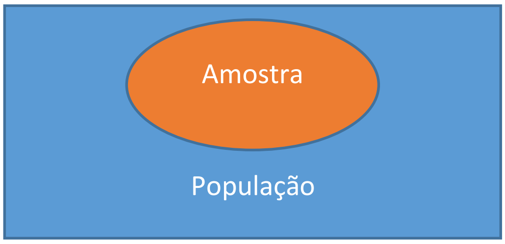
Resumo Probabilidade e Estatística



Variáveis → características de interesse da pesquisa

- Variáveis:
  - Qualitativas: não numéricas
    - Nominais: não é possível estabelecer ordem natural entre seus valores (A ou B, sim ou não)
    - Ordinais: o atributo tem ordenação natural (ruim, média ou boa, pequeno, médio ou grande)
  - Quantitativas: numéricas
    - Discretas: obtidas a partir de contagem (número de filhos)
    - Contínuas: obtidas por mensuração (altura)

Tabelas de frequência: forma resumida da tabela dos dados brutos.

- Variáveis discretas: Consiste em listar os possíveis valores da variável e fazer a contagem do número de ocorrências na tabela de dados brutos.
  - $n_i =$ frequência de ocorrência da variável i e n é a frequência total.
  - Frequência relativa: $f_i = \frac{n_i}{n}$
- Variáveis ordinais: acrescentar a frequência acumulada para termos pontos de corte com uma determinada frequência nos valores das variáveis
  - $f_{ac} =$ somatório das frequências de todos os valores da variável menores ou iguais ao valor considerado

Idade	n_i	f_i	f_{ac}
17	9	0,18	0,18
18	22	0,44	0,62
19	7	0,14	0,76
20	4	0,08	

Classes ou faixas: para variáveis quantitativas contínuas ou discretas

- Ex: de 40kg (inclusive) até 50kg (exclusive) => $40 \vdash 50$ *ou* $[40,50)$

Medidas de resumo: informações numéricas sobre um conjunto de dados

- Medidas de posição (tendência central)
  - Média:

$$\bar{x}_{obs} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Mediana: valor central depois que está ordenado
  Se n for ímpar: $md_{obs} = valorpos.\left(\frac{n+1}{2}\right)$
  Se n for par: $md_{obs} = \frac{valorpos.\left(\frac{n}{2}\right)+valorpos.\left(\frac{n+2}{2}\right)}{2}$
- Moda: é o valor mais frequente, com maior ocorrência.
  - Todos os valores com a mesma frequência de ocorrência: não tem moda
  - K valores tem a mesma frequência de ocorrência: tem k modas

- Medidas de dispersão
  - Amplitude: diferença entre o maior e o menor valor (Δ)
  - Desvio mediano: somatório dos módulos da distância de cada valor até a mediana

$$desvio\ mediano = \frac{1}{n}\sum_{i=1}^{n}|x_i - md_{obs}|$$

  - Desvio médio: somatório dos módulos da distância de cada valor até a média

$$desvio\ médio = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_{obs}|$$

  - Variância: somatório dos quadrados da distância de cada valor até a média:

$$var_{obs} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_{obs})^2$$

  - Desvio padrão: raiz quadrada do somatório dos quadrados da distância de cada valor até a média, ou seja, raiz quadrada da variância

$$dp_{obs} = \sqrt{var_{obs}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_{obs})^2}$$

Probabilidade

Fenômeno aleatório → Situação cujos resultados não podem ser previstos com certeza

Espaço amostral → conjunto de todos os resultados possíveis. Representado por Ω

└> Subconjuntos → eventos

- União de eventos: ocorrência de pelo menos um dos eventos A ou B: $A \cup B$
- Intersecção de eventos: ocorrência simultânea de A e B: $A \cap B$
- Eventos disjuntos: quando não tem elementos em comum: $A \cap B = \emptyset$
- Eventos complementares: se sua união é o espaço amostral e intersecção vazia:
- $A \cup B = \Omega$ e $A \cap B = \emptyset$
- Evento A ocorre mas o B não: $A \cap B^c$
- Nenhum deles ocorre: $A^c \cap B^c$

- Exatamente um deles ocorre: $(A^c \cap B) \cup (A \cap B^c)$
- Probabilidade da união de dois eventos: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Probabilidade condicional: uma informação anterior influencia em posteriores.

- Probabilidade de A dado que ocorreu B: $P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$
    - Se P(B) = 0, então $P(A|B) = P(A)$
- Eventos independentes: se a ocorrência ou não de B não altera a probabilidade de A:
$$P(A \cap B) = P(A)P(B)$$

Probabilidade total: $P(A) = \sum_{i=1}^{n} P(F_i) * P(A|F_i)$

Teorema de Bayes: basicamente é a probabilidade condicional de um cara acontecer dividido pela probabilidade total.

Variáveis aleatórias discretas

Função com valores numéricos, cujo valor é determinado por “fatores de chance”

Modelos:

- Uniforme: uma mesma probabilidade a cada um de seus valores, ou seja: $P(X = x_j) = \frac{1}{k}, \forall j = 1, 2, \dots, k$
- Binomial: quando a variável de interesse só pode assumir 2 valores:

n = nº de tentativas, p = prob. de sucesso em um evento, k = nº de sucessos

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0,1,2, \dots, n$$

OBS: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

- Geométrico: probabilidade depois de x erros sair um acerto: X=0 é sucesso na primeira tentativa, X=1 é sucesso na segunda etc etc, p = prob. de sucesso, k = nº de erros antes do acerto

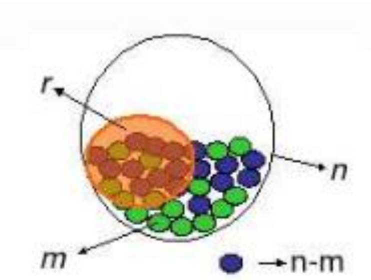
$$P(X = k) = p(1-p)^k$$

- Poisson: unidade de medida é contínua (tempo, área) e a variável aleatória é discreta (nº de chamadas por minuto, nº de clientes por hora).

Sendo k = nº de ocorrências, λ = nº médio de ocorrências em um determinado intervalo:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, k = 0,1,2, \dots$$

- Hipergeométrico: n = população, r = tamanho da amostra, m = sucessos na população, k = sucessos na amostra



$$P(X = k) = \frac{\binom{m}{k}\binom{n-m}{r-k}}{\binom{n}{r}}$$

Medidas de posição para variáveis aleatórias discretas

- Média: $E(X) = \sum_{i=1}^{k} x_i p_i$
- Mediana: $P(X \geq Md) \geq 0{,}5 \; e \; P(X \leq Md) \geq 0{,}5$
- Moda: valor (ou valores) que tem maior probabilidade de ocorrência

Medidas de dispersão

- Variância: (somatório dos desvios relativos a média)² * respectiva probabilidade

$$var(X) = \sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 * p_i$$

- Desvio padrão: raiz quadrada da variância. (σ)

Variáveis bidimensionais

Probabilidade conjunta: a cada par de informações (x_i, y_k), tem uma probabilidade $P(X = x_i, Y = y_k) = p(x_i, y_k)$

Distribuição marginal de probabilidade: é o somatório total das probabilidades.

fuma \ he	0	1	2	3	4	5	6	7	8	9	10	Total
sim	0,02	0	0,02	0	0	0,04	0,02	0	0	0	0,02	0,12
não	0,14	0,06	0,14	0,12	0,08	0,12	0,04	0,1	0,06	0	0,02	0,88
Total	0,16	0,06	0,16	0,12	0,08	0,16	0,06	0,1	0,06	0	0,04	1

– para a variável fuma:

Fuma(X)	P(X)
sim(1)	0,12
não(2)	0,88
Total	1

– para a variável horas semanais de exercícios (h.e.):

h.e.(Y)	0	1	2	3	4	5	6	7	8	9	10	Total
P(Y)	0,16	0,06	0,16	0,12	0,08	0,16	0,06	0,1	0,06	0	0,04	1

Probabilidade condicional: mesma coisa de variáveis com uma dimensão só:

- $P(X = x|Y = y) = \frac{P(X=x,\; Y=y)}{P(Y=y)}, P(Y = y) > 0$
  - Se P(Y=y) = 0, então $P(X = x|Y = y) = P(X = x)$

Independência de variáveis:

Se P(X,Y) = P(X) * P(Y)

Covariância: é o valor esperado do produto dos desvios de cada variável em relação a sua média. Calcula a variância do primeiro (cada um menos a média), depois multiplica a variância do x e do y, e depois calcula a média entre essa variância multiplicada.

Correlação entre variáveis aleatórias: $\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$. Próximo de 1 indica correlação forte!