problema3-ck4

Maelson Marques de Lima

21 de maio de 2016

# FPCC2-p2-c2

## ORGANIZANDO OS DADOS:

Descrição da atividade:

Escolha 2 das perguntas que você listou e construa o relatório final deste problema respondendo-as utilizando as ferramentas visuais e numéricas de análise descritiva e exploratória que você preferir. A partir dos resultados dessas 2 perguntas iniciais, derive mais duas e as responda.   
   
 Inicialmente a organização dos dados:

#Tabela mpvies  
movies <- read.csv("~/Rprojetos/Rprojects-fpcc2/bd-movies/movies.csv")  
#View(movies)  
  
#Tabela ratings  
ratings <- read.csv("~/Rprojetos/Rprojects-fpcc2/bd-movies/ratings.csv")  
  
#Dados Um --Base 01  
movies\_join\_ratings = merge(x = movies, y = ratings, by = "movieId", all.y = TRUE)  
#View(movies\_join\_ratings)  
  
#Dados Dois --Base 02 (Filmes Brasileiros Lançados - 1995 a 2014)  
filmes\_tratados<-read.csv("~/Rprojetos/Rprojects-fpcc2/bd-movies/filmes\_tratados.csv")  
#View(filmes\_tratados)  
  
  
##knitr::kable(filmes\_tratados[1:10, 1:11], caption = "Amostra de Filmes Tratados:")  
  
#Dados Três --BAse 03  
ratings\_filme<- read.csv("~/Rprojetos/Rprojects-fpcc2/bd-movies/ratings-por-filme.csv")   
#View(ratings\_filme)

**Minhas Perguntas**

\*\*Analisando a base de dados de filmes brasileiros, surgiram alguns questionamentos, inicialmente estes 2:\*\*  
  
   
 1. Qual gênero de filmes brasileiros gostam mais?  
   
 2. Qual gênero de filmes são mais rentáveis?  
   
 3. A média de público dos estados Nordestinos é maior que a de outras regiões?  
   
 4. Nos anos que houve decréscimo na produção de filmes, quais as estimativas?

## Respondendo questão 01:

**1. Qual gênero de filmes brasileiros gostam mais?**

##Inicilamente vamos entender a o atributo Gênero da Tabela de filmes tratados:  
  
table(filmes\_tratados$Genero)

Animação Documentário Ficção Videomusical   
 17 380 725 1

#Obtemos valores que podem atrapalhar na análise, pois a quantidade de filmes dos gêneros Animação e videomusical são muito baixos. mas Vídeos de animação tem valores altos de renda e público.Sendo assim, será removido apenas o gênero Videomusical.  
  
  
#filtro para remover estes gêneros:  
  
  
filmes\_tratados<-filter(filmes\_tratados,Genero!='Videomusical')  
  
#Agora limitar as informações com as quais irei tratar: Gênero e Publico e remover outliers(publico >=1000 ). Acredito que faça mais sentido, trabalhar com filmes que tiveram um público considerável.  
  
#names(filmes\_tratados)  
  
ft1<-filmes\_tratados[, c(3,8,11)]#Titulo, Genero e Publico  
  
#transforma caracteres  
ft1$Publico<-as.character(ft1$Publico)  
#estatística descritiva sobre a tabela agrupada:  
#View(ft1)  
  
ponto<-'\\.'  
  
for(i in 1:length(ft1$Publico)){  
 ft1[i,3]<-gsub(ponto,"",ft1[i,3])}  
  
  
  
#transforma numérico  
ft1$Publico<-as.numeric(ft1$Publico)  
  
ft1<-filter(ft1, ft1$Publico>1000)#miores que 1mil espectadores

\*\*Bootstrap:\*\*

Medias\_ft1<-ft1%>%group\_by(Genero)%>%summarise(mediap=mean(Publico), medianap=median(Publico))  
  
#Observem que os valores são bem próximos  
knitr::kable(Medias\_ft1, caption = "Média e Mediana da população")

Média e Mediana da população

|  |  |  |
| --- | --- | --- |
| Genero | mediap | medianap |
| Animação | 159868.75 | 45846 |
| Documentário | 16890.28 | 4495 |
| Ficção | 361431.58 | 33534 |

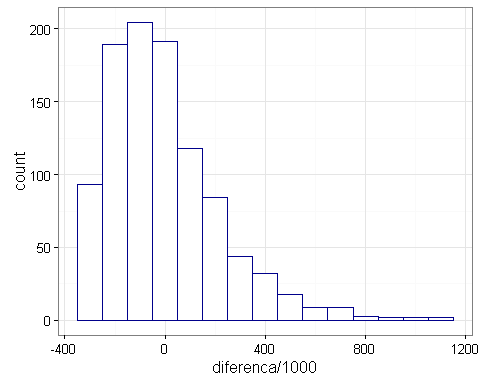
#Como podemos observar, Documentários tem uma média e Mediana muito baixa, sendo assim, executremos o IC apenas em Animações e Ficção. Além disso, existe uma dispareidade muito alta entre a média e a mediana. o que indica uma variância muito alta nestes valores, assim, é preferível tratar com a mediana.  
  
#Animação - Ficção (Mediana) -- ANIMAÇÃO > FICÇÃO  
dif\_mediana\_31 = Medias\_ft1[1,3] - Medias\_ft1[3,3]  
print(paste("Diferença na mediana observada:", round(dif\_mediana\_31),".Seria Melhor animação"))

[1] "Diferença na mediana observada: 12312 .Seria Melhor animação"

#Ficção - Animação (Média) -- FICÇÃO > ANIMAÇÃO  
dif\_media\_31=Medias\_ft1[3,2]-Medias\_ft1[1,2]  
print(paste("Diferença na media observada:", round(dif\_media\_31), ".Seria Melhor Ficção"))

[1] "Diferença na media observada: 201563 .Seria Melhor Ficção"

ft1\_31<-filter(ft1, Genero=='Animação'| Genero=="Ficção")  
  
# Retorna a diferença nas médias entre o grupos de y  
# nos índices com valor de x igual a baseline após embaralhar x.  
repete\_experimento\_de\_diff = function(x, y, baseline){  
 embaralhado = x[sample(NROW(x))]  
 baseline\_mean = mean(y[which(embaralhado == baseline)])  
 other\_mean = mean(y[which(embaralhado != baseline)])  
 return(other\_mean - baseline\_mean)  
}  
  
  
#EXPERIMENTO FICÇÃO > ANIMAÇÃO  
  
num\_experimentos = 1000#Rodando mil vezes  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = repete\_experimento\_de\_diff(ft1\_31$Genero, ft1\_31$Publico, "Ficção"))  
  
#Demora um absurdo!!!  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth = 100, colour = "darkblue", fill = "white")



#Tive que adicionar um valor alto para o binwith, as diferenças de valores variam muito em alguns casos.  
  
summary(experimentos)

i diferenca

Min. : 1.0 Min. :-344529  
 1st Qu.: 250.8 1st Qu.:-167230  
 Median : 500.5 Median : -41840  
 Mean : 500.5 Mean : -1879  
 3rd Qu.: 750.2 3rd Qu.: 106475  
 Max. :1000.0 Max. :1089514

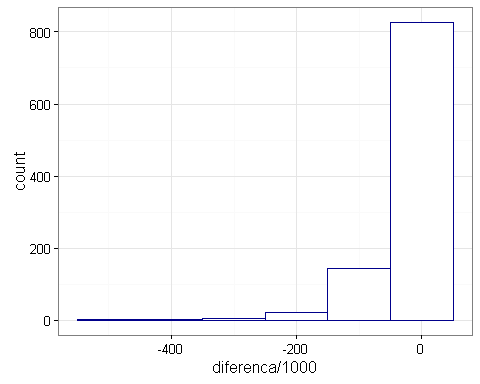
quantas\_aconteceram = length(which(experimentos$diferenca >= dif\_media\_31$mediap))  
probabilidade = quantas\_aconteceram / num\_experimentos  
  
print(paste("Este Evento ocorreu: ",quantas\_aconteceram,"vezes" ))

[1] "Este Evento ocorreu: 156 vezes"

print(paste("A probabilidade de ocorrência é de: ",probabilidade))

[1] "A probabilidade de ocorrência é de: 0.156"

#EXPERIMENTO aNIMAÇÃO > FICÇÃO -MEDIANA  
  
repete\_experimento\_de\_diff = function(x, y, baseline){  
 embaralhado = x[sample(NROW(x))]  
 baseline\_median = median(y[which(embaralhado == baseline)])  
 other\_median = median(y[which(embaralhado != baseline)])  
 return(other\_median - baseline\_median)  
}  
  
num\_experimentos = 1000#Rodando 100  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = repete\_experimento\_de\_diff(ft1\_31$Genero, ft1\_31$Publico, "Animação"))  
  
#Demora um absurdo!!!  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth = 100, colour = "darkblue", fill = "white")



summary(experimentos)

i diferenca

Min. : 1.0 Min. :-534808  
 1st Qu.: 250.8 1st Qu.: -29753  
 Median : 500.5 Median : -2018  
 Mean : 500.5 Mean : -19606  
 3rd Qu.: 750.2 3rd Qu.: 12566  
 Max. :1000.0 Max. : 32527

quantas\_aconteceram = length(which(experimentos$diferenca >= dif\_mediana\_31$medianap))  
probabilidade = quantas\_aconteceram / num\_experimentos  
  
print(paste("Este Evento ocorreu: ",quantas\_aconteceram,"vezes" ))

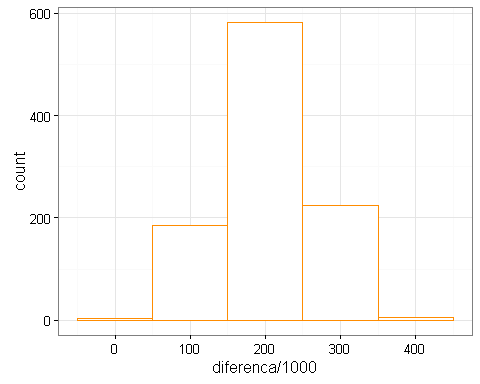
[1] "Este Evento ocorreu: 257 vezes"

print(paste("A probabilidade de ocorrência é de: ",probabilidade))

[1] "A probabilidade de ocorrência é de: 0.257"

### Garantir o IC:

#BOOTSTRAPPING IC FICÇÃO > ANIMAÇÃO --MÉDIA  
  
repeticoes = 1000 # Experimento com 1000 reptições  
  
#Gerando os valores para o bootrap para MÉDIA EM QUE FICÇÃO > ANIMAÇÃO  
exp\_com\_bootstrap <- function(x, y){  
 boot\_x <- sample(x, size = NROW(x), replace = TRUE) # aqui é o bootstrap  
 boot\_y <- sample(y, size = NROW(y), replace = TRUE) # de novo!  
 return(mean(boot\_x) - mean(boot\_y))  
}  
  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = exp\_com\_bootstrap(ft1\_31[ft1\_31$Genero=="Ficção",]$Publico,ft1\_31[ft1\_31$Genero=="Animação",]$Publico ))  
  
  
  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth = 100, colour = "darkorange", fill = "white")



summary(experimentos)

i diferenca

Min. : 1.0 Min. : 4351  
 1st Qu.: 250.8 1st Qu.:164787  
 Median : 500.5 Median :201841  
 Mean : 500.5 Mean :203145  
 3rd Qu.: 750.2 3rd Qu.:245889  
 Max. :1000.0 Max. :368720

# IC com 90%:   
#"Portanto em 90% com (margem +-0.5) das repetições os ficaram entre: "  
alpha = .1  
  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

5% 95%

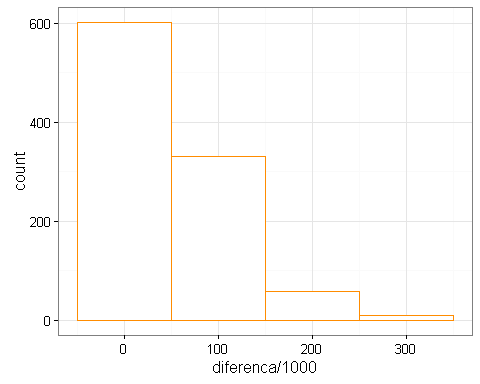
104143.8 295853.2

# IC com 95%:  
#Portanto em 95% das repetições, com margem de +-2,5 os valores p ficaram entre:  
alpha = .05  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

2.5% 97.5%

82649.78 310170.84

##BOOTSTRAPPING IC ANIMAÇÃO > FICAÇÃO --MEDIANA  
repeticoes = 1000 # Experimento com 1000 reptições  
  
#Gerando os valores para o bootrap para MÉDIA EM QUE FICÇÃO > ANIMAÇÃO  
exp\_com\_bootstrap <- function(x, y){  
 boot\_x <- sample(x, size = NROW(x), replace = TRUE) # aqui é o bootstrap  
 boot\_y <- sample(y, size = NROW(y), replace = TRUE) # de novo!  
 return(median(boot\_x) - median(boot\_y))  
}  
  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = exp\_com\_bootstrap(ft1\_31[ft1\_31$Genero=="Animação",]$Publico,ft1\_31[ft1\_31$Genero=="Ficção",]$Publico ))  
  
  
  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth = 100, colour = "darkorange", fill = "white")



summary(experimentos)

i diferenca

Min. : 1.0 Min. :-28714.5  
 1st Qu.: 250.8 1st Qu.: -236.8  
 Median : 500.5 Median : 13230.8  
 Mean : 500.5 Mean : 46171.1  
 3rd Qu.: 750.2 3rd Qu.: 76496.2  
 Max. :1000.0 Max. :333056.0

# IC com 90%:   
#"Portanto em 90% com (margem +-0.5) das repetições os ficaram entre: "  
alpha = .1  
  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

5% 95%

-11427.9 161039.2

# IC com 95%:  
#Portanto em 95% das repetições, com margem de +-2,5 os valores p ficaram entre:  
alpha = .05  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

2.5% 97.5%

-15292.02 191906.38

**Portanto, é possível dizer, que As pessoas gostam mais de filmes de Animação no brasil.**

## Respondendo questão 02:

**2. Qual gênero de filmes são mais rentáveis?**

**Tratamento**

#Agora limitar as informações com as quais irei tratar: Gênero e Renda e remover outliers(Renda >1000 ). Acredito que faça mais sentido, trabalhar com filmes que tiveram um Renda considerável.  
  
#names(filmes\_tratados)  
  
ft2<-filmes\_tratados[, c(3,8,10)]#Titulo, Genero e Renda  
  
  
#transforma caracteres  
ft2$Renda<-as.character(ft2$Renda)  
#estatística descritiva sobre a tabela agrupada:  
#View(ft2)  
  
ponto<-'\\.'  
  
for(i in 1:length(ft2$Renda)){  
 ft2[i,3]<-gsub(ponto,"",ft2[i,3])}  
  
virg<-'\\,'  
  
for(i in 1:length(ft2$Renda)){  
 ft2[i,3]<-gsub(virg,".",ft2[i,3])}  
  
#transforma numérico  
ft2$Renda<-as.numeric(ft2$Renda)  
#Pega apenas quem rendeu mais de mil reais.  
  
ft2<-filter(ft2, ft2$Renda>1000)#  
ft2$Renda<-round(ft2$Renda)  
  
media\_ft2<-ft2%>%group\_by(Genero)%>%summarise(mediar=mean(Renda),medianar=median(Renda))  
  
# E novamente observamos uma diferença entre os valores de Média e Mediana em relação à Animação e Ficção, com Documentário ficando para trás.  
  
knitr::kable(media\_ft2, caption = "Média e Mediana da Renda")

Média e Mediana da Renda

|  |  |  |
| --- | --- | --- |
| Genero | mediar | medianar |
| Animação | 1032396.0 | 333563 |
| Documentário | 97631.1 | 18789 |
| Ficção | 2638327.2 | 179521 |

**Iniciamos aqui as repetições para um probabilidade de ocorrências e depois um bootstrapping.**

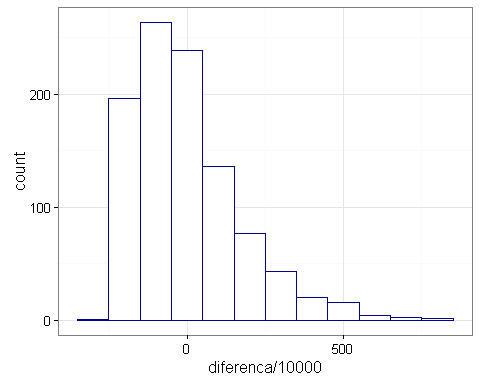
#Como podemos observar, Documentários tem uma média e Mediana muito baixa, sendo assim, executremos o IC apenas em Animações e Ficção. Além disso, existe uma dispareidade muito alta entre a média e a mediana. o que indica uma variância muito alta nestes valores.  
  
#Animação - Ficção (Mediana)  
dif\_mediana\_31 = media\_ft2[1,3] - media\_ft2[3,3]  
  
#Ficção - Animação (Média)  
dif\_media\_31=media\_ft2[3,2]-media\_ft2[1,2]  
#View(dif\_mediana\_31)  
#View(dif\_media\_31)  
  
print(paste("Diferença na mediana observada:", round(dif\_mediana\_31),".Seria Melhor animação"))

[1] "Diferença na mediana observada: 154042 .Seria Melhor animação"

print(paste("Diferença na media observada:", round(dif\_media\_31), "Seria Melhor Ficção"))

[1] "Diferença na media observada: 1605931 Seria Melhor Ficção"

ft2\_31<-filter(ft2, Genero=='Animação'| Genero=="Ficção")  
  
# Retorna a diferença nas médias entre o grupos de y  
# nos índices com valor de x igual a baseline após embaralhar x.  
repete\_experimento\_de\_diff = function(x, y, baseline){  
 embaralhado = x[sample(NROW(x))]  
 baseline\_mean = mean(y[which(embaralhado == baseline)])  
 other\_mean = mean(y[which(embaralhado != baseline)])  
 return(other\_mean - baseline\_mean)  
}  
  
  
#O experimento medindo as média para Ficção  
  
num\_experimentos = 1000#Rodando mil vezes  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = repete\_experimento\_de\_diff(ft2\_31$Genero, ft2\_31$Renda, "Ficção"))  
  
#Demora um absurdo!!!  
ggplot(experimentos, aes(x = diferenca/1e4)) +   
 geom\_histogram(binwidth = 100, colour = "darkblue", fill = "white")



#Tive que adicionar um valor alto para o binwith, as diferenças de valores variam muito em alguns casos.  
  
summary(experimentos)

i diferenca

Min. : 1.0 Min. :-2509433  
 1st Qu.: 250.8 1st Qu.:-1281607  
 Median : 500.5 Median : -366230  
 Mean : 500.5 Mean : -32751  
 3rd Qu.: 750.2 3rd Qu.: 835399  
 Max. :1000.0 Max. : 8264713

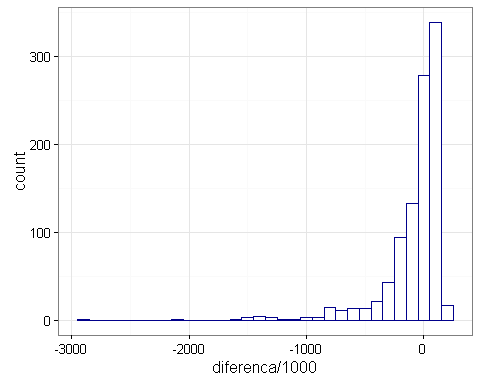
quantas\_aconteceram = length(which(experimentos$diferenca >= dif\_media\_31$mediar))  
probabilidade = quantas\_aconteceram / num\_experimentos  
  
print(paste("Este Evento ocorreu: ",quantas\_aconteceram,"vezes" ))

[1] "Este Evento ocorreu: 150 vezes"

print(paste("A probabilidade de ocorrência é de: ",probabilidade))

[1] "A probabilidade de ocorrência é de: 0.15"

#####  
  
#O mesmo experimento para mediana em relação à Animação  
  
repete\_experimento\_de\_diff = function(x, y, baseline){  
 embaralhado = x[sample(NROW(x))]  
 baseline\_median = median(y[which(embaralhado == baseline)])  
 other\_median = median(y[which(embaralhado != baseline)])  
 return(other\_median - baseline\_median)  
}  
  
num\_experimentos = 1000#Rodando 100  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = repete\_experimento\_de\_diff(ft2\_31$Genero, ft2\_31$Renda, "Animação"))  
  
#Demora um absurdo!!!  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth = 100, colour = "darkblue", fill = "white")



summary(experimentos)

i diferenca

Min. : 1.0 Min. :-2919955  
 1st Qu.: 250.8 1st Qu.: -135358  
 Median : 500.5 Median : -3652  
 Mean : 500.5 Mean : -80538  
 3rd Qu.: 750.2 3rd Qu.: 83405  
 Max. :1000.0 Max. : 163958

quantas\_aconteceram = length(which(experimentos$diferenca >= dif\_mediana\_31$medianar))  
probabilidade = quantas\_aconteceram / num\_experimentos  
  
print(paste("Este Evento ocorreu: ",quantas\_aconteceram,"vezes" ))

[1] "Este Evento ocorreu: 9 vezes"

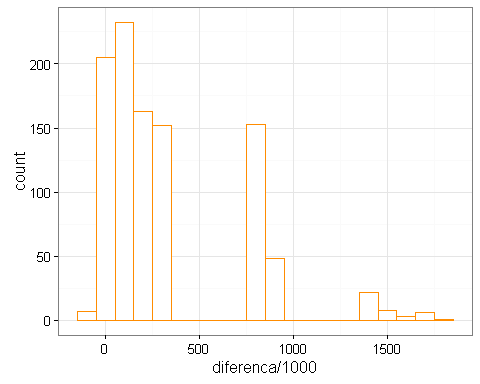
print(paste("A probabilidade de ocorrência é de: ",probabilidade))

[1] "A probabilidade de ocorrência é de: 0.009"

\*\*De acordo com os valores obtidos, podemos dizer que filmes de ficção tem maior rentabilidade do que filmes de animação.\*\*

### Garantir o IC:

##Intervalo de confiança para Animação > Ficção --MEDIANA  
  
repeticoes = 1000 # Experimento com 1000 reptições  
  
#Gerando os valores para o bootrap  
exp\_com\_bootstrap <- function(x, y){  
 boot\_x <- sample(x, size = NROW(x), replace = TRUE) # aqui é o bootstrap  
 boot\_y <- sample(y, size = NROW(y), replace = TRUE) # de novo!  
 return(median(boot\_x) - median(boot\_y))  
}  
  
  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = exp\_com\_bootstrap(ft2\_31[ft2\_31$Genero=="Animação",]$Renda,ft2\_31[ft2\_31$Genero=="Ficção",]$Renda ))  
  
  
  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth = 100, colour = "darkorange", fill = "white")



summary(experimentos)

i diferenca

Min. : 1.0 Min. : -82065  
 1st Qu.: 250.8 1st Qu.: 88681  
 Median : 500.5 Median : 161520  
 Mean : 500.5 Mean : 328063  
 3rd Qu.: 750.2 3rd Qu.: 302156  
 Max. :1000.0 Max. :1801439

# IC com 90%:   
#"Portanto em 90% com (margem +-0.5) das repetições os valores para esse genero (Animação) ficaram entre: "  
alpha = .1  
  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

5% 95%

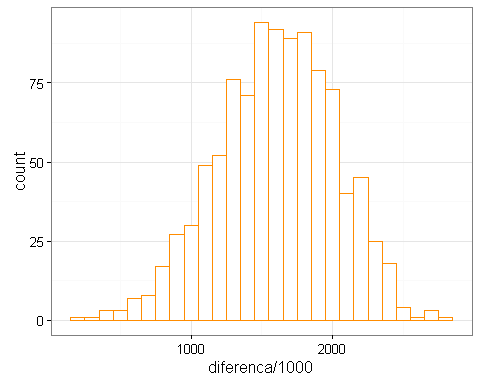
-5467.0 876561.2

# IC com 95%:  
#Portanto em 95% das repetições, com margem de +-2,5 os valores para esse gênero (Animação) ficaram entre:  
alpha = .05  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

2.5% 97.5%

-25614.17 1434717.78

###############  
  
  
  
##Intervalo de confiança para Ficção > aNIMAÇÃO --MÉDIA  
repeticoes = 1000 # Experimento com 1000 reptições  
  
#Gerenaod os valores para o bootrap  
exp\_com\_bootstrap <- function(x, y){  
 boot\_x <- sample(x, size = NROW(x), replace = TRUE) # aqui é o bootstrap  
 boot\_y <- sample(y, size = NROW(y), replace = TRUE) # de novo!  
 return(mean(boot\_x) - mean(boot\_y))  
}  
  
  
  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = exp\_com\_bootstrap(ft2\_31[ft2\_31$Genero=="Ficção",]$Renda,ft2\_31[ft2\_31$Genero=="Animação",]$Renda ))  
  
  
  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth = 100, colour = "darkorange", fill = "white")



summary(experimentos)

i diferenca

Min. : 1.0 Min. : 236667  
 1st Qu.: 250.8 1st Qu.:1314465  
 Median : 500.5 Median :1617748  
 Mean : 500.5 Mean :1603050  
 3rd Qu.: 750.2 3rd Qu.:1901958  
 Max. :1000.0 Max. :2750823

# IC com 90%:   
#"Portanto em 90% com (margem +-0.5) das repetições as médias ficaram entre: "  
alpha = .1  
  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

5% 95%

905313.5 2256783.8

# IC com 95%:  
#Portanto em 95% das repetições, com margem de +-2,5 as médias ficaram entre:  
alpha = .05  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

2.5% 97.5%

760463.4 2355223.6

## QUESTÃO 3

1. A média de público dos estados Nordestinos é maior que a de outras regiões?

filmes\_tratados<-filter(filmes\_tratados,Genero!='Videomusical')  
  
#Agora limitar as informações com as quais irei tratar: Estado e Público e remover outliers(publico >1000 ). Acredito que faça mais sentido, trabalhar com filmes que tiveram um público considerável.  
  
   
  
ft3<-filmes\_tratados[, c(3,6,11)]#Titulo, Genero e Publico  
  
View(ft3)  
#transforma caracteres  
ft3$Publico<-as.character(ft3$Publico)  
#estatística descritiva sobre a tabela agrupada:  
#View(ft1)  
  
ponto<-'\\.'  
  
for(i in 1:length(ft3$Publico)){  
 ft3[i,3]<-gsub(ponto,"",ft3[i,3])}  
  
  
  
#transforma numérico  
ft3$Publico<-as.numeric(ft3$Publico)  
  
ft3<-filter(ft3, ft3$Publico>1000)#miores que 1mil espectadores  
  
#  
medias\_ft3<-ft3%>%group\_by(UF)%>%summarise(mediap=mean(Publico), medianap=median(Publico))  
medias\_ft3<-filter(medias\_ft3, UF!='AM',UF!='MG / RJ' )#médias sem am e mg/rj  
#View(medias\_ft3)  
  
#Observem que os valores são bem próximos  
knitr::kable(medias\_ft3, caption = "Média e Mediana da população")

Média e Mediana da população

|  |  |  |
| --- | --- | --- |
| UF | mediap | medianap |
| BA | 12688.308 | 4047.0 |
| CE | 163504.222 | 3110.0 |
| DF | 16030.765 | 4386.0 |
| ES | 4371.500 | 4371.5 |
| GO | 2479.000 | 2479.0 |
| MA | 4880.000 | 4880.0 |
| MG | 20044.300 | 8496.0 |
| PE | 26391.500 | 9385.0 |
| PR | 16715.000 | 5168.0 |
| RJ | 392149.714 | 28869.0 |
| RS | 55345.513 | 9486.0 |
| SC | 8456.143 | 2920.0 |
| SE | 7517.000 | 7517.0 |
| SP | 149559.311 | 15934.0 |

#agora agupamento para receber as médias de cada região:  
  
#nordeste  
df1<-filter(ft3,UF=='BA'|UF=='PE'|UF=='CE'|UF=='SE'|UF=='MA')%>%summarise(Regiao='NORDESTE',somaP=sum(Publico),mediaP=mean(Publico))  
  
  
#sudeste  
df2<-filter(ft3,UF=='SP'|UF=='RJ'|UF=='MG'|UF=='ES')%>%summarise(Regiao='SUDESTE',somaP=sum(Publico),mediaP=mean(Publico))  
  
#sul  
df3<-filter(ft3,UF=='SC'|UF=='RS'|UF=='PR')%>%summarise(Regiao='SUL',somaP=sum(Publico),mediaP=mean(Publico))  
  
Media\_regiao<-df1  
Media\_regiao<-rbind(Media\_regiao,df2)  
Media\_regiao<-rbind(Media\_regiao,df3)  
  
#As médias da região sudeste são muito altas, o que já responderia a questão inicial,  
knitr::kable(Media\_regiao, caption = "Média e Mediana da população")

Média e Mediana da população

|  |  |  |
| --- | --- | --- |
| Regiao | somaP | mediaP |
| NORDESTE | 2071147 | 51778.68 |
| SUDESTE | 231471983 | 293002.51 |
| SUL | 2418248 | 41693.93 |

Como as médias das regiões sudeste já se mostram muito superiores às das outras regiões, Farei um pouco diferente e irei comparar apenas entre os estados do SP e CE que se mostram mais parecidos em suas médias, mas tem dispareidade nas medianas.

#filtro para CE e SP  
 ft3<-filter(ft3, UF=="CE" | UF=="SP")  
 #View(novo\_ft3)  
  
#Agrupando as médias de ft3  
medias\_ft3<-ft3%>%group\_by(UF)%>%summarise(mediap=mean(Publico), medianap=median(Publico))  
 #View(medias\_ft3)  
  
#Existem diferença entre as medianas e médias dos estados:   
  
#Ceará > São paulo--média  
dif\_media\_12=medias\_ft3[1,2]-medias\_ft3[2,2]  
print(paste("Diferença na media observada:", round(dif\_media\_12), ".Seria Melhor Ceará"))

[1] "Diferença na media observada: 13945 .Seria Melhor Ceará"

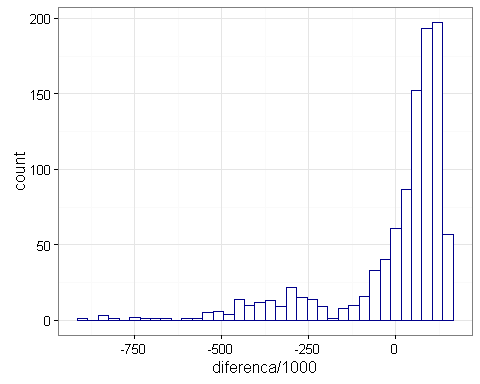
#São paulo> ceará -- mediana  
dif\_mediana\_21 = medias\_ft3[2,3] - medias\_ft3[1,3]  
print(paste("Diferença na mediana observada:", round(dif\_mediana\_21),".Seria Melhor São paulo"))

[1] "Diferença na mediana observada: 12824 .Seria Melhor São paulo"

**Iniciando as probabilidades**

**O experimento medindo as médias CE**

# Retorna a diferença nas médias entre o grupos de y  
# nos índices com valor de x igual a baseline após embaralhar x.  
repete\_experimento\_de\_diff = function(x, y, baseline){  
 embaralhado = x[sample(NROW(x))]  
 baseline\_mean = mean(y[which(embaralhado == baseline)])  
 other\_mean = mean(y[which(embaralhado != baseline)])  
 return(other\_mean - baseline\_mean)  
}  
  
  
#O experimento medindo as médias CE  
  
num\_experimentos = 1000#Rodando mil vezes  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = repete\_experimento\_de\_diff(ft3$UF, ft3$Publico, "CE"))  
  
#Demora um absurdo!!!  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth =30 , colour = "darkblue", fill = "white")



#Tive que adicionar um valor alto para o binwith, as diferenças de valores variam muito em alguns casos.  
  
summary(experimentos)

i diferenca

Min. : 1.0 Min. :-909250  
 1st Qu.: 250.8 1st Qu.: -16009  
 Median : 500.5 Median : 65980  
 Mean : 500.5 Mean : -1290  
 3rd Qu.: 750.2 3rd Qu.: 105707  
 Max. :1000.0 Max. : 150591

quantas\_aconteceram = length(which(experimentos$diferenca >= dif\_media\_12$mediap))  
probabilidade = quantas\_aconteceram / num\_experimentos  
  
print(paste("Este Evento ocorreu: ",quantas\_aconteceram,"vezes" ))

[1] "Este Evento ocorreu: 689 vezes"

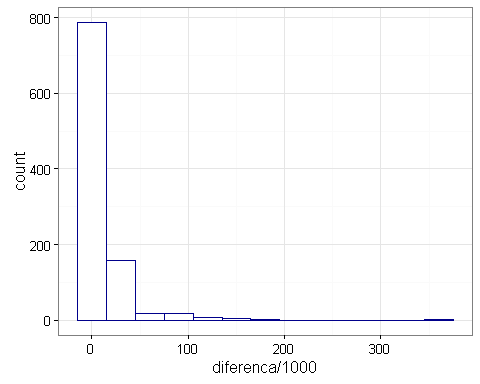
print(paste("A probabilidade de ocorrência é de: ",probabilidade))

[1] "A probabilidade de ocorrência é de: 0.689"

#De acordo com os valores obtidos, podemos dizer que filmes cearenses tem uma público maior do que filmes Paulistas de acordo com a Analisando a Média

**O experimento medindo as médianas SP**

################################  
  
  
  
#O experimento medindo as médianas SP  
  
repete\_experimento\_de\_diff = function(x, y, baseline){  
 embaralhado = x[sample(NROW(x))]  
 baseline\_median = median(y[which(embaralhado == baseline)])  
 other\_median = median(y[which(embaralhado != baseline)])  
 return(other\_median - baseline\_median)  
}  
  
  
num\_experimentos = 1000#Rodando mil vezes  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = repete\_experimento\_de\_diff(ft3$UF, ft3$Publico, "SP"))  
  
#Demora um absurdo!!!  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth =30 , colour = "darkblue", fill = "white")



#Tive que adicionar um valor alto para o binwith, as diferenças de valores variam muito em alguns casos.  
  
summary(experimentos)

i diferenca

Min. : 1.0 Min. :-14234  
 1st Qu.: 250.8 1st Qu.: -7597  
 Median : 500.5 Median : -84  
 Mean : 500.5 Mean : 8827  
 3rd Qu.: 750.2 3rd Qu.: 11623  
 Max. :1000.0 Max. :354762

quantas\_aconteceram = length(which(experimentos$diferenca >= dif\_mediana\_21$medianap))  
probabilidade = quantas\_aconteceram / num\_experimentos  
  
print(paste("Este Evento ocorreu: ",quantas\_aconteceram,"vezes" ))

[1] "Este Evento ocorreu: 225 vezes"

print(paste("A probabilidade de ocorrência é de: ",probabilidade))

[1] "A probabilidade de ocorrência é de: 0.225"

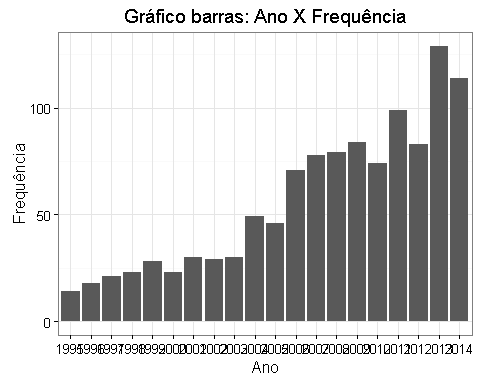
#De acordo com os valores obtidos, podemos dizer que filmes Paulista tem uma público maior do que filmes Cearenses se analisarmos a Mediana.

## QUESTÃO 4

1. Nos anos que houve decréscimo na produção de filmes, quais as estimativas ou Qual o gênero mais assistido?

Se observarmos curidadosamente as quantidades de lançamentos em cada ano, vamos perceber que em alguns anos houve decréscimo de lançamento. Acredito ser possível traçar um relacionamente entre alguns fatores de econômia, por exemplo. Mas no meu caso irei apenas demostrar quais os filmes que as pessoas preferiram nestes anos.

#tabela com valores eplotagem  
  
x<-table(filmes\_tratados$AnoLancamento)  
  
x<-as.data.frame(x)  
  
ggplot(data = x,aes(x= x$Var1, y= x$Freq) )+geom\_bar(stat = "identity")+ylab("Frequência")+xlab("Ano")+ggtitle("Gráfico barras: Ano X Frequência" )



#organizando dados  
  
ft4<-filmes\_tratados[, c(2,3,8,10,11)]#Titulo, Genero e Publico  
#View(ft4)  
  
#limpando e tratando os valores  
  
ft4$Publico<-as.character(ft4$Publico)  
  
  
#remove pontos de Publico  
ponto<-'\\.'  
  
for(i in 1:length(ft4$Publico)){  
 ft4[i,5]<-gsub(ponto,"",ft4[i,5])}  
  
ft4$Publico<-as.numeric(ft4$Publico)  
  
#remove pontos e virgulas de --renda  
ft4$Renda<-as.character(ft4$Renda)  
ponto<-'\\.'  
  
for(i in 1:length(ft4$Renda)){  
 ft4[i,4]<-gsub(ponto,"",ft4[i,4])}  
  
virg<-'\\,'  
  
for(i in 1:length(ft4$Renda)){  
 ft4[i,4]<-gsub(virg,".",ft4[i,4])}  
  
#transforma numérico  
ft4$Renda<-as.numeric(ft4$Renda)  
  
#Pega apenas quem rendeu mais de mil reaise teve mais de mil espectadores  
  
ft4<-ft4[, c(1,3,4,5)]  
ft4<-filter(ft4,Renda>1000, Publico>1000)  
  
  
media\_ft4<-group\_by(ft4,AnoLancamento)%>%summarise(mediaR=mean(Renda),mediaP=mean(Publico))  
View(media\_ft4)  
  
 knitr::kable(media\_ft4, caption = "Médias de Renda e Publico")

Médias de Renda e Publico

|  |  |  |
| --- | --- | --- |
| AnoLancamento | mediaR | mediaP |
| 1995 | 1048649.1 | 234179.14 |
| 1996 | 278656.1 | 62946.18 |
| 1997 | 788782.7 | 178614.90 |
| 1998 | 809421.9 | 188285.09 |
| 1999 | 1009949.4 | 243605.20 |
| 2000 | 1436821.4 | 288239.50 |
| 2001 | 1349197.0 | 264935.50 |
| 2002 | 1404177.9 | 256053.96 |
| 2003 | 4666351.1 | 768648.48 |
| 2004 | 2223506.1 | 336576.04 |
| 2005 | 1663813.3 | 236700.63 |
| 2006 | 1291148.7 | 176162.11 |
| 2007 | 1120537.8 | 143652.59 |
| 2008 | 1048302.8 | 136681.22 |
| 2009 | 2274010.5 | 278679.16 |
| 2010 | 3757844.3 | 421650.47 |
| 2011 | 2409900.4 | 261858.64 |
| 2012 | 3456776.4 | 340229.52 |
| 2013 | 3311759.2 | 306618.82 |
| 2014 | 2207233.2 | 188173.61 |

Irei trabalhar apenas com os anos em que houve diminuição da quantidade de filmes em relação ao seu ano anterior, neste caso são 06: 2000, 2002, 2005, 2010,2012, 2014.

ft4<-filter(ft4, AnoLancamento=='2000'| AnoLancamento=='2002'|AnoLancamento=='2005'|AnoLancamento=='2010'|AnoLancamento=='2012'|AnoLancamento=='2014' )  
  
#Médias de Renda e Publico dos anos de decréscimo  
medias\_ft4<-group\_by(ft4,AnoLancamento)%>%summarise(mediaR=mean(Renda),mediaP=mean(Publico))  
  
knitr::kable(medias\_ft4, caption = "Médias de Renda e Publico dos anos de decréscimo")

Médias de Renda e Publico dos anos de decréscimo

|  |  |  |
| --- | --- | --- |
| AnoLancamento | mediaR | mediaP |
| 2000 | 1436821 | 288239.5 |
| 2002 | 1404178 | 256054.0 |
| 2005 | 1663813 | 236700.6 |
| 2010 | 3757844 | 421650.5 |
| 2012 | 3456776 | 340229.5 |
| 2014 | 2207233 | 188173.6 |

#Aqui é possível perceber que dentro desses anos os filmes de ficção ciêntífica tiveram tanto renda quando Público bem maior que os demais.  
  
medias\_ft4<-group\_by(ft4, Genero)%>%summarise(mediaR=mean(Renda),mediaP=mean(Publico))  
#View(medias\_ft4)  
knitr::kable(medias\_ft4, caption = "Médias de Renda e Publico dos anos dos gêneros")

Médias de Renda e Publico dos anos dos gêneros

|  |  |  |
| --- | --- | --- |
| Genero | mediaR | mediaP |
| Animação | 858428.7 | 114146.83 |
| Documentário | 170079.6 | 19930.38 |
| Ficção | 3574002.9 | 399576.81 |

#Como foram poucos filmes de Animação (apenas 06), irei retirálos do experimento, e agir em cima apenas de documentário e Ficção.  
  
ft4<-filter(ft4,Genero!="Animação")

Agora veremos as estimativas de público e rendas em anos de decrescimo apenas para os gêneros Documentário e Ficção:

medias\_ft4<-group\_by(ft4,AnoLancamento)%>%summarise(mediaR=mean(Renda),mediaP=mean(Publico))  
  
knitr::kable(medias\_ft4, caption = "Médias de Renda e Publico dos anos de decréscimo")

Médias de Renda e Publico dos anos de decréscimo

|  |  |  |
| --- | --- | --- |
| AnoLancamento | mediaR | mediaP |
| 2000 | 1436821 | 288239.5 |
| 2002 | 1404178 | 256054.0 |
| 2005 | 1602021 | 228140.7 |
| 2010 | 3757844 | 421650.5 |
| 2012 | 3576557 | 351953.4 |
| 2014 | 2279624 | 194268.5 |

#Aqui é possível perceber que dentro desses anos os filmes de ficção ciêntífica tiveram tanto renda quando Público bem maior que os demais.  
  
medias\_ft4<-group\_by(ft4, Genero)%>%summarise(mediaR=mean(Renda),mediaP=mean(Publico))  
#View(medias\_ft4)  
  
medias\_ft4<-group\_by(ft4, Genero)%>%summarise(mediaR=mean(Renda),mediaP=mean(Publico))  
knitr::kable(medias\_ft4, caption = "Médias de Renda e Publico dos anos dos gêneros")

Médias de Renda e Publico dos anos dos gêneros

|  |  |  |
| --- | --- | --- |
| Genero | mediaR | mediaP |
| Documentário | 170079.6 | 19930.38 |
| Ficção | 3574002.9 | 399576.81 |

#Ficção - Documentário Média Público  
dif\_mediap\_21 = medias\_ft4[2,3] - medias\_ft4[1,3]  
print(paste("Diferença na mediana observada:", round(dif\_mediap\_21),".Seria Melhor Ficção"))

[1] "Diferença na mediana observada: 379646 .Seria Melhor Ficção"

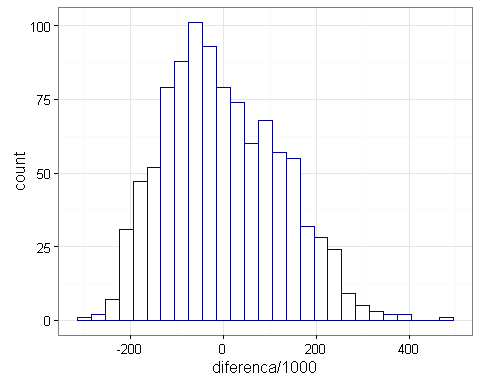
#Ficção - Documentário Média Renda  
dif\_mediar\_21=medias\_ft4[2,2]-medias\_ft4[1,2]  
print(paste("Diferença na media observada:", round(dif\_mediar\_21), ".Seria Melhor Ficção"))

[1] "Diferença na media observada: 3403923 .Seria Melhor Ficção"

**Iniciando as probabilidades**

PROBABILIDADE PARA MÉDIAS DE PÚBLICO:

# Retorna a diferença nas médias entre o grupos de y  
# nos índices com valor de x igual a baseline após embaralhar x.  
  
repete\_experimento\_de\_diff = function(x, y, baseline){  
 embaralhado = x[sample(NROW(x))]  
 baseline\_mean = mean(y[which(embaralhado == baseline)])  
 other\_mean = mean(y[which(embaralhado != baseline)])  
 return(other\_mean - baseline\_mean)  
}  
  
  
#O experimento medindo as médias Ficção --PÚBLICO  
  
num\_experimentos = 1000#Rodando mil vezes  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = repete\_experimento\_de\_diff(ft4$Genero, ft4$Publico, "Ficção"))  
  
#Demora um absurdo!!!  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth =30 , colour = "darkblue", fill = "white")



#Tive que adicionar um valor alto para o binwith, as diferenças de valores variam muito em alguns casos.  
  
summary(experimentos)

i diferenca

Min. : 1.0 Min. :-286720.6  
 1st Qu.: 250.8 1st Qu.: -92808.2  
 Median : 500.5 Median : -15281.3  
 Mean : 500.5 Mean : -462.2  
 3rd Qu.: 750.2 3rd Qu.: 91204.4  
 Max. :1000.0 Max. : 465036.3

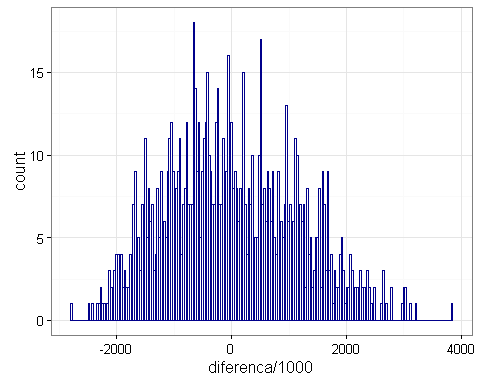
quantas\_aconteceram = length(which(experimentos$diferenca >= dif\_mediap\_21$mediaP))  
probabilidade = quantas\_aconteceram / num\_experimentos  
  
print(paste("Este Evento ocorreu: ",quantas\_aconteceram,"vezes" ))

[1] "Este Evento ocorreu: 2 vezes"

print(paste("A probabilidade de ocorrência é de: ",probabilidade))

[1] "A probabilidade de ocorrência é de: 0.002"

#Neste experimento apenas uma houve apenas uma vez em que a diferença no experimento foi maior do que a média.  
#Portanto é centro que o valor do público de filmes de Ficção é meior que os de Documentários  
  
##########################  
  
repete\_experimento\_de\_diff = function(x, y, baseline){  
 embaralhado = x[sample(NROW(x))]  
 baseline\_mean = mean(y[which(embaralhado == baseline)])  
 other\_mean = mean(y[which(embaralhado != baseline)])  
 return(other\_mean - baseline\_mean)  
}  
  
  
#O experimento medindo as médias Ficção --RENDA  
  
num\_experimentos = 1000#Rodando mil vezes  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = repete\_experimento\_de\_diff(ft4$Genero, ft4$Renda, "Ficção"))  
  
#Demora um absurdo!!!  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth =30 , colour = "darkblue", fill = "white")



#Tive que adicionar um valor alto para o binwith, as diferenças de valores variam muito em alguns casos.  
  
summary(experimentos)

i diferenca

Min. : 1.0 Min. :-2778449  
 1st Qu.: 250.8 1st Qu.: -845670  
 Median : 500.5 Median : -73955  
 Mean : 500.5 Mean : 6550  
 3rd Qu.: 750.2 3rd Qu.: 832153  
 Max. :1000.0 Max. : 3846993

quantas\_aconteceram = length(which(experimentos$diferenca >= dif\_mediar\_21$mediaR))  
probabilidade = quantas\_aconteceram / num\_experimentos  
  
print(paste("Este Evento ocorreu: ",quantas\_aconteceram,"vezes" ))

[1] "Este Evento ocorreu: 1 vezes"

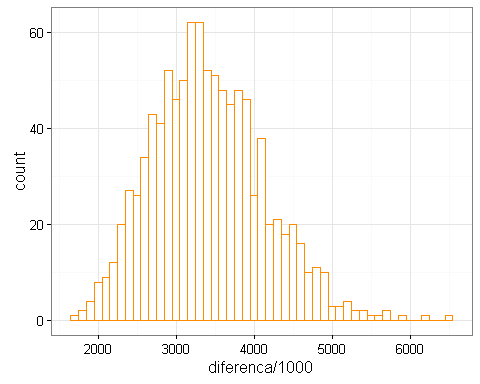
print(paste("A probabilidade de ocorrência é de: ",probabilidade))

[1] "A probabilidade de ocorrência é de: 0.001"

#Neste experimento apenas não houve vez em que a diferença no experimento foi maior do que a média.  
#Portanto é certo que o valor da Renda de filmes de Ficção é meior que os de Documentários

**Medindo o IC:**

##Intervalo de confiança para Ficção > aNIMAÇÃO --MÉDIA  
repeticoes = 1000 # Experimento com 1000 reptições  
  
#Gerenaod os valores para o bootrap  
exp\_com\_bootstrap <- function(x, y){  
 boot\_x <- sample(x, size = NROW(x), replace = TRUE) # aqui é o bootstrap  
 boot\_y <- sample(y, size = NROW(y), replace = TRUE) # de novo!  
 return(mean(boot\_x) - mean(boot\_y))  
}  
  
  
  
experimentos = data\_frame(i = 1:num\_experimentos)  
experimentos = experimentos %>%   
 rowwise() %>%   
 mutate(diferenca = exp\_com\_bootstrap(ft4[ft4$Genero=="Ficção",]$Renda,ft4[ft4$Genero=="Documentário",]$Renda ))  
  
  
  
ggplot(experimentos, aes(x = diferenca/1e3)) +   
 geom\_histogram(binwidth = 100, colour = "darkorange", fill = "white")



summary(experimentos)

i diferenca

Min. : 1.0 Min. :1651804  
 1st Qu.: 250.8 1st Qu.:2893248  
 Median : 500.5 Median :3351362  
 Mean : 500.5 Mean :3411904  
 3rd Qu.: 750.2 3rd Qu.:3869280  
 Max. :1000.0 Max. :6505020

# IC com 90%:   
#"Portanto em 90% com (margem +-0.5) das repetições as médias ficaram entre: "  
alpha = .1  
  
quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

5% 95%

2321566 4672177

# IC com 95%:  
#Portanto em 95% das repetições, com margem de +-2,5 as médias ficaram entre:  
alpha = .05  
  
 quantile(experimentos$diferenca, probs = c(alpha/2, 1 - alpha/2))

2.5% 97.5% 2154957 4904928

#Fim