LECTURE 2A – UNIVARIATE DISCRETE DISTRIBUTIONS

Bernoulli, Binomial, Multinomial, Geometric, Negative Binomial, and Poisson Distributions

See Book Chapter 4

Types of random variables

- We can divide random variables broadly into two types:
  - **Discrete** random variables have probabilities defined for discrete numerical values of the random variable
  - **Continuous** random variables have probabilities defined on ranges of values of the random variable
- Examples:
  - Gender can be made a (discrete) random variable by recasting it as: M = 0 and F = 1.
  - Income is a continuous random variable, though probabilities are only defined for ranges of its values

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Probability Distribution Function

- A **probability distribution __function__** is a mathematical function that provides the probabilities, given values of the random variable.

- A discrete random variable is said to have a *probability mass function* (pmf) that supplies probabilities for *each value* of the random variable. The probabilities are non-negative and always sum up to 1 for the sample space (or for all values of the random variable).
  - pmf for X_1 (Gender): $P(X_1 = 0) = p$; $P(X_1 = 1) = (1-p)$
  - The pmf directly gives you the probabilities based on the values of the random variable. i.e., it is often expressed as a function of the random variable.
  - The *cumulative probability distribution* (cdf) specifies the probability for the random variable being *less than or equal* to a particular value.

- A continuous random variable is said to have a *probability density function (pdf)* that (when integrated) supplies probabilities for *ranges of values* of the random variable. The probabilities are non-negative and the probability is 1 when the probability density function is integrated over the entire range of the random variable.
  - Pdf for X_5 (Income): $1/\sqrt{2\pi\sigma^2}\, e^{-(x-\mu)^2/2\sigma^2}$
  - To find the Probability (X_5 <= 0.25), you will *integrate the pdf* from –infinity to 0.25. This is also its cumulative distribution function (cdf)

Known Discrete (Probability) Distribution Functions

- We will discuss the pmf and cdf of several types of discrete random variables where, knowing the value of the random variable we can calculate its probability.
  - Bernoulli
  - Binomial
  - Multinomial
  - Negative Binomial
  - Geometric
  - Hypergeometric
  - Poisson
- In each case we will also look at the underlying events and the applications
- We can calculate parameters such as mean, median, mode, variance etc., for these known distribution functions.

BINOMIAL DISTRIBUTION

Discrete Distributions

The Bernoulli Trial & the Binomial Distribution

- The simplest discrete distribution is the Bernoulli trial and corresponds to an experiment with a "success" of probability p, and a "failure" of probability (1 – p)

- We can *convert the experiment's outcome to a random variable* X such that Success = 1 and Failure = 0.
  - Probability Mass Function: $f_{bernoulli} = p^x(1-p)^{1-x}$
  - Expected Value = p; Variance = p(1-p)

- The Bernoulli distribution (or trial) forms the basis of the binomial and geometric distributions

- **Binomial Distribution**:
  - Underlying experiment – n independent Bernoulli trials
  - Random variable X – Number of successes in n trials; X ~ Binomial(n, p).
    Note: ~ indicates "distributed as"
  - Probability Mass Function: $f_{binom} = \binom{n}{x}p^x(1-p)^{n-x}$
  - Expected Value = np; Variance = np(1-p)
  - When n = 1, we get the Bernoulli

- Because we have the *probability mass function* (pmf) for the random variable, <u>we do not need a table showing probabilities for each value of the random variable; the probability can be obtained from the pmf once we know the value of the random variable.</u>

Practice with calculator

Binomial Probability Mass Function (pmf)

- Binomial Distribution:
  - Random variable X – Number of successes in n trials; X ~ Binomial(n, p).
    Note: ~ indicates "distributed as"
  - Probability Mass Function: $f_{binom} = \binom{n}{x} p^x (1-p)^{n-x}$
  - Expected Value = np; Variance = np(1-p)
  - When n = 1, we get the Bernoulli

- Example:
  - Consider the Gender (X_2) random variable. Let us assume that the population proportion of male clients is .52 and female clients 0.48. Then, p = 0.52 (since male is our success state and X_2 = 1 for males and 0 for females).
  - The probability that we have X_2 =5 males in a data set of n=10 is given by:
  - $f_{binom} = \binom{n}{x} p^x (1-p)^{n-x}$ = (10!/(5!5!))$(0.52)^5(0.48)^5$ = 0.244

Practice with calculator

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Practice with calculator

Using the Calculator

X=Males	Prob
0	0.000649
1	0.007034
2	0.034289
3	0.099056
4	0.187793
5	0.244131
6	0.220396
7	0.136436
8	0.055427
9	0.013344
10	0.001446
Sum	1

- In our example n = 10, p=0.52
- pmf = $\binom{n}{x}p^x(1-p)^{n-x}$ = $\binom{10}{x}0.52^x(0.48)^{10-x}$
- A choice function $\binom{n}{x}$ means "the number of ways you can "*choose from N things taking x at a time*". Its value = $\frac{n!}{(x!)(n-x)!}$
  - n! or "n-factorial" = 1 x 2 x 3 x….x (n-1) x n. i.e., the product of all numbers up to and including n.
  - Example: $\binom{5}{2}$ means number of ways of choosing say two different people out of 5. Let us say we have **A**be, **B**eth, **C**arey, **D**ave and **E**rnie;
  - Then we have $\binom{5}{2}$ = $\frac{5!}{(2!)(3!)}$ = $\frac{1x2x3x4x5}{(1x2)(1x2x3)}$ =10
  - i.e., we have 10 pairs of people {AB, AC, AD, AE, BC, BD, BE, CD, CE, DE}. Note that CD is the same as DC etc.
- You can usually get factorials and combination function values from calculators or software such as Excel or you can compute them using the factorials.
- **TI-30 XIIS Binomial Probability**: YouTube video - https://www.youtube.com/watch?v=UYgsvTM9tBA

Using R-language

- In this course we will the R-language to do many of the tasks required.
- R is open-source.
- R-platforms such as R-Studio are freely available
- You can install R-Studio (and R-language) on your personal computers to do the assignments.
- *RStudio* is free for download at: https://www.rstudio.com/products/rstudio/download/#download along with the R-Language download.
  - **This is the preferred approach**. It is easy to save and re-open files in the R-program from your own PC or laptop,
- The Data and Program Files will be available in **Canvas** under the Module **R-files**.
- R-Studio is also available through the Spears Lab Virtual Machine
  - Go to http://desktop.okstate.edu and download VMWare Horizon Client. Then use your Okey Id and password to log into the server **SSB MSIS Lab.** Look for R-Studio under Start**.**
  - Your R-programs can be saved on One Drive available in the virtual machine or on memory sticks

Download ClassData.csv and Binomial.R from Canvas

1

R-Files

ClassData.csv (Data File) Click

MR100.csv (Data File For Multiple Regression)

ReadClassData.R

Binomial.R Click

2

ClassData.csv

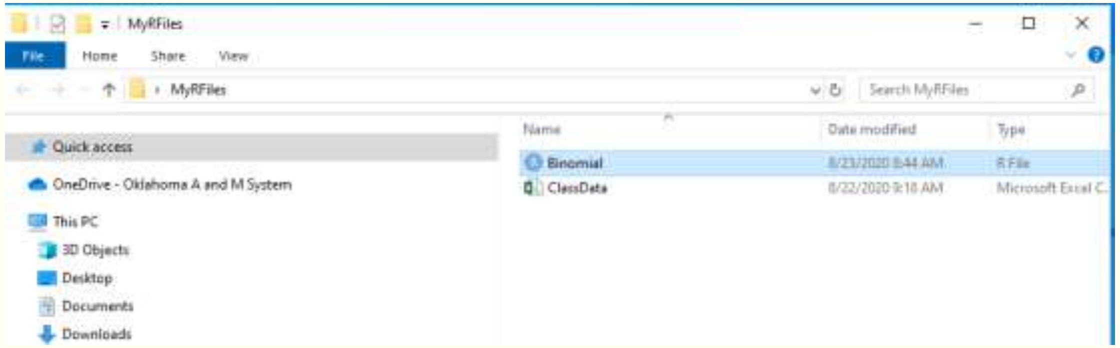
Download ClassData.csv (578 Bytes)

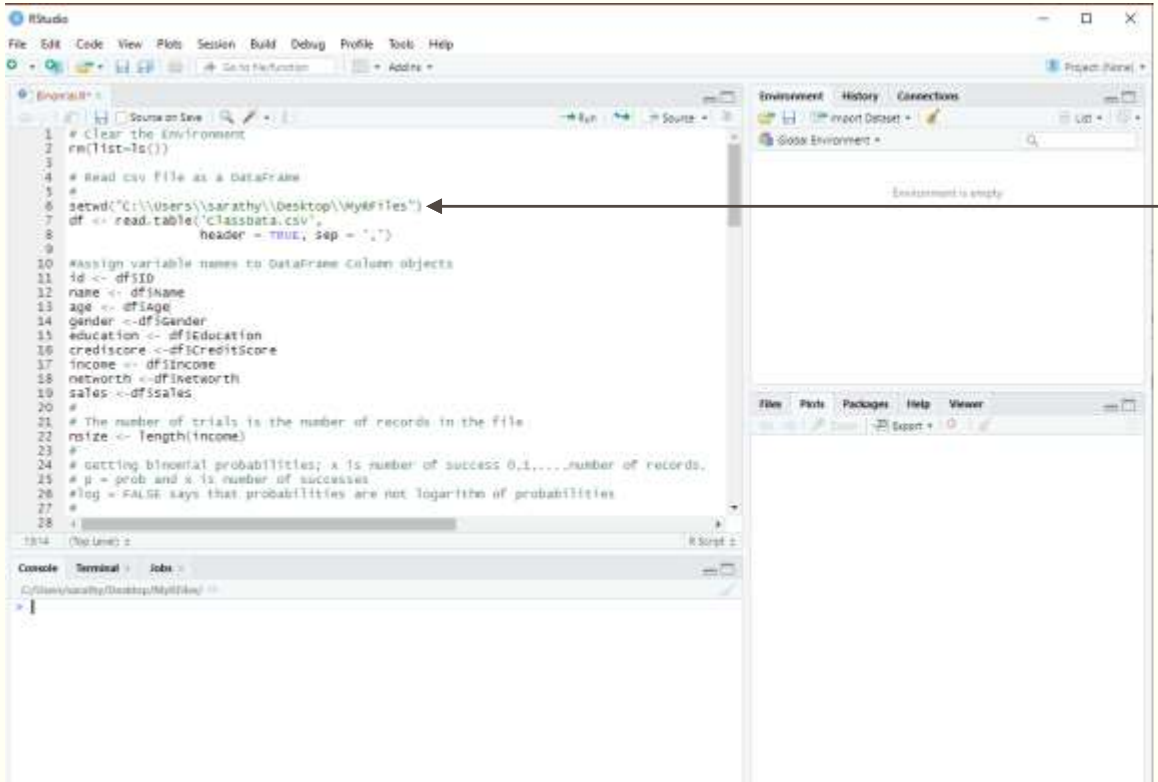
Clicking here will download the file to your browser's Downloads folder. Copy from there into a folder "MyRFiles" on the Desktop.

Do the Same for "Binomial.R".

ID	Name	Age	Gender	Education	CreditScor	Income	NetWorth	Sales
1	Adams, Jol	36	M	HS	350	38900	65924	1535
2	Ramesh, J	23	F	Bachelors	600	172000	178154	2196
3	Mendez, Ni	67	M	Bachelors	700	218000	265209	1287
4	Mendez, Jc	38	F	PhD	550	182000	85277	2143
5	Ritter, Jake	24	M	Masters	625	434000	193760	707
6	Rao, Eric	61	M	PhD	770	82000	314953	2170
7	Blake, Ann	26	F	HS	490	112000	192946	1229
8	Bishop, Ma	44	F	Masters	540	242000	339705	520
9	Ahmed, Mc	31	M	Masters	680	111000	185767	2326
10	Shultz, Dar	44	M	Bachelors	280	66000	97778	588

Binomial.R File opened from “MyRFiles” on Desktop





Open up Binomial.R by double-clicking on it or by opening R-Studio, suing File…Open File…and pointing to this file on your computer.

Make sure that the working directory in the setwd command is:

setwd("C:\\Users\\<yourid>\\Desktop\\MyRFiles")

Practice with calculator

Basic Commands for pmf, cdf and quantiles of Binomial in R

- In R,
  - *dbinom*(x, n, p, log = FALSE) gives the probability from the **pmf** P(X = x)
  - *pbinom*(x, n, p, log = FALSE) gives the cumulative probability (**cdf)** P(X ≤ x)
  - *qbinom*(percentile, n, p, lower.tail = TRUE, log = FALSE) gives x for P(X ≤ x) = percentile
- Working directly from R Console
  - You can enter the command directly into the R console (without using the Binomial.R script)
  - Type in the first command for dbinom(x, n, p,): **dbinom(0, 10, 0.52)**
  - Thereafter, use the Up Arrow key on your keyboard to recall the previous commands to modify the value for X and reissue the command.

```
Console   Terminal ×   Jobs ×
C:/Users/sarathy/Desktop/MyRFiles/
> dbinom(0, 10, 0.52)
[1] 0.0006492506
> dbinom(1, 10, 0.52)
[1] 0.007033548
> dbinom(2, 10, 0.52)
[1] 0.03428855
> dbinom(3, 10, 0.52)
[1] 0.09905581
> dbinom(4, 10, 0.52)
[1] 0.1877933
> dbinom(5, 10, 0.52)
[1] 0.2441313
> dbinom(6, 10, 0.52)
[1] 0.2203963
> dbinom(7, 10, 0.52)
[1] 0.1364358
> dbinom(8, 10, 0.52)
[1] 0.05542705
> dbinom(9, 10, 0.52)
[1] 0.01334355
> dbinom(10, 10, 0.52)
[1] 0.001445551
```

X=Males	Prob
0	0.000649
1	0.007034
2	0.034289
3	0.099056
4	0.187793
5	0.244131
6	0.220396
7	0.136436
8	0.055427
9	0.013344
10	0.001446
Sum	1

Working Directly on the RStudio Console – cdf values and quantiles/percentiles

Console | Terminal | Jobs

C:/Users/sarathy/Desktop/MyRFiles/

```
> pbinom(0, 10, 0.52)
[1] 0.0006492506
> pbinom(1, 10, 0.52)
[1] 0.007682799
> pbinom(2, 10, 0.52)
[1] 0.04197135
> pbinom(3, 10, 0.52)
[1] 0.1410272
> pbinom(4, 10, 0.52)
[1] 0.3288205
> pbinom(5, 10, 0.52)
[1] 0.5729517
> pbinom(6, 10, 0.52)
[1] 0.793348
> pbinom(7, 10, 0.52)
[1] 0.9297839
> pbinom(8, 10, 0.52)
[1] 0.9852109
> pbinom(9, 10, 0.52)
[1] 0.9985544
> pbinom(10, 10, 0.52)
[1] 1
```

- Use *pbinom*(x, n, p,) to get the **cdf** values P(X ≤ x)
- Note: You can use “Cntrl L” to clear the Console.

Console | Terminal | Jobs

C:/Users/sarathy/Desktop/MyRFiles/

```
> print("The 10th percetile is:")
[1] "The 10th percetile is:"
> qbinom(0.10,10,0.52, lower.tail = TRUE)
[1] 3
> print("The 50th percetile is or Median is:")
[1] "The 50th percetile is or Median is:"
> qbinom(0.50,10,0.52, lower.tail = TRUE)
[1] 5
```

- Use *qbinom*(percentile, n, p, lower.tail = TRUE) to get the **percentile** values X such that P(X ≤ x) = percentile.

Binomial.R File

```r
# Clear the Environment
rm(list=ls())

# Read csv file as a DataFrame
#
setwd("C:\\Users\\sarathy\\Desktop\\R-Code Examples")
df <- read.table('ClassData.csv',
                 header = TRUE, sep = ',')

#Assign variable names to DataFrame Column objects
id <- df$ID
name <- df$Name
age <- df$Age
gender <-df$Gender
education <- df$Education
crediscore <-df$CreditScore
income <- df$Income
networth <-df$Networth
sales <-df$Sales
#
# The number of trials is the number of records in the file
nsize <- length(income)
#
# Getting binomial probabilities; x is number of success 0,1,...,number of records.
# p = prob and x is number of successes
#log = FALSE says that probabilities are not logarithm of probabilities
#
prob = 0.52
x = 5
# dbinom gives pmf value - probability of x successes
#
print(paste("Probability of ",x," successes in ",nsize, " trials is:",
            round(dbinom(x, nsize, prob, log = FALSE),4) ))
# pbinom gives cdf value - probability of <= x successes
#
print(paste("Probability of <= ",x," successes in ",nsize, " trials is:",
            round(pbinom(x, nsize, prob, log = FALSE),4) ))
# qbinom gives the quantile value of x for the specified quantile.
#
print(paste("The value of x for the 35th quantile i.e., P(X <=x) = 0.35 is ",
            qbinom(0.35, nsize, prob, lower.tail = TRUE, log = FALSE)))
#
```

```
> print(paste("Probability of ",x," successes in ",nsize, " trials is:",
+             round(dbinom(x, nsize, prob, log = FALSE),4) ))
[1] "Probability of  5  successes in  10  trials is: 0.2441"
> # pbinom gives cdf value - probability of <= x successes
> #
> print(paste("Probability of <= ",x," successes in ",nsize, " trials is:",
+             round(pbinom(x, nsize, prob, log = FALSE),4) ))
[1] "Probability of <=  5  successes in  10  trials is: 0.573"
> # qbinom gives the quantile value of x for the specified quantile.
```

```
> print(paste("The value of x for the 35th quantile i.e., P(X <=x) = 0.35 is ",
+            qbinom(0.35, nsize, prob, lower.tail = TRUE, log = FALSE)))
[1] "The value of x for the 35th quantile i.e., P(X <=x) = 0.35 is  5"
```

Binomial.R File (continued)

Table of Probabilities, Empirical Mean, Variance and Standard Deviation

```
# Table of probabilities
#
# Setting vectors to hold intermediate results
#
# Because x=0,1,2,,,number of records, we need the vectors to have size 1 greater
# than number of records
#
result <- vector("numeric", nsize+1)
cum_result <- vector("numeric", nsize+1)
xPx <- vector("numeric", nsize+1)
x2Px <- vector("numeric", nsize+1)
rv_binom <-vector("numeric", 5)
#
# Each vector's index goes from 1 to 11. For example result[1] will hold probability x = 0
# and result[11] will hold probability x = 10
#
for (i in 0:nsize) {
  result[i+1] <- dbinom(i, nsize, prob, log = FALSE)
  cum_result[i+1] <- pbinom(i, nsize, prob, log = FALSE)
  print(paste("X = ",i, "probability = ",round(result[i+1],4),
              "cumumlative probability = ", round(cum_result[i+1],4), sep = " "))
}

# Checking the Empirical mean vs Theoretical mean = np and
# Checking the Empirical variance vs Theoretical variance = np(1-p)
# Checking the Empirical standard deviation vs Theoretical standard deviation = sqrt(np(1-p))
for (i in 0:nsize) {
  xPx[i+1] <- i*result[i+1]
  x2Px[i+1] <- i*xPx[i+1]
}

# Mean = sum of xPx
Exp_val = sum(xPx)
print(paste("The Expected Value (empirical mean) is",round(Exp_val, 4), sep = " "))

# Var = sum of X2Px - (sum(xPx)^2)
varian = sum(x2Px) - Exp_val*Exp_val
print(paste("The Empirical Variance is",round(varian, 4), sep = " "))
print(paste("The Empirical standard deviation is",round(sqrt(varian), 4), sep = " "))
```

```
[1] "X =  0 probability =  6e-04 cumumlative probability =  6e-04"
[1] "X =  1 probability =  0.007 cumumlative probability =  0.0077"
[1] "X =  2 probability =  0.0343 cumumlative probability =  0.042"
[1] "X =  3 probability =  0.0991 cumumlative probability =  0.141"
[1] "X =  4 probability =  0.1878 cumumlative probability =  0.3288"
[1] "X =  5 probability =  0.2441 cumumlative probability =  0.573"
[1] "X =  6 probability =  0.2204 cumumlative probability =  0.7933"
[1] "X =  7 probability =  0.1364 cumumlative probability =  0.9298"
[1] "X =  8 probability =  0.0554 cumumlative probability =  0.9852"
[1] "X =  9 probability =  0.0133 cumumlative probability =  0.9986"
[1] "X =  10 probability =  0.0014 cumumlative probability =  1"
```

```
> # Mean = sum of xPx
> Exp_val = sum(xPx)
> print(paste("The Expected Value (empirical mean) is",round(Exp_val, 4), sep = " "))
[1] "The Expected Value (empirical mean) is 5.2"
>
> # Var = sum of X2Px - (sum(xPx)^2)
> varian = sum(x2Px) - Exp_val*Exp_val
> print(paste("The Empirical Variance is",round(varian, 4), sep = " "))
[1] "The Empirical Variance is 2.496"
> print(paste("The Empirical standard deviation Variance is",round(sqrt(varian), 4), sep = " "))
[1] "The Empirical standard deviation Variance is 1.5799"
```

Binomial Distribution – Expected Value, Variance and Standard Deviation

Practice with calculator

- We will calculate the expected value and variance both ways – from formula and from data
- **From Formula**:
- Expected Value = np = 10*(0.52) = 5.2;
- Variance = np(1-p) = 5.2*0.48 = 2.496

From Data (see Lecture 1):

Expected Value (or **Mean**) E(X) = $\sum_{x=1 \text{ to } 10} xP(X=x)$;

$$\textbf{Variance} = V(X) = \sum_{x=1 \text{ to } 10} (x - E(X))^2 P(X=x)$$

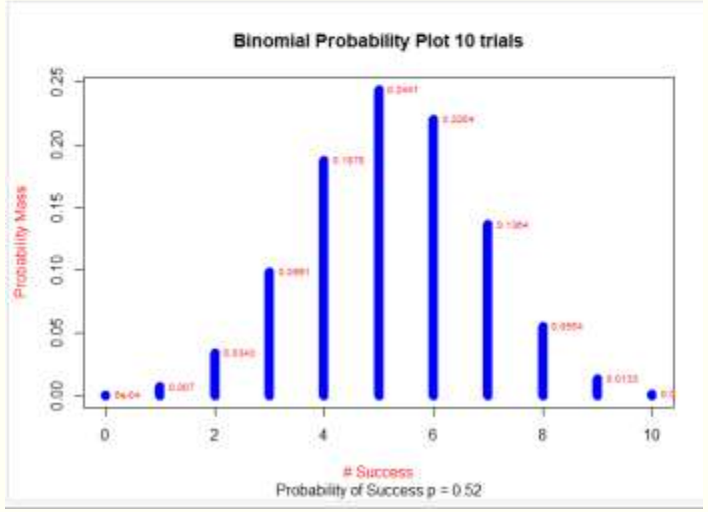
Standard Deviation = sqrt(2.496) = 1.58

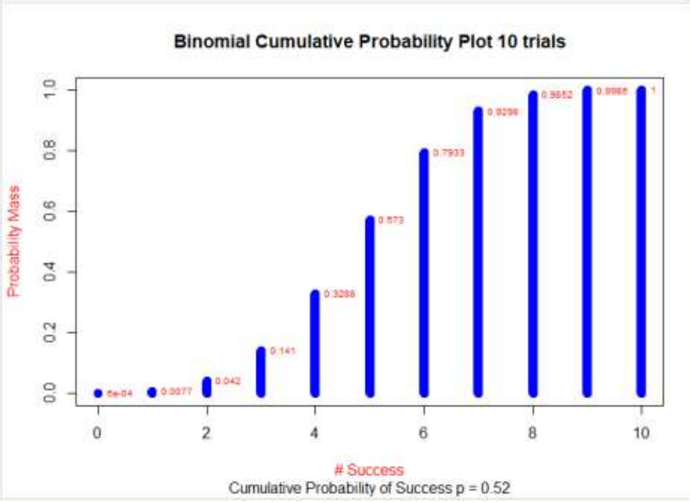
```
#table of probabilities
for (i in 0:nsize) {
  result[i+1] <- dbinom(i, nsize, prob, log = FALSE)
  xPx[i+1] <- i*result[i+1]
  x2Px[i+1] <- i*xPx[i+1]
  cum_result[i+1] <- pbinom(i, nsize, prob, log = FALSE)
  print(paste("X = ",i, "probability = ",round(result[i+1],4),
        "cumumlative probability = ", round(cum_result[i+1],4), sep = " "))
}

# Checking the mean = np and variance = np(1-p)

# Mean = sum of xPx
Exp_val = sum(xPx)
print(paste("The Expected Value is",round(Exp_val, 4), sep = " "))

# Var = sum of x2Px - (sum(xPx)^2)
varian = sum(x2Px) - Exp_val*Exp_val
print(paste("The Variance is",round(varian, 4), sep = " "))
```

X	Prob	E(X) = xP(x)	(x - E(X))^2	(x - E(X))^2 P(x)
0	0.000649	0	27.04	0.017555737
1	0.007034	0.00703355	17.64	0.124071794
2	0.034289	0.0685771	10.24	0.351114736
3	0.099056	0.29716742	4.84	0.479430104
4	0.187793	0.7511732	1.44	0.270422352
5	0.244131	1.22065645	0.04	0.009765252
6	0.220396	1.32237782	0.64	0.141053634
7	0.136436	0.95505065	3.24	0.442052014
8	0.055427	0.44341637	7.84	0.434548045
9	0.013344	0.12009193	14.44	0.192680837
10	0.001446	0.01445551	23.04	0.033305496
Sum	1	5.2		2.496

Binomial.R File (continued) – Histogram of pmf and cdf

```
# Plot the pmf
len_result <- length(result)
indx = len_result - 1
#
plot((0:indx),result[1:len_result],                 #--x-axis, y-axis
     type = "h",                                    #--type is histogram
     main = "Binomial Probability Plot 10 trials", #--Main title
     sub = "Probability of Success p = 0.52",      #--sub-title
     xlab = "# Success",                            #--X-Label
     ylab = "Probability Mass",                     #--Y-label
     col  = "blue",                                 #--color of pillars
     col.lab ="red",                                #-- color of X and Y labels
     lwd=10)                                        #--width of pillars
#
text((0:indx), result[1:len_result],                #-- adding text legend to pillars;x-axis;y-axis
     round(result[1:len_result], 4),                #-- value shown in pillars
     cex=0.6,                                       #-- size expansion for text font
     pos=4,                                         #-- 4 indicates show to right of pillar (choose 1,2,3,4)
     col="red")                                     #-- color of text

# Plot the cdf
len_result <- length(result)
indx = len_result - 1
plot((0:indx),cum_result[1:len_result],
     type = "h",
     main = "Binomial Cumulative Probability Plot 10 trials",
     sub = "Cumulative Probability of Success p = 0.52",
     xlab = "# Success", ylab = "Probability Mass",
     col  = "blue",
     col.lab ="red",
     lwd=10)
text((0:indx), cum_result[1:len_result], round(cum_result[1:len_result], 4), cex=0.6, pos=4, col="red")
#
```





Practice with calculator

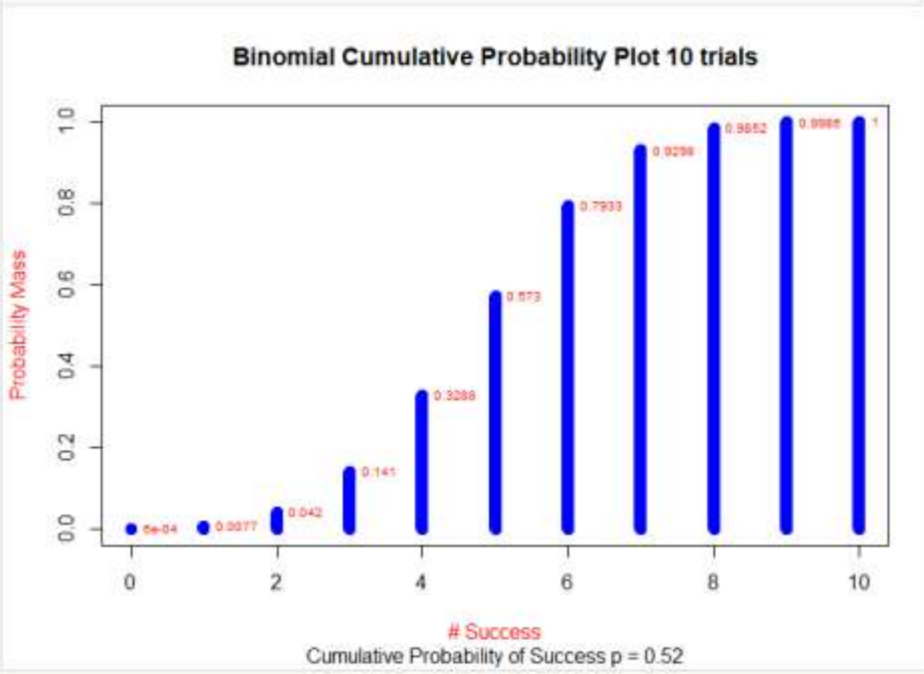
Binomial Cumulative Distribution Function (cdf)

- The probability that we have at least 3 *males* in our data set
  $P(X \geq 3)$
  $= 1 - P(X \leq 2) = 1 - 0.042 = 0.958$
  = print(1 - pbinom(2, 10, 0.52, log = FALSE))

- The probability that we have no more than 4 *females* in our data set
  $= P(X \geq 6\ ) = 1 - P(X \leq 5) = 1 - 0.573 = 0.4271$
  = print(1 - pbinom(5, 10, 0.52, log = FALSE))

- The probability that we have between 8 and 9 *females* in our data set
  $= P(X = 1)$ or $P(X = 2) = 0.007 + 0.0343 = 0.0413$
  = print(dbinom(1, 10, 0.52, log=FALSE) + dbinom(2, 10, 0.52, log=FALSE))

- The probability that all the people in our data set are *female*
  $= P(X = 0) = 0.0006$
  = print(dbinom(0, 10, 0.52, log=FALSE))




```
[1] "x =  0 probability =  6e-04 cumumlative probability =  6e-04"
[1] "x =  1 probability =  0.007 cumumlative probability =  0.0077"
[1] "x =  2 probability =  0.0343 cumumlative probability =  0.042"
[1] "x =  3 probability =  0.0991 cumumlative probability =  0.141"
[1] "x =  4 probability =  0.1878 cumumlative probability =  0.3288"
[1] "x =  5 probability =  0.2441 cumumlative probability =  0.573"
[1] "x =  6 probability =  0.2204 cumumlative probability =  0.7933"
[1] "x =  7 probability =  0.1364 cumumlative probability =  0.9298"
[1] "x =  8 probability =  0.0554 cumumlative probability =  0.9852"
[1] "x =  9 probability =  0.0133 cumumlative probability =  0.9986"
[1] "x =  10 probability =  0.0014 cumumlative probability =  1"
```

Practice with calculator

Book Problem 4.98 – Page 291

- Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number that participated in after-school sports all four years of high school.
  - In words, define the random variable X. X is "the number that participated in after-school sports all four years of high school."
  - List the values that X may take on. X = {0,1,2,….,60}
  - Give the distribution of X including parameters, if any. i.e., X ~ Binomial (60, 0.08) where "success" is a student participating in sports all 4 years and p=0.08 is the probability of success.
  - How many seniors are expected to have participated in after-school sports all four years of high school? Expected value = np = 60*0.08 = 4.8 seniors.
  - Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
    We look at P(X=0) from this binomial
    print(dbinom(0, 60, 0.08, log = FALSE)) = 0.0068.
    Since this is a low probability, yes, we would be surprised.
  - Based on numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically. 4 Seniors – see below.

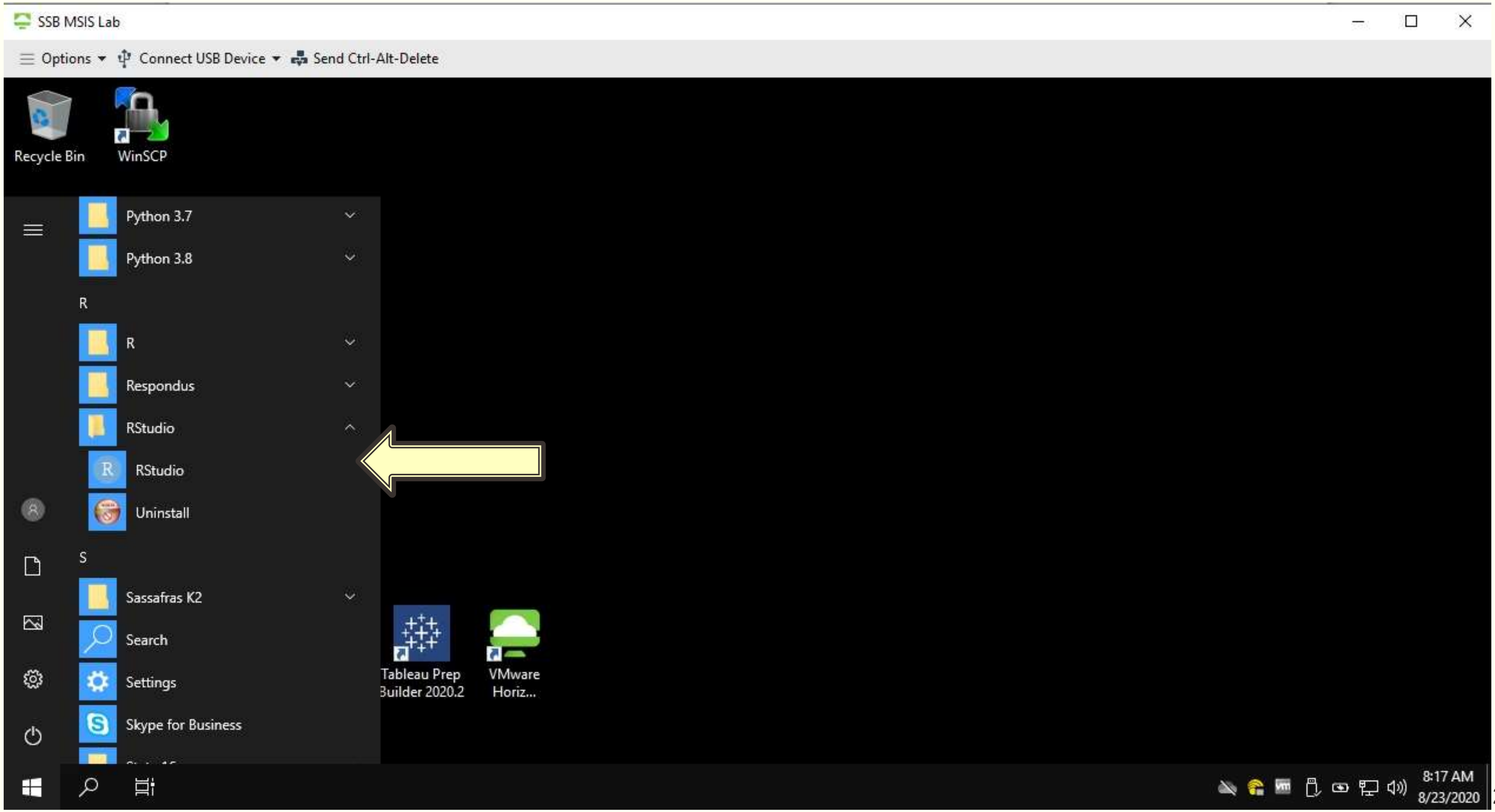
```
> print(dbinom(4, 60, 0.08, log = FALSE))
[1] 0.1873153
> print(dbinom(5, 60, 0.08, log = FALSE))
[1] 0.1824288
```

Practice with calculator

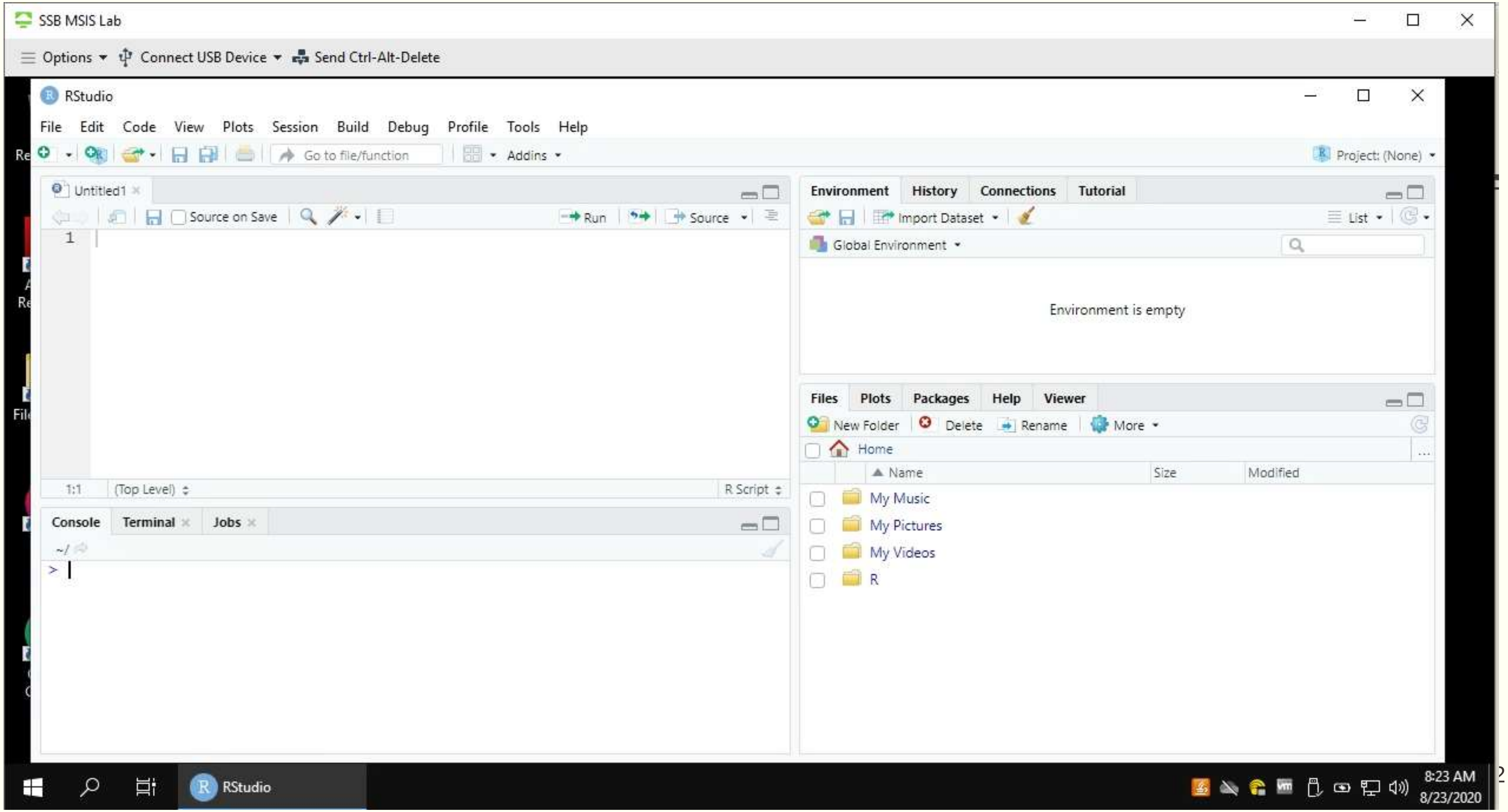
Problems

- A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.
  - In words, define the random variable X. The random variable X is: “the number of students who will attend Tet festivities”.
  - List the values that X may take on. $X = \{0,1,2,\ldots,12\}$
  - Give the distribution of X including parameters, if any. i.e., X ~ Binomial(12, 0.18) where “success” is if a student attends Tet.
  - How many of the 12 students do we expect to attend the festivities? The expected value is np = 12*0.18 = 2.16 students.
  - Find the probability that at most four students will attend.
    $P(X \leq 4) =$
    ```
    > print(pbinom(4, 12, 0.18, log = FALSE))
    [1] 0.9510694
    ```
  - Find the probability that at least three students will attend.
    $P(X \geq 3) = 1 - P(X \leq 2) =$
    ```
    > 1 - pbinom(2, 12, 0.18, log=FALSE)
    [1] 0.3702131
    ```
  - 50% of the time, up to how many students attend the festivities?
    ```
    > qbinom(0.5, 12, 0.18, log=FALSE)
    [1] 2
    ```

Accessing R-Studio from SSB MSIS Lab Virtual Machine



Accessing R-Studio from SSB MSIS Lab Virtual Machine



MULTINOMIAL DISTRIBUTION

Discrete Distributions

The Multinomial Distribution

- Multinomial Distribution:
  - Underlying experiment – n independent trials of **k-sided die**
  - Random variables (in this case k = 6 sides) $\{X_i: i=1,\ldots,k\}$
  - X_i – ***Number of times side i shows up*** in n trials; X ~ Multinomial (k, n, p_k)
  - Probability Mass Function: $f_{multinom} = \frac{n!}{x_1!x_2!,\ldots,x_k!}\, p_1^{x_1} p_2^{x_2}{}_{,\ldots,}\, p_k^{x_k}$
  - Mean = Expected Value $E(X_i) = np_i$;
  - Variance = $np_i\,(1-p_i)$;
  - Covariance $(X_i, X_j) = -np_ip_j$, $i \neq j$.

Example; Consider the Education random variable (X_3) that has a multinomial distribution.

Assume for k =1,…,4 we have PhD is k=1, Masters is k=2, Bachelors is k=3 and HS is k=4.

Assume the probabilities in the population are: p_1 = 0.1, p_2 = 0.2, p_3=0.3 and p_4=0.4.

http://stattrek.com/online-calculator/multinomial.aspx is an online multinomial calculator.

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Multinomial.R

- Let us use **R** to find the probability that in 10 records we have 2 PhDs, 2 Masters, 2 Bachelors and 4 HS
- In R, *dmultinom*(x_vec, n_size, prob_x, log = FALSE) gives the probability from the pmf P(X = x)
- In our example of 2 PhDs, 2 Masters, 2 Bachelors and 4 HS
  x_vec <- c(2, 2, 2, 4)
  n_size = 10
  prob_x <- c(0.1, 0.2, 0.3, 0.4);
- The answer is: 0.01742

```
#Multinomial Probability of 2 PhDs, 2 Masters, 2 Bachelors and 4 HS
x_vec <- c(2,2,2,4)
n_size = 10
prob_x <- c(0.1, 0.2, 0.3, 0.4)
m_result <- dmultinom(x_vec, n_size, prob_x, log = FALSE)
print(paste("The Multinomial Probability of 2 PhDs, 2 Masters, 2 Bachelors and 4 HS is",m_result,sep=" "))
[1] "The Multinomial Probability of 2 PhDs, 2 Masters, 2 Bachelors and 4 HS is 0.0174182400000001"
```

```
#If we select 5 customers, what is the probability that there will be 2 Masters and 3 HS?
x_vec <- c(0, 2, 0, 3)
n_size = 5
prob_x <- c(0.1, 0.2, 0.3, 0.4)
m_result <- dmultinom(x_vec, n_size, prob_x, log = FALSE)
print(paste("The Multinomial Probability of 0 PhDs, 2 Masters, 0 Bachelors and 3 HS is",m_result,sep=" "))

[1] "The Multinomial Probability of 0 PhDs, 2 Masters, 0 Bachelors and 3 HS is 0.0256"
```

The Multinomial Distribution

- **Note:** In this case, **we cannot** show the table of probabilities for all the values for X_3 since there could be 10 (PhD) x10 (Masters) x10 (Bachelors) x10 (HS) = 10^4 or 10,000 possible values.
- But we can certainly calculate the probabilities for each one set of values for the random variable.
- The mean (expected value) of Bachelor degree customers is $np_3 = 10 * 0.3 = 3$
- The variance of HS degree customers is $np_4(1- p_4) = 10 * 0.4 * 0.6 = 2.4$. The standard deviation = sqrt (2.4)
- The covariance (we will see later what covariance is) between PhD and Masters = $-np_1p_2 = -10*0.1*0.2 = -0.2$

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

The Multinomial Distribution

- The multinomial distribution (as well as the binomial distribution) can be used in sampling problems, where each trial can be considered a draw from a sample.
- The type of sampling modeled is **sampling with replacement** because it does not affect the probability of success in the population (p)
- This will be in contrast to the *hypergeometric distribution (*we shall see later), where we consider probabilities of values in a sample, where the probability of success changes with each item sampled. This is **sampling without replacement**.

Problem

- Suppose we have a bowl with 10 marbles - 2 red marbles, 3 green marbles, and 5 blue marbles. We randomly select **4 marbles** from the bowl, ***with replacement***.
  - In words, define the random variable X. X is the number of marbles of each color in the sample of 4.
  - List *some of the values* the values that X may take on. {0 red, 0 green, 4 blue}, {0 red, 1 green, 3 blue},…etc.
  - Give the distribution of X including parameters, if any. i.e., X ~ Multinomial (k, n, p_k) where k = 3, n = 4, p_1=0.2, p_2=0.3, p_3=0.5
  - What is the probability of selecting 2 green marbles and 2 blue marbles?

```
x_vec <- c(0, 2, 2)
n_size = 4
prob_x <- c(0.2, 0.3, 0.5)
m_result <- round(dmultinom(x_vec, n_size, prob_x, log = FALSE), 4)
print(paste("The Multinomial probability of selecting 2 green marbles and 2 blue marbles is",m_result,sep=" "))
```

```
[1] "The Multinomial probability of selecting 2 green marbles and 2 blue marbles is 0.135"
```

HYPERGEOMETRIC DISTRIBUTION

Discrete Distributions

The Hypergeometric Distribution

- Contrasted with the Binomial (and multinomial) distribution, the **hypergeometric distribution** represents **sampling without replacement** i.e., each draw affects the probability of success (and failure), because the population is not assumed infinite like the binomial.

- The hypergeometric distribution pmf is given by:

- P (X = x) = $f_{hypgeo} = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$ where:
    - N is the population size,
    - K is the number of "successes" in the population,
    - n is our sample size, and
    - x is the number of "successes" in the sample.

- The Expected value E(X) of the HyperGeometric Distribution is n(K/N) and variance is:

$$n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}$$

The Hypergeometric Distribution

- Suppose our data set with 10 records below represents a random sample of our customers and we are interested in the **probability** of obtaining the 4 Females ("successes") given that we have 100 customers, 60 of whom are Females. The hypergeometric distribution gives such a probability.

- The probability from hypergeometric distribution pmf P (X = x) = $f_{hypgeo} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = \frac{\binom{60}{4}\binom{40}{6}}{\binom{100}{10}} = 0.1081$

  - N = 100,
  - K is the number of "successes" or Females in the population = 60,
  - n is our sample size = 10
  - and k is the number of "successes" or Females in the sample (4)

- The Expected value of the HyperGeometric Distribution is n(K/N) = 6

- The variance is = 2.1818

Practice with calculator

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

HyperGeom.R

```
# Hypergeometric Distribution
#
#dhyper(x, m, n, k, log = FALSE) where
# x = number of successes in sample (k in our notes)
# m = number of successes in population (K in our notes)
# n = number of failures in population (N - K) in our notes
# k = sample size (n in our notes)
# probability of obtaining the x=4 Females ("successes"), m=60 of whom are Females, n=40 Males ("failures").
#
print(paste("probability of obtaining the x=4 Females, m=60 of whom are Females, n=40 Males ",
            round(dhyper(4, 60, 40, 10, log = FALSE),4)))
#
#
#table of probabilities
nsize <- length(income)
result <- vector("numeric", 11)
cum_result <- vector("numeric", 11)
xPx <- vector("numeric", 11)
x2Px <- vector("numeric", 11)
for (i in 0:nsize) {
  result[i+1] <- dhyper(i, 60, 40, 10, log = FALSE)
  xPx[i+1] <- i*result[i+1]
  x2Px[i+1] <- i*xPx[i+1]
  cum_result[i+1] <- phyper(i, 60, 40, 10, log = FALSE)
}
round(result[1:11], 4)
round(cum_result[1:11], 4)
#
# Mean = sum of xPx
Exp_val = sum(xPx)
print(paste("The Expected Value is",Exp_val, sep = " "))

# Var = sum of X2Px - (sum(xPx)^2)
varian = sum(x2Px) - Exp_val*Exp_val
print(paste("The Variance is",varian, sep = " "))
```

```
[1] "probability of obtaining the x=4 Females, m=60 of whom are Females, n=40 Males  0.1081"
```

```
> round(result[1:11], 4)
 [1] 0.0000 0.0009 0.0079 0.0369 0.1081 0.2076 0.2643 0.2204 0.1153 0.0342 0.0044
> round(cum_result[1:11], 4)
 [1] 0.0000 0.0010 0.0089 0.0457 0.1538 0.3614 0.6258 0.8462 0.9615 0.9956 1.0000
> #
> # Mean = sum of xPx
> Exp_val = sum(xPx)
> print(paste("The Expected Value is",Exp_val, sep = " "))
[1] "The Expected Value is 6"
>
> # Var = sum of X2Px - (sum(xPx)^2)
> varian = sum(x2Px) - Exp_val*Exp_val
> print(paste("The Variance is",varian, sep = " "))
[1] "The Variance is 2.18181818181817"
```

Practice with calculator

Book Problem 114 – Page 294

- Suppose that a technology task force is being formed to study technology awareness among instructors. Assume that 10 people will be randomly chosen to be on the committee from a group of 28 volunteers, 20 who are technically proficient and eight who are not. We are interested in the number on the committee who are **not** technically proficient.
  - In words, define the random variable X. X is the number on the committee who are **not** technically proficient.
  - List the values that X may take on. X = {0,1,…,8}
  - Give the distribution of X including parameters, if any. i.e., X ~ Hypergeometric(N=28, K=8, n=10, k=x); Note: that "success" here is **not** being technically proficient.
  - How many instructors do you expect on the committee who are not technically proficient? Expected value is n(K/N) = 10(8/28) = 2.857.
  - Find the probability that at least five on the committee are not technically proficient. P $(X \geq 5) = 1 - P(X \leq 4)$ =

```
> print(paste("probability of obtaining the x>=5 non-proficient ",
+             1 - round(phyper(4, 8, 20, 10, log = FALSE),4)))
[1] "probability of obtaining the x>=5 non-proficient  0.0772"
```

  - proficient. $P(X \leq 3)$ =

```
> print(paste("probability of obtaining the x<=3 non-proficient ",
+             round(phyper(3, 8, 20, 10, log = FALSE),4)))
[1] "probability of obtaining the x<=3 non-proficient  0.716"
```

NEGATIVE BINOMIAL & GEOMETRIC DISTRIBUTIONS

Discrete Distributions

The Bernoulli Trial & the Negative Binomial Distribution

- The negative binomial distribution gives the probability of n failures before the **rth** “success” (or r^{th} failure) in a sequence of Bernoulli trials. The number of Bernoulli trials is n+r.

- X is called a negative binomial random variable because, in contrast to the binomial random variable, *the number of successes (or failures) is fixed* and the number of Bernoulli trials (x+r) is a random variable because x is a random variable.

- There are many flavors of the negative binomial depending on what is defined as success, therefore what is p, and what the other parameters are.

- *We will use the version implemented in R-language*.

- If x is *the number of failures before r successes is reached* (so total number of trials is n = x + r), p is the probability of success in any trial, then X ~ NegBin(n, r , p) where n= r + x.

- Thus, $f_{negbin} = \binom{x + r - 1}{r - 1}(1\text{-}p)^x p^r = \frac{(x+r-1)!}{x!(r-1)!}\ (1-p)^x p^r$

- The Expected value = r*(1-p)/p i.e., the expected number of failures before r successes

- Variance = $r*(1\text{-}p)/p^2$.

Practice with calculator

The Bernoulli Trial & the Negative Binomial Distribution

- *As we look at new customers*, what is the probability that we will see 3 males before we see the second female customer in that group?
  - We define Female customer as success (and Male as "failure"), given p=probability of a Female in any trial (record) is 0.48?
- **Answer:**
  - To have x=3 failures (Males) before r=2 successes (Females) (n = 3+2), we have $\binom{4}{1}(.52)^3(0.48)^2$ = **0.1296**
  - To have x=0 failures (Males) before r=2 successes (Females) (n = 0+2), we have (0.48*0.48) = **0.2304**
  - Expected value = r(1-p)/p = 2*0.48/0.52 = 1.846; we expect to have about two males before we see two females and therefore 4 customers before we see two females.
  - Variance = $r*(1-p)/p^2$ = 1.846/0.52 = 3.551 and standard deviation is sqrt(3.551) = 1.884
- Note that it is not possible to enumerate the values that X can take, because theoretically X can take infinite values starting from 0, though after a while the probabilities for large X can become 0 depending on p.
- In the previous problem, using Chebchev's theorem we can see that by about 10 standard deviations from the mean (x = 20, n =22), the probabilities for X will become less than 0.01, if not sooner.

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	5,20
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

NegBinomial-Geometric.R

```
#
#Negative binomial - number of males we will see before the rth Female,
# assuming that p=Probabilty of seeing a female = 0.48
#
x <- c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)
r = 2
p_female=0.48
nsize = 16
result <- vector("numeric", 16)
cum_result <- vector("numeric", 16)
for (i in 0:15) {
  result[i+1] <- dnbinom(i, r, p_female, log = FALSE)
  cum_result[i+1] <- pnbinom(i, r, p_female, log = FALSE)
}
print("Number of failures before 2 successes")
x
print("Probabilities")
round(result[1:nsize], 4)
print("Cumulative Probabilities")
round(cum_result[1:nsize], 4)
#
```

```
[1] "Number of failures before 2 successes"
```

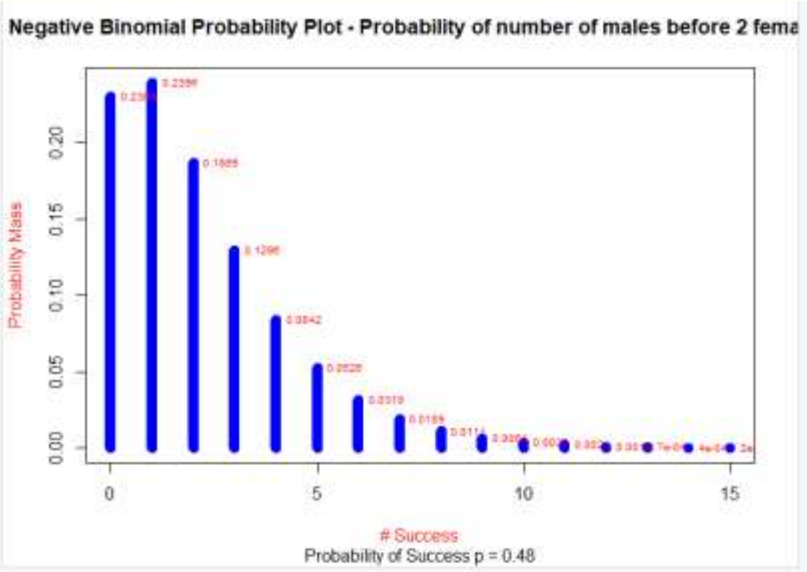
```
[1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
```

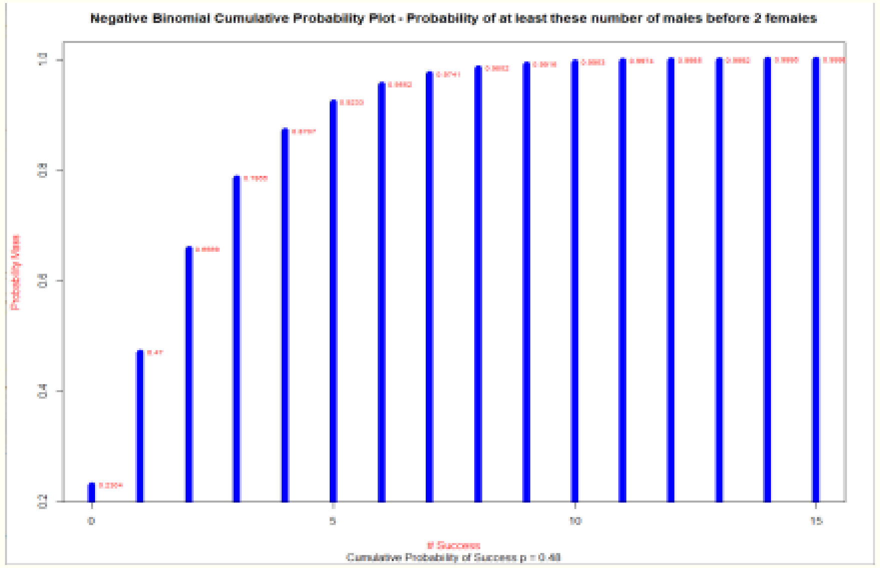
```
[1] 0.2304 0.2396 0.1869 0.1296 0.0842 0.0526 0.0319 0.0189 0.0111 0.0064 0.0037 0.0021 0.0012 0.0007 0.0004 0.0002
```

```
[1] "Cumulative Probabilities"
> round(cum_result[1:nsize], 4)
 [1] 0.2304 0.4700 0.6569 0.7865 0.8707 0.9233 0.9552 0.9741 0.9852 0.9916 0.9953 0.9974 0.9985 0.9992 0.9995 0.9998
```

NegBinomial-Geometric.R

```
#
#Plot the probability mass function
plot((0:indx),result[1:len_result], type = "h", main = "Negative Binomial Probability Plot
     - Probability of number of males before 2 females",
     sub = "Probability of Success p = 0.48", xlab = "# Success", ylab = "Probability Mass",
     col  = "blue", col.lab ="red", lwd=10)
text((0:indx), result[1:len_result], round(result[1:len_result], 4), cex=0.6, pos=4, col="red")

# Plot the cumulative probabilities
len_result <- length(result)
indx = len_result - 1
plot((0:indx),cum_result[1:len_result], type = "h", main = "Negative Binomial Cumulative Probability Plot
     - Probability of at least these number of males before 2 females",
     sub = "Cumulative Probability of Success p = 0.48", xlab = "# Success", ylab = "Probability Mass",
     col  = "blue", col.lab ="red", lwd=10)
text((0:indx), cum_result[1:len_result], round(cum_result[1:len_result], 4), cex=0.6, pos=4, col="red")
```





The Bernoulli Trial & the Geometric Distribution

- The Geometric distribution, is a specialization of the negative binomial, for **r=1**. It gives the probability of n Bernoulli trials before the first "success"
  - That is, we have (n-1) failures and 1 success
- X ~ Geom(n, p) with the understanding r=1 and x + 1 = n.
- P is the probability of "success", (1-p) the probability of failure.
- Thus, $f_{geom} = (1-p)^x p$; The Expected value = (1-p)/p. Variance = $(1-p)/p^2$.
- What is the probability that the next 5 customers are Male (the sixth customer is a Female)?
- $= (.52)^5(.48) = 0.0182$

Practice with calculator

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Practice with calculator

The Bernoulli Trial & the Geometric Distribution

- The Geometric distribution, is a specialization of the negative binomial, for **r=1**. It gives the probability of x failures before the first “success”
  - That is, we have x failures and 1 success and n = x+1 trials
- X ~ Geom(n, p); Thus, $f_{geom} = (1-p)^x p$; The Expected value = (1-p)/p. Variance = $(1-p)/p^2$.
- What is the probability that the next 5 customers are Male (the sixth customer is a Female)?

```
#
# Geometric Distribution
# What is the probability that the next 5 customers are Male (the sixth customer is a Female)?
#
r = 1
print(paste("The probability that the next 5 customers are Male (the sixth customer is a Female) is ",
            round(dnbinom(5, r, p_female, log = FALSE), 4), sep = " "))
```

```
[1] "The probability that the next 5 customers are Male (the sixth customer is a Female) is  0.0182"
```

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Practice with calculator

Book Problem 104 – Page 292

- A consumer looking to buy a used red Miata car will call dealerships until she finds a dealership that carries the car. She estimates the probability that any independent dealership will have the car will be 28%. We are interested in the number of dealerships she must call.
  - In words, define the random variable X. X is the number of dealerships she called that don't carry a red Miata. If the first dealership she calls has the car, then X = 0.
  - List the values that X may take on. X = {0, 1, 2,3, ….}
  - Give the distribution of X including parameters, if any. i.e., X ~ Geometric (n, p) where x + 1 = n.
  - On average, how many dealerships without cars does she call before she finds one with a car?
    Expected value of Geometric Distribution is (1-p)/p and since p = 0.28, it is 2.57.
  - Find the probability that she must call at most five dealerships. This means that the fifth dealership had a red Miata. Since we are modeling number of dealerships without the car (failures) we are looking at "at most" 4 dealerships.
    We need $P(X \leq 4) = 0.8065$

```
> print(paste("probability that she must call at most five dealerships ",
+             round(pnbinom(4, 1, 0.28, log = FALSE), 4)))
[1] "probability that she must call at most five dealerships  0.8065"
```

  - Find the probability that she must call four or five dealerships before she finds a car. This means we are looking at the probability of 3 failures or 4 failures.
    We need $P(X = 3) + P(X = 4) = 0.1045$

```
> print(paste("probability that she must call 4 or 5 dealerships ",
+             (round(dnbinom(3, 1, 0.28, log = FALSE) + round(dnbinom(4, 1, 0.28, log = FALSE)), 4))))
[1] "probability that she must call 4 or 5 dealerships  0.1045"
```

POISSON DISTRIBUTION

Discrete Distributions

Poisson Distribution

- The Poisson distribution gives the *probability of a specified number of events occurring in a fixed interval of time and/or space*.
  - *Example:* Suppose that a customer purchases 12 times a month on average, what is the probability that they will make 180 purchases over the next 12 months?
  - *Example:* Suppose a book has 3 errors per page on average. What is the probability of finding 10 pages with no errors?
- The Poisson distribution is actually *a limiting case of a Binomial distribution when the number of trials, n, gets very large and p, the probability of success, is small*.
- The probability of each event is independent of its previous occurrence just like in the Binomial where the probability of success in any trial is independent of the trial.
- The Poisson random variable X = the number of occurrences in the interval of interest. X ~ Poisson(λ)
- Pmf f_{pois} = P (X = k) = $\frac{e^{-\lambda}\lambda^{x}}{x!}$, x = 1,2,3,…
  - λ = average number of occurrences per time period, k is number of occurrences
  - E(X) = Var (X) = λ
- CDF F_{pois} = P (X <= x) = $e^{-\lambda}\sum_{0}^{k}\frac{\lambda^{x}}{x!}$, x = 1,2,3,…

Poisson Distribution

- Suppose that a customer purchases 12 times a month on average, what is the probability that they will make 180 purchases over the next 12 months?
  - Pmf $f_{pois} = P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$, k = 1,2,3,…
  - For our example, λ = 12 per month and x = 15 per month.
    - Exactly 15 purchases/month = $\frac{e^{-12}12^{15}}{(15)!}$ = 0.072
    - At least 15 purchases/month = $e^{-12}\sum_{0}^{15}\frac{12^{15}}{15!}$ = 0.844
    - Expected Value = Variance = 12/month

Practice with calculator

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

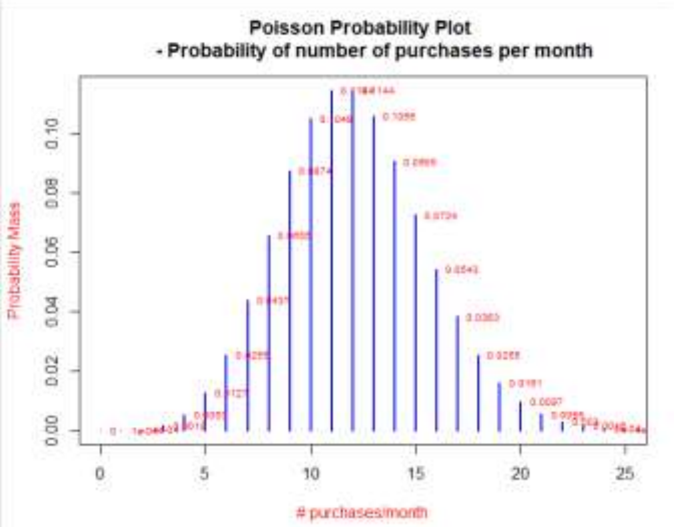
Poisson. R

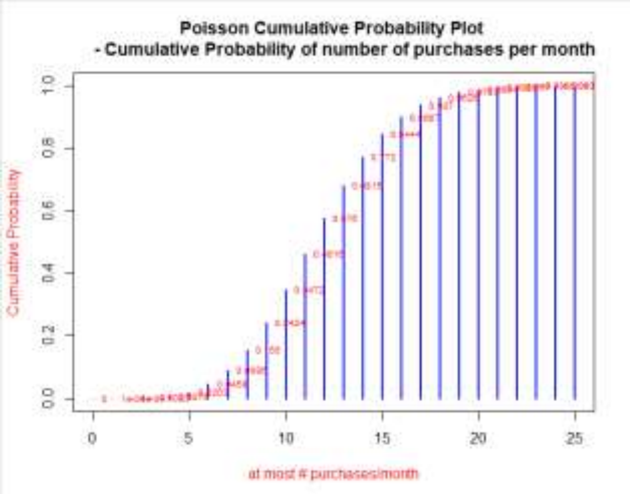
```
#Poisson Distribution
#
#dpois(x, lambda, log = FALSE) where
# x = number  of occurrences in a given time/space interval
# lamdda is mean number of occurrences in a given time/spce interval
#Example:
# Suppose that a customer purchases 12 (i.e., lamda = 12) times a month on average,
#what is the probability that they will make 180 purchases over the next 12 months? (i.e., x = 15/month)
print(paste("probability that they will make 180 purchases over the next 12 months or 15 per month?",
            round(dpois(15, 12, log = FALSE),4)))
#
print(paste("probability that they will make  at least 180 purchases over the next 12 months or 15 per month?",
            1 - round(ppois(14, 12, log = FALSE),4)))
#
#table of probabilities
nsize = 25
result <- vector("numeric", nsize+1)
cum_result <- vector("numeric", nsize+1)
xPx <- vector("numeric", nsize+1)
x2Px <- vector("numeric", nsize+1)
for (i in 0:nsize) {
  result[i+1] <- dpois(i, 12, log = FALSE)
  xPx[i+1] <- i*result[i+1]
  x2Px[i+1] <- i*xPx[i+1]
  cum_result[i+1] <- ppois(i, 12, log = FALSE)
}
round(result[1:nsize+1], 4)
round(cum_result[1:nsize+1], 4)
#
# Mean = sum of xPx
Exp_val = sum(xPx)
print(paste("The Empirical (mean) Expected Value is",round(Exp_val,4), sep = " "))
# Var = sum of X2Px - (sum(xPx)^2)
varian = sum(x2Px) - Exp_val*Exp_val
print(paste("The Empirical Variance is",round(varian,4), sep = " "))
```

```
[1] "probability that they will make 180 purchases over the next 12 months or 15 per month? 0.0724"
[1] "probability that they will make  at least 180 purchases over the next 12 months or 15 per month? 0.228"
[1] "The Empirical (mean) Expected Value is 11.9918"
[1] "The Empirical Variance is 11.9771"
```

Poisson. R

```
#
#Plot the probability mass function
plot((0:indx),result[1:len_result], type = "h", main = "Poisson Probability Plot
     - Probability of number of purchases per month",
     xlab = "# purchases/month", ylab = "Probability Mass",
     col  = "blue", col.lab ="red", lwd=2)
text((0:indx), result[1:len_result], round(result[1:len_result], 4), cex=0.6, pos=4, col="red")

# Plot the cumulative probabilities
len_result <- length(result)
indx = len_result - 1
plot((0:indx),cum_result[1:len_result], type = "h", main = "Poisson Cumulative Probability Plot
     - Cumulative Probability of number of purchases per month",
     xlab = "at most # purchases/month", ylab = "Cumulative Probability",
     col  = "blue", col.lab ="red", lwd=2)
text((0:indx), cum_result[1:len_result], round(cum_result[1:len_result], 4), cex=0.6, pos=4, col="red")
```





Practice with calculator

Book Problem 121 – Page 295

- **Problem:** The average number of children a Spanish woman has *in her lifetime* is 1.47. Suppose that one Spanish woman is randomly chosen.

  a. In words, define the Random Variable *X*. X = the number of children for a Spanish woman *in her lifetime*

  b. List the values that *X* may take on. X = {0, 1, 2, 3,...}

  c. Give the distribution of *X*. X ~ Poisson(1.47) assuming that the birth of each child is independent of the previous child.

  d. Find the probability that she has no children in her lifetime.

```
> print(paste("probability that she has no children in her lifetime",
+             round(dpois(0, 1.47, log = FALSE),4)))
[1] "probability that she has no children in her lifetime 0.2299"
```

  e. Find the probability that she has fewer children than the Spanish average.

```
> print(paste("probability that she has the same or fewer children than the Spanish average",
+             round(ppois(1.47, 1.47, log = FALSE),4)))
[1] "probability that she has the same or fewer children than the Spanish average 0.5679"
```

  f. Find the probability that she has more children than the Spanish average.

```
> print(paste("probability that she has more children than the Spanish average",
+             1 - round(ppois(1.47, 1.47, log = FALSE),4)))
[1] "probability that she has more children than the Spanish average 0.4321"
```

  g. Find the probability that she will have 2 more children, given that she already has 7 children
  =P(X=9| X=7) = P(X=9 AND X=7)/P(X=7); But given that each child birth is independent of the previous, P(X=9 AND X=7) = P(X=9)*P(X=7)/P(X=7) = P(X=9) = 0.

```
> print(paste("probability that she has 2 more children given she has 7",
+             round(dpois(9, 1.47, log = FALSE),5)))
[1] "probability that she has 2 more children given she has 7 2e-05"
```

Practice with calculator

Book Problem 119 – Page 295

- A manufacturer of Christmas tree light bulbs knows that 3% of its bulbs are defective. Find the probability that a string of 100 lights contains at most four defective bulbs using both the binomial and Poisson distributions.

- Solution;
  - As a binomial, “success” is a defective bulb, p=0.03, n =100 and we need P(X <= 4)

```
> print(pbinom(4, 100, 0.03, lower.tail = TRUE, log = FALSE))
[1] 0.8178548
```

  - As a Poisson, lambda = 3 (because we are looking at a 100 bulbs and we expect 3 to be defective, which is lambda) and we need P(X <= 4)

```
> print(ppois(4, 3, log = FALSE),4)
[1] 0.8153
```