

## Question 1. Seeing the CLT with simulation

Consider the following random variables

$$U \sim \text{Uniform}(5, 15)$$

$$V \sim \text{Exponential}(10)$$

$$W \sim \text{Binomial}(10, 0.4)$$

(a) What is the mean and variance of each?

### Uniform

The mean of an Uniform distributed variable is given by

$$\mu_U = E(U) = \frac{a+b}{2} = \frac{5+15}{2} = 10$$

The variance is given by

$$\sigma_U^2 = \text{Var}(U) = \frac{(b-a)^2}{12} = \frac{(15-5)^2}{12} = 8\frac{1}{3}$$

### Exponential

The mean of an Exponentially distributed variable is given by

$$\mu_V = E(V) = \frac{1}{\lambda} = \frac{1}{10} = 0.1$$

The variance is given by

$$\sigma_V^2 = \text{Var}(V) = \frac{1}{\lambda^2} = \frac{1}{10^2} = 0.01$$

### Binomial

The mean of a Binomial distributed variable is given by

$$\mu_W = E(W) = np = 10 \times 0.4 = 4$$

The variance is given by

$$\sigma_W^2 = np(1-p) = 10 \times 0.4 \times (1-0.4) = 2.4$$

(b) Consider now,

$$S_n = \sum_{i=1}^n X_i,$$

where  $X_i$  is either  $U_i$ ,  $V_i$  or  $W_i$  (distributed as  $U$ ,  $V$  or  $W$ ) and different  $X_i$  are assumed independent. What is the mean and variance of this random sum,  $S_i$ , (a function of  $n$ )? Answer this separately for  $U$ ,  $V$  and  $W$ .

In all cases this is a sum of random variables and so

$$\begin{aligned}\mu_{S_i} = E(S_i) &= E\left(\sum_{i=1}^n X_i\right) \\ &= \sum_{i=1}^n E(X_i)\end{aligned}$$

$$\begin{aligned}\sigma_{S_i}^2 = \text{Var}(S_i) &= \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \sum_{i=1}^n \text{Var}(X_i)\end{aligned}$$

As the variables will be independent identically distributed (i.i.d.) this leads to

$$\begin{aligned}\mu_{S_i} &= E(S_i) = nE(X) \\ \sigma_{S_i}^2 &= \text{Var}(S_i) = n\text{Var}(X)\end{aligned}$$

#### Uniform

Using the above working this means if  $S_i$  is made of Uniform random variables  $U$  then

$$\begin{aligned}\mu_{S_i} &= E(S_i) = nE(U) = 10n \\ \sigma_{S_i}^2 &= \text{Var}(S_i) = n\text{Var}(U) = 8\frac{1}{3}n\end{aligned}$$

#### Exponential

Using the above working this means if  $S_i$  is made of Exponential random variables  $V$  then

$$\begin{aligned}\mu_{S_i} &= E(S_i) = nE(V) = 0.1n \\ \sigma_{S_i}^2 &= \text{Var}(S_i) = n\text{Var}(V) = 0.01n\end{aligned}$$

#### Binomial

Using the above working this means if  $S_i$  is made of Binomial random variables  $W$  then

$$\begin{aligned}\mu_{S_i} &= E(S_i) = nE(W) = 4n \\ \sigma_{S_i}^2 &= \text{Var}(S_i) = n\text{Var}(W) = 2.4n\end{aligned}$$

(c) For  $X_i$  either  $U_i$ ,  $V_i$  or  $W_i$ , define

$$\tilde{Z}_n = \frac{S_n - E(S_n)}{\sqrt{\text{var}(S_n)}}.$$

Use the CLT to postulate the distribution of  $\tilde{Z}$  for non-small  $n$ .

Calculating the expected value for  $\tilde{Z}$  we get

$$\begin{aligned} E(\tilde{Z}) &= E\left(\frac{S_n - E(S_n)}{\sqrt{\text{var}(S_n)}}\right) \\ &= \frac{E(S_n) - E(S_n)}{\sqrt{\text{var}(S_n)}} \\ &= 0 \end{aligned}$$

as  $E(S_n)$  and  $\text{var}(S_n)$  would be constants for this random variable. Similarly calculating the variance we get

$$\begin{aligned} \text{var}(\tilde{Z}) &= \text{var}\left(\frac{S_n - E(S_n)}{\sqrt{\text{var}(S_n)}}\right) \\ &= \frac{\text{var}(S_n)}{(\sqrt{\text{var}(S_n)})^2} \\ &= 1 \end{aligned}$$

So all of the variables  $\tilde{U}_n$ ,  $\tilde{V}_n$  and  $\tilde{W}_n$  have mean zero and variance 1. This holds for all  $n$ . As  $n$  increases the CLT states that this random variable gets closer to the standard Normal distribution.

(d) Generate Monte Carlo estimates of  $P(|\tilde{Z}_n| > 2.0)$  using no less than  $10^6$  generations of  $\tilde{Z}_n$  for every  $n$ , (separately for each  $U$ ,  $V$  or  $W$ ). Compare your results to  $P(|Z| > 2.0)$  taken from a normal distribution table, where  $Z$  is a standard normal variable. Do this for  $n = 2, 5, 10, 15, 20$ . Tabulate your results neatly and explain your results.

The following Julia code will produce the required Monte Carlo simulations with the distributions being in rows and the number of points increasing left to right.

```
In [1]: using Distributions

@everywhere uniDist = Uniform(5,15)
@everywhere expDist = Exponential(1/10)
@everywhere binDist = Binomial(10,0.4)

output=pmap(d->map(n->sum(abs.((sum(rand(d,n,10^7),1).-n*mean(d))./sqrt(
n*var(d))) .> 2.0)/10^7,[2,5,10,15,20]),
[uniDist,expDist,binDist])
```

```
Out[1]: 3-element Array{Array{Float64,1},1}:
 [0.0337003, 0.042986, 0.0442306, 0.0446393, 0.0448919]
 [0.0465868, 0.0412807, 0.0414771, 0.0424985, 0.0432004]
 [0.0370493, 0.0594968, 0.0517734, 0.0368522, 0.051071]
```

Now calculate the probability of  $P(|Z| > 2)$  from a standard normal distribution

```
In [2]: 2*ccdf(Normal(),2)
```

```
Out[2]: 0.04550026389635841
```

Comparing this value to the table above, it can be seen that as the number of items in the sample increases the closer the probability calculated is to the Normal distribution. The only potential exception to this is the Binomial distribution which is close to the value but oscillates around it.

## Question 2. Sample Mean and Sample Variance

Suppose that a sample of size  $n = 20$  is selected at random from a normal population with mean 100 and standard deviation 8. Let  $\bar{X}$  be the sample mean and  $S^2$  being the sample variance.

(a) Calculate  $P(98 \leq \bar{X} \leq 102)$ .

To calculate this probability we need to find the difference of the cdf between the upper and lower bounds. To do this we need to standardise to the standard normal distribution and then look up the values in a table.

$$\begin{aligned} P(98 \leq \bar{X} \leq 102) &= P(\bar{X} \leq 102) - P(\bar{X} \leq 98) \\ &= P\left(\frac{\bar{X} - 100}{\frac{8}{\sqrt{20}}} \leq \frac{102 - 100}{\frac{8}{\sqrt{20}}}\right) - P\left(\frac{\bar{X} - 100}{\frac{8}{\sqrt{20}}} \leq \frac{98 - 100}{\frac{8}{\sqrt{20}}}\right) \\ &= P(Z \leq 1.118034) - P(Z \leq -1.118034) \\ &= 0.8682 - 0.1318 \\ &= 0.7364 \end{aligned}$$

(b) Find  $x$  such that  $P(|\bar{X} - 100| > x) = 0.01$

Here again we need to standardise the distribution to the standard normal distribution and then look up this value in the tables, remembering that as we are looking for a two-sided probability that we need to multiply the value in the table by two.

$$\begin{aligned} P(|\bar{X} - 100| > x) &= 0.01 \\ P\left(\frac{|\bar{X} - 100| - 0}{\frac{8}{\sqrt{20}}} > \frac{x - 0}{\frac{8}{\sqrt{20}}}\right) &= P(|Z| > 2.5758) \\ \frac{x}{\frac{8}{\sqrt{20}}} &= 2.5758 \\ x &= \frac{8}{\sqrt{20}} \times 2.5758 \\ x &= 4.6077 \end{aligned}$$

(c) Use Monte-Carlo simulation with  $10^5$  samples to verify that  $E[S^2] = 64$ . Then estimate: (i)  $P(|S^2 - 64| > 2)$  (ii)  $\text{var}(S^2)$ .

First we need to generate  $10^5$  random samples from the distribution and then take the variance of each sample. From there we can calculate the mean and other probabilities.

```
In [3]: using Distributions
samplesq2= rand(Normal(100,8),10^5,20);
varsampq2= var(samplesq2,2) #,2 calculates by row.
mean(varsampq2)
```

```
Out[3]: 64.08247747939495
```

```
In [4]: #P(|S^2-64|>2)
mean(abs.(varsampq2.-64).>2)
```

```
Out[4]: 0.92324
```

```
In [5]: #var(S^2)
        var(varsampq2)
```

```
Out[5]: 431.45111273153066
```

From these calculations it can be seen that the variance of a random variable is a random variable itself.

### Question 3. Choice of Sample Size

A normal population has a mean 30 and standard deviation 5. How large must the random sample be if you want the standard error of the sample average to be less than 0.5? Verify your result using Monte-Carlo simulation.

$$\begin{aligned}
 s.e.(\bar{x}) &= \frac{s}{\sqrt{n}} \\
 0.5 &= \frac{5}{\sqrt{n}} \\
 0.25 &= \frac{25}{n} \\
 n &= \frac{25}{0.25} \\
 &= 100
 \end{aligned}$$

To verify this using Monte-Carlo simulation we need to create a function that creates random samples from the distribution, increasing in sample size at each step until the standard error is the value we are looking for. The code for this is below:

```
In [6]: using Distributions

function testDistribution(meanPop, stdPop, targetStdErr)
    dist = Normal(meanPop, stdPop);
    n = 1;
    stdErr = 10^6;
    while stdErr > targetStdErr && n<10^6
        n += 1;
        distSamp = rand(dist,n);
        stdSamp = std(distSamp,mean=meanPop);
        stdErr = abs(stdSamp)/sqrt(n);
    end
    return n;
end

mean([testDistribution(30,5,0.5) for _ in 1:10^6])
```

```
Out[6]: 78.245309
```

This value is slightly lower than expected due to sampling variability in the random samples which would be expected.

## Question 4. Polymer Elasticity

The elasticity of a polymer is affected by the concentration of a reactant. When low concentration is used, the true mean elasticity is 65, and when high concentration is used, the mean elasticity is 75. The standard deviation of elasticity is 6 regardless of concentration. If two random samples of size 25 are taken, find the probability that  $\bar{X}_{high} - \bar{X}_{low} > 2$ .

Although it is not explicitly stated, as the question refers to true means, assume the variables have normal distribution. Thus

$$\begin{aligned}\bar{X}_{high} &\sim N\left(75, \left(\frac{6}{\sqrt{25}}\right)^2\right) = N(75, 1.2^2) \\ \bar{X}_{low} &\sim N\left(65, \left(\frac{6}{\sqrt{25}}\right)^2\right) = N(65, 1.2^2)\end{aligned}$$

Combining these in the linear combination stated in the question we get

$$\Delta = \bar{X}_{high} - \bar{X}_{low} \sim N(75 - 65, 2 \times 1.2^2) = N(10, 2 \times 1.2^2)$$

This can be done as the samples are assumed to be independent of each other so no covariance term is required when adding the variances. We can now calculate the probability of the event required.

$$\begin{aligned}P(\Delta > 2) &= P\left(\frac{\Delta - 10}{\sqrt{2 \times 1.2^2}} > \frac{2 - 10}{\sqrt{2 \times 1.2^2}}\right) \\ &= P(Z > -4.71) \\ &= 0.9999\end{aligned}$$

## Question 5. Building up Confidence

For a normal population with known variance  $\sigma^2$ , answer the following questions:

(a) What is the confidence level for the interval  $\bar{x} - 2.1\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 2.1\sigma/\sqrt{n}$ ?

To work out the confidence level here, first note that the interval is of the form  $\bar{x} \pm z_{\frac{1-\alpha}{2}}^* \sigma/\sqrt{n}$ . This means we only

need to look up the  $z_{1-\frac{\alpha}{2}}^*$  value in the tables and then calculate  $P\left(|Z| > z_{\frac{1-\alpha}{2}}^*\right)$ . So the probability  $P(|Z| > 2.1)$  which is 0.9643

(b) What is the confidence level for the interval  $\bar{x} - 2.39\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 2.39\sigma/\sqrt{n}$ ?

Here the probability  $P(|Z| > 2.39)$  is 0.9832.

(c) What is the confidence level for the interval  $\bar{x} - 1.85\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1.85\sigma/\sqrt{n}$ ?

Here the probability  $P(|Z| > 1.85)$  is 0.9357.

(d) What is the confidence level for the interval  $\bar{x} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1.96\sigma/\sqrt{n}$ ?

Here the probability  $P(|Z| > 1.96)$  is 0.95.

We can check these results using the following Julia code (although it is advisable to look them up in a table)

```
In [7]: for z = [2.1,2.39,1.85,1.96]
        println("The confidence level of z*=",z," is ", 1-2*ccdf(Normal(0,1)
        ,z));
        end
```

```
The confidence level of z*=2.1 is 0.9642711588743669
The confidence level of z*=2.39 is 0.9831516272013087
The confidence level of z*=1.85 is 0.9356864504087726
The confidence level of z*=1.96 is 0.9500042097035591
```

## Question 6. Beverage Machine

A postmix beverage machine is adjusted to release a certain amount of syrup into a chamber where it is mixed with carbonated water. A random sample of 20 beverages was found to have a mean syrup content of  $\bar{x} = 0.9$  fluid ounce and a standard deviation of  $s = 0.011$  fluid ounce. Find a 99% CI on the mean volume of syrup dispensed. State any assumptions that are made.

First we assume that the amount of syrup dispensed has an underlying Normal distribution and that all observations are independent meaning we can use a  $T$ -Distribution. Using a  $T$ -Distribution with  $20 - 1 = 19$  degrees of freedom we can look up the t-value related to  $P(T > t) = \frac{1-\alpha}{2} = \frac{1-0.99}{2} = 0.005$  which is 2.861. Now we can construct the confidence interval as follows

$$\bar{x} - t_{\frac{1-\alpha}{2}}^* \sigma / \sqrt{n} \leq \mu \leq \bar{x} + t_{\frac{1-\alpha}{2}}^* \sigma / \sqrt{n}$$

$$0.9 - 2.861 \times 0.011 / \sqrt{20} \leq \mu \leq 0.9 + 2.861 \times 0.011 / \sqrt{20}$$

$$0.892963 \leq \mu \leq 0.907037$$

So the beverage machine dispenses between 0.892863 and 0.907037 on average with 99% confidence.

## Question 7. P-Value

For the hypothesis test  $\mathbf{H}_0 : \mu = 10$  against  $\mathbf{H}_1 : \mu > 10$  and variance known, calculate the P-value for each of the following test statistics:

(a)  $z = 2.05$

Similar to Question 5 we just need to look up the probability  $P(Z > z)$  from the tables. Doing this we get that the P-value is  $P(Z > 2.05) = 0.0202$

(b)  $z = -1.84$

The P-value in this case is  $P(Z > -1.84) = 0.9671$ .

(c)  $z = 0.4$

The P-value in this case is  $P(Z > 0.4) = 0.3446$ .

(d)  $z = 0$

The P-value in this case is  $P(Z > 0) = 0.5000$ .

(e)  $z = -2.05$

The P-value in this case is  $P(Z > -2.05) = 0.9798$ .

(f)  $z = 3$

The P-value in this case is  $P(Z > 3) = 0.0013$ .

Now repeat (a)-(f) when the alternative hypothesis is  $\mathbf{H}_1 : \mu \neq 10$ .

Now we have to calculate  $P(Z \neq z) = 2 \times \min(P(Z < z), P(Z > z))$  as the test is now two sided.

(a)  $z = 2.05$

Doing this we get that the P-value is  $P(|Z| > |2.05|) = 0.0404$

(b)  $z = -1.84$

The P-value in this case is  $P(|Z| > |-1.84|) = 0.0658$ .

(c)  $z = 0.4$

The P-value in this case is  $P(|Z| > |0.4|) = 0.6892$ .

(d)  $z = 0$

The P-value in this case is  $P(Z > 0) = 1.0000$ .

(e)  $z = -2.05$

The P-value in this case is  $P(Z > -2.05) = 0.0404$ .

(f)  $z = 3$

The P-value in this case is  $P(Z > 3) = 0.0026$ .

This can be checked with the following Julia code



```
In [8]: for z=[2.05,-1.84,0.4,0,-2.05,3]
        println("P(Z>",z,")= ",round(ccdf(Normal(0,1),z),4)," , P(Z!=",z,")= ",
        round(2*ccdf(Normal(0,1),z),4))
        end

P(Z>2.05)=0.0202, P(Z!=2.05)=0.0404
P(Z>-1.84)=0.9671, P(Z!=-1.84)=1.9342
P(Z>0.4)=0.3446, P(Z!=0.4)=0.6892
P(Z>0.0)=0.5, P(Z!=0.0)=1.0
P(Z>-2.05)=0.9798, P(Z!=-2.05)=1.9596
P(Z>3.0)=0.0013, P(Z!=3.0)=0.0027
```

## Question 8. Sodium Content in Organic Cornflakes

The sodium content of twenty 300-gram boxes of organic cornflakes was determined. The data (in milligrams) is contained in (9-65.csv).

(a) Can you support a claim that mean sodium content of this brand of cornflakes differs from 120 milligrams? use  $\alpha = 0.05$ , state your hypothesis clearly, and the P-value and make a conclusion.

First so that we can calculate this hypothesis test we need to calculate the mean and variation for the data set.

```
In [9]: using DataFrames
        cornflake = readtable("9-65.csv",header=false)
        cornflake = cornflake[1] #To save dealing with indexes.

        meancorn = mean(cornflake)
        varcorn = var(cornflake)
        lencorn = length(cornflake)

        println("The mean of the cornflakes is ",meancorn)
        println("The variance of the cornflakes is ",varcorn)
        println("There are ", lencorn, " data points.")

        The mean of the cornflakes is 129.747
        The variance of the cornflakes is 0.7681273684210458
        There are 20 data points.
```

Now, we state the hypotheses

$H_0$  : The mean sodium content of this brand of cornflakes is equal to 120 milligrams.

$H_1$  : The mean sodium content of this brand of cornflakes is not equal to 120 milligrams

We now calculate the test statistic from a T-distribution with 19 (20-1) degrees of freedom. This assumes that the data has an underlying normal distribution and that the observations are independent. The test statistics is calculate as

$$\begin{aligned} t_{19} &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{129.747 - 120}{\sqrt{\frac{0.768...}{20}}} \\ &= 49.735... \end{aligned}$$

This is a very large t-value, so looking up the probability in the tables with 19 degrees of freedom give that  $P(|T| > t) < 0.0002$ . If we want an exact value we need to use software and we get,

```
In [10]: using Distributions
        2*ccdf(TDist(19),(mean(cornflake)-120)/sqrt(var(cornflake)/20))

Out[10]: 1.3747603780291543e-21
```

From this p-value we conclude that there is extremely strong evidence to support the alternative hypothesis, therefore we conclude that the mean sodium level of this brand of cornflakes is significantly different from 120 milligrams.

To use Julia to compute this hypothesis test we would run the following commands.

```
In [11]: using HypothesisTests
          OneSampleTTest(cornflake,120)

Out[11]: One sample t-test
-----
Population details:
  parameter of interest:  Mean
  value under h_0:       120
  point estimate:        129.747
  95% confidence interval: (129.33681871490268, 130.15718128509735)

Test summary:
  outcome with 95% confidence: reject h_0
  two-sided p-value:          1.3747603780291443e-21

Details:
  number of observations:  20
  t-statistic:             49.73582705870276
  degrees of freedom:      19
  empirical standard error: 0.1959754281052915
```

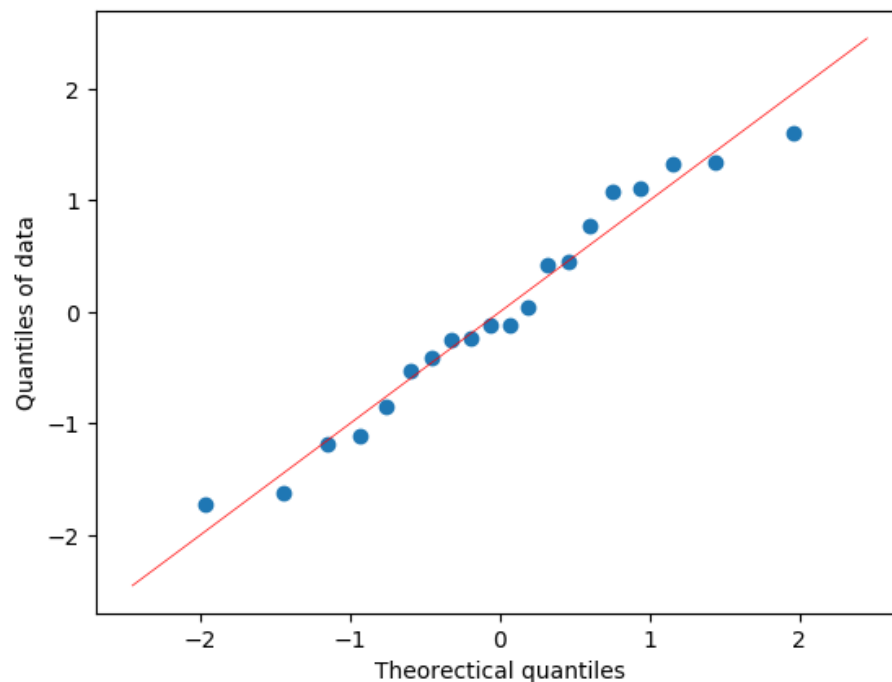
(b) Check that sodium content is normally distributed (e.g. using the code for Normal probability plots from Assignment 3).

Using the code from Assignment 3 we get the following plot

```
In [12]: 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

NormalProbabilityPlot(cornflake)
```



Looking at this plot, there are no strong deviations from the line and thus normality. With a small sample such as this this looks reasonable to assume that the assumption of the Normal Distribution is satisfied.

(c) Compute the power of the test if the true mean sodium content is 130.5 milligrams.

Given that we assume that the true mean is 120 milligrams it is highly likely that a mean that is 10.5 milligrams away will be detected as the variance is less than 1. This is a badly designed question and apologies are made for this. There are many ways to compute the power of this test, however using Monte-Carlo methods we can use the following code:

```
In [13]: using Distributions

function tStatisticUnderH1(testMean,n)
    data= rand(Normal(testMean,std(cornflake)),n);
    xBar= mean(data);
    s= std(data);
    tStatistic = (xBar - 120)/(s/sqrt(length(data)));
    return tStatistic
end

mean([abs(tStatisticUnderH1(130.5,20)) > 2.093 for _ in 1:10^6])

Out[13]: 1.0
```

As can be seen this gives a power of near 1 meaning the test will be able to detect this difference with high probability.

(d) What sample size would be required to detect a true mean sodium content of 130.1 milligrams if you wanted the power of the test to be at least 0.75? Explain your answer.

Given that the test value is still a large distance from the true mean of 120 milligrams the sample size will be quite small. Again if we run the function above by this time iterate through different samples sizes we can find the correct sample size.

```
In [14]: [(n,mean([abs(tStatisticUnderH1(130.1,n)) > quantile(TDist(n-1),0.975) f
or _ in 1:10^6])) for n in 2:3]

Out[14]: 2-element Array{Tuple{Int64,Float64},1}:
 (2, 0.799633)
 (3, 1.0)
```

So at least 2 samples would be required to detect a true mean content of 130.1 milligrams at a power of at least 0.75.

(e) Explain how the question in part (a) could be answered by constructing a two-sided confidence interval on the mean sodium content.

If we construct a two sided confidence interval we can see if the true mean lies within its bounds to determine if it is a plausible value. If it does not the mean of the sample is significantly different to that of the true mean. Here the confidence interval would be

```
In [15]: confint(OneSampleTTest(cornflake,120))

Out[15]: (129.33681871490268, 130.15718128509735)
```

As 120 milligrams is not within the confidence interval we can say that the sample is significantly different.