HW 3 - due 04/28 at 11:59 pm.

Math 181B, Spring 23, Rava

Follow closely the 'Hw guide' under Files in the folder 'Course Contents' on how to write, scan and submit your homework.
On any problem involving R, you should include your code and output as part of your answer. You may take a screenshot of the code/output, or write it by hand.
Be careful with notation, remember to define the parameters and the random variables you intend to use.

1 Exercise 1

[5 points] Let $(X_1, \ldots, X_t) \sim Multinom(n, p_1, \ldots, p_t)$. Prove that $(X_1, X_2, Y) \sim Multinom(n, p_1, p_2, 1 - p_1 - p_2)$, where $Y = n - X_1 - X_2 = X_3 + \ldots + X_t$.

2 Exercise 2

[6 points] You want to model the number of questions students ask in an hour of lecture. Based on other research, you learn this is well modeled by a Poisson distribution. You are curious to know if this is actually the case. You collect data on 200 random hours of lecture. The data are recorded in the dataset 'Question.csv', available on Canvas. Conduct an HT with $\alpha = 0.03$. When constructing the table of observed and expected counts, consider as the possible number of questions asked in an hour 0,1,2,3,4,5,6,7+ where with 7+ we mean seven or more questions. This binning guarantees that the assumption $E_i \geq 5$ for $i = 1, \ldots, 8$ is satisfied. Something along this line is always necessary when potentially the number of categories could be infinite. (Hint: you can use in R the function table to find the observed counts. Remember that you can always use ?table to ask R for help. You can use in R the functions dpois and ppois to compute the probabilities needed to compute the expected counts.)

3 Exercise 3

In class we studied the χ^2 goodness-of-fit test that is used to test assumptions such as $H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_t = p_{t0}$ vs $H_1 : p_i \neq p_{i0}$ for at least one i. One key assumption of the test is that all the expected counts under the null $E_i = np_{i0}$ are at least 5. If this assumption is not satisfied, the test statistic $D = \frac{\sum_{i=1}^{t}(X_i - np_{i0})^2}{np_{i0}}$ is not guaranteed to have approximately χ^2_{t-1} distribution under the null and therefore the test is not guaranteed to have probability of type I error equal to α. In this exercise, you will perform simulations in R to convince yourself that this is the case.

a) [4 points] In R sample $(X_1, X_2, X_3) \sim Multinom(50, 0.3, 0.5, 0.2)$. (Hint: you can use the function rmultinom(). Remember that you can always use ?rmultinom to ask R for help). Compute the value of

the test statistic $D = \frac{\sum_{i=1}^{t}(X_i - np_{i0})^2}{np_{i0}}$ to test $H_0 : p_1 = 0.3, p_2 = 0.5, p_3 = 0.2$. Repeat the process 1000 times and record in a vector the 1000 obtained D. (Advice: To do that create a function that samples the data and returns the realization of d. Use the function replicate to apply the function created 1000 times. Remember that you can always use ?replicate to ask R for help).
At this point, you should have a vector that records 1000 realizations of $D = \frac{\sum_{i=1}^{t}(X_i - np_{i0})^2}{np_{i0}}$, that according to the theory, has a χ^2_2 distribution. (Make sure that you know why the degrees of freedom are 2).
Plot a histogram of your 1000 realizations of D and superimpose to the histogram the density of the χ^2_2 distribution. (Hint: make sure to not plot a histogram of the frequencies. Use ?hist to understand how). Do your simulations agree with the theory?

b) [4 points] We are now going to use simulations to convince ourselves that indeed the χ^2 goodness-of-fit test described in class guarantees probability of type I error equal to α, as long as the expected counts under the null are at least 5. In R sample $(X_1, X_2, X_3) \sim Multinom(50, 0.3, 0.5, 0.2)$. Compute the value of the test statistic $D = \frac{\sum_{i=1}^{t}(X_i - np_{i0})^2}{np_{i0}}$ to test $H_0 : p_1 = 0.3, p_2 = 0.5, p_3 = 0.2$. Compute the $p - value = P(\chi^2_2 \geq d)$. You reject the null if $p - value < \alpha = 0.05$. Report 1/TRUE if you reject the null, 0/FALSE if you don't.
Repeat the process 1000 times and record in a vector the results of the 1000 tests. At this point, you should have a vector that records the results of 1000 tests. TRUE/1 means that you rejected the null, FALSE/0 means that you have failed to reject the null. According to the theory, the proportion of TRUE/1 should be around $\alpha = 0.05$. (Remember that in this case we are simulating data under the null).
Verify that this is the case and that indeed the procedure described in class guarantees probability of type I error equal to α.

c) [4 points] Repeat exercise a) and b) but now use $(X_1, X_2, X_3) \sim Multinom(20, 0.02, 0.01, 0.97)$. What do you notice?