Problema2 - Parte2

Arthur Sena

05/01/2015

library(ggplot2, quietly = T, warn.conflicts = F)  
library(dplyr, quietly = T, warn.conflicts = F)

# Descrição dos dados:

deputados <- read.csv("AnoAtual.csv")

Os dados usados nesse experimento foram obtidos do site da Câmara dos Deputados e seu conteúdo pode ser acessado clicando [aqui](http://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar/dados-abertos-cota-parlamentar). Tais dados são relativos aos gastos parlamentares registrados na Câmara dos Deputados. Abaixo, pode ser visto uma pequena amostra do seu conteúdo:

amostra <- select(deputados, txNomeParlamentar, ideCadastro, sgPartido, vlrLiquido)  
 head(amostra,n = 10)

## txNomeParlamentar ideCadastro sgPartido vlrLiquido  
## 1 LUIS CARLOS HEINZE 73483 PP 3700.00  
## 2 GERALDO THADEU 74151 PSD 3504.55  
## 3 CARLOS EDUARDO CADOCA 74474 PCdoB 1450.90  
## 4 JOÃO DERLY 178965 PCdoB 988.10  
## 5 MILTON MONTI 74787 PR 952.50  
## 6 LAERCIO OLIVEIRA 151208 SD 853.60  
## 7 NILMÁRIO MIRANDA 74751 PT 694.80  
## 8 JOÃO FERNANDO COUTINHO 178917 PSB 603.72  
## 9 GORETE PEREIRA 129618 PR 599.00  
## 10 JOSÉ AIRTON CIRILO 141464 PT 599.00

Todas as variáveis/colunas e suas, respectivas, descrições podem ser visualizadas no [link](http://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar/explicacoes-sobre-o-formato-dos-arquivos-xml).

# Analisando os dados

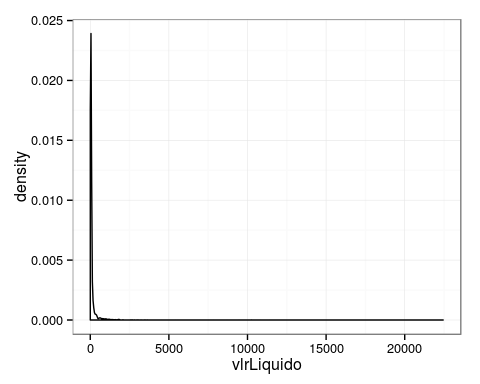
Com esses dados em mãos, nós podemos fazer algumas análises interessantes sobre eles. Por exemplo, vamos calcular a média da variável "vlrLiquido" para aquelas despesas relativa à serviços portais.

servicosPostais\_gastos <- deputados %>% select(txNomeParlamentar,vlrLiquido,txtDescricao,sgUF,sgPartido) %>% filter (txtDescricao == "SERVIÇOS POSTAIS")  
 mean(servicosPostais\_gastos$vlrLiquido)

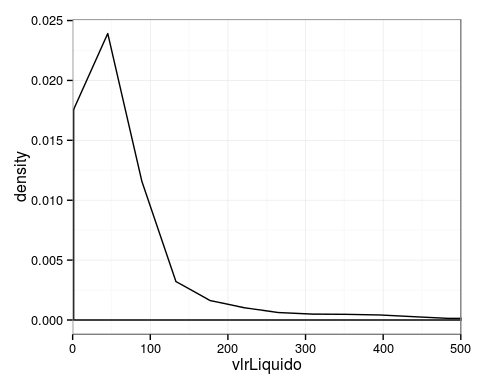
## [1] 118.2888

A fim de fazer uma análise mais detalhada, abaixo se encontra a função de distribuição de probabilidade, no qual representas chances de um determinado valor ser assumido pela variavel.

ggplot( servicosPostais\_gastos, aes(vlrLiquido)) +   
 geom\_density() +  
 theme\_bw()



ggplot( servicosPostais\_gastos, aes(vlrLiquido)) +   
 geom\_density() +coord\_cartesian(xlim=c(0, 500))+   
 theme\_bw()

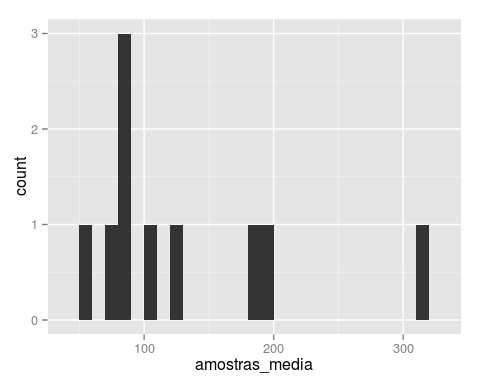


Visualizando distribuição de probabilidade do gráfico acima, podemos observar que a variável apresenta uma maior chance de obter valores abaixo de R$500,00 reais. A partir dos gráficos acima observados, podemos demontrar alguns conceitos interessantes de estatistica, como por exemplo: O teorema do limite central. Tal teorema afirma que "Qualquer que seja a distribuição da variável de interesse para grande amostras, a distribuição das médias amostrais serão aproximadamente normalmente distribuídas". A fim de um melhor entendimento sobre o teorema, podemos demontrá-lo utilizando os gastos de Serviços Postais dos deputados.

Primeiramente, vamos reunir 10 amostras de tamanho 100 da nossa população e calcular, para cada amostra, sua média. Feito isso, vamos visualizar a distribuição de probabilidade para as médias das nossas amostras.

# 10 amostras coletadas

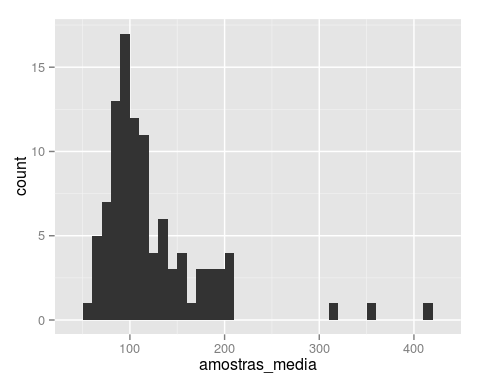
dist\_original <- servicosPostais\_gastos$vlrLiquido  
amostra\_tamanho <- 100  
num\_amostras <- 10  
  
amostras\_media <- c()  
for(i in seq(1, num\_amostras)){  
 uma\_amostra <- sample(dist\_original, amostra\_tamanho)  
 amostras\_media[i] <- mean(uma\_amostra)  
}  
  
ggplot(data.frame(amostras\_media), aes(amostras\_media)) + geom\_histogram(binwidth = 10)



Aparentemente, a distribuição acima, não se apresenta muito similar a uma distribuição normal. Contudo, se continuarmos a aumentar o números de amostras coletadas o Teorema Central do Limite poderá ser constatado.

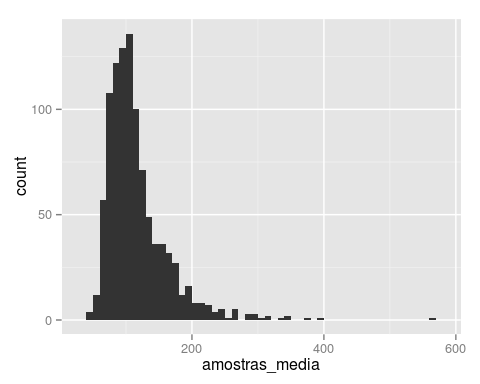
# 100 amostras coletadas

dist\_original <- servicosPostais\_gastos$vlrLiquido  
amostra\_tamanho <- 100  
num\_amostras <- 100  
  
amostras\_media <- c()  
for(i in seq(1, num\_amostras)){  
 uma\_amostra <- sample(dist\_original, amostra\_tamanho)  
 amostras\_media[i] <- mean(uma\_amostra)  
}  
  
ggplot(data.frame(amostras\_media), aes(amostras\_media)) + geom\_histogram(binwidth = 10)



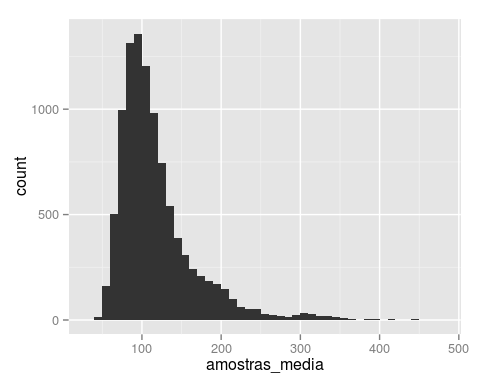
# 1000 amostras coletadas

dist\_original <- servicosPostais\_gastos$vlrLiquido  
amostra\_tamanho <- 100  
num\_amostras <- 1000  
  
amostras\_media <- c()  
for(i in seq(1, num\_amostras)){  
 uma\_amostra <- sample(dist\_original, amostra\_tamanho)  
 amostras\_media[i] <- mean(uma\_amostra)  
}  
  
ggplot(data.frame(amostras\_media), aes(amostras\_media)) + geom\_histogram(binwidth = 10)



# 10000 amostras coletadas

dist\_original <- servicosPostais\_gastos$vlrLiquido  
amostra\_tamanho <- 100  
num\_amostras <- 10000  
  
amostras\_media <- c()  
for(i in seq(1, num\_amostras)){  
 uma\_amostra <- sample(dist\_original, amostra\_tamanho)  
 amostras\_media[i] <- mean(uma\_amostra)  
}  
  
ggplot(data.frame(amostras\_media), aes(amostras\_media)) + geom\_histogram(binwidth = 10)



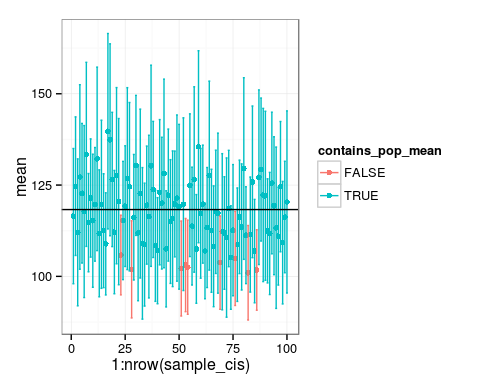
Podemos observar que à medida que aumentamos o número de amostras a forma da nossa distribuição se aproxima cada vez mais da forma de uma distribuição normal.

# Intervalos de Confiança

Outro conceito interessante, e que vale ser mencionado nessa análise, é o conceito de Intervalos de Confiança, no qual se destinam a mensurar um intervalo, relativo a média de cada amostra, onde a média da população se encontra. Ou seja, se eu tenho uma amostra com média X, então eu posso determinar um intervalo relativo à X, no qual a média da população pode estar inserida. Esse intervalo será determinado pode ser definido como: **(X - Erro, X + Erro)**.Onde, o erro será definido pela fórmula:

Assim sendo, podemos definir um intervalo de confiança para cada amostra coletada anteriormente e visualizar tais intervalos em um gráfico.

dist\_original <-(servicosPostais\_gastos$vlrLiquido)  
 amostra\_tamanho <- 2500  
 num\_amostras <- 100  
 pop\_mean <- mean(dist\_original)  
  
 calcula\_intervalo <- function(amostra,tamanhoAmostra){  
 temp <- as.data.frame(amostra)  
 se <- sd(temp[,1])/sqrt(tamanhoAmostra)  
 inferior <- mean(amostra) - (1.96\*se)  
 superior <- mean(amostra) + (1.96\*se)  
 list("mean"=mean(amostra) , "lower"=inferior,"upper"=superior)  
 }  
  
 sample\_cis <- data.frame(upper = c(), mean = c(), lower = c())  
 for(i in seq(1, num\_amostras)){  
 a\_sample <- sample(dist\_original, amostra\_tamanho)  
 interval <- calcula\_intervalo(a\_sample, amostra\_tamanho)  
 sample\_cis <- rbind(sample\_cis, data.frame(mean = interval["mean"], lower = interval["lower"],upper = interval["upper"]))  
 }  
 sample\_cis <- sample\_cis %>% mutate(contains\_pop\_mean = (upper >= pop\_mean & lower <= pop\_mean))   
  
 ggplot(sample\_cis, aes(x = 1:nrow(sample\_cis), y = mean, colour = contains\_pop\_mean)) +   
 geom\_point() +   
 geom\_errorbar(aes(ymin = lower, ymax = upper)) +   
# geom\_hline(aes(yintercept=mean(mean(dist\_original)))) +   
 geom\_hline(aes(yintercept=pop\_mean)) +   
 theme\_bw()



Cada ponto azul e vermelho acima representa o valor da média da amostra, onde as barras acima e abaixo dos pontos representa a extensão dos seus, respectivos, intervalos de confiança. Os pontos e barras vermelhos são aqueles no qual a média da população não se encontra inserida nos seus intervalos. A barra preta desenhada horizontalmente representa a média da população.

Eu calculei os intervalos com um nível de confiança de 95%, ou seja, a probabilidade de que , com várias repetições, os intervalos conterão o verdadeiro valor do parâmetro. O que é facilmente observado no gráfico, pois apenas seis intervalos não contiveram a média da população. É importante notar que as margens de erros dos intervalos são inversamente proporcionais ao tamanho das amostras, ou seja, quanto maior o tamanho menor a margem de erro.