BASE

ANÁLISES EXPLORATÓRIAS DE DADOS

Neste tutorial, pretendemos instrumentalizar os(as) usuários(as) a realizar várias técnicas de Análise Exploratória de Dados (AED).

Objetivos da Análise Exploratória de Dados (AED)

Apesar de ter sido criada para minimizar problemas com as análises frequentistas, a AED é bastante versátil e também pode ser utilizada no contexto de outras abordagens analíticas.

Dentre os principais **objetivos** de uma AED podemos listar os seguintes:

- Detectar erros de entrada de dados;
- Detectar pontos extremos (*outliers*) e anomalias;
- Compreender a estrutura dos dados coletados;
- Avaliar premissas de testes que serão utilizados posteriormente;
- Avaliar preliminarmente se os dados apresentam dependência/autocorrelação (espacial ou temporal).
- Avaliar se variáveis dentro de um conjunto abrangente apresentam colinearidade;
- Avaliar se os dados se adequam aos modelos que serão utilizados nas análises posteriores;

Apesar de alguns autores considerarem aceitável olhar para as relações entre os dados brutos, criar hipóteses a partir dessas observações e testá-las com o mesmo conjunto de dados, esse procedimento (conhecido por *"data dredging"*) não é considerado adequado pela maioria dos pesquisadores.

O procedimento mais correto é estabelecer *a priori* suas hipóteses, **com base nos contextos teóricos da sua área de pesquisa** e estabelecer *a priori* suas análises, **com base nas características das variáveis que serão analisadas**.

Se o procedimento de *"data dredging"* for realizado em um conjunto de dados, ele não deve ser usado para testar as hipóteses geradas. Em um mundo ideal, um novo conjunto de dados deve ser coletado.

Preparação dos dados e programa

1) Crie um diretório (pasta) e copie os arquivos de dados abaixo para esse diretório:

- univar.csv
- autocorr.csv
- bivar.csv
- fluxo_ppt.csv

Conhecendo os dados:

Note que para variáveis numéricas (contínuas ou discretas) são apresentados os valores Mínimo, Máximo, Média, Mediana, Primeiro quartil, Terceiro quartil, e, no caso de haver dados faltantes, será apresentado o número de dados faltantes, representados como “NA's”. Para variáveis categóricas (por exemplo, para o conjunto de dados univar.csv, a coluna “NIVEL_DISTURBIO”) são apresentados os níveis existentes e quantas observações cada um dos níveis possui. Se houver dados faltantes, será apresentado o número de “NA's”.

Veja se você entendeu o conjunto de dados. Antes mesmo de fazer as análises gráficas, você consegue pensar em como esses dados podem estar distribuídos?

ANÁLISES GRÁFICAS

Salve todos os gráficos que você criar a partir de agora

1) Histograma de frequência:

histograma

Inspecionando o seu Histograma

- O que está representado no eixo X e no eixo Y de cada um desses gráficos?
- Quais são os valores que delimitam as classes usadas no eixo X desse histograma?
- O número de classes parece adequado?

Vamos mudar então o número de classes no eixo X e ver como ficam os gráficos.

densidade

2) Gráfico de densidade

Ao invés de usarmos classes, podemos representar a distribuição por meio de uma linha, que é obtida usando a densidade estimada (por uma função conhecida como *kernel*) de valores para “janelas”

(*bandwidth*) muito pequenas. Vamos ver como ficam as distribuições das mesmas variáveis para as quais fizemos os histogramas.

boxplot

3) Box-plot ou Box-whiskers plot ou Five-numbers-summary

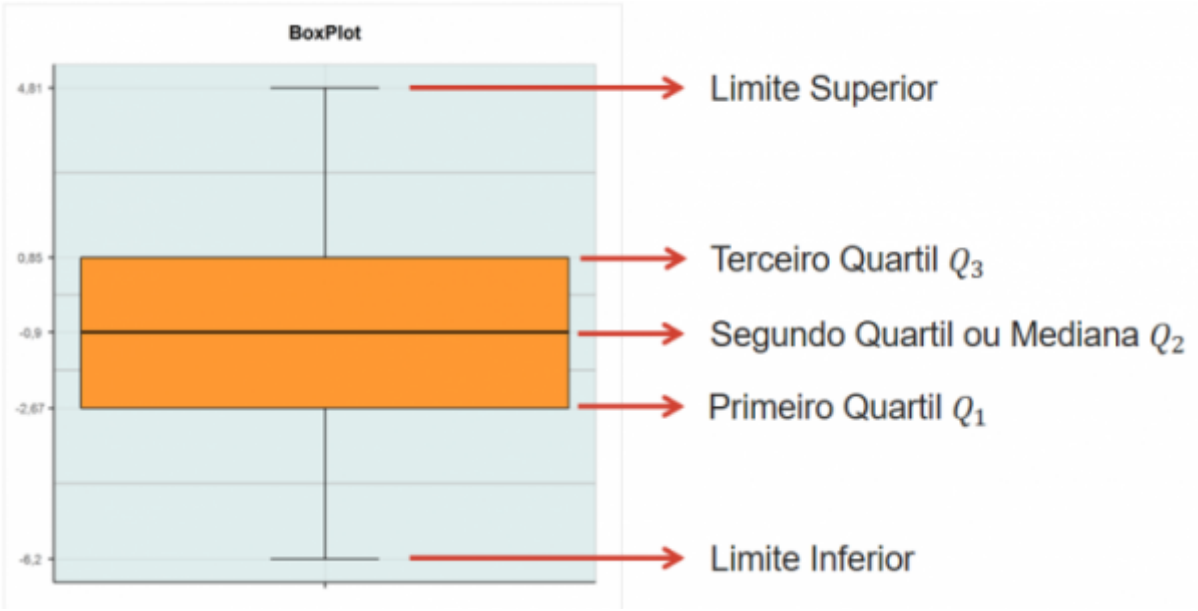
Um box-plot clássico utiliza os seguintes valores:

- - Mínimo
- - Primeiro quartil
- - Mediana
- - Terceiro quartil
- - Máximo

Todos esses dados estão no “Resumo dos dados”, que você obteve no lá no início, quando estava inspecionando os dados.

boxplot2

Com esses dados você poderia construir um box-plot simples, manualmente, conforme a figura abaixo:



outliers

Confira se os valores utilizados para a construção do box-plot são iguais aos que estavam no “Resumo dos dados”. Você percebe a diferença no Limite Superior?

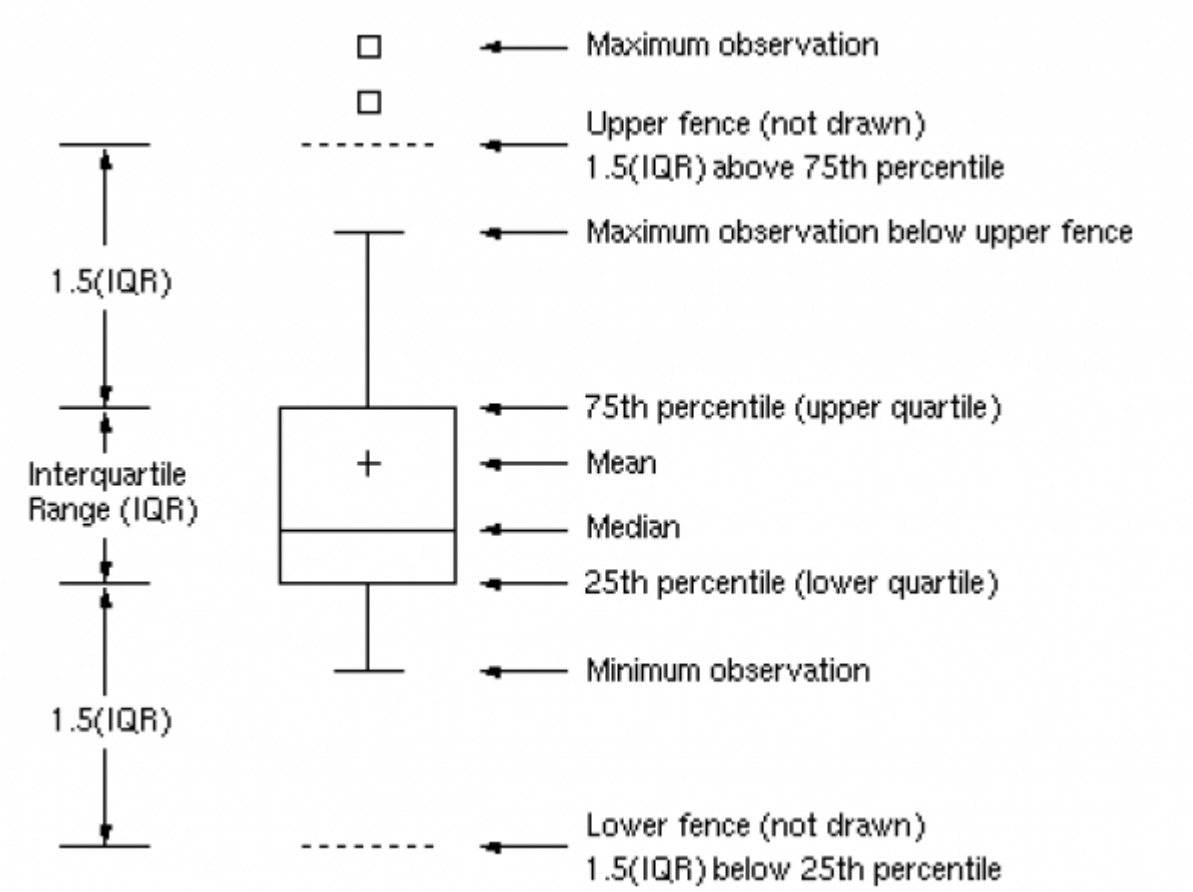
Muitas vezes, a opção padrão (*default*) de um programa estatístico faz um gráfico chamado **box-plot**

modificado. Esse **box-plot modificado** nos ajuda a identificar os pontos extremos que comumente chamamos de ***outliers***.

Ao invés de usarmos os valores de **máximo** e **mínimo** nas pontas das linhas verticais (tanto para cima quanto para baixo), usamos a equação *1.5*IQR* para definirmos o valor máximo (ou mínimo) que essa linha vertical **poderia** atingir (isso é o que chamamos de "fence").

- *IQR* é a distância entre o primeiro e o terceiro quartil (ou seja, a amplitude da caixa central do box-plot). Ou ainda, *IQR = Q3 - Q1*, onde Q3 é o valor do terceiro quartil e Q1 é o valor do primeiro quartil.

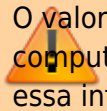
A aparência geral de um **box-plot modificado** é assim:



Exemplo de cálculo dos limites: Suponha uma situação em que sua variável resposta é medida em ml. Se Q3=30ml e Q1=20ml, então, IQR = 10ml e a linha vertical deve estar, no máximo, a 15ml (1.5*10ml) para cima ou para baixo a partir das bordas da caixa (quartis). Então, para se obter o valor do limite superior é preciso somar Q3 + 1.5*IQR [30 + (1.5*10) = 45ml] e para se obter o valor inferior do limite é preciso subtrair Q1 - I1.5*QR [20- (1.5*10) = 5 ml]. Nesse caso, os valores que estiverem aproximadamente abaixo de 5ml e acima de 45 ml serão considerados *outliers*.

Calcule o valor de IQR e os limites superior e inferior para a variável COMPRIMENTO_BICO, usando os valores obtidos inicialmente no "Resumo dos dados".

Existe ainda uma outra complicação para estabelecer os limites reais. Caso os valores exatos calculados pelas equações acima para os limites superiores e inferiores não existam no conjunto de dados, o que definirá o limite **real** da linha vertical superior será o valor mais alto existente no conjunto de dados, dentro do limite estabelecido. E, para baixo, o limite **real** da linha vertical inferior será o valor mais baixo dentro do limite estabelecido. Para encontrarmos esses valores reais precisamos ordenar os dados e buscar os valores reais mais próximos dos valores calculados conforme indicado acima. Nesse momento da atividade não vamos fazer isso, mas é importante deixar claro que, para que um valor seja considerado um *outlier*, ele deverá estar acima ou abaixo desses limites **reais**.



O valor a ser multiplicado por IQR pode variar de um autor para outro e de um programa computacional para outro, então é muito importante que as legendas dos gráficos tragam essa informação. Infelizmente, essa não é uma prática comum.

boxplot3

Várias informações podem ser obtidas a partir de um box-plot:

- Existem *outliers* no conjunto de dados?
- Eles estão entre os valores mais altos ou mais baixos?
- A distribuição dos dados é simétrica ou assimétrica?
- Se for assimétrica, os dados estão concentrados[1)] em valores acima da mediana (gerando uma distribuição assimétrica com cauda grande para a esquerda - *left-skewed*) ou abaixo da mediana (gerando uma distribuição assimétrica com cauda grande para a direita - *right-skewed*)?

A maioria dos programas produz boxplots mesmo se o seu conjunto de dados tiver menos de 5 valores. Porém, isso faz sentido? O que significariam os valores em cada ponto do boxplot?

boxplot4

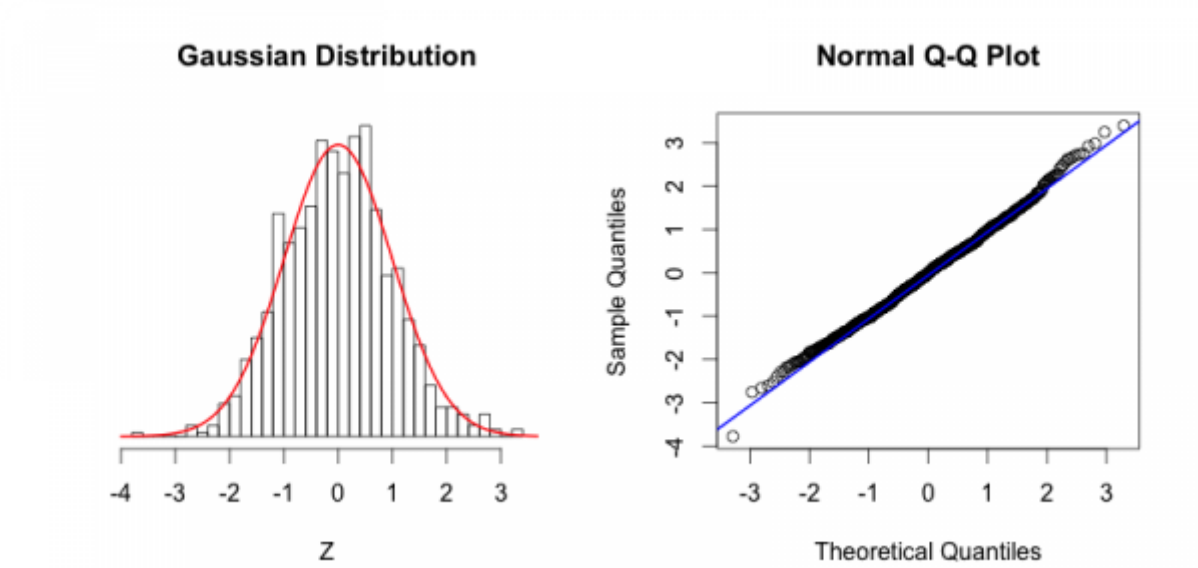
Todas essas informações nos ajudam a entender a distribuição dos nossos dados.

Para explorar grandes conjuntos de dados, existem algumas outras formas interessantes de gráficos. Veja aqui no blog da Melina Leite.

CHECANDO O AJUSTE DOS DADOS A UMA DISTRIBUIÇÃO

Agora vamos avaliar visualmente se uma variável se distribui de acordo com uma distribuição conhecida. Essa avaliação tem pouca utilidade para analisar dados brutos, porém será muito útil para avaliar os resíduos de uma análise. Então, optamos por ensiná-la usando dados brutos, somente para simplificar o entendimento.

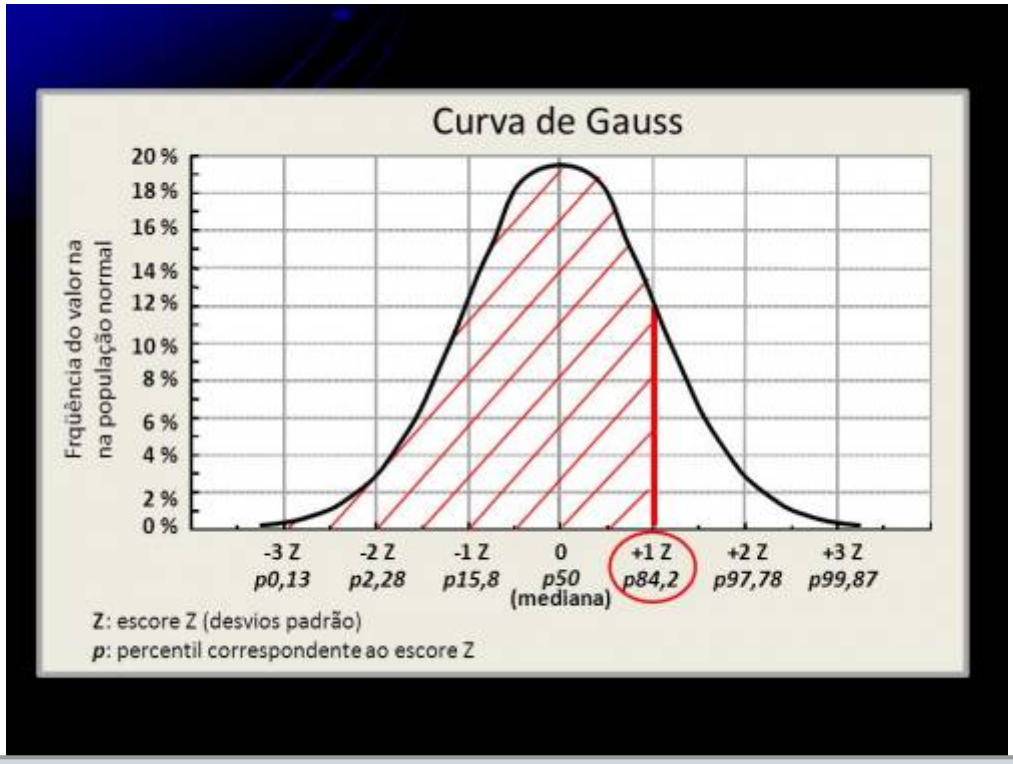
Gráfico quantil-quantil



A ideia desses gráficos do tipo **quantil-quantil** é expressar visualmente o quanto um conjunto de dados se aproxima de uma determinada distribuição. Eles podem ser usados para comparar as distribuições de dois conjuntos de dados diferentes (para saber se ambos vêm de uma mesma população) ou para comparar a distribuição de um conjunto de dados coletados com uma distribuição de probabilidade teórica conhecida (normal, binomial, etc). Nesse segundo caso, eles são também chamados de Gráficos de Probabilidade (*Probability Plots*). No exemplo abaixo, vamos comparar os dados coletados a uma **distribuição normal**.

Um **quantil** divide dados ordenados (do menor para o maior) em subconjuntos que têm dimensões iguais (mesmo número de observações em cada quantil). Os **quartis** calculados acima no **Resumo** dos dados são casos especiais de quantis. São chamados de quartis, pois dividem o conjunto de dados em 4 grupos (até 25%, 25-50%, 50-75% e 75-100%) com o mesmo número de dados em cada divisão. Os **percentis** são casos especiais de quantis que dividem o conjunto em 100 grupos.

Para facilitar a comparação dos valores é usada uma **distribuição normal padronizada**, que tem média=0 e desvio-padrão=1. Especificamente para a **distribuição normal padronizada**, quando o desvio é igual a +1 (ou seja, 1 desvio padrão acima da média), estamos no percentil de 84,2% (50% da média [2)] + 34,2% do desvio correspondente) e quando o desvio-padrão é igual a +2 estamos no percentil 97.8% (50% da média + 47.8% acumulando até o desvio-padrão = 2). Veja a figura abaixo:



Você já deve ter percebido que o valor do percentil 50% é a mediana, que é exatamente o segundo quartil (Q2) que é apresentado na tabela de **Resumo** dos dados (que aprendemos a obter no início dessa aula) e que é usado para construir a linha central de um boxplot.

Observação importante: Em uma distribuição normal, a mediana é exatamente igual à média!

Porém, para seguir adiante, esteja certo que você também compreende que, por exemplo, o valor que representa o percentil de 75% é o mesmo que o valor do terceiro quartil (Q3) usado para construir a parte de cima da caixa central de um boxplot. Esse valor do terceiro quartil-Q3 ou do percentil 75% também está disponível naquela tabela inicial de **Resumo** dos dados, conforme vimos em uma atividade anterior.

qqplot2

Uma vez compreendida a forma como esse gráfico foi construído, como os resultados devem ser interpretados?

Para cada uma das variáveis avalie:
a) É possível visualizar *outliers* nas variáveis?;
b) Os dados se ajustam bem à distribuição normal?;
c) Nos casos em que a distribuição for assimétrica, os dados estão concentrados em torno dos valores mais baixos ou mais altos

?}}preservefilenames::QQPlot_CaudaLonga_CaudaCurta.jpg

Existe uma página muito bacana que mostra algumas distribuições simuladas e como ficam os QQplots dessas distribuições. Veja aqui

IMPORTANTE: Vários tipos de análises têm a normalidade como premissa. Porém, é importante não confundir a normalidade dos dados brutos, com a normalidade dos erros, da variância ou dos resíduos das relações.

No tópico **Testes clássicos frequentistas** usamos essa análise gráfica para avaliarmos a normalidade dos resíduos da regressão linear.

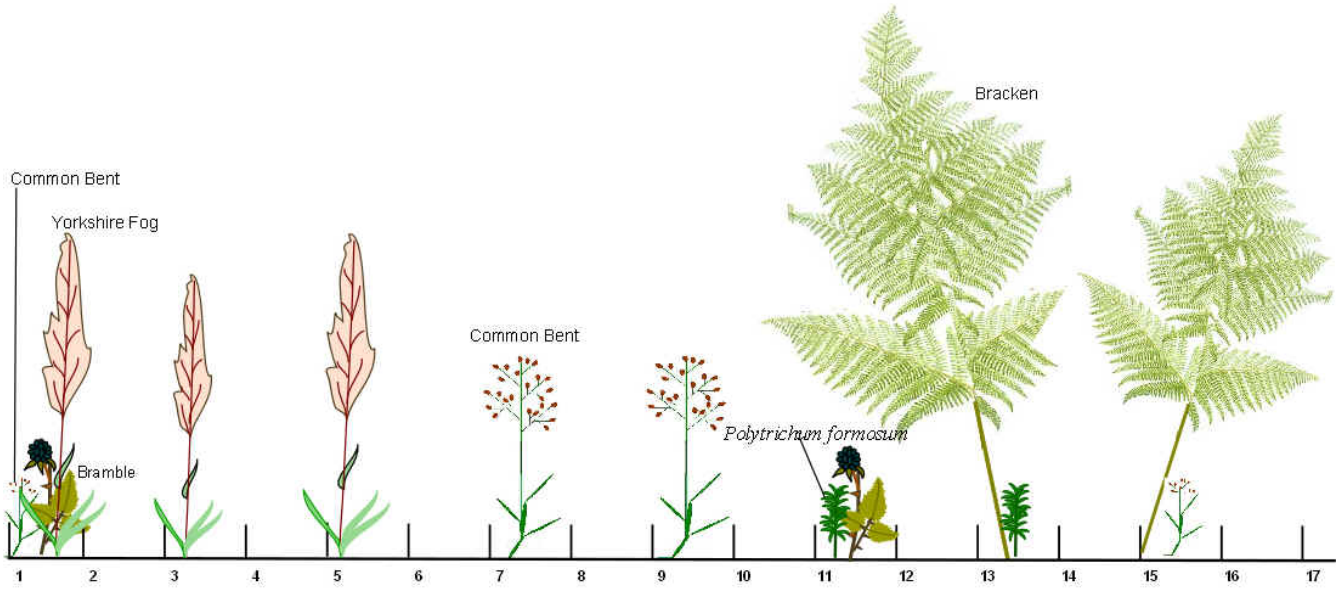
Entretanto, nessa atividade, em que analisamos a normalidade de uma variável isoladamente, estamos apenas compreendendo e descrevendo o tipo de dado que coletamos.

AVALIANDO AUTOCORRELAÇÃO

Dentre as premissas mais importantes dos testes estatísticos está a **independência** (espacial e/ou temporal) dos dados coletados. Existem diversas formas de avaliar o nível de autocorrelação entre os dados. Quando estamos lidando com dados distribuídos em apenas uma dimensão (transectos lineares ou series temporais direcionais), esse processo é mais simples. Porém, quando os dados estão distribuídos em duas dimensões (p.ex. posições x e y em uma parcela) ou em três dimensões (posições x e y e mais a profundidade, no caso de medidas em sistemas aquáticos) os métodos são mais complexos e fogem ao escopo desse tutorial. Se tiver interesse em entender alguns desses métodos para duas dimensões visite Padrões Multiescala

Porém, existe uma forma simples de visualizar os dados e obter uma primeira impressão sobre possíveis autocorrelações para dados coletados em transectos lineares e também para dados de séries temporais.

Imagine o transecto abaixo, no qual os números na linha inferior representam os locais de coleta e cuja informação coletada é o tamanho das folhas, que vai permitir o cálculo da variável “tamanho médio das folhas” das plantas presentes em cada ponto:



Considerando que as espécies de plantas, em geral, têm distribuição espacial agregada, você poderia se perguntar se os dados mais próximos espacialmente são mais parecidos entre si (i.e. positivamente autocorrelacionados) em função de uma maior similaridade na composição de espécies. Uma forma de avaliar isso é plotar o valor de um dado em relação ao seu antecessor, então, no eixo X teríamos os valores do segundo dado em diante e no eixo Y teríamos, correspondendo a cada valor do eixo X, o valor do dado anterior. Esse tipo de gráfico é chamado de "*lag-plot*"

O gráfico do tipo *lag-plot* apresenta a relação entre um determinado dado e o seu antecessor (temporal ou espacial) quando os dados são tomados em uma sequência unidimensional.

O que você esperaria que acontecesse em um gráfico desse tipo se os valores estiverem autocorrelacionados? E se não estiverem?

Vejamos como ficam esses gráficos para os dados do conjunto "autocorr.csv" que temos disponível para essa análise. Nesse arquivo temos os dados de dois transectos (x1 e x2) com 100 pontos cada.

autocorr2

Olhando para esses resultados, qual a sua conclusão?

No gráfico padrão (*default*) produzido por essa função *lag.plot()* estamos relacionando um determinado dado com o seu antecessor **imediato**, ou seja, o antecessor está a 1 unidade de distância em relação ao dado que colocamos no eixo X. Porém, alguns processos podem ocorrer em escalas diferentes de 1 unidade de distância e podemos querer checar se existe autocorrelação em

outras escalas. Para isso, usamos o argumentos "*lags*" e "*set.lags*" dentro dessa função.

Vamos ver como ficam os gráficos com lags=2 (ou seja, duas unidades de distância).

Primeiro para o transecto 1:

```
lag.plot(autocorr$x1, do.lines = FALSE, lags=2, set.lags=2, layout= c(1,1),
diag=FALSE)
```

Depois para o transecto 2:

```
lag.plot(autocorr$x1, do.lines = FALSE, lags=2, set.lags=2, layout= c(1,1),
diag=FALSE)
```

E agora, olhando para esses resultados, qual a sua conclusão?

ANALISANDO DADOS BIVARIADOS

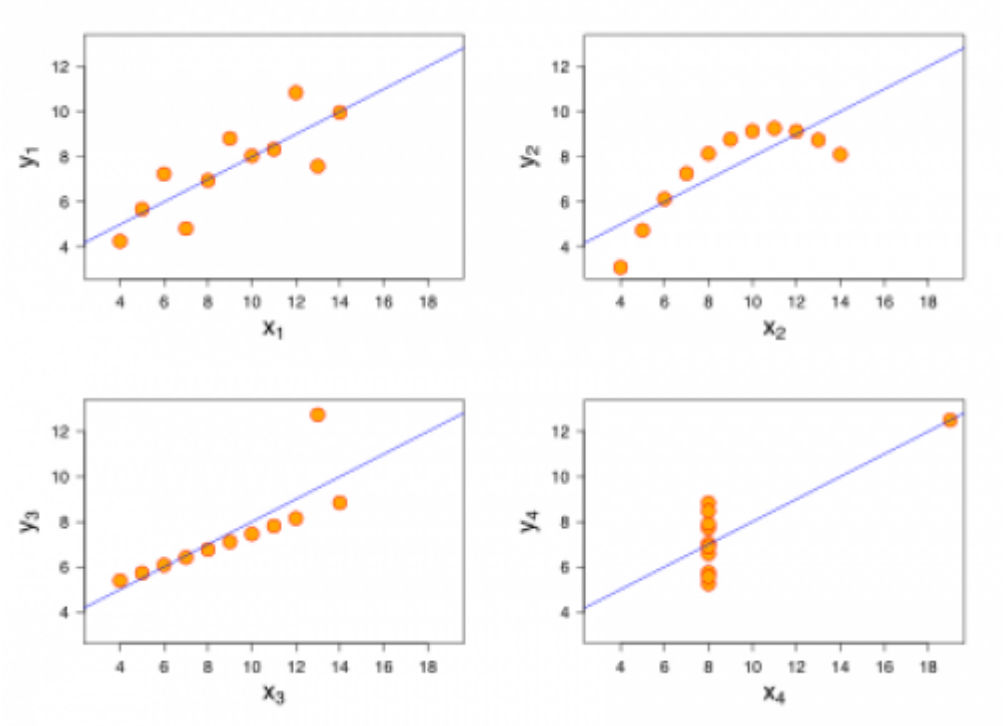
Muitas vezes não estamos interessados em compreender a distribuição de cada variável individualmente, mas sim, em avaliar se existe alguma relação entre duas ou mais variáveis e qual a forma dessa relação.

A significância (ou plausibilidade) dessa relação deve ser avaliada por testes estatísticos apropriados, definidos *a priori* e a partir das hipóteses estabelecidas *a priori*, mas é bastante recomendado que seja feita uma inspeção visual da distribuição das variáveis, da relação entre elas e da forma das relações.

Infelizmente, ainda é bastante comum que a interpretação de uma relação esteja apoiada apenas em alguns poucos "*números sintéticos*" (*p*, $$R^2$$, média, desvio) gerados a partir das análises estatísticas. Entretanto, esses números sozinhos podem não descrever adequadamente as relações observadas.

Em 1973, F. J. Anscombe publicou um conjunto de dados com o objetivo de indicar os problemas relacionados à interpretação desses números. O autor simulou conjuntos de dados com valores muito similares para os números sintéticos, mas cujos dados estavam distribuídos de formas muito diferentes. Essa publicação ressaltou a importância da inspeção visual dos dados antes de se proceder às análises estatísticas.

Veja abaixo a distribuição dos dados nos conjuntos apresentados por Anscombe. Eles são chamados de **Quarteto de Anscombe**:



E esses são os valores sintéticos comuns a todos esses conjuntos de dados:

Vamos agora aprender como avaliar os principais aspectos de uma relação entre variáveis, utilizando dados simulados.

Para todos os gráficos produzidos abaixo, avalie esses três aspectos:

1 - Qual a direção da relação (positiva ou negativa)?

2 - A relação se assemelha a uma reta?

3 - Como os valores da variável do eixo Y variam em relação a pequenos intervalos dos valores da variável do eixo X (muita ou pouca variação nos valores de Y dentro de

intervalos de valores de X)?

Para fazermos uma inspeção visual, vamos construir **Diagramas de dispersão**.

Usaremos um novo conjunto de dados para essas análises. Importe o conjunto “bivar.csv” do mesmo modo que foi feito anteriormente para os outros conjuntos de dados. Coloque o nome **“bivar”** nesse novo conjunto de dados.

Agora, faça um gráfico de dispersão (ou *Gráfico XY*) e descreva sua primeira impressão sobre a relação entre as variáveis.

bivariado2

Analisando esse gráfico, como você interpreta os 3 aspectos indicados acima?

Para tentar captar a tendência da relação, você poderia ir traçando pequenas linhas que buscassem a melhor relação entre os dados da variável y.l e da variável x.l ao longo de pequenos trechos da variável x.l, como se estivesse desenhando “à mão”. Existe uma função que faz isso. Ela se chama *lowess* ou *smooth line*.

Existem também outras opções que nos mostram mais informações sobre a relação entre duas variáveis e podem ajudar bastante no entendimento da relação.

bivariado3

Com essas opções adicionadas, a interpretação de cada um dos 3 aspectos analisados mudaria e/ou seria reforçada?

Agora vamos fazer o gráfico de dispersão para outras duas variáveis y.n (resposta) e x.n (preditora), já incluindo todas as opções indicadas acima.

bivariado4

Para esse gráfico, como você interpreta os 3 aspectos indicados acima?

Transformando os dados:

Algumas vezes, a relação que observamos entre duas variáveis não é linear, mas gostaríamos de analisar essa relação dentro do escopo de uma Análise de Regressão Linear, em função das facilidades de trabalhar com esse tipo de análise. Para isso, precisamos recorrer aos recursos de **transformação dos dados**.

Esses recursos podem ser utilizados para fazer com que a distribuição dos dados de uma variável (ou de ambas) seja mais similar a uma distribuição normal.

ATENÇÃO: Atualmente, existem muitas formas alternativas de realizar as análises sem que haja necessidade de transformação dos dados (ver o tópico *Modelos Lineares Generalizados*).

Para esse tutorial, vamos analisar o que acontece quando usamos uma transformação básica.

Logaritmo natural (ln)

Vamos analisar a relação entre as variáveis COMPRIMENTO_BICO e TAMANHO_SEMENTES e verificar se a relação parece linear.

transforma2

E agora, a relação parece mais linear?

Mas cuidado! Agora a relação linear é entre "log(COMPRIMENTO_BICO)" e "TAMANHO_SEMENTES", então é o "logaritmo do comprimento de bico" que aumenta a uma taxa constante em relação ao tamanho das sementes e não mais o "comprimento de bico".

Outras transformações que podem ser utilizadas:

- **Logaritmo base 10**: também para variáveis contínuas com valores extremos
- **Logaritmo natural de x+1**: quando a variável tem muitos zeros
- **Raiz quadrada**: para variáveis que representam contagens (p.ex.: número de indivíduos)
- **Arco seno**: para variáveis que representam proporções/porcentagens

1)

i.e. a altura da caixa é menor

2)

que é igual à mediana