In [24]:
```
%reset
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy
from scipy import stats
import statsmodels.api as sm
```

An article in Nature describes an experiment to investigate the effect on consuming chocolate on cardiovascular health (“Plasma Antioxidants from Chocolate,” 2003, Vol. 424, p. 1013). The experiment consisted of using three different types of chocolates: 100 g of dark chocolate, 100 g of dark chocolate with 200 ml of full-fat milk, and 200 g of milk chocolate. Twelve subjects were used, seven women and five men with an average age range of 32.2 ± 1 years, an average weight of 65.8 ± 3.1 kg, and body mass index of 21.9 ± 0.4 kg m−2 . On different days, a subject consumed one of the chocolate-factor levels, and one hour later total antioxidant capacity of that person’s blood plasma was measured in an assay. Data similar to those summarized in the article follow.

DC		DC + MK		MC	
118.8	115.8	105.4	100.0	102.1	102.8
122.6	115.1	101.1	99.8	105.8	98.7
115.6	116.9	102.7	102.6	99.6	94.7
113.6	115.4	97.1	100.9	102.7	97.8
119.5	115.6	101.9	104.5	98.8	99.7
115.9	107.9	98.9	93.5	100.9	98.6

icle follow.

Step 1:

Construct box plots to compare the factor levels.

In [25]:
```
# read in the data

DC=pd.Series([],name="DC")
DM=pd.Series([],name="DM")
MC=pd.Series([],name="MC")

Total=pd.Series([118.8,122.6,115.6,113.6,119.5,115.9,115.8,115.1,116.9,115.4,115.6,107.9,105.4
```

In [26]:
```
# create a pandas dataframe of the name of df using pd.concat; axis=1 is for columns

df=pd.concat([DC,DM,MC],axis='columns')
df
```

Out[26]:

DC	DM	MC

In [27]:
```
# Plot the data to visualize: Create a BoxPlot
```

In [28]:
```
# Describe the data
```

Step 2

Perform an ANOVA and provide an ANOVA table

The ANOVA table decomposes the variance into the following component sum of squares:

- Total sum of squares. The degrees of freedom for this entry is the number of observations minus one.
- Sum of squares for the factor. The degrees of freedom for this entry is the number of levels minus one. The mean square is the sum of squares divided by the number of degrees of freedom.
- Residual sum of squares. The degrees of freedom is the total degrees of freedom minus the factor degrees of freedom. The mean square is the sum of squares divided by the number of degrees of freedom.

The sums of squares summarize how much of the variance in the data (total sum of squares) is accounted for by the factor effect (batch sum of squares) and how much is random error (residual sum of squares). Ideally, we would like most of the variance to be explained by the factor effect.

The ANOVA table provides a formal F test for the factor effect. For our example, we are testing the following hypothesis.

```
H0: All individual chocolate means are equal
```

$\mu_{1} = \mu_{2} = \mu_{3}$ OR $\tau_{1} = \tau_{2} = \tau_{3} = 0$

```
Ha: At least one chocolate mean is not equal to the others
```

μ_{1} and/or μ_{2} and/or μ_{3} $\neq 0$ OR $\tau_{i} \neq 0$ for at least one i

Thus, if the null hypothesis is true, each observation consists of the overall mean μ plus a realization of the random error component εij . This is equivalent to saying that all N observations are taken from a normal distribution with mean μ and variance $\sigma 2$. Therefore, if the null hypothesis is true, changing the levels of the factor (treatment effects are defined as deviations form the overall mean) has no effect on the mean response. e.

The linear statistical model is:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

- Yij is a random variable denoting the (ij)th observation
- u is the overall mean (the mean parameter common to all treatments)
- τ_i is the treatment effect (the parameter associated with the ith treatment)
- ε_{ij} is a random error component

Assume that the errors are normally and independently distributed with mean=0 and variance = σ^2.

TABLE 13.2 Typical Data for a Single-Factor Experiment

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots\vdots\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

Use Python to run an ANOVA

In python: One-Way ANOVA

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html

In [ ]:
```
# Run the one-way ANOVA in python
```

Run this ANOVA analysis by hand

Fill out the following ANOVA table:

ANOVA				
SOURCE	**SS**	**dof**	**MS**	**F**
Treatment	$SS_{Treatment} = n\sum_{i=1}^{a}(\bar{y}_i - \bar{y})^2$	a - 1	$MS_{Treatment} = \frac{SS_{\ Treatment}}{dof_{\ Treatment}}$	$\frac{MS_{Treatment}}{MS_{Error}}$
Error	$SS_{Error} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$	a(n-1)	$MS_{Error} = \frac{SS_{\ Error}}{dof_{\ Error}}$	
Total	$SS_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2$	an - 1		

What is a?

What is n?

What is N?

In [14]:
```
# Find the Total SS
```

In [15]:
```
# Find the Tretament SS (the Between SS)
```

In [16]:
```
# Find the Treatment MS
```

In [17]:
```
# Find the Error SS (the Within SS)
```

In [18]:
```
# Find the MS Error
```

The test statistic is the F-statistic, calculate this F-statistic and test it

In [19]:
```
# Find the F statistic
```

F(A=.95)	df1								
df2	1	2	3	4	5	6	7	8	9
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040

The F statistic is the group mean square divided by the error mean square.

This statistic follows an F distribution with (a-1) and a(n-1) degrees of freedom. For our example, the critical F value (upper tail) for $\alpha = 0.05$, (a-1) = 2, and a(n-1) = 33 is ~3.3.

You can find the critical value for F and the p-value through python as well.

In [20]:
```
from scipy.stats import f
```

In [21]:
```
# critical F to compare against
```

In [22]:
```
# Find the exact p-value
```

In [23]:
```
# Make a conclusion
```

In [ ]:

In [ ]: