Unit 2 – Discrete Distributions
Practice Problems
Solutions – *R Users*

1. *Source: Rosner B. Fundamentals of Biostatistics, second edition. Boston: Duxbury Press, 1986. Chapter 4 problem 4.29, page 93.*

Suppose that the expected number of deaths due to bladder cancer for all workers at a tire plant on January 1, 1964, over the next 20 years (1/1/64-12/31/83) based on US mortality rates is 1.8. If the Poisson distribution is assumed to hold and there are 6 reported deaths due to bladder cancer among the tire workers, then how unusual is this event? In developing your answer, compute the probability of *exactly 6 deaths* using the appropriately defined Poisson distribution model.

Answer: The likelihood of this event is .0078, representing an unlikely event

```
# dpois(x=, lambda= ) for exact Poisson Distribution probability
# Solve for Pr [ Poisson(mean=1.8) = 6

# Tip for Learning R.  Surrounding command in parentheses tells R to show output.
(dpois(6,lambda=1.8))

## [1] 0.007808587
```

2. *Source: Rosner B. Fundamentals of Biostatistics, second edition. Boston: Duxbury Press, 1986. Chapter 4 problem 4.30, page 93-94.*

The rate of myocardial infarction (MI) in 50–59-year-old disease-free women is approximately 2 per 1000 per year or 10 per 1000 over 5 years. Suppose that 3 MI's are reported over 5 years among 1000 women initially disease-free who have been taking postmenopausal hormones.

(a). Use the binomial distribution to see if this experience represents an unusually small number of events based on the overall rate. That is, compute the probability of *3 or fewer events of MI* using the appropriately defined Binomial distribution model.

Answer: .01007

```
# pbinom(q=, size=, prob= ) for exact cumulative Binomial Distribution probability
# Solve for Pr [ Binomial(n=1000, p=.01) <= 3]

(pbinom(q=3,size=1000,prob=.01))

## [1] 0.01007265
```

(b). Answer exercise "a" using the Poisson approximation to the binomial distribution.

Answer: .01033

```
# ppois(q=, lambda= ) for cumulative Poisson Distribution probability
# Solve for Pr [ Poisson(mean=10) <= 3

(ppois(3,lambda=10))

## [1] 0.01033605
```

(c). Compare your answers to "a" and "b".

> The answers are similar! The closeness of these answers is what we might expect given that the probability of event is small and the number of trials is large.

3. *Source: Fisher LD and Van Belle G. Biostatistics: A Methodology for the Health Sciences. New York: Wiley, 1993. Chapter 6 problem 5, page 232.*

Smith, Delgado and Rutledge (1976) report data on ovarian carcinoma. Individuals had different numbers of courses of chemotherapy. The 5-year survival data for those with 1-4 and 10 or more courses of chemotherapy are:

	Five Year Status	
Courses	**Dead**	**Alive**
1-4	21	2
≥ 10	2	8

Using Fisher's Exact test, is there a statistically significant association ($p<.05$) in this table? In 1-2 sentences, write a clear interpretation of your hypothesis test.

Preliminary – Create an R object that is a 2x2 table (there are lots of ways to do this …)

```
# rbind( ) and as.table( ) to create 2x2 table of counts with counts entered row by row
# dimnames(tablename) <- list(
#    ROWNAME=c("Smoke","Dont Smoke"),
#    COLNAME=c("Abnormal","Normal"))
# tablename

q3table <- as.table(rbind(c(21,2),c(2,8)))
dimnames(q3table) <- list(
    "Courses"=c("1-4",">= 10"),
    "Five Year Status"=c("Dead","Alive"))
q3table

##        Five Year Status
## Courses Dead Alive
##   1-4     21     2
##   >= 10    2     8
```

Tip for Learning R. The R command **fisher.test( )** does not produce the odds ratio calculation (OR = ad/bc) that you are familiar with:

	Five Year Status	
Courses	**Dead**	**Alive**
1-4	21	2
≥ 10	2	8

Familiar odds ratio calculation: $OR = \frac{ad}{bc} = \frac{21 \cdot 8}{2 \cdot 2} = 42$

This familiar odds ratio calculation is called by various names: "sample OR", "unconditional MLE OR"
The other kind of odds ratio calculations are called conditional MLE OR.
These do not have straightforward formulae and are not covered in this course.

Solution I: Using command fisher.test( ) which does NOT calculate OR as OR = (ad/bc)

Answer:
The null hypothesis of "no association of number of courses of chemotherapy with 5-year survival" is rejected (assumption of the null hypothesis and its application to the data has led to a highly unlikely result; p-value = .0001). Examination of the data suggests ≥ 10 courses of chemotherapy, relative to 1-4 courses is associated with improved survival at 5-years.

```
# Tip Learning R.  Use the option alterntive = to specify one or two-sided p-value calculation
# fisher.test(TABLENAME,alternative="two.sided")
# fisher.test(TABLENAME,alternative="greater")
# fisher.test(TABLENAME,alternative="less")

fisher.test(q3table,alternative="greater")

##
##  Fisher's Exact Test for Count Data
##
## data:  q3table
## p-value = 0.0001255
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  4.927592      Inf
## sample estimates:
## odds ratio
##   34.05494   ← Dear class – This is not match the familiar OR = (ad)/(bc) = 42.
```

Solution II: Using command epi.2by2( ) in package {epiR} which DOES calculate OR as OR = (ad/bc)

Answer: The null hypothesis of "no association of number of courses of chemotherapy with 5-year survival" is rejected (assumption of the null hypothesis and its application to the data has led to a highly unlikely result; p-value = .0001). Examination of the data suggests ≥ 10 courses of chemotherapy, relative to 1-4 courses is associated with improved survival at 5-years (OR = 42) but is imprecise due to small cell frequencies (95% CI estimate of OR = 5.03 - 350.75).

```
# Using command epi.2by2 in package {epiR}
# The following assumes that you have previously installed {epiR}. One way to do this is:
# install.packages("epiR", dependencies=TRUE)

library(epiR)

q3data = matrix(c(21,2,2,8), ncol=2, byrow=TRUE,
              dimnames = list(c("1-4 courses", "10 or more courses"),
                              c("Dead", "Alive")))
q3data

##                     Dead Alive
## 1-4 courses           21     2
## 10 or more courses     2     8

epi.2by2(q3data)

##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           21            2         23              91.3
## Exposed -            2            8         10              20.0
## Total               23           10         33              69.7
##                   Odds
## Exposed +        10.50
## Exposed -         0.25
## Total             2.30
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                4.57 (1.31, 15.87)
## Odds ratio = (ad/bc)                          42.00 (5.03, 350.75)   This is OR = (ad)/(bc)
## Attrib risk *                                 71.30 (43.97, 98.64)
## Attrib risk in population *                   49.70 (20.36, 79.03)
## Attrib fraction in exposed (%)                78.10 (23.85, 93.70)
## Attrib fraction in population (%)             71.30 (9.47, 90.90)
## -------------------------------------------------------------------
##  X2 test statistic: 16.778 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

4. *Source: Vu J and Harrington D. Introductory Statistics for the Life and Biomedical Sciences, First Edition. OpenIntro, 2020. Chapter 3 problem 10, page 187.*

The US CDC estimates that 90% of Americans have had chickenpox by the time they reach adulthood. Consider a sample of 120 American adults.

(a). How many people in this sample would you expect to have had chickenpox in their childhood?

Answer: 108

```
This is Binomial distribution (n,p) calculation of the mean μ = np with:
n = Number of trials = 120
p = Pr [ event ] = .90


μ = (n)(p) = (120)(.90)
   = 108
```

(b). What is the standard deviation σ?

Answer: σ = 3.29

```
This is Binomial distribution (n,p) calculation of the standard deviation
```

$$\sigma = \sqrt{(n)\,(p)\,(1-p)}$$
$$= \sqrt{(120)\;(.90)\;(.10)}$$
$$= 3.29$$

(c). What is the probability that 105 or fewer people in this sample have had chicken pox in their childhood?

Answer: 0.22, representing a 22% chance, approximately

```
> # Binomial (n=120, p=.90)  Solution for Pr[ X <= 105]
> # Solution I - using sum (dbinom() )
> sum( dbinom(x=0:105,size=120,prob=.90, log=FALSE))
[1] 0.2181634
>
> # Solution II - using pbinom()
> pbinom(q=105,size=120,prob=.90)
[1] 0.2181634
```

5. *Source: Whitlock MC and Schluter D. The Analysis of Biological Data, Second Edition. WH Freeman, 2015. Chapter 8 problem 19, page 230.*

Hurricanes hit the United States often and hard, causing some loss of life and enormous economic costs. They are ranked in severity by the Saffir-Simson scale, which ranges from Category 1 to Category 5, with 5 being the worst. In some years, as many as three hurricanes that rate a Category 3 or higher hit the U.S. coastline. In other years, no hurricane of this severity hit the United States. The following table lists the number of years that had 0, 1, 2, 3, or more hurricanes of at least Category 3 in severity of the 100 years of the 20th century (Blake et al 2005).

Number of hurricanes with severity Category 3 or higher	**Number of Years Observed**
0	50
1	39
2	7
3	4
More than 3	0

(a). Calculate the mean number of severe (Category 3 or higher) hurricanes to hit the United States per year during the 100 years of the 20th century.

Answer: 0.65, representing less than 1 per year on average

```
> solution <- (0*50  +  1*39  +  2*7  +  3*4  +  0)/100
> cat("Mean number severe hurricanes per year during 20th century = ", solution )
Mean number severe hurricanes per year during 20th century =  0.65
```

(b). Define all terms in the probability distribution model that would be used to describe the distribution of hurricanes per year during the 100 years of the 20th century under the assumptions that hurricanes hit independently of each other and the probability of a hurricane hit is the same every year.

Answer: **Poisson (μ = 0.65)**

This is a Poisson model developed as follows
Unit of time = 1 year
Rate of event per unit of time = estimated mean # severe hurricanes/year = 0.65 yielding
Poisson μ = (duration surveillance) x (rate) = (100) x (0.65) = 65

```
# Check of Poisson (lambda = .65) probabilities of 0, 1, 2, 3 or 3+ severe hurricanes per years
> p0 <- dpois(x=0, lambda=.65, log=FALSE)
> p1 <- dpois(x=1, lambda=.65, log=FALSE)
> p2 <- dpois(x=2, lambda=.65, log=FALSE)
> p3 <- dpois(x=3, lambda=.65, log=FALSE)
> pmore <- 1 - sum(dpois(x=0:3, lambda=.65, log=FALSE))
>
> # Expected number of years over entire 20th century with 0, 1, 2, 3 or 3+ severe hurricanes
> e0 <- 100*p0
> e1 <- 100*p1
> e2 <- 100*p2
> e3 <- 100*p3
> emore <- 100*pmore
>
> observed_hurricanes <- c(50,39, 7, 4, 0)
> poisson_expected <- c(e0, e1, e2, e3, emore)
>
>
> mytable <- cbind(observed_hurricanes, poisson_expected)
> mytable

     observed_hurricanes poisson_expected
[1,]                  50       52.2045777
[2,]                  39       33.9329755
[3,]                   7       11.0282170
[4,]                   4        2.3894470
[5,]                   0        0.4447828
```

Sampling, as it will, produces observed outcomes that differ from the probability distribution model expected values. For example, our Poisson model predicts that there would be 52.2 years (out of 100) with 0 Category 3+ hurricanes.