CA660 Statistical Data Analysis (2013_2014)
M.Sc. (DA Major) - backgrounds various

Exercises 3 : Classical Inference: Many-sample tests +Alternative Estimation Approaches

Note: Calculations generally facilitated using R or MATLAB or SAS or SPSS or etc. for ANOVA type questions and similar. Suggest work through others and at least one ANOVA for practice.

1. In an upgrade on the power grid, the following is the distribution of the number of power failures reported in a city over 300 days. On the basis of these data, would it be reasonable to assume that the No. of positive counts is a Poisson random variable, with mean (λ) = 3.2?

   (a) Test at the 0.05 level of significance
   (b) If you are told that test requirements are usually that each expected frequency should be no less than 5 for chi-squared to be viable, do you need to adjust the distribution table here – i.e. collapse it at all?

No. Power Failures /day	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
No. of days	9	43	64	62	42	36	22	14	6	2

2. In order to assess measurement accuracy on new laboratory equipment, 100 Measurements, made on a standard item, were treated as a random sample and the following distribution was obtained. The view was that sensitivity was poor and measurements within the given range were 'equally likely': (check tables for the 'Uniform' distribution). Would you support the view stated, given the evidence available?

   (a) at the 0.01 level of significance?
   (b) at the 0.90 level of significance?
   (c) What can you say about the nature of statistical testing from your results on parts (a) and (b)?

Weight in g.	15.5-15.6	15.6-15.7	15.7-15.8	15.8-15.9	15.9-16.0	16.0-16.1	16.1-16.2	16.2-16.3	16.3-16.4	16.4-16.5
No. in range	5	9	7	10	12	8	13	15	13	8

3. Two production lines were quality-inspected and rated for performance over a random sample of records on 1000 tasks. Ratings, obtained, were found to be as follows.

		Rating		
Line	**Poor**	**Average**	**Good**	**Total**
A	98	256	188	542
B	43	202	213	458
Total	141	458	401	1000

   (a) Formulate and test a suitable hypothesis at the 0.025 level of Significance
   (b) Given that, for the test to be valid, ideally none but certainly no

more than 20% of expected frequencies must be less than 5, is this condition a problem here?

(c) If expected frequencies were lower than 5 in column 1, row 2 and column 2, row 1, how might you *sensibly collapse* the table and still test the principal hypothesis of interest? How would your test be affected? You can illustrate with the frequencies given, if helpful.

4. Market researchers question a sample of consumers to determine if four brands of headache remedy provide different levels of relief for patients with chronic headaches. Key question answers are weighted by lifestyle factors, to give an average score. Twelve subjects are randomly selected for each of the four factor levels and controlled also for use of blood pressure medication. The data are as given.

BP Med. Usage	**Headache Tablet Brand**			
	1	**2**	**3**	**4**
	8.68	6.23	4.92	7.43
None	8.14	6.73	5.21	6.76
	11.57	4.31	6.34	4.63
	8.98	5.80	5.00	5.87
In last 30 days	8.62	8.25	6.35	7.38
	3.35	7.88	8.41	5.58
	4.95	5.65	6.42	7.37
	5.16	6.47	4.70	7.28
Currently	7.89	3.78	7.72	5.46
	5.62	1.27	5.16	6.44
	7.08	0.83	2.60	6.02
	6.91	5.85	5.82	8.64

Use suitable software to perform a basic analysis of these data. Formulate, test and report on appropriate hypotheses and results.

5. To assess whether students' exam. performances vary with the advance description of an exam., a lecturer uses different descriptors in different classes and no descriptor for one class. He also feels that interpretation of descriptors may be gender-related. Students take a common exam. simultaneously. The data on exam. marks are as follows.

Gender	**Descriptor Used**		
	Easy	**Difficult**	**None**
Female	90	92	90
	85	76	85
	87	98	95
	85	65	75
Male	70	78	76
	72	70	70
	78	55	60

	65	50	62
	58	40	68

(a) Initially ignoring the gender effect, use SAS, R, SPSS or other to analyse these data. Formulate and test suitable hypotheses. What assumptions would you expect to apply? Indicate how you would use the information in the ANOVA table to test difference between performance means for Descriptors Difficult vs None.
(b) Omit the last row of the table (for the sake of illustration only) and analyse these data as a two-factor ANOVA, again formulating and testing suitable hypotheses.
(c) Indicate how you would analyse the complete data, using a regression (linear model) type approach. What are the independent variables here?

6. Yields of a particular substance are dependent on four properties of the original material, labelled X_1, X_2, X_3, X_4 here for convenience. It is required to use these data to provide a model equation for predicting yield (Y). Assume the observations Y_i to be independent and Normally distributed with constant variance. Perform the analysis, using suitable software and report briefly on your results.

X_1	X_2	X_3	X_4	Y
38.4	6.1	220	235	6.9
40.3	4.8	231	307	14.4
40.0	6.1	217	212	7.4
31.8	0.2	316	365	8.5
40.8	3.5	210	218	8.0
41.3	1.8	267	235	2.8
38.1	1.2	274	285	5.0
50.8	8.6	190	205	12.2
32.2	5.2	236	267	10.0
38.4	6.1	220	300	15.2
40.3	4.8	231	367	26.8
32.2	2.4	284	351	14.0
31.8	0.2	316	379	14.7
41.3	1.8	267	275	6.4
38.1	1.2	274	365	17.6
50.8	8.6	190	275	22.3
32.2	5.2	236	360	24.8
38.4	6.1	220	365	26.0
40.3	4.8	231	395	34.9
40.0	6.1	217	272	18.2
32.2	2.4	284	424	23.2
31.8	0.2	316	428	18.0
40.8	3.5	210	273	13.1
41.3	1.8	267	358	16.1
38.1	1.2	274	444	32.1
50.8	8.6	190	345	34.7
32.2	5.2	236	402	31.7
38.4	6.1	220	410	33.6
40.0	6.1	217	340	30.4
40.8	3.5	210	347	26.6
41.3	1.8	267	416	27.8
50.8	8.6	190	407	45.7

7. Three independent observations on a Poisson distribution with an unknown mean μ are 6,9,11. What is the log-likelihood function? Hence, give the MLE of μ.

8. [Optional]
Independent observations $x_1, x_2, ...x_n$ are taken from a Normal distribution with unknown mean μ and unknown variance σ^2 .Obtain the MLEs

9. Suppose we have paired observations (x_1, y_1), (x_2, y_2),.....(x_n, y_n) from the model

$$E\{Y\} = \hat{y} = \beta_0 + \beta_1 x$$

$$V(Y) = \sigma^2$$

where the Y's are independent and Normally distributed. Write down the log-likelihood and the form of the MLEs.

10. [Optional]
If we observe r successes in n trials and define θ as the probability of a success in any trial, then the likelihood and log-likelihood are, respectively

$$\ell(\theta) = k\theta^r (1-\theta)^{n-r}$$

$$L(\theta) = c + r\log\theta + (n-r)\log(1-\theta)$$

so that MLE given by:

$$\frac{dL}{d\theta} = \frac{r}{\theta} - \frac{(n-r)}{(1-\theta)}$$

$$\hat{\theta} = \frac{r}{n}$$

Obtain the minimum variance. State how this is related to the Information content.

11. Weights of product from randomly-selected cartons were given in mg. as

(Group 1)	0.6	3.6	7.1	8.1
(Group 2)	21.9	24.5	25.8	34.2

Test the null hypothesis that the medians are equal, using Wilcoxon-Mann-Whitney. (You might follow up by thinking about extending the example and how you could apply Kolmogorov-Smirnov to data from two groups).

12. Data, on serum-glucose values of mice, were recorded, where the design dealt with mice in 6 randomized blocks, each block containing one representative for each treatment. Results were as shown

	Saline on day 14		**Adrenaline on day 14**		
Block No.	**Uninfected (A)**	**Pertussis(B)**	**Uninfected (A)**	**Pertussis(B)**	**Block Total (Q)**
I	221	94	330	163	808
II	200	109	302	157	768
III	233	146	283	177	839
IV	180	141	273	139	733
V	198	124	307	148	777
VI	213	114	279	144	750
Treatment Total	**1245**	**728**	**1774**	**928**	**4675**

Apply the Friedman Test to these data, stating the null hypothesis clearly and reporting on your results. (Think further about contrasting use of Friedman here with the usual form of the Analysis of Variance).

13. Scores on a series of competency tests were aggregated for workers in three departments. The following results were reported for the three different environments. Perform a Kruskal-Wallis test, stating the null hypothesis clearly and reporting your results. What check would you perform for this, (and other), rank tests?

Section A Scores: 93 98 216 249 301 319 731 910

Section B Scores: 29 39 60 78 82 112 125 170 192 224 263 275 276 286 369 756

Section C Scores: 126 142 156 228 245 246 370 419 433 454 478 503

14. Concentration levels, of a particular substance, were measured in a contaminants' check at 16 locations, with results as given.
Use the Sign test to determine whether the null hypothesis that mean concentration = 0.65 (as opposed to *greater than* 0.65) is supported. What *property* is being tested here?

0.60 0.66 0.67 0.59 0.72 0.61 0.64 0.57 0.71 0.69 0.65 0.78 0.74 0.64 0.75 0.77

Would it have been more useful to use e.g. Wilcoxon Signed Rank?