MAST90105 Lab and Workshop 11 Solutions

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)
```

```
## [1] 2 9
```

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)
```

```
## [1] 0.9614258
```

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:

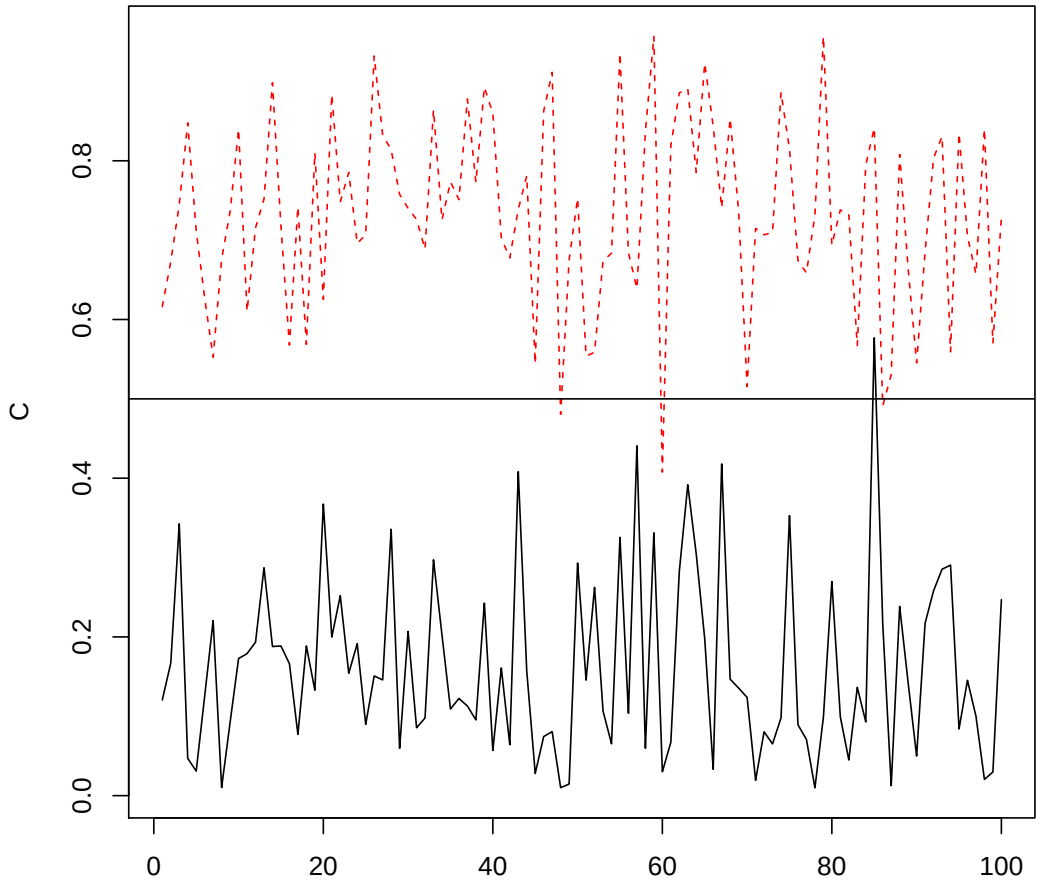
```
f = function(n) {
    X = runif(n)
    Y = sort(X)
    c(Y[2], Y[9])
}
f(11)
```

```
## [1] 0.1484990 0.7242913
```

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)

## [1] 0.96
```

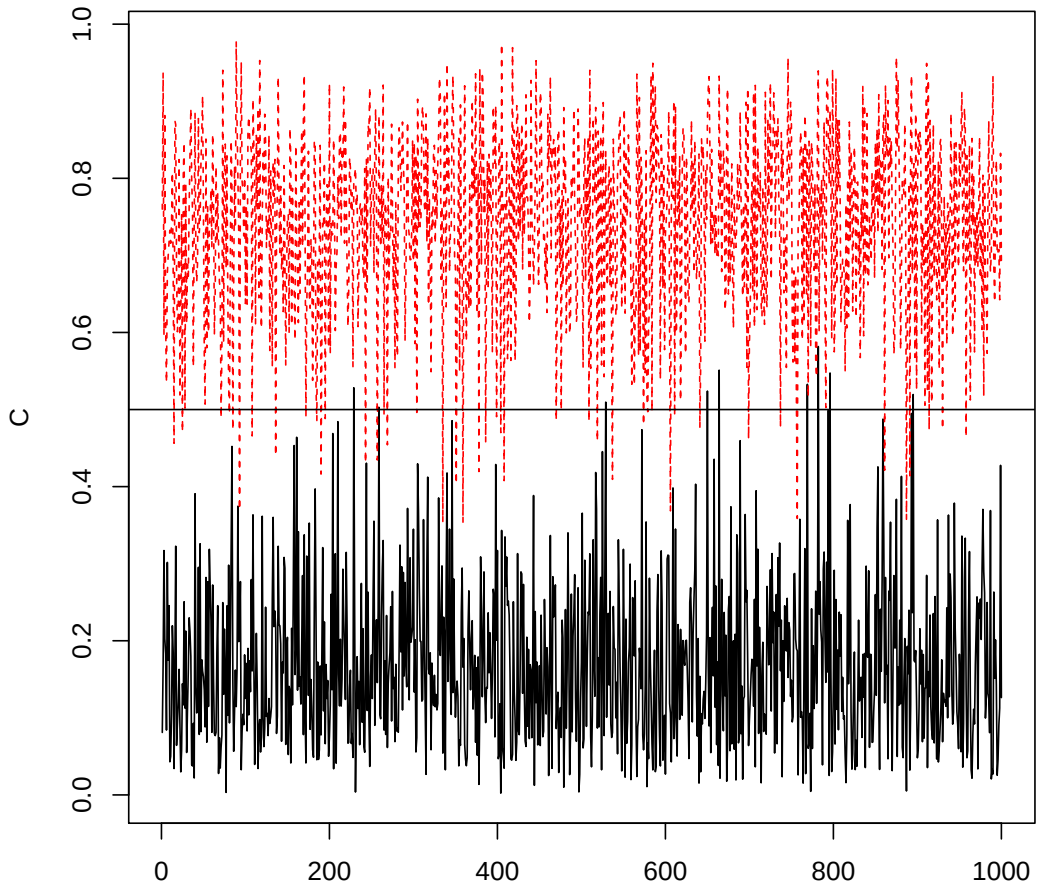
and hence compute the proportion of your simulated samples that contain the true mean value 1/2. Is this close to your answer in (b)? (The apply command

applies the function f to each row in t and t(A) computes the transpose of the matrix A).

- *This is reasonably close to the true value.*

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))
```



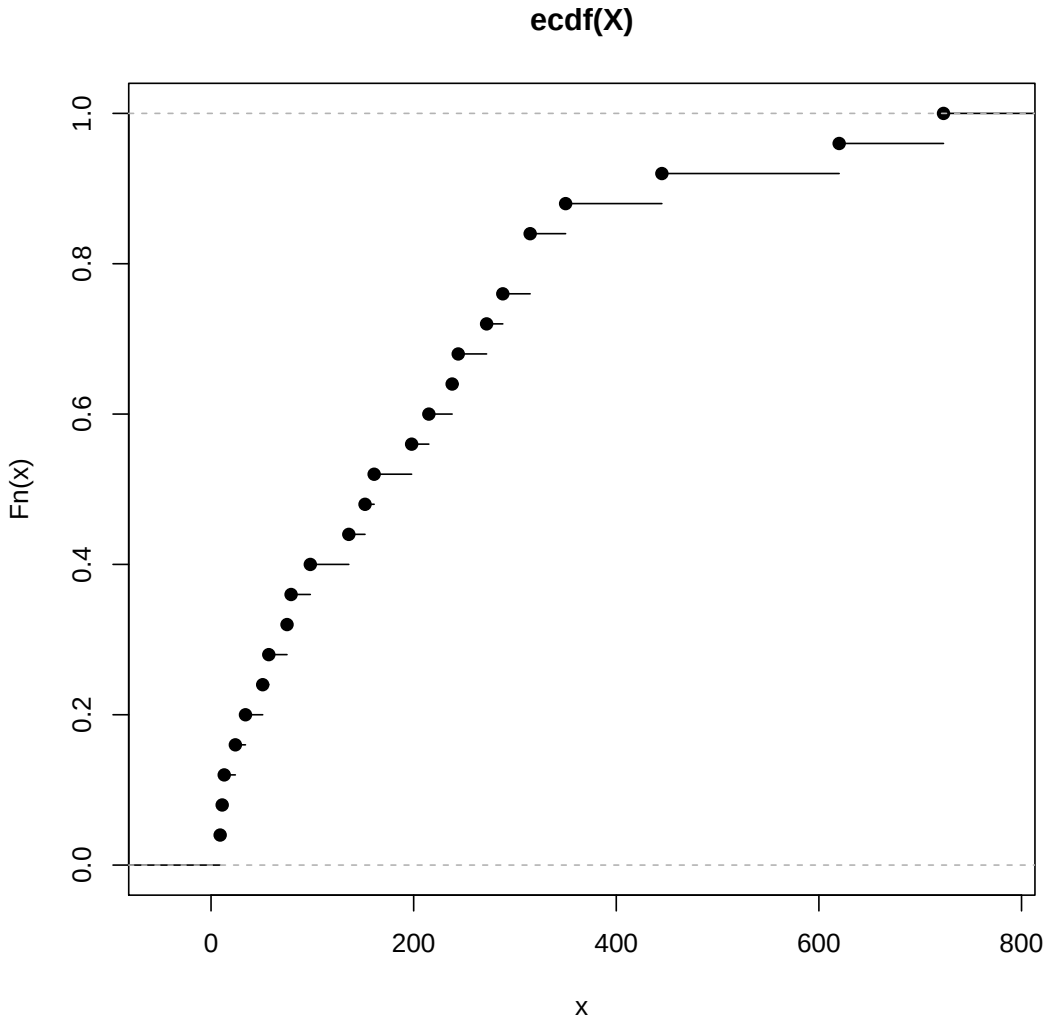

```
sum((C[, 1] < 0.5) & (C[, 2] > 0.5))/nrow(C)

## [1] 0.946
```

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))
```

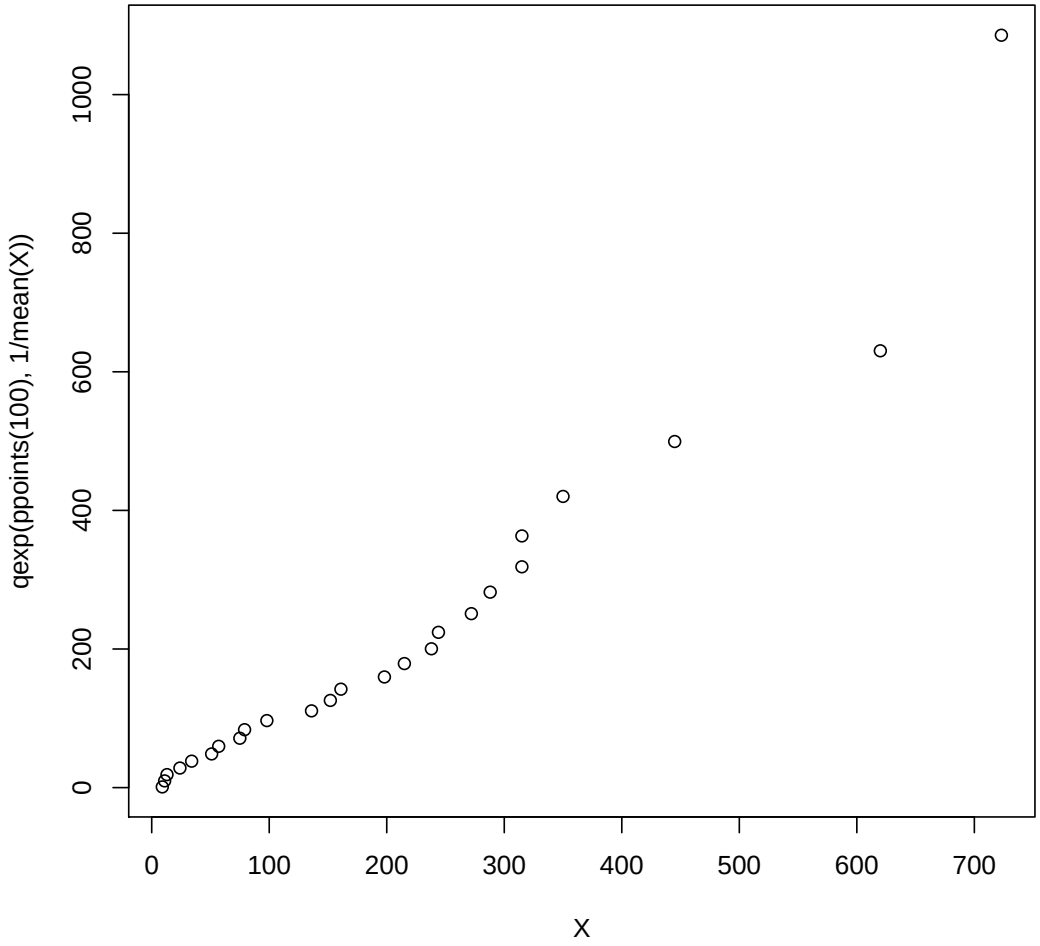


to plot the cumulative distribution function. Use the plot to give approximate point estimates of $\pi_{0.25}$, m and $\pi_{0.75}$.

- *The first quartile seems to be around 60, the median is around 160 and the third quartile around 300.*

b. Use the command

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

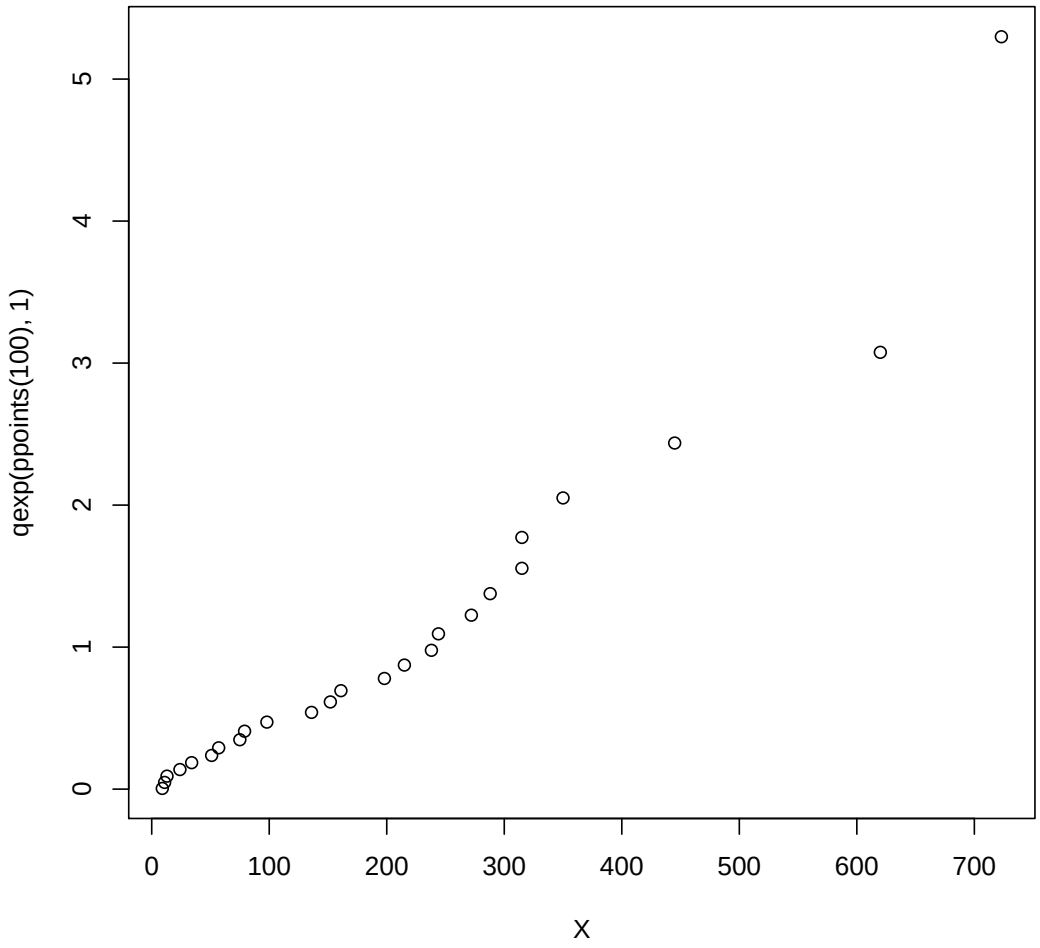


to obtain a quantile-quantile plot of X for the exponential distribution. What do you think?

- *There is a reasonable approximation to the line $y = x$ consistent with the exponential distribution with mean matching the sample, but the last points is an outlier.*

c. Use the command

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

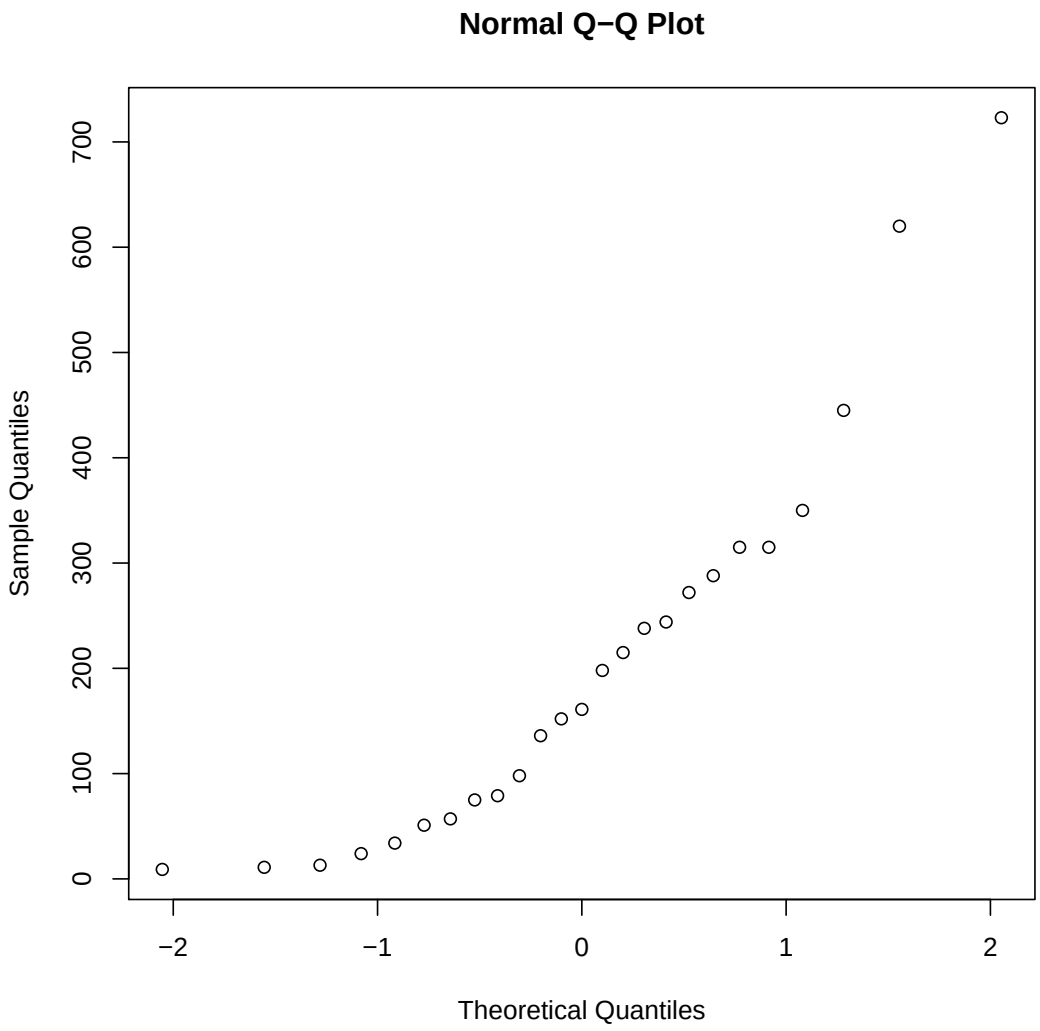


to obtain a quantile-quantile plot of X for the exponential distribution. How does this differ from your previous plot?

- *The plot looks similar - the only difference is the scale on the y-axis is changed by having mean 1.*

d. Use the command

```
qqnorm(X)
```



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

- *The normal plot does not look linear at both ends indicating a poor fit to the normal distribution.*

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)

##   25%   50%   75%
##  54.0 161.0 301.5
```

for the 25th percentile)

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

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

```
Y <- sort(X)
c(Y[3], Y[10])
## [1] 13 98
pbinom(9, 25, 0.25) - pbinom(2, 25, 0.25)
## [1] 0.8965632
```

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

```
c(Y[9], Y[17])
## [1]  79 244
pbinom(16, 25, 0.5) - pbinom(8, 25, 0.5)
## [1] 0.8922479
```

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

```
c(Y[16], Y[23])
## [1] 238 445
pbinom(22, 25, 0.75) - pbinom(15, 25, 0.75)
## [1] 0.8965632
```

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).

```
t.test(X, conf.level = 0.892)$conf.int

## [1] 142.7571 267.0829
## attr(,"conf.level")
## [1] 0.892
```

- *The confidence interval is wider for the median and in a different position - the data does not appear to come from a normal distribution, so in part this may reflect the fact that the population median is not the same as the population but the t-confidence interval is probably not appropriate.*

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 $H_1 : p \neq 0.2$.

- *Let $\hat{p}$ be the proportion yellow lollies counted. Assuming that the number of yellow lollies counted is Binomial(n, p) where n is the total number of lollies counted, we would reject H_0 if*

$$| z | = \frac{\hat{p} - 0.2}{\sqrt{0.2 \times 0.8/n}} > 1.96.$$

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?

- *0.05 - See R output.*

```
load("/Volumes/Maths & Stats/Lab-Materials/MAST90105MethodsofMathematicalStatis
phat = Data$y/Data$n
zabs = abs((phat - 0.2)/sqrt(0.8 * 0.2/Data$n))
sum(zabs > 1.96)/20

## [1] 0.05
```

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?

- *0.05*

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

- *0.9 - see R output. There is no contradiction with part b, because the width of the confidence interval is determined by the observed proportion, $\hat{p}$, not the null hypothesis value 0.2.*

```
ci <- 1.96 * sqrt(phat * (1 - phat)/Data$n)
sum((phat - ci < 0.2) * (phat + ci > 0.2))/20

## [1] 0.9
```

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

- *No - see R output. Note that this does not contradict the results in (b) because there were different total numbers counted each time.*

```
x = sum(Data$y)
N = sum(Data$n)
ptothat = x/N
(ptothat - 0.2)/sqrt(0.2 * 0.8/N)

## [1] -0.4324987
```

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).

```
pbinom(3.5, 10, 0.5)

## [1] 0.171875
```

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

```
1 - pbinom(3.5, 10, 0.25)

## [1] 0.2241249
```

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?

- *It is reasonably close - see R output.*

```
sum(rbinom(200, 10, 0.5) < 3.5)/200

## [1] 0.145
```

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?

- *It is reasonably close - see R output.*

```
sum(rbinom(200, 10, 0.25) > 3.5)/200

## [1] 0.225
```

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?

```
# alpha is prob of critical region under null
# hypothesis:
(alpha <- 1 - pbinom(1, size = 3, prob = 1/3))

## [1] 0.2592593

# beta is prob outside critical region under
# alternative hypothesis:
(beta <- pbinom(1, size = 3, prob = 2/3))

## [1] 0.2592593
```

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.

   - *There is a 30% of rejecting H_0 when it is true so the Type I error has not been well controlled - see R output.*

```
pbinom(6, 100, 0.08)

## [1] 0.303156
```

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

   - *A Type II error occurs if H_0 is not rejected but the alternative is true. There is a 10% of not rejecting H_0 if $p = 0.04$ so the type II error is quite well controlled - see R output. The R output also shows that a better decision rule would be to reject H_0 if $Y \leq 2$ having the Type I error 0.04 and Type II error 0.57. This illustrates that it is difficult to get both good Type I and Type II error when the probabilities are small.*

```r
1 - pbinom(6, 100, 0.04)
```

```
## [1] 0.1063923
```

```r
pbinom(3, 100, 0.08)
```

```
## [1] 0.0367059
```

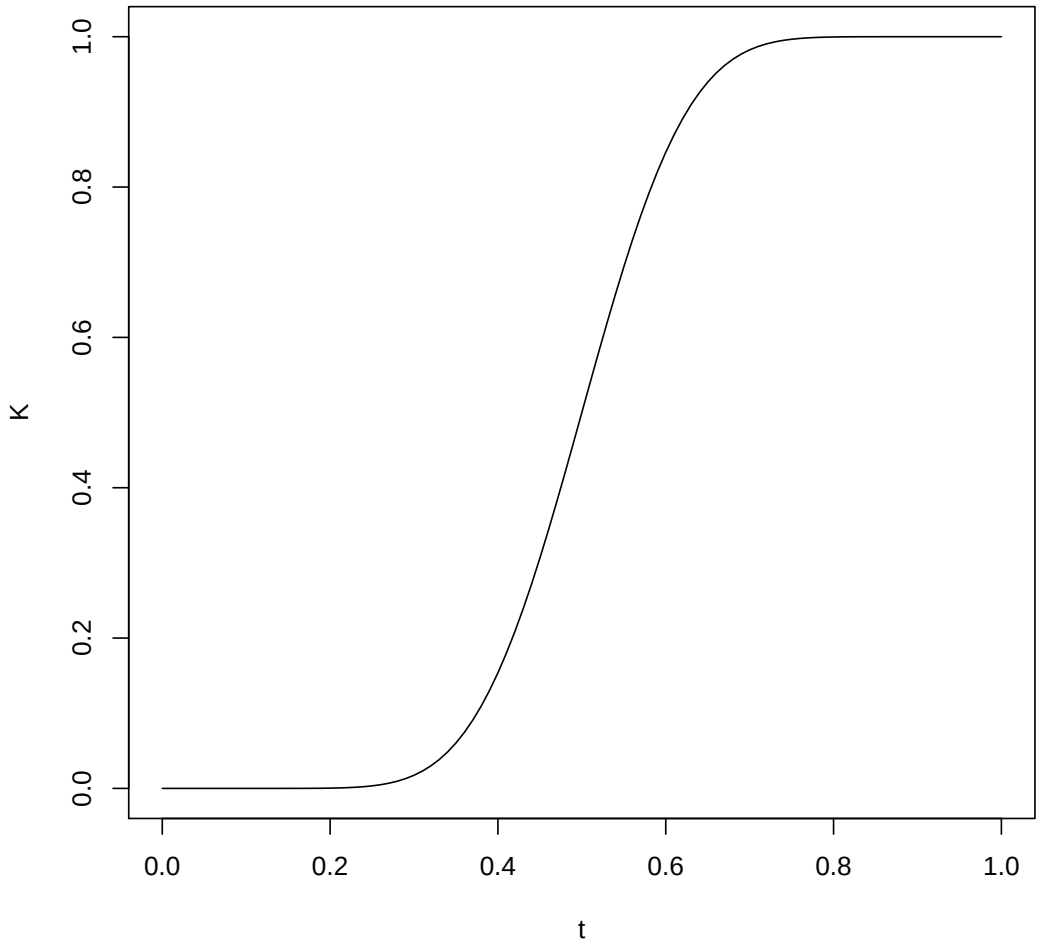
```r
1 - pbinom(3, 100, 0.04)
```

```
## [1] 0.5705244
```

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 $0 < p < 1$.

```r
Kfun <- function(p) {
    1 - pbinom(12, 25, p)
}
t <- seq(0, 1, 0.01)
K <- Kfun(t)
plot(t, K, type = "l")
```



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

```
Kfun(0.4)
```

```
## [1] 0.1537678
```

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

```
1 - Kfun(0.6)
```

```
## [1] 0.1537678
```

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.

   - *The likelihood is*

$$L(\theta) = \begin{cases} \exp[-\sum_{i=1}^{10}(x_i - \theta)] & \theta < \min(x_i) \\ 0 & \textit{otherwise} \end{cases}$$

   - *Clearly each $(x_i - \theta)$ is minimised (so that $-(x_i - \theta)$ is maximised) when θ is as large as possible. Hence $\hat{\theta} = \min(X_i) = Y_1$.*

   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.

   - *Firstly, $F(x) = \int_\theta^x \exp(-(t-\theta))dt = 1 - \exp(-(x-\theta))$, $x \geq \theta$. Then for $y \geq \theta$*

$$F_{Y_1}(y) = 1 - P(Y_1 > y) = 1 - P(\textit{all } X's > y) = 1 - (1 - F_{X_1}(y))^1 0 = 1 - \exp(-10(y-\theta))$$

   - *Differentiating gives the density g_1 for $y \geq \theta$ as*

$$g_1(y_1) = 10\exp(-10(y-\theta)) \tag{1}$$

   - *Hence, with $y = y_1 - \theta$*

$$\begin{aligned} E(Y_1) &= \int_\theta^\infty y_1 10\exp[-10(y_1-\theta)]dy_1 = \theta + \int_0^\infty y 10\exp(-10y)dy \\ &= \theta + 1/10 \end{aligned}$$

   and $Y_1 - 1/10$ is an unbiased estimator of θ as required.

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

   - *Firstly,*

$$P(\theta \leq Y_1 \leq \theta + c) = \int_\theta^{\theta+c} 10\exp[-10(y-\theta)]dy = \int_0^c 10\exp(-10y)dy = 1 - \exp(-10c)$$

   - *Hence we need to solve*

$$1 - \exp(-10c) = 19/20 \textit{ or } \exp(-10c) = 1/20 \textit{ so that } c = (1/10)\ln(20)$$

   - *Now, simple rearranging yields*

$$P(\theta \leq Y_1 \leq \theta + c) = P(Y_1 - c \leq \theta \leq Y_1)$$

   so that our 95% confidence interval is $[y_1 - 0.1\ln(20), y_1]$.

- *See Solutions to Lab and Workshop 10 Question 14.*

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?

   - *The pdf is*

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

   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.)

   - *The likelihood is*

$$f(u_1, \cdots, u_n | X = x) = \begin{cases} \frac{1}{x^n} & 0 < u_i < x, i = 1, \cdots, n \\ 0 & \textit{otherwise.} \end{cases}$$

   Hence the posterior pdf is proportional to $f(u_1, \cdots, u_n | X = x) f_X(x)$*, which is proportional to*

$$\begin{cases} \frac{1}{x^n}\frac{\beta}{x_0}\left(\frac{x_0}{x}\right)^{\beta+1} & 0 < u_i < x, i = 1, \cdots, n \quad \& \quad x > x_0, \\ 0 & \textit{otherwise.} \end{cases}$$

   This expression is proportional to a Pareto distribution with parameters $x_n = \max(u_1, \cdots, u_n, x_0), \beta_n = \beta + n$*. Hence this is the posterior pdf.*

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

   - *A* $100(1-\alpha)$ *% posterior probability interval for* X *is given by the posterior* $\alpha 2$ *quantile to the posterior* $1 - \alpha^2$ *quantile.*
   - *The* $\alpha/2$ *quantile,* $\pi_{\alpha/2}$ *for a Pareto* x_0, β *distribution is obtained by solving* $1 - \left(\frac{x_0}{x}\right)^\beta = \frac{\alpha}{2}$ *for* x*, giving* $\pi_{\alpha/2} = x_0\left(\frac{2}{2-\alpha}\right)^{1/\beta}$.
   - *Similarly, The* $1 - \alpha/2$ *quantile,* $\pi_{1-\alpha/2}$*, for a Pareto* x_0, β *distribution is given by* $\pi_{1-\alpha/2} = x_0\left(\frac{2}{\alpha}\right)^{1/\beta}$.
   - *Applying these to the posterior distribution calculated in part (b) gives a* $100(1-\alpha)$ *% posterior probability interval for* X *of* $\left(x_n\left(\frac{2}{2-\alpha}\right)^{1/\beta_n}, x_n\left(\frac{2}{\alpha}\right)^{1/\beta_n}\right)$.

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

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

## [1] 1.644854 2.326348

pnorm(2.269)

## [1] 0.9883658
```

- *Under* H_0, $E(Y) = 14.63$, $\text{Var}(Y) = 13.606 = 3.689^2$. *Hence*

$$z = \frac{23 - 14.63}{3.689} = 2.269$$

    a. $z > 1.645$ *so reject* H_0 *at 5% level of significance.*

    b. $z < 2.326$ *so don't reject* H_0 *at the 1% level of significance.*

    c. *p-value is* $P(Z \geq 2.269) = 1 - \Phi(2.269) = 1 - 0.9883 = 0.0117$

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 $\alpha = 0.05$?

- $\hat{p}_m = \frac{124}{894} = 0.1387$, $\hat{p}_f = \frac{70}{700} = 0.1$, *so that*

$$\hat{p}_m - \hat{p}_f \pm 1.96\sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{n_m} + \frac{\hat{p}_f(1-\hat{p}_f)}{n_f}} = [0.007, 0.07].$$

$\hat{p} = 194/1594 = 0.1217$. *Reject* H_0 *if* $\mid z \mid > 1.96$ *and*

$$\mid z \mid = \frac{\mid \hat{p}_m - \hat{p}_f \mid}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_m} + \frac{1}{n_f}\right)}} = 2.345 > 1.96$$

so we reject H_0.