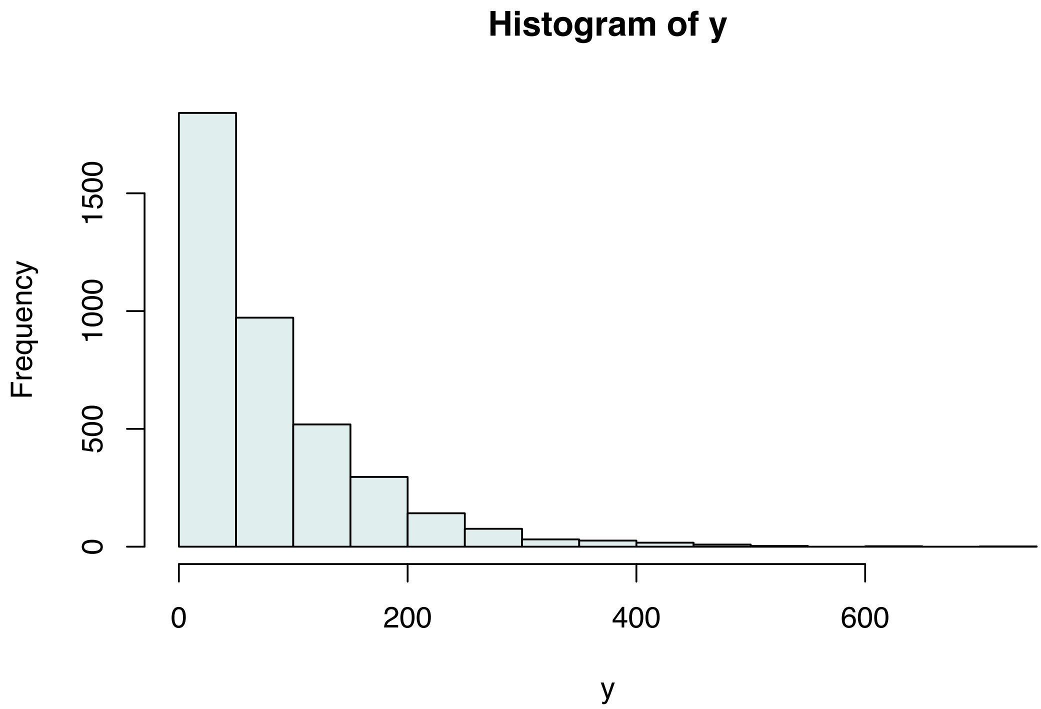
Tutorial 7 Solutions

STAT 3013/4027/8027

1. [based on Q 3.17]. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{gamma}(a, b)$ where $E[X] = ab$. Suppose that the data on on gamma-rays (measuring the interarrival times of 3,935 photons (units in seconds) can be modeled by a gamma distribution.

a. Let's load in the data and examine a histogram of the data.

```
data <- read.table("gamma-arrivals.txt")
y <- data$V1
hist(y, col="azure2")
```



Based on the histogram, an gamma(a,b) distribution does not seem like an unreasonable model for the data.

$$f(y) = \frac{1}{\Gamma(a)b^a} y^{a-1} exp(-y/b)$$

b. Let's first determine the **Method of Moments** for a and b. Our system of equations is:

$$
\begin{aligned}
E[Y] = ab &= \frac{1}{n}\sum_{i=1}^{n} y_i = m_1 \\
E[Y^2] = ab^2 + a^2b^2 &= \frac{1}{n}\sum_{i=1}^{n} y_i^2 = m_2
\end{aligned}
$$

Solving this system of equations we have:

$$
\tilde{a} = \frac{m_1^2}{m_2 - m_1^2} \quad \tilde{b} = \frac{m_2 - m_1^2}{m_1}
$$

```
n <- length(y)
m1 <- mean(y)
m2 <- sum(y^2)/n

a.mom <- m1^2/(m2-m1^2)
b.mom <- (m2-m1^2)/m1

a.mom
```

```
## [1] 1.012352
```

```
b.mom
```

```
## [1] 78.95989
```

c. Now let's consider the **Maximum Likelihood** estimators.

$$
L(a,b) = \prod_{i=1}^{n} \frac{1}{\Gamma(a)\ b^a} y_i^{a-1} exp\left(-y_i/b\right)
$$

Here we have the log-likelihood:

$$
\ell(a,b) = -n\ log\left(\Gamma(a)\right) - n\ a\ log(b) + (a-1)\sum_{i=1}^{n} log(y_i) - \sum_{i=1}^{n} y_i/b
$$

$$
\begin{aligned}
\frac{\partial \ell(a,b)}{\partial a} &= -n\psi(a) - nlog(b) + \sum_{i=1}^{n} log(y_i) && (1) \\
\frac{\partial^2 \ell(a,b)}{\partial a^2} &= -n\psi'(a) && (2)
\end{aligned}
$$

In Eqn (1), let's substitute in for $b = \sum_{i=1}^{n} y_i/(na)$. So we have an equation which only has a:

$$
\begin{aligned}
\frac{\partial \ell(a,b)}{\partial a} &= -n\psi(a) - nlog\left(\sum_{i=1}^{n} y_i/(na)\right) + \sum_{i=1}^{n} log(y_i) \qquad (3)\\
\frac{\partial^2 \ell(a,b)}{\partial a^2} &= -n\psi'(a) \qquad (4)
\end{aligned}
$$

- Where $\psi(a) =$ `digamma`(a) and $\psi'(a) =$ `trigamma`(a).

```
## Let's find the MLE of a using the N-R Approach.
## Then we can solve for b analytically
## Write some functions for U and H
U <- function(a){
  n <- length(y)
  out <-  -n* digamma(a) - n*log(sum(y)/(n*a)) + sum(log(y))
  return(out)
  }

H <- function(a){
    n <- length(y)
    out <- -n*trigamma(a)
    return(out)
    }

## Starting values - use MoM estimator
a <- mean(y)^2/( (n-1)*var(y)/n)

## set a stopping point
eps <- 1e-07
check <- 10

## Save the results.
out <- a

## Run the algorithm
while(check > eps){
a.new <- a - U(a)/H(a)
check <- sum(abs(a-a.new))
a <- a.new
out <- rbind(out, t(a))
    }

a.mle <- a
b.mle <- sum(y)/(n*a.mle)

##
a.mle
```

```
## [1] 1.026332
```

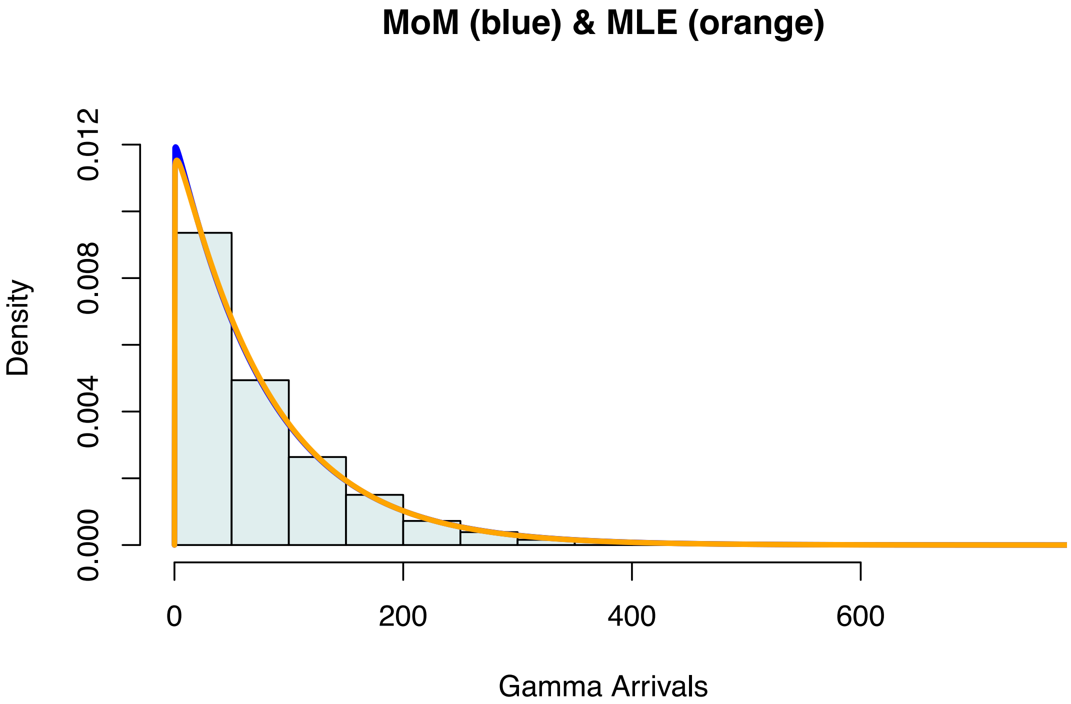
```
b.mle
```

```
## [1] 77.8844
```

The ML and MoM estimates are very similar.

d. Let's plot the two fitted densities on top of the histogram.

```
hist(y, col="azure2", freq=FALSE, ylim=c(0, 0.013), xlab="Gamma Arrivals",
     main="MoM (blue) & MLE (orange)")
x <- seq(0, 800, by=0.5)
lines(x, dgamma(x, shape=a.mom, scale=b.mom), lwd=3, col="blue")
lines(x, dgamma(x, shape=a.mle, scale=b.mle), lwd=3, col="orange")
```



We can see the results are very similar as the two lines essentiall overlap, except for a slight difference near 0.

2. Consider a random sample of twins pairs. Twin pairs may be identical or fraternal. Let u of these pairs consist of male pairs, v consist of female pairs, and w consist of opposite sex pairs. A simple model for these data is based on a Bernoulli distribution for each pair dictating whether it consists of identical or fraternal twins. Suppose that identical twins occur with probability p and fraternal twins with probability $1-p$. Once the decision is made as to whether the twins are identical or not, then sexes are are assigned to the twins. If the twins are identical, then one assignment of sex is made. If the twins are non-identical, then two independent assignments of sex are made. Suppose boys are chosen with probablity q and girls with probability $1-q$.

a-b. Write out the likelihood for these data. Formulate a "missing data" model and specify both steps of the EM algorithm to obtain the MLEs for p and q.

$$
\begin{aligned}
u &= \text{number of male pairs} = 22 = u_I + u_F \\
v &= \text{number of female pairs} = 21 = v_I + v_F \\
w &= \text{number of male/female pairs} = 25 \\
\\
p &= \text{probability of identical (I) twins} \\
(1-p) &= \text{probability of fraternal (F) twins} \\
\\
q &= \text{probability of a boy} \\
(1-q) &= \text{probability of a girl}
\end{aligned}
$$

- Let's determine the probabilities for u_I, u_F, v_I, v_F, w:

$$
\begin{aligned}
P(u_I) &= P(I)P(\text{Boy, Boy}|I) = pq \\
P(u_F) &= P(F)P(\text{Boy, Boy}|F) = (1-p)q^2 \\
P(v_I) &= P(I)P(\text{Girl, Girl}|I) = p(1-q) \\
P(v_F) &= P(F)P(\text{Girl, Girl}|F) = (1-p)(1-q)^2 \\
P(w) &= P(F)P(\text{Girl/Boy, Boy/Girl}|F) = 2(1-p)(1-q)q \\
& \text{Note: the 2 is there as you can have G/B and B/G}
\end{aligned}
$$

- Here we have a multinomial distribution, so we can write the likelihood as follows:

$$
L(\theta) = \frac{n!}{u_I!\ v!\ w!}\ [p_u]^u\ [p_v]^v\ [p_w]^w
$$

- We can re-write this in terms of missing data by separating out u_I, u_F and v_I, v_F (even though we only observe u and v!

$$
\begin{aligned}
L(\theta) &= \left(\frac{n!}{u_I!\ u_F!\ v_I!\ v_F!\ w!}\right)\ [pq]^{u_I}\ [(1-p)q^2]^{u_F}\ [p(1-q)]^{v_I}\ [(1-p)(1-q)^2]^{v_F}\ [2(1-p)(1-q)q]^w \\
\ell(\theta) &= log\left(\frac{n!}{u_I!\ u_F!\ v_I!\ v_F!\ w!}\right) + u_I\ log(pq) + u_F\ log((1-p)q^2) + v_I\ log(p(1-q)) \\
&= +v_F\ log((1-p)(1-q)^2) + w\ log(2(1-p)(1-q)q)
\end{aligned}
$$

$$
\begin{aligned}
\ell(\theta) &= log\left(\frac{n!}{u_I!\ u_F!\ v_I!\ v_F!\ w!}\right) + (u_I + v_I)\ log(p) + (u_F + v_F + w)\ log(1-p) \\
& +(u_I + 2u_F + w)\ log(q) + (v_I + 2v_F + w)\ log(1-q) + w\ log(2)
\end{aligned}
$$

- Let's calculate the **E-Step** for the **E-M Algorithm**:

$$\begin{aligned}
u_I* &= E[u_I|u] \\
&= u\ P(u_I|u) \\
&= u\ P(\text{Identical}|\text{Boy, Boy}) \\
&= u\ \frac{P(\text{Boy, Boy}|\text{Identical})P(\text{Identical})}{P(\text{Boy, Boy})} \\
&= u\ \frac{p^s q^s}{p^s q^s + (1-p^s)(q^s)^2} \\
&= u\ \frac{p^s}{p^s + (1-p^s)(q^s)}
\end{aligned}$$

- Where s is the current value in the iterative process.

$$\begin{aligned}
u_F* &= E[u_F] \\
&= u\ P(\text{Fraternal}|\text{Boy, Boy}) \\
&= u\ \frac{P(\text{Boy, Boy}|\text{Fraternal})P(\text{Fraternal})}{P(\text{Boy, Boy})} \\
&= u\ \frac{(1-p^s)(q^s)^2}{(1-p^s)(q^s)^2 + (p^s)(q^s)} \\
&= u\ \frac{(1-p^s)(q^s)}{p^s + (1-p^s)(q^s)}
\end{aligned}$$

$$\begin{aligned}
v_I* &= E[v_I] \\
&= v\ P(\text{Identical}|\text{Girl, Girl}) \\
&= v\ \frac{P(\text{Girl, Girl}|\text{Identical})P(\text{Identical})}{P(\text{Girl, Girl})} \\
&= v\ \frac{p^s}{p^s + (1-p^s)(1-q^s)}
\end{aligned}$$

$$\begin{aligned}
v_F* &= E[v_F] \\
&= v\ P(\text{Fraternal}|\text{Girl, Girl}) \\
&= v\ \frac{P(\text{Girl, Girl}|\text{Fraternal})P(\text{Fraternal})}{P(\text{Girl, Girl})} \\
&= v\ \frac{(1-p^s)(1-q^s)}{p^s + (1-p^s)(1-q^s)}
\end{aligned}$$

- Now let's get the **M-Step** for the **E-M Algorithm**:

$$\frac{\partial \ell}{\partial p} = \frac{u_I^s + v_I^s}{p} - \frac{u_F^s + v_F^s + w}{1-p} = 0$$

$$\frac{\partial \ell}{\partial q} = \frac{u_I^s + 2u_F^s + w}{q} - \frac{v_I^s + 2v_F^s + w}{1-q} = 0$$

$$\begin{aligned}
\hat{p} &= \frac{u_I^s + v_I^s}{u_I^s + v_I^s + u_F^s + 2v_F^s + w} \\
\hat{q} &= \frac{u_I^s + 2u_F^s + w}{u_I^s + 2u_F^s + v_I^s + 2v_F^s + 2w}
\end{aligned}$$

c. Carry out the algorithm for $u = 22$, $v = 21$, and $w = 25$.

```
## observed data
u <- 22
v <- 21
w <- 25


##
p.store <- NULL
q.store <- NULL

## starting value
p <- 0.5
q <- 0.5

##
check <- 10
eps <- 1e-10

##
while(check>eps){

  ## E-step
  u.I <- u* (p/ (p + (1-p)*q))
  u.F <- u* ( ((1-p)*q)/ ((p + (1-p)*q)))

  v.I <- v * (p / (p + (1-p)*(1-q) ))
  v.F <- v * ( ( (1-p)*(1-q))/ ((p + (1-p)*q)))

  ## M-step
  p.new <- (u.I+v.I)/(u+v+w)
  q.new <- (u.I + 2*u.F + w)/ (u.I + 2*u.F + v.I + 2*v.F + 2*w)
```

```
  ##
  check <- abs((p-p.new)) +  abs((q-q.new))

  p.store <- c(p.new, p.store)
  q.store <- c(q.new, q.store)

  p <- p.new
  q <- q.new

}

## MLEs
p
```

```
## [1] 0.2645349
```

```
q
```

```
## [1] 0.5099922
```

3. SI question 4.1 (a, b). **See the handwritten solutions.**

GJJ Q 4.1(a) $X_1, \dots, X_n \overset{iid}{\sim} \text{Poisson}(\theta)$

$$H_0: \theta = \theta_0 \;;\; H_1: \theta = \theta_1 \;;\; \theta_1 > \theta_0$$

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\Sigma x_i}}{\Pi x_i!}$$

$$\Rightarrow \quad \lambda = \frac{L(\theta_0 \mid \underline{x})}{L(\theta_1 \mid \underline{x})} = \frac{e^{-n\theta_0}\theta_0^{\Sigma x_i}}{e^{-n\theta_1}\theta_1^{\Sigma x_i}}$$

$$= \exp(-n\theta_0 + n\theta_1)\left(\frac{\theta_0}{\theta_1}\right)^{\Sigma x_i}$$

$$= \exp(n(\theta_1 - \theta_0))\left(\frac{\theta_0}{\theta_1}\right)^{\Sigma x_i}$$

$$C = \left\{ \exp(n(\theta_1-\theta_0))\left(\frac{\theta_0}{\theta_1}\right)^{\Sigma x_i} < k \right\}$$

Reject H_0 for small values

$$= \left\{ n\overbrace{(\theta_1 - \theta_0)}^{\text{positive}} + \Sigma x_i \log\left(\frac{\theta_0}{\theta_1}\right) < \log(k) \right\}$$

$$= \left\{ \Sigma x_i \log\left(\frac{\theta_0}{\theta_1}\right) < \log(k) - n(\theta_1 - \theta_0) \right\}$$

$$= \left\{ \Sigma x_i < [\log(k) - n(\theta_1 - \theta_0)] / \log\left(\frac{\theta_0}{\theta_1}\right) \right\}$$

$$= \left\{ \Sigma x_i < k^* \right\}$$

- Under $H_0: \theta = \theta_0 \Rightarrow X \sim \text{Poisson}(\theta_0)$

$\Sigma x_i \sim \text{Poisson}(n\theta_0)$

use MGF to show this

$$P(C) = P(\Sigma X_i < k^*) = \alpha = 0.05$$

- Sps $\theta_0 = 2$; $n = 10$

$\Rightarrow$ In R : > qpois(0.05, 20) = 13 = k^*

Due to the discreteness of the Poisson distribution

$\Rightarrow$ ppois(13, 20) = 0.066

$\therefore$ We have an ump test for $\alpha = 0.066$!

This is an exact Result!

- We can also rely on the CLT (Asymptotic Solution):

$$P(C) = P(\bar{X} < k^{**})$$

$$= P\left(\frac{\bar{X} - \theta_0}{\sqrt{\theta_0/n}} < k^{**}\right) = P(Z < k^{**}) = \alpha = 0.05$$

$\therefore$ k^{**} = qnorm(0.05) = -1.64

GJJ 4.1(b)

$X_1, \dots, X_n \overset{iid}{\sim}$ exponential(θ); $E(X) = \theta$.

$H_0: \theta = \theta_0$; $H_1: \theta = \theta_1$; $\theta_1 > \theta_0$.

$$L(\theta) = \prod_{i=1}^{n} \left(\frac{1}{\theta}\right) \exp(-x_i/\theta)$$

$$= \frac{1}{\theta^n} \exp(-\Sigma x_i/\theta)$$

$$\Rightarrow \quad \lambda = \frac{L(\theta_0 \mid \underset{\sim}{x})}{L(\theta_1 \mid \underset{\sim}{x})} = \frac{\frac{1}{\theta_0^n} \exp(-\Sigma x_i/\theta_0)}{\frac{1}{\theta_1^n} \exp(-\Sigma x_i/\theta_1)}$$

$$= \left(\frac{\theta_1}{\theta_0}\right)^n \exp\left(-\Sigma x_i/\theta_0 + \Sigma x_i/\theta_1\right)$$

$$= \left(\frac{\theta_1}{\theta_0}\right)^n \exp\left(\Sigma x_i \left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)\right)$$

$$\Rightarrow \quad C = \left\{ \left(\frac{\theta_1}{\theta_0}\right)^n \exp\left(\Sigma x_i\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)\right) < k \right\}$$

$$= \left\{ \exp\left(\Sigma x_i\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)\right) < k \Big/ \left(\frac{\theta_1}{\theta_0}\right)^n \right\}$$

$$= \left\{ \Sigma x_i \underbrace{\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)}_{\text{negative}} < \log\left(k \Big/ \left(\frac{\theta_1}{\theta_0}\right)^n\right) \right\}$$

$$= \left\{ \Sigma x_i \geq k^* \right\}$$

use MGF to show   under H_0

If $X \sim \exp(\theta) \Rightarrow \Sigma x_i \sim \text{gamma}(n, \theta_0)$

$$\Rightarrow \quad P(C) = P(\Sigma x_i \geq k^*) = (1 - P(\Sigma x_i < k^*)) = \alpha$$

$$= P(\Sigma x_i < k^*) = 1 - \alpha = 1 - 0.05 = 0.95$$

- Suppose $n = 10$, $\theta_0 = 2$:

  > qgamma(0.95, 10, 2) = 7.853

- The above test is an exact UMP test for $\alpha = 0.05$.

- We can also use the CLT (asymptotic):

$$P(C) = P(\Sigma x_i \geq k^*)$$

$$= P(\bar{x} \geq k^{**}) = P\left(\frac{\bar{x} - \theta_0}{\theta_0/\sqrt{n}} \geq k^{**}\right)$$

$$= P(Z \geq k^{**}) = 1 - P(Z \leq k^{**})$$

$$\Rightarrow P(Z \leq k^{**}) = 1 - \alpha = 1 - 0.05 = 0.95$$

$$> \text{qnorm}(0.95) = 1.645 = k^{**}$$