BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2014

Homework 3 (version 1) - Key

Assigned September 24; Due 11:59PM September 29

Problem 1 (Easy)

a. Using two sentences at most, explain why we can control the type I error but why we cannot directly control the type II error.

Type I error is set by the investigator as a function of the sampling distribution of the statistic under the null hypothesis, which is known given the null assumption. Type II error is a function of the Type I error and the sampling distribution of the statistic under the true parameter value, and since the later are not known, it cannot be set directly.

b. Note that a p-value is constructed in such a way that we reject the null hypothesis if our observed statistic is in an area under our 'statistic sampling distribution assuming the null is true' that has relatively low probability, i.e. the tails of the distribution. Using two sentences at most, provide an intuitive explanation as to why it makes sense to construct a p-value this way.

While we could, in theory, choose it reject the null for any pre-selected region of the sampling distribution of the statistic under the null hypothesis corresponding to the area of the pre-selected Type I error, in practice, we are interested in rejecting the null hypothesis in cases where the true parameter value is quite different from the null. By defining the p-value in the tails, we are increasing the chances of having a low p-value for cases when the true parameter value is far from the parameter value under the null hypotehsis.

Problem 2 (Medium)

See separate .html key for this problem.

Recall that the likelihood of an i.i.d. sample $\mathbf{x} = [x_1, x_2, ..., x_n]$ where each X_i is well-approximated by a normal probability model has the following form:

$$L(\mu, \sigma^2|\mathbf{x}) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (1)$$

and that the maximum likelihood estimators of the true values of μ and σ^2 are:

$$MLE(\hat{\mu}) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

$$MLE(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - MLE(\hat{\mu}))^2 \tag{3}$$

Finally note that in R you can generate samples that are well-approximated by a normal probability model using the function:

rnorm(n, μ, $\sqrt{\sigma^2}$)

where n is the size of the sample you want to generate and μ, σ^2 are the true parameter values of the probability model.

FOR THIS PROBLEM (!!) provide a separate text file with your R code used to generate your answers (where appropriate)!

a. A transformed version of the likelihood ratio test statistic for this case is the following:

$$-2ln(\Lambda) = -2ln\left(\prod_{i=1}^{n} \frac{\frac{1}{\sqrt{2\pi MLE(\hat{\sigma}^2)}} e^{\frac{-(x_i - H_0(\mu))^2}{2*MLE(\hat{\sigma}^2)}}}{\frac{1}{\sqrt{2\pi MLE(\hat{\sigma}^2)}} e^{\frac{-(x_i - MLE(\hat{\mu}))^2}{2*MLE(\hat{\sigma}^2)}}}\right) \tag{4}$$

Use the following rules (1. cancel fraction = 1, 2. $\prod e^{a_i} = e^{\sum a_i}$, 3. $ln(\frac{a}{b}) = lna - lnb$, 4. $ln(e^x) = x$, 5. multiply through by -2 and move constant outside the sum) to show that this equation can be written as follows (i.e. show each step and list the rule you used!):

$$-2ln(\Lambda) = \frac{1}{MLE(\hat{\sigma}^2)}\sum_{i=1}^{n}(x_i - H_0(\mu))^2 - \frac{1}{MLE(\hat{\sigma}^2)}\sum_{i=1}^{n}(x_i - MLE(\hat{\mu}))^2 \tag{5}$$

b. We are going to test the null hypothesis $H_0 : \mu = 0$. Assume that (unknown to you!!) that the null hypothesis is TRUE and that the true $\sigma^2 = 2$. Using R, simulate 100 samples of size $n = 100$ (i.e. assume each individual sample is the result of an independent experiment performed in a duplicate universe!), for each calculate $LRT = -2ln\Lambda$, and plot a histogram of the values you obtain using the R code 'hist(input, probability = TRUE)'.

c. Use the R code: 'curve(dchisq(x, 1), from=0, to = 20, add=TRUE)' to plot the pdf of a χ^2-distribution with 1 degree of freedom (d.f. = 1) over the histogram you produced in b.

d. Still assuming (still unknown to you!) that the null is TRUE, produce a single sample of $n = 100$, calculate the $LRT = -2ln\Lambda$ for this statistic, and use the R code: '1 - pchisq(LRT, 1)' to calculate the p-value. Did you reject the null at a type I error of $\alpha = 0.05$? Why did you expect / not expect your particular result (one sentence)?

e. Repeat question b AND c, but assume each sample is of size $n = 10,000$. Using one sentence at most, describe the difference between the histogram obtained in b, c and this new histogram?

f. Next, assume (AGAIN, unknown to you!) that the null is FALSE and that the true $\mu = 0.5$ and true $\sigma^2 = 2$. Using R, simulate 100 samples of size $n = 100$ (i.e. again, assume each individual sample is the result of an independent experiment performed in a duplicate universe), for each calculate $LRT = -2ln\Lambda$, and plot a histogram of the values you obtain AND plot a pdf of a χ^2-distribution (d.f. = 1) over this histogram.

g. Is your histogram similar to the pdf of a χ^2-distribution (d.f. = 1)? Why did you expect this would be the case?

h. Still assuming that (still unknown to you!) that the null is FALSE and that the true $\mu = 0.5$ and true $\sigma^2 = 2$. Produce a single sample of $n = 100$, calculate the $LRT = -2ln\Lambda$ for this statistic, and test the null hypothesis $H_0 : \mu = 0$ at $\alpha = 0.05$ using the R code: '1 - pchisq(LRT, 1)' to calculate the p-value. Did you reject the null? Why did you expect / not expect your particular result (one sentence)?

i. Finally, assume (you get the drill...) that the null is FALSE and that the true $\mu = 1$ and true $\sigma^2 = 2$. Using R, simulate 100 samples of size $n = 100$, for each calculate $LRT = -2ln\Lambda$, and plot a histogram of the values you obtain AND plot a pdf of a χ^2-distribution (d.f. = 1) over this histogram.

j. Explain why it makes sense how the histogram in i. is shifted compared to the distribution in f.? Explain why it therefore follows that the test in i. is more powerful than the test in f?

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 an infinite number of times in equivalent, alternative universes producing a sample for each, and we were to conduct the same hypothesis test on each sample, the resulting p-values would have a uniform distribution $Pr(pval(T(\mathbf{x}))) \sim unif[0, 1]$ if the null hypothesis was correct.

ANS: Define $Z = F_X(x)$, such that Z is the cdf of a random variable X. Note that by the definition of p-values and cdf's, Z can be thought of as a random variable that takes the value of a p-value (or more specifically 1 - p-value). We are now interested in the $Pr(Z)$, which we may equivalently phrase as determining $F_Z(z)$, the cdf of Z. Note that we can perform the following operations:

$$F_Z(z) = Pr(Z \leqslant z) = Pr(F_X(x) \leqslant z) = Pr(x \leqslant F_X^{-1}(z)) = F_X(F_X^{-1}(z)) = z \tag{6}$$

where the first equality is simply the definition of a cdf, the second equality follows after substitution, the third equality follows by applying the inverse of the cdf to both terms on either side of the greater than or equal to operator, the fourth equality follows from the definition of a cdf, and the fifth follows after applying a function to an inverse of the function, i.e. the input is returned. Since the $F_Z(z) = z$, this means the cdf for a particular value of z is a constant. This is the definition of a uniform distribution, i.e. the probability of any equivalently sized interval is the same. Since a Z is a random variable that can only take the values of a p-value, it only has non-zero probability between zero and one, thus $F_Z(z) \sim unif[0, 1]$ and the statement is proved.

AS AN INTUITIVE NOTE (!!): To see why it is critical that our p-values be uniform on [0,1] under H_0, consider the definition of a p-value: 'the statistic would take this value or more extreme with a probability of x' and next consider a case where this does not hold. For example, assume that we had a probability greater than 0.05 of producing a p-value with a value of 0.05 or less. Such a distribution would contradict the definition of a p-value and we would similarly have a contradiction of this definition for a definable region whenever the distribution is not uniform.