Questions

To each of the ‘is ... statistically sigifnicant’ question below, set up a hypothesis test. Write down • H0 and HA • What test you use in R: name of test, one-sided or two-sided • Your chosen signiﬁcance level • Report the p-value you obtained • Give a conclusion

**Titanic dataset**

Consider the Titanic dataset.

1. (Descriptive question) What are the survival rates of children? Of adult women? Of adults?
   1. Get the survive rate of children: 47.706%, adult women survive rate: 74.35294%, and adult survive rate: 31.26195%
2. Do children have signiﬁcantly better survival rate than adults?
   1. H0 : children does not have significantly better survival rate than adults
   2. Ha : Children have significantly better survival rate than adults. (One-sided)
   3. What test you use in R:

Since this is a categorical versus categorical, we use fisher’s exact test. From the table of survive vs. age, we can observe that children has higher survive rate than adult, so we want to use one-sided test and want to know if the survive rate of children is significantly higher.

* 1. Significant level alpha = 0.01
  2. P-value: 7.417e-06 = 0.000007417 = 0.0007417%
  3. Conclusion: If H0 (children does not have significantly higher rates of children), then we will observe (S >= s) the children have survived significantly more with probability 0.0007417%. This is pretty small, so we conclude that most likely, Ha(children have significantly better survival rate than adults) is better than the null hypothesis at explaining the data.

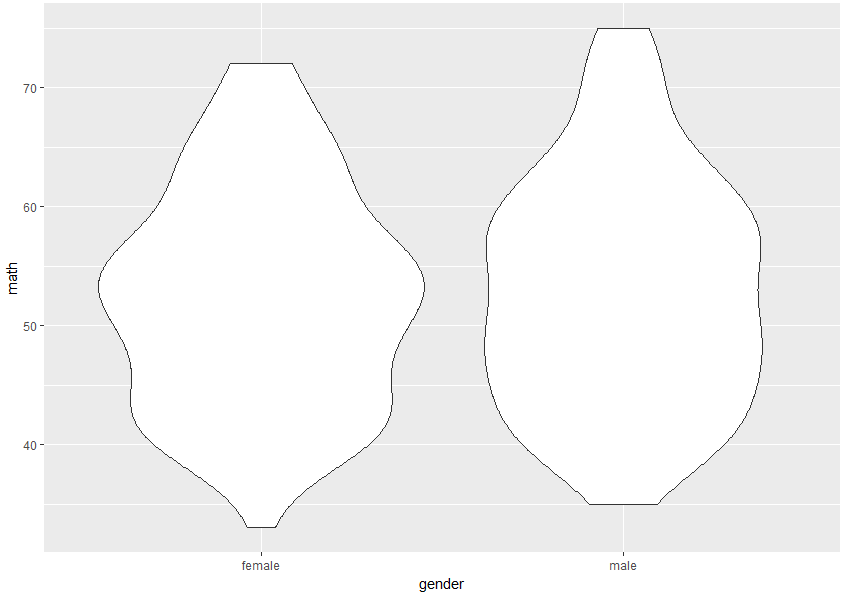
1. Are the survival rate of children signiﬁcantly diﬀerent from that of adult women?
   1. H0 : children does not have significantly different survival rate than adult women
   2. HA : Children have significantly different survival rate than adult women(Two-sided)
   3. What test you use in R:

Since the two data we obtained are the survival rate of children and the survival rate of adult women in question one. Thus, using z-test for different proportion would help us determine if the children does not have significantly different survival rate than adult women.

* 1. Significant level alpha = 0.01
  2. P-value: 0
  3. Conclusion: If H0 (children does not have significantly different survival rate than adult women), then we will observe that children have survived more than adult women with a probability extremely close to zero. This is smaller than the significant level, so we conclude that most likely, Ha(children have significantly different survival rate than adult women) is better than the null hypothesis at explaining the data.

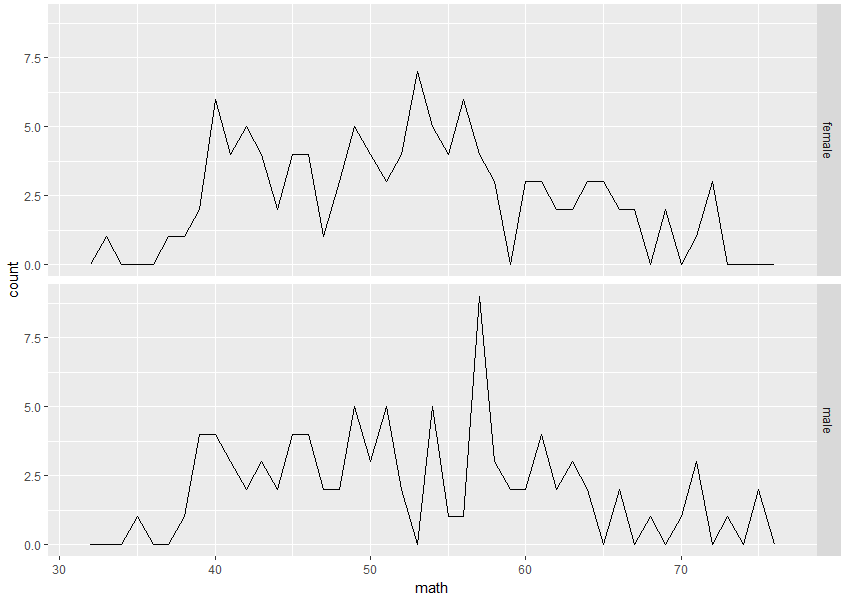
**hsb2 dataset**

1. Describe in graphs and numbers the distribution of math scores between male and female



In this violin graph we can see that comparing two gender math grades, the male maximum score is higher than the female has, and the lowest score of male is also higher than that of female. Nonetheless, two shapes are similar in the way that there are less number of people having higher scores and much more people in either gender in the middle to bottom part of the scores.

The other plot would be more clear in the population distribution.



From this graph, we can see there are more students in the median scores, however, we can realize that there are not that much difference between two genders by this graph.

Let us see the number by table which relate gender to math score which divides into five ranks by 20 points.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Math | | | | |
|  |  | [0,20) | [20,40) | [40,60) | [60,80) | [80,100) |
| Gender | Female | 0 | 5 | 78 | 26 | 0 |
| Male | 0 | 6 | 62 | 23 | 0 |

It could be easily observed that for both genders, the majority are in score 40 to 60, with some scoring 60 to 80. However, opposite to what shown in the first graph that male student have a generally higher score in math. By this table we could even notice that more female students are in 40 to 60 range and 60 to 80 range but one student less in 20 to 40 range. Hence, more detailed analysis needed for the math score among male and female students.

1. Is there a signiﬁcant diﬀerence in the median score between these two groups? Use a permutation test to ﬁnd out. (Math median)
   1. H0 : There does not exist a significant difference in the median score between these two groups.
   2. HA : The median score between these two groups are significantly different. (Two-sided)
   3. What test you use in R:

As the question stated, the permutation test is used to find out if the difference is significant.

* 1. Significant level alpha = 0.01
  2. P-value: 0.894
  3. Conclusion: If H0 (the difference in median score between these two group by gender is not significant), then we will observe that the difference between two medians are close to each other with a probability 0.894. This is larger than the significant level, so we conclude that we cannot reject the null hypothesis, and Ha(the median score between these two groups are significantly different) is worse than the null hypothesis at explaining the data.

1. Is there a signiﬁcant diﬀerence between male and female in the proportion of those who math score is 65 or more? Use a test of your choice
   1. Thus we let students have higher or equal to 65 math score in one catalog, and the rest in the other catalog.
   2. H0 : The difference between male and female in the proportion of those who math score is 65 or more is not significant.
   3. HA : The difference between male and female in the proportion of those who math score is 65 or more is significant. (Two-sided)
   4. What test you use in R:

Since this become a categorical versus categorical variables, and formed a 2 x 2 matrix with small numbers – less than a hundred population in side. Therefore, use fisher’s exact test to find out if the null hypothesis could be proved by data or not.

[Table: Math Score vs. Gender]

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Gender | |
|  |  | female | male |
| Math Score | [0,65) | 96 | 81 |
| [65,100) | 13 | 10 |

* 1. Significant level alpha = 0.01
  2. P-value: 1
  3. Conclusion: If H0 ( The difference between male and gemale in the proportion of those who math score is 65 or more is not significant), then we will observe that difference between math score female student have and the math score male students have are close to zero with a probability 1. This is larger than the significant level, so we conclude that we cannot reject the null hypothesis, and Ha(the difference between male and female students in proportion of those who math score is 65 or more is significant) is worse than the null hypothesis at explaining the data.

1. Does your analysis disprove or support the claim that ”top math students tend to be male”?

My analysis disprove that “top math students tend to be male.” Indeed, the math score is pretty independent from gender if we divide the math score into two group by a line of 65. From the earlier analysis about median, we also concluded that math scores for male and female students does not have a significant difference. Even from the graph, we could see there is only little difference between the distribution of female students’ math score and male students’. Overall, none of the analysis support the claim that math scores could be different depends on the gender of student.

**murders dataset**

The dataset murders contains the victim name, age, and location of every murder recorded in the Greater London area by the Metropolitan Police from January 1, 2006 to September 7, 2011.

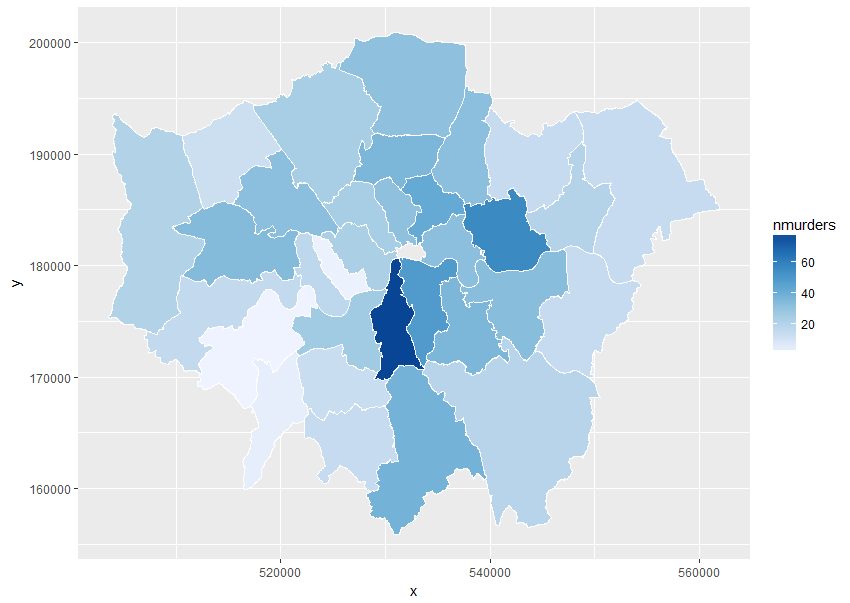
This dataset is included in the OIdata package. You can load this dataset in R with the following commands: install.packages("OIdata") library(OIdata) data(murders)

Now there will be a dataframe named murders in your R workspace. You can do usual commands, such as head(murders)

For information on the variables, type in R ?murders This documentation has helpful examples, including a code on how you can visualize the murder on the London map.

Our goal is to answer the question: do all boroughs have the same murder rate, or are there some ”bad neighborhoods”?

1. Produce a map to visualize the murders by borough.



1. Produce a table that counts the number of murders by borough.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Borough | Barking & Dagenham | Barnet | Bexley | Brent |
| No Murders | 20 | 24 | 14 | 32 |
| Borough | Bromley | Camden | Croydon | Ealing |
| No Murders | 19 | 24 | 38 | 34 |
| Borough | Enfield | Greenwich | Hackney | Hammersmith & Fulham |
| No Murders | 31 | 33 | 42 | 17 |
| Borough | Haringey | Harrow | Havering | Hillingdon |
| No Murders | 36 | 12 | 14 | 21 |
| Borough | Hounslow | Islington | Kensington & Chelsea | Kingston |
| No Murders | 15 | 31 | 2 | 4 |
| Borough | Lambeth | Lewisham | Merton | Newham |
| No Murders | 79 | 36 | 13 | 56 |
| Borough | Redbridge | Richmond | Southwark | Sutton |
| No Murders | 14 | 1 | 49 | 14 |
| Borough | Tower Hamlets | Waltham Forest | Wandsworth | Westminster |
| No Murders | 32 | 32 | 26 | 23 |

1. Is the count itself meaningful? What other statistics are we missing to compute the murder rate? Go online to ﬁnd them and compute the murder rate by borough.
   1. We need to know the population in each borough to know that the murder rate. Otherwise, even some of the borough seems have a relatively high number or murders, but if the population is also very large, hence the murder rate makes more sense.

|  |  |
| --- | --- |
| 1 BarkingDagenham  2 Barnet  3 Bexley  4 Brent  5 Bromley  6 Camden  7 Croydon  8 Ealing  9 Enfield  10 Greenwich  11 Hackney  12 HammersmithFulham  13 Haringey  14 Harrow  15 Havering  16 Hillingdon  17 Hounslow  18 Islington  19 KensingtonChelsea  20 Kingston  21 Lambeth  22 Lewisham  23 Merton  24 Newham  25 Redbridge  26 Richmond  27 Southwark  28 Sutton  29 TowerHamlet  30 Waltham Forest  31 Wandsworth  32 Westminster | 1.07E-04  6.71E-05  6.01E-05  1.02E-04  6.12E-05  1.09E-04  1.04E-04  1.00E-04  9.87E-05  1.29E-04  1.70E-04  9.32E-05  1.41E-04  4.99E-05  5.88E-05  7.62E-05  5.88E-05  1.50E-04  1.26E-05  2.49E-05  2.59E-04  1.30E-04  6.48E-05  1.80E-04  4.98E-05  5.33E-06  1.70E-04  7.33E-05  1.25E-04  1.23E-04  8.45E-05  1.05E-04 |

1. Is there a signiﬁcant diﬀerence in the rate of murders between boroughs? (be clear on what test you are using).
   1. H0 : There is no significant difference in the rate of murders between boroughs.
   2. HA : There is significant difference in the rate of murders between boroughs.
   3. What test you use in R:

Since this situation is comparison between all rates. Set an average to be 9.83E-5. Use t-test to compare each rate in borough to the average.

* 1. Significant level alpha = 0.01
  2. P-value: 0.3478
  3. Conclusion: If H0 ( There is no significant difference in the rate of murders between boroughs), then we will observe that difference between all boroughs murders rate could be random with a probability 0.3478. This is larger than the significant level, so we conclude that we cannot reject the null hypothesis, and Ha(the difference rate of murders between boroughs are significant) is worse than the null hypothesis at explaining the data.

Bonus question Answer the previous question by performing own permutation test. You will need to write a piece of R code to randomly reassign the location of the murders in the dataset. Be VERY clear on what test statistic you are using.

Bonus question Obtain a similar dataset for Austin, and perform the same analysis.