DA ASSIGNMENT PART(II)

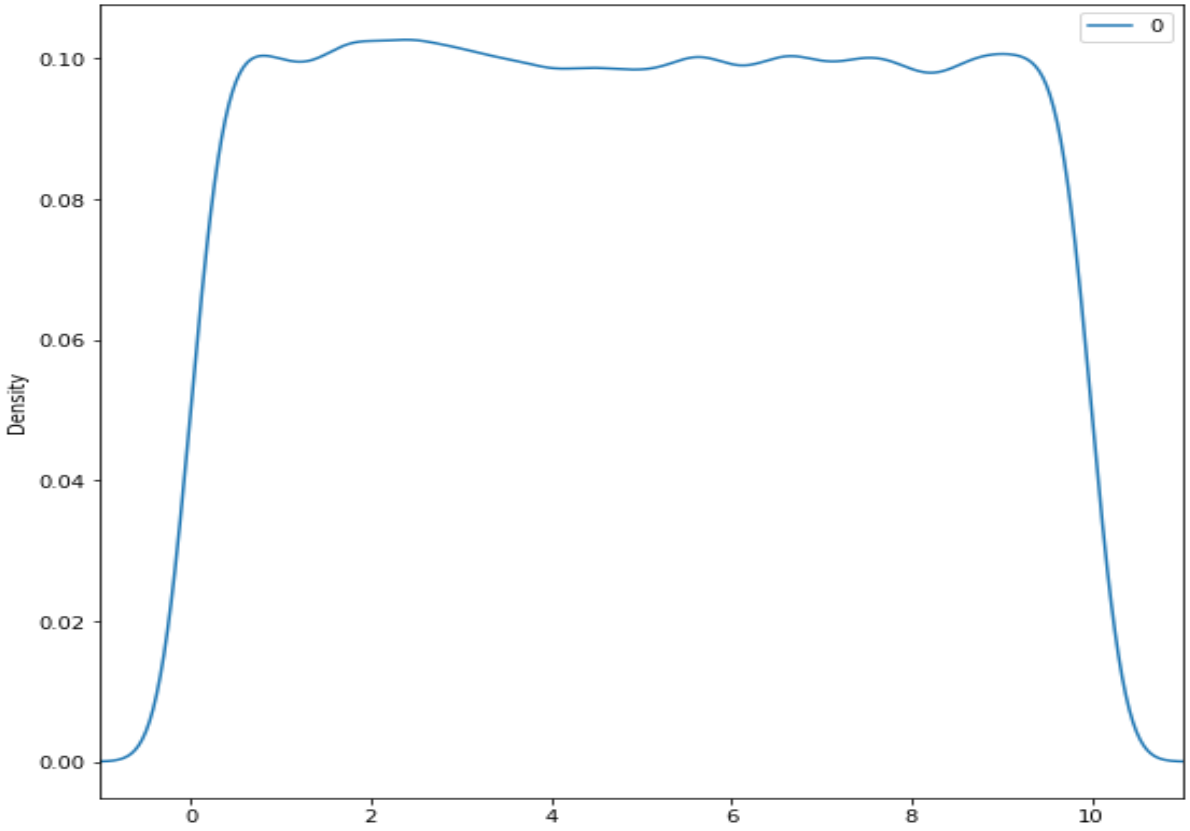
NAME: SNEHA SAVARKAR Subject code: OEIT6 UID: 2019120055

Probability Disribution

Many statistical tools and techniques used in data analysis are based on probability. Probability measures how likely it is for an event to occur on a scale from 0 (the event never occurs) to 1 (the event always occurs). When working with data, variables in the columns of the data set can be thought of as random variables: variables that vary due to chance. A probability distribution describes how a random variable is distributed; it tells us which values a random variable is most likely to take on and which values are less likely. In statistics, there are a range of precisely defined probability distributions that have different shapes and can be used to model different types of random events.

Uniform Distribution

The uniform distribution is a probability distribution where each value within a certain range is equally likely to occur and values outside of the range never occur. If we make a density plot of a uniform distribution, it appears flat because no value is any more likely (and hence has any more density) than another.



we generated 100,000 data points from a uniform distribution spanning the range 0 to 10. In the density plot, we see that the density of our uniform data is essentially level meaning any given value has the same probability of occurring. The area under a probability density curve is always equal to 1.

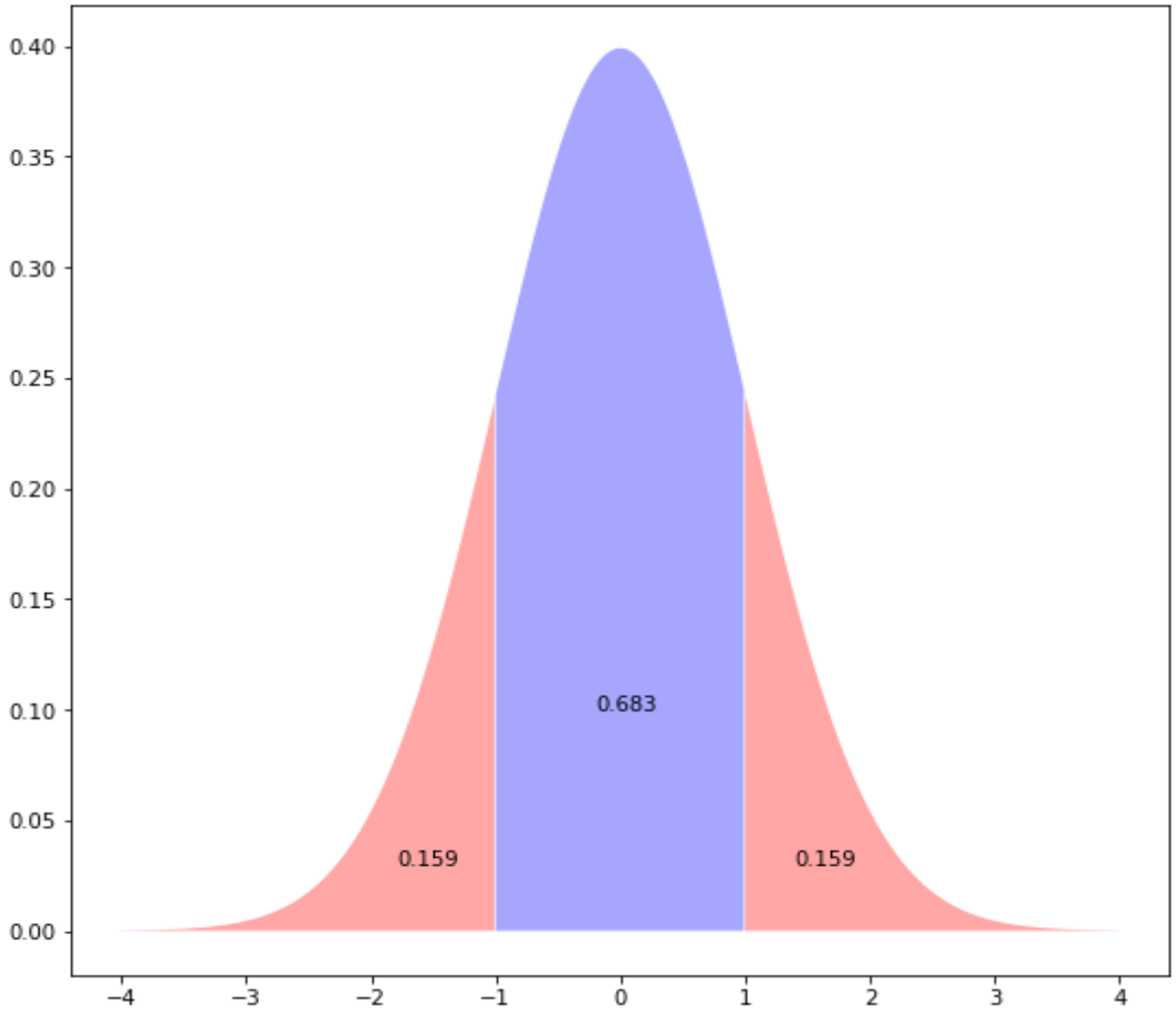
The Normal Distribution

The normal or Gaussian distribution is a continuous probability distribution characterized by a symmetric bell-shaped curve. A normal distribution is defined by its center (mean) and spread (standard deviation.). The bulk of the observations generated from a normal distribution lie near the mean, which lies at the exact center of the distribution: as a rule of thumb, about 68% of the data lies within 1 standard deviation of the mean, 95% lies within 2 standard deviations and 99.7% lies within 3 standard deviations.

Output: 0.15865525393145707 0.15865525393145707 0.6826894921370859

The output shows that roughly 16% of the data generated by a normal distribution with mean 0 and standard deviation 1 is below -1, 16% is above 1 and 68% lies between -1 and 1, which agrees with the 68, 95, 99.7 rule.

Plot :Normal distribution



The plot above shows the bell shape of the normal distribution, the area below and above one standard deviation and the area within 1 standard deviation of the mean.

Finding quantiles of the normal distribution :

-1.9599639845400545
1.959963984540054

The quantile output above confirms that roughly 5% of the data lies more than 2 standard deviations from the mean.

chi-squared goodness-of-fit test

The chi-squared goodness-of-fit test is an analog of the one-way t-test for categorical variables: it tests whether the distribution of sample categorical data matches an expected distribution.

calculate the chi-squared statistic with the following formula:

sum*((observed−expected)^2/expected)

In the formula, observed is the actual observed count for each categoryand expected is the ex pected count based on the distribution of the population for the corresponding category.

Let’s generate some imaginary data of automobile/ cars for India and Mumbai and perform a chi-square goodness of fit test to check whether they are different
In the formula, observed is the actual observed count for each category and expected is the expected count based on the distribution of the type for the corresponding category.

count 1982.942857
dtype: float64

The chi-squared test assumes none of the expected counts are less than 5.

Similar to the t-test where we compared the t-test statistic to a critical value based on the t-distribution to determine whether the result is significant, in the chi-square test we compare the chi-square test statistic to a critical value based on the chi-square distribution.

The scipy library shorthand for the chi-square distribution is chi2. Let's use this knowledge to find the critical value for 95% confidence level and check the p-value of our result:

Critical value
9.487729036781154
P value
[0.]

Since our chi-squared statistic exceeds the critical value, we'd reject the null hypothesis that the two distributions are the same.

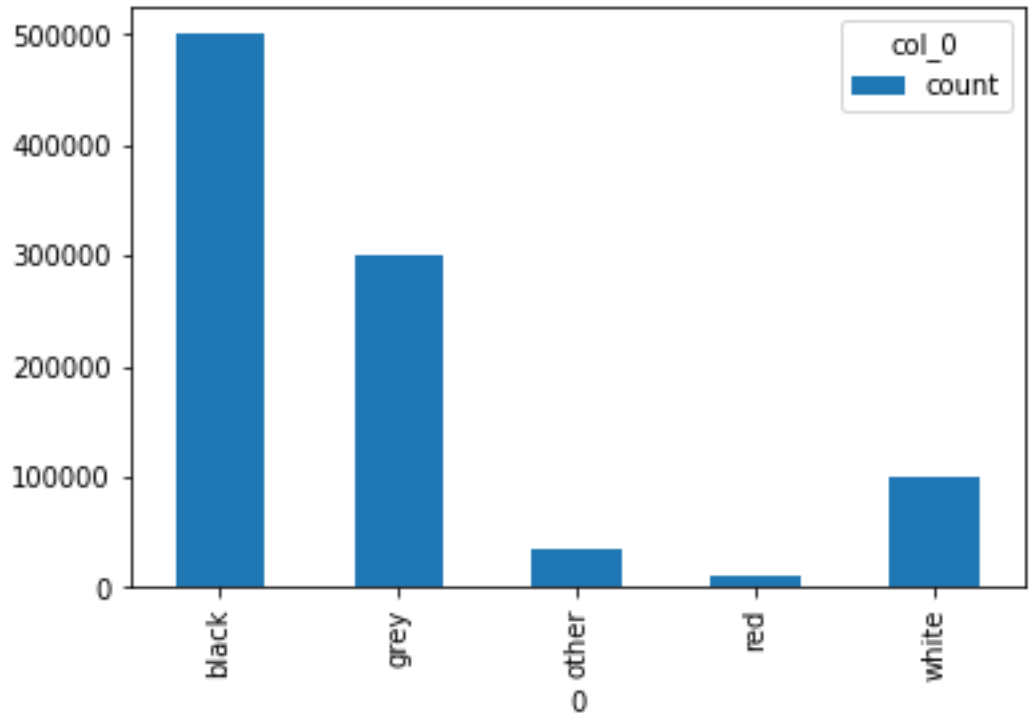
We can perform chi-squared goodness of fit test directly automatically using scipy function scipy.stats.chisquared():

Power_divergenceResult(statistic=array([1982.94285714]), pvalue=array([0.]))

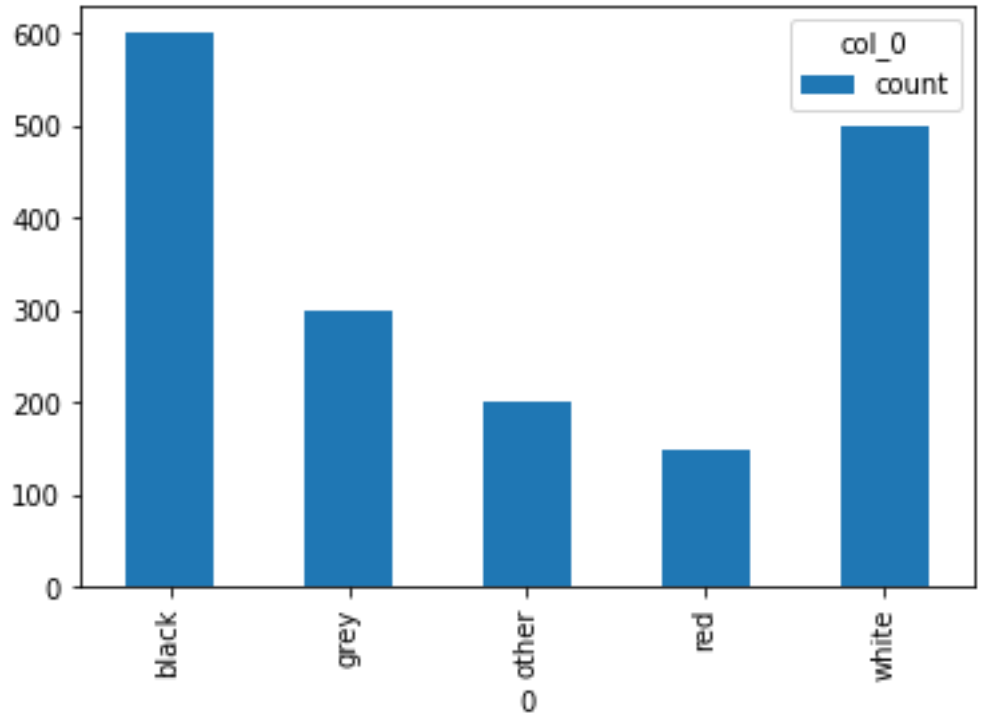
The test results agree with the values we calculated above as we get same value.

Graphs:

1) National Data



2)Sample(Mumbai) Data



The data which I have generated is categorical so the normal distribution doesn't effect the clarity of the data. Therefor the above graphs doesn't define anything but we can model it to be normal distribution by changing the columns of data.

Chi-Squared Test of Independence

Independence is a key concept in probability that describes a situation where knowing the value of one variable tells you nothing about the value of another. The chi-squared test of independence tests whether two categorical variables are independent. To make the data more elaborate lets add some more catogories to it as fuel type.

We did not use the race data to inform our generation of the category data so the variables are independent.

	petrol	diesel	ev	row_totals
white	50	94	107	251
red	8	15	15	38
black	96	212	189	497
grey	25	64	65	154
other	7	32	21	60
col_totals	186	417	397	1000

The above figure signifies the Data generated

	petrol	diesel	ev
white	46.686	104.667	99.647
red	7.068	15.846	15.086
black	92.442	207.249	197.309
grey	28.644	64.218	61.138
other	11.160	25.020	23.820

The critical value and the p-value for the above data:
7.169321280162059

This is to calculate critical value assuming 95% confidense or 5%significance and (5+3)-1 variable categories

Critical value
15.50731305586545

P value
0.518479392948842

As with the goodness-of-fit test, we can use scipy to conduct a test of independence quickly. Using stats.chi2_contingency() function to conduct a test of independence automatically given a fre uency table of observed counts:

(7.16932128016206, 0.518479392948842, 8, array([[46.686, 104.667, 99.647],
[7.068, 15.846, 15.086],
[92.442, 207.249, 197.309],
[28.644, 64.218, 61.138],
[11.16 , 25.02 , 23.82]]))

The output shows the chi-square statistic, the p-value and the degrees of freedom followed by the expected counts.

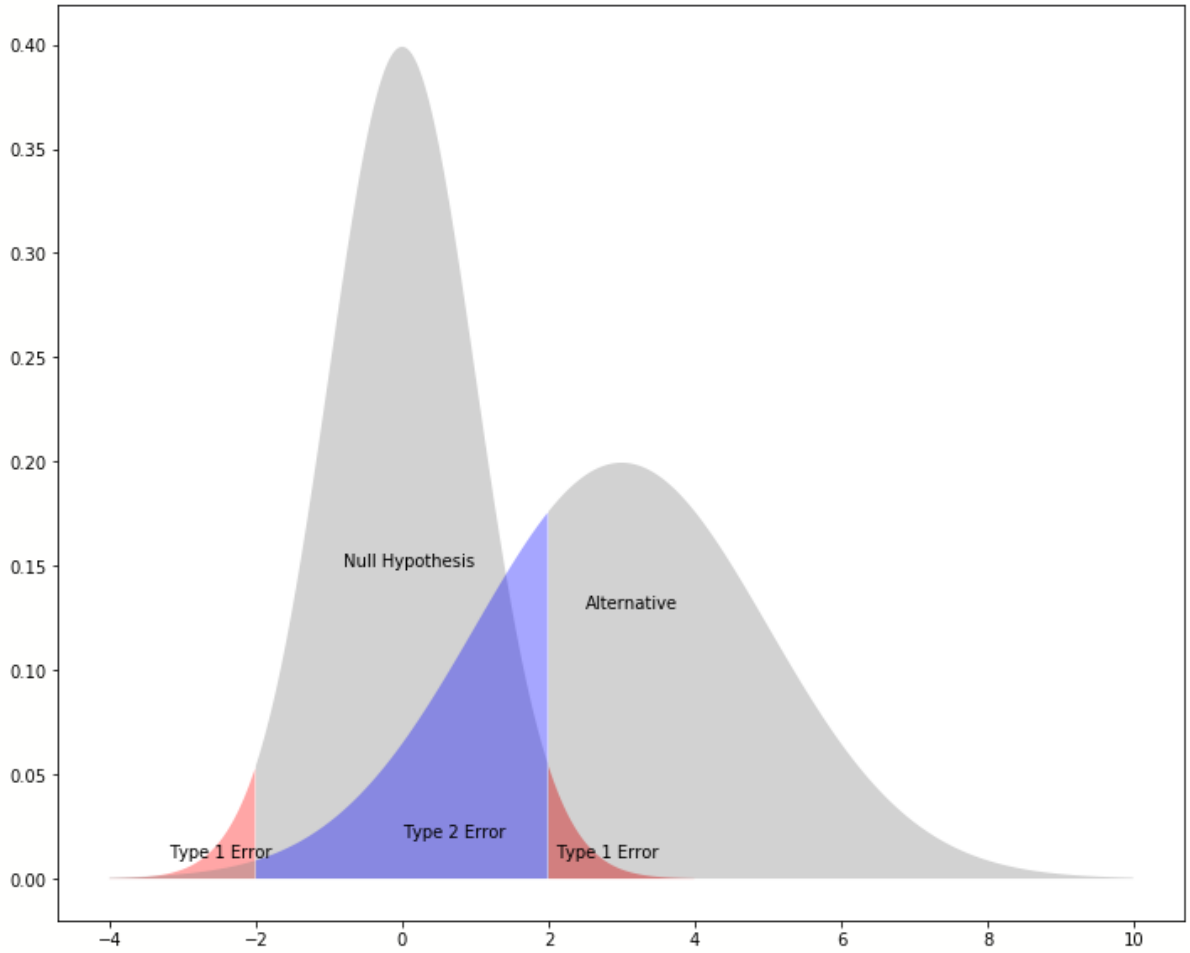
As expected, given the high p-value, the test result does not detect a significant relationship between the variables.As our hypothesis that the Race/color of the car is independent of its fuel type.

Chi-squared tests provide a way to investigate differences in the distributions of categorical variables with the same categories and the dependence between categorical variables.

HYPOTHESIS TESTING

Statistical hypothesis tests are based a statement called the null hypothesis that assumes nothing interesting is going on between whatever variables you are testing. The exact form of the null hypothesis varies from one type test to another: if you are testing whether groups differ, the null hypothesis states that the groups are the same.

The purpose of a hypothesis test is to determine whether the null hypothesis is likely to be true given sample data. If there is little evidence against the null hypothesis given the data, you accept the null hypothesis. If the null hypothesis is unlikely given the data, you might reject the null in favor of the alternative hypothesis: that something interesting is going on. The exact form of the alternative hypothesis will depend on the specific test you are carrying out.



In the plot above, the red areas indicate type I errors assuming the alternative hypothesis is not different from the null for a two-sided test with a 95% confidence level.

The blue area represents type II errors that occur when the alternative hypothesis is different from the null, as shown by the distribution on the right. Note that the Type II error rate is the area under the alternative distribution within the quantiles determined by the null distribution and the confidence level.