Quantitative Genomics and Genetics - Spring 2021
BTRY 4830/6830; PBSB 5201.01

Homework 4 (version 1)

Assigned March 26; Due 11:59PM April 6

Problem 1 (Easy)

a. If you reject a null hypothesis, explain why this does not mean that your null hypothesis is definitely wrong. Also explain why if you cannot reject a null hypothesis, you cannot interpret this result as evidence that the null hypothesis is correct.

b. Explain why we can precisely set the Type I error of a test but we cannot precisely control power.

Problem 2 (Medium)

Note that for this question, your answer should include R code that generates the appropriate answers. Use Rmarkdown and submit your .Rmd script (HTML optional) and note there will be penalties for scripts that fail to compile. Please note that you do not need to repeat code for each part (i.e., if you write a single block of code that generates the answers for some or all of the parts, that is fine, but do please label your output that answers each question!!).

Consider a 'height' type experiment where we can assume $X \sim N(\mu, \sigma^2)$ where the true $\mu = 0.5$ and $\sigma^2 = 1$ is known to you in this case (obviously in real cases, this will never be known). Assume that you will be generating samples of size n and that for each sample, the test statistic you will be calculating will be the mean of the sample. For parts [a, c, e, g] you are being asked to provide answers based on the distribution of the test statistic (the mean) for the null hypothesis or the true distribution for the experiment, where the R functions 'qnorm()' and 'pnorm()' can be used in the calculation of the answers. For parts [b, d, f, h] you are being asked to simulate $k = 1000$ samples of size n (i.e., 1000 separate samples) under the null hypothesis or the true distribution for the experiment and for each of the k samples, you are asked to calculate the mean statistic, where the function 'rnorm()' can be used to generate the samples. For parts [i, j], you are bring asked to calculate the LRT statistic for each of the 1000 samples of size n (i.e., a different statistic!) where each of the k samples is generated under the null hypothesis or the true distribution.

a. Consider $H_0 : \mu = 0$ where you know the true $\sigma^2 = 1$ for the purposes of this question (remember, not realistic!). Consider a sample size $n = 20$ and the mean of the sample as the

test statistic. What is the critical threshold for this statistic for a Type I error = $\alpha = 0.05$ when considering $H_A > 0$? What are the two values of the critical threshold for this statistic for a Type I error = $\alpha = 0.05$ when considering $H_A \neq 0$? HINT: use the function 'qnorm()'.

b. Simulate $k = 1000$ samples under the null hypothesis in part [a], calculate the mean for each sample, and plot a histogram of the means. NOTE: you do not need to output the samples or the means (!!), just the histogram (and provide your code). btw, no answer required, but take a look at how many of your sample statistics are beyond the critical thresholds you calculated in part [a]...

c. Consider the true distribution of the experiment $\mu = 0.5, \sigma^2 = 1$ for the purposes of this question (again, not realistic!). Consider a sample size $n = 20$ and the mean of the sample as the test statistic. What is the power of this test for the null hypothesis in part [a] with a Type I error = $\alpha = 0.05$ when considering $H_A > 0$? What is the power of this test for the null hypothesis in part [a] with a Type I error = $\alpha = 0.05$ when considering $H_A \neq 0$? HINT: use the function 'pnorm()' and your critical thresholds from part [a].

d. Simulate $k = 1000$ samples under the true distribution (as used in part [c]), calculate the mean for each sample, and plot a histogram of the means. NOTE: you do not need to output the samples or the means (!!), just the histogram (and provide your code). btw, no answer required, but take a look at how many of your sample statistics are beyond the critical thresholds you calculated in part [a] and see how this compares to the power you calculated in part [c]...

e. Repeat part [a] but consider $n = 100$.

f. Repeat part [b] but consider $n = 100$.

g. Repeat part [c] but consider $n = 100$.

h. Repeat part [d] but consider $n = 100$.

i. For the samples you generated in parts [b] and [f] calculate the $LRT = -2ln\frac{L(\hat{\mu}_0|\mathbf{x})}{L(\hat{\mu}_1|\mathbf{x})}$ statistic for $\hat{\mu}_1 \in (-\infty, \infty)$ and plot a histogram of these statistics. NOTE: you do not need to output the samples or each LRT (!!), just the two histograms (and provide your code). btw, no answer required, but see if you can figure out what distribution these histograms look like and consider whether these are different (and see if you can figure out why?).

j. For the samples you generated in parts [d] and [h] calculate the $LRT = -2ln\frac{L(\hat{\mu}_0|\mathbf{x})}{L(\hat{\mu}_1|\mathbf{x})}$ statistic for $\hat{\mu}_1 \in (-\infty, \infty)$ and plot a histogram of these statistics. NOTE: you do not need to output the samples or each LRT (!!), just the two histograms (and provide your code). btw, no answer required, but notice whether these are different than the distributions you calculated in part [i] and whether they differ from each other (and see if you figure out why?).

Problem 3 (Difficult)

Prove that for an experiment, producing a sample for which we are conducting a hypothesis test using the statistic $T(\mathbf{x})$, if we were to independently conduct this same experiment and produce the same number of experimental trials in each of an infinite set of equivalent, alternative universes (i.e.,

producing a sample for each) and for each sample we were to calculate the statistic and p-value, the resulting set of p-values would have a uniform distribution if the null hypothesis was correct, i.e., $Pr(pval(T(\mathbf{X}))) \sim unif[0,1]|H_0$ is true.