Milkyes Trial Data Analysis

Mo’men Mohamed

December 22, 2018

## What are we doing?

We are working on data collected from **Milkys** Farm as results of an experiment conducted on 40 Newly borne calves to measure the potentiality of **AXCELERA-C** to improve :

* Body weight gain
* weaning period reduction
* Health statues

## Why do we do this ?

This analysis is important for many reasons :

* Determining if there is any significant improvement due to using AXCELERA-C
* Giving insights about the best application of the Product
* directing the attention toward the precautions and issues which must be taken in consideration

## What is our stratigy and steps of Analysis ?

our work will be divided into 3 parts:

1- Data cleaning and processing –> to make the data ready for valid analysis

2- Data Analysis –> which divided into :

* Descriptive analysis –> data description and exploration
* Statistical testing –> for insuring the significant effect

### Data cleaning and processing

First of all we have to load all libraries which are necessary for our analysis

library(readxl)  
library(magrittr)  
library(dplyr)  
library(tidyr)  
library(ggplot2)  
library(psych)  
library(sjstats)

Then we will import the data and storing it in an object called **“Milkyes”**

Milkyes<-read\_excel(path = "./data sets/Milkyes.xlsx")  
head(Milkyes)

## # A tibble: 6 x 9  
## code type breed `birth date` `calving weight`  
## <dbl> <chr> <chr> <dttm> <dbl>  
## 1 15384 AXC brow~ 2018-10-17 00:00:00 39  
## 2 15385 AXC hols~ 2018-10-17 00:00:00 31  
## 3 15386 AXC hols~ 2018-10-17 00:00:00 29  
## 4 15387 AXC simm~ 2018-10-18 00:00:00 38  
## 5 15388 AXC hols~ 2018-10-18 00:00:00 31  
## 6 15389 AXC hols~ 2018-10-18 00:00:00 34  
## # ... with 4 more variables: `weight date` <dttm>, `weight 2` <dbl>,  
## # age <dbl>, `disease history` <chr>

Now we are going to look at the structure of the data for any modification

str(Milkyes)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 40 obs. of 9 variables:  
## $ code : num 15384 15385 15386 15387 15388 ...  
## $ type : chr "AXC" "AXC" "AXC" "AXC" ...  
## $ breed : chr "brown swiss" "holstein" "holstein" "simmental" ...  
## $ birth date : POSIXct, format: "2018-10-17" "2018-10-17" ...  
## $ calving weight : num 39 31 29 38 31 34 40 39 38 36 ...  
## $ weight date : POSIXct, format: "2018-12-15" "2018-12-15" ...  
## $ weight 2 : num 78 59 62 71 72 75 74 73 72 62 ...  
## $ age : num 59 59 59 58 58 58 58 58 58 57 ...  
## $ disease history: chr "en" "bl" NA "pn / bl" ...

This is very important to do some modification like :

* code : has to be *chr*
* type : has to be *factor*
* breed : has to be *factor*
* birth date : has to be *date*
* weight date : has to be *date*

Milkyes<-Milkyes%>%mutate(  
 code=as.character(code),  
 group=as.factor(type),  
 breed=as.factor(breed),  
 `birth date`=as.Date(`birth date`),  
 `weight date`=as.Date(`weight date`)  
)  
  
str(Milkyes)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 40 obs. of 10 variables:  
## $ code : chr "15384" "15385" "15386" "15387" ...  
## $ type : chr "AXC" "AXC" "AXC" "AXC" ...  
## $ breed : Factor w/ 3 levels "brown swiss",..: 1 2 2 3 2 2 2 2 2 1 ...  
## $ birth date : Date, format: "2018-10-17" "2018-10-17" ...  
## $ calving weight : num 39 31 29 38 31 34 40 39 38 36 ...  
## $ weight date : Date, format: "2018-12-15" "2018-12-15" ...  
## $ weight 2 : num 78 59 62 71 72 75 74 73 72 62 ...  
## $ age : num 59 59 59 58 58 58 58 58 58 57 ...  
## $ disease history: chr "en" "bl" NA "pn / bl" ...  
## $ group : Factor w/ 2 levels "AXC","CONT": 1 1 1 1 1 1 1 1 1 1 ...

I think that it will be useful to add a variable which is a factor of disease occurrence {0,1} in which {0} means no disease history and {1} means occurrence of disease

Milkyes<-Milkyes%>%mutate('D\_Occure'=sapply(`disease history`,FUN = function(x){  
 if(is.na(x)){0}  
 else{1}}))%>%mutate(D\_Occure=factor(D\_Occure,levels = c(0,1)))  
head(Milkyes)

## # A tibble: 6 x 11  
## code type breed `birth date` `calving weight` `weight date` `weight 2`  
## <chr> <chr> <fct> <date> <dbl> <date> <dbl>  
## 1 15384 AXC brow~ 2018-10-17 39 2018-12-15 78  
## 2 15385 AXC hols~ 2018-10-17 31 2018-12-15 59  
## 3 15386 AXC hols~ 2018-10-17 29 2018-12-15 62  
## 4 15387 AXC simm~ 2018-10-18 38 2018-12-15 71  
## 5 15388 AXC hols~ 2018-10-18 31 2018-12-15 72  
## 6 15389 AXC hols~ 2018-10-18 34 2018-12-15 75  
## # ... with 4 more variables: age <dbl>, `disease history` <chr>,  
## # group <fct>, D\_Occure <fct>

Now the Data is ready for analysis …

### Data Analysis

First of all lets illustrate and describe each variable :

#### Variables Describtion

* *code* – > the code number of each calf
* *group* –> weather the animal in control group or AXCELERA-C group
* *breed* –> the breed of the calf
* *birth date* –> the birth date of each animal
* *calving weight* –> the weight of each calf at birth
* *weight date* –> the date of second weight recording
* *weight 2* –> the second weight record after a period range from 53 to 62 day
* *age* –> age at *wieght 2*
* *disease history* –> type of disease affected the calves
* *D\_Occure* –> code illustrate disease occurrence or absence

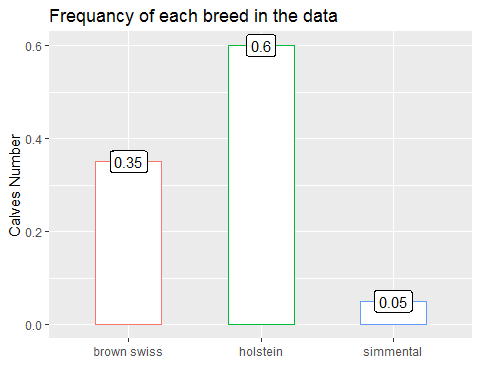
#### Exploring the Data

Now we will look at the data , exploring and graphing it

##### **Breed**

Lets first see the frequency and percentage of each breed representation

freq<-table(Milkyes$breed)%>%data.frame%>%mutate(perc=(Freq/sum(Freq)))  
freq%>%ggplot(aes(x=Var1,y=perc))+  
 geom\_col(aes(color=Var1),fill="white",width = .5,show.legend = FALSE)+  
 geom\_label(data=freq,aes(label=perc))+xlab(' ')+ylab("Calves Number")+  
 ggtitle(label = "Frequancy of each breed in the data")



We can see that the representation of Holstein is the most frequent **60%** then brown Swiss comes in the second level **30%** and the least rep. is Simmental **5%**

To Judge on these distribution as an issue we have to conduct chi\_square goodness of fit test as following :

goodness<-chisq.test(table(Milkyes$breed),p=rep(1/3,3),rescale.p = FALSE)  
goodness

##   
## Chi-squared test for given probabilities  
##   
## data: table(Milkyes$breed)  
## X-squared = 18.2, df = 2, p-value = 0.0001117

As *P-value* of chi square test is 1.116658110^{-4} and it is **< 0.05** , then we reject the the NULL hypothesis

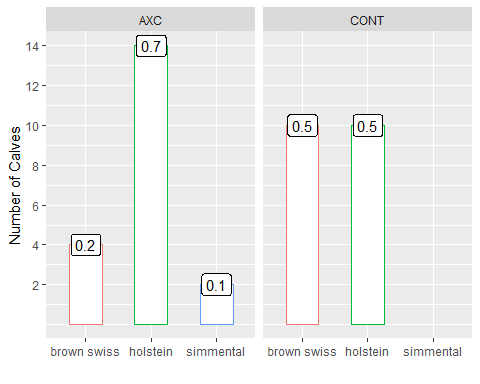
H0 : the breed of calves is equally distributed

and it is an issue in our sample

Another issue may face us if the distribution of breeds not independent from the two groups

but firstly lets visualize its distribution through groups :

grouped\_perc<-Milkyes%>%group\_by(group,breed)%>%summarize(Num=n())%>%mutate("%"=Num/sum(Num))  
  
Milkyes%>%ggplot(aes(x=breed))+  
 geom\_bar(aes(color=breed),fill="white",width = .5,show.legend = FALSE)+scale\_y\_continuous(breaks = 2\*c(1:7))+  
 facet\_wrap(.~group)+xlab("")+ylab("Number of Calves")+  
 geom\_label(data=grouped\_perc,aes(y=Num,label=`%`))



As we see here in the graph, there are a lack in the uniformity of the distribution of the breed at the two groups , and for more precision we can do a chi square for independence to insure that the distribution is independent from the group type and statistical uniform as following :

chi\_sq<-table(Milkyes$breed,Milkyes$group)%>%chisq.test()  
chi\_sq

##   
## Pearson's Chi-squared test  
##   
## data: .  
## X-squared = 5.2381, df = 2, p-value = 0.07287

Good News !!

we accept the Null hypothesis : H0: the breed distribution is independent from the group type as the *P-Value* 0.0728722 is **> 0.05**

**OK …**

##### **Calving Weight**

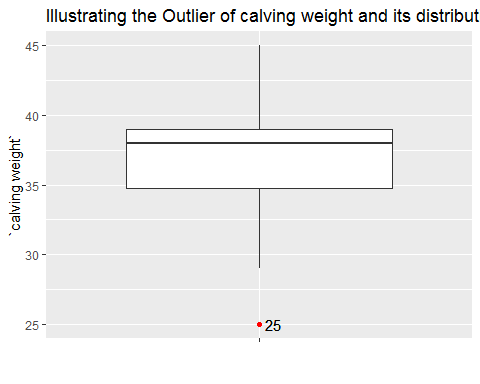
Now lets talk about the calving weight and its relationship with other factors like :

* breed
* incidence of disease occurrence
* Body weight at the second weight record

SO,

First of all lets test Normality and presence of outlier

Min<-Milkyes%>%filter(`calving weight` == min(`calving weight`))  
  
Milkyes%>%ggplot(aes("",`calving weight`))+  
 geom\_boxplot(outlier.colour = "red")+  
 geom\_text(data = Min,aes(label=`calving weight`),nudge\_x = .04)+  
 xlab("")+  
 ggtitle(label = "Illustrating the Outlier of calving weight and its distribution")

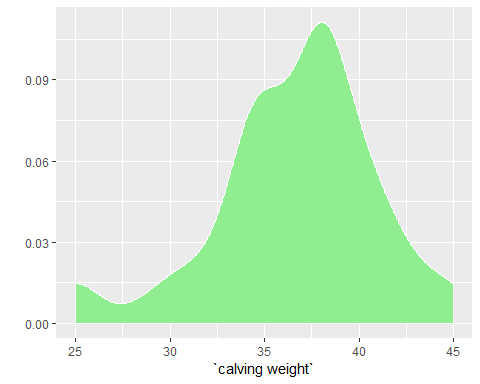


The Calving weight of *25 KG*  is an lower outlier which may affect on the results but we will keep it, as it may be useful in future analysis and create another object without outlier for statistical testing and replace the outlier with *NA*

Milkyes\_na<-Milkyes%>%mutate(`calving weight`=sapply(`calving weight`,function(x){  
 if(x == min(Milkyes$`calving weight`)){NA}  
 else{x}  
}))

Now lets take a look on Normality:

Milkyes%>%ggplot(aes(`calving weight`))+  
 geom\_density(fill="lightgreen",color="white",bins = 7)+  
 ylab("")



We here see that the Data roughly Normal with small left skew. Now we can see the summary of this variable :

summary(Milkyes$`calving weight`)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 25.00 34.75 38.00 36.58 39.00 45.00

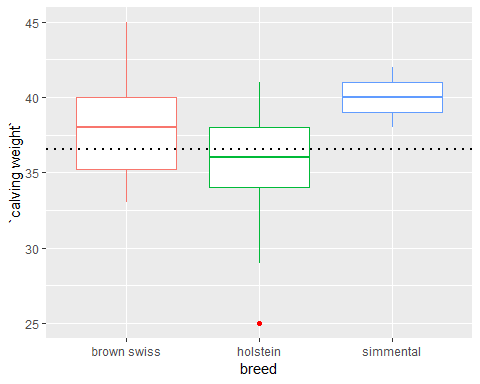
this summary illustrate the maximum and minimum weight and also the average

Now we will test the relationship with calving weight and other variables like:

###### *breed*

firstly we can compare between breeds in calving weight:

Milkyes%>%ggplot(aes(x=breed,y=`calving weight`))+  
 geom\_boxplot(aes(color=breed),show.legend = FALSE,outlier.color = "red")+  
 geom\_hline(aes(yintercept=mean(Milkyes$`calving weight`)),lty=3,lwd=.9)



From this box plot we can see that both *simmental* and *brown swiss* are above the average of calving weight unlike Holstein which is under the average

and this tell us that the calving weight might be affected by type of breed …

And to prove it statistically we can do one way ANOVA as below :

AOV<-aov(data = Milkyes,`calving weight`~breed)  
summary(AOV)

## Df Sum Sq Mean Sq F value Pr(>F)   
## breed 2 104.0 51.98 3.054 0.0592 .  
## Residuals 37 629.8 17.02   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Wow, the result here is so confusing as the P-Value is slightly more than **0.05**

So, we need here calculating Eta effect size

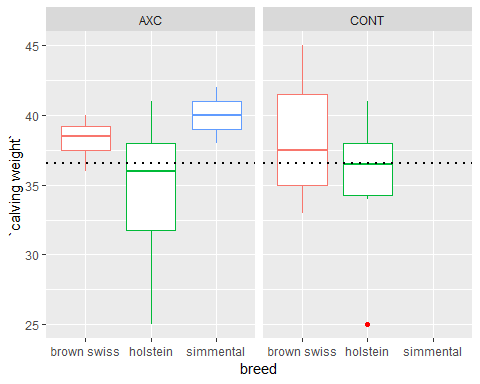
eta<-eta\_sq(AOV)  
(eta)

## term etasq  
## 1 breed 0.142

From Effect size Eta squared we can see that 14.2% of variation in calving weight is due to breed and it isn’t a percentage which we can neglect

Also to see the complete picture of this variation we need to insert another factor which is the group :

Milkyes%>%ggplot(aes(x=breed,y=`calving weight`))+  
 geom\_boxplot(aes(color=breed),show.legend = FALSE,outlier.color = "red")+  
 geom\_hline(aes(yintercept=mean(Milkyes$`calving weight`)),lty=3,lwd=.9)+  
 facet\_wrap(.~group)

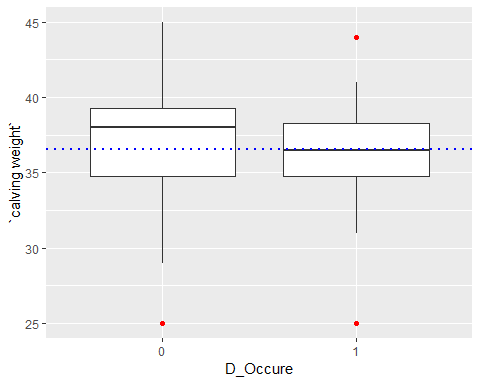


Great, in this plot we can see that the variation in AXCELERA-C group is larger than control one

###### *Incedence of Disease Occurance*

we can see this as following :

Milkyes%>%ggplot(aes(x=D\_Occure,y=`calving weight`))+  
 geom\_boxplot(outlier.color = "red")+  
 geom\_hline(aes(yintercept=mean(Milkyes$`calving weight`)),lty=3,lwd=.9,color="blue")



Actually to prove that the calving body weight has a significant effect on health statues you need to do t-test :

t.test(`calving weight`~ D\_Occure,data = Milkyes\_na)

##   
## Welch Two Sample t-test  
##   
## data: calving weight by D\_Occure  
## t = 0.50554, df = 34.518, p-value = 0.6164  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.747092 2.904986  
## sample estimates:  
## mean in group 0 mean in group 1   
## 37.47368 36.89474

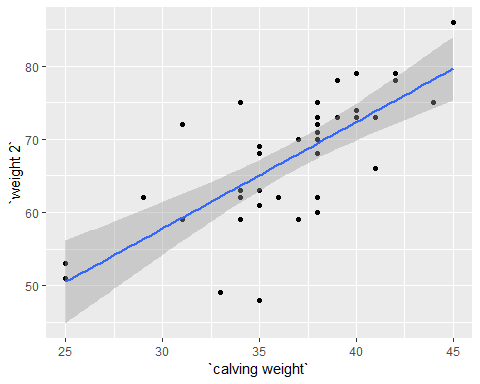
Good !! here we can exclude the calving weight as a factor affecting the incidence of disease in the future !!

###### *First weight*

Now we want to know the relationship between the calving weight and the second weight record

Lets plot and visualize it

Milkyes%>%ggplot(aes(x=`calving weight`,y=`weight 2`))+geom\_point()+geom\_smooth(method = 'lm')



You can see that there are a relationship between the calving weight and the second weight record

and we can statistically see that by calculating the correlation

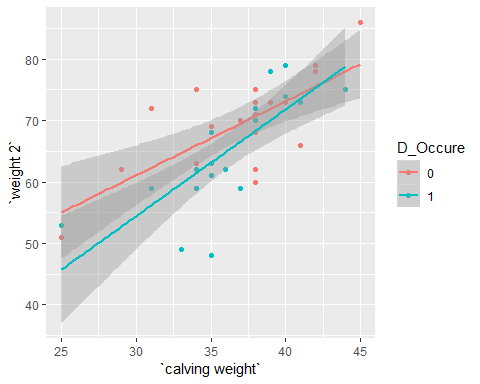
with(Milkyes,cor.test(`weight 2`,`calving weight`))

##   
## Pearson's product-moment correlation  
##   
## data: weight 2 and calving weight  
## t = 6.4199, df = 38, p-value = 1.514e-07  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5285694 0.8433290  
## sample estimates:  
## cor   
## 0.7213117

Great !! we have high correlation between these to variables !!

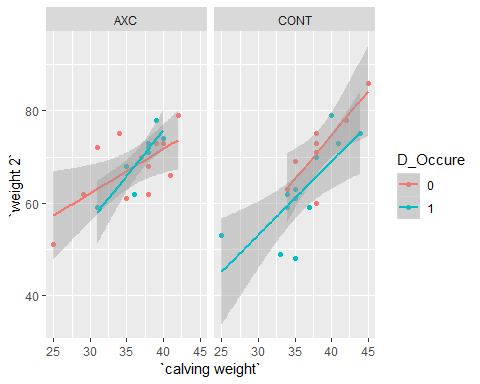
But do you think that this correlation might affected by the disease occurrence during this period !!

Milkyes%>%ggplot(aes(x=`calving weight`,y=`weight 2`,color=D\_Occure))+geom\_point()+geom\_smooth(method = "lm")



The intersection between “*lm*” lines here may be due to other factor like using *AXCELERA-C* so , lets see the effect of using it

Milkyes%>%ggplot(aes(x=`calving weight`,y=`weight 2`,color=D\_Occure))+geom\_point()+geom\_smooth(method = "lm")+  
 facet\_grid(.~group)



Great News !!!

this intersection is in the AXCELERA-C group and we can say that AXCELERA-C has an effect on infected animal to regain its body weight gain normally

but we can’t definitely conclude that Now

we will back to this point during discussing the difference between the two groups

##### **Groups differences**

**Now** we can talk about the core of our analysis *GROUPS* and its relation to the different variables

And we will discuss the difference between the two groups at:

* difference in calving body weight
* Incidence of disease occurrence
* difference in the second body weight record
* difference in average body weight

###### *difference in calving body weight*

I will test the difference in congruent with breed in each group :

ANOVA<-aov(`calving weight`~group\*breed,data = Milkyes)  
  
summary(ANOVA)

## Df Sum Sq Mean Sq F value Pr(>F)   
## group 1 5.6 5.62 0.313 0.5792   
## breed 2 99.5 49.77 2.773 0.0762 .  
## group:breed 1 0.4 0.44 0.024 0.8766   
## Residuals 35 628.2 17.95   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

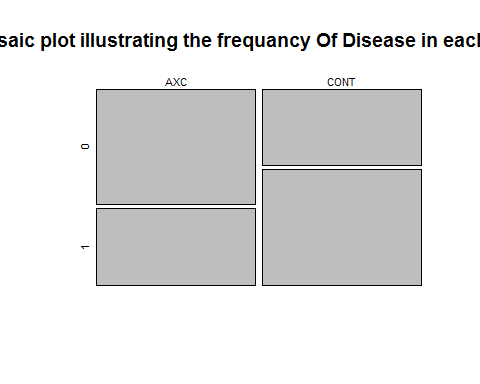
We here see the result of **Two way ANOVA** as there is no significant difference in calving weight between the two groups and also absence of moderating role of *breed* here

So, we can say that the two group are *matched* …

###### *Incedence of disease occurence*

After matching the two groups , Now lets compare between the two group in incidence of disease occurrence by *chi-squared* test after visualizing it :

table<-table(Milkyes$group,Milkyes$D\_Occure)  
plot(table,main="Mosaic plot illustrating the frequancy Of Disease in each group")



In this plot we see an obvious difference in cases of disease occurrence between the two group as the control group has more cases of disease

and now, it’s time of statistical testing

chisq.test(x = Milkyes$group,Milkyes$D\_Occure)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: Milkyes$group and Milkyes$D\_Occure  
## X-squared = 0.9, df = 1, p-value = 0.3428

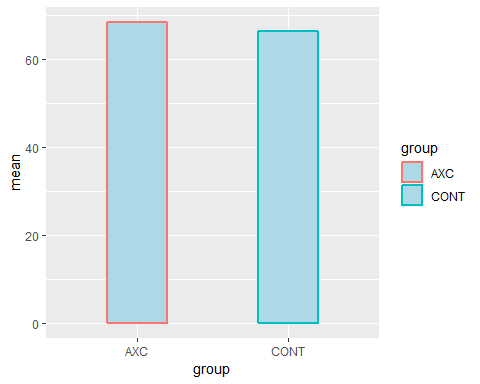
**Unfortunately** the difference between the two groups in incidence of disease occurrence is Not significant and P-value is > **0.05**

OK …

###### *difference in the second body weight record*

Lets test the difference in the second body weight record

Milkyes%>%group\_by(group)%>%summarize(mean=mean(`weight 2`))%>%ggplot(aes(x=group,y=mean,color=group))+  
 geom\_col(width = .4,fill="lightblue",lwd=.8)



Now lets use *t-test*

t.test(Milkyes$`weight 2`~Milkyes$group)

##   
## Welch Two Sample t-test  
##   
## data: Milkyes$`weight 2` by Milkyes$group  
## t = 0.77176, df = 33.774, p-value = 0.4456  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.512912 7.812912  
## sample estimates:  
## mean in group AXC mean in group CONT   
## 68.45 66.30

As we see here that the difference here is not significant also and the P-value is > **0.05**

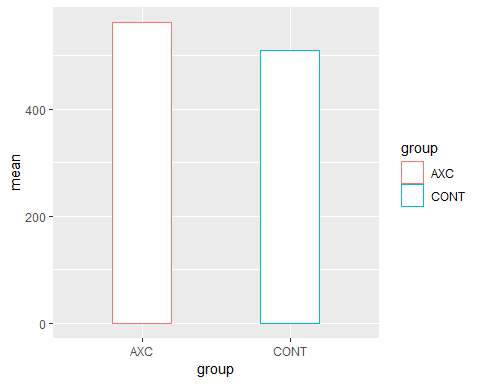
###### *difference in average body weight*

Lets make a variable called **DWG**

Milkyes<-Milkyes%>%mutate(DWG = round(((`weight 2`- `calving weight`)/as.numeric(`weight date`- `birth date`))\*1000,0)  
)

Then we can visualize it and test the difference between means

Milkyes%>%group\_by(group)%>%summarize(mean=mean(DWG))%>%ggplot(aes(x=group,y=mean,color=group))+geom\_col(width = .4,fill="white")



And we also will use t-test to know if the difference between the two groups is significant:

t.test(Milkyes$DWG~Milkyes$group)

##   
## Welch Two Sample t-test  
##   
## data: Milkyes$DWG by Milkyes$group  
## t = 1.6813, df = 35.361, p-value = 0.1015  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -10.91969 116.41969  
## sample estimates:  
## mean in group AXC mean in group CONT   
## 562.65 509.90

And here also we can see that the difference in average body weight is not significant

but there is another confirm needed by adding another factor \* incedence of disese occurence \* to this analysis and doing two way ANOVA

ANOVA2<-aov(DWG~D\_Occure\*group,data=Milkyes)  
  
summary(ANOVA2)

## Df Sum Sq Mean Sq F value Pr(>F)  
## D\_Occure 1 25654 25654 2.765 0.105  
## group 1 18922 18922 2.039 0.162  
## D\_Occure:group 1 23266 23266 2.507 0.122  
## Residuals 36 334033 9279

Also Here we can’t see any significant interaction between *incedence of disease occurence* and *DWG*

**Now** it is time of conclusion and recommendation

### Conclusion & Recommendations

#### Conclusion

As we saw in our analysis we have some conclusions To point here :

* There are a lack in distribution uniformity in our sample regarding breed
* the Distribution between group in each breed is statistically uniform and independent
* There are an observable difference in calving weight between each breed has to be taken in consideration
* AXCELERA-C group is less in number of cases which infected by disease but this difference isn’t significant
* the effect of calving body weight on Incidence of disease occurrence is not statistically significant
* There are a strong significant positive correlation between calving weight and second body weight record
* In Accelerate-C group there are an intersection between recovered calves after disease and disease free animal but this intersection is not significant
* the two groups are matched in calving body weight and number of breed “*statistically*”
* The difference in number of diseased animal is not significant
* The difference in the second body weight is not significant
* The Difference in the Average body weight is not significant between groups

#### Recomendation

* It is highly recommended to select a sample with a uniform distribution of breed type
* To do more Experiments on the potentiality of AXCELERA-C to buffer the bad effect of Disease decline growth rate

## REGARDS …..