

# Estatística e Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística  
Universidade de São Paulo  
pam@ime.usp.br  
<http://www.ime.usp.br/~pam>

## Sumário

- 1 As Origens
- 2 Inferência Bayesiana
- 3 Inferência Frequentista
- 4 Estatística
- 5 Estatística e Ciência de Dados
- 6 Aprendizado com Estatística e com Máquina
- 7 Sugestões

## Paradigma

- Modelo, padrão a ser seguido.
- Um pressuposto filosófico, uma teoria, um conhecimento, que origina o estudo de um campo científico.
- Aquilo que os membros de uma comunidade científica partilham.

Exemplos:

- Inferência frequentista, inferência bayesiana
- Data Mining, Neural Networks, Data Science

# Probabilidade

- Início em 1654 com Fermat (1601-1665), Pascal (1623-1662): jogos de dados
- Huygens (1629-1695): primeiro livro de probabilidade em 1657.
- Bayes (1702-1761): primeira versão do Teorema de Bayes, publicado em 1763.

# Probabilidade

- Início em 1654 com Fermat (1601-1665), Pascal (1623-1662): jogos de dados
- Huygens (1629-1695): primeiro livro de probabilidade em 1657.
- Bayes (1702-1761): primeira versão do Teorema de Bayes, publicado em 1763.

# Probabilidade

- Início em 1654 com Fermat (1601-1665), Pascal (1623-1662): jogos de dados
- Huygens (1629-1695): primeiro livro de probabilidade em 1657.
- Bayes (1702-1761): primeira versão do Teorema de Bayes, publicado em 1763.

# Gauss

- Gauss (1777-1856) inventou o método de mínimos quadrados (MQ) na última década do século 18 (1795) e o usou regularmente depois de 1801 em cálculos astronômicos.
- Legendre (1752-1833): publicou no apêndice de "Nouvelles Methodes pour la Détermination des Orbites des Còmetes". Nenhuma justificação.
- Gauss (1809): deu justificativa probabilística do método. Em "The Theory of the Motion of Heavenly Bodies".
- Gauss (1820): "Theoria Combinationis Observationum Erroribus Minimis Obnoxiae".

# Gauss

- Gauss (1777-1856) inventou o método de mínimos quadrados (MQ) na última década do século 18 (1795) e o usou regularmente depois de 1801 em cálculos astronômicos.
- Legendre (1752-1833): publicou no apêndice de "Nouvelles Methodes pour la Détermination des Orbites des Còmetes". Nenhuma justificação.
- Gauss (1809): deu justificativa probabilística do método. Em "The Theory of the Motion of Heavenly Bodies".
- Gauss (1820): "Theoria Combinationis Observationum Erroribus Minimis Obnoxiae".

## Gauss

- Gauss (1777-1856) inventou o método de mínimos quadrados (MQ) na última década do século 18 (1795) e o usou regularmente depois de 1801 em cálculos astronômicos.
- Legendre (1752-1833): publicou no apêndice de "Nouvelles Methodes pour la Détermination des Orbites des Còmetes". Nenhuma justificação.
- Gauss (1809): deu justificativa probabilística do método. Em "The Theory of the Motion of Heavenly Bodies".
- Gauss (1820): "Theoria Combinationis Observationum Erroribus Minimis Obnoxiae".

## Gauss

- Gauss (1777-1856) inventou o método de mínimos quadrados (MQ) na última década do século 18 (1795) e o usou regularmente depois de 1801 em cálculos astronômicos.
- Legendre (1752-1833): publicou no apêndice de "Nouvelles Methodes pour la Détermination des Orbites des Còmetes". Nenhuma justificação.
- Gauss (1809): deu justificativa probabilística do método. Em "The Theory of the Motion of Heavenly Bodies".
- Gauss (1820): "Theoria Combinationis Observationum Erroribus Minimis Obnoxiae".

## Bayes ou Laplace?

- Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
- 1812: *Théorie Analytique des Probabilités*: aplicações científicas e práticas.
- 1814: *Essais Philosophiques sur les Probabilités*: interpretação Bayesiana das probabilidades
- Inferência Bayesiana  $\leftrightarrow$  Inferência Laplaciana. Usada a partir de 1800.
- Fisher e Neyman: início do século 20.
- Jeffrey (1939). *Theory of Probability*. Considerado como o re-início da Inferência Bayesiana
- de Finetti, Savage, Lindley etc.

## Bayes ou Laplace?

- Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
- 1812: *Théorie Analytique des Probabilités*: aplicações científicas e práticas.
- 1814: *Essais Philosophiques sur les Probabilités*: interpretação Bayesiana das probabilidades
- Inferência Bayesiana  $\leftrightarrow$  Inferência Laplaciana. Usada a partir de 1800.
- Fisher e Neyman: início do século 20.
- Jeffrey (1939). *Theory of Probability*. Considerado como o re-início da Inferência Bayesiana
- de Finetti, Savage, Lindley etc.

## Bayes ou Laplace?

- Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
- 1812: *Théorie Analytique des Probabilités*: aplicações científicas e práticas.
- 1814: *Essais Philosophiques sur les Probabilités*: interpretação Bayesiana das probabilidades
- Inferência Bayesiana  $\leftrightarrow$  Inferência Laplaciana. Usada a partir de 1800.
- Fisher e Neyman: início do século 20.
- Jeffrey (1939). *Theory of Probability*. Considerado como o re-início da Inferência Bayesiana
- de Finetti, Savage, Lindley etc.

## Bayes ou Laplace?

- Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
- 1812: *Théorie Analytique des Probabilités*: aplicações científicas e práticas.
- 1814: *Essais Philosophiques sur les Probabilités*: interpretação Bayesiana das probabilidades
- Inferência Bayesiana  $\leftrightarrow$  Inferência Laplaciana. Usada a partir de 1800.
- Fisher e Neyman: início do século 20.
- Jeffrey (1939). *Theory of Probability*. Considerado como o re-início da Inferência Bayesiana
- de Finetti, Savage, Lindley etc.

## Bayes ou Laplace?

- Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
- 1812: *Théorie Analytique des Probabilités*: aplicações científicas e práticas.
- 1814: *Essais Philosophiques sur les Probabilités*: interpretação Bayesiana das probabilidades
- Inferência Bayesiana  $\leftrightarrow$  Inferência Laplaciana. Usada a partir de 1800.
- Fisher e Neyman: início do século 20.
- Jeffrey (1939). *Theory of Probability*. Considerado como o re-início da Inferência Bayesiana
- de Finetti, Savage, Lindley etc.

## Bayes ou Laplace?

- Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
- 1812: *Théorie Analytique des Probabilités*: aplicações científicas e práticas.
- 1814: *Essais Philosophiques sur les Probabilités*: interpretação Bayesiana das probabilidades
- Inferência Bayesiana  $\leftrightarrow$  Inferência Laplaciana. Usada a partir de 1800.
- Fisher e Neyman: início do século 20.
- Jeffrey (1939). *Theory of Probability*. Considerado como o re-início da Inferência Bayesiana
- de Finetti, Savage, Lindley etc.

## Bayes ou Laplace?

- Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
- 1812: Théorie Analytique des Probabilités: aplicações científicas e práticas.
- 1814: Essais Philosophiques sur les Probabilités: interpretação Bayesiana das probabilidades
- Inferência Bayesiana  $\leftrightarrow$  Inferência Laplaciana. Usada a partir de 1800.
- Fisher e Neyman: início do século 20.
- Jeffrey (1939). Theory of Probability. Considerado como o re-início da Inferência Bayesiana
- de Finetti, Savage, Lindley etc.

## Bayes ou Laplace?

- Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
- 1812: Théorie Analytique des Probabilités: aplicações científicas e práticas.
- 1814: Essais Philosophiques sur les Probabilités: interpretação Bayesiana das probabilidades
- Inferência Bayesiana  $\leftrightarrow$  Inferência Laplaciana. Usada a partir de 1800.
- Fisher e Neyman: início do século 20.
- Jeffrey (1939). Theory of Probability. Considerado como o re-início da Inferência Bayesiana
- de Finetti, Savage, Lindley etc.

## Fisher e Neyman

- Inferência Frequentista (testes de hipóteses, estimação, planejamento de experimentos e amostragem) foi iniciada por R. Fisher( 1890-1962) e J. Neyman (1894-1981).
- Fisher (1925): Statistical Methods for Research Workers. (14ª Edição: 1970)
- Fisher (1935): The Design of Experiments (8ª Edição: 1966)

## Fisher e Neyman

- Inferência Frequentista (testes de hipóteses, estimação, planejamento de experimentos e amostragem) foi iniciada por R. Fisher( 1890-1962) e J. Neyman (1894-1981).
- Fisher (1925): Statistical Methods for Research Workers. (14<sup>a</sup> Edição: 1970)
- Fisher (1935): The Design of Experiments (8<sup>a</sup> Edição: 1966)

## Fisher e Neyman

- Inferência Frequentista (testes de hipóteses, estimação, planejamento de experimentos e amostragem) foi iniciada por R. Fisher( 1890-1962) e J. Neyman (1894-1981).
- Fisher (1925): Statistical Methods for Research Workers. (14<sup>a</sup> Edição: 1970)
- Fisher (1935): The Design of Experiments (8<sup>a</sup> Edição: 1966)

## Gosset/Student

- W. Gosset (1876-1937): Em 1908 publicou sob o pseudônimo de Student um artigo que iniciou um novo paradigma em "Pequenas Amostras". Resultado provado por Fisher em 1912-1915.
- Student (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
- Fisher (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Royal Society, A*, 222, 308-368.
- Stigler: "The most influential article on Theoretical Statistics in the 20th Century"
- Hald: "For the first time in the history of Statistics a framework for frequency-based general theory of parametric statistical inference was clearly formulated."

## Gosset/Student

- W. Gosset (1876-1937): Em 1908 publicou sob o pseudônimo de Student um artigo que iniciou um novo paradigma em "Pequenas Amostras". Resultado provado por Fisher em 1912-1915.
- Student (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
- Fisher (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Royal Society, A*, 222, 308-368.
- Stigler: "The most influential article on Theoretical Statistics in the 20th Century"
- Hald: "For the first time in the history of Statistics a framework for frequency-based general theory of parametric statistical inference was clearly formulated."

## Gosset/Student

- W. Gosset (1876-1937): Em 1908 publicou sob o pseudônimo de Student um artigo que iniciou um novo paradigma em "Pequenas Amostras". Resultado provado por Fisher em 1912-1915.
- Student (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
- Fisher (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Royal Society, A*, 222, 308-368.
- Stigler: "The most influential article on Theoretical Statistics in the 20th Century"
- Hald: "For the first time in the history of Statistics a framework for frequency-based general theory of parametric statistical inference was clearly formulated."

## Gosset/Student

- W. Gosset (1876-1937): Em 1908 publicou sob o pseudônimo de Student um artigo que iniciou um novo paradigma em "Pequenas Amostras". Resultado provado por Fisher em 1912-1915.
- Student (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
- Fisher (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Royal Society, A*, 222, 308-368.
- Stigler: "The most influential article on Theoretical Statistics in the 20th Century"
- Hald: "For the first time in the history of Statistics a framework for frequency-based general theory of parametric statistical inference was clearly formulated."

## Gosset/Student

- W. Gosset (1876-1937): Em 1908 publicou sob o pseudônimo de Student um artigo que iniciou um novo paradigma em "Pequenas Amostras". Resultado provado por Fisher em 1912-1915.
- Student (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
- Fisher (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Royal Society, A*, 222, 308-368.
- Stigler: "The most influential article on Theoretical Statistics in the 20th Century"
- Hald: "For the first time in the history of Statistics a framework for frequency-based general theory of parametric statistical inference was clearly formulated."

## Neyman e Pearson

- Neyman and Pearson (1933a) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Royal Society*, A, 231, 289-331.
- Neyman and Pearson (1933b). On the testing of statistical hypothesis in relation to probabilities a priori. *Proc. Cambridge Philos. Society*, 24, 429-510.
- Livros de E. Lehmann sobre estimação e testes de hipóteses. 1967, 1983.
- Era do "Small Data".

## Neyman e Pearson

- Neyman and Pearson (1933a) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Royal Society*, A, 231, 289-331.
- Neyman and Pearson (1933b). On the testing of statistical hypothesis in relation to probabilities a priori. *Proc. Cambridge Philos. Society*, 24, 429-510.
- Livros de E. Lehmann sobre estimação e testes de hipóteses. 1967, 1983.
- Era do "Small Data".

## Neyman e Pearson

- Neyman and Pearson (1933a) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Royal Society*, A, 231, 289-331.
- Neyman and Pearson (1933b). On the testing of statistical hypothesis in relation to probabilities a priori. *Proc. Cambridge Philos. Society*, 24, 429-510.
- Livros de E. Lehmann sobre estimação e testes de hipóteses. 1967, 1983.
- Era do "Small Data".

## Neyman e Pearson

- Neyman and Pearson (1933a) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Royal Society*, A, 231, 289-331.
- Neyman and Pearson (1933b). On the testing of statistical hypothesis in relation to probabilities a priori. *Proc. Cambridge Philos. Society*, 24, 429-510.
- Livros de E. Lehmann sobre estimação e testes de hipóteses. 1967, 1983.
- Era do "Small Data".

# O que é Estatística?

- [1] Coleta de dados: amostras, planejamento de experimentos, estudos observacionais.
- [2] Modelagem e análise de dados.
- [3] Tomada de decisões.

## Eras da Estatística

A história da Estatística pode ser dividida em três eras:

- (1) A era de Quetelet (astrônomo, matemático, estatístico belga, 1796-1874) e seus sucessores, na qual o objetivo era obter grandes conjuntos de dados (censos) em ciências sociais.
- (2) O período clássico de Pearson (1857-1936), Fisher (1890-1962), Neyman (1894-1981), Hotelling (1895-1973) e seus sucessores, que desenvolveram a teoria de inferência ótima; métodos apropriados para *small data sets*.
- (3) A era da produção de dados em massa (*big data sets*), com novas tecnologias como *microarrays*, dados de alta frequência em finanças, dados astronômicos etc.

## Algoritmo e Inferência

### Análise Estatística:

- (a) algorítmica
- (b) inferencial

**Exemplo:** Considere estimar a média  $\mu = E(X)$  de uma v.a.  $X$ , definida sobre uma população.

Para uma AAS  $X_1, \dots, X_n$ , considere o estimador

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Este é o **algoritmo**.

Quão acurado e preciso é  $\bar{X}$ . Esta é a parte da **inferência**.

## Algoritmo e Inferência

Algoritmos: é o que os estatísticos fazem.

Inferência: porque os estatísticos usam os algoritmos. (Efrom, 2016).

Conjuntos de dados enormes (*Big Data*) requerem novas metodologias. Esta demanda está sendo atendida por algoritmos estatísticos baseados em computação intensiva.

## Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

## Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

## Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

## Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

## Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

## Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

## Ciência de Dados

- Atualmente, os termos *Data Science* (Ciência de Dados) e *Big Data* (Megadados) são utilizados em profusão, como se fossem conceitos novos, distintos daqueles com que os estatísticos lidam há cerca de dois séculos.
- Na década de 1980, numa palestra na Universidade de Michigan, EUA, C.F. Jeff Wu já sugeria que se adotassem os rótulos *Statistical Data Science*, ou simplesmente, *Data Science*, em lugar de *Statistics*, para dar maior visibilidade ao trabalho dos estatísticos.
- Talvez seja Tukey (1962, 1977), sob a denominação *Exploratory Data Analysis* (Análise Exploratória de Dados), o primeiro a dar importância ao que hoje se chama Ciência de Dados, sugerindo que se desse mais ênfase ao uso de tabelas, gráficos e outros dispositivos para uma análise preliminar de dados, antes que se passasse a uma **análise confirmatória**, que seria a **inferência estatística**.

## Ciência de Dados

- Outros autores, como Chambers (1993), Breiman (2001) e Cleveland (1985, 1993, 2001), também enfatizaram a preparação, apresentação e descrição dos dados como atividades preparatórias para inferência ou modelagem.
- Basta uma procura simples na Internet para identificar novos centros de Ciências de Dados (CD) em várias universidades ao redor do mundo, com programas de mestrado, doutorado e mesmo graduação.
- O interessante é que muitos desses programas estão alojados em escolas de Engenharia, Bioestatística, Ciência da Computação, Administração, Economia etc., e não em departamentos de Estatística.
- Paradoxalmente, há estatísticos que acham que Estatística é a parte menos importante de CD! Certamente isso é um equívoco. Como ressalta Donoho (2017), se uma das principais características de CD é analisar grandes conjuntos de dados (Megadados), há mais de 200 anos os estatísticos têm se preocupado com a análise de vastos conjuntos de dados provenientes de censos, coleta de informações meteorológicas, observação de séries de índices financeiros etc., que têm essa característica.

## Ciência de Dados

- Outro equívoco consiste em imaginar que a Estatística Clássica (frequentista, bayesiana etc.) trata somente de pequenos volumes de dados, conhecidos como *Small Data*.
- Essa interpretação errônea vem do fato de que muitos livros didáticos apresentam conjuntos de dados, em geral de pequeno ou médio porte, para que as metodologias apresentadas possam ser aplicadas pelos leitores, mesmo utilizando calculadoras ou aplicativos estatísticos (pacotes). Nada impede que essas metodologias sejam aplicadas a grandes volumes de dados a não ser pelas dificuldades computacionais inerentes.
- Talvez seja este aspecto computacional, aquele que mascara os demais componentes daquilo que se entende por CD, pois em muitos casos, o interesse é dirigido apenas para o desenvolvimento de algoritmos cuja finalidade é aprender a partir dos dados, omitindo-se características estatísticas.

## Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neurais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

## Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neurais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

## Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neurais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

## Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neurais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

## Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neurais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

## CD: Perspectiva da Estatística

- Estatística "serve" a Ciência guiando na coleta e análise de dados.
- Dados envolvem incertezas: como foram coletados, medidos ou como foram gerados. A modelagem estatística ajuda a quantificar e racionalizar incertezas de maneira sistemática.
- Conjuntos de dados são complexos: tipos diferentes de dependência (ao longo do tempo, sobre escalas espaciais, entre variáveis diferentes)
- Dados de alta dimensão: medimos milhares de variáveis para cada unidade amostral.

## CD: Perspectiva da Estatística

- Estatística "serve" a Ciência guiando na coleta e análise de dados.
- Dados envolvem incertezas: como foram coletados, medidos ou como foram gerados. A modelagem estatística ajuda a quantificar e racionalizar incertezas de maneira sistemática.
- Conjuntos de dados são complexos: tipos diferentes de dependência (ao longo do tempo, sobre escalas espaciais, entre variáveis diferentes)
- Dados de alta dimensão: medimos milhares de variáveis para cada unidade amostral.

## CD: Perspectiva da Estatística

- Estatística "serve" a Ciência guiando na coleta e análise de dados.
- Dados envolvem incertezas: como foram coletados, medidos ou como foram gerados. A modelagem estatística ajuda a quantificar e racionalizar incertezas de maneira sistemática.
- Conjuntos de dados são complexos: tipos diferentes de dependência (ao longo do tempo, sobre escalas espaciais, entre variáveis diferentes)
- Dados de alta dimensão: medimos milhares de variáveis para cada unidade amostral.

## CD: Perspectiva da Estatística

- Estatística "serve" a Ciência guiando na coleta e análise de dados.
- Dados envolvem incertezas: como foram coletados, medidos ou como foram gerados. A modelagem estatística ajuda a quantificar e racionalizar incertezas de maneira sistemática.
- Conjuntos de dados são complexos: tipos diferentes de dependência (ao longo do tempo, sobre escalas espaciais, entre variáveis diferentes)
- Dados de alta dimensão: medimos milhares de variáveis para cada unidade amostral.

## CD: Perspectiva da Computação

- Particularmente importante na análise de dados contemporâneos, onde frequentemente nos deparamos com a dicotomia entre acurácia e precisão estatística e recursos computacionais (tempo e memória).
- Exemplos: otimização, bootstrap, MCMC.
- Distribuição de conjuntos de dados enormes por múltiplos processadores (velocidade) e múltiplos equipamentos de armazenamento (memória).

## CD: Perspectiva da Computação

- Particularmente importante na análise de dados contemporâneos, onde frequentemente nos deparamos com a dicotomia entre acurácia e precisão estatística e recursos computacionais (tempo e memória).
- Exemplos: otimização, bootstrap, MCMC.
- Distribuição de conjuntos de dados enormes por múltiplos processadores (velocidade) e múltiplos equipamentos de armazenamento (memória).

## CD: Perspectiva da Computação

- Particularmente importante na análise de dados contemporâneos, onde frequentemente nos deparamos com a dicotomia entre acurácia e precisão estatística e recursos computacionais (tempo e memória).
- Exemplos: otimização, bootstrap, MCMC.
- Distribuição de conjuntos de dados enormes por múltiplos processadores (velocidade) e múltiplos equipamentos de armazenamento (memória).

## CD: Perspectiva Humana

- CD liga modelos estatísticos e métodos computacionais para resolver problemas específicos de outras disciplinas.
- Entender o domínio de um problema, decidir quais dados obter, como processá-los, explorar e visualizar os dados, selecionar um modelo estatístico e métodos computacionais apropriados, comunicar os resultados da análise.
- Estas habilidades não são usualmente ensinadas em disciplinas tradicionais de Estatística ou Computação, mas são adquiridas por meio da experiência e colaboração com outros pesquisadores.

## CD: Perspectiva Humana

- CD liga modelos estatísticos e métodos computacionais para resolver problemas específicos de outras disciplinas.
- Entender o domínio de um problema, decidir quais dados obter, como processá-los, explorar e visualizar os dados, selecionar um modelo estatístico e métodos computacionais apropriados, comunicar os resultados da análise.
- Estas habilidades não são usualmente ensinadas em disciplinas tradicionais de Estatística ou Computação, mas são adquiridas por meio da experiência e colaboração com outros pesquisadores.

## CD: Perspectiva Humana

- CD liga modelos estatísticos e métodos computacionais para resolver problemas específicos de outras disciplinas.
- Entender o domínio de um problema, decidir quais dados obter, como processá-los, explorar e visualizar os dados, selecionar um modelo estatístico e métodos computacionais apropriados, comunicar os resultados da análise.
- Estas habilidades não são usualmente ensinadas em disciplinas tradicionais de Estatística ou Computação, mas são adquiridas por meio da experiência e colaboração com outros pesquisadores.

## Aprendizado com Estatística

- O Aprendizado com Estatística (AE) pode ser **supervisionado** ou **não supervisionado**.
- No AE supervisionado, o objetivo é prever o valor de uma variável resposta (*outcome*) a partir de variáveis preditoras (*inputs*).
- A variável resposta pode ser quantitativa ou qualitativa. No caso de variáveis respostas quantitativas, um dos modelos estatísticos mais utilizados é o de **regressão**; quando a variável resposta é qualitativa, utilizam-se geralmente modelos de **regressão logística** para a análise.
- Adicionalmente, para variáveis qualitativas (categóricas ou discretas), com valores em um conjunto finito, os modelos mais comuns são os de classificação, em que a partir de um conjunto  $(x_i, y_i), i = 1 \dots, N$  de dados, chamado de **conjunto de treinamento**, obtemos, por exemplo, previsões para a variável resposta,  $y_i$ , ou uma regra de classificação, dependendo de ela ser quantitativa ou qualitativa.

## Aprendizado com Estatística

- Outros problemas em AE supervisionado envolvem expansões em bases de funções, como polinômios, *splines*, ondaletas, regressão logística não linear, modelos aditivos generalizados (MAG), modelos de regressão por suporte vetorial *support vector regression*, redes neurais etc.
- No caso de AE não supervisionado, temos apenas um conjunto de variáveis preditoras (*inputs*) e o objetivo é descrever associações e padrões entre essas variáveis. Nesse caso, não há uma variável resposta.
- Dentre as técnicas mais utilizadas nesta situação temos a **análise de agrupamentos**, a **análise de componentes principais** e a **análise de componentes independentes** (ambas proporcionando a redução da dimensionalidade dos dados), *self-organizing methods*, **escalamento multidimensional** etc.

## Aprendizado Automático=ML

- Inteligência Artificial (IA) é um tópico de extremo interesse e que aparece frequentemente nas mídias escritas e faladas. Normalmente o termo suscita questões do tipo: computadores no futuro vão se tornar inteligentes e a raça humana será substituída por eles? Ou que todos perderemos nossos empregos, por que seremos substituídos por robôs inteligentes? Pelo menos até o presente esses receios são infundados.
- Acredita-se que o artigo de Turing (1950) seja o primeiro a tratar do tema. A primeira frase do artigo diz:  
*I propose to consider the question, "Can machines think?"*
- De modo informal, a IA é um esforço para automatizar tarefas intelectuais usualmente realizadas por seres humanos

## Aprendizado Automático=ML

- A IA está intimamente ligada ao desenvolvimento da computação (ou programação de computadores) e até a década de 1980, a IA era entendida como na programação clássica: temos um sistema computacional (SC) (um computador ou um *cluster* de computadores ou nuvem etc.) no qual se alimentam dados e uma regra de cálculo e se obtém uma resposta.
- Exemplo: regressão, usando-se MQ para se obter os EMQ. Regra de cálculo é um algoritmo que resolve o problema e pode ser programado em alguma linguagem (Fortran, C, S etc.). A maioria dos pacotes computacionais existentes funcionam dessa maneira.
- A partir da década de 1990, o aprendizado com máquina (ou **aprendizado automático** (AA)) criou um novo paradigma. A programação clássica não resolve problemas mais complicados, como reconhecimento de imagens, voz, escrita etc.

## Aprendizado Automático=ML

- Então a ideia é **treinar** um SC no lugar de programá-lo. Isso significa que se apresentam muitos exemplos relevantes a determinada tarefa (**dados de treinamento**) ao SC, de modo que esse encontre uma estrutura estatística nesses exemplos, produzindo uma regra automatizada. Ou seja, no AA, a entrada é constituída de dados e respostas, e a saída é uma regra de cálculo. Com um novo conjunto de observações (**dados de teste**) procura-se obter a eficácia do método segundo algum critério.
- Existem atualmente, muitos procedimentos que são usados em AA (ou em AE): SVM (*support vector machines*), métodos baseados em árvores de decisão (árvores, florestas, *bagging*, *boosting*), redes neurais etc. O objetivo é obter algoritmos que tenham um alto valor preditivo em problemas de regressão, agrupamento, classificação e previsão.
- AA está fortemente relacionado com Estatística Computacional, que também trata de fazer previsões com o auxílio de computador. Tem relação forte com otimização, que fornece métodos, teoria e aplicações a este campo.

## Redes Neurais

- As contribuições pioneiras para a área de Redes Neurais (RN) (também denominadas redes neurais) foram as de McCulloch e Pitts (1943), que introduziram a ideia de RN como máquinas computacionais, de Hebb (1949), por postular a primeira regra para aprendizado organizado e Rosenblatt (1958), que introduziu o *perceptron*, como o primeiro modelo de aprendizado supervisionado.
- O algoritmo do perceptron (programado para o IBM 704) foi implementado por uma máquina, chamada Mark I, planejada para reconhecimento de imagens. O modelo consiste de uma combinação linear das entradas, incorporando um viés externo. A soma resultante é aplicada a um limitador, na forma de uma função degrau (ou uma sigmóide).
- Se  $\mathbf{x} = (+1, x_1, x_2, \dots, x_p)^\top$  contém as entradas,  $\mathbf{w} = (b, w_1, w_2, \dots, w_p)^\top$  são os pesos, a saída é dada por

$$v = \sum_{i=0}^p w_i x_i = \mathbf{w}^\top \mathbf{x}.$$

## Redes Neurais

- Atualmente, a RN mais simples consiste de entradas, de uma camada intermediária escondida e das saídas.
- Sejam  $\mathbf{X} = (X_1, \dots, X_p)^\top$ ,  $\mathbf{Y} = (Y_1, \dots, Y_M)^\top$  e  $\mathbf{W} = (W_1, \dots, W_K)^\top$  e sejam, os vetores de pesos  $\alpha_j$ ,  $j = 1, \dots, M$ ,  $\beta_k$ ,  $k = 1, \dots, K$ , de ordens  $p \times 1$  e  $M \times 1$ , respectivamente.

Essa rede neural simples pode ser representada pelas equações:

$$Y_j = f(\alpha_{0j} + \alpha_j^\top \mathbf{X}), j = 1, \dots, M, \quad (1)$$

$$W_k = \beta_{0k} + \beta_k^\top \mathbf{Y}, k = 1, \dots, K, \quad (2)$$

$$f_k(\mathbf{X}) = g_k(\mathbf{W}), k = 1, \dots, K. \quad (3)$$

- A função  $f$  é chamada **função de ativação** e geralmente é usada a sigmóide  $f(x) = 1/(1 + e^{-x})$ .

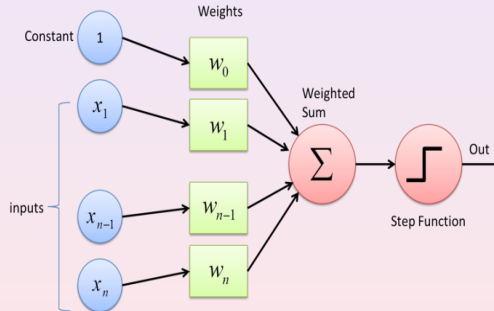
Os pesos  $\alpha_{0j}$  e  $\beta_{0k}$  têm o mesmo papel de  $b$  no perceptron e representam vieses. A saída final é  $g_k(\mathbf{W})$ . Em problemas de regressão,  $g_k(\mathbf{W}) = W_k$  e em problemas de classificação,  $g_k(\mathbf{W}) = e^{W_k} / \sum_i e^{W_i}$ , que corresponde a uma logística multidimensional.

Os  $Y_i$  constituem a camada escondida e não são observáveis.

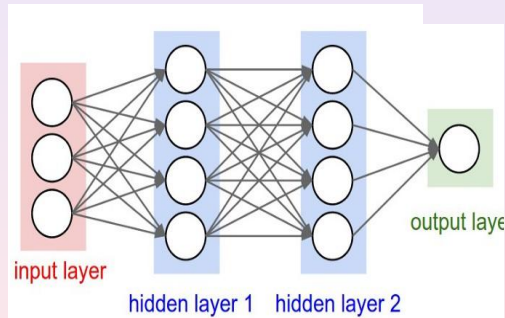
## Redes Neurais

- O ajuste de modelos de RN é feito minimizando a soma dos quadrados dos resíduos, no caso de regressão, onde a minimização é sobre os pesos. No caso de classificação, usamos a entropia. Nos dois casos é usado um algoritmo chamado de **backpropagation**. É necessário escolher valores iniciais e regularização (usando uma função penalizadora), porque o algoritmo de otimização é não convexo e instável.
- No caso de termos várias camadas intermediárias obtém-se o que é chamado aprendizado profundo (*deep learning*). A complexidade do algoritmo é proporcional ao número de observações, número de preditores, número de camadas e número de épocas de treinamento. Para detalhes sobre esses tópicos, veja Hastie et al. (2017) e Cholet (2018).
- Leo Breiman (U.C. Berkeley) distingue dois paradigmas em modelagem estatística: *data model* e *algorithmic model*, onde o segundo engloba os algoritmos usados em ML.

## Redes Neurais



# Redes Neurais



## Sugestões para BE

- um currículo balanceado: mais ênfase em coleta de dados, modelagem, computação para conjuntos de dados grandes e complexos (Análise Exploratória de Dados).
- treinamento inter-disciplinar, um "minor" em outras disciplinas.
- conhecimento de outras ciências: física, biologia, computação etc.
- uma disciplina em Introdução à Ciência de Dados.

## Disciplinas para Mestrado em Estatística com Ênfase em Data Science

- 1) Introdução à Data Science
- 2) Probabilidade e Inferência I (do mestrado em Estatística)
- 3) Disciplina da Computação ( Algoritmos para Data Science, Computer Systems for DS)
- 4) Machine/Statistical Learning

Poderiam ser 2 disciplinas, uma dada pelo Dep. Estatística, com base no livro de Hastie et al. (2009), e outra pelo Dep. Computação.

- 5) Big Data

Seria uma disciplina sobre Análise de Dados, univariados e multivariados, com uso de pacotes do R usuais e outros já desenvolvidos para DS.

## Referências

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Second Edition. Springer

Efron, B. (2010). *Large-Scale Inference*. Cambridge University Press.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P.A. and Singer, J.M. (2019). *Introdução à Ciência de Dados*. Versão Preliminar. IME-USP.