APSTA Week 11 Exercises

1. Sugar Packing (Theory)

A sugar packaging machine is filling the bags with sugar where the weight accurately follows normal distribution with the mean μ and variance σ^2, where $\sigma = 7$g. In the sample of 36 packages the sample mean for the weight is 507g. Compute the 97% confidence interval for μ.

Solution:
The formula for the $100(1-\alpha)\%$ confidence interval for normal distributions with known variance is:

$$\left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Since we want 93%, we set $\alpha = 0.03$. The mean is $\bar{x}_n = 507$, samples $n = 36$ and the standard deviation $\sigma = 7$. That makes our 97% confidence interval

$$\left(507 - z_{0.015}\frac{7}{6}, 507 + z_{0.015}\frac{7}{6}\right).$$

Now we just need to find the critical value $z_{0.015}$. It can be calculated by taking the inverse CDF of the normal distribution at $1 - 0.015$:

$$F(z_p) = 1 - p \Leftrightarrow F^{-1}(1-p) = z_p$$

This function exists in R as `qnorm`:

```
qnorm(1-0.015)
```

```
## [1] 2.17009
```

We therefore insert 2.17 in the formula.

$$\begin{aligned}\left(507 - 2.17\frac{7}{6}, 507 + 2.17\frac{7}{6}\right) &= \left(507 - \frac{15.19}{6}, 507 + \frac{15.19}{6}\right) \\ &= (504.47, 509.53)\end{aligned}$$

We are 97% confident that the this interval contains the true mean.

2. Bags of potatoes (Theory)

You bought 10 very large bags of potatoes. Assume that the 10 weights can be viewed as a realization of a random sample from a normal distribution with unknown parameters. Your measures give you the following data:
- Sample mean: 14.5 kg
- Sample standard deviation: 0.3 kg

Construct a 95% confidence interval for the expected weight of a bag.

Solution:
The formula for the confidence interval for a normal distribution with unknown variance is

$$\left(\bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}\right).$$

This time, we need to find the critical value of the t-distribution with $n-1$ degrees of freedom. We do the same as above, but this time with the `qt` command. We have $10-1=9$ degrees of freedom, and since we want the 95% confidence interval, our $\alpha = 0.05$.

```
qt(0.95, 9)
```

```
## [1] 1.833113
```

Therefore, our critical value is 1.83. We insert the values into the interval:

$$\left(14.5 - 1.83\frac{0.3}{\sqrt{10}}, 14.5 + 1.83\frac{0.3}{\sqrt{10}}\right) = (14.326, 14.674)$$

3. How many samples do we need? (Theory)

Assume that we measure a person's height (in meters) and that the measurements are normal distributed with standard deviation $\sigma = 0.01$. How many measurements do we to make, if we want a 99% confidence interval no wider than 0.001 meters for the mean μ? Please explain how you find the number of required measurements.

Solution:
We can determine the size of the confidence interval from a confidence level based on the formula for the confidence interval. The size for an interval with confidence level $1-\alpha$ is

$$2 \cdot z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

If we then impose a restriction of a maximum width, we can isolate the number of samples required:

$$2 \cdot z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq w \;\Leftrightarrow\; n \geq \left(\frac{2z_{\alpha/2}\sigma}{w}\right)^2.$$

The maximum width w is 0.001, and the standard deviation is $\sigma = 0.01$. We want a 99% confidence interval, so $\alpha = 0.01$. Finally, we need to find the critical value.

```
qnorm(1-(0.01/2))
```

```
## [1] 2.575829
```

The critical value is 2.576. We can then insert the numbers:

$$n \geq \left(\frac{2 \cdot 2.576 \cdot 0.01}{0.001}\right)^2 = 51.52$$

The minimum number of samples is 52.

4. Cats & confidence (R)

The `cats` data set (available in `UsingR`) contains the bodyweight and heart weight of adult cats, along with their sex. (a) Compute the mean and the 90% confidence intervals for the body weight and heart weight assuming normality of the samples. Do the computation separately for the female and male cats. (So you have to compute 4 confidence intervals). Is there any difference between the results?
(b) Compute the one-sided 95% confidence intervals for the mean body weight of male cats and compare to the results you obtained at (a).

Solution:
I start by creating a function for computing the confidence interval.

```
CI <- function(mean, crit_val, sd, samples) {
  return(
    c(
      mean - crit_val * (sd / sqrt(samples)),
      mean + crit_val * (sd / sqrt(samples))
    )
  )
}
```

Since the true variance is unknown, I need to use the t-distribution. Since I want to 90% confidence interval, $\alpha = 0.1$ and $p = \alpha/2 = 0.05$.

```
library(UsingR)
```

```
## Indlæser krævet pakke: MASS

## Indlæser krævet pakke: HistData

## Indlæser krævet pakke: Hmisc

## Indlæser krævet pakke: lattice

## Indlæser krævet pakke: survival

## Indlæser krævet pakke: Formula

## Indlæser krævet pakke: ggplot2

##
## Vedhæfter pakke: 'Hmisc'

## De følgende objekter er maskerede fra 'package:base':
##
##     format.pval, units

##
## Vedhæfter pakke: 'UsingR'

## Det følgende objekt er maskeret fra 'package:survival':
##
##     cancer
```

```r
p <- 0.05
crit_val <- qnorm(1-p)

cats_F <- cats[cats$Sex == "F",]
cats_M <- cats[cats$Sex == "M",]
sample_means <- c(
  mean(cats_F$Bwt), mean(cats_F$Hwt),
  mean(cats_M$Bwt), mean(cats_M$Hwt)
)
names(sample_means) <- c(
  "Female body mean", "Female heart mean",
  "Male body mean", "Male heart mean"
)
```

Now that I've calculated the means, I can now calculate the confidence intervals.

```r
sample_sds <- c(
  sd(cats_F$Bwt), sd(cats_F$Hwt),
  sd(cats_M$Bwt), sd(cats_M$Hwt)
)
names(sample_sds) <- c(
  "Female body sd", "Female heart sd",
  "Male body sd", "Male heart sd"
)

FB_ci <- CI(sample_means[1], crit_val, sample_sds[1], length(cats_F))
FH_ci <- CI(sample_means[2], crit_val, sample_sds[2], length(cats_F))
MB_ci <- CI(sample_means[3], crit_val, sample_sds[3], length(cats_M))
MH_ci <- CI(sample_means[4], crit_val, sample_sds[4], length(cats_M))

cat("90% confidence intervals:")
```

```
## 90% confidence intervals:
```

```r
cat("\nFemale body weight:", FB_ci)
```

```
##
## Female body weight: 2.09938 2.619769
```

```r
cat("\nFemale heart weight:", FH_ci)
```

```
##
## Female heart weight: 7.912811 10.49144
```

```r
cat("\nMale body weight:", MB_ci)
```

```
##
## Male body weight: 2.45605 3.34395
```

```
cat("\nMale heart weight:", MH_ci)
```

```
##
## Male heart weight: 8.908379 13.73698
```

The intervals seem to be centered on larger values for males than females. Whatever the "heart weight" is, it's also way larger than the body weight.
Now I need the 95% confidence interval for for the male body weight. This means $\alpha = 0.05$ and $p = \alpha/2 = 0.025$.

```
p <- 0.025
crit_val <- qnorm(1-p)

ci <- CI(sample_means[3], crit_val, sample_sds[3], length(cats_M))
cat("Male body weight 95%:", ci)
```

```
## Male body weight 95%: 2.371001 3.428999
```

As expected, the confidence interval has grown slightly to accommodate our increase confidence.

5. Bootstrapping confidence intervals (R)

Consider again the `cats` dataset of the previous exercise. Construct the 90% confidence intervals for the mean body weight of male cats by empirical bootstrap, using 500 bootstrap repetitions. Compare the result to those you got in Problem 1.
How do they compare with the intervals you found in the previous exercise?

Solution:
To bootstrap, we first need to create an ECDF, then create studentised bootstrap means.

```
Fn <- ecdf(cats_M$Bwt)
sample_mean <- sample_means[3]
bootstrap_studentised_means <- c()
for (i in 1:500) {
  bootstrap_sample <- sample(cats_M$Bwt, size = length(cats_M), replace = TRUE)
  bootstrap_mean <- mean(bootstrap_sample)
  bootstrap_sd <- sd(bootstrap_sample)

  bootstrap_studentised_means <- c(
    bootstrap_studentised_means,
    ( bootstrap_mean - sample_mean ) / ( bootstrap_sd / sqrt(length(cats_M)) )
  )
}
```

Now that we have our studentised bootstrap means, we find c_l^* and c_u^* such that

$$\mathrm{P}\left(c_l^* < \frac{\bar{X}_n^* - \mu^*}{S_n^*/\sqrt{n}} < c_u^*\right) \approx 1 - \alpha.$$

I do so using the inverse ECDF. This is simply the quantile function. Since I want a 90% confidence interval, I use the $0.05th$ and $0.95th$ quantiles:

```
cl <- quantile(bootstrap_studentised_means, .05)
cu <- quantile(bootstrap_studentised_means, .95)
```

These values can then be inserted into the following formula:

$$\left(\bar{x}_n - c_u^* \frac{s_n}{\sqrt{n}}, \bar{x}_n - c_l^* \frac{s_n}{\sqrt{n}}\right).$$

```
interval <- c(
  sample_mean - cu * ( sample_sds[3] / sqrt(length(cats_M)) ),
  sample_mean - cl * ( sample_sds[3] / sqrt(length(cats_M)) )
)
cat("The 90% bootstrapped confidence interval for male body weight is", interval)
```

```
## The 90% bootstrapped confidence interval for male body weight is 2.198747 4.249511
```

For easier viewing, here are the three intervals:

$$\begin{aligned}
\text{Male body weight, 90\%:} &\ (2.456, 3.344) \\
\text{Male body weight, 95\%:} &\ (2.371, 3.429) \\
\text{Male body weight, 90\% bootstrapped:} &\ (2.124, 3.841)
\end{aligned}$$

As we can see, the bootstrapped interval is significantly wider than the other two, even when comparing it to the 95% confidence interval. This just goes to show how much better your confidence becomes when you *know* what distribution your data is coming from.