STAT 135 CONCEPTS OF STATISTICS
HOMEWORK 5

Assigned July 22, 2021, due July 27, 2021

This homework pertains to materials covered in Lecture 9 and 10. The assignment can be typed or handwritten, with your name on the document, and **with properly labeled input code and computer output for those problems that require it**. To obtain full credit, please write clearly and show your reasoning. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the Week 5 Assignment on bCourses.

Note in this homework, we use the following abbreviations: Uniformly most powerful (UMP) test, likelihood ratio test (LRT).

Problem 1. An experiment was done to measure the effects of ozone, a component of smog. A group of 22 seventy-day-old rats were kept in an environment containing ozone for 7 days, and their weight gains were recorded. Another group of 23 rats of a similar age were kept in an ozone-free environment for a similar time, and their weight gains were recorded. The data (in grams) are given below.

$$\begin{aligned}\mathbf{X}_{\text{control}} = \{&41.0,\ 38.4,\ 24.9,\ 25.9,\ 21.9,\ 18.3,\ 13.1,\ 27.3,\ 28.5,\\ &-16.9,\ 17.4,\ 21.8,\ 15.4,\ 27.4,\ 19.2,\ 22.4,\ 17.7,\ 26.0,\\ &29.4,\ 21.4,\ 22.7,\ 26.0,\ 26.6\}\\ \mathbf{X}_{\text{ozone}} = \{&10.1,\ 6.1,\ 20.4,\ 7.3,\ 14.3,\ 15.5,\ -9.9,\ 6.8,\ 28.2,\\ &17.9,\ -12.9,\ 14.0,\ 6.6,\ 12.1,\ 15.7,\ 39.9,\ -15.9,\ 54.6,\\ &-14.7,\ 44.1,\ -9.0,\ -9.0\}\end{aligned}$$

Establish a null and an alternative hypotheses, and analyze the data using both Normal population assumption and the non-parametric Mann-Whitney test to determine the effect of ozone. Write a summary of your conclusions. [This problem is from Doksum and Sievers (1976) who provide an interesting analysis.]

Solution. This is an open question. We want to test

$$H_0 : \mu_c = \mu_o \text{ versus } H_1 : \mu_c \neq \mu_o$$

in which μ_c and μ_o are the population means of the control group and the ozone treatment group respectively.

For the parametric test, students can use both equal variance assumption and the more general test that we saw in homework 4. We look at the analyses from two of your classmates.

(1) Under the Normal assumption with equal variances, the rejection of the t-test is

$$R = \left\{ \frac{\left|\bar{X}_n - \bar{Y}_m\right|}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \geq t_{n+m-2}\left(\alpha/2\right) \right\},$$

for which we can calculate

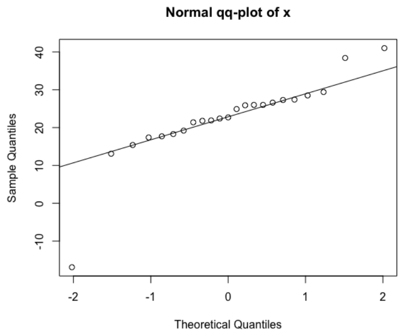
$$\bar{X}_n = 22.426,\ \bar{Y}_m = 11.009,\ S_p^2 = 236.0396,\ t_{n+m-2}\left(0.05/2\right) = 2.017,$$

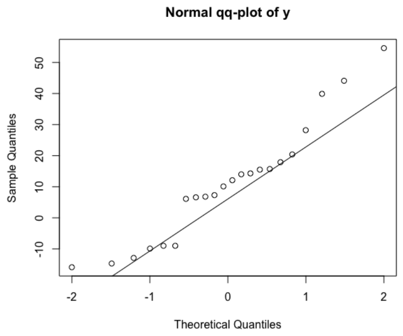
$$\frac{\bar{X}_n - \bar{Y}_m}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = 2.491887 > t_{n+m-2}\left(0.05/2\right).$$

Therefore, we reject the null hypothesis H_0, and conclude that there is significant evidence of effects of ozone on the weight gains.

Sam Menefee verified the normality of these two samples using qq-plots:

```
x <- c(41.0,38.4,24.9,25.9,21.9,18.3,13.1,27.3,
    28.5,-16.9,17.4,21.8,15.4,27.4,19.2,22.4,17.7,
    26.0,29.4,21.4,22.7,26.0,26.6)
y <- c(10.1,6.1,20.4,7.3,14.3,15.5,-9.9,6.8,
    28.2,17.9,-12.9,14.0,6.6,12.1,15.7,39.9,
    -15.9,54.6,-14.7,44.1,-9.0,-9.0)
qqnorm(x, main = "Normal qq-plot of x"); qqline(x)
qqnorm(y, main = "Normal qq-plot of y"); qqline(y)
```

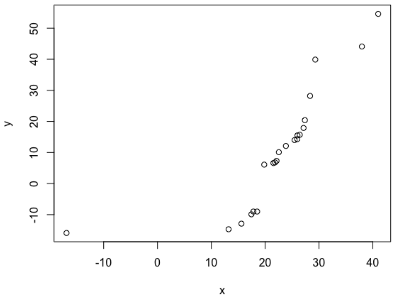




Since for both samples, the quantile-quantile points are closed to the theoretical qq-lines, the normal assumption is likely good.

He also directly compared the qq-plot between the two samples. If the two samples come from the same population, then the quantile-quantile points should be linear on the qq-plot.

```
qqnorm(x, y)
```



From this plot, we can see the points do not appear to linear — the two populations are different, and we can visually reject H_0.

(2) Under the Normal assumption without the equal variances, Dongwoong Seo utilized the result from the previous homework:

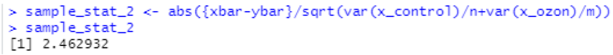
Assume $\begin{cases} X_i \overset{i.i.d}{\sim} N(\mu_X, \sigma_X^2) & \text{for } i=1,\dots,23 \\ Y_j \overset{i.i.d}{\sim} N(\mu_Y, \sigma_Y^2) & j=1,\dots,22 \end{cases}$ are independent.

Recall that $\dfrac{(\bar{X}_n - \bar{Y}_m) - \cancel{(\mu_X - \mu_Y)}}{\sqrt{\dfrac{S_1^2}{n} + \dfrac{S_2^2}{m}}} \xrightarrow{d} t(\min\{n-1, m-1\})$ from HW#4.

$\to = 0$ under H_0.

$\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X \neq \mu_Y. \end{cases}$ with $\alpha = 0.05$.

In this case, test statistic is $2.462932 > t_{\frac{\alpha}{2}}(43) = 2.016692$.



```
> sample_stat_2 <- abs({xbar-ybar}/sqrt(var(x_control)/n+var(x_ozon)/m))
> sample_stat_2
[1] 2.462932
```

Therefore, we reject H_0 again.

For the non-parametric Mann-Whitney test, we first calculate the rank sums of each sample:

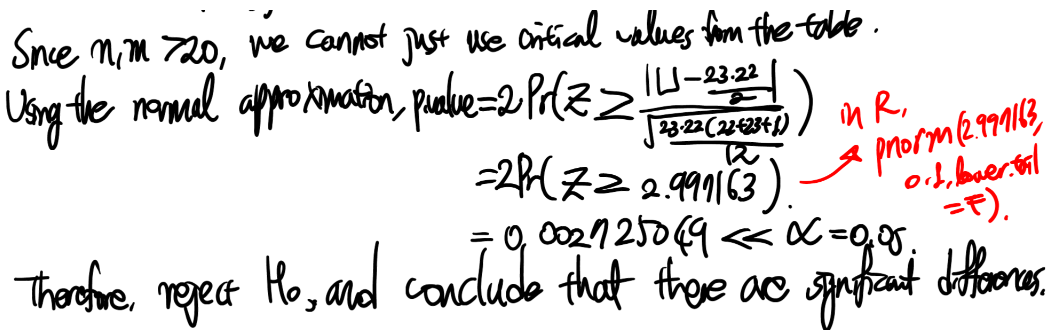
```
ranks_pool <- rank(c(x,y))
R1 <- sum(head(ranks_pool, length(x)))
R2 <- sum(tail(ranks_pool, length(y)))
```

Then

$$U_1 = mn + \frac{n(n+1)}{2} \quad - R_1 = 121,$$

$$U_2 \ = mn + \frac{m(m+1)}{2} - R_2 = 385.$$

Thus, $U = \min\{U_1, U_2\} = 121$. Since $n, m > 20$, the table of critical values cannot be used, we can calculated the p-value via Normal approximation. See Dongwoong Seo's solution:



Since $n, m > 20$, we cannot just use critical values from the table.

Using the normal approximation, $\text{pvalue} = 2\Pr\left(Z \geq \frac{\left|U - \frac{23\cdot 22}{2}\right|}{\sqrt{\frac{23\cdot 22(22+23+1)}{12}}}\right)$

$= 2\Pr(Z \geq 2.991163)$. $\longrightarrow$ in R, pnorm(2.991163, 0,1, lower.tail =F).

$= 0.002725069 \ll \alpha = 0.05$.

Therefore, reject H_0, and conclude that there are significant differences.

Problem 2. A cross-over trial investigated whether eating oat bran lowered serum cholesterol levels. Fourteen individuals were randomly assigned a diet that included either oat bran or corn flakes. After two weeks on the initial diet, serum cholesterol were measured and the participants were then "crossed-over" to the alternate diet. After two-weeks on the second diet, cholesterol levels were once again recorded.

Data appear below. The variables CORNFLK and OATBRAN in the table represent cholesterol levels (mmol/L) of the participant on the corn flake diet and the oat bran diet respectively.

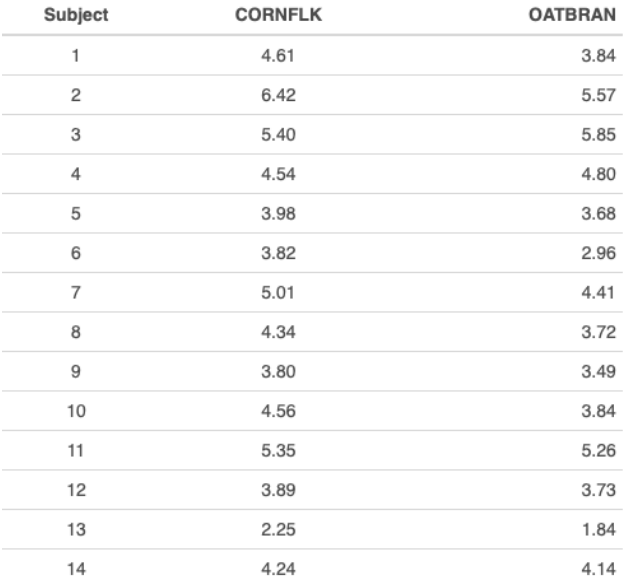
(*You can download this dataset from the Data_sets directory on bCourses*).

(1) Use normal theory to test the hypothesis that the cholesterol levels while on the corn flake diet is less than the oat bran diet ($\alpha = 0.05$);
(2) Perform Wilconxon signed rank test by hand to test the same hypothesis ($\alpha = 0.05$);
(3) Compare the p-value from the two tests. Are they very different?

Solution.

(1) The rejection region for testing $H_0 : \mu_c = \mu_o$ versus $H_1 : \mu_c < \mu_0$ under the Normal assumption should be

$$R = \left\{ \frac{\bar{D}_n}{\sqrt{\frac{1}{n} S_D^2}} \leq -t_{n-1}(\alpha) \right\},$$



Subject	CORNFLK	OATBRAN
1	4.61	3.84
2	6.42	5.57
3	5.40	5.85
4	4.54	4.80
5	3.98	3.68
6	3.82	2.96
7	5.01	4.41
8	4.34	3.72
9	3.80	3.49
10	4.56	3.84
11	5.35	5.26
12	3.89	3.73
13	2.25	1.84
14	4.24	4.14

for which we can calculate:

$$\bar{D}_n = 0.3629,\ S_D^2 = 0.1648,\ -t_{n-1}(\alpha) = -1.770933$$

$$\frac{\bar{D}_n}{\sqrt{\frac{1}{n}S_D^2}} = 3.3444.$$

Therefore, we fail to reject H_0. Also, we can get the p-value $=$ 0.9974, which is way larger than 0.05.

```
dat = read.table(file='~/Downloads/oatbran.txt',
    header=TRUE)
mean(dat$CORNFLK-dat$OATBRAN)
var(dat$CORNFLK-dat$OATBRAN)
qt(0.05, df=nrow(dat)-1)
pt(3.3444, nrow(dat)-1)
```

(2) For the Wilcoxon signed rank test, we calculate the positive rank sum from the differences `dat$CORNFLK-dat$OATBRAN`: $W_+ = 93$.

```
which_pos <- (dat$CORNFLK-dat$OATBRAN)>0
ranks_pool <- rank(abs(dat$CORNFLK-dat$OATBRAN))
W_plus <- sum(ranks_pool[which_pos])
```

Since the alternative is one-sided, the rejection region is $R = \{W_+ \leq c\}$, in which the critical value c can be looked up from the table for 14 paired observations: $c = 26$. Thus, our observed W_+ is not in the reject region, and we fail to reject the null hypothesis.

(3) From R, we can check the p-value for the Wilcoxon ranked sum test:

```
wilcox.test(dat$CORNFLK, dat$OATBRAN, paired = TRUE,
    alternative = 'less')
```

in which we get a p-value of 0.9966, which is comparable to 0.9974 from (1).

Problem 3. Let X equal the number of alpha particles emitted from barium-133 in 0.1 second and counted by a Geiger counter. One hundred observations of X produced the following table:

Category	X	Obs'd
1	0,1,2*	5
2	3	13
3	4	19
4	5	16
5	6	15
6	7	9
7	8	12
8	9	7
9	10,11,12*	4
		$n = 100$

It is claimed that X follows a Poisson distribution. Use a chi-square goodness-of-fit statistic to test whether this is true.

Solution. We want to test whether

$$H_0 : p_1(\lambda) = e^{-\lambda}\left(\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\frac{\lambda^2}{2!}\right), p_2(\lambda) = e^{-\lambda}\frac{\lambda^3}{3!}, \ldots, p_9(\lambda) = 1 - e^{-\lambda}\sum_{i=0}^{9}\frac{\lambda^i}{i!}$$

is true.

To obtain the maximum likelihood estimator for λ under H_0, one could use the sample mean $\bar{X}_n \approx \left(\frac{0+1+2}{3} * 5 + 3 * 13 + \ldots + 9 * 7 + \frac{10+11+12}{3} * 4\right)/100 = 5.56$. Or we can find the MLE using `optimize()` in R:

```
counts <- c(5,13,19,16,15,9,12,7,4)
log_lik_H0 <- function(lambda){
  c = ppois(12, lambda) # normalizing constant
  ll = 0
  for (i in 1:length(counts)){
    if(i==1){
      ll = ll+counts[i]*log(ppois(2, lambda)/c)
    } else if(i==length(counts)){
      ll = ll+counts[i]*log(ppois(12, lambda)/c-ppois(9, lambda)/c)
    } else{
      ll = ll+counts[i]*log(dpois(i+1, lambda)/c)
    }
  }
  return(ll)
}

optimize(log_lik_H0, interval = c(0,12), maximum = TRUE)

$maximum
[1] 5.617932

$objective
[1] -212.3556
```

Notice we used 9 categories for the null multinomial distribution with the last category being $\{10, 11, 12\}$. You can also specify the last category to be $\{10, 11, \ldots\}$. The former parametrization requires a normalization. Thus, the MLE under H_0 is $\hat{\lambda} = 5.617$. We then calculate the expected counts in each cell:

```
n <- 100
E <- rep(NA, 9)
lambda_hat <- 5.617
for (i in 1:9){
    c_hat <- ppois(12, lambda_hat) # normalizing constant
    if(i==1){
        E[i] <- n*ppois(2, lambda_hat)/c_hat
    } else if(i==length(O)){
        E[i] <- n*(ppois(12, lambda_hat)-ppois(9, lambda_hat))/c_hat
    } else{
        E[i] <- n*dpois(i+1, lambda_hat)/c_hat}
}
E
[1]  8.183940 10.795055 15.158956 17.029572 15.942517 12.792731
[9]  8.982096 5.509306
```

We know from Theorem A of Lecture 13 that

$$-2 \log \lambda(\mathbf{X}_n) = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{d} \chi^2_{m-2} = \chi^2_7.$$

Since $\sum_{i=1}^{m}(O_i - E_i)^2/E_i = 5.678958$, and $\chi^2_7(0.05) = 14.06714$, we fail to reject the null hypothesis that the number of alpha particle emissions is Poisson distributed. (*Or*, you can calculate the p-value which can be calculated to be 0.578, which is very probable for H_0 to be true.)

Problem 4. During phase 3 trial of the Pfizer/BNT 162b2 vaccine, some vaccine recipients were asked to complete diaries of their symptoms during the 7 days after vaccination.

Here is a table summarizing the number of recipients who experienced headaches after the second dose:

	Pfizer-BioNTech Vaccine	Placebo
No headache	1013	1597
Mild	538	321
Moderate	480	170
Severe	67	15

TABLE 1. Systemic reactions in persons aged 18-55 years. Data from CDC.gov.

Use the test for homogeneity to examine whether two columns in Table 1 as multinomial variables have equal probabilities of having headaches of a certain severity.

Solution. We want to test:

$$H_0 : \pi_{i,\text{Pfizer}} = \pi_{i,\text{Placebo}},\ i = \text{‘No headache’, ‘Mild’, ‘Moderate’ or ‘Severe’} .$$

Following the steps of the χ^2 test of homogeneity, we first calculate the expected counts for each cell:

```
Observed <- matrix(c(1013, 538, 480, 67, 1597, 321, 170, 15),
    ncol=2)
row_sum <- rowSums(Observed)
column_sum <- colSums(Observed)
I = length(row_sum); J = length(column_sum); n = sum(row_sum)

Expected <- matrix(NA, nrow = I, ncol=J)
for (i in 1:I){
  for (j in 1:J){
    Expected[i,j] = row_sum[i]*column_sum[j]/n
  }
}
Expected
         [,1]      [,2]
[1,] 1303.4468 1306.5532
[2,]  428.9888 430.0112
[3,]  324.6132 325.3868
[4,]  40.9512  41.0488
```

Then the likelihood ratio $-2\log\lambda(\mathbf{X}_n)$ and the critical value $\chi^2_{3\times 1}$ can be calculated as

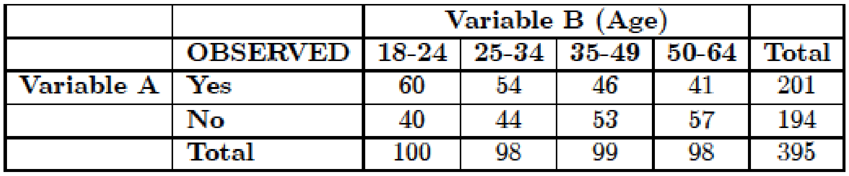
```
 sum((Observed-Expected)^2/Expected)
 [1] 366.3075

 qchisq(0.05, df = (I-1)*(J-1), lower.tail=FALSE)
 [1] 7.8147
```

```
pchisq(366.3075, df = (I-1)*(J-1), lower.tail = FALSE)
[1] 4.389154e-79
```

Since 366.3075 $>>$ 7.8147 and the p-value is almost zero, we reject H_0 with undeniable evidence that probabilities of having headaches of a certain severity are different between these two groups.

Problem 5. Is age independent of the desire to ride a bicycle? A random sample of 395 people were surveyed. Each person was asked their interest in riding a bicycle (Variable A) and their age (Variable B). The data that resulted from the survey is summarized in the following table:



		Variable B (Age)				
	OBSERVED	**18-24**	**25-34**	**35-49**	**50-64**	**Total**
Variable A	**Yes**	60	54	46	41	201
	No	40	44	53	57	194
	Total	100	98	99	98	395

Solution. We want to test

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \; i = 1, 2, \; j = 1, 2, 3, 4.$$

Following the steps of the χ^2 test of independence, we first calculate the expected counts for each cell:

```
Observed <- matrix(c(60, 54, 46, 41, 40, 44, 53, 57), ncol=4,
    byrow=TRUE)
row_sum <- rowSums(Observed)
column_sum <- colSums(Observed)
I = length(row_sum); J = length(column_sum); n = sum(row_sum)

Expected <- matrix(NA, nrow = I, ncol=J)
for (i in 1:I){
  for (j in 1:J){
    Expected[i,j] = row_sum[i]*column_sum[j]/n
  }
}
Expected
         [,1]     [,2]     [,3]     [,4]
[1,] 50.88608 49.86835 50.37722 49.86835
[2,] 49.11392 48.13165 48.62278 48.13165
```

Then the likelihood ratio $-2\log\lambda(\mathbf{X}_n)$ and the critical value $\chi^2_{1\times 3}$ can be calculated as

```
sum((Observed-Expected)^2/Expected)
[1] 8.0061
```

```
qchisq(0.05, df = (I-1)*(J-1), lower.tail=FALSE)
[1] 7.8147

pchisq(8.0061, df = (I-1)*(J-1), lower.tail = FALSE)
[1] 0.04588581
```

The observed test statistic 8.0061 is just above the critical value, we reject H_0 at the significance level 0.05 and conclude that the desire to ride a bike is dependent on age group. (You can also look at the p-value$= 0.046 < 0.05$.)