MÉTODOS ESTATÍSTICOS

Elementos da Teoria da Amostragem

Licenciatura em Engenharia Informática

Departamento de Matemática
Escola Superior de Tecnologia de Setúbal
Instituto Politécnico de Setúbal
2023-2024

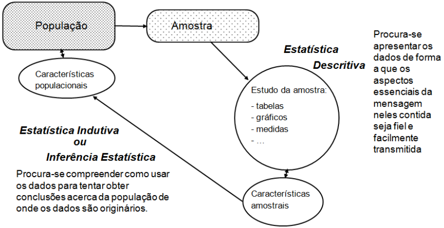
Inferência Estatística

Objetivo

A Inferência Estatística (ou Estatística Indutiva) tem por objetivo o ajustamento de modelos da teoria das probabilidades às observações decorrentes de processos aleatórios. De uma forma geral, pretende estimar uma quantidade ou testar uma hipótese, utilizando técnicas estatísticas convenientes, permitindo tirar conclusões acerca de uma População.

Para atingir este objetivo é necessário percorrer diversas etapas:

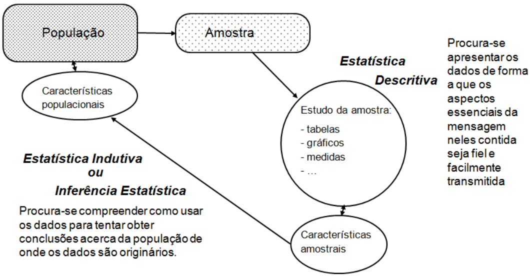
- recolha dos dados - teoria da amostragem,
- análise dos dados - estatística descritiva,
- interpretação e conclusões sobre a População - inferência estatística.



Vamos percorrer estas etapas de forma muito breve, pois o nosso foco é a Inferência Estatística.

Primeira etapa:

- recolha dos dados - **teoria da amostragem**,
- análise dos dados - estatística descritiva,
- interpretação e conclusões sobre a População - inferência estatística.



Teoria da Amostragem

Objetivo

A teoria da amostragem tem por objetivo retirar conclusões sobre uma dada População quando apenas parte dela foi observada, isto é, a partir de uma amostra. Portanto, estuda as relações existentes entre uma População e as amostras extraídas dessa população.

População

Conjunto vasto de elementos que estão sob estudo e, em relação aos quais, se deseja obter alguma informação (relativa a uma característica quantificável).

Amostra

Subconjunto da população que mantém as características da mesma.

Teoria da Amostragem

Como foi referido, a **Teoria da Amostragem** tem por objetivo retirar conclusões sobre uma dada **População**, quando apenas parte dela foi observada, isto é, a partir de uma **amostra**. Para tal é necessário definir um **Plano Amostral**.

O plano de amostragem definido é da máxima importância, dado que a amostra a constituir tem que ser necessariamente: significativa e representativa da população.

Existem vários processos de amostragem e, toda uma teoria sobre o assunto; limitar-nos-emos a referir cada um dos tipos de amostragem e a respetiva importância estatística.

Teoria da Amostragem - Métodos de Amostragem

Probabilística ou Aleatória

- Possuem fundamentação na estatística e nas leis das probabilidades.
- **Vantagens**: Maior confiança na validade dos resultados e na sua generalização.
- **Métodos de Amostragem**:
  - aleatória simples
  - aleatória sistemática
  - aleatória estratificada
  - aleatória por grupos
  - aleatória por etapas

Não Probabilística ou Dirigida

- Não apresentam fundamentação estatística, dependem apenas dos critérios do investigador.
- Menor confiança na validade dos resultados, mas custos e tempos despendidos mais baixos.
- **Métodos de Amostragem**:
  - conveniência
  - julgamento
  - quotas
  - "Bola de Neve"

Observação: Muitas vezes são combinados os 2 métodos de amostragem: **Amostragem Mista**. Quando se conhecem algumas informações da população, define-se uma característica dos elementos a incluir na amostra, deixando-se os restantes fatores ao acaso.

Amostra Aleatória e Estimadores

Amostra Aleatória

Diz-se que $A = (X_1, X_2, \ldots, X_n)$ é uma amostra aleatória de uma população X se $X_1, X_2, \ldots, X_n$ são n variáveis aleatórias independentes e com a mesma função densidade de probabilidade (variáveis aleatórias contínuas) ou função de probabilidade (variáveis aleatórias discretas) da população X. Então a função densidade de probabilidade conjunta ou função de probabilidade conjunta de $A = (X_1, X_2, \ldots, X_n)$ é dada por

$$f(x_1, x_2, \ldots, x_n) = f(x_1) \times f(x_2) \times \cdots \times f(x_n).$$

Estimadores

Seja $A = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma população X. Um estimador de X sobre A é qualquer função real de A, que não contenha parâmetros de valor desconhecido.

Parâmetros e Estimadores

Parâmetro

Característica de uma População, isto é, valor caracterizador da população que, embora possa ser desconhecido, é fixo.

Estimadores

Característica da Amostra, isto é, valor que caracteriza determinada amostra e que é variável de amostra para amostra, ou seja, é uma variável aleatória.

Parâmetros e Estimadores

Estimadores de Interesse

Parâmetros **População** X	**Estimadores** **Amostra** $A = (X_1, X_2, \ldots, X_n)$
Média: $\mu = E[X]$	Média amostral: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variância: $\sigma^2 = V[X] = E\left[(X-\mu)^2\right]$	Variância amostral: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$
Proporção (ou probabilidade de sucesso): $p = \frac{\text{nº de casos favoráveis na População}}{\text{nº de casos possíveis na População}}$	Proporção amostral: $p^* = \frac{\text{nº de casos favoráveis na Amostra}}{\text{nº de casos possíveis na Amostra}}$

Amostra Aleatória

Amostras Aleatórias Independentes

Diz-se que $A = (X_1, X_2, \ldots, X_n)$ e $B = (Y_1, Y_2, \ldots, Y_n)$ são duas amostras aleatórias independentes se não existe nenhum tipo de relação ou fator unificador entre os elementos das amostras. Ou seja, a probabilidade de um sujeito pertencer a ambas as amostras é nula.

Por exemplo, a amostra A refere-se a indivíduos do sexo feminino e a amostra B refere-se a indivíduos do sexo masculino.

Amostras Aleatórias Emparelhadas

Diz-se que $A = (X_1, X_2, \ldots, X_n)$ e $B = (Y_1, Y_2, \ldots, Y_n)$ são duas amostras aleatórias emparelhadas se as amostras são constituídas usando os mesmos sujeitos experimentais ou homólogos.

Por exemplo, a amostra A refere-se a uma medição feita num indivíduo antes de um tratamento e a amostra B refere-se a uma medição feita no mesmo indivíduo depois de um tratamento.

Também se consideram amostras emparelhadas quando se utilizam gémeos ou animais da mesma ninhada (há uma grande probabilidade dos resultados serem semelhantes).

Distribuição Amostrais

- Uma população é representada por uma variável aleatória X cujos parâmetros da correspondente distribuição são fixos, embora porventura desconhecidos.

- Quando é recolhida uma amostra aleatória a partir de uma população de interesse não existe a certeza absoluta de que esta seja representativa dessa população, só se sabe que esta foi recolhida sob critérios de aleatoriedade.

- A partir dessa amostra pode ser calculado o valor do estimador, por exemplo, a média ou a variância da amostra. Porém, se outras amostras forem recolhidas da mesma população não existe a garantia de que os valores dos estimadores calculados com essas amostras sejam todos iguais ao primeiro.

- Cada amostra aleatória retirada de uma população X irá dar origem a valores diferentes do mesmo Estimador, logo os Estimadores são variáveis aleatórias e portanto têm uma certa distribuição de probabilidade.

- À distribuição de probabilidade de um Estimador chama-se **Distribuição amostral** ou **Distribuição de amostragem**.

Distribuição Amostrais

- Nos próximos slides serão indicadas as distribuições de amostragem dos estimadores definidos como de interesse: $\overline{X}$, S^2 e p^*.

- A demonstração desses resultados não será apresentada. No entanto basta rever as propriedades referidas para as distribuições teóricas.

- Os resultados apresentados nos slides seguintes encontram-se no **Formulário** disponível no moodle.

Distribuição Amostral para a Média Amostral

Se uma amostra aleatória de dimensão n é proveniente de uma população de média μ e variância σ^2, então a **Média Amostral** é o Estimador representado pela variável aleatória

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

e a sua **distribuição amostral** é:

Condições		
População	**Parâmetros e Amostra**	**Distribuição Amostral**
Normal	σ conhecido	$Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
	σ desconhecido	$T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$
Qualquer	σ conhecido e $n \geq 30$	$Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \dot{\sim} N(0,1)$
	σ desconhecido e $n \geq 30$	$Z = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \dot{\sim} N(0,1)$

Obs.: Quando a População é Normal e $n \geq 30$, em vez $T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$ pode usar $Z = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \dot{\sim} N(0,1)$

Distribuição Amostral para a Diferença de Duas Médias Amostrais

Duas **amostras aleatórias independentes**: uma amostra aleatória de dimensão n_1 proveniente de uma população 1 com média μ_1 e variância σ_1^2 e a outra amostra aleatória de dimensão n_2 proveniente de uma população 2 com média μ_2 e variância σ_2^2. A **Diferença de Duas Médias Amostrais** é o Estimador representado pela variável aleatória

$$\overline{X}_1 - \overline{X}_2$$

e a sua **distribuição amostral** é:

Condições		
Populações	**Parâmetros e Amostras**	**Distribuição Amostral**
Normais	Amostras independentes σ_1 e σ_2 conhecidos	$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
	Amostras independentes σ_1 e σ_2 desconhecidos $\sigma_1 = \sigma_2$	$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sim t_{(gl_1)}$
	Amostras independentes σ_1 e σ_2 desconhecidos $\sigma_1 \neq \sigma_2$	$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{(gl_2)}$
Quaisquer	Amostras independentes σ_1 e σ_2 conhecidos $n_1 \geq 30$ e $n_2 \geq 30$	$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \dot{\sim} N(0,1)$
	Amostras independentes σ_1 e σ_2 desconhecidos $n_1 \geq 30$ e $n_2 \geq 30$	$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} N(0,1)$

Obs.:

Quando as Populações são Normais, $n_1 \geq 30$ e $n_2 \geq 30$, em vez $T \sim t_{(gl)}$ pode usar $Z \dot{\sim} N(0,1)$

com

- $gl_1 = n_1 + n_2 - 2$, ou seja

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}} \sim t_{(n_1+n_2-2)}$$

- $gl_2 \approx \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_2^4}{n_2^2(n_2-1)}}$

  habitualmente considera-se o inteiro mais próximo ou faz-se a correção de Welch-Satterthwaite

Distribuição Amostral para a Variância Amostral

Se uma amostra aleatória de dimensão n é proveniente de uma **população com distribuição Normal** de média μ e variância σ^2, então a **Variância Amostral** é o Estimador representado pela variável aleatória

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$

e a sua **distribuição amostral** é:

Condição	
População	**Distribuição Amostral**
Normal	$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$

Observação:

- Se $\mu = E[X]$ for um valor conhecido, a Variância amostral pode ser $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}$ e neste caso tem-se

$$X^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2_{(n)}$$

Distribuição Amostral para a Razão de Duas Variâncias Amostrais

Duas **amostras aleatórias independentes**: uma amostra aleatória de dimensão n_1 proveniente de uma população 1 com distribuição Normal de média μ_1 e variância σ_1^2 e a outra amostra aleatória de dimensão n_2 proveniente de uma população 2 com distribuição Normal de média μ_2 e variância σ_2^2. A **Razão de Duas Variâncias Amostrais** é o Estimador representado pela variável aleatória

$$\frac{S_1^2}{S_2^2} = \frac{\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_{1_i} - \overline{X}_1\right)^2}{\frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(X_{2_j} - \overline{X}_2\right)^2}$$

e a sua **distribuição amostral** é:

Condições		
Populações	**Amostras**	**Distribuição Amostral**
Normais	Amostras independentes	$F = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(n_1-1, n_2-1)}$

Distribuição Amostral para a Proporção Amostral

Se uma amostra aleatória de dimensão $n \geq 30$ é proveniente de uma população com distribuição Binomial com probabilidade de sucesso p, então a **Proporção Amostral** é o Estimador representado pela variável aleatória

$$p^* = \frac{\text{número de casos favoráveis na Amostra}}{\text{número de casos possíveis na Amostra}}$$

e a sua **distribuição amostral** é:

Condições	
Amostra	**Distribuição Amostral**
$n \geq 30$	$Z = \frac{p^* - p}{\sqrt{\frac{pq}{n}}} \dot{\sim} N(0,1)$

Observação: Numa população Binomial a probabilidade de sucesso p entende-se como um caso particular de um valor médio, então a distribuição amostral do estimador p^* é um caso particular da distribuição amostral de $\overline{X}$ quando $n \geq 30$.

Distribuição Amostral para a Diferença de Duas Proporções Amostrais

Sejam duas populações Binomiais cuja a probabilidade de sucesso da população 1 é p_1 e da população 2 é p_2 e sejam duas amostras aleatórias independentes.. Na população 1 foi recolhida uma amostra aleatória de dimensão n_1 e na população 2 de dimensão n_2. Então a **Diferença de Duas Proporções Amostrais** é o Estimador representado pela variável aleatória

$$p_1^* - p_2^*$$

e a sua **distribuição amostral** é:

Condições	
Amostras	**Distribuição Amostral**
Amostras independentes $n_1 \geq 30$ e $n_2 \geq 30$	$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \dot{\sim} N(0,1)$

Observação:

- Numa população Binomial a probabilidade de sucesso p entende-se como um caso particular de um valor médio, então a distribuição amostral do estimador $p_1^* - p_2^*$ é um caso particular da distribuição amostral de $\overline{X}_1 - \overline{X}_2$ quando $n_1 \geq 30$ e $n_2 \geq 30$.

Exemplo 1

O fabricante A afirma que os cabos produzidos na sua fábrica têm uma espessura média de $0.20\ mm$ e um desvio padrão de $0.04\ mm$ e o fabricante B afirma que os cabos produzidos na sua fábrica têm uma espessura média de $0.18\ mm$ e um desvio padrão de $0.05\ mm$. Um engenheiro pretende adquirir cabos de uma das fábricas e como a espessura dos cabos é fator preferencial, o engenheiro recolhe uma amostra aleatória de 61 cabos de cada uma das fábricas.

1. Qual a probabilidade da média da amostra recolhida na fábrica A ser superior à média referida para essa fábrica?

Exemplo 1

O fabricante A afirma que os cabos produzidos na sua fábrica têm uma espessura média de $0.20\ mm$ e um desvio padrão de $0.04\ mm$ e o fabricante B afirma que os cabos produzidos na sua fábrica têm uma espessura média de $0.18\ mm$ e um desvio padrão de $0.05\ mm$. Um engenheiro pretende adquirir cabos de uma das fábricas e como a espessura dos cabos é fator preferencial, o engenheiro recolhe uma amostra aleatória de 61 cabos de cada uma das fábricas.

1. Qual a probabilidade da média da amostra recolhida na fábrica A ser superior à média referida para essa fábrica?

- **População**

X_A = espessura, em mm, dos cabos produzidos pelo fabricante A

média populacional $= \mu_A = E[X_A] = 0.20$

desvio padrão populacional $= \sigma_A = \sqrt{V[X_A]} = 0.04$

- **Amostra Aleatória**

dimensão $= n_A = 61$

média amostral $= \overline{X}_A$

Pretende-se

$$P\left(\overline{X}_A > \mu_A\right) = P\left(\overline{X}_A > 0.20\right)$$

Como a População não tem distribuição conhecida, então é necessário que a amostra tenha dimensão $n \geq 30$ para se poder recorrer ao Teorema do Limite Central. Como $n_A = 61 \geq 30$, então pode-se considerar que a População é aproximadamente Normal (Teorema do Limite Central). Agora basta escolher a distribuição para a média amostral:

Condições		Distribuição Amostral
População	**Parâmetros e Amostra**	
Qualquer	σ conhecido e $n \geq 30$	$Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \dot{\sim} N(0,1)$
	σ desconhecido e $n \geq 30$	$Z = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \dot{\sim} N(0,1)$

então

$$Z = \frac{\overline{X}_A - \mu_A}{\frac{\sigma_A}{\sqrt{n_A}}} \dot{\sim} N(0,1)$$

logo

$$P\left(\overline{X}_A > \mu_A\right) = P\left(\overline{X}_A > 0.20\right) = P\left(\frac{\overline{X}_A - 0.20}{\frac{0.04}{\sqrt{61}}} > \frac{0.20 - 0.20}{\frac{0.04}{\sqrt{61}}}\right) =$$

$$= P(Z > 0) = 1 - P(Z \leq 0) = 1 - \Phi(0) \underset{Z \sim N(0,1)}{=} 1 - 0.5 = 0.5$$

Exemplo 1

O fabricante A afirma que os cabos produzidos na sua fábrica têm uma espessura média de $0.20\ mm$ e um desvio padrão de $0.04\ mm$ e o fabricante B afirma que os cabos produzidos na sua fábrica têm uma espessura média de $0.18\ mm$ e um desvio padrão de $0.05\ mm$. Um engenheiro pretende adquirir cabos de uma das fábricas e como a espessura dos cabos é fator preferencial, o engenheiro recolhe uma amostra aleatória de 61 cabos de cada uma das fábricas.

2. O engenheiro escolhe a fábrica a que corresponde a amostra com menor espessura média. Qual a probabilidade do engenheiro escolher a fábrica A?

Exemplo 1

O fabricante A afirma que os cabos produzidos na sua fábrica têm uma espessura média de $0.20\ mm$ e um desvio padrão de $0.04\ mm$ e o fabricante B afirma que os cabos produzidos na sua fábrica têm uma espessura média de $0.18\ mm$ e um desvio padrão de $0.05\ mm$. Um engenheiro pretende adquirir cabos de uma das fábricas e como a espessura dos cabos é fator preferencial, o engenheiro recolhe uma amostra aleatória de 61 cabos de cada uma das fábricas.

2. O engenheiro escolhe a fábrica a que corresponde a amostra com menor espessura média. Qual a probabilidade do engenheiro escolher a fábrica A?

- **População** A

X_A = espessura, em mm, dos cabos produzidos pelo fabricante A

média populacional $= \mu_A = E[X_A] = 0.20$

desvio padrão populacional $= \sigma_A = \sqrt{V[X_A]} = 0.04$

- **População** B

X_B = espessura, em mm, dos cabos produzidos pelo fabricante B

média populacional $= \mu_B = E[X_B] = 0.18$

desvio padrão populacional $= \sigma_B = \sqrt{V[X_B]} = 0.05$

- **Amostra Aleatória de** A

dimensão $= n_A = 61$

média amostral $= \overline{X}_A$

- **Amostra Aleatória de** B

dimensão $= n_B = 61$

média amostral $= \overline{X}_B$

- **Amostras Independentes**

Pretende-se

$$P\left(\overline{X}_A < \overline{X}_B\right) = P\left(\overline{X}_A - \overline{X}_B < 0\right)$$

Como as Populações não têm distribuição conhecida, então é necessário que as amostras tenham dimensão $n \geq 30$ para se poder recorrer ao Teorema do Limite Central. Como $n_A = 61 \geq 30$ e $n_B = 61 \geq 30$, então pode-se considerar que as Populações são aproximadamente Normais (Teorema do Limite Central). Agora basta escolher a distribuição para a diferença de duas médias amostrais:

Condições		
Populações	**Parâmetros e Amostras**	**Distribuição Amostral**
Quaisquer	Amostras independentes σ_1 e σ_2 conhecidos $n_1 \geq 30$ e $n_2 \geq 30$	$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \dot{\sim} N(0,1)$
	Amostras independentes σ_1 e σ_2 desconhecidos $n_1 \geq 30$ e $n_2 \geq 30$	$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} N(0,1)$

então

$$Z = \frac{\left(\overline{X}_A - \overline{X}_B\right) - \left(\mu_A - \mu_B\right)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \dot{\sim} N(0,1)$$

logo

$$P\left(\overline{X}_A < \overline{X}_B\right) = P\left(\overline{X}_A - \overline{X}_B < 0\right) =$$

$$= P\left(\frac{\overline{X}_A - \overline{X}_B - (0.20 - 0.18)}{\sqrt{\frac{0.04^2}{61} + \frac{0.05^2}{61}}} < \frac{0 - (0.20 - 0.18)}{\sqrt{\frac{0.04^2}{61} + \frac{0.05^2}{61}}}\right) =$$

$$= P(Z < -2.44) \underset{\text{v.a. contínua}}{=} \Phi(-2.44) \underset{Z \sim N(0,1)}{=}$$

$$= 0.0073$$

Exemplo 2

O fabricante C afirma que os cabos produzidos na sua fábrica têm uma espessura com um comportamento normal e com desvio padrão de $0.04\ mm$ e o fabricante D afirma que os cabos produzidos na sua fábrica têm uma espessura com um comportamento normal e com desvio padrão de $0.05\ mm$. Um engenheiro pretende adquirir cabos de uma das fábricas e como a espessura dos cabos é fator preferencial, o engenheiro recolhe uma amostra aleatória de 61 cabos de cada uma das fábricas.

1. Qual a probabilidade da variância da amostra recolhida na fábrica D ser inferior à variância referida para essa fábrica?

Exemplo 2

O fabricante C afirma que os cabos produzidos na sua fábrica têm uma espessura com um comportamento normal e com desvio padrão de $0.04\ mm$ e o fabricante D afirma que os cabos produzidos na sua fábrica têm uma espessura com um comportamento normal e com desvio padrão de $0.05\ mm$. Um engenheiro pretende adquirir cabos de uma das fábricas e como a espessura dos cabos é fator preferencial, o engenheiro recolhe uma amostra aleatória de 61 cabos de cada uma das fábricas.

1. Qual a probabilidade da variância da amostra recolhida na fábrica D ser inferior à variância referida para essa fábrica?

- **População**

$X_D =$ espessura, em mm, dos cabos produzidos pelo fabricante D, $\quad X_D \sim N(\mu_D, 0.05)$

média populacional $= \mu_D$ desconhecida

desvio padrão populacional $= \sigma_D = \sqrt{V[X_D]} = 0.05$

- **Amostra Aleatória**

dimensão $= n_D = 61$

variância amostral $= S_D^2$

Pretende-se

$$P\left(S_D^2 < \sigma_D^2\right) = P\left(S_D^2 < 0.05^2\right)$$

A População tem de ter obrigatoriamente distribuição Normal (ver propriedades da distribuição Qui-Quadrado), como é o caso. Agora basta ver qual é a distribuição para a variância amostral:

Condição População	Distribuição Amostral
Normal	$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$

então

$$X^2 = \frac{(n_D - 1)\, S_D^2}{\sigma_D^2} \sim \chi^2_{(n_D-1)} \Leftrightarrow X^2 \sim \chi^2_{(61-1)} \Leftrightarrow X^2 \sim \chi^2_{(60)}$$

logo

$$P\left(S_D^2 < \sigma_D^2\right) = P\left(S_D^2 < 0.05^2\right) = P\left(\frac{(61-1)\, S_D^2}{0.05^2} < \frac{(61-1)\, 0.05^2}{0.05^2}\right) =$$

$$= P(X^2 < 60) \underset{\text{v.a. contínua}}{=} F(60) \underset{X^2\sim\chi^2_{(60)}}{=} 0.5243$$

Exemplo 2

O fabricante C afirma que os cabos produzidos na sua fábrica têm uma espessura com um comportamento normal e com desvio padrão de $0.04\ mm$ e o fabricante D afirma que os cabos produzidos na sua fábrica têm uma espessura com um comportamento normal e com desvio padrão de $0.05\ mm$. Um engenheiro pretende adquirir cabos de uma das fábricas e como a espessura dos cabos é fator preferencial, o engenheiro recolhe uma amostra aleatória de 61 cabos de cada uma das fábricas.

2. Qual a probabilidade da variância da amostra recolhida na fábrica C ultrapassar a variância da amostra recolhida na fábrica D?

Exemplo 2

O fabricante C afirma que os cabos produzidos na sua fábrica têm uma espessura com um comportamento normal e com desvio padrão de $0.04\ mm$ e o fabricante D afirma que os cabos produzidos na sua fábrica têm uma espessura com um comportamento normal e com desvio padrão de $0.05\ mm$. Um engenheiro pretende adquirir cabos de uma das fábricas e como a espessura dos cabos é fator preferencial, o engenheiro recolhe uma amostra aleatória de 61 cabos de cada uma das fábricas.

2. Qual a probabilidade da variância da amostra recolhida na fábrica C ultrapassar a variância da amostra recolhida na fábrica D?

- **População** C

X_C = espessura, em mm, dos cabos produzidos pelo fabricante C,

$$X_C \sim N(\mu_C, 0.04)$$

média populacional $= \mu_C$ desconhecida
desvio padrão populacional $= \sigma_C = \sqrt{V[X_C]} = 0.04$

- **População** D

X_D = espessura, em mm, dos cabos produzidos pelo fabricante D,

$$X_D \sim N(\mu_D, 0.05)$$

média populacional $= \mu_D$ desconhecida

desvio padrão populacional $= \sigma_D = \sqrt{V[X_D]} = 0.05$

- **Amostra Aleatória de** C

dimensão $= n_C = 61$

variância amostral $= S_C^2$

- **Amostra Aleatória de** D

dimensão $= n_D = 61$

variância amostral $= S_D^2$

- **Amostras Independentes**

Pretende-se

$$P\left(S_C^2 > S_D^2\right) = P\left(\frac{S_C^2}{S_D^2} > 1\right)$$

As Populações têm de ter obrigatoriamente distribuição Normal (ver propriedades da distribuição F de Snedecor e da distribuição Qui-Quadrado), como é o caso. Agora basta ver qual é a distribuição para o quociente entre duas variâncias amostrais:

Condições		
Populações	**Amostras**	**Distribuição Amostral**
Normais	Amostras independentes	$\mathsf{F} = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(n_1-1, n_2-1)}$

então

$$\mathsf{F} = \frac{S_C^2}{S_D^2} \times \frac{\sigma_D^2}{\sigma_C^2} \sim F_{(n_C-1, n_D-1)} \Leftrightarrow \mathsf{F} \sim F_{(61-1, 61-1)} \Leftrightarrow \mathsf{F} \sim F_{(60,60)}$$

logo

$$P\left(S_C^2 > S_D^2\right) = P\left(\frac{S_C^2}{S_D^2} > 1\right) = P\left(\frac{S_C^2}{S_D^2} \times \frac{0.05^2}{0.04^2} > 1 \times \frac{0.05^2}{0.04^2}\right) =$$

$$= P(\mathsf{F} > 1.56) = 1 - P(\mathsf{F} \leq 1.56) = 1 - F(1.56) \underset{\mathsf{F} \sim F_{(60,60)}}{=}$$

$$= 1 - 0.9562 = 0.0438$$

Exemplo 3

Suponha que está em presença de duas populações Binomiais onde $p_1 = 0.6$ e $p_2 = 0.5$. Retirou-se da primeira população uma amostra de 50 observações e da segunda uma amostra com 40 observações.

1. Qual a probabilidade da proporção amostral da população 1 ser superior a 0.7?

Exemplo 3

Suponha que está em presença de duas populações Binomiais onde $p_1 = 0.6$ e $p_2 = 0.5$. Retirou-se da primeira população uma amostra de 50 observações e da segunda uma amostra com 40 observações.

1. Qual a probabilidade da proporção amostral da população 1 ser superior a 0.7?

- **População**

$$X_1 \sim \text{ Binomial}$$

proporção populacional $= p_1 = 0.6$

- **Amostra Aleatória**

dimensão $= n_1 = 50$

proporção amostral $= p_1^*$

Pretende-se

$$P\left(p_1^* > 0.7\right)$$

Como a População é Binomial, então é obrigatório recolher uma amostra de dimensão $n \geq 30$ para se poder recorrer ao Teorema do Limite Central. Como $n_1 = 50 \geq 30$, então agora basta ver qual é a distribuição para a proporção amostral:

Condições	
Amostra	**Distribuição Amostral**
$n \geq 30$	$Z = \frac{p^* - p}{\sqrt{\frac{pq}{n}}} \dot{\sim} N(0,1)$

então

$$Z = \frac{p_1^* - p_1}{\sqrt{\frac{p_1 q_1}{n_1}}} \dot{\sim} N(0,1) \Leftrightarrow Z = \frac{p_1^* - p_1}{\sqrt{\frac{p_1 \times (1-p_1)}{n_1}}} \dot{\sim} N(0,1)$$

logo

$$P(p_1^* > 0.7) = P\left(\frac{p_1^* - 0.6}{\sqrt{\frac{0.6 \times (1-0.6)}{50}}} > \frac{0.7 - 0.6}{\sqrt{\frac{0.6 \times (1-0.6)}{50}}}\right) = P(Z > 1.44) =$$

$$= 1 - P(Z \leq 1.44) = 1 - \Phi(1.44) \underset{Z \sim N(0,1)}{=} 1 - 0.9251 = 0.0749$$

Exemplo 3

Suponha que está em presença de duas populações Binomiais onde $p_1 = 0.6$ e $p_2 = 0.5$. Retirou-se da primeira população uma amostra de 50 observações e da segunda uma amostra com 40 observações.

2. Qual a probabilidade da proporção amostral da população 1 ser inferior à proporção amostral da população 2?

Exemplo 3

Suponha que está em presença de duas populações Binomiais onde $p_1 = 0.6$ e $p_2 = 0.5$. Retirou-se da primeira população uma amostra de 50 observações e da segunda uma amostra com 40 observações.

2. Qual a probabilidade da proporção amostral da população 1 ser inferior à proporção amostral da população 2?

- **População 1**

$$X_1 \sim \text{ Binomial}$$

proporção populacional $= p_1 = 0.6$

- **População 2**

$$X_2 \sim \text{ Binomial}$$

proporção populacional $= p_2 = 0.5$

- **Amostra Aleatória de 1**

dimensão $= n_1 = 50$

proporção amostral $= p_1^*$

- **Amostra Aleatória de 2**

dimensão $= n_2 = 40$

proporção amostral $= p_2^*$

- **Amostras Independentes**

Pretende-se

$$P\left(p_1^* < p_2^*\right) = P\left(p_1^* - p_2^* < 0\right)$$

Como as Populações são Binomiais, então é obrigatório recolher amostras de dimensão $n \geq 30$ para se poder recorrer ao Teorema do Limite Central. Como $n_1 = 50 \geq 30$ e $n_2 = 40 \geq 30$, então agora basta ver qual é a distribuição para a diferença de duas proporções amostrais:

Condições	
Amostras	**Distribuição Amostral**
Amostras independentes $n_1 \geq 30$ e $n_2 \geq 30$	$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \dot{\sim} N(0,1)$

então

$$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \dot{\sim} N(0,1) \Leftrightarrow Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{p_1 \times (1-p_1)}{n_1} + \frac{p_2 \times (1-p_2)}{n_2}}} \dot{\sim} N(0,1)$$

logo

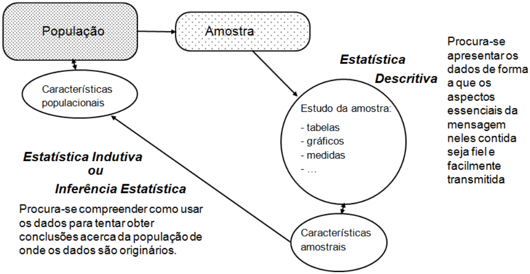
$$P(p_1^* < p_2^*) = P(p_1^* - p_2^* < 0) =$$

$$= P\left(\frac{(p_1^* - p_2^*) - (0.6 - 0.5)}{\sqrt{\frac{0.6 \times (1-0.6)}{50} + \frac{0.5 \times (1-0.5)}{40}}} < \frac{0 - (0.6 - 0.5)}{\sqrt{\frac{0.6 \times (1-0.6)}{50} + \frac{0.5 \times (1-0.5)}{40}}}\right) =$$

$$= P(Z < -0.95) \underset{\text{v.a. contínua}}{=} \Phi(-0.95) = 0.1711$$

Segunda etapa:

- recolha dos dados - teoria da amostragem,
- análise dos dados - **estatística descritiva**,
- interpretação e conclusões sobre a População - inferência estatística.

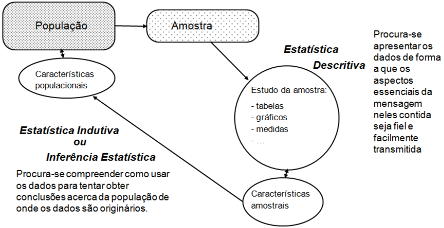


Já foi estudada no

capítulo 1 - Estatística Descritiva

Terceira etapa:

- recolha dos dados - teoria da amostragem,
- análise dos dados - estatística descritiva,
- interpretação e conclusões sobre a População - **inferência estatística**.



Próximos capítulos:

capítulo 4 - Elementos da Teoria da Estimação
capítulo 5 - Testes de Hipóteses Paramétricos
capítulo 6 - Testes de Hipóteses Não Paramétricos
capítulo 7 - Regressão Linear Simples