Statement of Problem: Computer Worksheet 2019/20 [Hospital Admissions]  
Download the dataset project\_data.csv.. This dataset contains genuine NHS data on the  
number of people attending Accident and Emergency or Minor Injury Units (or other) across Scotland  
in June 2019 available from NHS Scotland Open Data.  
Read the data into R and check that it has been read in correctly. The columns contain the following  
information.  
Month Year and Month  
TreatmentLocation: A unique code identifying the hospital  
DepartmentType: Emergency Department or Minor Injuries Unit/Other  
NumberOfAttendances: Total number of instances of a patient arriving at  
the department or unit  
NumberMeetingTarget: Number of instances where the patient is admitted, transferred or  
discharged within 4 hours of arrival  
This worksheet is going to look into two aspects of these data. First of all the number of people  
attending an emergency department and secondly to investigate whether the NHS target on waiting  
times (4 hours) is being met. We will consider only the data from Emergency Departments.  
  
Q1. Create a new data frame that contains the data for Emergency Departments. [1 mark]  
Let Xi denote the number of attendances at emergency department i = 1,...,n in June 2019, where n is the number of emergency departments in the data. Let λ denote the average monthly rate of attendance at an emergency department. We will assume Xi Po(λ) where the rate is the same at all hospitals.

#Replace column names

#Collect only data for emergency department

#Represent on dataframe

df=dataframe

Q2. Calculate a point estimate, λb, for the average monthly rate of attendances at emergency departments in Scotland in June 2019. [1 mark]

Point estimate: The probability is the arithmetic average of measurements of the experimental outcome, and the sampling distribution would be normal. There are different methods of calculating point estimates, including method of moments estimates, maximum likelihood methods, least squares estimation, Bayesian methods, and others.

*# I don’t know what this is: PointEstimate = df.averageattendance,λ*

Ans > x <- mean(Xi, na.rm=TRUE)  
n

Q3. Let X = XXi be the total number of attendances across all n emergency departments in i=1 June 2019. State the distribution of X. [1 mark]

#Calculate the probability of X

> pX <- X/Xi

#Calculate the standard deviation of Xi

Ans > sd <- sd(Xi, na.rm=TRUE)

#pnorm helps calculate the distribution

> pnorm(X, x, sd)

#Visualization of distribution

>   
n  
Q4. Simulate a large number of values (at least 1000) from the distribution of X = XXi.  
i=1. Using the simulated random deviates identify an approximate 95% confidence interval for λ. Interpret the interval in the context of the data. [2 marks]

Length/samplesize > n <- 1000

#Calculate z score for 95% confidence interval [qnorm(0.975)] multiply by sd/sqrt(n) to get error

> error <- qnorm(0.975) \* sd/sqrt(n) #addding 2.5 because we have 2.5% both tails or (1 + confidence)/2

Lower bound of confidence interval > lc <- λ - error

Upper bound of confidence interval > uc <- λ + error

Q5. (a) We have assumed Xi Po(λ) for i = 1,...,n. State the Normal approximation for  
Xi (as a function of λ). [1 mark]  
(b) Let us now assume that the number of attendances follow a Normal distribution such that Xi N(µ,σ2). Find point estimates for µ and σ2. [2 marks]  
(c) By considering your answers to parts (a) and (b), comment on the validity of the original Poisson assumption. [1 mark]  
(d) Calculate a 95% confidence interval for µ. Interpret the interval in the context of the data. [2 marks]  
  
This question is not directly related to the hospital data. The purpose of this question is to investigate confidence intervals. While it is the case that we normally have only one sample of data available to us in practice, we can use computer simulation to simulate taking many samples from a defined population.  
Q6. (a) Assume we have a population that is normally distributed with µ = 1000 and σ2 = 45000. Simulate 100 samples each of size m = 200 from this population and store the samples. [1 mark]  
(b) For each sample compute the sample mean, the lower and upper bound of a 90% confidence interval for µ (assuming both µ and σ2 are unknown) and store the results.  
[1 mark]  
(c) Count (using R commands) and comment on the number of your confidence intervals that contain the true parameter value µ = 1000. [1 mark]  
(d) Plot your 100 confidence intervals for µ with the estimate of the sample mean marked by a dot for each sample. Add to your plot a line indicating the true value of µ = 1000. Colour the intervals (and dots) that do not contain truth a different colour to those that do. (Hint: I suggest constructing the plot one step at a time using a series of commands rather than in one single step.) [2 marks]  
  
We return to the NHS data. In this section we will be investigating whether the target waiting time was met in June 2019. The NHS aims to admit, transfer or discharge 95% of emergency department patients within 4 hours of arrival.  
Q7. We wish to test, for each emergency department in turn, whether the target of 4 hours was met for 95% of patients in June 2019. State the null and alternative hypothesis for such tests in terms of an appropriate parameter. Choose and state your choice of an appropriate test statistic. State the distribution of your test statistic under the null hypothesis. (Do not use a normal approximation.) [2 marks]  
Q8. For each emergency department in turn, calculate a p-value for the hypothesis test you defined in Question 7. [2 marks]  
Q9. Summarise the findings of the hypothesis tests from the previous question. [3 marks]  
  
In this final section, we are going to take a look at the power of a hypothesis test.  
Q10. Consider the hypothesis test of whether Perth Royal Infirmary (Location T202H) met the target of 95% of patients waiting fewer than 4 hours (the test from the previous section).  
(a) Calculate the power of the hypothesis test (assuming a significance level of 5%) at p = 0.9 and p = 0.97, where p is the true proportion of patients who met the target.  
By considering many different values of p plot the power function of the test. need some space [3 marks]  
(b) Calculate the smallest sample size required such that the power of the test is greater than 0.7 for all values of the true proportion p beyond 0.01 of the null hypothesis when the significance level of the test is 5%. Use comments to describe what your code is doing. [2 marks]  
(c) Modify the code you used to answer part (b) to turn it into a function. The function should have a single argument that is the smallest difference between the null and the true value of p you wish to consider (this was 0.01 in Part (b)). The function should return the minimum sample size such that the power of the test is greater than 0.7 for all values beyond the smallest difference when the significance level of the test is 5%.  
Use your function to find the minimum sample size required when the smallest difference is 0.01 (you should obtain the same answer as in part (b), this will show you your function is working). Once you are certain the function is working as intended find the minimum sample size required when the smallest difference is 0.025 and 0.05.  
[2 marks]