Estatística

Conceitos básicos

População – É constutuida por todos os elementos que são passíveis de ser analisados de tamanho N

Amostragem – Subconjunto da população que é efectivamente analisado com um dado tamanho n

Amostra aleatória – Amostra onde cada elemento da população têm hipóteses iguais de ser escolhido para a amostra.

Estatística descritiva – Inclui gráficos e procedimentos numéricos que permitem sumarizar e processar dados por forma a transformá-los em informação.

Inferência Estatística – fornece as bases para prevêr, estimar e permite transformar informação em conhecimento

Estatística Descritiva

Classificação de variáveis
As variáveis podem ser numéricas ou por categorias.
No caso das numéricas existem variáveis discretas e continuas.

Distribuição de frequências
Tabela que têm na coluna esquerda intervalos e na coluna direita a frequência ou número de observações. Os intervalos são habitualmente do mesmo tamanho, abrangerem todo o intervalo e não serem sobrepostos

Tamanho da amostra vs. número de classes intervalos:

Até 50	5 a 6
50 a 100	6 a 8
> 100	8 a 10

Distribuição cumulativa de frequência – Mostra na coluna da direita o número total de observações cujo valores são menores do que o limite superior do intervalo.
Distribuição cumulativa da frequência – Mostra na coluna da direita o valor em percentagens acumuladas

Histograma – É um gráfico que consiste em barras verticais mostrando a frequência.
Ogiva – É uma linha onde são mostrados a percentagem relativa acumulada e que pode estar sobreposta à ogiva.

Caule e Folha – Diagrama alternativo ao histograma que permite manter informação sobre as observações individuais. Os números inicias são mantidos na coluna da esquerda e os restantes números surgem na coluna da direita

Diagrama de pareto – É um gráfico de barras que mostra as causas principais mostrando à esquerda as ocorrências mais frequentes.

Gráficos de linha - Pode mostrar a evolução de valores no tempo

Medidas de tendência central

Estatística – é uma medida descritiva calculado a partir de uma amostra
Parâmetro – é uma medida descrtiva calculada a partir da população

Média aritmética

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{x_1 + x_1 + ... + x_n}{n}$$

Média da população

$$\overline{\mu} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{x_1 + x_1 + ... + x_n}{N}$$

Mediana – A mediana é o valor para o qual 50% das observações são inferiores e 50% das observações são superiores. No caso da amostra conter um número par de observações a mediana é a média das 2 observações do meio.

Moda – A moda se existir é a observação mais frequente.

Forma da distribuição
Simetria – A forma de uma distribuição é simétrica se as observações forem balanceadas ou distibuidas de forma uniforme à volta da média.

Enviesamento - Uma distribuição é enviadas se as observações estão distribuidas de forma não simétrica. Um enviesamento positivo corresponde a a uma distribuição onde existem mais observações à esquerda da média

Média geométrica

$$\overline{X}_g = \sqrt[n]{x_1 * x_1 * ... * x_n}$$

É usada em especial para calcular médias de crescimentos

Medidas de variabilidade

Intervalo de variação - Diferença entre a observação maior e a observação menor

Variância simples – É a soma das diferenças quadradas entre cada observação e a média simples dividida pelo tamanho da amostra menos 1

$$s^2 = \frac{\sum_{i=1}^{n}(x_i + \overline{X})^2}{n-1}$$

Variância da população

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i + \mu)^2}{N}$$

Desvio padrão simples – É a raíz quadrada positiva da variância

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i + \overline{X})^2}{n-1}}$$

Desvio padrão da população

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i + \mu)^2}{N}}$$

Regra Empírica

Numa distribuição normal cerca de 68% das observações estão a desvio padrão da média, 95% estão a dois desvios padrões da média e quase todas as observações estão a 3 desvios padrões da média.

Coeficiente da Variação - É uma medida da dispersão relativa que exprime o desvio padrão como uma percentagem da média.

Coeficiente de variação simples

$$CV = \frac{s}{\overline{X}} x100\% \text{ , se } \overline{X} > 0$$

Coeficiente de variação da população:

$$CV = \frac{\sigma}{\mu} x100\% \text{ , se } \mu > 0$$

Percentis e quartis

Os percentis dividem as observações em centésimos e os quartis em quartos.

$$Q_1 = \frac{(n+1)}{4} \text{ e } Q_3 = \frac{3(n+1)}{4}$$

Intervalo interquartil - Diferença entre o terceiro e o primeiro quartil

Box and Whisker – Gráfico no qual são mostrados os valores de 5 medidas contendo: Uma caixa interna que vai do 1º ao 3º quartil. Uma linha que é desenhada na caixa correspondendo à mediana.
Os bigodes são as linhas 1º quartil ao mínimo e do 3º quartil ao máximo.

Descrição sumária de relações entre variáveis

Scatter Plot – Permite mostrar os valores por cada par de variáveis

Covariância simples

$$Cov(x, y) = S_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

Coeficiente de correlação simples

$$r_{x,y} = \frac{Cov(x, y)}{s_x s_y}$$

Relações lineares

$$Y = \beta_0 + \beta_1 X$$

Probabilidades

Experiência aleatória – É um processo que pode levar a dois ou mais resultados com incerteza sobre qual o resultado que irá ocorrer.
Espaço da amostra – Conjunto de resultados possíveis da experiência

Evento – Sub conjunto de resultados possíveis
Eventos mutuamente exclusivos – São eventos onde a ocorrência de um implica que o outro não ocorre
Eventos colectivamente exaustivos – São eventos que no seu conjunto abarcam todo o espaço de resultados.
Eventos complementares – São os eventos dentro dum espaço de amostra que não pertencem ao evento do qual são complementares

Definição clássica de probabilidade - É a proporção de vezes que um evento ocorrer assumindo que a possibilidade de ocorrer qualquer resultado é igual.

$$P(A) = \frac{N_A}{N}$$

Número de combinações – x itens tomados k de cada vez

$$C_x^n = \frac{n!}{x!(n-x)!}$$

Número de permutações – x itens tomados de n

$$P_x^n = \frac{n!}{(n-x)!}$$

Número de ordenações possíveis – x!

Definição subjectiva de probabilidade – É o grau em que um individuo acredita que um evento posso ocorrer.

Postulados das probabilidades –

$$0 \le P(A) \le 1$$

$$P(A) = \sum_i P(O_i)$$

P(S) =1

Regras das probabilidades –
Complementaridade :

$$P(\overline{A}) = 1 - P(A)$$

Adicção:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Condicionada:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(A)}$$

Independência estatística:

$$P(A \cap B) = P(A)P(B)$$

Probabilidades conjuntas – probabilidade de dois eventos acontecerem em simultâneo
Probabilidades marginais – probabilidade de um evento ocorrer dado que ocorre outro

Teorema de Bayes –

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Variáveis discretas aleatórias e distribuições de probabilidades

Uma variável aleatória é uma variável que assume um valor numérico determinado por uma experiência aleatória.

Uma variável aleatória é **discreta** se só pode assumir uma quantidade numerável de valores.
Uma variável aleatória é **continua** se pode assumir qualquer valor dentro de um intervalo.

Função de distribuição de probabilidade
P(x)=P(X=x)

Propriedades
$P(x) \geq 0$

$\sum_{x} P(x) = 1$

Função de probabilidade acumulada
$F(x_0) = P(X \leq x_0)$

$F(x_0) \geq 0$

$F(x_0) = \sum_{x \leq x_0} P(x)$

Valor esperado
$E(x) = \sum_{x} xP(x)$

Variância
$\sigma_x^2 = E\left[(X - \mu_x)^2\right] = \sum_{x} (x - \mu_x)^2 P(x)$

O desvio padrão é a raíz quadrada positiva da variância

Propriedade de funções lineares de uma variável aleatória

$\mu_Y = E(a + bX) = a + b\mu_x$

$\sigma_Y^2 = Var(a + bX) = b^2\sigma_x^2$

$\sigma_Y = b\sigma_x$

Média e variância normalizada

$Z = \frac{X - \mu_x}{\sigma_x}$

E(Z)=0
Var(Z)=1

Distribuição de Bernoulli

$E(X) = \pi$

$\sigma_x^2 = \pi(1-\pi)$

Distribuição binomial

$E(X) = n\pi$

$\sigma_x^2 = n\pi(1-\pi)$

Distribuição hiper geométrica – Probabilidade de tirar n objectos de N onde S é a probabilidade de sucesso

$$P(x) = \frac{C_x^S C_{n-x}^{N-S}}{C_n^N}$$

Distribuição de Poisson

1. Assume-se que a probabilidade de ocorrência é igual em diversos intervalos
2. Não pode haver mais que uma ocorrência em cada subintervalo
3. As ocorrências são independentes

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \ para\ x = 0,1,2,...$$

$\mu_x = \lambda$

$\sigma_x^2 = \lambda$

Soma de variáveis aleatórias

$E(X+Y) = \mu_x + \mu_y$

$E(X-Y) = \mu_x - \mu_y$

$Var(X+Y) = \sigma_x^2 + \sigma_y^2 + 2\,\mathrm{cov}(X,Y)$

Variáveis aleatórias continuas e distribuições de probabilidade

$F(x) = P(X \leq x)$

$P(a<X<b)=F(b)-F(a)$

Função de densidade de probabilidade

$f(x) \geq 0$

$F(x_0) = \int_{x_m}^{x_0} f(x)\,dx$

$\int f(x)\,dx = 1$

Distribuição normal

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

Propriedades da distribuição normal

$E(X) = \mu$

$Var(X) = \sigma^2$

$X \sim N(\mu, \sigma^2)$

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)$$

Aproximação de uma distribuição binomial a uma distribuição normal

Se $n\pi(1-\pi) > 9$ então :

$$P(a \leq X \leq b) = P\left(\frac{a-n\pi}{\sqrt{n\pi(1-\pi)}} < Z < \frac{b-n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

Se $5 < n\pi(1-\pi) < 9$ então :

$$P(a \leq X \leq b) = P\left(\frac{a-0.5-n\pi}{\sqrt{n\pi(1-\pi)}} < Z < \frac{b-0.5-n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

Distribuição exponencial

$f(t) = \lambda e^{-\lambda t}$

$F(t) = 1 - e^{-\lambda t}$, $E(t) = 1/\lambda$, $\sigma^2 = 1/\lambda^2$