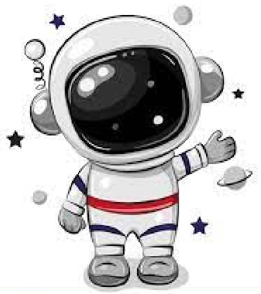
Spaceflights Data Analysis

Through the data of Astronauts

—

Team Millenium Falcons

Ninad Chaphekar	200010016
Jeet Mehta	200010036
Shambhu Kumar	200010072
Ripudaman Singh	200010063
Priyanshu Rajak	200010058
Harshit Singh	200010029



Overview

We all were interested in space missions and hence agreed that we would analyse spaceflights or space missions for our project but darker clouds lay ahead.

Even after long dreamy nights we were not able to find proper and complete data about every space mission from the start and were in turmoil until we found a .csv file of astronaut's missions having data about their individual missions.

What could be a better way to analyse space missions and other aspects of it over the time than analysing astronaut's data over the missions.

With the random variables below, we are analysing various relationships between the aspects of the astronauts in a particular mission.

Acknowledgement

- The dataset being used is obtained from kaggle.com, which is one of the databases for a wide range of topics.
- It is publicly accessible from aerospace.csis.org, which reaffirms its authenticity.

Random Variables and other details:

The random variables we are using are:

- Age of Astronaut during that mission
- Experience of astronauts
- Gender ratio of Astronaut [Male to total astronauts]

The Statistics used:

- $\overline{X}\ (Sample\ Mean) = \frac{x_1 + x_2 + ... + x_n}{n} (< \overline{X} > \; = \mu\ [True\ Mean])$
- $\overline{S}\ (Sample\ Variance) = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{(n-1)} (< S^2 > \; = \sigma^2\ [True\ Variance])$

The dataset is divided into 2 parts: The old era (1961 to 1990) and the new era (1991 to 2019) and all random variables will be sampled for both.

Reading, cleaning and manipulating data as per requirements

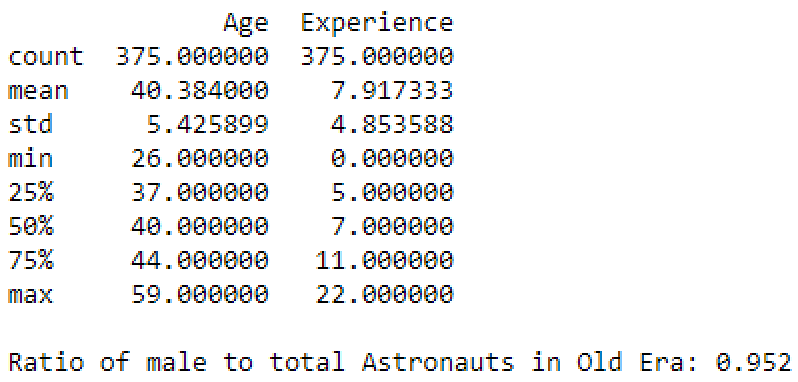
From the dataset we took,

1. Gender of the astronaut
2. Year of mission
3. Age of astronaut during the mission [difference between year of mission and year of birth]
4. Experience [difference between year of mission and year of selection]

True population parameters of Old/New/All Astronauts

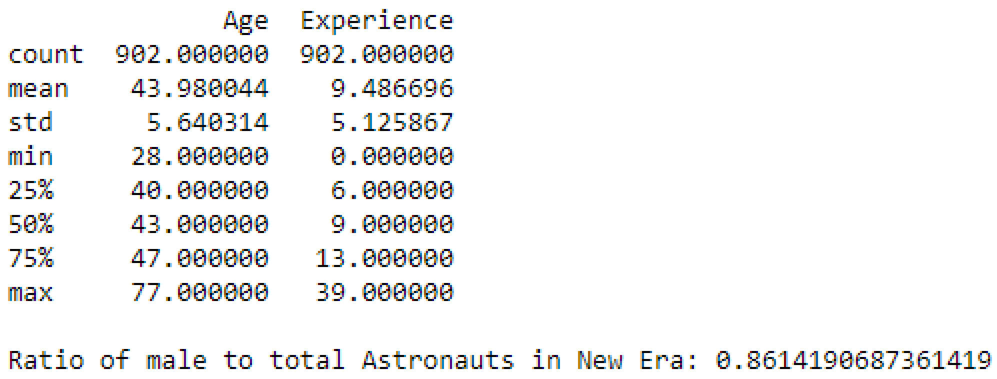
The 'pop_par' function returns the value of the population parameters μ[True mean] and σ[True Standard Deviation] and other relevant statistics. Since the dataset has been sorted for the purpose of analysis, it returns the required parameters and more for the new era, the old era and for the complete dataset.

Old Astronauts(Missions before 1990):



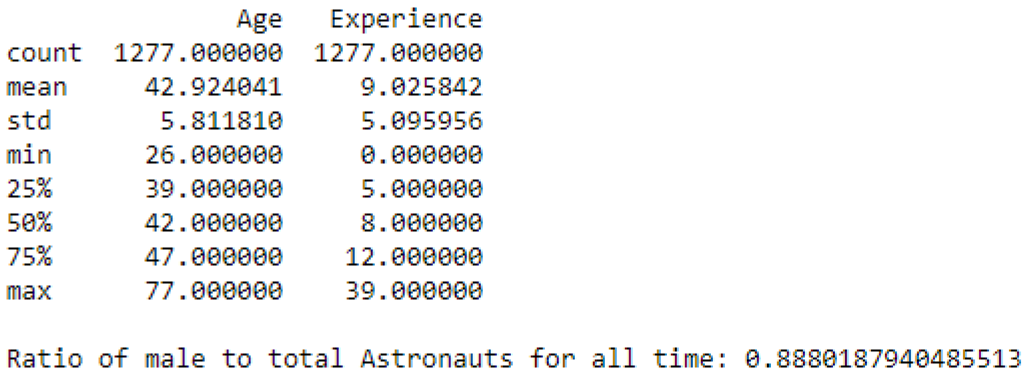
```
              Age  Experience
count  375.000000  375.000000
mean    40.384000    7.917333
std      5.425899    4.853588
min     26.000000    0.000000
25%     37.000000    5.000000
50%     40.000000    7.000000
75%     44.000000   11.000000
max     59.000000   22.000000

Ratio of male to total Astronauts in Old Era: 0.952
```

New Astronauts(Missions after 1990):



```
               Age  Experience
count   902.000000  902.000000
mean     43.980044    9.486696
std       5.640314    5.125867
min      28.000000    0.000000
25%      40.000000    6.000000
50%      43.000000    9.000000
75%      47.000000   13.000000
max      77.000000   39.000000

Ratio of male to total Astronauts in New Era: 0.8614190687361419
```

All the Astronauts:



```
                Age   Experience
count   1277.000000  1277.000000
mean      42.924041     9.025842
std        5.811810     5.095956
min       26.000000     0.000000
25%       39.000000     5.000000
50%       42.000000     8.000000
75%       47.000000    12.000000
max       77.000000    39.000000

Ratio of male to total Astronauts for all time: 0.8880187940485513
```

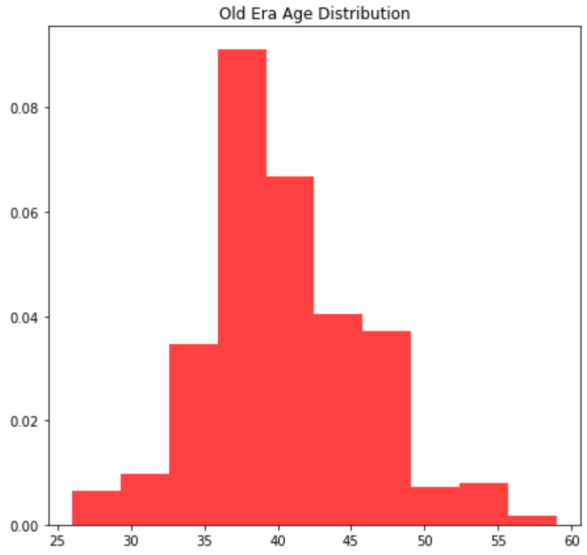
We can note that the average age of an astronaut in the new era is greater than the old era which will further be reinforced in checking one of the hypotheses later on.

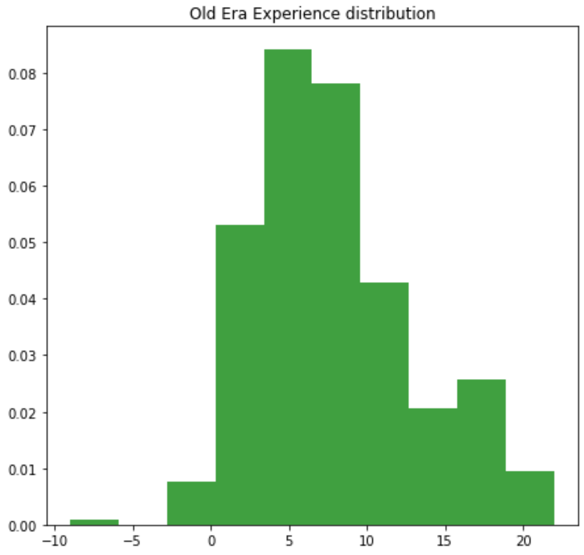
Further the ratio in both the eras also differs by a substantial amount and it is also verified in one of the hypotheses later on.

Marginal Distributions of Old/New/All Astronauts

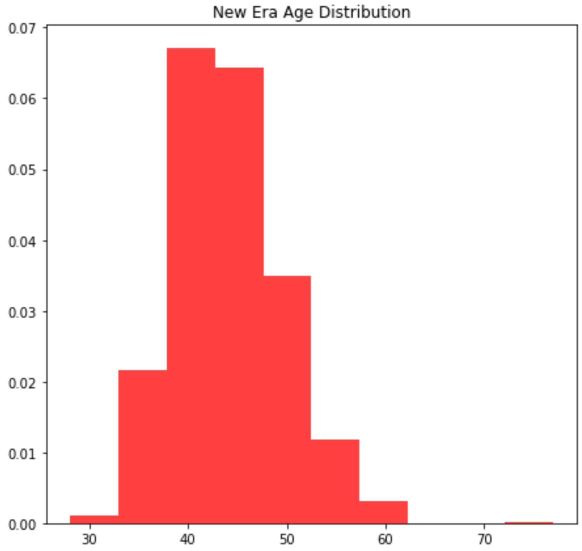
The marginal distributions for the age and experience are shown below for all the 3 timeframes. The histograms can be approximated to a normal distribution for both the random variables in all three cases as not only does the shape resemble a normal distribution, but during sampling we can also roughly approximate the sample variance distribution by a χ^2 distribution with appropriate degrees of freedom.

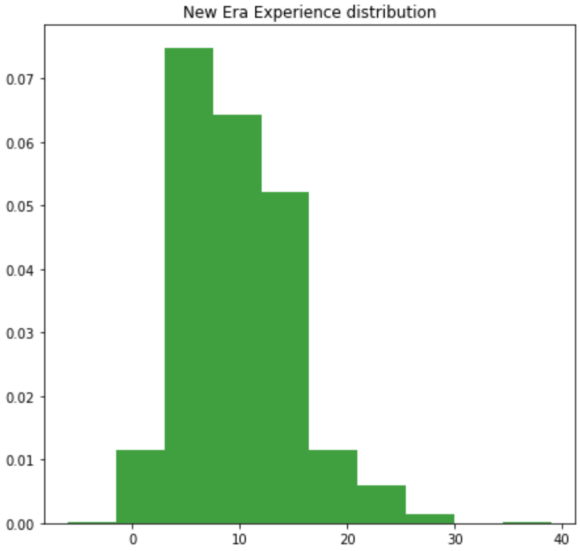
Old Era Astronauts(Missions before 1990):



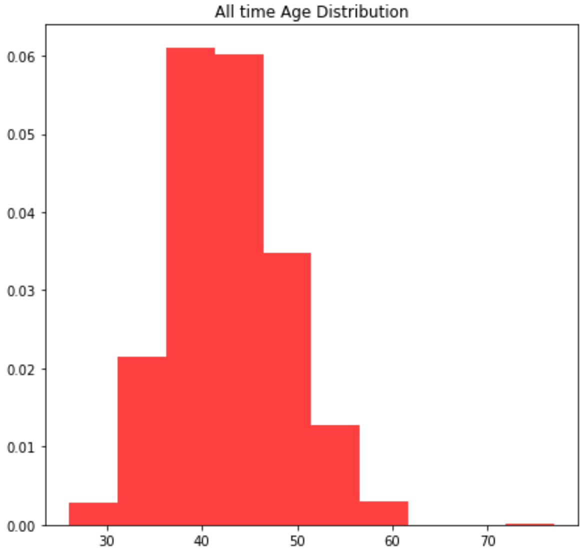


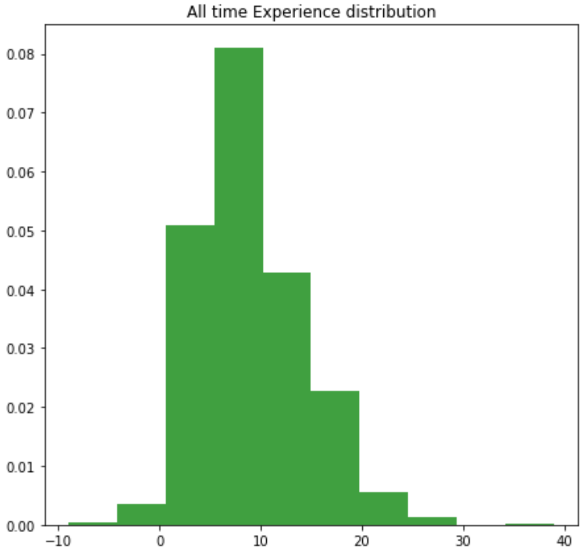
New Era Astronauts(Missions after 1990):





Astronauts of Both Eras:

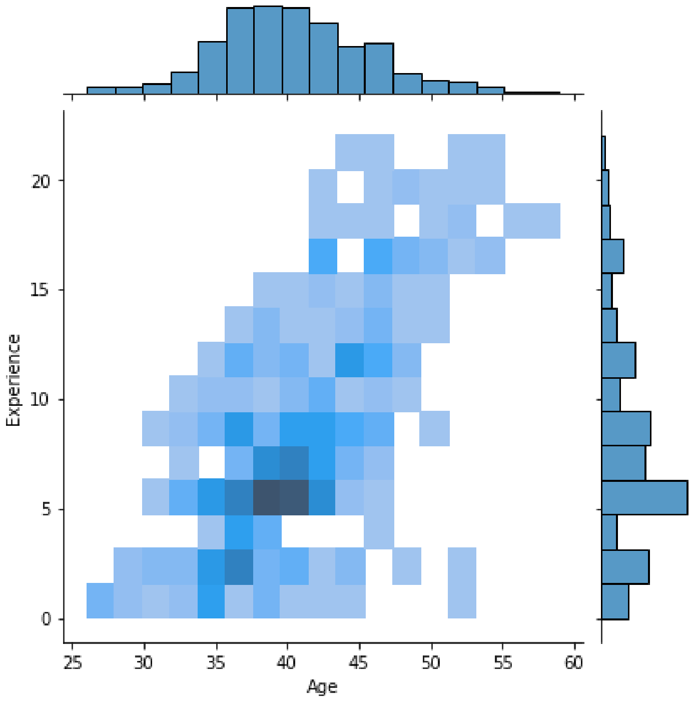




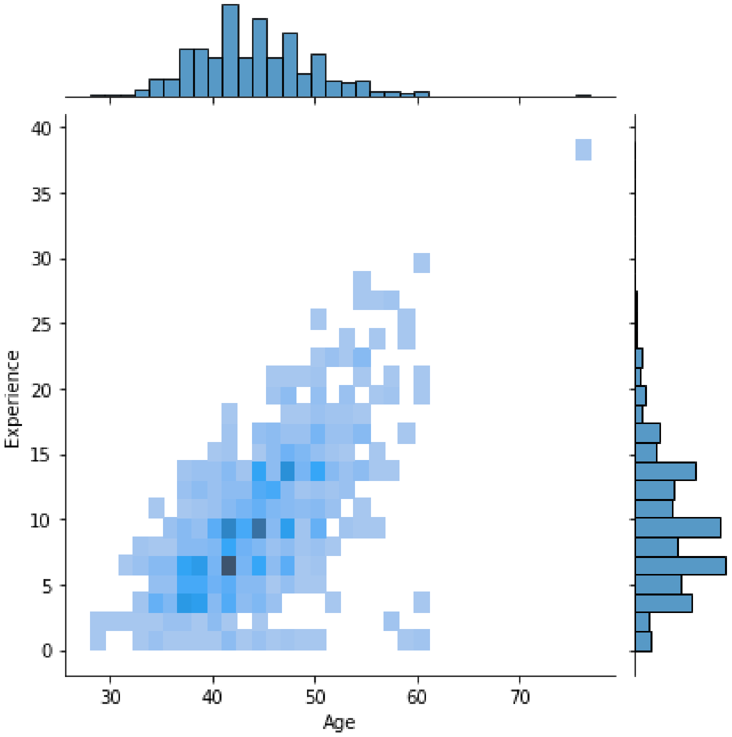
Joint Distribution

The joint distribution is plotted in the form of a 3-D histogram where the 3rd dimension is indicated by the hue of the blue color, with the darker shades indicating greater height meaning higher probability. In the plot we can verify that this occurs near the respective means of the 2 random variables for all 3 timeframes

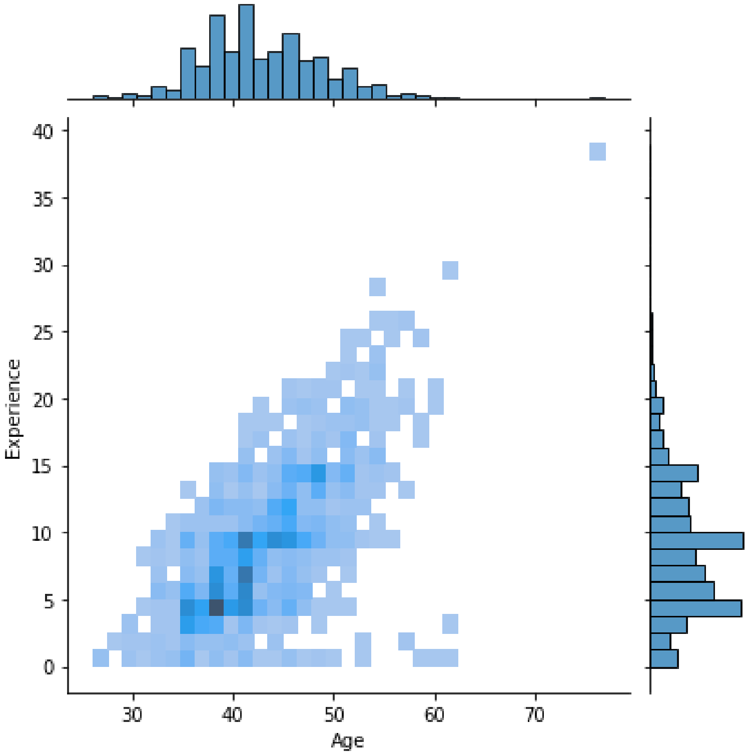
Joint Distribution for Old Era



Joint distribution for New Era



Joint Distribution for all time



Sampling Distribution

The Sampling distribution:

The ‘sampling_dist_est’ function returns the sampling distribution, point and interval estimates for specific sample size, number of samples and a confidence level. As seen with the interact function the sample means for both the random variables approximate to gaussian distribution with mean equal to the mean of the respective timeframe and standard deviation equal = $\frac{\sigma}{\sqrt{n}}$, where n represents the sample size, due to the central limit theorem.

The confidence intervals for the means are calculated by,

$$[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

for ((1 -α)*100)% confidence that the mean will lie in the calculated interval.

The ratio of males in a population is also sampled and for the respective timeframes, the peak occurs near the true mean.

For,

Sample Size = 25

Number of Samples = 100

Confidence level = 97

Old Astronauts(Missions before 1990):

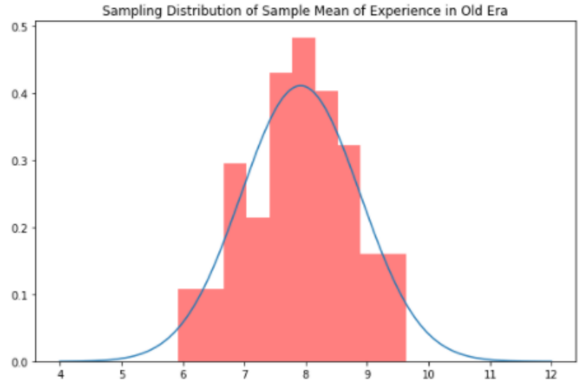
```
Point estimate for the sample age of old era is:
        Sample mean = 39.96
        Sample standard deviation = 5.480875842417889


Interval Estimate for true mean of age in old era for confidence level of 97.0 % is:
        (39.41164903483141 40.50835096516859 )
```

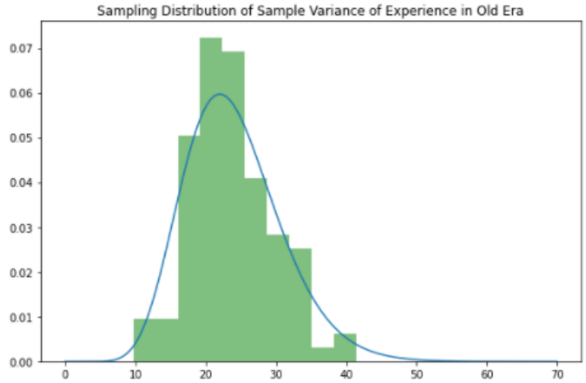
Point estimate for the sample Experience of old era is:

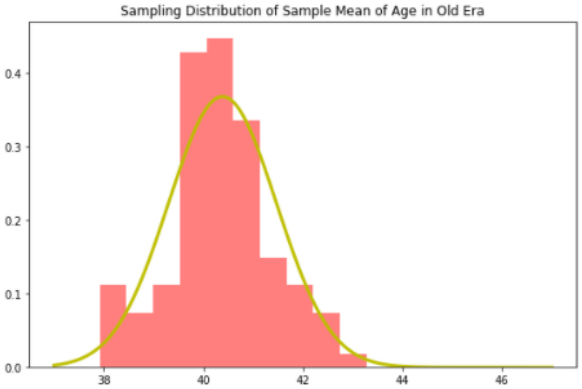
Sample mean = 6.92

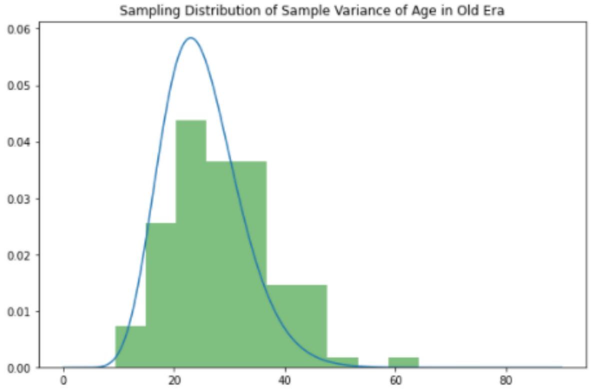
Sample standard deviation = 4.508510470950097

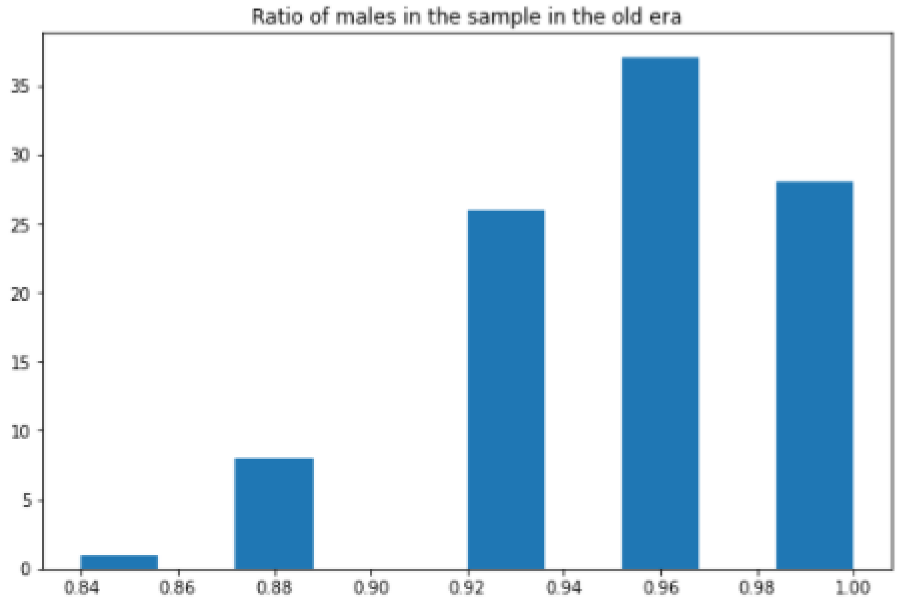
Interval Estimate for true mean of Experience in old era for confidence level of 97.0 % is:
(6.429487813528704 7.410512186471296)









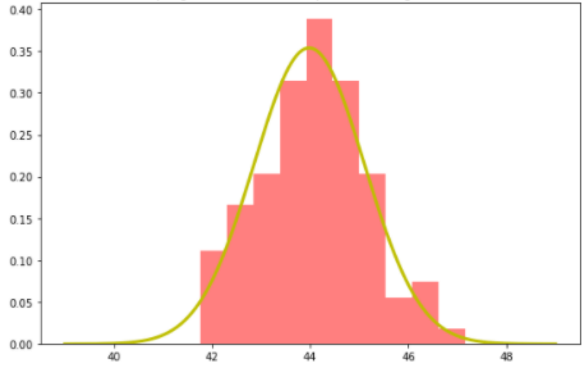


New Astronauts(Missions after 1990):

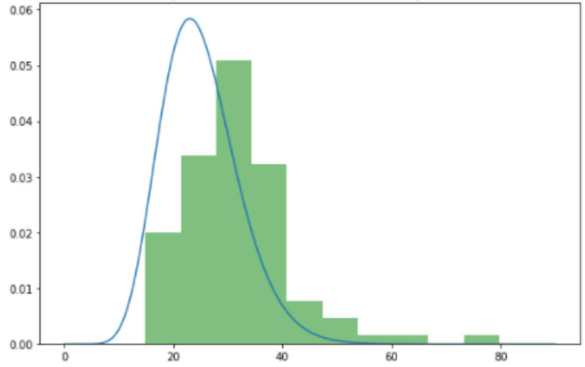
```
Point estimate for the sample age of new era is:
        Sample mean = 43.88
        Sample standard deviation = 5.246586191674227

Interval Estimate for true mean of age in new era for confidence level of 97.0 % is:
        ( 43.30953486108303 44.450465138916975 )
```

```
Point estimate for the sample Experience of new era is:
        Sample mean = 9.64
        Sample standard deviation = 5.536846274429756

Interval Estimate for true mean of Experience in new era for confidence level of 97.0 % is:
         ( 9.121566323800884 10.158433676199117 )
```

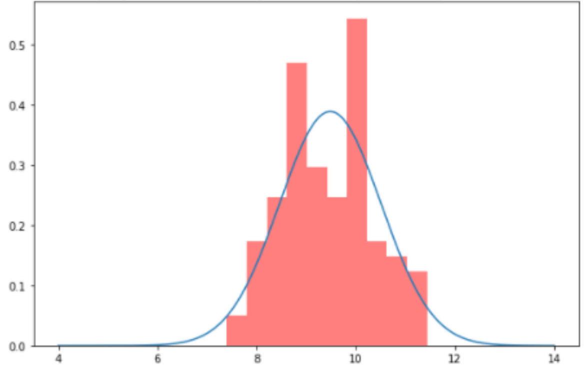
Sampling Distribution of Sample Mean of Age in New Era



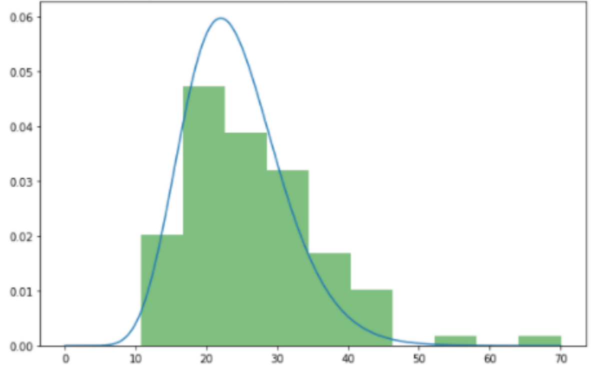
Sampling Distribution of Sample Variance of Age in New Era



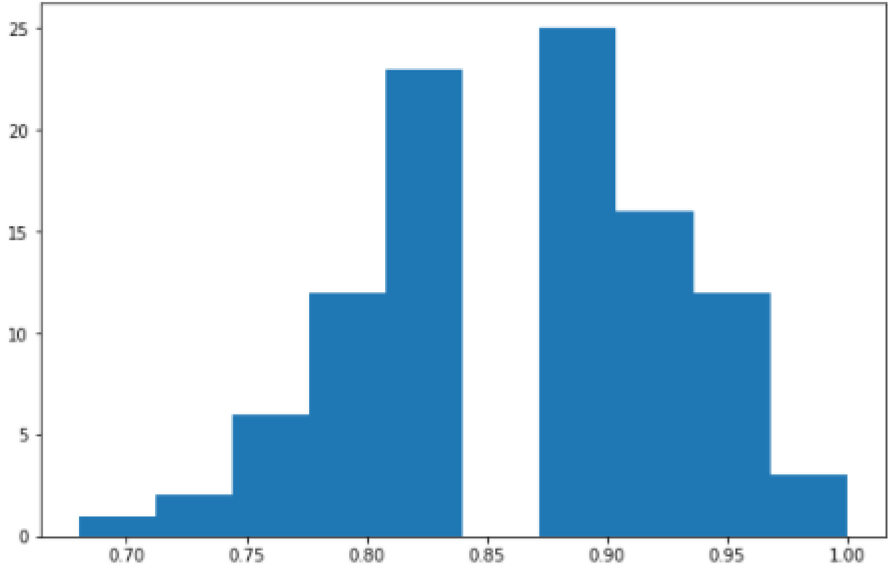
Sampling Distribution of Sample mean of Experience in space in New Era



Sampling Distribution of Sample Variance of Experience in New Era



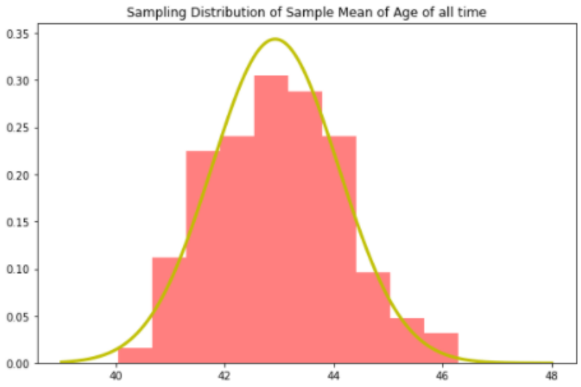
Ratio of males in the sample in the new era

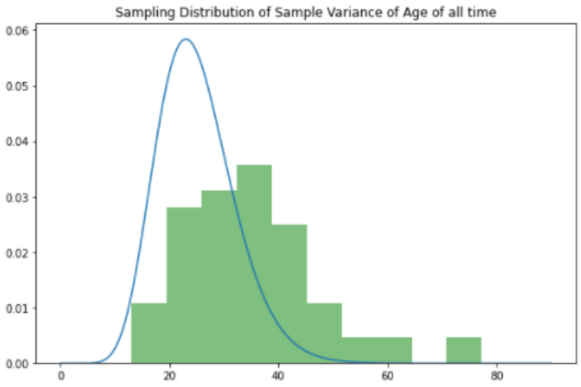


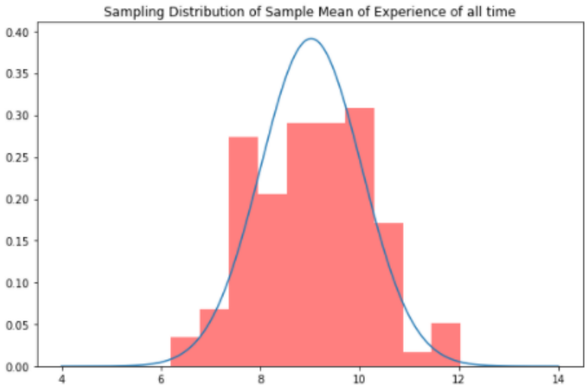
All the Astronauts:

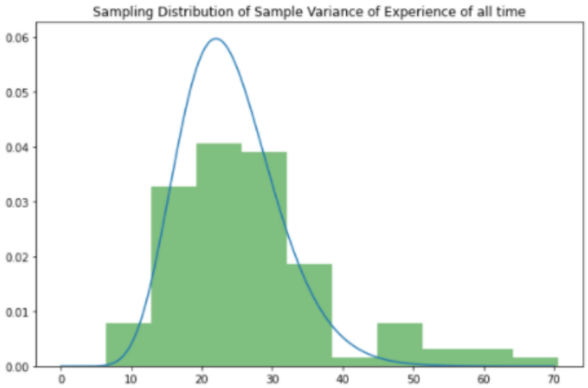
```
Point estimate for the sample age of all time is:
        Sample mean = 45.16
        Sample standard deviation = 8.644265922178317


Interval Estimate for true mean of age in all time for confidence level of 97.0 % is:
        ( 44.57209380795136 45.74790619204863 )
```

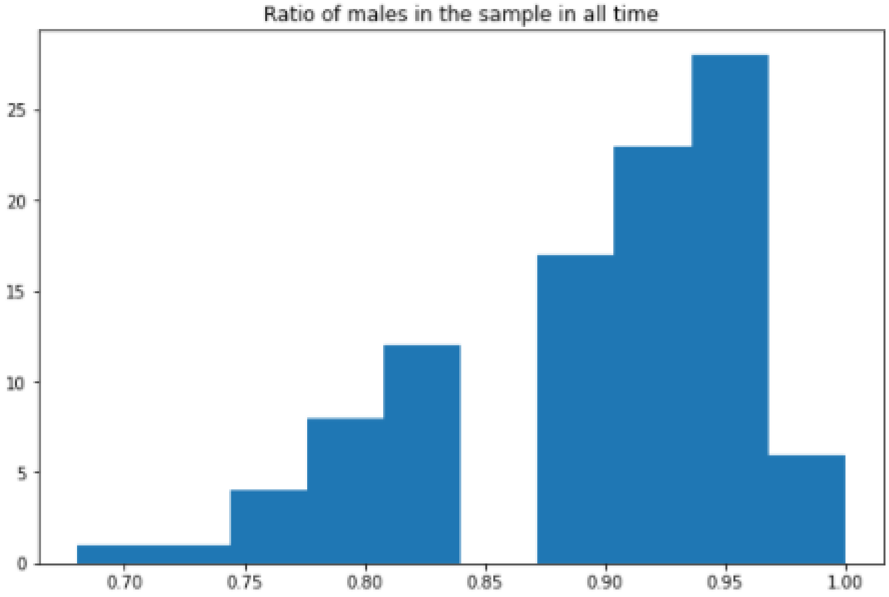
```
Point estimate for the sample Experience of all time is:
        Sample mean = 11.8
        Sample standard deviation = 7.621242243449117


Interval Estimate for true mean of Experience in all time for confidence level of 97.0 % is:
        ( 11.284507609063233 12.315492390936768 )
```











HYPOTHESIS TESTING:

After looking at all the various aspects of the data and seeing various plots, one might have some questions in their mind, like

- What is the probability at a particular mission, the astronaut is male?
- Is the average age in both the old and the new eras the same? If yes, why? And if not, why?
- Has the number of female astronauts really increased as we move from the old to the new era?

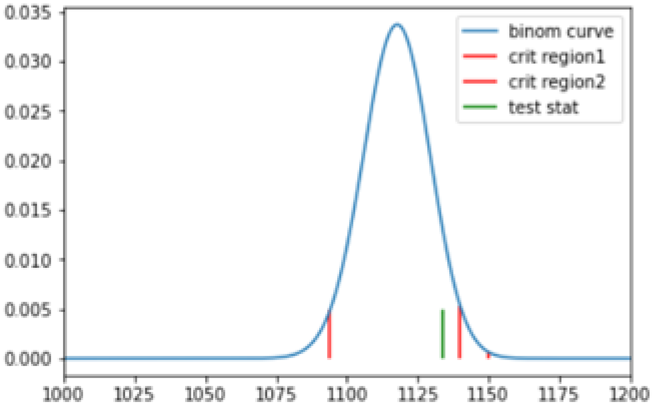
Thus, we perform hypothesis testing to get a rough idea of the answers to these questions.

Hypothesis 1 :

1. We need to estimate the probability that at a particular mission, the astronaut is male.
2. Our initial guess was, why not take $p_0 = 0.5$? It seems natural, but is horrendously wrong given the data has dominantly male astronauts.
3. By looking at the actual fraction, we have an estimate of $p_0 = \frac{1134}{1277} = 0.888$ so, we take a guess that $p = p_0 = 0.875$
4. $H_0 : p = p_0$

   $H_1 : p \neq p_0$

   Test statistic, total number of male astronauts = 1134 out of 1277
5. Then we perform the standard procedure, taking the level of significance $\alpha = 0.05$



6. Looking at this, we see that the test statistic corresponding to the total number of male astronauts lies outside the critical region, hence we fail to reject the null hypothesis.
7. From this, we can say with 95% confidence, that the $p_0$is 0.875 and is maybe higher.

Hypothesis 2:

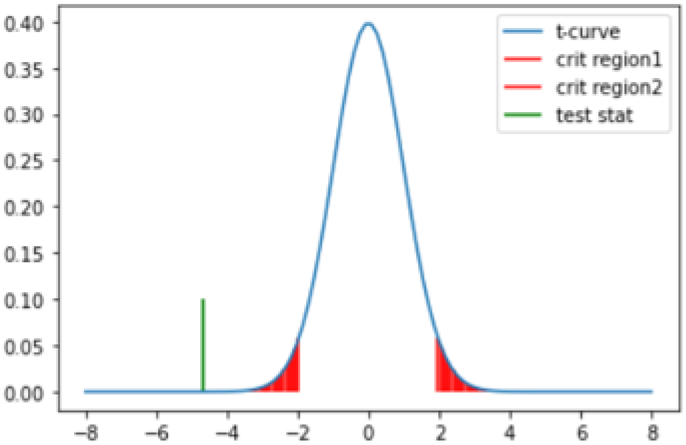
1. We need to check whether the true mean age of the astronauts in both the old and the new eras is equal or not.
2. Note that, since our original distribution of age is close to normal, we are assuming that the distributions in the old and the new era are also close to normal, and add up to give the original distribution.
3. Since we don’t know the true variances, we will work with the spool standard deviation S_p.

4. $H_0 : \mu_0 \ = \ \mu_n$

   $H_1 : \mu_0 \ \neq \ \mu_n$

   $[\mu_0 = True\ Mean\ of\ Old\ Era,\ \mu_n = True\ Mean\ of\ New\ Era]$

5. For the test statistic, we take a sample of size 100 and calculate their mean and standard deviation and use them.
6. Then we perform the standard procedure, taking the level of significance $\alpha \ = \ 0.01$



7. Looking at this, we see that the test statistic lies inside the critical region, hence the null hypothesis is rejected.
8. From this, we can say with 99% confidence that both the true means are not equal, and there has been a change.
9. A foreseeable conclusion: The mean age of the astronauts has changed, most probably because of the need for more experience needed for the challenging and ambitious age in the new era.

Hypothesis 3 :

1. We need to check whether the probability of the astronaut being a male in a particular mission has changed or not, in fact seeing whether it has decreased.
2. Here, we are taking a sample of 100 sets of 20 astronauts, to look at the number of male astronauts out of 100
3. Since we don't know the true variance, we are working with the spool standard deviation S_p.
4. We can make the assumption that the random variables are normally distributed, since they are binomial with large n but unknown mean and standard deviation.
5. $H_0 : \mu_0 \ = \ \mu_n$

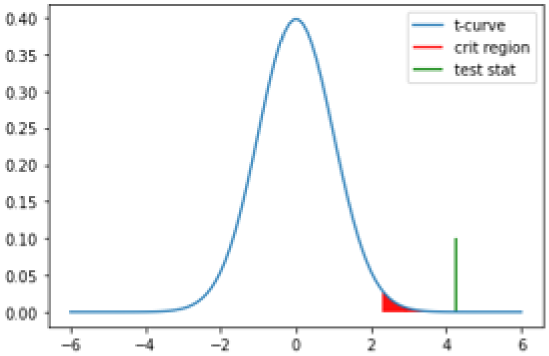
$H_1 : \mu_0 > \mu_n$

$[\mu_0 = True\ Mean\ of\ Old\ Era, \mu_n = True\ Mean\ of\ New\ Era]$

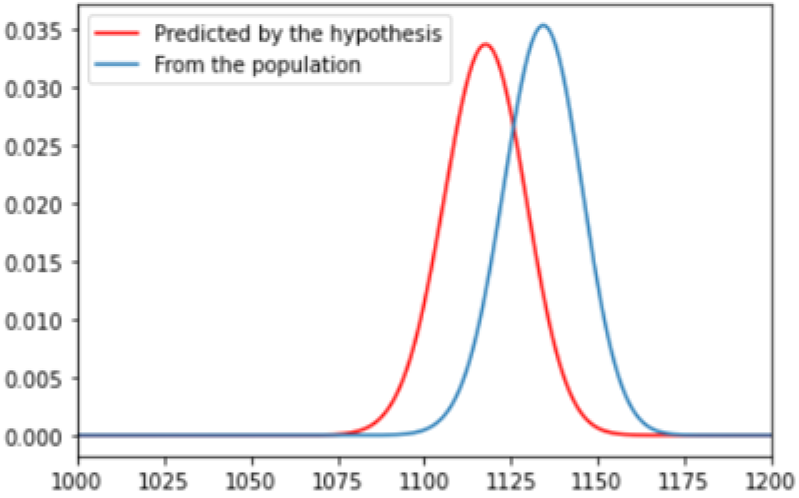
(A one sided test)

6. For the test statistic, we calculate the mean and standard deviation of both the samples and use them.
7. Then we perform the standard procedure, taking the level of significance

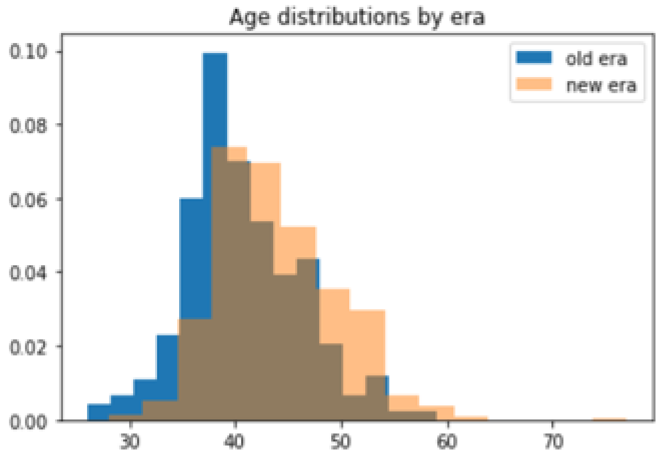
   $\alpha = 0.01$



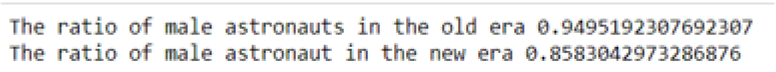
8. Looking at this, we see that the test statistic is inside the critical region, and thus we reject the null hypothesis.
9. A foreseeable conclusion: The likelihood of finding a female astronaut on a particular mission has changed, in fact it has increased as we go from the old to the new era.



From our Hypothesis 1



From this histogram, we can clearly see that the mean age is bigger for the new era, as we expected from our Hypothesis 2.

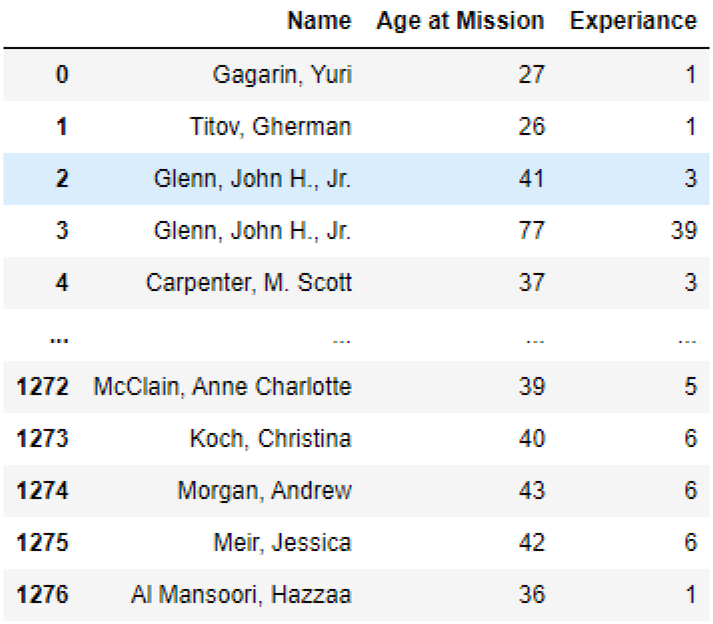


```
The ratio of male astronauts in the old era 0.9495192307692307
The ratio of male astronaut in the new era 0.8583042973286876
```

From the population for our Hypothesis 3.

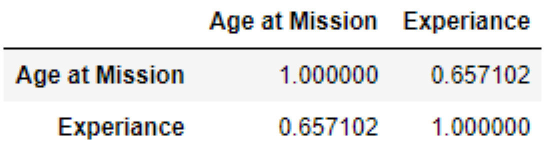
Relationship between Age and Experience of Astronauts

We can guess that the age and experience of an astronaut would be related. But can we say that they are linearly related?



	Name	Age at Mission	Experiance
0	Gagarin, Yuri	27	1
1	Titov, Gherman	26	1
2	Glenn, John H., Jr.	41	3
3	Glenn, John H., Jr.	77	39
4	Carpenter, M. Scott	37	3
...	...	...	...
1272	McClain, Anne Charlotte	39	5
1273	Koch, Christina	40	6
1274	Morgan, Andrew	43	6
1275	Meir, Jessica	42	6
1276	Al Mansoori, Hazzaa	36	1

The correlation gives us an idea if there is any linear relationship between age and experience.



	Age at Mission	Experiance
Age at Mission	1.000000	0.657102
Experiance	0.657102	1.000000

As the correlation between ‘age at mission’ and ‘experience’ is near to 0.66, we can say there is some positive correlation between age and experience.

Plotting the Regression Line along the Scatter Plot

For the regression line we need the slope and intercept and also a function which would tell us how close the data is to the linear regression we found, hence, we define functions to calculate the point estimates of slope and intercept and also a function to calculate the coefficient of determination to know how close the data is to the linear regression.

Equations for point estimates

$$S_{yx} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$b_1 = \frac{S_{yx}}{S_{xx}}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (\widehat{y_i} - \bar{y})^2$$

Coefficient of Determination (R^2),

$$R^2 = \frac{SSR}{SST}$$

```
Point estimate of slope for the linear regression is = 0.5761651673754774

Point estimate of y- intercept for the linear regression  is = -15.705495289365244

Coefficient of Determination is = 0.4317830090180474
```

From here we can see that the value of the coefficient of determination is very low which already indicates a weak linear relationship between the variables. For a better view of the linear regression, we calculate the interval estimates.

Interval Estimate

For interval estimate, we used a function to get point estimates of random samples which was further utilised to get the interval estimate of slope and intercept of the regression line according to the confidence level. For this, we took a random sample of size '100' from the original data. Then we obtained a point estimate of both slope and intercepts. We repeated this experiment '1000' times and thus obtained a set of point estimates for both slope and intercept.

We stored the values of point estimates of slope and intercept of all experiments in two separate arrays. To find the interval estimate of slope, we calculated the 95% confidence level of the values of slope obtained from the experiments. In a similar way, we found the interval estimate of the intercept.

For ,

Sample Size = 100

Number of Samples = 1000

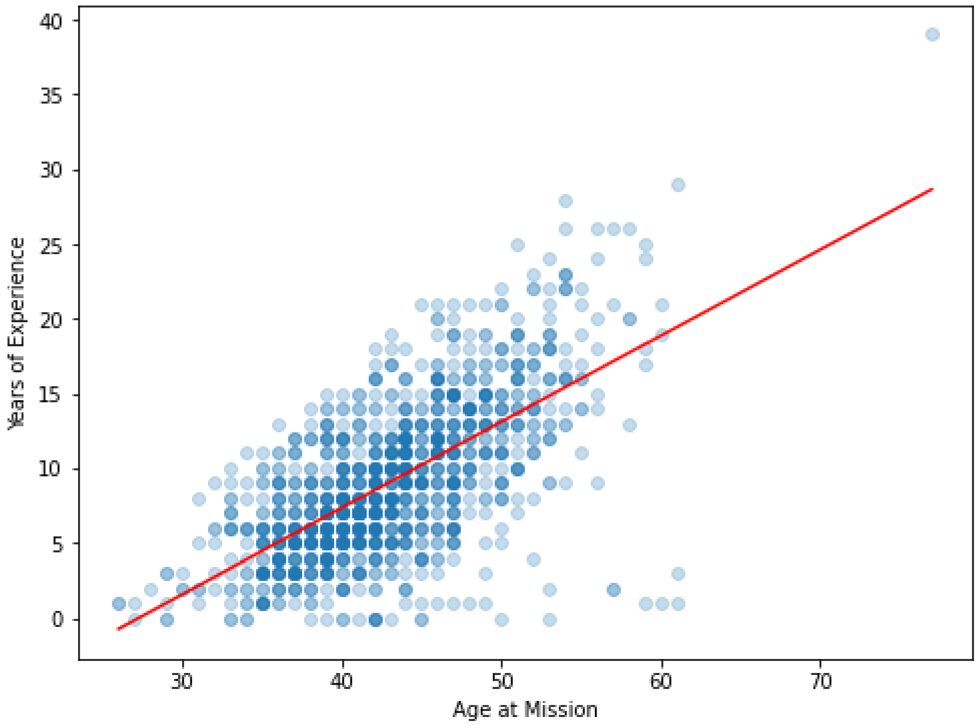
Confidence level = 95

```
Interval estimates of slope for the linear regression with confidence level  95.0  is =

 [0.39079523 0.74061009]


Interval estimates of y- intercept for the linear regression with confidence level  95.0  is =
[-22.53023298  -8.12914145]
```

As the interval estimate is spread over a large range and also the Coefficient of Determination calculated above is low, we can say that the age and experience of astronauts do not have a strong linear relationship.

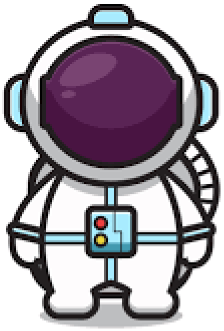
Plot

After knowing the slope and intercept from the point estimates we can plot the linear regression over the scatter plot as follows and graphically see how the data points deviate from the calculated linear regression.



Conclusions

- The true distribution of the age can be approximated to a normal distribution.
- The fraction of male astronauts is much more than female astronauts.
- The probability that the astronaut on a particular mission is male, irrespective of the timeframe, is 0.875 with 95% confidence.
- The true mean of age in the new and old era is quite different.
- The probability that the astronaut on a particular mission is male has decreased as we move from the old to new era. This implies that the number of female astronauts have risen over the years.
- Contrary to common belief, the age and experience of astronauts do not hold a strong linear relationship and we can’t surely say that an older astronaut is also more experienced.



Thank You!

Presentation Video Link