PROBLEMS IN BASIC STATISTICS

MANJUNATH KRISHNAPUR

Note: These are problems I gave as homeworks in the many times I taught the first course in UG probability and statistics at IISc. They are taken from various sources and there are also some that I made up.

Problem 1. (*) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Geo(p).

(1) Find the MLE (maximum likelihood estimate) for p. Is it unbiased?

(2) In the file `http://math.iisc.ernet.in/~manju/UGstatprob/geodata2` there are four columns and 100 rows. The first column has samples form i.i.d Geo(0.8) distribution (the second, third and fourth columns are from Geo(0.4), Geo(0.1) and Geo(0.05) distributions, respectively). In each of these four cases and using $n = 25, 50, 100$ (by taking the first n rows only), compute the MLE for p.

Problem 2. (*) In `http://math.iisc.ernet.in/~manju/UGstatprob/heightweight2.txt` you will find data on heights (second column) and weights (third column) of 200 individuals.

(1) Find the sample means, standard deviations and correlation between height and weight.

(2) Assume that the heights are normally distributed with mean μ and variance σ^2. Find the MLE for μ, σ^2.

(3) Do the same for the weight, assuming normal distribution again.

Problem 3. Suppose $X_1, \ldots, X_n$ are i.i.d. from Pois(λ) distribution.

(1) Find the MOM (method of moments) estimate and MLE for λ.

(2) Find the bias and mean squared error of your estimates.

(3) A historically popular data set (collected by von Bortkiewicz): Towards the end of 1800s, data was collected on casualties in the Prussian army by horse kicks. In 200 corps, the number of cavalrymen who died by horse kicks in a year was observed (actually it was 20 corps, but over 10 years). The data is as follows.

No. of casualties	0	1	2	3	4	5	6
No. of corps with observed casualties	109	65	22	3	1	0	0

Assume that the no. of casualties in a given corps in a year has Poisson distribution and that across different corps and years the data are independent. Estimate λ and find the expected frequencies of deaths to the actual observed values.

Problem 4. Let $X_1, \ldots, X_n$ be i.i.d. Unif$[a, b]$. The problem is to estimate a, b.

(1) Let $A_n = \min\{X_1, \ldots, X_n\}$ and $B_n = \max\{X_1, \ldots, X_n\}$. Show that (A_n, B_n) is MLE for (a, b).

(2) Based on A_n and B_n, find unbiased estimates for a and b respectively.

Problem 5. Let $X_1, \ldots, X_n$ be i.i.d. Beta$[a, b]$. The problem is to estimate a, b.

(1) Find MLE for (a, b).

(2) Find the m.s.e for your estimates.

Problem 6. (*) Let $X_1, \ldots, X_n$ be i.i.d. samples from a parametric family of discrete distributions. In each of the following cases, find the MLE for the unknown parameter(s).

(1) X_i are i.i.d. Bin(N, p) where N is known and p is unknown.

(2) X_i are i.i.d. Pois(λ) where λ is unknown.

(3) X_i are i.i.d. Geo(p) where p is unknown.

Problem 7. Let $X_1, \ldots, X_n$ be i.i.d. samples from a parametric family of densities. In each of the following cases, find the MLE for the unknown parameter(s).

(1) X_i are i.i.d. Gamma(ν, λ) where ν is known and λ is unknown.

(2) X_i are i.i.d. Unif$[a, b]$ where a, b are unknown.

(3) X_i are i.i.d. $N(\mu, \sigma^2)$ where μ, σ^2 are unknown.

Problem 8. (1) Let $X_1, \ldots, X_n$ be i.i.d. Unif$([0, 1])$. If M_n is a sample median, show that $\mathbf{P}\{|M_n - \frac{1}{2}| > \delta\} \to 0$ for any $\delta > 0$, as $n \to \infty$.

(2) If X_i are i.i.d from some density $f(x)$ (assume that the median is uniquely defined), deduce that the sample median M_n gets close to the population median $\mathbf{m}$ in the same sense, i.e., $\mathbf{P}\{|M_n - \mathbf{m}| > \delta\} \to 0$ for any $\delta > 0$, as $n \to \infty$.

(3) More generally, for any $0 < q < 1$, show that the sample q quantile $M_n^{(q)}$ is close to the population q-quantile in the same sense.

[**Hint:** In the first part, observe that $M_n \leq t$ if and only if more than half of the X_is are below t. As for the second part, try to deduce it from the first instead of re-doing the proof all over again!]

Problem 9. Let $X_1, \ldots, X_n$ be i.i.d. Exp(λ). Let $\theta = \log \lambda$. Let $\gamma = \int_0^\infty \log t \, e^{-t} \, dt$.

(1) Show that $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n (\gamma - \log X_i)$ is an unbiased estimate for θ.

(2) Compute the m.s.e of $\hat{\theta}$.

(3) Explain how you would give an $(1-\alpha)$-confidence interval for λ, based on $\hat{\theta}$. [**Hint:** If $X \sim \text{Exp}(\lambda)$, the distribution of $\log X + \log \lambda$ does not depend on λ.]

Problem 10. (*) In `http://math.iisc.ernet.in/~manju/UGstatprob/newcomb_lightspeed.txt` you will see the data from Simon Newcomb's experiment on the time taken (in nanoseconds) by light to travel 7442 meters at sea level.

(1) Compute the sample mean and sample standard deviation.

(2) Assuming normal distribution for the data, compute a confidence interval for the time taken. What confidence interval does it give for the speed of light (in meters per second)?

(3) Repeat the same after dropping the smallest two measurements (declared 'outliers').

Problem 11. This is the description of the data given in `http://math.iisc.ernet.in/~manju/UGstatprob/horsekicks.txt`. In each year from 1875 to 1894, the number of cavalrymen who died due to horse-kicks in the Prussian army of the time was counted. The data was collected in 14 different army-units (of equal size), which is what is indicated in the 14 columns following the year column.

Assume that the number of deaths per army-unit per year is a random variable having a Poisson(λ) and that the number of deaths in different units or in different years are independent.

Estimate λ from the given data. Compute the expected frequencies of deaths per units per year from your estimate (and compare with the actual figures).

Problem 12. In the file `http://math.iisc.ernet.in/~manju/UGstatprob/simulatednormaldata.txt` you will see data *simulated* on a computer from a normal distribution with unknown mean and variance. The problem is to test the hypotheses $H_0 : \mu = 2$ versus $H_1 : \mu \neq 0$.

(1) Use only the first n data points for $n = 20, 50, 100, 200, 400, 500, 800, 1000$, and carry out the test for each n at significance level 0.05. Report the p-values.

(2) Repeat the same tests but now assume that the variance is given to be 9.

Problem 13. Are real coins fair? Formulate this as a hypothesis testing problem and perform the test at 0.01 level of significance using the following data. Report the p-value.

(1) In an experiment reported in `http://www.stat.berkeley.edu/~aldous/Real-World/coin_tosses.html`, a real coin was tossed 20000 times. The number of heads observed was 10231.

(2) In another experiment reported on the same page, 10014 heads appeared in 20000 tosses. Repeat the test with this data.

Problem 14. In the file `http://math.iisc.ernet.in/~manju/UGstatprob/twomidtermgrades.txt` you see the scores obtained in two exams by your batch in the first and second midterms, respectively. Test the hypothesis that the overall performance is worse in the second mid-term than in the first.

Problem 15. In `http://math.iisc.ernet.in/~manju/UGstatprob/heightweight2.txt` you will find data on heights (second column) and weights (third column) of 200 individuals.

(1) Test the hypothesis that heights are normally distributed (this is the null hypothesis). Use χ^2-test with different choices of bins (i.e., do it with 10 bins and then with 15, etc. Each bin should have at least 5 observations).

(2) Do the same for weights.

Problem 16. *Benford's law* is the probability distribution with mass function given by $f(k) = \log_{10}(k+1) - \log_{10}(k)$ for $k = 1, 2, \ldots, 9$. It is observed that for various quantities that vary over several orders of magnitude, the first digit follows Benford's law. Here we give a few. In each case, conduct a χ^2-test at level 0.10, with the null hypothesis being that the distribution is indeed Benford's law. Compute the p-value in each case.

(1) Let $F_0, \ldots, F_{999}$ be the first 1000 Fibonacci numbers. This is a (non-random!) sequence of numbers defined by $F_0 = F_1 = 1$ and $F_k = F_{k-1} + F_{k-2}$ for $k \geq 2$. Although there is no randomness, extract the first digits, and compare against Benford's law by a χ^2-test.

(2) Do the same for the sequence of factorials, $1, 2, 6, 24, \ldots$ (go up to 100 or wherever your computer stops to compute the first digit).

(3) In `http://en.wikipedia.org/wiki/List_of_national_capitals_by_population` you will find the populations of the capitals of (almost) all countries in the world. For convenience, the list of populations is given in `http://math.iisc.ernet.in/~manju/UGstatprob/population.txt`. Again, compute the first digits and check the hypothesis that Benford's law applies. Report the p-value.

Quite challenging problems/theorems - optional!

We give two problems addressing two issues that we did not consider in class. One is that in many examples (eg., i.i.d. Exp(λ) data), the sample mean is the UMVU (*uniformly minimum variance unbiased estimate*). Second is that the maximum likelihood estimate is a reasonable choice, at least in the sense that for large sample of data, the MLE is close to the actual value of the parameter with high probability.

Problem 17. Let $f_\theta(x)$ is a collection of densities parameterized by a real number θ. Suppose $X_1, X_2, \ldots$ be i.i.d. samples from f_{θ_0} where θ_0 is fixed (we pretend it is unknown and give estimates for it). Let $\hat{\theta}_n$ be the MLE based on $X_1, \ldots, X_n$, i.e., $\hat{\theta}_n$ maximizes $\ell_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)$ (we write ℓ_n for simplicity but note that it depends on $X_1, X_2, \ldots, X_n$).

(1) Fix Define $\ell^*(\theta) = \mathbf{E}_{\theta_0}[\log f_\theta(X_1)]$. Show that $\ell^*(\theta) < \ell^*(\theta_0)$ with equality if and only if $\theta = \theta_0$. (This last statement is true only if we assume that the densities f_θ are distinct for distinct values of θ, which is a very reasonable assumption!). [**Hint:** For any two densities f and g, show that $\int_{\mathbb{R}} \log\left(\frac{f(x)}{g(x)}\right) g(x)dx \leq 0$.]

(2) Show that $\frac{1}{n}\ell_n(\theta) \xrightarrow{P} \ell^*(\theta)$ for all θ where the convergence is in the sense that $\mathbf{P}\{|\frac{1}{n}\ell_n(\theta) - \ell^*(\theta)| \geq \delta\} \to 0$ as $n \to \infty$, for every $\delta > 0$.

(3) $\hat{\theta}$ maximizes $\frac{1}{n}\ell_n(\theta)$ and θ_0 maximizes $\ell^*(\theta)$. Further the two functions $\frac{1}{n}\ell_n(\theta)$ and $\ell^*(\theta)$ are close to each other. Convince yourself that under some conditions (but not always) this implies that $\hat{\theta}$ and θ_0 must be close to each other. [**Note:** Obviously this last part is vague. The point is that one can impose various conditions on the densities f_θ which ensure that it works. It is enough to get the heuristic idea here.]

Problem 18. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). We want to show that among all unbiased estimates of p, the one with least variance (for any value of p) is $\bar{X}_n$. Let $T : \{0,1\}^n \to [0,1]$ be any unbiased estimate (i.e., if we see the data $(X_1, \ldots, X_n)$, the guess for p would be $T(X_1, \ldots, X_n)$).

(1) Define $S(x_1, \ldots, x_n) = \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} T(x_{\pi_1}, \ldots, x_{\pi_n})$, i.e., permute the arguments x_is in all possible ways and average out the values of T obtained. Show that
   (a) $S(X_1, \ldots, X_n)$ is an unbiased estimate of p.
   (b) $S(X_1, \ldots, X_n)$ depends on $\bar{X}_n$ only. That is, $S(x_1, \ldots, x_n) = S(y_1, \ldots, y_n)$ if $\bar{x} = \bar{y}$.
   (c) $\text{Var}(S(X_1, \ldots, X_n)) \leq \text{Var}(T(X_1, \ldots, X_n))$.

(2) From the second part above, we may write $S(X_1, \ldots, X_n)$ as $g(\bar{X}_n)$ for some function $g : [0,1] \to [0,1]$. By the unbiasedness, $\mathbf{E}_p[g(\bar{X}_n)] = p$ for all $p \in [0,1]$. Show that this implies that $g(\bar{X}_n) = \bar{X}_n$). [**Hint:** Recall that $X_1 + \ldots + X_n$ has binomial distribution.]

(3) Conclude that $\bar{X}_n$ is the UNMVU for this problem.

[**Remark:** The proof that $\bar{X}_n$ (or other specific estimates) are UMVU in other problems is somewhat more difficult, although the same as above once one understands conditional probability well.]

Problem 19. A large box contains 10000 marbles, of which some are red and the others are blue. To estimate the unknown proportion p of red balls, a sample of 100 marbles is drawn at random (with replacement) and it is observed that the number of red balls in the sample is 30. Construct a $1-\alpha$ confidence interval for p when (1) $\alpha = 0.01$, (2) $\alpha = 0.05$, (3) $\alpha = 0.10$. Repeat the same exercise when the number of red marbles in the sample is 40.

Problem 20. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, where μ, σ^2 are both unknown. Summary statistics of the data obtained in an experiment are given as follows:

$$n = 20, \quad \sum_{i=1}^{n} X_i = 60, \quad \sum_{i=1}^{n} X_i^2 = 240.$$

(1) Find a two-sided confidence interval for μ with confidence level 0.90.

(2) Find an upper bound for σ^2 with confidence level 0.90.

Problem 21. Let $X_1, \ldots, X_n$ be i.i.d. Exp(λ). Let $\theta = \log \lambda$. Let $\gamma = \int_0^\infty \log t \, e^{-t} \, dt$.

(1) Show that $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (\gamma - \log X_i)$ is an unbiased estimate for θ.

(2) Compute the m.s.e of $\hat{\theta}$.

(3) Explain how you would give an $(1 - \alpha)$-confidence interval for λ, based on $\hat{\theta}$. [**Hint:** If $X \sim \text{Exp}(\lambda)$, the distribution of $\log X + \log \lambda$ does not depend on λ.]

Problem 22. In each of the following cases, find the bias and m.s.e of the given estimate. The samples are $X_1, \ldots, X_n$, i.i.d. from the given distribution.

(1) Distribution is $N(\mu, \sigma^2)$, both parameters unknown. The estimate (for μ) if $\hat{\mu} = \bar{X}_n$.

(2) Distribution is Ber(p). The estimate for p is $\hat{p} = \bar{X}_n$.

(3) Distribution is Pois(λ). The estimate for λ is $\hat{\lambda} = \bar{X}_n$.

Problem 23. In the above problem, describe how you would construct a $1 - \alpha$ confidence interval for the unknown parameter in terms of $\bar{X}_n$. You may assume that n is large enough that central limit approximation is valid.

Problem 24. In `http://math.iisc.ernet.in/~manju/UGstatprob/newcomb_lightspeed.txt` you will see the data from Simon Newcomb's experiment on the time taken (in nanoseconds) by light to travel 7442 meters at sea level.

(1) Compute the sample mean and sample standard deviation.

(2) Assuming normal distribution for the data, compute a confidence interval for the time taken. What confidence interval does it give for the speed of light (in meters per second)?

[**Note:** You are being asked to assume that the measured times have a normal distribution. It is different from assuming that the measured speeds (i.e., reciprocals of times essentially) are normally distributed.]

Problem 25. A box contains N marbles of which m are red in colour and $N - m$ are blue. We are interested in estimating the proportion $p = m/N$ of red balls. A sample of size k is drawn from the box and the number of red balls in the sample if observed, call it X. Then, X/k is a reasonable estimate for p. What are its bias and m.s.e if

(1) the sampling is done with replacement?

(2) the sampling is done without replacement?

Before you do the calculations, can you guess in which case would the mean squared error be smaller?

Problem 26. In the file `http://math.iisc.ernet.in/~manju/UGstatprob/simulatednormaldata.txt` you will see data *simulated* on a computer from a normal distribution with unknown mean and variance. The problem is to test the hypotheses $H_0 : \mu = 2$ versus $H_1 : \mu \neq 0$.

(1) Use only the first n data points for $n = 20, 50, 100, 200, 400, 500, 800, 1000$, and carry out the test for each n at significance level 0.05. Report the p-values.

(2) Repeat the same tests but now assume that the variance is given to be 9.

Problem 27. Are real coins fair? Formulate this as a hypothesis testing problem and perform the test at 0.01 level of significance using the following data. Report the p-value.

(1) In an experiment reported in `http://www.stat.berkeley.edu/~aldous/Real-World/coin_tosses.html`, a real coin was tossed 20000 times. The number of heads observed was 10231.

(2) In another experiment reported on the same page, 10014 heads appeared in 20000 tosses. Repeat the test with this data.

Problem 28. In the file `http://math.iisc.ernet.in/~manju/UGstatprob/twomidtermgrades.txt` you see the scores obtained in two exams by a class of students in their first and second midterms in the UM201 course, respectively. Test the hypothesis that the overall performance is worse in the second mid-term than in the first.

The following problem may be omitted. It is a two sample test for Bernoulli (which we did not cover in class). But if interested, it is a problem where we have $X_1, \ldots, X_n$ i.i..d Ber(p_1) and $Y_1, \ldots, Y_m$ i.i.d. Ber(p_2) and we test $H_0 : \ p_1 = p_2$ versus $H_1 : \ p_1 < p_2$.

Problem 29. This gallup poll conducted in the USA has data on support for capital punishment for a person convicted of murder. In 2013, 60% of the 1028 people sampled favoured capital punishment. In a similar survey conducted in 2007, 1010 people were sample of which 69% favoured

capital punishment. Based on just these two surveys, would you agree that support for capital punishment in that country has gone down?

Set up the question as an appropriate hypothesis testing problem, carry out the test at $\alpha = 0.01$ level of significance, and report the p-value.