STAT 504 Homework 2
Due Sunday January 24 by 11:30pm Eastern
with no penalty grace period to Tuesday, January 26 by 11:30pm Eastern

This homework pertains to material covered in the first and second week, particularly distributions or one-way tables, maximum likelihood, and goodness-of-fit tests. The assignment should be typed, with properly labeled computer output and the input code provided. You are encouraged to reasonably collaborate, but the write up should be your own. Please make sure you show your work! Submit your homework on ANGEL in the Homework drop-off box.

1. (20 pts) For each of the following scenarios, say whether the binomial distribution is a reasonable probability model for X. If yes, explain why you think the assumptions of binomial are satisifed. If no, explain how you think the assumptions are violated. How could you test data to see if your answer was reasonable?

   (a) After warming up with a few throws, you throw a dart at a dartboard 20 times. Let X be the number of times you hit it out of 30 flips.

   Solution: X has the binomial distribution, since the following assumptions are satisfied: There are $n = 20$ trials. Each trial has two possible outcomes (hit or miss). The probability of hit, p, could reasonably be the same for each trial and each trial is independent (barring small corrections, which is why we have a warm-up period).

   (b) After a warm-up, you throw a dart at a dartboard 20 times at a competition, then 1000 times as practice, then again for 20 times at a competition. Let X be the number of times you hit it out of 40 competition throws.

   Solution: X does not have the binomial distribution, since although there might be $n = 40$ independent Bernoulli trials, they are not indentically distributed, and come from two different populations (since you will be better after the 1000 practice throws). You could test by comparing the means of the first 20 and last 20 throws, or by looking for overdispersion in the variance.

   (c) A survey firm draws a random sample of 15 households in State College. Two adults from each household are asked, "Are you generally optimistic about the upcoming years if Donald Trump is elected President?." Let X be the number out of 30 who said "Yes".

   Solution: Not a binomial distribution, since the two adults opinion in the same household are likely correlated. To see this, say that proportion q of households are optimistic, the firm drew only one household rather than 15, and that within each household all adults have the same answer. This could not be $Bin(2, p)$ for any p, since the result is $X = 2$ with probability q and $X = 0$ with probability $1 - q$, and

$X = 1$ is impossible. We can weaken that to a correlation close to one between household members and still get a deviation from the Binomial distribution.

(d) A survey firm draws a random sample of 30 households in State College containing two or more adults. In each one, one adult member of the household is asked, "Are you generally optimistic about the upcoming years if Donald Trump is elected President?." Let X be the number out of 30 who said "Yes".

Solution: Yes the binomial would be a reasonable probability model for X since there are n=30 independent, trials (selection of household). There are two possible outcomes (Yes (success) or No (failure)) on each trial. The probability of success, p=proportion of all households in State College that who are optimistic, is the same for each trial (random phone call).

2. (20 pts) For each of the following scenarios, say whether a Poisson model for the sample $(X_1, X_2, ..., X_n)$ may be appropriate.

(a) Let say that in the Pennsylvania state lottery game Mega Millions, any player may win with probability 1 out of 2 million independently of other players. Also, suppose that the number of players each week is always equal to a constant N. Let $X_1, X_2, ..., X_n$ be the number of Mega Millions winners in the first n weeks of 2010.

Solution: Poisson distribution is appropriate, since here X obeys $Bin(N, p)$ with N very large and p very small. Besides $\lambda = Np$, which denotes the probability of winning the lottery, is constant.

(b) Now suppose the scenario as above, but the number of players N varies substantially from week to week.

Solution: Poisson distribution is NOT appropriate, since here $\lambda = Np$ is not constant.

(c) Let $X_1, X_2, ..., X_{12}$ be the number of cars arriving at a toll booth on a major highway during 12 consecutive hours in a single day.

Solution: Poisson distribution is NOT appropriate, since there is no guarantee that the arrival rate is the same over time. It is possible that the arrival rate gets higher as it draws closer to rush hour.

(d) Let $X_1, X_2, ..., X_{48}$ be the number of cars arriving at the toll booth between hours of 1 and 2 pm on 48 Fridays in a single year (excluding holidays).

Solution: Poisson distribution is appropriate, because first, one can assume that the arrival rate is constant on every Friday between 1 pm to 2pm. Second, the arrivals of cars are independent of one another.

3. (20 pts) Census figures show that the racial/ethnic distribution of persons eligible for jury duty in a large urban county is given by this table:

Black	Hispanic	Other
20%	15%	65%

From this county, 182 persons were summoned to appear for jury duty on a particular day. The observed racial/ethnic distribution of this sample is shown below.

Black	Hispanic	Other
45	35	102

Is it likely that this is a random sample from the eligible population? If not, how does the sample differ from the population? (Be sure to calculate X^2, G^2, residuals, and comment on the appropriateness of the χ^2 approximation, that is of χ^2 sampling distribution.)

Solution: From the R output or SAS output below, p-value is less than 0.05, therefore we reject the null hypothesis that the model fits at 0.05 level. In other words, we can conclude that the selected jurors are unlikely to have been chosen as a random sample. Since $n = 182$ here is large enough to satisfy the rule of thumb $n\pi_j > 5$ for every j, chi-square test is appropriate. Also since chi-square X^2 and G^2 statistics are close, we can say that the large sample approximation is appropriates.

Residuals: from the R/SAS output the Pearson residuals of the black and Hispanic groups are positive, showing that the observed frequency of these groups are higher than expected under the model, while that of the other group is negative, showing the observed frequency of this group is lower than expected under the model. Also the absolute value of all the Pearson residuals are smaller than 2, showing that the model appear to fit the cells.

R Code:

```
# X2 test
> ex1=chisq.test(c(45,35,102),p=c(0.2,0.15,0.65))
> ex1

        Chi-squared test for given probabilities

data:  c(45, 35, 102)
X-squared = 6.4496, df = 2, p-value = 0.03976

> ex1$statistic
X-squared
 6.449563
> ex1$p.value
[1] 0.03976446
> ex1$residuals
```

```
[1]  1.425436  1.473701 -1.498633

#G2 test
> G2=2*sum(ex1$observed*log(ex1$observed/ex1$expected))
> G2
[1] 6.237534
> 1-pchisq(G2,2)
[1] 0.04421164
```

SAS Code:

```
data racial;

input race $ count pi;

datalines;

1 45 .20

2 35 .15

3 102 .65

;

run;

proc freq data=racial;

weight count;

tables race /nocum all chisq testp=(.2 .15 .65);

run;

data cal;

set racial;

ecount=182*pi;

res=(count-ecount)/sqrt(ecount);
```

```
devres=sqrt(abs(2*count*log(count/ecount)))*sign(count-ecount);

run;

proc print;

run;

proc sql;

select sum((count-ecount)**2/ecount) as X2,

1-probchi(calculated X2,2) as pval1,

2*sum(count*log(count/ecount)) as G2,

1-probchi(calculated G2,2) as pval2

from cal;
quit;
```

```
The SAS System
The FREQ Procedure
race Frequency Percent Test
Percent
1 45 24.73 20.00
2 35 19.23 15.00
3 102 56.04 65.00


Chi-Square Test for Specified Proportions
Chi-Square 6.4496
DF 2
Pr > ChiSq 0.0398



Sample Size = 182
The SAS System

Obs race count pi ecount res devres
1 1 45 0.20 36.4 1.42544 4.36903
2 2 35 0.15 27.3 1.47370 4.17041
3 3 102 0.65 118.3 -1.49863 -5.49938


The SAS System

 X2   pval1  G2             pval2
6.449563 0.039764 6.237534 0.044212
```

4. (20 pts) R. A. Fisher, in a now-famous analysis of the results of Mendel's breeding experiments, concluded that the data were too good to be true as the data fit Mendel's theory far better than one would expect. In one pea breeding experiment, Mendel hypothesized that four categories of peas: smooth yellow, wrinkled yellow, smooth green, and wrinkled green, should occur with relative frequencies 9:3:3:1. The observed frequencies reported by Mendel from this trial are shown below.

Type of pea	count
Smooth yellow	315
Wrinkled yellow	101
Smooth green	108
Wrinkled green	32

Perform a goodness-of-fit test to show that these data fit the theory too well. Perform your analysis in the software of your choice and make sure to included relevant edited portions of the output.

Solution: We are testing the null hypothesis $H_0 : \pi_1 = \frac{9}{15}, \pi_2 = \frac{3}{15}, \pi_3 = \frac{3}{15}, \pi_4 = \frac{1}{15}$. From the R or SAS output below we can see that here X^2 and G^2 are very small, with p-values almost equal to 1. This suggests that the data confirms Mendel's theory.

R Code

```
> ex1=chisq.test(c(315,101,108,32),p=c(9,3,3,1)/sum(c(9,3,3,1)))
> ex1
#Chi-squared test for given probabilities

data:  c(315, 101, 108, 32)
X-squared = 0.47, df = 3, p-value = 0.9254

>
> ex1$statistic
X-squared
 0.470024
> ex1$p.value
[1] 0.9254259
> ex1$residuals
[1]  0.1272283 -0.3183064  0.3672766 -0.4665039

> G2=2*sum(ex1$observed*log(ex1$observed/ex1$expected))
```

```
> G2
[1] 0.4754452
> 1-pchisq(G2,3)
[1] 0.9242519
```

SAS Code :

```
data pea;
input type $ count;
datalines;

smye 315

wrye 101

smgr 108

wrgr 32

;

proc freq order=data;
weight count;
tables type /all testp=(56.25, 18.75, 18.75, 6.25);
run;
```

The FREQ Procedure

type	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
smye	315	56.65	56.25	315	56.65
wrye	101	18.17	18.75	416	74.82
smgr	108	19.42	18.75	524	94.24
wrgr	32	5.76	6.25	556	100.00

Chi-Square Test for Specified Proportions

Chi-Square	0.4700
DF	3
Pr > ChiSq	0.9254

5. (20 pts) Below is a table of 100 values of a random variable X drawn from a discrete distribution. For example, X could be the number of cars passing a point on a road from 1:00pm to 1:10pm on 100 consecutive weekdays, or the number of orders arriving at an exchange from 1:00:00.00pm to 1:00:00.01pm during 100 consecutive weekdays.

x	count
0	22
1	36
2	20
3	16
4	3
≥ 5	3

Perform the X^2 and G^2 goodness-of-fit tests to see if these data are Poisson. The last category (≥ 5), indicates that outcomes where the count was 6 and higher were rare so we collapsed them to the category $(x \geq 5)$. Also compute and explain the MLE of the rate and the expected counts.

Solution: The null hypothesis is H_0 : The data are Poisson. From the R output below, the MLE for lambda is 1.51 and expected counts are listed below as ex. Also from the

p-value of X^2 and G^2 tests, we cannot reject the null hypothesis that the Poisson model fits.

R Code:

```
> ob=c(22,36,20,16,3,3)
> data=c(rep(0,22),rep(1,36),rep(2,20),rep(3,16),rep(4,3),rep(5,3))
> ## Remark:pay attention to the data structure here
> lambda=mean(data)
> lambda
[1] 1.51

> n=100
> pihat=dpois(0:4,lambda)
> pihat=c(pihat,1-sum(pihat))
> ex=round(n*pihat,3)
> ex
[1] 22.091 33.357 25.185 12.676  4.785  1.905

> # Pearsons chi-square test
> X2=sum((ob-ex)^2/ex)
> X2
[1] 3.444192
> 1-pchisq(X2,4)
[1] 0.4864147

> # G2 test
> G2=2*sum(ob*log(ob/ex))
> G2
[1] 3.46346
> 1-pchisq(G2,4)
[1] 0.483456
```

SAS code:

```
data q5;
input type $ count pi E;

datalines;
0 22 .2209 22.09
1 36 .3336 33.36
```

```
2 20 .2518 25.18

3 16 .1268 12.68

4 3 .0479 4.79

5 3 .0146 1.46

;

run;

/**** computation of residuals, deviance residuals, X2 and G2 ****/

data q5_res;

set q5;

res=(count-e )/sqrt(e );

devres=sqrt(abs(2*count*log(count/e )))*sign(count-e );

run;

proc print;

run;

proc sql;

select sum((count-e)**2/e) as X2,

1-probchi(calculated X2,98) as pval1,

2*sum(count*log(count/e)) as G2,

1-probchi(calculated G2,98) as pval2

from q5;

quit;
```

Obs	type	count	pi	E	res	devres
1	0	22	0.2209	22.09	-0.01915	-0.42383
2	1	36	0.3336	33.36	0.45708	2.34171
3	2	20	0.2518	25.18	-1.03229	-3.03524
4	3	16	0.1268	12.68	0.93235	2.72800
5	4	3	0.0479	4.79	-0.81787	-1.67556
6	5	3	0.0146	1.46	1.27451	2.07871

The SAS System

X2	pval1	G2	pval2
4.437483	1	5.04683	1