MXB107 Assessment 1

Harry Wright

Semester 2 2022

NOTE THIS ASSESSMENT IS DUE ON 4 September BY 11:59 PM.

For this Assessment, we will use the following dataset:

The dataset `episodes` included in the MXB107 package for R contains records for 704 episodes of the *Star Trek* aired between 1966 and 2005. (Type `?episodes` for a detailed description of the data.)

Part 1: Summarising Data

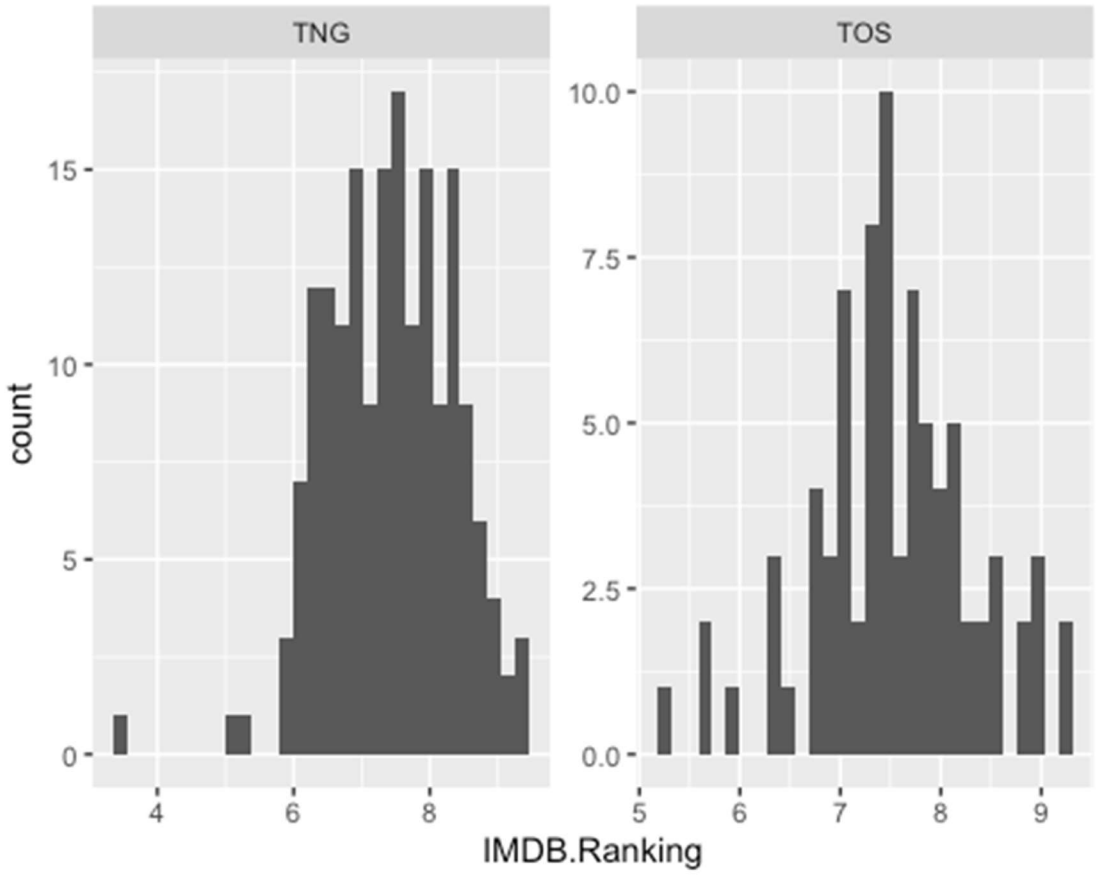
Question 1

a. Name three principles for good practice when creating graphical summaries of data.

Type your answer here:

1. The overarching title needs to describe the values and relationships in a graph/plot/summary.
2. The x and y axes need to be labelled with accurate descriptions
3. When comparing multiple data sets, their axes must match up and be consistent.

b. Identify three elements of the following graphical summary of data that should be corrected.



Type your answer here:

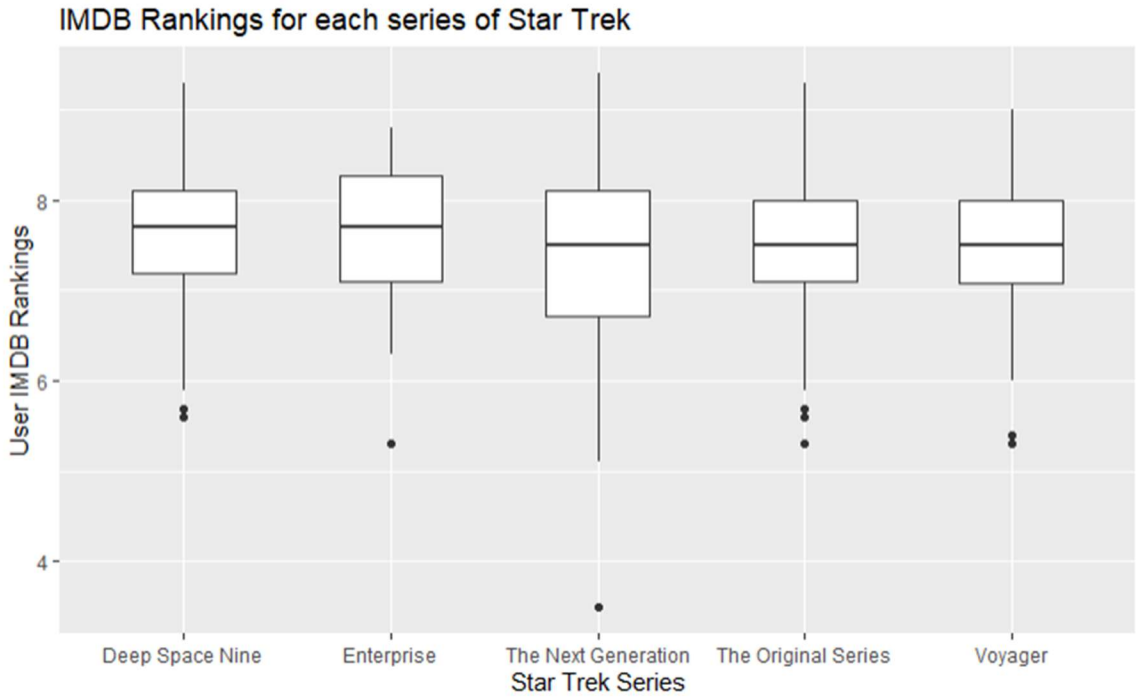
1. There is no overarching title for the two graphical summaries which describe the overall purpose for the graphical summary. The titles for each individual graph are very cryptic, which means a casual reader will not be able to determine what it's actually used for.
2. The axes for each graphical summary are confusing and do not provide any insight.
3. The two plots have axes which do not match which makes comparing them difficult.

c. Create a set of boxplots showing the IMDB rankings for each series of *Star Trek*. Discuss the results.

Show your code here:

```
##  The library MXB107 should be already loaded, if not type:
library(MXB107)
data(episodes)

#Plot the episodes data, using series name and IMDB ranking for axes
ggplot(episodes, aes(x=Series.Name, y=IMDB.Ranking))+
  geom_boxplot(width=.5)+
  xlab("Star Trek Series")+
  ylab("User IMDB Rankings")+
  ggtitle("IMDB Rankings for each series of Star Trek")
```

Results:



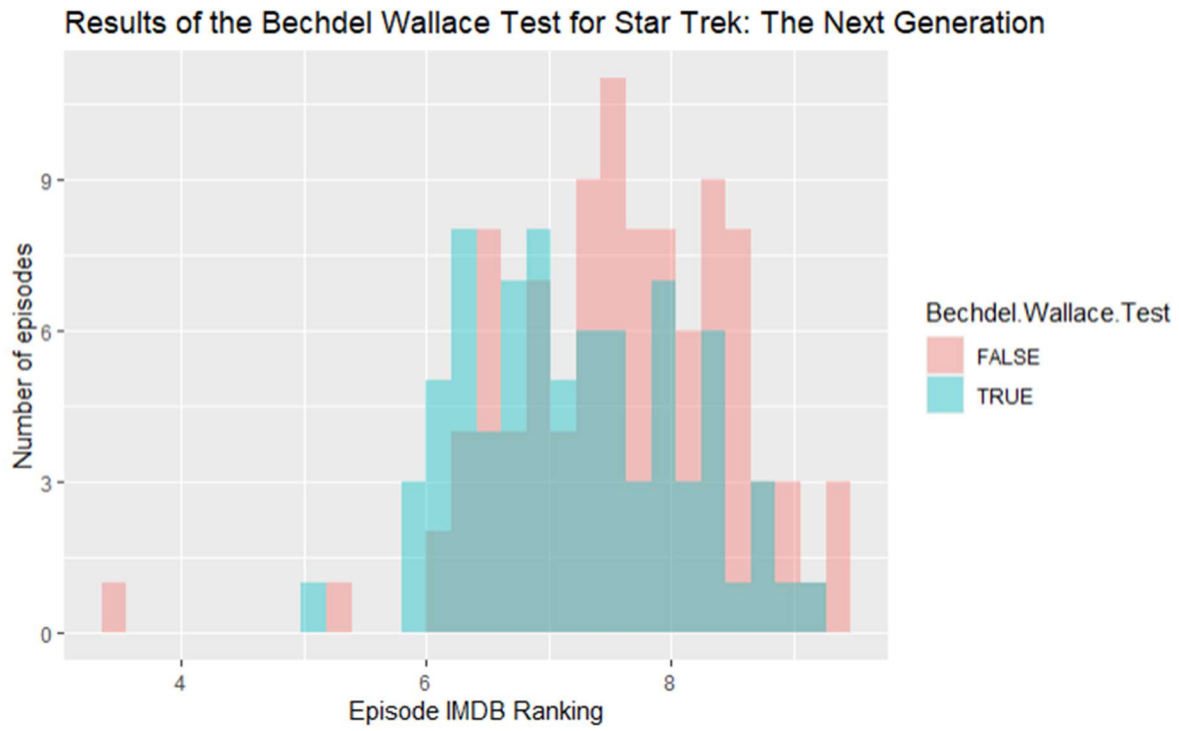
Type your answer here:

Each series stays pretty tightly around the 7.6(ish) mark and their outliers are pretty tightly packed around the 5-6 area, with the only exception being 'The Next Generation' which has the biggest extremes. The highest and lowest IMDB ranking.

d. Create a pair of histograms comparing the IMDB rankings for episodes of *Star Trek: The Next Generation* that pass the Bechdel-Wallace Test versus those that failed. Discuss the results.

Show your code here:

```
#Create a dataframe using a filter to only get episodes from TNG
TNG_episodes <- episodes%>%filter(Series=="TNG")

#Plot the TNG_episodes dataframe using the IMDB ranking as the x axes
and plotting the data against the results of Bechdel Wallace test for
TNG
ggplot(data=TNG_episodes, aes(x=IMDB.Ranking, fill =
Bechdel.Wallace.Test))+
  geom_histogram(position="identity",alpha=0.4)+
  xlab("Episode IMDB Ranking")+
  ylab("Number of episodes")+
  ggtitle("Results of the Bechdel Wallace Test for Star Trek: The Next
Generation")
```



Type your answer here:

The outliers from the boxplot earlier have appeared again. It ends up skewing the data towards the left (lower). One pattern is that despite the extremes of TNG episodes failing the test, the episodes passing the test seem to be quite tightly packed in terms of ranking.

Question 2

a. Identify and define three numerical summaries of centrality for data.

Type your answer here:

1. Mean: the average of the sum of a data sample
2. Median: the centre-most value in a data sample
3. Mode: the most frequently occurring value in a data sample

b. Identify and define three numerical summaries of dispersion for data.

Type your answer here:

1. Range: Difference between the largest and the smallest value in a sample
2. Standard Deviation: The square root of the variance
3. Variance: The average of the squared distance between observations and the mean

Question 3

a. For all 704 episodes of *Star Trek*, compute the standard deviation of their IMDB rankings using the definition of standard deviation and then use the empirical rule to estimate the standard deviation. Compare and discuss the results.

Show your code here:

```
#Define the mean of each episode's IMDB Ranking
episodes_mean <- mean(episodes$IMDB.Ranking)

#Define the standard deviation of each episode's IMDB Ranking
episodes_standard_deviation <- sd(episodes$IMDB.Ranking)

#find which values contain 68% of data
episodes_mean-episodes_standard_deviation;
episodes_mean+episodes_standard_deviation

#find which values contain 95% of data
episodes_mean-2*episodes_standard_deviation;
episodes_mean+2*episodes_standard_deviation

#find which values contain 99.7% of data
episodes_mean-3*episodes_standard_deviation;
episodes_mean+3*episodes_standard_deviation
```

Results:

episodes_mean $\approx$ 7.551

episodes_standard_deviation $\approx$ 0.7760

First 68% of data lies between: 6.774664 and 8.326756

Second 95% of data lies between: 5.998619 and 9.102802

Third 99.7% of data lies between: 5.222573 and 9.878847

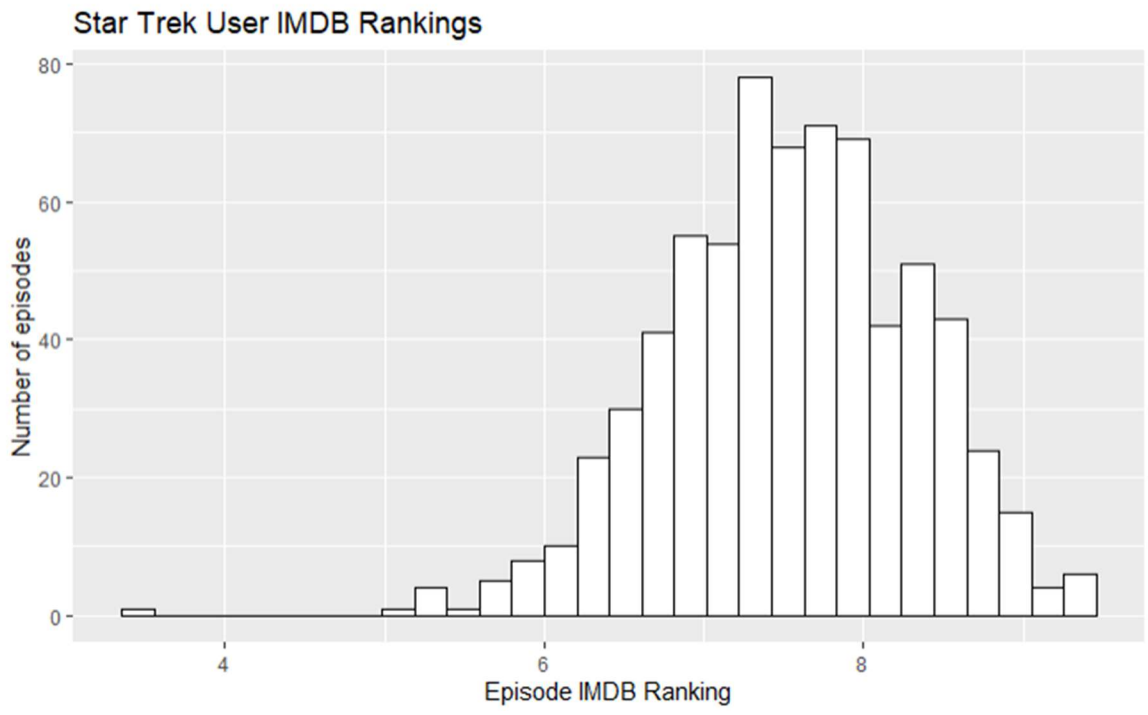
Type your answer here:

Considering that the Standard Deviation is less than 1, it shows that for the most part all 704 Star Trek episodes are relatively similar in terms of IMDB rankings. This isn't to say that ALL episodes follow this, as there are certainly some outliers.

b. For all 704 episodes of *Star Trek*, compute the mean and median of their IMDB rankings. Do the data appear to be skewed? Compute the skew of the data and plot a histogram of the episodes' IMDB rankings. Do they appear skewed? Compare and discuss the numerical results and your histogram.

Show your code here:

```
#Define the mean of each episode's IMDB Ranking
episodes_mean <- mean(episodes$IMDB.Ranking)

#Define the median of each episode's IMDB Ranking
episodes_median <- median(episodes$IMDB.Ranking)

#Define the standard deviation of each episode's IMDB Ranking
episodes_standard_deviation <- sd(episodes$IMDB.Ranking)

#Calculate the Skew of the Star Trek Episode's IMDB Rankings
num<-(episodes$IMDB.Ranking-episodes_mean)^3%>%mean()
den<-episodes_standard_deviation^3
skew<-num/den

#Plot a histogram of each episode's IMDB Rankings
ggplot(episodes, aes(x=IMDB.Ranking))+
  geom_histogram(fill = "white", colour = "black")+
  xlab("Episode IMDB Ranking")+
  ylab("Number of episodes")+
  ggtitle("Star Trek User IMDB Rankings")
```



Mean ≈ 7.551

Median $= 7.6$

Skew ≈ -0.3874

Type your answer here:

Since the 'skew' of the IMDB Rankings is a negative, it means that the "tail" of the histogra is longer on the left. This is confirmed when viewing the histogram plot, as there are values far down the left side. The mean and median of the episodes data are also very similar, which makes sense as they are calculations for the centrality of data.

Part 2: Computing Basic Probabilities for Events

Question 1

a. What is the classical definition of probability?

For a discrete, finite sample space, an event's probability is defined as the frequency of an outcome. Though the event must be bound on the interval (0,1) and the sum of the sample space must be 1. If the joint probability of the events is disjoint ($\emptyset$) then the conditional probability must equal zero.

Type your answer here:

b. What is the probability that a randomly selected episode of *Star Trek* will pass the Bechdel-Wallace Test?

Show your code here:

```
#To get the probability, divide the number of times the event occurs
by the size of the sample space
length(which(episodes$Bechdel.Wallace.Test==TRUE))/704
```

Probability a randomly selected episode will pass the Bechdel-Wallace Test ≈ 0.5199

Type your answer here:

The answer is ~0.5199 or ~52%, which means the probability of finding an episode which passes the test is slightly more than the probability of it failing the test.

Question 2

a. What is the definition of joint probability?

Type your answer here:

$$\Pr(AB) = \Pr(A) * \Pr(B) \ \ or \ \ \Pr(A \cap B) = \Pr(A)\Pr(B)$$

Joint probability is described as the conjoined area where the two sample spaces overlap.

b. What is the probability that an original series episode passes the Bechdel-Wallace Test?

Show your code here:

```
#Filter for ONLY the TOS episodes
TOS_episodes <- episodes%>%filter(Series=="TOS")

#To get the probability, divide the number of times the event occurs
by the size of the sample space, since it's 'The Original Series' that
means there are only 80 in the sample space.
length(which(TOS_episodes$Bechdel.Wallace.Test==TRUE))/80
```

length(which(TOS_episodes$Bechdel.Wallace.Test==TRUE))/80 = 0.0625

Type your answer here:

By filtering for only episodes of TOS you can determine the probability that any given episode FROM TOS ONLY will pass the Bechdel Wallace Test (BWT). The end result is 0.0625 or 6.25% chance a TOS episode will pass the BWT.

Question 3

a. What is the definition of conditional probability?

$$\Pr(A|B) = \frac{\Pr(AB)}{\Pr(B)}$$

When an event occurring, effects the probability of another event, this implies conditional probability. This doesn't just have to be an increase of chance though, as conditionally probable events can be disjoint, meaning one event consequently won't allow the other to occur.

b. What is the probability that an episode fails the Bechdel-Wallace Test, given that it is an episode from *Star Trek: Deep Space Nine*?

Show your code here:

```
#Create a table which contains only the series and their result with
the Bechdel Wallace Test
series_bechdel_table <-
addmargins(table(episodes$Series,episodes$Bechdel.Wallace.Test))

#Calculate the conditional probability that an episode fails the
Bechdel Wallace test given that it is an episode from Star Trek: Deep
Space Nine
series_bechdel_table[1,1]/series_bechdel_table[1,3]
```

series_bechdel_table:

```
     FALSE TRUE Sum
DS9     76  100 176
ENT     60   38  98
TNG    100   78 178
TOS     75    5  80
VOY     27  145 172
Sum    338  366 704
```

Probability episode fails Bechdel-Wallace given it's from Deep Space Nine is ≈ 0.4318

Type your answer here:

Using data from the question its possible to determine that ~0.4318 of episodes which are from Deep Space Nine will fail the Bechdel-Wallace Test.

Question 4

a. What is Bayes' Theorem

Type your answer here:

$$Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

b. Given that an episode passes the Bechdel-Wallace Test, what is the probability that it was from Season 3 of *Star Trek: Voyager*

Show your code here:

```
#Filter for only episodes from season 3 of star-trek voyager
VOY_3_episodes <- episodes%>%filter(Series=="VOY",Season==3)

#Calculate the conditional probability that an episode is from season
3 of star trek voyager GIVEN that it passed the test.
length(which(VOY_3_episodes$Bechdel.Wallace.Test==TRUE))/series_bechde
l_table[6,2]
```

series_bechdel_table:

	0	1	2	3	4	5	6	7	Sum
DS9	0	20	26	26	26	26	26	26	176
ENT	0	26	26	24	22	0	0	0	98
TNG	0	26	22	26	26	26	26	26	178
TOS	1	29	26	24	0	0	0	0	80
VOY	0	16	26	26	26	26	26	26	172
Sum	1	117	126	126	100	78	78	78	704

Probability episode is from Season 3 of Star Trek Voyager, given that it passed is ≈ 0.0464.

Type your answer here:

The probability is calculated by filtering for ONLY season 3 of star-trek voyager and inserting them into a data frame first. The condition probability that an episode is from season 3 of voyager given that it passed the test is found by dividing the number of season 3 voyager episodes by the number of episodes which passed the test overall.

c. Is this probability greater or less than the marginal probability that a randomly selected episode is from Season 3 of *Star Trek: Voyager*? Why?

Type your answer here:

You can calculate the probability of a season 3 Voyager episode using probability. In total there are 704 episodes, and 26 episodes in Season 3 of voyager, therefore:

$$\Pr(Random\ episode\ is\ from\ season\ 3\ of\ Voyager) = \frac{26}{704} \approx 0.0369$$

This result is indeed lower than the probability an episode is from season 3 of voyager given that it passed the test. This is because the sample space of episodes which pass the test are smaller than the total number of episodes.

Part 3: Modelling with Probability Distributions

Question 1

a. Define a Bernoulli random variable.

Type your answer here:

For $\Pr(X) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$

A single event which has two possible outcomes, success or failure. For example a coin toss which can only be heads or tails.

b. Assume I have a fair coin. What probability will I need more than two coin tosses to get a "heads"?

Show your code here:

```
##  The library MXB107 should be already loaded, if not type:

##  library(MXB107)

##  If after loading the library if the dataset episodes is not available,
type:

##  data(episodes)
```

Type your answer here:

The outcomes for two coin tosses is $S = \{HH, HT, TH, TT\}$. The probability you will need to throw more than two coins to get one heads is equal to 25%, as the only time this will occur is if you throw two Tails.

c. Define a geometrically distributed random variable and Write out the probability mass distribution for a geometric probability distribution. Define the process that gives rise to a geometrically distributed random variable in terms of Bernoulli trials.

Show your code here:

```
#Probability of hitting heads after n trials
dgeom(1:6,0.5)
```

Type your answer here:

Geometric distributions are used if you have independent trials and you perform them until success. Such as the number of times you throw a coin until you get heads. The PMD of a Geometric random variable is:

$$For\ \Pr(X = n), X \sim Geo(p) = (1-p)^{n-1}p$$

A process which gives rise to a geometrically distributed random variable in terms of Bernoulli trials is the probability of getting heads on a coin toss after n trials. This is because the process can only be success or failure.

dgeom(1:6,5):

n	1	2	3	4	5	6
$\Pr(X = x)$	0.25	0.125	0.0625	0.0313	0.0156	0.0078

d. If the overall proportion of *Star Trek* episodes that pass the Bechdel-Wallace Test is 0.52, then assume I begin watching episodes selecting them at random. How many episodes do I have to watch until the probability I see at least one episode that passes the Bechdel-Wallace Test is more than 95%?

Type your answer here:

The number of episodes you need to watch until one passes the test can be determined through Geometric distribution and (embarrassingly) guess and check. However, the problem with this is Geometric distributions are the probability until the FIRST success. Its best to rearrange the distribution for n as a compliment of the distribution:

$$\Pr(X = n) = 1 - ((1 - p)^{n-1}p)$$

Let $p = 0.48$, the compliment of an episode passing the BWT

Let $\Pr(X) = 0.95$, the desired probability

$$0.95 = 1 - ((1 - 0.48)^{n-1}0.48)$$

$$-0.05 = -0.48(1 - 0.48)^{n-1}$$

$$\frac{-0.05}{-0.48} = (1 - 0.48)^{n-1}$$

$$(n - 1)\ln(1 - 0.48) = \ln\left(\frac{0.05}{0.48}\right)$$

$$n = \frac{\ln\left(\frac{0.05}{0.48}\right)}{\ln(1 - 0.48)} + 1$$

$$n = 4.45874$$

∴ The viewer will need to watch an estimated $n \approx 5$ episodes (rounded UP) until the chances of them watching an episode which passes the BWT is greater / equal to 95%.

Question 2

a. I have a coin that comes up heads for a given coin toss with probability p. If I toss the coin n times, on average, how many heads should I get? What is the standard deviation for the random variable X = number of heads in n coin tosses?

Type your answer here:

$$E(x) = \sum_{\leq \in S} x \Pr(X) = 1 * p + (1 - p) * 0 = p$$

$$Var(x) = E(X^2) - \big(E(X)\big)^2$$

$$Standard\ Deviation = \sqrt{Var(x)}$$

b. Describe the binomial random variable in terms of Bernoulli trials. For what value of p is the variance for a binomial random variable maximised?

Type your answer here:

In terms of binomial variance, p=0.5 will result in the highest value. Anything else will end up resulting in a smaller variance.

c. What proportion of *Star Trek: The Original Series* episodes pass the Bechdel-Wallace Test? If I select ten episodes of *Star Trek: The Original Series* at random, what is the probability that I will see two or fewer episodes that pass the Bechdel-Wallace Test?

Show your code here:

```
#Filter for only episodes of TOS and put it into a dataframe
TOS_episodes <- episodes%>%filter(Series=="TOS")

#Create a table for TOS which contains their result from the Bechdel Wallace
Test which will be used for proportions
TOS_bechdel_table <- addmargins(table(TOS_episodes$Bechdel.Wallace.Test))

#Probability of an episode from TOS passing the Bechdel Wallace Test
TOS_BWT_pass_prob = TOS_bechdel_table[2]/TOS_bechdel_table[3]

#nCk combinatronic calculator
unordered_combinatronic <- function(n, k) {
  return((factorial(n))/(factorial(k)*factorial(n-k)))
}

#Binomial distribution function for less than or equal to i
bin_var_less <- function(n, i, p) {
  final_result = 0
  for(i in 0:i){
    result <- unordered_combinatronic(n,i)*(p^i)*((1-p)^(n-i))
    final_result = final_result+result
  }
  print(final_result)
}

#Find the probability that out of 10 episodes from TOS, 2 or less pass the
Bechdel Wallace Test
bin_var_less(10,2,TOS_BWT_pass_prob)
```

TOS_BWT_pass_prob = 0.0625

bin_var_less(10,2,0.0625) ≈ 0.979

Type your answer here:

By using binomial distribution, its possible to determine that out of 10 episodes from TOS, the probability of selecting two or less that pass the Bechdel Wallace Test is ~0.9790 or ~97.9%. These results are exclusively based off of the data and are exact.

Using probabilities:

$$X \sim Bin(10, 0.0625)$$

$$\Pr(X \leq 2) = \Pr(X = 2) + \Pr(X = 1) + \Pr(X = 0)$$

$$= \sum_{k=0}^{i=2} \binom{n = 10}{k} 0.0625^k (1 - 0.0625)^{10-k}$$

$$= 0.5245 + 0.3496 + 0.1049$$

$$= 0.979$$

d. Assume that I sample episodes at random from all 704 episodes of *Star Trek*, and the proportion of all episodes that pass the Bechdel-Wallace Test is 0.52. If I select 100 episodes at random from all the episodes of *Star Trek*, what is the probability that I see less than 50 episodes that pass the Bechdel-Wallace Test? Compute this using the binomial probability distribution, the Poisson probability distribution, and the Gaussian distribution. Compare and contrast the results.

Show your code here:

```
#nCk combinatronic calculator
unordered_combinatronic <- function(n, k) {
  return((factorial(n))/(factorial(k)*factorial(n-k)))
}

#The above function was proof of working the equation out, from this point on
I'll just use inbuilt commands for binomial
pbinom(49,100,0.52)

#Poisson distribution function for less than or equal to x
pois_dist_less <- function(rate,i) {
  final_result = 0
  for(k in 0:i) {
    result <- (rate^k)/factorial(k)
    final_result = (final_result+result)
  }
  final_result = exp(0-rate)*final_result
  print(final_result)
}

#The rate of a poisson distribution can be determined by the number of trials
multiplied by the probability of success (I think)
rate <- 100*0.52

#Poisson cumulative distribution of less than 50 episodes passing the BWT
pois_dist_less(rate,49)

#Normal distribution of less than 50 episodes passing the BWT, I'm not making
a function for this.
pnorm(49, mean=rate,sd=rate*(1-0.52))
```

pbinom ≈ 0.3082 (cumulative)

pois_dist_less ≈ 0.3721 (cumulative)

pnorm:≈ 0.3081 (Not using $\pm$ 0.5)

Type your answer here:

All three tests gave different probabilities, with binomial and Gaussian being the closest by far. This is to be expected as Gaussian/Normal distributions are very close to Binomial o

Question 3

a. Show that as $n \to \infty$ and $p \to 0$ the probability distribution of a random variable $X \sim Binom(n, p)$ converges to a Poisson probability distribution.

Type your answer here:

Poisson random variable takes on discrete values from 0 to infinity, with some rate or expected value lambda (λ). Consider a random variable X following a binomial probability distribution, defining lambda = np then the limit as $n -> \infty$ and $p \to 0$ is the Poisson probability distribution with rate lambda (λ).

Defining $\lambda = np$ implies that $p = \frac{\lambda}{n}$, which you can substitute;

$$p(x) = \frac{n!}{(n-x)!\,x!} * \left(\frac{\lambda}{n}\right)^x * \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

Expanding this results in:

$$= \frac{\lambda^x}{x!}\frac{n!}{(n-x)!}\frac{1}{n^x}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

Find a limit to continue:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n$$

As $n \to \infty$, $\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}$

All the other terms, except for $\frac{\lambda^x}{x!}$ goes to 1 as $n \to \infty$, this leaves:

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

This means that the mean and variance of $p(x)$ (Poisson Random variable) are:

$$E[X] = \lambda$$

$$Var[X] = \lambda$$

b. For *Star Trek: The Original Series* plot the probability distribution for the number of episodes out ten that would pass the Bechdel-Wallace Test. Use the Binomial and the Poisson distributions. Compare and discuss the results.

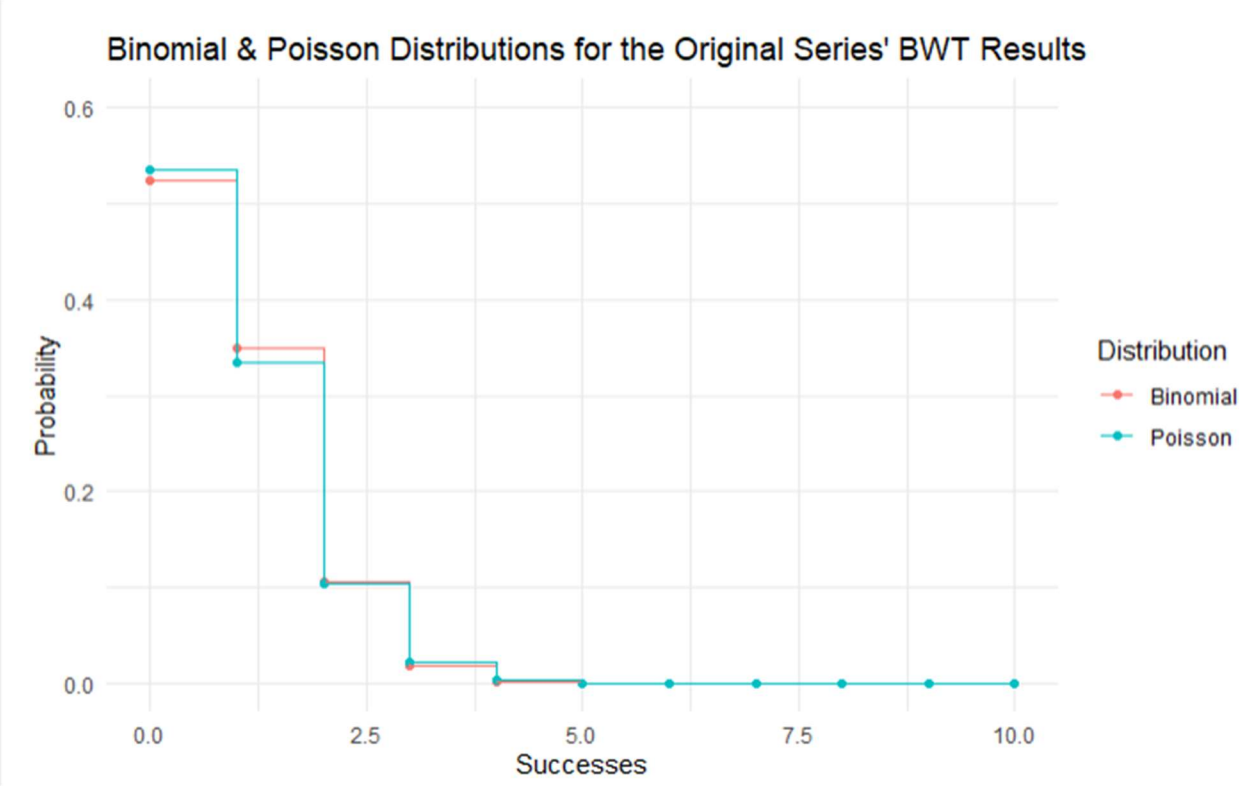
Show your code here:

```
#Filter for only episodes of TOS and put it into a dataframe
TOS_episodes <- episodes%>%filter(Series=="TOS")

#Create a table for TOS which contains their result from the Bechdel Wallace
Test which will be used for proportions
TOS_bechdel_table <- addmargins(table(TOS_episodes$Bechdel.Wallace.Test))

#Probability of an episode from TOS passing the Bechdel Wallace Test
TOS_BWT_pass_prob = TOS_bechdel_table[2]/TOS_bechdel_table[3]

prob_binomial <- dbinom(0:10,10, TOS_BWT_pass_prob)
prob_poisson <- dpois(0:10,(10* TOS_BWT_pass_prob))

df <- tibble(Successes = rep(0:10,2),Probability =
c(prob_binomial,prob_poisson),Distribution =
c(rep("Binomial",10+1),rep("Poisson",10+1)))

ggplot(df,aes(x=Successes,y=Probability))+
  geom_point(aes(color=Distribution, fill=Distribution))+
  geom_step(aes(color=Distribution))+
  theme_minimal()+
  xlim(0,10)+
  ylim(0,0.6)+
  ggtitle("Binomial & Poisson Distributions for the Original Series' BWT
Results")
```

TOS_bechdel_table:

FALSE	TRUE	Sum
75	5	80

TOS_BWT_pass_prob = 0.0625



Type your answer here:

When creating the probability distributions for Star Trek: The Original Series (TOS), it was assumed that using the overarching probability of 0.52 for passing the BWT would be incorrect. This was confirmed by filtering for only TOS episodes from the episodes dataframe, as only 5 episodes actually passed the test in TOS. The actual probability for a TOS episode to pass the test is only 0.0625, which was used for plotting the Binomial and Poisson distributions on-top-of each other.

They are shown to be incredibly similar when plotted on-top of each other, or when reviewing their results from one of the tables. Its shown that while it is quite likely you'll see at least one episode that passes the test, the probability quickly deteriorates.

c. What is the relationship between the Poisson and Exponential probability distributions?

Type your answer here:

Exponential distributions are just Poisson distributions with n=1. For $\Pr(X = i)$:

$$X \sim Pois(\lambda) = \frac{e^{-\lambda}\lambda^{i}}{i!}$$

$$X \sim Exp(\lambda) = \begin{cases} \lambda e^{-\lambda i} & i \geq 0 \\ 0 & i < 0 \end{cases}$$

Assume that the average episode is 45 minutes long. Given the probability that a given episode has a probability of passing the Bechdel-Wallace Test of $p = 0.52$, that is the equivalent 0.693 instances of passing the Bechdel-Wallace Test per hour of *Star Trek* viewing.

d. If I watch ten hours of *Star Trek* (assume the hours are completely random), what is the probability that I see more than seven instances of passing the Bechdel-Wallace Test.

Type your answer here:

$$\Pr(X > 7)\,, X \sim Exp(\lambda)$$

$Let\ \lambda = 0.693$ (Number of episodes which pass test per hour)

$$\Pr(X > 7) = 1 - \Pr(X \leq 7)$$

$$= 1 - (1 - e^{-\lambda})$$

$$= 1 - (1 - e^{-0.693*7}) = 1 - 0.9922$$

$$\approx 0.0078$$

e. What is the probability that I will have to watch more than three hours to see one instance of passing the Bechdel-Wallace Test

Type your answer here:

$$\Pr(X > 3) = 1 - \Pr(X \leq 2)$$

$$= 1 - (1 - e^{-\lambda i})$$

$$= 1 - (1 - e^{-0.693*2}) = 1 - 0.7499$$

$$\approx 0.2501$$

Question 4

a. Define the Z-score, or how we convert a Gaussian random variable to a Standard Gaussian random variable.

Type your answer here:

For $X \sim G(\mu, \sigma^2)$,

$$\Pr(X) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

Where $Z \sim G(0,1)$

$$Z = \frac{(x-\mu)}{\sigma}$$

b. For $X \sim N(4.3,2.7)$ find $Pr(X > 5)$

$$X \sim N(4.3,2.9), \Pr(X > 5)$$

Let $\mu = 4.3, \sigma^2 = 2.7$

$$\Pr(X > 5) = 1 - \Pr(X < 5) = 1 - \Pr\left(Z - \frac{x-\mu}{\sigma}\right)$$

$$1 - \Pr\left(Z - \frac{5-4.3}{\sqrt{2.7}}\right)$$

$$= 1 - \Pr(Z < 0.4260)$$

Show your code here:

```
#Calculate the cumulative standard normal function
1-pnorm(0.426,0,1)
```

1-pnorm(0.426,0,1) ≈ 0.3351

c. Assume that the IMDB rankings for *Star Trek* episodes follow a Gaussian distribution with $\mu = 7.55$ and $\sigma^2 = 0.60$. Based on the Gaussian distribution, what is the probability that a randomly selected episode will have an IMDB ranking of less than 7?

Show your code here:

```
#Calculate the cumulative normal function
pnorm(7,7.55,sqrt(0.6))
```

pnorm(7,7.55,$\sqrt{0.6}$) ≈ 0.2388

Type your answer here:

The number of episodes with an IMDB ranking of less than 7 is roughly equal to 0.2388 (23.88%) while $\mu = 7.55$, $\sigma^2 = 0.6$ and $\sigma = \sqrt{0.6}$.

d. Assume that the IMDB rankings for episodes of *Star Trek* follow a Gaussian distribution with $\mu = 7.55$ and $\sigma^2 = 0.60$. Based on the Gaussian distribution. What proportion of episodes have an IMDB ranking of over 7.9? What is the actual proportion of episodes with an IMDB ranking of over 7.9? Compare your results.

Show your code here:

```
#Find the actual proportion of episodes which have an IMDB ranking of
over 7.9 by dividing the table of episodes with an IMDB ranking
greater than 7.9, by the total of all 704 episodes.
length(which(episodes$IMDB.Ranking>7.9))/704

#Calculate the cumulative normal function
1-pnorm(7.9,7.55,sqrt(0.6))
```

Imdb_all_episodes[1,2]/Imdb_all_episodes[1,3] $\approx 216/704 \approx 0.3068$
1-pnorm(7.9,7.55,sqrt(0.6)) ≈ 0.3257

Type your answer here:

$$X \sim Normal(7.55, 0.6)$$

$$\Pr(X > 7.9) = 1 - \Pr(X < 7.9)$$

The real proportion of episodes over a 7.9 IMDB ranking is ~30.68%, while the calculated probability is 32.57%. While the real proportion is undeniably more accurate, the calculated value is still remarkably close.