# Computer practical 11

Solutions

*Statistics 1, Term 2*

Aim

- To learn how to implement statistical methods in R
  - To generate confidence intervals for a population mean
  - To validate assumptions underlying their use through inspecting the normality of the data
  - To employ transformations to make the data look more normal.

---

## Statistical techniques

Suppose *data* is a vector of data.

### Normal Quantile plots

Type:

```
> qqnorm(data)
```

### Overlaying the normal shape on a histogram

This is a bit more difficult. The following are general instructions. Read them and then try and use them later on. Type:

```
> hist(data, freq=FALSE)
> curve(dnorm(x, mean=mean(data), sd=sd(data)), add=T)
```

For more information, read the R help for the individual functions (hist, dnorm, curve).

### Obtaining two-sided confidence intervals

If you are not yet familiar with qnorm and qt, read their documentation.

Suppose alpha is the level of significance.

- If $\sigma$ is known, and n is large or we're sampling from normal, we must calculate

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}.$$

  Suppose sigma contains the value of $\sigma$. Type:

```
> mean(data)+qnorm(1-alpha/2)*sigma/sqrt(length(data))
> mean(data)-qnorm(1-alpha/2)*sigma/sqrt(length(data))
```

to get the confidence interval. The p-value of

$$H_0 : \mu = m$$

$$H_a : \mu \neq m$$

is calculated via

```
> 2*(1-pnorm(abs(mean(data)-m)/(sigma/sqrt(length(data)))))
```

- If $\sigma$ is unknown, but n is large, then simply replace $\sigma$ by the sample standard deviation s (sd(data)).
- If $\sigma$ is unknown, $n$ is small, and we are sampling from a normal distribution (use a normal quantile plot to check this!), then in addition to replacing $\sigma$ by sd(data), also use the t distribution with n-1 degrees of freedom, instead of the normal distribution:

  ```
  > mean(data)+qt(1-alpha/2, length(data)-1)*sd(data)/sqrt(length(data))
  > mean(data)-qt(1-alpha/2, length(data)-1)*sd(data)/sqrt(length(data))
  > 2*(1-pt(abs(mean(data)-m)/(sd(data)/sqrt(length(data))), + length(data)-1))
  ```

  Another way of obtaining a 95% confidence interval for a population mean based on a vector data is as follows:

  ```
  > confint(lm(data~1))
  ```

  What's actually happening is we are exploiting the regression facilities to fit a model which only has an intercept, i.e. the population mean. We can vary the confidence level by adding the optional level argument, as in

  ```
  > confint(lm(data~1), level=.99)
  ```

# Applications

## Dorsal lengths data

Dataset: octopod {durham}

Load the octopod data. The data are the dorsal lengths (in millimetres) of taxonomically distinct octopods. These data were also explored during computer practical 4.

Find a 95% confidence interval for the mean dorsal length. Write down the mean and standard error.

```
# LOAD THE PACKAGE AND THE DATA
library(durham)
data(octopod)
```

```
summary(octopod)
```

```
##      dorsal
##  Min.   :  5.00
##  1st Qu.: 19.00
##  Median : 32.50
##  Mean   : 44.67
##  3rd Qu.: 59.25
##  Max.   :190.00
```

- Hmmm. . . , the population variance is unknown, so I could use the confidence interval based on T-statistic. I can use the asymptotic (approximate) sampling distribution for T which is the Normal distribution because the sample size is large $n = 94$.

```
x <- octopod$dorsal
alpha <- 0.05
sigma <- sd(x)
n = length(x)
LL <- mean(x) -qnorm(1-alpha/2 )*sigma/sqrt(length(x))
UU <- mean(x) +qnorm(1-alpha/2 )*sigma/sqrt(length(x))
x.mean <- mean(x)
x.se <- sd(x)/sqrt(length(x))
```

- The 95 % (approximate) Confidence interval for the mean is (37.3801869, 51.9602386). The mean is 44.6702128 and the standard error is 3.7194693.
- Alternativelly, if I had condired the Student t as the sampling distribution, and hece the t-quantiles for the construvtion of the CI:

```
x <- octopod$dorsal
alpha <- 0.05
sigma <- sd(x)
n = length(x)
LL <- mean(x) -qt(1-alpha/2 , df=n-1)*sigma/sqrt(length(x))
UU <- mean(x) +qt(1-alpha/2 , df=n-1)*sigma/sqrt(length(x))
x.mean <- mean(x)
x.se <- sd(x)/sqrt(length(x))
```

- The 95 % Confidence interval (T statistic and t sampling distributon) for the mean is (37.2840839, 52.0563416). The mean is 44.6702128 and the standard error is 3.7194693. Note that, also:

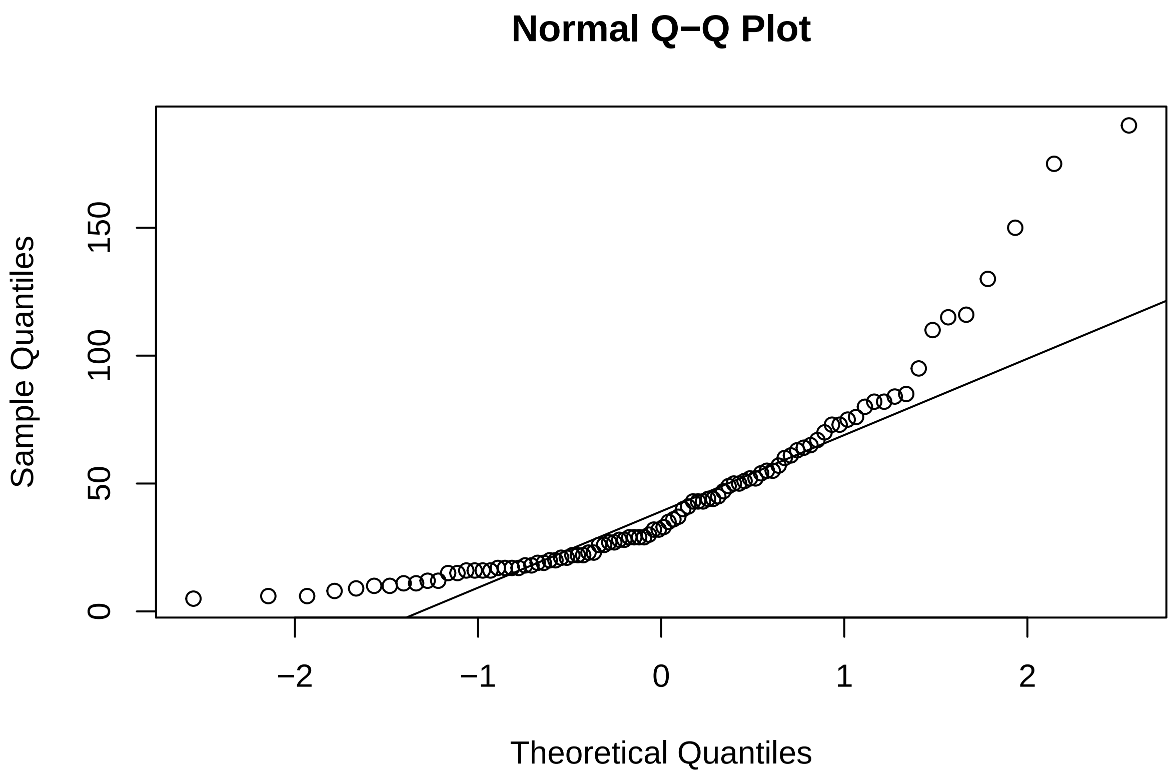
```
# this by default uses as sampling distribution the t distribution giving wider intervals:
confint(lm(x~1),level=1-alpha)
```

```
##                2.5 %   97.5 %
## (Intercept) 37.28408 52.05634
```

Assess the Normality of the distribution. Write down your interpretations.

- The q-q plot to assess if the data were drawn from a Normal distribution is presented below.

```
# INSERT YOUR CODE HERE ...
qqnorm(x)
qqline(x)
```



- The qq plot indicates significant evidence that the sample was not drawn from a Normally distributed population, eg. we can see evidence that the distribution may be right skewed.

Clearly, the octopod data are heavily skewed with a long tail to the right. What is the implication of this for the confidence interval you calculated above?

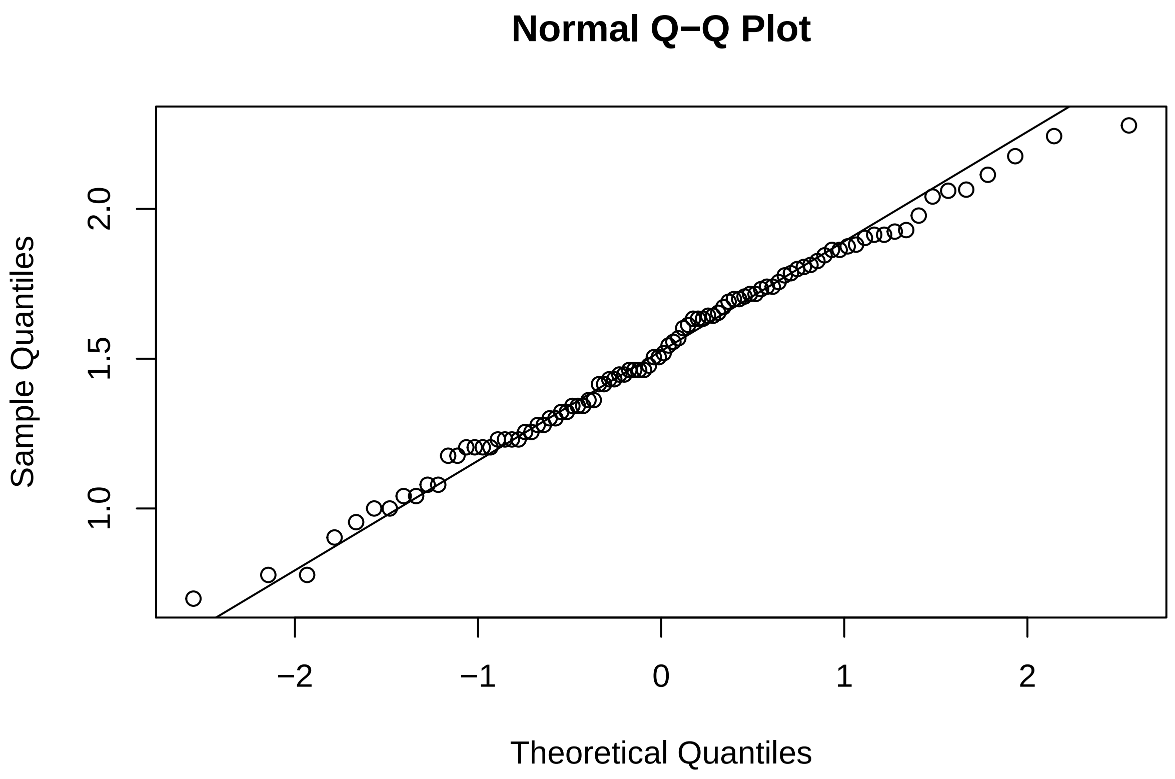
- If we accept that the sample size is large enough, (n=94) then due to the Central Limit Theorem, the (approximate) CI (using the Normal quantiles) is valid. The CI using the t quantiles, of course gives wider intervals that the one using Normal quantiles (aka it is more conservative). The reason is because t- distribution has havier tails than the Normal distribution… However, once again these CI are approximations, thery are not exact.
- We cannot say that the CI (statistic T and sampling distribution t) is a valid exact CI. This is because, one of the assumptions in order for this CI to be valid is that the sample is drawn from a Normally distributed population. This assumption is possibly violated here, according to the evidence from the QQ-plot.

For heavily skewed positive data, a logarithmic transformation often works well. Assess the normality for the logarithm (base 10, use log10) of the octopod data.

Has the transformation made the distribution more Normal in shape?

- Let's re-state this… 'Can we assume that the dorsal length (in $log_{10}$ scale), as resulted after applying a $log_{10}$ transformation to the original data, is Normally distributed?'

```
# INSERT YOUR CODE HERE ...
x.log <- log10(x)
qqnorm(x.log)
qqline(x.log)
```



- The QQ-plot of the log sample values does not indicate serious evidence against Normality. The Normality assumption of the t-test, for the data in $\log_{10}$ scale is not violated, so we can perform the test cautiously.

Find a 95% confidence interval for the mean of the logarithm to base 10 of dorsal length.

- We use the Confidence Interval based on the T statistic with sampling distribution the t distribution, on the $\log_{10}$ scaled data.

```
# INSERT YOUR CODE HERE ...
library(durham)
data(octopod)
x <- log10(octopod$dorsal)
alpha <- 0.05
sigma <- sd(x)
n = length(x)
LL <- mean(x) -qt(1-alpha/2, df=n-1)*sigma/sqrt(length(x))
UU <- mean(x) +qt(1-alpha/2, df=n-1)*sigma/sqrt(length(x))
x.mean <- mean(x)
x.se <- sd(x)/sqrt(length(x))
```

- The 95 % Confidence interval for the mean of the dorsal lengths (in $\log_{10}$ millimeters) is (1.4537683, 1.5934842). The mean is 1.5236263 and the standard error is 0.0351787.

What is the advantage of base 10 logarithms over natural logarithms when analysing data?

- Maybe $\log_{10}(\cdot)$ is easier show the order of the values transformed. . . , e.g. $\log_{10}(0.01) = \log_{10}(1 \times 10^{-2}) =$ -2.
- We are more familiar with orders of 10 rather than e.g. order of $e$. . . . .
- On the other hand, perhaps $\log_e(\cdot)$ is more convenient to use when we deal with equations involving $\exp(\cdot)$; e.g. the PDF of the Normal distribution.

### Cherry tree volumes data

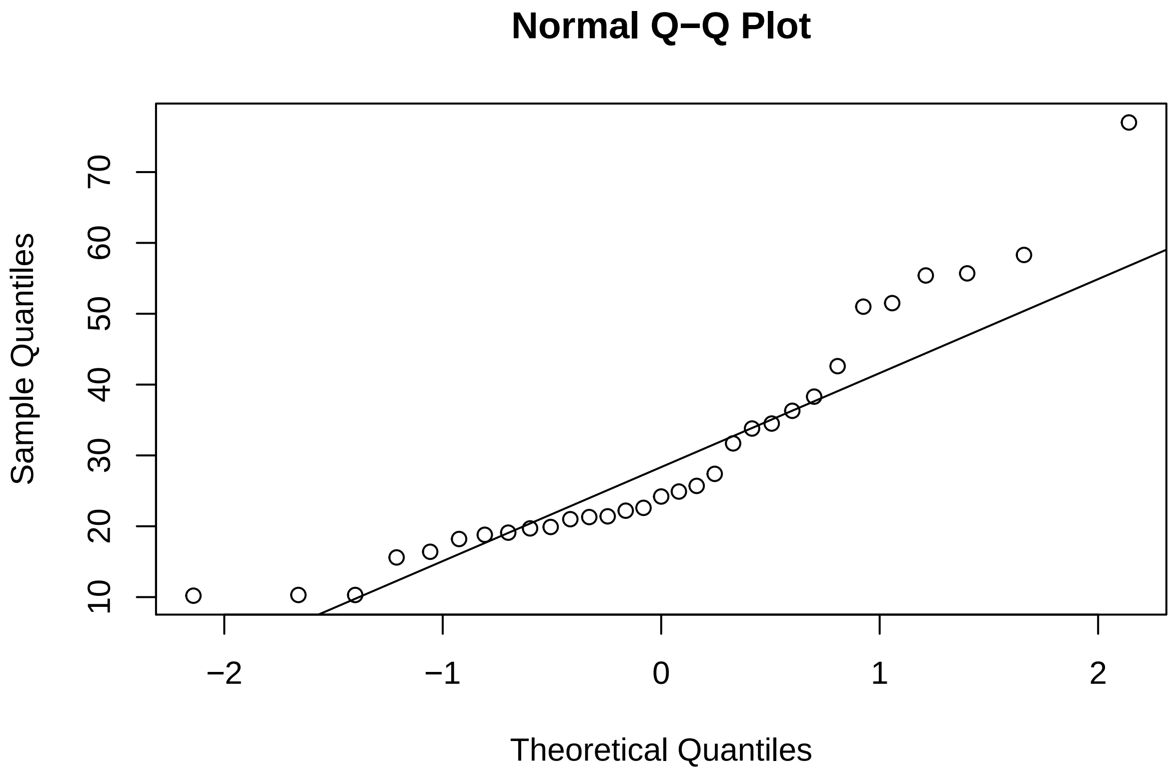
Dataset: *trees* {datasets}

The trees data consists of heights, diameters and volumes of cherry trees. We will ignore the heights and diameters.

Assess the Normality of the volumes. Write down your interpretations.

```
data(trees)
```

```
x <- trees$Volume
qqnorm(x)
qqline(x)
```



- The qq-plot suggest strong evidence that the cherry tree volume does not not follow a Normal distribution because the point marks deviate systematically from the straight line.

Find a 95% confidence interval for the mean cherry tree volume.

- I will use the approximating CI based on T statistics, and Normal sampling distribution.
  - The population variance is unknown, so I use the T statistic.
  - The Normality assumption is possibly violated. The QQ-plot suggests strong evidence that the sample is not drawn from a Normally distributed population. However, the sample size is rather large n=31, so I can use the approximating CI based on the Normal sampling distribution.

```
x <- trees$Volume
LL <- mean(x) -qnorm(1-alpha/2)*sigma/sqrt(length(x))
UU <- mean(x) +qnorm(1-alpha/2)*sigma/sqrt(length(x))
x.mean <- mean(x)
x.se <- sd(x)/sqrt(length(x))
```

- The 95 % Confidence interval for the mean of the tree volume is (30.0509042, 30.2910313). The mean is 30.1709677 and the standard error is 2.9523244.

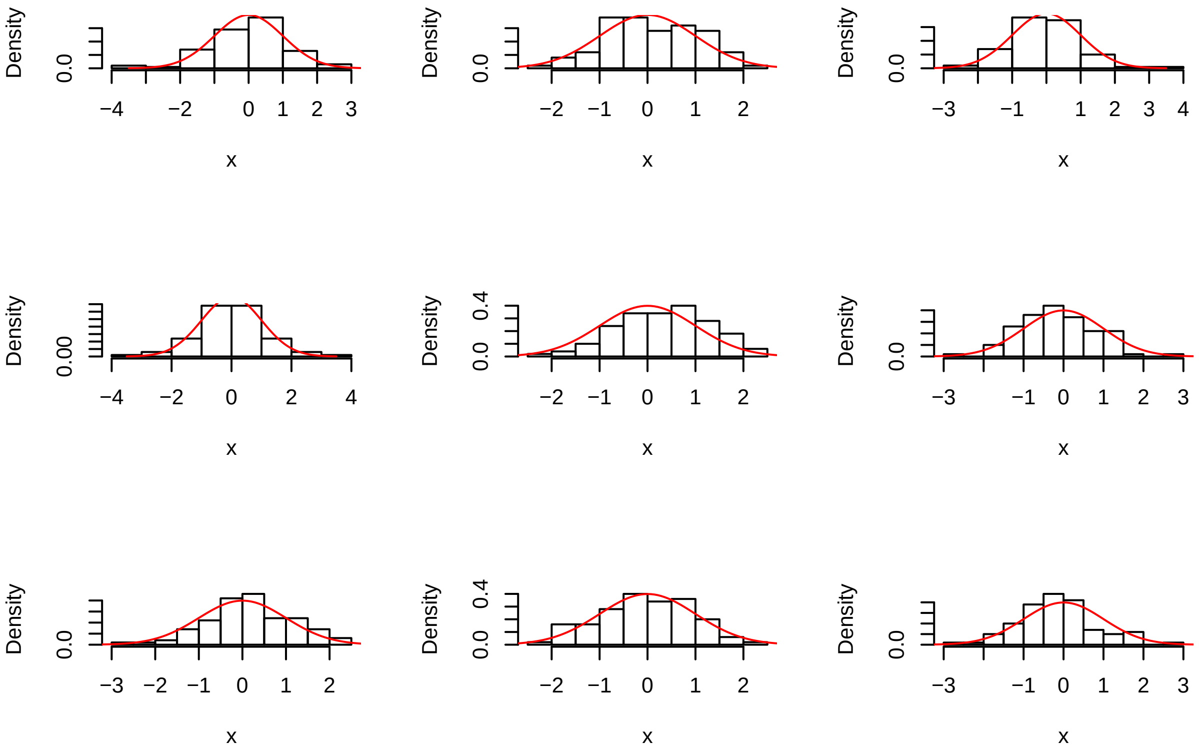
## Sampling distributions

It's easy to take samples from distributions in R. For a finite list of numbers, one can use the sample function. For standard distributions, R has special functions which you can use. For the normal distribution, function is called rnorm and to take a sample of size n from a normal distribution with mean mu and standard deviation sigma, one types

```
> rnorm(n, mu, sigma)
```

Try taking a sample of size 30 from the standard normal. Repeat the command to see that the answer changes each time.

```
par(mfrow=c(3,3))
for (i in 1:9) {
  n <- 100
  mu <- 0
  sigma <- 1
  x <- rnorm(n, mu, sigma)
  hist(x, main=' ', freq=FALSE)
  myfun<-function(y,a=mu, b=sigma) {
                  dnorm(y, mean=a, sd=b)
                  }
  curve(myfun, from=mu-3.5*sigma, to=mu+3.5*sigma, add=T, col='red') # look up curve in help
}
```



To study sampling distributions, one needs to take multiple samples. An easy way to do this is to take a single large sample and then arrange it as a matrix. For example, to take 10000 samples of size 15 from the normal distribution with 5 and standard deviation 2, one sample in each row of the matrix, type

```
> samples = matrix(rnorm(10000*15, 5, 2), nrow=10000)
```

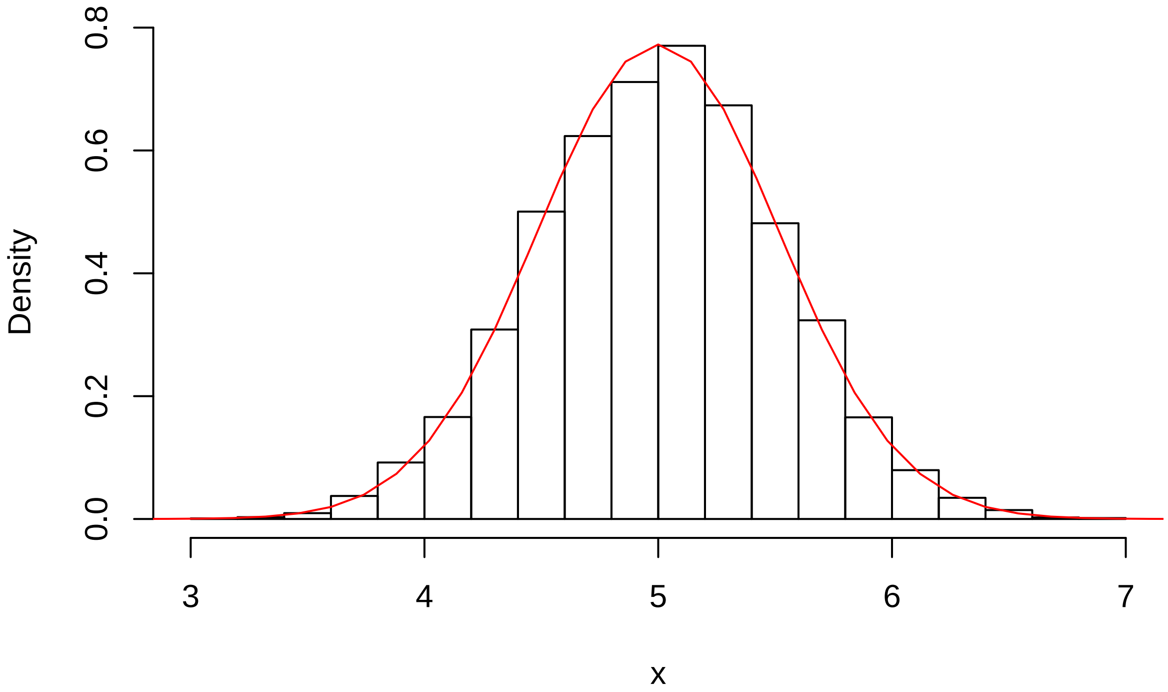
Now one can study the sampling distribution. One can obtain the mean of each sample by typing:

```
> means = apply(samples, 1, mean)
```

Produce a histogram of the means. What distribution should it have?

- Possibly a distribution with PDF close to that of a Normal distribution with mean 5 and standard deviation 0.5163978.

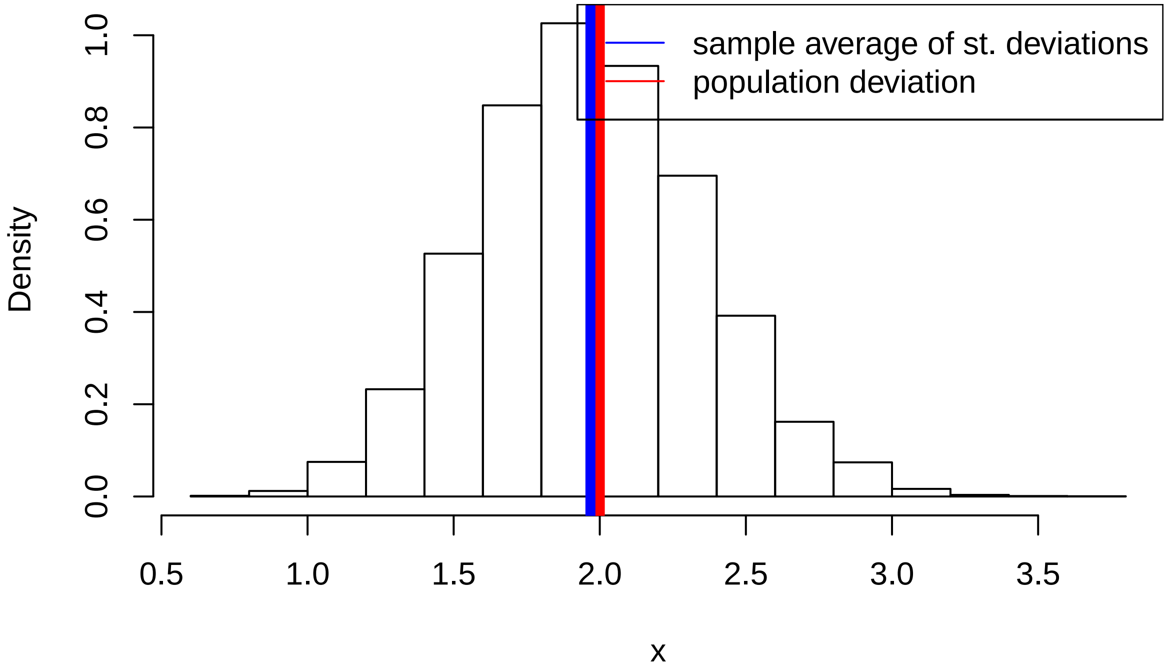
```
mu = 5
sigma = 2
m = 15
n = 10000
samples = matrix(rnorm(n*m, mu, sigma), nrow=n)
means = apply(samples, 1, mean) # look up apply in help
x <- means
hist(x, main=' ', freq=FALSE)
myfun<-function(y,a=mu, b=sigma/sqrt(m)) {
                  dnorm(y, mean=a, sd=b)
                  }
curve(myfun,from=mu-3.5*sigma, to=mu+3.5*sigma, add=T, col='red')
```



Produce a histogram of the standard deviations. Describe the distribution. How much variation is there relative to the original population standard deviation? How does the mean of these standard deviations compare to the population standard deviation?

```
sigma_vec = apply(samples, 1, sd)
x <- sigma_vec
hist(x, main=' ', freq=FALSE)
abline(v=sigma, col='red', lwd=5)
```

```r
abline(v=mean(x), col='blue', lwd=5)
legend('topright',
       legend=c('sample average of st. deviations', 'population deviation'),
       col=c('blue','red'),
       lty=c(1,1))
```



- The sample average of the sample standard deviation is close to the population standard deviation.

How can you obtain all 10000 confidence intervals?

```r
mu = 5
sigma = 2
m = 15
n = 10000
samples = matrix(rnorm(n*m, mu, sigma), nrow=10000)
x <- samples
# make a function that returns the lower and upper intervals
alpha = 0.05
myfun <- function(y,a=alpha) { confint(lm(y~1),level=1-a) }
# evaluate this function at each single value
cis = apply(x, 1, myfun) #
```

What proportion of the confidence intervals contain the population mean? How does it compare to the theoretical expected proportion?

```r
LLv <- cis[1,]
UUv <- cis[2,]
# Find the number of occurances where mu is suvh that mu>L and mu<U
Noccur <- sum(( mu > LLv ) * ( mu < UUv ))
Ntotal <- n # the total number of CI
PropOccur <- Noccur/Ntotal
```

- It is 95.12%. It was expected to be around 95% because we have evaluated the 95% Confidence Intervals.

Now repeat all of this exercise using a uniform distribution (see the R help for the *runif* function). To what extent do you expect the same outcomes and to what extent do you expect them to differ?

```
m = 15
n = 10000
samples = matrix(runif(n*m, 0, 1), nrow=10000)
x <- samples
# make a function that returns the lower and upper intervals
alpha = 0.05 # state the level of significance
myfun <- function(y,a=alpha) { confint(lm(y~1),level=1-a) }
# evaluate this function at each value
cis = apply(x, 1, myfun) # compute the CI
LLv <- cis[1,]
UUv <- cis[2,]
mu = 0.5 # the mean of U(0,1) is (1-0)/2
# Find the number of occurances where mu is suvh that mu>L and mu<U
Noccur <- sum(( mu > LLv ) * ( mu < UUv ))
Ntotal <- n # the total number of CI
PropOccur <- Noccur/Ntotal
```

- It is 94.83%. It was expected to be around 95% because we have evaluated the 95% Confidence Intervals.

## Save me

Generate the document as a Notebook, PDF, Word, or HTML by choosing the relevant option (from the pop-up menu next to the Preview button). Then save your Markdown code by choosing the relevant option (from the task bar menu).

Save the *.Rmd script, so that you can edit it later.