MAST90105: Lab and Workshop Problems for Week 11

The Lab and Workshop this week covers problems arising from Module 7.5 and 8.1. The problems have been assigned to groups this week.

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 $Y_1, \ldots, Y_n$ 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 $H_0 : p = 0.2$?

4. Let $X \sim \text{binomial}(1, p)$ and let $X_1, \ldots, X_{10}$ be a random sample of size 10. Consider a test of $H_0 : p = 0.5$ against $H_1 : p = 0.25$. Let $Y = \sum_{i=1}^{10} X_i$. Define the critical region as $C = \{y : y < 3.5\}$.

   a. Find the value of α the probability of a Type I error. Do not use a normal approximation. (Hint: Use pbinom).

   b. Find the value of β, the probability of a Type II error. Do not use a normal approximation.

   c. Simulate 200 observations on Y when $p = 0.5$. Find the proportion of cases when H_0 was rejected. Is this close to α?

   d. Simulate 200 observations on Y when $p = 0.25$. Find the proportion of cases when H_0 was not rejected. Is this close to β?

5. A ball is drawn from one of two bowls. Bowl A contains 100 red balls and 200 white balls; Bowl B contains 200 red balls and 100 white balls. Let p denote the probability of drawing a red ball from the bowl. Then p is unknown as we don't know which bowl is being used.To test the simple null hypothesis $H_0 : p = 1/3$ against the simple alternative that $p = 2/3$, three balls are drawn at random with replacement from the selected bowl. Let X be the number of red balls drawn. Let the critical region be $C = \{x : x = 2, 3\}$. Using R, what are the probabilities α and β respectively of Type I and Type II errors?

6. Let $Y \sim$ binomial$(100, p)$. To test $H_0 : p = 0.08$ against $H_1 : p < 0.08$, we reject H_0 and accept H_1 if and only if $Y \leq 6$. Using R or Mathematica,

   a. Determine the significance level α of the test.

   b. Find the probability of a Type II error if in fact $p = 0.04$.

7. Let p be the probability a tennis player's first serve is good. The player takes lessons to increase p. After the lessons he wishes to test the null hypothesis $H_0 : p = 0.4$ against the alternative $H_1 : p > 0.4$. Let y be the number out of $n = 25$ serves that are good, and let the critical region be defined by $C = \{y : y \geq 13\}$.

   a. Define the power function to be $K(p) = P(Y \geq 13; p)$. Graph this function for $0 < p < 1$.

   b. Find the value of $\alpha = K(0.40)$

   c. Find the value of β when $p = 0.6$, $(\beta = 1 - K(0.6))$

2 Workshop

8. Let $X_1, \ldots, X_{10}$ be a random sample of size $n = 10$ from a distribution with p.d.f. $f(x; \theta) = \exp(-(x - \theta))$, $\theta \leq x < \infty$.

   a. Show that $Y_1 = \min(X_i)$ is the maximum likelihood estimator of θ.

   b. Find the p.d.f. of Y_1 and show that $E(Y_1) = \theta + 1/10$ so that $Y_1 - 1/10$ is an unbiased estimator of θ.

   c. Compute $P(\theta \leq Y_1 \leq \theta + c)$ and use this to construct a 95% confidence interval for θ.

9. A random variable X is said to have a Pareto distribution with parameters, x_0 and β, if its cdf is

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_0}{x}\right)^\beta & x > x_0 \\ 0 & x \leq x_0 \end{cases}$$

   a. What is the pdf of X?

b. Suppose $U_1, \cdots, U_n$ are a random sample from the uniform distribution on $(0, X)$ where X is the unknown parameter. Suppose that X has a Pareto prior distribution with parameters x_0, β. Calculate the posterior distribution of X. (Hint: Consider carefully the values of the posterior pdf which are strictly positive, noting that both the joint distribution of the sample and the prior distribution pdf's have to be positive.)

c. Find a $100(1 - \alpha)$ % posterior probability interval for X.

10. If a newborn baby has a birth weight that is less than 2500 grams we say the baby has a low birth weight. The proportion of babies with birth weight is an indicator of nutrition for the mothers. In the USA approximately 7% of babies have a low birth weight. Let p be the proportion of babies born in the Sudan with low birth weight. Test the null hypothesis $H_0 : p = 0.07$ against the alternative $H_1 : p > 0.07$. If $y = 23$ babies out of a random sample of $n = 209$ babies had low birth weight, , using a suitable approximation, what is your conclusion at the significance levels

    a. $\alpha = 0.05$?

    b. $\alpha = 0.01$?

    c. Find the p-value of this test. (Recall the p-value is the probability of the observed value or something more extreme when the null hypothesis is true).

    Helpful R output

```r
qnorm(c(0.95, 0.99))

## [1] 1.644854 2.326348

pnorm(2.269)

## [1] 0.9883658
```

11. Let p_m and p_f be the respective proportions of male and female white crowned sparrows that return to their hatching site. Give the endpoints for a 95% confidence interval for $p_m - p_f$, given that 124 out of 894 males and 70 out of 700 females returned. (*The Condor*, 1992 pp.117-133.). Does this agree with the conclusion of the test of $H_0 : p_m = p_f$ against $H_1 : p_m \neq p_f$ with $\alpha = 0.05$?