CA660 Statistical Data Analysis (2013_2014)

M.Sc. (DA Major) - backgrounds various

Exercises 2 : Probability Distributions and Applications

includes conditionals, Decision-making
+ Classical Inference: Sampling Distributions,
Estimation and Hypothesis Testing on One and Two samples

Note: For estimation questions, do not worry about the f.p.c. unless explicitly asked for.
The questions are mainly straightforward and designed to reinforce points made in the lectures.

1. The Table shows the distribution of *X*=No. Bacteria/square obtained for counts of the root nodule bacteria (RT) in a Petroff-Hauser counting chamber. Complete the table for Expected number, assuming a Poisson. Comment on the observed distributional form.

No. Bacteria /Square (*X*)	Number of Squares Observed	Expected
0	34	?
1	68	?
2	112	?
3	94	?
4	55	?
5	21	?
6	12	?
7+	4	?
	400	

2 [Optional – for interest]Assume that the average number of crossovers is *m* in a genome segment, flanked by loci *A* and *B* with crossovers treated as Poisson events. Given that recombinant classes are observed only when an odd number of crossovers occur in the interval, obtain the expected recombination fraction, (defined as the probability of recombinant genotypes in the progeny) in terms of the expected number of crossovers (map distance).

For ℓ loci on a genome segment, and with r_ℓ, r_i , the recombinant fractions between two genes or genetic markers flanking the whole segment and between two markers flanking a sub-segment respectively, obtain the form of Haldane's mapping function, by arguing from the simple cases. [Hint: think about what each of the Poisson probabilities for x=0,1,2 etc. represents and how these sum).

3. Suppose length of service in a large firm distributes Normally, mean = 11.5 and variance =9. For a staff member selected at random, what is the probability he/she has served
   (i) between 8.5 and 14.5 years
   (ii) over 10 years
   (iii) under 12 years

4. Find the distribution function (or cumulative distribution function) *F(x),* median and mode for the following p.d.f.'s. [Note: range of *x* for part iii]

i. $f(x) = \dfrac{1}{2\sqrt{x}}, \quad 0 \le x \le 1$

ii. $f(x) = \dfrac{1}{4}\left(\dfrac{3}{4}\right)^{x-1}, \quad x = 1,2,...$

iii. $f(x) = 1 - |1 - x|, \quad 0 \le x \le 2$

iv. $f(x) = 6x(1-x), \quad 0 \le x \le 1$

5. Given the *prior distribution* for the proportion ***p*** of people with a given product preference is:

p	0.1	0.2
f(p)	0.6	0.4

Find the Bayes estimate for the proportion of people with that preference, if a random sample of size 2 gives just 1 person indicating that he/she has it.

[Hint: Clearly X = No. with preference ~ Binomial for given p and basic probability rules mean that $f(x,p)=f(x/p)\, f(p)$]

6. Consider the example, given in class, on setting the price for marketing a new computer tablet. The decision to be made is what price to charge, given possible states of nature, which represent when the competition may catch up/launch a similar product. If a decision-maker had *Perfect Information*, e.g. from a consultant, (s)he would always take *correct* action for the state of nature that applies.

Hence, under the maximum expected payoff criterion for decision-making,
Expected Value of Perfect Information = (Average Payoff using a *Perfect Predictor*) – (Average Payoff for whatever Action is actually selected)

Suppose the situation can be summarised:

States of nature	**Max. payoff (millions)**	**P{S_i} taken as**
S_1 < 6 months	250 (for A_1)	0.1
S_2 6-12 months	320 (for A_1)	0.5
S_3 12-18 months	410 (for A_4)	0.3
S_4 > 18 months	550 (for A_4)	0.1

(a) Obtain the expected payoff using a perfect predictor and hence the *Expected Value of Perfect Information*: (can be interpreted as the maximum amount a decision-maker is prepared to pay for making the correct decision.
(b) If *sample information* only available, what other input would you need?

7. If not using expected payoff as criterion for decision, but some other basis such as taking a gamble, could look at *risk* = v*ariance* of expected payoffs – as seen in class. Alternatively, could assign a *Utility value* to decision made, where this is, typically, something other than direct monetary value / profit. Can interpret as a way of combining given *attitude to risk* with each alternative profit or loss

Steps:

1. Assign utility values to smallest and largest payoff , U – range 0 to 100 convenient, so have U(Min) = 0, U(Max) =100 (*Relative* values important)
2. Utility Value for any payoff (**F**) to be considered = U(**F**) = P x 100.
   [P = what the probability would have to be of getting that payoff with *certainty* to be *equally attractive* to the decision maker as getting Max payoff (with prob 1-P)].

   Note: this probability relates to willingness to take a risk, not to Prob{S_i}

Can show *attitude to risk* on Utility vs Profit line. If above line = risk *avoider* and vice versa.

So, for a simple illustration, U values might be:

	P{S_1} = 0.7	P{S_2} = 0.3	
A_1 Gilt-edged	50	50	**A_1: Exp. Utility = (50)(0.7) +(50)(0.3) = 50**
A_2 Gold found	0	100	**A_2: Exp. Utility = (0)(0.7) +(100)(0.3) = 30**

Larger expected utility associated with gilt-edged (as risk avoider)

(a) Given the above and the table for the various payoffs for the example in Q 6, i.e. Min = 80, Max

= 550, assign U(80) =0, U(550) = 100. Hence, determine utility values for the other fourteen possible payoffs.

[Hint: easiest way to do this is to sketch the curve of U vs Profit (Payoff). Typically, pick values between min and max payoff and ask the question "for a payoff of e.g.. 200, what value of P would make getting that payoff with certainty equally as attractive as a payoff of 550, with probability P and a payoff of 80, with probability 1-P. If the decision maker says e.g. P=0.55, then clearly U(200) = 0.55x100. then, just plot with hypothetical values like this and read off actual A_i, S_i values from the curve]

(b) So, for the computer tablet example and the probabilities of the states of nature as given (see Q 6 above), obtain the expected utilities for *each* of the actions and determine the best action, based on *maximising expected utility*. Contrast with the result based on maximising expected payoff.

8. An oil company must decide whether to drill (a_1) or not to drill (a_2) in a particular place in the Celtic sea. The well may turn out to be dry (θ_1), wet (θ_2), or soaking (θ_3). On the basis of other drillings in the location, the company believes that the probabilities for these states are as follows:

$P(\theta_1) = 0.5$
$P(\theta_2) = 0.3$
$P(\theta_3) = 0.2$

The cost of drilling is M€70,000. If the well turns out to be wet the revenue will be M€120,000 and if it turns out to be soaking the revenue will be M€270,000. (There is no revenue for a dry well). Should the company drill or not?

(a) Obtain the solution, based on maximising expected value without further information.

(b) At a cost of M€10,000, the company could take seismic soundings that will help to determine the underlying geological structure at the site. The soundings will show whether the ground has no structure (O_1), an open structure (O_2) or a closed structure (O_3). In the past, oil wells which have been dry, wet or soaking have had the following conditional probabilities associated with seismic test outcomes:

Conditional probabilities – $P(O_k \mid \theta_j)$:

Geological Structure / **Well type**	O_1 *(None)*	O_2 *(Open)*	O_3 *(Closed)*
θ_1 *(Dry)*	0.6	0.3	0.1
θ_2 *(Wet)*	0.3	0.4	0.3
θ_3 *(Soaking)*	0.1	0.4	0.5

By consideration of the joint probabilities $P(O_k \cap \theta_j)$ and conditional probabilities $P(\theta_j \mid O_k)$ and using the maximum expected value criterion, would you advise the company to drill?

[Hint: you may find it helpful to draw the decision tree, showing the various decision points/ associated costs and probabilities for the problem and calculate the expected values at eth various decision points]

9 . (a) There are three values in a population, (100, 300 and 500). We are interested in investigating the sampling distribution of the mean ($\bar{x}$) for samples of size 2. How many possible distinct random samples are there? Assume sampling **with replacement**. Find the mean of each.

(b) Assuming each sample is equally likely to occur, give probabilities for results in (a), and hence give the *sampling distribution* of $\bar{x}$. On the basis of this, form a hypothesis about the mean of the population.

(c) If we change the sample size to 3 in (a), how many possible random samples are there now? What is the sampling distribution now?

(d) For the random samples generated in (a), what are the sample variances, s^2 ? Give the sampling distribution of s^2.

(e) For the sampling distributions in (b) and (c), find the probability that the mean is within 75, (either above or below), of 300. Which sample size gives more tightly clustered x values; (better precision in statistical terms)?

Note: samples here are, of course, very small – for purposes of illustration only.

10 (a) If X is Normally distributed with $\mu = 200$ and $\sigma = 50$, find $P\{X > 275\}$

(b) What proportion of the X values in (a) are within (i) 1.96 and (ii) 2.58 standard deviations of the mean?

(c) Use the following sampling distribution of ($\bar{x}$) to determine the proportion of $\bar{x}$ values within 37.5 units of 162.5, (include the boundaries of this interval).

$\bar{x}$	100	125	150	175	200	225
$P\{\bar{x}\}$	0.10	0.15	0.25	0.25	0.15	0.10

11. (a) Find the mean of the sampling distribution of the means in Question 9, part(b). Is this a *biased* or *unbiased estimate* of the population mean?

(b) Find the standard error (S.E.) of the sampling distribution of the means in the Question 9. How does this compare with the standard deviation of the population in that problem?

(c) As the sample size, n, increases do you expect the resulting $\bar{x}$ values to be closer to or further from the population mean μ ? In other words, what happens to the S.E.?

12. (a) A researcher selects a random sample of 400 substance weights from the population of weights on a long-running experiment. The population mean is 485 g. and its S.D. is 80 g. Find $P\{\bar{x} > 500\}$, i.e. prob. that a mean from the sampling distribution of the means is greater than 500.

(b) If you are told that the underlying observations (X = scores) are Normally distributed and a weight recorded is selected at random, what is the probability that it is greater than 500? (i.e. looking now for $P\{X > 500\}$.

(c) Comparing your answers to (a) and (b), comment.

13. (a) For Question 9, part (a), what difference would sampling *without replacement* make, i.e. from a *finite* population?

(b) What is the S.E. of the mean for this new sampling distribution?

(c) Compare the S.E.'s obtained here and in Question 11, part (b) and comment.

14 (a) A researcher needs estimates of the mean time required to complete certain processes in the lab. in order to ascertain hours of lab. time required. Randomly selecting 40 records on each of three processes, (s)he obtains a mean time of 12 hours for each. From past experience the S.D.'s for the three sections are taken to be 2hrs. , 4hrs and 6 hrs. respectively. Find a 90% confidence interval for the mean time for each section.

(b) Which of the samples in (a) is likely to provide the most accurate information about its population mean? Justify your answer.

(c) How would your confidence intervals be affected if you
(i) increased the sample size? (ii) wanted a higher level of confidence?

15. In 500 yields of a given experiment, percentage of "successes" (i.e. satisfactory outcome) given by

(a) 64% method A
(b) 83% method B
(c) 83% method A + extended heating period
(d) 51% method A + reduced heating period
(e) 49% method C

Find and interpret a 98% confidence interval for the proportion of successes for each estimate. Can we be 98% confident that over 50% of the population are successes for each of the above?

16. (a) A new piece of equipment is suspected of having faulty temperature control. The desired temperature is supposed to be 68 degrees and a random sample of 45 readings are collected. The

mean of these is 69.178 with a S.E. of 0.483. Formulate and test a suitable hypothesis.

(b) If a sample of only 20 readings had given the same mean and S.E., how would your calculation have been affected?

(c) If you were only concerned about temperatures being too low, how would your calculation be affected?

17. (a) A lab. manager claims that faulty readings on a piece of equipment are generated about 10% of the time, but a researcher believes it to be higher. A random sample of 400 records contain 60 errors. What evidence do you have at the 0.05 level of significance to support the lab. manager's statement or otherwise?

(b) A further random sample of the same size, revealed 45 errors. What would your conclusions have been on the basis of this sample?

(c) Is there a significant difference between the two results obtained?

18. To determine if a new process will improve productivity, two production lines are randomly selected, one to use the new, the other to maintain the current approach. Sample sizes are 36 working days for each line. The sample mean of work accomplished in a given period in the first sample is 17.83 tasks and the S.D. is 3.12 tasks. The sample mean in the second case is 20.1 tasks, with an S.D. of 5.47 tasks. Formulate and test a suitable hypothesis.

What would change if the samples had been smaller, say 15 ?

19. Eight randomly selected cultures are screened by Method A and also by Method B, which is supposedly better. Scores are assigned for both and are given below. Formulate and test a suitable hypothesis. Comment on any assumptions made.

Culture →		1	2	3	4	5	6	7	8
Scores	Method A	25	32	16	17	16	20	25	24
	Method B	30	35	30	17	22	19	29	29

Note: Samples here are small for a t-test. For purpose of illustration only.