Passo-a-passo

ETAPA 6. GRÁFICOS DESCRITIVOS

Prof. Pedro Feliú

Instituto de Relações Internacionais
Universidade de São Paulo

INTRODUÇÃO

Nesta etapa executaremos comandos referentes à descrição estatística dos dados por meio de gráficos. Desta forma, esta etapa 6 complementa as etapas 4 e 5. Iniciaremos com o gráfico de barras.

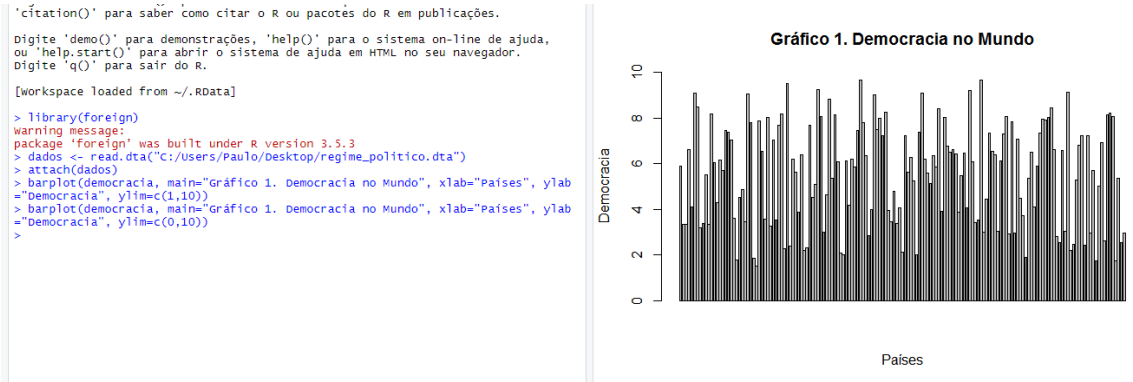
PASSO 1: Gráfico de Barras

Diferentemente das etapas anteriores, vamos utilizar outro comando para importar o próximo banco de dados utilizado nessa etapa: "regime_politico.dta". O motivo é mostrar alternativas de importação para que aqueles que quiserem usar o R e não o Rstudio, tenham a primeira ferramenta básica para mexer no programa. Embora os comandos sejam os mesmos, a interface do Rstudio permite a importação de dados por meio de cliques apenas. Já o R não, só na base dos comandos. Para cumprir essa tarefa, precisaremos de um pacote do R "foreign", instale o pacote e depois utilize o comando library(foreign). A função que utilizaremos para importar um arquivo em stata (.dta) é read.dta. Baixe o arquivo "regime_politico.dta" e salve na área de trabalho. Digite o seguinte comando, adaptando os nomes dos diretórios para o seu computador:

```
install.packages("foreign")
library(foreign)
dados <- read.dta("C:/Users/Paulo/Desktop/regime_politico.dta")
attach(dados)
```

Cumpridos os comandos acima, agora vamos utilizar o comando **barplot** que gerará o gráfico de barras. Junto com o comando **barplot**, incluímos também os comandos referentes à nomeação dos eixos do gráfico (**xlab** e **ylab**, para os eixos horizontal e vertical, respectivamente) e também de seu título (**main**). Isso tudo para a variável "democracia" do banco de dados importado. Essa variável vai de 1 a 10, onde 1 é o regime mais autoritário e 10 o mais democrático. Segue o comando abaixo:

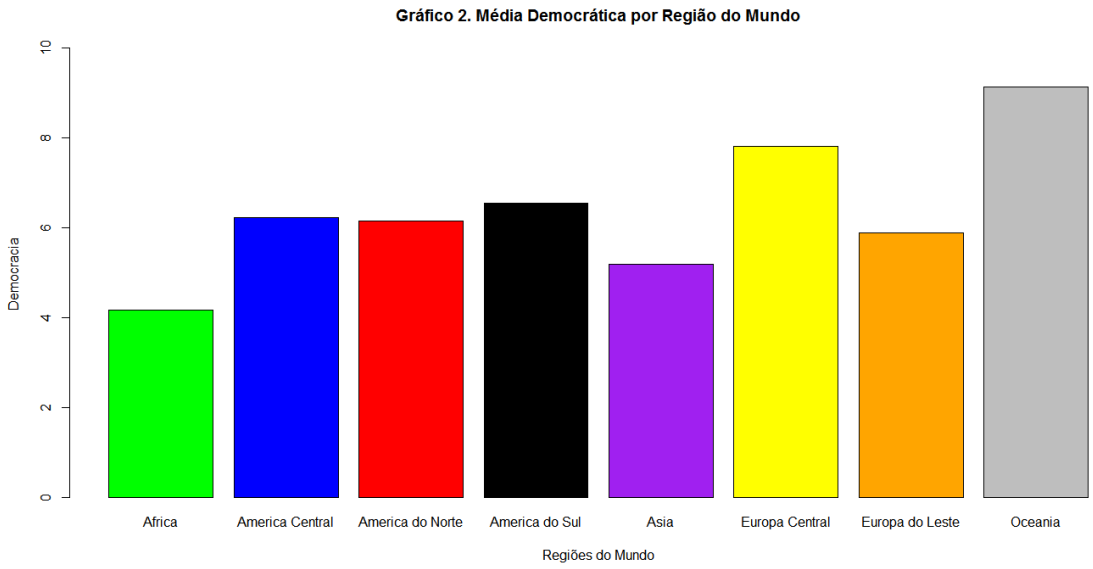
```
barplot(democracia, main="Gráfico 1. Democracia no Mundo", xlab="Países", ylab="Democracia", ylim=c(0,10))
```



O gráfico acima não ficou muito bom para visualizar, pois são muitos países, onde cada barra representa um país. Vamos realizar alguns comandos para construir barras que representem as regiões do mundo e não os países, agregando informação e facilitando a visualização do gráfico. Primeiro, vamos utilizar a função **tapply** já usada anteriormente, criando um objeto chamado "p", contendo as médias da democracia por região do mundo. Depois iremos gerar o gráfico com a função **barplot**, incluindo seis cores para cada uma das regiões do mundo com a função **col=c("cores")**. Segue o comando abaixo:

p<-tapply(democracia,regiao,mean)

barplot(p, col=c("green","blue","red","black","purple","yellow","orange"), main="Gráfico 2. Média Democrática por Região do Mundo", xlab="Regiões do Mundo", ylab= "Democracia ", ylim=c(0,10))



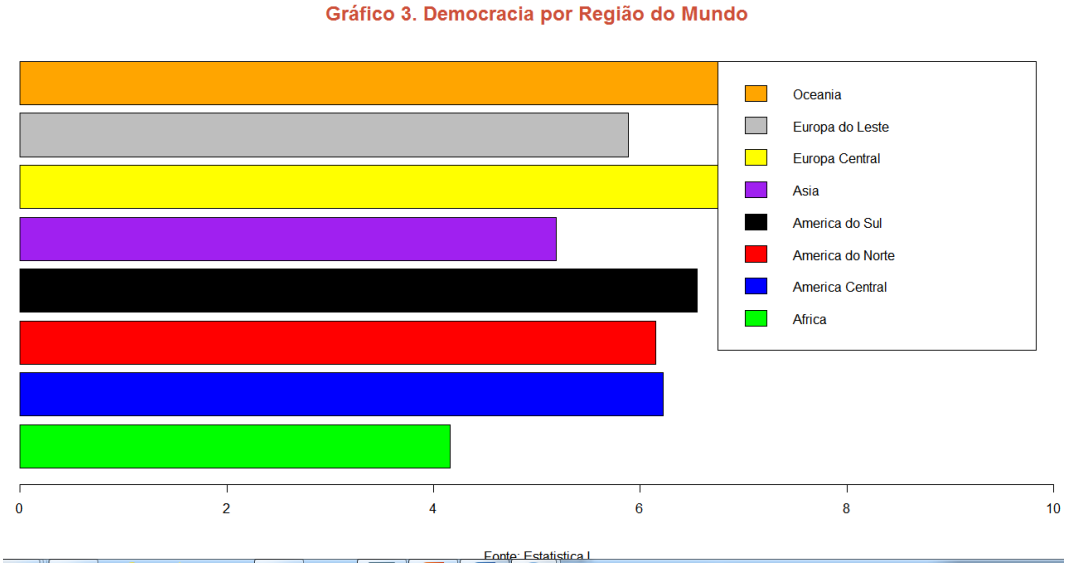
Agora visualizamos acima o gráfico de barras da democracia por região do mundo, a partir das médias dos países por região. Percebemos uma leve desigualdade democrática entre as regiões do mundo, onde a Europa Central e Oceania possuem médias mais elevadas, enquanto a África, Ásia e Europa do Leste revelam médias inferiores. Notem que são oito regiões do mundo e precisamos digitar no comando oito cores para que cada região seja representada por uma delas. Para ver as opções de cores do R há um comando bastante útil:

colors()

```
> colors()
  [1] "white"           "aliceblue"       "antiquewhite"    "antiquewhite1"   "antiquewhite2"   "antiquewhite3"   $
  [7] "antiquewhite4"   "aquamarine"      "aquamarine1"     "aquamarine2"     "aquamarine3"     "aquamarine4"     $
 [13] "azure"           "azure1"          "azure2"          "azure3"          "azure4"          "beige"           $
 [19] "bisque"          "bisque1"         "bisque2"         "bisque3"         "bisque4"         "black"           $
 [25] "blanchedalmond"  "blue"            "blue1"           "blue2"           "blue3"           "blue4"           $
 [31] "blueviolet"      "brown"           "brown1"          "brown2"          "brown3"          "brown4"          $
 [37] "burlywood"       "burlywood1"      "burlywood2"      "burlywood3"      "burlywood4"      "cadetblue"       $
 [43] "cadetblue1"      "cadetblue2"      "cadetblue3"      "cadetblue4"      "chartreuse"      "chartreuse1"     $
 [49] "chartreuse2"     "chartreuse3"     "chartreuse4"     "chocolate"       "chocolate1"      "chocolate2"      $
 [55] "chocolate3"      "chocolate4"      "coral"           "coral1"          "coral2"          "coral3"          $
 [61] "coral4"          "cornflowerblue"  "cornsilk"        "cornsilk1"       "cornsilk2"       "cornsilk3"       $
 [67] "cornsilk4"       "cyan"            "cyan1"           "cyan2"           "cyan3"           "cyan4"           $
 [73] "darkblue"        "darkcyan"        "darkgoldenrod"   "darkgoldenrod1"  "darkgoldenrod2"  "darkgoldenrod3"  $
 [79] "darkgoldenrod4"  "darkgray"        "darkgreen"       "darkgrey"        "darkkhaki"       "darkmagenta"     $
 [85] "darkolivegreen"  "darkolivegreen1" "darkolivegreen2" "darkolivegreen3" "darkolivegreen4" "darkorange"      $
 [91] "darkorange1"     "darkorange2"     "darkorange3"     "darkorange4"     "darkorchid"      "darkorchid1"     $
 [97] "darkorchid2"     "darkorchid3"     "darkorchid4"     "darkred"         "darksalmon"      "darkseagreen"    $
[103] "darkseagreen1"   "darkseagreen2"   "darkseagreen3"   "darkseagreen4"   "darkslateblue"   "darkslategray"   $
[109] "darkslategray1"  "darkslategray2"  "darkslategray3"  "darkslategray4"  "darkslategrey"   "darkturquoise"   $
[115] "darkviolet"      "deeppink"        "deeppink1"       "deeppink2"       "deeppink3"       "deeppink4"       $
[121] "deepskyblue"     "deepskyblue1"    "deepskyblue2"    "deepskyblue3"    "deepskyblue4"    "dimgray"         $
[127] "dimgrey"         "dodgerblue"      "dodgerblue1"     "dodgerblue2"     "dodgerblue3"     "dodgerblue4"     $
[133] "firebrick"       "firebrick1"      "firebrick2"      "firebrick3"      "firebrick4"      "floralwhite"     $
[139] "forestgreen"     "gainsboro"       "ghostwhite"      "gold"            "gold1"           "gold2"           $
[145] "gold3"           "gold4"           "goldenrod"       "goldenrod1"      "goldenrod2"      "goldenrod3"      $
[151] "goldenrod4"      "gray"            "gray0"           "gray1"           "gray2"           "gray3"           $
[157] "gray4"           "gray5"           "gray6"           "gray7"           "gray8"           "gray9"           $
```

Todos os nomes de cores aparecerão no console com os seus respectivos números. Teste o comando anterior com diferentes cores. Vamos fazer mais um comando com diferentes funções ainda no gráfico de barras.

barplot(p, names.arg=F, main="Gráfico 3. Democracia por Região do Mundo", col=c("green","blue","red","black","purple","yellow", "grey", "orange"), legend= rownames(p), beside=T, horiz=T, axes=T, sub="Fonte: Estatistica I", cex.main=1.5, col.main=" tomato3", xlim=c(0,10))



Vamos ver em detalhes as funções utilizadas:

names.arg=F – indica se queremos o rótulo das variáveis, no caso as regiões do mundo, onde F indica ausência. Como pusemos legenda neste gráfico, não é necessário rótulo.
legend=rownames(p) – indica a criação da legenda, com os nomes das linhas do objeto p (o objeto criado a partir das médias dos pib/capita dos países por região do mundo)
beside=T – indica que a legenda ficará ao lado no gráfico
horiz=T – transpõe da vertical para a horizontal as barras do gráfico
axes=T – é o default do R, mas se quiser retirar os eixos coloque F
sub=”fonte” – inseri uma fonte ao gráfico

cex.main – estabelece o tamanho da fonte do gráfico. Esse comando pode ser utilizado com os eixos (xlab, ylab), assim como os outros componentes do gráfico.
col.main – colore o título do gráfico, assim como qualquer item do gráfico (col.lab, col.axis, col.sub, etc).
xlim=c(0,10) – estabeleceo eixo X de 0 a 10.

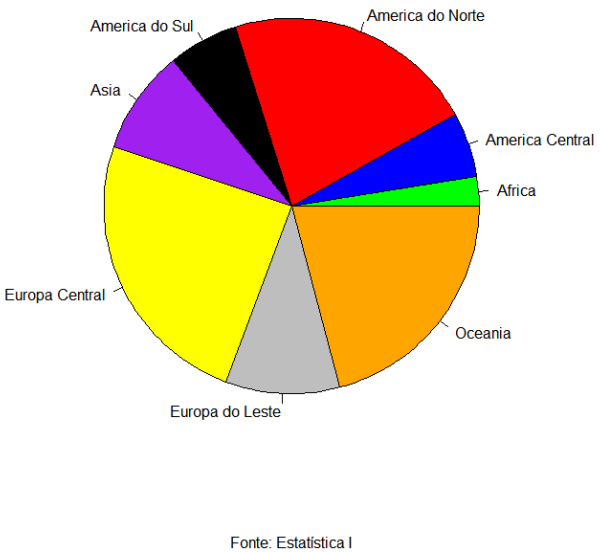
Notem que a legenda ficou em cima do gráfico. Para arrumar isso podemos utilizar o comando abaixo que permite mover a legenda com o mouse/cursor do computador.

```
legend(locator(1),legend=rownames(p),
fill=c("green","blue","red","black","purple","yellow", "grey",
"orange"),bty="n")
```

PASSO 2: Gráfico de Pizza

Digite os seguintes comandos, agora com a variável "pib" (PIB/capita dos países) do mesmo banco de dados, utilizando a função **pie** e a mesma função tapply:

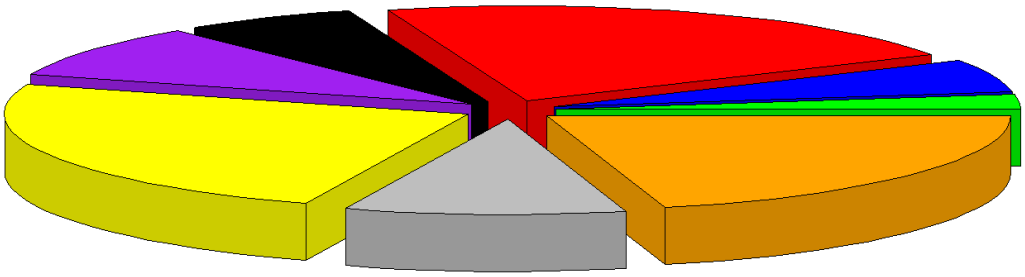
```
t<-tapply(pib,regiao,mean)
pie(t, main="Gráfico 4. PIB per Capita por Região do Mundo",
col=c("green","blue","red","black","purple","yellow", "grey", "orange"),
sub="Fonte: Estatística I", cex.main=2.5, col.main="darkblue")
```



Realizamos agora um gráfico de pizza 3D. Instale e carregue o pacote **plotrix** e digite os comandos abaixo:

```
library(plotrix)
pie3D(t, main="Gráfico 5. PIB per Capita por Região do Mundo",
col=c("green","blue","red","black","purple","yellow", "grey", "orange"),
explode=0.1,cex.main=1.0, col.main="tomato")
```

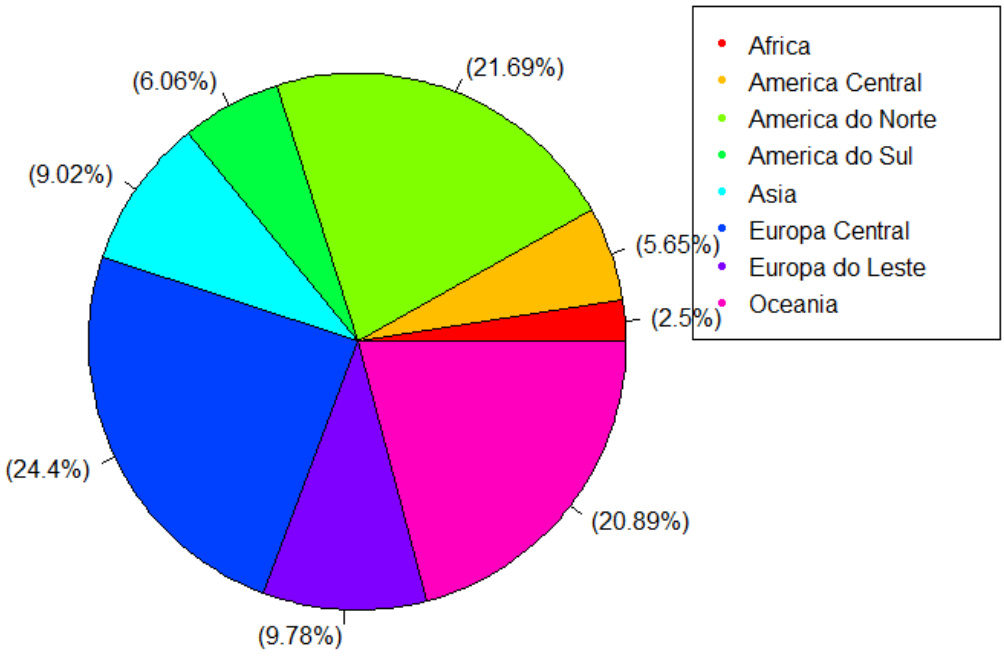
Gráfico 5. PIB per Capita por Região do Mundo



A única função nova neste exemplo, **explode**, determina a distância das fatias da pizza acima. Vamos agora voltar ao modo 2D, incluindo algumas funções novas, como a porcentagem no gráfico de pizzas com a função **labels**, utilizar um novo comando para distribuir as cores (**rainbow**) e o comando **pch** para editar a legenda de nosso gráfico inserida após a construção do mesmo. Digite os seguintes comandos:

```
porc<-round(t*100/sum(t),2)
rotulos<-paste("(",porc,"%)",sep="")
pie(t, main="Gráfico 6. PIB per Capita por Região do Mundo",labels=rotulos, col=rainbow(8))
legend(1,1,names(t),col = rainbow(8),pch=rep(15,5))
```

Gráfico 6. PIB per Capita por Região do Mundo

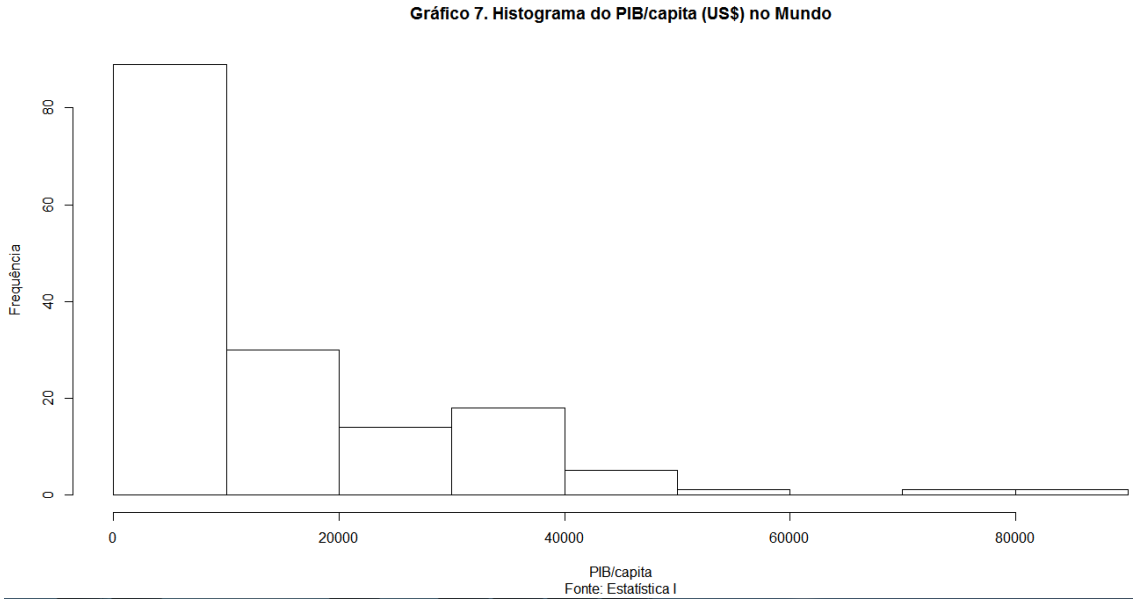


O primeiro comando digitado acima cria um objeto que determina as porcentagens do pib per capita de cada região em relação ao total. Chamamos por conveniência esse objeto de "porc". Utilizamos a função **round** e multiplicamos nosso objeto criado anteriormente "t" por 100, dividindo tudo pela soma dos valores de "t". Depois da vírgula, colocamos o número 2 para indicar ao R que queremos apenas duas casas decimais para as porcentagens. Após esse comando, criamos outro objeto, chamado convenientemente de "rotulos". Utilizamos a função **paste** para concatenar as porcentagens a cada região do mundo, incluindo a função **sep** indicando que as aspas separam os nomes. Em seguida utilizamos a função já conhecida **pie**, mas agora com outra novidade, a função **col=rainbow(8)**, que delega ao R a escolha das cores do gráfico, indicando a quantidade de cores que queremos (no nosso caso 8 regiões). Finalmente, esse gráfico já criado necessita uma legenda. Assim, geramos o comando **legend**. Os números "1,1" indicam a quantidade de legendas que queremos expor. A função **names** indica que queremos utilizar os nomes das regiões do mundo do objeto t e **pch** indica o tipo de figura na legenda (quadrado, ponto, etc, cada um com um respectivo número). Notem que o gráfico 5 (3D) está sem porcentagem ou legenda. Para treinar, coloquem porcentagem e legenda no gráfico 5, alterando os valores de **pch** para ilustrar o funcionamento deste comando. Utilizem também o comando de posicionar a legenda com o cursor Passemos agora para o histograma.

PASSO 3: Histograma

Manteremos o mesmo banco de dados e variável (pib) para utilizar o comando **hist** que gerará o histograma. Segue o comando abaixo:

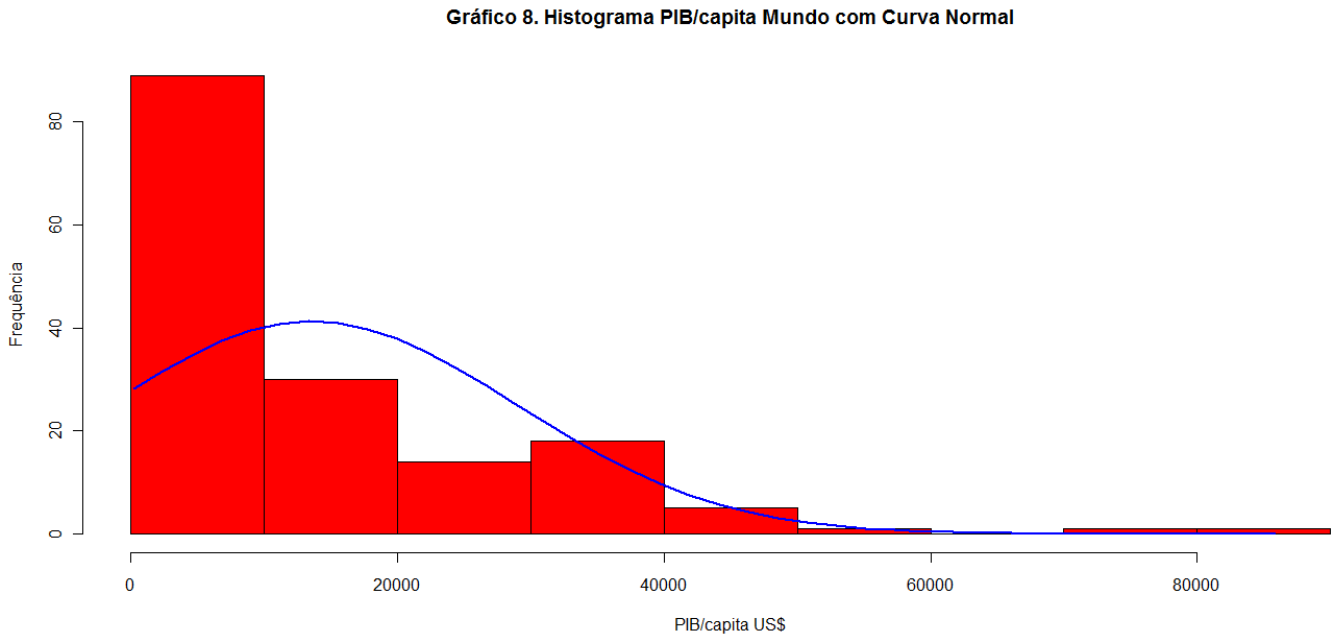
hist(pib, main="Gráfico 7. Histograma do PIB/capita (US$) no Mundo", xlab="PIB/capita", ylab="Frequência", sub="Fonte: Estatística I")



O histograma é um gráfico descritivo composto por retângulos justapostos cuja base de cada um deles corresponde ao intervalo dos dados (no caso o PIB/capita) e a sua altura à frequência (no caso de países). A construção de histogramas tem caráter preliminar em qualquer estudo e é um importante indicador da distribuição de dados. Podem indicar se

uma distribuição aproxima-se de uma função normal, como pode indicar assimetria na distribuição dos dados. Para facilitar a visualização e interpretação deste gráfico vamos incluir uma curva normal. Seguem os comandos:

```
h<-hist(pib, breaks=10, col="red", xlab="PIB/capita US$", ylab="Frequência",
main="Gráfico 8. Histograma PIB/capita Mundo com Curva Normal")
xfit<-seq(min(pib),max(pib),length=159)
yfit<-dnorm(xfit,mean=mean(pib),sd=sd(pib))
yfit <- yfit*diff(h$mids[1:2])*length(pib)
lines(xfit, yfit, col="blue", lwd=2)
```



Nos comandos acima, criamos primeiro um objeto "h", que é justamente o histograma em que iremos incluir a curva normal[1]. A única novidade nesta primeira linha é a função **breaks**, que delimita o tamanho dos retângulos do gráfico. Depois criamos o objeto "xfit", que cria a sequência de dados para construção da curva normal com a função **seq** e **length** (que diz até onde os dados vão, são 159 países). Depois utilizamos o comando de criação da curva normal **dnorm,** por meio da média (mean) e desvio padrão (sd) da distribuição**.** Executamos em seguida o comando **diff** que atrela a curva normal ao gráfico do histograma. Finalmente, plotamos a curva normal com a função **lines**, como pode ser observado no gráfico 8 acima. A função **lwd**, nova, determina o tamanho da linha da curva normal. Tentem modificar esse valor para ver como funciona o comando.
O gráfico 8 acima demonstra, mais uma vez, a assimetria na distribuição da renda dos países no mundo. A curva azul é assimétrica e indica que parte substancial dos países (mais de 80 países) se encontra entre 0 e 10.000 dólares per capita de renda. Pouquíssimos países, em contrapartida, possuem renda superior à 60.000 dólares. Passemos agora para o passo final, correspondente à criação de um gráfico do tipo box plot (caixa de dispersão).

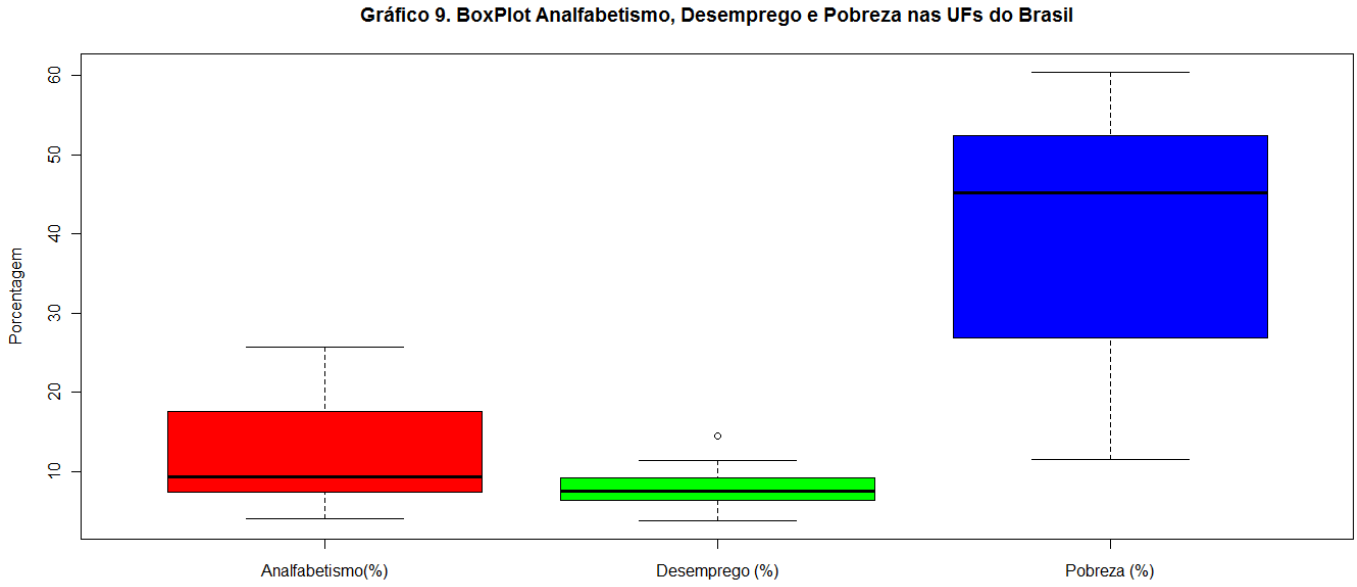
[1] A curva normal é caracterizada por uma média igual à mediana de uma distribuição de dados, sendo pressuposto para importantes testes estatísticos. Um interessante uso da Distribuição Normal é a aproximação para o cálculo de outras distribuições quando o número de observações é muito grande. Neste ponto, vale citar o Teorema do Limite Central: "toda soma de variáveis aleatórias independentes de média finita e variância limitada é aproximadamente Normal, desde que o número de termos da soma seja suficientemente grande". Para mais informações checar a biblioteca de estatística: curva normal.

PASSO 4: Box Plot

Neste passo, será utilizado o banco de dados "brasil.dta" na pasta banco de dados do moodle. Vamos importar o banco de dados da mesma forma nessa etapa, por meio de comando, como nos passos anteriores. Utilizaremos o comando **boxplot** para gerar o gráfico, gráfico de dispersão que permite a análise gráfica de quartis que foi feita na etapa passada:

```
library(foreign)
brasil <- read.dta("C:/Users/Paulo/Desktop/brasil.dta")
attach(brasil)

boxplot(analf, desemp, pobres, names=c("Analfabetismo(%)", "Desemprego (%)", "Pobreza (%)"), main="Gráfico 9. BoxPlot Analfabetismo, Desemprego e Pobreza nas UFs do Brasil", ylab="Porcentagem", border="black", col=rainbow(3))
```



No comando acima não há novidades. O boxplot é um gráfico que possibilita representar a distribuição de um conjunto de dados, no nosso caso a porcentagem de analfabetos, desempregados e pobres nas unidades da federação do Brasil (estados), com base em alguns parâmetros descritivos: a mediana (o traço escuro dentro da caixa), o quartil inferior (onde se encontram 25 % dos estados com a menor porcentagem de analfabetos, desempregados e pobres na sua população), o quartil superior (onde se encontram 25% dos estados com a maior porcentagem de analfabetos, desempregados e pobres na sua população) e do intervalo interquartil. Quanto maior a caixa do gráfico, maior será a dispersão do dado, ou seja, a sua heterogeneidade. Assim, pobreza é o dado que varia mais entre os estados brasileiros, enquanto desemprego menos. 50% dos estados brasileiros (mediana – traço preto na caixa) possuem aproximadamente 45% a 60% de sua população pobre. Já no analfabetismo, metade dos estados possui entre 10% e aproximadamente 5% de analfabetos em sua população. Notem na caixa verde, do desemprego, que há uma bolinha acima, representando um Estado. Esse estado

representa um dado aberrante (*outlier*), ou seja, ele possui um desemprego tão maior que os demais estados que ele não "cabe" no intervalo interquartil (os dois traços horizontais nas extremidades das distribuições). No caso, esse estado é o Maranhão. Na próxima etapa veremos como fazer um gráfico de dispersão e densidade.