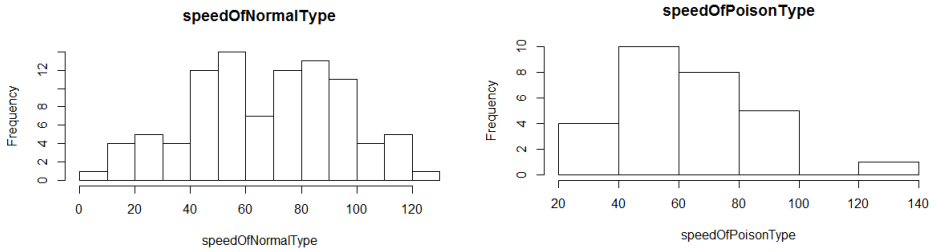
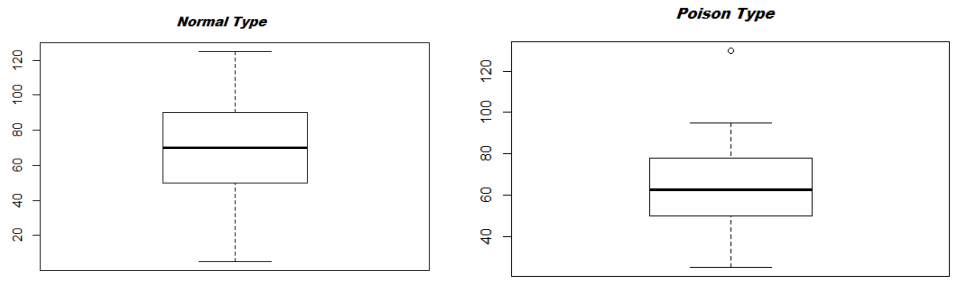
1 Vergelijken van een eigenschap tussen twee types

2 Normaliteit van twee variabelen

De eerste vraag gaat over twee gemiddelde waarden te vergelijken.De eigenschap waarvoor ik gemiddelde waarden moet vergelijken,is speed en twee types zijn PoisonType en Normaaaltype. Voordat wij een hypothese test opstellen , is het noodzakelijk om de normaliteit te controleren omdat de methode die wij gaan gebruiken werkt enkel met gegevens die normaal verdeeeld zijn. Op de basis van een histogram is het gemakkelijk te zien welk soort verdeling een variable heeft.Histogrammen geven een overzicht van de verdeling van variabelen.

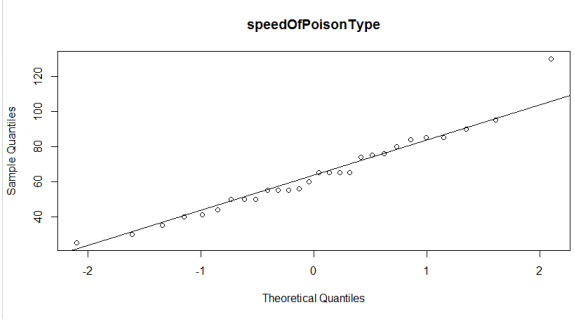


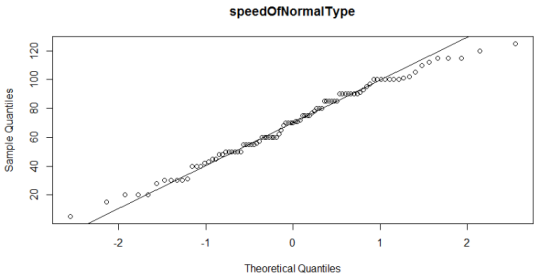
Als wij naar deze histogramen kijken , dan merken wij op dat het histogram van Normaal type een bimodale verdeling heeft en Poison type een rechts scheve verdeling volgt en wij kunnen ook zien dat de steekproef van Normaltype meeer gegevens bevat dan poison type. Op de basis van deze histogramen kunnen wij niet besluiten dat de gegevens normaal verdeeld zijn want als wij aantal klassen veranderen dan krijgen een heel andere figuur en wij kunnen ook niet zien met de behulp van histogram dat er uitschieters aanwezig zijn in onze gegevens. Het is dus beter om een boxplot te maken om te zien dat er geen uitschieters zijn.

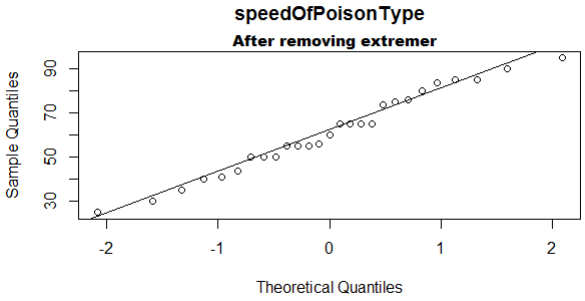


Met de behulp van bovenstaande figuren kunnen is het gemakkelijk te zien dat er een uitschieter aanwezig is in PoisonType.Als wij dit vergelijken met de histogram dan zien wij dat er ook een punt is dat een beetje afwijkt van andere gegevens. Bovenste whisker is ook een beetje korter dan de onderste whisker en mediaan is ook een beetje verschoven naar onderkant dus de verdeling is niet symmetrich.De boxplot van normaaltype laat zien dat deze verdeling ook niet echt symmetrisch is. De mediaan ligt in het midden maar de bovenste whisker

is korter dan de onderste whisker.Wij weten op dit moment dat de gegevens niet symmetrish verdeeld zijn.Om een echt idee te krijgen hoe ver gegevens liggen dan de verwachte plaats , maken wij gebruik van kwantielplot. Om de verwachte plaats te kennen tekenen wij ook een rechte(verwachte plaats) samen met kwantielplot.







Als wij naar kwantielplot zien merken wij op dat er een uitschieter aanwezig is en als wij deze uitschieter verwijderen vinden dan een punt dat onder de qqline ligt ,wij gaan deze uitschieter niet verwijderen als het onze resultaat niet beinvloedt en de varianties zijn niet bekend dus wij gaan t-test gebruiken en t-test zal niet zwaar beinvloed worden door een uitschieter(eigenschap van centrale limiet stlling).Er ligt ook een punt onder de rechte en op dezelfde afstand bij de figuur van poison type maar deze steekproef bevat meer gegevens daardoor werd het niet als een uitschieter aangeduid in boxplot.De gegevens zijn niet normaal verdeeld nog bestaat er een transformatie dus wij gaan nu controleren dat de normaliteit aanvardbaar is of niet! Wij gaan gebruik maken van Shapiro-Wilk test.Voodat wij een besslising nemen op de basis van hypothesetest, leggen wij uit wat het betekent in sectie: Betekenis van Hypothese op pagina 7.Als wij Shapiro-Wilk test uitvoeren op poison type dan vinden wij een P-waarde gelijk aan P-value = 0.3425.De betekenis van P-waarde is uitgelegd in sectie: Betekenis van P-value op pagina 7 . Omdat de nul-hypothese van deze test is dat de steekproef uit een normaal verdeelde populatie komt en P-Value is groter dan 0.05 , dus wij kunnen null-hypothese niet verwerpen. P waarde bij NormaalType is gelijk aan 0.3233 wat groter is dan 0.05 dus null-hypothese wordt niet verworpen.Wij hebben dus niet genoeg bewijzen gevonden om null-hypothese te verwerpen.Nu gaan wij variantie berekenen om over spreiding te

weten.

2.1 Vergelijking van Varianties

De variantie van Poison type is gelijk aan: 512.1799 en voor NormalType is gelijk aan: 728.3151.Als de variantie geijk zijn dan kunnen wij gepoolde varianties gebruiken. Wij stellen een hypothese op:
$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$
Wij berekenen de F waarde op de basis van varianties die gevonden hebben.
$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$
$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$
$F = \frac{S_1^2}{S_2^2}\ {}_{H_0}F_{n-1,n-2}$
F-waarde ligt niet te ver van 1 wat wij verwachten onder null-hypothese. Ratio of variance 0.7032 ligt in 95% betrouwbaarheidsinterval,$[0.3997 : - : 1.3792]$ en P-waarde 0.2991 is ook groter dan $\alpha = 0.05$ dat betekent dat er niet genoeg bewijzen zijn om null-hypothese te verwerpen dus wij nemen aan dat de varianties gelijk zijn.

2.2 Het gemiddelde van een variable voor twee types vergelijken

De hypothese is van de vorm:
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$
Het gemiddelde van NormalType is gelijk aan 69.65 en gemiddelde van poison Type is gelijk aan 63.57. Omdat wij juist een test uitgevoerd hebben voor varianties en besloten dat de varianties gelijk zijn , gebruiken wij gepoolde variantie.Gepoolde variantie wordt gegeven door: $S_p^2 = \frac{(n-1)S_1^2+(n-2)S_2^2}{n1+n2-2}$ en de test-statistiek wordt gegeven door $T = \frac{\bar{y_1}-\bar{y_2}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ Eerst berkenen wij gepoolde variantie en dit is gelijk aan(na het nemen van sqrt) $S_p = 26.28$ en $T = \frac{(69.65-63.57}{26.28\sqrt{1/93+1/28}})$.
$T = -1.0738$ en $\alpha = 0.05$ en $\alpha/2 = 0.025$ en de aanvaardingsgebied wordt gegeven door $[-t_{119,0.025}, t_{119,0.025}] = [-1.984, 1.984]$ en de t-waarde behoort tot de aanvardingsgebied .De betrouwbaarheidsinterval is gelijk aan
$[\bar{y_1} - \bar{y_2} \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}] = [-5.15, 17.31]$ en P-waarde is gelijk aan: $p = 0.28$ .Op $\alpha = 0.05$ kunnen wij de nullhypothese niet verwerpen omdat P-waarde groter is dan 0.05.

2.3 conclusie

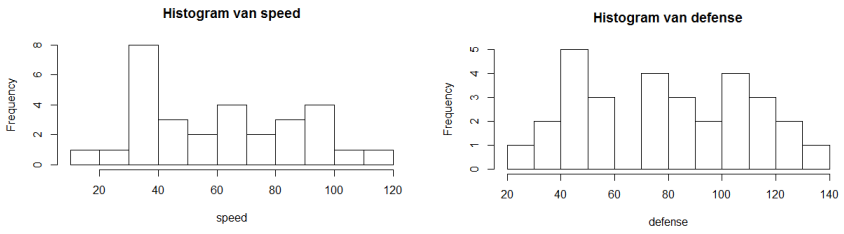
Omdat gemiddelde van een variable erg gevoelig is aan uitschieters,moeten wij ook controleren dat de uitschieter het resultaat beinvloedt of neit. Als wij een test zonder uitschieter uitvoeren , dan vinden dat de t-waarde gelijk is aan:

$T = -1.527$ en p-waarde is gelijk aan :$p = 0.129$. Wij concluderen dat wij op $\alpha = 0.05$ significantienieveau null-hypothese niet kunnen verwerpen en t-waarde behoort ook tot de aanvaardingsgebied.

3 Vraag 2: Correlatie tussen 2 eigenschappen van een bepaald type

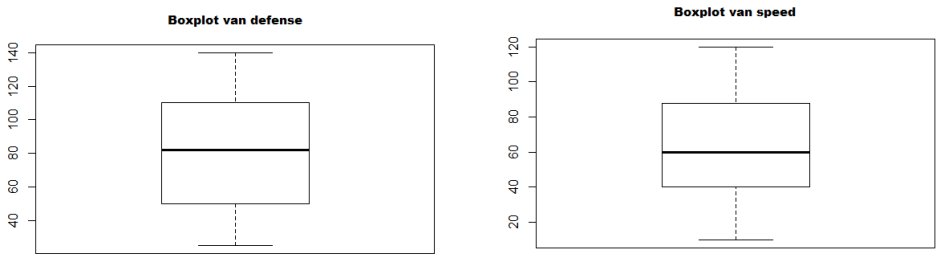
3.1 van twee variabelen apart

De tweede vraag gaat over correlatie tussen twee variabelen.Voor de groep 141 moet ik correlatie tussen speed en defense van ground type nagaan.Voordat wij een hypothese opstellen,moeten wij de normaliteit van beide variabelen controleren.Als een van de twee variabelen niet normaal verdeeld is, dan kunnen ze niet bivariate normaal verdeeld zijn dus wij moeten hun normaliteit apart controleren. Eerst maken wij een histogram om te zien hoe ze eigenlijk verdeeld zijn(welk soort verdeling, bimodaal, symmetrisch,scheef etc) daarna zullen wij zien met de behulp van boxplot dat er geen uitschieters zijn, een kwantielplot om te zien hoeveel ze van hun normaliteit afwijken en dan kunnen wij bivariate normaliteit controleren.



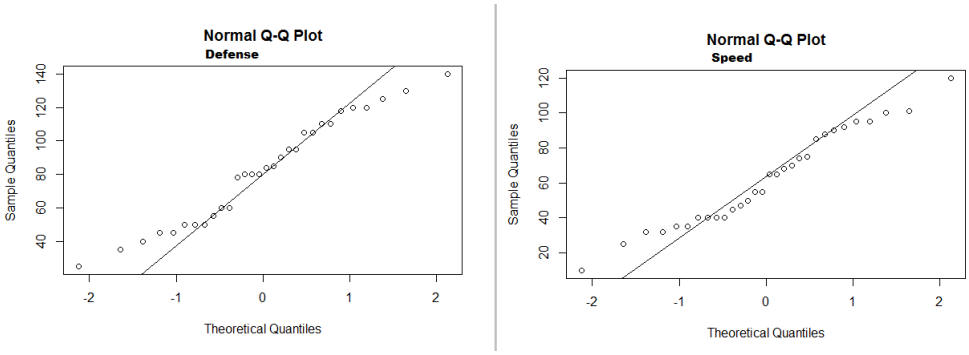
Als wij naar deze histogrammen kijken, merken wij op dat speed iets lager ligt dan defense en ook bij de histogram van defense missen er gegevens op een plaats, maar histogrammen zijn gebaseerd op klassen dus als het kan zijn dat als het aantal klassen toenemen , zien wij nog meer plaatsen waar geen gegevens zijn.De verdeling van defense ziet er beter uit dan speed. Normaliteit wordt beter gecontroleerd door een test en door kwantiel-plot maar wij maken eerst boxplots van beide eigenschappen.

3.2 Box plots



Het lijkt inderdaad dat de gegevens in defense beter verdeeld zijn dan in speed.Mediaan ligt in het midden beide whiskers zijn ongeveer gelijk en er is ook geen uitschieter. De boxplot van speed laat zien dat mediaan een beetje verschoven is naar onderkant dus het is een rechscheve verdeling.Er is geen uitschieter in de verdeling van speed.Wij controleren nu normaliteit op de basis van kwantiel-plot wat een beter inzicht geeft en dan voeren wij ook een Shapiro-wilk test.

3.3 kwantiel-plot



In de kwantiel-plot van defense merken wij op dat er op een plaats enkele punten ontbreken wat wij ook bij de histogram van defense gemerkt hadden en er liggen ook enkele punten ver van de qqline,maar het zijn zeker geen uitschieters want anders konden wij dat al zien in de boxplots. Kwantiel-plot van speed laat zien dat de punten tussen eerste een derde kwantile liggen dichter bij qqline maar wij moeten ook niet vergeten dat mediaan een beetje naar onderkant ligt in de boxplot van speed, aan de andere kant(in kwantiel-plot van defense) kunnen wij ook zien dat er punten tussen eerste en derde kwantiel in kwantiel-plot van defense liggen verder dan qqline en het zal de normailiteit heel erg beinvloeden omdat die punten liggen tussen eerste en derde kwantiel.

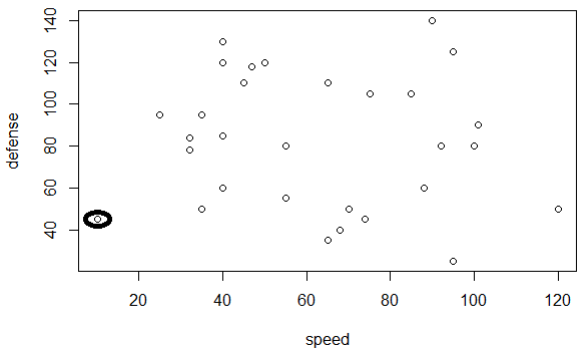
3.4 Shapiro-wilk test

Voeren wij nu Shapiro test op speed en wij vinden de p-waarde: $p = 0.38$ wat groter is dan $\alpha = 0.05$ dus wij kunnen de normaliteit van speed niet verwerpen

op $\alpha = 0.05$. De P-waarde voor defense is :$p = 0.31$ en het is ook groter dan $\alpha = 0.05$ en wij kunnen de normaliteit van defense ook niet verwerpen.P-waarde van defense is keiner dan de p-waarde van speed omdat wij al gezien hebben dat mediaan van defense ligt in midden(tussen eerste en derde kwantiel) waar gegevens op meer afstand liggen dan qqline.

3.5

Omdat de normaliteit van beide variabelen aanvaardbaar is, kunnen wij hun bivariate normaliteit controleren door een scatter plot te maken.



Uit de scatter-plot blijkt het dat er een klein negatief associatie. Als wij een punt dat zwart gemarkeerd verwijderen dan kunnen wij nog beter ellips tekenen maar de normaliteit is nog aanvardbaar en als wij deze punt gewoon verwijderen dan kunnen wij belangrijke informatie verliezen. Als wij pearson correlatie coefficient bereken dan vinden wij de waarde -0.078 wat inderdaad laat zien dat er een heel zwakke en negatieve correlatie is tussen beide variabelen maar het is maar een steekproef dus wij moeten een test-statstiek gebruiken om een beslissing te nemen op een significantie-niveau.

3.6 Speed en Defense zijn linear afhankelijk

Wij hebben juist gezien dat de pearson coefficient heel dicht bij 0 ligt dus wij stellen een hypothese op. De hypothese is:
$H_0 : p = 0$
$H_1 : p \neq 0$ waarbij H_0 beteket speed en defense zijn onafhankelijk en H_1 betekent speed en defense zijn linear afhankelijk.De test-statistiek is:$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ $H_0 t_{n-2}$.De correlatie-coefficient dat wij al berekend hebben is gelijk aan -0.078 en het is de waarde van R.Test-statistiek $t = \frac{(-0.078)\sqrt{30-2}}{\sqrt{1-(-0.078)^2}}$ $H_0 t_{n-2}$. De t-waarde is gelijk aan $t = -0.414$. We verwerpen de null-hypothese als $|t| > t_{n-2,\alpha/2}$ of als de p-waarde is kleiner dan de significantieniveau. De gegeven significantieniveau is gelijk aan $\alpha = 0.05$. P-waarde dat wij vinden bij $t = -0.414$ is gelijk aan P-value $= 0.93$.Omdat P-value is niet kleiner dan de significantieniveau , kunnen wij null-hypothese niet verwerpen en ook $|t = 0.41| < 2.048$.Wij besluiten dat

speed en Defense zijn onafhankelijk.

4 Betekenis van Hypothese

Null-hypothese zegt dat de parameter een bepaalde waarde aanneemt en het kan enkel verwerpen worden als de gegevens doen vermoeden dat er zeer onwaarschijnlijk verwachtingen zijn. Alternatieve hypothese stelt dat de gewenste waarde valt in bepaalde gebied(bereik) maar het kan ook negatie zijn van null-hypothese(bij two-tailed test) in dit geval alternatieve hypothese maakt twee gebieden op de basis van gewenste waarde en alternatieve hypothese stelt dat de waarde in linker of rechter gebied zal vallen maar niet gelijk aan de waarde wat null-hypothese stelt.

5 Betekenis van P-value

Voordat wij een significantie-test uitvoeren moeten wij een significantie-niveau kiezen, het vertelt hoe kleiner de p-waarde moet zijn om null-hypothese te verwerpen.Als de P-waarde kleiner of gelijk is aan significantie-niveau dan zeggen wij dat wij genoeg bewijzen hebben om null-hypothese te verwerpen.De p-waarde is kans , onder de null-hypothese, op de t waarde die in de richting van alternatieve-hypothese ligt.

6 Script

```
#variable for easy acces to Speed of Normal type
speedOfNormalType = Data$Speed[Data$Type == "Normal"]
#variable for easy acces to Speed of Poison type
speedOfPoisonType = Data$Speed[Data$Type == "Poison"]


#Histogram of poison type
hist(speedOfPoisonType,5,main = "speedOfPoisonType")
#Histogram of Normal Type, hier zijn verschillende klassen genomen want de grootte verschilt
#en dus om een beter overzicht te krijgen
hist(speedOfNormalType,5,main = "speedOfNormalType")


#BoxPlot of Normal Type
boxplot(speedOfNormalType)
#BoxPlot of Poison type
boxplot(speedOfPoisonType)


#Varaince of normal type
var(speedOfNormalType)
#varaince of Poison type
var(speedOfPoisonType)

#quantile plot of Poison type
qqnorm(speedOfPoisonType,main = "speedOfPoisonType")
#qq-line voor Poison Type
qqline(speedOfPoisonType)
```

```
#kwantiel-plot voor Normal Type
qqnorm(speedOfNormalType,main = "speedOfNormalType")
#qq-line voor NormalType
qqline(speedOfNormalType)


#Shapiro-test voor Speed of poisontype
shapiro.test(speedOfPoisonType)
#Shapiro-test voor Speed of speedOfNormalType
shapiro.test(speedOfNormalType)



#F-test voor speedOfPoisonType en normal type(variance test)
var.test(speedOfPoisonType, speedOfNormalType)


#Mean van SpeedOfNormalType
mNormal = mean(speedOfNormalType)
mNormal
#Mean van PoisonType
mPoison = mean(speedOfPoisonType)
mPoison


#uitschieter verwijderen uit Poisontype om het resultaat te kunnen zien in kwantiel-plot maar
#het was niet verwijderd.
d = speedOfPoisonType[-15]
```

```r
#gepoolde variantie
gepooldeS = (((93-1)*var(speedOfNormalType))+((28-1)*var(d)))/(92+27-2)
k = sqrt(gepooldeS)


#T-test voor de eerste vraag
T-value = ((mNormal - mPoison)/(k*sqrt(1/93 + 1/27)))
T-value

#p-waarde voor de t-value van eerste vraag
p-waarde =2*pt(abs(T),93+28-2,lower.tail = FALSE)
p-waarde


#Tweede Vraag


#defense , variable for easy acces
defense = Data$Defense[Data$Type == "Ground"]
#Speed , variable for easy acces
speed = Data$Speed[Data$Type == "Ground"]

#Histogram van speed
hist(speed, 9, main="Histogram van speed")
#Histogram van defense
hist(defese, 9, main="Histogram van defense")

#boxplot van speed
boxplot(speed)
#boxplot van defense
boxplot(defense)

#kwantiel-plot voor speed
qqnorm(speed)
#qqlin voor speed
qqline(speed)

#kwantiel-plot voor defense
qqnorm(defense)
#qqline voor defense
qqline(defense)


#Normality test voor speed
shapiro.test(speed)
#Normality test voor defense
shapiro.test(defense)



#correlatie tussen speed en defense
cor(speed, defense)

#kwantiel-plot voor defense en speed
plot(defense,speed,type="p")


#t-test werd op rekenmachine uitgevoerd en de waarde zijn al ingevuld in sectie waarin dat besproken is.
```

7 Uitvoer van programma

```
> #variable for easy acces to Speed of Normal type
> speedOfNormalType = Data$Speed[Data$Type == "Normal"]

> #variable for easy acces to Speed of Poison type
> speedOfPoisonType = Data$Speed[Data$Type == "Poison"]

> #Histogram of poison type
> hist(speedOfPoisonType,5,main = "speedOfPoisonType")

> #Histogram of Normal Type, hier zijn verschillende klassen genomen want de grootte verschilt
> #en dus om een beter overzicht te krijgen
> hist(spe .... [TRUNCATED]

> #BoxPlot of Normal Type
> boxplot(speedOfNormalType)

> #BoxPlot of Poison type
> boxplot(speedOfPoisonType)

> #Varaince of normal type
> var(speedOfNormalType)
[1] 728.3151

> #varaince of Poison type
> var(speedOfPoisonType)
[1] 512.1799

> #quantile plot of Poison type
> qqnorm(speedOfPoisonType,main = "speedOfPoisonType")

> #qq-line voor Poison Type
> qqline(speedOfPoisonType)

> #kwantiel-plot voor Normal Type
> qqnorm(speedOfNormalType,main = "speedOfNormalType")

> #qq-line voor NormalType
> qqline(speedOfNormalType)

> #Shapiro-test voor Speed of poisontype
> shapiro.test(speedOfPoisonType)

        Shapiro-Wilk normality test

data:  speedOfPoisonType
W = 0.95967, p-value = 0.3425


> #Shapiro-test voor Speed of speedOfNormalType
> shapiro.test(speedOfNormalType)

        Shapiro-Wilk normality test

data:  speedOfNormalType
W = 0.98417, p-value = 0.3233


> #F-test voor speedOfPoisonType en normal type(variance test)
> var.test(speedOfPoisonType, speedOfNormalType)

        F test to compare two variances

data:  speedOfPoisonType and speedOfNormalType
F = 0.70324, num df = 27, denom df = 92, p-value = 0.2991
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3997972 1.3729493
sample estimates:
ratio of variances
         0.7032394
```

```
> #Mean van SpeedOfNormalType
> mNormal = mean(speedOfNormalType)

> mNormal
[1] 69.65591

> #Mean van PoisonType
> mPoison = mean(speedOfPoisonType)

> mPoison
[1] 63.57143

> #uitschieter verwijderen uit Poisontype om het resultaat te kunnen zien in kwantiel-plot maar
> #het was niet verwijderd.
> d = speedOfPoisonType[-1 .... [TRUNCATED]

> d
 [1] 55 80 41 56 76 50 65 85 55 90 25 50 35 60 40 55 65 74 84 65 95 50 85 65 75 30 44

> #gepoolde variantie
> gepooldeS = (((93-1)*var(speedOfNormalType))+((28-1)*var(d)))/(92+27-2)

> k = sqrt(gepooldeS)

> #T-test voor de eerste vraag
> T = ((mNormal - mPoison)/(k*sqrt(1/93 + 1/27)))

> T
[1] 1.087669

> #p-waarde voor de t-value van eerste vraag
> p =2*pt(abs(T),93+28-2,lower.tail = FALSE)

> p
[1] 0.2789393
```

```
> #defense , variable for easy acces
> defense = Data$Defense[Data$Type == "Ground"]

> #Speed , variable for easy acces
> speed = Data$Speed[Data$Type == "Ground"]

> #Histogram van speed
> hist(speed, 9, main="Histogram van speed")

> #Histogram van defense
> hist(defense, 9, main="Histogram van defense")

> #boxplot van speed
> boxplot(speed)

> #boxplot van defense
> boxplot(defense)

> #kwantiel-plot voor speed
> qqnorm(speed)

> #qqlin voor speed
> qqline(speed)

> #kwantiel-plot voor defense
> qqnorm(defense)

> #qqline voor defense
> qqline(defense)

> #Normality test voor speed
> shapiro.test(speed)

        Shapiro-Wilk normality test

data:  speed
W = 0.96355, p-value = 0.3804
```

```
> #Normality test voor defense
> shapiro.test(defense)

        Shapiro-Wilk normality test

data:  defense
W = 0.96029, p-value = 0.3152


> #correlatie tussen speed en defense
> cor(speed, defense)
[1] -0.07872677

> #kwantiel-plot voor defense en speed
> plot(speed,defense,type="p")

> #t-test werd op rekenmachine uitgevoerd en de waarde zijn al ingevuld in sectie waarin dat besproken
is.
>
```