MAST90105: Lab and Workshop Problems for Week 12

The Lab and Workshop this week covers problems arising from Module 7.5 and 8.1.

1 Lab

1. Let $X \sim U(0,1)$ and consider a random sample of size 11 from X. Recall that if m is the median and $Y_1, \dots, Y_n$ are the order statistics then

$$P(Y_i < m < Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k}.$$

We will check this formula using R by computing some confidence intervals for the median of X.

a. Use the R command:

```
qbinom(c(0.025, 0.975), size = 11, prob = 0.5)
```

to compute quantiles of the binomial(11,0.5) distribution. (Ie. in the first case we find $\pi_{0.975}$ so that $P(X \leq \pi_{0.975}) \approx 0.975$. It is approximate as the distribution is discrete. However, it gives a guide to the endpoints of the confidence interval.)

b. Being careful about the correct evaluation points, use the *pbinom* command in R determine $P(Y_2 < m < Y_9)$?

```
pbinom(8, 11, 0.5) - pbinom(1, 11, 0.5)
```

c. Use the R command:

```
X <- runif(11)
```

to simulate 11 observations from X.

d. Use the *sort* command to compute the order statistics and store them in a new variable Y and hence compute Y_2 and Y_9.

e. Automate this in a function and check *f(11)* to see that it works:

```
f = function(n) {
    X = runif(n)
    Y = sort(X)
    c(Y[2], Y[9])
}
f(11)
```

Enter *f(11)* to check it works.

f. Enter the following R commands:

```
t = as.matrix(rep(11, 100))  #t needs to be a matrix for apply to work.
C = t(apply(t, 1, f))  #this is a trick to avoid programming
matplot(C, type = "l")
abline(c(0.5, 0))
sum((C[, 1] < 0.5) & (C[, 2] > 0.5))/nrow(C)
```

and hence compute the proportion of your simulated samples that contain the true mean value 1/2. Is this close to your answer in (b)? (The apply command applies the function f to each row in t and t(A) computes the transpose of the matrix A).

g. To get more precision, repeat with

```
t = as.matrix(rep(11, 1000))
C = t(apply(t, 1, f))  #this is a trick to avoid programming
matplot(C, type = "l")
abline(c(0.5, 0))
sum((C[, 1] < 0.5) & (C[, 2] > 0.5))/nrow(C)
```

2. The following 25 observations give the time in seconds between submissions of computer programs to a printer queue.
79 315 445 350 136 723 198 75 161 13 215 24 57 152 238 288 272 9 315 11 51 98 620 244 34

a. The cumulative distribution function allows us to use graphical methods to approximate the percentiles. Store the above data a vector X in R, and use the command

```
X <- c(79, 315, 445, 350, 136, 723, 198, 75, 161, 13,
    215, 24, 57, 152, 238, 288, 272, 9, 315, 11, 51,
    98, 620, 244, 34)
plot(ecdf(X))
```

to plot the cumulative distribution function. Use the plot to give approximate point estimates of $\pi_{0.25}$, m and $\pi_{0.75}$.

b. Use the command

```
qqplot(X, qexp(ppoints(100), 1/mean(X)))
```

to obtain a quantile-quantile plot of X for the exponential distribution. What do you think?

c. Use the command

```
qqplot(X, qexp(ppoints(100), 1))
```

to obtain a quantile-quantile plot of X for the exponential distribution. How does this differ from your previous plot?

d. Use the command

```
qqnorm(X)
```

to obtain a normal quantile-quantile plot of X. What do you think?

e. Give point estimates of $\pi_{0.25}$, m and $\pi_{0.75}$. (Use the command:

```
quantile(X, c(0.25, 0.5, 0.75), type = 6)
```

for the 25th percentile)

f. Find the following confidence intervals and give the confidence level.

i. (y_3, y_{10}), a confidence interval for $\pi_{0.25}$.

ii. (y_9, y_{17}), a confidence interval for the median m.

iii. (y_{16}, y_{23}), a confidence interval for $\pi_{0.75}$.

g. Find a t interval for the mean μ of the same confidence as that constructed for the median. Compare these two confidence intervals. Are the results surprising? (Your quantile plots and a histogram or stem and leaf plot may help).

3. *The data is in the file Lab11.RData in the LMS and Lab Folder.* Let p be the proportion of yellow lollies in a packet of mixed colours. It is claimed that $p = 0.2$.

a. Define a test statistic and critical region with a significance level of $\alpha = 0.05$ to test $H_0 : p = 0.2$ against $H_1 : p \neq 0.2$.

b. To perform the test, each of 20 students counted the number of yellow lollies and the total number of lollies in a 48.1 gram packet. The results were:

y	n	y	n
8.00	56.00	10.00	57.00
13.00	55.00	8.00	59.00
12.00	58.00	10.00	54.00
13.00	56.00	11.00	55.00
14.00	57.00	12.00	56.00
5.00	54.00	11.00	57.00
14.00	56.00	6.00	54.00
15.00	57.00	7.00	58.00
11.00	54.00	12.00	58.00
13.00	55.00	14.00	58.00

If each student made a test of $H_0 : p = 0.2$ at the 5% level of significance, what proportion of students rejected the null hypothesis?

c. If the null hypothesis were true, what proportion of students do you expect to reject the null hypothesis at the 5% level of significance?

d. For each of the 20 ratios in part (b) an approximate 95% confidence interval can be constructed. What proportion of these intervals contains $p = 0.2$?

e. If the 20 results are pooled do we reject $H_0 : p = 0.2$?

2 Workshop

4. Develop a function to simulate the distribution of the Wilcoxon two sample statistic based on sample sizes n, m by drawing random samples, using the R command `sample`.

```r
f <- function(x) {
    sum(sample(x[1] + x[2], size = x[2]))
}
W <- function(x, r) {
    t <- matrix(rep(x, r), byrow = TRUE, nrow = r,
        ncol = 2)
    apply(t, 1, f)
}
# This can be tested out on the sample sizes of 8
# and 8, used in the cinammon packet filling
# example as follows.  The code produces 10,000
# values of the W statistic It checks the mean and
# variance of the simulated samples It then finds
# the emprical probability that W is at least 87
w <- W(c(8, 8), 10000)
# compare to theoretical values
c(mean(w), 8 * 17/2)

## [1] 68.029 68.000

c(sd(w), sqrt(8 * 8 * 17/12))

## [1] 9.551967 9.521905

# empirical p-value compared to normal
# approximation
c(sum(w >= 87)/10000, 1 - pnorm((87 - 4 * 17)/sqrt(64 *
    17/12)))

## [1] 0.02520000 0.02299968

# try again with a larger number of repititions
w <- W(c(8, 8), 1e+06)
# compare to theoretical values
c(mean(w), 8 * 17/2)

## [1] 67.9866 68.0000

c(sd(w), sqrt(8 * 8 * 17/12))

## [1] 9.523474 9.521905

# empirical p-value
c(sum(w >= 87)/1e+06, 1 - pnorm((87 - 4 * 17)/sqrt(64 *
    17/12)))

## [1] 0.02496800 0.02299968
```

5. Vitamin B_6 is one of the vitamins in a multiple vitamin pill manufactured by a pharmaceutical company. The pills are produced with a mean of 50 milligrams of vitamin B_6 per pill. The company believes there is a deterioration of 1 milligram per month, so that after 3 months they expect that $\mu = 47$. A consumer group suspects that $\mu < 47$ after 3 months.

   a. Define a critical region to test $H_0 : \mu = 47$ against $H_1 : \mu < 47$ at the $\alpha = 0.05$ significance level based on a random sample of of size $n = 20$. ($t_{0.05}(19) = 1.729$, $t_{0.05}(20) = 1.724$, $t_{0.025}(19) = 2.093$, $t_{0.025}(20) = 2.086$).

   b. If the 20 pills yielded a mean of $\bar{x} = 46.94$ with standard deviation of $s = 0.15$, what is your conclusion?

   c. What is the approximate p-value of this test?

$$\frac{\bar{X}-M}{S/\sqrt{n}} \sim t(n-1)$$

Critical region: $\frac{\bar{X}-M}{S/\sqrt{n}} < t_{19}^{0.05}$ $\Rightarrow$ $\frac{\bar{x}-47}{S/\sqrt{20}} < -t_{0.05}(19) = -1.729$

critical region $C: \{t: t < t_{0.05}(19)\}$

$\bar{x} < \frac{1.729 \cdot S}{\sqrt{20}} + 47$

critical region: $\frac{46.94-47}{0.15/\sqrt{20}} \approx -1.789$

reject H_0

approximate:

Pvalue $= P(t^{(19)} \leq -1.789) \underset{\shortparallel}{\leq} 0.05$

$(P(t(19) \leq -1.729 \mid H_0 \text{ is true})$

$0.025 \leq$

$P(t(19) \leq -2.093 \mid H_0 \text{ istrue})$

c. What is the approximate p-value of this test?

6. Let X be the forced vital capacity (FVC) in liters for a female college student. Assume that $X \sim N(\mu, \sigma^2)$ approximately. Suppose it is known that $\mu = 3.4$ litres. A volleyball coach claims the FVC of volleyball players is greater than 3.4. She plans to test this using a random sample of size $n = 9$.

   a. Define the null hypothesis. $H_0: M = 3.4$

   b. Define the alternative hypothesis. $H_1: M > 3.4$

   c. Define a critical region for which $\alpha = 0.05$. Illustrate this on a figure. ($t_{0.025}(8) = 2.306$, $t_{0.05}(8) = 1.859$, $t_{0.01}(8) = 2.896$)

   d. Calculate the value of the test statistic if $\bar{x} = 3.556$ and $s = 0.167$.

   e. What is your conclusion?

   f. What is the approximate p-value of this test?

$C\{t: t \geq t_{0.05}^{(8)}\} \Rightarrow \{\frac{\bar{X}-M}{S/\sqrt{n}} \geq 1.859\} \Rightarrow \{\frac{\bar{X}-3.4}{S/\sqrt{9}} \geq 1.859\}$

$\frac{3.556-3.4}{} \approx 2.802 > 1.859$ reject

$0.167/\sqrt{9}$

$$\Pr(t \geq 2.896) \leq \Pr\{t \geq 2.802\} \leq \Pr(t \geq 2.306).$$

$$0.01 \leq \text{p-value} \leq 0.025$$

7. Among the data collected for the World Health Organisation air quality monitoring project is a measure of suspended particles in $\mu g/m^3$. Let X and Y equal the concentration of suspended particles in the city centres of Melbourne and Houston. Using $n = 13$ observations of X and $m = 16$ observations of Y, we shall test $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X < \mu_Y$.

   a. Define the test statistic and the critical region assuming the variances are equal. Let $\alpha = 0.05$

   b. If $\bar{x} = 72.9$, $s_x = 25.6$, $\bar{y} = 81.7$ and $s_y = 28.3$, calculate the value of the test statistic and state your conclusion.
   ($t_{0.025}(27) = 2.052$, $t_{0.05}(27) = 1.703$, ($t_{0.1}(27) = 1.314$, ($t_{0.25}(27) = 0.684$)

   c. Give limits for the p-value of this test.

a. $$\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \overset{H_0}{\sim} t(m+n-2)$$

$$S_p = \sqrt{\frac{(n-1)S_x^2+(m-1)S_y^2}{m+n-2}}.$$

$$T = \frac{\bar{x}-\bar{y}}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \overset{H_0}{\sim} t(m+n-2)$$

$$\parallel$$

$$-1.703$$

$$\frac{\bar{x}-\bar{y}}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(29-2) = t(27). \leq -t_{0.05}^{(27)} = 1.703$$

Critical region: $$\frac{\bar{x}-\bar{y}}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \leq -1.703$$

$$S_p = \sqrt{\frac{12\times 25.6^2 + 15\times 28.3^2}{27}}$$

$$\frac{72.9-81.7}{\sqrt{\frac{1}{16}+\frac{1}{13}}}. \approx -0.869. > -1.703$$

not enough evidence to reject H_0

$\times$ $p(T < -1.703|H_0) < p(T < -0.869 \mid H_0 \text{ is true}) < P(T < -0.684 \mid H_0 \text{ is true})$

$\parallel$

$0.05 < \text{p-value} < 0.25$

limit: $P[T < -1.514|16]$ 0.1] more closer.

5. Vitamin B_6 is one of the vitamins in a multiple vitamin pill manufactured by a pharmaceutical company. The pills are produced with a mean of 50 milligrams of vitamin B_6 per pill. The company believes there is a deterioration of 1 milligram per month, so that after 3 months they expect that $\mu = 47$. A consumer group suspects that $\mu < 47$ after 3 months.

   a. Define a critical region to test $H_0 : \mu = 47$ against $H_1 : \mu < 47$ at the $\alpha = 0.05$ significance level based on a random sample of of size $n = 20$. ($t_{0.05}(19) = 1.729$, $t_{0.05}(20) = 1.724$, $t_{0.025}(19) = 2.093$, $t_{0.025}(20) = 2.086$).

   b. If the 20 pills yielded a mean of $\bar{x} = 46.94$ with standard deviation of $s = 0.15$, what is your conclusion?

   c. What is the approximate p-value of this test?

6. Let X be the forced vital capacity (FVC) in liters for a female college student. Assume that $X \sim N(\mu, \sigma^2)$ approximately. Suppose it is known that $\mu = 3.4$ litres. A volleyball coach claims the FVC of volleyball players is greater than 3.4. She plans to test this using a random sample of size $n = 9$.

   a. Define the null hypothesis.

   b. Define the alternative hypothesis.

   c. Define a critical region for which $\alpha = 0.05$. lllustrate this on a figure. ($t_{0.025}(8) = 2.306$, $t_{0.05}(8) = 1.859$, $t_{0.01}(8) = 2.896$)

   d. Calculate the value of the test statistic if $\bar{x} = 3.556$ and $s = 0.167$.

   e. What is your conclusion?

   f. What is the approximate p-value of this test?

7. Among the data collected for the World Health Organisation air quality monitoring project is a measure of suspended particles in $\mu g/m^3$. Let X and Y equal the concentration of suspended particles in the city centres of Melbourne and Houston. Using $n = 13$ observations of X and $m = 16$ observations of Y, we shall test $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X < \mu_Y$.

   a. Define the test statistic and the critical region assuming the variances are equal. Let $\alpha = 0.05$

   b. If $\bar{x} = 72.9$, $s_x = 25.6$, $\bar{y} = 81.7$ and $s_y = 28.3$, calculate the value of the test statistic and state your conclusion.
   ($t_{0.025}(27) = 2.052$, $t_{0.05}(27) = 1.703$, ($t_{0.1}(27) = 1.314$, ($t_{0.25}(27) = 0.684$)

   c. Give limits for the p-value of this test.

8. It is claimed that the median weight m of certain loads of candy is 40,000 pounds.

   a. Use the following data and the Wilcoxon test statistic at an approximate significance level of $\alpha = 0.05$ to test the null hypothesis $H_0 : m = 40,000$ against

$H_1 : m < 40,000.$

41195, 39485, 41229, 36840, 38050, 40890, 38345, 34930, 39245, 31031, 40780, 38050, 30906

It may help to complete the following table. Ties are assigned the average rank.

	X	$X - m$	Rank	Sign
1	41195	1195	5	+
2	39485	−515	1	−
3	41229	1229	6	+
4	36840	−3160	~~9~~ 10	−
5	38050	−1950	8.5	−
6	40890	890	4	+
7	38345	−1655	7	−
8	34930	−5070	~~10~~ 11	−
9	39245	−755	2	−
10	31031	−8969	~~11~~ 12	−
11	40780	780	3	+
12	38050	−1950	8.5	−
13	30906	−9094	~~12~~ 13	−

$W = +5 - 1 + 6 - \cancel{9}^{-10} - 8.5 + 4$
$- 7 - 1\cancel{2}^{1} - 2 - \cancel{11}^{12} + 3$
$- 8.5 \cancel{+12} - 13$
$= 4 + 6 - 17.5 - 3$
$- 12 - 8 - 8.5 - 12$
$= 10 - 20.5 - 20$
$- 20.5$
$= 10 - 41 - 20$
$= -51 - 4 = -55$

$$\frac{W - 0}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \approx N(0,1)$$

$$\frac{-55}{\sqrt{\frac{13 \times 14 \times 27}{6}}} = \frac{-55}{\sqrt{13 \times 63}} \approx \cancel{1.702} \quad -1.922$$

≤ -1.645

enough evidence to reject

$-1.959 < -1.922 < -1.64$

$0.025 < p\text{-value} < 0.05$

Sign test

$Y = \sum_{i=1}^{n} \mathbb{I}(X_i - m < 0) = 9$ $\quad Y \sim Bin(13, \frac{1}{2})$

p-value: $P(Y \geq 9 \mid H_0 \text{ is true}) = 1 - \text{pbinom}(8, 13, 0.5)$
$= 1 - 0.866577$

$1334 > 0.05$

$= 0.1354$

no enough evidence to reject.

$H_1 : m < 40,000$.

41195, 39485, 41229, 36840, 38050, 40890, 38345, 34930, 39245, 31031, 40780, 38050, 30906

It may help to complete the following table. Ties are assigned the average rank.

	X	$X - m$	Rank	Sign
1	41195			
2	39485			
3	41229			
4	36840			
5	38050			
6	40890			
7	38345			
8	34930			
9	39245			
10	31031			
11	40780			
12	38050			
13	30906			

b. What is the approximate p-value?

```
qnorm(seq(0.9, 0.975, 0.025))

## [1] 1.281552 1.439531 1.644854 1.959964
```

c. Use the sign test to test the same hypothesis.

```
pbinom(6:12, 13, 0.5)

## [1] 0.5000000 0.7094727 0.8665771 0.9538574
## [5] 0.9887695 0.9982910 0.9998779
```

d. Compare the results of the two tests.

9. A 1-pound bag of candy-coated chocolate covered peanuts contained 224 pieces of candy coloured brown, orange, green and yellow. Test the null hypothesis that the machine filling these bags treats the four colours of candy equally likely. That is test

$$H_0 : p_B = p_O = p_G = p_Y = \frac{1}{4}.$$

The observed values were 42 brown, 64 orange, 53 green, and 65 yellow. You may select the significance level or give an appropriate p-value.
$(\chi^2_{0.025}(3) = 9.348,\ \chi^2_{0.05}(3) = 7.815,\ \chi^2_{0.10}(3) = 6.251.$

10. In a biology laboratory the mating of two red eye fruit flies yielded $n = 432$ offspring among which 254 were red-eyed, 69 were brown-eyed, 87 were scarlet-eyed, and 22 were white-eyed. Use these data to test, with $\alpha = 0.05$, the hypothesis that the ratio among the offspring would be 9:3:3:1 respectively.
$(\chi^2_{0.025}(3) = 9.348,\ \chi^2_{0.05}(3) = 7.815,\ \chi^2_{0.10}(3) = 6.251).$

11. We wish to determine if two groups of nurses distribute their time in six different categories about the same way. That is, the hypothesis under consideration is $H_0 : p_{i1} = p_{i2},\ i = 1 \ldots, 6$. To test this, nurses are observed at random throughout several days, each observation resulting in a mark in one of the six categories. The summary data is given in the following frequency table

	Category						
	1	2	3	4	5	6	Total
Group I	95	36	71	21	45	32	300
Group II	53	26	43	18	32	28	200

Use a chi-square test with $\alpha = 0.05$.

```
qchisq(seq(0.9, 0.975, 0.025), 5)

## [1]  9.236357 10.008315 11.070498 12.832502
```

12. A random sample of 1000 individuals from a rural area had 620 in favour of the election of a certain candidate, whilst a random sample of 1000 individuals from an urban area had 550 in favour of the same candidate. At the 5% level, test the hypothesis that area and opinion about the candidate are independent.

9. A 1-pound bag of candy-coated chocolate covered peanuts contained 224 pieces of candy coloured brown, orange, green and yellow. Test the null hypothesis that the machine filling these bags treats the four colours of candy equally likely. That is test

$$H_0 : p_B = p_O = p_G = p_Y = \frac{1}{4}.$$

The observed values were 42 brown, 64 orange, 53 green, and 65 yellow. You may select the significance level or give an appropriate p-value.
($\chi^2_{0.025}(3) = 9.348$, $\chi^2_{0.05}(3) = 7.815$, $\chi^2_{0.10}(3) = 6.251$.

$4 \overline{)224}$ 56.

	brown	orange	green	yellow
O	42	64	53	65
prob	0.25	0.25	0.25	0.25
E	56	56	56	56

$14^2 + 8^2 + 9 + 9^2$

$$\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2(3)$$

3^2 9 8 14^2

$$\frac{(42-56)^2}{56} + \frac{(64-56)^2}{56} + \frac{(53-56)^2}{56} + \frac{(65-56)^2}{56}$$

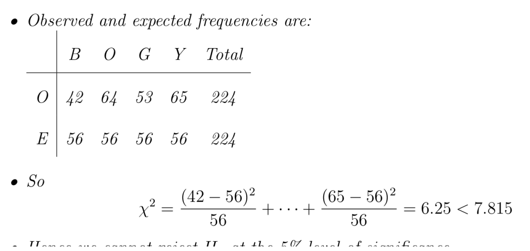
$$= \frac{14^2 + 8^2 + 9 + 9^2}{56}$$

$$= \frac{156+64+9+81}{56} = \frac{90+220}{56} = \frac{310}{56}$$

≈ 5.536

$5.536 < 6.251$

P-value > 0.1



- *Observed and expected frequencies are:*

	B	O	G	Y	Total
O	42	64	53	65	224
E	56	56	56	56	224

- *So*

$$\chi^2 = \frac{(42-56)^2}{56} + \cdots + \frac{(65-56)^2}{56} = 6.25 < 7.815$$

- *Hence we cannot reject H_0 at the 5% level of significance*