Problem 1: Which of the following statements are true? _____

An Instagram study has determined that IF a user has a tag in their story, 12% of the time the user tags a friend and 35% of the time the user tags a geographical location.

A. The probability that a person with 560 tagged stories in their story feed, has tagged a friend in 65 or more of these stories is 0.63. (TRUE)

B. The probability that a person with 560 tagged stories in their story feed, has tagged a place in 65 or less of these stories is 0.41 (FALSE)

C. Since the sample size is large we can find the probability in part B with a normal approximation where the variable can be approximated with mean and standard deviation. (TRUE)

D. Since the success probability is low, we cannot find the probability in part B with a normal approximation. (FALSE)

Problem 2: Which of the following statements are true? _____

Assume in a University Graduate program, on average five students drop below a grade of 80% (identified as “struggling students”) in every semester. A typical semester contains 14 weeks. Assume that this program has 125 students. We want to find the probability that there will be no more than two struggling students in seven weeks.

a. We can assume that identifying a struggling student is a Bernoulli trial with success probability p = 5/14 (FALSE)

b. The trials are independent, then X = number of struggling students in a week is binomial distributed with n=125, p=5/14. (FALSE)

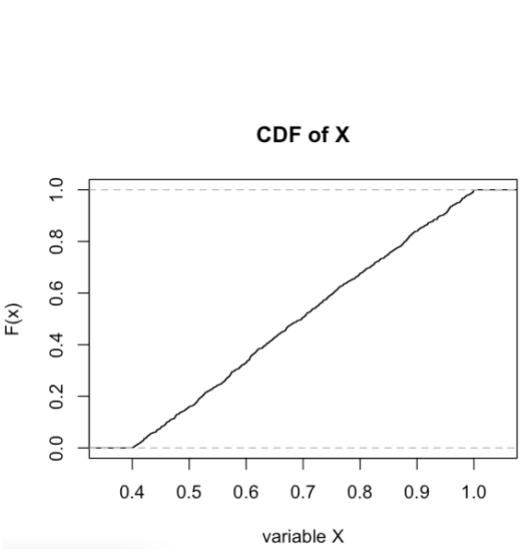
c. The probability that there will be no more than two struggling students in seven weeks is 0.543(TRUE because this is a poisson process. X = number of struggling students in 7 weeks, Lamda is the average struggling students in 7 weeks: 5*7/14)

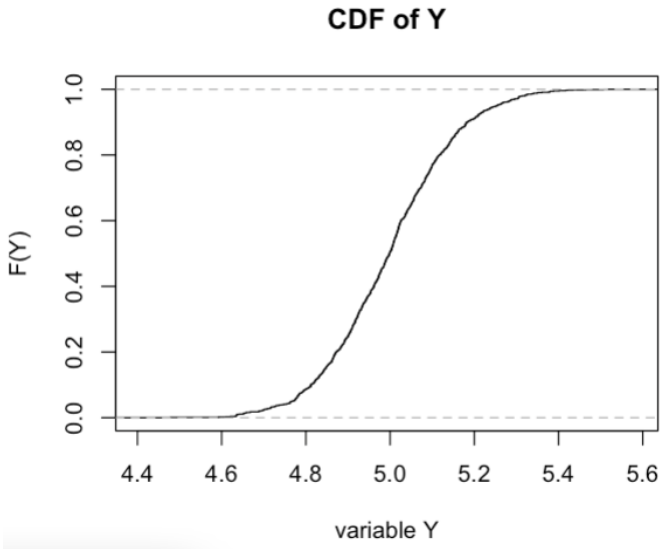
d. Since N is large and p is small, we can approximate this with a poisson distribution where lambda is Np = 44.64.(FALSE)

Problem 3: Which of the following statements are true ?

A. An empirical cumulative distribution function (CDF) is a non-parametric estimator of the underlying CDF of a random variable and ECDF (Empirical CDF) can be used to confirm a distribution of a variable. (TRUE)

B. According to the following ECDF plot, the X variable is uniformly distributed from X ~ Uni(0.4,1) and the Y variable has a continuous distribution. (TRUE)



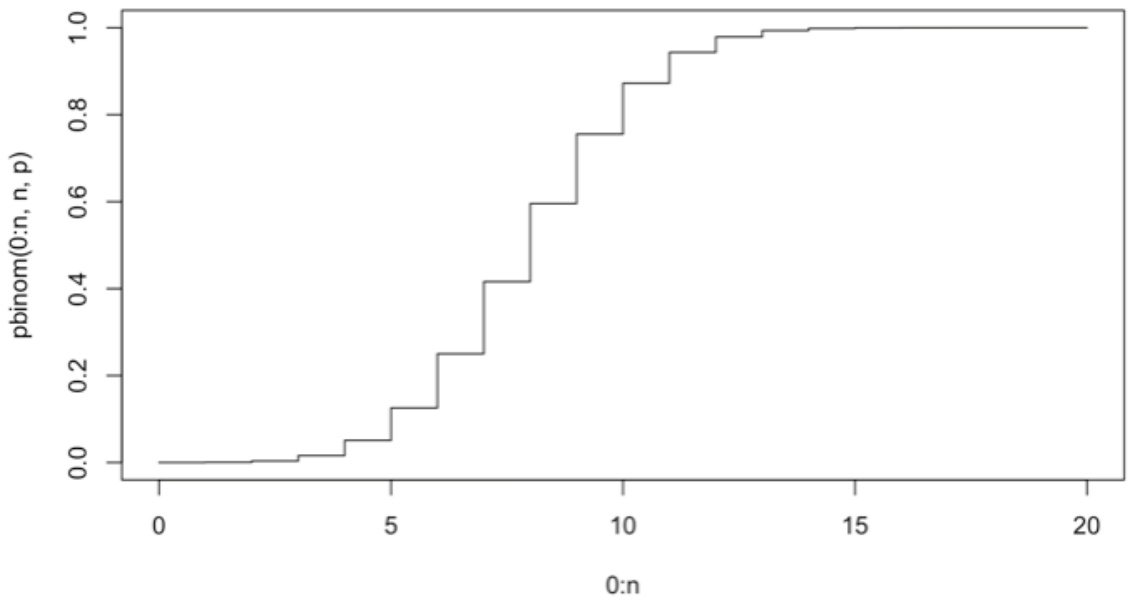


C. From the following results of the K-S test it can be seen that the Y variable has a normal distribution with mean 5 and standard deviation 0.15. (TRUE)

```r
ks.test(y,"pnorm",5,0.15)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  y
## D = 0.023801, p-value = 0.6227
## alternative hypothesis: two-sided
```

D. By looking at the following CDF plot it can be seen that P(X<8) is 0.6. (FALSE it is less than or equal to)



Problem 4: Which of the following statements are true?
Spotify's API gives you the ability to extract several audio features of a song. The available features that also have been used for this analysis are: https://medium.com/@boplantinga/what-do-spotifys-audio-features- tell-us-about-this-year-s-eurovision-song-contest-66ad188e112a
The first step was registering my application in the API Website(https://developer.spotify.com/documentat ion/web-api/) and getting the keys (Client ID and Client Secret) for future requests. The Spotify Web API has different URIs (Uniform Resource Identifiers) to access playlists, artists, or tracks information.

Using this Spotify API, we can get audio features of songs for different musicians. Assuming that there are only 400 musicians belonging to the high "Danceability" category (Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.) and 132 of them are pop music artists and the rest are Rock music artists.

A. If I randomly extracted 20 songs that belong to this Danceability category, the probability that more than 10 of them will belong to Rock music artists is 0.918. (TRUE)

B.If I randomly extracted 20 songs that belong to this Danceability category, the probability that more than 10 of them will belong to Pop music artists is 0.968. (FALSE)

C. Now imagine I am randomly picking balls without replacement, from an urn containing red(132) balls and blue(268) balls, the probability of choosing exactly 15 red balls out of the 40 balls I picked, will be 0.112. (TRUE)

D. Now imagine I am randomly picking balls with replacement, from an urn containing red(132) balls and blue(268) balls, the probability of choosing exactly 15 red balls out of the 40 balls I picked, will be 0.891. (FALSE).

Problem 5. Which of the following statements are true?

A common daily lottery game involves the drawing of three digits from 0 to 9 independently with replacement and independently from day to day. Lottery watchers often get excited when all three digits are the same, an event called triples.

A. If X is the number of days without triples before the first triple is observed, then X has a geometric distribution with *p*(the probability of obtaining triples) is 10/1000. (TRUE)

B. The Probability that there were more than 10 days without triples before the first triple is observed is 0.904 (FALSE)
c. The probability that there were 100 days until observing 4 triples is 0.038. (FALSE)
d. In a summary in Binomial distribution we are looking at How many successes; in Geometric distribution we are looking at How many failures until a success and in Negative Binomial we are looking at Number of failures until r successes. (TRUE)

Problem 6. Which of the following statements are true?

Consider DSAN students arriving at the restaurant Tombs in Georgetown, DC. Since the students have different schedules there's a possibility of students going to this restaurant all day long for meals.

A. If 5 students arrive every 2 hours, then the number of DSAN students arriving at the restaurant is a Poisson process with the rate parameter lambda is 2.5. (FALSE cannot exactly say without considering the units of the variable X)

Part A seems like a trick question to me: = in poisson process we need to consider the units of the variable

B. If 5 students arrive every 2 hours, then the probability that the time is less than 30 minutes until the next student arrives at the restaurant is 0.713. (TRUE)
P(X<30) = pexp(30,2.5/60)

```
> pexp(30,2.5/60)
[1] 0.7134952
```

C. The homogeneous Poisson process is based on a constant rate of events and In contrast to the homogeneous Poisson process,the intensity function (rate function) of an inhomogeneous Poisson process is a nonconstant function. (TRUE)
C. The exponential distribution is memoryless because it has the same distribution regardless of how much time has already elapsed if you haven't observed anything during that time period.(TRUE)

Problem 7. Which of the following statements are true?
A. Summation of independent random uniform variables has a Uniform distribution (FALSE)
B. Summation of independent random Normal variables has a Normal distribution (TRUE)
C. If X has a standard normal distribution, squared of X has a Chi Square Distribution with m degrees of freedom. (FALSE df is 1)
D. Exponential distribution and the chi-square distribution are coming from the Gamma distribution.(TRUE)

Problem 8. In each of the following examples, try to indicate whether the Poisson process would be a good model.

a. Number of days until a hurricane in Florida. (FALSE because it is time between events ~exp)
b. Number of times a chicken lays its eggs in a given day.(TRUE)
c. Number of trains arriving at Rosslyn metro station (FALSE because it is not random, they come at a certain time)
d. Number of DSAN students winning the second year scholarship each year (TRUE)

Problem 9.

Which of the following statements is/are true?

A. X is normally distributed with mean 3 and standard deviation 5. According to the following output the probability of X being equal to 5 is 0.0736. (FALSE dnorm gives the density not the probability)

```
> dnorm(5,3,5)
[1] 0.07365403
```

B. Assume that a team researches the age of people in a city who watch the TV series 'The Crown" on Netflix. It has found that the age follows a normal distribution with mean 43 and variance 471. The probability that this TV series is watched by adults between 40 years to 60 years old is 0.338 (TRUE)
B. If the above experiment has expanded to other cities then assume that the age became harder to explain by the normal distribution but still it had a symmetric distribution with heavier tails. In this case we may be able to model the age by a t-distribution. (TRUE)
B. On average, I can work on my laptop for 24hrs (if not used for intense computational tasks) before the battery dies. If I forget to bring my charger to campus, the probability that I will be able to use the laptop in my 2 and a half hours class without the battery dying(in the class I will be only using the laptop for teaching purposes) is 90%. (TRUE ~Exp with lambda 1/24)

Problem 10.

Which of the following statements are true?

A. For continuous variables with a continuous probability distribution, the likelihood and the probability density function values at a given point are not the same. (FALSE it's the same)
B. For a discrete variable with a discrete probability distribution, the exact probabilities are found by the probability mass function. (TRUE)

E. For continuous variables with a continuous probability distribution, the probabilities are found by integrating the probability density function. (TRUE)
E. For continuous variables with a continuous probability distribution, the probability density function can be found always by differentiating the cumulative density function (CDF) (FALSE not, if it is not differentiable)