variance, and significance level 0.05.)

ii. How big a sample do you estimate would be needed to detect a difference equal to 0.5 with probability 99 per cent?

iii. Draw the power curve of the test against the mean difference, using the estimate of variance from the experiment.

iv. Find a 90 per cent confidence interval for mean difference between dyes A and B.

---

16. Consider testing the null hypotheses $H_0 : \lambda = \lambda_0$ against the alternative $\lambda < \lambda_0$ for a Poisson distribution, where λ denotes the mean rate per unit time at which events occur.

One experiment for testing H_0 is to observe the number of events a which occur in time interval $[0,h]$. Let $\lambda_0 = 4$, $\alpha = 0.05$ (error type I).

i. Describe the test procedure when $h = 2$ and 6.

ii. Plot the power function for $h = 2$ and 6.

Another experiment for testing H_0 is to observe the time T until k events occur. Again, let $\lambda_0 = 4$, $\alpha = 0.05$ (error type I).

i. Find the test procedure when $k = 5$ and 10.

ii. Plot the power function for $k = 5$ and 10.

iii. Plot $E(T \mid \lambda)$ as a function of λ for $k = 5$ and 10, where $E(T \mid \lambda)$ denotes the expected duration of the experiment when the true value equals λ.

iv. Compare the two types of experiment.

---

17. Suppose that X, Y, Z are three independent random variables such that $E(X) = E(Y) = E(Z) = m$ and $Var(X) = 1$, $Var(Y) = 2$, $Var(Z) = 3$. Given observations X, Y, Z, find an unbiased linear estimate of m with minimum variance.

Solution: X.

---

18. Suppose that X, Y, Z are three independent random variables with mean m and variance σ^2.

i. Is $\frac{X}{2} + \frac{Y}{4} + \frac{Z}{4}$ an unbiased estimate of m?

ii. Is $\frac{X}{3} + \frac{Y}{3} + \frac{Z}{6}$ an unbiased estimate of m?

iii. Find a linear unbiased estimator with minimum variance.

iv. What is the linear estimator with minimum mean-square error when the coefficient of variation $\frac{\sigma}{m}$ is known to be equal to 1?

Solution:

(i) Yes, (ii) no, (iii) $\frac{X+Y+Z}{3}$.

---

19. Consider a lake with N fish of a certain kind. Let n fish be chosen, tagged and thrown back in the lake. Next, a random sample, with replacement, of k fish is taken from the lake. Let T be the number of tagged fish in the sample.

i. What is the distribution of T ?

ii. How can T be used for estimating N ?

iii. What is the approximate variance of the estimate obtained in (ii)?

Solutions:
(i) $b\left(k, \frac{n}{N}\right)$, (ii) $N = \frac{kn}{T}$.

---

20. The lifetime of an item has an exponential distribution with mean m. One hundred items were placed on life test for 200 hours. At the end of this period, the items were examined and 40 were defective.

i. What is the maximum likelihood estimate of m on the basis of these data?

ii. What is the approximate variance of estimate in (i)?

iii. Find a two-sided confidence interval for m with confidence coefficient 0.90.

iv. Find a lower-sided confidence interval for m with confidence coefficient 0.90.

Solutions:
(i) 391.52, (ii) 1532.90.

---

21. A certain plastic gave the following deflections Y in a tension machine when loaded with the amounts shown, x. The loads are exactly determined, and the distributions of Y for fixed x may be assumed normal, with constant variance over their x range.

x	2	3	4	5	6	7	
Y	7.5	18.6	19.0	23.9	32.5	30.0	
x	8	9	10	11	12	13	14
Y	40.5	50.0	40.0	56.3	60.5	62.5	70.0

(i) Find the least-squares line to fit these data (estimate slope and intercept).

(ii) Test the hypothesis that the slope is zero.

Solutions:
(i) $y = 4.9x + 0.13$, (ii) $r = 0.98$.

---

22. Suppose the random variable X has the density function

$$f(x)=\begin{cases}(1+\lambda)x^{\lambda}, \text{ ha } 0<x<1,\\ 0, \text{ otherwise.}\end{cases}$$

Show that the maximum likelihood estimate of λ based on a given random sample of size n is $\hat{\lambda}=-\left(1+\frac{n}{\sum_{i=1}^{n} \ln x_i}\right)$.

---

23. A sample believed to be from a normal population consists of the 20 numbers

14.56, 20.55, 14.1, 21.2, 24.57, 24.13, 9.68, 19.09, 14.51, 21.77, 14.72, 16.53, 24.92,

21.4, 14.95, 31.43, 17.86, 12.72, 18.54, and 23.92.

Find a 95% confidence interval for the mean m and the variance σ^2 of the population.

Solutions:
$16.600 < m < 21.516, \quad 15.959 < \sigma^2 < 58.865.$

Digitális Egyetem, Copyright © Fegyverneki Sándor, 2011