Estatística Aplicada

Sumário

Estatística Aplicada

Objetivo do estudo: Analisar uma base de dados com variáveis categóricas e escalares com a utilização de softwares como: Excel, R, Action, Gretl.

Estudo de caso: CHOICE IS YOURS

A prestadora de serviços de investimentos *Choice Is Yours*, ajuda os clientes no que concerne as suas respectivas opções de investimentos. A Choice Is Yours avalia investimentos tão diversificados quanto imóveis, investimentos com capital privado, derivativos, e vários tipos especializados de fundos mútuos. Você foi contratado para assessorar investidores que buscam investir em fundos mútuos, que agregam em uma única cesta o dinheiro de muitos investidores e investem o dinheiro dessas pessoas em uma combinação de ações, títulos e outros certificados de valores.

Os analistas da empresa organizaram uma amostra com dados sobre os fundos mútuos e foi solicitado que você apresentasse os dados sobre esses fundos de modo tal que ajude os clientes a fazer opções bem fundamentadas no que concerne a investimentos.

Bse de dados: *Choice_is_Yours_Fundos.xlsx*

A base de dados contém as variáveis apresentadas na Tabela 1. São quatro variáveis categóricas e cinco variáveis numéricas relacionadas aos findos mútuos.

Tabela 1. Variáveis contidas na base de dados Choice_is_Yours_Fundos.xlsx.

Variável	Tipo	Classes (Categorias)
Categoria	Categórica	Gde.Vol.Cap Médio Cap Baixo Cap
Objetivo	Categórica	Crescimento Valorização
Ativos	Numérica	
Comissões	Categórica	Sim Não
Prop.Despesas	Numérica	
Risco	Categórica	Baixo Médio Alto
Retorno 2005	Numérica	
Retorno 3 Anos	Numérica	
Retorno 5 Anos	Numérica	

Nas seções seguintes são apresentadas as análises dos dados que fundamentam a apresentação A1.

1 Tabelas Dinâmicas para explorar dados multivariados no Excel

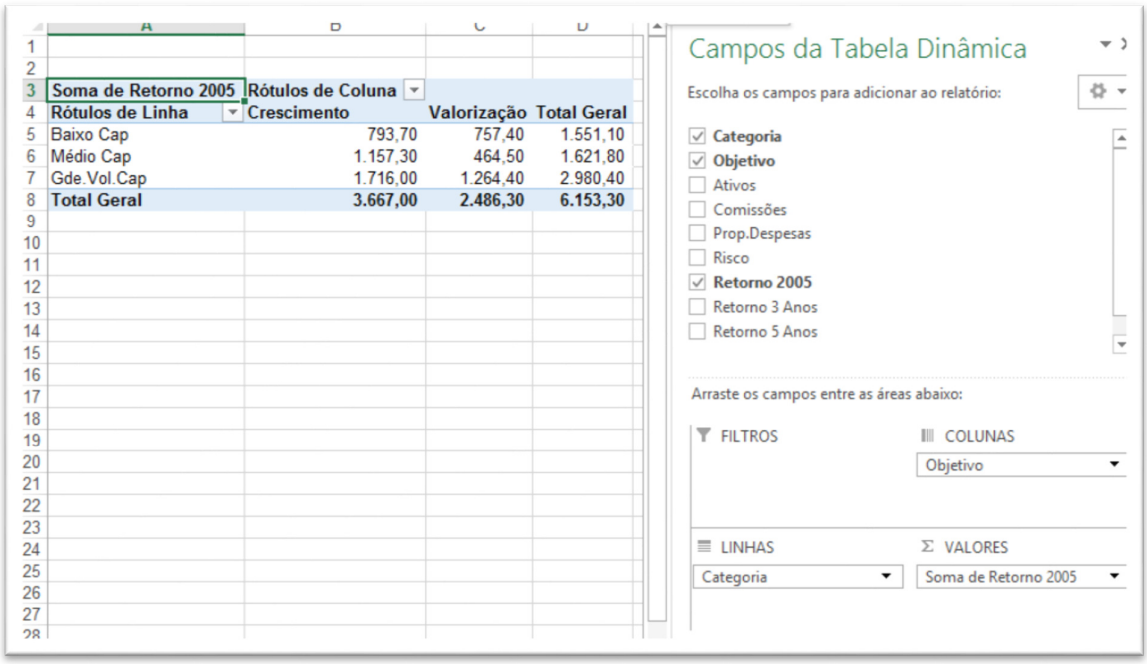
Uma tabela dinâmica é uma ferramenta do Excel para criar tabelas e resumir dados.

Utilize o arquivo ***Choice_is_Yours_Fundos.xlsx***.

Para criar uma tabela com resumo dos dados selecione: **Inserir → Tabelas → Tabela Dinâmica**. Na sequência, na caixa de diálogo “Criar Tabela Dinâmica” :

1. Clique em **Selecionar uma tabela ou intervalo** e informe o intervalo para **Tabela/Intervalo**.
2. Selecione a opção **Nova planilha** e clique em **OK**.
3. Aparece um painel de tarefas “Lista de campos da Tabela Dinâmica”. Marque as variáveis categóricas e arraste-as para linha ou coluna. Marque a variável numérica e arraste-a para o campo com soma. No caso de selecionar as variáveis **Categoria**, **Objetivo** e **Retorno 2005**, a tabela e os campos da tabela dinâmica podem ser vistos na Figura 1.

Figura 1. Planilha com Tabela dinâmica. Observe as variáveis que estão marcadas e os campos para os quais foram arrastadas.



4. Para formatar: clique À direita da Tabela Dinâmica e em Opções da Tabela Dinâmica.
5. Atenção: uma tabela dinâmica não atualiza os dados de forma automática. Se a base de dados for alterada então deve-se atualizar a tabela dinâmica: **Clicar na tabela com o botão direito → Atualizar**.

2 Estatísticas da base de dados com funções do Excel

Primeiramente, criamos algumas variáveis na planilha Dados: Selecione todos os dados da planilha (neste caso marque a região A1:I839) e depois digite CTRL + SHIFT + F3. Com isto serão criados nomes para as variáveis com base na linha superior, ou seja desmarque tudo, deixe apenas [x] Linha Superior.

Estatísticas para variáveis categóricas

Para obter frequências de ocorrências das variáveis categóricas utilize a função: CONT.SE, por exemplo:

EXCEL – CONT.SE: Para a variável "Categoria" digite: =CONT.SE(Categoria;Gde.Vol.Cap), ou ainda, digite as categorias possíveis para a variável Categoria numa coluna, por exemplo, B10 = Gde.Vol.Cap, B11 = Médio Cap, e B12 = Baixo Cap. Depois digite o comando: =CONT.SE(Categoria;B10). Este comando resulta na contagem de células que são iguais a "Gde.Vol.Cap", ou seja, 443 casos. O resultado é apresentado na Tabela 2.

Nas Tabelas 3 a 5 são presentadas as estatísticas para as variáveis Objetivo, Comissões e Risco.

Tabela 2. Estatísticas descritivas para variável Categoria.

Categoria	Contagem	Fequência Rel	Percentual
Gde.Vol.Cap	443	0,5286	52,9%
Médio Cap	166	0,1981	19,8%
Baixo Cap	229	0,2733	27,3%

Tabela 3. Estatísticas descritivas para variável Objetivo.

Objetivo	Contagem	Fequência Rel	Percentual
Crescimento	480	0,5728	57,3%
Valorização	358	0,4272	42,7%

Tabela 4. Estatísticas descritivas para variável Comissões.

Comissões	Contagem	Fequência Rel	Percentual
Sim	317	0,3783	37,8%
Não	521	0,6217	62,2%

Tabela 5. Estatísticas descritivas para variável Risco.

Risco	Contagem	Fequência Rel	Percentual
Alto	346	0,4129	41,3%
Médio	245	0,2924	29,2%
Baixo	247	0,2947	29,5%

Estatísticas para variáveis numéricas.

Dentre as várias estatísticas possíveis selecionamos a média, desvio padrão e as medidas utilizadas para construir um Boxplot, ou seja: mínimo, primeiro quartil, mediana, terceiro quartil e máximo.

EXCEL – MÉDIA, DESVPAD: Média e desvio padrão para a variável "Retorno_2005":

- =MÉDIA(Retorno_2005),
- =DESVPAD(Retorno_2005).

EXCEL – MENOR, MAIOR: Mínimo e máximo para a variável "Retorno_2005":

- =MENOR(Retorno_2005;1), o número 1 indica o primeiro menor.
- =MAIOR(Retorno_2005;1), o número 1 indica o primeiro maior,
- =MÍNIMO(Retorno_2005),
- =MÁXIMO(Retorno_2005).

EXCEL – MED: Mediana para a variável "Retorno_2005":

- =MED(Retorno_2005).

EXCEL – QUARTIL: Quartis para a variável "Retorno_2005":

- =QUARTIL(Retorno_2005;1), o número 1 indica o primeiro quartil,
- =QUARTIL(Retorno_2005;3), o número 3 indica o terceiro quartil.

Observação: A função QUARTIL(dados;n) fornece também o mínimo (n=0), a mediana (n=2) e o máximo (n=3).

Nas Tabelas 6 a 5 são apresentadas as estatísticas para as variáveis Ativos, Prop.Despesas, Retorno_2005, Retorno_3_Anos e Retorno_5_Anos.

Tabela 6. Estatísticas descritivas para variável Ativos.

Ativos				
	Média	1.755,60	**Máximo**	71.536,40
	Desvio Padrão	5.770,27	**Quartil 3**	1.070,13
	Coef. Variação	328,7%	**Mediana**	385,30
			Quartil 1	143,85
			Mínimo	42,10

Tabela 7. Estatísticas descritivas para variável Prop.Despesas.

Prop.Despesas				
	Média	1,21	**Máximo**	2,93
	Desvio Padrão	0,40	**Quartil 3**	1,41
	Coef. Variação	32,7%	**Mediana**	1,20
			Quartil 1	0,99
			Mínimo	0,15

Tabela 8. Estatísticas descritivas para variável Retorno_2005.

Retorno 2005				
	Média	7,34	**Máximo**	25,30
	Desvio Padrão	4,53	**Quartil 3**	10,30
	Coef. Variação	61,6%	**Mediana**	6,90
			Quartil 1	4,30
			Mínimo	-5,10

Tabela 9. Estatísticas descritivas para variável Retorno_3_Anos.

Retorno 3 Anos	**Média**	17,80	**Máximo**	42,30
	Desvio Padrão	4,95	**Quartil 3**	21,08
	Coef. Variação	27,8%	**Mediana**	17,20
			Quartil 1	14,10
			Mínimo	6,70

Tabela 10. Estatísticas descritivas para variável Retorno_5_Anos.

Retorno 5 Anos	**Média**	3,38	**Máximo**	26,50
	Desvio Padrão	6,97	**Quartil 3**	7,40
	Coef. Variação	206,5%	**Mediana**	2,50
			Quartil 1	-1,60
			Mínimo	-26,50

3 Gráficos no Excel

Gráficos para variáveis categóricas

Para as variáveis categóricas utilize gráficos de Colunas, Barras, Pizza, Rosca, ou gráfico dinâmico. Para inserir um gráfico: **Inserir → Gráficos**.

Considerando a variável Risco, podemos obter os gráficos apresentados nas Figuras 2 e 3.

Figura 2. Gráficos para a distribuição de frequências das categorias da variável Rico, gráfico de Pizza à esquerda e gráfico de Rosca à direita.

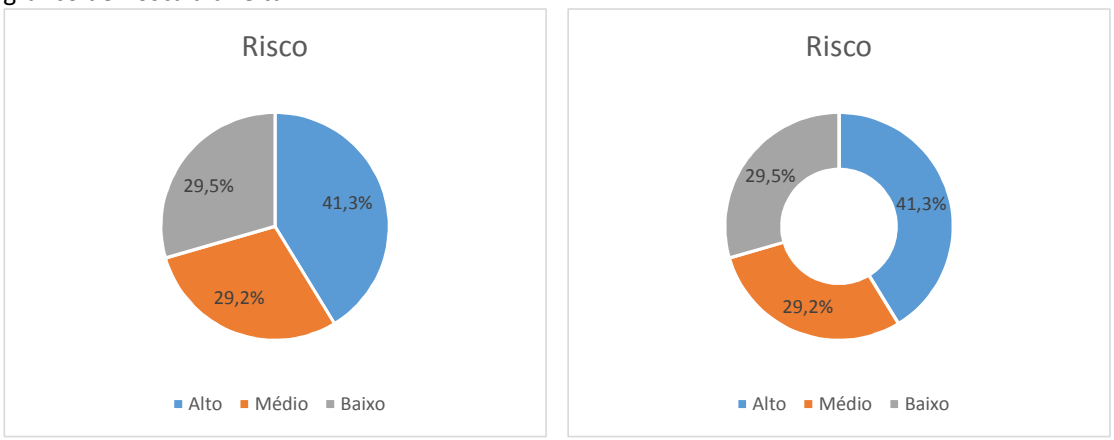
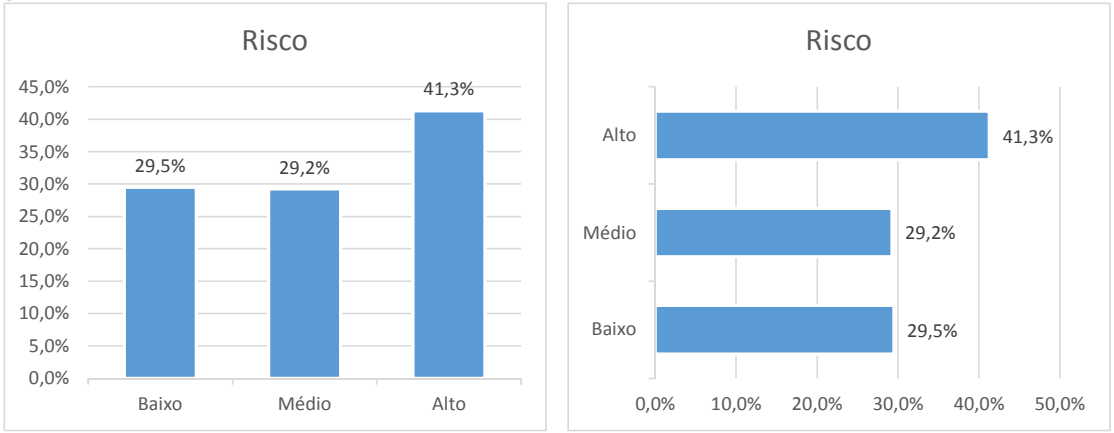
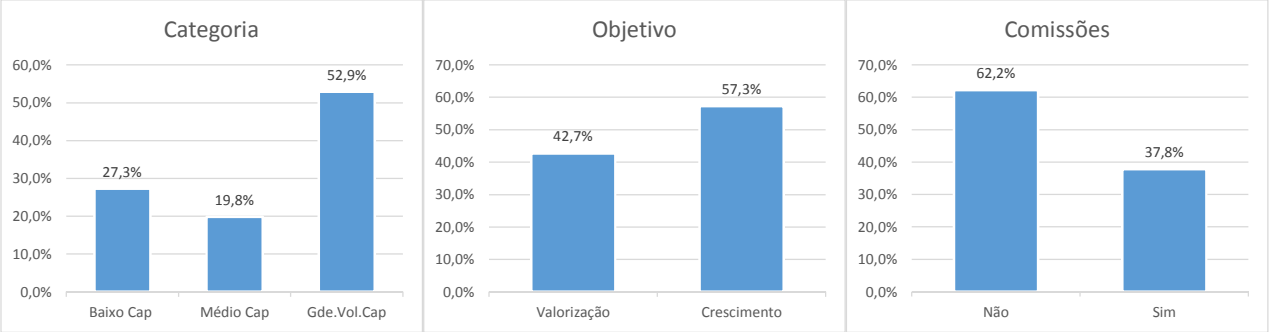


Figura 3. Gráficos para a distribuição de frequências das categorias da variável Rico, gráfico de Colunas à esquerda e gráfico de Barras à direita.



A seguir, na Figura 4, são apresentados gráficos de colunas para as variáveis qualitativas: Categoria, Objetivo, Comissões.

Figura 4. Gráficos para a distribuição de frequências das categorias da variável Categoria (esquerda), Objetivo (centro) e Comissões (direita).



Gráficos para variáveis numéricas

Para as variáveis numéricas, podemos considerar gráficos que apresentem distribuição dos valores, como por exemplo, histogramas ou Boxplot. No Excel não há gráfico Boxplot, mas um gráfico Boxplot pode ser feito no Excel a partir de um **Gráfico de linhas com marcadores**. Para tanto ordene os cinco números do gráfico Boxplot de acordo com a Figura 5. Selecione a coluna com os valores da Tabela para Boxplot e depois insira o gráfico: **Inserir → Gráficos → Gráfico de Linhas com Marcadores**.

O gráfico de linhas precisa ser "alterado": considere os passos:

- em **Ferramentas de Gráfico → Design → Dados** clique em **Alternar Linha/Coluna**.
- Linha do Boxplot: **Ferramentas de Gráfico → Design → Adicionar Elemento Gráfico → Linhas** e selecione **Linhas de Máximo/Mínimo**.
- Caixa do Boxplot: **Ferramentas de Gráfico → Design → Adicionar Elemento Gráfico → Barras Superiores Inferiores** e selecione **Barras Superiores Inferiores**.

Figura 5. Organização de dados para Boxplot da variável Ativos

Ativos	**Média**	1.755,60	**Máximo**	71.536,40	
	Desvio Padrão	5.770,27	**Quartil 3**	1.070,13	
	Coef. Variação	328,7%	**Mediana**	385,30	
			Quartil 1	143,85	
			Mínimo	42,10	

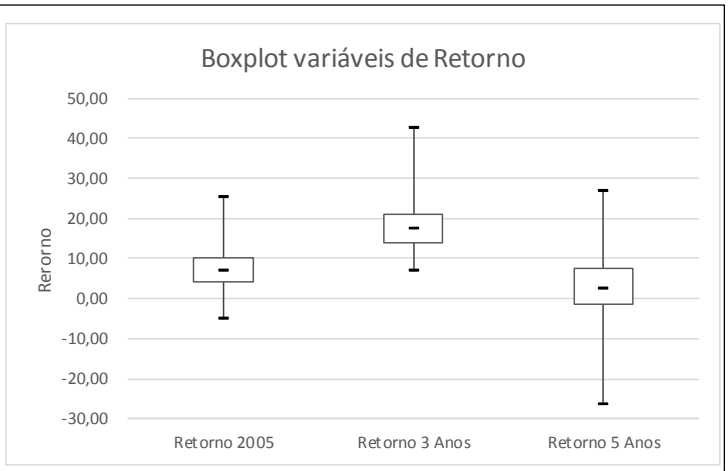
Tabela para Boxplot	
Quartil 3	1.070,13
Máximo	71.536,40
Mediana	385,30
Mínimo	42,10
Quartil 1	143,85

Na Figura 6 são apresentadas gráficos boxplots para as variáveis Prop.Despesas e Ativos, no caso de Ativos a escala do eixo-y é apesentada em Log. Na Figura 7 são apresentados os gráficos boxplots para as variáveis de retorno, os valores para os gráficos foram organizados na tabela apresentada também na Figura 7.

Figura 6. Gráficos Boxplot (diagrama de caixa) para as variáveis Prop.Despesas (esquerda) e Ativos (direita).

Figura 7. Organização de resultados (esquerda) e Gráfico com três boxplots para as variáveis de retorno (direita).

	Retorno 2005	Retorno 3 Anos	Retorno 5 Anos
Quartil 3	10,30	21,08	7,40
Máximo	25,30	42,30	26,50
Mediana	6,90	17,20	2,50
Mínimo	-5,10	6,70	-26,50
Quartil 1	4,30	14,10	-1,60



3 Gráficos Dinâmicos para explorar dados multivariados no Excel

Um gráfico dinâmico é uma ferramenta do Excel para criar gráficos e resumir dados.

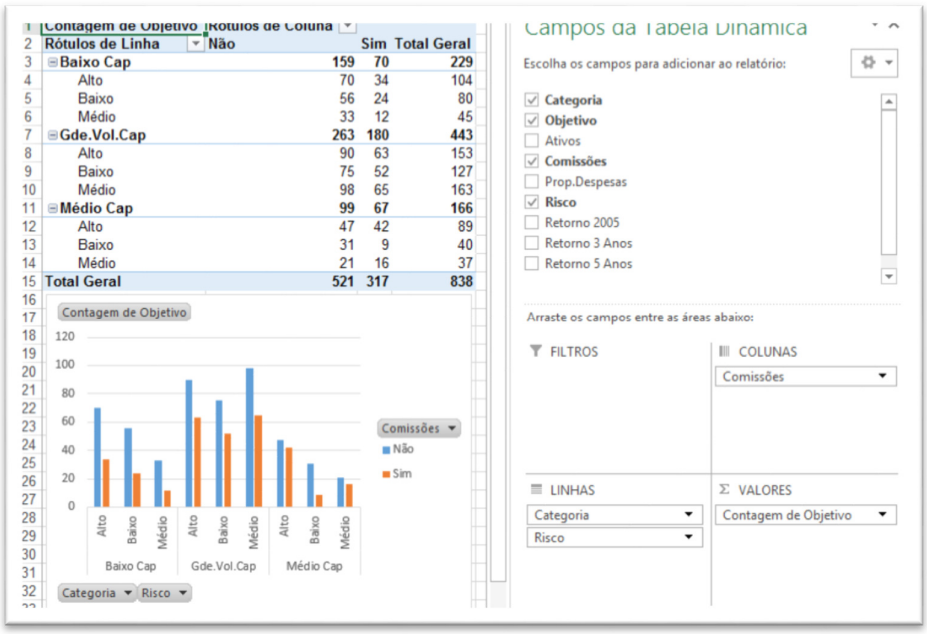
Utilize a planilha ***Dados*** do arquivo ***Choice_is_Yours_Fundos.xlsx***.

Para criar um gráfico dinâmico com resumo dos dados selecione: **Inserir → Gráficos → Gráfico Dinâmico**. Na sequência, na caixa de diálogo "Criar Gráfico Dinâmico" :

6. Clique em **Selecionar ou intervalo** e informe o intervalo para **Tabela/Intervalo**.
7. Selecione a opção **Nova planilha** e clique em **OK**.

Aparece um painel de tarefas "Campos do Gráficos Dinâmicos". Marque as variáveis categóricas e arraste-as para linha ou coluna. Veja o resultado do gráfico na Figura 8.

Figura 8. Gráfico dinâmico, exemplo.



4 Estatística descritiva no software R.

A base de dados foi gravada em arquivo do tipo csv, o arquivo ***Choice_is_Yours_Fundos.csv***. A utilização do software R fica facilitada com o arquivo de comandos ***Choice_is_Yours_Fundos.R***.

No ambiente do software R, mude de diretório: **Arquivo → Mudar dir...** Depois, abra o arquivo com comandos, no ambiente do R selecione: **Arquivo → Abrir script** .O script contém linhas com comentários (linhas que começam com #) e linhas com comandos que serão executados pelo software.

Para executar um comando clique na linha de comando (sem selecionar trechos da linha) e digite **CTRL** + **R**. A seguir são apresentados alguns dos "comandos".

R Leitura de dados: **dados <- read.table("Choice_is_Yours_Fundos.csv",header=TRUE,sep=";",dec=",")**

dados é um "*data frame*" com todas as variáveis. Para obter as estatísticas descritivas utilize ***summary***:

A Figura 9 apresenta o resultado do comando ***summary***, observe que para as variáveis categóricas são apresentadas frequências e para as variáveis numéricas, mínimo, Q1, mediana, Q3, média, máximo.

Figura 9. Principais estatísticas descritivas obtidas no R pelo comando ***summary***.

```
> summary(dados)
       Categoria          Objetivo       Ativos        Comissões Prop.Despesas
 Baixo Cap  :229   Crescimento:480   Min.    :   42.1   Não:521   Min.   :0.150
 Gde.Vol.Cap:443   Valorização:358   1st Qu.:  143.8   Sim:317   1st Qu.:0.990
 Médio Cap  :166                     Median :  385.3             Median :1.200
                                     Mean   : 1755.6             Mean   :1.211
                                     3rd Qu.: 1070.1             3rd Qu.:1.410
                                     Max.   :71536.4             Max.   :2.930
   Risco      Retorno.2005    Retorno.3.Anos  Retorno.5.Anos
 Alto :346   Min.   :-5.100   Min.   : 6.70   Min.   :-26.500
 Baixo:247   1st Qu.: 4.300   1st Qu.:14.10   1st Qu.: -1.600
 Médio:245   Median : 6.900   Median :17.20   Median :  2.500
             Mean   : 7.343   Mean   :17.80   Mean   :  3.376
             3rd Qu.:10.300   3rd Qu.:21.07   3rd Qu.:  7.400
             Max.   :25.300   Max.   :42.30   Max.   : 26.500
> |
```

O comando ***summary*** apresenta as principais medidas para cada variável, mas podemos estudar cada variável através de comandos específicos.

IMPORTANTE: Para que o software reconheça o nome das variáveis de forma simplificada devemos anexar as variáveis do *data frame* utilizando o comando ***attach***. Assim, execute o comando **attach(dados)**.

: table(x); table(x,y); table(y,x); table(x,y,z).

R Para contagens ou tabulação cruzada com variáveis categóricas utilize o comando ***table***, por exemplo: **table(Categoria)**, **table(Risco)**, **table(Risco,Categoria)**, **table(Categoria,Objetivo,Risco)**. Veja alguns resultados na Figura 10.

Figura 10. Tabelas com contagem de categorias obtidas pelo comando ***table*** no R.

```
> table(Categoria)
Categoria
 Baixo Cap  Gde.Vol.Cap   Médio Cap
       229         443         166
> table(Risco)
Risco
 Alto Baixo Médio
  346   247   245
> table(Categoria,Risco)
             Risco
Categoria     Alto Baixo Médio
  Baixo Cap    104    80    45
  Gde.Vol.Cap  153   127   163
  Médio Cap     89    40    37
> table(Risco,Categoria)
       Categoria
Risco   Baixo Cap  Gde.Vol.Cap Médio Cap
  Alto         104         153        89
  Baixo         80         127        40
  Médio         45         163        37
> |
```

R Para estatísticas básicas utilize funções estatísticas como, ***mean*** (média), ***sd*** (desvio padrão), ***sum*** (soma de valores), ***median*** (mediana), ***quantile*** (quantil). Veja os comandos e os resultados na Figura 11.

Figura 11. Estatísticas descritivas de variáveis numéricas com funções do R.

```
> # Estatísticas descritivas ----------------------------------------------------
> min(Retorno.2005)               # Mínimo
[1] -5.1
> max(Retorno.2005)               # Máximo
[1] 25.3
> mean(Retorno.2005)              # Média
[1] 7.34284
> sd(Retorno.2005)                # Desvio Padrão
[1] 4.525173
> sum(Retorno.2005)               # Soma de valores
[1] 6153.3
> median(Retorno.2005)            # Mediana
[1] 6.9
> quantile(Retorno.2005, p=0.75) # Quantil
 75%
10.3
> quantile(Retorno.2005, p=c(0.0, 0.25, 0.5, 0.75, 1.0)) # Números para boxplots
  0%  25%  50%  75% 100%
-5.1  4.3  6.9 10.3 25.3
> #---------------------------------------------------------------------------
```

R Para obter as estatísticas descritivas de acordo com categorias (de forma análoga à Tabela Dinâmica do Excel) utilize o comando **tapply**. Veja alguns exemplos na Figura 12.

Figura 12. Estatísticas descritivas por grupo pelo comando tapply no R.

```
> # Estatísticas descritivas por grupo ------------------------------
> # Média de Retorno.2005 por categorias de Risco:
> tapply(Retorno.2005,Risco,mean)
   Baixo    Médio     Alto
7.441040 7.184615 7.363673
> # Média de Retorno.2005 por categorias de Risco e Objetivo:
> tapply(Retorno.2005,list(Objetivo,Risco),mean)
               Baixo    Médio     Alto
Crescimento 7.442169 7.475000 8.156061
Valorização 7.414286 7.164502 6.438053
> #-------------------------
> # Desvio padrão de Retorno.2005 por categorias de Risco:
> tapply(Retorno.2005,Risco,sd)
   Baixo    Médio     Alto
4.988275 3.801691 4.522162
> # Desvio padrão de Retorno.2005 por categorias de Risco e Objetivo:
> tapply(Retorno.2005,list(Objetivo,Risco),sd)
               Baixo    Médio     Alto
Crescimento 4.855809 5.171267 4.428641
Valorização 7.745853 3.702425 4.472941
> #---------------------------------------------------------------------
```

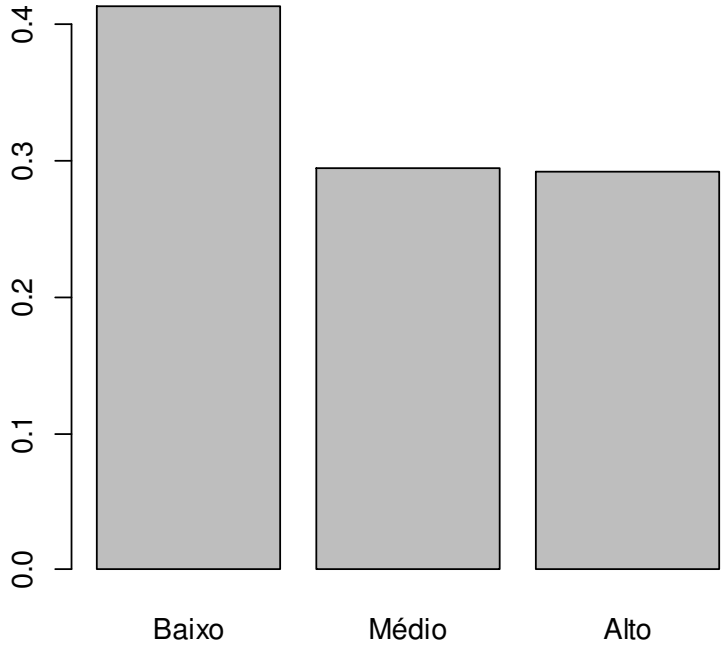
5 Gráficos no software R.

R **Gráficos para as variáveis categóricas**. A partir de tabelas podemos criar gráficos com o comando **barplot**. Para tanto o resultado da tabela é gravado numa variável e depois utiliza-se o comando **barplot**. Por exemplo, para a variável **Risco**, podemos utilizar os seguintes comandos:

```
> tRisco <- table(Risco)                # Cálculo da frequência em cada categoria
> prop.tRisco <- prop.table(tRisco)     # Transforma a tabela em frequências relativas
```

> barplot(prop.tRisco) # Faz o gráfico de barras das frequências (ver Figura 13)

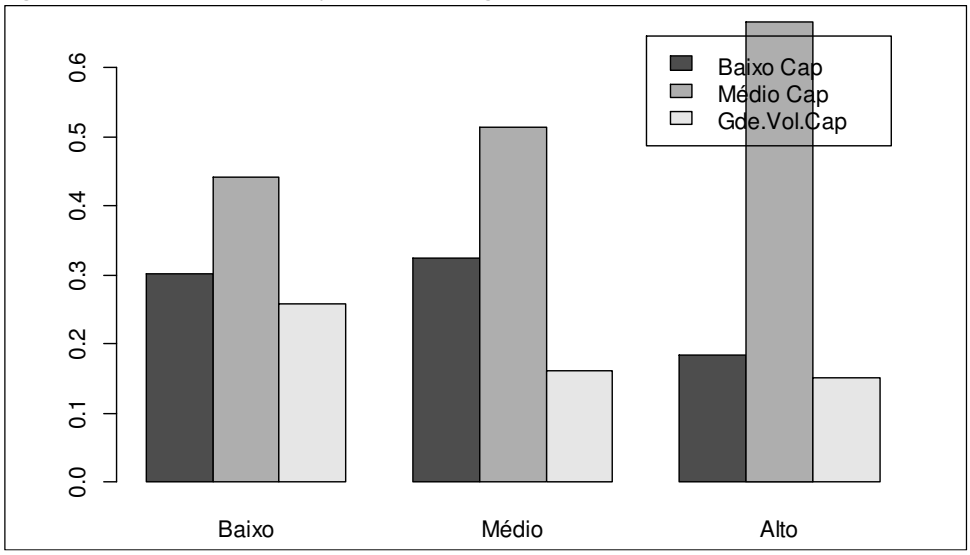
Figura 13. Resultado de **barplot** para a variável Risco, comando: **barplot(prop.table(table(Risco)))**.



R **Gráficos para duas variáveis categóricas**. O procedimento é análogo, armazenar a tabela e utilizar barplot. Veja o resultado na Figura 14.

>Cat.Risco <- table(Risco,Categoria)
>barplot(prop.table(t(Cat.Risco),2),beside=T,legend.text=colnames(Cat.Risco))

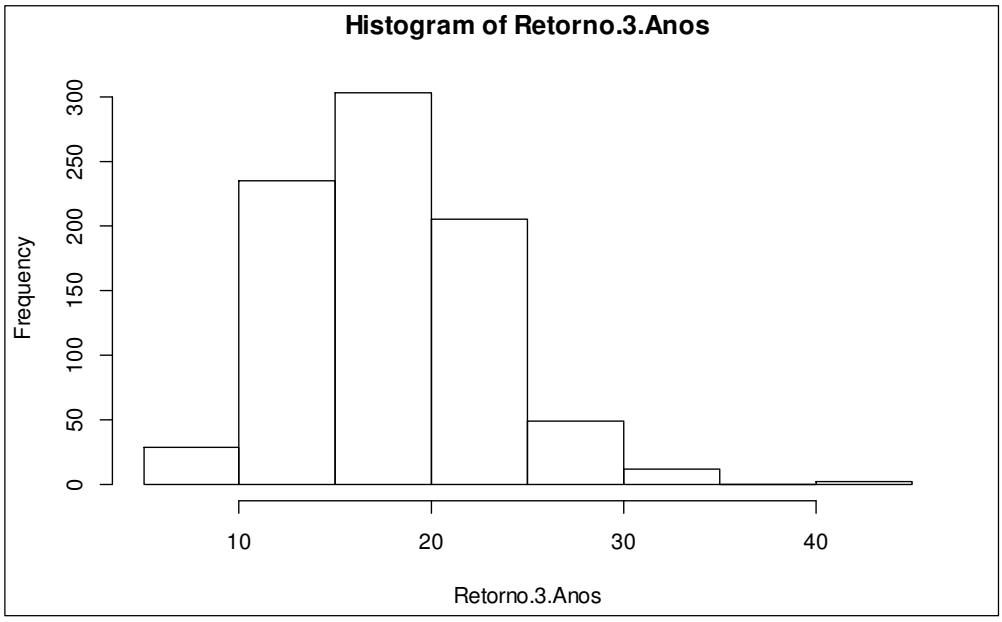
Figura 14. Gráfico de barras para duas categorias.



R **Gráficos para as variáveis numéricas**. Selecionando dois tipos de gráficos: Histograma e Boxplot.

R **Histograma.** Para fazer um histograma utilize o comando hist, por exemplo: **hist(Retorno.3.Anos)** fornece o gráfico apresentado na Figura 15.

Figura 15. Histograma para a variável Retorno.3.Anos.

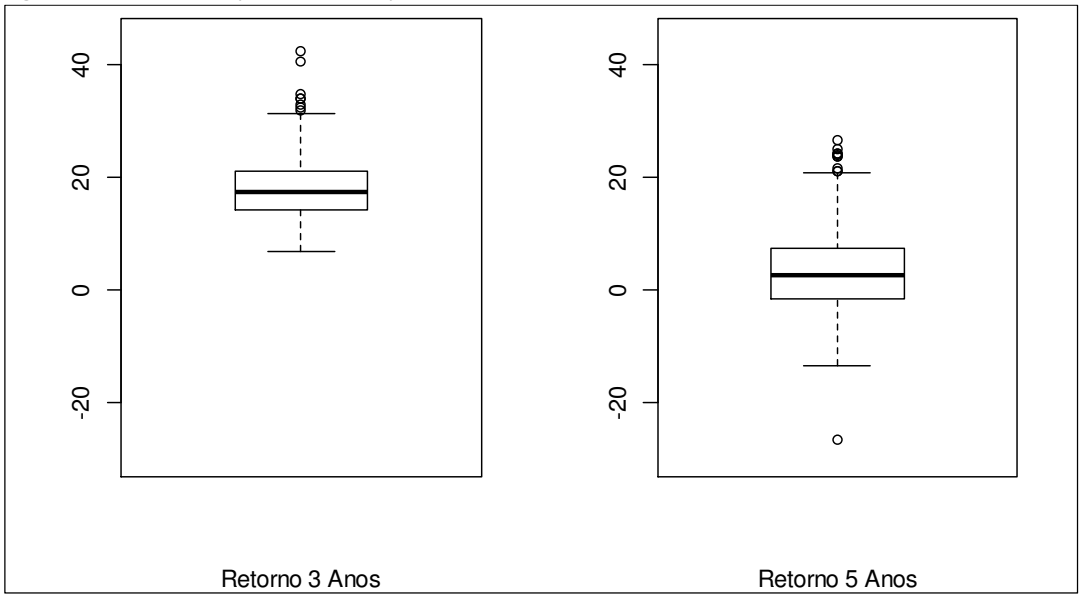


Boxplots para duas variáveis distintas. Para fazer boxplot, utilize diretamente o comando boxplot(variável).

No caso abaixo os comandos produzem dois boxplots para variáveis distintas, veja figura 16.

```
>#  Boxplot para duas variáveis distintas, por ex.
>#  Retorno.3.Anos e Retorno.5.Anos ----------
>par(mfrow=c(1,2))                                  # parâmetro para fazer gráficos em uma linha e duas colunas
>ylim <- range(-30,45)                              # limites para o eixo-y
>boxplot(Retorno.3.Anos,xlab="Retorno 3 Anos",ylim=ylim)     # Boxplot da esquerda, para Retorno.3.Anos
>boxplot(Retorno.5.Anos,xlab="Retorno 5 Anos",ylim=ylim)     # Boxplot da direita, para Reorno.5.Anos
>par(mfrow=c(1,1))                                  # parâmetro para retornar gráficos 1x1
```

Figura 16. Resultado para dois Boxplot de duas variáveis distintas.

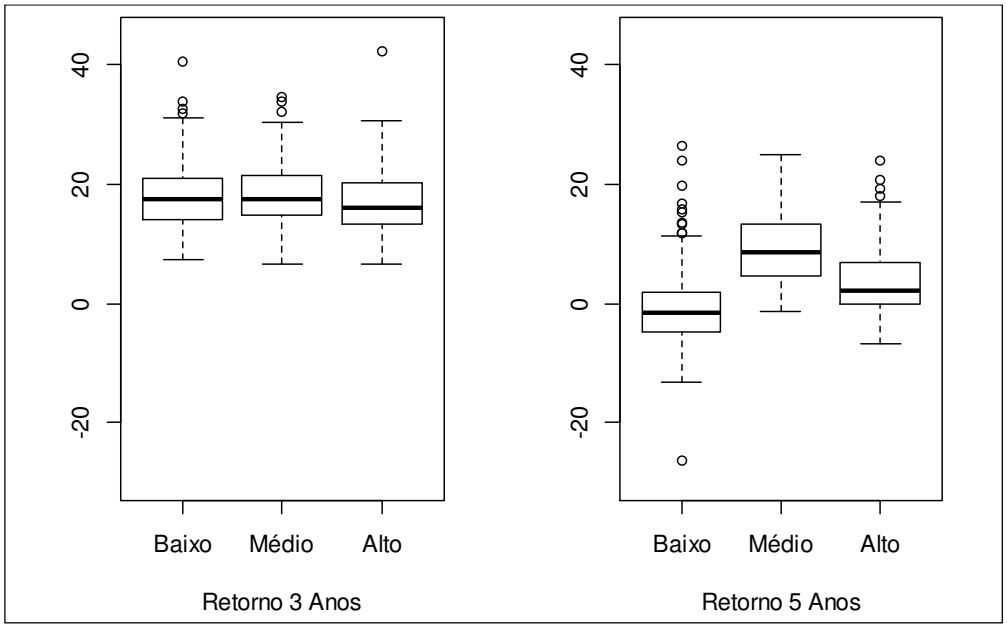


Boxplots para uma variável de acordo com categorias distintas de uma segunda variável (categórica). O comando é o mesmo, mas agora adicionamos a variável categórica, por exemplo Risco. O resultado é apresentado na Figura 17.

```
>#  Boxplot para duas variáveis distintas, por ex.
```

```
>#  Retorno.3.Anos e Retorno.5.Anos ----------
>par(mfrow=c(1,2))                                  # parâmetro para fazer gráficos em uma linha e duas colunas
>ylim <- range(-30,45)                              # limites para o eixo-y
>boxplot(Retorno.3.Anos~Risco,xlab="Retorno 3 Anos",ylim=ylim)      # Variável Risco inserida
>boxplot(Retorno.5.Anos~Risco,xlab="Retorno 5 Anos",ylim=ylim)      # Variável Risco inserida
>par(mfrow=c(1,1))                                  # parâmetro para retornar gráficos 1x1
```

Figura 17. Boxplots para vriáveis Retorno 3 Anos e Retorno 5 Anos de acordo com níveis de Risco.



6 Instalação de pacotes adicionais no R

Em geral, no software R, temos a necessidade de instalar pacotes adicionais para realizar algumas análises específicas. Um pacote bastante útil é o lattice, que possibilita a construção de gráficos em análise multivariada. Para instalar um pacote siga os passos abaixo:

- Na janela do software R selecione: **Pacotes → Instalar pacote(s)**.
- Selecione o CRAN Mirror (espelho CRAN), por exemplo, **→ Brasil SP(2) → [ok]**
- Selecione o nome do pacote e depois [ok], por exemplo: **→ lattice → [ok]**
- Agora, para utilizar o pacote sempre digite o comando: **library("lattice")**