Tema 3: Estimação por Intervalos de Confiança

Estatística Aplicada I

Catarina S. Nunes

CatarinaS.Nunes@uab.pt

Departamento de Ciências e Tecnologia

Universidade Aberta

UNIVERSIDADE ABERTA www.uab.pt

Conteúdos Teóricos

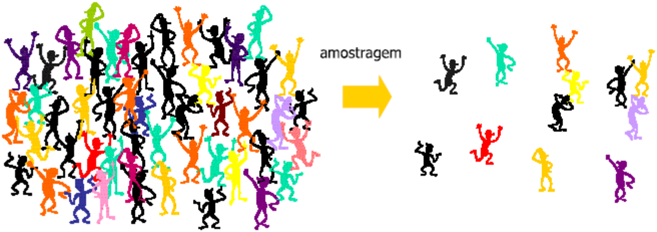
- Principais distribuições amostrais: proporções; média; diferenças; variância; razão de variâncias.
- Intervalos de confiança para a média populacional
- Intervalos de confiança para a variância e para o desvio padrão populacionais
- Intervalos de confiança para a diferença entre médias populacionais
- Intervalos de confiança para a razão de variâncias populacionais
- Intervalos de confiança para proporções
- Intervalos de confiança para a diferença de proporções

Estatística Inferencial

- Estatística Inferencial (ou Indutiva): a partir de uma amostra inferir sobre as características de uma população
- Podemos inferir determinadas características de uma população se extraírmos uma amostra representativa desta:

População: colecção de unidades individuais (pessoas ou resultados experimentais) com uma ou mais características comuns, que se pretendem estudar.

Amostra: Conjunto de dados ou observações, recolhidos a partir de um subconjunto da população, que se estuda com o objectivo de tirar conclusões para a população de onde foi recolhida

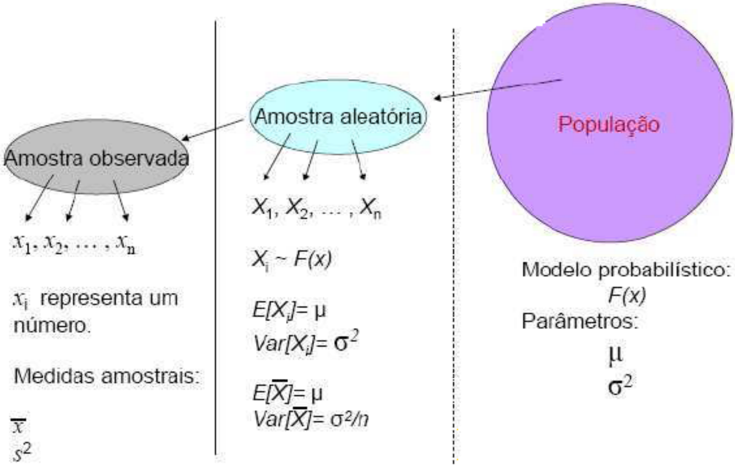


Amostragem

Amostragem: Processo pelo qual se extraem dados de uma população

- Amostragem aleatória simples: cada elemento da amostra é retirado aleatoriamnte de toda a população (com ou sem reposição) $\Rightarrow$ cada amostra tem a mesma probabilidade de ser recolhida;
- Amostragem estratificada: subdividir a população em, pelo menos, dois subgrupos distintos que partilham alguma característica e, em seguida, recolher uma amostra de cada um dos subgrupos (estratos);
- Amostragem por clusters: dividir a população em clusters (grupos); selecionar aleatoriamente alguns desses clusters; escolher todos os membros dos clusters selecionados.

Amostra Aleatória



ABERTA
www.uab.pt

Parâmetro versus Estatística

- **Parâmetro**: medida utilizada para descrever a distribuição da população
  - o valor esperado (média) μ e a variância σ^2 (ou desvio padrão σ) são parâmetros de uma distribuição Normal - $N \sim (\mu, \sigma^2)$
  - a probabilidade de sucesso p é um parâmetro da distribuição Binomial - $B \sim (n, p)$
- **Estatística**: função de uma amostra aleatória que não depende de parâmetros desconhecidos; é uma v.a. que estima (pontualmente) um parâmetro populacional as vezes é chamada de simplesmente de estimador
  - média amostral: $\bar{X} = 1/n \sum_{i=1}^{n} X_i$
  - variância amostral: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- **Estimativa pontual**: valor obtido a partir da amostra (valor do estimador) que se destina a fornecer valores aproximados do parâmetro

Estimação por intervalos

Estimação pontual (estatísticas) versus Estimação por intervalo (intervalos de confiança)

- Uma estimativa pontual de um parâmetro não contém informação sobre a precisão do valor obtido;
- A variância e o erro quadrático médio do estimador fornecem também alguma informação;
- Uma forma mais completa de abordar a questão consiste em construir estimativas na forma de intervalos e conhecer a probabilidade de o intervalo conter o verdadeiro valor do parâmetro.

Intervalo de Confiança

Um intervalo de confiança para um parâmetro θ, a um grau de confiança $1-\alpha$, é o um intervalo aleatório (L_{inf}, L_{sup}) tal que:

$$P(L_{inf} < \theta < L_{sup}) = 1-\alpha, \quad \alpha \in (0,1)$$

onde α deve ser um valor muito reduzido por forma a termos confianças elevadas.

- Valores usuais para o grau de confiança são 95%, 99% e 90%;
- Para cada amostra que se observa obtém-se (em geral) um intervalo de confiança diferente para o mesmo parâmetro;
- Dizer que um intervalo tem grau de confiança $1-\alpha$ é o mesmo que dizer que se observarmos muitas amostras distintas, os intervalos que se obtêm contêm o verdadeiro valor do parâmetro $(1-\alpha)\times 100\%$ das vezes.

Intervalo de Confiança (IC)

1. IC para a média (valor esperado) μ de uma população Normal com variância (σ^2) conhecida
2. IC para a média (valor esperado) μ de uma população Normal com variância (σ^2) desconhecida
3. IC para a variância σ^2 de uma população Normal
4. IC para a diferença de médias de duas populações Normais $\mu_1 - \mu_2$; duas amostras independentes, variâncias conhecidas (σ_1^2, σ_2^2)
5. IC para a diferença de médias de duas populações Normais $\mu_1 - \mu_2$; duas amostras independentes, variâncias desconhecidas mas supostamente iguais ($\sigma_1^2 = \sigma_2^2$)
6. IC para a razão de variâncias $\frac{\sigma_1^2}{\sigma_2^2}$
7. IC para uma proporção p
8. IC para a diferença entre duas proporções $p_1 - p_2$

ABERTA www.uab.pt

1. IC para a média μ de uma população Normal com variância (σ^2) conhecida

Um intervalo de confiança para a média μ de uma população Normal com variância (σ^2) conhecida, a um grau de confiança $1-\alpha$, é:

$$\left]\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} ; \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right[$$

onde:

- n é o tamanho da amostra
- $\bar{X}$ é a média amostral: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $z_{1-\frac{\alpha}{2}}$ é o quartil de ordem $1-\frac{\alpha}{2}$ da distribuição Normal Reduzida (ou Padrão); $Z \sim N(0,1)$ e $P\left(Z < z_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$

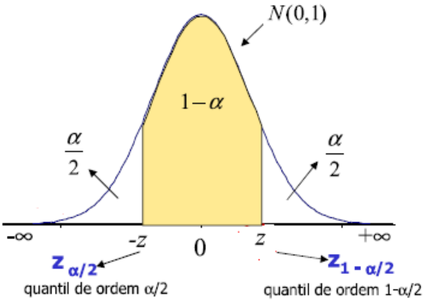
1. Interpretação do IC para a média μ

$X \sim N(\mu, \sigma^2)$ com μ desconhecido mas σ^2 conhecida

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Variável fulcral:

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \text{ (Distribuição Normal Reduzida)}$$



AbERTA
www.uab.pt

1. Interpretação do IC para a média μ

$$P(-z < Z < z) = 1-\alpha \Leftrightarrow P\left(-z < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z\right) = 1-\alpha$$

$$\Leftrightarrow P\left(-z\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\bar{X}-z\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Downarrow$$

$$IC_{1-\alpha}(\mu) =]\bar{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} ; \bar{X}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}[$$

- Quanto maior o nível de confiança maior a largura do intervalo;
- Quanto maior a variância maior a largura do intervalo;
- Quanto maior a amostra (maior o *n*) menor a largura do intervalo.

1. Interpretação do IC para a média μ

Para uma amostra aleatória de tamanho 50, seguindo uma distribuição Normal de média $\mu = 10$ e variância $\sigma^2 = 4 \longrightarrow X \sim N(10, 4)$, determinamos o IC para μ com 95% de grau de confiança ($\alpha = 0.05$):

$$P\left(\bar{X} - 1.96\frac{2}{\sqrt{50}} < \mu < \bar{X} + 1.96\frac{2}{\sqrt{50}}\right) = 95\%$$

$$P\left(\bar{X} - 0.5544 < \mu < \bar{X} + 0.5544\right) = 95\%$$

$$\Downarrow$$

$$IC_{95\%}(\mu) =]\bar{X} - 0.5544; \bar{X} + 0.5544[$$

Interpretação: 95% dos possíveis ICs obtidos a partir de uma amostra de tamanho 50, conterão de fato o verdadeiro valor da média $\mu = 10$

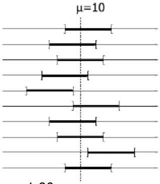


Tabela da Distribuição normal reduzida: $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$

IC & Grau de Confiança

Como poderia obter intervalos de confiança mais estreitos, ou seja, com limites mais próximos da média verdadeira?

Diminuindo o grau de confiança

Diminuindo o grau de confiança de 99% para 95%, aumentamos o risco de estar errados: de 1% de risco passamos a 5% de risco, ou seja temos mais mais possibilidades (5/100 em vez de 1/100) de que o IC não contenha a média populacional. Ao aumentar o risco, o intervalo deve ser mais preciso.

- 90% grau de confiança - existem 10 possibilidades em 100 que o IC não contenha a média populacional
- 95% grau de confiança - existem 5 possibilidades em 100 que o IC não contenha a média populacional
- 90% grau de confiança - existe 1 possibilidade em 100 que o IC não contenha a média populacional

IC & Dimensão da Amostra

Como poderia obter intervalos de confiança mais estreitos, ou seja, com limites mais próximos da média verdadeira?

Aumentando a dimensão da amostra

ABERTA www.uab.pt

1. IC para a média μ de uma população genérica com variância (σ^2) conhecida

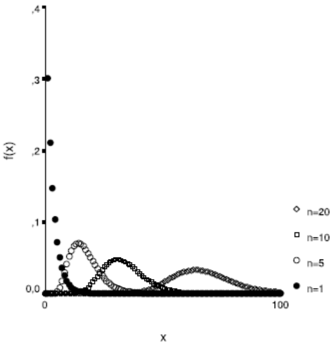
Se uma distribuição qualquer tiver média μ (desconhecida) e variância σ^2 (conhecida) e se forem válidas as condições do Teorema do Limite Central ($n >> 30$), podemos obter um IC aproximado para a média μ

IC para μ, a um grau de confiança $1-\alpha$, é: $\left]\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}};\bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$

Teorema do Limite Central

Seja $X_1, X_2, \ldots, X_n$ uma sucessão de variáveis aleatórias idd (independentes e identicamente distribuidas) com $E(X_i) = \mu$, $VAR(X_i) = \sigma^2, i = 1, \ldots, n$. Então,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \approx N(0,1), n \to \infty$$



1. IC para μ com variância conhecida (σ^2) - Resumo

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \begin{cases} \sim N(0,1) \;\; se \;\; X \sim N(\mu,\sigma^2) \\ \approx N(0,1) \;\; se \;\; X \;\; qualquer \;\; e \;\; n \geq 30 \end{cases}$$

$$\Rightarrow IC_{1-\alpha}(\mu) =]\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}[$$

$$\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- quanto maior $z_{1-\frac{\alpha}{2}} \Rightarrow$ IC menos preciso;
- quanto maior $n \Rightarrow$ menor o erro padrão ($\sigma/\sqrt{n}$), e IC mais preciso;
- se aumentarmos o grau de confiança $\Rightarrow$ a precisão diminui por porque aumenta o valor de $z_{1-\frac{\alpha}{2}}$:

$1-\alpha$	$z_{1-\frac{\alpha}{2}}$
90%	1.65
95%	1.96
99%	2.58

1. IC para a média μ de uma população Normal com variância (σ^2) desconhecida

Um intervalo de confiança para para a média μ de uma população Normal com variância (σ^2) desconhecida, a um grau de confiança $1-\alpha$, é:

$$\left]\bar{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} ; \bar{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right[$$

onde:

- n é o tamanho da amostra
- $\bar{X}$ é a média amostral $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- S^2 é a variância amostral corrigida $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- $t_{n-1,1-\frac{\alpha}{2}}$ é o quartil de ordem $1-\frac{\alpha}{2}$ da distribuição t de Student com $n-1$ graus de liberdade; $T \sim t_{n-1}$ e $P(T < t_{n-1,1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$

Estes intervalos têm uma largura maior do que se o valor de σ^2 fosse conhecido, refletindo a incerteza acrescida pelo desconhecimento deste parâmetro.

ABERTA www.uab.pt

2. Interpretação do IC para a média μ

$X \sim N(\mu, \sigma^2)$ com μ desconhecido e σ^2 desconhecida

Variável fulcral:

$$T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1} \text{ (Distribuição } t\text{-Student com } n-1 \text{ graus de liberdade)}$$

$$P\left(-t < T < t\right) = 1-\alpha \Leftrightarrow P\left(-t < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < t\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\bar{X} - t\frac{S}{\sqrt{n}} < \mu < \bar{X} + t\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$\Downarrow$$

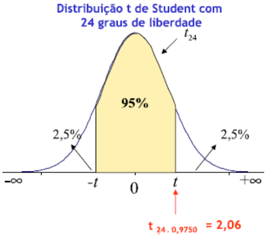
$$IC_{1-\alpha}(\mu) =]\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}; \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}[$$

2. Interpretação do IC para a média μ

Exemplo:

Uma v.a. X tem uma distribuição Normal com média μ e variância σ^2 desconhecidas. Retira-se uma amostra de 25 valores e calcula-se a média amostral e variância amostral. Construa um IC de 95% para μ supondo que $\bar{X} = 12.7$ e $S^2 = 16$

$$IC_{1-\alpha}(\mu) =]\bar{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} ; \bar{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}[$$



$$IC_{95\%}(\mu) =]\bar{X} - t_{24,0.975} \frac{S}{\sqrt{n}} ; \bar{X} + t_{24,0.975} \frac{S}{\sqrt{n}}[$$
$$=]12.7 - 2.06 \frac{4}{\sqrt{25}} ; 12.7 + 2.06 \frac{4}{\sqrt{25}}[$$
$$=]12.7 - 1.648 ; 12.7 + 1.648[$$
$$IC_{95\%}(\mu) =]11.052 ; 13.648[$$

ABERTA www.uab.pt

1. IC para a média μ de uma população com variância (σ^2) desconhecida

Uma abordagem para uma v.a. X com qualquer distribuição mas aproximada pela Normal, amostras grandes $\Rightarrow$ usar a dintribuição normal reduzida.

Um intervalo de confiança para a média μ de uma população qualquer com variância (σ^2) desconhecida, amostras grandes ($n >> 30$), a um grau de confiança $1-\alpha$, é:

$$IC_{1-\alpha}(\mu) \approx]\bar{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}; \bar{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}[$$

onde:

- X qualquer mas $n >> 30$
- $Z = \frac{\bar{X}-\mu}{S/\sqrt{n}} \approx N(0,1)$
- $z_{1-\frac{\alpha}{2}}$ é o quartil de ordem $1-\frac{\alpha}{2}$ da distribuição Normal reduzida $N(0,1)$

3. IC para a variância σ^2 de uma população Normal

Um intervalo de confiança para a variância σ^2 de uma população Normal, a um grau de confiança $1-\alpha$, é:

$$\left]\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}};\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right[$$

onde:

- n é o tamanho da amostra
- S^2 é a variância amostral corrigida $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$
- $\chi^2_{n-1,1-\frac{\alpha}{2}}$ é o quartil de ordem $1-\frac{\alpha}{2}$ da distribuição Chi-Quadrado (χ^2) com $n-1$ graus de liberdade;

ABERTA
www.uab.pt

3. Interpretação do IC para a variância σ^2

$X \sim N(\mu, \sigma^2)$ com σ^2 desconhecida

Variável fulcral:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \approx \chi^2_{n-1} \text{ (Distribuição } \chi^2 \text{ com } n-1 \text{ graus de liberdade)}$$

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}} < \chi^2 < \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = 1-\alpha$$

$$\Downarrow$$

$$IC_{1-\alpha}(\sigma^2) = \left]\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}; \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right[$$

UNIVERSIDADE ABERTA www.uab.pt

3. IC para o desvio padrão σ

De um modo geral (para uma amostra de dimensão n qualquer) o intervalo de confiança para o desvio padrão populacional σ, a um grau de confiança $1-\alpha$, pode ser calculado simplesmente a partir do IC para a variância, é:

$$\left]S\sqrt{\frac{(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}}\,;\,S\sqrt{\frac{(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}}\right[$$

No caso de populações Normais com amostras de grandes dimensões, a distribuição amostral do desvio padrão pode ser considerada Normal.

O intervalo de confiança para o desvio padrão σ, com base numa amostra de dimensão grande (n muito grande), a um grau de confiança $1-\alpha$, pode ser:

$$\left]S - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{2n}}\,;\,S + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{2n}}\right[$$

4. IC para a diferença de médias de duas populações Normais $\mu_1 - \mu_2$; duas amostras independentes, variâncias conhecidas (σ_1^2, σ_2^2)

Um intervalo de confiança para diferença de médias $\mu_1 - \mu_2$ de duas populações Normais com variâncias conhecidas σ_1^2, σ_2^2 obtido a partir de duas amostras independentes, a um grau de confiança $1-\alpha$, é:

$$\left]\bar{X}_1 - \bar{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; \bar{X}_1 - \bar{X}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right[$$

onde:

- n_1 é o tamanho da amostra de X_1 e n_2 é o tamanho da amostra de X_2
- $\bar{X}_1$ e $\bar{X}_2$ são as médias amostrais de X_1 e X_2, respetivamente
- $z_{1-\frac{\alpha}{2}}$ é o quartil de ordem $1-\frac{\alpha}{2}$ da distribuição Normal reduzida $N(0,1)$;

2. Interpretação do IC para a diferença de médias

Dadas duas v.a. independentes X_1 e X_2 de populações normais $N(\mu_1, \sigma_1^2)$ e $N(\mu_2, \sigma_2^2)$, respetivamente, podemos considerar a estatística:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$\Longleftrightarrow Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$\Downarrow$$

$$IC_{1-\alpha}(\mu_1 - \mu_2) =]\bar{X}_1 - \bar{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; \bar{X}_1 - \bar{X}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}[$$

5. IC para a diferença de médias de duas populações Normais $\mu_1 - \mu_2$; duas amostras independentes, com variâncias desconhecidas (σ_1^2, σ_2^2), mas suposto iguais

Um intervalo de confiança para diferença de médias $\mu_1 - \mu_2$ de duas populações Normais com variâncias desconhecidas σ_1^2, σ_2^2, mas iguais, obtido a partir de duas amostras independentes, a um grau de confiança $1 - \alpha$, é:

$$\left]\bar{X}_1 - \bar{X}_2 - t_{n_1+n_2-2,1-\frac{\alpha}{2}} S^*; \bar{X}_1 - \bar{X}_2 + t_{n_1+n_2-2,1-\frac{\alpha}{2}} S^*\right[$$

onde:

- $S^* = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
- S_1^2 e S_2^2 são as variâncias amostrais de X_1 e X_2, respetivamente
- $t_{n_1+n_2-2,1-\frac{\alpha}{2}}$ é o quartil de ordem $1 - \frac{\alpha}{2}$ da distribuição t-Student com $n_1 + n_2 - 2$ graus de liberdade

5. IC para a diferença de médias de duas populações quaisquer $\mu_1 - \mu_2$; duas amostras independentes, com variâncias desconhecidas (σ_1^2, σ_2^2), e n_1 e $n_2 >> 30$ (amostras grandes)

Um intervalo de confiança para diferença de médias $\mu_1 - \mu_2$ de duas populações quaisquer com variâncias desconhecidas σ_1^2, σ_2^2, e amostras muito grandes, obtido a partir de duas amostras independentes, a um grau de confiança $1 - \alpha$, pode ser aproximado por:

$$\left]\bar{X}_1 - \bar{X}_2 - z_{1-\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}; \bar{X}_1 - \bar{X}_2 + z_{1-\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right[$$

onde:

- S_1^2 e S_2^2 são as variâncias amostrais de X_1 e X_2, respetivamente
- $z_{1-\frac{\alpha}{2}}$ é o quartil de ordem $1 - \frac{\alpha}{2}$ da distribuição Normal reduzida $N(0, 1)$

5. IC para a diferença de médias de duas populações Normais $\mu_1 - \mu_2$, com variâncias desconhecidas (σ_1^2, σ_2^2), e amostras emparelhadas de igual dimensão

Um intervalo de confiança para diferença de médias $\mu_1 - \mu_2 = \mu_D$ de duas populações Normais com variâncias desconhecidas σ_1^2, σ_2^2, obtido a partir de duas amostras emparelhadas de tamanho n, a um grau de confiança $1 - \alpha$, é:

$$\left]\bar{D} - t_{n-1,1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}} ; \bar{D} + t_{n-1,1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}\right[$$

onde:

- $D = X_1 - X_2$
- $\bar{D}$ e S_D^2 são respetivamente a média e a variância amostral das diferenças
- $t_{n-1,1-\frac{\alpha}{2}}$ é o quartil de ordem $1 - \frac{\alpha}{2}$ da distribuição t-Student com $n-1$ graus de liberdade

5. Interpretação do IC para $\mu_1 - \mu_2$, com variâncias desconhecidas (σ_1^2, σ_2^2), e amostras emparelhadas de igual dimensão

Sejam $X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim N(\mu_2, \sigma_2^2)$.

As amostras são dependentes na medida em que têm a mesma dimensão (n) e cada observação X_{1i} depende da observação X_{2i}, mas os pares (X_{1i}, X_{2i}) e (X_{1j}, X_{2j}) são independentes $(i \neq j)$. Este tipo de amostras chamam-se amostras emparelhadas.

Dadas duas amostras aleatórias emparelhadas $(X_{11}, X_{12}, \ldots, X_{1n})$, $(X_{21}, X_{22}, \ldots, X_{2n})$ provenientes de populações Normais, consideram-se as diferenças:

$$D_i = X_{1i} - X_{2i} \sim N(\mu_D, \sigma_D^2)$$

onde μ_D é igual ás diferenças das médias das populações e σ_D^2 representa a variância das diferenças D_i.

A variável:

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

onde S_D representa o desvio padrão amostral das diferenças.

6. IC para a razão de variâncias $\frac{\sigma_1^2}{\sigma_2^2}$ de duas populações normais

Um intervalo de confiança para a razão de variâncias $\frac{\sigma_1^2}{\sigma_2^2}$ de duas populações normais, a um grau de confiança $1-\alpha$, é:

$$\left] \frac{1}{F_{n_1-1;n_2-1,1-\frac{\alpha}{2}}} \frac{S_1^2}{S_2^2} ; \frac{1}{F_{n_1-1;n_2-1,\frac{\alpha}{2}}} \frac{S_1^2}{S_2^2} \right[$$

onde:

- $n1$ e n_2 são o tamanho das amostras de X_1 e X_2, respetivamente
- S_1^2 e S_2^2 são a variância amostral de X_1 e X_2, respetivamente
- $F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ e $F_{n_1-1,n_2-1,\frac{\alpha}{2}}$ são valores da distribuição F de Snedecor com n_1-1 graus de liberdade no numerador e n_2-1 graus de liberdade do denominador

6. Interpretação do IC para a razão de variâncias $\frac{\sigma_1^2}{\sigma_2^2}$

$X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim N(\mu_2, \sigma_2^2)$ com σ_1^2 e σ_2^2 desconhecidas

Variável fulcral:

$$F = \frac{\frac{S_1^2}{S_2^2}}{\frac{\sigma_1^2}{\sigma_2^2}} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}$$ segue uma distribuição F de Snedcor com $n_1 - 1$ e $n_2 - 1$ graus de liberdade.

$$P\left(F_{n_1-1,n_2-1,\frac{\alpha}{2}} < F < F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}\right) = 1 - \alpha \Leftrightarrow$$

$$P\left(F_{n_1-1,n_2-1,\frac{\alpha}{2}} < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}\right) = 1 - \alpha$$

$$\Downarrow$$

$$IC_{1-\alpha}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left]\frac{1}{F_{n_1-1;n_2-1,1-\frac{\alpha}{2}}}\frac{S_1^2}{S_2^2}; \frac{1}{F_{n_1-1;n_2-1,\frac{\alpha}{2}}}\frac{S_1^2}{S_2^2}\right[$$

6. Interpretação do IC para a razão de variâncias $\frac{\sigma_1^2}{\sigma_2^2}$

Note-se que a distribuição F goza da seguinte propriedade:

$$F_{n_1-1,n_2-1,\frac{\alpha}{2}} = \frac{1}{F_{n_2-1,n_1-1,1-\frac{\alpha}{2}}}$$

e portanto o intervalo de confiança para a razão de variâncias também pode ser escrito como:

$$IC_{1-\alpha}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left] \frac{1}{F_{n_1-1;n_2-1,1-\frac{\alpha}{2}}} \frac{S_1^2}{S_2^2} ; F_{n_2-1;n_1-1,1-\frac{\alpha}{2}} \frac{S_1^2}{S_2^2} \right[$$

7. IC para uma proporção p

Um intervalo de confiança para uma proporção populacional p (*e.g.* proporção de indivíduos com uma certa característica na população), a um grau de confiança $1-\alpha$, é:

$$\left]\hat{p}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}};\hat{p}+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right[$$

onde:

- n é tamanho da amostra
- $\hat{p}=\frac{X}{n}$ é a proporção de individuos com uma certa característica numa amostra aleatória de dimensão n
- $z_{1-\frac{\alpha}{2}}$ é o quartil de ordem $1-\frac{\alpha}{2}$ da distribuição Normal reduzida $N(0,1)$

7. Interpretação do IC para p

Seja X a frequência absoluta do número de ocorrências de sucesso (de uma certa característica) numa amostra aleatória de tamanho n, então X é uma variável aleatória Binomial $X \sim B(np, np(1-p))$. Um estimador para a proporção de sucessos p da população é: $\hat{p} = \frac{X}{n}$

Variável fulcral:

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$$

$$P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow P\left(-z_{1-\frac{\alpha}{2}} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

substituindo nos limites do intervalo, p pelo estimador amostral $\hat{p}$, infere-se então:

$$IC_{1-\alpha}(p) =]\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}[$$

7. Interpretação do IC para p - Exemplo

Em 100 acessos a páginas de internet escolhidas ao acaso, 30 são as páginas nacionais. Determine um IC a 95% para a proporção de acessos a páginas nacionais.

X - número de acessos a páginas internet nacionais $\Rightarrow X \sim B(100, p)$

p - proporção de acessos a páginas nacionais (em geral) $\Rightarrow p$ desconhecido

Utilizando esta amostra determinamos um IC aproximado para p a 95%:

$$IC_{1-\alpha}(p) =]\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}[$$

$$\hat{p} = \frac{X}{n} = \frac{30}{100} = 0.3 \qquad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.3\times 0.7}{100}} = 0.04582$$

$$z_{1-\frac{\alpha}{2}} = z_{0.9750} = 1.96$$

$$IC_{95\%}(p) =]0.3 - 1.96 \times 0.04582; 0.3 + 1.96 \times 0.04582[=]0.2102; 0.3898[$$

8. IC para a diferença entre duas proporções $p_1 - p_2$

Um intervalo de confiança para a diferença de proporções $p1 - p_2$, a um grau de confiança $1-\alpha$, é:

$$\left]\left(\hat{p}_1 - \hat{p}_2\right) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}};\left(\hat{p}_1 - \hat{p}_2\right) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right[$$

onde:

- n_1 e n_2 são o tamanho das amostras independentes de X_1 e X_2, respetivamente
- $\hat{p}_1 = \frac{X_1}{n_1}$ e $\hat{p}_2 = \frac{X_2}{n_2}$, valores observados para a proporção de sucessos nas amostras independentes das duas populações
- $z_{1-\frac{\alpha}{2}}$ é o quartil de ordem $1-\frac{\alpha}{2}$ da distribuição Normal reduzida $N(0,1)$

8. Interpretação do IC para $p_1 - p_2$

Seja X_1 a frequência absoluta do número de ocorrências de sucesso numa amostra aleatória de tamanho n_1 de uma população, então X_1 é uma variável aleatória Binomial $X_1 \sim B(n_1 p_1, n_1 p_1(1-p_1))$. E seja X_2 a frequência absoluta do número de ocorrências de sucesso numa amostra aleatória de tamanho n_2 de uma população, então X_2 é uma variável aleatória Binomial $X_2 \sim B(n_2 p_2, n_2 p_2(1-p_2))$. Sendo as duas amostras independentes.

Um estimador para a diferença entre proporções de sucessos $p_1 - p_2$ é:

$$(\hat{p_1} - \hat{p_2}) \approx N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

Variável fulcral:

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$$

seguindo o mesmo raciocínio que para o IC de uma proporção, temos:

$$P\left((\hat{p_1} - \hat{p_2}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} < p_1 - p_2 < (\hat{p_1} - \hat{p_2}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

$$IC_{1-\alpha}(p_1 - p_2) = \left](\hat{p_1} - \hat{p_2}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} ; (\hat{p_1} - \hat{p_2}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\right[$$