## Spot the Differences

Without running the scripts, can you tell what the output will be? If you have some Python or programming

background, this section should take very little time.

### For Loops

```
# Script 1
list_num = [1,2,3]
for num in list_num:
    total = 0
    total += num
    print total
```

Solution:

```
1
2
3
```

```
# Script 2
list_num = [1,2,3]
total = 0
for num in list_num:
    total += num
    print total
```

Solution:

```
1
3
6
```

```
# Script 3
list_num = [1,2,3]
total = 0
for num in list_num:
    total += num
print total
```

Solution:

```
6
```

**For Loops in Functions**

```
# Script 1
def my_function1(my_list):
    output = []
    for item in my_list:
        output.append(item)
        return item

print my_function1(['cat', 'bad', 'dad'])
```

Solution:

```
‘cat’
‘bad’
‘dad’
```

```
# Script 2
def my_function2(my_list):
    output = []
    for item in my_list:
        output.append(item)
        return output

print my_function2(['cat', 'bad', 'dad'])
```

Solution:

[‘cat’]

[‘cat’,’bad’]

['cat', 'bad', 'dad']

```
# Script 3
def my_function3(my_list):
    output = []
    for item in my_list:
        output.append(item)
    return item

print my_function3(['cat', 'bad', 'dad'])
```

Solution:

'dad'

```
# Script 4
def my_function4(my_list):
    for item in my_list:
        output = []
        output.append(item)
    return output

print my_function4(['cat', 'bad', 'dad'])
```

Solution:

```
[‘dad’]
```

```
# Script 5
def my_function5(my_list):
    output = []
    for item in my_list:
        output.append(item)
    return output

print my_function5(['cat', 'bad', 'dad'])
print my_function5(['cat', 'bad', 'dad'])
```

Solution:

```
['cat', 'bad', 'dad']
['cat', 'bad', 'dad']
```

```
# Script 6
output = []
    def my_function6(my_list):
        for item in my_list:
            output.append(item)
        return output

print my_function6(['cat', 'bad', 'dad'])
print my_function6(['cat', 'bad', 'dad'])
```

Solution:

```
['cat', 'bad', 'dad']
```

```
['cat', 'bad', 'dad', 'cat', 'bad', 'dad']
```

## Make a function

Functions, blocks of reusable code, keep your code modular, well organized and easily maintainable. You should try to keep your code organized in functions. Take a look at each of the following snippets of code and organize them into functions.

1. We want a function that takes a list of numbers and returns that list where 10 was added to each number.

```
list_num = [1,2,3]
list_add_10 = []
for num in list_num:
list_add_10.append(num + 10)
print list_add_10
```

Solution:

```
    Def myfunction_1(list):
            list_add_10 = []
            for num in list:
                list_add_10.append(num + 10)
            return  list_add_10

    Print myfunction_1([1,2,3])
```

2. We want a function that takes in a list of strings and returns the list with the length of the words.

```
list_words = ['great', 'job', 'so', 'far']
list_length_words = []
for word in list_words:
list_length_words.append(len(word))
print list_length_words
```

Solution:

```
    Def myfunction_2(list):
            list_length_words = []
            for word in list:
                list_length_words.append(len(word))
            return  list_length_words


    Print myfunction_2(['great', 'job', 'so', 'far'])
```

## Challenge 1

Write a function that looks at the number of times given letters appear in a document. The output should be in a dictionary.

```
def letter_counter(path_to_file, letters_to_count):
''' Returns the number of times specified letters appear in a file

Parameters
-----------
path_to_file: str
Relative or absolute path to file of interest
```

~~~
letters_to_count: str
String containing the letters to count in the text

Returns
--------
letter_dict: dict
- key: letter
- value: the count of that letter in the file
The counting is case insensitive

Example
--------
```file.txt
This is the file of interest. Count my vowels!
```
>>> letter_counter('file.txt', 'aeiou')
{'i': 4, 'e':4, 'o':2, 'u':1}
'''
pass
~~~

**Solution:**

```
from collections import Counter
array1 =[]
def letter_counter(path_to_file, letters_to_count):
    with open(path_to_file) as f:
        array1= Counter(letter for line in f
                  for letter in line.lower()
                  if letter in letters_to_count.lower())
        return dict(array1)

letter_counter('C:\\Users\\curl\\xtek.txt.txt','Aeiou')
```

## Challenge 2

Write a function that removes one occurrence of a given item from a list. Do not use methods .pop() or .remove()! If the item is not present in the list, output should be ‘The item is not in the list’.

```
def remove_item(list_items, item_to_remove):
''' Remove first occurrence of item from list

Parameters
----------_
list_items: list
item_to_remove: object
The object to be removed form list_items

Returns
--------
- if the item is in the list: list
list with first occurrence of item removed
- if the item is not in the list: str
'The item is not in the list'

```

```
Example
--------
>>>list_items = [1,3,7,8,0]
>>>remove_item(list_items, 7)
[1,3,8,0]
'''
pass
```

**Solution:**

```
def remove_item(list_items, item_to_remove):

    if(item_to_remove in list_items):
        return  [i for i in  list_items   if i != item_to_remove]
    else:
        return 'The item is not in the list'

list_items = [1,3,7,8,0]
remove_item(list_items, 7)
remove_item(list_items, 100)
```

## Challenge 3

The simple substitution cipher basically consists of substituting every plaintext character for a different ciphertext character. The following is an example of one possible cipher from http://practicalcryptography.com/ciphers/simple-substitution-cipher/:

- Plain alphabet : abcdefghijklmnopqrstuvwxyz
- cipher alphabet: phqgiumeaylnofdxjkrcvstzwb

```
def cipher(text, cipher_alphabet, option='encipher'):
''' Run text through a particular cipher alphabet

Parameters
-----------
text: str
Either the plain text to encipher, or the cipher text to decrypt
cipher_alphabet: dict
Dictionary specifying {'original_letter': 'cipher_letter'}
option: str (default 'encipher')
'encipher' (accept plain text and output cipher text)
'decipher' (accept cipher text and output plain text)

Returns
--------
cipher text by default,
plain text if option is set to decipher

>>> d = dict(zip('abcdefghijklmnopqrstuvwxyz',
'phqgiumeaylnofdxjkrcvstzwb'))
>>> cipher('defend the east wall of the castle',
d)
'giuifg cei iprc tpnn du cei qprcni'
>>> cipher('giuifg cei iprc tpnn du cei qprcni',
d,
option='decopher')
'defend the east wall of the castle'
'''
pass
```

Solution:

```
def cipher(text, cipher_alphabet, option='encipher'):
    result = ''
    if(option == 'encipher'):
        result=  encrypt(cipher_alphabet,text)
    else:
        result=  decrypt(cipher_alphabet,text)
    return result

def encrypt(key, plaintext):
    return ''.join(' ' if l== ' ' else  key[l] for l in plaintext)

def decrypt(key, ciphertext):
    flipped = {v: k for k, v in key.items()}
    return ''.join(' ' if l== ' ' else flipped[l] for l in ciphertext)

d = dict(zip('abcdefghijklmnopqrstuvwxyz', 'phqgiumeaylnofdxjkrcvstzwb'))
cipher('defend the east wall of the castle', d)
cipher('giuifg cei iprc tpnn du cei qprcni',d,option='decipher')
```

## Challenge 4

Implement a function that counts the number of isograms in a list of strings.

- An isogram is a word that has no repeating letters, consecutive or non-consecutive.
- Assume the empty string is an isogram and that the function should be case insensitive.

```
def count_isograms(list_of_words):
''' Count the number of strings without repeating characters in a list

Parameters
-----------
list_of_words: list of strings

Returns
-------
count of isograms (as integer)

>>>count_isograms(['conduct', letter', 'contract', 'hours', 'interview'])
1
'''
pass
```

Solution:

```
def count_isograms(list_of_words):
    return [len(set(x.lower())) == len(x.lower()) for x in list_of_words].count(True)

count_isograms(['Conduct', 'letter', 'contract', 'hours', 'interview'])
```

## Challenge 5

Write a function that returns a list of matching items. Items are defined by a tuple with a letter and a number and we consider item 1 to match item 2 if:

1. Both their letters are vowels (aeiou), or both are consonnants and,
2. The sum of their numbers is a multiple of 3

(1,2) contains the same information as (2,1), the output list should only contain one of them.

```
def matching_pairs(data_list):
'''
Parameters
----------
data_list: as list of tuples (letter, number)

Returns
-------
A list of the matching pair referenced by their index (index_A, index_B).
Each pair should appear only once. (A,B) is the same as (B,A)

>>> data = [('a', 4), ('b', 5), ('c', 1), ('d', 3), ('e', 2), ('f',6)]
>>> matching_pairs(data)
[(0,4), (1,2), (3,5)]
'''
pass
```

**Solution:**

```
 import itertools
 def matching_pairs(data_list):
   vowel = ['a','e','i','o','u']
   d= [i  for i,t2 in enumerate(data_list) if t2[0] in vowel and sum(t2[1] for t2 in data_list) % 3 == 0   ]
   e =[i for i,t2 in enumerate(data_list)if t2[0] not in vowel and sum(t2[1] for t2 in data_list) % 3 == 0  ]
   return list( zip(d[0::2],d[1::2])) + list( zip(e[0::2], e[1::2]))

data = [('a', 4), ('b', 5), ('c', 1), ('d', 3), ('e', 2), ('f',6)]
matching_pairs(data)
```

# Getting Ready for the SQL Assessment

## Simple Queries on a Single Table

1. Use the WHERE clause to show the countries with a flag ratio of 2:3 (i.e. w_prop = 2 and l_prop = 3).

```
Select country from flags where w_prop = 2 and l_prop = 3
```

2. Use IN to check if an item is in a list and show the countries on a continent that is either Europe or North America.

```
select country
from countries
where contient IN ('Europe’, 'North America')
```

3. Use BETWEEN xxx AND xxx to show names of flags and countries that have width proportion higher than 1 but lower than 8.

```
select name, country
from flags
where w_prop between 2 and 8
```

4. Use LIKE 'X%' to show countries that have an name that starts with ‘U’.

```
select country from countries where country like 'U%'
```

5. Use CASE to show countries, their capital and a column to indicate whether the continent is ‘Eurasia’ (i.e. Europe or Asia) or ‘Americas’ (North or South America). Add a filter to select countries with capitals that are at least 7 character long.

```
select country, capital ,
    CASE WHEN contient IN ( 'Europe' ,'Asia') THEN 'Eurasia'
       WHEN  contient IN  ( 'North America' , 'South America') THEN 'Americas'
       ELSE 'Other'
       END AS Region
from countries
where LEN(capital) >= 7
```

## Build Queries with Aggregates

Aggregates include commands such as DISTINCT, COUNT, SUM, GROUP BY, HAVING, and ORDER BY. Try using these commands on the following questions!

8

1. Use DISTINCT to list the continents in the countries table - each continent should appear only once.

```
select distinct(contient) from countries
```

2. Use COUNT to see how many countries are in Europe.

```
select count(country) as CountriesInEurope from countries where contient= 'Europe'
```

3. Use GROUP BY to count how many countries are in each continent, with continents alphabetically ordered (hint: use ORDER BY).

```
 select contient, count(country) as CountriesInEachContinent from countries GROUP BY contient
order by contient asc
```

4. Use HAVING to determine which continents are represented at least twice in the countries table.

```
select contient, count(country) as CountriesInEachContinent from countries GROUP BY contient having
count(country) >= 2 order by contient asc
```

## Build Complex Queries on Multiple Tables

1. Use JOIN to display the capital, the country, and the flag name.

```
select c.capital, f.country , f.name as flagname from flags f INNER JOIN countries c on
f.country  = c.country
```

2. Use JOIN and WHERE to display the continents associated to the flags in the flags table when the flag has a name (i.e. not ‘NA’).

```
select c.contient , f.name as flagname from flags f INNER JOIN countries c on
f.country  = c.country where  f.name != 'NA'
```

3. Use JOIN and HAVING to display continents that have at least 2 countries represented as well as the average adoption date of the flag (as avg_date).

```
select c.contient, AVG(f.adoption_date) as avg_date, count(c.country) as CountriesInEachContinent

from flags f INNER JOIN countries c On f.country = c.country

GROUP BY c.contient having count(c.country) >= 2 order by c.contient asc
```

## Counting: permutations, combinations

**1. Permutations**

1. How many ways can you arrange the numbers 1, 2, 3, 4 and 5?

5!

2. How many ways can you arrange 1, 1, 2, 3, 4?

5! / 2! = 60

3. How many ways can you arrange two 3s and three 5s?

5! / 2! 3! = 10

**2. Combinations**

1. How many different poker hands (5 cards) can you have? A deck holds 52 cards.

52C5 = 2598960

2. There are five flavors of ice cream: Stracciatella, Mint chocolate chip, Cookies and Cream, Butter Pecan, Pistachio and Pralines and cream. How many three scoop ice-creams can you make if all the scoops must be different flavors?

5C3 = 10 ways

*Extra Credit*: what happens if you can take several scoops of the same flavor?

If repletion is allowed, then 5X5X5 …. .If  n is the number of scoops , then 5^n is the solution.

Some links: http://bit.ly/2iNIXSF, http://bit.ly/2jXlDiI

## Probability

**1. Probability of an event**

1. In a deck of cards (52 cards), what's the probability of #1 picking a queen? #2A heart? #3Of picking a card that's not a queen nor a heart?

#1 4/52

#2 13/52

#3  0.692

2. If I do not replace the cards, what is the probability of #1 picking 2 kings? #2  4 diamonds? #3How do these probabilities evolve if I replace the cards after each draw?

#1 4C2/52C2 =1/221

#2 13X12X11X10 / 52X51X50X49 =11/4165

#3 picking 2 kings with replacement  = 4/52 X 4/52  and picking 4 diamonds with replacement = (13 /52 )^4

**2. Probability of 2 or more events**

**Conditional probability**

1. What is the probability that the total of two dice is less than four, knowing that the first die is a 2?

Total sets = (1,1) (1,2) (2,1)

Probability of getting toatal 4 given first die is 2 = 1/3

2. 90% of candidates to a Web developer position can code both in Javascript and HTML. 70% of these candidates can code in Javascript and 50% can code in HTML. What is the probability that a candidate can code in HTML knowing that he can code in Javascript?

Let event h be coding HTML, j be the coding javascript.

P(hnj ) = 0.9 , P(h) = 50%  = 0.5, P(j) = 70% = 0.7

The event h and j are independent because knowing to code in Javascript does not depend on whether you know HTML or not.

P(h|j) = P(h n j ) / P(j) = P(h) x P(j) / P(j)
= P(j) = 0.7.

**Independent and dependent events**

1. Number of kids dressed as pumpkins or ghosts on Halloween night and the amount of candy they

received:
Amount of Candy less than 10 10 - 20 20 - 30 greater than 30
Pumpkins 5 10 60 25
Ghosts 15 40 80 15
• What is the probability that a kid dressed as a pumpkin gets 20 or more pieces of candy?

60+25/ total number of candies => 85/250 => .34
Howabout if he dresses as a ghost?
80+15/ total number of candies => 95/250 => .38
• What is the probability that a kid obtains less than 10 pieces of candy?
20 / 250 => 2/25 = .08
• What is the probability that two siblings, one dressed as a ghost and one dressed as a pumpkin, each receive 20 to 30 pieces of candy?
P(AnB) = P(A) .P(B) => 80/250 X 60/250 = 64/625 = 0.0768

2. You toss a fair die twice. What is the probability of getting less than 3 on the first toss and an even number on the second?
P(A) = 2/6
P(B) = 3/6
P(AnB) = 2/6 X 3/6 = 6/36 = 1/6

**Mutually exclusive events**
Let's consider a population from which we draw a sample of 40 individuals. The probability of your sample having no-one with glasses is 26%. The probability of having only one individual wearing glasses is 32%. What is the probability of
(a) Obtaining not more than one individual wearing glasses in a sample?
P(X<=1) = P(X=0)+P(X=1) = .26 +.32 = .58
(b) Obtaining more than one individual wearing glasses in a sample?
P(X>1) = 1 – P(X<=1) = .42

**Bayes' Theorem**
1. To detect a medical condition, patients are given two tests. 25% of the patients receive positive results on both tests and 42% of the patients receive positive results on the first test. What percent of those who have positive results on the first test passed also had positive result on the second test?

P( first test positive and second test positive) = 0.25

P(First test positive) = 0.42

P(Getting second test positive given that first test is positive)
= P( Second Positive | First positive)

P (Second positive n First Positive)
= -------------------------------------------------- = 0.25 / 0.42 = 0.595
P(First Positive)

2. Extra Credit: A jar contains red and blue marbles. You draw two marbles one after the other without replacing the first marble in the jar. You know that:
• The probability of selecting a blue marble and then a red marble is 30%.
• The probability of selecting a red marble on the first draw is 50%.
You first draw a red marble. What is the probability of selecting a blue marble on the second draw?

P( blue | red) = P( blue n red) /P(red) = 0.3 / 0.5 = 3/5 = 0.6

## Probability distributions Problems

Common problems relying on discrete (Binomial, Geometric, Poisson) or continuous (Uniform, Normal, Exponential) probability distributions.

Here are some exercises (http://bit.ly/2j7GK25) with their solutions as video.

**Binomial distribution**

1. Fair coin: Imagine you were to flip a fair coin 10 times. What would be the probability of getting 5 heads?

10C5 X (1/2)^5 X (1/2) ^5 = 252

2. Unfair coin: You have a coin with which you are 2 times more likely to get heads than tails. You flip the coin 100 times.

What is the probability of getting 20 tails?

P(H) = 2P(T)

P(H) = p=> p= 2(1-p) => P(H) = 2/3 and P(T) = 1/3

100C20 X (1/3)^20 X (2/3)^80 = 0.00125

What is the probability of getting atleast one heads?

P(H=0 , H=1) = P(H=0)+P(H=1) = 100C0 X (2/3)^0 X (1/3)^100 +100C1 X (2/3)^1 X (1/3)^99 < 0.000001

**Geometric distributions**

Suppose you have an unfair coin, with a 68% chance of getting tails. What is the probability that the first head will be on the 3rd trial?

P(X= 3) = P(TTH) = 0.68 X0.68 X0.32 = 0.147968

**Poisson distribution**

On average 20 taxis drive past your office every 30 minutes. What is the probability that 30 taxis will drive by in 1 hour?

- On average 40 taxis drive per one hour . So (μ) 40 is the average number of events per interval.
- P(K=30) = P(x; μ) = $(e^{-\mu})$ $(\mu^{x})$ / x! => 0.01847

**Exponential distribution**

Let *X*, the number of years a computer works, be a random variable that follows an exponential distribution with a lambda of 3 years. You just bought a computer, what is the probability that the computer will work in 8 years?

$P(X=8) = 3e^{-24} = 1.132e^{-10}$

*Extra Credit*: Let *X* be a random variable that now follows an exponential distribution with a half-life of 6 years. Find the parameter of the exponential distribution.

T1/2 = 6

6 = ln(2)/m

6 = 0.69314718056 / m
m= 0.115

What is the probability *P*(*X* > 10) and

P(X > 10) = 1- P(X <= 10) = 1- [1-$e^{-mx}$] where m is the decay parameter

= 1- [1-$e^{-0.11x}$] = $e^{-(0.11)(10)}$ = 0.3328

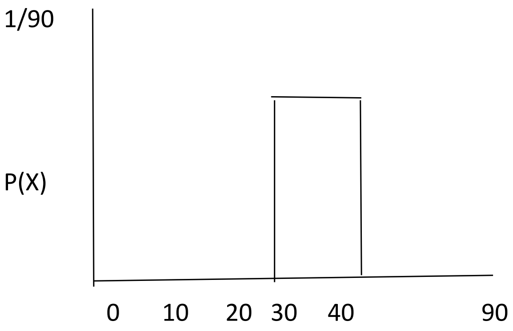
the conditional probability *P*(*X* > 20|*X* > 10)?

*P*(*X* > 20|*X* > 10) =P(X > 10) (by memory less property) = 0.3328

**Uniform distribution**

Let the random variable *X* be the angle of a slice of pizza. The angle *X* has a uniform distribution on the interval [0*,* 90].

What is the probability that your slice of pizza will have an angle between 30 and 40°?



P(30 <= X <= 40) = 10 X 1/90 = 1/9

*Extra Credit*: *X* is uniform on the interval [*a, b*], can you derive the expected value *E*(*X*)?

a is the minimum value and b is the maximum value of the random variable X.

f(x) = 1 /b-a for a<=x<=b

E(X) =

$$\int_a^b x\, f(x)\, dx =$$

=

$$\int_a^b x\ 1/(b-a)\, dx$$

$= 1/(b-a)\ [x^2/2]_a^b$

$= 1/2(b-a)\ (b^2 – a^2)$

E(x) = b+a / 2

The variance *V* (*X*)?

E(X^2] – E[X]^2

= E(X^2] – (b+a / 2 )^2

$$= \int_a^b x^2 f(x)\,dx\ - (b+a)^2 / 4$$

$= [1/b-a\ \ x^3 / 3]_a^b\ \ - (b+a)^2 / 4$

$= b^3 - a^3 /3(b-a) - (b+a)^2 / 4$

= (b-a) ^ 2  / 12

**Normal distribution**

1. Suppose *X* has a standard normal distribution. Compute *P*(*X* > 9), *P*(1 < *X* < 3) and *P*(*X* > −3).
Mean = 0 , Standard Deviation = 1
P(X > 9)
For a standard normal variable X , the values range from -3 to 3. P(X > 9) approximately equal to 0.003.

*P*(1 < *X* < 3)
Z score of 1 = 1 => P(Z < 1)  = 0.3413
Z score of 3 = 3 => P(Z < 3)  = 0.4987

*P*(1 < *X* < 3) = P(Z<3) – P(Z <1)
= 0.4987 - 0.3413 = 0.1574

*P*(*X* > −3)

Z score of 3  = 3  => P(Z < -3) = 0.9987

P(X > -3)

Z score of -3 = -3

P(Z > -3 ) = P(Z <3) = 0.9987

2. The weight in pounds of individuals in a population of interest has a normal distribution, with a mean of 150 and a standard deviation of 40.

What is the expected range of values that describe the weight of 68% of the population (Hint: use the empirical rule)?

The expected range of values fall within 110 and 190 pounds.

Of the people who weigh more than 170 pounds, what percent weigh more than 200 pounds (Hint: this is conditional probability)?

Z score for 170 = 0.5 => P(Z< 0.5) = 0.1915 . We need area above 0.5 so prob of individuals weighing above 170 is 0.3085

Z score of 200 = 1.25 => P(Z< 1.25) = 0.3944. We need area above 1.25 so prob of individuals weighing above 200 is 0.1056

P(X > 200 | X > 170) = P(X > 200 n X> 170 ) / P(x>170) = P(X > 200) / P(X > 170) = 0.1056 / 0.3085 = 0.3423

## Descriptive Statistics

**3 Measures of Average**

Give the mean, median and mode of the following data:
(20, 45, 68, 900, 57, 45, 33, 35, 45, 22)

Mean = 127

Median = Middle value of an ordered dataset ascending {20, 22, 33,35,45,45,45,57,68,900}= 45+45/2 = 45

Mode = 45

Do you think the mean is a good summary statistic? Why or why not?

Mean is good statistic though not the best. On an evenly distributed dataset, mean could tell us about the overall data set and the spread.

The mean is not a good statistic in the below situations:

- It is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value.
- If the data is skewed, mean loses its ability to provide best central location.

In the above data set outlier pulls the mean to the right, so the mean is not the central location for the given data set. Median and mode is the better statistic in this example.

**Variance, Range, IQR**

Give the mean, the variance, the standard deviation, the range and the interquartile of range of the following data:
(20, 45, 68, 900, 57, 45, 33, 35, 45, 22)

Ordered => {20, 22, 33,35,45,45,45,57,68,900}

Mean= 127

Variance = 66585.6

Standard Deviation = 258.0419

Range = 900 – 20 = 880

{20, 22, 33,35,45|,45,45,57,68,900}

IQR = 57 – 33 = 24

**Discrete random variables**

Give the expression of the mean and the variance for a discrete random variable *X*.
X = x1, x2, x3 …..xn

Mean =$\Sigma$xi / n

Variance = $\Sigma(\text{xi} - \mu)^2/\text{n}$

**Continuous random variables**

Give the expression of the mean and the variance for a continuous random variable *X*.

Expected value of X = E(X) = $\mu$

E(X) = $\int_a^b x\, f(x)dx$

Variance  = E((X- $\mu)^2$)

## Inferential Statistics

**1. Confidence intervals**

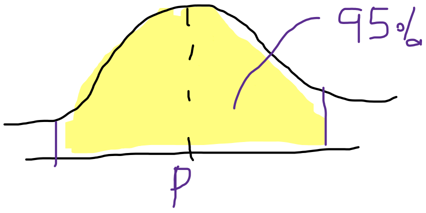
1. We are polling to get the approval rate of the president. Out of a population of 4 million, 6014 were surveyed and 3485 expressed their approval. Construct a 95% confidence interval for the approval rate of the president.

Sample size n = 6014

Using Bernoulli's distribution, mean = p = 1X 3485 + 0 X 2529 / 6014 = 0.5794

$S^2$ = 3485 X(1-0.5794)^2 + 2529 X ( 0 – 0.5794) ^2 / 6013 = 0.2422

S= Sqrt(0.2422) = 0.4922



To find the 95% confidence interval , we need to find the sample standard deviation of the sampling distribution
$\sigma_{\bar{\bar{x}}} = \sigma$   / Sqrt(n)

$\sigma$   is unknown but the best estimate would be the sample standard deviation.

$\sigma_{\bar{\bar{x}}} = \sigma$   / Sqrt(n)    = 0. 4922/ Sqrt(6014) = 0.0063

Z value for a 95% confidence is 1.96

95% chance that a random sample mean is within 1.96 $\sigma_{\bar{\vec{x}}}$ of p.

1.96 $\sigma_{\bar{\vec{x}}}$ = 0.0063 X 1.96 = 0.01234

Confident that 95% chance that p is within 0.01234 of the 0.5794.

So, 95% confidence interval of the approval rate of the president is 0.5794 – 0.01234 to 0.5794 + 0.01234, which is , 0.5670 to 0.5917.

2. The weight of a random sample of 100 individuals from a population of interest was surveyed and yielded a sample average weight of 150 pounds and sample standard deviation of 20 pounds. Construct a 95% confidence interval for the average weight of the population.

For a population with unknown mean $\mu$ and unknown standard deviation, a confidence interval for the population mean, based on a simple random sample (SRS) of size n, is $\overline{x}$ ± t* $\frac{s}{\sqrt{n}}$ , where t* is the upper (1-C)/2 critical value for the t distribution with n-1 degrees of freedom, t(n-1).

t* = 0.025, $\tilde{x}$ = 150, s = 20, n= 100

150 ±0.025 (20 / 10)

= 150- 0.05 to 150 + 0.05 => 95% confident that the p is within 149.95 to 150.05

**2. General hypothesis testing**

1. What is the definition of a significance level?

The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true.

Of a p-value?

P-values are the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis.

2. Would you use a one tailed or two tailed tests in the following cases:

- Investigating if women are paid less than men.

Left tailed test because we are interested in finding if the women are paid lesser.

- Comparing the click-through rate of website when the 'subscribe' button is green vs. when it is blue.

One tailed test.

3. A man goes to trial. In a hypothesis testing framework, let's define the null hypothesis as *Not Guilty* and the alternative hypothesis as *Guilty*.

- What type of error is made when the man is actually not guilty but verdict returned is guilty?

Type 1

- What type of error is made when the man is actually guilty but verdict returned is not guilty?

Type 2

**3. One sample hypothesis testing**

1. We want the test the hypothesis that at least 68% of the Canadian population (aged 18+) went to the movies at least once in the past 12 months with a significance level of 5%. We surveyed 4,000 respondents and found 3,012 did go at least once to the movies in the past 12 months. How would your conclusion compare if you only had 40 respondents, 30 of which went to the movies at least once in the past 12 months

Case 1:

Hypotheses:

Null Hypothesis : p >= 68% (claim) ; Alternate Hypothesis : p < 68%

$\bar{p}$= 3012 / 4000 = 0.753

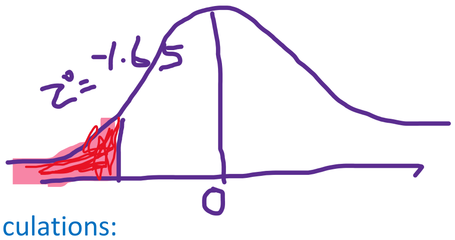
Assume null hypothesis is true,

$P_{H0}$ = 0.68

Distribution:

As the standard deviation of the sample proportion can be calculated and the sample size is large  and assumed to be normal, the left tailed z distribution will be chosen.

One tailed Z- distribution with significance level of 5 % is as follows:



Calculations:

p= 0.68

$\sigma_{H0}$ = Sqrt(0.68 X 0.32) = 0.466

$\sigma\ \ \bar{p} = \sigma_{H0}$ / Sqrt(4000) = 0.466 / 63.24 = 0.0073

Z =$\bar{p}$  - $u_{\bar{p}}$  / $\sigma\ \ \bar{p}$   = 9.58

Comparison:

9.58 > -1.65

Conclusion:

As the actual z value of the sample falls outside the significance level of 5%, **we accept the null hypothesis.**

95% confident that at least 68% of the Canadian population (aged 18+) went to the movies at least once in the past 12 months.

Case 2:

Hypotheses:

Null Hypothesis : p >= 68% ; Alternate Hypothesis : p < 68%
$\bar{p}$= 3012 / 4000 = 0.753

Assume null hypothesis is true,

$P_{H0}$ = 0.68

Distribution:

As the standard deviation of the sample proportion can be calculated and assumed to be normal, the left tailed z distribution will be chosen.

Calculations:

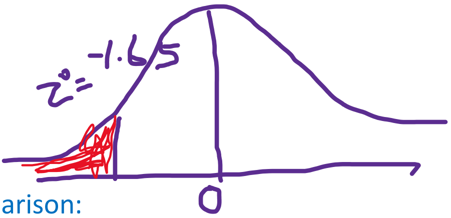
$u_{H_0}$= 0.68

$\sigma_{H0}$ = Sqrt(0.68 X 0.32) = 0.466

$\sigma\ \bar{p}$ = $\sigma_{H0}$ / Sqrt(40) = 0.466 / 6.324 = 0.073

Z =$\bar{p}$ - $u_{\bar{p}}$ / $\sigma\ \bar{p}$ = 0.75 – 0.68 / 0.073 = 0.96

One tailed Z- distribution with significance level of 5 % is as follows:



Comparison:

0.96 > -1.65

Conclusion:

As the actual z value of the sample falls outside the significance level of 5%, **we accept the null hypothesis.**

95% confident that at least 68% of the Canadian population (aged 18+) went to the movies at least once in the past 12 months.

Comparing case 1 and case 2:

We can say that both are taken from the same population based on the z statistics. Both the samples given are inclining into accepting the null hypothesis.

2. We want to test the hypothesis that the average weight in North America is at least 175 pounds. The mean of weights of the 100 individuals sampled is 178 pounds, with a sample standard deviation of 8 pounds. What are you conclusions?

Hypotheses:

Null Hypothesis : $\mu$ >= 175 ; Alternate Hypothesis : $\mu$ < 175

$\bar{x}$= 178 S= 8

Assume null hypothesis is true,

$u_{H_0}$= 175

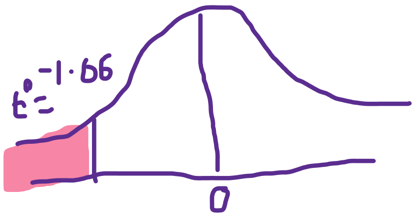
$\sigma_{H0}$ = 8

### Distribution:

Since this is a test about means where the population standard deviation is unknown I use the t-distribution.

Let significance level be 0.05. Degrees of Freedom = n – 1. The t critical = - 1.66

One tailed t- distribution with significance level of 5 % is as follows:



### Calculations:

$\sigma$ $\bar{x}$ = S / Sqrt(100) = 8/10 = 0.8

t =$\bar{x}$ - $u_{H_0}$ / $\sigma$ $\bar{x}$= 178 – 175 / 0.8 = 3.75

### Comparison:

3.75 > -1.66

### Conclusion:

As the actual z value of the sample falls above the significance level of 5%, **we accept the null hypothesis.**

95% confident that the average weight in North America is at least 175 pounds.

Some links: http://bit.ly/2jmht5d

3. We want to investigate the claim that on average, sea turtles lay 110 eggs in a nest. Volunteers have gone out and counted the number of eggs in 20 nest. What do you conclude?

- Data: 101, 120, 154, 89, 97, 132, 126, 105, 94, 111, 98, 90, 88, 115, 99, 85, 131, 127, 116

### Hypotheses:

Null Hypothesis : $\mu$ = 110 (claim); Alternate Hypothesis : $\mu \neq$ 110

$\bar{x}$= 109.36 S= 18.71

Assume null hypothesis is true,

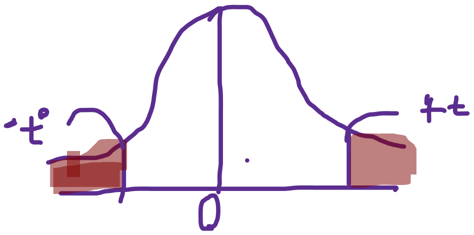
$u_{H_0}$= 110
$\sigma_{H0}$ = 18.71

Distribution:

Since this is a test about means where the population standard deviation is unknown I use the t-distribution.

Let significance level be 0.05. Degrees of Freedom = 19. The t critical = +/- 2.093

Two tailed t – distribution with significance value of 0.05 is as follows:



Calculations:

$\sigma$ $\bar{x}$ = 18.71 / Sqrt(20) = = 4.18

t =$\bar{x}$ - $u_{H_0}$ / $\sigma$ $\bar{x}$= 109.36 – 110 / 4.18 = -0.15

Comparison:

-2.093 < -0.15 < 2.093

Conclusion:

Fail to reject null hypothesis. Thus, it can be concluded that on an average, sea turtles lay 110 eggs in a nest.

Some links: http://bit.ly/2j7KpN2

**3. Two sample hypothesis testing**

1. Is there a meaningful difference between the proportion of teenagers vs that of adults that go to the movies at least once per month?

- Data:

– 1000 teenagers are surveyed, 780 answer positively. p1

– 1000 adults are surveyed, 620 answer positively.p2

Hypotheses:

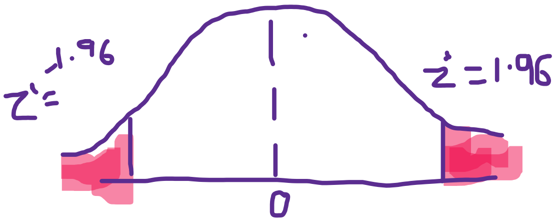
Null hypothesis: $P_1 = P_2$

Alternative hypothesis: $P_1 \neq P_2$

Distribution:

As the data given is proportions, the distribution is two tailed Z distribution.

Let us assume the significance level to be 0.05 with critical z value = +/- 1.96



Calculations:

Using sample data, we calculate the pooled sample proportion (p) and the standard error (SE). Using those measures, we compute the z-score test statistic (z).

$p = (p_1 * n_1 + p_2 * n_2) / (n_1 + n_2)$

p1 = 780 / 1000 = 0.780
p2 = 620 / 1000 = 0.620

p= 0.780 X 1000 + 0.620 X 1000 / 2000 = 780 + 620 / 2000 = 1400 / 2000 = 0.7

$SE = \text{sqrt}\{ p * ( 1 - p ) * [ (1/n_1) + (1/n_2) ] \}$

SE = sqrt( 0.7 X0.3 X [1/1000 + 1/ 1000]) = 0.0204

Z = p1 – p2 / SE = 0.780 – 0.620 / 0.0204 = 7.84

where $p_1$ is the sample proportion in sample 1, where $p_2$ is the sample proportion in sample 2, $n_1$ is the size of sample 1, and $n_2$ is the size of sample 2.

Comparison:

The decision rule is: Reject H0 if Z < -1.960 or if Z > 1.960.

7.8 > 1.96

Conclusion

Reject null hypothesis. There a meaningful difference between the proportion of teenagers vs that of adults that go to the movies at least once per month.

Some links: http://bit.ly/2j7GUXg
2. Is there a meaningful difference between the average wingspan of bald eagles vs that of crowned eagles?
• Data for bald eagles (in ft):
[7.4, 7.7, 6.0, 6.7, 8.3, 6.5, 6.9, 7.7, 7.8, 7.3, 6.9, 6.5, 6.3, 4.8, 8.0, 6.8,
5.8, 6.9, 6.3, 6.3, 6.4, 5.1, 6.9, 7.6, 5.6, 6.5, 6.7, 7.8, 6.6, 6.9, 7.0, 6.4, 7.4,
6.0, 7.0, 5.3, 5.8, 6.4, 7.1, 5.5, 7.0, 6.7, 5.8, 6.1, 7.1, 7.9, 7.7, 6.2, 5.3, 6.4,
6.9, 5.9, 7.8, 5.6, 5.0, 5.5, 6.4, 7.1, 8.6, 9.3, 6.8, 7.6, 7.2, 7.1, 5.8, 5.9, 5.1,
6.6, 6.8, 5.7, 6.3, 7.3, 6.3, 7.2, 7.7, 6.0, 7.2, 5.9, 7.2, 7.0, 7.4, 6.5, 7.8, 5.9,
6.3, 6.3, 8.3, 5.9, 6.9, 7.8]
• Data for crowned eagles (in ft):
[5.3, 5.6, 5.8, 5.3, 5.6, 4.9, 5.7, 5.4, 5.8, 5.4, 6.0, 5.4, 5.1, 5.4, 5.2, 5.7,
4.8, 5.8, 5.7, 5.1, 5.3, 5.4, 5.7, 6.6, 5.0, 5.4, 5.3, 5.5, 5.2, 5.6, 5.2, 5.9, 5.7,
5.8, 5.5, 5.2, 4.0, 5.8, 5.2, 6.2, 5.4, 4.6, 5.3, 5.8, 6.3, 4.8, 5.6, 5.4, 5.2, 5.4,
5.1, 6.0, 6.1, 5.4, 5.4, 5.3, 5.0, 6.0, 5.0, 5.8, 5.1, 5.3, 4.8, 5.6, 5.7, 6.1, 5.0,
6.4, 5.1, 4.6, 5.3, 6.0, 4.8, 5.4, 4.3, 5.4, 5.1, 4.7, 6.0, 5.5, 5.4, 5.6, 5.2, 5.8,
5.3, 4.9, 5.3, 5.5, 5.7, 4.7, 6.0, 5.6, 4.9, 5.4, 4.3, 5.5, 4.9, 5.3, 5.6, 6.0]

Hypotheses:

Null hypothesis: u1 = u2

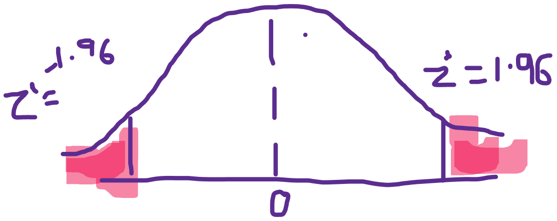
Alternative hypothesis: u1 ≠ u2

where u1, u2 are the means of the samples.

Distribution:

The distribution is two tailed Z distribution as the samples are large enough to have a normal distribution.

Let us assume the significance level to be 0.05 with critical z value = +/- 1.96



Calculations:

From the data sets, we can derive:

u1 = 6.7 u2 = 5.4 ; sd1 = 0.86 sd2= 0.45 n1=90 n2 =100

Z = u1 -u2 / sqrt(sd1^2 / n1 + sd2^2 /n2) = 6.7 – 5.4 / sqrt(0.008+0.002) = 13

Comparison:

The decision rule is: Reject H0 if Z < -1.960 or if Z > 1.960.

13 > 1.96

Conclusion:

Reject null hypothesis. There is a meaningful difference between the average wingspan of bald eagles vs that of crowned eagles

Some links: http://bit.ly/2jva7OY

14

## Relationship between two quantitative variables

1. Dataset

x 0 1 2 3 5

y 1 2.1 3.2 4 6.1

(a) Plot corresponding the scatter plot.

(b) Find the least square regression line *y* = *ax* + *b*. Add it to your plot.

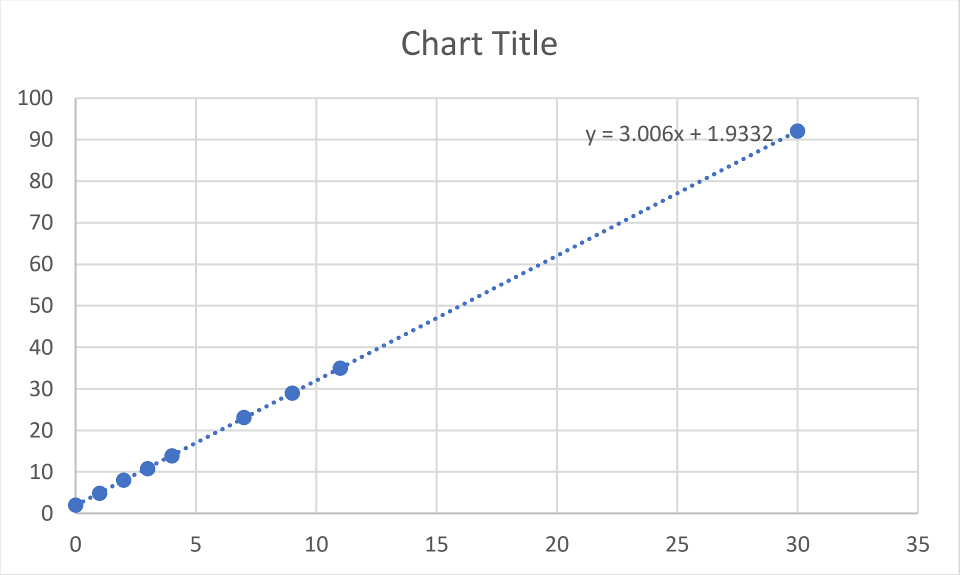
(c) Estimate the value of *y* when *x* = 4.

```
from scipy import stats
import numpy as np
import matplotlib.pyplot as plt


line = plt.figure()
x =[0, 1, 2, 3, 5]
y = [1, 2.1, 3.2, 4, 6.1]

fig, ax = plt.subplots()
ax.scatter(x,y,color="red")
plt.show()
fit = np.polyfit(x, y, 1)

# Find the slope and intercept of the best fit line
slope, intercept = np.polyfit(x, y, 1)
```

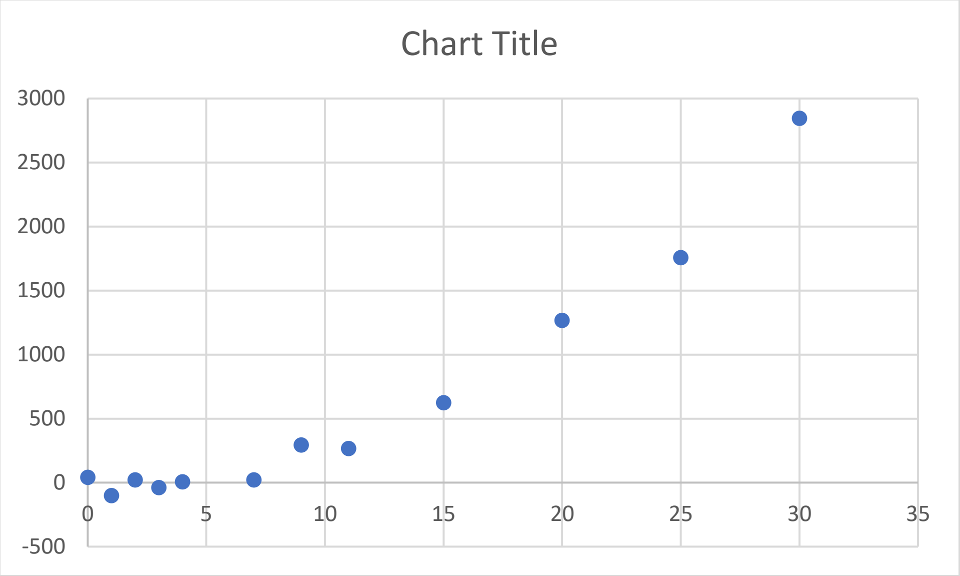
```
# Create a list of values in the best fit line
abline_values = [slope * i + intercept for i in x]

# Plot the best fit line over the actual values
plt.plot(x, y, '--')

plt.plot(x, abline_values, 'b')
plt.title(slope)

# if x=4 , y will be :
y4 =  [slope * 4 + intercept]

print(slope, intercept,y4)
plt.show()
```

*Extra Credit*: Can you do these steps in Python?

2. Dataset
x 0 1 2 3 4 7 9 11 30
y 2. 4.9 8. 10.8 13.9 23.1 29. 35. 92.1
(a) Find the least square regression line for the given data points.
(b) Plot the given points and the regression line on the same graph.

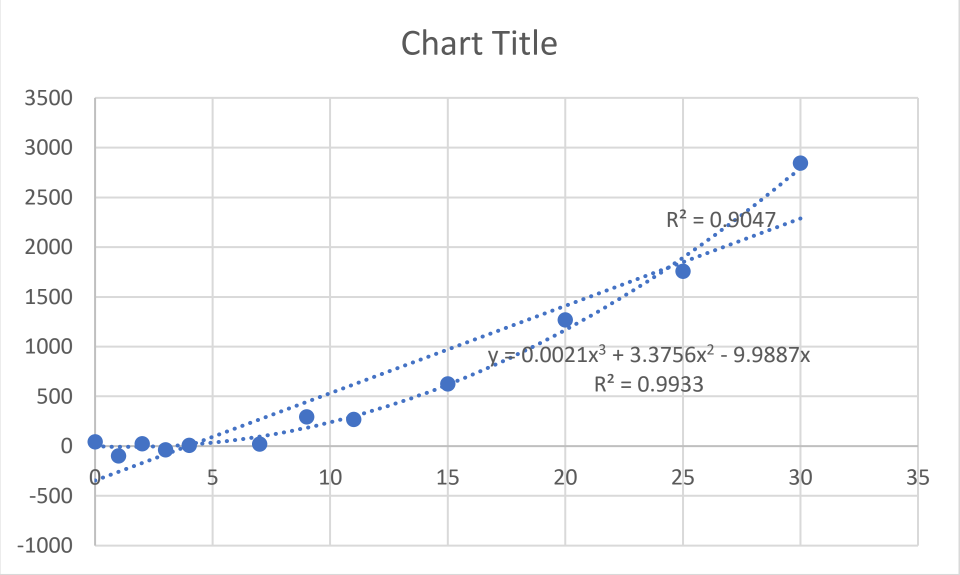


3. We have the following (x,y) points:
[(0, 42.0), (1, -101.0), (2, 21.0), (3, -38.0), (4, 5.0), (7, 20.0), (9, 293.0),
(11, 266.0), (15, 625.0), (20, 1266.0), (25, 1757.0), (30, 2844.0)]
(a) Plot the data.



(b) How do you think a linear model would perform? How about a 100 degree polynomial model? How would you figure out which of these models was preferable? (c) How would you model the relationship between these features?

The Linear regression line y = mx + b has most of the data points below the line and there is outlier, thus y = mx+ c is not the best fit line for this data set.



This line is not overfitting, thus there is room for additional data. The R Squared determines if the line is best fit and it looks like the Polynomial model with degree 3 fits best.

The polynomial with degree 100 will be too much away from all the points thus cannot be the right fit for this data.

The relationship can be found to be y = 0.0021 x^3 + 3.3756 x^2 -9.9887x.

One unit of increase in independent variable will decrease the dependent variable by approximately 7 units.

4. We have a dataset that gives the height and age of a sample of people. The range of age spans from 1 to 60 years. We decide to compute the correlation coefficient to model to understand the relationship between these features.
(a) Do you expect the correlation coefficient to be positive or negative?

Correlation coefficient will be positive but weak as the linear model would not fit the data well. Positive because there is increase in height as there is increase in age and weak because the model is non linear.

(b) What are some of the limitation of this approach?

- We are only considering LINEAR relationships.
- r and least squares regression are NOT resistant to outliers.
- There may be variables other than x which are not studied, yet do influence the response variable.
- A strong correlation does NOT imply cause and effect relationship.
  Correlation is not and cannot be taken to imply causation. Even if there is a very strong association between two variables we cannot assume that one causes the other.
  For example, suppose we found a positive correlation between watching violence on T.V. and violent behavior in adolescence. It could be that the cause of both these is a third (extraneous) variable - say for example, growing up in a violent home - and that both the watching of T.V. and the violent behavior are the outcome of this.
- Correlation does not allow us to go beyond the data that is given. For example, suppose it was found that there was an association between time spent on homework (1/2 hour to 3 hours) and number of G.C.S.E. passes (1 to 6). It would not be legitimate to infer from this that spending 6 hours on homework would be likely to generate 12 G.C.S.E. passes.

Some links: http://bit.ly/2jXyDF6, http://bit.ly/2jqXuRp, http://bit.ly/2jxlCFA

## Modeling

1. What is Linear Regression and Logistic Regression? How are they different?

If the outcome Y is a dichotomy with values 1 and 0, define p = E(Y|X), which is just the probability that Y is 1, given some value of the regressors X. Then the linear and logistic probability models are:

p = a0 + a1X1 + a2X2 + … + akXk    (linear)

ln[p/(1-p)] = b0 + b1X1 + b2X2 + … + bkXk       (logistic)

The linear model assumes that the probability p is a linear function of the regressors, while the logistic model assumes that the natural log of the odds p/(1-p) is a linear function of the regressors.

The major advantage of the linear model is its interpretability. In the linear model, if a1 is (say) .05, that means that a one-unit increase in X1 is associated with a 5 percentage point increase in the probability that Y is 1. Just about everyone has some understanding of what it would mean to increase by 5 percentage points their probability of, say, voting, or dying, or becoming obese.

The logistic model is less interpretable. In the logistic model, if b1 is .05, that means that a one-unit increase in X1 is associated with a .05 increase in the log odds that Y is 1.

2. Describe cross-validation and its role in model selection.
Cross-validation is primarily a way of measuring the predictive performance of a statistical model.

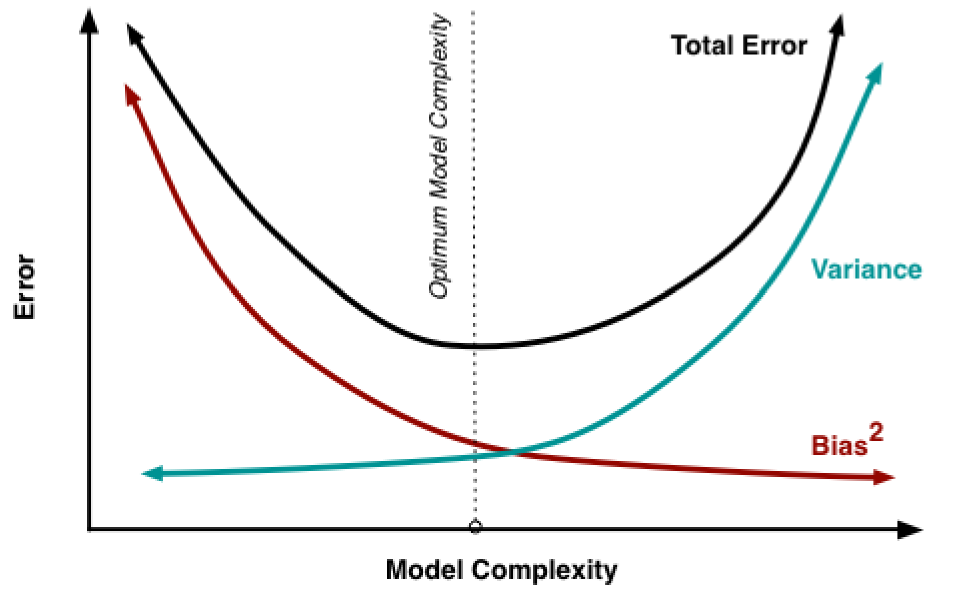
Cross-validation is performed by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Cross-validation error, estimates prediction error at any fixed value of the tuning parameter, and so by using it, we are implicitly assuming that achieving minimal prediction error is our goal, thus can assist in selecting the true model.

3. Generally speaking, as we increase the complexity of the model we are evaluating, how is the behavior of the model's bias and variance changing?

Bias has a negative first-order derivative in response to model complexity while variance has a positive slope.



4. A bank that grants auto loans is building a model, using historical sales data, to predict the price that a used car will sell for. Why is the average error between the predicted and actual price NOT an appropriate for evaluating the performance of the model?

The most commonly used scale-dependent measures are Mean Absolute Error (MAE), Mean Squared Error (MSE) and RMSE.

MAE may produce biased results when extremely large outliers exist in data sets.

MSE is more vulnerable to outliers since it gives extra weight to large errors.
RMSE is also sensitive to forecasting outliers.

The mean errors measurement of accuracy doesn’t work so well for time series data as there may be trends and other patterns in the data, making the mean a poor comparison.

The historical sales data will have different trends and patterns with large number of outliers making the mean error method inappropriate.

5. In linear regression, how should coefficients be interpreted? What is the difference between the size of a coefficient versus its statistical significance?

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model.

Y = mx + B is a linear regression model with coefficients m, the slope and B, Y- intercept.

That is, dependent variable = independent variable X slope + Y intercept. For example, weight = constant + slope x height, weight = -114.3 + 106 height.

The equation shows that the coefficient for height in meters is 106 kilograms. The coefficient indicates that for every additional meter in height you can expect weight to increase by an average of 106 kilograms.

6. Name two ways to measure the accuracy of a linear regression model.

1.Mean squared error
2. Cross validation

Some links: http://stanford.io/1Ry9D60