

# Hidrologia Estatística

Apostila de Walter Collischonn

Adaptado ao Python por Cláudio Bielenki Jr

Novembro 2019

```
In [1]: from __future__ import division
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
```

## 1 Introdução

A estatística reúne os métodos para coleta, resumo, apresentação e análise de dados bem como na obtenção de conclusões válidas e na tomada de decisões razoáveis baseadas em tais análises. A estatística é importante em quase todas as áreas da ciência, mas é especialmente importante na Hidrologia, nas ciências da terra como Geografia e Geologia e nas Engenharias.

No passado, tratar uma grande quantidade de dados numéricos era uma tarefa tediosa, cansativa e sujeita a erros de cálculo diversos. O desenvolvimento dos computadores e de diversos programas permite que grande quantidade de dados possa ser analisada rapidamente. Por outro lado, o uso de computadores e programas sofisticados não exclui a possibilidade de ocorrerem erros na interpretação dos resultados, especialmente no caso de pessoas sem o mínimo conhecimento teórico.

A Estatística pode ser dividida em três áreas:

- Estatística Descritiva
- Probabilidade
- Inferência Estatística

A Estatística Descritiva é utilizada nas etapas iniciais de análise, quando os dados são coletados e verificados pela primeira vez. Para uma primeira análise dos dados são usados métodos para apresentação dos dados e métodos para resumo dos dados. A Estatística Descritiva pode ser definida como um conjunto de técnicas destinadas a descrever e resumir os dados, a fim de que possamos tirar conclusões a respeito de características de interesse.

A Probabilidade pode ser pensada como a teoria matemática utilizada para se estudar a incerteza oriunda de fenômenos de caráter aleatório. Esta área não será aprofundada neste curso, mas apenas abordada de forma superficial em casos de problemas simples.

A Inferência Estatística é o conjunto de técnicas que possibilitam tirar conclusões sobre fenômenos que atingem um grande conjunto de dados a partir da análise de um pequeno conjunto de dados, como no caso das pesquisas eleitorais, que buscam prever o resultado de uma eleição em que votam vários milhões de eleitores a partir da entrevista de apenas alguns milhares de eleitores. Assim, o objetivo da Inferência Estatística é obter respostas corretas de questões específicas, atendendo a um determinado grau de acerto.

Para alguns estudiosos a estatística é uma arte; para outros a estatística é a simples aplicação do bom senso. Em qualquer caso, a estatística ajuda a tomar decisões com informações incompletas, tendo presente que o sucesso da decisão dependerá da habilidade do analista para compreender os resultados das informações contidas nos dados. A primeira parte do processo decisório é a estatística descritiva e a outra é a inferência estatística.

## 2 Coleta de dados

A Estatística lida com dados, números dentro de um contexto. Entretanto, a utilização de estatística é mais do que trabalhar com números, pois embora a organização dos números e a construção de gráficos possa ser mecanizada com softwares e modelos, as idéias e bons julgamentos, por enquanto, não podem ser automatizados. O analista deve ter o hábito de perguntar, por exemplo, o que mostram os resultados dentro de um determinado contexto? Quais as respostas que os dados podem dar a perguntas específicas?

Tenha em mente que durante a apresentação da disciplina Estatística é realizada uma análise explanatória de dados conhecidos, não havendo, em geral, nenhuma pergunta em mente. Entretanto, na prática diária da estatística são procuradas respostas a perguntas específicas, por exemplo, quais indivíduos (medições, pessoas, animais, taxas de juros e outras coisas) devem ser estudados? Que variáveis devem ser medidas? Nesses casos, em geral, os dados devem ser gerados. Este é o caso na hidrologia, em que devem ser medidos os dados de chuva, vazão, nível dos rios, granulometria dos sedimentos etc.

Os dados requeridos pela análise são obtidos pesquisando dados disponíveis, ou gerando novos dados. Em hidrologia e hidrometria podem ser realizadas análises estatísticas sobre dados já existentes, como as séries de dados de chuva em um determinado posto pluviométrico, ou sobre dados novos, como no caso de uma série de testes de esvaziamento de um tanque, ou de medição de capacidade de infiltração do solo.

## Classificação dos dados

Como o procedimento estatístico a ser aplicado dependerá da natureza dos dados ou das observações de cada variável, deve-se desenvolver a habilidade de distinguir os tipos de dados possíveis e suas unidades de medida. Quanto a sua natureza, as observações ou dados se classificam em quantitativas discretas e contínuas, qualitativas nominais e ordinais, de corte transversal e séries temporais.

**Dados quantitativos.** Refere-se a quantidades medidas numa escala numérica, em geral, acompanhadas de alguma unidade de medida e podem ser de dois tipos discretos ou contínuos.

- **Dados discretos.** Referem-se aos valores numéricos que assumem somente números inteiros positivos 0,1,2,3 .... Os dados discretos resultam, em geral, de contagens: a quantidade de vendas diárias de uma empresa. o número de filhos das famílias de uma

quantidade de vendas de uma empresa, o número de movimentos da conta corrente dos clientes de um banco comercial, a quantidade de peças defeituosas em um lote de produção, o número de transações financeiras com erro de lançamentos, o número de acidentes nas estradas durante as férias anuais de verão etc.

- **Dados contínuos.** Referem-se aos valores numéricos que assumem qualquer valor do conjunto dos números reais. Os dados contínuos resultam, em geral, de medições que podem ter grande precisão: a estatura dos alunos de uma turma, o consumo mensal de energia elétrica de uma casa, o tempo de espera na fila do banco, o tempo de espera na parada de ônibus, o coeficiente de escoamento de um vertedor.

**Dados qualitativos.** Refere-se às observações não numéricas e são classificadas em nominais e ordinais.

- **Dados nominais.** Estes dados não tem ordenamento nem hierarquia. Por exemplo, o sexo dos funcionários registrados no cadastro de uma empresa, o estado civil, o nome das empresas que tem ações negociadas na bolsa de valores, etc.
- **Dados ordinais.** Estes dados são semelhantes aos nominais, mas incluem uma ordem, uma hierarquia. Por exemplo, o cargo dos funcionários numa empresa: presidente, diretor, gerente, etc. Ou a escala em um questionário de avaliação de satisfação de um cliente: ótimo, bom, regular, ruim, péssimo. O grau de nebulosidade de um dia: claro, nublado, encoberto.

## Tipos de variáveis

As variáveis podem ser obtidas de duas formas, dependendo de como os valores são obtidos ao longo do tempo, ou se o tempo desempenha algum papel na análise: séries temporais ou cortes transversais numa data ou período.

**Séries temporais.** As observações são dados de uma mesma variável em diferentes períodos de tempo: o valor do PIB anual de um país, a taxa mensal de desemprego numa região, as cotações diárias de uma ação, a rentabilidade mensal de uma empresa, a demanda de energia elétrica diária na região Sudeste medida às dezoito horas etc.

**Corte transversal numa data ou período.** Se na coleta dos dados não for considerada a sequência temporal; por exemplo, amostras da quantidade produzida e do preço médio dos produtos, ou das vendas e do investimento em propaganda, a média de apartamentos vendidos durante o último mês pelas primeiras dez imobiliárias da cidade, o número de operações fechadas por cinco ações numa determinada data etc.

## População e amostra

População é o conjunto total unidades elementares de pessoas, objetos ou coisas sobre as quais se querem obter informações. Um subconjunto de unidades elementares selecionadas de uma população é denominado amostra.

Uma população pode ser formada por todos os habitantes de um país, ou de um estado, ou de um município etc. Um exemplo de pesquisa de uma população completa é o censo demográfico do Brasil realizado pelo IBGE. A análise das vendas de um segmento da economia, por exemplo, o de montadoras de carros, durante o mesmo ano é outro exemplo de população. Entretanto, nem sempre é conveniente obter informações de todas as pessoas, objetos ou coisas de uma

população. Os resultados de uma pesquisa de intenção de voto de todos os eleitores do país numa eleição presidencial não conseguiriam captar do que os partidos políticos necessitam, pois o tempo necessário para coletar todas as opiniões comprometeria os resultados, além de ser muito cara para a finalidade que se propõe. Em alguns casos, a restrição de consultar toda a população é econômica, como é o caso da determinação da vida útil das lâmpadas que obrigaria a testar todas as lâmpadas produzidas, não restando nenhuma para a venda! Dessa maneira, o procedimento recomendado é escolher uma amostra representativa de um lote de lâmpadas produzidas.

Uma amostra aleatória de tamanho  $n$ , retirada de uma população é uma das muitas possíveis e igualmente prováveis combinações de  $n$  unidades elementares que podem ser retiradas de uma população. Portanto, qualquer amostra de tamanho  $n$  tem a mesma probabilidade de ser selecionada.

Na hidrologia é comum termos variáveis na forma de séries temporais, como por exemplo, as chuvas totais anuais em um determinado local. Neste caso a população seriam todos os anos (a idade da Terra), enquanto uma amostra seriam, por exemplo, 30 anos de dados.

Mas na hidrologia também são comuns os dados na forma de cortes transversais, onde o tempo não importa. As características do solo, por exemplo, variam no espaço. Para caracterizar perfeitamente a porosidade do solo esta característica deveria ser medida em uma infinidade de pontos (população) cobrindo toda a região de interesse. É claro que isso é impossível de realizar, assim é necessário fazer apenas umas poucas medições, talvez dez ou vinte, para estimar as características. Este conjunto também seria chamado amostra.

### 3 Apresentação de dados

A apresentação dos dados de forma organizada é necessária para obter informações a partir de uma amostra, especialmente se a amostra for relativamente grande, o que sempre é desejável. A organização dos dados é normalmente feita na forma de gráficos e tabelas.

#### Tabela de dados discretos

A forma mais simples de apresentar dados estatísticos é na forma de tabelas. Qualquer análise estatística normalmente parte da elaboração de uma tabela com os dados. Uma das tabelas mais utilizadas é a tabela de freqüências. A freqüência do valor de uma variável é o número de repetições deste valor dentro da amostra. Considere a seqüência de números a seguir, que correspondem ao número de dias de chuva no mês de janeiro em um determinado local nos últimos 26 anos.

14	12	13	11	12	13	16	14	14	15	17	14	11
13	14	15	13	12	14	13	14	13	15	16	12	12

Podemos obter algumas informações apenas olhando para estes números, como o máximo (17) e o mínimo (11), mas temos dificuldades em aprofundar a análise. Uma tabela de freqüências absolutas mostra quantas vezes ocorreu cada um dos valores da variável número de dias de chuva em janeiro. Olhando a tabela de freqüências absolutas observamos facilmente que o valor

mais freqüente é 14, e que o valor 17 ocorreu apenas uma vez. Muitas vezes as tabelas de freqüência absoluta são transformadas em tabelas de freqüência relativa, dividindo as freqüências absolutas pelo tamanho da amostra.

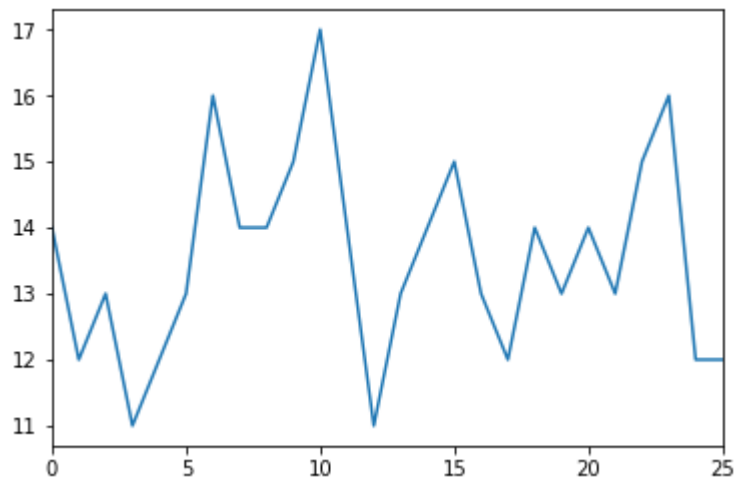
```
In [2]: dias_chuva = [14, 12, 13, 11, 12, 13, 16, 14, 14, 15, 17, 14, 11, 13, 14, 15, 13,
```

```
In [3]: len(dias_chuva)
```

```
Out[3]: 26
```

```
In [4]: pd.Series(dias_chuva).plot()
```

```
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0xae20d90>
```



## Frequência Absoluta

```
In [5]: Sdias_chuva = pd.Series(dias_chuva)  
pd.DataFrame(Sdias_chuva.value_counts(sort=False), columns=['Contagem'])
```

```
Out[5]:
```

	Contagem
11	2
12	5
13	6
14	7
15	3
16	2
17	1

## Frequência Relativa

```
In [6]: pd.DataFrame(Sdias_chuva.value_counts(normalize=True, sort=False), columns=['Porcentagem'])
```

```
Out[6]:
```

	Porcentagem
11	0.076923
12	0.192308
13	0.230769
14	0.269231
15	0.115385
16	0.076923
17	0.038462

Esta tabela permite observações ainda mais aprofundadas, por exemplo, podemos verificar que em pouco mais de um quarto (26,92%) dos anos da amostra o mês de janeiro apresentou 14 dias de chuva. Em alguns casos o interesse da análise dos dados reside em conhecer os valores da variável que são maiores que um determinado limite, por exemplo, o número de anos em que janeiro teve mais de 15 dias de chuva. Neste caso é útil elaborar a tabela de freqüências acumuladas. A freqüência acumulada do valor de uma variável é a soma das freqüências absolutas ou relativas desde o valor inicial da variável. No exemplo anterior relativo ao número de dias de chuva em janeiro podemos elaborar a tabela com freqüências acumuladas (tanto absolutas como relativas):

## Freqüência Relativa Acumulada

```
In [7]: pd.DataFrame(Sdias_chuva.value_counts(normalize=True, sort=False).cumsum(), columns=['Porcentagem'])
```

```
Out[7]:
```

	Porcentagem
11	0.076923
12	0.269231
13	0.500000
14	0.769231
15	0.884615
16	0.961538
17	1.000000

## Tabela de dados contínuos

No caso de dados contínuos recomenda-se trabalhar com intervalos de valores para a contagem de freqüência. Isto ocorre porque é praticamente impossível contar a freqüência de dados contínuos, e a tabela teria um número excessivo de linhas e freqüências baixas para cada um dos

valores, inviabilizando qualquer análise. Considere os dados de chuva anual em um determinado local (medidos em mm), dados na tabela a seguir.

1421	1234	1326	1187	1281	1311	1600	1489	1492	1522	1709	1490	1101
1393	1414	1505	1333	1201	1444	1380	1477	1329	1540	1603	1267	1299

Para criar uma tabela de freqüências como a anterior, teríamos que ter 609 linhas, para cada um dos valores entre 1101 e 1709. Isto não é prático e não faz sentido. Assim, na contagem de freqüência somamos o número de anos em que o valor da variável está em cada intervalo. Podemos definir, por exemplo, intervalos de 100 mm: de 1000 a 1100; de 1100 a 1200; de 1200 a 1300; de 1300 a 1400; de 1400 a 1500; de 1500 a 1600; de 1600 a 1700; e de 1700 a 1800.

```
In [8]: chuva_anual = [1421, 1234, 1326, 1187, 1281, 1311, 1600, 1489, 1492, 1522, 1709,
```

```
In [9]: dfchuva_anual = pd.DataFrame(chuva_anual)
dfchuva_anual.columns=['Chuva Anual']
dfchuva_anual
```

Out[9]:

	Chuva Anual
0	1421
1	1234
2	1326
3	1187
4	1281
5	1311
6	1600
7	1489
8	1492
9	1522
10	1709
11	1490
12	1101
13	1393
14	1414
15	1505
16	1333
17	1201
18	1444
19	1380
20	1477
21	1329
22	1540
23	1603
24	1267
25	1299

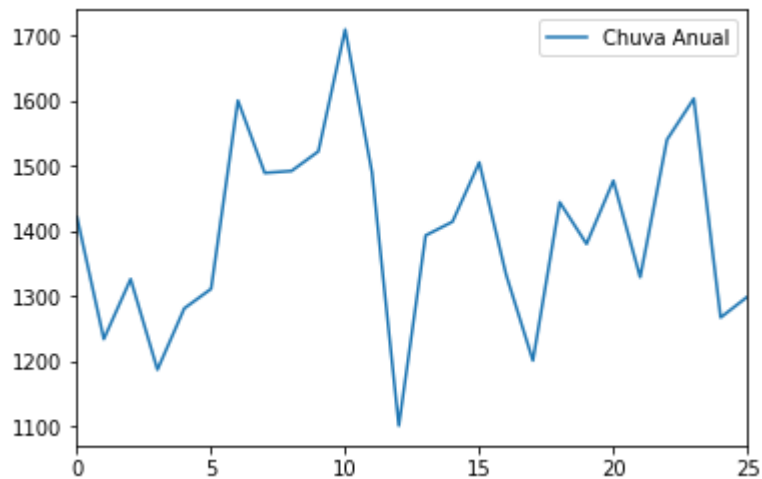


```
In [10]: print dfchuva_anual.size  
print dfchuva_anual['Chuva Anual'].max()  
print dfchuva_anual['Chuva Anual'].min()  
print dfchuva_anual['Chuva Anual'].mean()
```

```
26  
1709  
1101  
1398.0
```

```
In [11]: dfchuva_anual.plot()
```

```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0xc5a38b0>
```

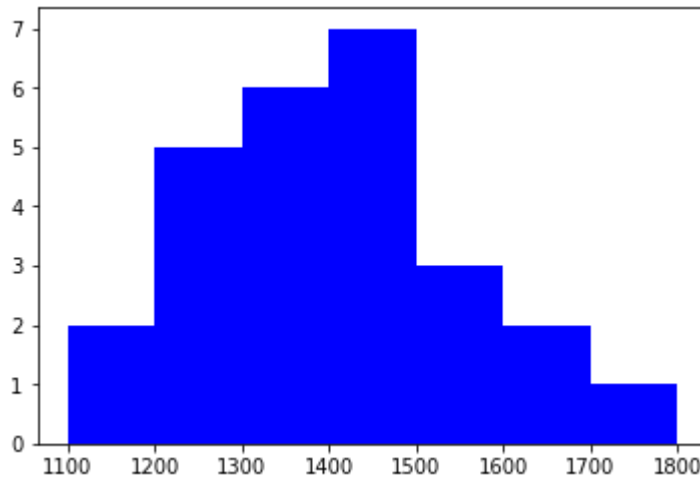


O número (k) de intervalos em que deve ser dividida uma amostra é arbitrário, mas um critério que pode ser utilizado é dado por:  $k = \sqrt{n}$  onde n é o tamanho da amostra e k deve ser arredondado para o valor mais próximo.

## Histograma

O histograma é um gráfico que permite visualizar as tabelas de frequência de forma rápida, utilizando barras verticais para representar as frequências. O histograma pode ser feito representando frequências relativas ou absolutas (a figura será igual, somente os valores serão diferentes).

```
In [12]: hist = plt.hist(dfchuva_anual['Chuva Anual'], bins=range(1100, 1900,100), color=
```



```
In [13]: hist
```

```
Out[13]: (array([2., 5., 6., 7., 3., 2., 1.]),  
          array([1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800]),  
          <a list of 7 Patch objects>)
```

A partir de um histograma podemos ver facilmente qual é a faixa normal de precipitações anuais neste local e quais são os valores que mais ocorrem. Com base nesta figura seria fácil dizer que um ano com 2500 mm de chuva foi um ano extremamente chuvoso, e que um ano com apenas 600 mm de chuva foi um ano extremamente seco. Por outro lado, um ano com 1250 mm de chuva seria um ano razoavelmente normal, pelo que nos mostra o histograma.

## Frequência Absoluta

```
In [14]: tabfreqabs = pd.DataFrame(hist[0],index=hist[1][: -1])
tabfreqabs
```

Out[14]:

	0
1100	2.0
1200	5.0
1300	6.0
1400	7.0
1500	3.0
1600	2.0
1700	1.0

## Frequência Relativa

```
In [15]: tabfreqrel = pd.DataFrame(hist[0]/hist[0].sum(),index=hist[1][: -1])
tabfreqrel
```

Out[15]:

	0
1100	0.076923
1200	0.192308
1300	0.230769
1400	0.269231
1500	0.115385
1600	0.076923
1700	0.038462

## Frequência Relativa Acumulada

```
In [16]: tabfreqrelacum = pd.DataFrame(tabfreqrel.cumsum())
tabfreqrelacum
```

Out[16]:

	0
1100	0.076923
1200	0.269231
1300	0.500000
1400	0.769231
1500	0.884615
1600	0.961538
1700	1.000000

## 4 Resumo de dados estatísticos

Para tentar conhecer as características de uma população, extraímos uma amostra desta população e analisamos esta amostra. A análise pode começar organizando os dados como descrito no capítulo anterior. A partir daí procuramos extrair mais informação utilizando formas de resumir os dados da amostra em alguns poucos valores que representam de forma razoavelmente fiel a variabilidade dos dados da amostra. Entre os valores que são usados para resumir os dados de uma amostra estão a média, o desvio padrão, a moda, o coeficiente de variação e o coeficiente de assimetria. Outra ferramenta útil para resumir os dados de uma amostra é a análise baseada no ordenamento, baseada no simples ato de colocar em ordem crescente ou decrescente todos os valores da amostra. A partir do ordenamento da amostra é possível obter informações que resumem a amostra como a mediana, os percentis ou quantis, e, especialmente, os quartis, que são bastante utilizados.

### Média

A média é o valor obtido pela soma de todos os valores dos dados da amostra dividida pelo tamanho da amostra, como apresentado na equação abaixo:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

onde  $x_i$  são os valores;  $n$  é o tamanho da amostra (número de valores que devem ser somados); e  $\bar{x}$  é a média. Calculadoras permitem obter o valor da média de uma sequência de valores de forma muito rápida.

### Mediana

A mediana é o valor que é superado por 50% dos pontos da amostra. A média e a mediana podem ter valores relativamente próximos, porém não iguais. A mediana pode ser obtida organizando os  $n$  valores  $x_i$  da amostra em ordem crescente. Sendo  $x_i$  com  $i = 1$  a  $n$ , os valores de  $x$  organizados em ordem decrescente, a mediana é obtida por:

$$Mediana = x_p \text{ com } p = \frac{n-1}{2} + 1 \text{ se } n \text{ for ímpar,}$$

$$\text{e } Mediana = \frac{x_p + x_{p+1}}{2} \text{ se } n \text{ for par.}$$

Ao contrário da média, a mediana não é uma função encontrada em qualquer calculadora. Para calcular a mediana de um conjunto grande de dados o ideal é utilizar uma planilha de cálculo no computador, como o programa Excel, por exemplo.

## O desvio padrão

O desvio padrão é uma medida de dispersão dos valores de uma amostra em torno da média. O desvio padrão é dado por:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

o quadrado do desvio padrão  $s^2$  é chamada variância da amostra.

## A moda

A moda é o valor que mais freqüente da amostra. Assim como a média e a mediana, a moda indica uma tendência dos valores aleatórios.

A moda pode ser utilizada para valores numéricos ou para dados não numéricos, como categorias, cores, nomes de pessoas ou modelos de carros. Também é possível definir a moda para intervalos de valores numéricos.

```
In [17]: print pd.Series(dias_chuva).max()
print pd.Series(dias_chuva).min()
print pd.Series(dias_chuva).mean()
print pd.Series(dias_chuva).std()
print pd.Series(dias_chuva).mode()
```

```
17
11
13.538461538461538
1.5292029095125141
0    14
dtype: int64
```

```
In [18]: dfchuva_anual.describe()
```

Out[18]:

	Chuva Anual
<b>count</b>	26.000000
<b>mean</b>	1398.000000
<b>std</b>	144.559469
<b>min</b>	1101.000000
<b>25%</b>	1302.000000
<b>50%</b>	1403.500000
<b>75%</b>	1491.500000
<b>max</b>	1709.000000

Exemplo: Um conjunto de dois dados foi lançado 20 vezes, revelando os resultados da soma dos valores apresentados na tabela abaixo. Qual é o valor da moda?

```
In [19]: ValorSoma = range(3, 12)
ValorSoma
```

Out[19]: [3, 4, 5, 6, 7, 8, 9, 10, 11]

```
In [20]: Nocorr = [1, 3, 1, 4, 5, 2, 3, 0, 1]
```

```
In [21]: tabela = pd.DataFrame(Nocorr, ValorSoma)
tabela.columns=['Contagem']
tabela.index.name='Soma'
tabela
```

Out[21]:

Contagem	
Soma	
3	1
4	3
5	1
6	4
7	5
8	2
9	3
10	0
11	1

```
In [22]: moda = tabela['Contagem'].idxmax()
moda
```

Out[22]: 7

O valor da moda é 7, porque este é o valor da soma mais freqüente na amostra. De certa forma isto poderia ser esperado, uma vez que a soma 7 pode ocorrer pela combinação de 1+6; 2+5; 3+4; 4+3; 5+2 e 6+1; enquanto os outros valores, como o 2, o 6 e o 12, tem um número de combinações possíveis menores.

A moda também pode ser calculada para variáveis contínuas, desde que se definam intervalos para a contagem de freqüência.

Exemplo: Os pesos de 15 alunos de uma turma foram medidos, com os resultados apresentados na tabela abaixo. Qual é o valor da moda?

Aluno	Peso
1	81.1
2	73.2
3	61.5
4	64.3
5	55.7
6	62.9
7	73.4
8	70.0
9	71.5
10	78.0
11	78.1
12	73.9
13	68.5
14	66.9
15	59.0

```
In [23]: import os  
os.getcwd()
```

```
Out[23]: 'c:\\HidrologiaEstatistica'
```

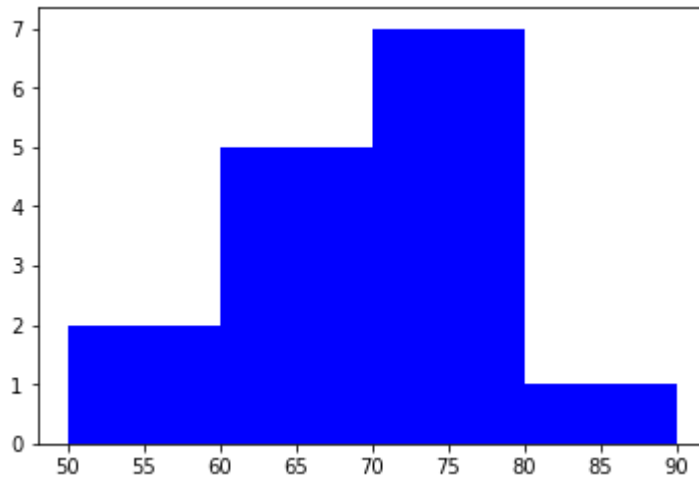
```
In [24]: pesos = pd.read_excel('pesos.xlsx', index_col=0)  
pesos
```

```
Out[24]:
```

	Peso
Aluno	
1	81.1
2	73.2
3	61.5
4	64.3
5	55.7
6	62.9
7	73.4
8	70.0
9	71.5
10	78.0
11	78.1
12	73.9
13	68.5
14	66.9
15	59.0



```
In [25]: hist_pesos = plt.hist(pesos['Peso'], bins=range(50, 100,10), color='b')
```



```
In [26]: hist_pesos
```

```
Out[26]: (array([2., 5., 7., 1.]),
          array([50, 60, 70, 80, 90]),
          <a list of 4 Patch objects>)
```

```
In [27]: tabfreqabs = pd.DataFrame(hist_pesos[0],index=hist_pesos[1][:-1], columns = ['Contagem'])
```

```
Out[27]:
```

	Contagem
50	2.0
60	5.0
70	7.0
80	1.0

```
In [28]: moda = tabfreqabs['Contagem'].idxmax()
print ('Assim, a moda é o intervalo de peso entre ' + str(moda) + ' e ' + str(moda+10))
```

Assim, a moda é o intervalo de peso entre 70 e 80 Kg

## O coeficiente de variação

O coeficiente de variação é uma relação entre o desvio padrão e a média. O coeficiente de variação é uma medida da variabilidade dos valores em torno da média, relativamente a própria média.

$$cv = \frac{s}{\bar{x}}$$

Exemplo: O seguinte conjunto de valores apresenta a chuva anual ocorrida em uma cidade ao longo de 30 anos. Calcule a média, o desvio padrão e o coeficiente de variação destes dados.

```
In [29]: chuva_anual130 = pd.DataFrame([1671, 1485, 1766, 1565, 2082, 1370, 1926, 2042, 1691, 1491, 2024, 1305, 1644, 1908, 1913, 1485, 1693, 1313, 1567, 1493, 1357, 2023, 1390, 1641, 1585, 1526, 1962, 1672, 1404, 1352],  
chuva_anual130
```

Out[29]:

	0
0	1671
1	1485
2	1766
3	1565
4	2082
5	1370
6	1926
7	2042
8	1691
9	1491
10	2024
11	1305
12	1644
13	1908
14	1913
15	1485
16	1693
17	1313
18	1567
19	1493
20	1357
21	2023
22	1390
23	1641
24	1585
25	1526
26	1962
27	1672
28	1404
29	1352

```
In [30]: chuva_anual30.describe()
```

```
Out[30]:
```

	0
<b>count</b>	30.000000
<b>mean</b>	1644.866667
<b>std</b>	241.949087
<b>min</b>	1305.000000
<b>25%</b>	1485.000000
<b>50%</b>	1613.000000
<b>75%</b>	1872.500000
<b>max</b>	2082.000000

```
In [31]: coeficienteV = chuva_anual30[0].std()/chuva_anual30[0].mean()
print coeficienteV
```

```
0.14709343459657212
```

```
In [32]: print ('A média é de ' + str(chuva_anual30[0].mean()) + ' mm por ano, o desvio pa
str(coeficienteV))
```

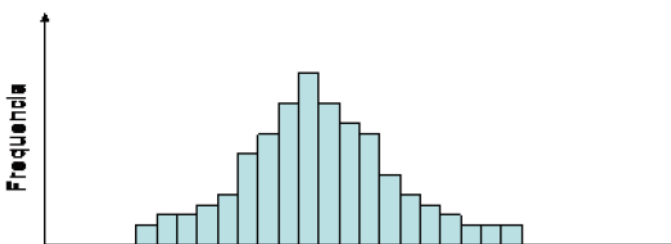
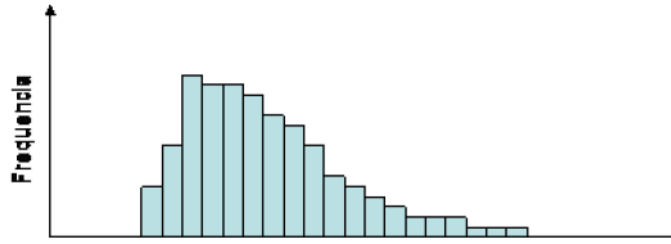
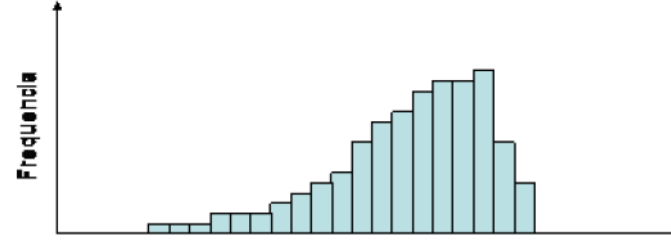
```
A média é de 1644.8666666666666 mm por ano, o desvio padrão é de 241.9490874534
1493 mm por ano e o coeficiente de variação é de0.14709343459657212
```

## O coeficiente de assimetria

O coeficiente de assimetria é um valor que caracteriza o quanto uma amostra de dados é assimétrica com relação à média. Uma amostra é simétrica com relação à média se o histograma dos dados revela o mesmo comportamento de ambos os lados da média.

$$G = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n \cdot S^3}$$

A assimetria é chamada positiva quando o valor de G é positivo e a assimetria é negativa quando o valor de G é negativo. Algumas variáveis importantes na hidrologia, como as vazões máximas anuais em rios, apresentam uma assimetria positiva.

Assimetria	Valor de G	Exemplo de histograma
Nula	0 ou próximo de zero	 <p>Um histograma com barras azuis representando a frequência de dados. A distribuição é simétrica e centrada no meio, com uma única curva de frequência que se eleva no centro e desce igualmente para ambos os lados.</p>
Positiva	$G > 0$	 <p>Um histograma com barras azuis. A distribuição é assimétrica à direita, com uma cauda longa estendendo-se para o lado direito. O pico da frequência está deslocado para a esquerda.</p>
Negativa	$G < 0$	 <p>Um histograma com barras azuis. A distribuição é assimétrica à esquerda, com uma cauda longa estendendo-se para o lado esquerdo. O pico da frequência está deslocado para a direita.</p>

O cálculo da assimetria de uma amostra é um pouco mais complexo do que o da média e do desvio padrão. A maior parte das calculadoras simples não permite calcular diretamente o coeficiente de assimetria.

## Quartis e quantis

Quantis separam a amostra de forma semelhante à mediana, porém em intervalos diferentes. Enquanto a mediana separa a amostra em dois grupos, com 50% dos dados com valores inferiores e 50% dos dados com valores superiores à mediana, os quartis e os quantis dividem a amostra em grupos de tamanhos diferentes. O primeiro Quartil é o valor que separa a amostra em dois grupos em que 25% dos pontos tem valor inferior ao quartil e 75% tem valor superior ao quartil. O terceiro Quartil é o valor que separa a amostra em dois grupos em que 75% dos pontos tem valor inferior ao quartil e 25% tem valor superior ao quartil. Já o segundo quartil é a própria mediana.

Além dos três quartis, que separam a amostra em quatro, podem ser definidos quantis arbitrários, que dividem a amostra arbitrariamente em frações diferentes. Por exemplo, o quantil 90 % divide a amostra em dois grupos. O primeiro (90% dos dados) tem valores inferiores ao quantil 90% e o segundo (10% dos dados) tem valores superiores ao quantil 90%.

## Aplicações em hidrologia

As variáveis hidrológicas como chuva e vazão têm como característica básica uma grande variabilidade no tempo. Para analisar a vazão de rio e a sua variabilidade temporal é necessário utilizar alguns valores estatísticos que resumem, em grande parte, o comportamento hidrológico do rio ou da bacia. Entre as estatísticas mais importantes estão: a vazão média, a vazão média mensal, a vazão média específica, as vazões mínimas e as vazões máximas de cada ano.

A vazão média é a média de toda a série de vazões diárias registradas, e é muito importante na avaliação da disponibilidade hídrica total de uma bacia. A vazão mediana é a vazão que é superada em 50% dos dias da série. Normalmente, a vazão média e a vazão mediana têm valores próximos, porém não iguais. A vazão média específica é a vazão média dividida pela área de drenagem da bacia.

As vazões médias mensais representam o valor médio da vazão para cada mês do ano, e são importantes para analisar a sazonalidade de um rio. O gráfico abaixo apresenta as vazões médias mensais do rio Cuiabá na seção da cidade de Cuiabá, com base nos dados de 1967 a 1999.

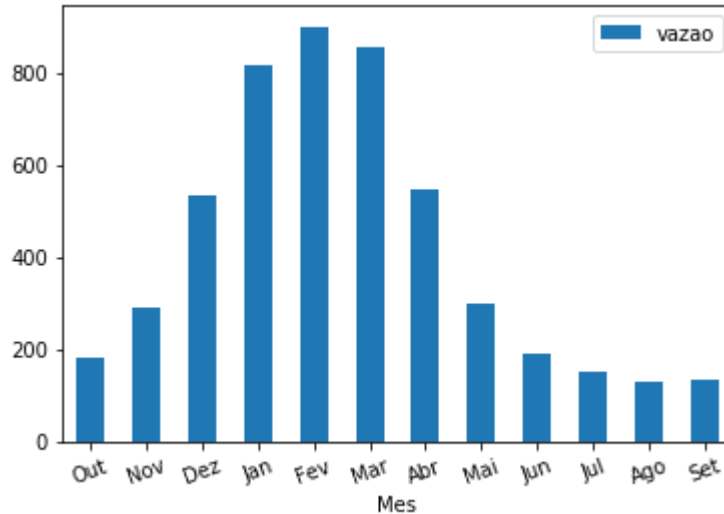
```
In [33]: vazoes = pd.read_excel('vazoes_cuiaba.xlsx', index_col='data')
vazoes_sel = vazoes['1-1-1967':'12-1-1999'].copy()
print vazoes_sel.head()
print vazoes_sel.tail()
```

	vazao
data	
1967-01-01	240.20
1967-02-01	540.52
1967-03-01	509.66
1967-04-01	430.34
1967-05-01	211.45

	vazao
data	
1999-08-01	93.36
1999-09-01	91.22
1999-10-01	105.05
1999-11-01	170.73
1999-12-01	161.67

```
In [34]: meses = ['Out', 'Nov', 'Dez', 'Jan', 'Fev', 'Mar', 'Abr', 'Mai', 'Jun', 'Jul', 'Ago', 'Set']
vazoes_sel['Mes'] = vazoes_sel.index.month
vazoes_mes = vazoes_sel.groupby('Mes').mean()
#print vazoes_mes
vazoes_mes = vazoes_mes.reindex([10, 11, 12, 1, 2, 3, 4, 5, 6, 7, 8, 9])
#print vazoes_mes
ax = vazoes_mes.plot.bar().set_xticklabels(meses, rotation=20)
```



Observa-se nesta figura que há uma sazonalidade marcada, com estiagem no inverno e vazões altas no verão. As maiores vazões mensais médias ocorrem em Fevereiro e as menores em Agosto, o que é consequência direta da sazonalidade das chuvas, que ocorrem de forma concentrada no período de verão.

Utilizando os dados de vazão média mensal da tabela abaixo, calcule as médias das vazões em cada mês.

Ano	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
1981	132	154	122	98	67	34	23	18	15	19	39	88
1982	126	191	122	75	62	43	27	19	13	22	45	101
1983	119	155	180	94	45	37	24	19	13	29	53	99
1984	108	125	170	91	83	56	29	15	15	21	39	88
1985	172	167	199	108	70	23	18	12	10	14	28	68
1986	145	187	100	93	60	30	20	18	20	23	49	108

```
In [35]: vazoes2 = pd.read_excel('vazoes2.xlsx', index_col='Ano')
print vazoes2
vazoes2.describe().mean()
```

	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Ano												
1981	132	154	122	98	67	34	23	18	15	19	39	88
1982	126	191	122	75	62	43	27	19	13	22	45	101
1983	119	155	180	94	45	37	24	19	13	29	53	99
1984	108	125	170	91	83	56	29	15	15	21	39	88
1985	172	167	199	108	70	23	18	12	10	14	28	68
1986	145	187	100	93	60	30	20	18	20	23	49	108

```
Out[35]: Jan    104.209491
Fev    125.845947
Mar    117.341360
Abr     71.865770
Mai     50.660494
Jun     30.196966
Jul     18.891902
Ago     13.640026
Set     11.957499
Out     17.376182
Nov     33.378806
Dez     71.262456
dtype: float64
```

## 5 A curva de permanência de vazões

A elaboração da curva de permanência é uma das análises estatísticas mais simples e mais úteis na hidrologia. A curva de permanência auxilia na análise dos dados de vazão com relação a perguntas como as destacadas a seguir.

- O rio tem uma vazão aproximadamente constante ou extremamente variável entre os extremos máximo e mínimo?
- Qual é a porcentagem do tempo em que o rio apresenta vazões em determinada faixa de valores?
- Qual é a porcentagem do tempo em que um rio tem vazão suficiente para atender determinada demanda?

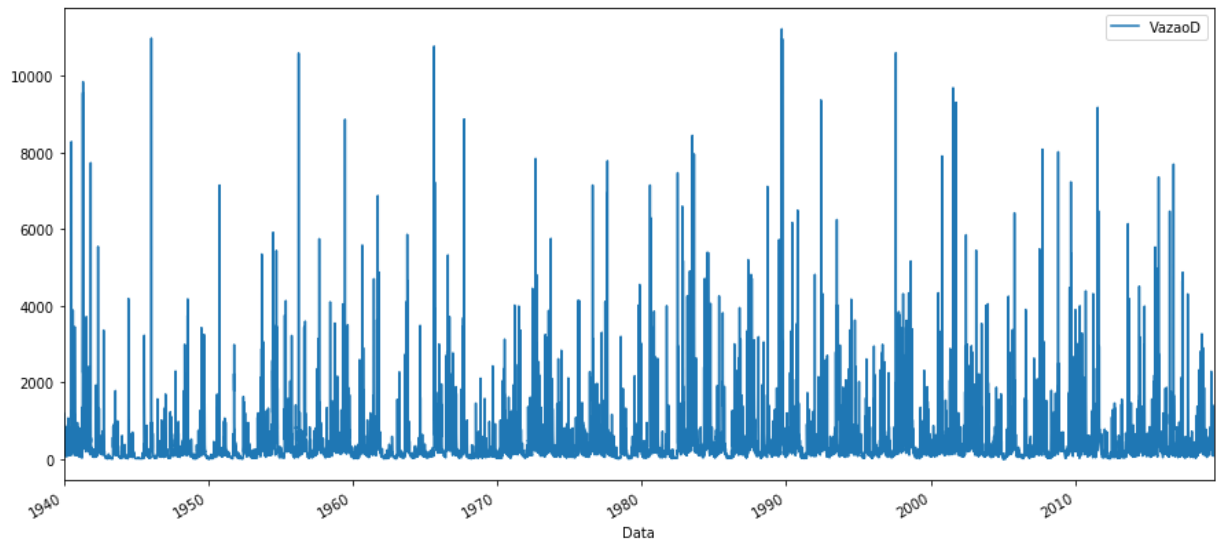
A curva de permanência expressa a relação entre a vazão e a frequência com que esta vazão é superada ou igualada. A curva de permanência pode ser elaborada a partir de dados diários ou dados mensais de vazão.

A figura a seguir apresenta o hidrograma de vazões diárias do rio Taquari, em Muçum (RS), e a curva de permanência que corresponde aos mesmos dados apresentados no hidrograma. Observa-se que a vazão de 1000 m<sup>3</sup>.s<sup>-1</sup> é igualada ou superada em menos de 10% do tempo. Apesar de apresentar picos de cheias com 7000 m<sup>3</sup>.s<sup>-1</sup> ou mais, na maior parte do tempo as vazões do rio Taquari neste local são bastante inferiores a 500 m<sup>3</sup>.s<sup>-1</sup>.



```
In [36]: vazoesD_taquari = pd.read_excel('86510000_vazaoD.xlsx', index_col='Data')
vazoesD_taquari.plot(figsize=(15,7))
```

```
Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0xca27070>
```



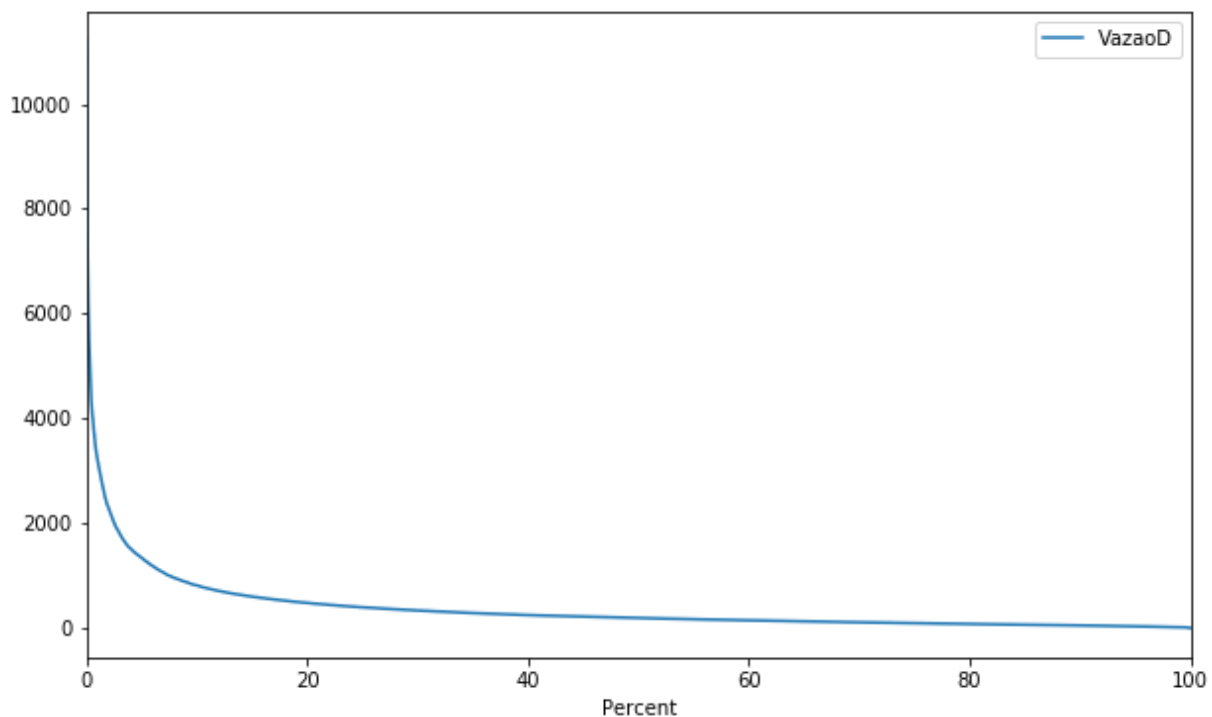
```
In [37]: NrObs = vazoesD_taquari.size
NrSup = range(1,NrObs+1)
```

```
In [38]: CP = vazoesD_taquari.copy()
CP.sort_values(by='VazaoD', ascending=False, inplace=True)
CP['NrSup'] = NrSup
CP['Percent'] = (CP['NrSup']/NrObs)*100
print CP.head()
print CP.tail()
```

	VazaoD	NrSup	Percent
Data			
1989-09-12	11213.45	1	0.003480
1946-01-26	10972.51	2	0.006960
1989-09-24	10952.52	3	0.010440
1965-08-19	10761.02	4	0.013919
1997-08-04	10591.96	5	0.017399
	VazaoD	NrSup	Percent
Data			
2005-02-02	0.0	28733	99.986081
2005-02-03	0.0	28734	99.989560
2005-02-04	0.0	28735	99.993040
2005-02-05	0.0	28736	99.996520
2005-01-08	0.0	28737	100.000000

```
In [39]: fig, ax = plt.subplots()
CP.plot(x='Percent', y='VazaoD', ax=ax, figsize=(10,6))
```

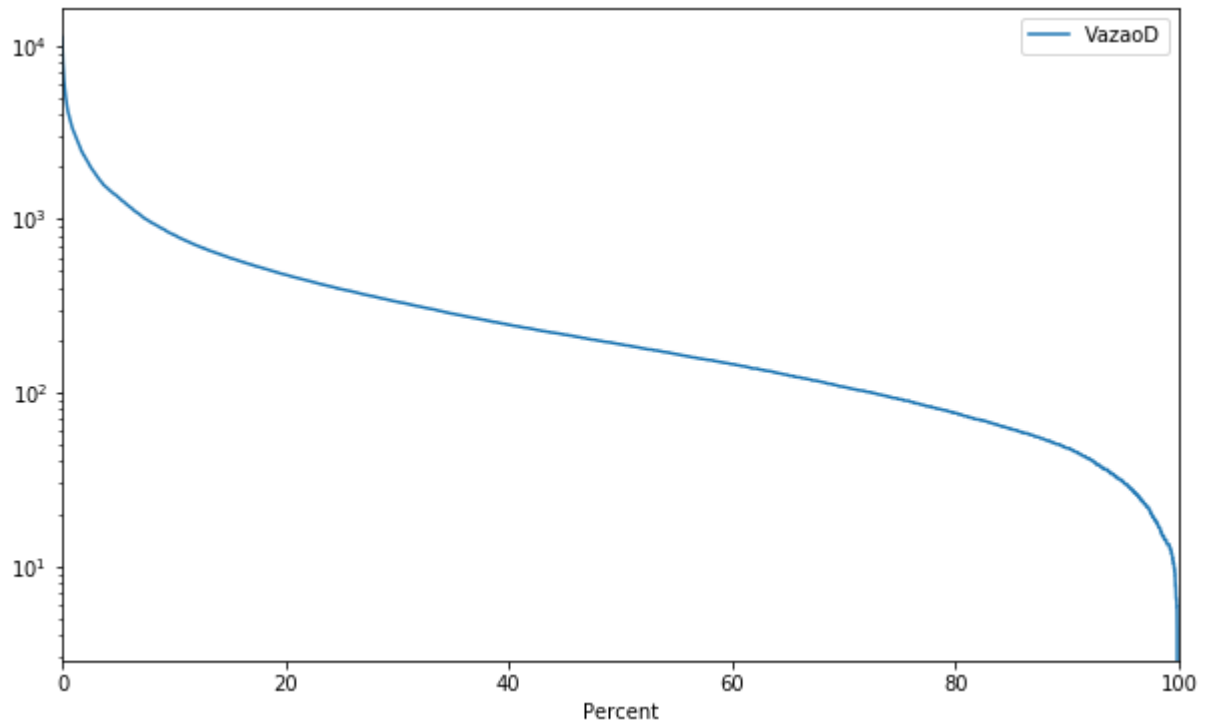
```
Out[39]: <matplotlib.axes._subplots.AxesSubplot at 0xca97a10>
```



Para destacar mais as faixas de vazões mais baixas a curva de permanência é apresentada com eixo vertical logarítmico, como mostra o gráfico abaixo.

```
In [40]: fig, ax = plt.subplots()
ax.set_yscale('log')
CP.plot(x='Percent', y='VazaoD', ax=ax, figsize=(10, 6))
```

```
Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0xc4d3c70>
```

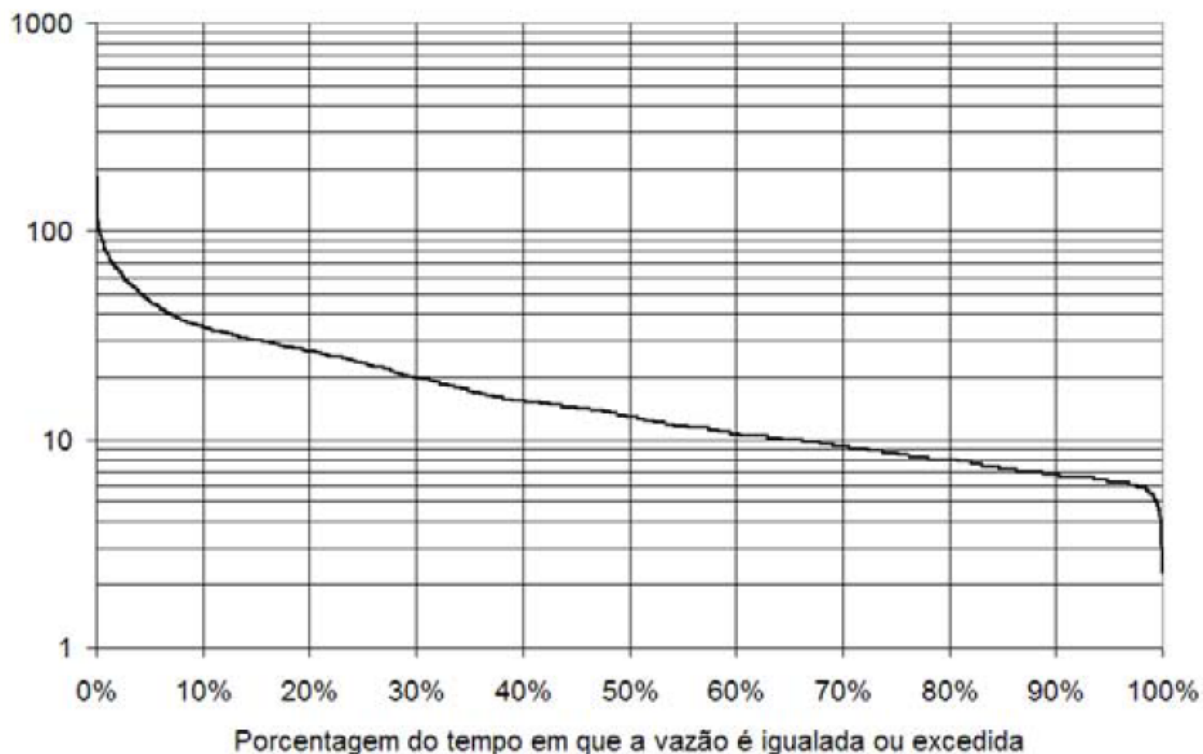


Alguns pontos da curva de permanência recebem atenção especial:

- A vazão que é superada em 50% do tempo (mediana das vazões) é a chamada Q50.
- A vazão que é superada em 90% do tempo é chamada de Q90 e é utilizada como referência para legislação na área de Meio Ambiente e de Recursos Hídricos em muitos Estados do Brasil.
- A vazão que é superada em 95% do tempo é chamada de Q95 e é utilizada para definir a Energia Assegurada de uma usina hidrelétrica.

## EXEMPLO

1) Os dados de vazão do rio Descoberto em Santo Antônio do Descoberto (GO) foram organizados na forma de uma curva de permanência, como mostra o gráfico abaixo. Um empreendedor solicita outorga de  $2,5 \text{ m}^3.\text{s}^{-1}$  num ponto próximo no mesmo rio. Considerando que a legislação permite outorgar apenas 20% da  $Q_{90}$  a cada solicitante, responda: é possível atender a solicitação?



Observa-se na curva de permanência que a vazão  $Q_{90}$  é de  $7 \text{ m}^3.\text{s}^{-1}$  aproximadamente. Portanto a máxima vazão que pode ser outorgada para um usuário individual neste ponto corresponde a:

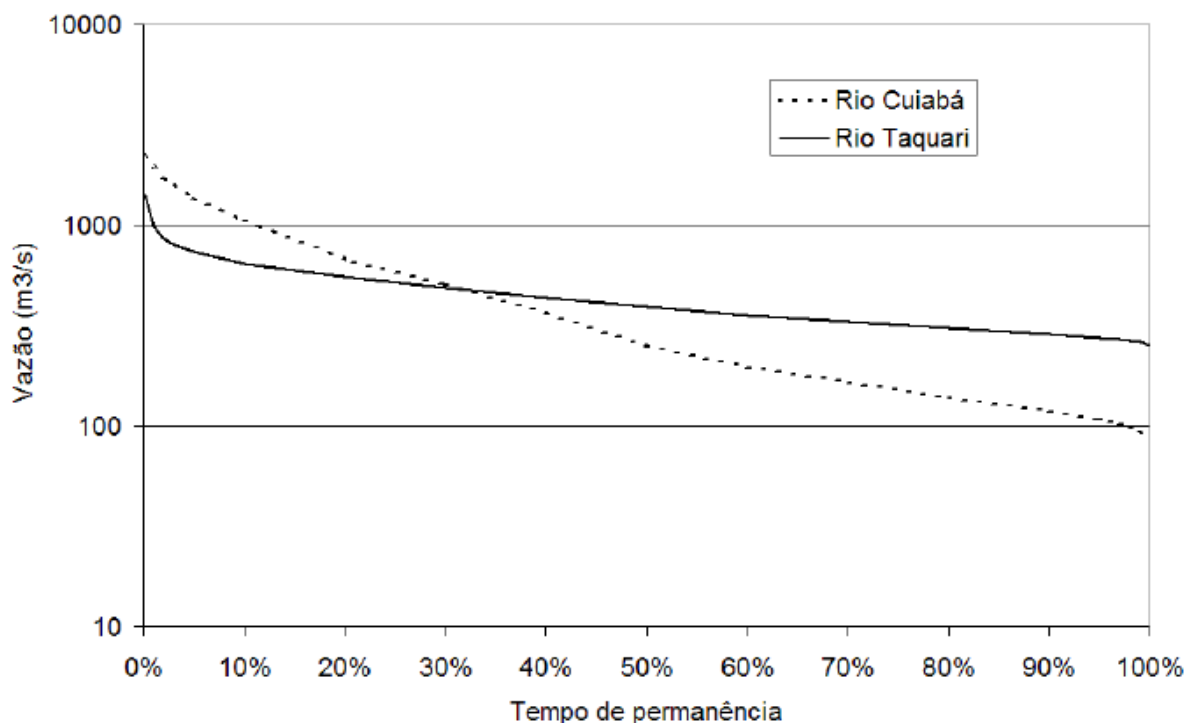
$$Q_{\text{max}} = 0,2 \cdot 7 = 1,4 \text{ m}^3.\text{s}^{-1}$$

Como o empreendedor solicitou  $2,5 \text{ m}^3.\text{s}^{-1}$  não é possível atender sua solicitação.

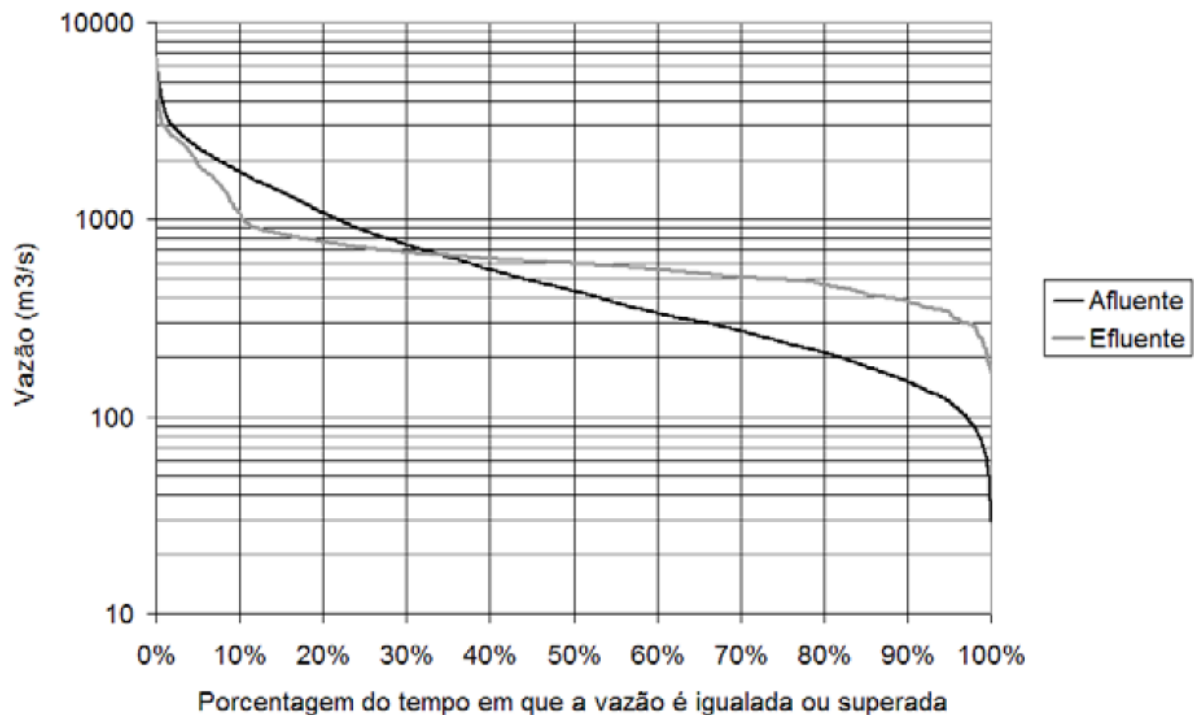
A curva de permanência também é útil para diferenciar o comportamento de rios e para avaliar o efeito de modificações como desmatamento, reflorestamento, construção de reservatórios e extração de água para uso consuntivo.

O gráfico abaixo apresenta as curvas de permanência dos rios Cuiabá, em Cuiabá (MT), e Taquari, em Coxim (MS), baseadas nos dados de vazão diária de 1980 a 1984. As duas bacias tem áreas de drenagem de tamanho semelhante. A bacia do rio Cuiabá tem, aproximadamente,  $22.000 \text{ km}^2$ , e a do rio Taquari cerca de  $27.000 \text{ km}^2$ . O relevo e a precipitação média anual são semelhantes. A vazão média do rio Cuiabá é de  $438 \text{ m}^3.\text{s}^{-1}$  neste período, enquanto a vazão média do rio Taquari é de  $436 \text{ m}^3.\text{s}^{-1}$ , ou seja, são praticamente idênticas. Entretanto, observa-se que as vazões mínimas são mais altas no rio Taquari do que no rio Cuiabá e as vazões máximas são maiores no rio Cuiabá.

O rio Cuiabá apresenta maior variabilidade das vazões, que se alternam rapidamente entre situações de baixa e de alta vazão, enquanto o rio Taquari permanece mais tempo com vazões próximas da média. Esta diferença ocorre basicamente porque a geologia da bacia do rio Taquari favorece mais a infiltração da água no solo, e esta água chega ao rio apenas após um longo período em que fica armazenada no subsolo. A vazão do rio Taquari é naturalmente regularizada pelos aquíferos existentes na bacia, enquanto que na bacia do rio Cuiabá este efeito não é tão importante.



O gráfico abaixo apresenta as curvas de permanência de vazão afluente (entrada) e efluente (saída) do reservatório de Três Marias, no rio São Francisco (MG). Este reservatório tem um grande volume e uma grande capacidade de regularização, permitindo reter grande parte das vazões altas que ocorrem durante o período do verão, aumentando a disponibilidade de água no período de estiagem. Como resultado observa-se que a vazão Q90 é alterada de 148 m³.s<sup>-1</sup> para 379 m³.s<sup>-1</sup> pelo efeito de regularização do reservatório, enquanto a vazão Q95 é alterada de 120 m³.s<sup>-1</sup> para 335 m³.s<sup>-1</sup>.



Portanto o efeito da regularização da vazão sobre a curva de permanência é torná-la mais horizontal, com valores mais próximos da mediana durante a maior parte do tempo.

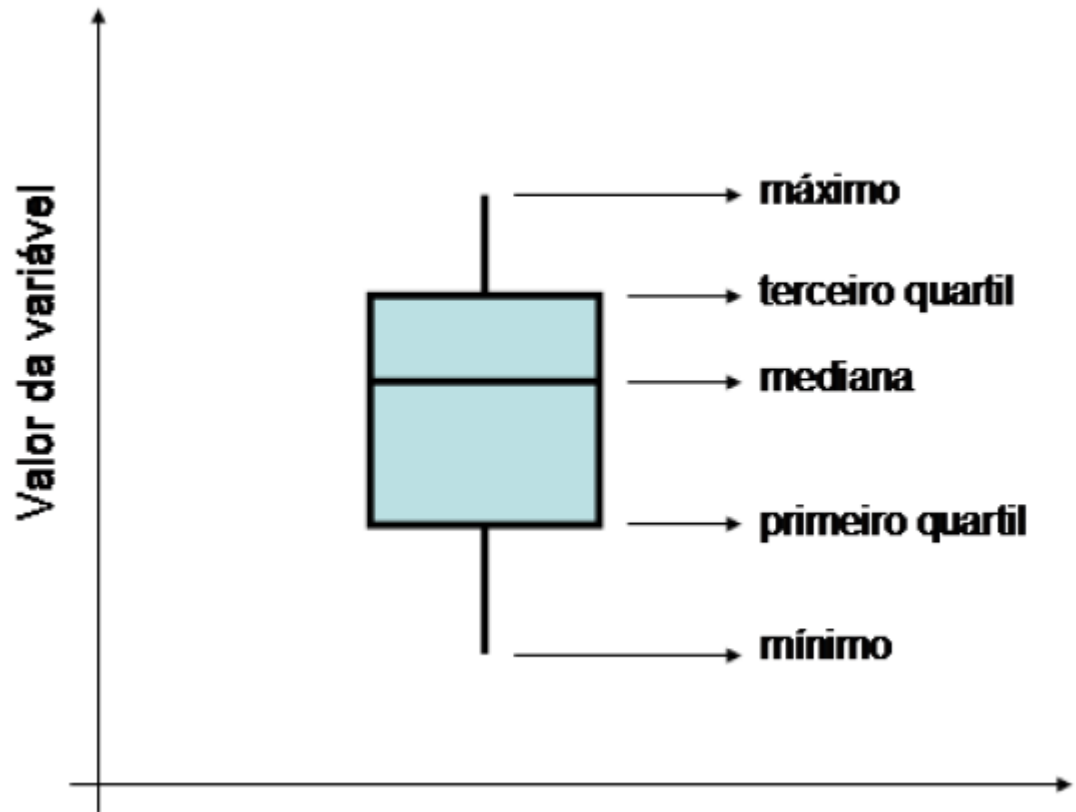
## 6 O Box-Plot

O Box plot, também conhecido como gráfico de caixa, é uma forma simples de representar graficamente a faixa de variação de uma variável, bem como algumas características de seu histograma. O Box-plot é uma representação gráfica envolvendo os quartis, a mediana, os valores máximo e mínimo.

Para elaborar o Box-plot define-se uma caixa em que os limites superior e inferior são dados pelo terceiro e pelo primeiro quartil, respectivamente. A mediana é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até os valores máximo e mínimo.

O primeiro Quartil é o valor que separa a amostra em dois grupos, em que 25% dos pontos tem valor inferior ao primeiro quartil e 75% tem valor superior ao quartil.

O segundo quartil é a mediana e o terceiro quartil é o valor que separa a amostra em dois grupos, em que 75% dos pontos tem valor inferior ao terceiro quartil e 25% tem valor superior ao terceiro quartil. A Figura abaixo apresenta um exemplo de um Box-plot.



Exemplo:

Elabore um Box plot para representar os dados referentes ao seguinte problema: Um técnico em hidrologia realizou 20 ensaios de infiltração em uma área de solo utilizado para agricultura, obtendo os seguintes resultados em mm/hora: 48; 35; 37; 52; 43; 29; 61; 33; 44; 55; 69; 43; 22; 35; 38; 57; 53; 67; 62; 48.

Solução: Estes valores podem ser organizados em ordem crescente ou decrescente permitindo encontrar os seguintes valores: a mediana é 46; o primeiro quartil (25%) é 36,5; o terceiro quartil é 55,5; o valor máximo é 69 e o mínimo é 22.

```
In [41]: import matplotlib.cbook as cbook
ensaio = [48, 35, 37, 52, 43, 29, 61, 33, 44, 55, 69, 43, 22, 35, 38, 57, 53, 67]
stats = cbook.boxplot_stats(ensaio)[0]
bxp = plt.boxplot(ensaio, showmeans=True)
print 'Mínimo: ' + str(stats['whislo'])
print 'Q25: ' + str(stats['q1'])
print 'Mediana: ' + str(stats['med'])
print 'Média: ' + str(stats['mean'])
print 'Q75: ' + str(stats['q3'])
print 'Máximo: ' + str(stats['whishi'])
```

Mínimo: 22

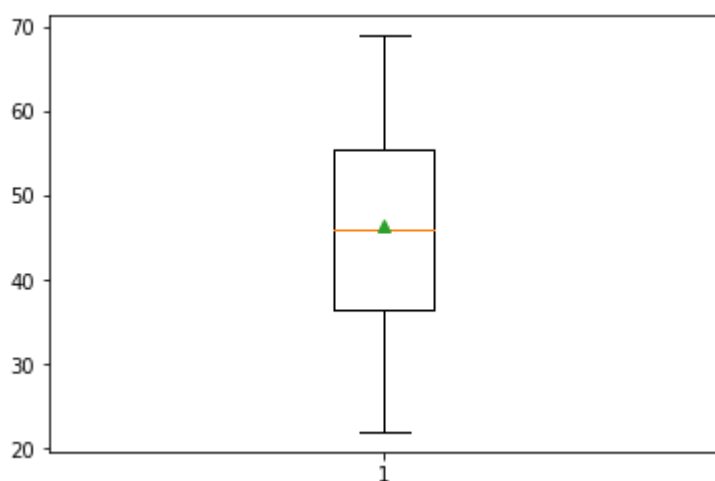
Q25: 36.5

Mediana: 46.0

Média: 46.55

Q75: 55.5


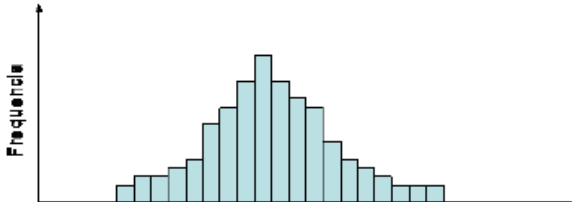
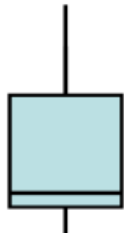
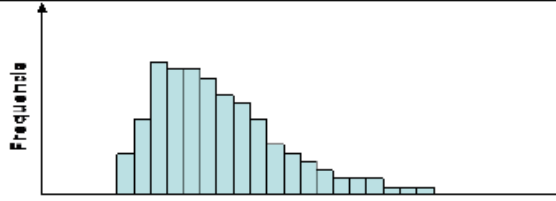
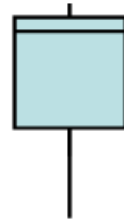
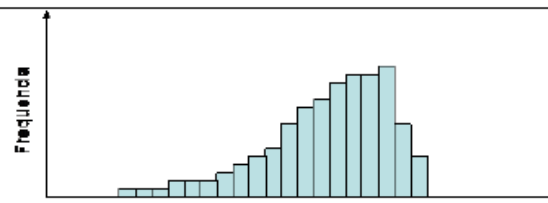
Máximo: 69



Uma importante aplicação do Box-plot é a avaliação rápida de algumas características da distribuição estatística dos dados da amostra. Neste sentido o Box-plot é quase tão útil como o histograma. Por exemplo, é relativamente simples avaliar se a distribuição dos dados é simétrica ou assimétrica observando a forma do Box-plot. Se o traço representando a mediana está no centro da caixa definida pelo primeiro e pelo terceiro quartil, a distribuição é simétrica. Se o traço representando a mediana está deslocado para cima, então a distribuição tem assimetria negativa. Se o traço da mediana está mais próximo do lado inferior da caixa, então o Box-plot mostra que os dados tem distribuição com assimetria positiva.

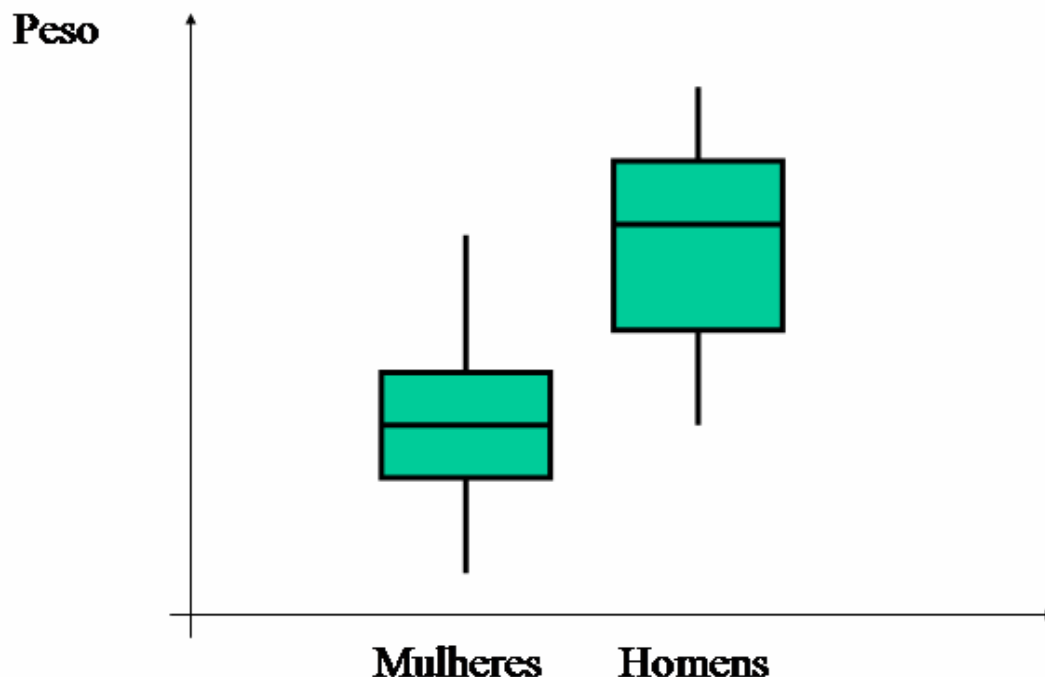
Os traços verticais que representam os valores máximo e mínimo também ajudam a avaliar a assimetria da distribuição pelo Box-plot. Se o valor do máximo está muito distante da caixa, então é provável que a distribuição tenha assimetria positiva. Se o valor do mínimo é que se afasta mais da caixa, então a distribuição tem assimetria negativa. Caso a distância do máximo e do mínimo até a caixa seja a mesma, então a distribuição é simétrica.



Assimetria	Valor de G	Box-plot	Exemplo de histograma
Nula	0 ou próximo de zero		
Positiva	$G > 0$		
Negativa	$G < 0$		

Outra aplicação muito importante do Box-plot é a comparação rápida entre duas amostras, ou de uma amostra com um ou mais valores individuais.

# Diferenças com Box-Plot



## 7 Relações entre variáveis

Nos capítulos anteriores foram apresentados métodos para analisar dados de uma variável pertencente a uma população. Outro tipo de análise importante na hidrologia é como uma variável se relaciona com as outras, da mesma população. Existem formas de medir o grau de associação entre variáveis e existem métodos para prever o valor de uma variável A, desde que se conheça o valor da variável B que mantém uma relação com a variável A.

Existem muitos exemplos de relações entre variáveis: a velocidade da água de um rio tem relação com a concentração de sedimentos; o nível da água de um rio tem relação com a vazão que está passando por ele; a altura das ondas em um lago tem relação com a velocidade do vento; a temperatura média do ar em Porto Alegre tem relação com o dia do ano; as notas dos alunos do CTH em Estatística tem relação com o número de horas por semana que eles dedicam ao estudo; e assim por diante.

## Análise gráfica de relações

Em alguns casos é útil elaborar o gráfico relacionando duas variáveis de um experimento para identificar possíveis relações entre estas variáveis.

Sabe-se que existe uma relação entre a temperatura do ar e a altitude de um determinado local. Podemos testar esta relação com os dados de várias estações meteorológicas do Rio Grande do Sul, coletados ao longo de 10 anos ou mais entre 1957 e 1977. A tabela abaixo apresenta os

dados de altitude e de temperatura de 22 estações meteorológicas no RS. São apresentados os dados de temperatura média anual e de temperatura média das máximas no mês de Dezembro.

Podemos explorar se existe uma relação entre os dados de altitude e temperatura elaborando um gráfico relacionando estas duas variáveis, com a altitude no eixo horizontal e a temperatura no eixo vertical. Com base neste gráfico observamos que existe uma tendência de que as temperaturas sejam mais baixas em locais mais altos.

Estação	Altitude (m)	Temperatura média anual (°C)	Temperatura média das máximas de Dezembro (°C)
Bajé	214	18.7	28.4
Encruzilhada do Sul	420	17.6	26.1
Erexim	760	18.7	27.3
Farroupilha	702	17.3	26.5
Guaíba	46	19.6	28.9
Ijuí	448	20.5	30
Jaguarão	11	18.3	28.2
Júlio de Castilho	514	18.6	28
Osório	32	19.8	27.8
Passo Fundo	709	18.4	28
Quaraí	100	19.5	30
Rio Grande	16	18.8	26.9
Santa Maria	153	19.7	29.3
Santana do Livramento	210	18.7	28.5
Santo Augusto	380	20.1	29.7
São Borja	99	21	30.4
São Gabriel	109	19.3	27
Taquari	76	20.2	29.4
Tramandaí	3	19.6	25.8
Uruguaiana	74	20.2	30.7
Vacaria	955	16.4	25.4
Veranópolis	705	17.5	26.1

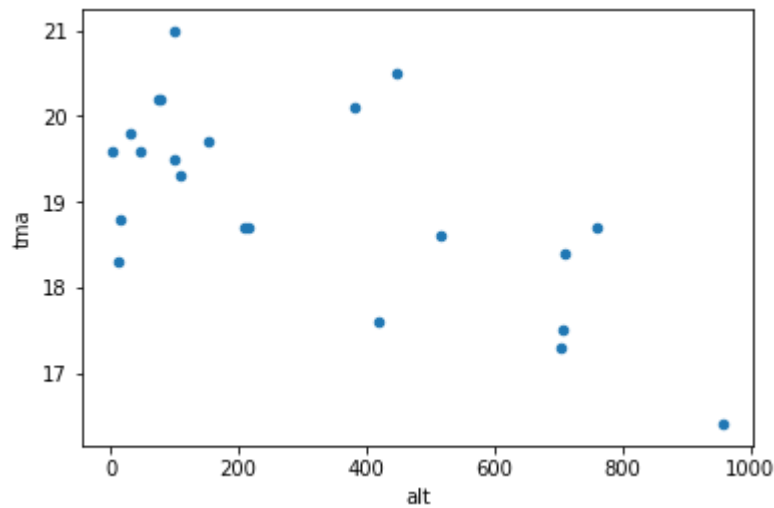
```
In [42]: dados = pd.read_excel('alt_temp.xlsx')
dados.columns = ['est', 'alt', 'tma', 'tmmd']
dados
```

Out[42]:

	est	alt	tma	tmmd
0	Bajé	214	18.7	28.4
1	Encruzilhada do Sul	420	17.6	26.1
2	Erexim	760	18.7	27.3
3	Farroupilha	702	17.3	26.5
4	Guaíba	46	19.6	28.9
5	Ijuí	448	20.5	30.0
6	Jaguarão	11	18.3	28.2
7	Júlio de Castilho	514	18.6	28.0
8	Osório	32	19.8	27.8
9	Passo Fundo	709	18.4	28.0
10	Quaraí	100	19.5	30.0
11	Rio Grande	16	18.8	26.9
12	Santa Maria	153	19.7	29.3
13	Santana do Livramento	210	18.7	28.5
14	Santo Augusto	380	20.1	29.7
15	São Borja	99	21.0	30.4
16	São Gabriel	109	19.3	27.0
17	Taquari	76	20.2	29.4
18	Tramandaí	3	19.6	25.8
19	Uruguaiana	74	20.2	30.7
20	Vacaria	955	16.4	25.4
21	Veranópolis	705	17.5	26.1

```
In [43]: dados.plot.scatter('alt', 'tma')
```

```
Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0xe37cb70>
```



## Análise de relações em tabelas

Também podemos explorar relações entre variáveis qualitativas e quantitativas através de uma tabela.

Considere um conjunto de medições de capacidade de infiltração de solos com o método dos anéis concêntricos. Foram medidos dados em diferentes tipos de solos, que foram classificados como Arenosos ou Argilosos. Uma tabela apresenta a frequência de dados com capacidade de infiltração em faixas: menor que 10; entre 10 e 20; e maior do que 20 mm/hora. A tabela mostra que foram medidos 82 locais, dos quais 39 em locais de solo arenoso e 43 em locais de solo argiloso. Em 36 locais a capacidade de infiltração medida foi inferior a 10 mm/hora, em 21 a capacidade ficou entre 10 e 20 e em 25 pontos a capacidade de infiltração foi superior a 20 mm/hora.

Dos 36 pontos de capacidade de infiltração baixa, 29 ocorreram em solos argilosos, indicando que existe uma relação entre a classe de solo e a capacidade de infiltração. Da mesma forma, dos 25 pontos com capacidade de infiltração alta (>20 mm/hora), apenas 2 ocorreram em solos argilosos e 23 em solos arenosos.

	Taxa	Solo	Contagem
1	<10	Areia	7
2	(10-20)	Areia	9
3	>20	Areia	23
1	<10	Argila	29
2	(10-20)	Argila	12
3	>20	Argila	2

```
In [44]: tabela = pd.read_excel('infiltracao.xlsx')
print tabela
print
cross = pd.crosstab(tabela.Taxa, tabela.Solo, values=tabela.Contagem, aggfunc=np
print cross
```

```

      Taxa  Solo  Contagem
0  1  <10  Areia         7
1  2  (10-20) Areia         9
2  3  >20  Areia        23
3  1  <10  Argila        29
4  2  (10-20) Argila       12
5  3  >20  Argila         2
```

```

Solo      Areia  Argila  All
Taxa
1  <10         7      29   36
2  (10-20)     9      12   21
3  >20        23       2   25
All          39      43   82
```

A mesma tabela pode ser elaborada com valores percentuais, deixando ainda mais clara a relação entre capacidade de infiltração e tipos de solos.

```
In [45]: nova = pd.DataFrame(cross.Areia/cross.All, columns=['Areia'])
nova['Argila'] = cross.Argila/cross.All
nova['All'] = cross.All/cross.All
print nova.round(3) * 100
```

```

      Areia  Argila  All
Taxa
1  <10     19.4    80.6  100.0
2  (10-20)  42.9    57.1  100.0
3  >20     92.0     8.0  100.0
All      47.6    52.4  100.0
```

## O coeficiente de correlação

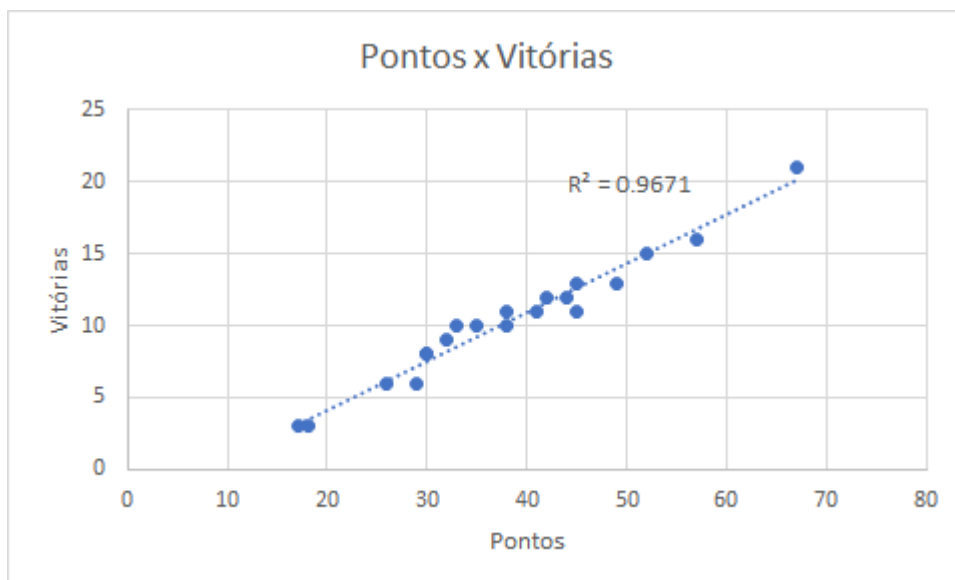
Podemos medir o grau ou a intensidade da relação entre duas variáveis utilizando a estatística. O coeficiente de correlação de Pearson ( $r$ ), definido nas equações abaixo, permite avaliar o quanto duas variáveis estão relacionadas. O valor de  $r$  pode estar entre -1 e 1, e a interpretação dos valores de  $r$  é dada na tabela que segue.

$$r = \frac{\left( \sum_{i=1}^n x_i \cdot y_i \right) - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left[ \sum_{j=1}^n x_j^2 - n \cdot \bar{x}^2 \right] \cdot \left[ \sum_{j=1}^n y_j^2 - n \cdot \bar{y}^2 \right]}}$$

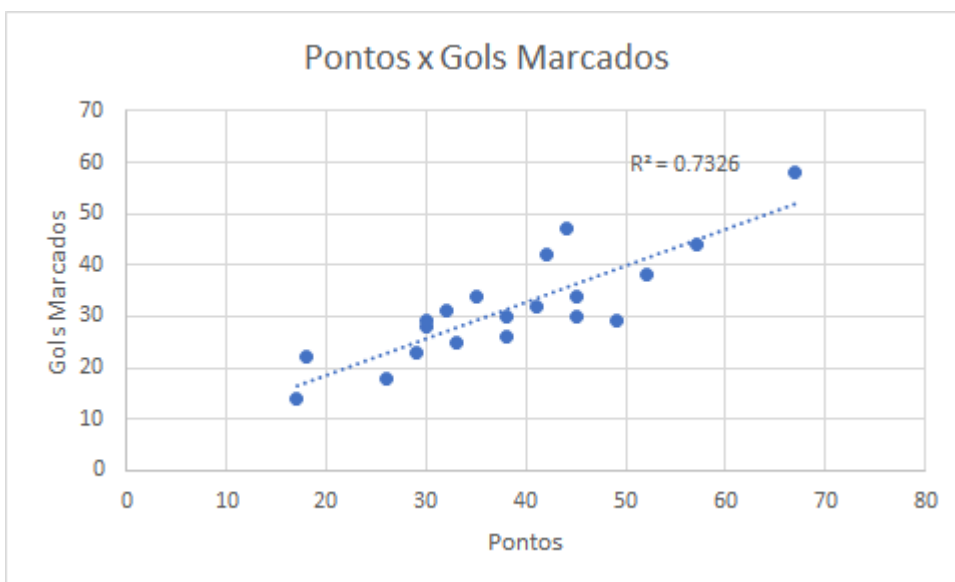
$$r = \frac{n \cdot \left( \sum_{i=1}^n x_i \cdot y_i \right) - \left( \sum_{i=1}^n x_i \right) \cdot \left( \sum_{i=1}^n y_i \right)}{\sqrt{n \cdot \left( \sum_{i=1}^n x_i^2 \right) - \left( \sum_{i=1}^n x_i \right)^2} \cdot \sqrt{n \cdot \left( \sum_{i=1}^n y_i^2 \right) - \left( \sum_{i=1}^n y_i \right)^2}}$$

Valor do coeficiente de correlação de Pearson (r)	Interpretação
+1	Existe uma relação linear perfeita e positiva entre as variáveis. À medida que x aumenta y também aumenta.
+0,8 até +1	Existe forte correlação positiva entre as variáveis. À medida que x aumenta y também aumenta.
+0,4 até +0,8	Existe uma correlação positiva moderada entre as variáveis. À medida que x aumenta y também aumenta.
-0,4 a +0,4	Existe pouca correlação linear entre os dados. Os dados podem não ter relação nenhuma ou pode ser necessário avaliar relações não lineares.
-0,4 até -0,8	Existe uma correlação negativa moderada entre as variáveis. À medida que x aumenta y diminui.
-0,8 até -1	Existe forte correlação negativa entre as variáveis. À medida que x aumenta y diminui.
-1	Existe uma relação linear perfeita e negativa entre as variáveis. À medida que x aumenta y diminui.

O gráfico abaixo apresenta a relação entre pontos ganhos e número de vitórias dos times da primeira divisão de futebol do Brasil na 28ª rodada do campeonato de 2019. O coeficiente de correlação é de 0,96, o que indica uma forte correlação entre pontos e número de vitórias.



O gráfico abaixo apresenta a relação entre pontos ganhos e número de gols marcados na mesma situação. O coeficiente de correlação é de 0,73, o que indica uma forte correlação entre pontos e número de vitórias, mas a relação é mais fraca do que no caso anterior.



## Regressão linear simples

Constatada uma correlação relativamente alta entre duas variáveis, pode ser interessante encontrar uma equação que representa adequadamente a relação entre estas variáveis.

A equação mais simples que pode ser explorada é a equação de uma linha reta. Tomando como exemplo a relação entre altitude e temperatura média analisada antes, podemos tentar traçar manualmente uma linha reta relacionando as duas variáveis diretamente sobre a figura. O problema é que duas pessoas diferentes vão obter retas diferentes, assim é necessário definir matematicamente a equação da reta que melhor representa os dados.

O formato básico de uma equação de reta, ou equação linear é:



$$y = a.x+b \text{ ou } y=b.x+a \text{ ou } y=m.x+n$$

Procurando a resposta para a pergunta - Qual é a linha reta que melhor representa os pontos? – chegou-se a conclusão que o ideal é escolher uma linha que minimiza os erros. Pode-se mostrar que é melhor trabalhar com erros médios quadrados, ao invés de erros simples ou dos módulos dos erros. Assim, a equação escolhida é a que minimiza o somatório dos erros quadrados.

Considerando que a equação da reta é  $y=a+b.x$ ; então o erro cometido ao utilizar esta equação para representar um ponto qualquer  $x_i, y_i$ , é dado por:

$$\text{erro}_i = (a+b.x_i) - y_i$$

e o somatório dos quadrados é:

$$SQ = \sum_{i=1}^n ((a + b \cdot x_i) - y_i)^2$$

Derivando esta equação com relação a 'a' e depois 'b' obtemos duas novas equações cujo valor deve ser zero (para entender por que, estude Cálculo I e Cálculo II). Estas duas novas equações representam um sistema de duas equações e duas incógnitas a e b. Resolvendo este sistema chegamos aos valores:  $y = a + b \cdot x$

$$b = \frac{n \cdot \left( \sum_{i=1}^n x_i \cdot y_i \right) - \left( \sum_{i=1}^n x_i \right) \cdot \left( \sum_{i=1}^n y_i \right)}{n \cdot \left( \sum_{i=1}^n x_i^2 \right) - \left( \sum_{i=1}^n x_i \right)^2}$$

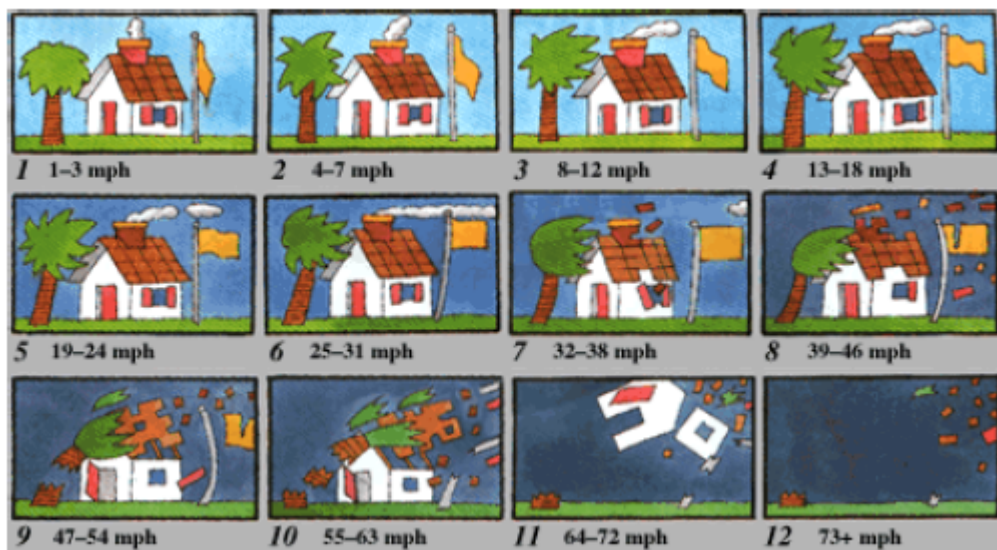
---


$$a = \bar{y} - b \cdot \bar{x}$$

Exemplo:

A escala de Beaufort classifica o vento em diferentes categorias e relaciona as velocidades do vento com os efeitos que causa no oceano e na terra. Foi criada pelo meteorologista britânico Francis Beaufort no início do século XIX.

Um dos aspectos interessantes da escala de Beaufort é a relação entre velocidade do vento e altura de ondas no mar aberto (Tabela: Velocidade do vento e altura das ondas na escala de Beaufort.). Tente desenvolver uma equação que relacione estas duas variáveis.



Vento (nós)	Altura de onda (m)
0.5	0
2	0.1
5	0.5
8.5	1
13.5	2
19	3
24.5	4
30.5	5
37	6
44	7
51.5	9

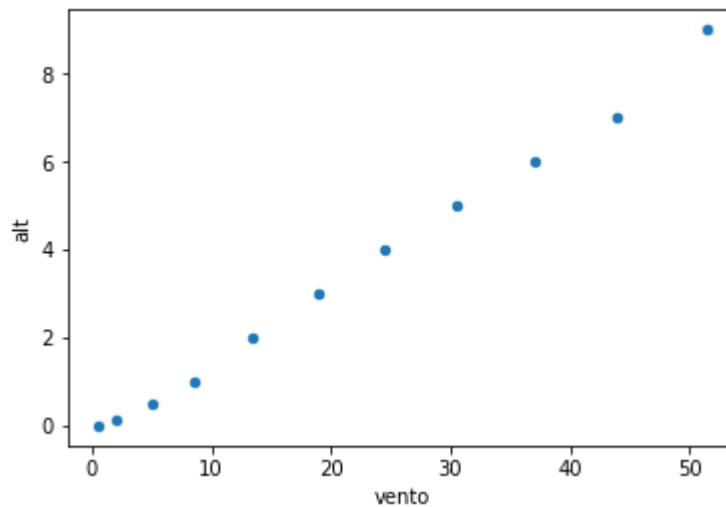
```
In [46]: dados = pd.read_excel('ventos_ondas.xlsx')
dados.columns = ['vento', 'alt']
dados
```

Out[46]:

	vento	alt
0	0.5	0.0
1	2.0	0.1
2	5.0	0.5
3	8.5	1.0
4	13.5	2.0
5	19.0	3.0
6	24.5	4.0
7	30.5	5.0
8	37.0	6.0
9	44.0	7.0
10	51.5	9.0

```
In [47]: dados.plot.scatter('vento', 'alt')
```

Out[47]: <matplotlib.axes.\_subplots.AxesSubplot at 0xe48f4d0>



```
In [48]: dados['xy']=dados['vento']*dados['alt']
dados['x2']=dados['vento']*dados['vento']
N = dados['vento'].count()
somatorio_x = dados['vento'].sum()
somatorio_y = dados['alt'].sum()
somatorio_xy = dados['xy'].sum()
somatorio_x2 = dados['x2'].sum()
print N, somatorio_x, somatorio_y, somatorio_xy, somatorio_x2
```

11 236.0 37.6 1339.2 8132.5

```
In [49]: dados.append(dados.sum().rename('Total'))
```

```
Out[49]:
```

	vento	alt	xy	x2
0	0.5	0.0	0.0	0.25
1	2.0	0.1	0.2	4.00
2	5.0	0.5	2.5	25.00
3	8.5	1.0	8.5	72.25
4	13.5	2.0	27.0	182.25
5	19.0	3.0	57.0	361.00
6	24.5	4.0	98.0	600.25
7	30.5	5.0	152.5	930.25
8	37.0	6.0	222.0	1369.00
9	44.0	7.0	308.0	1936.00
10	51.5	9.0	463.5	2652.25
<b>Total</b>	236.0	37.6	1339.2	8132.50

```
In [50]: b = ((N*somatorio_xy)-(somatorio_x*somatorio_y))/((N*somatorio_x2)-(somatorio_x**2))
print b
```

```
0.17349940020437482
```

```
In [51]: media_x = dados['vento'].mean()
media_y = dados['alt'].mean()
print media_x, media_y
```

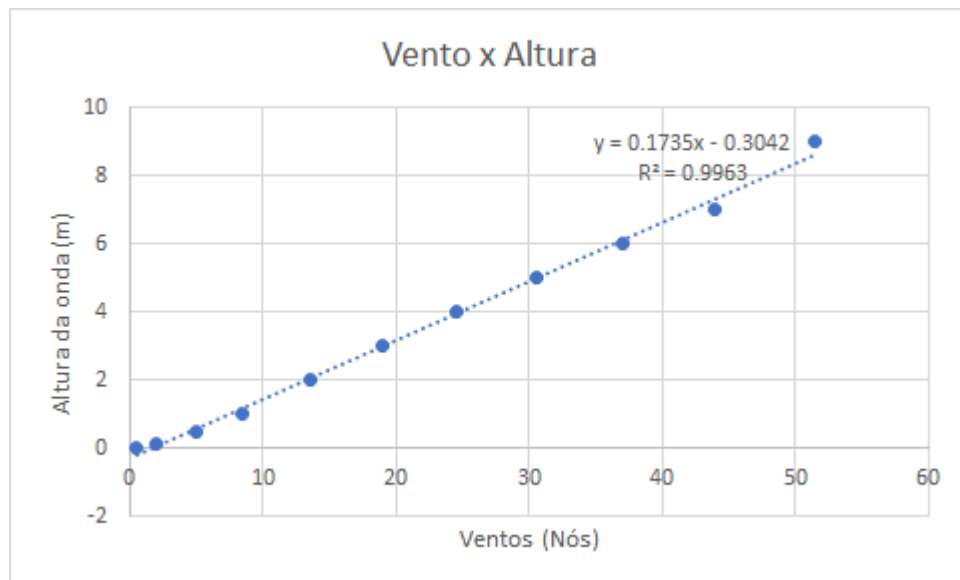
```
21.454545454545453 3.4181818181818184
```

```
In [52]: a = media_y - (b * media_x)
print a
```

```
-0.30416894983931364
```

```
In [53]: print ('A reta ajustada é Y = {1:.2f} * X {0:.2f}'.format(a, b))
```

```
A reta ajustada é Y = 0.17 * X -0.30
```



```
In [54]: import statsmodels.api as sm
x = dados.vento
X = sm.add_constant(x)
y = dados.alt
model = sm.OLS(y, X)
results = model.fit()
#print(results.summary())
```

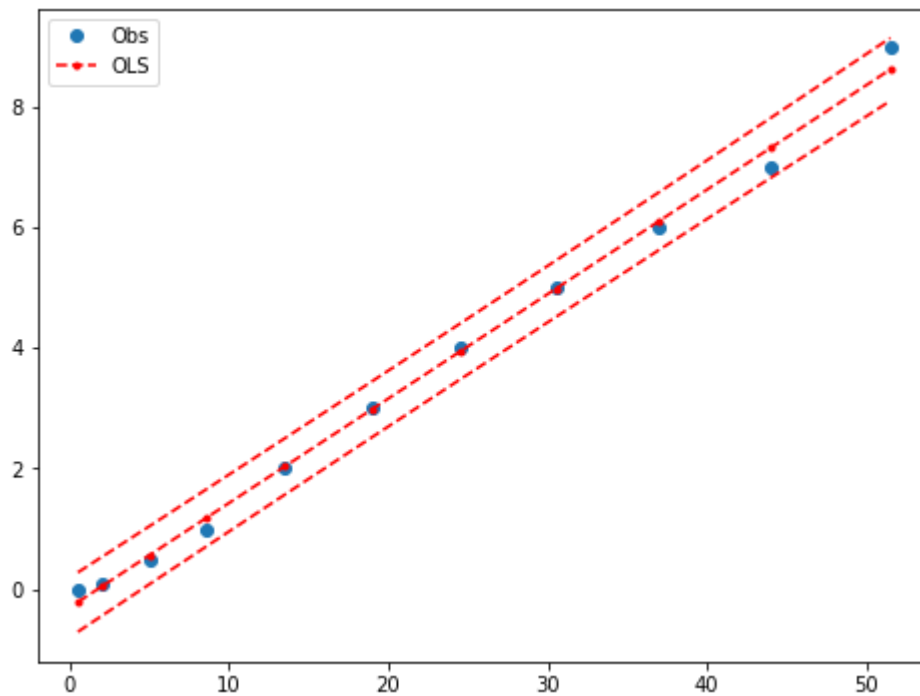
c:\python27\arcgis10.6\lib\site-packages\numpy\core\fromnumeric.py:2389: Future Warning: Method .ptp is deprecated and will be removed in a future version. Use numpy.ptp instead.

```
return ptp(axis=axis, out=out, **kwargs)
```

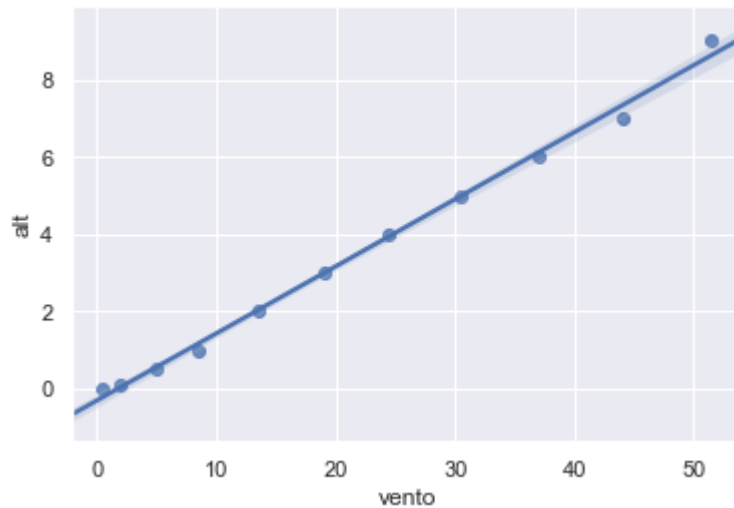
```
In [55]: print(r'Parametros: ', results.params)
print('R2: ', results.rsquared)
```

```
('Parametros: ', const    -0.304169
vento      0.173499
dtype: float64)
('R2: ', 0.9962651569818115)
```

```
In [56]: from statsmodels.sandbox.regression.predstd import wls_prediction_std  
prstd, iv_l, iv_u = wls_prediction_std(results)  
  
fig, ax2 = plt.subplots(figsize=(8,6))  
  
ax2.plot(x, y, 'o', label="Obs")  
ax2.plot(x, results.fittedvalues, 'r--.', label="OLS")  
ax2.plot(x, iv_u, 'r--')  
ax2.plot(x, iv_l, 'r--')  
ax2.legend(loc='best');
```



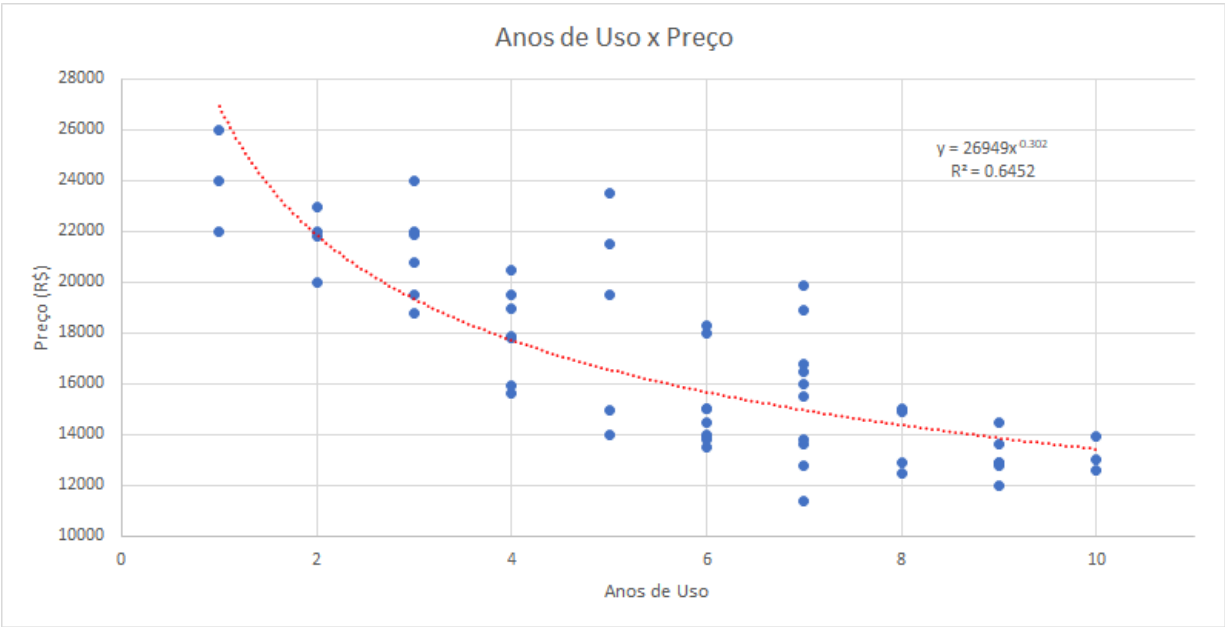
```
In [57]: import seaborn as sns; sns.set(color_codes=True)
ax = sns.regplot(y=dados.alt, x=dados.vento)
```



## Regressão não lineares

Algumas variáveis podem apresentar um alto grau de relação, facilmente visível num gráfico, porém eventualmente esta relação pode ser não linear. A tabela a seguir apresenta preços de automóveis usados do modelo Fiat Palio. Considerando que estes dados correspondem ao ano de 2007, podemos contar os anos de uso desde a data de fabricação até 2007 e elaborar um gráfico.

Ano	Preço	Ano	Preço
2005	R\$ 19990.00	2000	R\$ 11400.00
2003	R\$ 18990.00	1998	R\$ 13600.00
2006	R\$ 23990.00	2004	R\$ 21900.00
2005	R\$ 22990.00	2001	R\$ 14500.00
2005	R\$ 21990.00	2001	R\$ 13900.00
2004	R\$ 23990.00	2004	R\$ 18800.00
2006	R\$ 26000.00	2000	R\$ 15500.00
2004	R\$ 20800.00	1999	R\$ 14900.00
2006	R\$ 21990.00	1997	R\$ 12600.00
2002	R\$ 23500.00	2002	R\$ 19500.00
1998	R\$ 14500.00	2000	R\$ 16500.00
2003	R\$ 17800.00	2001	R\$ 13800.00
1998	R\$ 12800.00	1998	R\$ 12900.00
1999	R\$ 15000.00	2003	R\$ 15600.00
2002	R\$ 21500.00	1998	R\$ 12000.00
1999	R\$ 12900.00	1997	R\$ 13900.00
2004	R\$ 21990.00	2001	R\$ 13500.00
2001	R\$ 18000.00	2002	R\$ 14950.00
2001	R\$ 15000.00	1999	R\$ 12500.00
2001	R\$ 15000.00	2003	R\$ 20500.00
2000	R\$ 16800.00	2000	R\$ 12800.00
2001	R\$ 14000.00	1997	R\$ 13000.00
2001	R\$ 18300.00	2003	R\$ 17900.00
2000	R\$ 13600.00	2003	R\$ 15900.00
2000	R\$ 18900.00	2000	R\$ 19900.00
2005	R\$ 21800.00	2003	R\$ 19500.00
2000	R\$ 16000.00	1998	R\$ 12900.00
2000	R\$ 13800.00	2004	R\$ 19500.00
2002	R\$ 14000.00		





Observa-se que a relação é negativa, isto é, quanto maior o tempo de uso, menor o preço do carro.

Entretanto, observa-se também que carros muito velhos aproximam-se de um patamar entre 12 e 15 mil reais a partir do oitavo ano. Neste caso, um outro tipo de equação deveria ser avaliado para estimar a relação entre tempo de uso e preço.

O Excel pode ser utilizado para testar várias equações de regressão a um conjunto de dados. Isto pode ser feito usando a ferramenta Solver ou Adicionar Linha de Tendência sobre um gráfico. Alguns critérios devem ser lembrados, entretanto:

- É melhor evitar polinômios de grau alto, mesmo que o  $R^2$  seja mais alto.
- Nunca usar polinômio com grau igual ou superior ao número de pontos.
- $R^2$  alto nem sempre significa bom ajuste.
- É desejável fazer um gráfico de resíduos da regressão.

## 8 Probabilidade e Distribuições de probabilidade

Na tomada de decisões na vida profissional, e até mesmo na vida pessoal, estamos constantemente sendo obrigados a fazer escolhas sem ter certeza sobre diversos aspectos e variáveis que gostaríamos de saber. Muitas vezes estamos considerando implicitamente a probabilidade de termos sucesso ou insucesso. O exemplo mais evidente e mais antigo são os jogos de azar, como o jogo de dados ou o lançamento de uma moeda.

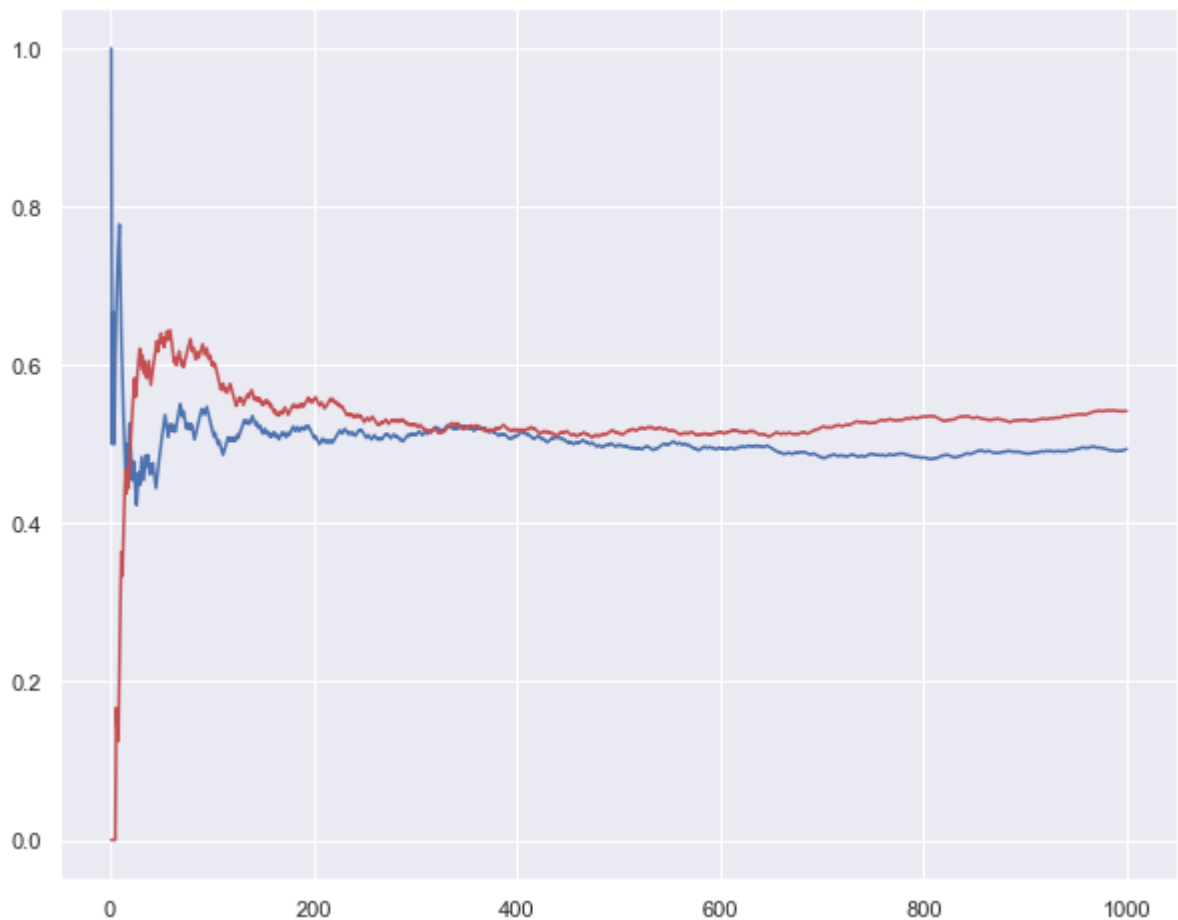
### A idéia da probabilidade

O comportamento do acaso é imprevisível a curto prazo mas tem um padrão regular e relativamente previsível a longo prazo.

Quando você lança uma moeda, há apenas dois resultados possíveis: cara ou coroa. O gráfico abaixo mostra os resultados de lançar uma moeda em duas seqüências de 1000 vezes cada uma. Para cada número de lançamentos o gráfico apresenta a proporção dos lançamentos que resultaram em uma cara.

```
In [58]: import random
lance = []
prob1, prob2 = [],[]
cara1, cara2 = 0, 0
for i in range(0, 1000, 1):
    lance.append(i + 1)
    teste1 = random.getrandbits(1)
    if teste1 == 1:
        cara1 = cara1+1
    prob1.append(cara1/lance[i])
    teste2 = random.getrandbits(1)
    if teste2 == 1:
        cara2 = cara2+1
    prob2.append(cara2/lance[i])
fig, ax = plt.subplots(figsize=(10,8))
plt.plot(lance, prob1)
plt.plot(lance, prob2, color = 'r')
```

Out[58]: [<matplotlib.lines.Line2D at 0xf9795b0>]



Nas duas seqüências a proporção varia entre 0,4 e 0,6 por algum tempo e lentamente vai se aproximando de 0,5, a medida que aumenta o número de lançamentos. Isto significa que a proporção lançamentos que resultam em caras é muito variável no começo. A medida que são feitos mais e mais lançamentos, contudo, a proporção de caras aproxima-se de 0,5 e lá permanece.

Se uma terceira seqüência for testada, a proporção de caras, a longo prazo, novamente se estabilizará em 0,5. Dizemos que 0,5 é a probabilidade de ocorrer uma cara.

Aleatório em Estatística não é sinônimo de “descontrolado”, mas uma descrição de um tipo de ordem que emerge apenas em longo prazo. Poderíamos suspeitar que uma moeda tem probabilidade 0,5 de aparecer cara apenas porque a moeda tem dois lados, o que significaria que há metade de probabilidade de ocorrência de cada um dos lados. Estas suspeitas não são sempre corretas. No lançamento de um percevejo, por exemplo, há dois resultados possíveis: cair com a ponta para cima, ou cair deitado. Entretanto a probabilidade não é igual para os dois resultados. O percevejo cai deitado com uma probabilidade maior do que com a ponta para cima.

É importante destacar que os lançamentos de um dado ou de uma moeda são eventos independentes, isto é, o resultado de um lançamento não interfere nas probabilidades do próximo. Algumas pessoas que gostam de jogos de azar tendem a esquecer este detalhe, acreditando em “ondas de sorte” ou “marés de azar”.

## A distribuição normal

Muitos fenômenos aleatórios na natureza seguem a distribuição de probabilidade conhecida como distribuição normal, ou gaussiana. A distribuição normal é descrita em qualquer livro introdutório de estatística e se aplica a muitos tipos de informações da natureza. Um gráfico da função densidade de probabilidade da distribuição normal tem uma forma de sino e é simétrico com relação à média, que é o valor central. A forma em sino indica que existe uma probabilidade maior de ocorrerem valores próximos à média do que nos extremos mínimo e máximo.

A função densidade de probabilidade (PDF) da distribuição normal é uma expressão que depende de dois parâmetros: a média e o desvio padrão da população, conforme a equação seguinte:

$$f_x(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_x}} \cdot \exp \left[ -\frac{1}{2} \cdot \left( \frac{x - \mu_x}{\sigma_x} \right)^2 \right]$$

onde  $\mu_x$  é a média da população e  $\sigma_x$  é o desvio padrão da população. Para o caso mais simples, em que a média da população é zero e o desvio padrão igual a 1, a expressão acima fica simplificada:

$$f_z(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left[-\frac{z^2}{2}\right]$$

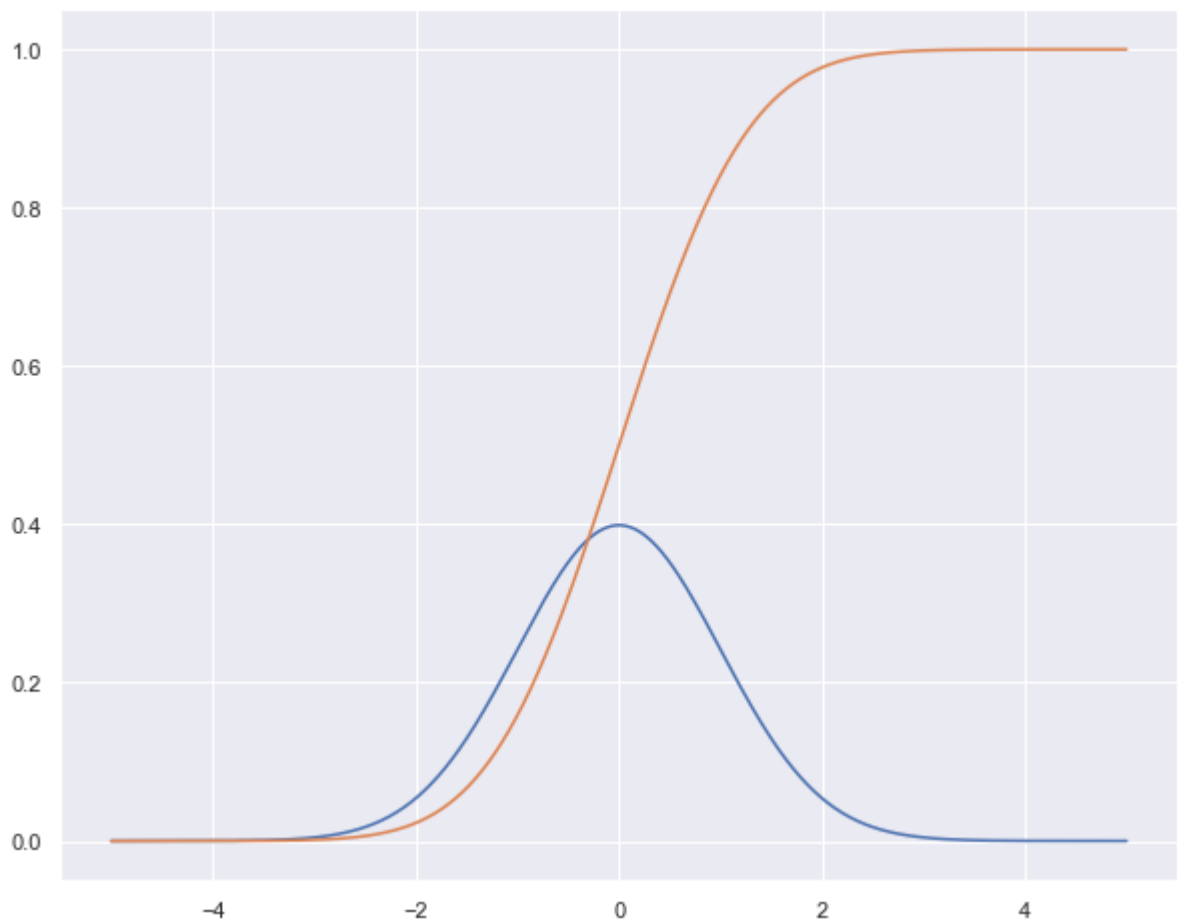
onde  $z$  é uma variável aleatória com média zero e desvio padrão igual a 1.

O gráfico desta última é apresentado abaixo. A área total sob a curva é igual a 1.

```
In [59]: import scipy.stats as ss

def plot_normal(x_range, media=0, varianca=1, cdf=False, **kwargs):
    x = x_range
    desvio = np.sqrt(varianca)
    if cdf:
        y = ss.norm.cdf(x, media, desvio)
    else:
        y = ss.norm.pdf(x, media, desvio)
    plt.plot(x, y, **kwargs)

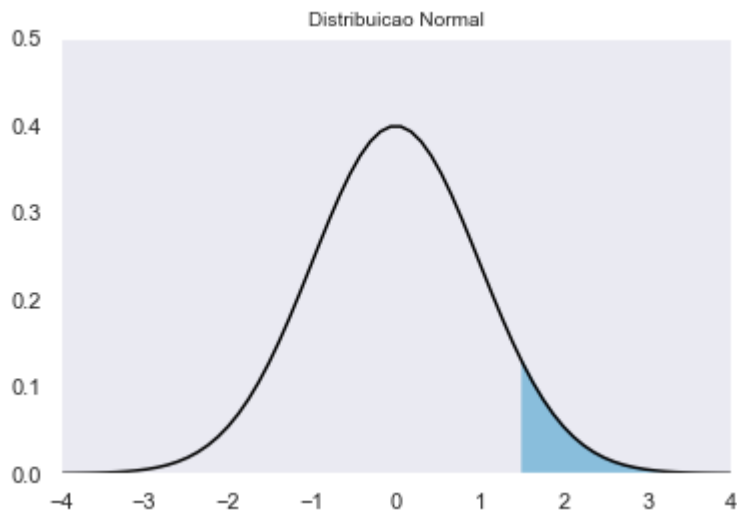
fig, ax = plt.subplots(figsize=(10,8))
x = np.linspace(-5, 5, 5000)
plot_normal(x)
plot_normal(x, cdf=True)
```

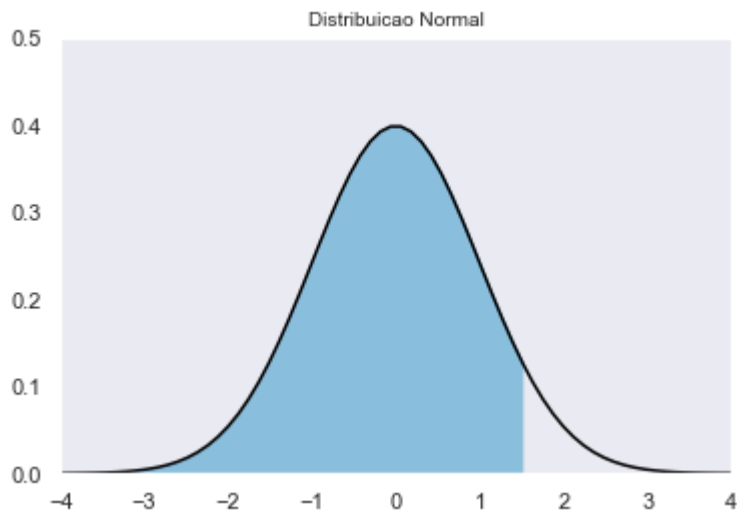


A área hachurada representa a probabilidade de ocorrência de um valor maior do que  $z$  (figura de cima) ou menor do que  $z$  (figura de baixo). A área sob a curva pode ser calculada por integração analítica, mas resulta numa série infinita. Por este motivo, as aplicações práticas são mais comuns na forma de tabelas que relacionam o valor de  $z$  com a probabilidade de ocorrer um valor maior do que  $z$  ou menor do que  $z$ . Existem, também, tabelas que fornecem valores da área entre 0 e  $z$ , ou de  $-z$  a  $z$ .

No programa Excel é possível obter os valores das probabilidades utilizando a função `DIST.NORMP(z)`, que dá a probabilidade de ocorrer um valor inferior a  $z$ .

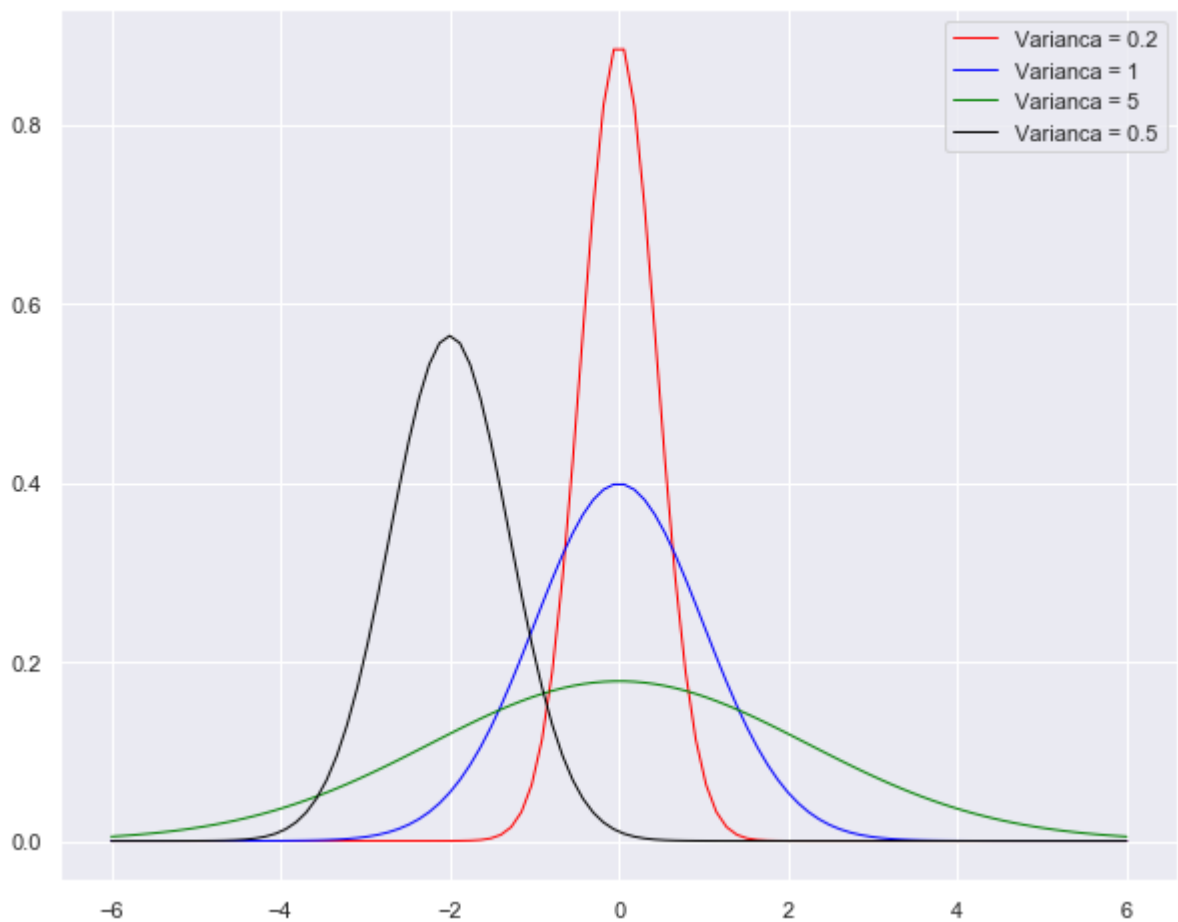
```
In [60]: x_min = -6.0
x_max = 6.0
mean = 0.0
std = 1.0
x = np.linspace(x_min, x_max, 100)
y = ss.norm.pdf(x,mean,std)
plt.plot(x,y, color='black')
pt1 = 1.5
pt2 = 6
ptx = np.linspace(pt1, pt2, 100)
pty = ss.norm.pdf(ptx,mean,std)
plt.fill_between(ptx, pty, color='#89bedc', alpha='1.0')
plt.grid()
plt.xlim(-4,4)
plt.ylim(0,0.5)
plt.title('Distribuicao Normal',fontsize=10)
plt.show()
plt.plot(x,y, color='black')
pt1 = -6
pt2 = 1.5
ptx = np.linspace(pt1, pt2, 100)
pty = ss.norm.pdf(ptx,mean,std)
plt.fill_between(ptx, pty, color='#89bedc', alpha='1.0')
plt.grid()
plt.xlim(-4,4)
plt.ylim(0,0.5)
plt.title('Distribuicao Normal',fontsize=10)
plt.show()
```





```
In [61]: fig, ax = plt.subplots(figsize=(10,8))
plot_normal(x, 0, 0.2, color='red', lw=1, ls='-', label='Varianca = 0.2')
plot_normal(x, 0, 1, color='blue', lw=1, ls='-', label='Varianca = 1')
plot_normal(x, 0, 5, color='green', lw=1, ls='-', label='Varianca = 5');
plot_normal(x, -2, 0.5, color='black', lw=1, ls='-', label='Varianca = 0.5');
plt.legend()
```

Out[61]: <matplotlib.legend.Legend at 0xf971150>

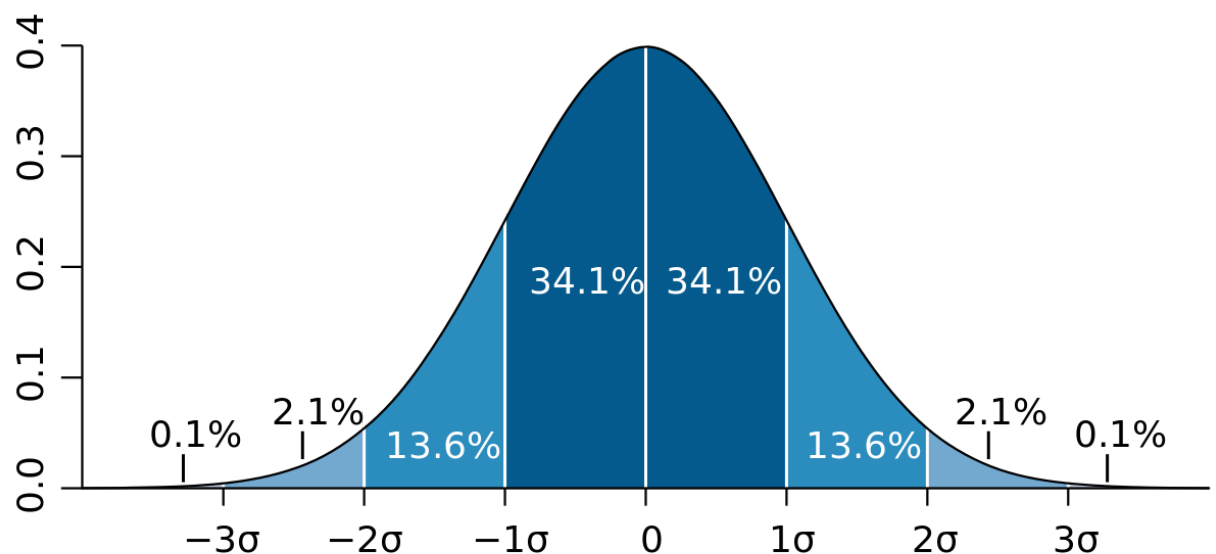


**A regra 68-95-99,7**

Algumas propriedades da distribuição normal são muitas vezes utilizadas na análise de dados estatísticos. É importante conhecer, por exemplo, a probabilidade de uma observação ocorrer em determinado intervalo de valores. Assim, pode-se dizer que, para uma distribuição normal com média  $\mu$  e com desvio padrão  $\sigma$ :

- 68% das observações estão a menos de  $\sigma$  da média  $\mu$ ;
- 95% das observações estão a menos de  $2\sigma$  da média  $\mu$ ;
- 99,7% das observações estão a menos de  $3\sigma$  da média  $\mu$ .

Lembrando destes valores é possível visualizar distribuições normais sem ter de constantemente fazer cálculos ou consultar tabelas.



Por exemplo, se em um vestibular os escores dos alunos se distribuem de forma normal (seguem a distribuição normal) com média 600 e desvio padrão igual a 100, podemos inferir que aproximadamente 68% dos alunos tem escores entre 500 e 700, e que cerca de 99,7% dos alunos tem escores entre 300 e 900.

## A distribuição normal padrão

Como sugere a regra 68-95-99,7, todas as distribuições normais compartilham algumas propriedades. Na verdade, todas as distribuições normais são idênticas, se medirmos em unidades de tamanho  $\sigma$  em torno da média  $\mu$  como centro. A mudança para estas unidades é chamada padronização. Para padronizar o valor de uma observação, subtrai-se a média e divide-se o resultado pelo desvio padrão, como expresso na equação que segue:

$$z = \frac{x - \mu}{\sigma}$$



onde  $x$  é o valor da observação oriunda de uma distribuição com média  $\mu$  e desvio padrão  $\sigma$ ; e  $z$  é o valor padronizado da observação. Um valor padronizado  $z$  nos diz de quantos desvios padrão a observação original está afastada da média, e em que direção. As observações maiores do que a média tem  $z$  positivo enquanto as observações menores do que a média tem  $z$  negativo.

A padronização de uma variável que tem uma distribuição normal qualquer gera uma nova variável que tem uma distribuição normal padrão. A distribuição normal padrão tem média igual a zero e desvio padrão igual a 1.

## Cálculos com a distribuição normal

Uma área sob uma curva de densidade de probabilidade é uma proporção das observações em um determinado intervalo da distribuição. Podemos responder qualquer pergunta acerca de qual proporção de observações está em uma determinada faixa de valores, determinando a área sob a curva. Como todas as distribuições normais são iguais quando padronizadas, podemos determinar áreas sob qualquer curva normal utilizando uma única tabela que forneça as áreas sob a curva para a distribuição normal padrão. No final do capítulo está apresentada uma tabela deste tipo.

Exemplo: As alturas das mulheres adultas seguem uma distribuição normal com média 164,2 e desvio padrão 6,8 cm.

Qual é a proporção de todas as mulheres que tem altura superior a 180 cm?

$$Z = (180 - 164,2) / 6,8 = 2,32$$

Utilizando as tabelas da distribuição normal também podemos encontrar qual é o valor da variável que cumpra um certo requisito de probabilidade. Por exemplo, no exemplo anterior poderia ser encontrado a altura para a qual 10 % das mulheres são mais altas.

```
In [62]: z = (180-164.2)/6.8
print '%5.2f' %(z)
prob = (1-ss.norm.cdf(z,0,1))*100
print '%5.2f' %(prob)
```

```
2.32
1.01
```

```
In [63]: prob1 = (1-ss.norm.cdf(180,164.2,6.8))*100
print '%5.2f' %(prob1)
```

```
1.01
```

```
In [64]: altu = ss.norm.ppf(0.9, 164.2, 6.8)
print '%5.2f' %(altu)
```

```
172.91
```

```
In [65]: prob2 = (1-ss.norm.cdf(altu,164.2,6.8))*100  
print '%5.2f' %(prob2)
```

10.00

```
In [66]: altu = ss.norm.ppf((100-prob1)/100, 164.2, 6.8)  
print '%5.2f' %(altu)
```

180.00

## Exercícios

1) O nível de colesterol no sangue é importante, pois níveis elevados de colesterol podem aumentar o risco de doença do coração. A distribuição de níveis de colesterol no sangue em uma grande população de pessoas da mesma idade e do mesmo sexo é aproximadamente Normal. Para meninos de 14 anos de idade, a média é de 170 miligramas de colesterol por decilitro de sangue (mg/dl) e o desvio padrão é de 30 mg/dl. Os níveis de colesterol acima de 240 mg/dl podem exigir cuidados médicos. Que percentual de meninos com 14 anos tem mais que 240 mg/dl de colesterol?

```
In [67]: prob3 = (1-ss.norm.cdf(240,170,30))*100  
print '%5.2f' %(prob3) + '%'
```

0.98%

2) Considerando que a chuva média anual em Caxias é 1600 mm por ano, e que o desvio padrão é de 250 mm, qual é a probabilidade de que no ano que vem a chuva anual seja superior a 2500 mm?

```
In [68]: prob4 = (1-ss.norm.cdf(2500,1600,250))*100  
print '%5.2f' %(prob4) + '%'
```

0.02%

3) Considere que os salários dos trabalhadores do Unguistão tem uma distribuição normal com média 1500 dólares e desvio padrão de 300 dólares, e que no país vizinho, o Danistão, os salários também tem distribuição normal, mas com média 2000 dólares e desvio padrão de 350 dólares. O sr. Al Amunamed mora no Unguistão e recebe 1800 dólares e o sr. Mustafah Umarah mora no Danistão e recebe 2100 dólares. Comparando com os seus vizinhos do mesmo país, qual dos dois é o que recebe mais? Utilize a padronização dos valores dos salários.

```
In [69]: prob_Amunamed = (ss.norm.cdf(1800,1500,300))*100  
print '%5.2f' %(prob_Amunamed) + '%'  
prob_Mustafah = (ss.norm.cdf(2100,2000,350))*100  
print '%5.2f' %(prob_Mustafah) + '%'
```

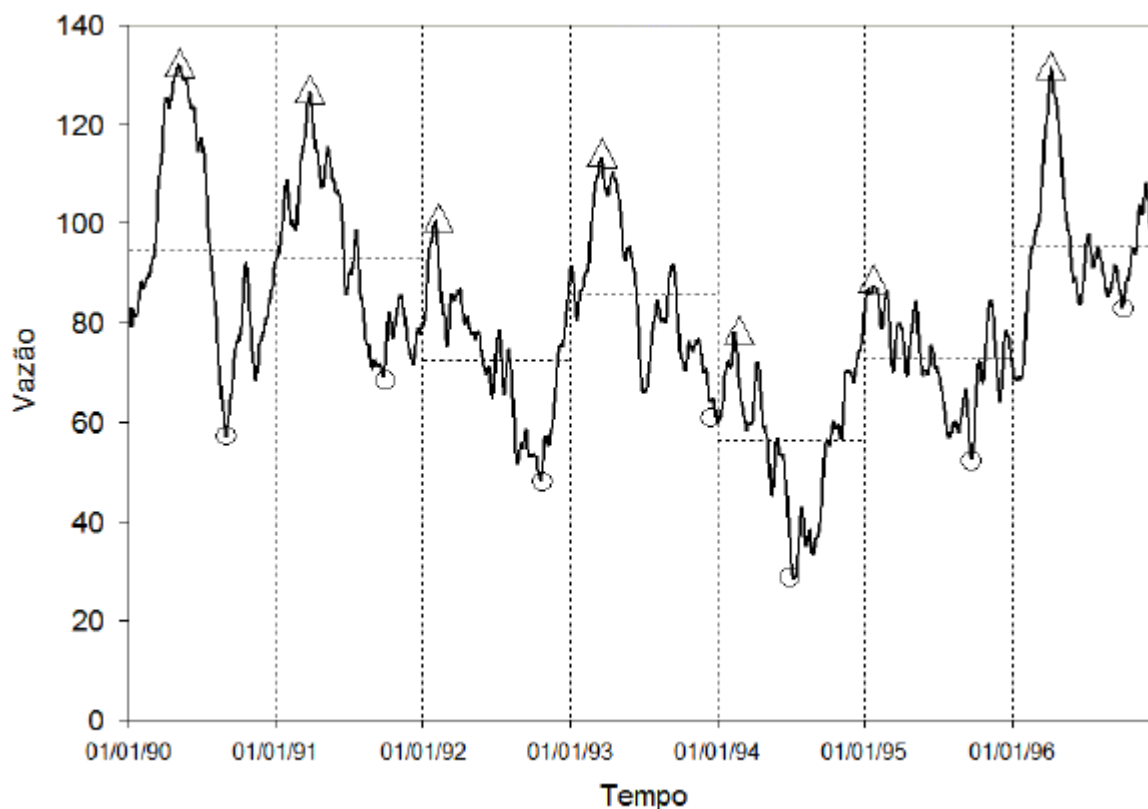
84.13%

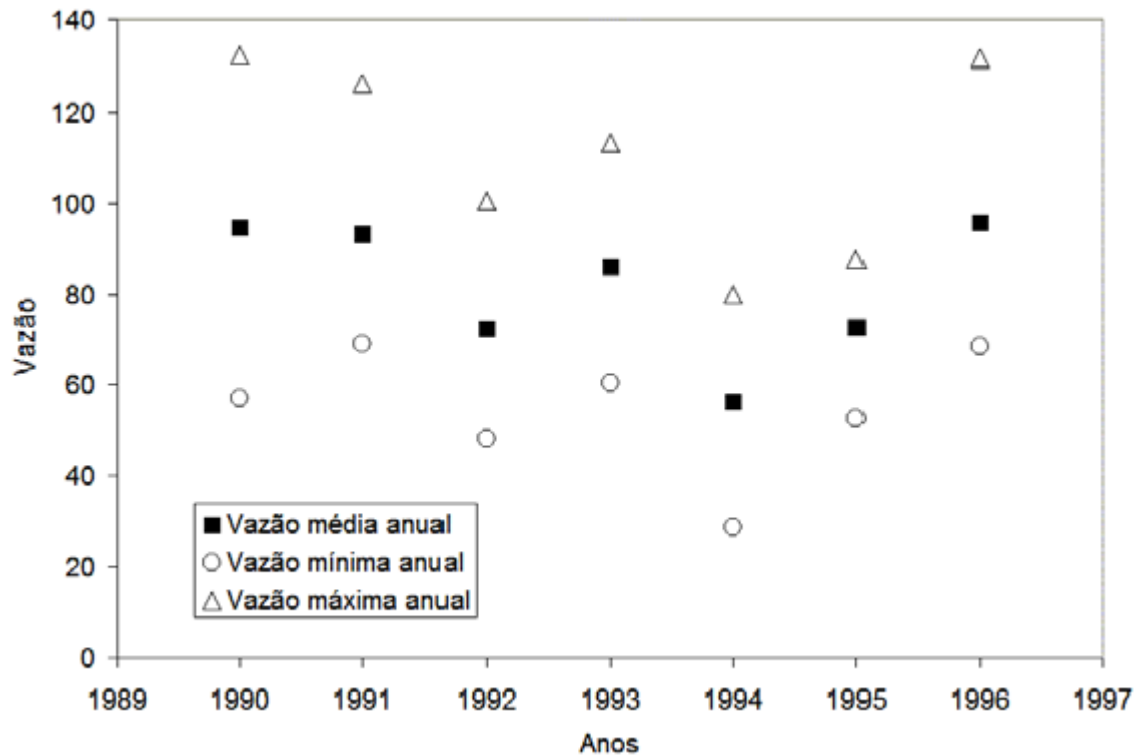
61.25%

## Vazões mínimas e máximas

A vazão de um rio é uma variável que se modifica de forma contínua no tempo, e pode ser representada em um hidrograma, que é o gráfico que relaciona os valores de vazão com o tempo, como na figura abaixo.

Diversas análises estatísticas de dados hidrológicos são realizadas de forma mais conveniente sobre valores discretos no tempo, ao contrário das seqüências contínuas. A partir de uma seqüência contínua de vazões é possível identificar séries temporais de valores discretos, como, por exemplo, as vazões médias anuais, as vazões máximas anuais e as vazões mínimas anuais, conforme representado na figura. As séries discretas que são obtidas a partir da observação de alguns anos de dados de vazão são tratadas como amostras do comportamento de um rio ou de uma bacia. A população, neste caso, seriam todos os anos de existência de um rio. A vazão é considerada uma variável aleatória porque depende de fenômenos climáticos complexos e de difícil previsibilidade a partir de um certo horizonte.

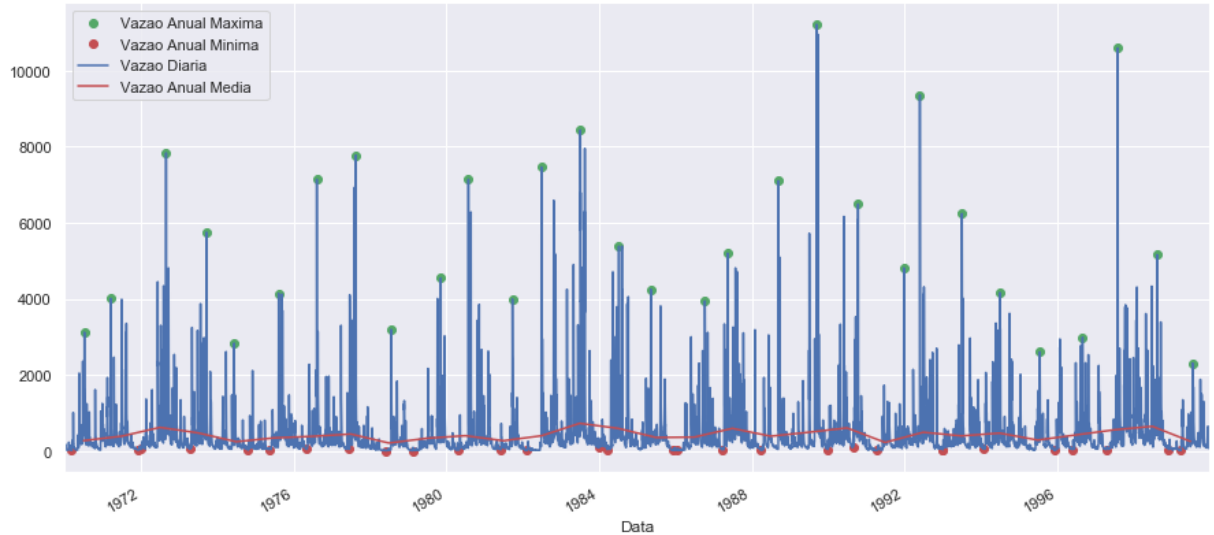




Vamos fazer conforme o exemplo da figura para a série de vazoes do Rio Taquari, na estação 86510000, para o período de 1970-1999

```
In [70]: vazoesD_taquari = pd.read_excel('86510000_vazaoD.xlsx', index_col='Data')
VazaoDT = vazoesD_taquari['1-1-1970':'12-31-1999'].copy()
VazaoAMed_taq = VazaoDT.groupby(VazaoDT.index.year).mean()
VazaoAMax_taq_valor = VazaoDT.groupby(VazaoDT.index.year).max()
VazaoAMax_taq_data = VazaoDT.groupby(VazaoDT.index.year).idxmax()
VazaoAMin_taq_valor = VazaoDT.groupby(VazaoDT.index.year).min()
VazaoAMin_taq_data = VazaoDT.groupby(VazaoDT.index.year).idxmin()
```

```
In [71]: data = pd.date_range('1/1/1970', periods = 30, freq = 'A-JUN')
fig, ax = plt.subplots(figsize=(15,7))
VazaoAMed_taq.set_index(data, inplace=True)
plt.plot_date(x=VazaoAMax_taq_data.VazaoD, y=VazaoAMax_taq_valor.VazaoD, marker=
plt.plot_date(x=VazaoAMin_taq_data.VazaoD, y=VazaoAMin_taq_valor.VazaoD, marker=
VazaoDT.plot(ax=ax, label='Vazao Diaria')
plt.plot(VazaoAMed_taq.VazaoD, color = 'r')
ax.legend(["Vazao Anual Maxima", "Vazao Anual Minima", "Vazao Diaria", "Vazao An
```



## Risco, probabilidade e tempo de retorno

Séries temporais discretas são convenientes para avaliar riscos em hidrologia. Risco é muitas vezes entendido como um sinônimo de probabilidade, mas em hidrologia é mais adequado considerar o risco como a probabilidade de ocorrência de um evento multiplicada pelos prejuízos que se espera da ocorrência deste evento.

Projetos de estruturas hidráulicas sempre são elaborados admitindo probabilidades de falha. Por exemplo, as pontes de uma estrada são projetadas com uma altura tal que a probabilidade de ocorrência de uma cheia que atinja a ponte seja de apenas 1% num ano qualquer. Isto ocorre porque é muito caro dimensionar as pontes para a maior vazão possível, por isso admite-se uma probabilidade, ou risco, de que a estrutura falhe. Isto significa que podem ocorrer vazões maiores do que a vazão adotada no dimensionamento. A probabilidade admitida pode ser maior ou menor, dependendo do tipo de estrutura. A probabilidade admitida para a falha de uma estrutura hidráulica é menor se a falha desta estrutura provocar grandes prejuízos econômicos ou mortes de pessoas. Assim, a probabilidade de falha admitida para um dique de proteção de uma cidade é a probabilidade de que ocorra uma cheia em que o nível da água supere o nível de proteção do dique. Diques que protegem grandes cidades deveriam ser construídos admitindo uma probabilidade menor de falha do que diques de proteção de pequenas áreas agrícolas. A tabela abaixo apresenta o tempo de retorno em anos adotado, normalmente, para diferentes tipos de estrutura.

Estrutura	TR (anos)
Bueiros de estradas pouco movimentadas	5 a 10

<b>Estrutura</b>	<b>TR (anos)</b>
Bueiros de estradas muito movimentadas	50 a 100
Pontes	50 a 100
Diques de proteção de cidades	50 a 200
Drenagem pluvial	2 a 10
Grandes barragens (vertedor)	10000
Pequenas barragens	100

O risco também pode estar relacionado a situações de vazões mínimas. Por exemplo, considere uma cidade que utilize a água de um rio para abastecimento da população. Dependendo do tamanho da população e das características do rio, existe um sério risco de que, num ano qualquer, ocorram alguns dias em que a vazão do rio é inferior à vazão necessária para abastecer a população.

No caso da análise de vazões máximas, são úteis os conceitos de probabilidade de excedência e de tempo de retorno de uma dada vazão. A probabilidade anual de excedência de uma determinada vazão é a probabilidade que esta vazão venha a ser igualada ou superada num ano qualquer. O tempo de retorno desta vazão é o intervalo médio de tempo, em anos, que decorre entre duas ocorrências subseqüentes de uma vazão maior ou igual. O tempo de retorno é o inverso da probabilidade de excedência como expresso na seguinte equação:

$$TR = \frac{1}{P}$$

onde TR é o tempo de retorno em anos e P é a probabilidade de ocorrer um evento igual ou superior em um ano qualquer. No caso de vazões mínimas, P refere-se à probabilidade de ocorrer um evento com vazão igual ou inferior.

A equação acima indica que a probabilidade de ocorrência de uma cheia de 10 anos de tempo de retorno, ou mais, num ano qualquer é de 0,1 (ou 10%).

A vazão máxima de 10 anos de tempo de retorno (TR = 10 anos) é excedida em média 1 vez a cada dez anos. Isto não significa que 2 cheias de TR = 10 anos não possam ocorrer em 2 anos seguidos. Também não significa que não possam ocorrer 20 anos seguidos sem vazões iguais ou maiores do que a cheia de TR=10 anos.

Existem duas formas de atribuir probabilidades e tempos de retorno às vazões máximas e mínimas: métodos empíricos e métodos analíticos.

Probabilidades empíricas podem ser estimadas a partir da observação das variáveis aleatórias. Por exemplo, a probabilidade de que uma moeda caia com a face “cara” virada para cima é de 50%. Esta probabilidade pode ser estimada empiricamente lançando a moeda 100 vezes e contando quantas vezes cada uma das faces fica voltada para cima.

O problema das probabilidades empíricas é que quando o tamanho da amostra é pequeno, a estimativa tende a ser muito incerta. Suponha, por exemplo, que apenas 6 lançamentos sejam feitos para estimar a probabilidade de que uma moeda caia com a face “cara” voltada para cima. É possível que seja estimada uma probabilidade muito diferente de 50%.

Para contornar este problema é comum supor que os dados hidrológicos sejam aleatórios e que sigam uma determinada distribuição de probabilidade analítica, como a distribuição normal, por exemplo. Esta metodologia analítica permite explorar melhor as amostras relativamente pequenas de dados hidrológicos, como se descreve na sequência deste capítulo.

## Chuvas anuais e a distribuição normal

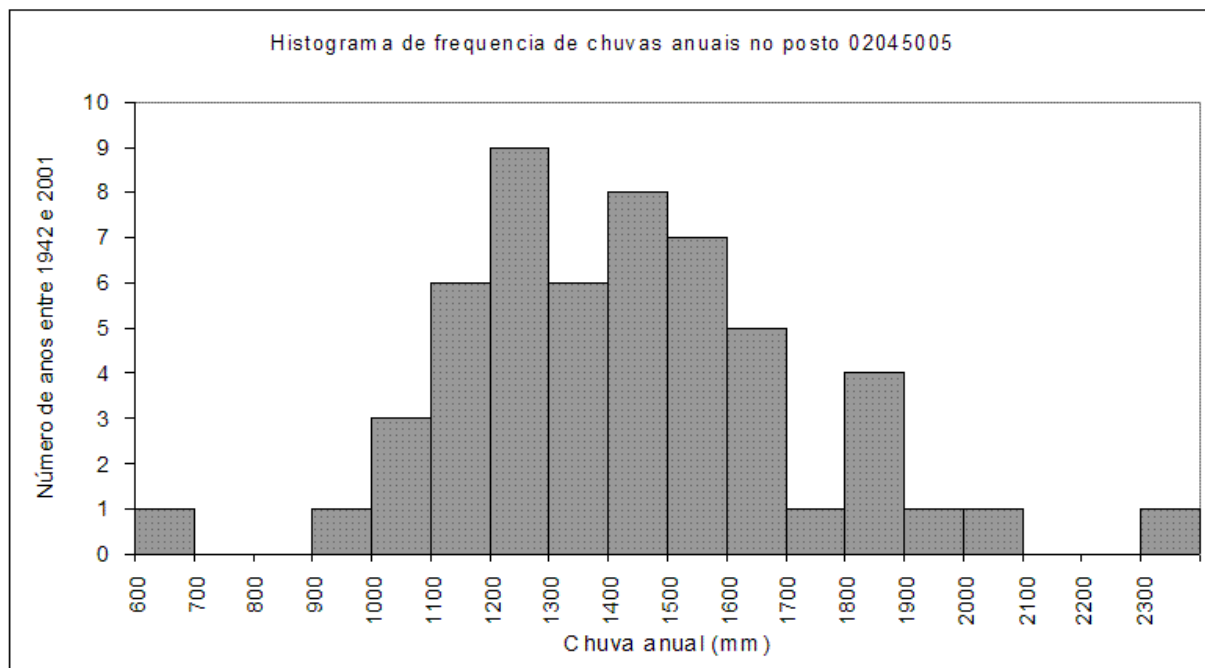
O total de chuva que cai ao longo de um ano pode ser considerado uma variável aleatória com distribuição aproximadamente normal. Esta suposição permite explorar melhor amostras relativamente pequenas, com apenas 20 anos, por exemplo.

A distribuição normal é descrita em qualquer livro introdutório de estatística e se aplica a muitos tipos de informações da natureza. Um gráfico da função densidade de probabilidade da distribuição normal tem uma forma de sino e é simétrica com relação à média, que é o valor central. A forma em sino indica que existe uma probabilidade maior de ocorrerem valores próximos à média do que nos extremos mínimo e máximo.

Lembrando a relação entre probabilidades e tempos de retorno, é interessante saber os valores de  $z$  que correspondem a alguns valores específicos de probabilidade, como 0,1 0,01 e 0,001. Estes valores correspondem aos tempos de retorno de 10, 100 e 1000 anos. Apresentamos uma tabela de probabilidades da distribuição normal, indicando os valores de  $z$  correspondentes aos tempos de retorno de 2 a 10000 anos.

$z$	Probabilidade	TR
0.000	0.5	2
0.842	0.2	5
1.282	0.1	10
1.751	0.04	25
2.054	0.02	50
2.326	0.01	100
2.878	0.002	500
3.090	0.001	1000
3.719	0.0001	10000

Considere, por exemplo, a chuva anual em um determinado local. Anos com chuva próxima da média são relativamente freqüentes, enquanto anos muito chuvosos ou muito secos são menos freqüentes. Em muitos locais as chuvas anuais seguem, aproximadamente uma distribuição normal, como mostra a figura abaixo.



A probabilidade de ocorrência de chuvas anuais superiores a 2000 mm, por exemplo, pode ser estimada a partir da análise dos dados de n anos, e da suposição de que os dados seguem uma distribuição normal.

1) As chuvas anuais no posto pluviométrico localizado em Lamounier, em Minas Gerais (Código 02045005) seguem, aproximadamente, uma distribuição normal, com média igual a 1433 mm e desvio padrão igual a 299 mm. Qual é a probabilidade de ocorrer um ano com chuva total superior a 2000 mm?

Calculando temos que:

Considerando que a média e o desvio padrão da amostra disponível sejam boas aproximações da média e do desvio padrão da população, pode se estimar o valor da variável reduzida z para o valor de 2000 mm:

$$z = \frac{x - \mu_x}{\sigma_x} \cong \frac{x - \bar{x}}{s} = \frac{2000 - 1433}{299} = 1,896$$

de acordo com a Tabela A, no final do capítulo, a probabilidade de ocorrência de um valor maior do que  $z=1,896$  é de aproximadamente 0,0287 (valor correspondente a  $z=1,9$ ). Portanto, a probabilidade de ocorrer um ano com chuva total superior a 2000 mm é de, aproximadamente, 2,87%. O tempo de retorno correspondente é de pouco menos de 35 anos.

Isto significa que, em média, um ano a cada 35 apresenta chuva total superior a 2000 mm neste local.

Utilizando o python temos que:



```
In [72]: prob5 = (1-ss.norm.cdf(2000,1433,299))
print '%5.2f' %(prob5*100) + '%'
TR = 1/prob5
print '%5.2f' %(TR) + ' anos'
```

2.90%  
34.53 anos

2) As chuvas anuais no posto pluviométrico localizado em Lamounier, em Minas Gerais (Código 02045005) seguem, aproximadamente, uma distribuição normal, com média igual a 1433 mm e desvio padrão igual a 299 mm. Qual é a probabilidade de ocorrer um ano com chuva total inferior a 550 mm?

A distribuição normal é simétrica. A probabilidade de ocorrer um valor superior a  $z$  é igual à probabilidade de ocorrer um valor inferior a  $-z$ . Assim,

$$z = \frac{x - \mu_x}{\sigma_x} \cong \frac{x - \bar{x}}{s} = \frac{550 - 1433}{299} = -2,95$$

de acordo com a Tabela A, no final do capítulo, a probabilidade de ocorrência de um valor maior do que  $z=2,95$  está entre 0,0012 e 0,0019. Portanto, a probabilidade de ocorrer um ano com chuva total inferior a 550 mm é de, aproximadamente, 0,15%. O tempo de retorno correspondente é de pouco menos de 636 anos. Isto significa que, em média, um ano a cada 636 apresenta chuva total inferior a 550 mm neste local.

```
In [73]: prob6 = (ss.norm.cdf(550,1433,299))
print prob6
print '%5.2f' %(prob6*100) + '%'
TR = 1/prob6
print '%5.2f' %(TR) + ' anos'
```

0.0015726065157136023  
0.16%  
635.89 anos

## Vazões máximas

Selecionando apenas as vazões máximas de cada ano em um determinado local, é obtida a série de vazões máximas deste local e é possível realizar análises estatísticas relacionando vazão com probabilidade. As séries de vazões disponíveis na maior parte dos locais (postos fluviométricos) são relativamente curtas, não superando algumas dezenas de anos.

Analisando as vazões do rio Cuiabá no período de 1984 a 1992, por exemplo, podemos selecionar de cada ano apenas o valor da maior vazão, e analisar apenas as vazões máximas (tabela 7.3). Reorganizando as vazões máximas para uma ordem decrescente, podemos atribuir uma probabilidade de excedência empírica a cada uma das vazões máximas da série, utilizando a fórmula de Weibull:

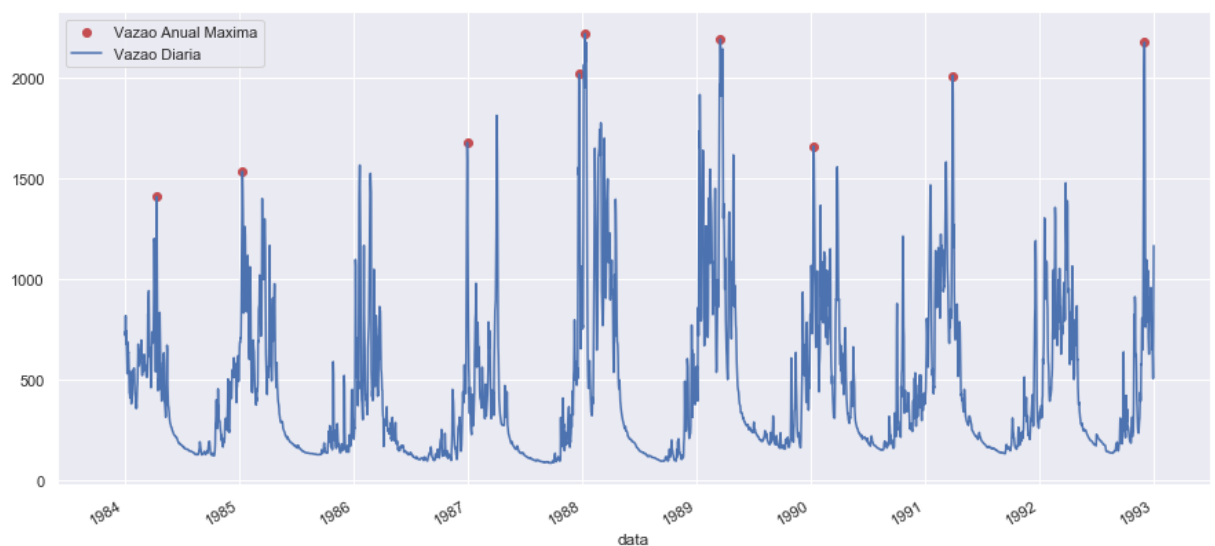
$$P = \frac{m}{N+1}$$

onde N é o tamanho da amostra (número de anos); e m é a ordem da vazão (para a maior vazão m=1 e para a menor vazão m=N). O resultado é apresentado na tabela abaixo e visualizado na figura a seguir.

```
In [74]: import datetime
vazoesCuiaba = pd.read_excel('vazaoD_Cuiaba.xlsx', index_col='data')
vazoes_sel = vazoesCuiaba['1-1-1984':'12-31-1992'].copy()
VazaoAMax_data = vazoes_sel.groupby(vazoes_sel.index.year).idxmax()
VazaoAMax = vazoes_sel.groupby(vazoes_sel.index.year).max()
fig, ax = plt.subplots(figsize=(15,7))

plt.plot_date(x=VazaoAMax_data.VazaoD, y=VazaoAMax.VazaoD, marker='o', color='r')
vazoes_sel.plot(ax=ax, label='Vazao Diaria')
ax.legend(["Vazao Anual Maxima", "Vazao Diaria"])
ax.set_xlim([datetime.date(1983, 6, 1), datetime.date(1993, 6, 30)])
print VazaoAMax
```

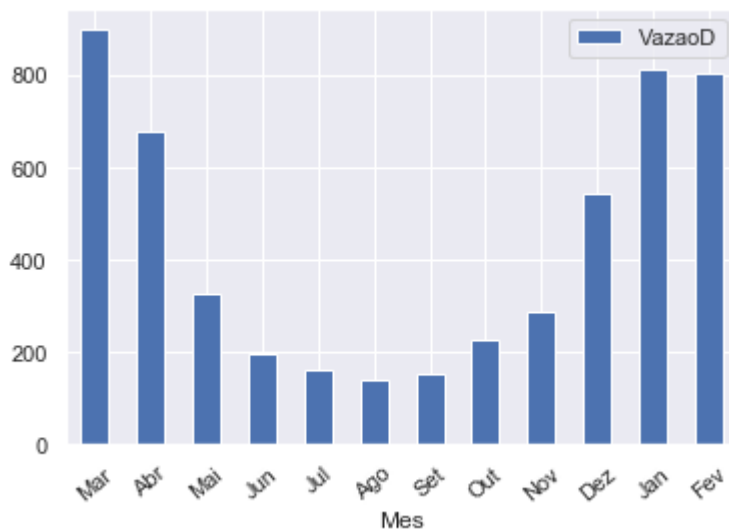
	VazaoD
data	
1984	1408.80
1985	1534.20
1986	1677.40
1987	2018.40
1988	2218.00
1989	2190.00
1990	1658.96
1991	2009.34
1992	2174.24



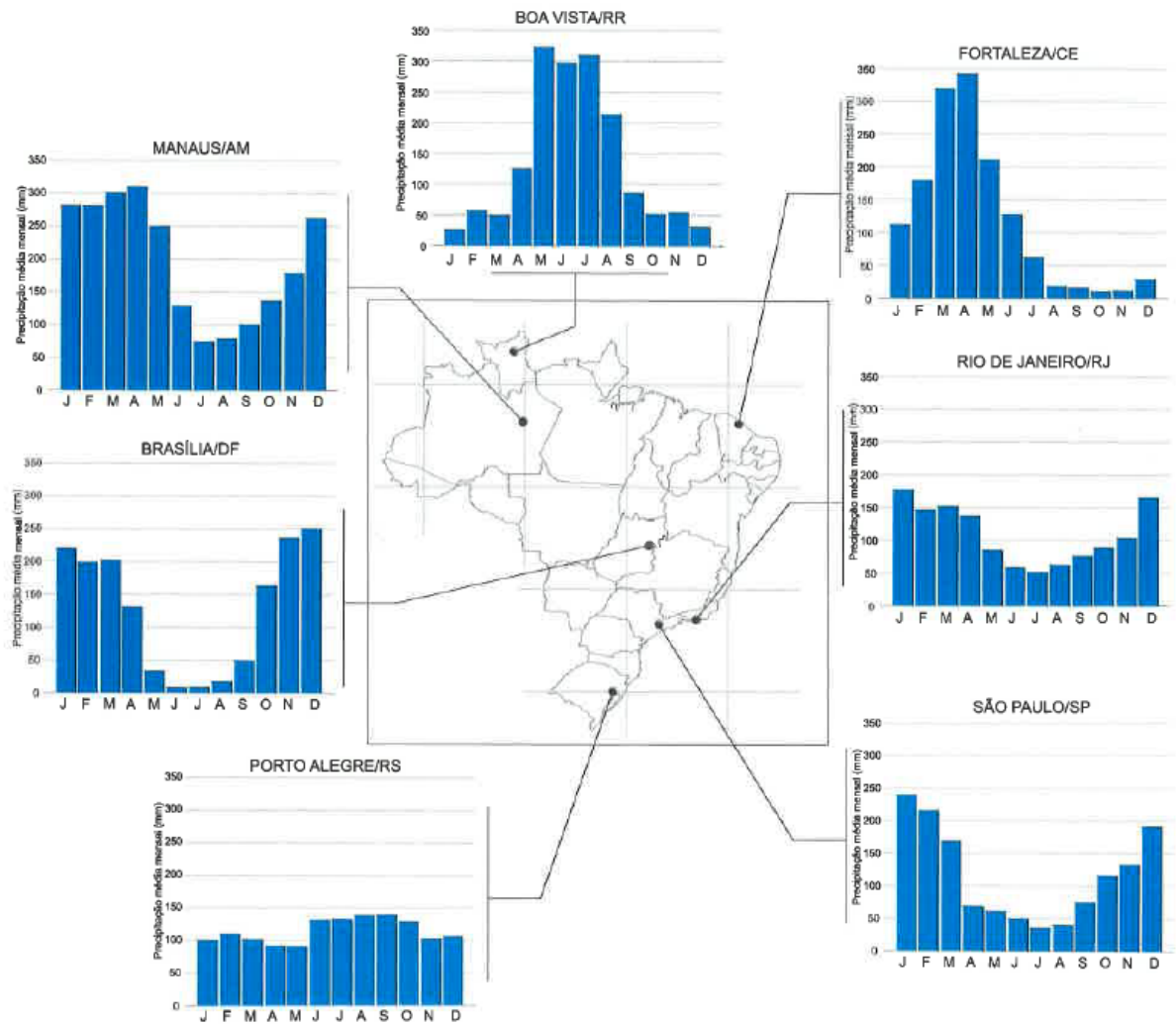
## Ano hidrológico

Na análise de vazões máximas e vazões mínimas é conveniente a definição do ano hidrológico, em contraposição ao ano civil. O ano hidrológico é definido como o período de 12 meses que se estende do início do período chuvoso até o final do período seco (Tucci, 1993). Para o rio Cuiabá, a sazonalidade das vazões é apresentada na figura abaixo.

```
In [75]: meses = ['Mar', 'Abr', 'Mai', 'Jun', 'Jul', 'Ago', 'Set', 'Out', 'Nov', 'Dez', 'Jan', 'Fev']
vazoes_sel['Mes'] = vazoes_sel.index.month
vazoes_mes = vazoes_sel.groupby('Mes').mean()
#print vazoes_mes
vazoes_mes = vazoes_mes.reindex([3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2])
#print vazoes_mes
ax = vazoes_mes.plot.bar().set_xticklabels(meses, rotation=40)
```



O ano hidrológico depende das características do clima da região. Em outras regiões do Brasil o ano hidrológico pode ser definido usando um intervalo diferente de 12 meses. Na figura abaixo é apresentada a distribuição temporal típica das chuvas em diferentes regiões do Brasil. Essa figura pode auxiliar a definir o ano hidrológico em diferentes regiões do país. No extremo norte o ano hidrológico segue um comportamento oposto ao do centro do Brasil, porque o período de maior pluviosidade corresponde aos meses de maio a agosto. Já na região Sul, o ano hidrológico é menos definido, porque a distribuição temporal das chuvas não apresenta grande sazonalidade, como mostra a figura.

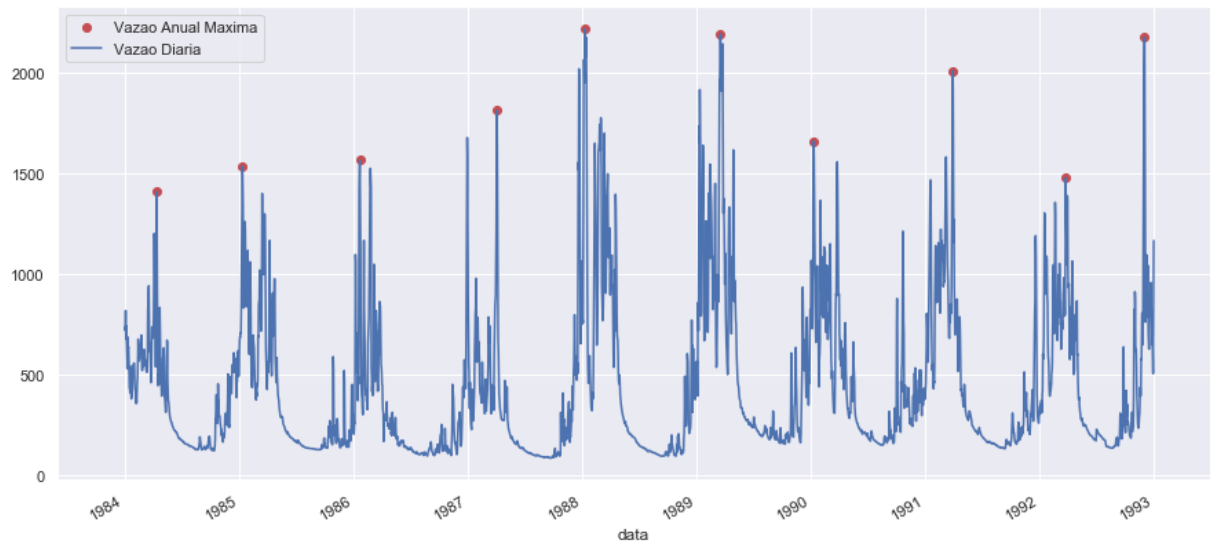


A principal motivação para a definição do ano hidrológico diferente do ano civil é que nas análises de frequência de valores máximos ou mínimos anuais é adotada a hipótese que os valores em anos sucessivos são independentes entre si. Caso o ano hidrológico seja mal definido, pode ocorrer uma situação em que a vazão máxima de um ano  $k$  seja encontrada no final do ano  $k$ , e a vazão máxima do ano  $k+1$  seja encontrada no início do ano  $k+1$ , o que significa que os dois valores são, na verdade, parte do mesmo evento de cheia. Nesse caso as duas vazões máximas não podem ser consideradas independentes.

```
In [76]: import datetime
vazoesCuiaba = pd.read_excel('vazaoD_Cuiaba.xlsx', index_col='data') #index_col=
vazoes_sel = vazoesCuiaba['1-1-1984':'12-31-1992'].copy()
VazaoAMax_data = vazoes_sel.groupby(vazoes_sel.index.to_period('A-AUG')).idxmax()
VazaoAMax = vazoes_sel.groupby(vazoes_sel.index.to_period('A-AUG')).max()
fig, ax = plt.subplots(figsize=(15,7))

plt.plot_date(x=VazaoAMax_data.VazaoD, y=VazaoAMax.VazaoD, marker='o', color='r')
vazoes_sel.plot(ax=ax, label='Vazao Diaria')
ax.legend(["Vazao Anual Maxima", "Vazao Diaria"])
ax.set_xlim([datetime.date(1983, 6, 1), datetime.date(1993, 6, 30)])
print VazaoAMax
```

```
VazaoD
data
1984 1408.80
1985 1534.20
1986 1565.00
1987 1811.70
1988 2218.00
1989 2190.00
1990 1658.96
1991 2009.34
1992 1476.54
1993 2174.24
```



O problema da estimativa empírica de probabilidades é que não é possível extrapolar a estimativa para tempos de retorno maiores. Por exemplo, se é necessário estimar a vazão máxima de 100 anos de tempo de retorno, mas existem apenas 18 anos de dados observados, as probabilidades empíricas permitem estimar vazões máximas de TR próximo de 18 anos.

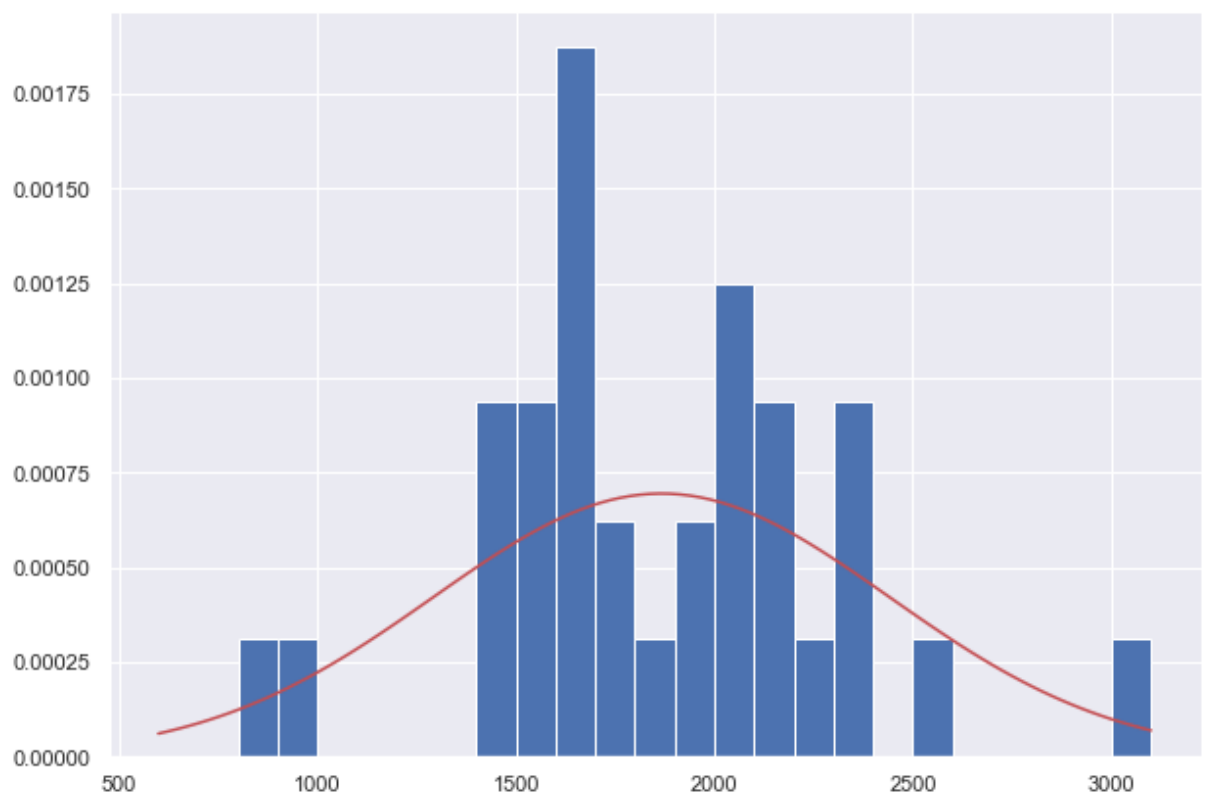
Para extrapolar as estimativas de vazão máxima é necessário supor que as vazões máximas anuais seguem uma distribuição de probabilidades conhecida, como no caso das chuvas anuais. Infelizmente, porém, as vazões máximas não seguem a distribuição normal. Histogramas de vazões máximas anuais tendem a apresentar uma forte assimetria positiva (longa cauda na direção dos maiores valores), o que invalida o uso da distribuição normal.

```
In [77]: vazoes_sel = vazoesCuiaba['1-1-1967':'12-31-1999'].copy()
VazaoAMax_data = vazoes_sel.groupby(vazoes_sel.index.to_period('A-AUG')).idxmax()
VazaoAMax = vazoes_sel.groupby(vazoes_sel.index.to_period('A-AUG')).max()
fig, ax = plt.subplots(figsize=(10,7))
media = VazaoAMax['VazaoD'].mean()
desvio = VazaoAMax['VazaoD'].std()
x = np.linspace(600, 3100, 100)
y = ss.norm.pdf(x,media,desvio)
plt.hist(VazaoAMax['VazaoD'], bins=range(600, 3200,100), color='b', normed=True)
plt.plot(x,y, color='r')
```

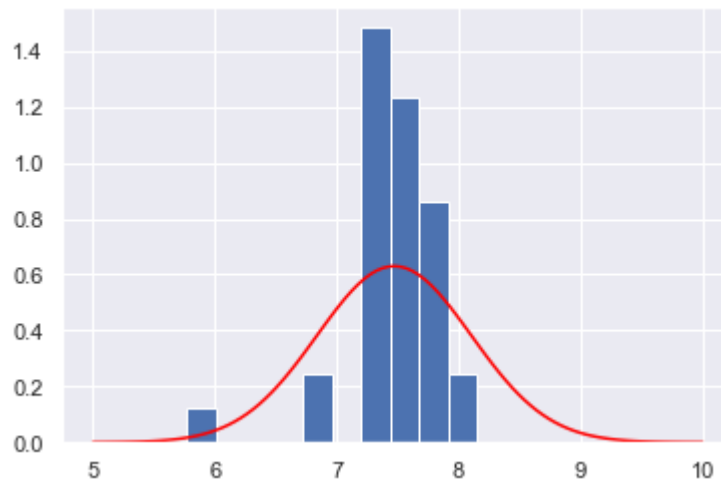
c:\python27\arcgis10.6\lib\site-packages\matplotlib\axes\\_axes.py:6571: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.

warnings.warn("The 'normed' kwarg is deprecated, and has been "

Out[77]: [<matplotlib.lines.Line2D at 0x16c03fd0>]



```
In [78]: VazaoAMax['logVazao']=np.log(VazaoAMax['VazaoD'])  
x = np.linspace(5, 10, 100)  
ax = plt.hist(VazaoAMax['logVazao'], bins=10, color='b', normed=True)  
plot_normal(x,VazaoAMax['logVazao'].mean(), VazaoAMax['logVazao'].std(), color='r')
```



```
In [79]: VazaoAMax['Rank'] = (VazaoAMax['VazaoD'].size + 1) - VazaoAMax['VazaoD'].rank()
VazaoAMax.sort_values("Rank", inplace = True)
VazaoAMax['Prob'] = VazaoAMax['Rank']/(VazaoAMax['VazaoD'].size + 1)
VazaoAMax['TR'] = 1/VazaoAMax['Prob']
print VazaoAMax
```

	VazaoD	logVazao	Rank	Prob	TR
data					
1995	3479.34	8.154598	1.0	0.028571	35.000000
1974	3007.00	8.008698	2.0	0.057143	17.500000
1997	2574.15	7.853275	3.0	0.085714	11.666667
1979	2393.70	7.780596	4.0	0.114286	8.750000
1980	2371.50	7.771278	5.0	0.142857	7.000000
1982	2323.40	7.750787	6.0	0.171429	5.833333
1988	2218.00	7.704361	7.0	0.200000	5.000000
1989	2190.00	7.691657	8.0	0.228571	4.375000
1993	2174.24	7.684434	9.0	0.257143	3.888889
1981	2143.80	7.670335	10.0	0.285714	3.500000
1978	2093.50	7.646593	11.0	0.314286	3.181818
1970	2062.40	7.631626	12.0	0.342857	2.916667
1991	2009.34	7.605562	13.0	0.371429	2.692308
1973	2001.20	7.601502	14.0	0.400000	2.500000
1996	1997.76	7.599782	15.0	0.428571	2.333333
1968	1918.81	7.559460	16.0	0.457143	2.187500
1987	1811.70	7.502021	17.0	0.485714	2.058824
1984	1793.20	7.491757	18.0	0.514286	1.944444
1999	1742.27	7.462944	19.0	0.542857	1.842105
1994	1680.54	7.426870	20.0	0.571429	1.750000
1998	1669.74	7.420423	21.0	0.600000	1.666667
1990	1658.96	7.413946	22.0	0.628571	1.590909
1977	1643.60	7.404644	23.0	0.657143	1.521739
1983	1638.50	7.401536	24.0	0.685714	1.458333
1972	1612.40	7.385479	25.0	0.714286	1.400000
1969	1588.48	7.370533	26.0	0.742857	1.346154
1986	1565.00	7.355641	27.0	0.771429	1.296296
1985	1534.20	7.335764	28.0	0.800000	1.250000
1992	1476.54	7.297457	29.0	0.828571	1.206897
1975	1466.20	7.290429	30.0	0.857143	1.166667
1976	1433.80	7.268084	31.0	0.885714	1.129032
1967	984.62	6.892256	32.0	0.914286	1.093750
1971	897.40	6.799502	33.0	0.942857	1.060606
2000	321.53	5.773091	34.0	0.971429	1.029412

Para superar este problema existem outras distribuições de probabilidade que são, normalmente, utilizadas para a análise de vazões máximas. A mais simples destas distribuições é a denominada log-normal. Nesta distribuição a suposição é que os logaritmos das vazões seguem uma distribuição normal.

Se o objetivo da análise é determinar a vazão de 100 anos de tempo de retorno em um determinado local, por exemplo, a sequência de etapas para a estimativa supondo que os dados correspondem a uma distribuição log-normal é a seguinte:

- Obter vazões máximas de N anos
- Calcular os logaritmos das vazões máximas



- Calcular a média e o desvio padrão dos logaritmos das vazões máximas
- Obter o valor de z para a probabilidade correspondente ao tempo de retorno de 100 anos
- Obter o valor do logaritmo da vazão de tempo de retorno de 100 anos a partir da variável aleatória z
- Obter o valor da vazão através da função inversa do logaritmo.

Esta sequência de etapas fica mais clara na aplicação em um exemplo.

## EXEMPLOS

1) As vazões máximas anuais do rio Guaporé no posto fluviométrico Linha Colombo são apresentadas na tabela abaixo. Utilize a distribuição log-normal para estimar a vazão máxima com 100 anos de tempo de retorno.

ANO	MAXIMA	ANO	MAXIMA	ANO	MAXIMA	ANO	MAXIMA	ANO	MAXIMA	ANO	MAXIMA
1940	953	1950	1192	1960	falha	1970	365	1980	653	1990	falha
1941	1171	1951	356	1961	718	1971	671	1981	537	1991	falha
1942	723	1952	246	1962	503	1972	1785	1982	945	1992	falha
1943	267	1953	1093	1963	falha	1973	726	1983	1650	1993	1115
1944	646	1954	840	1964	457	1974	397	1984	1165	1995	639
1945	365	1955	622	1965	915	1975	480	1985	888		
1946	1359	1956	falha	1966	742	1976	falha	1986	728		
1947	411	1957	598	1967	840	1977	673	1987	809		
1948	480	1958	646	1968	331	1978	760	1988	945		
1949	365	1959	953	1969	320	1979	780	1989	1380		

Este exemplo apresenta uma situação muito comum na análise de dados hidrológicos: as falhas. As falhas são períodos em que não houve observação. As falhas são desconsideradas na análise, assim o tamanho da amostra é N=48. Utilizando logaritmos de base decimal, a média dos logaritmos das vazões máximas é 2,831 e o desvio padrão é 0,206. Para o tempo de retorno de 100 anos a probabilidade de excedência é igual a 0,01. Na tabela B, ao final do capítulo, pode-se obter o valor de z correspondente (z=2,326). A vazão máxima de TR=100 anos é obtida por:

$$z \cong \frac{x - \bar{x}}{s}$$

$$2,326 \cong \frac{x - 2,831}{0,206}$$

$$x = 2,326 \cdot 0,206 + 2,831 = 3,31$$

$$Q = 10^{3,31} = 2041$$

*Portanto, a vazão máxima de 100 anos de tempo de retorno é 2041 m<sup>3</sup>/s.*

```
In [80]: VazaoMax= pd.read_excel('Linha_Colombo_Max_Anual.xlsx', index_col='ANO')
VazaoMax.dropna(inplace=True)
VazaoMax['logVazao']=np.log10(VazaoMax['MAXIMA'])
VazaoMax['Rank'] = (VazaoMax['MAXIMA'].size + 1) - VazaoMax['MAXIMA'].rank()
VazaoMax.sort_values("Rank", inplace = True)
VazaoMax['Prob'] = VazaoMax['Rank']/(VazaoMax['MAXIMA'].size + 1)
VazaoMax['TR'] = 1/VazaoMax['Prob']
mediaLog = VazaoMax['logVazao'].mean()
desvioLog = VazaoMax['logVazao'].std()
Vazao = 10**(ss.norm.ppf(q=0.99, loc=mediaLog, scale=desvioLog))
print '%5.2f' %(Vazao) + ' m3/s'
```

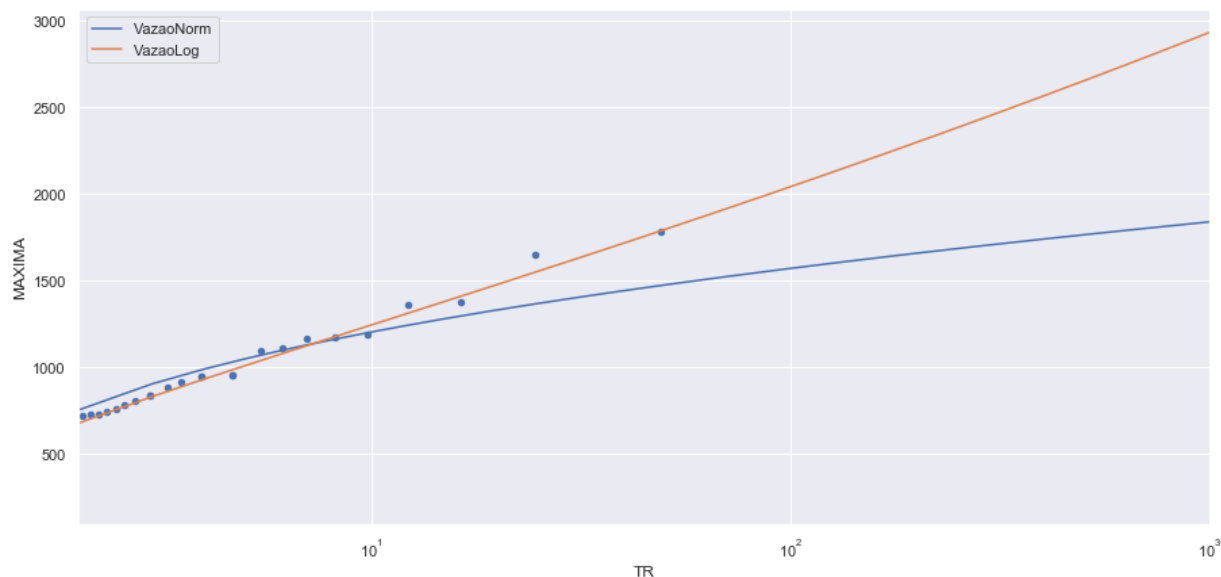
2041.82 m3/s

Este procedimento pode ser repetido para outros valores de TR, e o resultado pode ser apresentado na forma de um gráfico, relacionando vazão com tempo de retorno, como na figura 7.13. Nesta figura fica claro, também, que a suposição de uma distribuição log-normal é muito mais adequada do que a suposição de uma distribuição normal.

```

In [81]: TR = pd.Series(np.arange(2, 1001, 1))
Prob = pd.Series(np.zeros((999), dtype=np.float64))
i=0
for item in TR:
    teste = 1-(1/item)
    Prob[i]=(teste)
    i=i+1
VazaoNorm = pd.Series(np.zeros((999), dtype=np.float64))
VazaoLog = pd.Series(np.zeros((999), dtype=np.float64))
mediaNorm = VazaoMax['MAXIMA'].mean()
desvioNorm = VazaoMax['MAXIMA'].std()
i=0
for item in Prob:
    VazaoNorm[i] = (ss.norm.ppf(q=item, loc=mediaNorm, scale=desvioNorm))
    i=i+1
i=0
for item in Prob:
    VazaoLog[i] = 10**(ss.norm.ppf(q=item, loc=mediaLog, scale=desvioLog))
    i=i+1
fig, ax = plt.subplots(figsize=(15,7))
DF = pd.DataFrame()
DF['VazaoNorm']=VazaoNorm
DF['VazaoLog']=VazaoLog
DF['TR']=TR
DF.plot(x="TR", y="VazaoNorm", ax=ax)
DF.plot(x="TR", y="VazaoLog", ax=ax)
VazaoMax.plot.scatter(x='TR', y= 'MAXIMA', ax=ax)
ax.set_xscale('log')

```



Os métodos de estimativa de vazões máximas apresentados neste texto são relativamente simples e a forma de apresentação é resumida. Para realizar análises de vazões máximas mais rigorosas normalmente é necessário testar três ou mais distribuições de probabilidade teóricas, e avaliar qual é a distribuição que melhor se adequa aos dados. Metodologias mais aprofundadas podem ser encontradas em Tucci (1993), Maidment (1993) e Wurbs e James (2001).

## Vazões mínimas

A análise de vazões mínimas é semelhante à análise de vazões máximas, exceto pelo fato que no caso das vazões mínimas o interesse é pela probabilidade de ocorrência de vazões iguais ou menores do que um determinado limite.

No caso da análise utilizando probabilidades empíricas, esta diferença implica em que os valores de vazão devem ser organizados em ordem crescente, ao contrário da ordem decrescente utilizada no caso das vazões máximas.

A aplicação da análise estatística para vazões mínimas é analisada através de um exemplo.

### EXEMPLO

5) A tabela abaixo apresenta as vazões mínimas anuais observadas no rio Piquiri, no município de Iporã (PR). Considerando que os dados seguem uma distribuição normal, determine a vazão mínima de 5 anos de tempo de retorno. A distribuição normal se ajusta bem aos dados observados?

Ano	Mínima
1980	202
1981	128.6
1982	111.4
1983	269
1984	158.2
1985	77.5
1986	77.5
1987	166
1988	70
1989	219.6
1990	221.8
1991	111.4
1992	204.2
1993	196
1994	172
1995	130.4
1996	121.6
1997	198
1998	320.6
1999	101.2
2000	118.2
2001	213

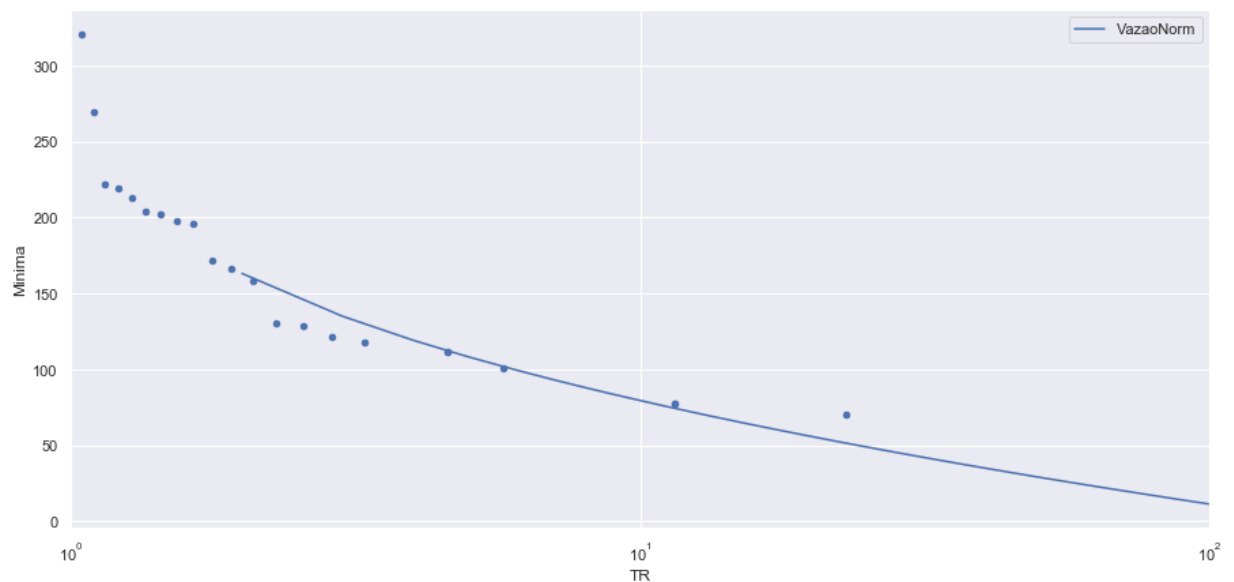
```
In [82]: VazaoMin= pd.read_excel('PiquiriMinimas.xlsx', index_col='Ano')
VazaoMin['Rank'] = VazaoMin['Minima'].rank()
VazaoMin['Rank'] = (VazaoMin['Rank']//1)
VazaoMin.sort_values("Rank", inplace = True)
VazaoMin['Prob'] = VazaoMin['Rank']/(VazaoMin['Minima'].size + 1)
VazaoMin['TR'] = 1/VazaoMin['Prob']
mediaNorm = VazaoMin['Minima'].mean()
desvioNorm = VazaoMin['Minima'].std()
print VazaoMin
```

	Minima	Rank	Prob	TR
Ano				
1988	70.0	1.0	0.043478	23.000000
1985	77.5	2.0	0.086957	11.500000
1986	77.5	2.0	0.086957	11.500000
1999	101.2	4.0	0.173913	5.750000
1991	111.4	5.0	0.217391	4.600000
1982	111.4	5.0	0.217391	4.600000
2000	118.2	7.0	0.304348	3.285714
1996	121.6	8.0	0.347826	2.875000
1981	128.6	9.0	0.391304	2.555556
1995	130.4	10.0	0.434783	2.300000
1984	158.2	11.0	0.478261	2.090909
1987	166.0	12.0	0.521739	1.916667
1994	172.0	13.0	0.565217	1.769231
1993	196.0	14.0	0.608696	1.642857
1997	198.0	15.0	0.652174	1.533333
1980	202.0	16.0	0.695652	1.437500
1992	204.2	17.0	0.739130	1.352941
2001	213.0	18.0	0.782609	1.277778
1989	219.6	19.0	0.826087	1.210526
1990	221.8	20.0	0.869565	1.150000
1983	269.0	21.0	0.913043	1.095238
1998	320.6	22.0	0.956522	1.045455

```
In [83]: TR = pd.Series(np.arange(1, 101, 1))
Prob = pd.Series(np.zeros((100), dtype=np.float64))
i=0
for item in TR:
    teste = (1/item)
    Prob[i]=(teste)
    i=i+1
i=0
for item in Prob:
    VazaoNorm[i] = (ss.norm.ppf(q=item, loc=mediaNorm, scale=desvioNorm))
    i=i+1

fig, ax = plt.subplots(figsize=(15,7))
DF = pd.DataFrame()
DF['VazaoNorm']=VazaoNorm
DF['TR']=TR
DF.plot(x="TR", y="VazaoNorm", ax=ax)

VazaoMin.plot.scatter(x='TR', y= 'Minima', ax=ax)
ax.set_xscale('log')
```



Normalmente, as análises estatísticas de vazões mínimas são realizadas sobre as vazões mínimas de 7 dias, 15 dias ou 30 dias de duração. Neste caso, para cada ano do registro histórico encontra-se a vazão mínima média de 7 dias (médias móveis de 7 dias). O restante do procedimento de análise é semelhante ao apresentado aqui.

## A distribuição binomial

A distribuição de probabilidades binomial é adequada para avaliar o número (x) de ocorrências de um dado evento em N tentativas.

As seguintes condições devem existir para que seja válida a distribuição binomial:

- 1) são realizadas N tentativas;

2) em cada tentativa o evento pode ocorrer ou não, sendo que a probabilidade de que o evento ocorra é dada por  $P$  enquanto a probabilidade de que o evento não ocorra é dada por  $1-P$  ;

3) a probabilidade de ocorrência do evento numa tentativa qualquer é constante e as tentativas são independentes, isto é, a ocorrência ou não do evento na tentativa anterior não altera a probabilidade de ocorrência atual.

Estas propriedades ficam mais claras considerando o exemplo de um dado de seis faces. A probabilidade de obter um “seis” num lançamento qualquer é de  $1/6$ . A probabilidade de não obter um “seis” num lançamento qualquer é de  $5/6$ . Se um dado é lançado uma vez, resultando em um “seis”, isto não altera a probabilidade de obter um “seis” no lançamento seguinte.

De acordo com a probabilidade binomial, a probabilidade de que um evento ocorra  $x$  vezes em  $N$  tentativas, é dada pela equação abaixo:

$$P_x(X = x) = \frac{N!}{x!(N-x)!} \cdot P^x \cdot (1-P)^{N-x}$$

Nesta equação  $P_x(X=x)$  é a probabilidade de que o evento ocorra  $x$  vezes em  $N$  tentativas.  $P$  é a probabilidade que o evento ocorra numa tentativa qualquer e  $(1-P)$  é a probabilidade que o evento não ocorra numa tentativa qualquer.

#### EXEMPLOS

1) Calcule a probabilidade de obter exatamente 5 “coroas” em 10 lançamentos de uma moeda. Neste caso  $x=5$  e  $N=10$ . A probabilidade de obter “coroa” num lançamento qualquer é de 50%, ou  $1/2$ . A probabilidade de obter exatamente 5 “coroas” pode ser calculada pela equação apresentada anteriormente.

$$P_x(X = 5) = \frac{10!}{5!(10-5)!} \cdot \left(\frac{1}{2}\right)^5 \cdot \left(1 - \frac{1}{2}\right)^{10-5} = 0,246$$

*Portanto, a probabilidade de obter exatamente 5 “coroas” em 10 lançamentos é de 24,6%.*

```
In [84]: print '%5.2f' %(ss.binom.pmf(5,10,0.5)*100) + ' %'
```

24.61 %

2) A probabilidade da vazão de 10 anos de tempo de retorno seja igualada ou excedida num ano qualquer é de 10%. Qual é a probabilidade que ocorram duas cheias iguais ou superiores à cheia de  $TR = 10$  anos em dois anos seguidos?

Neste caso  $x=2$  e  $N=2$ . A probabilidade de ocorrer a cheia num ano qualquer é de 10%, ou  $1/10$ . A probabilidade de ocorrer exatamente 2 cheias em 2 anos pode ser calculada pela equação da probabilidade binomial.

$$P_x(X = 2) = \frac{2!}{2!(2-2)!} \cdot \left(\frac{1}{10}\right)^2 \cdot \left(1 - \frac{1}{10}\right)^{2-2} = \left(\frac{1}{10}\right)^2 = 0,01$$

*Portanto, a probabilidade de ocorrerem exatamente 2 cheias em 2 anos é 1%.*

```
In [85]: print '%5.2f' %(ss.binom.pmf(2,2,0.1)*100) + ' %'
```

1.00 %

## 10 Distribuição amostral e erros

Após extrair uma amostra de uma população sabemos as respostas dos indivíduos da amostra. Mas muitas vezes não basta ter as informações sobre a amostra. Queremos inferir a partir dos dados amostrais alguma conclusão sobre a população mais ampla que a amostra representa. A inferência estatística fornece métodos para extrair conclusões sobre uma população a partir dos dados amostrais.

Quando obtemos dados de uma amostra não podemos ter certeza de que nossas conclusões são corretas. Uma amostra diferente poderia conduzir a conclusões diferentes. A inferência estatística usa a linguagem de probabilidade para expressar o grau de confiança das conclusões.

Considere, por exemplo, os dados da seguinte tabela, que expressam a perda de cálcio (%) de 47 mães durante três meses de amamentação. Considere que as 47 mães são a população. Com a média de uma amostra de 20 mães, extraída desta população, será estimada a média desta população. No entanto, desta mesma população pode-se tomar muitas amostras diferentes de mesmo tamanho ( $n=20$ ). Estas amostras provavelmente terão médias um pouco diferentes entre si. Qual delas é a melhor estimativa da média da população?

-4.7	-2.5	-4.9	-2.7	-0.8	-5.3	-8.3	-2.1	-6.8	-4.3
2.2	-7.8	-3.1	-1	-6.5	-1.8	-5.2	-5.7	-7	-2.2
-6.5	-1	-3	-3.6	-5.2	-2	-2.1	-5.6	-4.4	-3.3
-4	-4.9	-4.7	-3.8	-5.9	-2.5	-0.3	-6.2	-6.8	1.7
0.3	-2.3	0.4	-5.3	0.2	-2.2	-5.1			

Usando os dados das duas primeiras linhas ( $n=20$ ) a média de perda de cálcio é de menos 4,025 %, já usando os dados das linhas 3 e 4, a média é de -3,705. Nenhuma das duas estimativas é perfeita, já que a média da população, neste caso, é de -3,587.

É claro que, à medida que aumenta o tamanho da amostra, a média da amostra se aproxima cada vez mais da média da população. A lei dos grandes números diz que a média da amostra  $\bar{x}$  será próxima da média da população  $\mu$  se tomarmos uma amostra suficientemente grande. Entretanto,  $\bar{x}$  raramente será igual a  $\mu$  logo, nossa estimativa terá algum erro.

## Intervalos de confiança para as médias



O intervalo de confiança indica o quanto eu posso confiar no valor da média obtida da amostra. Por exemplo: A partir dos dados da amostra, eu tenho 95% de certeza que a média está entre 10,54 e 11,21 mg/l.

O intervalo de confiança para a estimativa da média pode ser calculado por:

$$\mu = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}}$$

Isto significa que dada uma amostra com média  $\bar{x}$  tirada de uma população com desvio padrão  $\sigma$ , espera-se que a média da população da qual foi extraída a amostra esteja dentro de um intervalo dado por:

$$\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}}$$

O valor de  $Z_{\alpha/2}$  depende do grau de confiança que se deseja para a estimativa. Para uma confiança de 95% o valor de  $Z_{\alpha/2}$  é de 1,960. Outros valores são dados na tabela abaixo. Estes valores de  $Z_{\alpha/2}$  são obtidos da distribuição normal.

Nível de confiança	Valor crítico de $Z_{\alpha/2}$
90%	1.645
95%	1.960
99%	2.576

## Intervalos de confiança para as médias com amostras pequenas

Com amostras pequenas ( $n < 30$ ), ou nos casos em que não se conhece o desvio padrão da população, mas apenas o da amostra, o intervalo de confiança para as médias pode ser obtido:

$$\mu = \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{N}}$$

Isto significa que dada uma amostra com média  $\bar{x}$  e desvio padrão  $s$ , espera-se que a média da população da qual foi extraída a amostra esteja dentro de um intervalo dado por:

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{N}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{N}}$$

O valor de  $t_{\alpha/2}$  depende do grau de confiança que se deseja para a estimativa e do tamanho da amostra. Com amostras pequenas a distribuição normal tem menor validade. Neste caso é mais

adequado utilizar a distribuição de Student (t). Quando o tamanho da amostra supera 30 ( $N > 30$ ) a distribuição de Student e a distribuição normal ficam muito próximas.

Os valores de  $t_{\alpha/2}$  da distribuição de Student são dados em uma tabela.

Exemplos:

1) A amostra (7; 4; 2; 5; 7) foi obtida de uma população com distribuição normal. Estime a média da população com intervalo de confiança 90%.

Solução: A média da amostra é 5. O desvio padrão é 2,12. O tamanho da amostra é 5. O número de graus de liberdade é  $N-1 = 4$ .

Então o valor de  $t_{\alpha/2}$  é de aproximadamente 2,132 (este valor pode ser obtido da tabela para a confiança de 90% com 4 graus de liberdade).

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{N}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{N}}$$

$$5 - 2,132 \cdot \frac{2,12}{\sqrt{5}} < \mu < 5 + 2,132 \cdot \frac{2,12}{\sqrt{5}}$$

Assim, a média da população deve estar entre 5-2,02 e 5+2,02. Podemos dizer isso com 90% de confiança.

```
In [86]: amostra = np.array([7, 4, 2, 5, 7])
media = amostra.mean()
desvio = amostra.std(ddof=1)
escala = desvio/np.sqrt(len(amostra))
ss.t.interval(0.90, 4, loc=media, scale=escala)
```

```
Out[86]: (2.9775525645435152, 7.022447435456483)
```

2) Foram realizadas 12 medições para calcular o coeficiente C de um orifício a partir de dados de vazão observados em um laboratório utilizando a equação abaixo. Os valores obtidos em cada uma das medições são apresentados na tabela abaixo. Determine o intervalo de confiança para o valor médio de C. Considere um grau de confiança de 95%.

$$Q = C \cdot A \cdot \sqrt{2 \cdot g \cdot h}$$

Medição	Coeficiente
1	0.613
2	0.598
3	0.62
4	0.61
5	0.615
6	0.623
7	0.604
8	0.609
9	0.611
10	0.608
11	0.603
12	0.61

```
In [87]: coeficientes= pd.read_excel('coeficientes.xlsx', index_col='Medicao')
media = coeficientes['Coeficiente'].mean()
print media
desvio = coeficientes['Coeficiente'].std(ddof=1)
print desvio
escala = desvio/np.sqrt(coeficientes['Coeficiente'].count())
print escala
ss.t.interval(0.95, 4, loc=media, scale=escala)
```

```
0.6103333333333333
0.006984832051515546
0.0020163473325934132
```

Out[87]: (0.6047350556513756, 0.6159316110152909)

3) Foram obtidas 15 amostras de sedimentos do leito de um rio. O diâmetro mediano característico  $d_{50}$  foi calculado para cada uma das amostras (veja na tabela abaixo). Determine o intervalo de confiança para o valor médio de  $d_{50}$  do leito deste rio com base nestes dados. Considere um grau de confiança de 90%.

Amostra	Diâmetro $d_{50}$ (mm)
1	0.739
2	1.098
3	0.820
4	0.910
5	1.015
6	0.623
7	1.504
8	1.609
9	0.811
10	0.758
11	0.923
12	0.918
13	1.001
14	1.200
15	0.998

```
In [88]: diametros= pd.read_excel('diametro.xlsx', index_col='Amostra')
media = diametros['Diametro'].mean()
print media
desvio = diametros['Diametro'].std(ddof=1)
print desvio
escala = desvio/np.sqrt(diametros['Diametro'].count())
print escala
ss.t.interval(0.90, 14, loc=media, scale=escala)
```

```
0.9951333333333332
0.2715121325009955
0.07010413116464113
```

```
Out[88]: (0.8716582165533607, 1.1186084501133056)
```

4) Um estagiário do curso técnico de hidrologia fez uma série de medições de velocidade com um molinete chegando ao valor médio de 0,74 m/s. O seu chefe ficou desconfiado do resultado e enviou outro estagiário para repetir as medições. Este obteve os resultados da tabela ao lado. O valor encontrado pelo primeiro estagiário é suspeito? Resolva o problema usando o intervalo de confiança de 90% para a média.

Medição	Velocidades (m/s)
1	0.739
2	0.739
3	0.720
4	0.710
5	0.739
6	0.723
7	0.739
8	0.739
9	0.711
10	0.758
11	0.723
12	0.721
13	0.721
14	0.734
15	0.748

```
In [89]: velocidades= pd.read_excel('velocidades.xlsx', index_col='Medicao')
media = velocidades['Velocidades'].mean()
print media
desvio = velocidades['Velocidades'].std(ddof=1)
print desvio
escala = desvio/np.sqrt(velocidades['Velocidades'].count())
print escala
ss.t.interval(0.90, 14, loc=media, scale=escala)
```

```
0.7309333333333333
0.01372415319618596
0.003543561117975167
```

```
Out[89]: (0.724692023219506, 0.7371746434471607)
```

## Testes de hipóteses para diferenças de médias

### Testes Paramétricos sobre a Média de uma Única População Normal

A premissa básica dos testes, descritos a seguir, é a de que as variáveis aleatórias independentes  $\{X_1, X_2, \dots, X_N\}$ , componentes de uma certa amostra aleatória simples, foram todas extraídas de uma única população normal, de média  $\mu$  desconhecida. O conhecimento ou o desconhecimento da variância populacional  $\sigma^2$  determina a estatística de teste a ser usada.

•  **$H_0: \mu = \mu_1$  contra  $H_1: \mu = \mu_2$ . Atributo de  $\sigma^2$ : conhecida.**

Estatística de teste:

$$Z = \frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{N}}}$$

Distribuição de probabilidades da estatística de teste: Normal  $N(0,1)$

Tipo de Teste: unilateral a um nível de significância  $\alpha$

Decisão:

Se  $\mu_1 > \mu_2$ , rejeitar  $H_0$  se

$$\frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{N}}} < -Z_{1-\alpha}$$

Se  $\mu_1 < \mu_2$ , rejeitar  $H_0$  se

$$\frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{N}}} > +Z_{1-\alpha}$$

•  **$H_0: \mu = \mu_1$  contra  $H_1: \mu = \mu_2$ . Atributo de  $\sigma^2$ : desconhecida e estimada por  $S^2_X$ .**

Estatística de teste:

$$T = \frac{\bar{X} - \mu_1}{\frac{S_X}{\sqrt{N}}}$$

Distribuição de probabilidades da estatística de teste: t de Student com  $\nu = N-1$  ou  $t_{N-1}$

Tipo de Teste: unilateral a um nível de significância  $\alpha$

Decisão:

Se  $\mu_1 > \mu_2$ , rejeitar  $H_0$  se

$$\frac{\bar{X} - \mu_1}{\frac{S_X}{\sqrt{N}}} < -t_{1-\alpha, \nu=N-1}$$

Se  $\mu_1 < \mu_2$ , rejeitar  $H_0$  se

$$\frac{\bar{X} - \mu_1}{\frac{S_X}{\sqrt{N}}} > +t_{1-\alpha, \nu=N-1}$$

•  **$H_0: \mu = \mu_0$  contra  $H_1: \mu \neq \mu_0$ . Atributo de  $\sigma^2$ : conhecida.**

Estatística de teste:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{N}}}$$

Distribuição de probabilidades da estatística de teste: Normal  $N(0,1)$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão:

$$\left| \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{N}}} \right| > Z_{1-\alpha/2}$$

•  **$H_0: \mu = \mu_0$  contra  $H_1: \mu \neq \mu_0$ . Atributo de  $\sigma^2$ : desconhecida e estimada por  $S^2_X$ .**

Estatística de teste:

$$T = \frac{\bar{X} - \mu_0}{\frac{S_X}{\sqrt{N}}}$$

Distribuição de probabilidades da estatística de teste: t de Student com  $\nu=N-1$  ou  $t_{N-1}$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão:

$$\left| \frac{\bar{X} - \mu_0}{\frac{S_X}{\sqrt{N}}} \right| > t_{1-\alpha/2, \nu=N-1}$$

### Exemplos

1) Considere as vazões médias do mês de Julho do Rio Paraopeba em Ponte Nova do Paraopeba (VMM\_Paraopeba.xlsx), para o período de 1938 a 1999. Teste a hipótese de que a média populacional do mês de Julho é 47,65 m<sup>3</sup>/s, a um nível de significância  $\alpha = 5\%$ .

Solução: A premissa básica é a que as vazões médias do mês de Julho, em Ponte Nova do Paraopeba, seguem uma distribuição Normal. A amostra de 62 observações fornece  $\bar{X}=44,526$  e  $S_X=12,406$  m<sup>3</sup>/s, não havendo nenhuma informação adicional sobre a variância populacional.

Nesse caso, a hipótese nula é  $H_0: \mu = 47,65$  contra a hipótese alternativa  $H_1: \mu \neq 47,65$ . Trata-se, portanto, de um teste bilateral ao nível  $\alpha = 5\%$ , com a estatística de teste dada por:

$$\frac{\bar{X} - 47,65}{\frac{S_X}{\sqrt{N}}}$$

, a qual possui uma distribuição t de Student com 61 graus de liberdade.

Substituindo os valores amostrais, resulta que o valor absoluto da estimativa de T é igual a 1,9828. A tabela de t de Student, fornece  $t_{0.975, \nu=61}$ .

Como  $1,9828 < 1,9996$ , a hipótese  $H_0$  não deve ser rejeitada, em favor de  $H_1$ . Em outras palavras, com base na amostra disponível, não há evidências de que a média populacional difira significativamente de 47,65 m<sup>3</sup>/s, ou seja, que a diferença existente entre a média amostral  $\bar{X}=44,526$  e a média hipotética  $\mu=47,65$  deve-se unicamente a flutuações aleatórias das observações.

```
In [90]: paraopeba= pd.read_excel('VMM_ParaopebaAnoCivil.xlsx')
# paraopeba['Jul'].describe()
mediaA=paraopeba['Jul'].mean()
desvio=paraopeba['Jul'].std()
N=paraopeba['Jul'].count()
media = 47.65
T=abs((mediaA-media)/(desvio/np.sqrt(N)))
T
```

Out[90]: 1.982844199727758

```
In [91]: Tcritico = ss.t.ppf(0.975, N-1, loc=0, scale=1)
print Tcritico
```

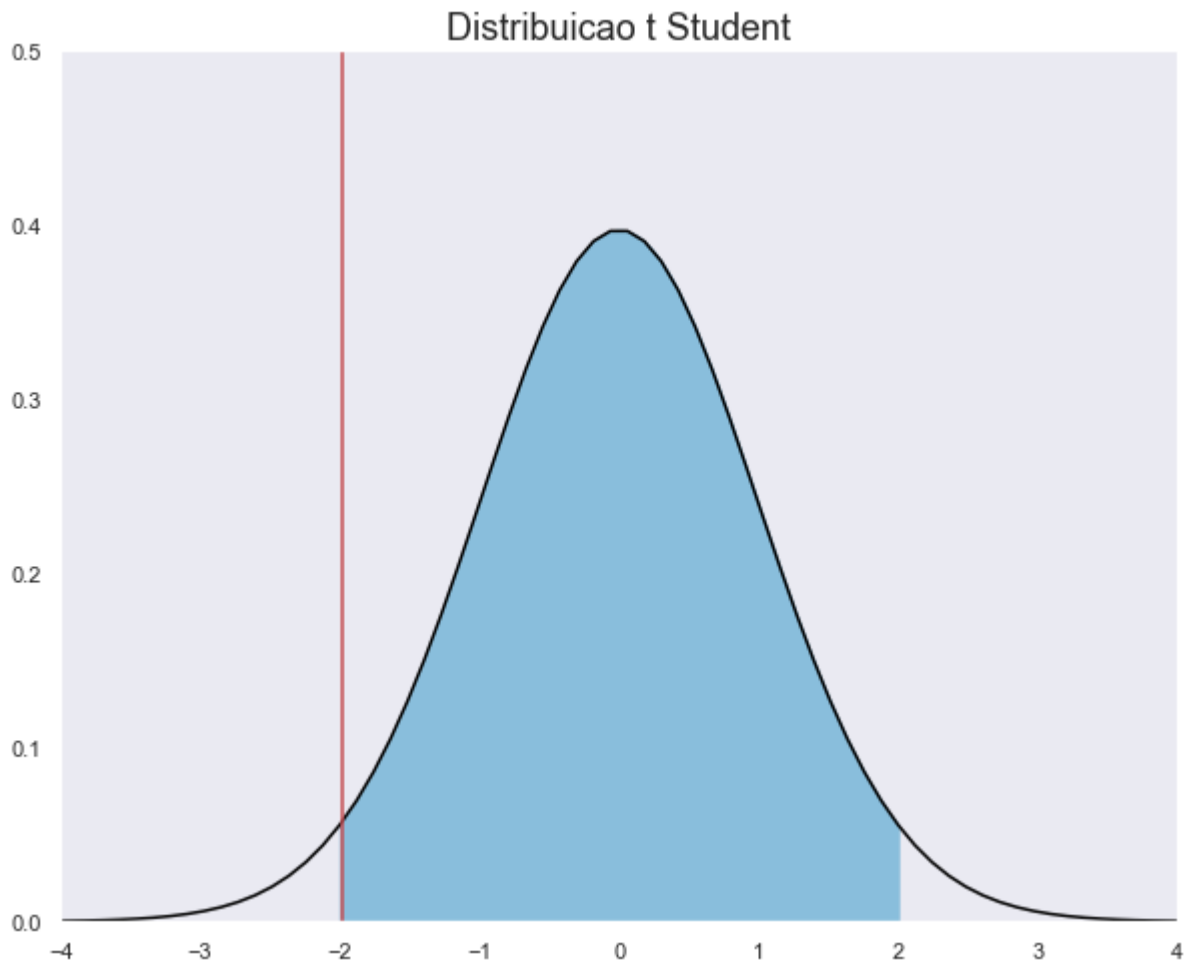
1.9996235841149779

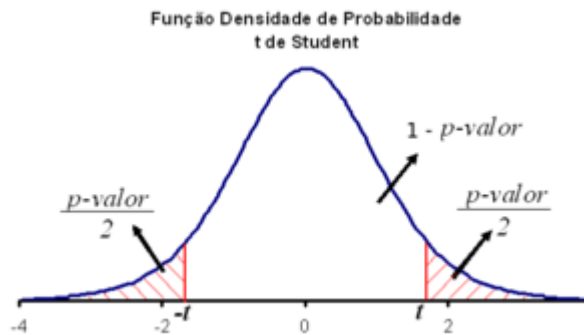
```
In [92]: T = ss.ttest_1samp(paraopeba['Jul'],[media]).statistic[0]
Pvalor = ss.ttest_1samp(paraopeba['Jul'],[media]).pvalue[0]
print T
print Pvalor
if Pvalor<0.05:
    print("Rejeitar a hipótese nula")
else:
    print("Aceitar a hipótese nula")
```

-1.982844199727758  
0.05189282301496914  
Aceitar a hipótese nula



```
In [93]: fig, ax = plt.subplots(figsize=(10,8))
x_min = -6.0
x_max = 6.0
df=61
x = np.linspace(x_min, x_max, 100)
y = ss.t.pdf(x,df)
plt.plot(x,y, color='black')
pt1 = -Tcritico
pt2 = Tcritico
ptx = np.linspace(pt1, pt2, 100)
pty = ss.t.pdf(ptx, df,mean,std)
plt.fill_between(ptx, pty, color='#89bedc', alpha='1.0')
plt.grid()
plt.xlim(-4,4)
plt.ylim(0,0.5)
plt.title('Distribuicao t Student',fontsize=18)
plt.axvline(x=T, color='r', linestyle='--')
plt.show()
```





2) Repita o exemplo anterior, supondo que a variância populacional  $\sigma^2$  seja conhecida e igual a 153,9183 (m<sup>3</sup>/s).

Solução: A premissa básica continua sendo a de que as vazões médias do mês de Julho, em Ponte Nova do Paraopeba, seguem uma distribuição Normal. O fato de que a variância populacional é conhecida altera a estatística de teste.

Nesse caso, trata-se de um teste bilateral ao nível  $\alpha = 5\%$ , com a estatística de teste dada por:

$$Z = \frac{\bar{X} - 47,65}{\frac{\sigma}{\sqrt{N}}}$$

a qual possui uma distribuição  $N(0,1)$ .

Substituindo os valores amostrais, resulta que o valor absoluto da estimativa de Z é igual a 1,9828. A tabela normal, fornece  $Z_{0,975} = 1,96$

Como  $1,9828 > 1,96$ , a hipótese  $H_0$  deve ser rejeitada, em favor de  $H_1$ .

Portanto, sob as condições estipuladas para esse caso, é significativa a diferença entre a média amostral  $\bar{X} = 44,526$  e a média hipotética  $\mu = 47,65$ .

```
In [94]: paraopeba= pd.read_excel('VMM_ParaopebaAnoCivil.xlsx')
# paraopeba['Jul'].describe()
mediaA=paraopeba['Jul'].mean()
desvio=np.sqrt(153.9183)
N=paraopeba['Jul'].count()
media = 47.65
Z=abs((mediaA-media)/(desvio/np.sqrt(N)))
Z
```

Out[94]: 1.9828444541747863

```
In [95]: Tcritico = ss.norm.ppf(0.975, loc=0, scale=1)
print Tcritico
```

1.959963984540054

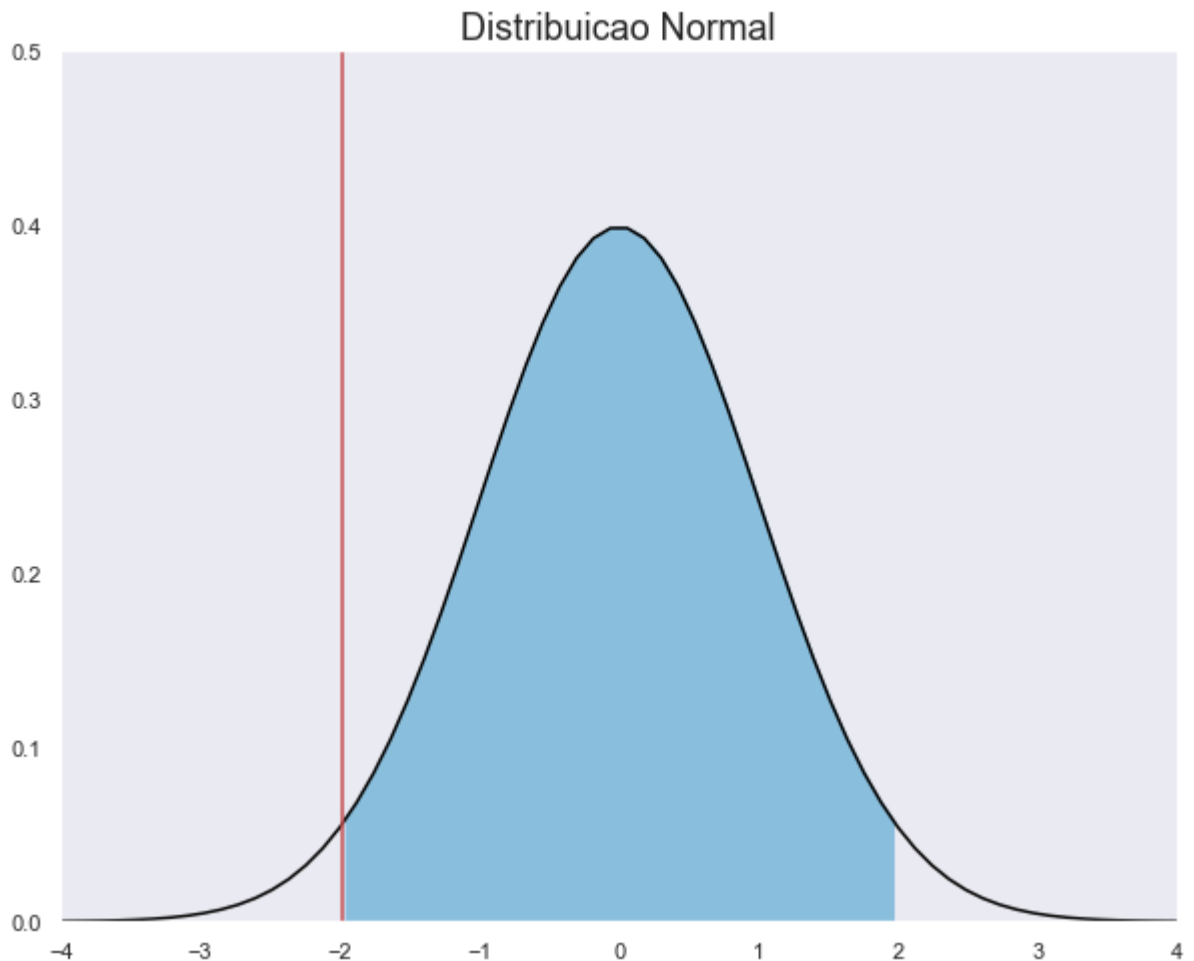
```
In [96]: from statsmodels.stats import weightstats
ztest, Pvalor = weightstats.ztest(x1=paraopeba['Jul'],value=[media])
print ztest[0]
print Pvalor[0]
if Pvalor<0.05:
    print("Rejeitar a hipótese nula")
else:
    print("Aceitar a hipótese nula")
```

-1.982844199727758

0.047384835165471685

Rejeitar a hipótese nula

```
In [97]: fig, ax = plt.subplots(figsize=(10,8))
x_min = -6.0
x_max = 6.0
df=61
x = np.linspace(x_min, x_max, 100)
y = ss.norm.pdf(x)
plt.plot(x,y, color='black')
pt1 = -Tcritico
pt2 = Tcritico
ptx = np.linspace(pt1, pt2, 100)
pty = ss.norm.pdf(ptx)
plt.fill_between(ptx, pty, color='#89bedc', alpha='1.0')
plt.grid()
plt.xlim(-4,4)
plt.ylim(0,0.5)
plt.title('Distribuicao Normal',fontsize=18)
plt.axvline(x=ztest[0], color='r', linestyle='--')
plt.show()
```



## Testes Paramétricos sobre as Médias de Duas Populações Normais

A premissa básica dos testes, descritos a seguir, é a de que as variáveis aleatórias independentes  $\{X_1, X_2, \dots, X_N\}$  e  $\{Y_1, Y_2, \dots, Y_M\}$ , componentes de duas amostras aleatórias simples de tamanhos iguais a  $N$  e  $M$ , foram extraídas de duas populações normais, de respectivas médias  $\mu_X$  e  $\mu_Y$ , desconhecidas. O conhecimento ou o desconhecimento das variâncias populacionais  $\sigma_X^2$  e  $\sigma_Y^2$ , assim como a condição de igualdade entre elas, determinam a estatística de teste a ser usada. Os testes descritos a seguir são tomados como bilaterais, podendo ser transformados em unilaterais pela modificação de  $H_1$  e de  $\alpha$ .

•  **$H_0: \mu_X - \mu_Y = \delta$  contra  $H_1: \mu_X - \mu_Y \neq \delta$ .**

Atributo de  $\sigma_X^2$  e  $\sigma_Y^2$ : conhecidas.

Estatística de teste:

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta}{\sqrt{\frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M}}}$$

Distribuição de probabilidades da estatística de teste: Normal  $N(0,1)$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão:

Rejeitar  $H_0$  se

$$\left| \frac{(\bar{X} - \bar{Y}) - \delta}{\sqrt{\frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M}}} \right| > Z_{1-\frac{\alpha}{2}}$$

•  **$H_0: \mu_X - \mu_Y = \delta$  contra  $H_1: \mu_X - \mu_Y \neq \delta$ .**

Atributo de  $\sigma^2_X$  e  $\sigma^2_Y$ : supostamente iguais, mas desconhecidas. Estimadas por  $S^2_X$  e  $S^2_Y$ .

Estatística de teste:

$$T = \frac{(\bar{X} - \bar{Y}) - \delta}{\sqrt{(N-1) \times S^2_X + (M-1) \times S^2_Y}} \times \sqrt{\frac{N \times M \times (N+M-2)}{N+M}}$$

Distribuição de probabilidades da estatística de teste: t de Student com  $\nu = N + M - 2$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão:

Rejeitar  $H_0$  se

$$\left| \frac{(\bar{X} - \bar{Y}) - \delta}{\sqrt{(N-1) \times S^2_X + (M-1) \times S^2_Y}} \times \sqrt{\frac{N \times M \times (N+M-2)}{N+M}} \right| >$$

•  **$H_0: \mu_X - \mu_Y = \delta$  contra  $H_1: \mu_X - \mu_Y \neq \delta$ .**

Atributo de  $\sigma^2_X$  e  $\sigma^2_Y$ : supostamente desiguais, mas desconhecidas. Estimadas por  $S^2_X$  e  $S^2_Y$ .

Estatística de teste:

$$T = \frac{(\bar{X} - \bar{Y}) - \delta}{\sqrt{\left(\frac{S^2_X}{N}\right) + \left(\frac{S^2_Y}{M}\right)}}$$

Distribuição de probabilidades da estatística de teste: segundo Casella e Berger (1990), a distribuição de T pode ser aproximada por uma distribuição t de Student com

$$\nu = \frac{\left[\left(\frac{S^2_X}{N}\right) + \left(\frac{S^2_Y}{M}\right)\right]^2}{\left[\frac{\left(\frac{S^2_X}{N}\right)^2}{N-1} + \frac{\left(\frac{S^2_Y}{M}\right)^2}{M-1}\right]}$$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão:

Rejeitar  $H_0$  se

$$\left| \frac{(\bar{X} - \bar{Y}) - \delta}{\sqrt{\left(\frac{S^2_X}{N}\right) + \left(\frac{S^2_Y}{M}\right)}} \right| > t_{1-\frac{\alpha}{2}, \nu}$$

Exemplo

3) Considere as vazões médias do mês de Julho do Rio Paraopeba em Ponte Nova do Paraopeba, listadas no Anexo 1, separando-as em duas amostras iguais de mesmo tamanho: a amostra denotada por X, para o período de 1938 a 1968, e a amostra Y, para o período de 1969 a 1999. Teste a hipótese de que, considerados os períodos de 1938-1968 e de 1969-1999, as médias populacionais do mês de Julho não sofreram alterações importantes, a um nível de significância  $\alpha = 5\%$ .

Solução: A premissa básica é a que, considerados os períodos de 1938- 1968 e de 1969-1999, as vazões médias do mês de Julho, em Ponte Nova do Paraopeba, seguem duas distribuições normais de médias  $\mu_X$  e  $\mu_Y$ , com variâncias  $\sigma_X$  e  $\sigma_Y$  supostamente desiguais e desconhecidas.

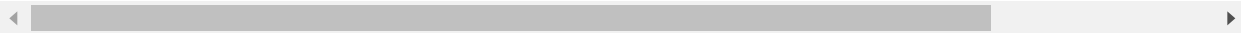
A amostra de 31 observações, para o período de 1938 a 1968, fornece  $\bar{X} = 45,08m^3/s$  e  $\sigma_X = 11,505m^3/s$ , enquanto, para o período restante, esses valores resultam ser  $\bar{Y} = 43,97m^3/s$  e  $\sigma_Y = 13,415m^3/s$ . Nesse caso, a hipótese nula é  $H_0 : \mu_X - \mu_Y = \delta = 0$  contra a hipótese alternativa  $H_1 : \mu_X - \mu_Y = \delta \neq 0$ . Como as variâncias são supostamente desiguais e devem ser estimadas pelas variâncias amostrais, a estatística de teste é

$$T = \frac{(47.08 - 43.97)}{\sqrt{(\frac{11.50^2}{31}) + (\frac{13.41^2}{31})}}$$

a distribuição de probabilidades da qual pode ser aproximada por uma t de Student com

$$v = \frac{[(11.50^2/31) + (13.41^2/31)]^2}{[\frac{(11.50^2/31)^2}{31-1} + \frac{(13.41^2/31)^2}{31-1}]}$$

que corresponde a 58 graus de liberdade. Substituindo os valores amostrais, resulta que o valor absoluto da estimativa de T é igual a 0,3476. A tabela de t de Student, fornece  $t_{0.975,v=58} = 2.00$ . Como  $0,3476 < 2,00$ , a hipótese  $H_0$  não deve ser rejeitada, em favor de  $H_1$ . Em outras palavras, com base nas amostras disponíveis, não há evidências de que as médias populacionais, dos períodos considerados, difiram significativamente entre si, ao nível de  $\alpha = 5\%$ .



```
In [98]: paraopeba = pd.read_excel('VMM_ParaopebaAnoCivil.xlsx')
X = paraopeba[paraopeba['Ano']<=1968]['Jul']
Y = paraopeba[paraopeba['Ano']>1968]['Jul']
mediaX = X.mean()
mediaY = Y.mean()
Sx = X.std()
Sy = Y.std()
N = X.count()
M = Y.count()
print mediaX, Sx, N
print mediaY, Sy, M
T = abs((mediaX-mediaY)/np.sqrt((Sx**2/N)+(Sy**2/M)))
gl = (((Sx**2/N)+(Sy**2/M))**2)/((((Sx**2/N)**2)/(N-1))+(((Sy**2/M)**2)/(M-1)))
print T, gl
Tcritico = ss.t.ppf(0.975, gl, loc=0, scale=1)
print Tcritico
```

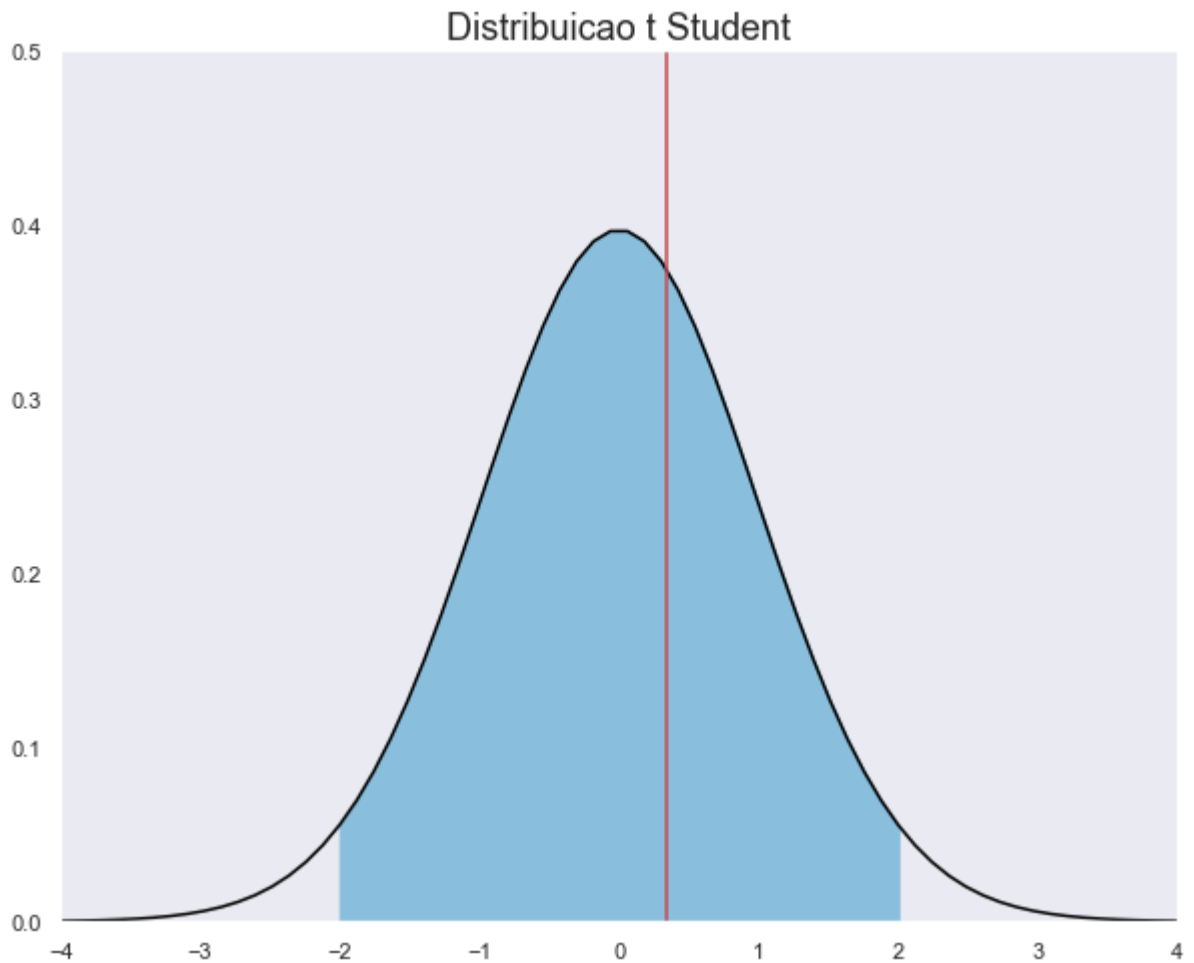
```
45.0774193548387 11.50500498847985 31
43.97419354838709 13.415413218680854 31
0.3475621960654423 58.637650097454554
2.001254128676434
```

```
In [99]: Tteste, Pvalor = ss.ttest_ind(X, Y)
print Tteste
print Pvalor
if Pvalor<0.05:
    print("Rejeitar a hipótese nula")
else:
    print("Aceitar a hipótese nula")
```

```
0.3475621960654423
0.7293853104275272
Aceitar a hipótese nula
```



```
In [100]: fig, ax = plt.subplots(figsize=(10,8))
x_min = -6.0
x_max = 6.0
df=gl
x = np.linspace(x_min, x_max, 100)
y = ss.t.pdf(x,df)
plt.plot(x,y, color='black')
pt1 = -Tcritico
pt2 = Tcritico
ptx = np.linspace(pt1, pt2, 100)
pty = ss.t.pdf(ptx, df,mean,std)
plt.fill_between(ptx, pty, color='#89bedc', alpha='1.0')
plt.grid()
plt.xlim(-4,4)
plt.ylim(0,0.5)
plt.title('Distribuicao t Student',fontsize=18)
plt.axvline(x=Tteste, color='r', linestyle='-')
plt.show()
```



## Testes de hipóteses para diferenças das variâncias

### Testes Paramétricos sobre a Variância de uma Única População Normal

A premissa básica dos testes, descritos a seguir, é a de que as variáveis aleatórias independentes  $\{X_1, X_2, \dots, X_N\}$ , componentes de uma certa amostra aleatória simples, foram todas extraídas de uma única população normal, de variância  $\sigma^2$  desconhecida. O conhecimento ou o desconhecimento da média populacional  $\mu$  determina a estatística de teste a ser usada. Os testes são tomados como bilaterais, podendo ser transformados em unilaterais pela modificação de  $H_1$  e de  $\alpha$ .

- $H_0 : \sigma^2 = \sigma_0^2$  contra  $H_1 : \sigma^2 \neq \sigma_0^2$ .

Atributo de  $\mu$ : conhecida.

Estatística de teste:

$$Q = \frac{\sum_{i=1}^N (X_i - \mu)^2}{\sigma_0^2} = N \frac{S_0^2}{\sigma_0^2}$$

Distribuição de probabilidades da estatística de teste:  $\chi^2$  com  $\nu = N$ , ou  $\chi_N^2$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão:

Rejeitar  $H_0$  se

$$N \frac{S_0^2}{\sigma_0^2} < \chi_{\frac{\alpha}{2}, N}^2 \text{ ou } N \frac{S_0^2}{\sigma_0^2} > \chi_{1-\frac{\alpha}{2}, N}^2$$

- $H_0 : \sigma^2 = \sigma_0^2$  contra  $H_1 : \sigma^2 \neq \sigma_0^2$ .

Atributo de  $\mu$ : desconhecida, estimada por  $\bar{X}$ .

Estatística de teste:

$$K = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{\sigma_0^2} = (N - 1) \frac{S_0^2}{\sigma_0^2}$$

Distribuição de probabilidades da estatística de teste:  $\chi^2$  com  $\nu = N - 1$ , ou  $\chi_{N-1}^2$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão:

Rejeitar  $H_0$  se

$$(N - 1) \frac{S_0^2}{\sigma_0^2} < \chi_{\frac{\alpha}{2}, N-1}^2 \text{ ou } (N - 1) \frac{S_0^2}{\sigma_0^2} > \chi_{1-\frac{\alpha}{2}, N-1}^2$$

Exemplo

4) Considere novamente as vazões médias do mês de Julho do Rio Paraopeba em Ponte Nova do Paraopeba, para o período de 1938 a 1999. Teste a hipótese nula de que a variância populacional  $\sigma_0^2$ , das vazões médias do mês de Julho, é de  $150 (m^3/s)^2$  contra a hipótese alternativa  $H_1$ :  $\sigma_0^2 > 150(m^3/s)^2$ , a um nível de significância  $\alpha = 5\%$ .

Solução: Novamente, a premissa básica é a que as vazões médias do mês de Julho, em Ponte Nova do Paraopeba, seguem uma distribuição Normal.

A amostra de 62 observações fornece  $\bar{X} = 44.526$ , e  $S_X = 12.406 m^3/s$ , não havendo nenhuma informação adicional sobre a média populacional.

Nesse caso, a hipótese nula é  $H_0 : \sigma_0^2 = 150$  contra a hipótese alternativa  $H_1 : \sigma_0^2 > 150$ .

Trata-se, portanto, de um teste unilateral ao nível  $\alpha = 5\%$ , com a estatística de teste dada por

$K = (N - 1) \frac{S_X^2}{\sigma_0^2}$ , a qual possui uma distribuição  $\chi^2$  com 61 graus de liberdade. Substituindo os

valores amostrais, resulta que o valor de K é igual a 62,593. A tabela de  $\chi^2$ , fornece

$\chi_{0.95, 61}^2 = 80.232$ . Como  $62,593 < 80,232$ , a hipótese  $H_0$  não deve ser rejeitada, em favor de  $H_1$ .

Em outras palavras, com base na amostra disponível, não há evidências de que a variância populacional supere significativamente o valor de  $150 (m^3/s)^2$ , ou seja, que a diferença existente entre a variância amostral  $S_X^2 = 153,918$  e a variância  $\sigma_0^2 = 150$  deve-se unicamente a flutuações aleatórias das observações.

```
In [101]: paraopeba= pd.read_excel('VMM_ParaopebaAnoCivil.xlsx')
mediaA=paraopeba['Jul'].mean()
desvioA=paraopeba['Jul'].std()
N=paraopeba['Jul'].count()
gl=N-1
variancaP = 150
print mediaA, desvioA, N
K = (N-1)*(desvioA**2/variancaP)
print K
Tcritico = ss.chi2.ppf(0.95,gl)
print Tcritico
pValor = 1 - ss.chi2.cdf(x=K, df=gl)
print pValor

from scipy.stats.distributions import chi2
chi2.sf(K,gl)
```

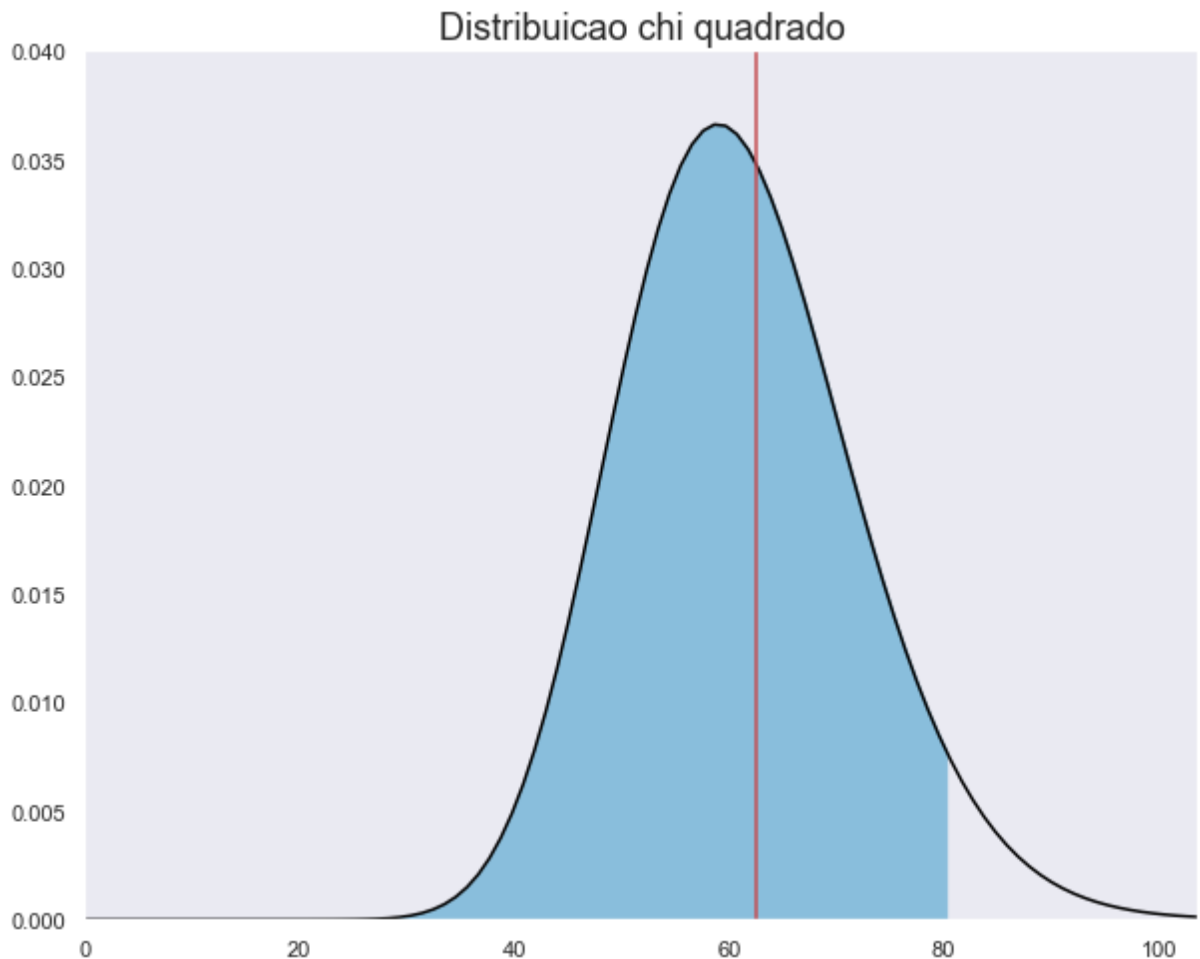
```
44.52580645161289 12.406383014517509 62
62.593458064516156
80.23209784876272
0.41944959772328483
```

```
Out[101]: 0.4194495977232848
```

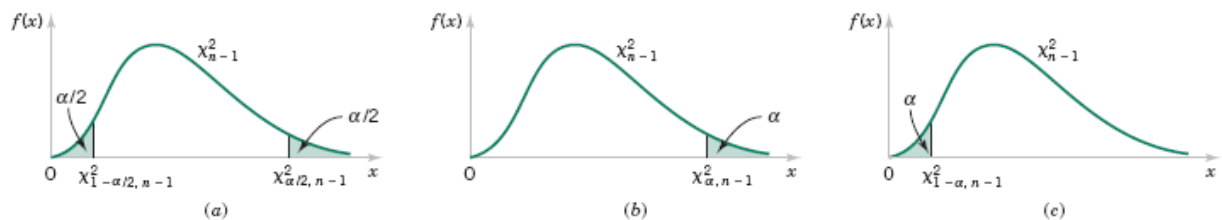
```

In [102]: fig, ax = plt.subplots(figsize=(10,8))
df=N-1
x_min = 0
x_max = 1.7*df
x = np.linspace(x_min, x_max, 100)
y = ss.chi2.pdf(x,df)
plt.plot(x,y, color='black')
pt1 = 0
pt2 = Tcritico
ptx = np.linspace(pt1, pt2, 100)
pty = ss.chi2.pdf(ptx, df)
plt.fill_between(ptx, pty, color='#89bedc', alpha='1.0')
plt.grid()
plt.xlim(x_min,x_max)
plt.ylim(0,0.04)
plt.title('Distribuicao chi quadrado',fontsize=18)
plt.axvline(x=K, color='r', linestyle='--')
plt.show()

```



A distribuição de referência para o teste  $H_0 : \sigma^2 = \sigma_0^2$



com valores da região crítica para (a),  $H_1 : \sigma^2 \neq \sigma_0^2$  (b),  $H_1 : \sigma^2 > \sigma_0^2$  e (c)  $H_1 : \sigma^2 < \sigma_0^2$ .

## Testes Paramétricos sobre as Variâncias de Duas Populações Normais

A premissa básica dos testes, descritos a seguir, é a de que as variáveis aleatórias independentes  $\{X_1, X_2, \dots, X_N\}$  e  $\{Y_1, Y_2, \dots, Y_N\}$ , componentes de duas amostras aleatórias simples de tamanhos iguais a  $N$  e  $M$ , foram extraídas de duas populações normais, de respectivas variâncias  $\sigma_X^2$  e  $\sigma_Y^2$  desconhecidas. O conhecimento ou o desconhecimento das médias populacionais  $\mu_X$  e  $\mu_Y$  determina a estatística de teste a ser usada. Os testes são tomados como bilaterais, podendo ser transformados em unilaterais pela modificação de  $H_1$  e de  $\alpha$ .

$$\bullet H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1 \text{ contra } \frac{\sigma_X^2}{\sigma_Y^2} \neq 1$$

Atributos de  $\mu_X$  e  $\mu_Y$ : conhecidas

Estatística de teste:

$$\varphi = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

Distribuição de probabilidades da estatística de teste: F de Snedecor com  $\nu_1 = N$  e  $\nu_2 = M$ , ou  $F_{N,M}$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão:

Rejeitar  $H_0$  se  $\varphi < F_{N,M,\alpha/2}$  ou se  $\varphi > F_{N,M,1-\alpha/2}$

$$\bullet H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1 \text{ contra } \frac{\sigma_X^2}{\sigma_Y^2} \neq 1$$

Atributos de  $\mu_X$  e  $\mu_Y$ : desconhecidas, estimadas por  $\bar{X}$  e  $\bar{Y}$

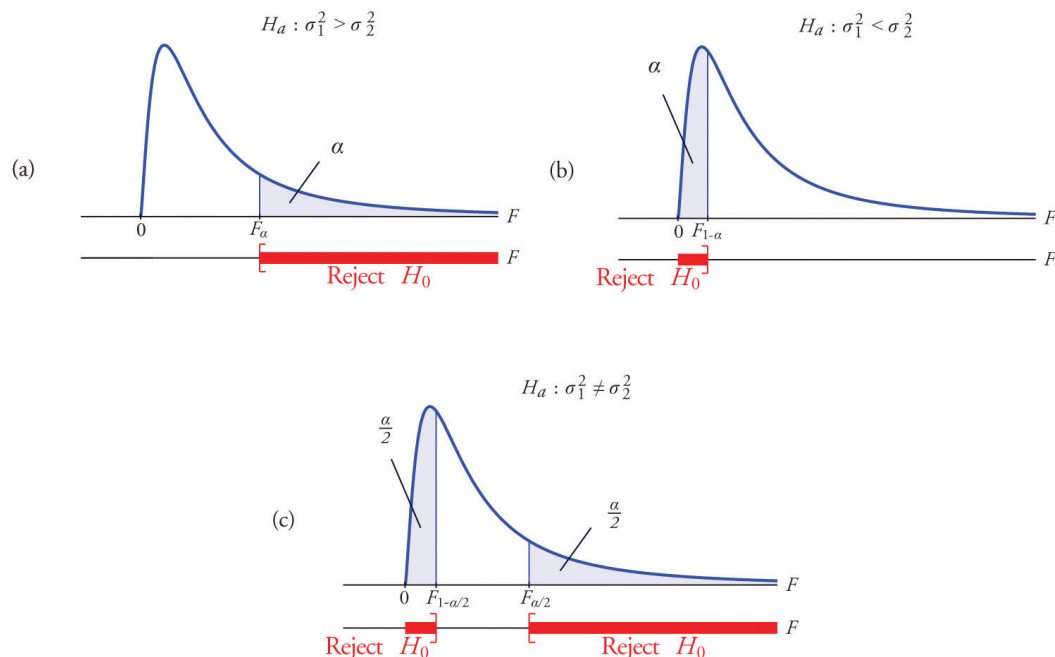
Estatística de teste:

$$f = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

Distribuição de probabilidades da estatística de teste: F de Snedecor com  $\nu_1 = N - 1$  e  $\nu_2 = M - 1$ , ou  $F_{N-1, M-1}$

Tipo de Teste: bilateral a um nível de significância  $\alpha$

Decisão: Rejeitar  $H_0$  se  $f < F_{N-1, M-1, \alpha/2}$  ou se  $f > F_{N-1, M-1, 1-\alpha/2}$



### Exemplo

5) – Um certo constituinte de um efluente foi analisado 7 e 9 vezes por meio dos procedimentos X e Y, respectivamente. Os resultados das análises apresentaram os seguintes desvios-padrão:  $S_X = 1,9$  e  $S_Y = 0,8$  mg/l. Teste a hipótese de que o procedimento Y é mais preciso do que o procedimento X, ao nível de significância  $\alpha = 5\%$ . (adap. de Kottegoda e Rosso, 1997)

Solução: Supondo tratar-se de duas populações normais, a hipótese nula a ser testada é

$H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$  contra a hipótese alternativa  $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1$  ou  $\sigma_X^2 > \sigma_Y^2$ .

Trata-se, portanto, de um teste unilateral com  $\alpha = 0,05$ . A estatística de teste é  $f = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ ,

a qual segue uma distribuição F de Snedecor com  $\nu = 7-1 = 6$  e  $\nu_2 = 9-1 = 8$  graus de liberdade para o numerador e denominador, respectivamente. Substituindo os valores amostrais, resulta que  $f = 5,64$ . Da tabela de F, lê-se que  $F_{6,8,0.05} = 3,58$ . Como  $5,64 > 3,58$ , a decisão é de rejeitar a hipótese nula em favor da hipótese alternativa, ao nível de significância  $\alpha = 0,05$ . Em outras palavras, conclui-se que a variância dos resultados do procedimento Y é menor do que a de seu concorrente, tratando-se, portanto, de um método de análise mais preciso.

```
In [109]: Sx = 1.9
          Sy = 0.8
          N = 7
          M = 9
          gl1 = N-1
          gl2 = M-1
          f = Sx**2/Sy**2
          print f
          fTeste = ss.f.isf(0.05, gl1, gl2)
          print fTeste
          pValor = 1-ss.f.cdf(f, gl1, gl2)
          print pValor
```

```
5.640625
3.5805803197614603
0.01439913972528184
```



```
In [115]: fig, ax = plt.subplots(figsize=(10,8))
x_min = 0
x_max = 1.7*gl1
x = np.linspace(x_min, x_max, 100)
y = ss.f.pdf(x,gl1, gl2)
plt.plot(x,y, color='black')
pt1 = 0
pt2 = fTeste
ptx = np.linspace(pt1, pt2, 100)
pty = ss.f.pdf(ptx, gl1, gl2)
plt.fill_between(ptx, pty, color='#89bedc', alpha='1.0')
plt.grid()
plt.xlim(x_min,x_max)
plt.ylim(0,0.7)
plt.title('Distribuicao Fisher',fontsize=18)
plt.axvline(x=f, color='r', linestyle='-')
plt.show()
```

