Instituto Superior de Engenharia do Porto

ESTAT

2019/2029

LEEC

C3 - Amostragem

Teorema do limite central

Sejam $X_1, X_2, \ldots, X_n$ um conjunto de v.a.'s independentes e identicamente distribuídas (iid) com médias e variâncias (limitadas) dadas, respetivamente, por respetivamente, por $\mu_i \; e \; \sigma_i^2 \quad i = 1,2,\ldots.n$

Então, a v.a.

$Y = \sum_{i=1}^{n} a_i X_i$ aproxima-se (em distribuição) da v.a. $N(\sum_{i=1}^{n} a_i \mu_i ; \sum_{i=1}^{n} a^2{}_i \sigma^2{}_i); \quad a_i \in R : i = 1,2,\ldots,n$

à medida que n cresce.
Por conseguinte

$$Z = \frac{Y - \sum_{i=1}^{n} a_i \mu_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \sigma_i^2}} \to N(0,1) \quad \text{quando} \quad n \to +\infty$$

Nota: Na prática, considera-se que a qualidade da aproximação é suficientemente boa quando **n≥30**

Teorema do limite central: corolários

Corolário 1

Sejam $X_1, X_2, \ldots, X_n$ um conjunto de v.a's <u>i.i.d</u>, com médias e variâncias iguais, respectivamente, a μ_i e σ_i^2 :i=1,2,…,n.

Então, a v.a.

$$Y = X_1 + X_2 + \ldots + X_n \sim N\left(\sum_{i=1}^{n} \mu_i ; \sum_{i=1}^{n} \sigma_i^{\,2}\right) \quad \text{quando} \quad n \to +\infty$$

Corolário 2

Sejam $X_1, X_2, \ldots, X_n$ um conjunto de v.a.'s <u>i.i.d</u>, com médias e variâncias iguais, respectivamente, a μ e σ^2

Então, a v.a.

$$Y = \sum_{i=1}^{n} X_i \sim N(n\mu ; n\sigma^2) \quad \text{quando} \quad n \to +\infty$$

e, por conseguinte $$Z = \frac{Y - n\mu}{\sigma\sqrt{n}} \to N(0,1) \quad \text{quando} \quad n \to +\infty$$

Teorema do limite central: exemplo

Admite-se que o erro cometido em cada operação de medição é uma v.a. com média μ=0 mm e desvio padrão σ=5 mm. Para a realização de um trabalho de medição realizaram-se 50 operações.

Calcule a probabilidade do erro de medição acumulado em 50 operações exceder 2 cm.

Definindo:
X_i-"erro cometido na *i-ésima* operação (mm)", i=1,2,…,50
Y-"Erro total de 50 operações (mm)" vem que

$$Y = \sum_{i=1}^{50} X_i \sim N(50 \cdot 0; 50 \cdot 5^2)$$

Pelo corolário do teorema do limite central (n≥30)

$$ou\ seja,\ Y \sim N(0; 35.36^2)$$

$$R: P(Y > 20) = 1 - P(Y \le 20) = 1 - \Phi\left(\frac{20-0}{35.36}\right) = 1 - \Phi(0.57) = 0.2843$$

Amostragem

Em estatística estudam-se fenómenos aleatórios associados a populações.
As populações podem ser finitas ou infinitas.

- No caso das **populações finitas,** embora seja possível obter a informação pertinente, a tarefa da recolha de dados de toda população, em geral, não é viável.

- Para as populações infinitas a tarefa revela-se impossível restando como alternativa a **amostragem**.

- A **amostragem** traduz-se num conjunto de métodos de seleção de elementos de uma população de modo a estimar as propriedades e características inerentes a toda a população.

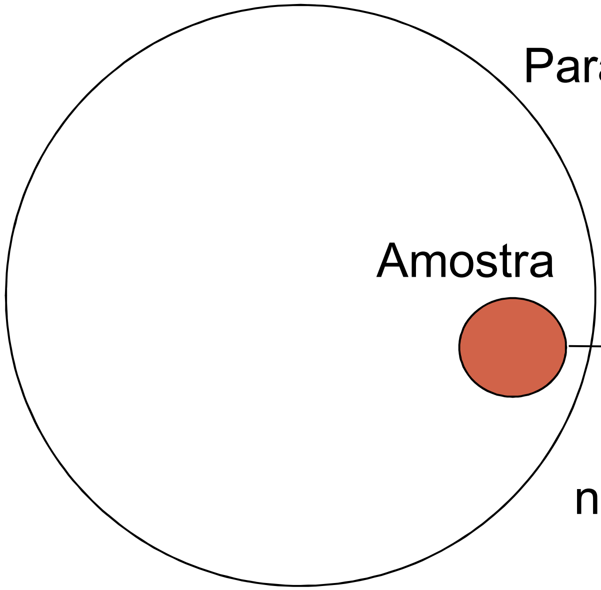
- O interesse da amostragem e enorme e passa, em grande parte, pelas vantagens que oferece: custo da obtenção dos dados; tempo de recolha e processamento dos dados; disponibilidade de recursos computacionais e humanos.

- Existem técnicas de amostragem aleatória (cada elemento da população tem uma probabilidade conhecida de ser selecionado para a amostra) e não aleatória.

Amostragem

Muitas aplicações da estatística a problemas reais consistem na recolha de amostras de populações e subsequente cálculo de certas medidas descritivas (média, proporção, etc…) para a obtenção de informações/conclusões sobre as características das populações sob estudo.

População: caracterizada por uma v.a. X com uma dada distribuição.



Parâmetros: μ

σ^2

p

estatísticas

$\bar{x}$

n – dimensão da amostra

s^2

$\hat{p}$

Amostragem: amostra aleatória e estatística

Def: Amostra aleatória

Uma amostra aleatória de uma população, representada por uma v.a. X, é um conjunto de v. a.'s independentes $\{X_1,X_2,\ldots,X_n\}$, com distribuição igual à da população.

Def: Estatística teste (ou estimador)

É uma v.a. que é função de uma amostra aleatória, ou seja

$$\text{EstatísticaTeste} = G(X_1,X_2,\ldots,X_n)$$

Exemplos:

Máximo da amostra: $\text{Max}(X_1,X_2,\ldots,X_n)$

Média amostral $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ n – dimensão da amostra

Proporção amostral $\hat{P}$

As amostras classificam-se em grandes amostras, se $n \geq 30$

Pequenas amostras se $n<30$

Amostragem: Distribuição da média amostral

Considere-se a extracção de uma amostra aleatória de tamanho n de uma população e que:

i) A <u>população tem distribuição Normal</u> de média μ e variância σ^2

A média dessa amostra

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

tem distribuição normal com média $E(\bar{X})$ e variância $V(\bar{X})$

Pois, uma vez que cada elemento X_i, i=1,2,…,n da amostra tem distribuição $N(\mu;\sigma^2)$, podemos utilizar o teorema da aditividade da Normal.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

$$V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Resumindo:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\mu;\frac{\sigma^2}{n}\right)$$

Amostragem: Distribuição da média amostral

ii) A <u>população não tem distribuição Normal</u>

A média dessa amostra

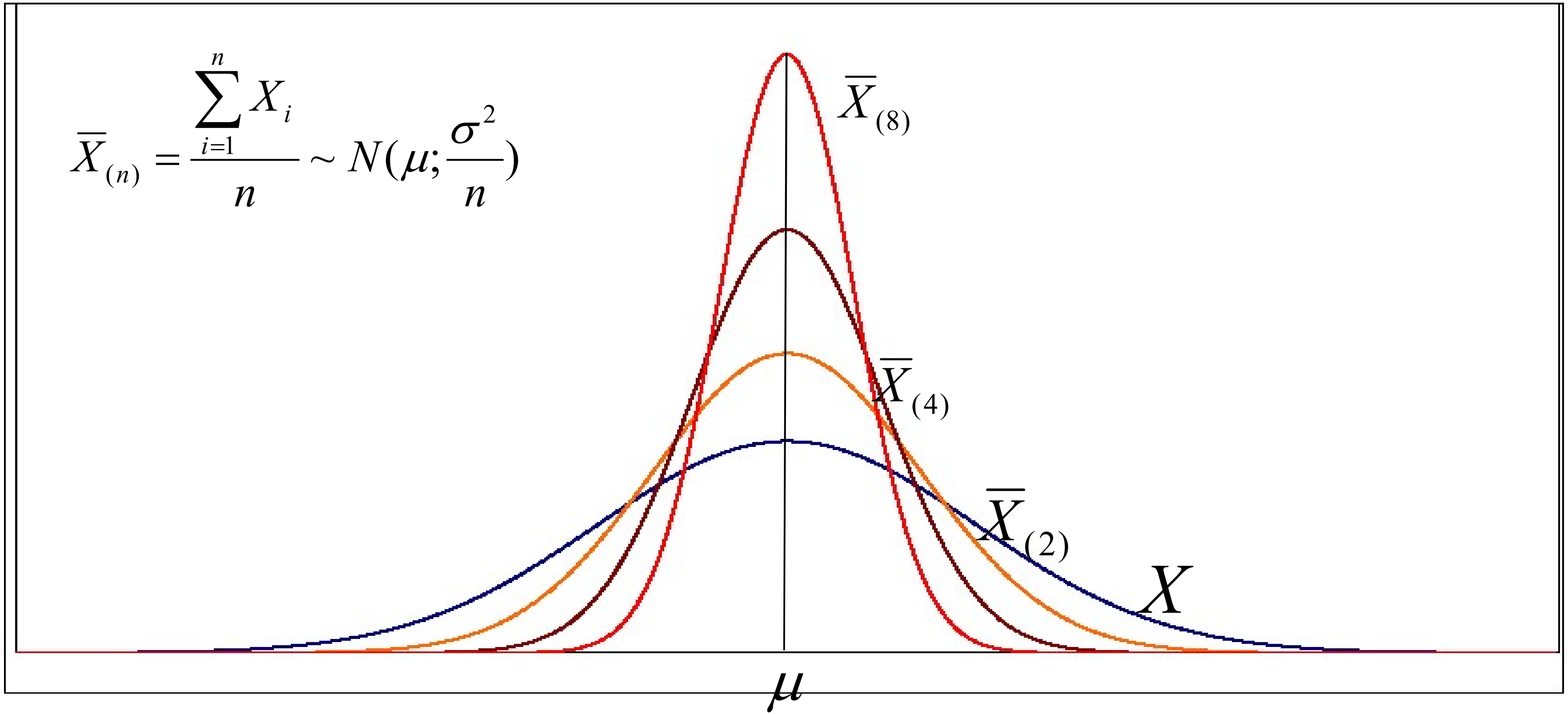
$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ tem distribuição com média $E(\overline{X})$ e variância $V(\overline{X})$

Neste caso é necessário utilizar o teorema do limite central. Porém este só justifica uma aproximação à curva normal se n≥30 (se tivermos uma grande amostra)

Se n≥30 (grande amostra) $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$

Se n<30 (pequena amostra) $\overline{X}$ não tem distribuição normal

Distribuição da média amostral: dimensão da amostra e dispersão



Obs: Quando a dimensão da amostra aumenta a dispersão diminui

Distribuição da média amostral: exemplo

Exemplo 1: Admite-se que a resistência à tração das peças produzidas por um fornecedor é uma v.a. N(120 kg; 25 kg^2).
Um cliente, interessado em realizar um grande negócio, combinou com o fornecedor a realização de um ensaio de tração a 40 peças escolhidas aleatoriamente.

Calcule a probabilidade de se realizar negócio sabendo que o comprador aceita o negócio caso se obtenha uma resistência média superior a 118 kg.

Definindo:

$\overline{X}$ - "Resistência média de 40 peças (kg)" e
X_i - "Resistência da i-ésima peça (kg)"; X_i~N(120;25) i=1,2,…,40

$$\overline{X} = \sum_{i=1}^{40} \frac{X_i}{40} \sim N(120; 25/40)$$

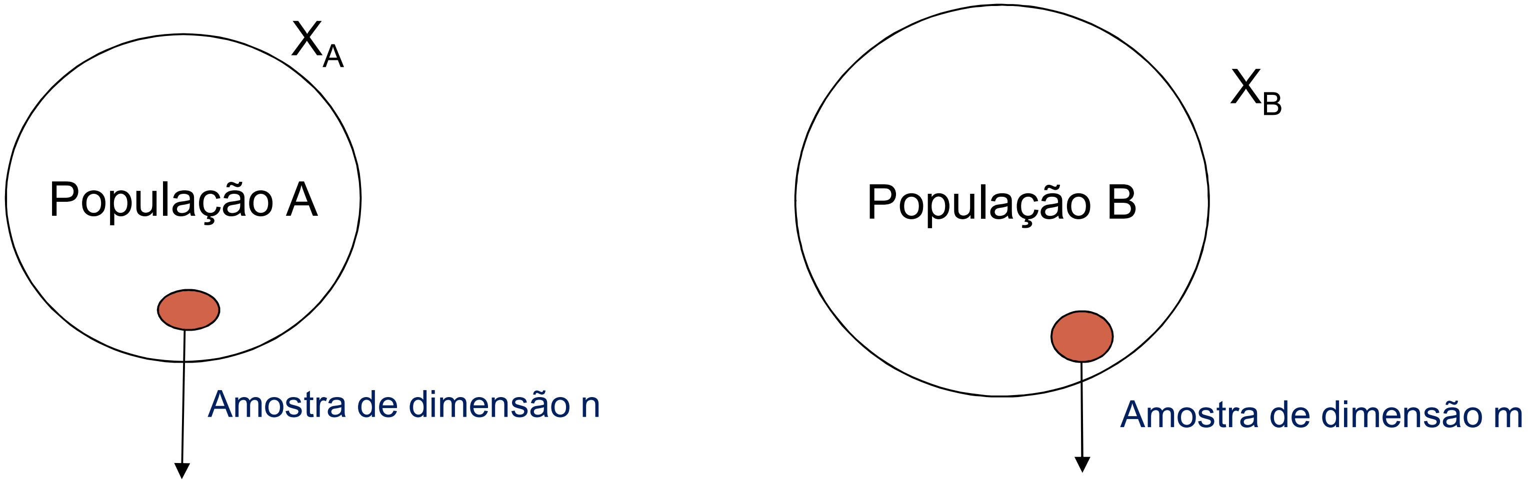
Teorema da Aditividade da distrib. Normal

$$R: P(\overline{X} > 118) = 1 - P(\overline{X} \leq 118) = 1 - \Phi\left(\frac{118-120}{\sqrt{25/40}}\right)$$

$$= 1 - \Phi(-2.53) = 1 - (1 - \Phi(2.53)) = \Phi(2.53) = 0.9943$$

Distribuição da diferença de médias amostrais

Duas populações sem relação estatística: X_A e X_B independentes



Distrib. amostra aleatória A

$$\overline{X}_A = \frac{\sum_{i=1}^{n} X_{Ai}}{n} \sim N(\mu_A; \frac{\sigma_A^2}{n})$$

Distrib. Amostra aleatória B

$$\overline{X}_B = \frac{\sum_{i=1}^{m} X_{Bi}}{m} \sim N(\mu_B; \frac{\sigma_B^2}{m})$$

Então,

$$\overline{X}_A - \overline{X}_B \sim N(\mu_A - \mu_B; \frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m})$$

Distribuição da diferença de médias amostrais: exemplo

Exemplo 2: Suponha que se realizou uma amostra de 40 peças a cada um de dois fornecedores.

fornecedor A
produz peças cuja resistência é N(120 kg;25 kg2)
Fornecedor B
produz peças cuja resistência é uma v.a. com média 122 kg e uma variância de 49 kg2.

Calcule a probabilidade da resistência média das 40 peças do fornecedor A exceder a resistência média de 40 peças do fornecedor B.

$$\overline{X}_A \sim N(120;\frac{25}{40}) \quad \overline{X}_B \sim N(122;\frac{49}{40})$$

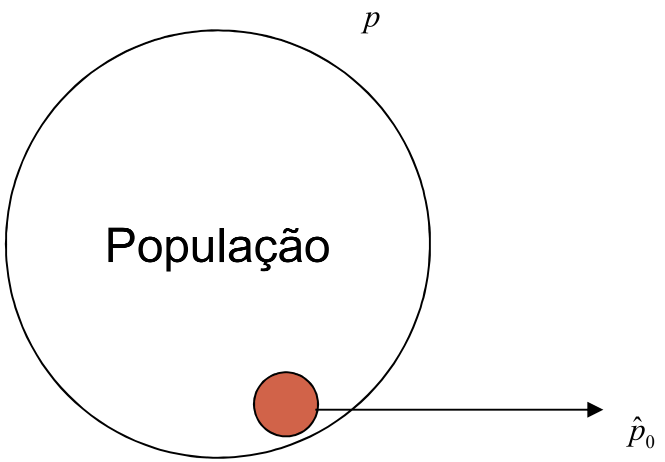
$$R: P(\overline{X}_A > \overline{X}_B) = P(\overline{X}_A - \overline{X}_B > 0) = P(Y > 0) = (*)$$

$$Y = \overline{X}_A - \overline{X}_B \sim N(120-122;\frac{25}{40}+\frac{49}{40}) \Leftrightarrow Y \sim N(-2;1.85)$$

$$(*) = 1 - P(Y < 0) = 1 - \Phi\left(\frac{0-(-2)}{\sqrt{1.85}}\right) = 1 - \Phi(1.47) = 0.0708$$

Distribuição da proporção amostral

De uma população, com uma percentagem p de elementos que têm determinada característica em estudo, são recolhidas amostras aleatórias de dimensão n e calculada a correspondente percentagem observada $\hat{p}_o$



Qual a distribuição da estatística $\hat{P}$?

Uma vez recolhida uma amostra aleatória de tamanho n (n>>30) é de interesse teórico e prático determinar a lei da <u>distribuição da percentagem</u> de elementos da amostra que têm a característica em estudo.

Distribuição da proporção amostral

Assim, definindo X: “Número de indivíduos, em n, que têm a característica em causa”, verifica-se que, usando o teorema do limite central

$$X = \sum_{i=1}^{n} X_i \sim N(np; npq),$$

tendo em conta que

$X_i \sim Be(p), \quad with\ E(X_i) = p\ and\ V(X_i) = pq, i = 1,2,\ldots,n \ \ (q = 1 - p)$

e notando que that $\{X_1, X_2, \ldots, X_n\}$ *is a large random sample with* $n \geq 30)$

A variável que representa a % de indivíduos com a característica em análise, em n elementos amostrados, é dada por

$$\hat{P} = \frac{X}{n} \sim N(p; \frac{pq}{n})$$

pois

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{np}{n} = p$$

$$V(\hat{P}) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{npq}{n^2} = \frac{pq}{n}$$

Distribuição da proporção amostral: exemplo

Exemplo3: O fornecedor A produz componentes com uma taxa de 5% de defeituosos.
O comprador retira uma amostra de 50 peças do total produzido pelo fornecedor A e calcula a % de componentes defeituosos.

a) Calcule a probabilidade de que, na referida amostra, se observe uma percentagem de componentes defeituosos superior a 6%.

b) Qual deveria ser a dimensão mínima da amostra a retirar para que a probabilidade de se obterem mais de 6% de componentes defeituosos seja inferior a 1%?

Distribuição da proporção amostral: exemplo

Resolução:

a) A v.a. que representa a % de peças defeituosas observada na amostra de 50 componentes é

$$\hat{P} \sim N(p; \frac{pq}{n}) \quad ou\ seja,\ \hat{P} \sim N(0.05; \frac{0.05*0.95}{50})$$

A probabilidade pedida é

$$P(\hat{P} > 0.06) = 1 - P(\hat{P} < 0.06) = 1 - P\left(Z < \frac{0.06 - 0.05}{\sqrt{\frac{0.05 \times 0.95}{50}}} \right) = 1 - \Phi(0.33) = 0.3707$$

b)

$\hat{P}_n$ – "Proporção de componentes defeituosos, em n" $\quad \hat{P}_n \sim N(0.05; \frac{0.05 \times 0.95}{n})$

$n = ?$

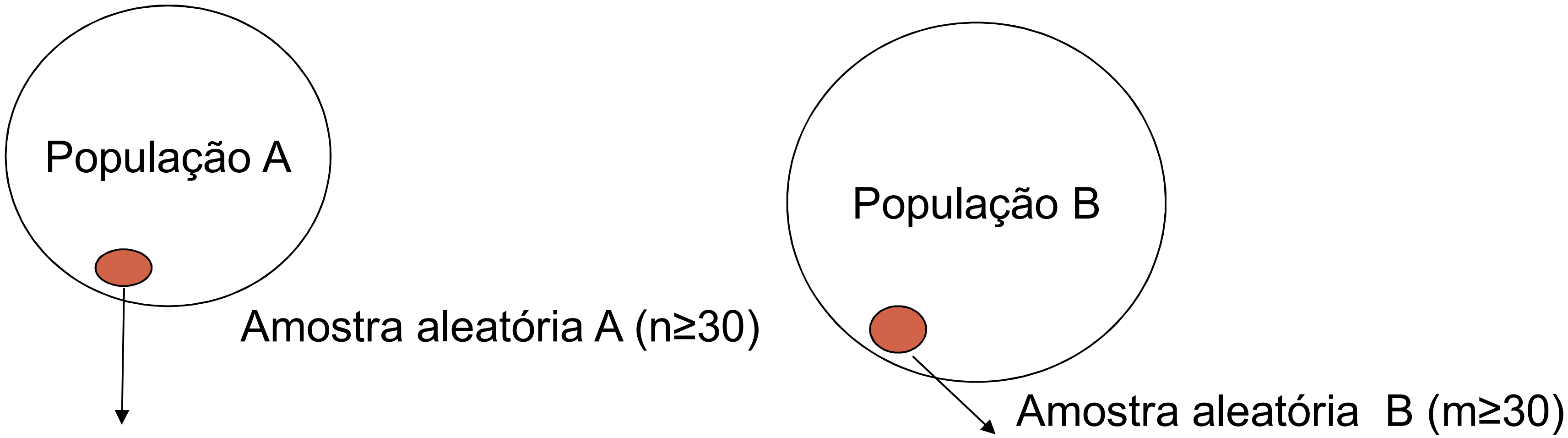
$$P(\hat{P}_n > 0.06) = 0.01 \Leftrightarrow P(\hat{P}_n \leq 0.06) = 0.99 \Leftrightarrow P(Z \leq \frac{0.06 - 0.05}{\sqrt{\frac{0.05 \times 0.95}{n}}}) = 0.99$$

$$\Leftrightarrow P(Z \leq 0.0459\sqrt{n}) = 0.99 \Leftrightarrow \phi(0.0459\sqrt{n}) = 0.99 \Leftrightarrow 0.0459\sqrt{n} = \phi^{-1}(0.99) = 2.33$$

$R: n \geq 2579$

Distribuição da diferença de proporções

Populações não relacionadas estatisticamente:
2 amostras independentes entre si



$$\hat{P}_A \sim N(p_A; \frac{p_A q_A}{n})$$

$$\hat{P}_B \sim N(p_B; \frac{p_B q_B}{m})$$

Então,

$$\hat{P}_A - \hat{P}_B \sim N(p_A - p_B; \frac{p_A q_A}{n} + \frac{p_B q_B}{m})$$

Distribuição da diferença de proporções amostrais. Exemplo:

Exemplo 4: No quadro seguinte indica-se a % de peças defeituosas produzidas por 2 fornecedores

Fornecedor	% de defeituosas
A	6%
B	4%

Será provável que, retirando duas amostras aleatórias, de 60 elementos a cada um dos fornecedores se obtenha uma % de defeituosas superior na amostra do fornecedor B?

A v.a. que representa a diferença de percentagens observadas nas duas amostras é

$$\hat{P} = \hat{P}_B - \hat{P}_A \sim N(0.04 - 0.06; \frac{0.04 * 0.96}{60} + \frac{0.06 * 0.94}{60})$$

$$\hat{P} \sim N(-0.02; 0.04^2)$$

A probabilidade pedida é

$$P(\hat{P} > 0) = 1 - P(\hat{P} < 0) = 1 - P\left(Z < \frac{0 - (-0.02)}{0.04}\right) = 1 - \Phi(0.5) = 0.3085$$

R:Existe uma probabilidade não negligenciável de tal acontecer. Aprox. 31%