CHAPTER 1

*1. A teacher wishes to know whether the males in his/her class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared. Is this an example of descriptive or inferential statistics?*

This is an example of **Descriptive statistics** as the data is not collected from a sample but rather from the entire population which in this case is all students in the teacher's class.

*3. If you are told that you scored in the 80th percentile,R from just this information would you know exactly what that means and how it was calculated? Explain.*

There are three methods of defining the percentile. The first, assumes that X percentile can be defined as the lowest score that is greater than X% of the scores, the second assumption is the smallest score that is greater than or equal to X% of the scores. There is even a third interpretation that takes weighted average of the percentiles computed in the above 2 definitions.

However, in general, the Percentile divides the entire space into 100 equal portions. So if you scored in the $80^{th}$ percentile, it would mean that you scored more than 80% of students in the class or in other words, 20% of the class scored better than you. Steps for calculating percentile :

1) Arrange the scores of the 100 students in increasing order.
2) Then compute the percentile using the following formula :

$$\mathbf{P = (n/N) \times 100}$$

Where P = the percentile, n = number of students with score lesser than my score in the ordered list, N = total number of students

*5. Give an example of an independent and a dependent variable.*

An independent variable is the cause and the dependent variable is the effect. Therefore, **body temperature** is an independent variable and classifying the **condition** as Normal, Fever, Hypothermia etc. based on the body temperature would be the dependant variable.

*7. Specify the level of measurement used for the items*

*Rating of the quality of a movie on a 7-point scale -* **Ordinal**
*Age -* **Ratio**
*Country you were born in -* **Nominal**

*Favourite Colour -* **Nominal**
*Time to respond to a question –* **Ratio** (here time is Ratio as it is a duration and it has a meaningful zero)

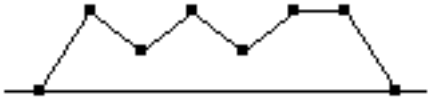
*9. The formula for finding each student's test grade (g) from his or her raw score (s) on a test is as follows: g = 16 + 3s*
*Is this a linear transformation? If a student got a raw score of 20, what is his test grade?*

**Yes**, this is a linear transformation as the transformation involves multiplying by one constant and then adding a second constant. If a student got a raw score of 20, then his/her test grade = 16 + 60 = **76**

*11. Which of the frequency polygons has a large positive skew? Which has a large negative skew?*

A.

B.



C.

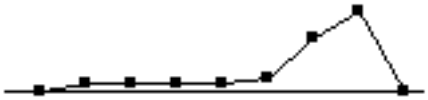


Image A has a large positive skew and it can be described as *skewed to the right* as it has a longer tail extending in the positive direction.

Similarly, Image C has a negative skew and it can be described as *skewed to the left* as it has a longer tail extending in the negative direction.

CHAPTER 2

*1. Name some ways to graph quantitative variables and some ways to graph qualitative variables.*

Graphs for Quantitative variables :

- Histograms
- Line chart
- Scatter plot
- Stem and leaf displays
- Frequency polygons
- Box plots
- Dot plots

Graphs for Qualitative variables :

- Pie chart
- Bar graph
- Pareto charts

*3. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the total number of pieces correctly remembered from three chess positions. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots?*

| Non-players | Beginners | Tournament players |
|---|---|---|
| 22.1 | 32.5 | 40.1 |
| 22.3 | 37.1 | 45.6 |
| 26.2 | 39.1 | 51.2 |
| 29.6 | 40.5 | 56.4 |
| 31.7 | 45.5 | 58.1 |
| 33.5 | 51.3 | 71.1 |

| | | |
|---|---|---|
| 38.9 | 52.6 | 74.9 |
| 39.7 | 55.7 | 75.9 |
| 43.2 | 55.9 | 80.3 |
| 43.2 | 57.7 | 85.3 |

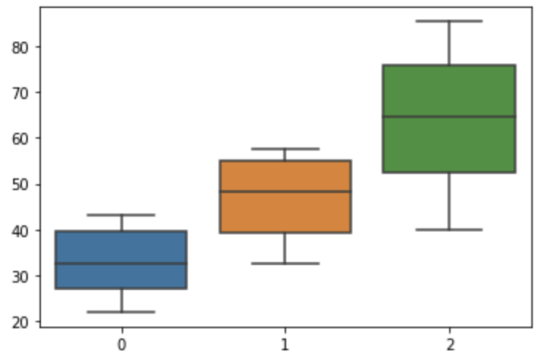
Descriptive Statistics for the three groups :

```
NON PLAYERS :              BEGINNERS :              TOURNAMENT PLAYERS :
count    10.000000         count    10.000000      count    10.000000
mean     33.040000         mean     46.790000      mean     63.890000
std       8.033292         std       9.030621      std      15.621456
min      22.100000         min      32.500000      min      40.100000
25%      27.050000         25%      39.450000      25%      52.500000
50%      32.600000         50%      48.400000      50%      64.600000
75%      39.500000         75%      54.925000      75%      75.650000
max      43.200000         max      57.700000      max      85.300000
```

```
import seaborn as sns
data = { "non_players" : [22.1,22.3,26.2,29.6,31.7,33.5,38.9,39.7,43.2,43.2],
         "beginners" :[32.5,37.1,39.1,40.5,45.5,51.3,52.6,55.7,55.9,57.7],
         "tournament_players" : [40.1,45.6,51.2,56.4,58.1,71.1,74.9,75.9,80.3,85.3]}
df = pd.DataFrame(data)

sns.boxplot(data=[df["non_players"], df["beginners"], df["tournament_players"]])
```

```
<AxesSubplot:>
```




From the plot we can observer that the mean number of pieces correctly remembered increases with increase in the experience level of the chess players. It can also be observed that with increase in the level of experience, the range for the middle 50% ( i.e., between 25$^{th}$ percentile and 75$^{th}$ percentile – Interquartile range) also increases.

*5. In a box plot, what percent of the scores are between the lower and upper hinges?*

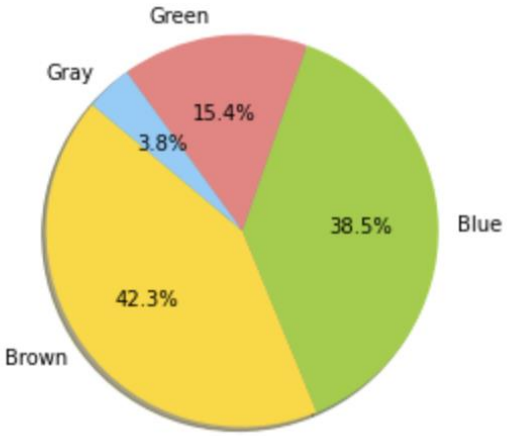
50% of the scores are between the lower and upper hinges. The range of scores between the lower quartile and upper quartile is called Inter Quartile Range. All values between the $25^{th}$ and $75^{th}$ percentile lie between the lower and upper hinges of a box plot.

*7. For the data from the 1977 Stat. and Biom. 200 class for eye colour, construct:*

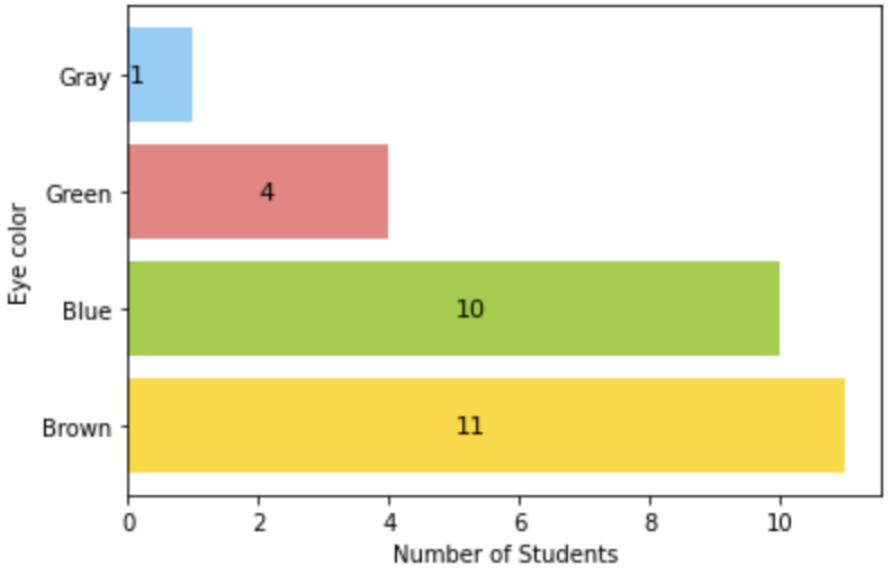
a. *pie graph*
b. *horizontal bar graph*
c. *vertical bar graph*
d. *a frequency table with the relative frequency of each eye colour*

| Eye Color | Number of students |
|---|---|
| Brown | 11 |
| Blue | 10 |
| Green | 4 |
| Gray | 1 |

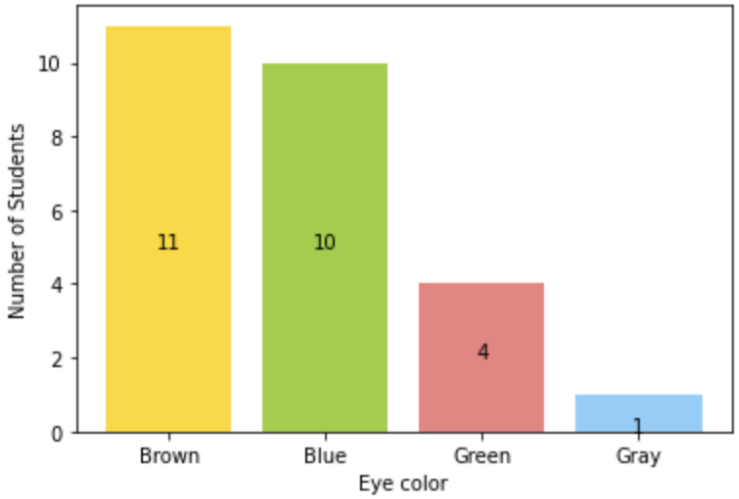
a. Pie graph




b. Horizontal bar graph



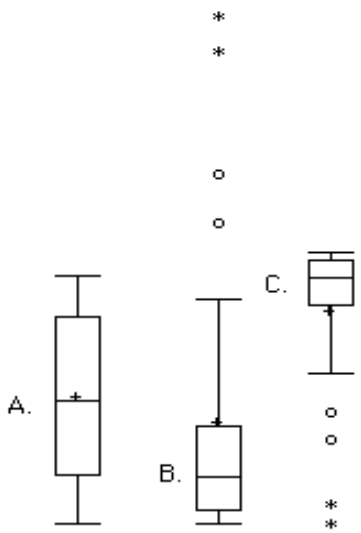

c. Vertical bar graph




d. Frequency table with relative frequency of each eye color

| Eye Color | Frequency f | Relative Frequency, Rf = f/26 | Rf(%) |
|---|---|---|---|
| Brown | 11 | 0.423 | 42.3% |
| Blue | 10 | 0.385 | 38.5% |
| Green | 4 | 0.154 | 15.4% |
| Grey | 1 | 0.038 | 3.8% |
| *Total* | *26* | *1.000* | *100%* |

*9. Which of the box plots below has a large positive skew? Which has a large negative skew?*




Graph A has a symmetric distribution, as the values are equally distributed around the center of the distribution.

Graph B has large positive skew because it has a longer whisker in the positive direction. Also its mean is greater than the median, indicating a positive skew.

Graph C has large negative skew because it has a long whisker in the negative direction with its median closer to the upper quartile. Also its mean is lesser than its median, indicating a negative skew.

CHAPTER 3

*1. Make up a dataset of 12 numbers with a positive skew. Use a statistical program to compute the skew. Is the mean larger than the median as it usually is for distributions with a positive skew? What is the value for skew?*

```
import pandas as pd

s = pd.Series([10, 26, 30, 42, 51, 67, 32, 12, 78, 51, 26, 21])
s.describe()
```

```
count    12.000000
mean     37.166667
std      21.156702
min      10.000000
25%      24.750000
50%      31.000000
75%      51.000000
max      78.000000
dtype: float64
```

The given dataset s with 12 numbers has a positive skew. It can be seen that the mean (37.16) is greater than the median (31.0) as it usually is for distributions with a positive skew.

Computing the skew using the Pearson's measure of skew :

```
#Pearson's measure of skew = 3(Mean - Median)/SD
Mean = statistics.mean(s)
Median = statistics.median(s) #50%
SD = statistics.stdev(s)

skew = (3*(Mean-Median))/SD
print("Skew :", skew)
```

```
Skew : 0.8744273830666093
```

The skew value calculated is **0.874**. Positive values for the skewness indicate data that are skewed right, meaning the right tail is long relative to the left tail. Hence, the given data set is said to have a positive skew.

*3. Make up three data sets with 5 numbers each that have:*

*(a) the same mean but different standard deviations.*

*(b) the same mean but different medians.*

*(c) the same median but different means.*

a) Same mean but different standard deviations

Set 1 : {5,10,15,20,25} Mean = 50, SD = 7.90

Set 2 : {7,14,9,2,43} Mean = 50, SD = 16.23
Set 3 : {53,2,13,4,3} Mean = 50, SD = 21.69

b) Same mean but different medians
Set 1 : {12,24,40,56,21} Mean = 30.6, Median = 24
Set 2 : {42,47,8,33,23} Mean = 30.6, Median = 33
Set 3 : {20,65,11,30,27} Mean = 30.6, Median = 27

c) Same median but different means
Set 1 : {7,14,21,28,35} Mean = 21, Median = 21
Set 2 : {2,10,21,39,42} Mean = 22.8, Median = 21
Set 3 : {10,13,21,67,82} Mean = 38.6, Median = 21

*5. A sample of 30 distance scores measured in yards has a mean of 7, a variance of 16, and a standard deviation of 4. (a) You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, variance, and standard deviation? (b) You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, variance, and standard deviation?*

Given : no. of observations = 30, mean = 7, variance = 16, S.D = 4

a. To convert all distances from yards to feet, we multiply each score in the sample by 3.

New Mean = Old Mean * Scaling constant = 7*3 = **21**
The variance(σ2) is rescaled by multiplying by the scaling constant squared.
New Variance = Old variance * Square of the Scaling constant = 16*3*3 = **144**
New Standard Deviation = Old S.D * Scaling constant = 4*3 = **12**

b. After multiplying the original scores by 3 and subtracting 4 from each. The new values are as follows:
Multiplying each score by 3, scales the mean by 3 and subtracting 4 from each score subtracts 4 from the mean.
New Mean =(Old Mean * 3)-4 = 7*3 = 21-4 = **17**
var (X + c) = var (X), where X is a variable and c is a constant.
Subtracting/Adding a constant value, c, to a random variable X does not change the variance.
New Variance = Old Variance * 3 = 16*3 = **144**
Subtracting/Adding a constant value will not affect the Standard Deviation.
New Standard Deviation = Old S.D * 3= 4*3 = **12**

*7. For the given test score, which measures of variability (range, standard deviation, variance) would be changed if the 22.1 data point had been erroneously recorded as 21.2?*

*15.2*
*18.8*
*19.3*
*19.7*
*20.2*
*21.8*
*22.1*
*29.4*

**Calculation on correct data:** [15.2, 18.8, 19.3, 19.7, 20.2, 21.8, 22.1, 29.4]

Range : 29.4 - 15.2 = **14.2**

S.D : $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$ = **4.067796**

Variance : square of the S.D = **4.067796** * **4.067796** = **16.54696**

**Calculation on erroneous data:** [15.2, 18.8, 19.3, 19.7, 20.2, 21.8, 21.2, 29.4]

Range : 29.4 - 15.2 = **14.2** ( Remains unchanged as Range is the difference between the limits and the limits were not erroneous)

S.D : **4.039448**

Variance : square of the S.D = **16.3171401**

Therefore, the Range will remain unchanged, but the Standard Deviation and Variance would in case of error.

*9. For the numbers 1, 3, 4, 6, and 12:*

a. *Find the value (v) for which Σ(X-v)2 is minimized.*

b. *Find the value (v) for which Σ|x-v| is minimized.*

a. The mean is the value that minimizes the sum of the squared deviations. Mean of the given numbers = (1+3+4+6+12)/5 = 26/5 = 5.2
**The value (v) for which Σ(X-v)2 is minimized is 5.2**

b. The median is the value that minimizes the sum of absolute deviations. Median of the given numbers = 4
**The value (v) for which Σ|x-v| is minimized is 4**

*11. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the total number of pieces correctly remembered from three chess positions. Compare the performance of each group. Consider spread as well as central tendency.*

| Non-players | Beginners | Tournament players |
|---|---|---|
| 22.1 | 32.5 | 40.1 |
| 22.3 | 37.1 | 45.6 |
| 26.2 | 39.1 | 51.2 |
| 29.6 | 40.5 | 56.4 |
| 31.7 | 45.5 | 58.1 |
| 33.5 | 51.3 | 71.1 |
| 38.9 | 52.6 | 74.9 |
| 39.7 | 55.7 | 75.9 |
| 43.2 | 55.9 | 80.3 |
| 43.2 | 57.7 | 85.3 |

```
import numpy
data = { "non_players" : [22.1,22.3,26.2,29.6,31.7,33.5,38.9,39.7,43.2,43.2],
         "beginners" :[32.5,37.1,39.1,40.5,45.5,51.3,52.6,55.7,55.9,57.7],
         "tournament_players" : [40.1,45.6,51.2,56.4,58.1,71.1,74.9,75.9,80.3,85.3]}
df = pd.DataFrame(data)

#Measures of central tendency are : mode, median, mean
#Measures of spread are : range, quartiles and the interquartile range, variance and standard deviation.
result = {"Measure" : ["mean", "median", "mode", "range", "25%", "50%", "75%", "IQR", "variance","S.D","Coefficient of
          "Non Players" : [statistics.mean(data["non_players"]),statistics.median(data["non_players"]),statistics.mode(d
          "Beginners" : [statistics.mean(data["beginners"]),statistics.median(data["beginners"]),statistics.mode(data["b
          "Tournament Players" : [statistics.mean(data["tournament_players"]),statistics.median(data["tournament_players
```

```
final = pd.DataFrame(result)
final
```

| | Measure | Non Players | Beginners | Tournament Players |
|---|---|---|---|---|
| 0 | mean | 33.040000 | 46.790000 | 63.890000 |
| 1 | median | 32.600000 | 48.400000 | 64.600000 |
| 2 | mode | 43.200000 | 32.500000 | 40.100000 |
| 3 | range | 21.100000 | 25.200000 | 45.200000 |
| 4 | 25% | 27.050000 | 39.450000 | 52.500000 |
| 5 | 50% | 32.600000 | 48.400000 | 64.600000 |
| 6 | 75% | 39.500000 | 54.925000 | 75.650000 |
| 7 | IQR | 12.450000 | 15.475000 | 23.150000 |
| 8 | variance | 64.533778 | 81.552111 | 244.029889 |
| 9 | S.D | 8.033292 | 9.030621 | 15.621456 |
| 10 | Coefficient of var. | 24.313837 | 19.300322 | 24.450549 |

From the numbers calculated, we can see that the mean or measure of central tendency is widely different for the 3 groups of players and hence we can conclude that the average number of pieces remembered correctly increases with the increase in the level of player expertise.
The S.D shows that more datapoints from the Tournament players pool lie close to the mean compared to the Beginners and Non-Players. From the coefficient of variation, we can understand the spread. Higher the coefficient of variation, the greater the level of dispersion around the mean.

*13. The best way to describe a skewed distribution is to report the mean.*

Result: **False**. For positively skewed distribution, the mean is greater than the median and likewise for negatively skewed distribution, the median is greater than the mean. So, a skewed distribution cannot be reported with just the mean as the median would also be required.

*15. Compare the mean, median, trimean in terms of their sensitivity to extreme scores*

The mean is highly sensitive to extreme values. The median is not really affected by extreme scores as outliers only move the median one or two positions to the left or right. The trimean is calculated by adding the 25th percentile to twice the 50th percentile and the 75th percentile and dividing the sum by 4. The trimean is a good measure of central tendency and is almost as resistant to extreme scores like the

median. Therefore, in decreasing order of their sensitivity, **mean > trimean > median**

*17. A set of numbers is transformed by taking the log base 10 of each number. The mean of the transformed data is 1.65. What is the geometric mean of the untransformed data?*

Mean of n numbers = $(x_1 + x_2 + x_3 + \ldots\ldots + x_n)/n$

After taking log base 10 of each number, the mean is = $\log_{10} x_1 + \log_{10} x_2 + \log_{10} x_3 + \ldots\ldots + \log_{10} x_n$

We know that log a + log b = log ab and if $\log_{10} b = x$, then $b = 10^x$

Given,

$\log_{10} (x_1 . x_2 . x_3 . \ldots\ldots x_n)/n = 1.65$

$\log_{10} (x_1 . x_2 . x_3 . \ldots\ldots x_n) = 1.65n$

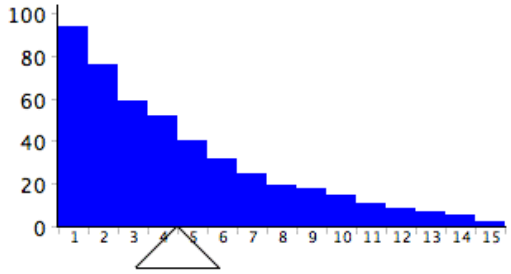
$(x_1 . x_2 . x_3 . \ldots\ldots x_n) = 10^{1.65n}$ ------- (1)

Geometric mean = $(x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{1/n}$

From (1),

GM = $(10^{1.65n})^{1/n} = 10^{1.65}$ = **44.6683592**

*19. The histogram is in balance on the fulcrum. What is the mean, median, and mode of the distribution?*




From the figure, we can notice that the fulcrum is at 4.5. Since this is where the histogram is in balance, the mean = **4.5**.

Mode is the data point with the maximum frequency and since it appears from the graph that data point 1 occurs close to 95 times and hence the mode = **1**.

Since this is a histogram, to compute the median, we need to identify the number of values in each bucket. Approximately,

| **Bucket** | **# Values** | **Cum. sum** |
|---|---|---|
| 1 | 95 | 95 |
| 2 | 75 | 170 |
| 3 | 58 | 228 |
| 4 | 55 | 283 |
| 5 | 40 | 323 |
| 6 | 35 | 358 |
| 7 | 25 | 383 |

| | | |
|---|---|---|
| 8 | 20 | 403 |
| 9 | 19 | 422 |
| 10 | 15 | 437 |
| 11 | 13 | 450 |
| 12 | 12 | 462 |
| 13 | 10 | 472 |
| 14 | 7 | 479 |
| 15 | 5 | 484 |

Since we have approximately 484 values, the median would be in the cumulative sum bucket approx. 248 which lies in bucket 4. Hence the median is **4**.

CHAPTER 4

*1. Describe the relationship between variables A and C. Think of things these variables could represent in real life.*

From the graph it is evident that as A increases, C decreases. We can conclude that the two variables have a negative association.

Some real-life examples for the same would be:

- Atmospheric pressure (C) decreases with the increase in altitude (A)
- Sale of Movie Tickets (C) decreases with increase in number of days from Date of Movie Release (A)

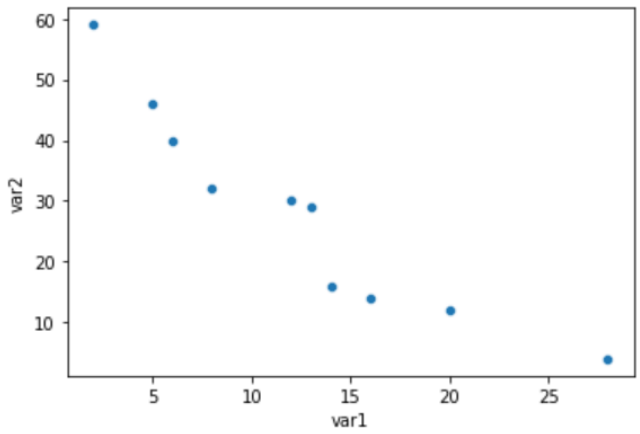
*3. Make up a data set with 10 numbers that has a negative correlation.*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| var1 | 2 | 5 | 6 | 8 | 12 | 13 | 14 | 16 | 20 | 28 |
| var2 | 59 | 46 | 40 | 32 | 30 | 29 | 16 | 14 | 12 | 4 |

```
import pandas as pd
import seaborn as sns

# initialize data of lists.
plot_data = {'var1': [2,5,6,8,12,13,14,16,20,28],
        'var2': [59,46,40,32,30,29,16,14,12,4]}
# Create DataFrame
df = pd.DataFrame(plot_data)
sns.scatterplot(x="var1", y="var2", data = df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f9d16076d8>
```



It can be inferred from the plot that the data set has a negative correlation. Furthermore, the correlation matrix we can affirm that the data set has a **negative correlation**.

```
r = np.corrcoef(plot_data['var1'], plot_data['var2'])
print("Correlation coefficient : \n", r)
```

```
Correlation coefficient :
 [[ 1.        -0.9374443]
 [-0.9374443  1.       ]]
```

*5. Would you expect the correlation between High School GPA and College GPA to be higher when taken from your entire high school class or when taken from only the top 20 students? Why?*

The correlation between High School GPA and College GPA would be higher when taken only for top 20 students as top performing students in school are more likely to perform well in college as well. Since the correlation taken between GPAs for the entire high school class would operate on a wider pool of students with differing characteristics, it would most likely be lower.

*7. For this same class, the relationship between the amount of time spent studying and the amount of time spent socializing per week was also examined. It was determined that the more hours they spent studying, the fewer hours they spent socializing. Is this a positive or negative association?*

This is a **negative association** because when one factor (no. of hours spent studying) increases the other (no. of hours spent socializing) decreases.

*9. Students took two parts of a test, each worth 50 points. Part A has a variance of 25, and Part B has a variance of 36. The correlation between the test scores is 0.8. (a) If the teacher adds the grades of the two parts together to form a final test grade, what would the variance of the final test grades be? (b) What would the variance of Part A - Part B be?*

Given: Var(A) = 25, Var(B) = 36, correlation(A,B) = 0.8

$$\sigma_{X\pm Y}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2\rho\sigma_X\sigma_Y$$

a) Var(A+B) = Var(A) + Var(B) + 2*Correlation(A,B) . SD(A) . SD(B)
Var(A+B) = 25 + 36 + 2*0.8*5*6 = **109**

b) Var(A-B) = Var(A) + Var(B) - 2*Correlation(A,B) . SD(A) . SD(B)
Var(A+B) = 25 + 36 - 2*0.8*5*6 = **13**

*11. True/False: It is possible for variables to have r=0 but still have a strong association.*

**True.** r=0 means that there is no linear relationship between X and Y. However, it is possible for the variables to have a strong non-linear association.

*13. True/False: After polling a certain group of people, researchers found a 0.5 correlation between the number of car accidents per year and the driver's age. This means that older people get in more accidents.*

**False**. Correlation coefficients whose magnitude are between 0.3 and 0.5 indicate variables which have a low correlation. Hence, we cannot confirm that older people get in more accidents from the given information.

*15. True/False: To examine bivariate data graphically, the best choice is two side by side histograms.*

**False.** Though bivariate data can be examined using two side by side histograms, it is not the best choice. The Scatter plot is better suited to reveal the relationship between two variables.

# ANGRY MOODS CASE STUDY

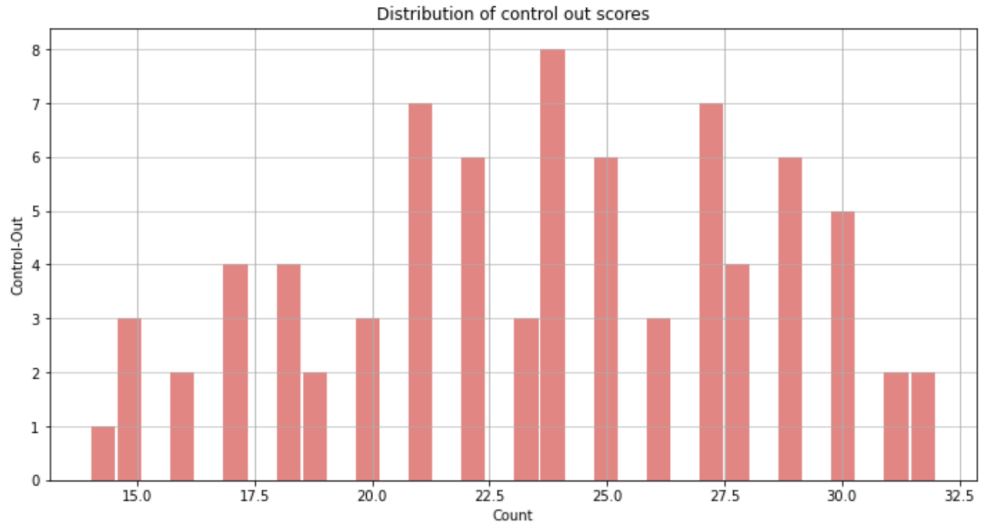
*__AM#10__ Plot a histogram of the distribution of the Control-Out scores.*

```
# import libraries
import pandas as pd
import matplotlib.pyplot as plt

# read by default 1st sheet of an excel file
angry_moods = pd.read_excel('angry_moods.xls')

# Increase size of plot in jupyter
plt.rcParams["figure.figsize"] = (12,6)

angry_moods['Control-Out'].plot.hist(grid=True, bins=angry_moods['Control-Out'].max(), rwidth=0.9,
                    color='lightcoral')
plt.title('Distribution of control out scores')
plt.xlabel('Count')
plt.ylabel('Control-Out')
plt.grid(axis='y', alpha=0.75)
```




*AM#11 What is the overall mean Control-Out score? What is the mean Control-Out score for the athletes? What is the mean Control-Out score for the non-athletes?*

```
#Sports : 1 = athletes, 2 = non-athletes

# Importing the statistics module
import statistics

print("Overall mean of the Control-Out score : ",
      statistics.mean(angry_moods['Control-Out']))
```

```
Overall mean of the Control-Out score :  23.692307692307693
```

```
print("Overall mean of the Control-Out score of athletes: ",
      statistics.mean(list(angry_moods.loc[angry_moods['Sports'] == 1]['Control-Out'])))
```

```
Overall mean of the Control-Out score of athletes:  24.68
```

```
print("Overall mean of the Control-Out score of non-athletes: ",
      statistics.mean(list(angry_moods.loc[angry_moods['Sports'] == 2]['Control-Out'])))
```

```
Overall mean of the Control-Out score of non-athletes:  23.22641509433962
```

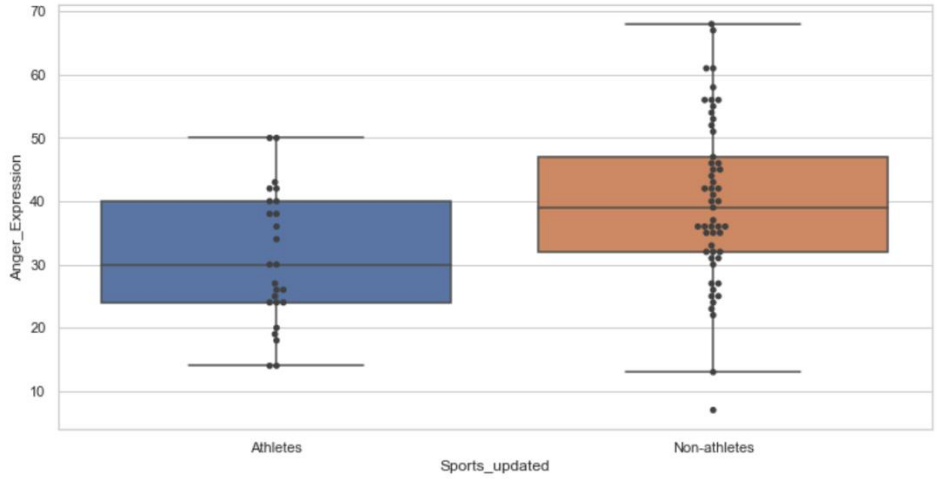
*AM#17 Plot parallel box plots of the Anger Expression Index by sports participation. Does it look like there are any outliers? Which group reported expressing more anger?*

```
angry_moods["Sports_updated"] = np.where(angry_moods["Sports"].isin([1]), 'Athletes', 'Non-athletes')


sns.set(style="whitegrid")

ax = sns.boxplot(x="Sports_updated", y="Anger_Expression", data=angry_moods, showfliers = False)
ax = sns.swarmplot(x="Sports_updated", y="Anger_Expression", data=angry_moods, color=".25")

plt.show()
```




From the box plot it is evident that there are outliers for non-athletes as there are data points lying outside the whiskers (1.5 IQR). Also it can be observed that non-athletes have a greater mean Anger_Expression and can be seen to express more anger.

*AM#18 Plot parallel box plots of the Anger Expression Index by gender.*

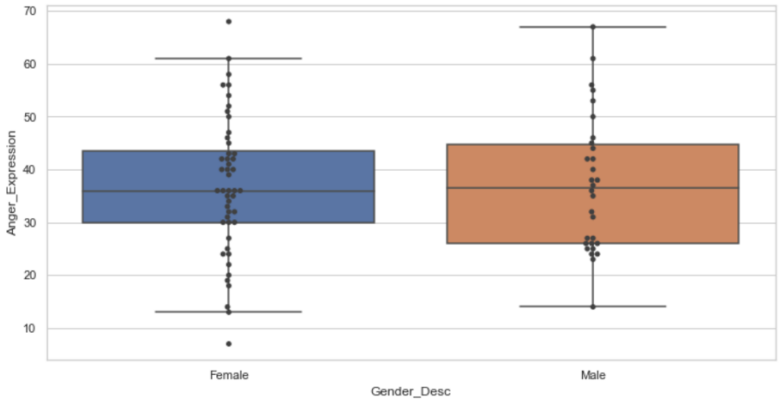
```
#Gender : 1 = males, 2 = females

angry_moods["Gender_Desc"] = np.where(angry_moods["Gender"].isin([1]), 'Male', 'Female')


sns.set(style="whitegrid")

ax = sns.boxplot(x="Gender_Desc", y="Anger_Expression", data=angry_moods, showfliers = False)
ax = sns.swarmplot(x="Gender_Desc", y="Anger_Expression", data=angry_moods, color=".25")

plt.show()
```




*AM#20 What is the correlation between the Control-In and Control-Out scores? Is this correlation statistically significant at the 0.01 level?*

*We first define the hypothesis*

- **H0**: There is no significant mean difference between the Control-In scores and Control-Out scores ($\mu 1 = \mu 2$)
- **H1**: There is significant mean difference between the Control-In scores and Control-Out scores ($\mu 1 != \mu 2$)

```
import scipy as sp
import pandas as pd
from scipy import stats

Control_in = angry_moods["Control-In"]
Control_Out = angry_moods["Control-Out"]

Data_core = {"Control-in": Control_in, "Control-Out": Control_Out}
DF = pd.DataFrame(data=Data_core)

Pearson = DF.corr()
print(Pearson)
```

```
             Control-in  Control-Out
Control-in     1.000000     0.719283
Control-Out    0.719283     1.000000
```

The 2 variables, Control-In and Control-Out have a strong positive correlation. The mean of Control-In (21.96) is very close to the mean of Control-Out (23.69). We perform a two-sample t test :

t = 9.024689, df = 76, p-value = 0

The alpha value is given as 0.01. A pvalue is the probability that the null hypothesis is true.

The *p*value (0) is smaller than the significance level ($\alpha$ =0.01), we reject the null hypothesis. We conclude that the correlation is **statistically significant** for alpha value 0.01. Or in other words “we conclude that there is a significant linear correlation between Control-In and Control-Out scores” at the 0.01 level.

*AM#21 Would you expect the correlation between the Anger-Out and Control-Out scores to be positive or negative? Compute this correlation.*

```
import scipy as sp
import pandas as pd
from scipy import stats

Anger_Out = angry_moods["Anger-Out"]
Control_Out = angry_moods["Control-Out"]

Data_core = {"Anger-Out": Anger_Out, "Control-Out": Control_Out}
DF = pd.DataFrame(data=Data_core)

Pearson = DF.corr()
print(Pearson)
```

```
             Anger-Out  Control-Out
Anger-Out     1.000000    -0.582683
Control-Out  -0.582683     1.000000
```

The correlation between Anger-Out and Control-Out scores is negative. However, the correlation value of -0.5 is a weak negative correlation and hence we cannot deduce any concrete relationship between the 2 variables.