Applied Statistics MATH 661 Assignment #7

November 6, 2019

Andres Castellano

1 Task 1 Describing a relationship between two variables: IPC 2.147, 2.148, 2.149

1.1 IPC 2.147 Population in Canadian Provinces and Territories

1.1.1 a) A brief description of the data.

The scatter plot suggests the two variables are possibily inversely proportional. That is, as the percentage of the population under 15 years of age increases, the percentage of the population over 65 years of age decreases. There is one possible outlier in the data set over (>30% , <5%). The data is approximately linarly related, but the rate of change would differe drastically depending on whether or not the possible outlier is included.

1.1.2 b) Find the Correlation between the two variables.

[1]:
```
data = {'Province or Territory' : ['Alberta', 'British Columbia', 'Manitoba',␣
↪'New Brunswick', 'Newfoundland & Labrador',
                                  'Northwest Territories', 'Nova Scotia',␣
↪'Nunavaut', 'Ontario', 'Prince Edward Island',
                                  'Quebec', 'Saskatchewan', 'Yukon'],
       'Population' : [4124.7, 4631.3, 1282.0, 753.0, 527.0, 43.6, 942.7, 36.6,␣
↪13678.7, 146.3, 8214.7, 1125.4, 36.5],
       '% 15 & Under' : [18.3, 14.6, 18.7, 14.6, 14.4, 21.4, 14.1, 31.1, 16.0,␣
↪15.9, 15.4, 18.9, 16.6],
       '% 65 & over' : [11.4, 17.0, 14.6, 18.3, 17.7, 6.6, 18.3, 3.7, 15.6, 17.
↪9, 17.1, 14.5, 10.5]}

import pandas as pd
import seaborn as sns
data_frame = pd.DataFrame(data, index=data['Province or Territory'],
                          columns=pd.Index(['Population','% 15 & Under','% 65 &␣
↪over']))
t_framed=pd.DataFrame(data_frame.loc[['Northwest␣
↪Territories','Yukon','Nunavaut']]) # Territories
```

```
p_framed=pd.DataFrame(data_frame.drop(['Northwest␣
↪Territories','Yukon','Nunavaut']))
data_frame.corr()
```

[1]:

	Population	% 15 & Under	% 65 & over
Population	1.000000	-0.259210	0.248544
% 15 & Under	-0.259210	1.000000	-0.882948
% 65 & over	0.248544	-0.882948	1.000000

The correlation coefficient of -0.8829 is a good description of the relationship of the data. As stated before, the two variables appear to be inversely or negatively correlated. In addition, because the above calculation includes a possible outlier, we expected a good correlation but not great that is, close to |1| but not too close.

1.2 IPC 2.148 Nunavaut

1.2.1 a) Do I think Nunavaut is an outlier?

[2]:
```
data_frame
```

[2]:

	Population	% 15 & Under	% 65 & over
Alberta	4124.7	18.3	11.4
British Columbia	4631.3	14.6	17.0
Manitoba	1282.0	18.7	14.6
New Brunswick	753.0	14.6	18.3
Newfoundland & Labrador	527.0	14.4	17.7
Northwest Territories	43.6	21.4	6.6
Nova Scotia	942.7	14.1	18.3
Nunavaut	36.6	31.1	3.7
Ontario	13678.7	16.0	15.6
Prince Edward Island	146.3	15.9	17.9
Quebec	8214.7	15.4	17.1
Saskatchewan	1125.4	18.9	14.5
Yukon	36.5	16.6	10.5

[3]:
```
data_frame.describe()
```

[3]:

	Population	% 15 & Under	% 65 & over
count	13.000000	13.000000	13.000000
mean	2734.038462	17.692308	14.092308
std	4088.042568	4.580295	4.720604
min	36.500000	14.100000	3.700000
25%	146.300000	14.600000	11.400000
50%	942.700000	16.000000	15.600000
75%	4124.700000	18.700000	17.700000
max	13678.700000	31.100000	18.300000

[4]:
```
UpperLimit = 16.3 + (17.7-11.4)*1.5
print(UpperLimit)
```

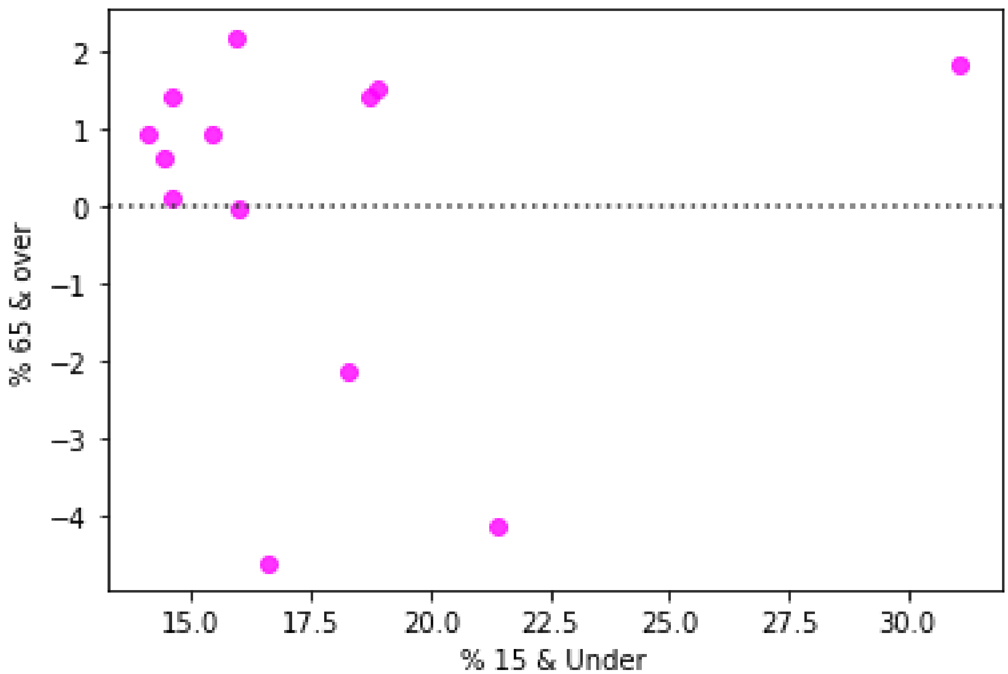
```
25.75
```

Yes, I believe Nunavaut is an outlier as it falls outside of the Median + 1.5xIQR.

1.2.2 b) Make a residual plot and comment on the size of the residual for Nunavaut. Use this iformation to expand on answer from part a.

[5]:
```
sns.residplot(data_frame['% 15 & Under'], data_frame['% 65 & over'],␣
↪color='magenta')
```

[5]: `<matplotlib.axes._subplots.AxesSubplot at 0x15c4839d0f0>`



The Residual value for Nunavaut implies that that data point is not similarly behaved compared to the rest of the data. Possibly an outlier.

1.2.3 c) Find the correlation values excluding the Nunavaut datapoint.

[6]:
```
data_frame=data_frame.drop(['Nunavaut'])
data_frame.corr()
```

[6]:

	Population	% 15 & Under	% 65 & over
Population	1.000000	-0.181900	0.159714
% 15 & Under	-0.181900	1.000000	-0.843924
% 65 & over	0.159714	-0.843924	1.000000

Surpirsingly, the correlation value obtained excluding Nunavaut is worse than the value obtained including Nunavaut: -0.8439 and -0.8829 respectively.

1.2.4 d) Nunavaut may not be an outlier after all.

Even though graphical and numercal analysis would suggest it is, it may just be a case that it does follow the same distribution as the other provinces but there is not enough data to see this. Perhaps analyzing the data by province is not the right approach.

1.3 IPC 2.149 "Split" data into provinces and Territories

[7]:
```
t_framed #Territories
```

[7]:

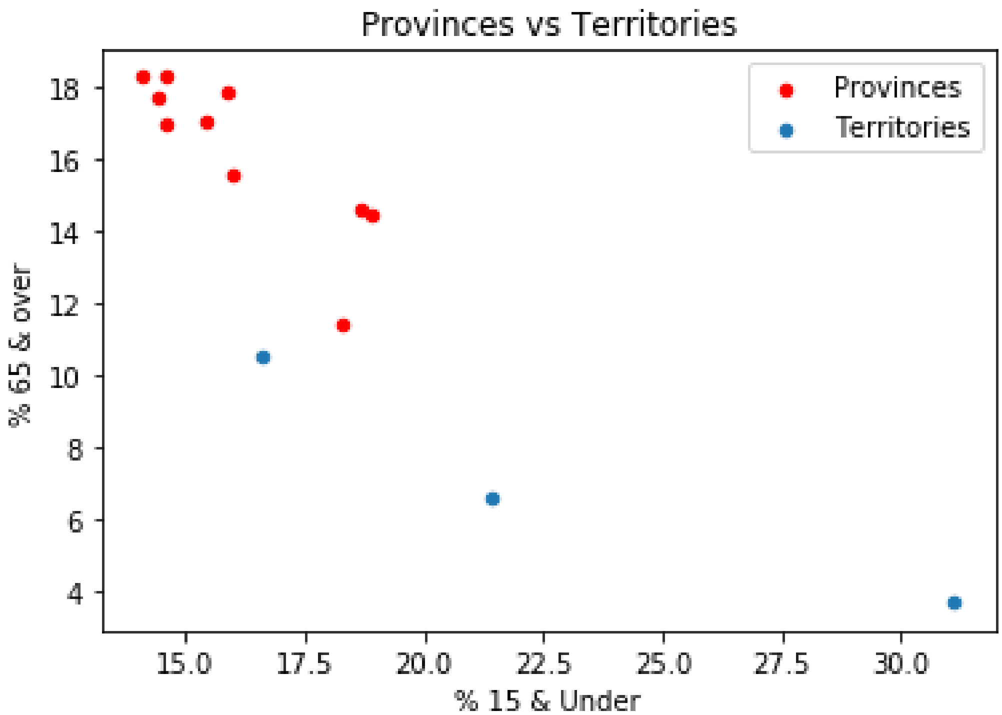
	Population	% 15 & Under	% 65 & over
Northwest Territories	43.6	21.4	6.6
Yukon	36.5	16.6	10.5
Nunavaut	36.6	31.1	3.7

[8]:
```
p_framed #Provinces
```

[8]:

	Population	% 15 & Under	% 65 & over
Alberta	4124.7	18.3	11.4
British Columbia	4631.3	14.6	17.0
Manitoba	1282.0	18.7	14.6
New Brunswick	753.0	14.6	18.3
Newfoundland & Labrador	527.0	14.4	17.7
Nova Scotia	942.7	14.1	18.3
Ontario	13678.7	16.0	15.6
Prince Edward Island	146.3	15.9	17.9
Quebec	8214.7	15.4	17.1
Saskatchewan	1125.4	18.9	14.5

[9]:
```
ax1 = p_framed.plot.scatter(x='% 15 & Under', y='% 65 & over',␣
↪c='r',label='Provinces');
ax2 = t_framed.plot.scatter(x='% 15 & Under', y='% 65 & over',␣
↪ax=ax1,label='Territories',title='Provinces vs Territories');
```



1.3.1 b) Splitting The Data into Provinces and Territories provides a better picture of how the data is distributed.

---

2 Task 2 Probability of an Event: IPC 4.135 and IPC 4.136

2.1 IPC 4.135

Multiplication Rule applies to independent events. P = 0.006, notP= 1-.006=.994

Thus probability of first win on the tenth day is equal to the probability of no wins (notP) in the first 9 days and a win (P) on the 1oth day.

[10]:
```
Probability = ((.994)**9)*(.006)
print('The proability of the 1st win on the 10th day is ',Probability)
```

```
The proability of the 1st win on the 10th day is  0.005683668109920798
```

3 Task 3 Marginal and Conditional Probabilities

3.1 IPC 4.136

[11]:
```
edu_data = {'Type' : ['Two-Year','Four-Year','Total'],
           'Public' : [1000/5167,2774/5167,3774/5167],
           'Private' : [721/5167, 672/5167,1393/5167],
           'Total' : [1721/5167,3446/5167,5167/5167]}
edu_frame = pd.DataFrame(edu_data, index=edu_data['Type'],
                        columns=pd.Index(['Public','Private','Total']))
edu_frame
```

[11]:

	Public	Private	Total
Two-Year	0.193536	0.139539	0.333075
Four-Year	0.536869	0.130056	0.666925
Total	0.730404	0.269596	1.000000

In the U.S, the majority (53.68%) of higher education institutions are Four-Year Public institutions. Two year Public institutions account for 19.35% of all institutions, two year private institutions account for 13.95% and four year private institutions are 13.00% of all institutions. 73.04% of all institutions are Public and 26.95% are Private, 33.30% are Two-Year institutions and 66.69% are Four-Year institutions. ***

4 Task 4 Mechanics of Confidence Intervals IPC 6.12, 6.13,6.14

4.1 IPC 6.12

4.1.1 a) Give 95% Confidence Interval.

A confidence level of 95% requires the non-varying population mean to be contained in the interval $\mu \pm 2\sigma$ Since the margin is given as 5, the confidence interval is [73,83]

4.1.2 b) If a 99% confidence level was desired…

the margin of error would have to be greater in order to make up for increased confidence. Generally, confidence level and interval size are inversely proportional.

4.2 IPC 6.13

[12]:
```
from math import sqrt
sigma9 = 20/sqrt(9)
sigma25 = 20/sqrt(25)
sigma81 = 20/sqrt(81)
sigma100= 20/sqrt(100)
intervals = {'Sample Size' : [9,25,81,100],
            'Lower Bound' : [78-1.96*sigma9,78-1.96*sigma25,78-1.
 ↪96*sigma81,78-1.96*sigma100],
            'Upper Bound' : [78+1.96*sigma9,78+1.96*sigma25,78+1.
 ↪96*sigma81,78+1.96*sigma100]}
```

```
Interval_Frame = pd.DataFrame(intervals, index=intervals['Sample Size'],
                              columns=pd.Index(['Lower Bound','Upper␣
 ↪Bound'],name='Sample Size'))
Interval_Frame
```

[12]:

Sample Size	Lower Bound	Upper Bound
9	64.933333	91.066667
25	70.160000	85.840000
81	73.644444	82.355556
100	74.080000	81.920000

The table above suggests that as the sample size increases, the confidence interval shrinks, this is explained by the relationship between sample standard deviation, population standandard deviation and sample size.

sample standard deviation \$ = $\sigma / \sqrt{n}$\$

From the equation, it is clear that as n increases, sample standard deviation decreases and thus confidence interval shrinks, population standard deviation \$ σ\$*remainsconstantasitshould*.

4.3 IPC 6.14 Effect of confidence level of interval length.

[13]:
```
sigma64 = 20/sqrt(64)
intervals = {'Confidence Level' : ['80%','90%','95%','99%'],
             'Lower Bound' : [78-1.28*sigma64,78-1.645*sigma64,78-1.
 ↪96*sigma64,78-2.576*sigma64],
             'Upper Bound' : [78+1.28*sigma64,78+1.645*sigma64,78+1.
 ↪96*sigma64,78+2.576*sigma64]}
Interval_Frame = pd.DataFrame(intervals, index=intervals['Confidence Level'],
                              columns=pd.Index(['Lower Bound','Upper␣
 ↪Bound'],name='Confidence Level'))
Interval_Frame
```

[13]:

Confidence Level	Lower Bound	Upper Bound
80%	74.8000	81.2000
90%	73.8875	82.1125
95%	73.1000	82.9000
99%	71.5600	84.4400

As the confidence level increases, the confidence interval extends.

[ ]:
```
```