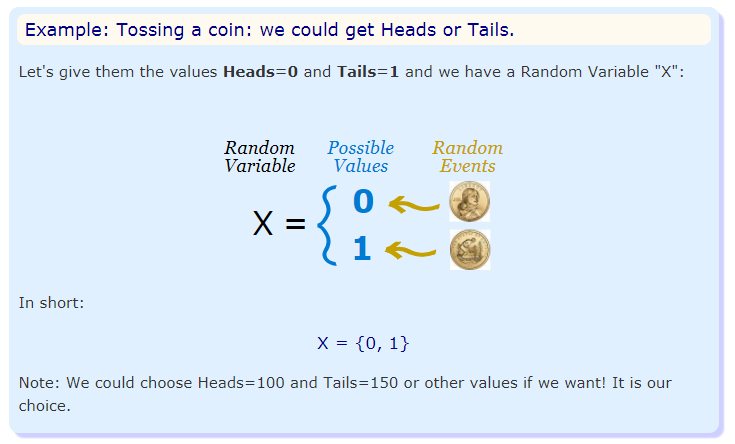
**Distributions in Pandas**

**Distribution**: Set of all possible Random Variables

**Random Variable**: A Random Variable is a set of possible values from a random experiment

**Random Variable Example:**



In the above fig:

* We have an experiment (Tossing a coin)
* We give values to each event
* The set of values is a random variable

**Types of Distributions:**

1. Discrete Distributions
2. Continuous Distributions

**Discrete Distributions:**

A **discrete distribution** describes the **probability** of occurrence of each value of a **discrete** random variable.

A **discrete** random variable is a random variable that has countable values, such as a list of non-negative integers.

**Example:**

Suppose we toss a coin three times.

The sample space consists of 8 equally likely possibilities:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| HHH | HHT | HTH | HTT | THH | THT | TTH | TTT |

Suppose the random variable X counts the number of heads seen in three tosses.

We can't see a negative number of heads in three tosses, nor can we see more than 3, so the values that X can assume are: 0, 1, 2, or 3

(By the way, the probability distribution for this random variable is called discrete because there are a countable number of values that our random variable can assume. In this case there are 4 of them.)

We can summarize these results with the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of Heads(X) | 0 | 1 | 2 | 3 |
| Probability p(x) | 1/8 | 3/8 | 3/8 | 1/8 |

It is easy to see that:   
The probability of seeing no heads is 1/8  
The probability of seeing exactly one head is 3/8  
The probability of seeing exactly two heads is 3/8, and  
The probability of seeing exactly three heads is 1/8

**Types of Discrete Distribution**

1. **Binomial Distribution**
2. **Poisson Distribution**

**Binomial Distribution:** The binomial is a type of distribution that has **two possible outcomes** (the prefix “bi” means two, or twice). (or) Describes the distribution of a variable that can only take two values (0/1, False/True, Heads/Tails)

**The binomial is defined by two parameters: the probability of success in any given trial and the number of trials. The binomial distribution tells you how likely it is to achieve a given number of successes in n trials of the experiment.**

**For example:** A coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

**Using IDLE**

n=10 # number of trials

p = 0.5 #probability of each trial

>>> import numpy as np

>>> import pandas as pd

>>> s=np.random.binomial(10,0.5,6)

>>> s

array([3, 3, 4, 6, 5, 6])

6 is the size

# result of flipping a coin 10 times, tested 6 times.

**A real world example.**

**A company drills 9 wild-cat oil exploration wells, each with an estimated probability of success of 0.1. All nine wells fail. What is the probability of that happening?**

Let’s do 20 trials of the model, and count the number that generate zero positive results.

>>> s=sum(np.random.binomial(9,0.1,20)==0)/20

>>> s

0.40000000000000002

**2.Poisson Distribution**

The  Poisson Distribution models the probability of seeing a certain number of successes within a time interval**.**

**Parameters: Two**

**lambda:**Expectation of interval should be >= 0

**Size:(Optional)** a big list of integers.

>>> import numpy as np

>>> s=np.random.poisson(5)

>>> s

10

>>> s=np.random.poisson(5,8)

>>> s

array([11, 7, 5, 6, 6, 5, 4, 5])

We can specify lambda as say 5 and how many numbers are desired (the second argument) and get a big list of integers

>>> s=np.random.poisson(15,8)

>>> s

array([15, 19, 6, 19, 19, 13, 13, 10])

>>> len(s)8

**Continuous Distribution:**

A continuous distribution describes the probabilities of the possible values of a Continuous random variable. A continuous random variable is a random variable with a set of possible values (known as the range) that is infinite and uncountable.

**For Example:**

We might have a random variable that takes on any value between 3 and 4.Our random variable might take a value 3.1(or)3.12(or)3.126789(or)3.92(or)what you have(or)any of the infinite number of values in between 3&4.

**Types of Continuous Distributions:**

1.uniform Distribution

2.normal Distribution

3.Bimodal Distribution

**1.Uniform Distribution:**

A uniform distribution, sometimes also known as a rectangular distribution, is a distribution that has constant probability.

**Example:**    
Suppose in a quiz there are 30 participants. A question is given to all 30 participants and the time allowed to answer it is 25 seconds. Find the probability of participants responds within 6 seconds?

**Solution:**

Given Interval of probability distribution = [0 seconds, 25 seconds]

Density of probability = 1/25−0= 1/25

Interval of probability distribution of successful event = [0 seconds, 6 seconds]

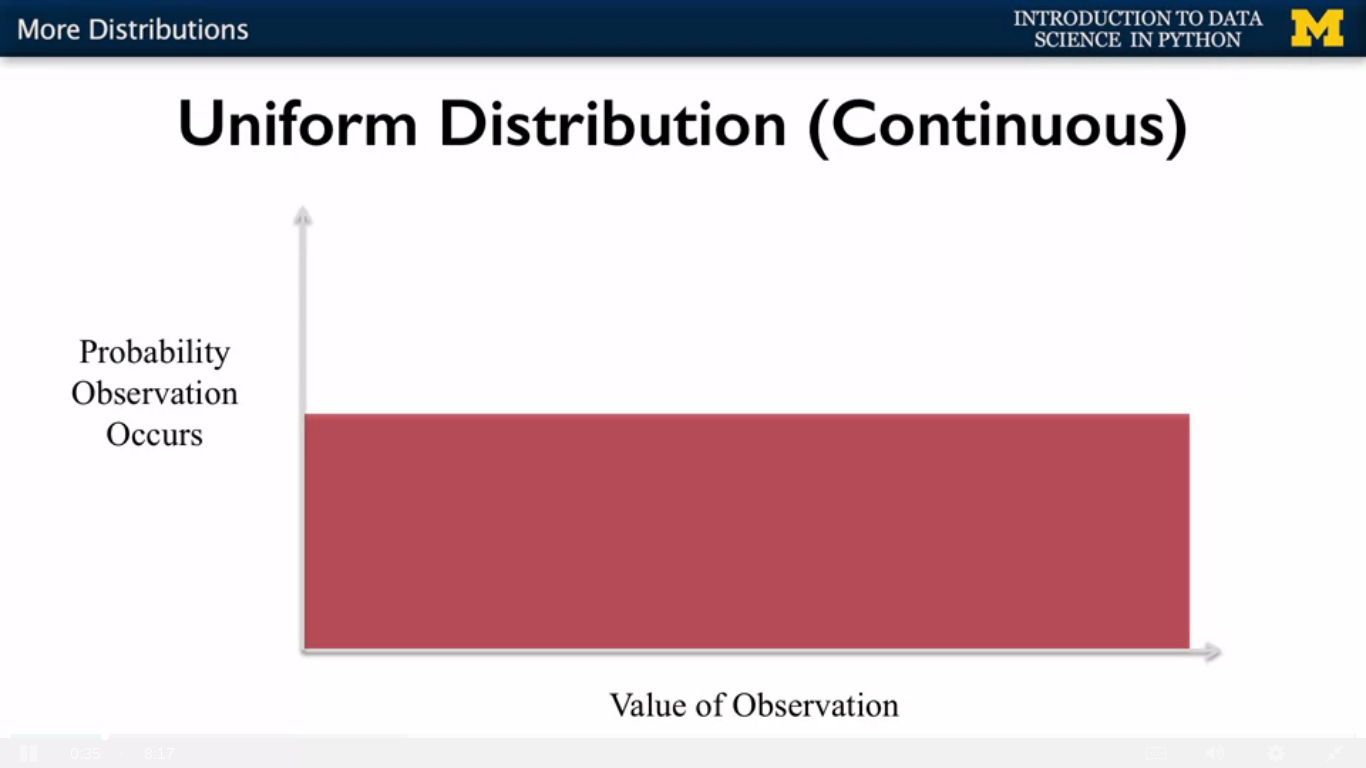
The probability P(x<6)

The probability ratio = 6/25

There are 30 participants in the quiz

Hence the participants likely to answer it in 6

seconds = 6/25 × 30 ≈ 7



where the x axis is the value of the observation and the y axis represents the probability that a given observation will occur. If all numbers are equally likely to be drawn when you sample from it, this should be graphed as a flat horizontal line. And this flat line is actually called the uniform distribution.

**Using IDLE:**

>>> import numpy as np

>>> np.random.uniform(0,1)

0.10732019770244017

>>> np.random.uniform(0,9)

8.997966957365728

>>> np.random.uniform(1,12)

4.734263394746756

>>> np.random.uniform(1,12)

9.706877203677132

**2.Normal Distribution:**

The **normal distribution** is a probability distribution that (roughly) describes many common datasets in the real world.

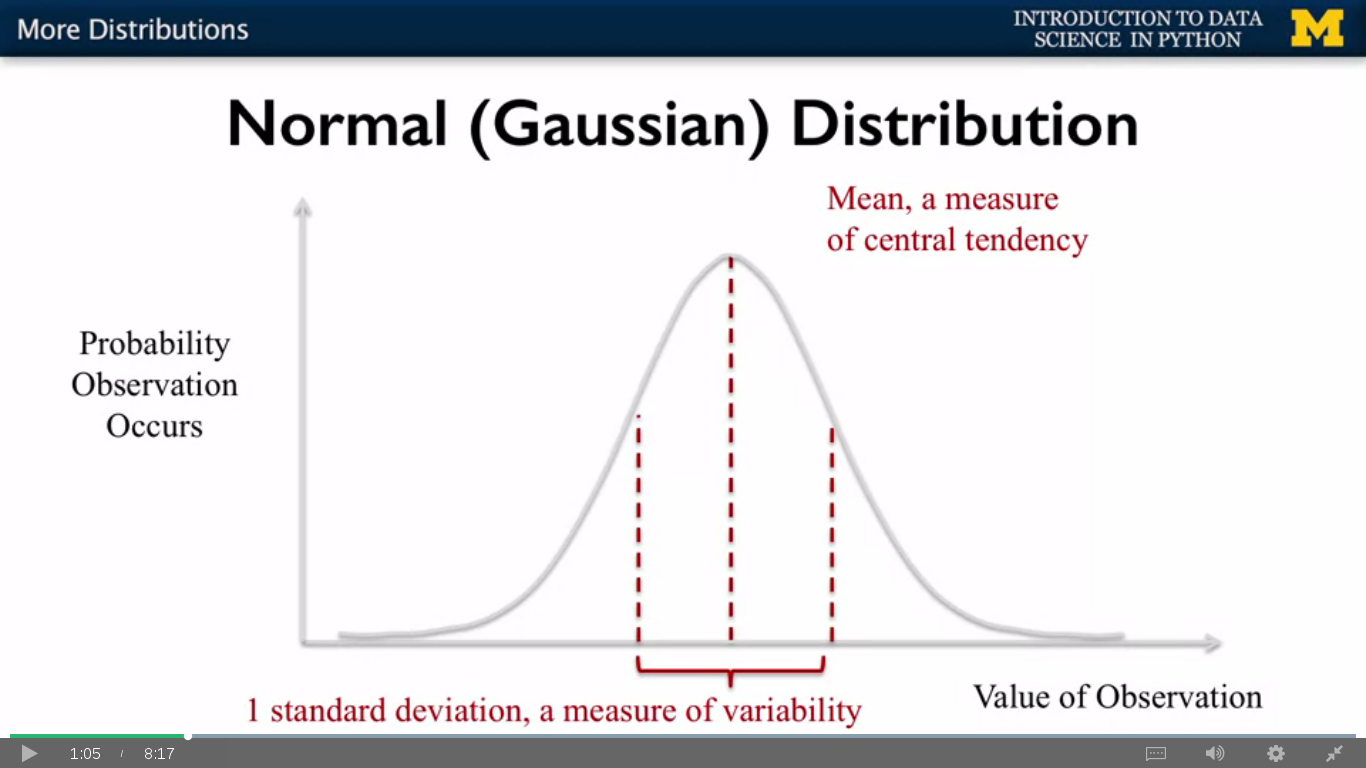
* Heights of people.
* Measurement errors.
* Blood pressure.
* Salaries.

(or)

The normal distribution which is also called Gaussian Distribution or sometimes, a Bell Curve. This distribution looks like a hump where the number which has the highest probability of being drawn is a zero, and there are two decreasing curves on either side of the X axis. One of the properties of this distribution is that the mean is zero, not the two curves on either side are symmetric

In this normal distribution use the term expected value. I think that most of us are familiar with the mean is the sum of all the values divided by the total number of values.

**For Example:** rolling a die three times might give you 1, 2 and 6, the mean value is then 4.5.



**For Example:**

Let's draw 1,000 samples from a normal distribution with an expected value of 0.75 and a standard deviation of 1.

>>> import numpy as np

>>> np.random.normal(0.75)

-0.2428122179855765

**calculating Standard Deviation:**

>>> distribution=np.random.normal(0.75,size=1000)

>>> np.sqrt(np.sum((np.mean(distribution)-distribution)\*\*2)/len(distribution))

1.0532206819487331

(or)

>>> distribution=np.random.normal(0.75,size=1000)

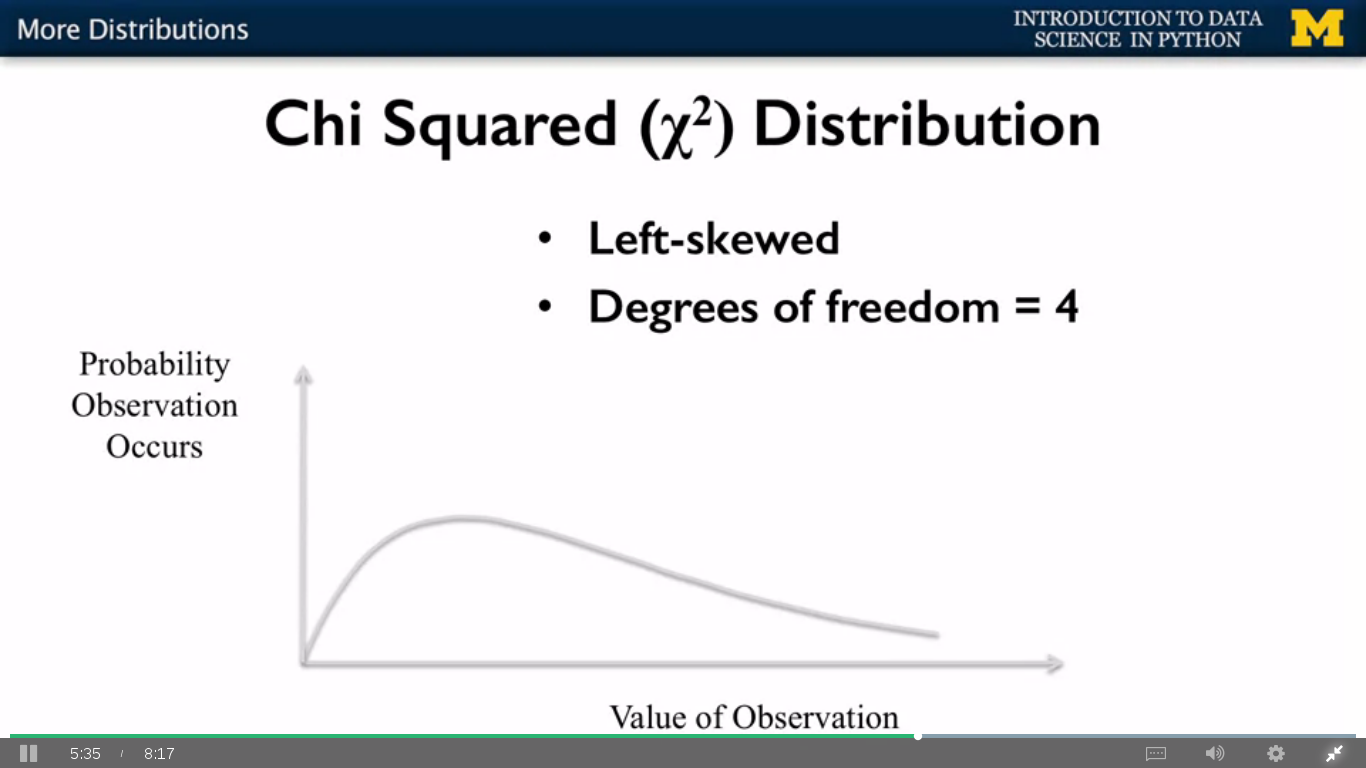
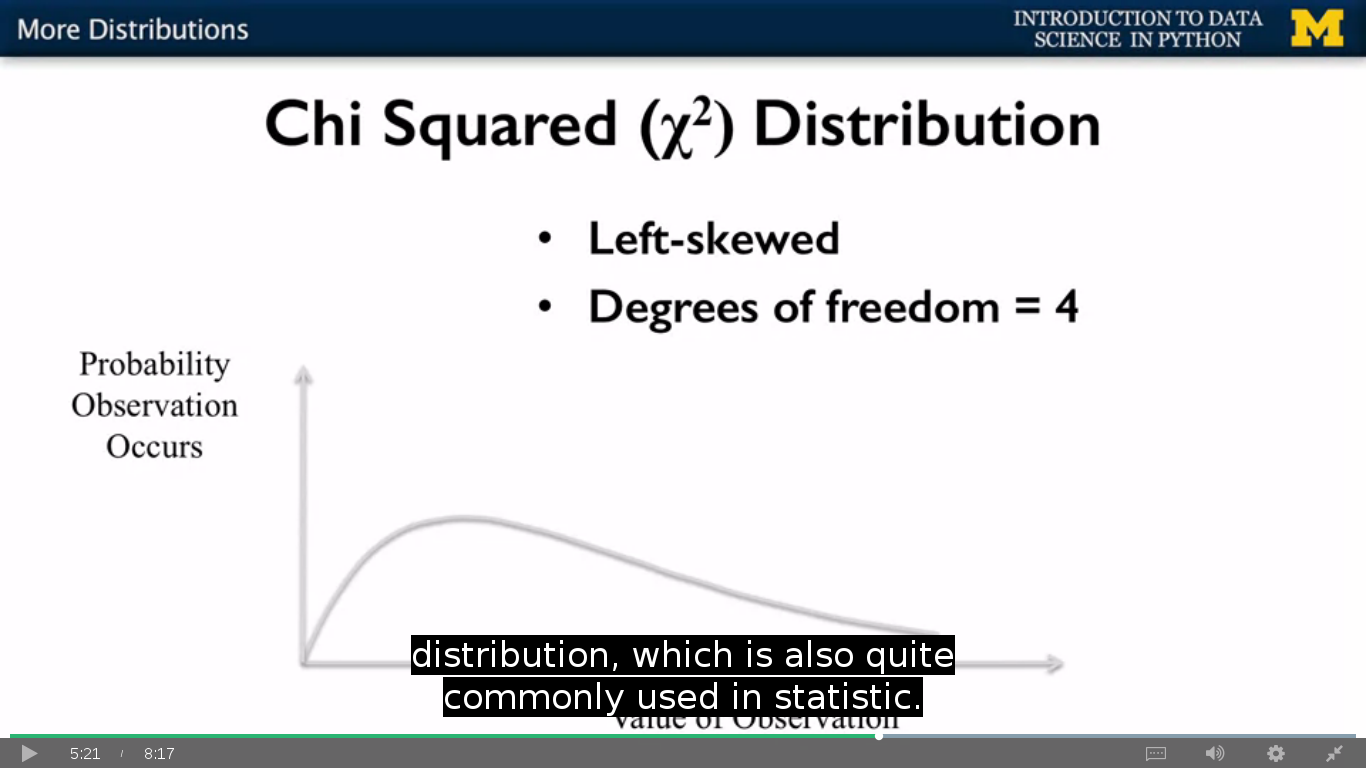
>>> np.std(distribution)

1.0532206819487331

**Chi-Square Distribution**

The Chi Squared Distribution has only one parameter called the degrees of freedom. The degrees of freedom is closely related to the number of samples that you take from a normal population.

Chi square distributions vary depending on the degrees of freedom



We could also move out of the normal distributions and push the peak of the curve one way or the other. And this is called the skew

>>> import scipy.stats as stats

>>> chi=np.random.chisquare(2,size=1000)

>>> stats.skew(chi)

1.799838145713491

First we'll sample 1,000 values from a Chi Squared distribution with degrees of freedom 2. Now we can see that the skew is quite large. Now if we re-sample changing degrees of freedom to 5 ,we see that the skew has decreased significantly

>>> import scipy.stats as stats

>>> chi=np.random.chisquare(5,size=1000)

>>> stats.skew(chi)

1.3226418259225006

**3.Bimodel Distribution**:

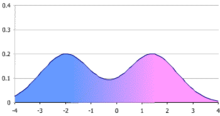
Data distributions in statistics can have **one peak**, or they can have **several peaks**. The type of distribution you might be familiar with seeing is the normal distribution which has one peak. The bimodal distribution has two peaks.

The “bi” in bimodal distribution refers to “two” and modal refers to the peaks. The term “mode” refers to the most common number. However, if you think about it, the peaks in any distribution *are* the most common number(s). The two peaks in a bimodal distribution also represent two local maximums; these are points where the data points stop increasing and start decreasing.

(or)

a **bimodal distribution** is a continuous distribution with two different modes. These appear as distinct peaks in the [probability function](https://en.wikipedia.org/wiki/Probability_density_function)

**Example**



A simple bimodal distribution, in this case a [m](https://en.wikipedia.org/wiki/Mixture_distribution)ixture of two normal distributions with the same variance but different means. The figure shows the probability function, which is an equally-weighted average of the bell-shaped probability functions of the two normal distributions. If the weights were not equal, the resulting distribution could still be bimodal but with peaks of different heights.

**Hypothesis Testing:**

Hypothesis testing in statistics is a way for you to test the results of a survey (or) experiment to see if you have meaningful results.

* A hypothesis is a statement that we can test.

**Example**: Educational technology and learning analytics. Thus, we might expect that those students who sign up quite quickly after the course is launched with higher performance than those students who signed up after the MOCK has been around for a while.

**Types Of Hypothesis:**

1.Alternative Hypothesis

2.Null Hypothesis

In this example, we have samples from two different groups which we want to compare. The early sign ups and the late sign ups. When we do hypothesis testing, we hold out that our hypothesis as the alternative and we create a second hypothesis called the null hypothesis, which in this case would be that there is no difference between groups.

>>> import pandas as pd

>>> df=pd.read\_csv("/home/acer/Desktop/APDS-Videos/csvfiles/grades.csv")

>>> df.head()

student\_id assignment1\_grade \

0 B73F2C11-70F0-E37D-8B10-1D20AFED50B1 92.733946

1 98A0FAE0-A19A-13D2-4BB5-CFBFD94031D1 86.790821

2 D0F62040-CEB0-904C-F563-2F8620916C4E 85.512541

3 FFDF2B2C-F514-EF7F-6538-A6A53518E9DC 86.030665

4 5ECBEEB6-F1CE-80AE-3164-E45E99473FB4 64.813800

>>> len(df)

2315

>>> early=df[df['assignment\_submission']<='2015-12-31']

>>> late=df[df['assignment1\_submission']>'2015-12-31']

>>> early.mean()

assignment1\_grade 74.972741

assignment2\_grade 67.252190

assignment3\_grade 61.129050

assignment4\_grade 54.157620

assignment5\_grade 48.634643

assignment6\_grade 43.838980

dtype: float64

>>> late.mean()

assignment1\_grade 74.017429

assignment2\_grade 66.370822

assignment3\_grade 60.023244

assignment4\_grade 54.058138

assignment5\_grade 48.599402

assignment6\_grade 43.844384

dtype: float64

**P-Hacking (or) Dredging:**

As we run more and more tests, we're more likely to find a positive result just because of the number of T tests we have run. When a data scientist run many tests

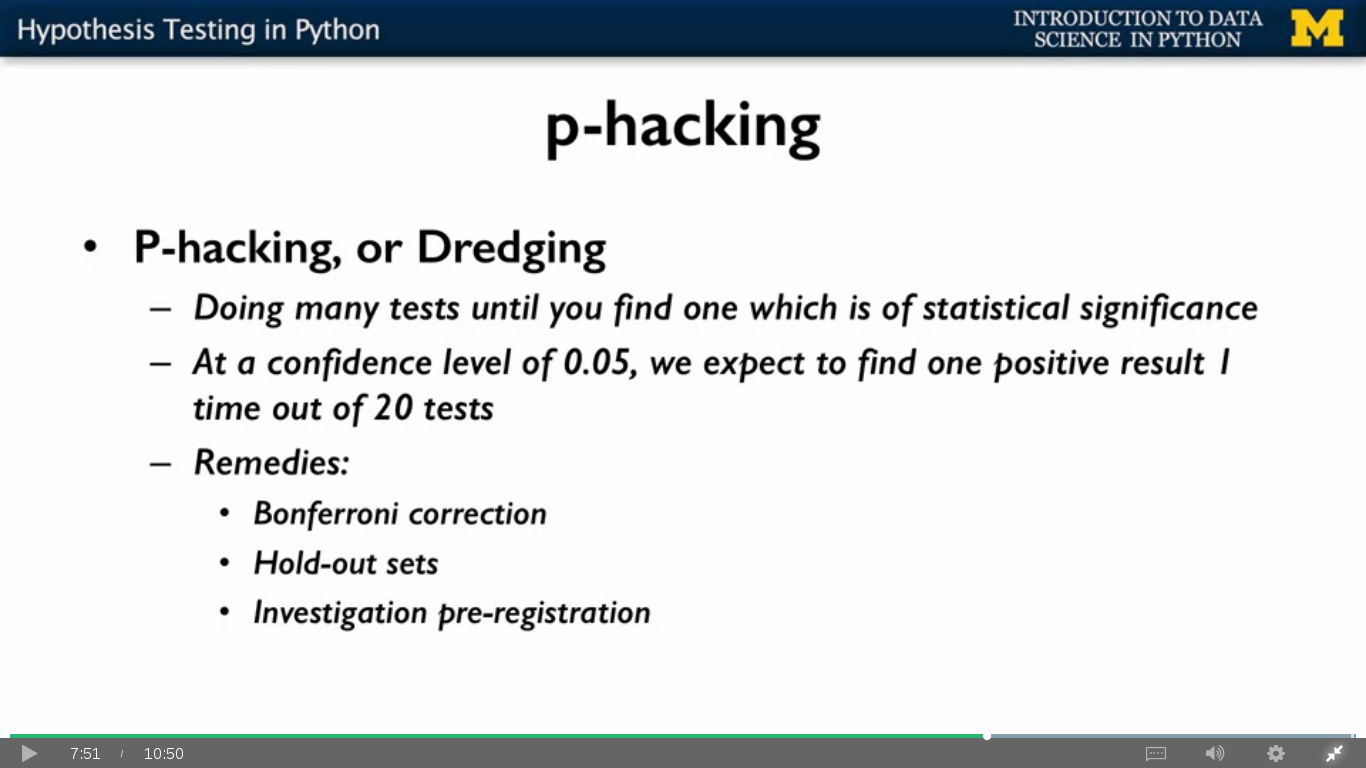
in this way, it's called p-hacking or dredging and it's a serious methodological issue

There are a couple of different ways you can deal with p-hacking

1.Bonferroni Correction

2.Hold- out sets

3.Investigation pre-registration



http://hamelg.blogspot.in/2015/11/python-for-data-analysis-part-24.html