|  |
| --- |
|  |

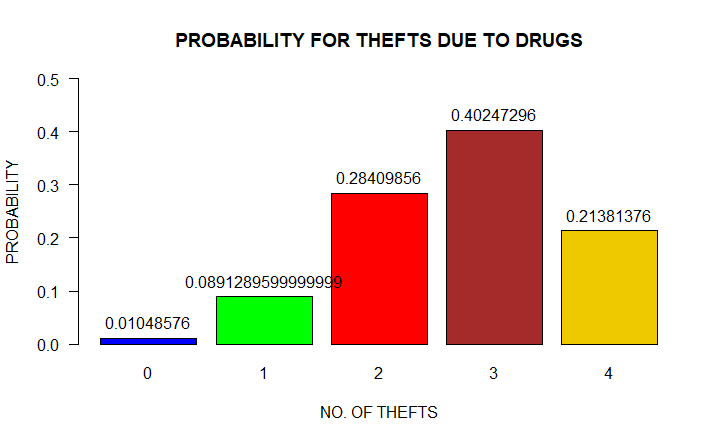
Q1

1. In a certain part of a city, the need to buy drugs is known to be responsible for 68% of all thefts. What is the probability that three of the next four thefts are due to the need for drugs?

Ans: The probability that three of the next four thefts are due to the need for drugs is 0.402.

1. Use the information from above (i) to answer this question. Provide a bar plot that shows the probabilities for zero, one, two, three and four thefts being due to drugs? The plot should be worthy of use in a report.

Ans: The bar plot below shows the probabilities from zero to four thefts being due to drugs.

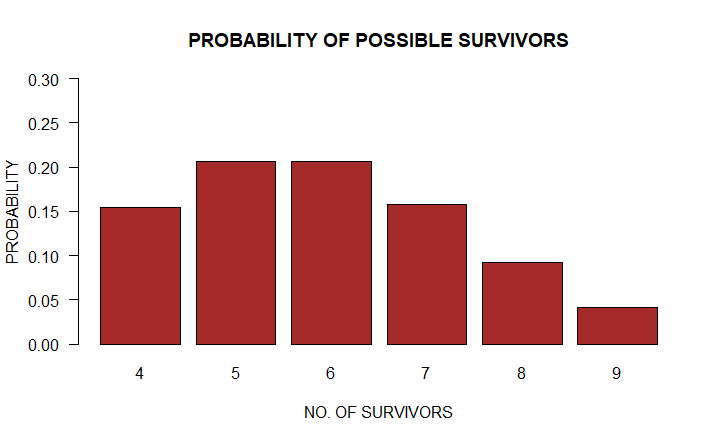


1. The probability that a patient recovers from a lymphoma is 0.4. If 14 people are known to have this disease, what is the probability that either 4, 5, 6, 7, 8 or 9 of these people will survive?

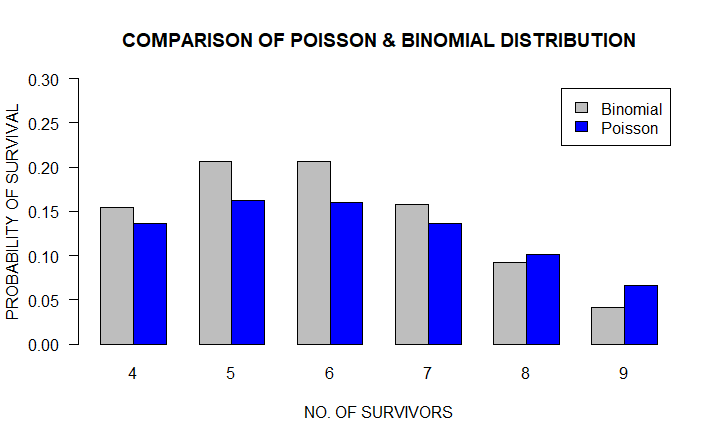
Ans: If 14 people are known to have lymphoma, the probability that either 4, 5, 6, 7, 8 or 9 people will survive is 0.858.

1. Use the information from above (iii) to answer this question. Provide a bar plot that shows the probabilities for each number of possible survivors? The plot should be worthy of use in a report.

Ans: Bar plot below showing probabilities for each no. of possible survivors.



1. Generate a Poisson approximation to (iv). Provide a comparative plot for this approximation? Make use of a legend in your comparative plot and make this graph worthy of use in a report. Describe the closeness of the distributions and provide an explanation of the result? Use no more than a total of five sentences.

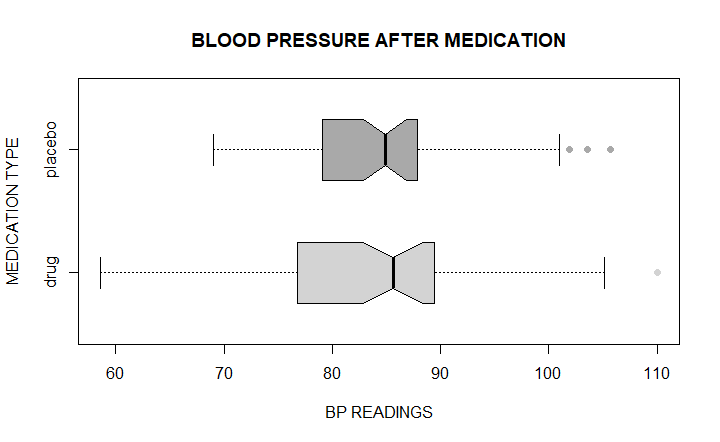


Ans: The barplot on the previous page compares probabilities for each no. of survivors when calculated thru two different methods, Binomial approximation and Poisson Approximation. It can be seen from the barplot that when the sample size is small, which is less than 8, the probabilities for each no. of survivor from Binomial approximation is higher than the probabilities from Poisson approximation for the same no. of survivors. However, when the no. of survivors is greater than 7, the probabilities from Poisson are higher than that of Binomial approximation. Overall, there is not much difference in probabilities from each approximation for each no. of survivors.

Q2) A new medication had been designed to reduce the chance of obtaining the flu. However it is suspected that the medication has a side effect of increasing the blood pressure of its consumer. To examine this suspicion an experiment was designed and necessary data collected. Your task is to analyse this data.

1. Compute the "mean after medication blood pressure" for those that received the medication and the "mean after medication blood pressure" for those that received the placebo. Also produce a boxplot for "after medication blood pressure", with respect to the two medications used? Make sure the boxplot is worthy of use in a report.

Ans: The mean blood pressure after medication for those who received the drug is 83.855 while mean blood pressure after medication for those who received placebo is 84.921. The box plot graph below also summarizes the information for “Blood pressure after medication”.



1. Using the above results (i), does the active drug appear to increase blood pressure? Is there a flaw in the approach used to make this judgement, if so, how could it have been avoided? Your response must consist of no more than six simple sentences.

Ans: from the results in part 1 of this question, there is no significant difference in the mean blood pressures for those who took the active drug and those who took placebo. Therefore, it can be determined that the active drug does not cause an increase in blood pressure.

The plot above shows some outliers for both drug and placebo; these outliers affect the value and quality of mean. Hence in this case, mean value of blood pressure is not the right parameter to judge the side effect of the drug. A better way to is to build a hypothesis and then perform tests to get a more accurate result.

1. Answer this question in light of any potential flaws identified above (ii). Using an appropriate set of hypotheses, determine if the active drug results in an increase in blood pressure, when compared to using no drug?

Ans: The objective of hypothesis testing is to determine whether active drug causes an increase in blood pressure or not.

Hypothesis:

Null Hypothesis: H0 => X1 = X2

Alternate Hypothesis: H1 => X2 > X1

Where:

* X1 is the mean blood pressure before medication,
* X2 is the mean blood pressure after medication

To briefly explain the equation, null hypothesis states that the mean blood pressures for before medication and after medication are equal. On the other hand, alternate hypothesis suggests that post medication mean blood pressure is greater than the mean blood pressure before medication. To test the hypothesis, we perform a paired t-test which will provide us with mean differences.

Assumptions:

* Data normally distributed
* Sample size is large enough
* Equal variances

Result:

t-value = 5.449

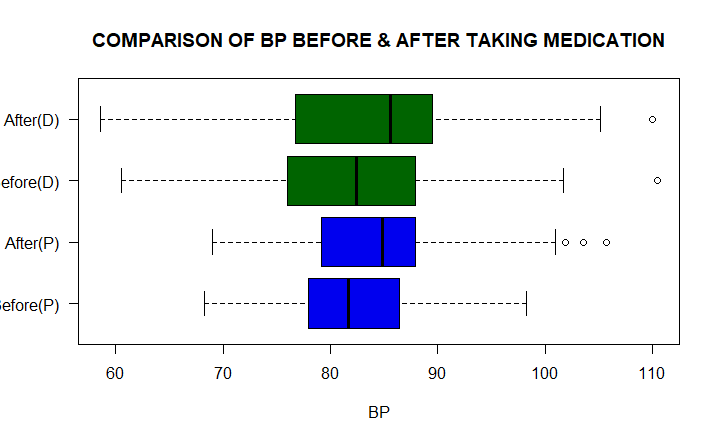
p-value = 1.855e-07

at 95% confidence interval = 1.639289 Inference

Mean of the differences = 2.357719

Alternate Hypothesis: true difference in means is greater than 0.

The associated p-value for t-value of 5.449 is quite small which suggests that there is some relation between drug and increased blood pressure. Also, the true difference in means is greater than 0. Therefore we reject the null hypothesis and claim that active drug does cause an increase in blood pressure after medication.



The boxplot above endorses the results from hypothesis testing that active drug tends to cause an increase in blood pressure.

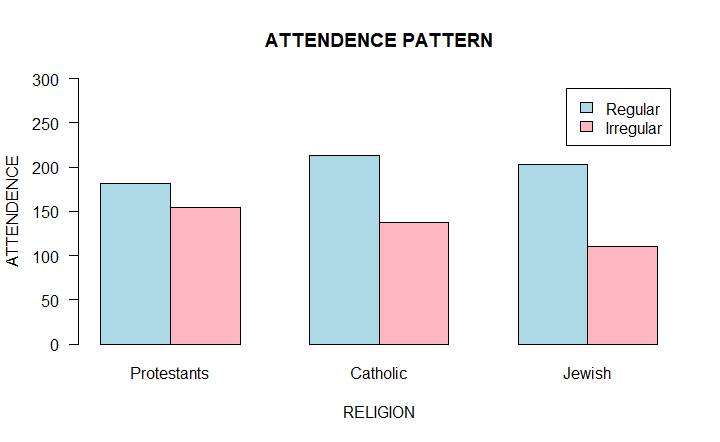
1. What is the 90% confidence interval for the measured difference, according to the hypothesis test conducted above (iii)? Provide a brief interpretation of this interval and what it could mean with respect to the change in blood pressure affected by the two medications?

Ans: The 90% confidence interval for the measured difference in means is 1.799 which means that the upper range for the population mean is 85.64 while the lower range is 82.05. If multiple samples were drawn from the same population and a 90% CI calculated for each sample, we are 90% confident that the population mean would be found within this range.

1. Briefly describe the role of a placebo in experiments such as the above?

Ans: Placebo is used as a control variable in the experiments to be used with the actual variable to test its effectiveness and evaluate its performance. In this case, placebo will only provide a psychological effect to the consumer that he has taken a drug but no physiological effect.

Q3i) Produce a bar plot that clearly shows the distribution of religion versus attendance pattern. Make sure the bar plot is worthy of use in a report. Provide a brief interpretation of what the plot shows?



Ans: The bar plot above shows the attendance pattern of member from different religions. A common pattern in all three religions is that regular church members are more than irregular church members for all religions. There is little difference in regular and irregular members for Protestants religion while the same difference for Catholic and Jewish is high.

ii) Create suitable hypotheses to test whether a difference exists with respect to religion and attendance patterns. The hypotheses themselves should be brief and to the point. In addition, also provide a sentence or two explanation for each of them.

Ans: The objective of this hypothesis test is to determine whether there is a significant relationship between a religion and attendance of its church members.

Hypothesis:

Null Hypothesis: H0 => There is no relationship between religion and attendance pattern of its members

Alternate Hypothesis: H1 => There is significant relationship between religion and attendance pattern

We perform a test of independence to test our hypothesis since both variables are categorical.

iii) Using the above hypotheses (ii), execute two versions of a hypothesis test:

* using a test that exploits a theoretical distribution
* using a simulation that involves many repetitions

Briefly describe the details of these two tests? You don't need to interpret the results here, rather, just how the tests were performed and the high-level purpose of the steps used.

Ans: Chi square test gave us a p-value of 0.01947.

Simulation technique with 2000 repetitions gave us a p-value of 0.021.

For a test that exploits a theoretical distribution, we performed a Pearson’s chisq.test which gave us a small pvalue of less than 0.05, rejecting the null hypothesis. For the simulation part, we used a replicate function with 2000 repetitions. In this simulation, we first sampled the attendance pattern and sampled the religion. After that, we produced a table of sampled data and calculated expected values using “outer” function. We then ran chi-square test for this simulation and calculated the p-value for a sample with 2000 repetitions. The p-value was 0.021 which is smaller than 0.05. Hence the results from simulation technique also rejected the null hypothesis.

Q4) **Relationship between blood pressure and age for Pima females**

1. Produce a plot that shows the relationship between age and blood pressure, showing the independent variable on the x-axis. Include a simple linear regression model for this data in the above plot. Make this plot worthy of use in a report; not a work of art, just something intelligible.

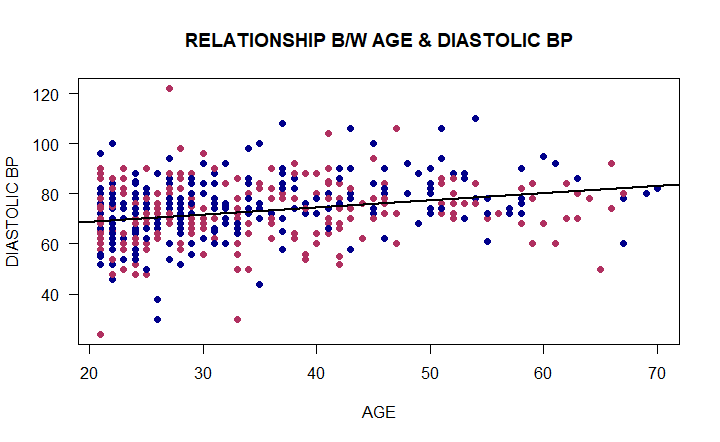
Ans: Linear regression model -> Y= a +bx

Coefficient = 0.2852

Intercept = 63.2590

Therefore, the equation is:

Y(Diastolic)=63.25898+0.28516\*Age



(ii) Using an appropriate method, calculate the correlation between the plotted variables above (i) and briefly provide an interpretation. Include a brief statement why you considered the chosen method was appropriate.

Ans: We use Pearson’s coefficient method to test the correlation between the two variables as it is an effective method when the variables are continuous. The pearson’s coefficient is 0.277 which means there is a positive correlation between the two variables.

(iii) Using the above (i) linear regression model, predict the diastolic blood pressure of a 39 year old female. Although this is a point estimate for the population, briefly consider the limitations of this estimate for 39 year old females in general (Pima population)?

Ans: We use the predict function to predict the diastolic blood pressure of a 39 year old female. The predict blood pressure is 74.38.

The limitations of this estimate is that every time a sample is taken from the population, the point estimate changes. This could lead to inaccurate results.

iv) Determine a 90% confidence interval for the diastolic blood pressure of 39 year old females. Considering the results of (iii), why does this estimate provide better accuracy with respect to 39 year old females in general?

**Ans:** The 90% confidence interval of the mean is 1.135 which means that the upper range for the population mean is 75.515 while the lower range is 73.245. This sample provides a better accuracy because if multiple samples were drawn from the same population and a 90% CI calculated for each sample, we are 90% confident that the population mean would be found within this range.