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. 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.

Analisando os dados

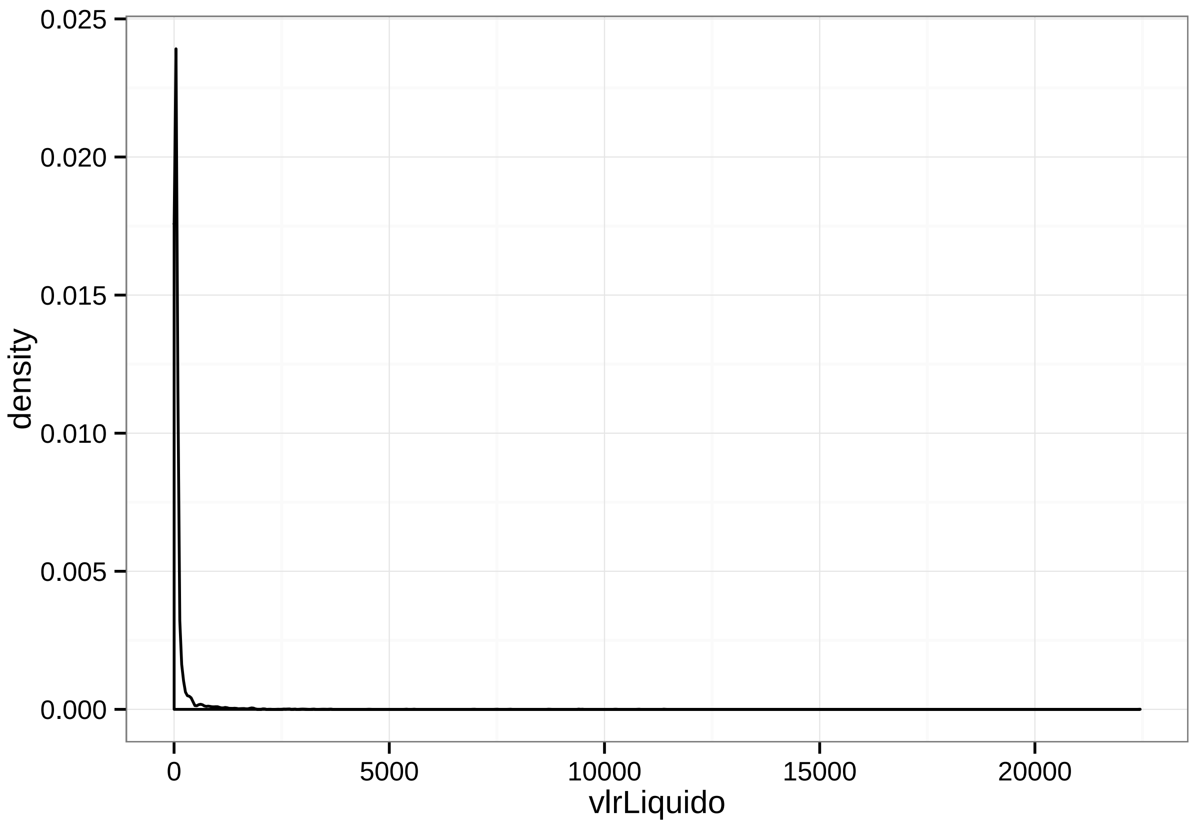
Com esses dados em mãos, nós podemos fazer algumas análises interessantes sobre eles. Por exemplo, vamos calcular a média da variável "vlrLiquido" para aquelas despesas relativa à serviços portais.

```
  servicosPostais_gastos <- deputados %>% select(txNomeParlamentar,vlrLiquido,txtDescricao,sgUF,sgPartic
  mean(servicosPostais_gastos$vlrLiquido)
```

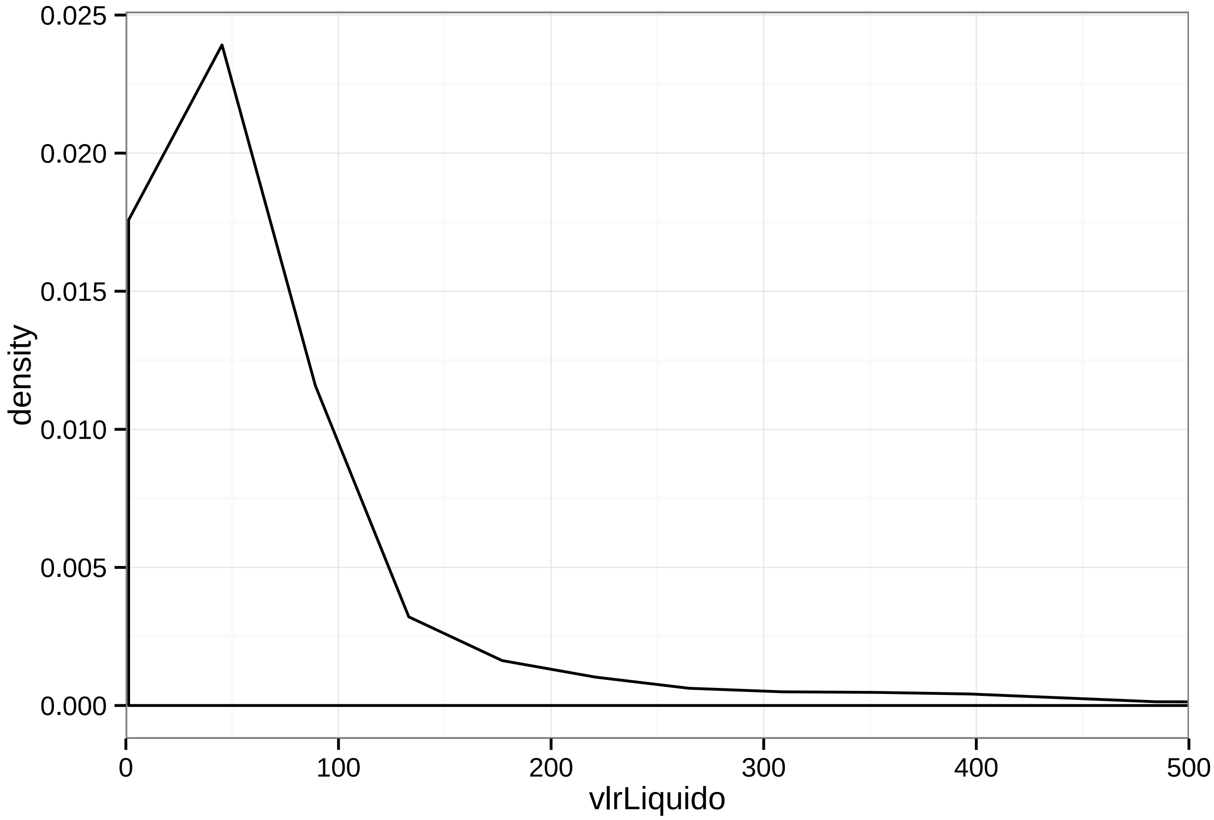
```
## [1] 118.2888
```

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.

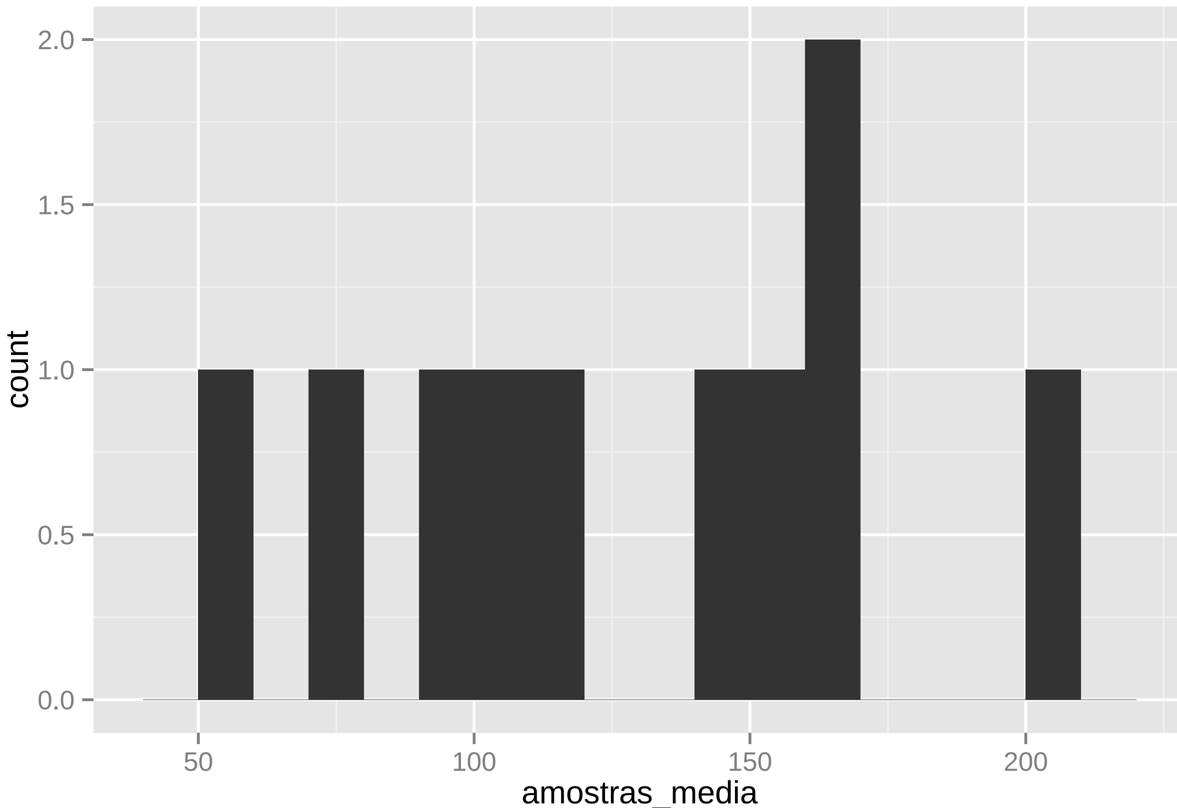
Primeiramente, vamos reunir 10 amostras de tamanho 100 da nossa população e calcular, para cada amostra, sua média. Feito isso, vamos visualizar a distribuição de probabilidade para as médias das nossas amostras.

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)
```



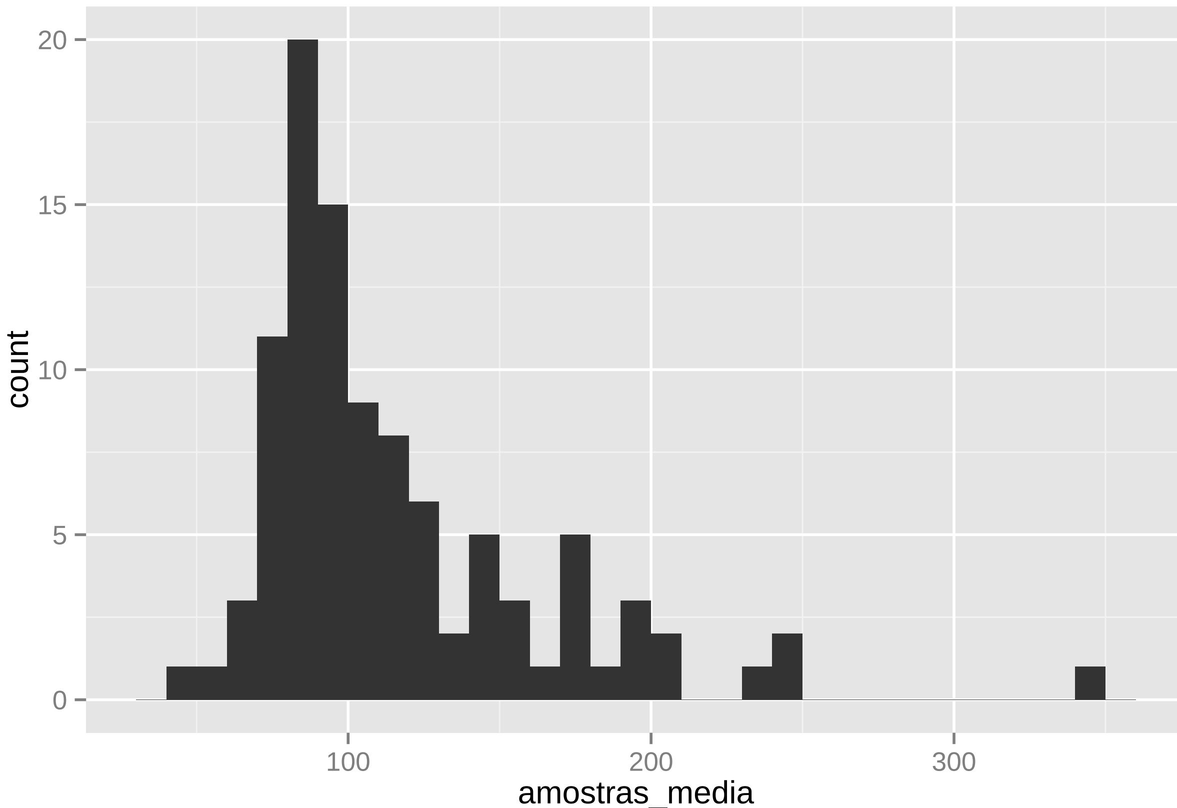
Aparentemente, a distribuição acima, não se apresenta muito similar a uma distribuição normal. Contudo, se continuarmos a aumentar o números de amostras coletadas o Teorema Central do Limite poderá ser constatado.

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)
```

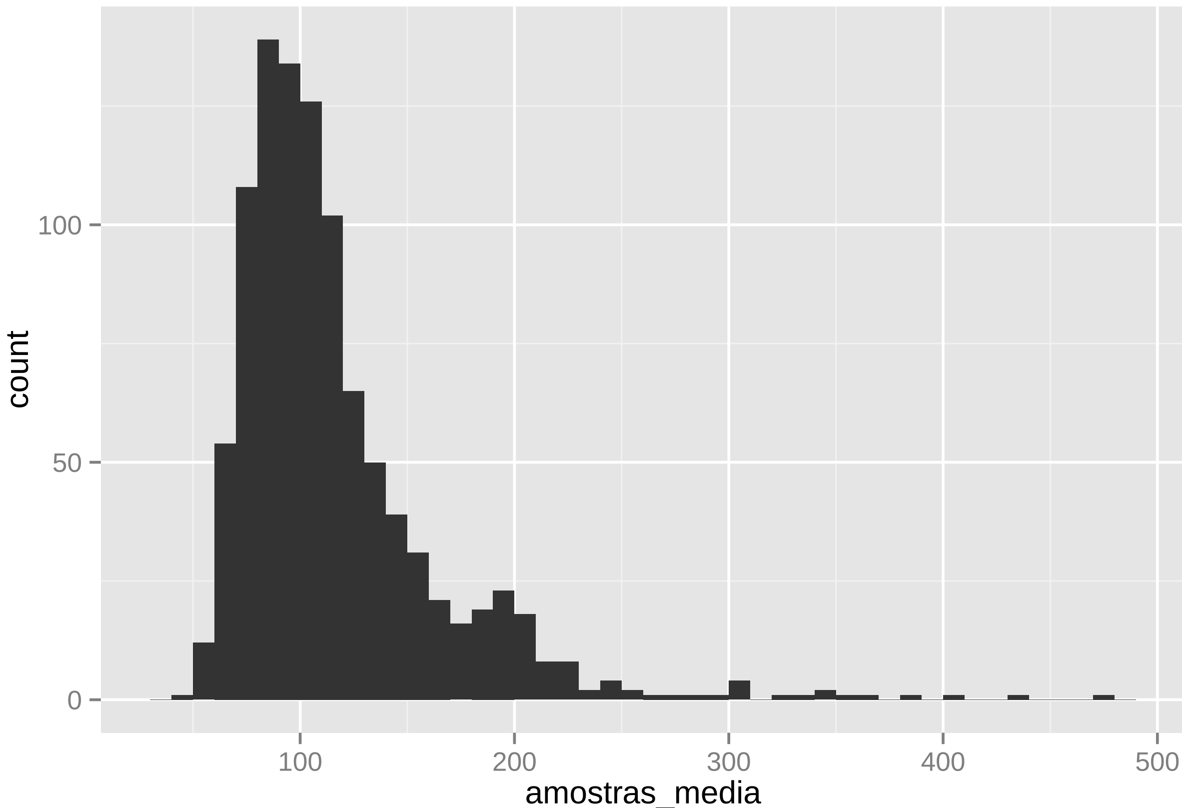


1000 amostras coletadas

```
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)
```

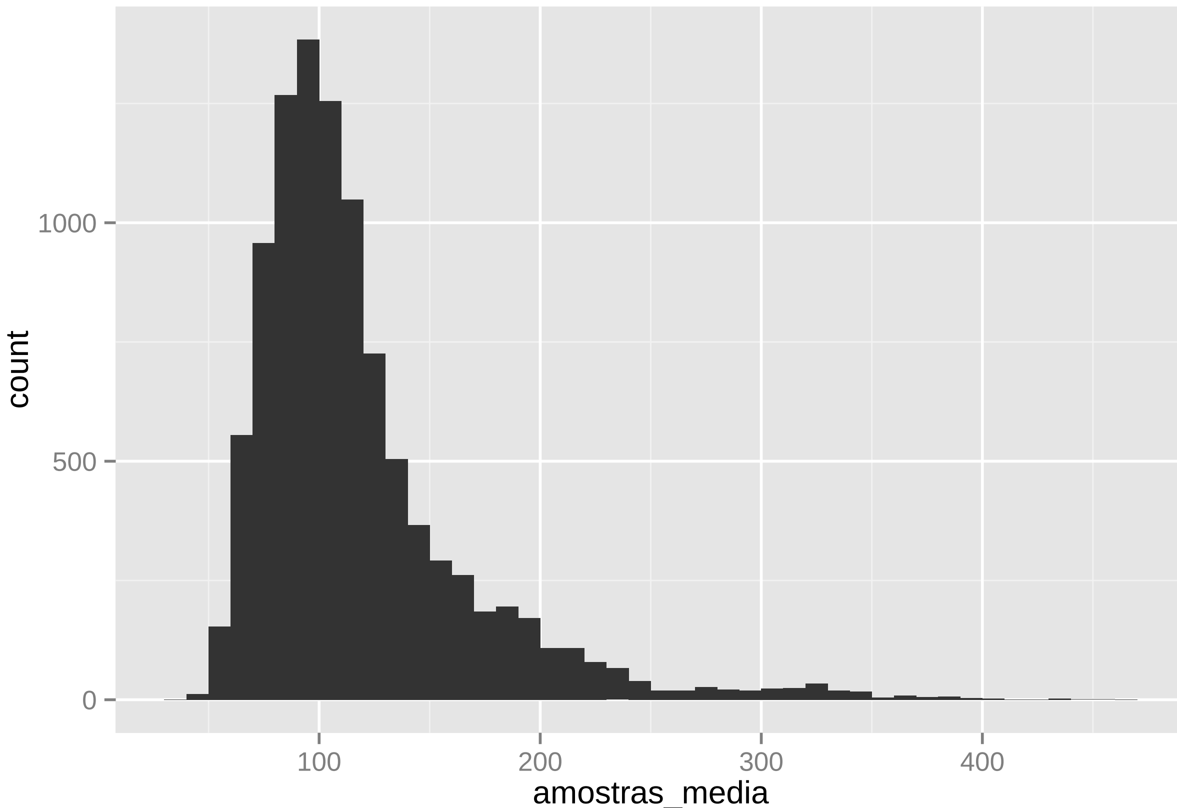


10000 amostras coletadas

```
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.