CPEQS

# AULA 05
# Inferência a partir de duas amostras


## Ernesto F. L. Amaral


**04 de outubro de 2013**


**Centro de Pesquisas Quantitativas em Ciências Sociais (CPEQS)**
**Faculdade de Filosofia e Ciências Humanas (FAFICH)**
**Universidade Federal de Minas Gerais (UFMG)**


**Fonte:**
**Triola, Mario F. 2008. "Introdução à estatística". 10 ª ed. Rio de Janeiro: LTC. Capítulo 9 (pp.360-407).**

# ESQUEMA DA AULA

– Inferências sobre duas proporções.

– Inferências sobre duas médias: amostras independentes.

# VISÃO GERAL

– Os capítulos anteriores (estimação de valores de parâmetros populacionais e teste de hipóteses) envolveram métodos para uma única amostra, usada para se fazer inferência sobre um único parâmetro populacional.

– Na prática, há muitas situações em que desejamos comparar dois conjuntos de dados amostrais.

– Portanto, este capítulo estende os métodos abordados anteriormente para situações que envolvem comparações de duas amostras em vez de apenas uma.

# INFERÊNCIAS SOBRE DUAS PROPORÇÕES

# INFERÊNCIAS SOBRE DUAS PROPORÇÕES

– Objetivo é de usar duas proporções amostrais:

  – Para **teste de afirmativa** sobre duas proporções populacionais.

ou

  – Para construção de estimativa de **intervalo de confiança** da diferença entre proporções populacionais correspondentes.

# REQUISITOS

– No **teste de hipótese** sobre duas proporções populacionais ou na construção de um **intervalo de confiança** para diferença entre duas proporções populacionais, temos estes requisitos:

  – Temos proporções de duas **amostras aleatórias simples independentes** (valores amostrais selecionados de uma população não estão relacionados ou emparelhados com valores amostrais selecionados da outra população).

  – Para cada uma das duas amostras, o número de sucessos é, pelo menos, cinco e o número de fracassos também.

# NOTAÇÃO PARA DUAS PROPORÇÕES

– Para a população 1, fazemos:

  – $p_1$ = proporção populacional

  – $n_1$ = tamanho da amostra

  – $x_1$ = número de sucessos na amostra

  – Proporção amostral: $\hat{p}_1 = \dfrac{x_1}{n_1}$

  – $\hat{q}_1 = 1 - \hat{p}_1$

– A população 2 possui o mesmo tipo de notação.

# PROPORÇÃO AMOSTRAL COMBINADA

– A proporção amostral combinada é simbolizada por *p*-barra e é dada por:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

– O complementar de *p*-barra é dado por:

$$\bar{q} = 1 - \bar{p}$$

# ESTATÍSTICA DE TESTE PARA DUAS PROPORÇÕES

– Hipótese nula ($H_0$): $p_1 = p_2$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

– Onde: $p_1 - p_2 = 0$ (pressuposto na hipótese nula)

$$\hat{p}_1 = \frac{x_1}{n_1} \quad \text{e} \quad \hat{p}_2 = \frac{x_2}{n_2}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad \text{e} \quad \bar{q} = 1 - \bar{p}$$

– Os valores de *p* e valores críticos são encontrados com base no valor calculado do escore *z* (Tabela A-2).

# ESTIMATIVA DE INTERVALO DE CONFIANÇA

– A estimativa de intervalo de confiança para $p_1 – p_2$ é:

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

– Onde a margem de erro *E* é dada por:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

# TESTES DE HIPÓTESES

– Consideraremos testes de hipóteses sobre duas proporções populacionais:

$$H_0: p_1 = p_2$$

– Sob a suposição de proporções iguais, a melhor estimativa da proporção comum é obtida pela combinação de ambas amostras em uma amostra grande, de modo que *p*-barra se torna uma estimativa mais óbvia da proporção populacional comum.

# EXEMPLO

– Pensar que a política é importante na vida é maior entre homens do que entre mulheres?

| politica | homem 0 | 1 | Total |
|---|---|---|---|
| 0 | 22,394 | 19,634 | 42,028 |
| 1 | 15,262 | 17,822 | 33,084 |
| Total | 37,656 | 37,456 | 75,112 |

– $n_0$ = 37.656

– $n_1$ = 37.456

– $H_0$: $p_0$=$p_1$

– $H_1$: $p_1$>$p_0$

– $\alpha$ = 0,05

$$\bar{p} = \frac{x_0 + x_1}{n_0 + n_1} = \frac{15.262 + 17.822}{37.656 + 37.456} = 0{,}44046224$$

$$z = \frac{(\hat{p}_1 - \hat{p}_0) - (p_1 - p_0)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_0}}}$$

$$z = \frac{\left(\frac{17.822}{37.456} - \frac{15.262}{37.656}\right) - (0)}{\sqrt{\frac{(0{,}44046224)(0{,}55953776)}{37.456} + \frac{(0{,}44046224)(0{,}55953776)}{37.656}}} = 19{,}46$$

– $P$≈0 é menor do que $\alpha$=0,05. Rejeitamos hipótese nula. Há evidência de que política é mais importante dentre homens.

# DESVIO PADRÃO EXATO ≠ ESTIMADO

– Podemos construir uma estimativa de intervalo de confiança da diferença entre proporções populacionais ($p_1$–$p_2$).

– Se um intervalo de confiança não inclui o zero, temos evidência que sugere que $p_1$ e $p_2$ tenham valores diferentes.

– O desvio padrão usado para intervalos de confiança é diferente do desvio padrão usado para o teste de hipótese.

  – O **teste de hipótese** usa desvio padrão **exato**, baseado na suposição de que não há diferença entre proporções.

  – O **intervalo de confiança** usa um desvio padrão baseado em valores **estimados** das proporções populacionais.

# INTERVALOS DE CONFIANÇA

– Se desejo é de estimar diferença entre duas proporções, utilize o **intervalo de confiança**.

– Se desejo é de testar alguma afirmativa sobre duas proporções, use um método de **teste de hipótese**.

– **NÃO** teste a igualdade de duas proporções populacionais pela determinação da existência de sobreposição de dois intervalos de confiança individuais.

– A análise da sobreposição de dois intervalos de confiança individuais é mais conservadora (menos rejeição de $H_0$) do que estimativa de um intervalo de confiança $p_1 – p_2$.

# EXEMPLO DE INTERVALO DE CONFIANÇA

– Use os dados do exemplo anterior para construir intervalo de 95% de confiança para a diferença entre as proporções.

– α = 0,05

– $z_{\alpha/2}$ = 1,96

– Margem de erro:

– $\hat{p}_0$= 15.262/37.656 = 0,4053

– $\hat{p}_1$= 17.822/37.456 = 0,4758

– $\hat{p}_1 - \hat{p}_0$ = 0,0705

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_0\hat{q}_0}{n_0} + \frac{\hat{p}_1\hat{q}_1}{n_1}}$$

$$E = 1{,}96\sqrt{\frac{\left(\frac{15.262}{37.656}\right)\left(\frac{22.394}{37.656}\right)}{37.656} + \frac{\left(\frac{17.822}{37.456}\right)\left(\frac{19.634}{37.456}\right)}{37.456}} = 1{,}96 * 0{,}0036 = 0{,}0071$$

– Intervalo de confiança:

$$(\hat{p}_1 - \hat{p}_0) - E < (p_1 - p_0) < (\hat{p}_1 - \hat{p}_0) + E$$

$$(0{,}4758 – 0{,}4053) – 0{,}0071 < (p_1 – p_0) < (0{,}4758 – 0{,}4053) + 0{,}0071$$

$$0{,}0634 < (p_1 – p_0) < 0{,}0776$$

# INTERPRETAÇÃO DO RESULTADO

– Limites do intervalo de confiança não contêm zero, sugerindo que há diferença significante entre as duas proporções populacionais.

– Temos 95% de confiança que porcentagem de homens que pensam que política é importante é maior do que porcentagem de mulheres que pensam que política é importante por uma quantidade entre 6,34% e 7,76%.

# INFERÊNCIAS SOBRE DUAS MÉDIAS: AMOSTRAS INDEPENDENTES

# DEFINIÇÕES DE AMOSTRAS

- **Amostras independentes:** valores amostrais de uma população não estão relacionados ou combinados com os valores amostrais selecionados da outra população.
  - Ex.: grupo de tratamento e grupo de controle.

- **Amostras dependentes:** membros de uma amostra podem ser usados para determinar os membros da outra amostra.
  - Consistem em dados emparelhados dependentes, tais como dados de marido/mulher.
  - Dependência pode ocorrer com amostras relacionadas por associações como membros de uma família.
  - Ex.: dados coletados antes e depois de política pública.

# INFERÊNCIAS SOBRE DUAS MÉDIAS

– Serão apresentados métodos para uso de dados amostrais provenientes de **duas amostras independentes** para:
  – Teste de hipóteses sobre duas médias populacionais.
  – Construção de estimativas de intervalos de confiança para diferença entre duas médias populacionais.

– Esses métodos podem ser aplicados a situações em que:
  – Desvios padrões das duas populações são **desconhecidos** e **diferentes**: métodos mais realistas, têm melhor desempenho e serão apresentados a seguir.
  – Desvios padrões das duas populações são **conhecidos**.
  – Desvios padrões das duas populações são **desconhecidos**, mas **se supõe que sejam iguais**.

# $\sigma_1$ E $\sigma_2$ DESCONHECIDOS E DIFERENTES

- Ao usar duas amostras independentes para testar afirmativa sobre diferença ($\mu_1 - \mu_2$) ou para construir intervalo de confiança utilize este requisitos:
  - $\sigma_1$ e $\sigma_2$ são desconhecidos e não se faz suposição sobre igualdade entre eles.
  - Duas amostras são independentes.
  - Amostras aleatórias simples.
  - Uma ou ambas destas condições são satisfeitas:
    - Duas amostras são grandes ($n_1 > 30$ e $n_2 > 30$).
    - Amostras provêm de populações com distribuições normais:
      - Em amostras pequenas, procedimentos funcionam se não houver *outliers*.

# TESTE DE HIPÓTESE PARA DUAS MÉDIAS

– Para obter estatística do teste de hipótese para duas médias com amostras independentes, utilize:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

– Ao determinar valores críticos ou valores *P*, é preciso obter o número de graus de liberdade (gl):

  – No livro, gl é o menor número entre $n_1$–1 e $n_2$–1.

  – Nos pacotes estatísticos:

$$gl = \frac{(A + B)^2}{\frac{A^2}{n_1 - 1} + \frac{B^2}{n_2 - 1}} \quad \text{onde: } A = \frac{s_1^2}{n_1} \quad \text{e} \quad B = \frac{s_2^2}{n_2}$$

# INTERVALO DE CONFIANÇA PARA $\mu_1 - \mu_2$

– Intervalo de confiança para a diferença $\mu_1 - \mu_2$ é:

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

– Onde:

$$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

– Graus de liberdade é o mesmo usado para teste de hipótese.

# EXPLORANDO CONJUNTOS DE DADOS

– Antes de realizar teste de hipótese ou construir intervalo de confiança, devemos explorar as duas amostras:
  – Encontrar estatísticas descritivas para ambos conjuntos de dados (*n*, média e desvio padrão).
  – Fazer diagramas de caixa para os dois conjuntos de dados com a mesma escala.
  – Fazer histogramas do dois conjuntos de dados para comparar suas distribuições.
  – Identificar valores extremos (*outliers*).