Advanced Institute for Artificial Intelligence

Introdução à Estatística

https://advancedinstitute.ai

Introdução à Estatística

Análise Bidimensional

Referências e Fontes das Imagens

- Estatística Básica (Book)
- Think Stats (Book)
- Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python (Book)

Introdução

- Frequentemente estamos interessados em analisar o **comportamento conjunto de duas ou mais variáveis aleatórias**
- Encontrar as **possíveis relações ou associações entre as duas variáveis**
  - Detectadas por meio de **métodos gráficos** e **medidas numéricas**

Introdução

- Frequentemente estamos interessados em analisar o **comportamento conjunto de duas ou mais variáveis aleatórias**
- Encontrar as **possíveis relações ou associações entre as duas variáveis**
  - Detectadas por meio de **métodos gráficos** e **medidas numéricas**
  - E.g.: existe relação entre a altura de pessoas e a região onde essa pessoa nasceu?

Introdução

- Frequentemente estamos interessados em analisar o **comportamento conjunto de duas ou mais variáveis aleatórias**
- Encontrar as **possíveis relações ou associações entre as duas variáveis**
  - Detectadas por meio de **métodos gráficos** e **medidas numéricas**
  - E.g.: existe relação entre a altura de pessoas e a região onde essa pessoa nasceu?
    - Qual a frequência esperada de uma pessoa dessa população ter, digamos, mais de 170 cm?
    - Qual a frequência esperada de alguém nascido no Nordeste (ou no Sul) ter mais de 170 cm?

Introdução

- Frequentemente estamos interessados em analisar o **comportamento conjunto de duas ou mais variáveis aleatórias**
- Encontrar as **possíveis relações ou associações entre as duas variáveis**
  - Detectadas por meio de **métodos gráficos** e **medidas numéricas**
  - E.g.: existe relação entre a altura de pessoas e a região onde essa pessoa nasceu?
    - Qual a frequência esperada de uma pessoa dessa população ter, digamos, mais de 170 cm?
    - Qual a frequência esperada de alguém nascido no Nordeste (ou no Sul) ter mais de 170 cm?
    - Respostas diferentes indicam **uma provável associação**

Introdução

- Frequentemente estamos interessados em analisar o **comportamento conjunto de duas ou mais variáveis aleatórias**
- Encontrar as **possíveis relações ou associações entre as duas variáveis**
  - Detectadas por meio de **métodos gráficos** e **medidas numéricas**
  - E.g.: existe relação entre a altura de pessoas e a região onde essa pessoa nasceu?
    - Qual a frequência esperada de uma pessoa dessa população ter, digamos, mais de 170 cm?
    - Qual a frequência esperada de alguém nascido no Nordeste (ou no Sul) ter mais de 170 cm?
    - Respostas diferentes indicam **uma provável associação**
- Incorporar conhecimento para **melhorar o entendimento sobre os comportamentos das variáveis**;

Introdução

- Conhecer o **grau de dependência entre duas variáveis**
  - Prever melhor o resultado de uma delas ao **conhecer a outra**;
  - E.g.: Estimar a renda média de uma família de São Paulo com a informação adicional sobre a classe social a que ela pertence;
    - **Dependência entre as duas variáveis**: renda familiar e classe social
- Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter três situações:
  - As duas variáveis são qualitativas
  - As duas variáveis são quantitativas; e
  - Uma variável é qualitativa e outra é quantitativa;

Introdução

- Conhecer o **grau de dependência entre duas variáveis**
  - Prever melhor o resultado de uma delas ao **conhecer a outra**;
  - E.g.: Estimar a renda média de uma família de São Paulo com a informação adicional sobre a classe social a que ela pertence;
    - **Dependência entre as duas variáveis**: renda familiar e classe social
- Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter três situações:
  - As duas variáveis são qualitativas
  - **As duas variáveis são quantitativas;** e
  - Uma variável é qualitativa e outra é quantitativa;

Gráficos de Dispersão (*Scatterplots*)

- A maneira **mais simples de verificar a relação** entre duas variáveis é um **gráfico de dispersão**;

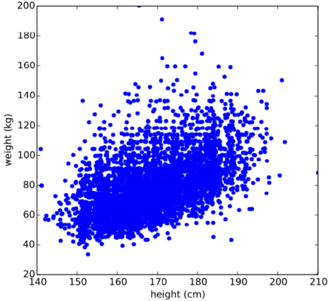
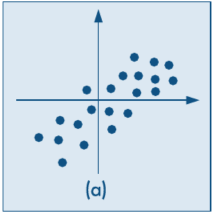
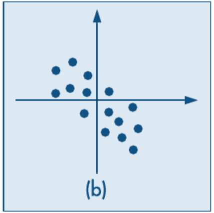


Figure: Gráfico de Dispersão - Peso vs Altura

Gráficos de Dispersão (*Scatterplots*)

- Tipos de associações entre duas variáveis
  - (a) Associação linear direta (ou **positiva**)
    - Soma do produto das coordenadas será sempre positivo
  - (b) Dependência linear inversa (ou **negativa**)
    - Soma dos produtos das coordenadas será negativa





Medidas de Dependência

- Indica que conforme **uma variável muda de valor**, a **outra variável tende a mudar em uma direção específica**;
  - Possível usar o **valor de uma variável para prever o valor da outra**;
- **Covariância**: uma medida da tendência de duas variáveis variarem juntas;
  - Possui unidade;
  - Difícil de interpretar, e.g., 113 quilogramas-centímetros (???)
- **Correlação**: quantificar a força da relação entre duas variáveis;
  - Normalização pelo desvio padrão;
  - Sem unidade associada;

Covariância

- Utilizamos os desvios: $dx_i = x_i - \overline{x}$
- Se $\mathcal{X}$ e $\mathcal{Y}$ variam juntos, seus desvios tendem a ter o mesmo sinal
- Se os multiplicarmos $dx_i\ dy_i$, o produto é positivo quando os desvios têm o mesmo sinal e negativo quando têm sinais opostos;
- Somar os produtos dá uma medida da tendência de variar em conjunto;
  - Normalizar pelo tamanho da amostra

Covariância

- Utilizamos os desvios: $dx_i = x_i - \overline{x}$
- Se $\mathcal{X}$ e $\mathcal{Y}$ variam juntos, seus desvios tendem a ter o mesmo sinal
- Se os multiplicarmos $dx_i\ dy_i$, o produto é positivo quando os desvios têm o mesmo sinal e negativo quando têm sinais opostos;
- Somar os produtos dá uma medida da tendência de variar em conjunto;
  - Normalizar pelo tamanho da amostra

$$Cov(\mathcal{X}, \mathcal{Y}) = \frac{1}{n} \sum_{i=1}^{n} dx_i\ dy_i$$

Correlação

- Normalização da covariância pelo desvio padrão;
- Produção de medida sem unidade;
  - Comparação entre dois pares de variáveis de unidades diferentes;
- Cálculo do Z-score
  - Variação entre -1 e 1;
- **Correlação de Pearson**
  - Dependência **Linear** (!!!)

Correlação

- Normalização da covariância pelo desvio padrão;
- Produção de medida sem unidade;
  - Comparação entre dois pares de variáveis de unidades diferentes;
- Cálculo do Z-score
  - Variação entre -1 e 1;
- **Correlação de Pearson**
  - Dependência **Linear** (!!!)

$$Corr(\mathcal{X}, \mathcal{Y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{dp(\mathcal{X})} \right) \left(\frac{y_i - \overline{y}}{dp(\mathcal{Y})} \right) = \frac{Cov(\mathcal{X}, \mathcal{Y})}{dp(\mathcal{X})\ dp(\mathcal{Y})}$$

Relações não Lineares

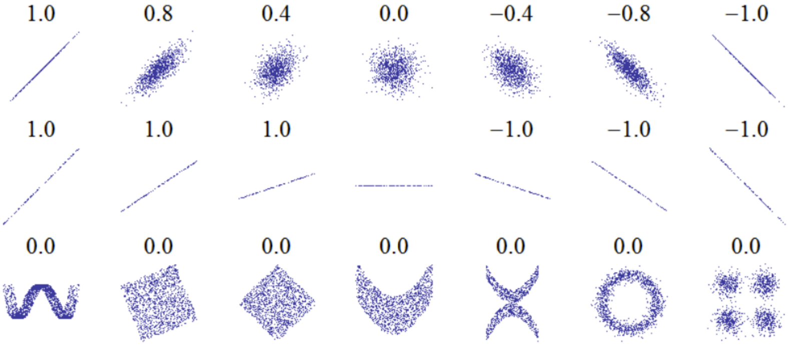


Figure: Exemplos de Correlações

Correlação e Causalidade

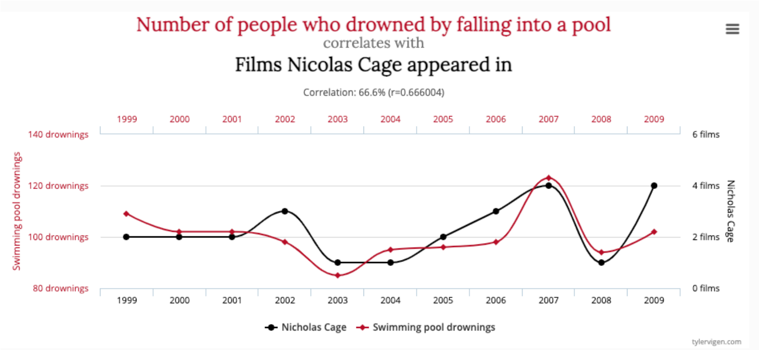
- Erro comum a ser evitado;

Correlação e Causalidade

- Erro comum a ser evitado;
- ***Correlation does not imply causation!***

Correlação e Causalidade

- Erro comum a ser evitado;
- ***Correlation does not imply causation!***



Introdução à Estatística

Distribuições de Probabilidade

Probabilidades

- Distribuição de frequências é importante para avaliarmos a variabilidade das observações de um fenômeno;
  - Medidas de posição e variabilidade;
  - **Estimativas de quantidades desconhecidas**, associadas a populações das quais os dados foram extraídos na forma de **amostras**;
- Frequências (relativas) são **estimativas de probabilidades** de ocorrências de certos eventos;
- Criar um modelo teórico que **reproduza de maneira razoável a distribuição das frequências** de quando o fenômeno é observado diretamente;

Probabilidades

- **Espaço amostral** Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão: $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$

Probabilidades

- **Espaço amostral** Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão: $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$
- **Probabilidade,** $P(\omega)$, para cada ponto amostral. A probabilidade do que chamaremos de um evento aleatório ou simplesmente evento.

Probabilidades

- **Espaço amostral** Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão: $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$
- **Probabilidade,** $P(\omega)$, para cada ponto amostral. A probabilidade do que chamaremos de um evento aleatório ou simplesmente evento.
- E.g.: Lançamos uma moeda duas vezes. Se *C* indicar cara e *R* indicar coroa, então um espaço amostral será: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, sendo $\omega_1 = (C, C)$, $\omega_2 = (C, R)$, $\omega_3 = (R, C)$ e $\omega 4 = (R, R)$

Probabilidades

- **Espaço amostral** Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão: $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$
- **Probabilidade,** $P(\omega)$, para cada ponto amostral. A probabilidade do que chamaremos de um evento aleatório ou simplesmente evento.
- E.g.: Lançamos uma moeda duas vezes. Se *C* indicar cara e *R* indicar coroa, então um espaço amostral será: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, sendo $\omega_1 = (C, C)$, $\omega_2 = (C, R)$, $\omega_3 = (R, C)$ e $\omega 4 = (R, R)$
- No caso de querermos descobrir a probabilidade do evento $\mathcal{A}$ que consiste de termos duas faces iguais, teríamos:

$$P(\mathcal{A} = P(\{\omega_1, \omega_4\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Função Probabilidade

Exemplo

Um empresário pretende estabelecer uma firma para montagem de um produto composto de uma esfera e um cilindro. As partes são adquiridas em fábricas diferentes (A e B), e a montagem consistirá em juntar as duas partes e pintá-las. O produto acabado deve ter o comprimento (definido pelo cilindro) e a espessura (definida pela esfera) dentro de certos limites, e isso só poderá ser verificado após a montagem. Para estudar a viabilidade de seu empreendimento, o empresário quer ter uma ideia da distribuição do lucro por peça montada.

Função Probabilidade

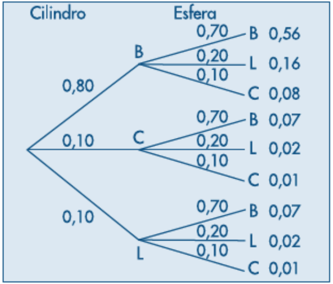
Exemplo - Cont.

Sabe-se que cada componente pode ser classificado como bom, longo ou curto, conforme sua medida esteja dentro da especificação, maior ou menor que a especificada, respectivamente. Além disso, foram obtidos dos fabricantes o preço de cada componente ($5,00) e as probabilidades de produção de cada componente com as características bom, longo e curto. Se o produto final apresentar algum componente com a característica C (curto), ele será irrecuperável, e o conjunto será vendido como sucata ao preço de $5,00. Cada componente longo poderá ser recuperado a um custo adicional de $5,00. Se o preço de venda de cada unidade for de $25,00, como seria a distribuição de frequências da variável X: lucro por conjunto montado?

Função Probabilidade

Produto		Fábrica A Cilindro	Fábrica B Esfera
Dentro das especificações	bom (B)	0,80	0,70
Maior que as especificações	longo (L)	0,10	0,20
Menor que as especificações	curto (C)	0,10	0,10

Função Probabilidade



Produto	Probabilidade	Lucro por montagem (X)
BB	0,56	15
BL	0,16	10
BC	0,08	–5
LB	0,07	10
LL	0,02	5
LC	0,01	–5
CB	0,07	–5
CL	0,02	–5
CC	0,01	–5

Função Probabilidade

- $\mathcal{X}$ pode assumir um dos seguintes valores:
  - **15**, se ocorrer o evento $A_1 = \{BB\}$;
  - **10**, se ocorrer o evento $A_2 = \{BL, LB\}$;
  - **5**, se ocorrer o evento $A_3 = \{LL\}$;
  - **–5**, se ocorrer o evento $A_4 = \{BC, LC, CB, CL, CC\}$
- Cada um desses eventos tem uma probabilidade associada:
  - $P(A_1) = 0,56, \;\; P(A_2) = 0,23, \;\; P(A_3) = 0,02, \;\; P(A_4) = 0,19$

Função Probabilidade

- A função $(x, p(x))$ é chamada função de probabilidade da v.a. $\mathcal{X}$:

x	$p(x)$
15	0,56
10	0,23
5	0,02
–5	0,19
Total	1,00

Função Probabilidade

Valor Médio de uma Variável Aleatória

- Qual o lucro médio por conjunto montado que o empresário espera conseguir?

$$(0,56)(15) + (0,23)(10) + (0,02)(5) + (0,19)(5) = 9,85.$$

Função Probabilidade

Valor Médio de uma Variável Aleatória

- Qual o lucro médio por conjunto montado que o empresário espera conseguir?

$$(0,56)(15) + (0,23)(10) + (0,02)(5) + (0,19)(5) = 9,85.$$

- Dada a v.a. $\mathcal{X}$ discreta, assumindo os valores $x_1, \ldots, x_n$, chamamos valor médio ou esperança matemática de $\mathcal{X}$ ao valor

$$E(X) = \sum_{i=1}^{n} x_i P(\mathcal{X} = x_i) = \sum_{i=1}^{n} x_i p_i$$

Função Densidade de Probabilidade

- Para o caso de variáveis contínuas
- Cálculo de probabilidade para um dado intervalo;
- Valor $=0$ em um ponto arbitrariamente pequeno
- Teoricamente, qualquer função f, que seja não negativa e cuja área total sob a curva seja igual à unidade, **caracterizará uma v.a. contínua**;
- E.g., Considerando $f(x) = 2x$, a probabilidade de $\mathcal{X}$ assumir um valor menor que $1/2$ é:

$$P(0 \leq X \leq 1/2) = \frac{1}{2}\left(\frac{1}{2} \times 1\right) = \frac{1}{4}$$

Função Densidade de Probabilidade

Valor médio de uma v.a. contínua

- Sendo $f()$, não negativa e $\int_{-\infty}^{\infty} f(x)dx = 1$, dizemos que f define a v.a. contínua $\mathcal{X}$
- Podemos dizer também que $P(a \leq \mathcal{X} \leq b) = \int_{a}^{b} f(x)dx$
- Por completude, temos que o valor médio da v.a. $\mathcal{X}$ é $E(\mathcal{X}) = \int_{-\infty}^{\infty} xf(x)dx$
- Por extensão temos a variância para uma v.a. contínua $\mathcal{X}$ definida como:

$$Var(\mathcal{X}) = E[((X) - E(\mathcal{X}))^2] = \int_{-\infty}^{\infty} (x - E(\mathcal{X}))^2 f(x)dx.$$

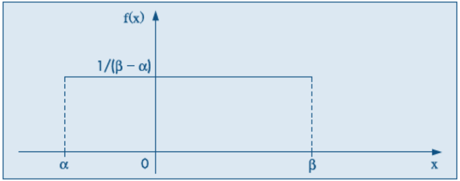
Introdução à Estatística

Modelos Probabilísticos para Variáveis Aleatórias Contínuas

Distribuição de Probabilidade Contínua

- A v.a. $\mathcal{X}$ tem distribuição uniforme no intervalo $[\alpha, \beta]$ se sua função densidade de probabilidade é dada por:

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta-\alpha}, & \text{se } \alpha \leqslant x \leqslant \beta, \\ 0, & \text{caso contrário.} \end{cases}$$



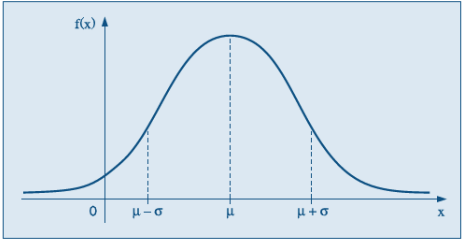
- $E(\mathcal{X}) = \dfrac{\alpha+\beta}{2}$
- $Var(\mathcal{X}) = \dfrac{(\beta-\alpha)^2}{2}$

Distribuição de Probabilidade Normal

- A v.a. $\mathcal{X}$ tem distribuição normal com parâmetros μ e σ^2, $-\infty \leq \mu \leq \infty$, $0 \leq \sigma^2 \leq \infty$ e $-\infty \leq x \leq \infty$ se sua função densidade de probabilidade é dada por:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- $E(\mathcal{X}) = \mu$
- $Var(\mathcal{X}) = \sigma^2$
- $\mathcal{X} \sim N(\mu, \sigma^2)$



Distribuição de Probabilidade Normal

- Normal Padrão ($\mu = 0$, $\sigma^2 = 1$)
  - Função Densidade de Probabilidades:

$$\Phi(z) = \frac{1}{2\sqrt{(2\pi)}} e^{-z^2/2}, \; -\infty < z < \infty$$

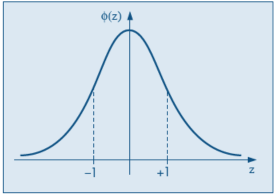


Figure: Função Densidade de Probabilidades para Normal Padrão ($\mathcal{Z} \sim N(0, 1)$)

Dúvidas?