Data Analytics
Full-Time Bootcamp

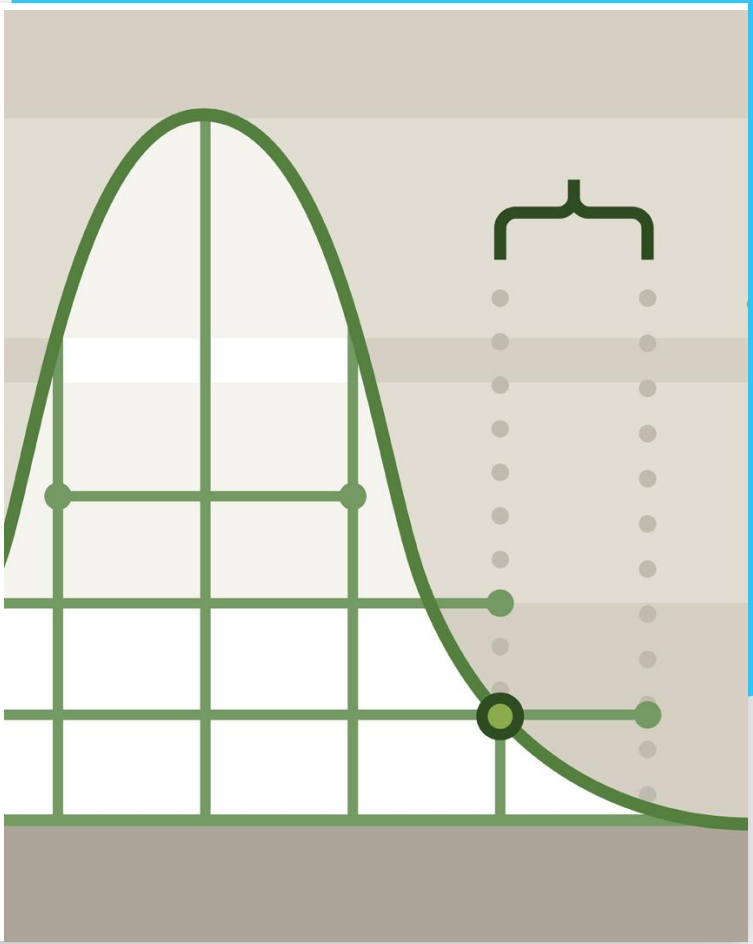
GOODNESS OF FIT

PREVIOUSLY, IN DATA ANALYTICS FULL TIME...

HYPOTHESIS TESTING

For a given observed sample X, we check how unlikely it is to produce X in a world where our null hypothesis is true (p-value). If our observation is overwhelmingly unlikely ($p<\alpha$), we prefer to reject the notion that we are living in a world where the null is true.

On the other hand, if our observation X is not that unlikely in a world where the null holds, we do not reject the null.



AND NOW, RESUMING OUR PREVIOUSLY SCHEDULED PROGRAM

GOODNESS OF FIT AND ASSOCIATION TESTING

Suppose that a sample is taken from a population and the members can be uniquely classified according to a pair of discrete characteristics A and B. The hypothesis to be tested is of no association in the population between possession of characteristic A and possession of characteristic B. For example, a travel agency may want to know if there is any relationship between a client's gender and the method used to make an airline reservation.

Essentially we will posit as null hypothesis that the presence of an aspect of characteristic A (say gender being male) does not influence the distribution of characteristic B (method of reservation), or vice versa.

GOODNESS OF FIT AND ASSOCIATION TESTING

Let’s take an example of market differentiation. Makers of products want their products to be distinctly perceived from the competition so let’s say 3 car makers want to understand how their brand is perceived and if they are sufficiently differentiated. A survey of 513 car owners where they are asked to identify 3 brands with the notions of “Sportive” or “Safe” returns the following results

Brand	**Sportive**	**Safe**	**Total**
BMW	256	74	***330***
Mercedes	41	42	***83***
Lexus	66	34	***100***
Total	***363***	***150***	***513***

GOODNESS OF FIT AND ASSOCIATION TESTING

We start with the answer

```
import scipy.stats as st
st.chi2_contingency(table)
```

Yay! (?)

To the colab!

Brand	Sportive	Safe	Total
BMW	256	74	***330***
Mercedes	41	42	***83***
Lexus	66	34	***100***
Total	***363***	***150***	***513***

What is this doing?

GOODNESS OF FIT AND ASSOCIATION TESTING

The null hypothesis would be that the brand does not influence perception. To compute the expected observations for each box, we first ignore the actual granular values and find the *marginal proportions*

Brand	Sportive	Safe	Total
BMW			***330***
Mercedes			***83***
Lexus			***100***
Total	***363***	***150***	***513***

330/513 =64.3%

Brand	Sportive	Safe	Total
BMW			***64.3%***
Mercedes			***16.2%***
Lexus			***19.4%***
Total	***70.7%***	***29.2%***	***100%***

GOODNESS OF FIT AND ASSOCIATION TESTING

Important fact of life:
If A and B are independent, the probability of A and B occurring simultaneously is P(A)*P(B)

Now we find the individual proportions for each pair under the null hypothesis. This assumes each car is rated on each characteristic independently, so the proportions do not change when you compare car brand or characteristic descriptive

Brand	Sportive	Safe	Total
BMW	70.7%*64.3%		***64.3%***
Mercedes			***16.2%***
Lexus			***19.4%***
Total	***70.7%***	***29.2%***	***100%***

Brand	Sportive	Safe	Total
BMW	45.5%	18.8%	***64.3%***
Mercedes	11.5%	4.7%	***16.2%***
Lexus	13.7%	5.7%	***19.4%***
Total	***70.7%***	***29.2%***	***100%***

GOODNESS OF FIT AND ASSOCIATION TESTING

We now get back to our expected values by remembering that we had 513 cars

513*45.5%

Brand	Sportive	Safe	Total
BMW	45.5%	18.8%	***64.3%***
Mercedes	11.5%	4.7%	***16.2%***
Lexus	13.7%	5.7%	***19.4%***
Total	***70.7%***	***29.2%***	***100%***

Brand	Sportive	Safe	Total
BMW	233.4	96.4	***330***
Mercedes	59.0	24.1	***83***
Lexus	70.3	29.2	***100***
Total	***363***	***150***	***513***

GOODNESS OF FIT AND ASSOCIATION TESTING

Now we have our Observed table and our Expected table under H_0

Observed Table

Brand	Sportive	Safe	Total
BMW	256	74	***330***
Mercedes	41	42	***83***
Lexus	66	34	***100***
Total	***363***	***150***	***513***

Expected Table (under H_0)

Brand	Sportive	Safe	Total
BMW	233.4	96.4	***330***
Mercedes	59.0	24.1	***83***
Lexus	70.3	29.2	***100***
Total	***363***	***150***	***513***

GOODNESS OF FIT - FIT TO DISTRIBUTION

This formula measures how much Observed reality differs from Expectation under H_0

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Where O_i is the number of actual observations in category i and E_i is the expected number of observations in that category.

Brand	Sport	Safe
BMW	256	74
Merc	41	42
Lexus	66	34

Brand	Sport	Safe
BMW	233.4	96.4
Merc	59.0	24.1
Lexus	70.3	29.2

$(256\text{-}233.4)^2/233.4$

Brand	Sport	Safe
BMW	2.19	5.20
Merc	0.15	13.30
Lexus	0.26	0.79

GOODNESS OF FIT - FIT TO DISTRIBUTION

Brand	Sport	Safe
BMW	256	74
Merc	41	42
Lexus	66	34

Brand	Sport	Safe
BMW	233.4	96.4
Merc	59.0	24.1
Lexus	70.3	29.2

Brand	Sport	Safe
BMW	2.19	5.20
Merc	0.15	13.30
Lexus	0.26	0.79

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

= 2.19+5.20+0.15+13.30+0.26+0.79 = 21.89

This measures “how much observed reality differs from the H_0”. Is this deviation “enough” for us to discard the hypothesis? That’s where the p-value comes in.