APSTA Week 13 Exercises

1. One sample t-test (T)

We perform a t-test for the null hypothesis $H_0 : \mu = 10$ by means of a dataset consisting of $n = 16$ elements with sample mean 11 and sample variance 4. We use significance level 0.05.
a. Should we reject the null hypothesis in favor of $H_1 : \mu \neq 10$?
b. What if we test against $H_1 : \mu > 10$?

Solution:
The test statistic we use is

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

where $\bar{x}_n = 11$, $\mu_0 = 10$, $s_n = \sqrt{4}$ and $n = 16$.
This test statistic has a $t(n-1)$ distribution. Since

$$\mathrm{P}(T \leq -t_{n-1,\alpha/2} \quad \text{or} \quad T \geq t_{n-1,\alpha/2}) = \alpha,$$

we can reject the null hypothesis of T goes outside of these bounds. Here, $\alpha = 0.05$.

$$\mathrm{P}(T \leq -t_{15,0.025} \quad \text{or} \quad T \geq t_{15,0.025}) = 0.05.$$

The critical value can be found using the `qt` function:

```
qt(1-0.05/2, 15)
```

```
## [1] 2.13145
```

Now we calculate the realisation of the test statistic.

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} = \frac{11-10}{\sqrt{4}/\sqrt{16}} = 2.$$

Since $2 \ngeq 2.132$, we cannot reject the null hypothesis.
To check the one-sided probability in the hypothesis $H_1 : \mu > 10$, we use $t_{n-1,\alpha}$ instead of $t_{n-1,\alpha/2}$.

```
qt(1-0.05, 15)
```

```
## [1] 1.75305
```

Since $2 \geq 1.753$, we reject H_0 in favour of $H_1 : \mu > 10$.
(This makes sense because the two-sided version takes into account the extremes on both sides. By removing the lower extreme, we can decrease the upper extreme under the same confidence.)

2. Major League Baseball (R)

The data set OBP (Using R) contains on-base percentages for the 2002 Major League Baseball season. Do a significance test to see wheter the mean on-base percentage is 0.330 against a two-sided alternative.

Solution:

```
# Import the dataset
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
```

```
summary(OBP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2016  0.3033  0.3265  0.3297  0.3534  0.5817
```

We want to test the alternative hypothesis that the mean μ is not equal to $\mu_0 = 0.330$. Therefore, we're testing

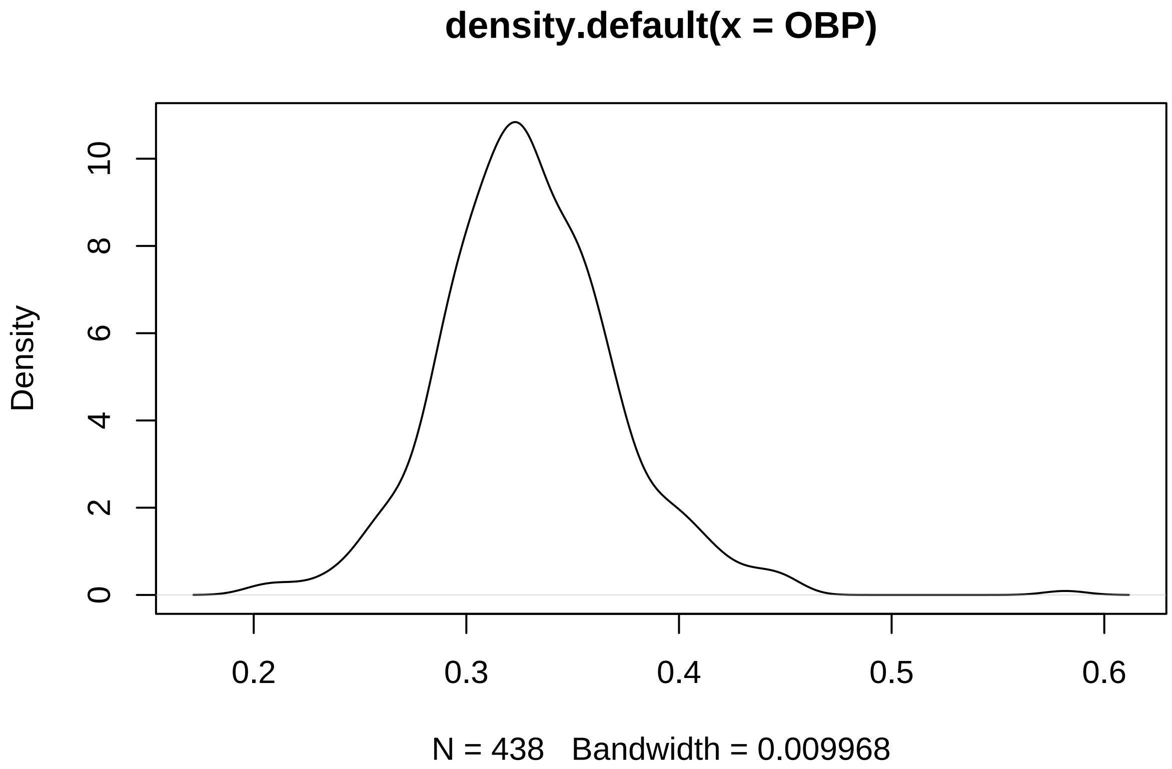
$$H_0 : \mu = 0.330 \quad \text{against} \quad H_1 : \mu \neq 0.330$$

Let's see if the data looks normally distributed:

```
length(OBP)
```

```
## [1] 438
```

```
plot(density(OBP))
```



```
shapiro.test(OBP)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OBP
## W = 0.97092, p-value = 1.206e-07
```

The size of the dataset is not large enough for a Z-test, and the shapiro test tells us that the data is not very normally distributed. We will therefore bootstrap the studentised mean.

```
n <- length(OBP)
sqrt_n <- sqrt(n)
x_bar <- mean(OBP)

t_s_samples <- c()
for (i in 1:10000) {
  bootstrap_sample <- sample(OBP, size=n, replace=TRUE)

  bootstrap_x_bar <- mean(bootstrap_sample)
  bootstrap_sd <- sd(bootstrap_sample)
  t_s <- (bootstrap_x_bar - x_bar) / (bootstrap_sd / sqrt_n)

  t_s_samples <- c(t_s_samples, t_s)
}
```

Now we can find c_l^* and c_u^* such that $\mathrm{P}(T^* \leq c_l^* \quad \text{or} \quad T^* \geq c_u^*) = \alpha$. At a significance of 0.05, $\alpha = 0.05$.

```
c_s <- quantile(t_s_samples, probs=c(0.05/2, 1-0.05/2))
t <- (x_bar - 0.330) / (sd(OBP) / sqrt_n)

cat("Test statistic realisation:", t)
```

```
## Test statistic realisation: -0.1684576
```

```
cat("\nLower/upper bound @ significance 0.05:", c_s)
```

```
##
## Lower/upper bound @ significance 0.05: -2.032491 1.932538
```

This value is not outside of the critical region, so we do not reject the null hyothesis.

3. Easter eggs (T & R)

Assume that you got six similar Easter eggs with $20g$ of chocolate reported in each. After taking one more lecture in Applied Statistics, you want to further investigate whether it is plausible that the eggs really contain $20g$ chocolate or if the egg producer is cheating. You weight the eggs and obtain the following six observations for the chocolate weight:

$$\begin{gathered}\text{Chocolate contents (g)}\\ [20.1, 19.1, 18.2, 20.2, 19.6, 19.1]\end{gathered}$$

You may assume that you measurement is a realization of a random sample from a normal distribution $N(\mu, \sigma^2)$, where μ represents the true average contents.
(a) Formulate the appropriate null hypothesis and alterantive hypothesis.

Solution:
The null hypothesis is that the mean is truly equal to $20g$ as they claim, that is, $H_0 : \mu = 20$. We are interested in whether or not the company is scamming us, so we only care if there is less chocolate than advertised, that is $H_1 : \mu < 20$.

(b) Which test is appropriate for testing the hypothesis? Explain why.

Solution:
We assume the measurements are from an $N(\mu, \sigma^2)$ distribution, and we only have a single sample, so we use the single-sample t-test.

(c) Compute the value of the test statistic and report your conclusion at significance level $\alpha = 0.05$.

Solution:
The test statistic is the studentised mean:

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

First we want to know the mean, standard deviation and square root of n:

```r
data <- c(20.1, 19.1, 18.2, 20.2, 19.6, 19.1)
cat("Mean:", mean(data))
```

```
## Mean: 19.38333
```

```r
cat("\nStandard deviation:", sd(data))
```

```
##
## Standard deviation: 0.7467708
```

```r
cat("\nsqrt(n):", sqrt(length(data)))
```

```
##
## sqrt(n): 2.44949
```

And then we calculate our realisation of the test statistic:

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} = \frac{19.38\bar{3} - 20}{0.747/2.45} = -\frac{0.61\bar{6}}{0.305} = -2.023$$

Finally, we need to find the critical value $t_{n-1,\alpha/2}$ where $\alpha = 0.05$. This can be done using `qt`:

```r
ct <- -qt(1-0.05, 6-1)
cat("Critical value:", ct)
```

```
## Critical value: -2.015048
```

As our value of -2.023 is less than the critical value -2.015, the reject the null hypothesis. The manufacturer is actually putting less than $20g$ of chocolate in the eggs.

(d) Compute the corresponding left ail p-value. Is it likely to observe these measurements under the null hypothesis?

Solution:
The p-value is calculated by $\mathrm{P}(T \leq t)$. We can calculate this using the `pt` function:

```r
n <- length(data)
t <- (mean(data)-20) / (sd(data)/sqrt(n))
p <- pt(t, n-1)
cat("p-value:", p)
```

```
## p-value: 0.04951219
```

So the p-value is only very slightly below the significance level 0.05.

4. Two-sample t-test (T)

The data in Table 28.3 (pp. 425, Dekking et al. (2010)) represents salaries (in pounds Sterling) in 72 randomly selected advertisements in The Guardian (April 6, 1992). When the range was given in the advertisement, the midpoint of the range is reproduced in the table. The data are salaries corresponding to two kinds of occupations ($n = m = 72$): (1) Creative, media, and marketing and (2) education. The sample mean and sample variance of the two datasets are, respectively:
(1) $\bar{x}_{72} = 17410$ and $s_x^2 = 41258741$,
(2) $\bar{y}_{72} = 19818$ and $s_y^2 = 50744521$,

Supposed that the datasets are modeled as realizations of normal distributions with expectations μ_1 and μ_2, which represent the salaries for occupations (1) and (2).
(a) Test the null hypothesis that the salary for both occupations is the same at level $\alpha = 0.05$ under the assumption of equal variances. Formulate the proper null and alternative hypotheses, compute the value of the test statistic, and report your conclusion.

Solution:
We want to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq= \mu_2$. We assume equal variance and normality, so we perform the two sample t-test with test statistic

$$T_p = \frac{\bar{X}_n - \bar{Y}_m}{S_p}$$

where

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right).$$

The test statistic has a $t(n+m-2)$ distribution. To test at level $\alpha = 0.05$ (and since $n = 5 = 72$) we need the critical value $t_{142,0.05/2}$.

```r
ct <- qt(1-0.05/2, 142)
cat("Critical value:", ct)
```

```
## Critical value: 1.976811
```

The critical region is therefore any value outside of $[-1.976811, 1.97681]$. We calculate the test statistic:

$$t_p = \frac{\bar{x}_{72} - \bar{y}_{72}}{s_p} = \frac{17410 - 19818}{\frac{71(41258741)+71(50744521)}{142} \cdot \frac{2}{72}} = -0.00189$$

This value is outside the critical region, so the null hypothesis is not rejected.

(b) Do the same without the assumption of equal variances.

Solution:
With no equal variance, we use the Welch's test. The test statistic is

$$T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d}$$

where

$$S_d^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}.$$

What's interesting about this test statistic is its degrees of freedom. It has v degrees of freedom, where

$$v = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}.$$

With a $t(v)$ distribution, our critical value is $t_{v,0.05/2}$ and we need to calculate v. I've opted to do this in R and not by hand because I don't hate myself enough.

```
var_x <- 41258741
var_y <- 50744521
n <- 72
m <- n

v <- ( ( (var_x/n) + (var_y/m) )^2 ) / ( ( (var_x^2)/((n^2) * (n-1)) ) + ( (var_y^2/((m^2) * (m-1)) ) )
cat("Degrees of freedom:", v)
```

```
## Degrees of freedom: 140.5064
```

And the $t_{140.5064,0.05/2}$ critical value is:

```
qt(1-0.05/2, v)
```

```
## [1] 1.976992
```

with the critical region being the complement of $[-1.976992, 1.976992]$. Our realisation of the test statistic is

$$t_d = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} = \frac{17410 - 19818}{\sqrt{\frac{41258741}{72} + \frac{50744521}{72}}} = -2.130204.$$

This value is in the critical region, and we therefore reject the null hypothesis.

(c) As a comparison, one carries out an empirical bootstrap simulation for the nonpooled studentized mean difference. The bootstrap approximations for the critical values are $c_l^* = -2.004$ and $c_u^* = 2.133$. Report your conclusions about the salaries on the basis of the bootstrap results.

Solution:
The test statistic for the nonpooled studentised mean difference is

$$T_d = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{S_d}$$

where

$$S_d^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}.$$

Under the H_0, $\mu_X - \mu_Y = 0$, so our realisation is

$$t_d = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} = -2.130204.$$

This is less than c_l^*, and therefore we reject the null hypothesis. The salaries for the two occupations are different.

5. Bootstrapping in two-sample tests (R)

For the babies (UsingR) data set, the variable age contains the recorded mom's age and dage contains the dad´s age for several different cases in the sample. Do a significance test of the null hypothesis of equal age against a one-sided alternative that dads are older in the sample population. Use a non-normal model with bootstrapping.

Solution:
As it isn't specified in the question, I assume unequal variance in the mom's and dad's ages. The test statistic to be bootstrapped is

$$T_d = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{S_d}$$

where

$$S_d^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}.$$

```r
moms <- babies$age # X
dads <- babies$dage # Y

x_bar <- mean(moms)
y_bar <- mean(dads)
x_var <- var(moms)
y_var <- var(dads)
n <- length(moms)
m <- length(dads)

t_d_s_samples <- c()
for (i in 1:10000) {
  bootstrapped_moms <- sample(moms, n, replace=TRUE)
  bootstrapped_dads <- sample(dads, m, replace=TRUE)

  x_bar_s <- mean(bootstrapped_moms)
  y_bar_s <- mean(bootstrapped_dads)
  x_sd_s <- sd(bootstrapped_moms)
  y_sd_s <- sd(bootstrapped_dads)

  s_d_s_sq <- ((x_sd_s^2) / n) + ((y_sd_s^2) / m)
  t_d_s <- ( (x_bar_s - y_bar_s) - (x_bar - y_bar) ) / sqrt(s_d_s_sq)

  t_d_s_samples <- c(t_d_s_samples, t_d_s)
}
```

If the dads are older, that means that the numbers will tend towards the negative. That means we need to find the critical value c_l^* such that $\mathrm{P}(T_d^* \leq c_l^*) = \alpha$ where I choose $\alpha = 0.05$.

```r
c_l_s <- quantile(t_d_s_samples, 0.05)
cat("Critical value:", c_l_s)
```

```
## Critical value: -1.605522
```

Now we just compare it with the actual test statistic.

```r
t_d <- (x_bar - y_bar) / sqrt( (x_var/n) + (y_var/m) )
cat("Realised test statistic:",t_d)
```

```
## Realised test statistic: -11.0671
```

This is significantly lower than the critical value, so the dads are definitely older than the moms.