**1.2**A screenshot of a social media post

Description automatically generated

R code:

*p=dbinom(2,size=10,prob=0.3)*

*paste("The probability mass at the value X =2 for a binomial B(10,0.3) is",as.character(p1))*

*p\_cumulative1=sum(dbinom(0:2,size=10,prob=0.3))*

*paste("The cumulative distribution corresponding to P(X<=2) for a binomial B(10,0.3) is",as.character(p\_cumulative1))*

*p\_cumulative2=pbinom(2,size=10,prob=0.3)*

*paste("This may be verified by function 'pbinom' to be",p\_cumulative2)*

**1.3**

This function returns the portion(probability) of Poisson variables that are larger than or equal to 'm' amongst all 'n' variables characterized by the same 'lambda'.

R code:

*f\_poisson <- function(m,n,lambda){*

*p\_vec = rpois(n,lamba)*

*m\_prob = mean(p\_vec>=m)*

*return(m\_prob)*

*}*

**1.4**

A screenshot of a social media post

Description automatically generated

R code:

*f\_poisson <- function(m=10,n=1000,lambda=5){*

*p\_vec = rpois(n,lambda)*

*m\_prob = mean(p\_vec>=m)*

*return(m\_prob)*

*}*

*"If no argument is provided, then..."*

*f\_poisson()*

*"We may also override the arguments..."*

*f\_poisson(m=15,n=500,lambda=8)*

**1.5**

If we assume that stays 0.5, then for 100 trials of Poisson simulation we can see that .

If we want to prove that the probability is smaller than , then we need to increase granularity of the simulation, with at least 106 trials. The user may feel free to change the “*trials*” variable below to 106, but the previous statement remains true.

A screenshot of a cell phone

Description automatically generated

R Code:

*#Assume that the false positive rate 'r' stays the same at 0.01*

*p <- 0.01*

*#Assume that the number of patient samples stays the same at 50*

*n <- 50*

*l\_protein\_position <- 100*

*#The 'lambda' parameter for the Poisson distribution is thus 0.5*

*lambda <- n\*p*

*#If we simulate 100 trials, then the probability of finding a maximum greater than or equal to 9 is thus...*

*trials <- 100*

*maxes <- replicate(trials,{max(rpois(l\_protein\_position,lambda))})*

*prob <- formatC(mean(maxes>=9),format = "e", digits=8)*

*prob*

*#Let's investigate the trend, part I*

*X <- c(1:10)*

*Y <- c()*

*for (x in X)*

*{*

*y=as.double(mean(maxes>=x),format="e",digit=4)*

*Y <- c(Y,y)*

*}*

*plot(X,Y,xlab = "x",ylab = "P (X > x)")*

**1.8**

a.

We can use the *letterFrequency* function in Biostrings to obtain

|  |  |  |  |
| --- | --- | --- | --- |
| A | C | G | T |
| 4335 | 1225 | 2055 | 6179 |

b.

Under the equal probability assumption, the distribution with 3 degrees of freedom should have a distribution with the following characteristics from a 106 simulation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Q | Median | Mean | 3rd Q | Max. |
| 0.00029 | 1.21009 | 2.36654 | 3.00133 | 4.10932 | 29.42178 |

The theoretical results from using the *dchisq()* function also comfirmed the correctness of the simulation above, giving a median of approximately 2.3814 and a mean of 3.

Given that the 99% percentile of the distribution is around 11.3351.

The probability that the *C.elegans* data is consistent with the uniform model is close to zero.

In conclusion, we can be fairly certain the *C.elegans* data did not come from a uniform distribution.

A screenshot of a social media post

Description automatically generated

R Code:

*seqnames(BSgenome.Celegans.UCSC.ce2)*

*M <- BSgenome.Celegans.UCSC.ce2[["chrM"]]*

*ACGT\_freq <- Biostrings::letterFrequency(letters=c("A","C","G","T"),M)*

*s = sum(ACGT\_freq)*

*#Obtain the expectation value assuming As Cs Gs and Ts are equally distributed*

*pvec = rep(1/4, 4)*

*expected\_stat = pvec\*s*

*equal\_distribution = rmultinom(1000000, prob = pvec, size = s)*

*#Chi2 statistics*

*stat = function(observation, expectation = expected\_stat){*

*return(sum((observation-expectation)^2/expectation))*

*}*

*S = apply(equal\_distribution,2,stat)*

*#The observed frequencies of 'A', 'C', 'G', 'T' are respectively...*

*ACGT\_freq*

*#The expected frequencies of 'A', 'C', 'G', 'T' under equal distribution assumption are respectively*

*expected\_stat*

*#From above, its Chi2 statistic is...*

*stat(ACGT\_freq)*

*#Under equal probability assumption, the distribution should have the following characteristics ..*

*summary(S)*

*#The 99% percentile is...*

*quantile(S,0.99)*

*hist(S, breaks = 50, main="",freq=FALSE, xlab="")*

*abline(v = stat(ACGT\_freq), col = "red")*