FIAP

MBA

MBA EM DATA SCIENCE & AI

APPLIED STATISTICS

AULA 3
Probabilidade
Amostragem

Probabilidade Condicional

Em diversas situações práticas, a probabilidade de ocorrência de um evento A se modifica quando dispomos de informação sobre a ocorrência de um outro evento associado.

A probabilidade condicional de A dado B é a probabilidade de ocorrência do evento A, sabido que o evento B já ocorreu. Pode ser determinada dividindo-se a probabilidade de ocorrência de ambos os eventos A e B pela probabilidade de ocorrência do evento B, como é mostrado a seguir:

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)} \ , \ P(B) > 0}$$

Da definição de probabilidade condicional, deduzimos a regra do produto de probabilidades que é uma relação bastante útil:

$$P(A \cap B) = P(A \mid B).P(B) \ , \ P(B) > 0$$

Teorema de Bayes

$$P(A|B) = \frac{P(B|A) \; x \; P(A)}{P(B)}$$

- **P(A|B):** É a probabilidade *a posteriori* de A, ou seja, a probabilidade de A ocorrer, dado que B ocorreu. Após receber novos dados (como pegar a jaca), você atualiza sua crença inicial. Agora, você tem uma nova estimativa do peso.

- **P(B|A):** É a probabilidade de B ocorrer, dado que A ocorreu.

Exemplo – Diagnóstico de uma Doença

Imagine que há uma doença rara que afeta 1% da população. Há um teste para detectar essa doença, e o teste tem uma taxa de verdadeiro positivo (sensibilidade) de 99% e uma taxa de falso positivo de 5%. Se uma pessoa testar positivo, qual é a probabilidade de ela realmente ter a doença?

Exemplo – Diagnóstico de uma Doença

- ***P*(*D*):** Probabilidade de ter a doença = 0.01
- **P(-D):** Já que a probabilidade de ter a doença é 0.01, a probabilidade de não ter a doença é 1-P(D) = 0.99
- **P(Pos|D):** Probabilidade de um teste positivo dado que tem a doença = 0.99
- **P(Pos|-D):** Probabilidade de um teste positivo dado que não tem a doença = 0.05

Exemplo – Diagnóstico de uma Doença

O que o problema quer encontrar é **P(D|Pos)**, desta forma:

$$P(D|Pos) = \frac{P(Pos|D)\ x\ P(D)}{P(Pos)}$$

Exemplo – Diagnóstico de uma Doença

Para calcular a **P(Pos)**, devemos considerar tanto os verdadeiros positivos (resultado é positivo e a pessoa realmente possui a doença) quanto os falsos positivos (resultado é positivo mas a pessoa não possui a doença):

$$\boldsymbol{P(Pos) = P(Pos|D)\ x\ P(D) + P(Pos|-D)\ x\ P(-D)}$$

Substituindo os valores, chega-se que P(Pos) = 0.0594. Substituindo este valor no Teorema de Bayes, temos:

$$\boldsymbol{P(D|Pos) = \frac{0.99\ x\ 0.01}{0.0594} = 0.1667 = 16.67\%}$$

Aplicações de Bayes

1. **Machine Learning:** O teorema bayesiano é utilizado em algoritmos de aprendizado de máquina supervisionados, como por exemplo o Naive Bayes, atuando na determinação das probabilidades das classes.

2. **Análise estatística:** Inferência Bayesiana, com a atualização das probabilidades com base no conhecimento de novas evidências.

Distribuição Binomial

Considere um experimento aleatório consistindo em n tentativas independentes e a probabilidade de ocorrer **sucesso** em cada uma das **n tentativas** é sempre igual a p e de **fracasso é q**, onde **p + q = 1**. A probabilidade de **sucesso e fracasso são as mesmas para cada tentativa**.

Definição: Seja X o número de sucesso em n tentativas, então X pode assumir os valores 0, 1, 2,..., n. Nesta condição a v.a. X tem distribuição Binomial com parâmetro n e p, isto é, $X \sim B(n; p)$.

Considere que se $X \sim B(n; p)$, então a média e a variância de X são definidos por:

a) Média de X: $E(X) = np$.

b) Variância de X: $\sigma^2 = npq$, onde $q = 1 - p$.

Distribuição Binomial

Considere um experimento aleatório consistindo em n tentativas independentes e a probabilidade de ocorrer **sucesso** em cada uma das **n tentativas** é sempre igual a p e de **fracasso é q**, onde **p + q = 1**. A probabilidade de **sucesso e fracasso são as mesmas para cada tentativa**.

Definição: Seja X o número de sucesso em n tentativas, então X pode assumir os valores 0, 1, 2,..., n. Nesta condição a v.a. X tem distribuição Binomial com parâmetro n e p, isto é, $X \sim B(n; p)$.

Considere que se $X \sim B(n; p)$, então a média e a variância de X são definidos por:

a) Média de X: $E(X) = np$.

b) Variância de X: $\sigma^2 = npq$, onde $q = 1 - p$.

Distribuição Binomial

A função de probabilidade da variável aleatória X ~ B(n; p) é dada por

$$P(X = x) = \binom{n}{x} . p^x . (1 - p)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

Onde $\binom{n}{k}$ representa o coeficiente binomial calculado por $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Distribuição Binomial

Utilizando o Python:

Distribuição Binomial para o cálculo de $P(X=x)$
binom.pmf(x,n,p)

Distribuição Binomial para o cálculo de $P(0 \leq X \leq x) = P(X \leq x)$
binom.cdf(x,n,p)

Distribuição Binomial para o cálculo de $P(X>x)$
binom.sf(x,n,p)

Importante:
Para usar as funções de cálculo de probabilidade para a distribuição binomial no Python é necessário primeiramente que você importe a função *binom:*
from scipy.stats **import** binom

Exemplo

Suponha que **5%** de todas as peças que saiam de uma linha de produção sejam defeituosas. Se 10 dessas peças forem escolhidas e inspecionadas, pede se

Observe que temos:

- Um experimento com somente duas opções de resposta (peças defeituosas ou não defeituosas)
- Um número fixo e independente de vezes que o experimento será repetido 10 amostras)
- A probabilidade de peças defeituosas é constante $p = 0{,}05$ e, consequentemente, a probabilidade de peças não defeituosas $q = 1 - p = 0{,}95$.

Dessa forma, podemos dizer que o modelo binomial se adapta bem à situação proposta no exemplo, ou seja, $X \sim B(10; 0{,}05)$.

Exemplo

a) Qual é a probabilidade de obtermos **exatamente 7** peças defeituosas?

$$P(X=7)=\binom{10}{7}.0{,}05^7.(1-0{,}05)^{10-7}=8{,}03789\times10^{-8}$$

```python
from scipy.stats import binom
```

```python
binom.pmf(7, 10, 0.05)
```

```
8.037890624999999e-08
```

Exemplo

b) Qual é a probabilidade de obtermos no máximo 2 peças defeituosas?

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0{,}9884$$

```
binom.pmf(0, 10, 0.05) + binom.pmf(1, 10, 0.05) + binom.pmf(2, 10, 0.05)
```

```
0.9884964426207032
```

Exercício

Cada amostra de ar tem 10% de chance de conter uma certa molécula rara. Considere que as amostras sejam independentes com relação à presença da molécula rara. Encontre a probabilidade de que em 18 amostras:

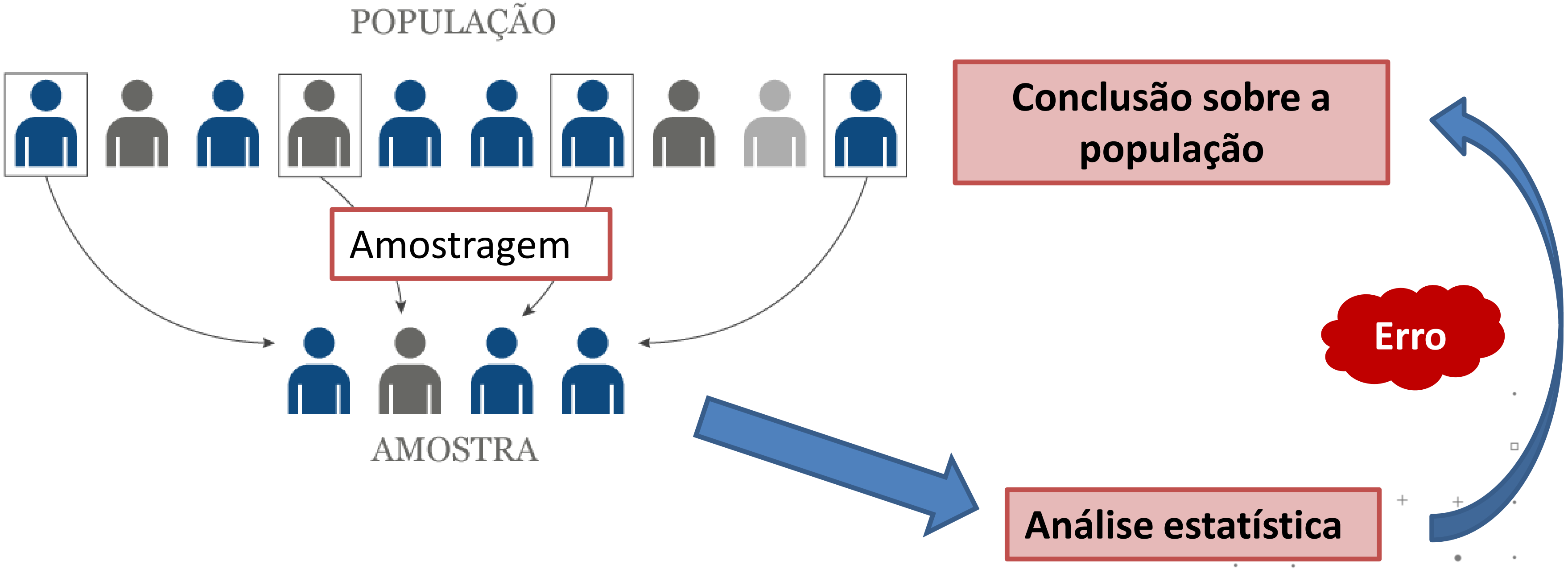
a) Exatamente 2 contenham a molécula rara.

b) No mínimo 4 amostras contenham a molécula rara.

c) De 3 a 7 amostras contenham a molécula rara.

d) O número médio e a variância de moléculas raras.

Amostragem


POPULAÇÃO
Amostragem
AMOSTRA
Conclusão sobre a população
Erro
Análise estatística

Amostragem

- Pesquisa eleitoral
- Pesquisa com clientes
- Controle de qualidade de produtos
- Desenvolvimento de modelos estatísticos
  - Amostra de desenvolvimento (Treino)
  - Amostra de validação (Teste/OOS)

Amostragem

O que é necessário garantir?

- Que a amostra seja representativa da população A amostra deve possuir as mesmas características básicas da população, no que diz respeito às variáveis que desejamos pesquisar.

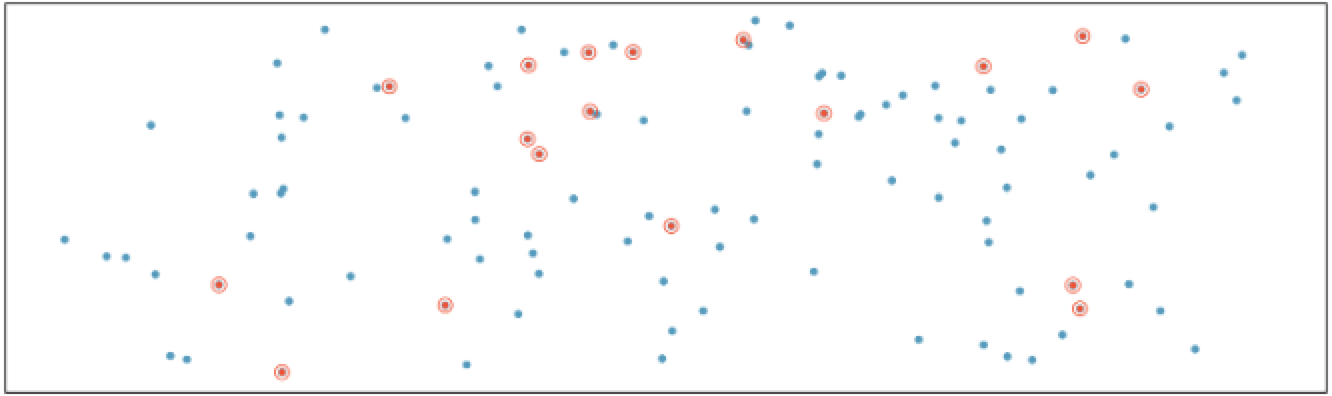
Tipos de amostragem

- PROBABILÍSTICA
  - ALEATÓRIA SIMPLES
  - SISTEMÁTICA
  - ESTRATIFICADA
  - CONGLOMERADO
- NÃO PROBABILÍSTICA (INTENCIONAL)
  - COTAS
  - PROCURA
  - ...

Tipos de amostragem

- PROBABILÍSTICA
  - ALEATÓRIA SIMPLES
  - SISTEMÁTICA
  - ESTRATIFICADA
  - CONGLOMERADO
- NÃO PROBABILÍSTICA (INTENCIONAL)
  - COTAS
  - PROCURA
  - ...

Aleatória simples

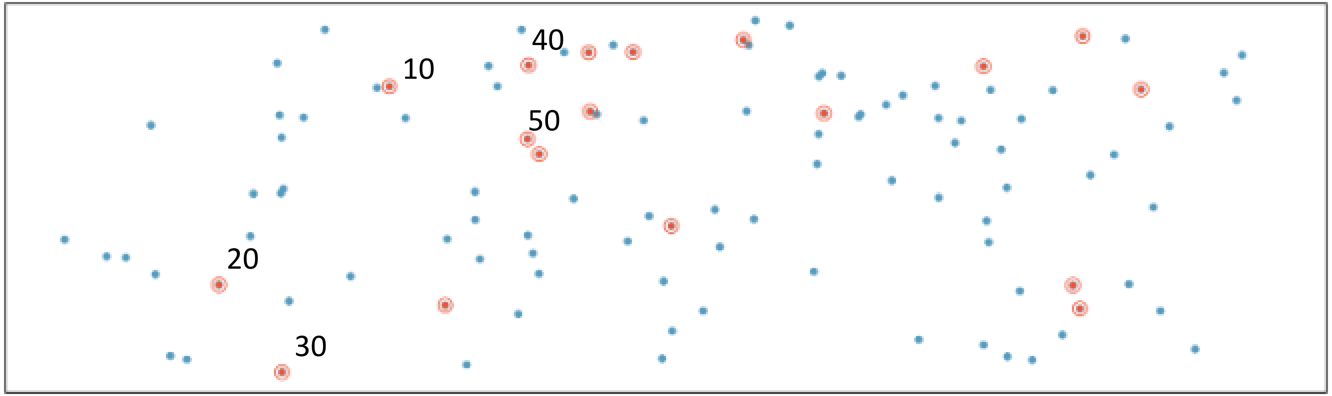


Sorteio de forma aleatória.

Aleatória simples

- Devemos utilizar essa abordagem quando **não** temos que garantir representatividade de nenhum grupo em específico.

Sistemática

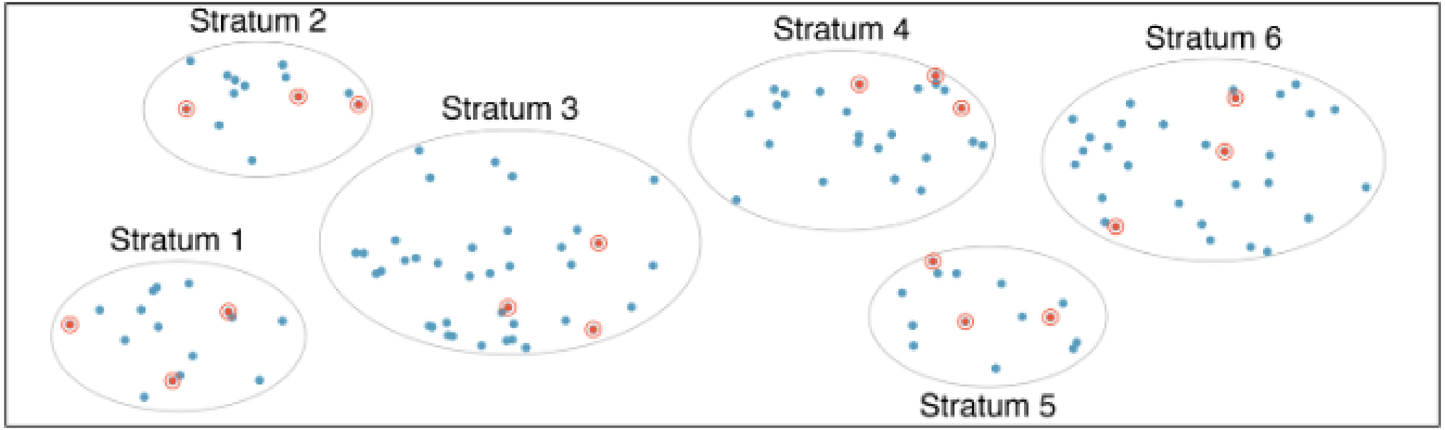


Sorteio baseado em uma estratégia. Ex: Selecionar a cada 10.

Sistemática

- Técnica bastante utilizada em controle de qualidade de processos industriais, onde não há uma especificação de qual elemento será coletado.

Estratificada

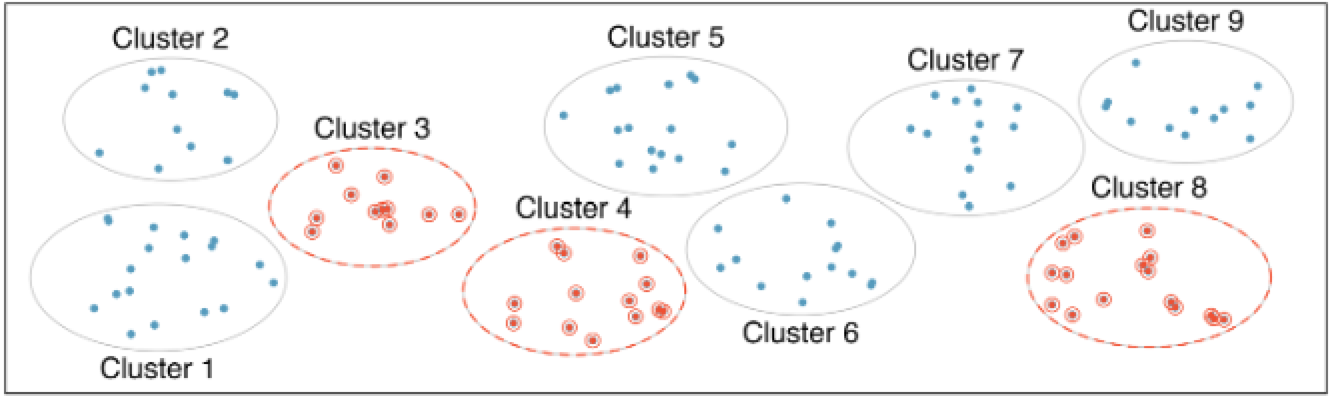


Sorteio de indivíduos dentro dos estratos

Estratificada

- Queremos garantir que cada estrato tenha uma quantidade de representantes pré-definida.
- **Ex:** Em problemas de Fraude onde a ocorrência é muito baixa. Gostaríamos de garantir uma proporção maior de ocorrência.

Conglomerados



Sorteio de clusters e não dos indivíduos.

Exercício (Claims.csv)

O arquivo **claims.csv** contém uma amostra aleatória de 996 apólices de seguros de veículos referente ao período 2004-2005. As variáveis do arquivo estão na seguinte ordem : (i) **valorv** (valor do veículo em 10000 dolares australianos), (ii) **expos** (exposição do veículo), (iii) **nsinistros** (número de sinistros no período), (iv) **csinistros** (custo total dos sinistros em dolares australianos), (v) **tipov** (tipo do veículo em 11 categorias), (vi) **idadev** (idade do veículo em 4 categorias), (vii) **sexoc** (sexo do condutor principal), (viii) **areac** (área de residência do condutor principal) e (ix) **idadec** (idade do condutor principal em 6 categorias).

Exercício (Claims.csv)

Utilizar a base ‘claims.csv’ e faça amostragens:

- Aleatória simples (200)
- Estratificada (100 pelo estrato sexo)

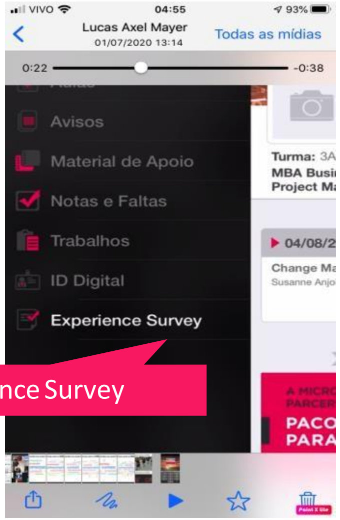
Compare a variável cmsinistros = csinistros/nsinistros por tipo de amostragem usando boxplot.

```
df = pd.read_csv('claims.csv',  delimiter=';', decimal=',')
```

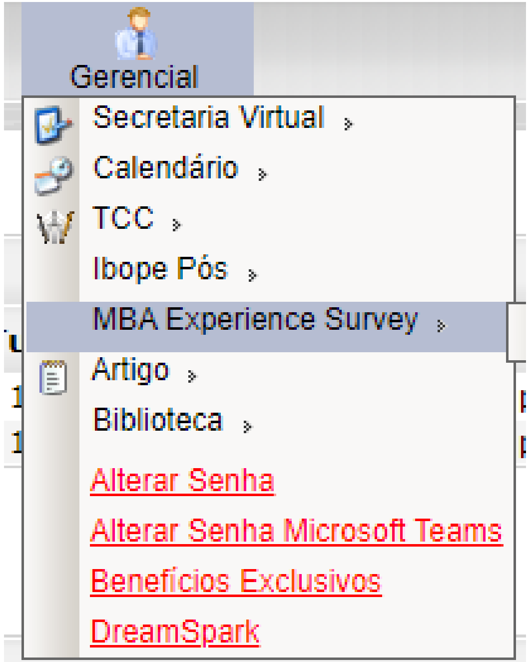
O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)



Experience Survey



OBRIGADO

in /lafphd

profleandro.ferreira@fiap.com.br

FIAP MBA+

FIAP