Homework 6

Instructions: This homework is due in class on Friday May 24.

Please read the following guidelines for presenting your work and follow them diligently.

- Write your full name clearly on the top right of the first page. **Staple** pages on the left hand corner. Write neatly in complete sentences.
- You are required to work all the problems, however, only 5 will be graded. The page numbers below refer to the fifth edition of the text.
- Answer the questions in the order in which they are posed. Clearly number the questions as I have.
- You must first work independently on the homework. Please post questions on the discussion board or come to office hours once you have tried the problems.
- Be sure to show/explain your work thoughtfully. How you write your answers is important.
- If you use R to make plots or as a calculator, it is enough to simply include the output (e.g., appropriately labeled plot) in the main part of your homework without the R code.

---

1. Generate 10^4 values of the discrepancy statistic:

$$D_{skew} \quad = \quad (n-1)^{-3/2} \sum_{i=1}^{n} R_i^3,$$

for the data set in

x	14.0	9.4	12.1	13.4	6.3	8.5	7.1	12.4	13.3	9.1

which is assumed to arise from a $N(\theta, \sigma^2)$ distribution with both parameters unknown.

   (a) Graph the density histogram of the 10^4 values of D_{skew}.

   (b) Indicate how you would use the density histogram of D_{skew} to investigate the normality assumption and implement. What do you conclude?

2. The table shown here is taken from all 674 homicide convictions in the state of FL between 1967 and 1980 in which the suspect (defendant) was either black or white and the victim was either black or white.

   Each homicide was cross-classified according to three attributes: race of defendant, race of victim and whether the the case resulted in a death sentence.

Victim's race	Defendant's race	Death penalty? yes	no
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

(a) Are "race of defendant" and "death penalty" independent? Use Fisher's exact method on all $n = 674$ cases to conduct a size $\alpha = 0.05$ test.

(b) Repeat the analysis from part (a), but just using the data for homicides where

i. the victim was white.

ii. the victim was black

(c) Are there any differences in the conclusions of parts (a) and (b) and how might you reconcile these differences?

3. Suppose $X_1, X_2, \ldots, X_n \sim i.i.d. Bernoulli(\theta)$. From STAT 341, you know that $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ. Show that the conditional distribution of $p(X_1|Y)$ does not depend on θ. (This problem is meant to illustrate in a simple model that conditioning on a sufficient statistic for a parameter removes dependence on the parameter.)

4. For the birthweight data, investigate whether it is reasonable to assume a normal sampling model for the birthweight of babies using the following three approaches. Do the following separately for mothers who are smokers and also for non-smokers.

(a) Using a discrepancy statistic (give the observed value of the statistic, a density histogram based on simulation and a p-value)

(b) A residual plot using standardized residuals

(c) A normal probability plot using standardized residuals.

Be sure to write a concluding paragraph synthesizing your observations on whether normality is a reasonable assumption for these data. (Include only graphs and other relevant results in the main part of your homework. All code for this problem can be relegated to an appendix.)

5. The following ordered set of 27 P-values (from Kaati et al., *Eur Hum Genetics* 2007) were the result of testing many independent subgroups of a sample.

0.01 0.01 0.02 0.04 0.04 0.05 0.07 0.07 0.10 0.19 0.24 0.27 0.34 0.37
0.44 0.50 0.53 0.54 0.55 0.61 0.70 0.77 0.80 0.80 0.82 0.94 0.99

Investigate the hypothesis that the P-values all come from a uniform distribution on [0,1] using:

(a) a probability plot

(b) a χ^2 goodness of fit after grouping the data using a partition of five equal-length intervals.

```
#code for part 2 (a)
x<-seq(0,68,1)
y<-dhyper(x,m=483,n=191,k=68)    #hypergeometric PMF
plot(x,y,type="h",main="Hypergeometric PMF",xlab="x11",ylab="P(X11=x)")
sum(y[y <= dhyper(x=53,m=483,n=191,k=68)])    #to find p-value
```