

Exercises Sheet 7

Revision MLE and Likelihood Ratio, Non-Parametric

1. 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 .
2. Independent observations x_1, x_2, \dots, x_n are taken from a Normal distribution with unknown mean μ and unknown variance σ^2 . Obtain the MLEs
3. Suppose we have paired observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 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.

4. 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.

5. The Backcross data (class example Week 3, Week 8 etc.) can also be analysed using a generalized log-likelihood approach. An estimate of the parameter is needed to construct the tests in this case. The parameter in this case is the recombination fraction for each of the 4 crosses. If you are given that this generalized approach gives the log-likelihood function for multiple populations having different recombination fractions between two loci as:

$$L = \sum_{i=1}^c \sum_{j=1}^{n_{1i}} \sum_{k=1}^{n_{2i}} f_{0ijk} \log p_{ijk}$$

where subscript i denotes cross and c the number of crosses, j and k denote genotypic classes for locus A and B respectively, f_{ijk} and p_{ijk} are the observed counts and expected frequencies of the genotypic class and n_{1i} and n_{2i} are the number of genotypic categories for locus A and locus B respectively. The Total G-statistic (L-LRTS) for linkage is then the sum over the four crosses and the G-statistics for Pooled and Heterogeneity are defined in the usual way. The recombination fraction estimates for crosses 1 to 4 are given, respectively, to be:

$$\hat{\theta} = 0.479, 0.383, 0.383, 0.392$$

Obtain the log-likelihoods for computing the L-LRTS for each of the four crosses (Note, under H_0 , $\theta = 0.5$).

Hence obtain the G-statistics for linkage for each of the crosses, as well as the Total, Pooled and Heterogeneity G-statistics. Comment briefly on your results.

6. Expected genotypic frequencies for a Backcross ($AaBb \times aabb$) progeny are shown below, where θ is the recombination fraction between A and B, f_{ij} is the observed genotypic count for the i th genotype of locus A and the j th genotype of B.

Give the form of the log-likelihood function, the MLE of θ , (where you can assume N individuals in the sample), the average Information content for an individual and hence the variance of the estimated recombination fraction for the sample size given.

Genotype	Observed Count (f_{ij})	Expected Frequency (p_{ij})
$AaBb$	f_{11}	$0.5(1-\theta)$
$Aabb$	f_{12}	0.5θ
$AaBb$	f_{21}	0.5θ
$aabb$	f_{22}	$0.5(1-\theta)$

7. Weights of water from a pipetting experiment were given in mg. As

Drainage (group 1)	0.6	3.6	7.1	8.1
Blow-out (group 2)	21.9	24.5	25.8	34.2

Test the null hypothesis that the medians are equal, using Wilcoxon-Mann-Whitney.

(This example comes from Wardlaw and it is suggested that you follow it up by reading the similar extended example on how you might apply Kolmogorov-Smirnov to data from two groups).

8. 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

Block No.	Saline on day 14		Adrenaline day 14		Block Total(Q)
	Uninfected (A)	Pertussis(B)	Uninfected(C)	Pertussis(D)	
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.

(Again, this example is taken from Wardlaw and it is of interest to contrast the use of Friedman here with the usual form of the Analysis of Variance).

9. Scores on a series of tests were aggregated for researchers in each of three Labs. 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 previous, rank tests?

Lab. A Scores: 93 98 216 249 301 319 731 910

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

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

10. The blood-concentration level, of a particular substance, was measured in 16 laboratory rats, 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).

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

Addendum: Assignment 2

This addition has been included as some people seem to be having trouble starting Assignment 2, though no information was given as to what has actually been tried so far. It is a little difficult to offer further clues, given this lack of information, but to *emphasise* some of those that you have been given in terms of a possible approach.

1. Classification of the Marker data is essentially by codes 1 and 2 for two different genotypes.
2. In the background information, you should find that QTL methodology for the double haploid progeny is the same as that for Backcross; (interpretation of genetic effects will differ).
3. The information from a classical backcross can be used to explain the rationale for a single marker analysis, in terms of co-segregation patterns for marker genotypes AA, Aa.
3. Information also includes the fact that single-marker analysis is based on comparisons between marker genotypic means through a t-test, ANOVA etc. Single-marker analysis is done by analysing one marker at a time.
4. The test between marker genotypic classes, μ_{AA} and μ_{Aa} also gives mean differences between Steptoe and Morex.
5. ANOVA for a single QTL analysis using Backcross is similar to that given in class, assume N progeny size, b number of replications, and sample sizes n_1 , n_2 for two classes of genotype - can define a constant $c = N - (n_1^2 + n_2^2)/N$ to keep the look of the ANOVA simpler - you have the terms from the model form given. The dof are the same for the ANOVA for single QTL analysis and typical single marker analysis. The issue then is what the Expected mean squares are measuring?
6. Similarly, for the regression, expected QTL genotypic frequencies are given for the co-segregation patterns, so the expected means, variances and covariance needed for estimating the regression coefficient for the two variables are, from the information given:

$$E(\bar{x}) = 0.5 \times 1 + 0.5 \times (-1) = 0$$

$$E(\hat{s}_x^2) = 0.5 \times 1^2 + 0.5 \times (-1)^2 = 1$$

for x mean and variance and similarly for y , using information from the same table. It should not be difficult to see that the expectation of the slope is the expectation for the difference between the two marker classes.