Séries Temporais: introdução, motivação e notação

Prof. Caio Azevedo

Introdução

- Estatística: área do conhecimento/Ciência que trata de metodologias (análise de dados) apropriadas para se coletar, organizar e analisar dados.
- A Estatística incorpora elementos de Probabilidade, Matemática (Cálculo Diferencial e Integral, Álgebra de Matrizes, Teoria da Medida, Análise Funcional etc), Computação e Ciência de dados (envolve outras formas de análise de dados), desenvolvendo (novas) metodologias.
- A Estatística é uma ferramenta muito importante na resolução de problemas levantados em diversas áreas: Biologia, Psicometria, Educação, Medicina, Física, Computação entre outras.
- É importante que o Estatístico participe de todas as etapas de um estudo (pesquisa/consultoria).

Etapas para a resolução de um problema

1. Determinação do problema/objeto de estudo (incluindo a população de interesse).
2. Determinação dos objetivos (gerais e específicos).
3. Determinação do tamanho da amostra-delineamento amostral/experimental.

4. Levantamento dos dados: entrevistas, experimento, coleta de dados etc.
5. **Análise Descritiva**.
6. **Análise Inferencial (Modelos de regressão)**.
7. Conclusões e elaboração dos relatórios/artigos/trabalhos pertinentes.

Pode-se retornar a pontos anteriores ou mesmo avançar, desconsiderando-se alguns dos pontos acima, consoante a necessidade.

Pré-requisitos

- Cálculo diferencial e integral: Cálculo I, Cálculo II, Cálculo III.
  - Probabilidade I : página do curso de Probabilidade I
  - Probabilidade II : página do curso de Probabilidade II
- Inferência: página do curso de ME 419/ME 420 (Graduação), MI 402 (Mestrado)
- Análise de regressão: página do curso de ME 613 (Graduação), MI 406 (Mestrado)

Séries Temporais: Conceitos Introdutórios

- Séries temporais, sob um enfoque informal, consiste em conjuntos de observações (univariadas e multivariadas), mensuradas (ordenadas) ao longo do tempo. Por "tempo" podemos denotar o tempo cronológico, espaço, distância, dentre outras "grandezas" (parâmetro físico).
- Formalmente, séries temporais são realizações (amostragem, experimentação, amostragem/experimentação), de um (ou mais de um) Processo(s) Estocástico(s) (ou de um mecanismo/processo gerador).

Séries Temporais: Conceitos Introdutórios

- À rigor, séries temporais, como os fenômenos naturais, de uma forma geral, são regidos por mecanismos/processos geradores que, embora inteligíveis (passíveis de serem compreendidos e modelados), são inacessíveis.
- Dessa forma, a utilização de Processos Estocásticos/Modelos para Séries temporais (como em qualquer outra área) são aproximações da realidade.
- Denotemos - ST (st): série temporal ou séries temporais.

Um pouco sobre modelagem...

All models are wrong

but some are useful

George E.P. Box

Exemplos

- Dados coletados sequencialmente ao longo tempo são extremamente comuns:
  - Em finanças: observações diárias de taxas de juros, preços diários de fechamento de ações, índices mensais de preços, etc.
  - Em meteorologia: temperaturas máximas e mínimas diárias, precipitação anual e índices de seca.
  - Em agricultura: registro anual de preços por colheita, taxas de erosão de solo, taxas de exportação, etc.
  - Em ciências biológicas: registro de atividade elétrica do coração em intervalos de milissegundo.
  - Entre outras áreas.

Notações gerais

- $Y_t, [Y(t)]$, $t \in T$ em que Y representa os valores possíveis (variáveis aleatórias) do processo estocástico (variável de interesse/resposta) e T é o conjunto de índices de interesse.
- $y_t[y(t)]$, $t \in T$, em que y representa os valores observados do processo estocástico (amostra observada da variável aleatória/resposta).
- T: é o conjunto de índices de interesse (tempo), por exemplo: $T = \mathcal{N}$, $T = \mathcal{N}^+$, $T = \Re$.

Natureza da(s) resposta(s) (Y)

- Y pode ser uma va (variável aleatória) contínua, discreta ou mista.
- Y também pode ser um vetor aleatório, nesse caso utilizaremos
  - $\boldsymbol{Y} = (Y_1, .., Y_p)'$ e, consequentemente: $\boldsymbol{Y}_t = (Y_{1t}, .., Y_{pt})'$, ou $\boldsymbol{Y}(t) = (Y_1(t), ..., Y_1(t))'$
  - E o equivalente para os valores observados, ou seja: $\boldsymbol{y} = (y_1, .., y_p)'$ e, consequentemente: $\boldsymbol{y}_t = (y_{1t}, .., y_{pt})'$, ou $\boldsymbol{y}(t) = (y_1(t), ..., y_p(t))'$
- Assim, podemos ter séries temporais univariadas e multivariadas.

Natureza dos índices (T)

- Discreta: $T = \{t_1, t_2, ..., t_n\}$. Ex: Exportações mensais do Brasil de 1970 a 1980 $\{01/1970, 02/1970, ..., 11/1980, 12/1980\}$.

  Notação usual: Y_t.

- Contínua: $T = \{t : t_1 < t < t_2\}$ Ex: Registro da maré no Rio de Janeiro durante 1 ano, $T = [0, 24]$ se unidade de tempo é a hora.

  Notação usual: $Y(t)$.

- Em geral, no curso, $T = \{1, 2, ..., n\}$ ou $T = \{1, 2, ..., N\}$.

Características e aspectos de ST

- Uma das características principais de uma ST é o fato das observações (variáveis aleatórias) apresentarem algum tipo de estrutura de dependência. Veremos outras características ao longo do curso.
- Também pelo fato acima, é de suma importância saber a ordem de coleta das observações.

Características e aspectos de ST

- Negligenciar estruturas de dependência, via de regra, leva à inferências errôneas (veremos, dentre outros, um exemplo simples, mais adiante).
- Por outro lado, levar em consideração todas (ou o maior número possível de) as características, nos permite fazer análises como: previsão (valores passados-calibração ou futuros), verificação da existência de mudança de regime, presença de outliers etc, de forma (mais) apropriada.

Objetivos

- **Descrever/Compreender/modelar o mecanismo gerador da série**:
  - Descrever apenas o comportamento da série; neste caso, a construção do gráfico da série (gráfico da série temporal, ou "time-series plot", a verificação da existência de tendências, ciclos e variações sazonais, a construção de histogramas e diagramas de dispersão podem ser ferramentas úteis.
  - Investigar (compreender/modelar) o processo gerador da série temporal (na verdade, encontrar um modelo que represente de forma adequada, tal processo); por exemplo, analisando uma série de valores mensais de vendas de automóveis no Brasil, podemos querer saber como estes valores de vendas foram gerados.

Objetivos

- **Predizer o comportamento futuro da série**: estas podem ser a curto prazo, como para série de vendas, produção ou estoque, ou a longo prazo, como para séries de produtividade.
- **Deduzir comportamentos anteriores ao período observado*** (relacionado à Teoria das Múltiplas Hipóteses Concorrentes / Inferência para a melhor explicação).

Objetivos

- Compreender o mecanismo da ST possibilita:
  - Descrever efetivamente o comportamento da série.
  - Encontrar periodicidades na série (neste caso, a análise espectral pode ser de grande utilidade).
  - Tentar obter razões para o comportamento da série (possivelmente através de variáveis auxiliares).
  - Controlar a trajetória da série.

Big Data/Data Science e ST

- Em geral, principalmente com o advento do "Big data" (ver também material enviado via Moodle) e o (res)surgimento da nomenclatura "Data Science", os conjuntos de dados, das mais diversas áreas, costumam apresentar estruturas temporais (por diversas razões, tais estruturas podem não ser consideradas, ou o ser de forma pouco apropriada).
- Até o momento, essencialmente, estudou-se metodologias para a análise de observações ou independentes ou com dependência multivariada (Métodos em Análise Multivariada). Mas não com dependência temporal.

Motivação

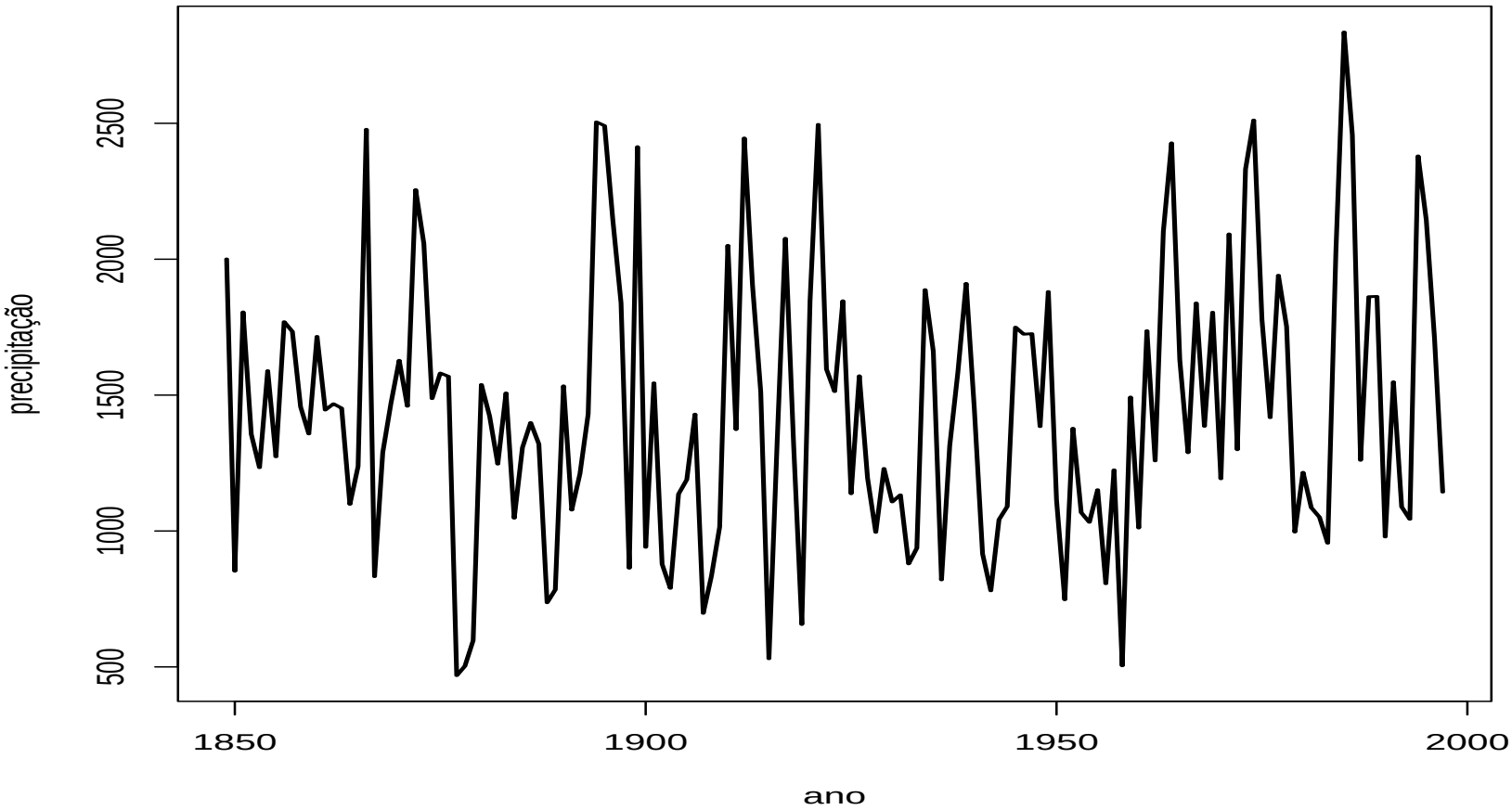


Figura: Precipitação anual da cidade de Fortaleza/CE: 1849 a 1997

Motivação

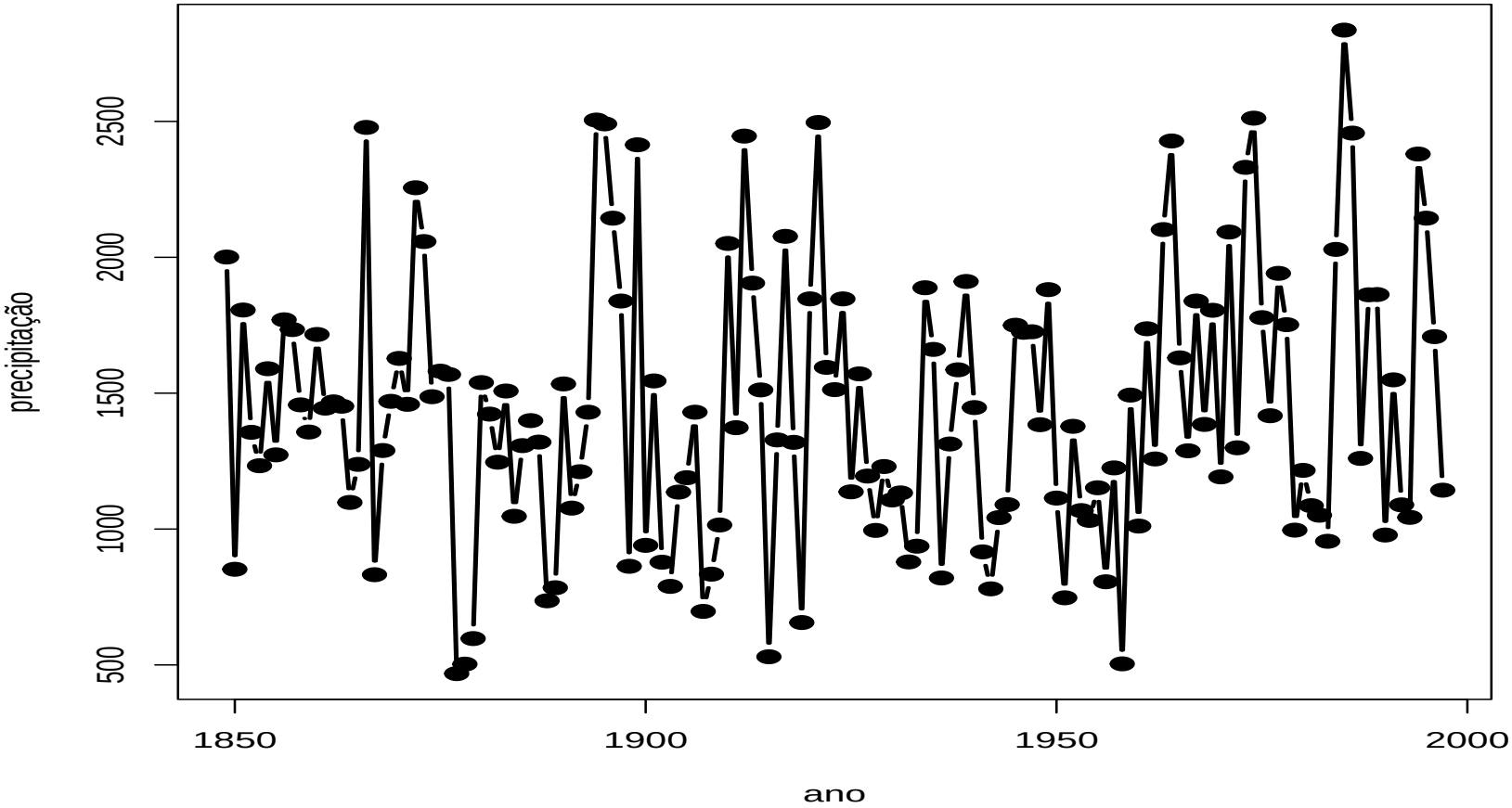
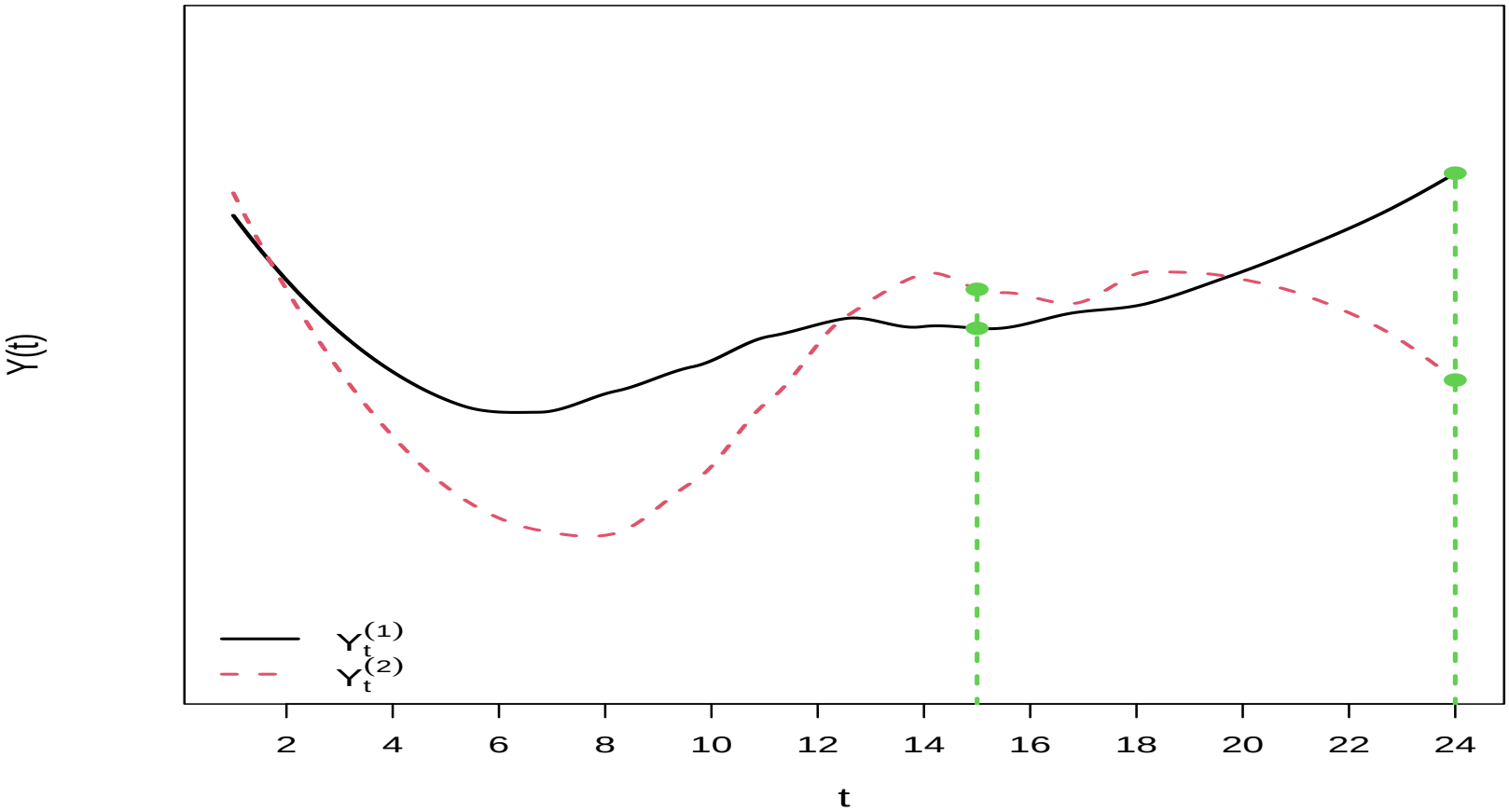


Figura: Precipitação anual da cidade de Fortaleza/CE: 1849 a 1997

Uma ilustração acerca da natureza das ST

- Suponha que queiramos medir a temperatura do ar, de dado local, durante 24 horas (de forma "contínua"), em diferentes dias (slide seguinte).
- Vamos denotar por $Y_t^{(i)}$ a temperatura do local na hora $t \in A \subset [0, 24]$ no dia $i, i = 1, 2, ..., 365$.
- Na prática, pode-se tanto estar interessado em uma única medida que represente cada dia (por exemplo, a mediana/média/total), ou em alguns instantes específicos (por exemplo, manhã, tarde e noite, ST trivariada) ou na curva completa (dados funcionais), por exemplo.
- Em relação ao item anterior, usualmente, consideraremos neste curso, a primeira opção.

Uma ilustração acerca da natureza das ST



Comentários

- Essas (duas) curvas são chamadas trajetórias do processo físico (fenômeno natural) que está sendo observado.
- O que usualmente se chamada de Processo Estocástico, nada mais é que o (um) conjunto de todas as possíveis trajetórias que poderíamos observar.
- Em geral, denotaremos uma trajetória qualquer por $Y_t^{(j)}$. Para cada t fixado, teremos os valores de uma variável aleatória Y_t, para a qual atribuímos alguma distribuição de probabilidade.
- Na realidade, o que chamamos de série temporal é uma parte de uma trajetória, dentre muitas que poderiam ter sido observadas. Por exemplo, considerar a temperatura medida em vários dias, na mesma hora.

Continuação

- Relembrando: $Y_1, ..., Y_n$ (variáveis aleatórias) e $y_1, ..., y_n$ (valores observados).
- Suponha que desejamos estimar a média da Estrutura Verdadeira (mecanismo gerador) responsável pela geração da série temporal observada (precipitação de Fortaleza). Para isso utilizaremos a média amostral $\widehat{\mu} = \frac{1}{n}\sum_{i}^{n} Y_i$ (n = 149).

Continuação

- Considere as seguintes opções:
  1. Desconsiderar a independência e supor que: $\mathcal{E}(Y_i) = \mu$ e $\mathcal{V}(Y_i) = \sigma^2$, $\forall i$.
  2. Considerar a dependência e supor que $\mathcal{E}(Y_i) = \mu$, $\mathcal{V}(Y_i) = \sigma^2$ e $Cov(Y_i, Y_j) = \rho\sigma^2$, $j = i - 1$, $\rho \in (0, 1)$, $\forall i$.

Continuação

- É possível provar que (exercício): $\mathcal{E}_1(\widehat{\mu}) = \mathcal{E}_2(\widehat{\mu}) = \mu$, $\mathcal{V}_1(\widehat{\mu}) = \dfrac{\sigma^2}{n}$ e $\mathcal{V}_2(\widehat{\mu}) = \frac{1}{n^2}\left(n\sigma^2 + (n-1)\rho\sigma^2\right) = \dfrac{\sigma^2}{n} + \left(\dfrac{n-1}{n^2}\right)\rho\sigma^2$, em que $\mathcal{E}_j(), \mathcal{V}_j(.)$ indica que o estimador está sendo avaliado sob a suposição $j, j = 1, 2$.
- Assim, $g(n) = \mathcal{V}_2(\widehat{\mu}) - \mathcal{V}_1(\widehat{\mu}) = \dfrac{n-1}{n^2}\rho\sigma^2 > 0$.

Continuação

- Portanto, dado que a ST em questão, possivelmente, apresenta auto-correlação ($Cov(Y_i, Y_j) \neq 0$), utilizar o procedimento usual (assumir independência) levar-nos-á a subestimar o erro-padrão do estimador.
- Dessa forma, embora a subestimação seja mitigada à medida que $n \to \infty$ (pois $\lim_{n\to\infty} g(n) = 0$), a inferência sob a suposição 2 é mais apropriada do que aquela feita sob a suposição.

Características empíricas das ST

- Tendência:
  - Evolução a longo do prazo, comportamento (nível) médio.
  - Pode ser linear, quadrática, não linear, etc.
  - Muitas vezes, a especificação (apropriada) da tendência não é simples.

Tendência

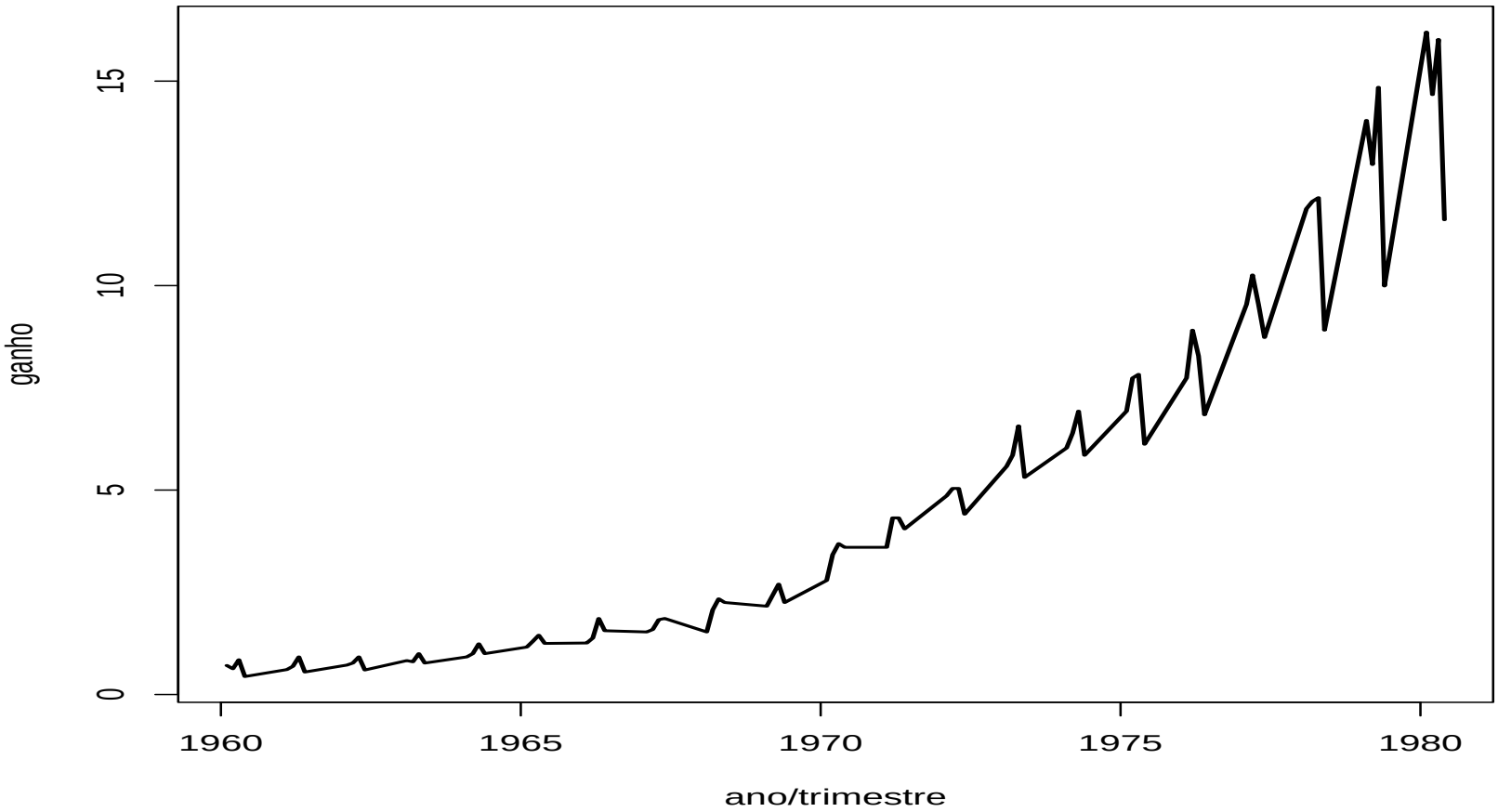


Figura: Ganhos trimestrais de ações (em dólares) da Johnson & Johnson: 196080

Características empíricas das ST

- Sazonalidade:
  - Flutuações cíclicas relacionadas a algum "calendário" (variação periódica - semanal, mensal, anual, etc).
  - Padrão regular anual - Por exemplo: se temos dados trimestrais, o comportamento dos trimestres é similar em cada ano.
  - Não existe uma definição precisa (consenso) de sazonalidade. Uma delas é: efeito (desvio) com relação à tendência (média).

Sazonalidade

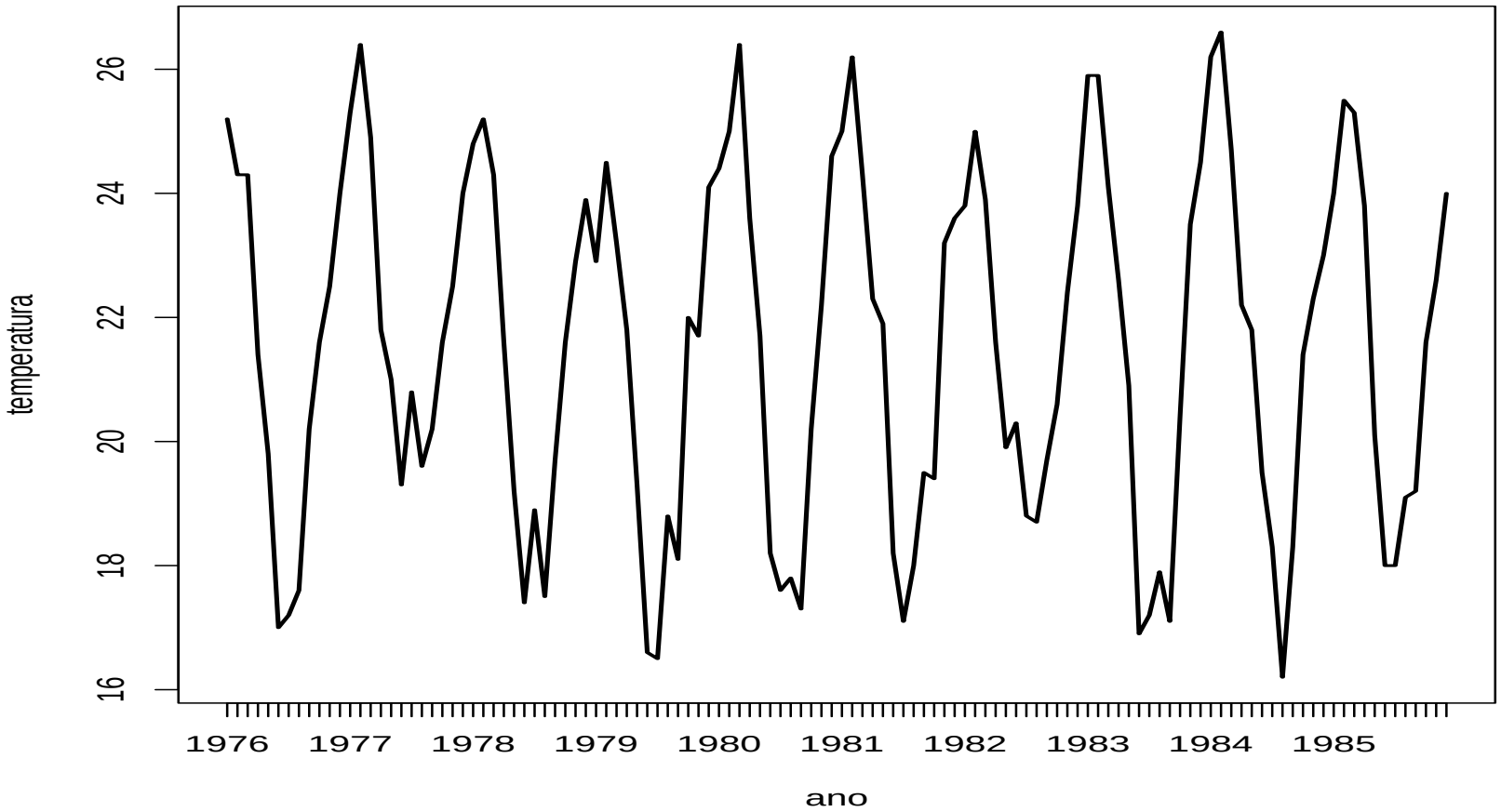


Figura: Temperaturas mensais ($^{\circ}C$) da cidade de Cananéia/SP: 1976 a 1985

Características empíricas das ST

- Ciclos:
  - Outros tipos de periodicidades.
  - Variações que apesar de periódicas não são associadas facilmente a nenhuma medida temporal.
  - Exemplos: ciclos econômicos (recessões a cada 7 anos), periodicidade do El Niño, ciclos de epidemias.

Ciclos

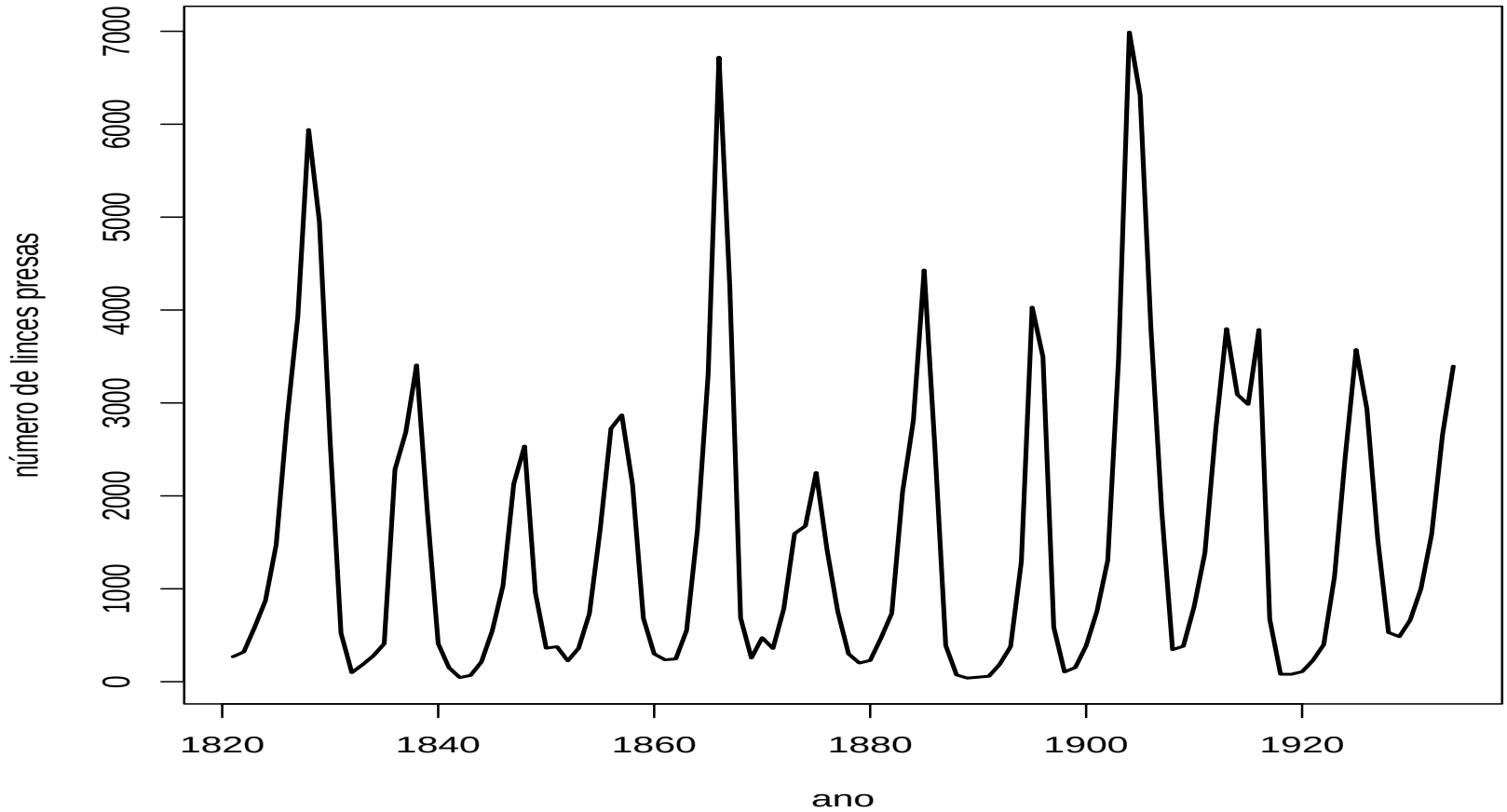


Figura: Número anual de linces presos em armadilhas no distrito do rio Mackenzie (Noroeste do Canadá) - 1821 a 1934

Características empíricas das ST

- Valores discrepantes ou outliers:
  - Observações fora do padrão da série.
  - É muito importante sua detecção, pois tais valores podem ter um efeito significativo (em termos estatísticos e/ou do problema) na análise estatística da série.
  - Estas observações devem ser estudadas mas não retiradas.

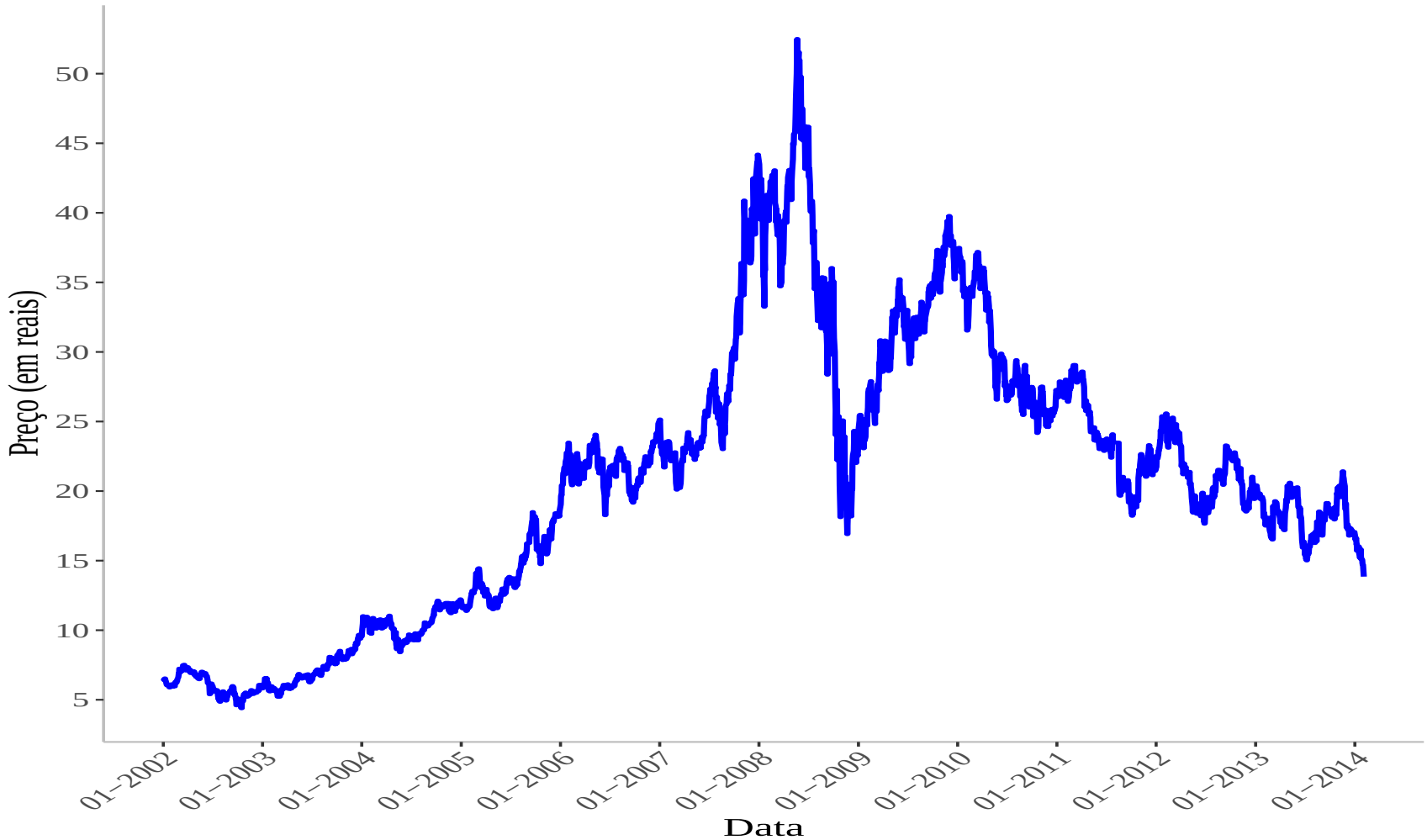


Figura: Preço (em R$) diário de fechamento da (ação) PETR4 PN, 02/01/2002 a 04/02/2014

?????

Figura: Produção anual de petróleo (em milhões de toneladas) da Arábia Saudita - 1965 a 2013

?????

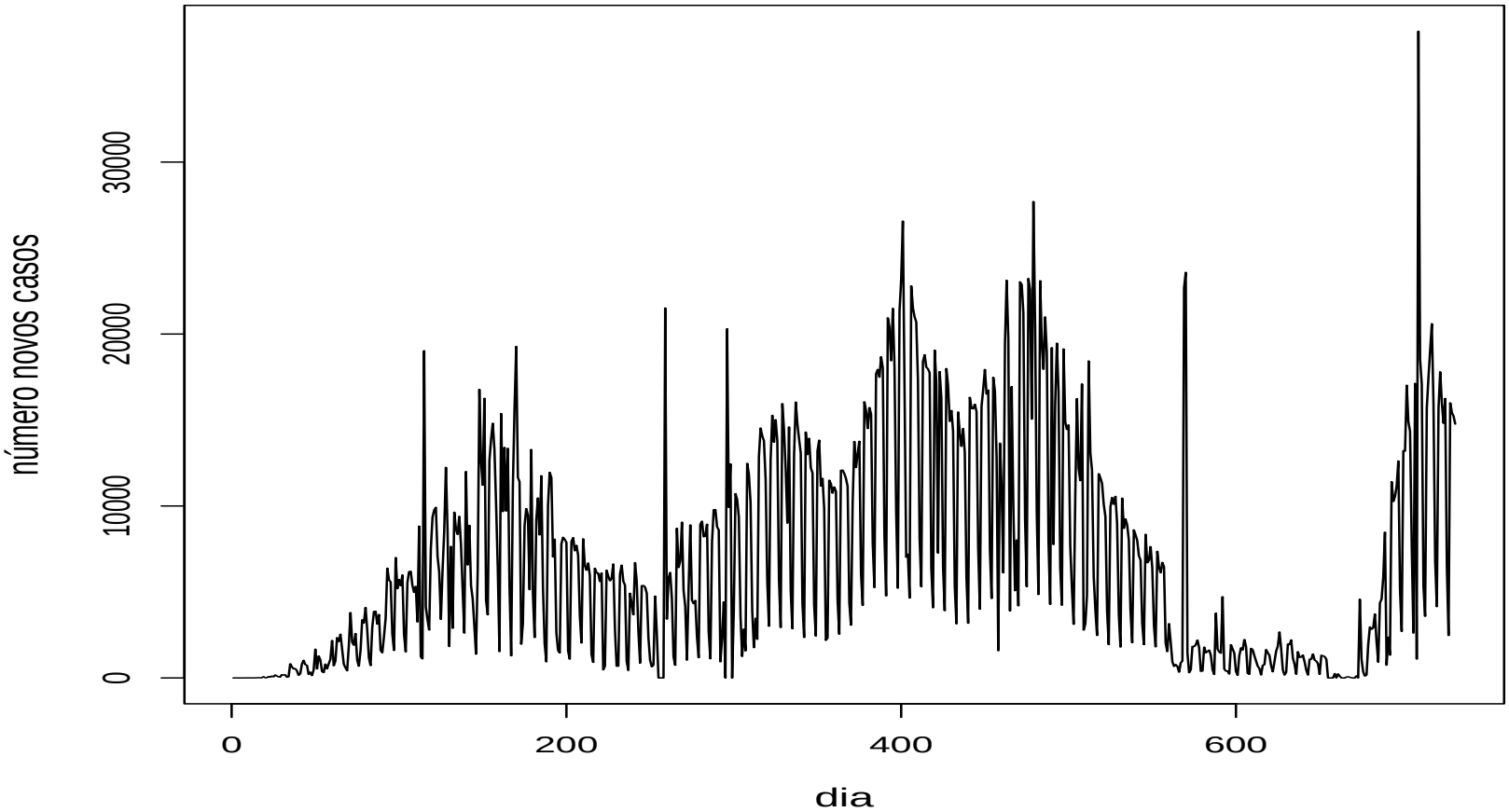


Figura: Número de novos casos diários de COVID 19 no estado de SP - 26/02/2020 a 26/02/2022

Tipos de Modelagem

- Há, basicamente, dois enfoques usados na análise de séries temporais:
  - **Domínio temporal**: Procedimento baseado no fato de que a correlação (dependência) entre valores adjacentes da série temporal é melhor explicada em termos de uma regressão dos valores passados da série e de um ruído. Os modelos propostos são paramétricos (com um número finito de parâmetros).
  - **Domínio da frequência**: Procedimento baseado no fato de que uma série temporal pode ser decomposta como uma superposição linear de senos e cossenos com períodos diferentes. Os modelos propostos são não-paramétricos.

Séries Temporais $\times$ Dados Longitudinais $\times$ Dados multivariados

- Séries temporais: uma ou mais de uma variável medida de poucos (ou somente um) indivíduo(s) (que também podem fazer papel de variável) ao longo de muitos instantes no tempo.
- Dados longitudinais: uma ou mais de uma variável medida de muitos indivíduos ao longo de (poucas) condições (mesmo sendo o tempo, profundidade).
- Dados multivariados: mais de uma variável (de diferentes naturezas) medidas, em geral, em um único instante (cohort).

Exemplo de dados longitudinais: Concentração de bilirrubina em recém-nascidos saudáveis

- Os dados correspondem a um estudo realizado na Escola Paulista de Medicina (UNIFESP), em que foi medida a concentração de bilirrubina (μ mol/L) em 89 recém-nascidos a termo (gestação entre 37 e 42 semanas) saudáveis em aleitamento materno durante 1, 2, 3, 4, 5, 6, 8, 10 e 12 dias após o nascimento.
- O objetivo era explicar a variação da concentração de bilirrubina em função da idade.

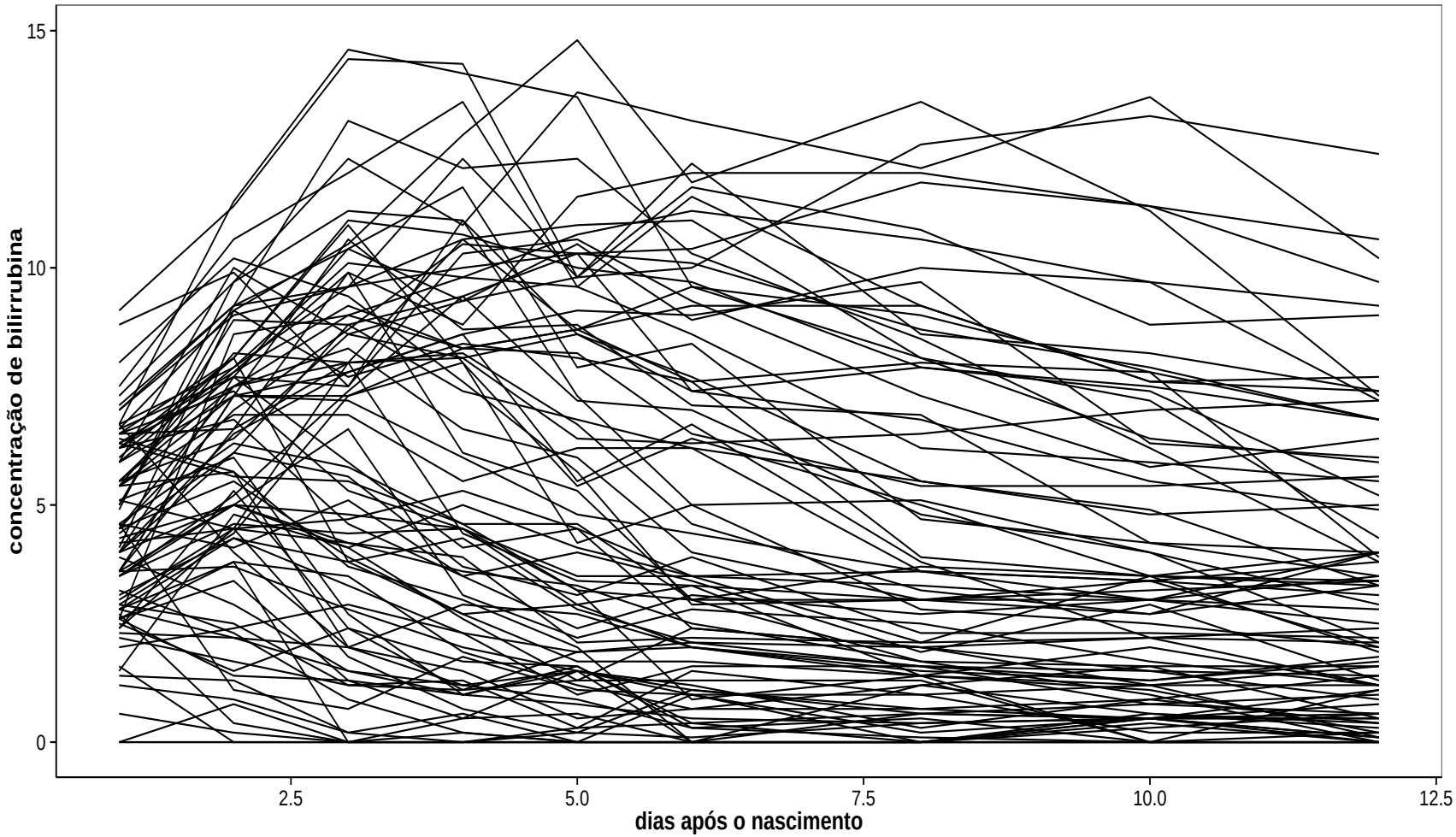
Exemplo 1: cont.

- A bilirrubina é uma substância amarelada encontrada na bile, que permanece no plasma sanguíneo até ser eliminada na urina. Quanto mais bilirrubina eliminada na urina, mais amarela ela se torna. Excesso de bilirrubina (hiperbilirrubinemia) pode indicar problemas no fígado, baço, nos rins ou na vesícula biliar.
- Estudo irregular, balanceado e completo (89 observações para cada condição de avaliação e 9 por indivíduo).

Banco de dados (longitudinal)

RN	Dia	Bilirrubina
1	1	2,70
1	2	0,40
⋮	⋮	⋮
1	12	0,80
⋮	⋮	⋮
89	1	2,60
89	2	1,40
⋮	⋮	⋮
89	12	0,60

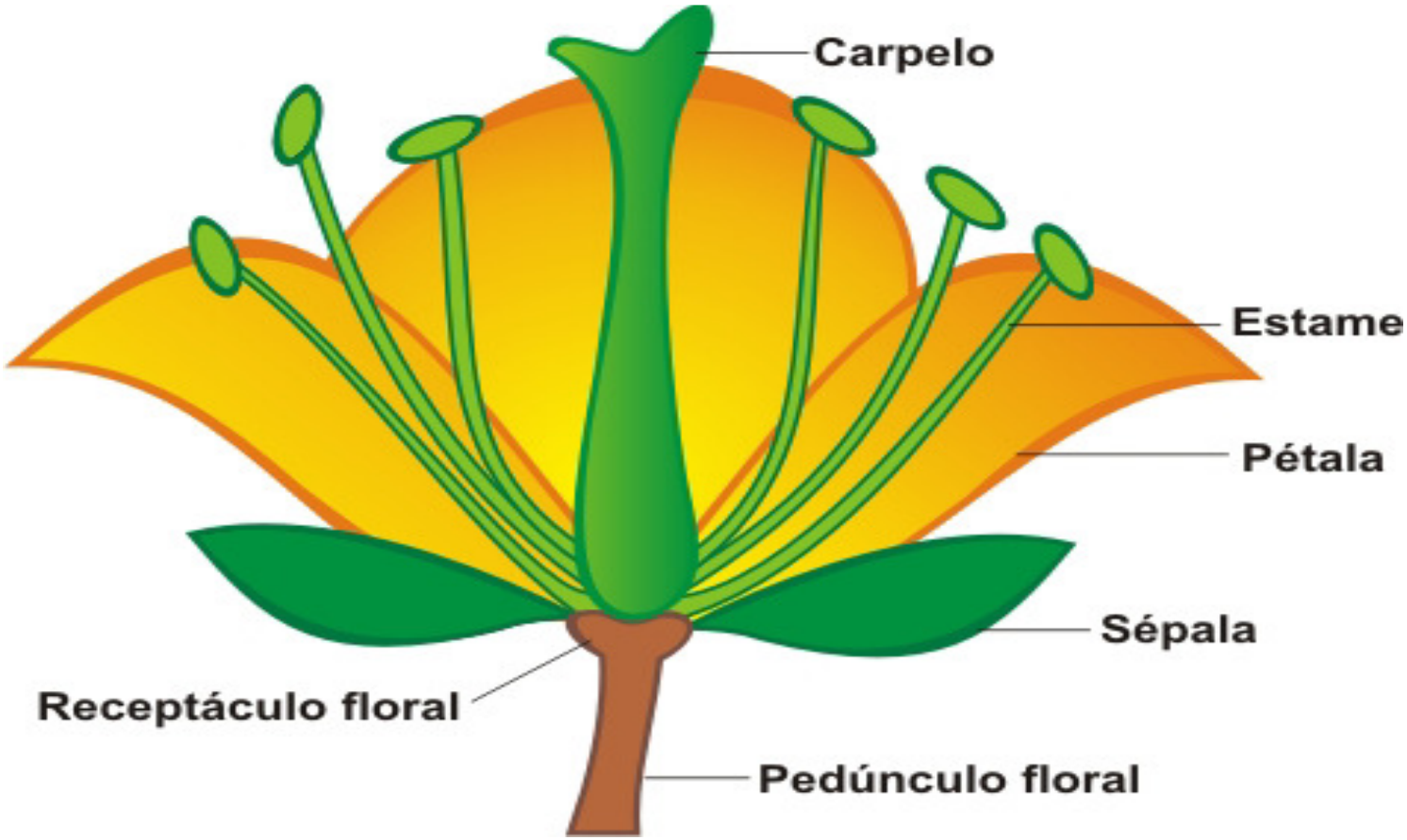
Perfis individuais: amostra completa



Exemplo de dados multivariados: Iris de Fisher

- Os dados consistem de 50 unidades amostrais de três espécies (setosa, virginica, versicolor) de íris (uma espécie de planta), ou seja, temos um total de 150 unidades amostrais.
- De cada uma delas mediu-se quatro variáveis: comprimento e largura da sépala (CS, LS) e comprimento e largura da pétala (CP,LP).
- Objetivo original: quantificar a variação morfológica em relação à essas espécies com bases nas quatro variáveis de interesse.

Cont.


Carpelo
Estame
Pétala
Sépala
Receptáculo floral
Pedúnculo floral

Planta	Tipo	Comp. Sep	Larg. Sep.	Comp. Pet.	Larg. Pet
1	setosa	5,10	3,50	1,40	0,20
2	setosa	4,90	3,00	1,40	0,20
⋮	⋮	⋮	⋮	⋮	⋮
50	setosa	5,00	3,30	1,40	0,20
51	versicolor	7,00	3,20	4,70	1,40
52	versicolor	6,40	3,20	4,50	1,50
⋮	⋮	⋮	⋮	⋮	⋮
100	versicolor	5,70	2,80	4,10	1,30
101	virginica	6,30	3,30	6,00	2,50
102	virginica	5,80	2,70	5,10	1,90
⋮	⋮	⋮	⋮	⋮	⋮
150	virginica	5,90	3,00	5,10	1,80

Comentários

- Concentrar-nos-emos (eventualmente considerando situações fora dos contextos abaixo) em:
  - ST univariadas.
  - Tempo discreto.
  - Domínio do tempo (com algumas metodologias não paramétricas).
  - Inferência Frequentista.
- Expectativas acerca do aprendizado do aluno:
  - Reconhecer, adequadamente uma ST, e suas principais características.
  - Ser capaz de utilizar as principais técnicas descritivas e inferenciais relativas a ST univariadas.
  - Interpretar e reportar adequadamente os resultados obtidos.

(Referências para) Outros assuntos

- ST multivariadas.
- Tempo contínuo.
- Domínio da frequência.
- Inferência Bayesiana.