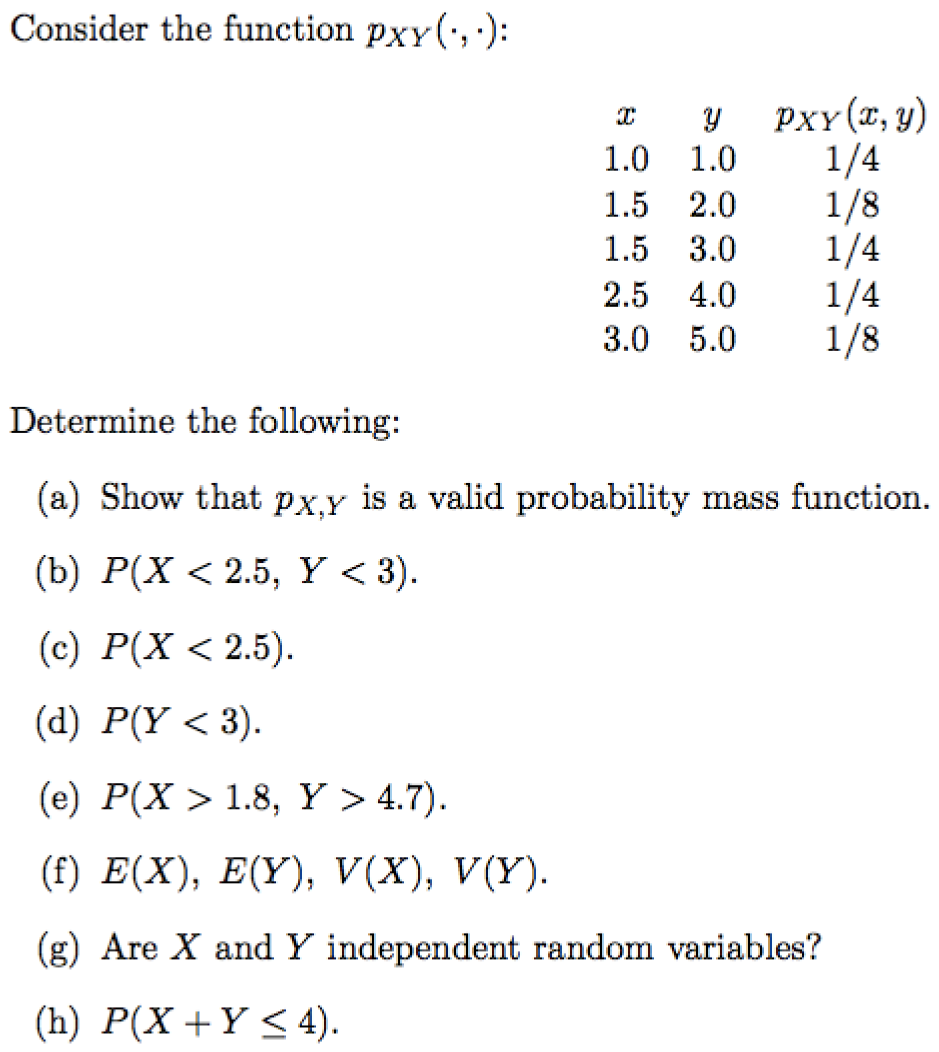
STAT2201, Semester 1, 2017

Solution for Assignment 3

Questions marked for grade (each 20%): Q2, Q3, Q6, Q8

Question 1 - Joint Probability Mass Function



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

x	y	$p_{XY}(x,y)$
1.0	1.0	1/4
1.5	2.0	1/8
1.5	3.0	1/4
2.5	4.0	1/4
3.0	5.0	1/8

Determine the following:

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

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

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

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

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

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

(g) Are X and Y independent random variables?

(h) $P(X + Y \leq 4)$.

Solution:

(a) Carry out, $\sum_{x,y}\ p(x,y)$ on all values of x and y in the support and see it is 1.

(b) $P(X < 2.,5, Y < 3) = p(1,1) + p(1.5,2) = 3/8$

(c) $P(X < 2.5) = p(1,1) + p(1.5,2) + p(1.5,3) = 5/8$

(d) $P(Y < 3) = p(1,1) + p(1.5,2) = 3/8$

(e) $P(X > 1.8, Y > 4.7) = p(3,5) = 1/8$

(f) $E(X) = \sum_x x\, P(X=x) = 1 * 1/4 + 1.5 * (1/8 + 1/4) + 2.5 * 1/4 + 3.0 * 1/8 = 1.8125$ (notice that the coefficient of 1.5 is $p(1.5,2) + p(1.5,3)$)

$E(Y) = \sum_y y\, P(Y=y) = 1 * 1/4 + 2 * 1/8 + 3 * 1/4 + 4 * 1/4 + 5 * 1/8 = 2.875$

$E(X^2) = \sum_x x\, P(X=x) = 1^2 * 1/4 + 1.5^2 * (1/8 + 1/4) + 2.5^2 * 1.4 + 3.0^2 * 1/8 = 10.9687$

$E(Y^2) = \sum_y y\, P(Y=y) = 1^2 * 1/4 + 2^2 * 1/8 + 3^2 * 1/4 + 4^2 * 1/4 + 5^2 * 1/8 = 10.125$

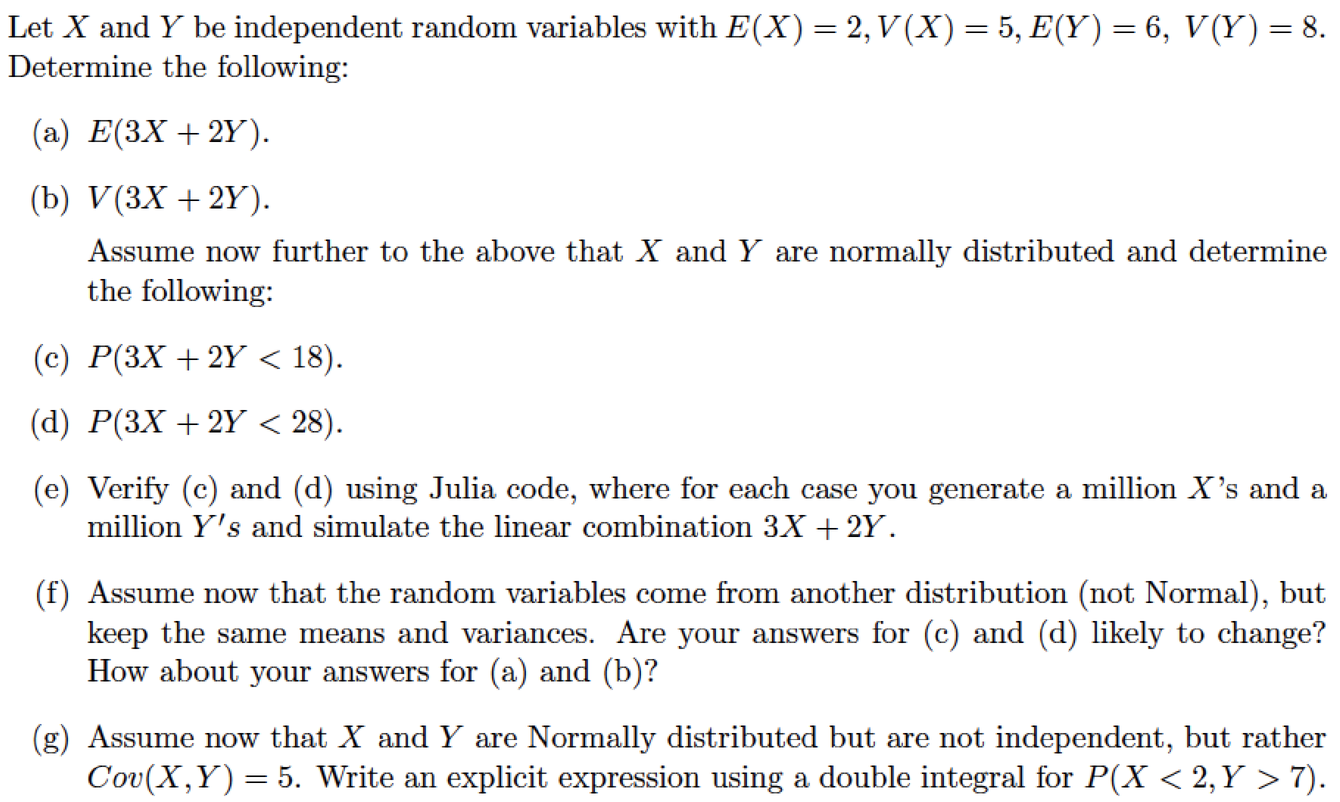
$V(X) = E(X^2) - E(X)^2 = 7.68359375$

$V(Y) = E(Y^2) - E(Y)^2 = 2.703125$

(g) No. The random variables are not independent. Look for example at the fact that if $Y = 1.0$ you know that $X = 1.0$.

(h) $P(X + Y \leq 4) = p(1,1) + p(1.5,2) = 3/8$

Question 2 - More Fun With Two Random Variables



Let X and Y be independent random variables with $E(X)=2, V(X)=5, E(Y)=6,\ V(Y)=8$. Determine the following:

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

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

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

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

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

(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 $3X+2Y$.

(f) Assume now that the random variables 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)?

(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 $P(X<2,Y>7)$.

Solution:

(a) $E(3X+2Y)=3E(X)+2E(Y)=3*2+2*6=18$

(b) $V(3X+2Y)=9V(X)+4V(Y)=9*5+4*8=77$

(c) Since they are normally distributed, linear combinations are also normally distributed. Define $W=3X+2Y$, then $W\sim N(18,77)$. In this case $P(W<18)=1/2$ because 18 is the mean.

(d)

In [14]:

```
using Distributions
cdf(Normal(18,sqrt(77)),28)
```

Out[14]:

```
0.8727747086768317
```

Or using a table, $P(W<28)=P(Z\leq\frac{28-18}{\sqrt{77}})=P(Z\leq 1.1396)=\phi(1.1396)=0.8727$

(e) First let's just check the mean and variance are as we expect (this wasn't asked for, but good to do)

In [20]:

```
data  = [3*rand(Normal(2,sqrt(5))) + 2 * rand(Normal(6,sqrt(8))) for _ in 1:10^6];
mean(data),var(data)
```

Out[20]:

```
(18.023643888289378,76.98991504620085)
```

In [21]:

```
sum([w < 18 for w in data])/10^6    ,    sum([w < 28 for w in data])/10^6
```

Out[21]:

```
(0.499125,0.871936)
```

(f) In general, the answers to (c) and (d) depend on the distribution. So yes, the answers are likely to change. As opposed to that, the answer to (a) and (b) will not change because the mean and variance calculations did not assume any specific distributional form.

(g) We have, $\mu_x = 2, \mu_y = 6, \sigma_x = \sqrt{5}, \sigma_y = \sqrt{8}$

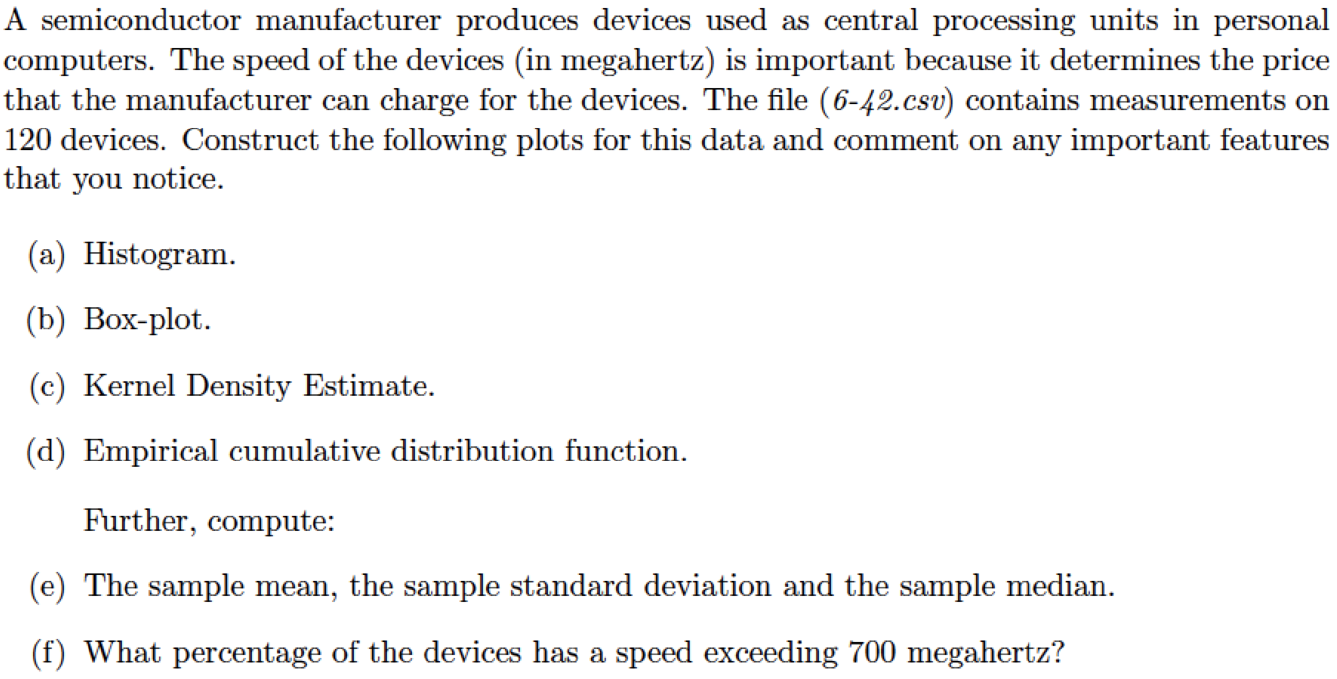
Now the correlation coefficient, $\rho = \frac{Cov(X,Y)}{\sigma_x \, \sigma_y} = .7906$

With these five parameters, we have a well specified bivariate normal density, $f_{X,Y}(x,y)$ as for example in Chapter 4 of the condensed lecture notes.That is, plug in the 5 parameters of the bivarate normal in the joint density expression.

Now,

$$P(X < 2, Y > 7) = \int_{x=-\infty}^{2} \int_{y=7}^{\infty} f_{X,Y}(x,y) \, dx \, dy$$

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 file (*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.

(b) Box-plot.

(c) Kernel Density Estimate.

(d) Empirical cumulative distribution function.

Further, compute:

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

(f) What percentage of the devices has a speed exceeding 700 megahertz?

Solution:

In [31]:

```
using DataFrames
table = readtable("6-42.csv",header = false)
data = table[1]
length(data)
```

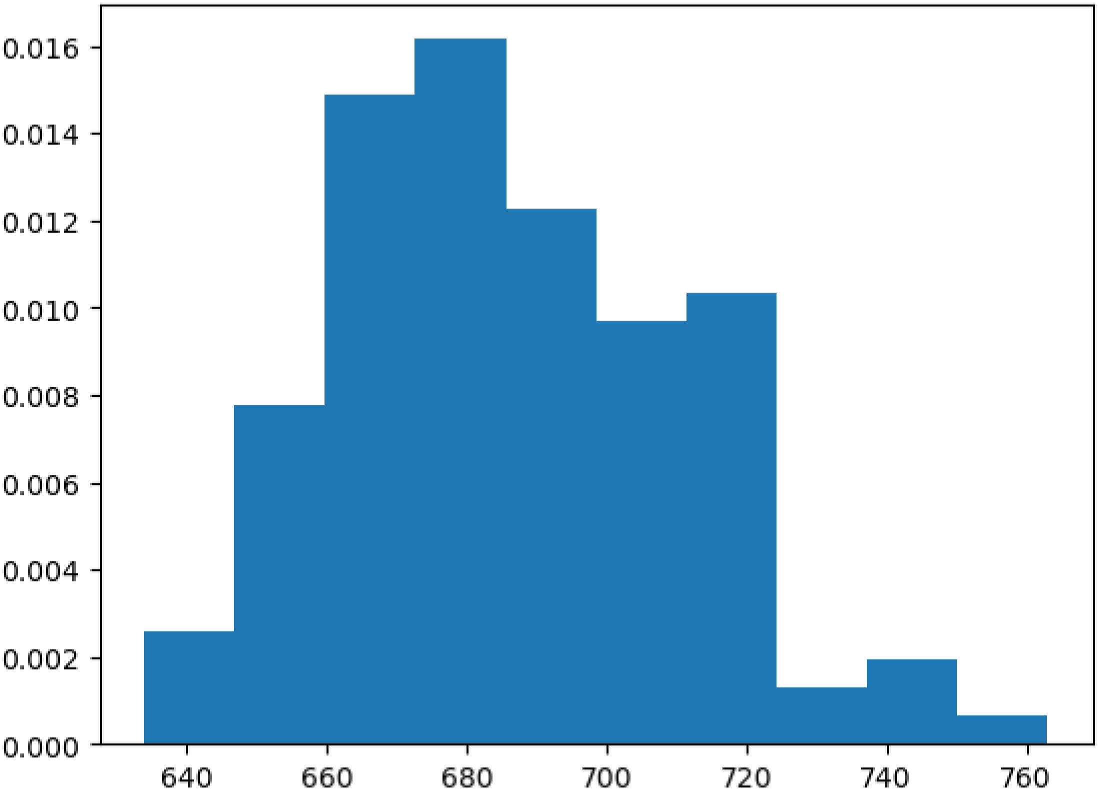
Out[31]:

```
120
```

(a)

In [35]:

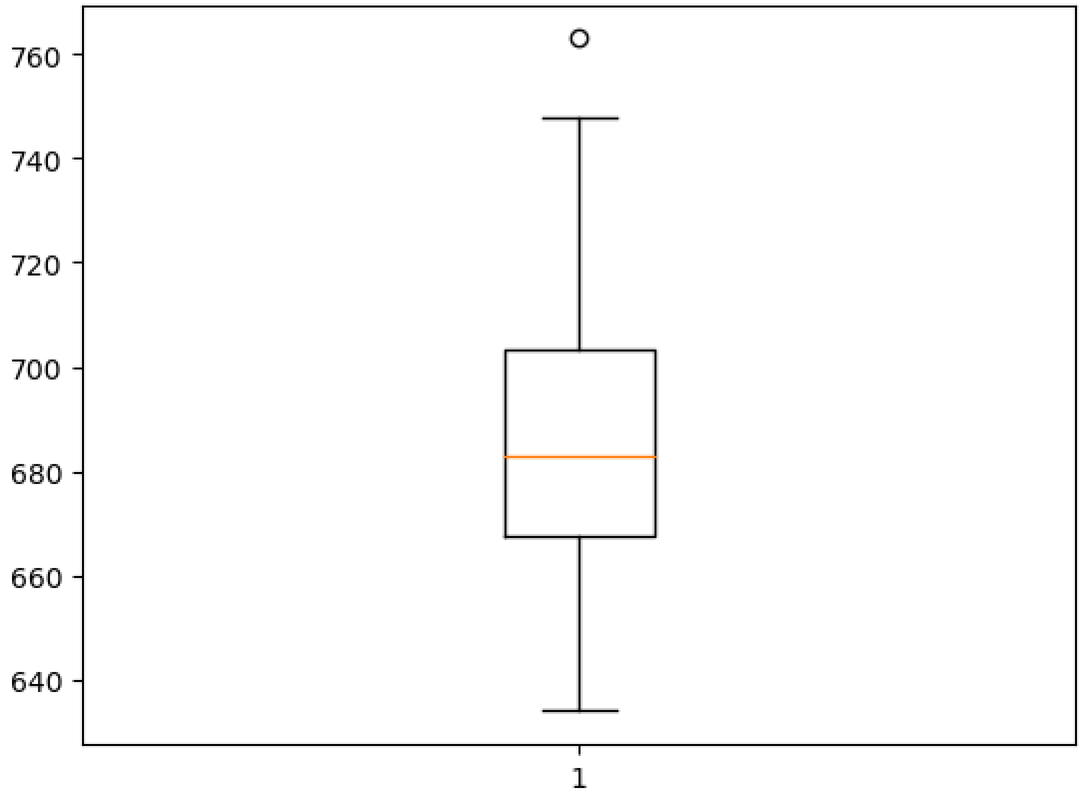
```
using PyPlot
PyPlot.plt[:hist](data,normed = true);
```



(b)

In [37]:

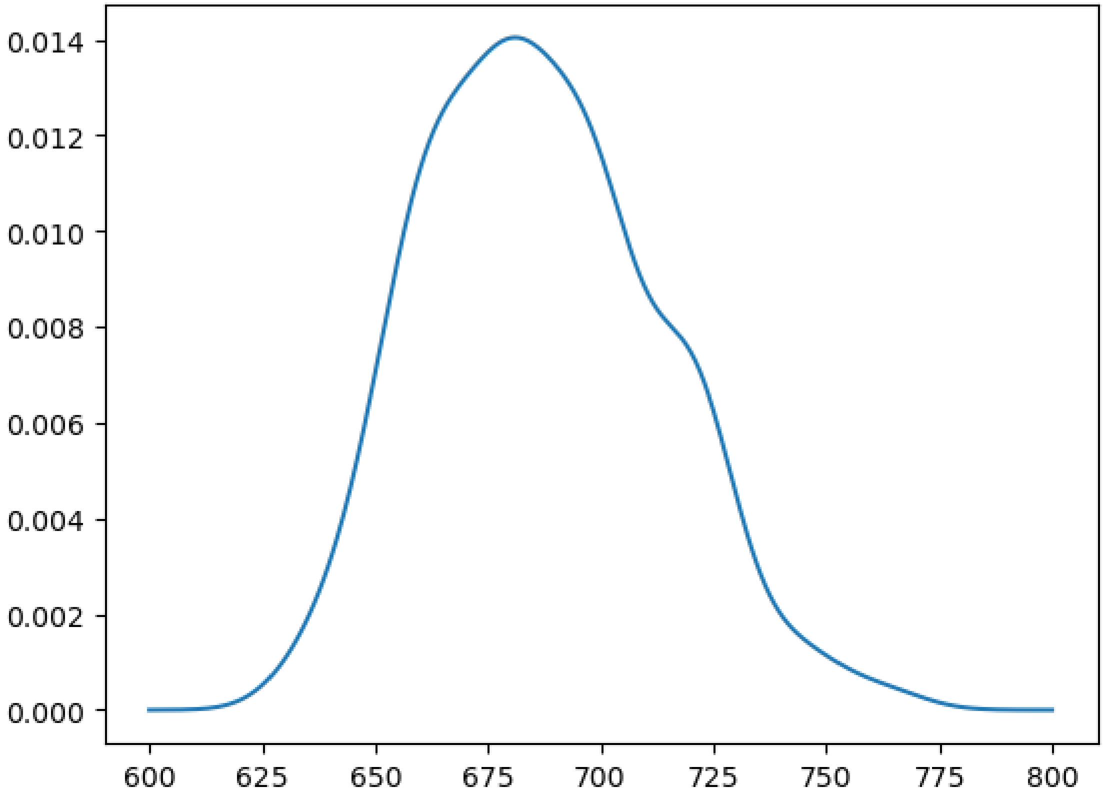
```
PyPlot.boxplot(data);
```



(c)

In [42]:

```
using KernelDensity
k = kde(data)
support = 600:0.5:800
p = pdf(k,support)
PyPlot.plot(support,p);
```



(d)

In [45]:

```
mean(data),std(data),median(data)
```

Out[45]:

```
(686.775,25.668036723592586,683.0)
```

(e) Percentage of devices > 700mhz

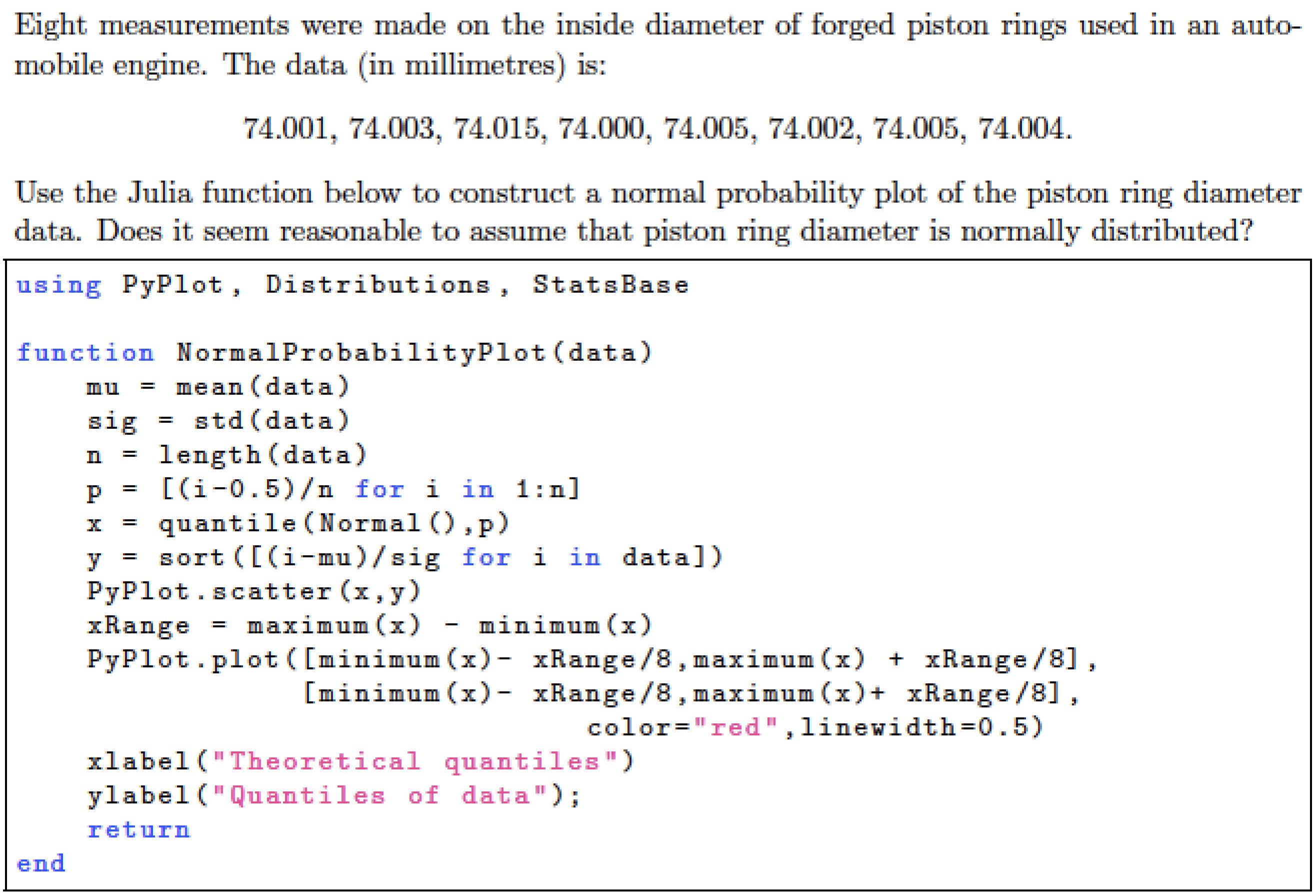
In [47]:

```
100*sum([x > 700 for x in data])/length(data)
```

Out[47]:

```
29.166666666666668
```

Question 4 - The Thickest Rod



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

74.001, 74.003, 74.015, 74.000, 74.005, 74.002, 74.005, 74.004.

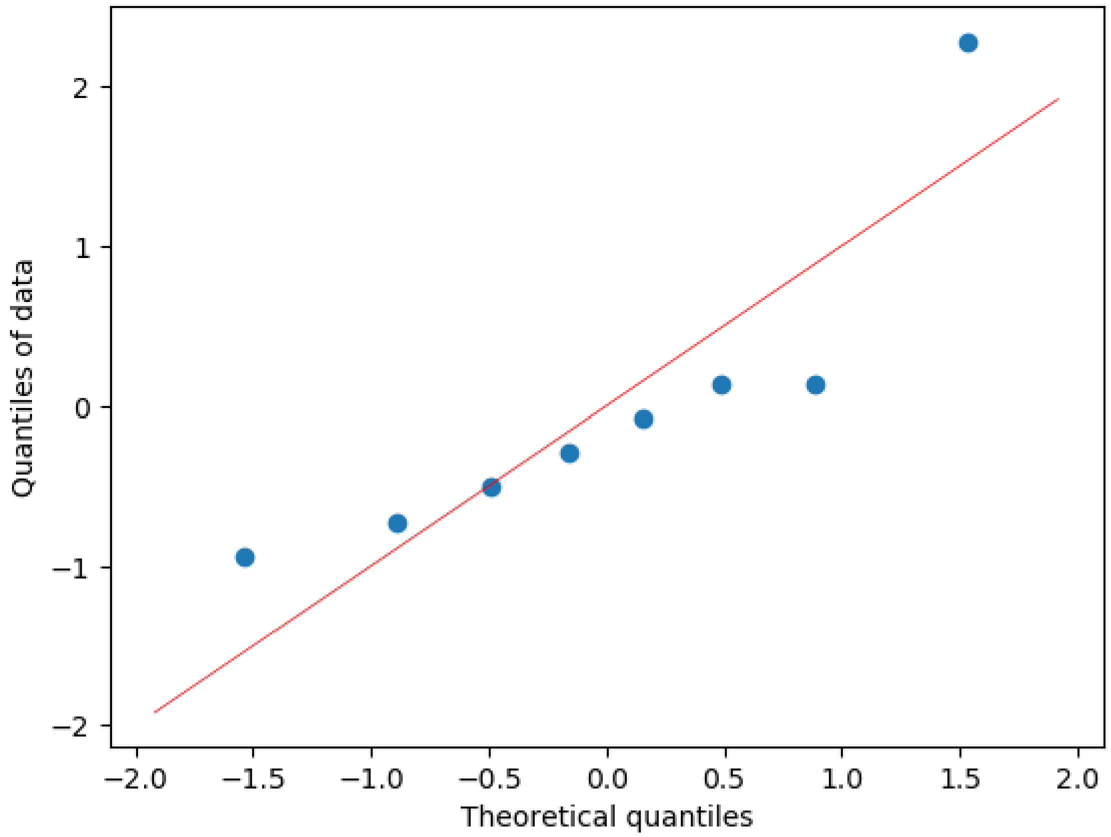
Use the Julia function below to construct a normal probability plot of the piston ring diameter data. Does it seem reasonable to assume that piston ring diameter is normally distributed?

```
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
```

Solution:

In [57]:

```
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
dat = [74.001, 74.003, 74.015, 74.000, 74.005, 74.002, 74.005, 74.004];
NormalProbabilityPlot(dat)
```

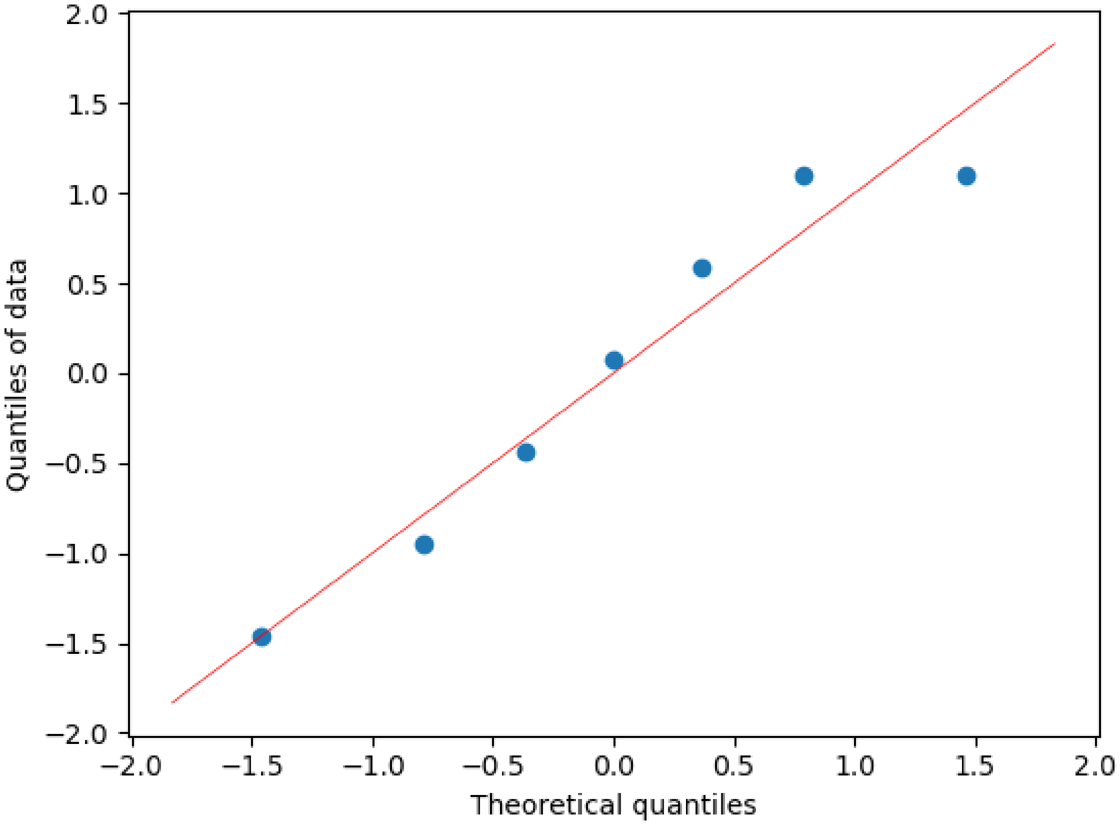


```
WARNING: Method definition NormalProbabilityPlot(Any) in module Main at In
[56]:4 overwritten at In[57]:4.
```

With so few observations it isn't immediatly clear if there is evidence for the data being normally distributed or not. Observe that the highest valued observation (74.015) in the normally probability plot skews the plot. Here is a plot without that observation:

In [58]:

```
dat = [74.001, 74.003, 74.000, 74.005, 74.002, 74.005, 74.004];
NormalProbabilityPlot(dat)
```



Now it appears that the data is NOT normally distributed. Even though it is hard to determine with so few observations, there appears to be a pattern that is clearly not following the straight line.

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 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.

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

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

Solution:

(a)

In [64]:

```
table = readtable("6-122.csv",header=false)
data = table[1]
length(data)
```

Out[64]:

```
29
```

In [65]:

```
mean(data), std(data), median(data)
```

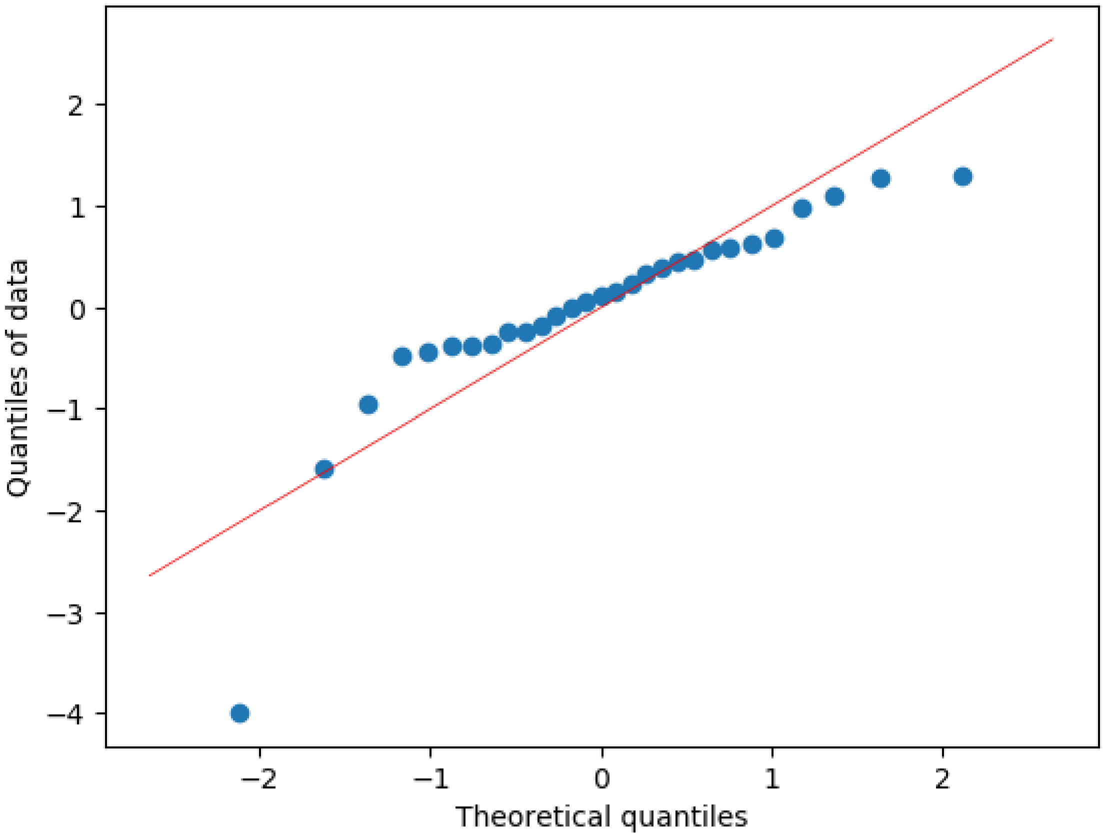
Out[65]:

```
(5.419655172413792,0.3388792743169407,5.46)
```

(b)

In [66]:

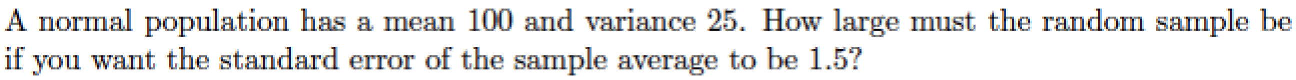
```
NormalProbabilityPlot(data)
```



Yes, there appears to be a low outlier.

(c) Yes, given the fact that there is such an outlier, the sample median is probably a better estimate since it is robust to outliers.

Question 6 - Normal Confidence Interval



A normal population has a mean 100 and variance 25. How large must the random sample be if you want the standard error of the sample average to be 1.5?

Solution:

The sample mean has a variance of $25/n$ and a standard error of $5/\sqrt{n}$. So we want, $5/\sqrt{n} \leq 1.5$ or $(5/1.5)^2 \leq n$.

In [67]:

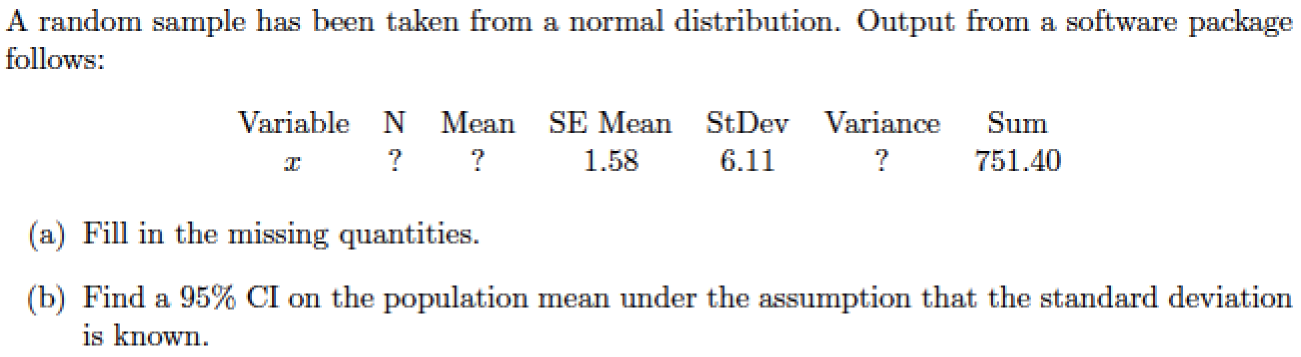
```
(5/1.5)^2
```

Out[67]:

```
11.111111111111112
```

Hence we need $n = 12$ observations.

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	?	751.40

(a) Fill in the missing quantities.

(b) Find a 95% CI on the population mean under the assumption that the standard deviation is known.

Solution:

(a) Variance = StDev^2 = 37.33

SE Mean = StDev/$\sqrt{N}$. Hence $\sqrt{N}$ = StDev/SEMean. Hence $N = 15$.

Mean = Sum/N = 50.09

(b) Here is the $z_{1-\alpha/2}$ quantile:

In [72]:

```
quantile(Normal(),0.975)
```

Out[72]:

```
1.9599639845400583
```

In [73]:

```
1.96*1.58
```

Out[73]:

```
3.0968
```

This is the 1.96 appearing in the t-table with infinite degrees of freedom. The confidence interval is then, 50.09 ± 3.0968. Or,

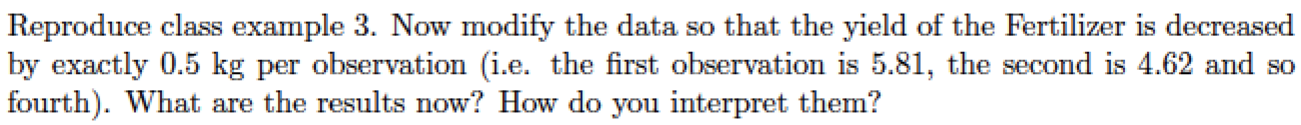
In [74]:

```
(50.09-3.0968,50.09+3.0968)
```

Out[74]:

```
(46.9932,53.186800000000005)
```

Question 8 - More on the 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?

Solution:

In [85]:

```
using Combinatorics , DataFrames

data = readtable("Fertilizer.csv")
control = data[1]
fertilizer = data[2]-0.5   #HERE WE DECREASE BY 0.5

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

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

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

```
Number of combinations : 184756
```

Out[85]:

0.5118751217822425

Now with the control and the fertilizier being so close to each other, the P-value is at 0.511. Hence there is no strong evidence for rejection of H_0 : MEANS EQUAL.

In [ ]: