Applied Statistics

Exam Submission - 01/06-2022

Submission information

- Name: Mie Jonasson
- Student-ID: 20431
- Date-of-birth: 31/07-2001
- Course: Applied Statistics
- Course Code: BSAPSTA1KU

Question 1

(a) **Sample space**

Since we are throwing the fair four-sided die until we get our first '4', it is possible to get it at the first throw, but also to get other things for a long time before getting a '4'. It is not possible to obtain negative numbers, since we are denoting the sample space as the *number of throws*.

This can also be stated as a geometric distribution!

A natural choice for the sample space would therefore be the set of **all positive integers**:

$$\Omega = \{1, 2, 3, 4, 5, 6, ...\}$$

(b) **Events**

Event A is the set of all even positive integers:

$$A = \{2, 4, 6, 8, 10, ...\}$$

Event B is all positive integers that are greater than or equal to 3:

$$B = \{3, 4, 5, 6, 7, 8, 9, ...\}$$

(c) **Probabilities**

For this we use that the experiment can be described as a geometric distribution, with parameter $p = \frac{1}{4}$ (fair foursided dice). A geometric distribution has a probability mass function given by:

$$p_X(k) = P(X = k) = (1 - p)^{k-1} * p$$

We see that $B^C = \Omega - B = \{1, 2\}$ Furthermore, when we are looking at $A \cap B^C$ we are looking at the *intersect* between the sets, which in this case only contains a single element: 2

We can now simply calculate the probability of getting the first '4' on the 2nd throw:

$$p_X(2) = P(X = 2) = \left(\frac{3}{4}\right)^{2-1} * \frac{1}{4} = \frac{3}{4} * \frac{1}{4} = \frac{3}{16}$$

Question 2

We are looking at a discrete random variable with probability density function:

$$f_X(x) = \begin{cases} \frac{3}{4}(1-x^2), & x \in [-1,1], \\ 0, & \text{otherwise.} \end{cases}$$

(a) **Expected Value of X**

Expected value can be calculated by:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Since X has 0-probabilities outside the bounds $[-1, 1]$ we will only need to calculate the integral within this interval (since the integral over the rest of the sample space is 0):

$$\begin{aligned} E[X] &= \int_{-1}^{1} \left(x * \frac{3}{4}(1-x^2) \right) dx \\ &= \int_{-1}^{1} \left(\frac{3}{4} * x - \frac{3}{4} * x^3 \right) dx \end{aligned}$$

Now we integrate and insert boundaries:

$$E[X] = \left[\frac{3}{4} * \frac{1}{2} * x^2 - \frac{3}{4} * \frac{1}{4} * x^4 \right]_{-1}^{1}$$

$$= \left[\frac{3}{8} * x^2 - \frac{3}{16} * x^4 \right]_{-1}^{1}$$

$$= \left(\frac{3}{8} * 1^2 - \frac{3}{16} * 1^4 \right) - \left(\frac{3}{8} * (-1)^2 - \frac{3}{16} * (-1)^4 \right)$$

$$= \left(\frac{6}{16} - \frac{3}{16} \right) - \left(\frac{6}{16} - \frac{3}{16} \right)$$

$$= \frac{3}{16} - \frac{3}{16} = 0$$

I.e. the expected value of the random variable is $E[X] = 0$

(b) **Variance of X**

Firstly we obtain $E[X^2]$ by same means as in problem a, using:

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx$$

We use the same boundaries as before by same argument:

$$E[X^2] = \int_{-1}^{1} \left(x^2 * \frac{3}{4}(1-x^2) \right) dx$$
$$= \int_{-1}^{1} \left(\frac{3}{4} * x^2 - \frac{3}{4} * x^4 \right) dx$$

Now we integrate:

$$E[X^2] = \left[\frac{3}{4} * \frac{1}{3} * x^3 - \frac{3}{4} * \frac{1}{5} * x^5 \right]_{-1}^{1}$$
$$= \left[\frac{3}{12} * x^3 - \frac{3}{20} * x^5 \right]_{-1}^{1}$$
$$= \left(\frac{3}{12} * 1^3 - \frac{3}{20} * 1^5 \right) - \left(\frac{3}{12} * (-1)^3 - \frac{3}{20} * (-1)^5 \right)$$
$$= \left(\frac{3}{12} - \frac{3}{20} \right) - \left(-\frac{3}{12} - (-\frac{3}{20}) \right)$$
$$= \frac{3}{12} - \frac{3}{20} + \frac{3}{12} - \frac{3}{20} = \frac{6}{12} - \frac{6}{20} = \frac{5}{10} - \frac{3}{10} = \frac{2}{10}$$

Now we can use the formula for the variance given by:

$$Var[X] = E[X^2] - E[X]^2 = \frac{1}{5} - 0^2 = \frac{1}{5}$$

I.e. the variance of X is $\frac{1}{5}$

(c) $P(X > 0.5)$

This is the integral:

$$P(X > 0.5) = \int_{0.5}^{\infty} \left(\frac{3}{4}(1-x^2) \right)$$

We can set the upper boundary to 1, as everything past that point will have probability 0.

We now integrate:

$$P(X > 0.5) = \left[\frac{3}{4} * x - \frac{3}{4} * \frac{1}{3} * x^3 \right]_{0.5}^{1}$$
$$= (\frac{3}{4} * 1 - \frac{3}{4} * \frac{1}{3} * 1^3) - (\frac{3}{4} * 0.5 - \frac{3}{4} * \frac{1}{3} * (0.5)^3)$$
$$= \frac{3}{4} - \frac{3}{12} - (\frac{3}{8} - \frac{3}{12} * \frac{1}{8})$$
$$= \frac{3}{4} - \frac{1}{4} - \frac{3}{8} + \frac{1}{32} = \frac{16}{32} - \frac{12}{32} + \frac{1}{32} = \frac{5}{32}$$

We can now conlude that: $P(X > 0.5) = \frac{5}{32}$

Question 3

We are looking at the continuous distribution with probability density function:

$$f_\alpha(x) = \begin{cases} \mathrm{e}^{-(x-\alpha)} & \text{for } x \geq \alpha, \\ 0, & x < \alpha, \end{cases}$$

(a) Likelihood Function

The likelihood function for parameter α can be calculated by the following function:

$$L(\alpha) = P(X_1 = x_1, ...X_n = x_n) = p_\alpha(x_1) * p_\alpha(x_2) * ... * p_\alpha(x_n)$$

Now we simply just insert the probability density function given, in place of $p_\alpha(x_i)$:

$$L(\alpha) = \mathrm{e}^{-(x_1-\alpha)} * \mathrm{e}^{-(x_2-\alpha)} * ... * \mathrm{e}^{-(x_n-\alpha)}$$

(b) Maximum Likelihood Estimate

We want to choose the parameter α such that our sample is the most likely! I.e. we want to maximize the likelihood function

To make it easier for ourselves to do computations, we look at the loglikelihood-function, which shares the same maximum as the likelihood function:

$$\ell(\alpha) = ln(L(\alpha)) = ln(\mathrm{e}^{-(x_1-\alpha)}) + ln(\mathrm{e}^{-(x_2-\alpha)}) + ... + ln(\mathrm{e}^{-(x_n-\alpha)})$$

This can be cut down, since $ln(\mathrm{e}^c) = c$:

$$\ell(\alpha) = -(x_1 - \alpha) + (-(x_2 - \alpha)) + ... + (-(x_n - \alpha))$$

This can be cut down even more, since we see that we add α n times, and subtract each sample once:

$$\ell(\alpha) = n * \alpha - \sum_{i=1}^{n} x_i$$

Oops... we see that this is a linear relationship that is constantly increasing the likelihood.. To maximize this function we see that we want α to be as large as possible.

BUT the fallpit here is, that if ANY of our $x_1, x_2, ..., x_n$ has a 0-probability, then $L(\alpha) = 0$ (by 0-multiplication). Looking at the probability density function, we note that 0-probabilities occur if and only if $x < \alpha$.

This means, that we want to make our parameter α as big as possible, without any $x_i < \alpha$. The natural choice for this job is of course the minimum.

Thus we can now conclude, that the maximum likelihood estimate for α is

$$\alpha = min(x_1, x_2, ..., x_n)$$

Question 4

(a) Simple Statistics

```
summary(Langren1644$Longitude)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.74   20.99   25.81   24.55   27.88   30.13
```

```
sd(Langren1644$Longitude)
```

```
## [1] 4.196055
```

We can conclude that the longtitudes observed span between 17.74 and 30.13. Furthermore we see that the mean of 24.55 is less than the median of 25.81, which in turn means that it is not a symmetric distribution, it is leaning to the right with possible outliers in the lower half.

(b) Confidence Interval

Since we do not know the true variance of the assumend normal distribution, we use the fact that the studentized mean has a $t(n-1)$ distribution. The studentized mean is given by $\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$ and the Confidence interval can then be estimated by:

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

I created a function for this task:

```
confidence_interval_normal <- function(sample,conf) {
  a <- (1 - conf)
  m <- mean(sample)
  s <- sd(sample)
  n <- length(sample)
  t <- qt(a/2, n-1, lower.tail = F)
  return(c(m-t*s/(sqrt(n)),m+t*s/(sqrt(n))))
}
```

Now we run the function:

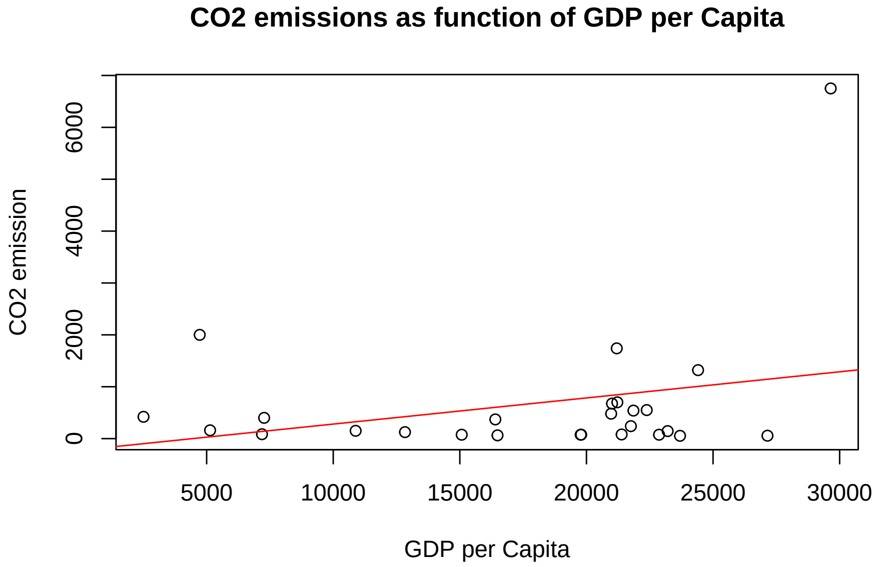
```
confidence_interval_normal(iq,0.95)
```

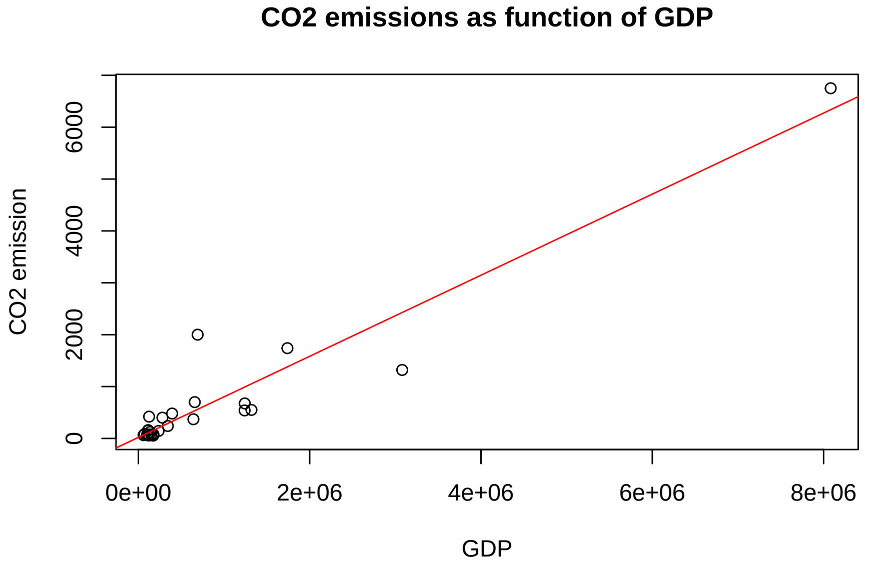
```
## [1]  98.90868 103.79132
```

This means that we can say with a 95% confidence level that the mean IQ is between 98.9 and 103.8

(c) Linear Regression

We fit a linear model on the data, and then we plot it together with the scatterplot. I modelled both the 'perCapita' GDP and total GDP to explain CO2-emissions, since the latter seemed to have a better linear relationship:





From this we see that the 'perCapita' GDP does not seem to be correlated with the CO2 emissions, but on the other hand, the total GDP of a country does seem to have a somehow nice linear relationship.

We quickly look at the correlation coefficient to deepen our knowledge:

```
cor(emissions$CO2, emissions$perCapita)
```

```
## [1] 0.2757962
```

```
cor(emissions$CO2, emissions$GDP)
```

```
## [1] 0.9501753
```

These numbers confirm our initial suspicion: GDP and CO2 emission have a correlation coefficient in the proximity of 1, indicating a linear relationship, whilst percapita GDP and CO2 emmision have a relatively low correlation coefficient.

Question 5

(a) Formulating hypotheses

Our null-hypothesis is, that there is the same amount of ice cream in both brands packages:

$$H_0 : \mu_A = \mu_B$$

Since we are specifically interested in whether there is less ice cream in brand A than brand B packages, our alternative hypothesis is:

$$H_A : \mu_A < \mu_B$$

(b) Why Bootstrap?

We use bootstrapping whenever we cannot assume that the data we get is normally distributed. This is because the standard t-test is modelled after normally distributed samples, and we have to estimate det distribution of the test statistic T in all other cases, except when we are dealing with a large sample.

(c) Bootstrap program

Before making the program, i will determine whether i am dealing with samples that have different or similar variances:

```r
sample_a <- c(99.2,100.5,98.9,99.6,97.7)
sample_b <- c(100.5,100.1,99.1,98.2,98.3)
var(sample_a)
```

```
## [1] 1.047
```

```r
var(sample_b)
```

```
## [1] 1.078
```

These variances are similar, thus i will be using pooled bootstrap estimation:

Firstly i create seperate functions for the basic entities of the calculations of the test statistic of any two-sample test, along with a function that returns a bootstrap sample:

```r
variance_p <- function(sample1,sample2) {
  n <- length(sample1)
  m <- length(sample2)
  sx <- var(sample1)
  sy <- var(sample2)
  return(((n-1)*sx+(m-1)*sy)/(n+m-2) * (1/n + 1/m))
}
test_t_p <- function(sample1,sample2) {
  sp <- sqrt(variance_p(sample1,sample2))
  m1 <- mean(sample1)
  m2 <- mean(sample2)
  return((m1-m2)/sp)
}
get_boot <- function(data) {
  sample(data, length(data), replace=T)
}
```

The two first functions correspond to the following formulas:

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

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

Now i will define two more functions; one that calculates the bootstrap estimate of T_p along with a function that takes 2 samples and a confidence level, prints the critical value of a ‘lesser than’ test and returns the bootstrapped test-statistics:

```
tp_boot <- function(b_sample_a,b_sample_b,ma,mb) {
  mab <- mean(b_sample_a)
  mbb <- mean(b_sample_b)
  b_sp <- sqrt(variance_p(b_sample_a,b_sample_b))
  return(((mab-mbb)-(ma-mb))/b_sp)
}
two_sample_less <- function(samplea,sampleb,conf) {
  ma <- mean(samplea)
  mb <- mean(sampleb)
  boot_ts <- c()
  for (i in 1:10000) {
    boot_a <- get_boot(samplea)
    boot_b <- get_boot(sampleb)
    boot_ts <- c(boot_ts, tp_boot(boot_a,boot_b,ma,mb))
  }
  a <- 1-conf
  cat('Critical value (lower bound): ',quantile(boot_ts,a))
  return(boot_ts)
}
```

The first function corresponds to the following formula:

$$t_p^* = \frac{(\bar{x}_n^* - \bar{y}_m^*) - (\bar{x}_n - \bar{y}_m)}{s_p^*}$$

And the second function corresponds to calculating this value t_p^* for 10,000 bootstrap samples, printing the critical value and returning all the t_p^* in a vector

(d) **Conclusions at** $\alpha = 0.05$

Now we can simply calculate the test statistic T_p of our samples, and compare it to our functions critical value. At the same time we will use the returned vector of bootstrap test statistics to estimate the tail probability (p-value) of our test statistic;

```
t <- test_t_p(sample_a,sample_b) ; t
```

```
## [1] -0.0920358
```

```
boots <- two_sample_less(sample_a,sample_b,0.95)
```

```
## Critical value (lower bound):  -1.780776
```

```
cat('P-value of test-statistic t: ',mean(boots < t))
```

```
## P-value of test-statistic t:  0.4559
```

We see that we do not have strong enough evidence against our null-hypothesis, so we **do not** reject H_0 on the basis of our samples. It is also noteable that the p-value is at 45%, meaning that there is a 45% chance of getting a sample at least as "extreme" as ours if H_0 is true, i.e. very likely!

For sanity check we quickly look at the means of our samples:

```
mean(sample_a)
```

```
## [1] 99.18
```

```
mean(sample_b)
```

```
## [1] 99.24
```

These means are indeed very similar, and thus we feel no need to refute our function.