QBS 120 - Problem Set 9

Rob Frost

1. (Based on Rice 12.1) Repeat problem 12.1 but simulate 1000 separate data sets with each data set containing 7 batches of 10 values.

   (a) For each of the simulated data sets, test the H_0 of no difference in the mean for the seven levels using an F test and compute the associated p-value. Hint: to get the p-value from an R aov() analysis, you can use: summary(aov(...))[[1]][["Pr(>F)"]][1]

   (b) Generate a probability plot for these p-values relative to U(0,1).

   (c) Does the distribution match your expectations? Explain.

   (d) For how many simulations would the H_0 be rejected at $\alpha = 0.05$? Does this match your expectations? Explain.

   (e) After controlling for FWER using the Bonferroni method, how many simulations are significant? Does this match your expectations?

   (f) After controlling for the FDR using the Benjamini & Hochberg method, how many simulations have FDR values ≤ 0.05. Does this match your expectations?

2. Answer the following questions using the pancreatic cancer gene expression example from the last problem set. For these questions, assume that statistical tests are performed for all measured genes . Answer a) - c) for both Experiments A and B. Answer d) - h) for just Experiment A.

   (a) Which approach to MHC (FWER or FDR) do you think will provide the best balance of type I error control and power? Justify your answer.

   (b) For FWER control, is Bonferroni, Holm, Hochberg, Hommel or Westfall & Young preferrable? Justify your answer.

   (c) For FDR control, is Benjamini & Hochberg or Benjamini & Yekutieli preferrable? Justify your answer.

   (d) Simulate data for Experiment A and test each gene for DE according to the H_0 and H_A specified above. Assume that $n = 50$, $\mu_{M0} = 0.5$ and $\mu_{M1} = 1.75$

   (e) Plot the distribution of raw p-values. Is the distribution of p-values consistent with the global H_0? Explain.

   (f) How many significant DE genes are there at a FWER of 0.05 using Bonferroni, Holm and Hochberg? Do these results match your expectations? What is your empirical average power in this case?

   (g) How many significant DE genes are there at an FDR of 0.05 using BH and BY methods? Do these results match your expectations? What is your empirical average power in this case?

(h) Proposed a statistically valid p-value weight that could be used with a weighted FDR (wFDR) analysis to improve statistical power. Perform wFDR analysis using this weight. Was power improved relative to unweighted FDR?

3. (Based on Rice 13.1) Adult-onset diabetes is known to be highly genetically determined. A study was done comparing frequencies of a particular allele in a sample of such diabetics and a sample of nondiabetics. The data are shown in the following table:

	Diabetic	Normal
Bb or bb	12	4
BB	39	49

(a) Are the frequencies of the alleles significantly different in the two groups? Test using both the Fisher Exact test and a chi-squared test (you can use R functions). Comment on the difference in p-values between the two methods.

(b) Explain in your own words how hypothesis testing is performed using each approach.

4. (Based on Rice 13.21) Do the following for problem 13.1:

(a) Estimate the odds ratio (OR) using the unconditional MLE (what is described in Rice) without using special-purpose R packages.

(b) Estimate the OR using the four different estimation methods implemented by the oddsratio() method in R epitools package. Confirm that your estimate from part a) matches the output for the "wald" method.

(c) Use the parametric bootstrap to estimate the sampling distribution of the OR following the approach outlined in Rice Section 13.6. Plot a kernel density estimate of this sampling distribution.

(d) Calculate the 95% CI using the percentile method. Compare your results to the OR and CI estimates from the R fisher.test() and oddsratio() methods. Do they differ? If so, why?