STA3030f_KPNTHA001

Project1

Thabo Kopane

HandInDate: 9 March 2018

Through our project, we were tasked with collecting data of incoming communications either by social media or email clients amongst students on campus. I decided to ask students on Jammie plaza how many emails (both on personal and outlook - overlap if both) the previous day. I tried my best to randomize the students and I did not ask them what they were studying and none of the following analysis is influenced by the students' choice of degrees or majors.

1.1 The observed data

I went out on Jammie plaza on a mid-day on a Tuesday, between 13h00 and 14h00 and asked students who had time to check how many emails that they had received, in total, the previous day. This is the data

```
emails <- c(7, 13, 5, 10, 21, 22, 12, 18, 13, 14)
```

1.2 exploratory analysis

The following is the exploratory data, to give an indication of the underlying data.

```
##mean of sample
(xbar <- mean(emails))
```

```
## [1] 13.5
```

```
#sdev of samples
(stdev <- sd(emails))
```

```
## [1] 5.562773
```

```
#median of samples
(med <- median(emails))
```

```
## [1] 13
```

```
#variance of samples
(variance <- var(emails))
```

```
## [1] 30.94444
```

```
#Min and Max respectively
(MIN <- min(emails))
```

```
## [1] 5
```

```
(MAX <- max(emails))
```

```
## [1] 22
```

```
#The interquartile range
(iqr <- IQR(emails))
```

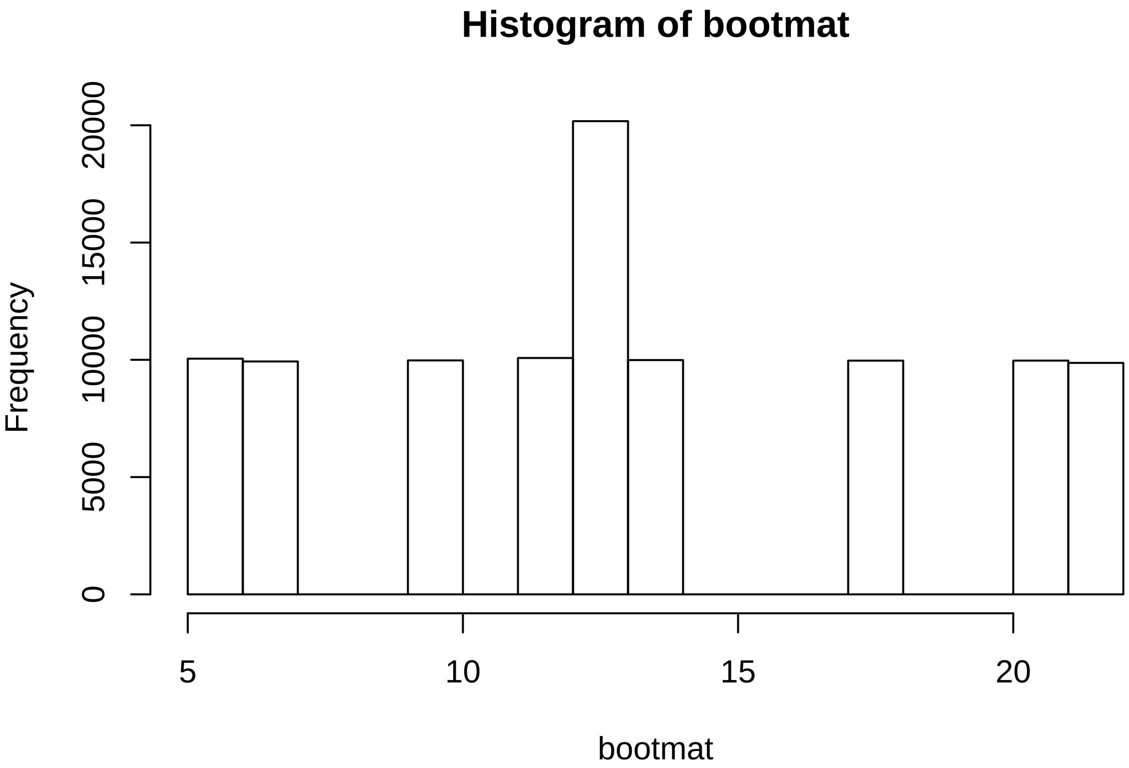
```
## [1] 6.5
```

The interquartile range is marginally higher than the minimum recorded observation and way lower than the Maximum observation, I could say that based in this - maybe our data might be skewed to the left of the mean - but that is too soon for our observation.

1.3 a) part1: bootstrapping for the CI

A for loop to create 10000 items to bootstrap

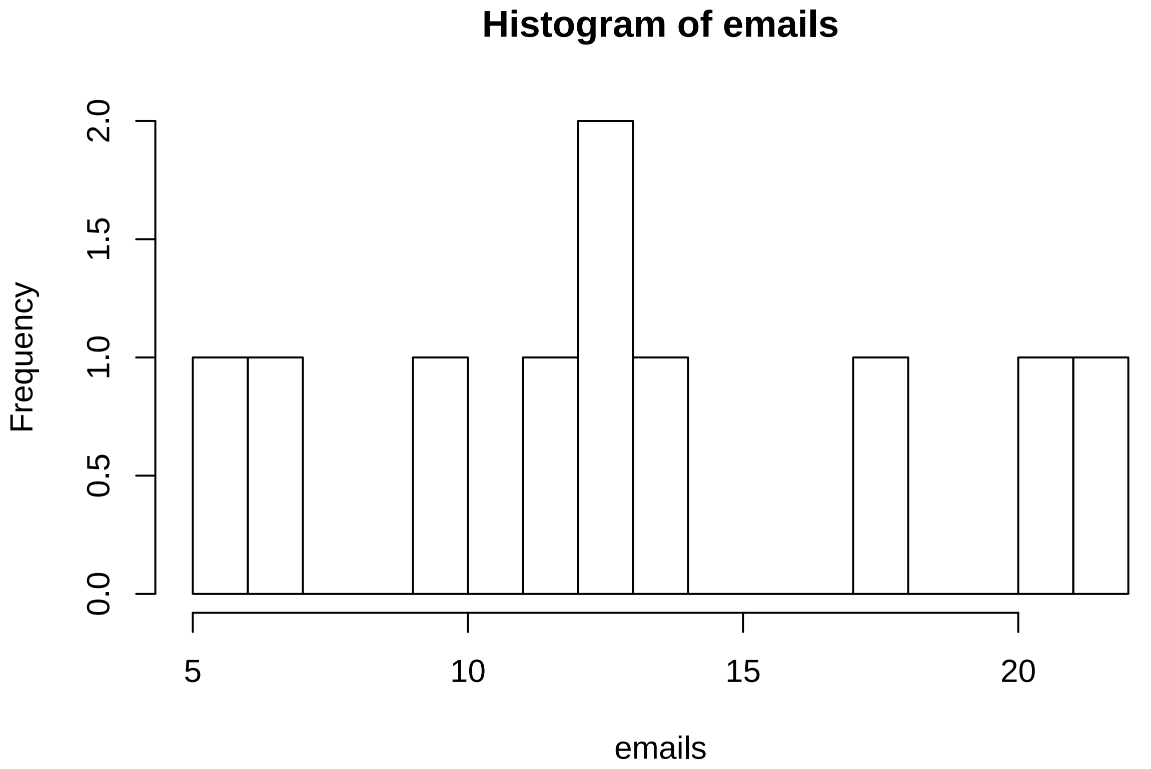
```
n = 10000
bootmat <- matrix(NA, nrow = n, ncol = 10)
for(i in 1:n){
  bot <- sample(emails, 10, replace = TRUE)

  #bootmat[i] <- sample(emails, 10, replace = TRUE)
  bootmat[i,] <- bot
}



#histogram for the bootrapped sample
(hist(bootmat, breaks = 20))
```



```
## $breaks
##  [1]  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
##
## $counts
##  [1] 10050  9930     0     0  9975     0 10078 20176  9989     0     0
## [12]     0  9966     0     0  9967  9869
##
```

```
## $density
##  [1] 0.10050 0.09930 0.00000 0.00000 0.09975 0.00000 0.10078 0.20176
##  [9] 0.09989 0.00000 0.00000 0.00000 0.09966 0.00000 0.00000 0.09967
## [17] 0.09869
##
## $mids
##  [1]  5.5  6.5  7.5  8.5  9.5 10.5 11.5 12.5 13.5 14.5 15.5 16.5 17.5 18.5
## [15] 19.5 20.5 21.5
##
## $xname
## [1] "bootmat"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
##histogram for the sample data
(hist(emails, breaks = 20))
```



```
## $breaks
##  [1]  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
##
## $counts
##  [1] 1 1 0 0 1 0 1 2 1 0 0 0 1 0 0 1 1
##
## $density
##  [1] 0.1 0.1 0.0 0.0 0.1 0.0 0.1 0.2 0.1 0.0 0.0 0.0 0.1 0.0 0.0 0.1 0.1
##
## $mids
##  [1]  5.5  6.5  7.5  8.5  9.5 10.5 11.5 12.5 13.5 14.5 15.5 16.5 17.5 18.5
## [15] 19.5 20.5 21.5
##
## $xname
## [1] "emails"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
##5number summary
(summary(bootmat))
```

```
##        V1              V2              V3             V4
##  Min.   : 5.00   Min.   : 5.00   Min.   : 5.0   Min.   : 5.00
##  1st Qu.:10.00   1st Qu.:10.00   1st Qu.:10.0   1st Qu.:10.00
##  Median :13.00   Median :13.00   Median :13.0   Median :13.00
##  Mean   :13.48   Mean   :13.49   Mean   :13.4   Mean   :13.53
##  3rd Qu.:18.00   3rd Qu.:18.00   3rd Qu.:18.0   3rd Qu.:18.00
##  Max.   :22.00   Max.   :22.00   Max.   :22.0   Max.   :22.00
##        V5              V6              V7              V8
##  Min.   : 5.00   Min.   : 5.00   Min.   : 5.00   Min.   : 5.00
##  1st Qu.:10.00   1st Qu.:10.00   1st Qu.:10.00   1st Qu.:10.00
##  Median :13.00   Median :13.00   Median :13.00   Median :13.00
##  Mean   :13.48   Mean   :13.55   Mean   :13.51   Mean   :13.52
##  3rd Qu.:18.00   3rd Qu.:18.00   3rd Qu.:18.00   3rd Qu.:18.00
##  Max.   :22.00   Max.   :22.00   Max.   :22.00   Max.   :22.00
##        V9             V10
##  Min.   : 5.00   Min.   : 5.00
##  1st Qu.:10.00   1st Qu.:10.00
##  Median :13.00   Median :13.00
##  Mean   :13.52   Mean   :13.35
##  3rd Qu.:18.00   3rd Qu.:18.00
##  Max.   :22.00   Max.   :22.00
```

```
##Bootstrap means
meanBS <- apply(bootmat, MARGIN = 1, FUN = mean)
```

The bootstrapped sample and the sample of the population have an almost identical histogram, that looks normal - normal in the sense that the graph would be symmetric.

1.3 a) part2 and b) Confidence Intervals

Bootstrapping uses the Bootstrap principle, which estimates the behaviour of the bootstrapped sample around the parameter value - which we do not know and we are trying to estimate. here we are comparing two methods of finding the CI of the sample

95 percent quantile based on bootstrap resampling. we assume that the bootstrapped sample adheres to the Bootstrap principle

```
##the quantiles - percentile
quants <- quantile(bootmat, probs = c(0.025, 0.975))

##the lowe and upper limits
```

```
(lowerl <- 2*xbar - quants[2])
```

```
## 97.5%
##     5
```

```
(upperl <- 2*xbar - quants[1])
```

```
## 2.5%
##   22
```

95 percent CI, based on the standard normal theory. We do not know the underlying distribution. We use the t-distribution to find the confidence interval.

```
##based on the t-distribution.
m= length(emails)
##lower limit
(lwler <- xbar +qt(.025, df=m-1) * stdev/sqrt(m))
```

```
## [1] 9.520632
```

```
##Upper kimit
(upperLimit <- xbar +qt(0.0975, df=m-1) * stdev/sqrt(m))
```

```
## [1] 11.03708
```

The original sample that was collected had a mean of 13.5, and when we do the confidence interval with bootstrapping, we can see that:

```
## deviation from the "true" mean
(xbar - lowerl)
```

```
## 97.5%
##   8.5
```

```
(xbar - upperl)
```

```
## 2.5%
## -8.5
```

The difference or the deviation is exactly 8.5%.

when we do the standard normal theory, the interval deviation:

```
## deviation from the "true" mean
(xbar - lwler)
```

```
## [1] 3.979368
```

```
(xbar - upperLimit)
```

```
## [1] 2.462918
```

The difference is narrower on the t-distribution. if we are to use the bootstrap to estimate the true mean, it would lie between [5,22] and thus we would coclude that we are we claim to be 95% confident that the mean could be in [5, 22]. If we are to use the t-distribution, it would lie between [9.5206, 11.037] and we could conclude by saying that we claim to be 95% confident that the mean could be in [9.5206, 11.037].

The t-distribution's CI is not accurate, we are adviced against using the word "accurate" - if we are to compare the confidence intervals, the sample CI is a slightly better estimate of where the mean could lie because it's the actual data and not an estimate of the actual data.

1.4 Interquartile range confidence inteval

```
num=10000
#because bigger than 10000 is overkill
bootsIQR <- matrix(NA, ncol = 10, nrow = num)

for(i in 1:num){
  botsamp <- sample(emails, 10, replace = T)

  botIQR <- IQR(botsamp)
  bootsIQR[i,] <- botIQR
}

##Quants for the matrix
quantiling <- quantile(bootsIQR, probs = c(0.025, 0.975))
quantiling
```

```
##  2.5% 97.5%
##  1.00 13.25
```

```
(lowerquartile <- 2*iqr - quantiling[2])
```

```
## 97.5%
## -0.25
```

```
(upperquartile <- 2*iqr - quantiling[1])
```

```
## 2.5%
##   12
```

Based on this we can see that the CI for the interquartile range is [.25, 12].

1.5 Standard Errors

The following is an estimate of the bias, the standard error and the sample error of the bootstats

```
##mean and sd
(sxbar <- mean(bootmat))
```

```
## [1] 13.48393
```

```
(ssdev <- sd(bootmat))
```

```
## [1] 5.266458
```

```
#Bias estimate
(esbias <- sxbar - xbar)
```

```
## [1] -0.01607
```

```
#Corrected bias
(corrbias <- 2*xbar - sxbar)
```

```
## [1] 13.51607
```

```
#Estimate of standard erro -stdev
ssdev
```

```
## [1] 5.266458
```

```
##Sample error
samperror <- bootmat - xbar
```

1.6 Hypothesis testing

Are some people getting more messages per day than the average person, or less or no difference.

H0: mu = 13.5 people are getting the same amount of emails a day. H1: mu >13.5 some people are getting more emails per day.

1.7 Testing for significance

```
##AMOUNT OF entries where the mean is 13.5
mean13.5 <- (meanBS >13.5)
(countme <- sum(mean13.5))
```

```
## [1] 4869
```

```
##p-value based on observation
(p_value <- countme/n)
```

```
## [1] 0.4869
```

out of 10000 values in our bootstrap sample, 4896 items exceed 13.5 - that could mean that our observed p-value is 0.4896. Based on this observation, our value is not significant.

now we calculate the p-value with t-stats and see if it corresponds to the above p-value

```
#we use the means and sdev from the bootstats
(tstats <- (sxbar - xbar)/(ssdev/sqrt(m)))
```

```
## [1] -0.009649332
```

```
degF <- m-1

(t_p_value <- 1 - pt(tstats,degF))
```

```
## [1] 0.5037442
```

the p-value is certainly higher, this means that the data is not significant enough to reject the null hypothesis. there mean of emails appear to equal 13.5 a day. So people are generally getting the same amount of emails a day.