Homework5_Hwang

Charles Hwang

3/25/2022

Charles Hwang

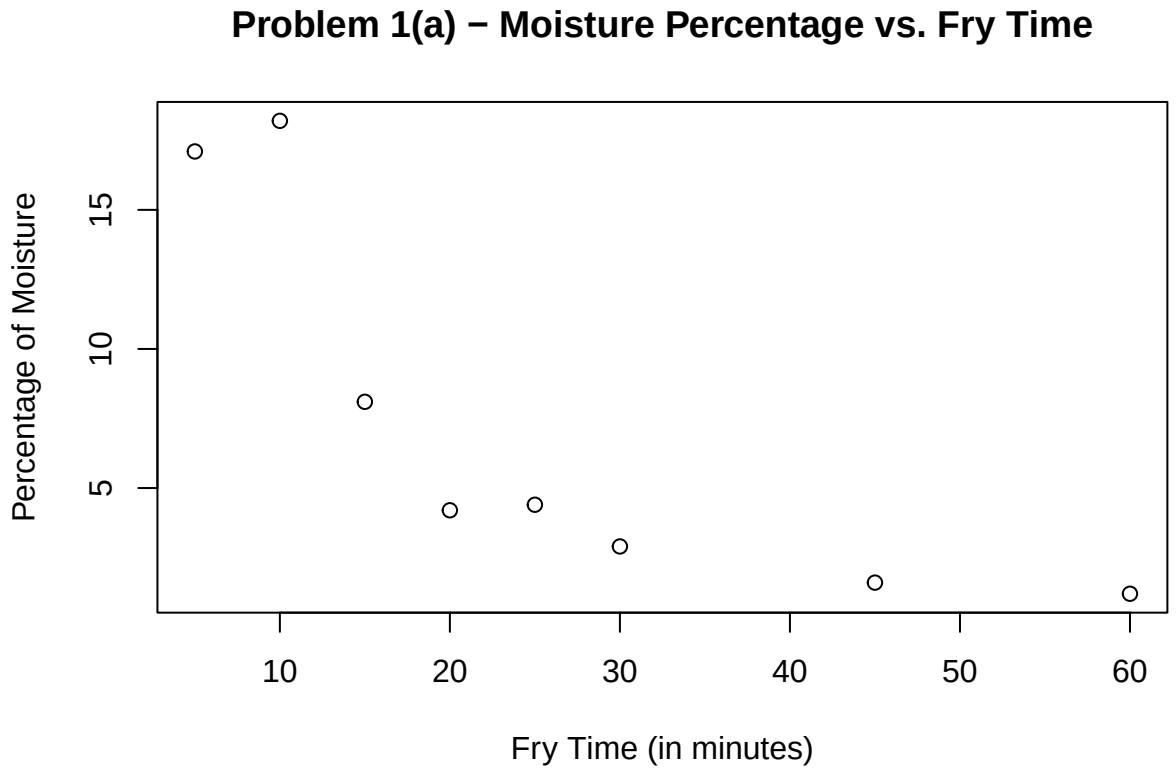
Dr. Perry

STAT 451-001

25 March 2022

Problem 1

```
rm(list=ls())
x<-c(5,10,15,20,25,30,45,60)
y<-c(17.1,18.2,8.1,4.2,4.4,2.9,1.6,1.2)
plot(x,y,xlab="Fry Time (in minutes)",ylab="Percentage of Moisture",main="Problem 1(a) - Moisture Percer
```




```
# There appears to be a negative quadratic or negative exponential relationship in this data.
cor(x,y,method="pearson")                                    # Problem 1(b)
```

```
## [1] -0.7984417
```

```r
# This may not be an appropriate statistic because the Pearson correlation coefficient is
# used to measure the strength and direction of a linear relationship, which we do not
# have in this data.
# H0: rho = 0                                                    # Problem 1(c)
# HA: rho =/= 0
r<-rep(NA,10000)
set.seed(2503,sample.kind="Rounding")
for (i in 1:10000){ys<-sample(y,length(y),replace=FALSE)
r[i]<-cor(x,ys)}
sum(abs(r)>=abs(cor(x,y)))/10000
```

```
## [1] 0.0224
```

```r
# We reject H0 at the alpha = 0.10 level. There is sufficient evidence (p = 0.0224)
# that the Pearson correlation coefficient is significant.
# Ranks: (1,7), (2,8), (3,6), (4,4), (5,5), (6,3) (7,2), (8,1)  # Problem 1(d)
cor(rank(x),rank(y))
```

```
## [1] -0.952381
```

```r
cor(x,y,method="spearman")
```

```
## [1] -0.952381
```

```r
# We can see the two values are the same (rho = -0.952381). This value seems like a
# better correlation coefficient for this data because Spearman's rank correlation
# coefficient is able to assess nonlinear relationships, which we have in this data.
# H0: rho = 0                                                    # Problem 1(e)
# HA: rho =/= 0
cor.test(x,y,method="spearman",exact=TRUE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  x and y
## S = 164, p-value = 0.001141
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.952381
```

```r
cor.test(x,y,method="spearman",exact=TRUE)$statistic
```

```
##   S
## 164
```

```r
cor.test(x,y,method="spearman",exact=TRUE)$p.value
```

```
## [1] 0.001140873
```

```r
# We reject H0 at the alpha = 0.05 level. There is sufficient evidence (p = 0.001140873)
# that the Spearman's rank correlation coefficient is significant.
t<-data.frame(rep("",8),c("C",rep("",7)),c("D","D",rep("",6)),c(rep("D",3),rep("",5)),c(rep("D",3),"C",
2*sum(t=="C")/choose(length(y),2)-1                              # Problem 1(f)
```

```
## [1] -0.8571429
```

```r
cor(x,y,method="kendall") # "C" for concordant pair, "D" for discordant pair
```

```
## [1] -0.8571429
```

We can see the two values are the same (τ = -0.8571429), as intended.

Problem 2

Problem 2(a)

H_0 : There is no association between diabetes categorization and sex

H_A : There is an association between diabetes categorization and sex

```
sd<-data.frame(c(rep("M",6),rep("F",6)),c("ND",rep("PD",3),"D","D",rep("ND",4),"D","D"))
names(sd)<-c("Sex","Diabetes Categorization")        # Problem 2(b)
chisq.test(table(sd),correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(sd)
## X-squared = 4.8, df = 2, p-value = 0.09072
```

```
chisq.test(table(sd),correct=FALSE)$statistic
```

```
## X-squared
##       4.8
```

```
# We reject H0 at the alpha = 0.10 level. There is sufficient evidence (p = 0.09071795)
# that there is an association between diabetes categorization and sex.
mean(chisq.test(table(sd),correct=FALSE)$expected<5) # Problem 2(c)
```

```
## [1] 1
```

```
# We can see that 100 percent of cells have expected values less than 5, which is a
# clear violation of the assumptions for a chi-squared test.
library(gtools)                                      # Problem 2(d)
com<-combinations(nrow(sd),sum(sd=="M"))
nrow(com)
```

```
## [1] 924
```

```
X<-rep(NA,nrow(com))                                 # Problem 2(e)
set.seed(2503)
for (i in 1:nrow(com)){sdv<-sd
  ind<-c(com[i,],setdiff(1:12,com[i,]))
  sdv$Sex<-sdv$Sex[ind]
  X[i]<-chisq.test(table(sdv))$statistic}
sum(X>=chisq.test(table(sd),correct=FALSE)$statistic)/nrow(com)
```

```
## [1] 0.961039
```

We fail to reject H_0 at the $\alpha = 0.10$ level. There is insufficient evidence ($p = 0.961039$) that there is an association between diabetes categorization and sex.

Problem 2(f)

The p-value from the standard χ^2-test was 0.090718, while the p-value from the permutation test was 0.961039.

It does not appear the standard χ^2-test would have been sufficient. We saw in problem 2(c) the expected values are all less than 5 which violates the assumptions for the standard χ^2-test.

Problem 3

```
co<-data.frame(c(rep("L",6),rep("H",9)),c(rep("N",4),"NN","NN",rep("N",8),"NN")) # 3(a)
names(co)<-c("Cont. Level","Distance")
nrow(combinations(nrow(co),sum(co=="L")))
```

```
## [1] 5005
```

```
m1<-sum(co=="N")                                                    # Problem 3(b)
m2<-sum(co=="NN")
n1<-sum(co=="L")
th<-choose(m1,3)*choose(m2,n1-3)/choose(nrow(co),n1)
fo<-choose(m1,4)*choose(m2,n1-4)/choose(nrow(co),n1) # Use this for problem 3(c)
fi<-choose(m1,5)*choose(m2,n1-5)/choose(nrow(co),n1)
si<-choose(m1,6)*choose(m2,n1-6)/choose(nrow(co),n1)
P<-data.frame(th,fo,fi,si)
names(P)<-c("x = 3","x = 4","x = 5","x = 6")
row.names(P)<-"P((L,N) = x)"
P
```

```
##                   x = 3     x = 4     x = 5     x = 6
## P((L,N) = x) 0.04395604 0.2967033 0.4747253 0.1846154
```

```
# H0: There is no relationship between contamination level and distance  # Problem 3(c)
# HA: There is a relationship between contamination level and distance
1-fo # Calculating P((L,N)=/=4) = 1 - P((L,N)=4)
```

```
## [1] 0.7032967
```

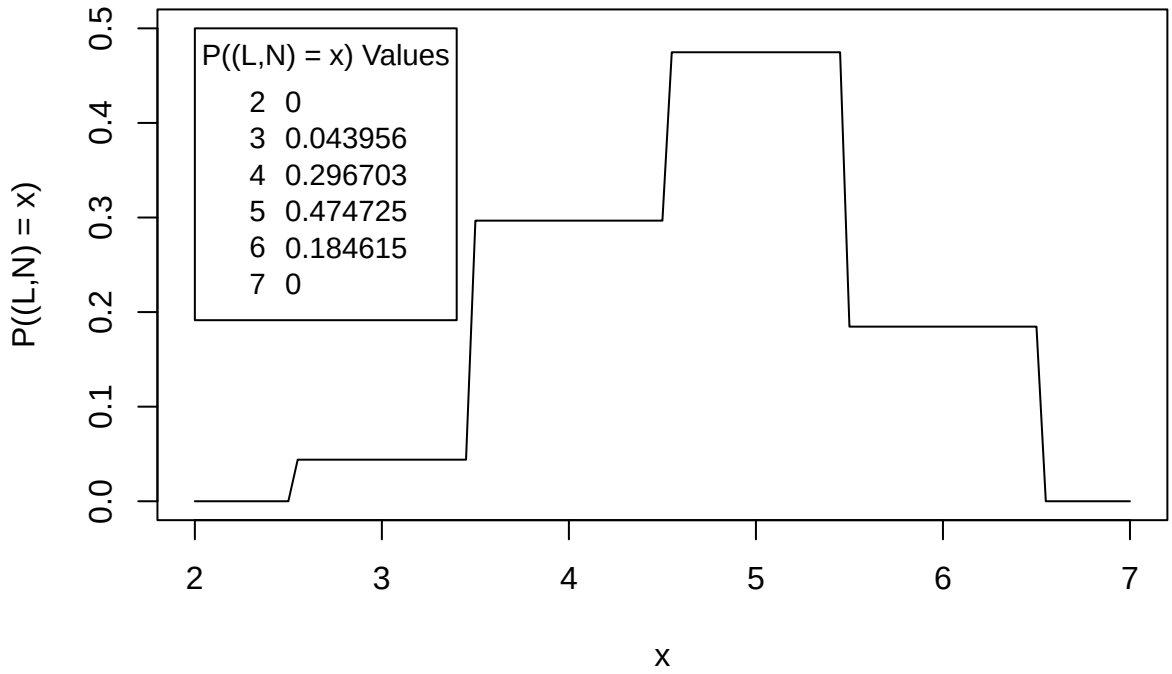
```
# We fail to reject H0 at the alpha = 0.05 level. There is insufficient evidence (p = 0.71)
# that there is a relationship between contamination level and distance.
fisher.test(table(co))                                              # Problem 3(d)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(co)
## p-value = 0.5253
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.1466385 264.2650545
## sample estimates:
## odds ratio
##   3.622029
```

```
th+fo+si
```

```
## [1] 0.5252747
```

```
# We can see that R calculated the p-value by adding P((L,N) = 4) to all
# P((L,N) = x) < P((L,N) = 4) for all whole numbers x. For example, from problem 3(b),
# we can see p = P((L,N) = 3) + P((L,N) = 4) + P((L,N) = 6).
plot(0,0,type="n",xlab="x",ylab="P((L,N) = x)",main="Problem 3(d) - (Rough) Histogram of Probabilities
curve(choose(m1,x)*choose(m2,n1-x)/choose(nrow(co),n1),add=TRUE)
legend(2,0.5,title="P((L,N) = x) Values",pch="234567",legend=round(c(0,th,fo,fi,si,0),digits=6),cex=0.9
```



We can see the probabilities for (L, N) for each value of x visualized in the "histogram".

Problem 4

Problem 4(a)

$H_0 : P_W = P_S$

$H_A : P_W > P_S$

```
de<-data.frame(c(rep("Y",31),rep("N",29)),c(rep("Y",18),rep("N",13),rep("Y",4),rep("N",25)))
names(de)<-c("Winter","Summer")                # Problem 4(b)
table(de)[c("Y","N"),c("Y","N")]
```

```
##       Summer
## Winter  Y  N
##      Y 18 13
##      N  4 25
```

```
# We can see the table is ordered in the correct direction.
chisq.test(table(de),correct=FALSE)$statistic # Problem 4(c)
```

```
## X-squared
##  12.64595
```

```
chisq.test(table(de),correct=FALSE)$p.value    # Problem 4(d)
```

```
## [1] 0.0003763804
```

Problem 4(e)

We reject H_0 at the $\alpha = 0.05$ level. There is sufficient evidence ($\text{p} = 3.7638037 \times 10^{-4}$) that $P_W > P_S$.