Empirical software engineering

Lab 1: Descriptive statistics, regression and hypothesis testing
Group 9

Anton Lutteman   Daniel Olsson   Gerson Silva Filho   Johan Mejborn

15 November 2020

Exercise 1 - Time to Develop

a) Descriptive data:

Mean = 244.625
Median = 231
Standard deviation = 83.4672591
Variance = 6966.7833333

b) What is being calculated?

What is being calculated is the **sample** standard deviation, since the company has provided the time spent only for 16 features chosen at random. The population would be if we had the time for all the features developed.

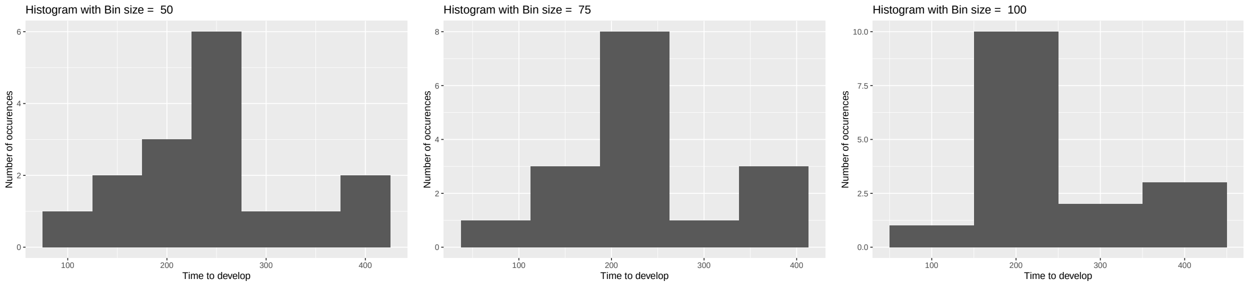
c) Hypothesis

The hypothesis is one tailed.

h0: mean(Time) <= 225 Hours
h1: mean(Time) > 225 Hours

d) Histogram

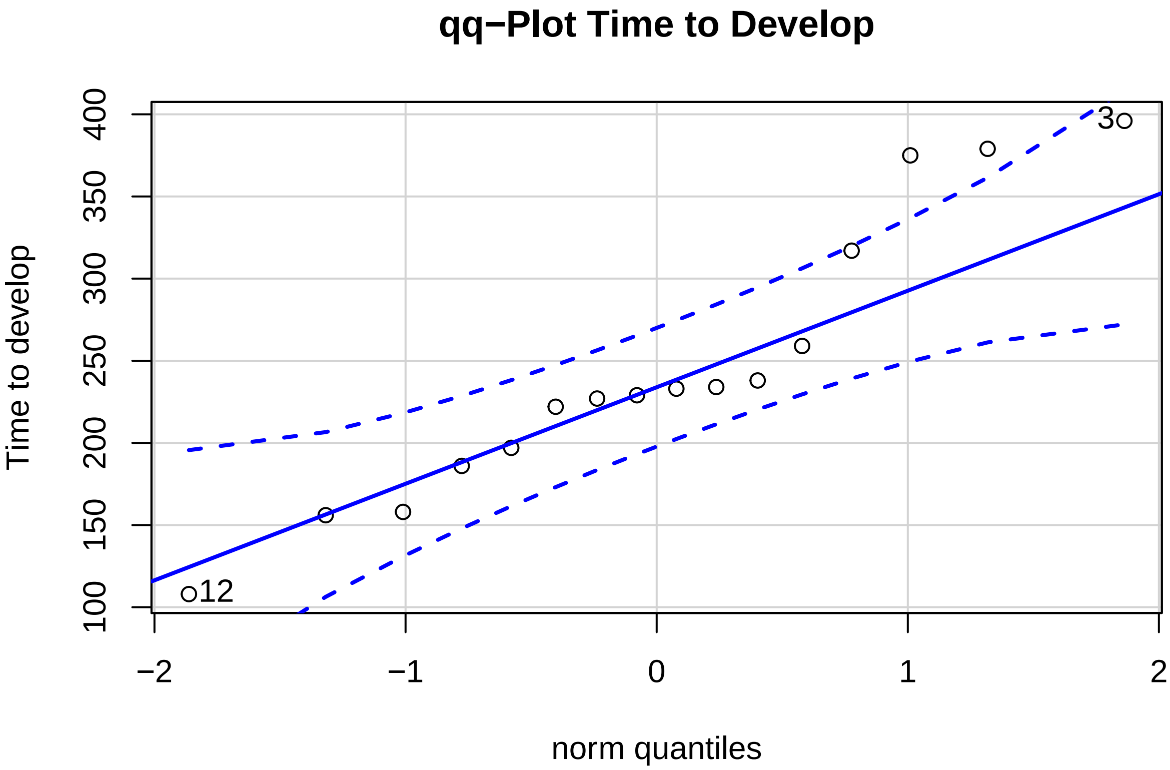
```r
dbin<-data.frame(d1)
histogram_with_bin <- function(bin_size) {
  title<- paste("Histogram with Bin size = ", bin_size)
  ggplot(data = dbin, aes(x=d1))+
  geom_histogram(binwidth = bin_size)+
  labs(title=title, x=’Time to develop’, y=’Number of occurences’)
}

histogram_with_bin(50)
histogram_with_bin(75)
histogram_with_bin(100)
```



Changing the amount of bins doesn't change the perspective drastically, but we can observe that most of the values are concentrated between 200 and 250, using the bin size of 50. When we used bigger values we would assume that most values were between 100 and 300 which gives us much less precision on that.

e) qq-Plot

```
car::qqPlot(d1,main = 'qq-Plot Time to Develop', ylab= 'Time to develop')
```



```
## [1]  3 12
```

f) Shapiro-Wilk test

```
shapiro.test(d1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d1
## W = 0.92033, p-value = 0.1708
```

Observing the QQ-plot of the data we can observe that data doesn't follow a normal distribution since there are points out of the 95% confidence interval that we are using in the plot.

The Shapiro-Wilk test also reinforce this non-normality of the data since it's W value is below 1. The p-value of 0.1708 is considered safe for this experiment since we don't have a huge sample.

g) One sample T-Test

```
t.test(d1, mu=225 ,conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  d1
## t = 0.94049, df = 15, p-value = 0.3619
## alternative hypothesis: true mean is not equal to 225
## 95 percent confidence interval:
##  200.1484 289.1016
## sample estimates:
## mean of x
##   244.625
```

Exercise 2 - Performance

a) Descriptive Statistics

```
psych::describeBy(df2$Time,df2$Group)
```

```
##
##  Descriptive statistics by group
## group: timeOptimized
##    vars  n mean   sd median trimmed  mad   min   max range  skew kurtosis   se
## X1    1 10   16 0.03  16.01   16.01 0.02 15.96 16.04  0.08 -0.43    -1.19 0.01
## ---------------------------------------------------------------------------------
## group: timeOriginal
##    vars  n  mean   sd median trimmed  mad   min   max range  skew kurtosis   se
## X1    1 10 16.02 0.03  16.02   16.02 0.04 15.96 16.05  0.09 -0.43    -1.28 0.01
```

b) Type of data

```
str(df2)
```

```
## tibble [20 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Group: Factor w/ 2 levels "timeOptimized",..: 2 1 2 1 2 1 2 1 2 1 ...
##  $ Time : num [1:20] 16 16 16 16 16 ...
```

```
df2$Group <- as.factor(df2$Group)
df2$Time <- as.numeric(df2$Time)

str(df2)
```

```
## tibble [20 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Group: Factor w/ 2 levels "timeOptimized",..: 2 1 2 1 2 1 2 1 2 1 ...
##  $ Time : num [1:20] 16 16 16 16 16 ...
```

c) Linear Model

```
lm(Time ~ Group,df2)
```

```
##
## Call:
## lm(formula = Time ~ Group, data = df2)
##
## Coefficients:
##       (Intercept)  GrouptimeOriginal
##             16.00               0.01
```

d) Is the factori "Group" statistically significant for this model?

```
t.test(formula = Time~Group,
       data=df2, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  Time by Group
## t = -0.79894, df = 18, p-value = 0.4347
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03629652  0.01629652
## sample estimates:
## mean in group timeOptimized  mean in group timeOriginal
##                      16.005                      16.015
```