Statistics first HW

FIRST PROBLEM

∗∗1∗∗

```
library(readxl)
library(e1071)
library(stringr)
set.seed(1)
ceo <- read_excel("ceo.xls")
salary <- ceo$salary
```

−−*a*−−

```
mean(salary)
```

```
## [1] 2027.517
```

This value describes average salary of all CEOs.

```
mean(salary, trim=0.1)
```

```
## [1] 1710.092
```

We removed 10% from both sides of sample and computed new mean value. This is done in order to remove extreme values from a sample which may make big impact on **mean**. The **10%-trimmed mean** is smaller than the **mean**, which says us there are more "big" extreme values then "small" ones.

```
median(salary)
```

```
## [1] 1600
```

This value tells us about element in the middle of the sample. 50% of salaries are lower than median and 50% are higher.

```
print(quantile(salary, prob=c(0.25, 0.75)))
```

```
##    25%    75%
## 1084.0 2347.5
```

25% of salaries are lower than 1084.0$ and 25% of salaries are higher than 2347.5$

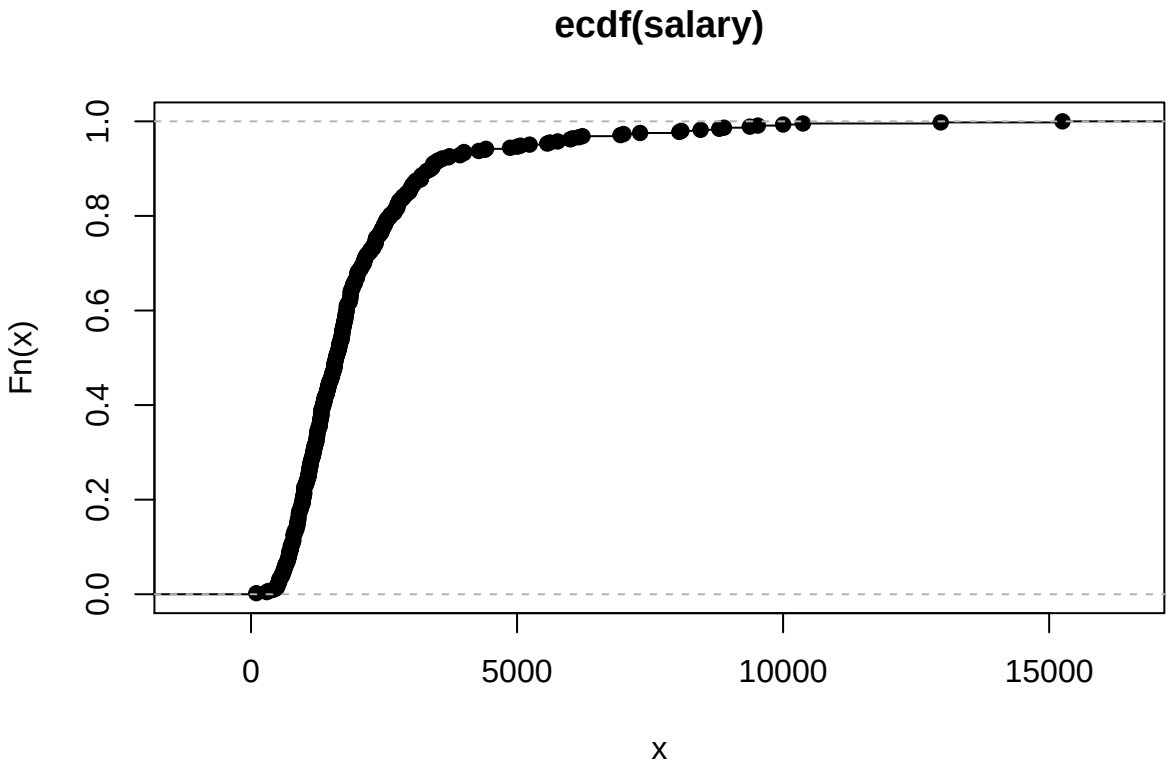
```
print(quantile(salary, prob=c(0.10, 0.90)))
```

```
##    10%    90%
##  750.0 3384.4
```

10% of salaries are lower 750.0$ and 10% of salaries are higher than 3384.4$.

−−*b*−−

```
salary_ecdf = ecdf(salary)
plot(salary_ecdf)
```




```
quantile(salary, 0.2)
```

```
##   20%
## 976.2
```

$(\hat{F})^{-1}(0.2)$ is the same as second decile. This value tells us that ~20% of CEOs have salary less then 976.2.

```
quantile(salary, 0.8)
```

```
##  80%
## 2613
```

$(\hat{F})^{-1}(0.8)$ equals to eights decile. This value may be read as "~20% of CEOs have salary more then 2613".

```
salary_ecdf(1000)
```

```
## [1] 0.2237136
```

This tells that 22.37% of CEOs have salary smaller then 1000$ (or 77.63% of CEOs have salary bigger than 1000$).

```
1 - salary_ecdf(5000)
```

```
## [1] 0.05369128
```

And this tells you that only ~5.37% of CEOs have salary bigger than 5000$.

$- - c - -$

```r
attach(mtcars)
```

```r
par(mfrow=c(1, 2))
boxplot(salary, main="Boxplot of salary")
hist(salary)
```

From this plots we can see that median is a bit higher than ~1500$, interquartile range is much smaller than whole range from min to max and there are a lot of outliers in the top, so we can make a decision that distribution is not normal with a long tail on the right.

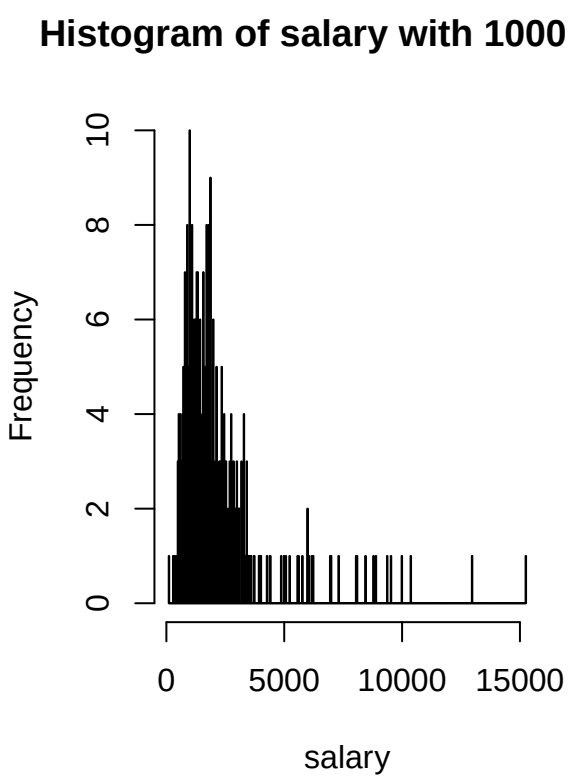
```r
skewness(salary)
```

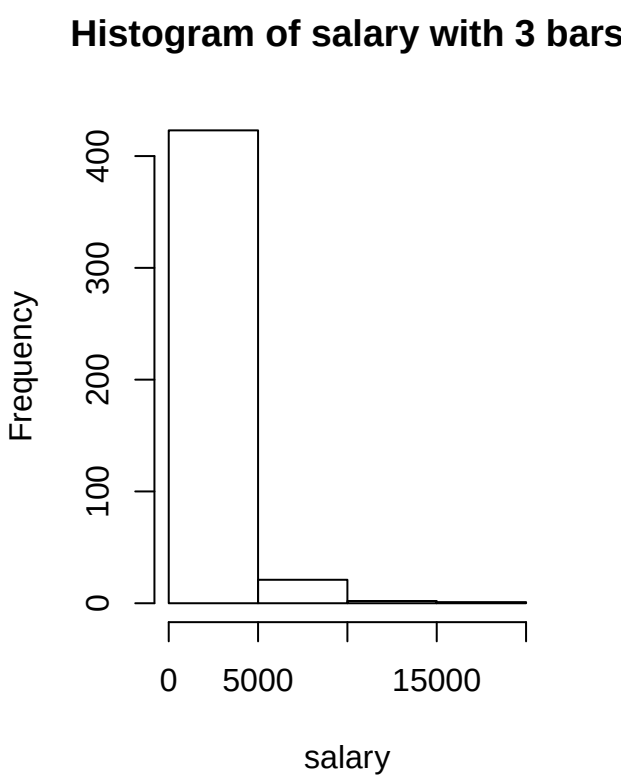
```
## [1] 3.379632
```

This value only assures my words that a tail on the right is fatter than a tail on the left.

$$--d--$$

```r
par(mfrow=c(1, 2))
hist(salary, breaks=1000, main="Histogram of salary with 1000 bars")
hist(salary, breaks=3, main="Histogram of salary with 3 bars")
```



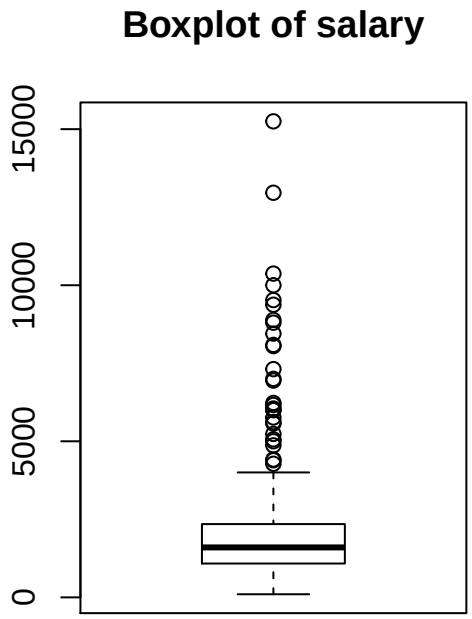


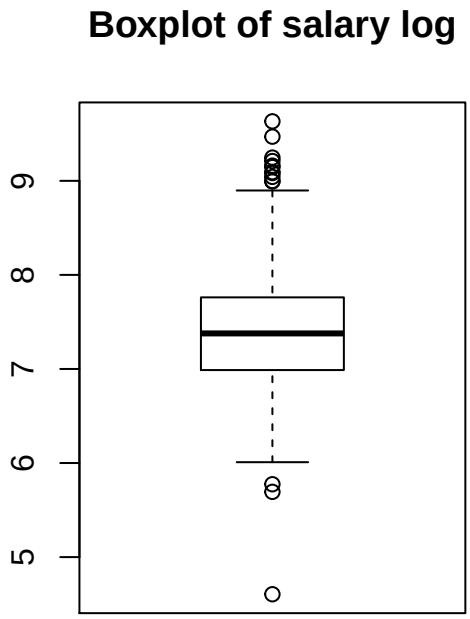
A default number of bars in `histo` is calculated by Sturges' formula as **ceiling(log2(k) + 1)**, and for our dataset ceiling(log2(447) = 10. On this plots we can see that too much bars on the histogram are as bad as too few. It is not giving an easy-readable information about data being observed, so it's nearly useless.

– – *e* – –

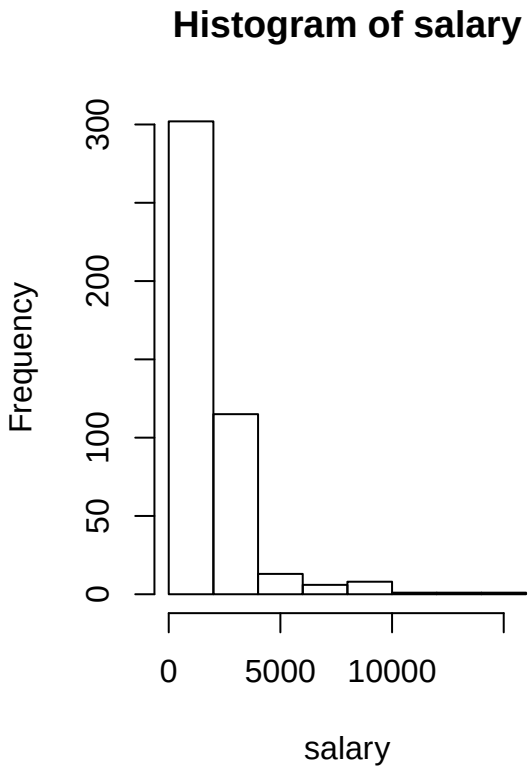
```
salary_log = log(salary)
```

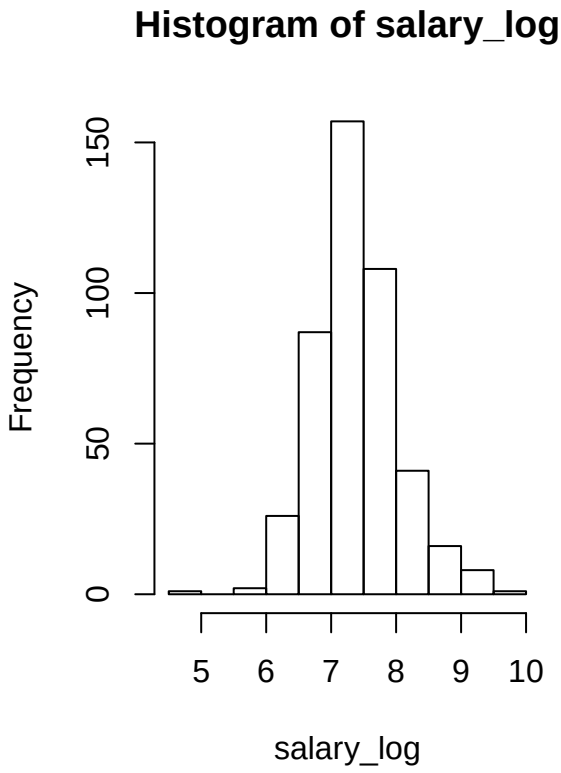
```
par(mfrow=c(1, 2))
boxplot(salary, main="Boxplot of salary")
boxplot(salary_log, main="Boxplot of salary log")
```






```
par(mfrow=c(1, 2))
hist(salary)
hist(salary_log)
```






```
mean(salary_log)
```

```
## [1] 7.391898
```

```
median(salary_log)
```

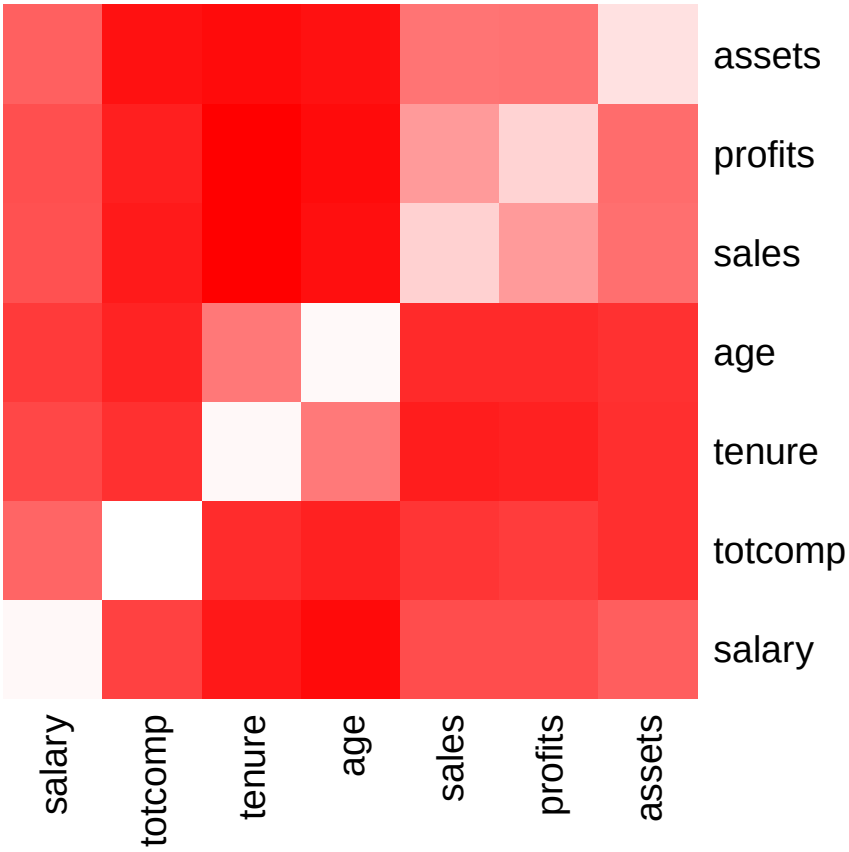
```
## [1] 7.377759
```

From plots above we can make a conclusion that lg(salary) looks much more like normal distribution, because central bar is the highest one and it is the line of symmetry for histogram, and because in boxplot median is also near the central line of the plot. Close values of mean and median only confirm these conclusions.

$$**2**$$

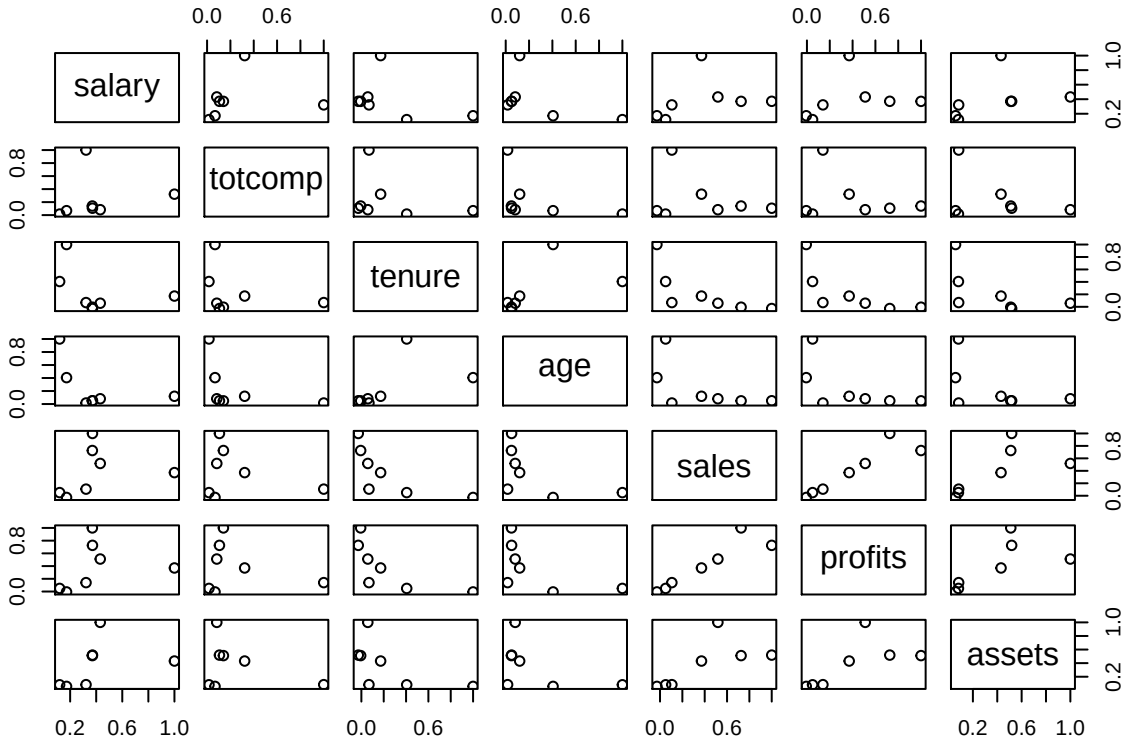
$$--a--$$

```
ceo = ceo[1:7] # remove NaNs
correlations_pearson = cor(ceo, method='pearson')
correlations_spearman = cor(ceo, method="spearman")
heatmap(correlations_pearson, Rowv=NA, Colv=NA, col = colorRampPalette(c("red", "white"))(n = 299))
```

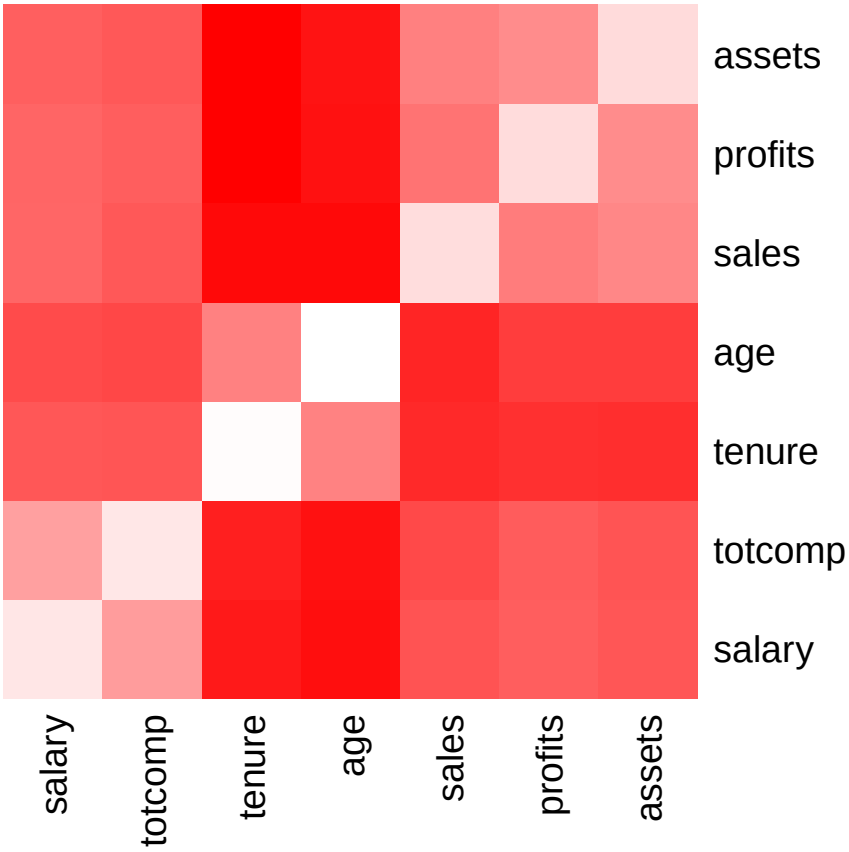


First of all we see that all features are correlating with themselves (white colour). After that we can see that the most correlating features are (light-red colour): sales&profits, sales&assets and tenure&age. Everything else correlates with small coefficients (no correlation, red colour).

```
pairs(correlations_pearson)
```



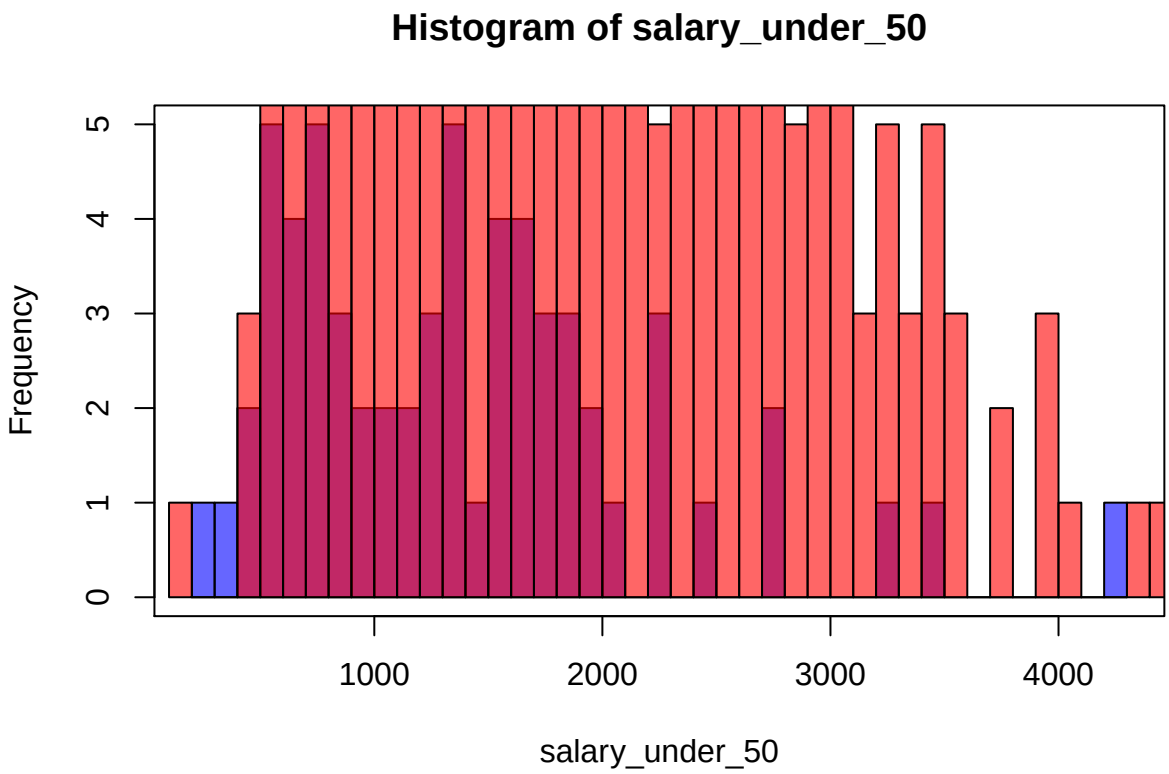

```
heatmap(correlations_spearman, Rowv=NA, Colv=NA, col = colorRampPalette(c("red", "white"))(n = 299))
```



From this heatmap we also see that all diagonal elements correlate fully. As well we see that totcomp&salary, sales&assets, assets&profits and age&tenure correlate better than everything else.

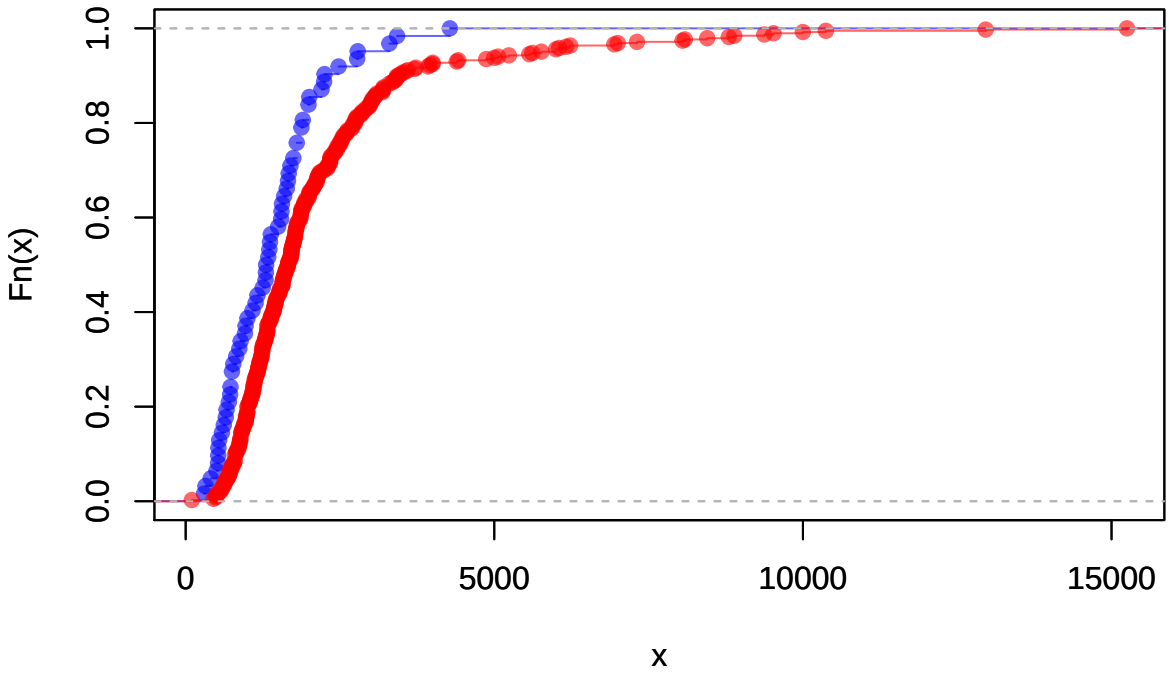
```
salary_under_50 = ceo[ceo$age < 50,]$salary
salary_above_50 = ceo[ceo$age >= 50,]$salary
```

```
hist(salary_under_50, col=rgb(0,0,1,0.6), breaks=length(salary_under_50)/2)
hist(salary_above_50, col=rgb(1,0,0,0.6), add=T, breaks=length(salary_above_50)/2)
box()
```



From this plot we can see that there are much more CEOs with age more than 50 years. Also we see, that their salaries are higher. But on the other hand it is possible that person is younger than 50 years and has big salary - we can see this by one blue element in the right of the plot. Moreover - minimal salary is person's who is older than 50. On the other hand, we can see that majority of young CEOs have low salaries while older CEOs are distributed from the one side to another on salaries axe. It means that in the beginning young CEO may expect to have low salary, but he has chances to become rich or to remain poor, nothing is guaranteed.

```
salary_under_50_ecdf = ecdf(salary_under_50)
salary_above_50_ecdf = ecdf(salary_above_50)
plot(salary_under_50_ecdf, col=rgb(0,0,1,0.6), main='', ylim=c(0, 1), xlim=c(min(salary), max(salary)))
par(new=T)
plot(salary_above_50_ecdf, col=rgb(1,0,0,0.6), main='', ylim=c(0, 1), xlim=c(min(salary), max(salary)))
```



On this plot we can clearly see that younger CEOs have salary smaller than older CEOs. As well, we can see that salaries of younger CEOs are more squeezed while older CEOs' salaries are stretched.

∗ ∗ 3 ∗ ∗

```
salaries_and_age <- ceo
salaries_and_age$salary[as.numeric(ceo$salary) < 2000] <- "S1"
salaries_and_age$salary[as.numeric(ceo$salary) >= 2000 & as.numeric(ceo$salary) < 4000] <- "S2"
salaries_and_age$salary[as.numeric(ceo$salary) >= 4000] <- "S3"
salaries_and_age$age[as.numeric(ceo$age)<50] <- "A1"
salaries_and_age$age[as.numeric(ceo$age)>=50] <- "A2"
```

```
absolute_frequencies <- table(salaries_and_age$salary, salaries_and_age$age)
addmargins(absolute_frequencies)
```

```
##
##        A1  A2 Sum
##   S1   52 248 300
##   S2    9 107 116
##   S3    1  30  31
##   Sum  62 385 447
```

```
relative_frequencies <- prop.table(absolute_frequencies)
addmargins(relative_frequencies)
```

```
##
##                A1          A2         Sum
##   S1  0.116331096 0.554809843 0.671140940
```

```
##    S2  0.020134228 0.239373602 0.259507830
##    S3  0.002237136 0.067114094 0.069351230
##    Sum 0.138702461 0.861297539 1.000000000
```

1) n12 = 9 It's number of CEOs who earns from 2000$ to 4000$ and is younger than 50 years.
2) h12 = 0.116 It's how many percents of all CEO earn from 2000$ to 4000$ and is younger than 50 years. 1 = 100%, so 0.116 = 11.6%
3) n1. = 300. It shows how many CEOs have salary from 2000$ to 4000$.
4) h1. = 0.67 It's how many percents of all CEO earn from 2000$ to 4000$.

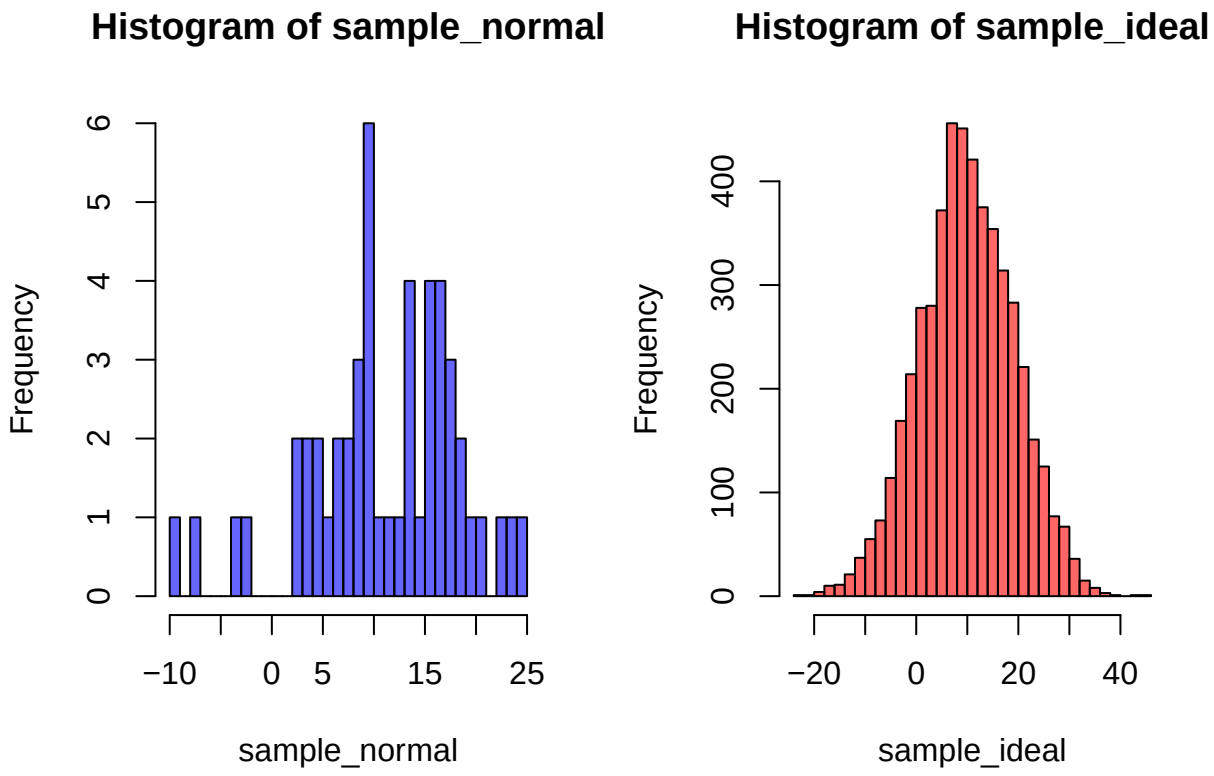
SECOND PROBLEM

∗∗1∗∗

```
sample_normal <- rnorm(50, 10, 9)
sample_ideal <- rnorm(5000, 10, 9)
```

– – *a* – –

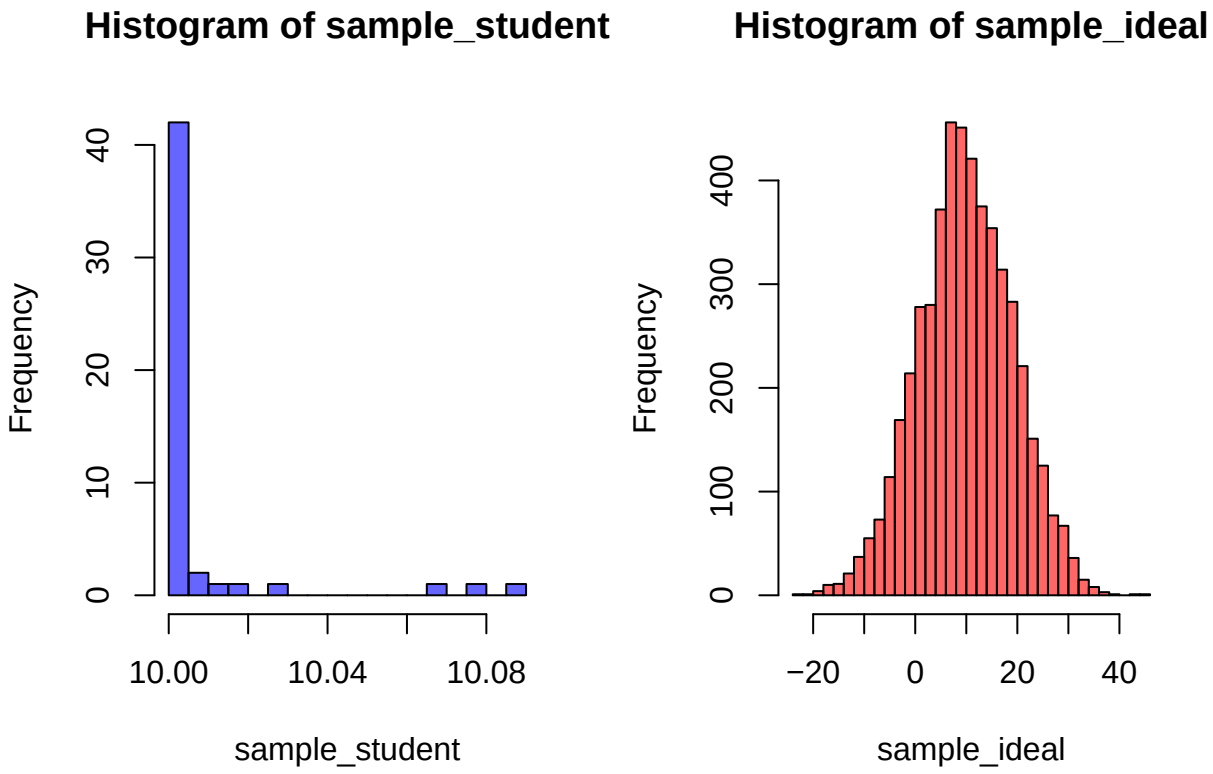
```
par(mfrow=c(1, 2))
hist(sample_normal, breaks=length(sample_normal)/2, col=rgb(0,0,1,0.6))
hist(sample_ideal, breaks=length(sample_normal)/2, col=rgb(1,0,0,0.6))
```



As for me, from the plot (blue) it is not visible that the distribution of the sample is normal.

– – *b* – –

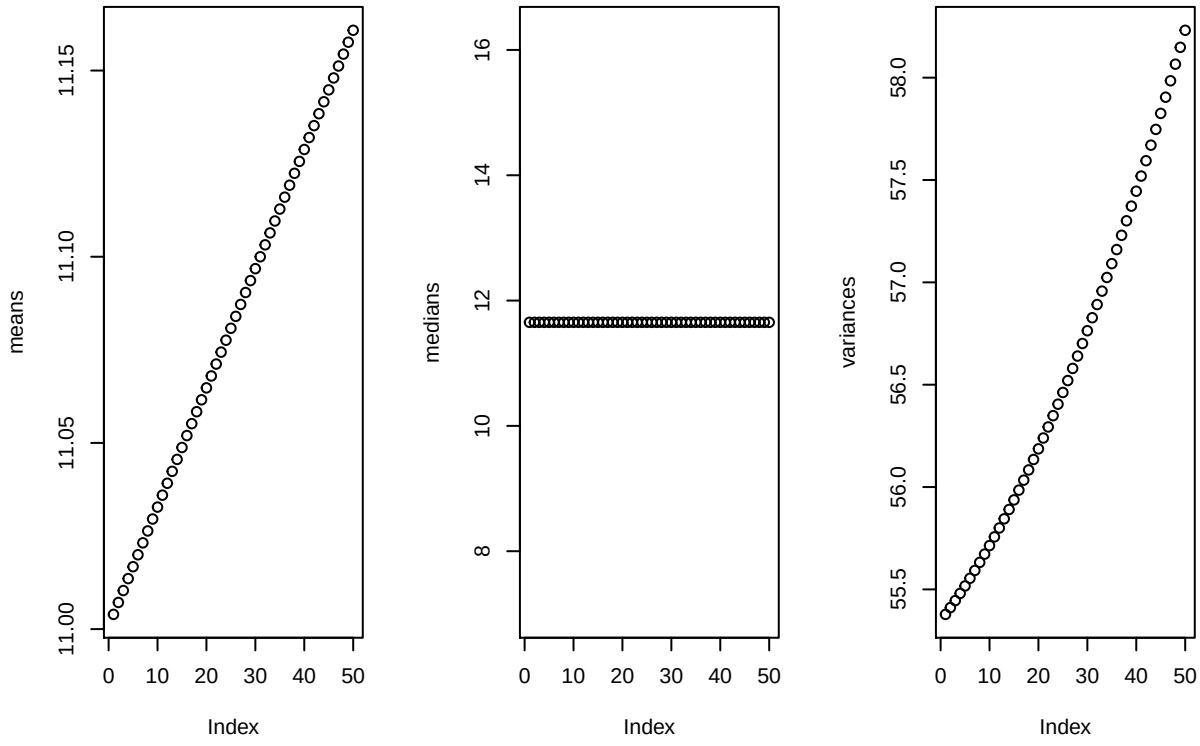
```
sample_student <- dt(sample_normal, 5)
sample_student <- 10 + 3 * sqrt(3/5) * sample_student
par(mfrow=c(1, 2))
hist(sample_student, breaks=length(sample_normal)/2, col=rgb(0,0,1,0.6))
hist(sample_ideal, breaks=length(sample_normal)/2, col=rgb(1,0,0,0.6))
```



$$**2**$$

$$--a--$$

```
means = c()
medians = c()
variances = c()
for (i in 0:49) {
  new_element <- 16 + (8/49) * i
  modified_sample <- c(sample_normal, new_element)
  means = c(means, mean(modified_sample))
  medians = c(medians, median(modified_sample))
  variances = c(variances, var(modified_sample))
}
par(mfrow=c(1, 3))
plot(means)
plot(medians)
plot(variances)
```

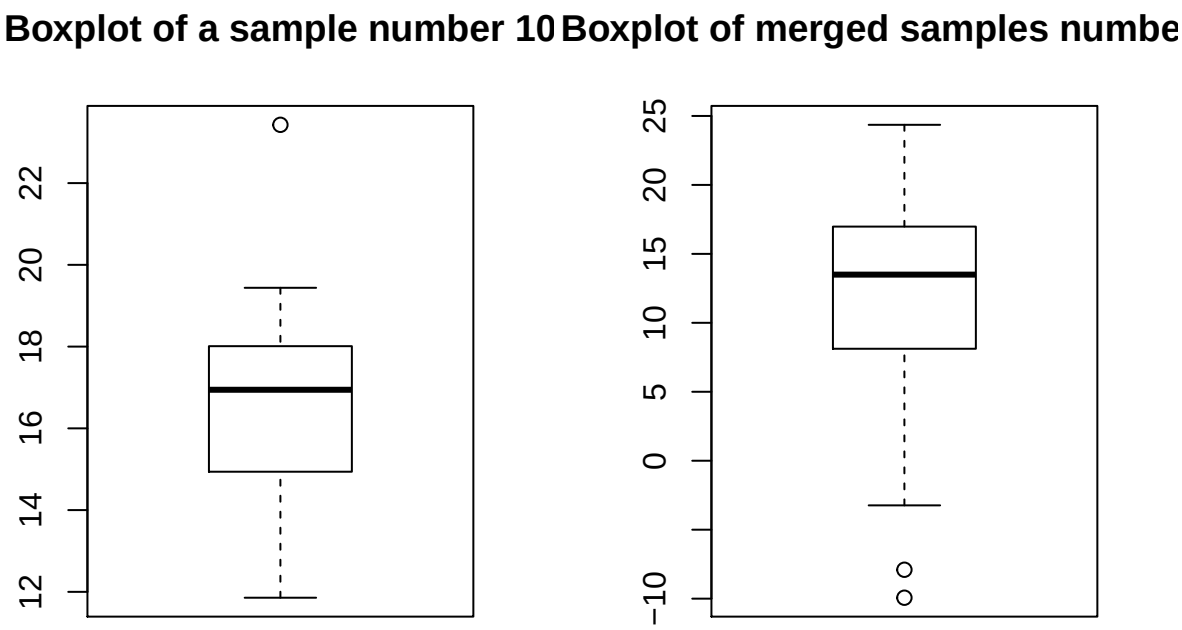


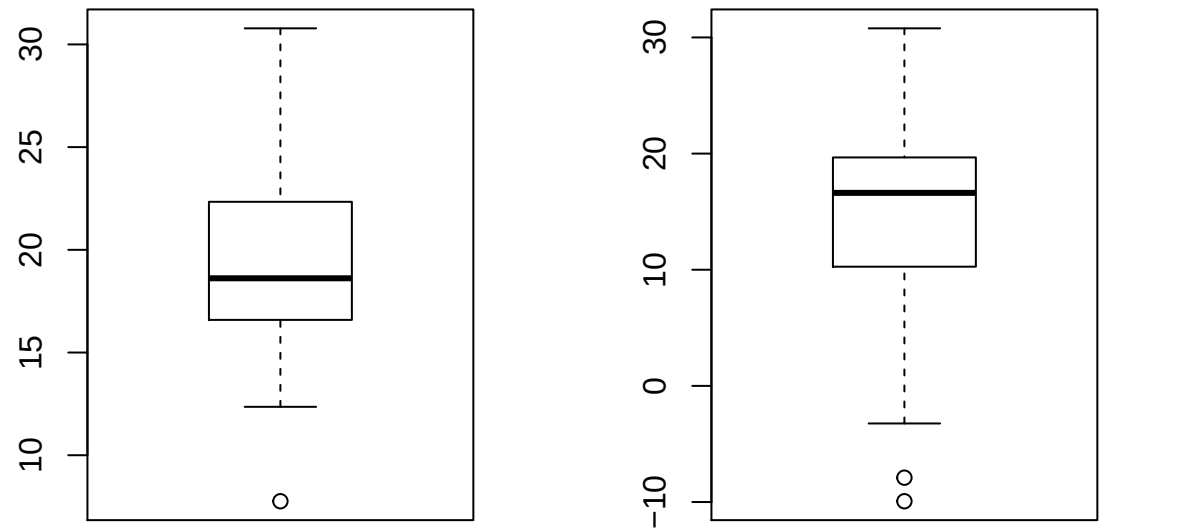
What I can see from the above plots: 1) A mean is growing, but not significantly (from ~10.15 to ~10.35). Also, it's growing linearly. 2) A meadian is staying the same. That's because we each time add only one element and each time to the same side of the sample. 3) A variance is also growing. This means that each element increases a range of values in the sample. The variance is growing not linearly.

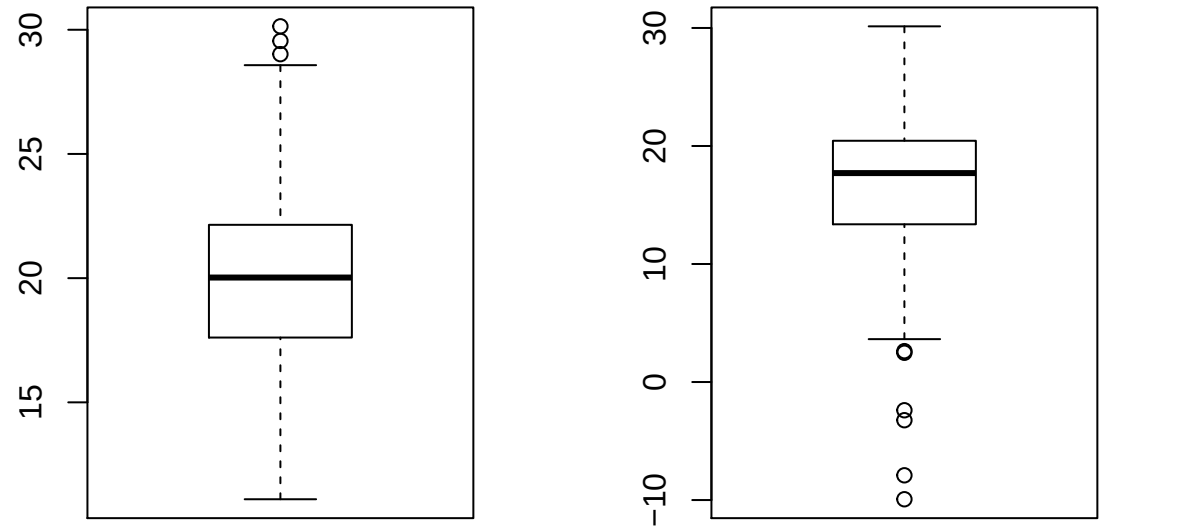
$$--b,\ c--$$

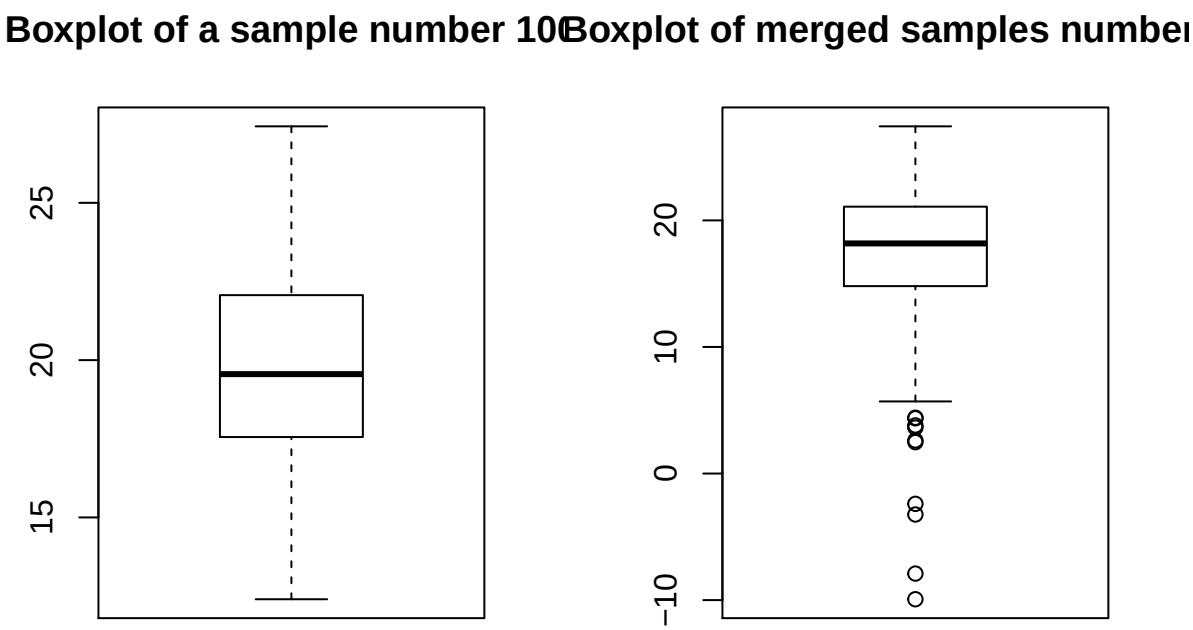
```
lower_bound <-  10
upper_bound <- 100

for(m in lower_bound:upper_bound) {
 new_sample <- rnorm(m, 20, 4)
 new_sample_merged <- c(sample_normal, new_sample)

 filename = str_pad(paste(m, ".png", sep = ""), pad = c("0"), 7, "left")
 png(filename)

 plot(new_sample_merged, xlim = c(0, 150), ylim = c(0, 30), main = "Red -> mean, blue -> median, green
 abline(v=mean(new_sample_merged),col="red")
 abline(v=median(new_sample_merged),col="blue")
 abline(v=var(new_sample_merged),col="green")

 dev.off()
 if(m %in% c(10, 50, 75, 100)){
   par(mfrow=c(1, 2))
   boxplot(new_sample)
   title(paste("Boxplot of a sample number", m))
```

```
   boxplot(new_sample_merged)
   title(paste("Boxplot of merged samples number", m))
 }
}
```




Boxplot of a sample number 50
Boxplot of merged samples numbe
10
15
20
25
30
−10
0
10
20
30


Boxplot of a sample number 75
Boxplot of merged samples numbe
30
25
20
15
30
20
10
0
−10



$--d--$

$**3**$

```
m = 10
sample_to_add <- rnorm(m, 20, 4)
```

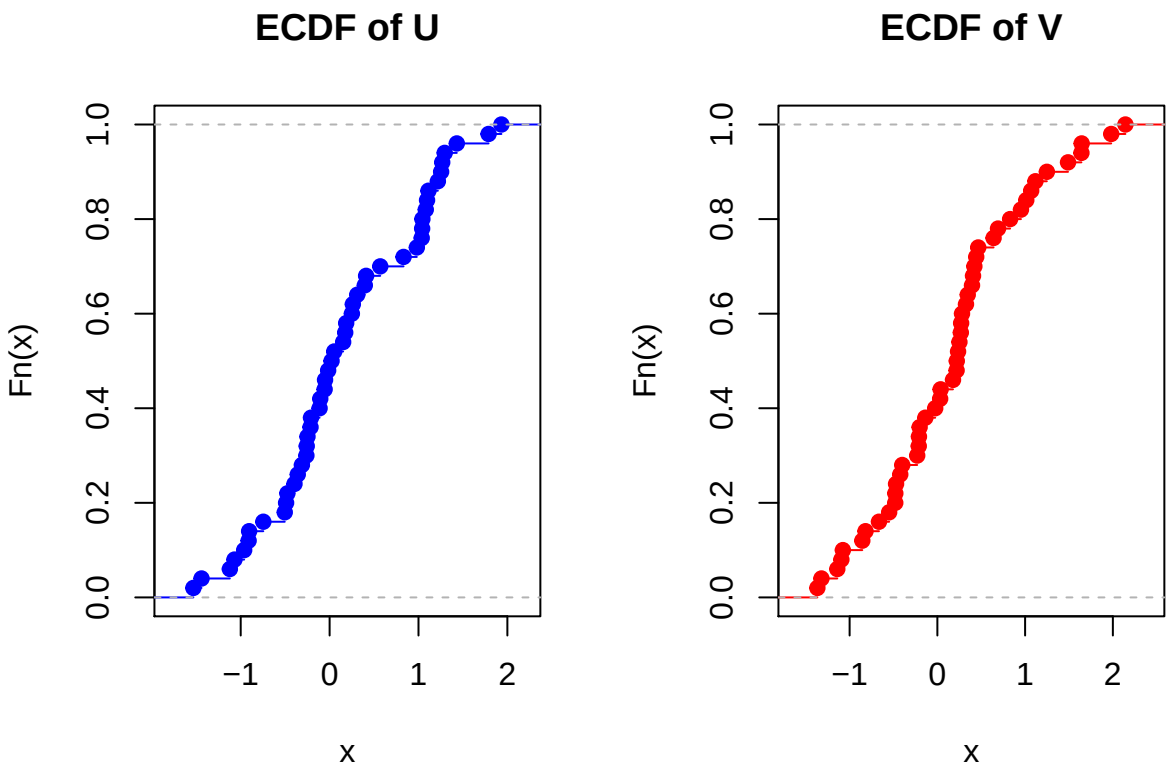
```
size <- 50
mean <- 0
sd <- 1
u = rnorm(size, mean, sd)
set.seed(2)
v = rnorm(size, mean, sd)
p = 0.95
u1 = u
v1 = p * u + sqrt(1 - p * p) * v
cor(u1, v1, method="pearson")
```

```
## [1] 0.9078329
```

```
cor(u1, v1, method="spearman")
```

```
## [1] 0.89503
```

```
par(mfrow = c(1, 2))
plot(ecdf(u1), main = "ECDF of U", col="blue")
plot(ecdf(v1), main = "ECDF of V", col="red")
```



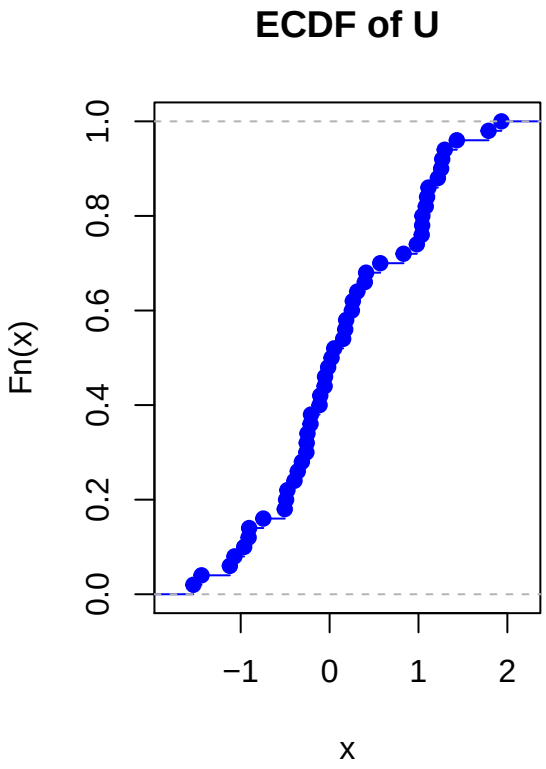

```
v1_squared <- v1 * v1
cor(u1, v1_squared, method="pearson")
```

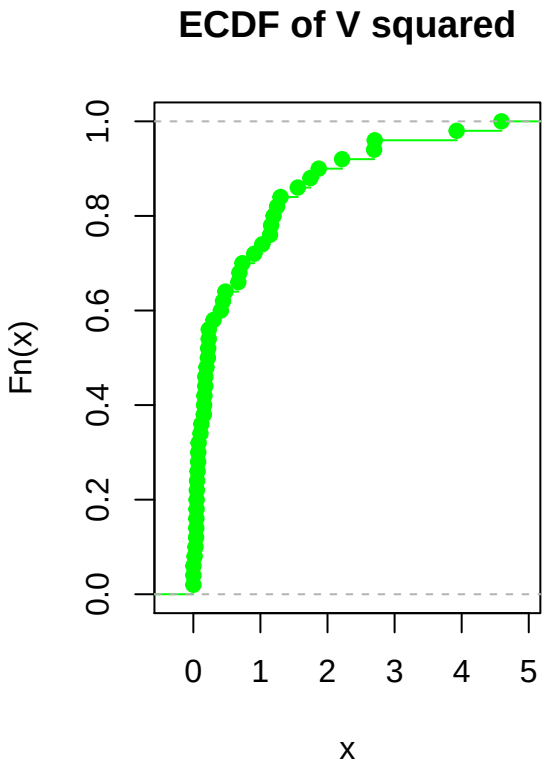
```
## [1] 0.3884649
```

```
cor(u1, v1_squared, method="spearman")
```

```
## [1] 0.1605282
```

```
par(mfrow = c(1, 2))
plot(ecdf(u1), main = "ECDF of U", col = "blue")
plot(ecdf(v1_squared), main = "ECDF of V squared", col = "green")
```


ECDF of U
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−1
0
1
2
x


ECDF of V squared
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
x