## S&DS 242/542: Homework 3

Due Wednesday, February 5, at 1PM

1. Testing gender ratios (based on Rice 9.45). In a classical genetics study, Geissler (1889) studied hospital records in Saxony and compiled data on the gender ratio. The following table shows the number of male children in 6115 families having 12 children:

Number of male children	Number of families
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3

Let  $X_1, \ldots, X_{6115}$  denote the number of male children in these 6115 families. (Thus the table indicates that 7 values of  $X_1, \ldots, X_{6115}$  are equal to 0, that 45 values are equal to 1, etc.)

(a) Suggest two reasonable test statistics  $T_1$  and  $T_2$  for testing the null hypothesis

 $H_0: X_1, \dots, X_{6115} \stackrel{IID}{\sim} \text{Binomial}(12, 0.5).$ 

This is intentionally open-ended: try to pick  $T_1$  and  $T_2$  to detect different possible alternatives to the above null hypothesis. Compute the values of  $T_1$  and  $T_2$  on the above data.

(b) Perform a simulation of the null distributions of  $T_1$  and  $T_2$ , and plot histograms of the simulated null distributions. For either of your tests, can you reject  $H_0$  at the significance level  $\alpha = 0.05$ ? Include both your code and the histograms with your homework submission.

[You may perform 1000 simulations using a for loop as in Homework 2. In each simulation, sample  $X_1, \ldots, X_{6115} \sim \text{Binomial}(12, 0.5)$ , and compute  $T_1$  and  $T_2$  from these samples.

In addition to what was introduced in Homework 2, the following commands may be helpful in R:

To generate a numeric vector of 6115 independent Binomial(12, 0.5) samples:

X = rbinom(6115, 12, 0.5)

To count the number of elements of a vector X that are equal to, say, 8:

count = length(which(X==8))

To create a new numeric vector with fixed values, say, (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12):

k = c(0,1,2,3,4,5,6,7,8,9,10,11,12)

To evaluate the binomial coefficients  $\binom{12}{k}$  for each value of k above:

bincoef = choose(12,k)

If you would like a primer to other commands in R, Section 2.3 of your ISLR textbook gives a friendly introduction.]

[Not for submission, but just food for thought: It is interesting to think about why  $H_0$  may not hold even if, biologically, the probability of having a male child is exactly 50%. For example, what happens if families choose whether or not to have another child based on the numbers of male/female children they already have, and Geissler's study selected only those families with exactly 12 children?]

## 2. Distribution of the p-value.

(a) Consider any hypothesis test that rejects  $H_0$  for large values of a test statistic Z having null distribution  $\mathcal{N}(0, 1)$ . Recall that the associated p-value of this test is the right tail probability  $P = 1 - \Phi(Z)$  where  $\Phi(\cdot)$  is the standard normal CDF.

Show that if  $H_0$  is true, then  $P \sim \text{Uniform}([0, 1])$ . That is, viewing the p-value itself as the test statistic, its null distribution is Uniform([0, 1]). [Hint: Compute the CDF of P.]

(b) If instead  $H_1$  is true, would P tend to take larger or smaller values as compared with its distribution under  $H_0$ ? How would you perform a level- $\alpha$  test of  $H_0$  vs.  $H_1$  using P as the test statistic?

## 3. The $t_n$ distribution for large n.

(a) Let  $U_n \sim \frac{1}{n} \cdot \chi_n^2$ . Using the Law of Large Numbers and the Continuous Mapping Theorem, show that  $\sqrt{U_n} \to 1$  in probability as  $n \to \infty$ .

(b) A result called Slutsky's lemma states: If two sequences of random variables  $X_1, X_2, X_3, \ldots$ and  $Y_1, Y_2, Y_3, \ldots$  are such that  $X_n \to c$  in probability and  $Y_n \to Y$  in distribution, then  $X_n Y_n \to cY$  in distribution. Furthermore, if  $c \neq 0$ , then  $Y_n/X_n \to Y/c$  in distribution.

Suppose  $T_n \sim t_n$  (the t-distribution with *n* degrees-of-freedom). Using Slutsky's lemma and your result in part (a), show that  $T_n \to \mathcal{N}(0, 1)$  in distribution as  $n \to \infty$ .

## 4. The $t_1$ distribution.

(a) Let  $T \sim t_1$  (the t-distribution with 1 degree-of-freedom). Explain why T has the same distribution as  $\frac{X}{|Y|}$  where  $X, Y \stackrel{IID}{\sim} \mathcal{N}(0, 1)$ , and why T also has the same distribution as  $\frac{X}{Y}$ .

(b) Applying a change-of-variables from (X, Y) to  $(T, U) = (\frac{X}{Y}, Y)$ , show that the  $t_1$  distribution has the PDF  $f(t) = \frac{1}{\pi} \cdot \frac{1}{t^2+1}$ . Use this to show that  $\mathbb{E}[T^2] = \infty$ .

 $[t_1 \text{ is also called the Cauchy distribution. It is "heavy-tailed", meaning that the PDF <math>f(t)$  decays quite slowly to 0 as  $t \to \pm \infty$ .]

You may wish to use the change-of-variables formula: If (T, U) = g(X, Y) where  $g(\cdot)$  is a 1-to-1 map,  $f_{X,Y}(x, y)$  is the joint PDF of (X, Y), and  $f_{T,U}(t, u)$  is the joint PDF of (T, U), then

$$f_{T,U}(t,u) = f_{X,Y}(g^{-1}(t,u)) \cdot \left| \det \begin{pmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \end{pmatrix} \right|$$