Chapter 3: Confidence intervals and hypothesis testing

3.1 a) $N = 4513$ $s = 25$
$n = 100$ $\sigma =$ unknown
$\bar{x} = 145$ $\mu =$ unknown

3.1 b) $\mu =$ 145 seconds

Explanation: The best estimate for the population mean μ is the sample mean $\bar{x}$.

3.1 c) $SE_\mu = \frac{s}{\sqrt{n}} = \frac{25}{\sqrt{100}} = 2.5$

3.1 d) z-value: 2.576

Explanation: In table 2 of the formula sheet, the cumulative probability in that lies the closest to 0.995 (split the risk over two sides) is 0.9949. That value can be found at a z-value of 2.576.

Lower bound: $\bar{x} - z_{0.995} \times SE_\mu = 145 - 2.567 \times 2.5 = 138.56$
Upper bound: $\bar{x} + z_{0.995} \times SE_\mu = 145 - 2.567 \times 2.5 = 151.44$

3.1 e) $\mu_0 = 150$ (the to be tested limit of 150 seconds for the actual population mean call duration).

$H_0 : \mu \geq \mu_0$ $H_0 : \mu < \mu_0$

3.1 f) Upper bound: $\bar{x} + z_{0.99} \times SE_\mu = 145 - 1.645 \times 2.5 = 149.11$

3.1 g) The upper bound of the confidence interval for μ is lower than μ_0. H_0 is rejected with 99% confidence. μ is shown to be significantly lower than 150 seconds. There is a risk of 5% for a type-I error.

3.2 a)

i	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	3.03	-0.0863	0.0074
2	3.45	0.3338	0.1114
3	3.94	0.8238	0.6786
4	2.34	-0.7763	0.6026
5	3.34	0.2238	0.0501
6	2.53	-0.5863	0.3437
7	2.88	-0.2363	0.0558
8	3.42	0.3038	0.0923

$\sum x_i = 24.93 \qquad \sum (x_i - \bar{x})^2 = 1.9418$

$\bar{x} = 3.116 \qquad s^2 = 0.277$

3.2 b) Hartley's F: $\frac{s^2_{max}}{s^{min}} = \frac{1.113}{0.227} = 4.018$

3.2 c) $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \qquad H_0 : \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$

3.2 d) Hartley's F_{max}: 6.94

3.2 e) The value $F = 4.018$ is lower than critical value $F_{max} = 6.94$. H_0 is not rejected. There is no indication the variance for these months is not homogeneous. There is a risk of a type-II error.

3.3 a)

```
# Be sure to set your working directory when providing a relative path
dataset5 <- read.csv('populations.csv')
```

3.3 b) The code creates four random samples of size 90 from the columns **P1** - **P4** in the data frame called **dataset5** . The seed makes sure that you can recreate the same random samples again, so that if you close R and continue tomorrow we get the same samples.

```
set.seed(54321) # You can replace 54321 with your own seed number

sample1 <- sample(dataset5$P1, size = 90)
sample2 <- sample(dataset5$P2, size = 90)
sample3 <- sample(dataset5$P3, size = 90)
sample4 <- sample(dataset5$P4, size = 90)
```

3.3 c) Standard error sample 1: 29.24
Standard error sample 2: 7.64
Standard error sample 3: 30.61
Standard error sample 4: 18.09

```
# Means
x1 <- mean(sample1) # Mean: 456.78
x2 <- mean(sample2) # Mean: 511.02
x3 <- mean(sample3) # Mean: 790.32
x4 <- mean(sample4) # Mean: 533.37

# Standard deviations
sd1 <- sd(sample1) # Standard deviation: 277.38
sd2 <- sd(sample2) # Standard deviation: 72.43
sd3 <- sd(sample3) # Standard deviation: 290.46
sd4 <- sd(sample4) # Standard deviation: 171.58

# Standard errors se = sd / sqrt(n)
se1 <- sd1 / sqrt(length(sample1)) # Standard error: 29.24
se2 <- sd2 / sqrt(length(sample2)) # Standard error: 7.64
se3 <- sd3 / sqrt(length(sample3)) # Standard error: 30.61
se4 <- sd4 / sqrt(length(sample4)) # Standard error: 18.09
```

3.3 d) The value 1.644854 comes from the standard normal distribution with $\mu = 0$ and $\sigma = 1$. This is the z-value for a 95% one-sided confidence interval (or a 90% two-sided confidence interval).

3.3 e)

```
# Gives the left-tailed probability (z-value) for 95% confidence
qnorm(p = 0.95) # 1.645
```

3.3 f) In a 95% confidence interval there is 2.5% of the risk at the lower bound and 2.5% of the risk at the upper bound. You can therefore use the 97.5% quantile of the standard normal distribution to get the z-value for a two-sided 95% confidence interval and use it for both upper and lower bound.

```
z <- qnorm(p = 0.975)
```

3.3 g)

```
# Lower bounds
lb1 <- x1 - z * se1 # Lower bound: 399.48
lb2 <- x2 - z * se2 # Lower bound: 496.06
lb3 <- x3 - z * se3 # Lower bound: 730.31
lb4 <- x4 - z * se4 # Lower bound: 497.92

# Upper bounds
ub1 <- x1 + z * se1 # Upper bound: 514.10
ub2 <- x2 + z * se2 # Upper bound: 525.99
ub3 <- x3 + z * se3 # Upper bound: 850.33
ub4 <- x4 + z * se4 # Upper bound: 568.81
```

3.3 h)

	sample1	sample2	sample3	sample4
ub	514.10	525.99	850.33	568.81
x	456.78	511.01	790.32	533.37
lb	399.48	496.06	730.31	497.92

	P1	P2	P3	P4
mu	495.54	500.08	748.96	556.43
mu in interval?	YES	YES	YES	YES

```
# Population means
mu1 <- mean(dataset5$P1) # Mean: 495.54
mu2 <- mean(dataset5$P2) # Mean: 500.08
mu3 <- mean(dataset5$P3) # Mean: 748.96
mu4 <- mean(dataset5$P4) # Mean: 556.43
```

3.3 i) Yes, in this case all population means are inside the intervals.

3.3 j) The population mean will fall inside the interval 95 out of a 100 times (95% confidence). This means that about 1 in 20 confidence intervals will not have the true population mean between their lower and upper bound.

3.4 a)

```
# install.packages('car')
library(car)

# This create a 2x2 layout
layout(matrix(c(1, 2, 3, 4), byrow = TRUE, nrow = 2))

hist(sample1, col = 'gray')
hist(sample2, col = 'gray')
hist(sample3, col = 'gray')
hist(sample4, col = 'gray')

# This resets the layout to the default (1 plot)
layout(1)
```

3.4 b) `Sample 2 looks like it might come from a normal distribution.`

3.4 c)

```
# This create a 2x2 layout
layout(matrix(c(1, 2, 3, 4), byrow = TRUE, nrow = 2))

qqPlot(sample1, distribution = 'norm')
qqPlot(sample2, distribution = 'norm')
qqPlot(sample3, distribution = 'norm')
qqPlot(sample4, distribution = 'norm')

# This resets the layout to the default (1 plot)
layout(1)
```

3.4 d) The **sample1** histogram looks normal in the middle, but has too many low and too many high values. This can be seen in the qqplot by the dots on the left of the diagonal at the bottom and the dots on the right at the top.

The **sample2** histogram looks normal so the dots in the qqplot are almost everywhere on the diagonal. Only in the right part of the middle the histogram frequency is a bit too low, this is reflected in the dots below the diagonal around zero in the qqplot.

The **sample3** histogram is too high on the sides (or too low in the middle) to be normal. This can be seen in the qqplot by the dots on the left of the diagonal at the bottom and the dots on the right at the top.

sample4 is negatively skewed. This can be seen in the qqplot from the arch shape; the left of the histogram is too low causing the dots on the right of the diagonal and the right of the histogram is too high also causing dots to the right of the diagonal.

3.4 e) H_0: The sample is normally distributed
H_1: The sample is not normally distributed

3.4 f) Samples 1, 3, and 4 are not normally distributed, since the p-value is lower than 1% (for 99% confidence). Sample 2 is normally distributed, since the p-value is higher than 1%.

```
shapiro.test(sample1) # p-value: 0.0022
shapiro.test(sample2) # p-value: 0.949
shapiro.test(sample3) # p-value: 0.0068
shapiro.test(sample4) # p-value: 0.0001
```

3.4 g) When the sample is not normally distributed, the sample can be used to estimate the population mean.

Explanation: This method to estimate the population mean assumes the distribution of sample means to be normally distributed. It does not assume a normally distributed sample. The Central Limit Theorem states that any sample large enough ($n \geq 30$) will have a normal distribution of sample means, so you can use this method here ($n = 90$) without problems.

3.5 a)

```
library(car)
data(iris)

plot(x = iris$Species, y = iris$Sepal.Length,
    col = 'grey', main = 'Sepal Length')

plot(x = iris$Species, y = iris$Sepal.Width,
    col = 'grey', main = 'Sepal Width')
```

3.5 b) Looking at the width of the range and quartile ranges: For sepal length the variance for setosa looks much smaller than for the other two species, for sepal width all variances look similar.

3.5 c) $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ $H_0 : \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$

3.5 d) The p-value for the sepal length is lower than 10%. This implies that H_0 is rejected. This means that the variance of the sepal length over the species is not homogeneous. There is a 5% change of a type-I error.

The p-value for the sepal with is not lower than 10%. This implies that H_0 is not rejected. This means that there is no indication that the variance of the sepal width over the species is not homogeneous. There is a risk of a type-II error.

```
# Levene's Test for Homogeneity of Variance
leveneTest(y = iris$Sepal.Length,
group = iris$Species) # p-value: 0.002259

leveneTest(y = iris$Sepal.Width,
group = iris$Species) # p-value: 0.5555
```