ST227 - Introductory exercise with solutions

Exercise 1

Suppose we have two factories, say *Factory A* and *Factory B*, that use certain machines to produce goods.

The probability of failure of such a machine in any given week is 0.1 in Factory A and 0.2 in Factory B.

1.a

In R, write a procedure that simulates the time to failure for a machine in Factory A

```
set.seed(123)

rnos <- runif(100)
time_to_failure <- which(rnos < 0.1)[1]

time_to_failure
```

```
## [1] 6
```

1.b

Estimate the expected time to failure by repeating your procedure from 1.a 30 times and taking the average of your results.

```
set.seed(123)
reps <- 30

death <- numeric(reps)

for(i in 1:reps){
    death[i] <- which(runif(100) < 0.1)[1]
}

mean(death)
```

```
## [1] 9.9
```

1.c

The expected lifespan in this case is inversely proportional to the prbability of failure. Provide an argument why this is the case and compare your simulation result from 1.b with the expected lifespan.

The distribution of the time to failure is a geometric random variable with parameter equal to the failure probability, say p. The mean of a geometric random variable is the reciprocal of this parameter, i.e. $1/p$. We can also see this as follows: In any random week, the machine either fails, with probability p, or not. If it does not fail, the prbability of failure is independent of the past, and thus its expected lifespan at this point must be equal to the overall expected lifespan. That is,

$$\mathrm{E}(T) = p + (1-p)(1 + \mathrm{E}(T))\,,$$

which solves to $T = 1/p$.

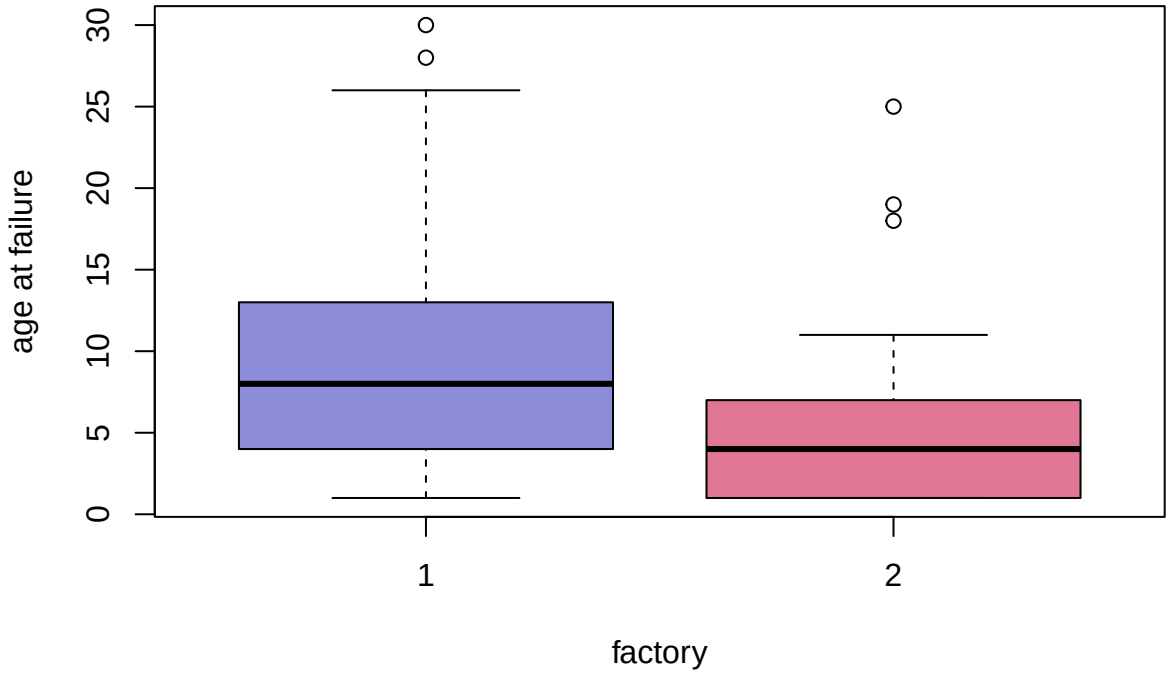
```
true_failure_prob <- 0.1
simulated_failure_prob <- 1 / mean(death)

cbind(true_failure_prob, simulated_failure_prob)
```

```
##      true_failure_prob simulated_failure_prob
## [1,]               0.1              0.1010101
```

1.d

Repeat steps 1.a and 1.b for Factory B. Using a boxplot, or any other appropriate visualisation method of your choosing, visually compare the distribution of the simulated times to failure of Factories A and B.

```
death2 <- numeric(reps)

for(i in 1:reps){
    death2[i] <- which(runif(100) < 0.2)[1]
}

deaths <- c(death, death2)
factory <- factor(c(rep(1,30),rep(2,30)))
cols <- c(rgb(0.1,0.1,0.7,0.5), rgb(0.8,0.1,0.3,0.6))

plot(factory, deaths, xlab="factory", ylab="age at failure",col=cols)
```



1.e (OPTIONAL)

Use a Gamma GLM to test for difference between the time to failure data from Factory A and B. What are your conclusions?

```
model1 <- glm(deaths ~ factory, family = Gamma)

summary(model1)
```

```
##
## Call:
## glm(formula = deaths ~ factory, family = Gamma)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.10101    0.01704   5.928 1.79e-07 ***
## factory2     0.07650    0.03446   2.220   0.0303 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.8538053)
##
##     Null deviance: 51.203  on 59  degrees of freedom
## Residual deviance: 46.496  on 58  degrees of freedom
## AIC: 363.37
##
## Number of Fisher Scoring iterations: 6
```

We reject the null hypothesis of no difference at a significance level of 5% but not at 1%.

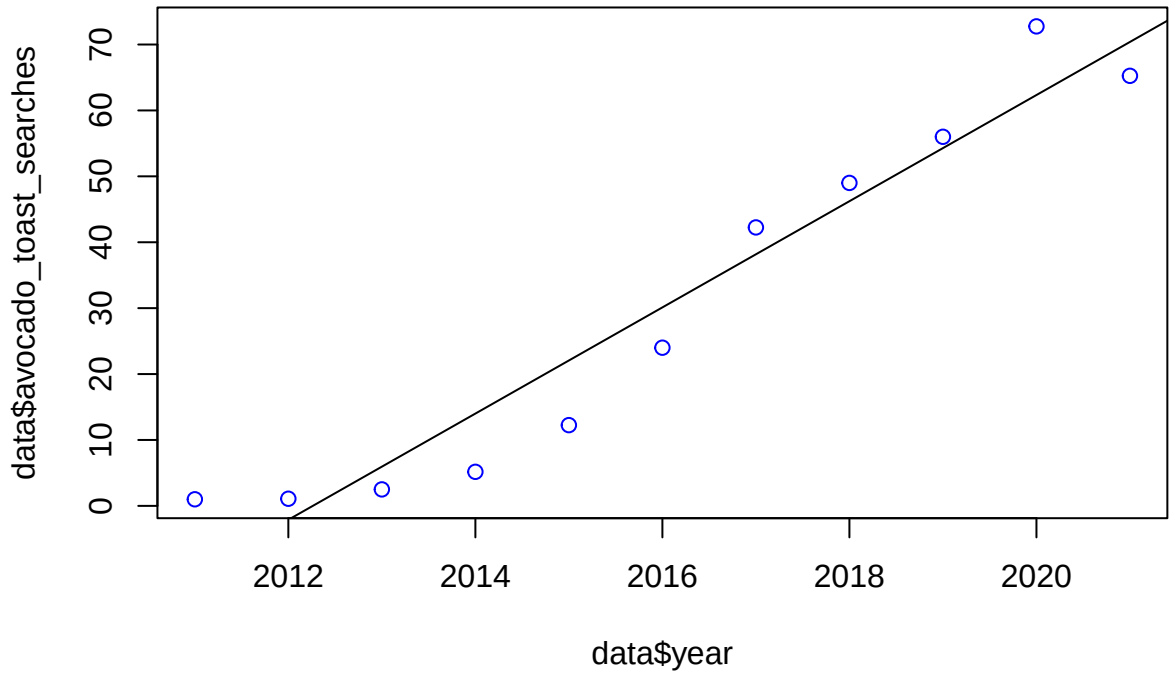
Exercise 2

Load the `ST227_example_data.xlsx` using the library `readxl`. Run a linear regression using `avocado_toast_searches` as your response and `year` as covariate. Interpret the results.

```
library(readxl)

data <- read_xlsx("ST227_example_data.xlsx")
data <- as.data.frame(data)

model2 <- lm(avocado_toast_searches ~ year, data)

summary(model2)
```

```
##
## Call:
## lm(formula = avocado_toast_searches ~ year, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.808 -5.627  1.720  3.636 11.163
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.621e+04  1.462e+03  -11.09 1.51e-06 ***
## year         8.055e+00  7.251e-01   11.11 1.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.605 on 9 degrees of freedom
## Multiple R-squared:  0.932,  Adjusted R-squared:  0.9245
## F-statistic: 123.4 on 1 and 9 DF,  p-value: 1.482e-06
```

```r
plot(data$year, data$avocado_toast_searches, col = "blue")
abline(model2)
```



The intecept term is negative, suggesting negative search results for low values of year, which is not meaningful. There seems to be a strong time trend in `avocado_toast_searches`.