UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

Problems on Statistical Inference Unit 1

Some problems adapted from Solved Exercises and Problems of Statistical Inference by David Casado

1. Cats weigh 5kg on average with standard deviation 2 kg. The distribution is positively skewed. Describe the distribution of the average of a sample of 25 cats.

2. Female height in Benin have average height 160 cm with variance 400. The heights of 25 women are measured
   (a) Describe the distribution of the mean of such a sample
   (b) Find the probability that the mean of such a sample is above 165 cm

3. A sample of 50 *centenarians* found that the age at which they died has Σx=5200 and Σx^2=540,000
   Find an estimate for the mean life expectancy and standard deviation of all centenarians

4. Donkey's ears are known to have standard deviation 5 cm. A farmer samples 30 donkeys and finds they have ears with average length 20 cm
   (a) Construct a 95% confidence interval for the mean length of this farmer's donkeys
   (b) A veterinarian claims the donkeys have mean length below 22 cm. Is there evidence to support this claim?

5. Every week a dairy farm samples 50 eggs to check if any have imperfections. The data for the first 10 weeks is given below.

Week	1	2	3	4	5	6	7	8	9	10
Number of eggs with imperfections	3	2	5	4	3	5	3	6	11	8

   (a) Calculate an estimate for the proportion of eggs with imperfections
   (b) Find the upper and lower control limits

6. The weight of 30 chestnuts in grams is found to have Σx=75 and Σx^2=300. Calculate an estimate for the mean and standard deviation of the weight of chestnuts.

7. The number of sesame seeds on five big mac buns was: 150, 120, 130, 127, 133. Determine a 95% confidence interval for the mean number of sesame seeds on a big mac bun, and interpret it in context. From a long term study the variance is known to be 124.5 (so does not need to be calculated).

8. Suppose that you manage a bank where the amounts of daily deposits and daily withdrawals are given by independent random variables with normal distributions. For deposits, the mean is £12,000 and the Standard deviation is £4,000; for withdrawals, the mean is £10,000 and the standard deviation is £5,000.
   (a) For a week, calculate the probability that the five withdrawals will add up to more than £55,000.

(b) For a particular day, calculate the probability that withdrawals will exceed deposits by more than £5,000.

9. We have reliable information that suggests the probability distribution with density function $f(x;\theta)=2\theta^2(\theta-x)$, $x\in[0,\theta]$, as a model for studying the population quantity X. Let $X=(X_1,\ldots,X_n)$ be a simple random sample. Try to apply the maximum likelihood method to find and estimator of the parameter θ.

Solution:

$$\ell(\theta;x)=ln\left(\prod_{i=1}^{n}2\theta^2(\theta-x_i)\right)=\sum_{i=1}^{n}\left(ln(2\theta^2)+ln(\theta-x_i)\right)$$

$$\frac{\ell(\theta;x)}{\partial\theta}=\frac{2n}{\theta}+\sum_{i=1}^{n}\frac{1}{\theta-x_i}=0\rightarrow 2n+\sum_{i=1}^{n}\frac{\theta}{\theta-x_i}=0\rightarrow\quad ?$$

$\hat{\theta}$ *can not be found*

10. Let X be a random variable following the Rayleigh distribution, whose with probability function is given below. Apply the maximum likelihood method to find and a ML estimator of the parameter θ.

$$f(x;\theta)=\frac{x}{\theta^2}e^{-\frac{x^2}{2\theta^2}},\quad x\geq 0,\quad(\theta>0)$$

such that $\boldsymbol{E}(X)=\theta\sqrt{\frac{\pi}{2}}$ *and* $Var(X)=\frac{4-\pi}{2}\theta^2$.

Solution:

$$\ell(\theta;x)=ln\left(\prod_{i=1}^{n}\frac{x_i}{\theta^2}exp\left(-\frac{x_i^2}{2\theta^2}\right)\right)=\sum_{i=1}^{n}\left(ln\left(\frac{x_i}{\theta^2}\right)-\frac{x_i^2}{2\theta^2}\right)$$

$$\frac{\partial\ell(\theta;x)}{\partial\theta}=-\frac{2n}{\theta}+\sum_{i=1}^{n}\frac{x_i^2}{\theta^3}=0\rightarrow -2n+\sum_{i=1}^{n}\frac{x_i^2}{\theta^2}=0\rightarrow\quad\sum_{i=1}^{n}x_i^2=2n\,\theta^2\rightarrow$$

$$\hat{\theta}=\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{2n}}$$

Second order conditions have to be proof (hessian matrix has to be negative defined).

$$\frac{\partial^2\ell(\theta;x)}{\partial\theta^2}=\frac{2n}{\theta^2}-3\sum_{i=1}^{n}\frac{x_i^2}{\theta^4}$$

$$\frac{\partial^2\ell(\hat{\theta};x)}{\partial\theta^2}=\frac{(2n)^2}{\sum_{i=1}^{n}x_i^2}-3\sum_{i=1}^{n}\frac{(2n)^2x_i^2}{\left(\sum_{i=1}^{n}x_i^2\right)^2}=\frac{(2n)^2\sum_{i=1}^{n}x_i^2}{\left(\sum_{i=1}^{n}x_i^2\right)^2}-3\sum_{i=1}^{n}\frac{(2n)^2x_i^2}{\left(\sum_{i=1}^{n}x_i^2\right)^2}\rightarrow$$

$$=\frac{(2n)^2\sum_{i=1}^{n}x_i^2}{\left(\sum_{i=1}^{n}x_i^2\right)^2}-3(2n)^2\sum_{i=1}^{n}\frac{x_i^2}{\left(\sum_{i=1}^{n}x_i^2\right)^2}=-\frac{2(2n)^2}{\sum_{i=1}^{n}x_i^2}<0$$

11. To forecast the yearly inflation (in percent, %), a simple random sample has been gathered: 1.5, 2.1, 1.9, 2.3, 2.5, 3.2, 3.0. It is assumed that the variable inflation follows a normal distribution.
(a) By using these data, construct a 99% confidence interval for the mean of the inflation.
(b) Experts have the opinion that the previous interval is too wide, and they want a total length of a unit.
Find the level of confidence for this new interval.
(c) Construct a confidence interval of 90% for the standard deviation.

12. In the library of a university, the mean duration (in days, d) of the borrowing period seems to be 20d. A simple random sample of 100 books is analyzed, and the values 18d and 8d2 are obtained for the sample mean and the sample variance, respectively. Construct a 99% confidence interval for the mean duration of the borrowings to check if the initial population value is inside.

13. The accounting firm Price Waterhouse periodically monitors the U.S. Postal Service's performance. One parameter of interest is the percentage of mail delivered on time. In a simple random sample of 332,000. mailed items, Price Waterhouse determined that 282,200 items were delivered on time (Tampa Tribune, March 26, 1995.) Use this information to estimate with 99% confidence the true percentage of items delivered on time by the U.S. Postal Service.

14. Two independent groups, A and B, consist of 100 people each of whom have a disease. A serum is given to group A but not to group B, which are termed treatment and control groups, respectively; otherwise, the two groups are treated identically. Two simple random samples have yielded that in the two groups, 75 and 65 people, respectively, recover from the disease. To study the effect of the serum, build a 95% confidence interval for the difference ηA–ηB. Does the interval contain the case ηA = ηB?

15. The lengths (in millimeters, mm) of metal rods produced by an industrial process are normally distributed with a standard deviation of 1.8mm. Based on a simple random sample of nine observations from this population, the 99% confidence interval was found for the population mean length to extend from 194.65mm to 197.75mm. Suppose that a production manager believes that the interval is too wide for practical use and, instead, requires a 99% confidence interval extending no further than 0.50mm on each side of the sample mean. How large a sample is needed to achieve such an interval?

16. The mark of an aptitude exam follows a normal distribution with standard deviation equal to 28.2. A simple random sample with nine students yields the following results:
$\sum_{i=1}^{9} x_i$ =1,098 and $\sum_{i=1}^{9} x_i^2$ =138,148
a) Find a 90% confidence interval for the population mean μ.
b) Discuss without calculations whether the length of a 95% confidence interval will be smaller, greater
or equal to the length of the interval of the previous section.
c) How large must the minimum sample size be to obtain a 90% confidence interval with length (distance between the endpoints) equal to 10?

17. A 64-element simple random sample of petrol consumption (litres per 100 kilometers, u) in private cars has been taken, yielding a mean consumption of 9.36u and a standard deviation of 1.4u. Then:
a) Obtain a 96% confidence interval for the mean consumption.
b) Assume both normality (for the consumption) and variance σ2 = 2u2. How large must the sample be if, with the same confidence, we want the maximum error to be a quarter of litre?

18. A company makes two products, A and B, that can be considered independent and whose demands follow the distributions N(μA, σA=70u) and N(μB, σB=60u), respectively. After analyzing 500 shops, the two simple random samples yield means of a = 156 and b = 128.
(a) Build 95 and 98 percent confidence intervals for the difference between the population means.

(b) What are the absolute errors? If sales are measured in the unit u = number of boxes, what is the unit of measure of the margin of error?
(c) A margin of error equal to 10 is desired, how many shops are necessary?
(d) If only product A is considered, as if product B had not been analyzed, how many shops are necessary to guarantee a margin of error equal to 10?

19. The lifetime of a machine (measured in years, y) follows a normal distribution with variance equal to $4y^2$. A simple random sample of size 100 yields a sample mean equal to 1.3y. Test the null hypothesis that the population mean is equal to 1.5y, by applying a two-tailed test with 5 percent significance level. What is the type I error? Calculate the type II error when the population mean is 2y.

20. A company produces electric devices operated by a thermostatic control. The standard deviation of the temperature at which these controls actually operate should not exceed 2.0ºF. For a simple random sample of 20 of these controls, the sample quasi-standard deviation of operating temperatures was 2.39ºF.

    a. Stating any assumptions, you need (write them), test at the 5% level the null hypothesis that the population standard deviation is not larger than 2.0ºF against the alternative that it is.
    b. Calculate the type II error at $\sigma^2=4.5ºF^2$.

21. Let X = (X1,...,Xn) be a simple random sample with 25 data taken from a normal population variable X. The sample information is summarized in

$$\sum_{j=1}^{25} x_j=105 \qquad and \qquad \sum_{j=1}^{25} x_j^2=579.24$$

(a) Should the hypothesis H0: $\sigma2 = 4$ be rejected when H1: $\sigma2 > 4$ and $\alpha = 0.05$?
(b) And when H1: $\sigma2 \neq 4$ and $\alpha = 0.05$?

22. The distribution of a variable is supposed to be normally distributed in two independent biological populations. The two population variances must be compared. After gathering information through simple random samples of sizes $n_X = 11$, $n_Y = 10$, respectively, we are given the value of the estimators

$$S_X^2=\frac{1}{n_X-1}\sum_{j=1}^{n_X}(x_j-\bar{x})^2=6.8 \qquad s_Y^2=\frac{1}{n_Y}\sum_{j=1}^{n_Y}(y_j-\bar{y})^2=7.1$$

For $\alpha = 0.1$, test:
(a) H0: $\sigma_X = \sigma_Y$ against H1: $\sigma_X < \sigma_Y$
(b) H0: $\sigma_X = \sigma_Y$ against H1: $\sigma_X > \sigma_Y$
(c) H0: $\sigma_X = \sigma_Y$ against H1: $\sigma_X \neq \sigma_Y$

23. Two simple random samples of 700 citizens of Italy and Russia yielded, respectively, that 53% of Italian people and 47% of Russian people wish to visit Catalonia within the next ten years. a) Should we conclude, with confidence 0.99, the Italians' desire is higher than the Russians'? b) Determine the critical region and make a decision. What is the type I error? c) Calculate the p-value and apply the methodology based on the p-value to make a decision.

24. The fog index is used to measure the reading difficulty of a written text: The higher the value of the index, the more difficult the reading level. We want to know if the reading difficulty index is different for three magazines: Scientific American, Fortune, and the New Yorker. Three independent random samples of 6 advertisements were taken, and the fog indices for the 18 advertisements were measured, as recorded in the following table

SCIENTIFIC AMERICAN	*FORTUNE*	*NEW YORKER*
15.75	12.63	9.27
11.55	11.46	8.28
11.16	10.77	8.15
9.92	9.93	6.37
9.23	9.87	6.37
8.20	9.42	5.66

Apply an analysis of variance to test whether the average level of difficulty is the same in the three magazines.

25. The following table is based on data from the U.S. Department of Labor, Bureau of Labor Statistics.

	Police	**Cashiers**	**Taxi Drivers**	**Guards**	
Homicide	82	107	70	59	
Cause of death other than homicide	92	9	29	42	
					490

a) Use the data in the table, coming from a simple random sample, to test the claim that occupation is independent of whether the cause of death was homicide. Use a significance α = 0.05 and apply a nonparametric chi-square test.
b) Does any particular occupation appear to be most prone to homicides? If so, which one?

26. World War II Bomb Hits in London. To carry out an analysis, South London was divided into 576 areas. For the variable N ≡ number of bombs in the k-th area (any), a simple random sample (x1,...,x576) was gathered and grouped in the following table:

EMPIRICAL							
Number of Bombs	**0**	**1**	**2**	**3**	**4**	**5 or more**	
Number of Regions	229	211	93	35	7	1	$n = 576$

By applying the chi-square goodness-of-fit methodology,
(1) Test at 95% confidence whether N can be supposed to follow a Poisson distribution.
(2) Test at 95% confidence whether N can be supposed to follow a Poisson distribution with λ = 0.8.

27. To test if a coin is fair or not, it has independently been tossed 100,000 times (the outputs are a simple random sample), and 50,347 of them were heads. Should the fairness of the coin, as null hypothesis, be rejected when α = 0.1?
    (a) Apply a parametric test.
    (b) Apply the nonparametric chi-square goodness-of-fit test.

28. From a previous pilot study, concerning the monthly amount of money (in $) that male and female students of a community spend on cell phones, the following hypotheses are reasonably supported:
    i. The variable amount of money follows a normal distribution in both populations.
    ii. The population means are μM = $14.2 and μF = $13.5, respectively.
    iii. The two populations are independent.
    Two independent simple random samples of sizes nM = 53 and nF = 49 are considered, from which the
    following statistics have been calculated:

$$S_M^2 = \$^2 4.99 \qquad and \qquad S_F^2 = \$^2 5.02$$

Then,

A) Calculate the probability P($\overline{M} - \overline{F}$ ≤1.27). Repeat the calculations with the supposition that σM and σF are equal.

B) Build a 95% confidence interval for the quotient σ_M/σ_F.

29. The electric light bulbs of manufacturer X have a mean lifetime of 1400 hours (h), while those of manufacturer Y have a mean lifetime of 1200h. Simple random samples of 125 bulbs of each brand are tested. From these datasets the sample variances Sx^2 = 156h2 and Sy^2 = 159h2 are computed. If manufacturers are supposed to be independent and their lifetimes are supposed to be normally distributed:
a) Build a 99% confidence interval for the quotient of standard deviations σX/σY. Is the value σX/σY=1, that is, the case σX=σY, included in the interval?
b) By using the proper statistic T, find k such that P($\overline{X} - \overline{Y}$ ≤k)= 0.4.

30. In 1990, 25% of births were by mothers of more than 30 years of age. This year a simple random sample of size 120 births has been taken, yielding the result that 34 of them were by mothers of over 30 years of age.
a) With a significance of 10%, can it be accepted that the proportion of births by mothers of over 30 years of age is still ≤ 25%, against that it has increased? Select the statistic, write the critical region and make a decision.
b) Obtain a 90% confidence interval for the proportion. Use it to make a decision about the value of π, which is equivalent to having applied a two-sided (nondirectional) hypothesis test in the first section.

31. Assume that the height (in centimeters, cm) of any student of a group follows a normal distribution with variance 55cm2. If a simple random sample of 25 students is considered, calculate the probability that the sample sample variance will be bigger than 64.625cm2.

32. A poll of 1000 individuals, being a simple random sample, over the age of 65 years was taken to determine the percent of the population in this age group who had an Internet connection. It was found that 387 of the 1000 had one. Find a 95% confidence interval for π.

33. Is There Intelligent Life on Other Planets? In a 1997 an institute survey of 935 randomly selected Americans, 60% of the sample answered “yes” to the question “Do you think there is intelligent life on other planets?” Let's use this sample estimate to calculate a 90% confidence interval for the proportion of all Americans who believe there is intelligent life on other planets. What are the margin of error and the length of the interval?

34. Plastic sheets produced by a machine are constantly monitored for possible fluctuations in thickness (measured in millimeters, mm). If the true variance in thicknesses exceeds 2.25 square millimeters, there is cause for concern about product quality. The production process continues while the variance seems smaller than the cutoff. Thickness measurements for a simple random sample of 10 sheets produced in a particular shift were taken, giving the following results:
(226, 226, 227, 226, 225, 228, 225, 226, 229, 227)
Test, at the 5% significance level, the hypothesis that the population variance is smaller than 2.25mm2.
Suppose that thickness is normally distributed. Calculate the type II error β(σ^2 =2).

35. If 132 of 200 male voters and 90 of 159 female voters favor a certain cantidate running for governor of Illinois, find a 99% confidence interval for the difference between the actual proportions of male and female voters who favor the candidate.

36. A research worker wants to determine the average time it takes a mechanic to rotate the tires of a car, and she wants to be able to assert with 95% confidence that the mean of her sample is off by at most 0.50 minute. If she can presume from past experience that $\sigma = 1.6$ minutes (min), how large a sample will she have to take?

37. To estimate the average tree height of a forest, a simple random sample with 20 elements is considered yielding

$$\bar{x} = 14.70\,u \quad and \quad S = 6.34\,u$$

where u denotes a unit of length. If the population variable height is supposed to follow a normal distribution, find a 95 percent confidence interval. What is the absolute error?