Data Analytics
Full-Time Bootcamp

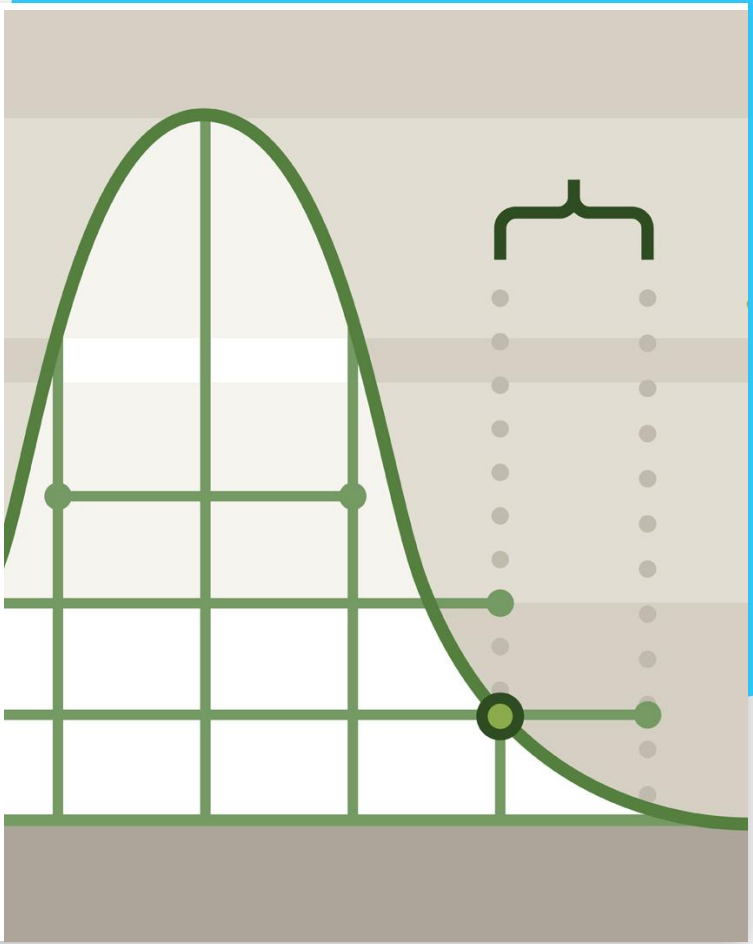
GOODNESS OF FIT

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 - FIT TO DISTRIBUTION

Suppose you have some phenomenon that you repeat and you count the *frequency* of its outcomes, i.e., the number of times each different event happens.

E.g.

- The number of times you get 0,1,2,3,... passengers missing an airflight
- The number of times you open 1,2,3,... cans of tuna until you find one that is rotten
- The number of times 0,1,2,3,... goals are scored in the last 10 minutes of a football match
- The number of times people buy cola, pepsi, water out of 100 sales

Obs 1: Some of these, we have seen, follow a particular distribution (binomial, geometric, poisson...)

Obs 2: Notice you can only do this for discrete (or if you prefer categorical) observations

GOODNESS OF FIT - FIT TO DISTRIBUTION

If we don’t have a strong theoretical reason to claim that a certain phenomenon follows a particular distribution and we suspect it does, can we show that using data? After all, if we can assume something follows a particular well-studied distribution, we get a lot of results and stats packages that are directly applicable to our phenomenon.

We will test if a certain set of observations can reasonably be said to come from a specific distribution by using a Goodness-of-Fit test.

The guiding principle is the following: If our phenomenon does follow a certain distribution, we would expect it’s outcomes to follow certain patterns. If our observed data does not stray too far from these patterns (i.e. if it is a “good fit” to the patterns), we can assume it comes from the posited distribution.

GOODNESS OF FIT - FIT TO DISTRIBUTION

Based on the data from all vending machines, Unilever believes that 40% of sales from their machines are Coca-cola, 30% are Pepsi and the remaining 30% are water. They have a stocking system quite optimized for these ratios, but we know not all vending locations necessarily follow this distribution.

We have one machine and we noticed that form the last 300 sold bottles we sold the following:

	Sold
Coca-cola	120
Pepsi	80
Water	100

Do we have evidence here to disprove that our machine follows Unilever's posited distribution for vending machines? Should we develop our own stocking system or can we import the one created by Unilever?

We will use the hypothesis testing logic: we assume that the machine follows Unilever's distribution and see if observed reality strains the credibility of that assumption at a significance level of, say 5%

GOODNESS OF FIT - FIT TO DISTRIBUTION

$H_0 : Dist_{machine} \sim (40\%,30\%,30\%)$ vs $H_1 : Dist_{machine} \neq (40\%,30\%,30\%)$

We first build a table of expected observations under H_0

Expected sales under H_0

	Dist	Expected
Coca-cola	40%	40%*300 = 120
Pepsi	30%	30%*300 = 90
Water	30%	30%*300 = 90

Observed sales from sample

	Sold
Coca-cola	120
Pepsi	80
Water	100

GOODNESS OF FIT - FIT TO DISTRIBUTION

If the null hypothesis was true, we would think that the observed data in each category would be close in value to the expected numbers in each category. In such circumstances the data provides a close fit to the assumed population distribution of probabilities.

A test of the null hypothesis is based on an assessment of the closeness of this fit and is generally referred to as a goodness-of-fit test. Now, in order to test the null hypothesis, it is natural to look at the magnitudes of the discrepancies between what is observed and what is expected. The larger these discrepancies are in absolute value, the more suspicious we are of the null hypothesis.

GOODNESS OF FIT - FIT TO DISTRIBUTION

The test statistic for this problem is

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

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

i	E_i	O_i
Coca-cola	120	120
Pepsi	90	80
Water	90	100

$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
0	0
100	1.11
100	1.11

And our statistic is

0 + 1.11 + 1.11 = 2.22

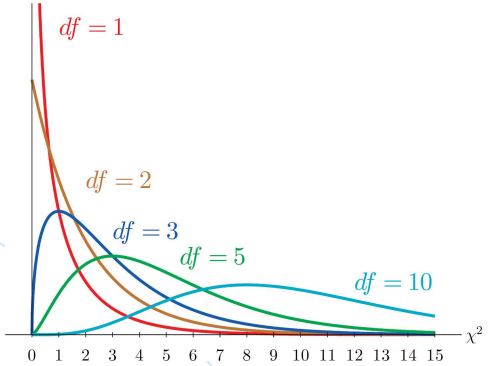
Is this number big or small?

This is not a t-stat nor a z-stat

GOODNESS OF FIT - FIT TO DISTRIBUTION

This statistic follows a Chi-squared distribution. Just like T-distributions, there is no single Chi-squared. Instead there is one for each $\mathbb{n}$ (called degrees of freedom), but now $\mathbb{n}$ does not stand for the number of observations -1. Instead it stands for the number of categories -1.

In this case, our $\mathbb{n}$=3-1=2



We can of course go to a table and check the value for a Chi-squared with 2 degrees of freedom and a α=5% significance level.

But we can also run the function

```
import scipy.stats as st
st.chi2.sf(abs(stat),n)
```

To automatically get the p-value

In this case there are no considerations on one-tailed versus 2-tailed 😁

TO THE COLAB!

HOW TO COMPUTE THE EXPECTED OBSERVATIONS

As discussed on the Colab, we could have done the previous test just by running

```
import scipy.stats as st
st.chisquare([120,80,100], f_exp=[120,90,90])
```

The difficulty in the Chi2 test is often going to be how to figure out the expected outcomes under the null

Say that we observed the following count, out of 200 observations

Value	1	2	3	4	5	6	7	8	9	10	>10
Count	55	40	38	28	10	6	9	1	2	4	7

And we posit they come from a geometric distribution with parameter p=0.3. We have our observed values, but how can we get the expected ones?

HOW TO COMPUTE THE EXPECTED OBSERVATIONS

Well, under the Hypothesis that this distribution is a Geometric with parameter 0.3, we actually know the probability of the distribution yielding 1, 2, 3,...

```
from scipy.stats import geom
g = geom(0.3)
```

Value	1	2	3	4	5	6	7	8	9	10	>10
P(X=Value)	`g.pmf(1)`	`g.pmf(2)`	`g.pmf(3)`	`g.pmf(4)`	`g.pmf(5)`	`g.pmf(6)`	`g.pmf(7)`	`g.pmf(8)`	`g.pmf(9)`	`g.pmf(10)`	`g.sf(10)`
P(X=Value)	0.3	0.21	0.15	0.1	0.07	0.05	0.04	0.02	0.02	0.01	0.03

And so the expected value of 200 observations

Value	1	2	3	4	5	6	7	8	9	10	>10
E_{value}	0.3*200	0.21*200	0.15*200	0.1*200	0.07*200	0.05*200	0.04*200	0.02*200	0.02*200	0.01*200	0.03*200
E_{value}	60	42	29	21	14	10	7	5	3	2	6

TO THE COLAB!

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

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

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

The test statistic for this problem is still

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

Where O_i is the number of actual observations in category i and E_i is the expected number of observations in that category. Notice that we now have 6 categories (3 car brands * 2 characteristics)

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

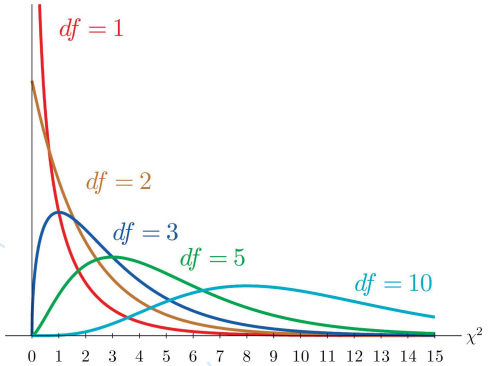
$(256-233.4)^2/233.4$

Brand	Sport	Safe
BMW		
Merc		
Lexus		

GOODNESS OF FIT - FIT TO DISTRIBUTION

This statistic still follows a Chi-squared distribution. But in these tests, m (called degrees of freedom) is the number of rows -1, multiplied by the number of columns-1

In this case, our m=(3-1)*(2-1)=2*1=2



We can of course go to a table and check the value for a Chi-squared with 2 degrees of freedom and a α=5% significance level.

But we can also run the function

```
import scipy.stats as st
st.chi2.sf(abs(stat),n)
```

To automatically get the p-value

In this case there are no considerations on one-tailed versus 2-tailed 😁

TO THE COL...

You may notice that I very conspicuously am not going to the colab yet.

This is because, as usual, there is a function that does all of this for us

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

Yay! (?)

To the colab!