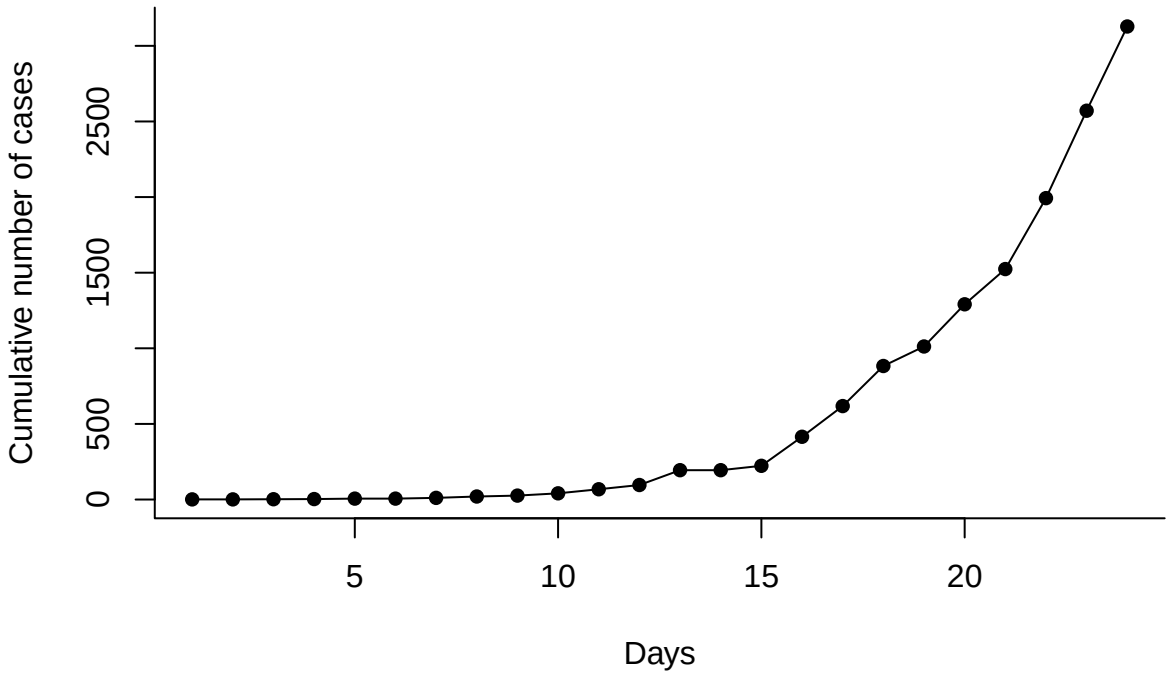
Problem Set 2

Hanyu Wang, Zhiyu Wang

1. What is the growth rate of the epidemic over the first 24 days (through March 31, 2020)? Show how you calculated it.

```r
# Calculate the cumulative number of cases up to time t for t = 1 to 24 days.
CumCasesE <- cumsum(covidCT$newcasesCT[1:24])

# Make a plot
plot(CumCasesE,
  type = "o", bty = "l", pch = 16,
  ylab = "Cumulative number of cases", xlab = "Days"
)
```



Since $R_0 > 1$ (epidemic spread), the number of new infections is expected to grow exponentially initially:

$$I(t) = I(0)\ e^{rt}, \text{ where } r \text{ is the intrinsic growth rate}$$

We can estimate r using Poisson regression:

$$log(I_t) = log(I_0) + rt$$

```r
# Create a variable, t (time)
t <- 1:24

# Fit a linear regression to the log of the cumulative number of cases
mod1 <- glm(log(CumCasesE) ~ t, family = poisson(link = "log"))  # you could also use lm(log(CumCasesE)
```

```
## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.693147
## Warning in dpois(y, mu, log = TRUE): non-integer x = 1.098612
## Warning in dpois(y, mu, log = TRUE): non-integer x = 1.791759
## Warning in dpois(y, mu, log = TRUE): non-integer x = 1.791759
## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.397895
## Warning in dpois(y, mu, log = TRUE): non-integer x = 2.995732
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.258097
## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.713572
## Warning in dpois(y, mu, log = TRUE): non-integer x = 4.219508
## Warning in dpois(y, mu, log = TRUE): non-integer x = 4.564348
## Warning in dpois(y, mu, log = TRUE): non-integer x = 5.267858
## Warning in dpois(y, mu, log = TRUE): non-integer x = 5.267858
## Warning in dpois(y, mu, log = TRUE): non-integer x = 5.407172
## Warning in dpois(y, mu, log = TRUE): non-integer x = 6.028279
## Warning in dpois(y, mu, log = TRUE): non-integer x = 6.426488
## Warning in dpois(y, mu, log = TRUE): non-integer x = 6.783325
## Warning in dpois(y, mu, log = TRUE): non-integer x = 6.919684
## Warning in dpois(y, mu, log = TRUE): non-integer x = 7.163172
## Warning in dpois(y, mu, log = TRUE): non-integer x = 7.329094
## Warning in dpois(y, mu, log = TRUE): non-integer x = 7.597396
## Warning in dpois(y, mu, log = TRUE): non-integer x = 7.852050
## Warning in dpois(y, mu, log = TRUE): non-integer x = 8.048149
```

```r
# Result of the linear regression model
summary(mod1)
```

```
##
## Call:
## glm(formula = log(CumCasesE) ~ t, family = poisson(link = "log"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.19897    0.27396   0.726    0.468
## t            0.08880    0.01557   5.705 1.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 45.8615  on 23  degrees of freedom
## Residual deviance:  9.6694  on 22  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

```
r1 <- coef(mod1)["t"]         # slope (growth rate)
r1
```

```
##          t
## 0.08879803
```

Answer: The growth rate of the epidemic over the first 24 days (through March 31, 2020) is 0.08879803.

2. Show how Lipsitch et al (2003) derived their expression for calculating R0 from the growth rate of the epidemic (which they call λ) by solving for the maximum eigenvalue

$$\begin{vmatrix} -\frac{1}{L} - \lambda & \frac{R_0}{D} \\ \frac{1}{L} & -\frac{1}{D} - \lambda \end{vmatrix} = 0 \longrightarrow (-\frac{1}{L} - \lambda)(-\frac{1}{D} - \lambda) - (\frac{R_0}{D})(\frac{1}{L}) = 0$$

Remember, they define V = D+L as the serial interval and f =L/V as the ratio of the mean latent period to serial interval (see Ref/Note 7 in Lipsitch et al (2003)).

$$(\frac{1}{L} + \lambda)(\frac{1}{D} + \lambda) - \frac{R_0}{DL} = 0 \longrightarrow \lambda^2 + \lambda(\frac{1}{D} + \frac{1}{L}) + \frac{1 - R_0}{DL} = 0$$

Since $V = D + L$, $f = \frac{L}{V}$, we have $L = Vf$, $D = V - L = V - fV = V(1 - f)$:

$$\frac{1}{D} + \frac{1}{L} = \frac{1}{V(1-f)} + \frac{1}{Vf} = \frac{1}{V}(\frac{1}{1-f} + \frac{1}{f}) = \frac{1}{V} \cdot \frac{1}{f(1-f)}$$

$$\frac{1 - R_0}{DL} = \frac{1 - R_0}{V(1-f) \cdot Vf} = \frac{1 - R_0}{V^2 f(1-f)}$$

Therefore, $(\frac{1}{L} + \lambda)(\frac{1}{D} + \lambda) - \frac{R_0}{DL} = 0 \longrightarrow \lambda^2 + (\frac{1}{Vf(1-f)})\,\lambda + \frac{1-R_0}{V^2 f(1-f)} = 0$

$$\lambda^2 V^2 f(1-f) + V\lambda = R_0 - 1$$

Answer: $R_0 = \lambda^2(1-f)fV^2 + \lambda V + 1$, where V is the serial interval and f is the ratio of the latent period to the serial interval.

3. Calculate R_0 from the growth rate of the epidemic assuming the mean latent period is 2.8 days and the mean serial interval is 6.3 days.

```
# Let's set f and v
f <- 2.8 / 6.3 # ratio of the latent period to the serial interval
v <- 6.3 # duration of the serial interval

# The equation from Lipsitch et al (2003) is:
R0 <- (r1^2) * (1 - f) * f * (v^2) + r1 * v + 1
R0
```

```
##        t
## 1.636701
```

Answer: R_0 is 1.636701.

4. Let's assume only 1 out of every 10 cases of COVID-19 was reported during the early stages of the pandemic. Would this underreporting bias your estimate of R0 above? Why or why not?

Answer: No, this underreporting would *NOT* bias the estimate of R_0, since the proportion of cases reported remains constant over time (only 1 out of every 10 cases of COVID-19 was reported = always 10% of cases are reported).

The growth rate r is estimated from the exponential growth pattern of cases, then the observed cases are simply a constant fraction of the true cases:

$$\text{Proportion of cases reported} = \text{a constant fraction} = c = 0.1$$

$$I_{obs}(t) = c \cdot I_{true}(t)$$

$$I_{obs}(t) = 0.1 \cdot I_{true}(t)$$

When we estimate r using Poisson regression and take logarithms:

$$\log(I_t) = \log(I_0) + rt$$

Observed cases:

$$\log(I_{obs}(t)) = \log(0.1) + \log(I_{true}(0)) + rt$$

The constant $\log(0.1)$ is absorbed into the intercept of the regression model ($\log(0.1) + \log(I_{true}(0))$), while the slope ($r$ be used to find the growth rate) remains unchanged. Therefore, the estimated growth rate and consequently R_0 would be unbiased.

5. What is the probability that this epidemic occurred following a single introduction of one infectious individual (assuming an exponentially distributed infectious period)?

We could use the probability of a large outbreak in a branching process.
let $1 - s$ be the probability of a big epidemic, s is the probability a big epidemic dose NOT occur.
Under the assumption of exponentially distributed infectious periods (which corresponds to $\tau_I = 1$, where τ_I is the coefficient of variation = SD/Mean), the probability of no epidemic can be derived mathematically.
For $\tau_I = 1$:

$$s = (\frac{1}{1 + (1-s)R_0\tau_I^2})^{\tau_I^{-2}} = (\frac{1}{1 + (1-s)R_0 * 1})$$

$$s(1 + (1-s)R_0) = 1$$

$$s + s(1-s)R_0 = 1$$

$$s(1-s)R_0 = 1 - s$$

$$sR_0 = 1$$

$$s = \frac{1}{R_0}$$

So, using R_0=1.636701 from Q3, we can get s (the Probability a big epidemic dose NOT occur):

```
# Probability a big epidemic dose NOT occur
s<-1/R0
s
```

```
##        t
## 0.610985
```

$$\text{Probability a big epidemic dose NOT occurs} = s = \frac{1}{1.636701} \approx 0.610985$$

Then we can get the probability of a big epidemic:

```
# Probability of a big epidemic
prob_epidemic<-1-s
prob_epidemic
```

```
##        t
## 0.389015
```

$$\text{Probability a big epidemic} = 1 - s = 1 - 0.610985 \approx 0.389015$$

Answer: The probability that this epidemic occurred following a single introduction of one infectious individual (assuming an exponentially distributed infectious period) would be 0.389015. This means if a single infectious individual entered the population, there was nearly a 38.90% chance it would trigger a major epidemic like the one observed.

6. Calculate R_j for the first 85 days of the epidemic (March 8 to May 31) using the method of Wallinga & Teunis and plot the results as a bar plot. (Show your code.)

```
# Extract the serint (serial interval data) from the covidCT
obsV <- as.numeric(covidCT$serint)
obsV
```

```
##  [1] 11 12  1  4  4  4 11  6  4  2 10  5  1  3  5 16 13  3  7  7  9  4  6  5  3
## [26]  3 23  4  6  4 15  6  9  9 12  6  2  4  1  4  8  5  4  4  9  9  9 15
```

```
# Gamma distribution
library(MASS) # Load this library: MASS
gpars <- fitdistr(obsV, densfun = "gamma", start = list("shape" = 1, "rate" = 1))[[1]] # fitdistr finds
```

```
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
```

```r
gpars
```

```
##     shape      rate
## 2.4261611 0.3561302
```

```r
# Probability density of serial interval
# dgamma() calculates the probability of infection after 1, 2, ..., 200 days
# shape and rate come from MLE of your serial interval data
g <- dgamma(1:200, shape = gpars[1], rate = gpars[2])

# Create a 200x200 matrix to store infection probabilities
# Rows = day when new case occurs (i = infected day)
# Columns = day of potential infector (j = source day)
# p[i,j] = probability that a case on day i was infected by a case on day j
p <- matrix(0, nrow = 200, ncol = 199)

# Loop over days starting from day 2 (day 1 has no previous cases)
for (i in 2:length(covidCT$newcasesCT)) {
  if (covidCT$newcasesCT[i] > 0) {for (j in 1:(i - 1)) {
    if (covidCT$newcasesCT[j] > 0) {
      p[i, j] <- g[i - j] / (g[seq(i - 1, 1, -1)] %*% covidCT$newcasesCT[1:(i - 1)])}}}}
```

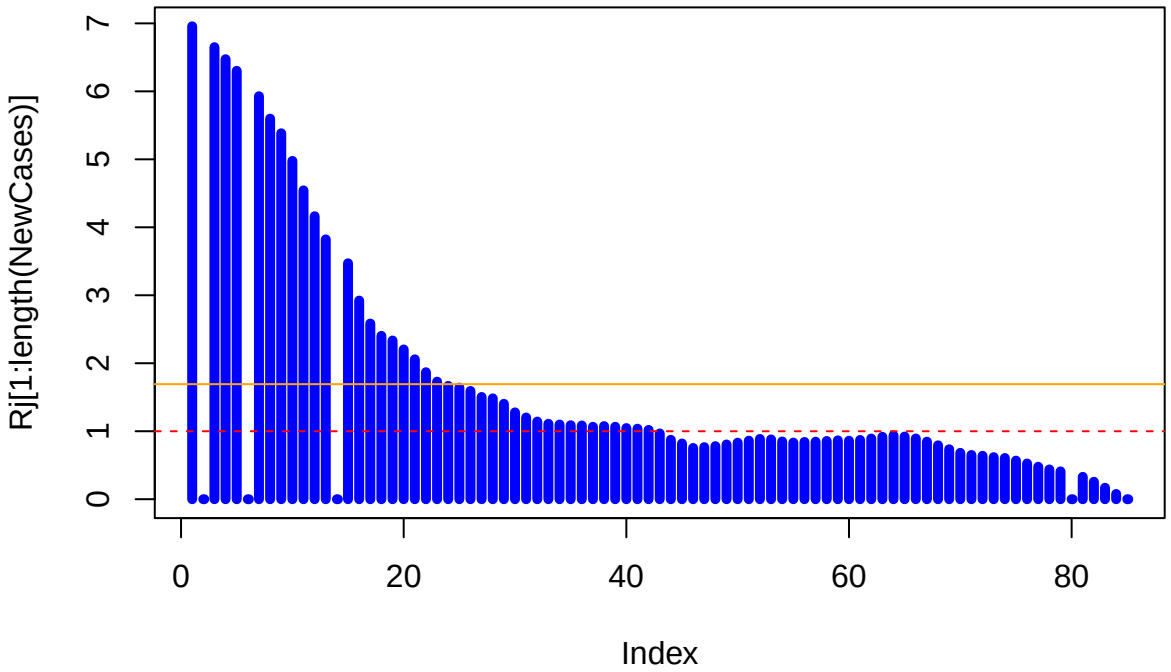
```r
# Extract new cases data for first 85 days
NewCases <- covidCT$newcasesCT[1:85]
Rj <- rep(NA, length(NewCases))
# Loop over each day t
for (t in 1:length(NewCases)) {
  # Rj[t] = expected number of secondary cases caused by cases on day t
  Rj[t] <- p[t:length(NewCases), t] %*% NewCases[t:length(NewCases)]
}

Rj # The number of secondary cases that the people on that day caused (effective reproduction number on
```

```
##  [1] 6.95656923 0.00000000 6.64962827 6.47241984 6.30170627 0.00000000
##  [7] 5.92761103 5.59828142 5.38275288 4.97657194 4.54240358 4.16234800
## [13] 3.82139862 0.00000000 3.47053719 2.92149515 2.58391938 2.40261282
## [19] 2.33405840 2.20237340 2.05759956 1.86754001 1.72941568 1.66618191
## [25] 1.64107707 1.59262258 1.50347000 1.48315734 1.40322606 1.27688488
## [31] 1.20173533 1.14183332 1.10922634 1.09823915 1.08800933 1.08445434
## [37] 1.06479725 1.07499133 1.06521322 1.04603944 1.03715886 1.01904791
## [43] 0.96692335 0.87384208 0.81900505 0.75352414 0.76566481 0.78006625
## [49] 0.80476964 0.83162237 0.86108289 0.88663387 0.87940227 0.84894995
## [55] 0.83277542 0.84196045 0.84607636 0.85592660 0.86510069 0.86014286
## [61] 0.87172783 0.89182321 0.91691937 0.93699498 0.92154565 0.89257661
## [67] 0.84651961 0.79223368 0.73450685 0.68135909 0.65009021 0.63834734
## [73] 0.61809703 0.60550881 0.56759456 0.52445111 0.47683924 0.43715507
## [79] 0.40902613 0.00000000 0.32852223 0.25648460 0.16799263 0.07674649
## [85] 0.00000000
```

```r
# Compute mean Rj across days with non-zero values
# (we exclude zero values, e.g., last day where no further infections can occur)
m<-mean(Rj[1:85][Rj[1:85] > 0])
m
```

```
## [1] 1.692189
```

```r
# Make a plot of Rj
plot(Rj[1:length(NewCases)], type = "h", col = "blue", lwd = 5)
abline(h = 1, col = "red", lty = 2) # R0=1
abline(h = m, col = "orange") # Mean
```



Answer: Using the Wallinga & Teunis with a Gamma serial-interval (shape = 2.4261611, rate = 0.3561302), we computed daily case reproduction numbers R_j for the first 85 days. Based on the graph, R_j was well above 1 early on and then declined toward/below 1, indicating control of transmission.

7. What is $R_{j=1}$ for the first case? Explain why it might be smaller/larger than the value of R_0 you calculated in question 3. (There are multiple correct answers.)

The effective reproductive number of case j is then:

$$R_j = \sum_i p_{ij}$$

Then $R_{j=1}$ represents the sum of probabilities that subsequent cases were infected by the first case:

$$R_1 = \sum_i p_{i1}$$

```r
# Calculate Rj for the first case (Day 1)
Rj_first <- Rj[1]
Rj_first
```

```
## [1] 6.956569
```

Answer: $R_{j=1}$ for the first case = 6.956569, and R_0 from Question 3 = 1.636701. So $R_{j=1}$ is significantly larger than the R_0, indicating explosive initial transmission. This is likely due to the early epidemic advantage: the population was 100% susceptible with no prior immunity, and no intervention measures were initially implemented (no mask-wearing, no social distancing, and no isolation protocols were in place). These conditions created an ideal environment for rapid viral transmission during the initial phase.

Extra credit **Middle East respiratory syndrome (MERS) is caused by a closely related coronavirus, and is associated with severe disease and a high case fatality risk. It is a zoonotic infection associated with exposure to camels, but is capable of limited human-to-human transmission. Breban et al (Lancet 2013) estimated the value of R0 for MERS-CoV based on the following observed transmission trees:**

Number of cases	Scenario 1	Scenario 2
1	17	11
2	4	2
3	3	3
4	1	1
5	0	2
24	1	1

(a) What is the maximum likelihood estimate of R_0 under Scenario 1 if you assume that the infectious period is exponentially distributed?

Since MERS has limited human-to-human transmission, the basic reproduction number should lie below the epidemic threshold ($R_0 < 1$). The number of secondary infections generated by one infectious individual in a fully susceptible population, should be smaller than 1.

If the infectious period is exponentially distributed, then the offspring distribution is *Geometric*, and the probability of observing n cases following (and including) 1 primary case should be $Pr(n|R_0) = \frac{(2n-2)!}{n!(n-1)!}\frac{R_0^{n-1}}{(R_0+1)^{2n-1}}$.

The distribution have $mean = \frac{1}{1-R_0}$.

Hence, $\frac{n}{p} = \frac{1}{1-R_0}$.

$n(1-R_0) = p \quad \longrightarrow \quad R_0 = \frac{n-p}{n} = 1 - \frac{p}{n}$, where n is the total number of cases resulting from (and including) p primary cases.

```
n <- 1*17 + 2*4 + 3*3 + 4*1 + 5*0 + 24*1

p <- 17 + 4 + 3 + 1 + 0 + 1

R0_a <- 1 - p/n
R0_a
```

```
## [1] 0.5806452
```

Therefore, $R_0 \approx 0.5806452$.

(b) Given their estimate of $R_0 = 0.69$ for Scenario 2, and assuming a constant infectious period, what is the most likely scenario leading to the 5 cases observed in Dammann (i.e. how many separate transmission trees were there and of what size)?

Since epidemic spread but extinction, R_0, the number of secondary infections generated by one infectious individual in a fully susceptible population, should be smaller than 1. $R_0 = 0.69 < 1$. If the infectious period is constant (same fro every one), then the offspring distribution is *Poisson*, and the probability of observing n cases following (and including) 1 primary case should be $Pr(n|R_0) = \frac{(nR_0)^{n-1}exp(-nR_0)}{n!}$.

```
R0_b <- 0.69

p_pois <- rep(NA, 5)
```

```
for (i in 1:5){
  p_pois[i] <- (i*R0_b)^(i-1)*exp(-i*R0_b)/factorial(i)
}

p_pois
```

```
## [1] 0.50157607 0.17358920 0.09011558 0.05544511 0.03747824
```

$Pr(n = 1|R_0 = 0.69) =\approx 0.5016$, $Pr(n = 2|R_0 = 0.69) \approx 0.1736$, $Pr(n = 3|R_0 = 0.69) \approx 0.0901$, $Pr(n = 4|R_0 = 0.69) \approx 0.0554$, $Pr(n = 5|R_0 = 0.69) \approx 0.0375$.

The scenarios leading to the 5 cases observed in Dammam can be:

$1 \to 1 \to 1 \to 1 \to 1 \to 1$: $Pr(n = 5|R_0 = 0.69)^5 \approx 0.0317$

```
p_pois[1]^5
```

```
## [1] 0.03174564
```

$1 \to 2 \to 1 \to 1 \to 1$ or $1 \to 1 \to 2 \to 1 \to 1$ or $1 \to 1 \to 1 \to 2 \to 1$ or $1 \to 1 \to 1 \to 1 \to 2$: $Pr(n = 2|R_0 = 0.69)Pr(n = 1|R_0 = 0.69)^3 \approx 0.0219$

```
p_pois[2]*p_pois[1]^3
```

```
## [1] 0.02190449
```

$1 \to 1 \to 2 \to 2$ or $1 \to 2 \to 1 \to 2$ or $1 \to 2 \to 2 \to 1$: $Pr(n = 2|R_0 = 0.69)^2 Pr(n = 1|R_0 = 0.69) \approx 0.0151$

```
p_pois[2]^2*p_pois[1]
```

```
## [1] 0.0151141
```

$1 \to 3 \to 1 \to 1$ or $1 \to 1 \to 3 \to 1$ or $1 \to 1 \to 1 \to 3$: $Pr(n = 3|R_0 = 0.69)Pr(n = 1|R_0 = 0.69)^2 \approx 0.0227$

```
p_pois[3]*p_pois[1]^2
```

```
## [1] 0.02267115
```

$1 \to 3 \to 2$ or $1 \to 2 \to 3$: $Pr(n = 3|R_0 = 0.69)Pr(n = 2|R_0 = 0.69) \approx 0.0156$

```
p_pois[3]*p_pois[2]
```

```
## [1] 0.01564309
```

$1 \to 4 \to 1$ or $1 \to 1 \to 4$: $Pr(n = 4|R_0 = 0.69)Pr(n = 1|R_0 = 0.69) \approx 0.0278$

```
p_pois[4]*p_pois[1]
```

```
## [1] 0.02780994
```

$1 \to 5$: $Pr(n = 5|R_0 = 0.69) \approx 0.0375$

```
p_pois[5]
```

```
## [1] 0.03747824
```

The most likely scenario:

```
max(p_pois[1]^5, p_pois[2]*p_pois[1]^3, p_pois[2]^2*p_pois[1], p_pois[3]*p_pois[1]^2, p_pois[3]*p_pois[
```

```
## [1] 0.03747824
```

Answer: The most likely scenario leading to the 5 cases observed in Dammam is a single transmission tree of size 5.