HW7

Code ▾

Problem 7A

Test the goodness of fit of the data to the genetic model given in Problem 55 of Chapter 8.

Hide

```
f34 <- c(1997,906,904,32)

l_f34 <- function(t) {
  sum_l_f34 <- dmultinom(x=f34, size=sum(f34), prob=c((.25*(2+t)), (.25*(1-t)), (.25*(1-t)), (.25*t)))
  return (sum_l_f34)
}
theta_o <- optimize(l_f34, c(0,1), maximum=TRUE); theta_o
```

```
$maximum
[1] 0.03571143

$objective
[1] 6.249532e-06
```

Hide

```
th <- theta_o$maximum

probs <- c(.25*(2+th),.25*(1-th),.25*(1-th),.25*th)

chisq.test(f34,p=probs)
```

```

	Chi-squared test for given probabilities

data:  f34
X-squared = 2.0154, df = 3, p-value = 0.5692
```

Hide

```
p_value = 1-pchisq(2.0155,2)
p_value
```

```
[1] 0.3650394
```

Since p_value = 0.3650394, the fit is not rejected. Model is good.

Problem 7B

Yip et al. (2000) studied seasonal variations in suicide rates in England and Wales during 1982–1996, collecting counts shown in the following table:

Hide

```
suic_f = c(1362,1244,1496,1452,1448,1376,1370,1301,1337,1351,1416,1226)
chisq.test(suic_f)
```

```

	Chi-squared test for given probabilities

data:  suic_f
X-squared = 53.786, df = 11, p-value = 1.292e-07
```

Hide

```
suic_m = c(3755,3251,3777,3706,3717,3660,3669,3626,3481,3590,3605,3392)
chisq.test(suic_m)
```

```

	Chi-squared test for given probabilities

data:  suic_m
X-squared = 74.56, df = 11, p-value = 1.646e-11
```

Both datasets show seasonality as we see dips and rises at the same times throughout the year. Mainly dips in December and February, and rises between March through May.

Problem 7C

In the Current Population Survey of March 2005, men were classified by employment status and marital status. Here are the data; you may assume they come from a simple random sample of U.S. men. “Once married” means, “widowed, divorced, or separated

a. H_0 : Employment and marital status are independant. H_A : They are dependant.

Hide

```
emp <- data.frame (
  Employed = c(790, 98, 209),
  Unemployed = c(56, 11, 27),
  not_in_labor = c(21, 7, 12)
)

# chisq.test(emp)$exp
chisq.test(emp)
```

```
Chi-squared approximation may be incorrect
```

```


	Pearson's Chi-squared test

data:  emp
X-squared = 13.369, df = 4, p-value = 0.009609
```

Since p_value = 0.009609, the variables are dependant, we reject the null hypothesis.

b. 772.6231 is the expected count under the null hypothesis. P(married_and_employed) = P(married)P(employed) since under the null they are independent. R gives a warning message because the expected count in one of the values is close to 0. With such a large n (n=1232), we should use the poisson approx rather than the normal approx that is used in chi-squared test.

Problem 7D

Is R calculating by…

Hide

```
exp_emp <- data.frame (
  exp_Employed = c(772.6231, 66.204712, 28.172218),
  exp_Unemployed = c(103.3729, 8.857839, 3.769293),
  exp_not_in_labor = c(221.0041, 18.937449, 8.058489)
)

sum((emp-exp_emp)^2/exp_emp)
```

```
[1] 1532.186
```

Hide

```
2*sum(emp*log(emp/exp_emp))
```

```
[1] 888.8899
```

Not sure why my values are so large in R, did the calculations on hand and it seems that R calculates through $\sum (O_i - E_i)^2/E_i$

Problem 7E

Use the datea in Problem 7C to construct an approximate 95% confidence interval for the proportion of unemployed men among all U.S. men in 2005.

Hide

```
n = 1232
p_hat = (56+11+27)/n
se = sqrt(p_hat * 0.9237/n)

(0.0763-1.96*se)
```

```
[1] 0.06147568
```

Hide

```
(0.0763+1.96*se)
```

```
[1] 0.09112432
```

Therefore, the 95% CI is (0.061, 0.091).

Problem 7F

If possible use the data in Problem 7C to construct an approximate 95% confidence interval for the difference between the proportion of employed and unemployed men among all U.S. men in 2005.

Hide

```
p_hat_x = 0.890422
p_hat_y = 0.0762987
p_hat = p_hat_x - p_hat_y
se = 0.0158

(p_hat-1.96*se)
```

```
[1] 0.7831553
```

Hide

```
(p_hat+1.96*se)
```

```
[1] 0.8450913
```

Therefore, the 95% CI is (0.7831553, 0.8450913).

Problem 7G

Consider testing goodness of fit for a multinomial distribution with two cells. Denote the number of observations in each cell by X1 and X2 and let the hypothesized probabilities be p1 and p2. Pearson’s chi-square statistic is equal to…

$X_1 + X_2 = n$ and $p_1 + p_2 = 1$, chi squared stat is:

$$\frac{(X_1-np_1)^2}{np_1}+\frac{(X_2-np_2)^2}{np_2}=\frac{(1-p_1)(X_1-np_1)^2+p_1(n-X_1-n(1-p_1))^2}{np_1(1-p_1)}$$

$$=\frac{(X_1-np_1)^2}{np_1(1-p_1)}$$

X_1 is binom so it has exp np_1 and var $np_1(1-p_1)$

When n is large, X_1 is approximately normal, so the stat is approx the square of a normal which is a chi-squared with 1 degree of freedom.

Problem 7G

Generate an i.i.d. sample of size 1000 from the binomial distribution with parameters n = 5 and p = 0.4. Compute the counts in the categories 0, 1, 2, 3, 4, and 5. You shouldnt have to get each count separately (try the R function table())

Hide

```
x <- rbinom(1000,5,0.4)
p_hat <- mean(x)/5
obs <- hist(x, plot=FALSE, breaks=seq(-0.5,5.5,by=1))$counts
chisq.test(obs, p=dbinom(0:5,5,p_hat))
```

```


	Chi-squared test for given probabilities

data:  obs
X-squared = 9.3629, df = 5, p-value = 0.09543
```

But like in problem 7A, the df is wrong since df should = 4, so…

Hide

```
1-pchisq(9.3629, 4)
```

```
[1] 0.05264187
```

So, the p-value = 0.0526. Since the p-value is above 0.05, we fail to reject the null hypothesis.

Problem 7I

Now repeat the sampling in the previous problem 2000 times. That is, generate 2000 independent samples, each of size 1000, from the binomial distribution with n = 5 and p = 0.4. For each of your samples, get the counts in each category 0, 1, 2, 3, 4, and 5.

Hide

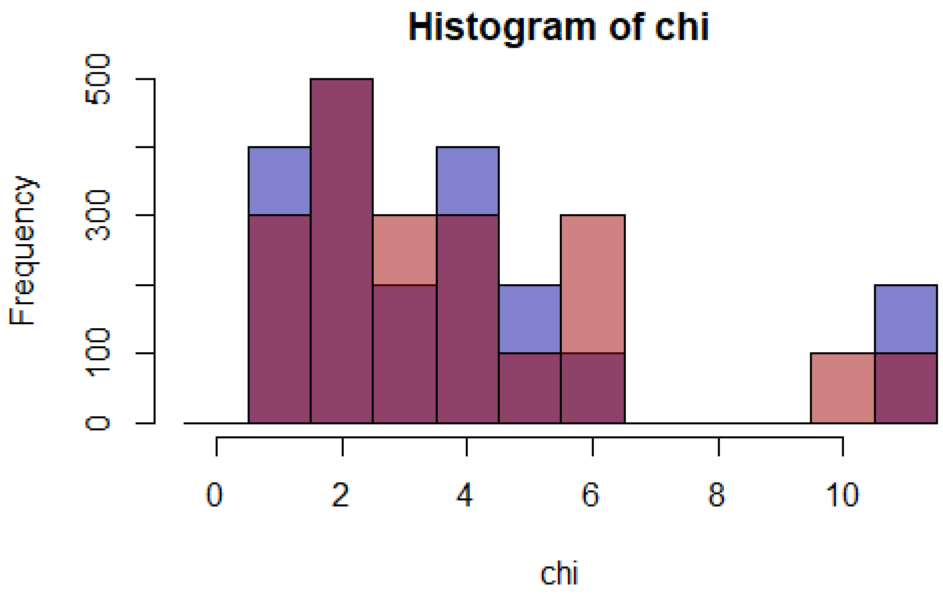
```
x <- rbinom(20000,5,0.4)
sample <- matrix(x,nrow=1000,ncol=2000)
p_hats <- colMeans(sample)/5
obs <- apply(sample,2,table)
exp <- sapply(p_hats,dbinom,x=0:5,size=5)*1000

chi <- apply((obs-exp)**2/exp,2,sum)
chi2 <- apply(obs*log(obs/exp),2,sum)*2

chi_hist <- hist(chi, plot=FALSE, breaks=seq(-0.5,11.5,by=1))
chi2_hist <- hist(chi2, plot=FALSE, breaks=seq(-0.5,11.5,by=1))
```

Hide

```
mycol <- rgb(0, 0, 155, max = 255, alpha = 125, names = "blue50")
mycol2 <- rgb(155, 0, 0, max = 255, alpha = 125, names = "red50")
plot(chi_hist, col=mycol)
plot(chi2_hist, col=mycol2,add=TRUE)
```


Histogram of chi
Frequency
chi