

## Assignment 3-Solutions

### Question 1. - Joint Probability Mass Function

Consider the function  $p_{XY}(\cdot, \cdot)$  :

$x$	$y$	$p_{XY}(x, y)$
1.0	1.0	1/8
1.5	2.0	1/4
1.5	3.0	1/8
2.5	4.0	1/4
3.0	4.0	1/4

Determine the following:

(a) Show that  $p_{X,Y}$  is a valid probability mass function.

If  $\sum_{x,y} p_{XY} = 1$  then it is a valid probability mass function, therefore the calculation

$$\begin{aligned}\sum_{x,y} p_{XY} &= p_{XY}(1.0, 1.0) + p_{XY}(1.5, 2.0) + p_{XY}(1.5, 3.0) + p_{XY}(2.5, 4.0) + p_{XY}(3.0, 4.0) \\ &= \frac{1}{8} + \frac{1}{4} + \frac{1}{8} + \frac{1}{4} + \frac{1}{4} \\ &= 1\end{aligned}$$

So  $p_{XY}$  is a valid probability mass function.

(b)  $P(X < 2.5, Y < 3)$ .

The cases where  $X < 2.5$  are  $p_{XY}(1.0, 1.0)$ ,  $p_{X,Y}(1.5, 2.0)$  and  $p_{X,Y}(1.5, 3.0)$ . The first two of these cases also satisfy the condition  $Y < 3$ , so they should be added together to get the probability required.

$$P(X < 2.5, Y < 3) = p_{XY}(1.0, 1.0) + p_{X,Y}(1.5, 2.0) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8} = 0.375$$

(c)  $P(X < 2.5)$ .

Here all three cases when  $X < 2.5$  are valid as there is no condition on  $Y$

$$P(X < 2.5) = p_{XY}(1.0, 1.0) + p_{X,Y}(1.5, 2.0) + p_{XY}(1.5, 3.0) = \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2} = 0.5$$

(d)  $P(Y < 3)$ .

The cases where  $Y < 3$  are the same as part (b),  $p_{XY}(1.0, 1.0)$  and  $p_{X,Y}(1.5, 2.0)$ .

$$P(Y < 3) = p_{XY}(1.0, 1.0) + p_{X,Y}(1.5, 2.0) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8} = 0.375$$

(e)  $P(X > 1.8, Y > 4.7)$ .

Looking at the probability mass function there are no cases where  $Y > 4.7$  is satisfied so  $P(X > 1.8, Y > 4.7) = 0$

(f)  $E(X)$ ,  $E(Y)$ ,  $V(X)$ ,  $V(Y)$ .

To work out the expected value of  $X$  calculate  $\sum x p_{XY}(x)$  as follows

$$\begin{aligned} E(X) &= \sum_x x p_{XY}(x, \cdot) \\ &= 1.0 \times \frac{1}{8} + 1.5 \times \frac{1}{4} + 1.5 \times \frac{1}{8} + 2.5 \times \frac{1}{4} + 3.0 \times \frac{1}{4} \\ &= 2.0625 \end{aligned}$$

**Note:** 1.5 is included twice as the probability mass function includes the number twice. Similarly the same can be done for the expected value of  $Y$

$$\begin{aligned} E(Y) &= \sum_y y p_{XY}(\cdot, y) \\ &= 1.0 \times \frac{1}{8} + 2.0 \times \frac{1}{4} + 3.0 \times \frac{1}{8} + 4.0 \times \frac{1}{4} + 4.0 \times \frac{1}{4} \\ &= 3.0 \end{aligned}$$

Again note that 4.0 has been included twice due to appearing in the probability mass function twice. To calculate the variance of  $X$ , first calculate  $E(X^2)$

$$\begin{aligned} E(X^2) &= \sum_x x^2 p_{XY}(x, \cdot) \\ &= 1.0^2 \times \frac{1}{8} + 1.5^2 \times \frac{1}{4} + 1.5^2 \times \frac{1}{8} + 2.5^2 \times \frac{1}{4} + 3.0^2 \times \frac{1}{4} \\ &= 4.78125 \end{aligned}$$

Now the variance can be calculated by

$$V(X) = E(X^2) - E(X)^2 = 4.78125 - 2.0625^2 = 0.527344$$

Similarly by calculating  $E(Y^2)$

$$\begin{aligned} E(Y^2) &= \sum_y y^2 p_{XY}(\cdot, y) \\ &= 1.0^2 \times \frac{1}{8} + 2.0^2 \times \frac{1}{4} + 3.0^2 \times \frac{1}{8} + 4.0^2 \times \frac{1}{4} + 4.0^2 \times \frac{1}{4} \\ &= 10.25 \end{aligned}$$

the variance of  $Y$  can be calculated as

$$V(Y) = E(Y^2) - E(Y)^2 = 10.25 - 3.0^2 = 1.25$$

This can be checked in Julia using the following code

```
In [1]: x = [1.0, 1.5, 1.5, 2.5, 3.0]
        y = [1.0, 2.0, 3.0, 4.0, 4.0]
        p = [1/8, 1/4, 1/8, 1/4, 1/4]
```

```
EX = sum(x.*p)
EY = sum(y.*p)
EX2 = sum(x.^2.*p)
EY2 = sum(y.^2.*p)
```

```
VX = EX2-EX^2
VY = EY2-EY^2
```

```
[EX, EY, EX2, EY2, VX, VY]
```

```
Out[1]: 6-element Array{Float64,1}:
 2.0625
 3.0
 4.78125
10.25
 0.527344
 1.25
```

(g) Are  $X$  and  $Y$  independent random variables?

For  $X$  and  $Y$  to be independent it should not be possible to predict the value of one variable from the value of the other. This is not true as for  $X = 1.0$  the only  $Y$  value possible is 1.0.

(h)  $P(X + Y \leq 5)$

Here the sum of the value of  $X$  and  $Y$  needs to equal 5, so the final two cases are not counted.

$$P(X + Y \leq 5) = p_{XY}(1.0, 1.0) + p_{XY}(1.5, 2.0) + p_{XY}(1.5, 3.0) = 0.5$$

## Question 2. More Fun With Two Random Variables

Let  $X$  and  $Y$  be independent random variables with  $E(X) = 3$ ,  $V(X) = 4$ ,  $E(Y) = 5$ ,  $V(Y) = 9$ . Determine the following

(a)  $E(2X + 3Y)$ .

This is a linear combination of two random variables so this can be reduced to a linear combination of the expected values given above. This is done as such:

$$E(2X + 3Y) = E(2X) + E(3Y) = 2E(X) + 3E(Y) = 2 \times 3 + 3 \times 5 = 21$$

(b)  $V(2X + 3Y)$ .

Here a similar idea as part (a) can be used, however remember that the variance is a squared quantity and so the operations will need to also be squared. The calculation is:

$$V(2X + 3Y) = V(2X) + V(3Y) = 2^2V(X) + 3^2V(Y) = 4 \times 4 + 9 \times 9 = 97$$

Assume now further to the above that  $X$  and  $Y$  are normally distributed and determine the following:

(c)  $P(2X + 3Y > 18)$ .

First the linear combination of random variables needs to be standardised to the Standard Normal Distribution. This  $z$  value can then be looked up to determine the probability required.

$$\begin{aligned} P(2X + 3Y > 18) &= P\left(Z > \frac{18 - 21}{\sqrt{97}}\right) \\ &= P(Z > -0.30) \\ &= 0.6196 \end{aligned}$$

(d)  $P(2X + 3Y < 28)$ .

In a similar approach to the previous part, first standardise the random variable and then determine the probability required

$$\begin{aligned} P(2X + 3Y < 28) &= P\left(Z < \frac{28 - 21}{\sqrt{97}}\right) \\ &= P(Z < 0.74) \\ &= 0.7614 \end{aligned}$$

(e) Verify (c) and (d) using Julia code, where for each case you generate a million  $X$ 's and a million  $Y$ 's and simulate the linear combination  $2X + 3Y$ .

The code below will produce the results required. Remember that `Normal(mu, std)` rather than the lecture notes definition. Also that a vector of random numbers can be calculated by `rand(dist, N)`. The `.` operator here produces element wise operations.

```
In [2]: using Distributions
XNormDist = Normal(3, sqrt(4));
YNormDist = Normal(5, sqrt(9));

X = rand(XNormDist, 10^6);
Y = rand(YNormDist, 10^6);

linCombs = 2.*X+3.*Y;

mean(linCombs.>18)
```

Out[2]: 0.618317

```
In [3]: X = rand(XNormDist, 10^6);
Y = rand(YNormDist, 10^6);

linCombs = 2.*X+3.*Y;

mean(linCombs.<28)
```

Out[3]: 0.761042

(f) Assume now that the random variable come from another distribution (not Normal), but keep the same means and variances. Are your answers for (c) and (d) likely to change? How about your answers for (a) and (b)

As (c) and (d) are probabilities which are dependent on the distribution used they will change. However as the answers for (a) and (b) do not depend on the distribution and are only linear combinations of the mean and variance respectively they will remain the same.

This can be demonstrated if the distribution is change to a Uniform Distribution (the only one covered in the course that will allow the means and variances to remain the same). To calculate the start and end points, first calculate the relationship between  $a$  and  $b$  (the start and end points) using the Expected Value and Variance formulae.

$$\begin{aligned} E(X') &= 3 = \frac{a+b}{2} \\ 6 &= a+b \\ 6-b &= a \end{aligned}$$

$$\begin{aligned} V(X') &= 4 = \frac{(b-a)^2}{12} \\ 48 &= (b-a)^2 \end{aligned}$$

Combining these equations the values for  $a$  and  $b$  can be calculated

$$\begin{aligned} 48 &= (b - (6 - b))^2 \\ &= (2b - 6)^2 \\ &= 4(b - 3)^2 \\ 12 &= (b - 3)^2 \\ &= b^2 - 6b + 9 \\ 0 &= b^2 - 6b - 3 \\ b &= \frac{6 \pm \sqrt{6^2 + 4 \times 3}}{2} \\ &= \frac{6 \pm 4\sqrt{3}}{2} \\ &= 3 + 2\sqrt{3} \quad \text{Taking larger value} \end{aligned}$$

$$\begin{aligned} a &= 6 - b \\ &= 6 - 3 - 2\sqrt{3} \\ &= 3 - 2\sqrt{3} \end{aligned}$$

The same process can be done to create  $Y'$  with  $E(Y') = 5$  and  $V(Y') = 9$ , finding the end points to be  $[5 - 3\sqrt{3}, 5 + 3\sqrt{3}]$ . Using Julia to simulate these distributions as before:

```
In [4]: XUniformDist=Uniform(3-2*sqrt(3), 3+2*sqrt(3))
        YUniformDist=Uniform(5-3*sqrt(3), 5+3*sqrt(3))

        X = rand(XUniformDist,10^6);
        Y = rand(YUniformDist,10^6);

        linCombs= 2.*X+3.*Y;

        mean(linCombs.>18)
```

Out[4]: 0.596642

```
In [5]: X = rand(XUniformDist,10^6);
        Y = rand(YUniformDist,10^6);

        linCombs= 2.*X+3.*Y;

        mean(linCombs.<28)
```

Out[5]: 0.724571

As predicted, these values are slightly different to those obtained in (c) and (d).

(g) Assume now that  $X$  and  $Y$  are Normally distributed but are not independent, but rather  $Cov(X, Y) = 5$ . Write an explicit expression using a double integral for  $P(X < 2, Y > 7)$ .

From the previous parts,  $\mu_X = 3$ ,  $\mu_Y = 5$ ,  $\sigma_X = \sqrt{4} = 2$  and  $\sigma_Y = \sqrt{9} = 3$ . Now calculate the correlation coefficient  $\rho$

$$\begin{aligned}\rho &= \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{5}{2 \times 3} \\ &= \frac{5}{6}\end{aligned}$$

For the bivariate normal distribution

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}$$

Substituting in the values above and simplifying

$$\begin{aligned}f_{X,Y}(x, y) &= \frac{1}{2\pi \times 2 \times 3\sqrt{1-\frac{25}{36}}} \times \exp\left\{\frac{-1}{2\left(1-\frac{25}{36}\right)}\left[\frac{(x-3)^2}{2^2} - \frac{2 \times \frac{5}{6}(x-3)(y-5)}{2 \times 3} + \frac{(y-5)^2}{3^2}\right]\right\} \\ &= \frac{1}{12\pi\sqrt{\frac{11}{36}}} \times \exp\left\{\frac{-18}{11}\left[\frac{(x-3)^2}{4} - \frac{5(x-3)(y-5)}{18} + \frac{(y-5)^2}{9}\right]\right\} \\ &= \frac{1}{2\pi\sqrt{11}} \times \exp\left\{-\frac{9(x-3)^2}{22} + \frac{5(x-3)(y-5)}{11} - \frac{2(y-5)^2}{11}\right\} \\ &= \frac{1}{2\pi\sqrt{11}} \times \exp\left\{-\frac{1}{22}[(3(x-3) - 2(y-5))^2 + 2(x-3)(y-5)]\right\}\end{aligned}$$

Then the definite integral is as follows

$$\begin{aligned}P(X < 2, Y > 7) &= \int_{x=-\infty}^2 \int_{y=7}^{\infty} \frac{1}{2\pi\sqrt{11}} \times \exp\left\{-\frac{1}{22}[(3(x-3) - 2(y-5))^2 + 2(x-3)(y-5)]\right\} dx dy \\ &= \frac{1}{2\pi\sqrt{11}} \int_{x=-\infty}^2 \int_{y=7}^{\infty} \exp\left\{-\frac{1}{22}[(3(x-3) - 2(y-5))^2 + 2(x-3)(y-5)]\right\} dx dy\end{aligned}$$

### Question 3. Rise of the machines

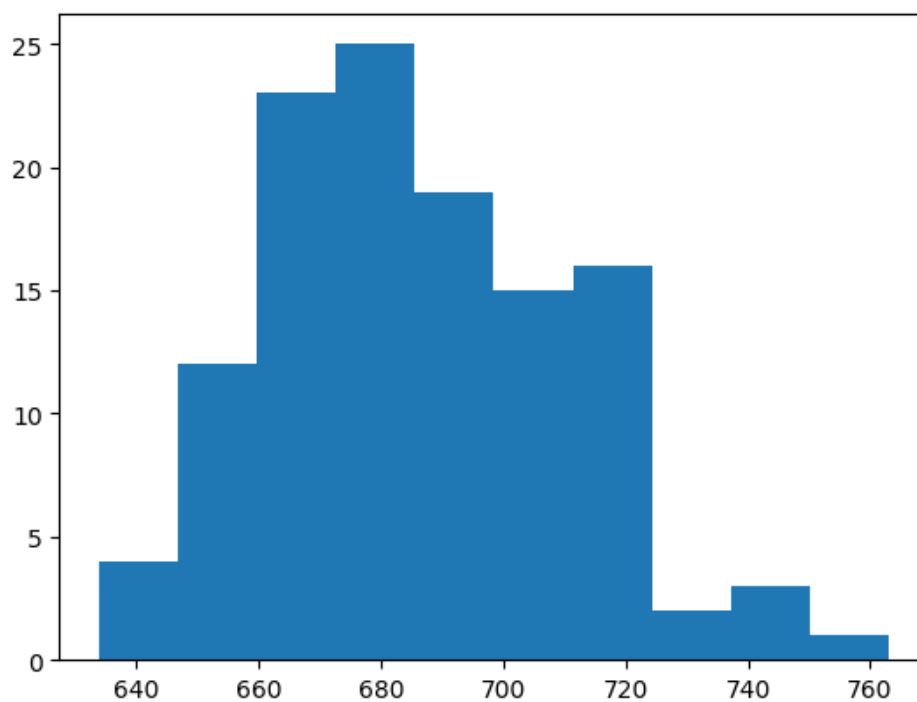
A semiconductor manufacturer produces devices used as central processing units in personal computers. The speed of the devices (in megahertz) is important because it determines the price that the manufacturer can charge for the devices. The files (6-42.csv) contains measurements on 120 devices. Construct the following plots for this data and comment on any important features that you notice.

(a) Histogram

Looking at the histogram it can be seen that it is slightly right skewed with a peak around 670 Megahertz. Also the histogram shows a fairly wide peak.

```
In [6]: using DataFrames, StatsBase, PyPlot, Distributions, KernelDensity
speeds = readtable("6-42.csv",header=false)
speeds = speeds[1]

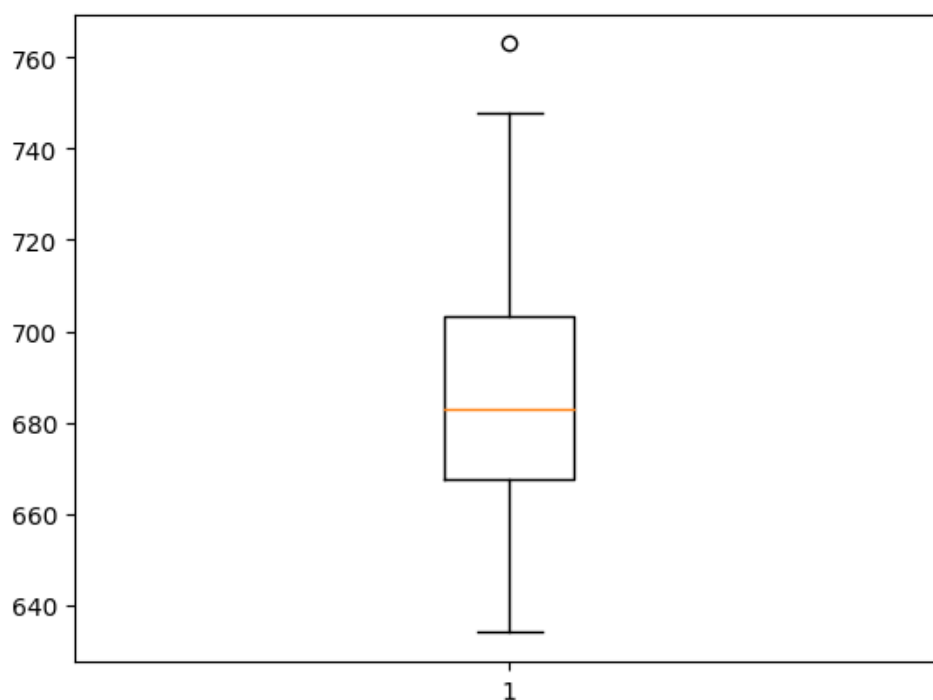
PyPlot.plt[:hist](speeds);
```



#### (b) Boxplot

As with the histogram it can be seen that there is a right skew to the data, and that this skewness is also present in the interquartile range. There is one outlier above 760 Megahertz that should be checked.

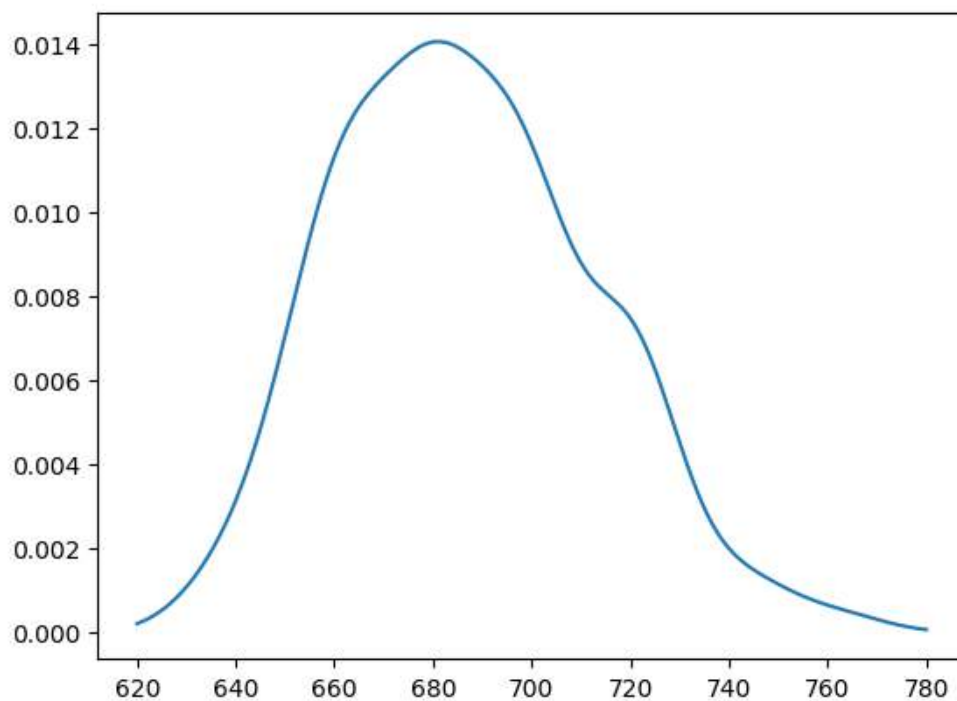
```
In [7]: PyPlot.boxplot(speeds);
```



### (c) Kernel Density Estimate

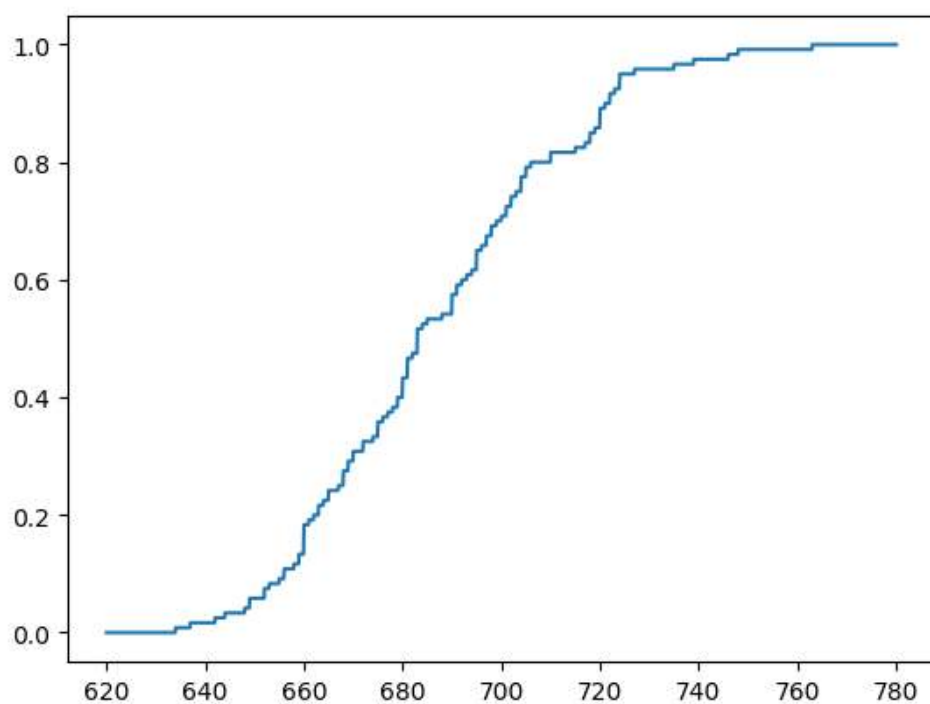
As with the past two plots a right skewness can be observed. The graph is not symmetric with a slight bump at approximately 700 to 740 Megahertz.

```
In [8]: speedKDE= kde(speeds)
        grid = 620:0.1:780
        PyPlot.plot(grid,pdf(speedKDE,grid));
```



### (d) Empirical cumulative distribution function

```
In [9]: estimatedCDF=ecdf(speeds)
        PyPlot.plot(grid,estimatedCDF(grid));
```





Further, compute:

(e) The sample mean, the sample standard deviation and the sample median.

```
In [10]: println("The sample mean is ", mean(speeds), " MHz.")
println("The sample standard deviation is ", std(speeds), " MHz.")
println("The sample median is ", median(speeds), " MHz.")

The sample mean is 686.775 MHz.
The sample standard deviation is 25.668036723592586 MHz.
The sample median is 683.0 MHz.
```

As can be seen from the above summary statistics there is a slight right skewness to the data with the median being lower than the mean.

(f) What percentage of the devices has a speed less than 750 megahertz?

Here count the number of times the speed is less than 750 megahertz and then divide by the total number of speeds. The `mean` function will allow the sum to be taken and then divided by the total number of entries.

```
In [11]: mean(speeds.<750)

Out[11]: 0.9916666666666667
```

Alternatively, the estimated cumulative density function can also provide this information

```
In [12]: estimatedCDF(750)

Out[12]: 0.9916666666666667
```

So 99.16% of the devices have a speed less than 750 megahertz.

## Question 4. The thickest rod

Eight measurements were made on the inside diameter of forged piston rings in an automobile engine. The data (in millimetres) is:

74.004, 73.999, 74.021, 74.001, 74.006, 74.002, 74.005.

Use the Julia function below to construct a normal probability plot of the piston ring diameter data. Does it seem reasonable to assume that piston ring diameter is normally distributed? How about if you remove a single observation that is potentially an outlier?

```
In [13]: using PyPlot, Distributions, StatsBase

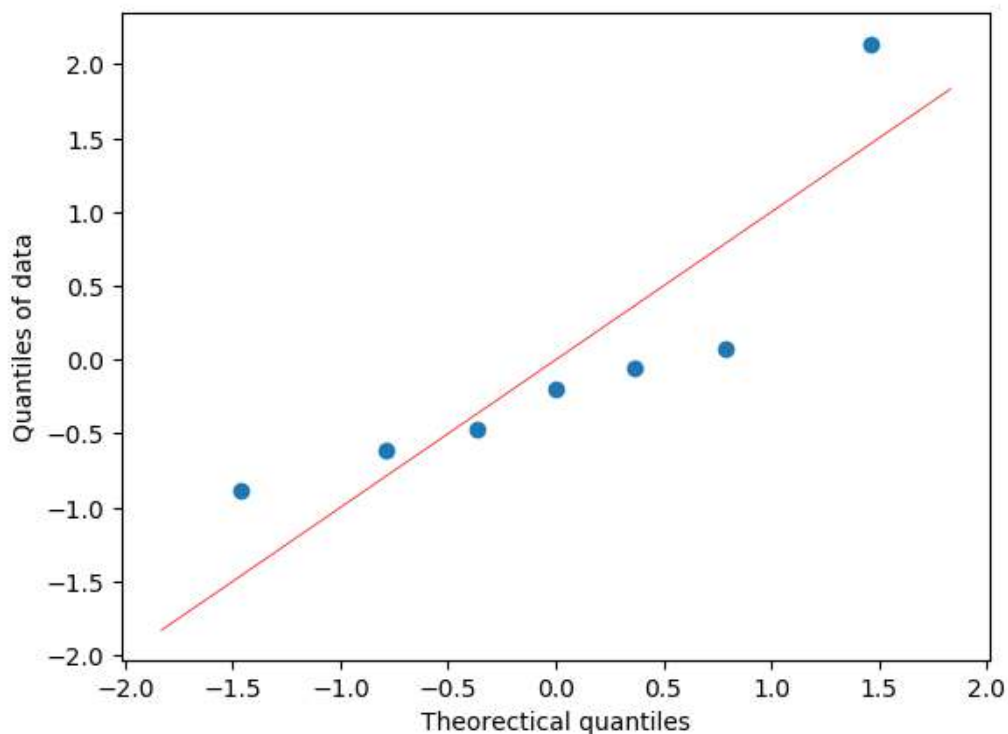
function NormalProbabilityPlot(data)
    mu = mean(data)
    sig = std(data)
    n = length(data)
    p = [(i-0.5)/n for i in 1:n]
    x = quantile.(Normal(),p)
    y = sort([(i-mu)/sig for i in data])
    PyPlot.scatter(x,y)
    xRange = maximum(x) - minimum(x)
    PyPlot.plot([minimum(x) - xRange/8,maximum(x) + xRange/8],[minimum(x) - xRange/8,maximum(x) + xRange/8],
        color="red",linewidth=0.5)
    xlabel("Theoretical quantiles")
    ylabel("Quantiles of data")
    return
end
```

Out[13]: NormalProbabilityPlot (generic function with 1 method)

First read in the data by constructing a vector of the values. Then input this vector into the function above.

```
In [14]: rods = [74.004,73.999,74.021,74.001,74.006,74.002,74.005]

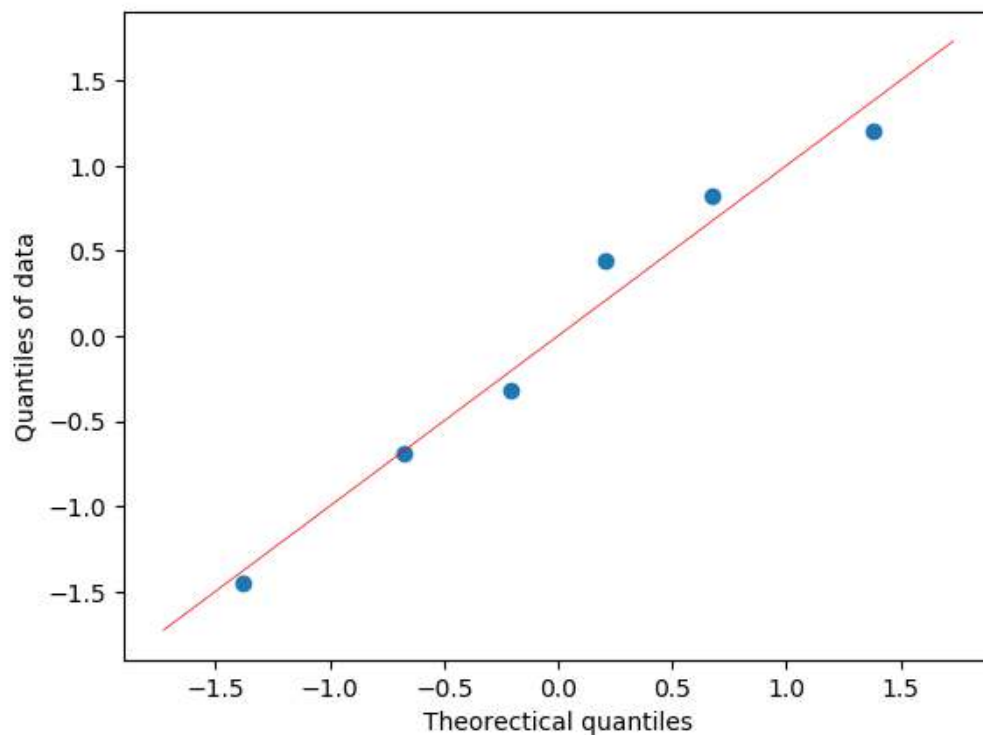
NormalProbabilityPlot(rods)
```



As the data does not follow the line of the plot, the data can not be said to be distributed normally. There is a potential outlier at the highest point, so this should be removed and then the plot created again.

```
In [15]: rods2 = [74.004, 73.999, 74.001, 74.006, 74.002, 74.005]
```

```
NormalProbabilityPlot(rods2)
```



With such a small data set it is hard to determine any properties, however a wave pattern is appearing suggesting that the data is not normally distributed.

## Question 5. A non-flat earth

In 1789, Henry Cavendish estimated the density of the Earth by using a torsion balance. His 29 measurements are in the file (6-122.csv), expressed as a multiple of the density of water.

(a) Calculate the sample mean, sample standard deviation, and median of the Cavendish density data

First read in the data from the file provided using `readtable` then run the Julia functions for mean, standard deviation and median.

```
In [16]: cavendish = readtable("6-122.csv", header=false)
cavendish = cavendish[1]

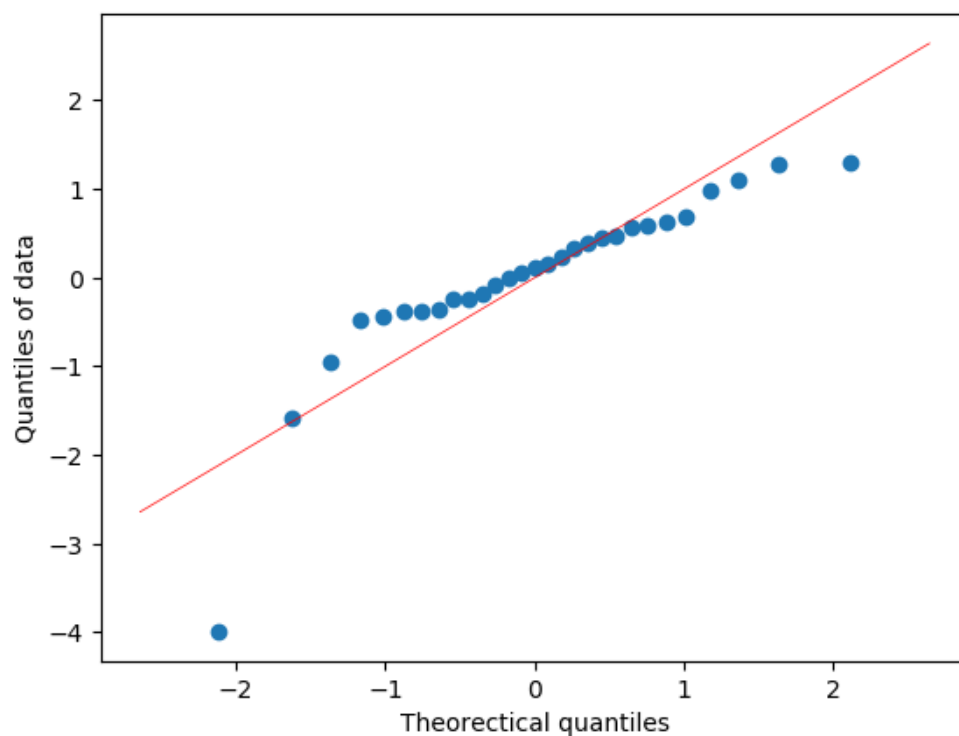
println("The sample mean is ", mean(cavendish))
println("The sample standard deviation is ", std(cavendish))
println("The sample median is ", median(cavendish))
```

```
The sample mean is 5.419655172413792
The sample standard deviation is 0.3388792743169407
The sample median is 5.46
```

(b) Construct a normal probability plot of the data. Comment on the plot. Does there seem to be a "low" outlier in the data?

Using the function defined in Question 4 the following plot is obtained

```
In [17]: NormalProbabilityPlot(cavendish)
```



There does appear to be a low outlier on the plot at approximately -4. Also the plot would be linear in the middle if this were removed with some deviation towards the end of the line.

(c) Would the sample median be a better estimate of the density of the earth than the sample mean? Why?

With the presence of an outlier in the data the median would be a better estimate of center as it is robust against the presence of outliers.

## Question 6. Normal Confidence Interval

A normal population has a mean 122.3 and variance 36. How large must the random sample be if you want the standard error of the sample average to be 1.5?

The standard error of the mean is given by

$$S. E. (\bar{x}) = \frac{s}{\sqrt{n}}$$

where  $n$  is the sample size. Substituting the values in

$$\begin{aligned} 1.5 &= \frac{\sqrt{36}}{\sqrt{n}} \\ 1.5^2 &= \frac{36}{n} \\ n &= \frac{36}{2.25} \\ n &= 4^2 = 16 \end{aligned}$$

So the random sample must be at least 16 items large.

## Question 7. Fill in the blanks

A random sample has been taken from a normal distribution. Output from a software package follows:

Variable	N	Mean	SE Mean	StDev	Variance	Sum
$x$	?	?	1.58	6.11	?	701.20

(a) Fill in the missing quantities

First the variance is simply the standard deviation squared so will equal 37.33. Recall from the previous question that  $N$  is the variance divided by the standard error of the mean squared so is 14.95. As all numbers are rounded to two decimal places the ceiling of this value should be taken making  $N = 15$ . From this the mean can be calculated from dividing the sum by 15. This leads to the complete table:

Variable	N	Mean	SE Mean	StDev	Variance	Sum
$x$	15	46.75	1.58	6.11	37.33	701.20

(b) Find a 99% CI on the population mean under the assumption that the standard deviation is known.

A confidence interval is calculated using the following formula

$$\left[ \bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}} \right]$$

Using the value from the table above:

$$[46.75 - 2.5758 \times 1.58, 46.75 + 2.5758 \times 1.58]$$

$$[42.68, 50.82]$$

## Question 8. More on randomization test

Reproduce class example 3. Now modify the data so that the yield of the Fertilizer is decreased by exactly 0.5 kg per observation (i.e the first observation is 5.81, the second is 4.62 and so fourth). What are the results now? How do you interpret them?

First reproduce class example 3, the code being given below and make sure the value agrees with the given solution of 2.39%.

```
In [18]: using Combinatorics
fert=readtable("Fertilizer.csv")
control = fert[1]
fertilizer = fert[2]

x = collect(combinations([control;fertilizer],10))

println("Number of combinations: ", length(x))

pvalue = sum([mean(c) >= mean(fertilizer) for c in x])/length(x)

Number of combinations: 184756
```

```
Out[18]: 0.023972157873086666
```

Now subtract 0.5 from each value of the fertilizer data, this can be done with elementwise operations.

```
In [19]: fertilizer = fertilizer.-0.5
```

```
Out[19]: 10-element DataArrays.DataArray{Float64,1}:  
 5.81  
 4.62  
 5.04  
 5.0  
 4.87  
 4.79  
 4.42  
 5.65  
 5.3  
 4.76
```

Recreate the combinations of the values and check to see how many are larger than the result observed.

```
In [20]: x = collect(combinations([control;fertilizer],10))  
  
println("Number of combinations: ", length(x))  
  
pvalue = sum([mean(c) >= mean(fertilizer) for c in x])/length(x)  
  
Number of combinations: 184756
```

```
Out[20]: 0.5118751217822425
```

Here it can be seen that the probability of obtaining the result observed or a greater effect is 51.19%. This means there is no longer evidence to reject the null hypothesis that the mean yield is the same for both groups.