Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Probabilidade e Estatística 2
1^o semestre de 2020
Gustavo L. Gilardoni

Comparação de Duas Médias ou Proporções

2 de Setembro de 2020

0 Introdução

Ao longo destas notas de aula pensamos e discutimos os dois exemplos seguintes. É sugerido que o aluno reflita antes de ler as notas sobre como poderiam ser realizados experimentos para tentar responder a questões levantadas neles.

Exemplo 1. Uma locadora de carros possui uma grande frota de veículos todos da mesma marca, modelo e ano. Deseja testar dois tipos de combustíveis, digamos gasolina do tipo A e do tipo B, para saber qual deles oferece maior rendimento, medido como kilômetros/litro. Como poderíamos fazer um experimento para tentar responder essa pergunta?

Exemplo 2. Uma fábrica de calçados pode usar dois materiais, digamos A e B, para fazer as solas de um modelo de tênis. Os dois materiais têm custo semelhante e a empresa vai decidir usar aquele que tiver menos desgaste ao uso. A empresa quer testar os tênis em condições reais de uso. Para isso existe uma lista de filhos de funcionários da empresa que poderiam usar tênis do material A e do material B durante um período de tempo específico, ao fim do qual seria medido o desgaste como a diferencia entre a espesura das solas antes e após o uso. Como poderíamos fazer um experimento para tentar responder essa pergunta?

1 Inferência para uma única população (Revisão)

A idéia central no problema de inferência para uma única média μ estudada em PE1 pode ser descrita da seguinte forma: Suponha que queremos fazer inferência para um parâmetro θ e que existe um estimador $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, cujo desvío padrão estimado (ou erro de estimação) $\hat{\sigma}_{\hat{\theta}}$ é tal que a distribuição (ou a distribuição limite quando $n \to \infty$) de

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} = \frac{\hat{\theta}(X_1, \ldots, X_n) - \theta}{\hat{\sigma}_{\hat{\theta}}}$$

não depende dos parâmetros do modelo (no jargão estatístico, diz-se que essa quantidade é um *pivote* para θ, no sentido que (i) a relação $\theta \mapsto (\hat{\theta} - \theta)/\hat{\sigma}_{\hat{\theta}}$ é monótona e (ii) a sua distribuição não depende dos parâmetros do modelo). Nesse caso, usando a distribuição conhecida de $(\hat{\theta} - \theta)/\hat{\sigma}_{\hat{\theta}}$, podemos achar por exemplo valores a_I e a_S tais que

$$1 - \alpha = P\left(a_I \le \frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} \le a_S\right) = P\left(\hat{\theta} - a_S\, \hat{\sigma}_{\hat{\theta}} \le \theta \le \hat{\theta} + a_I\, \hat{\sigma}_{\hat{\theta}}\right)$$

ou a_C tal que

$$\alpha = P\left(\frac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} > a_C\right).$$

A primeira dessas igualdades justifica usar o intervalo

$$\hat{\theta} - a_S\hat{\sigma}_{\hat{\theta}} \le \theta \le \hat{\theta} - a_I\hat{\sigma}_{\hat{\theta}}$$

como um *intervalo de confiança* (IC) de nível $100(1-\alpha)\%$ para θ, enquanto a segunda justifica usar a região crítica

$$\frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}} > a_C$$

(equivalentemente, $\hat{\theta} > \theta_0 + a_C\hat{\sigma}_{\hat{\theta}}$) para testar a $H_a:\ \theta > \theta_0$

Revisamos então alguns dos ICs e testes estudados em PE1:

- **Distribuição Normal, variância conhecida**: Suponha que $X_1,\ldots,X_n$ é uma amostra aleatória da distribuição Normal com média μ e variância **conhecida** σ_0^2. Nesse caso, $\theta=\mu$, $\hat{\theta}=\hat{\mu}=\bar{X}=n^{-1}\sum_{i=1}^n X_i$, $\hat{\sigma}_{\bar{X}}=\sigma_{\bar{X}}=\sqrt{\mathrm{Var}(\bar{X})}=\sigma_0/\sqrt{n}$ e

$$\frac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} = \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma_0} \sim \ \mathrm{Normal}(0,1)\,.$$

  Podemos definir portanto $a_I=-z_{\alpha/2}$, $a_S=z_{\alpha/2}$ e $a_C=z_\alpha$, onde z_α e o valor que deixa área ou probabilidade α a sua direita na distribuição N(0,1) (para achar z_α, usamos a tabela da Normal padrão ou, numa sessão do **R**, digite '`> qnorm(1-`α`)`'). Assim, o IC acima fica

$$\bar{x} - z_{\alpha/2}\,\frac{\sigma_0}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\,\frac{\sigma_0}{\sqrt{n}}\,,$$

  enquanto a região crítica para testar $H_a:\ \mu > \mu_0$ é

$$z_{obs} = \sqrt{n}\,\frac{\bar{x}-\mu_0}{\sigma_0} > z_\alpha$$

  ou, equivalentemente, $\bar{x} \ge \mu_0 + z_\alpha\,\frac{\sigma_0}{\sqrt{n}}$.

- **Qualquer distribuição, n grande**: Suponha que $X_1,\ldots,X_n$ é uma amostra aleatória de uma distribuição (qualquer uma!) com média μ e **variância finita** desconhecidas, que n é grande (no jargão estatístico pensa-se que $n\to\infty$). Então, se $S^2=(n-1)^{-1}\sum_{i=1}^n(X_i-\bar{X})^2$ é a variância da amostra, o Teorema Central do Limite (TCL) garante que a distribuição de $\sqrt{n}\,(\bar{X}-\mu)/S$ é aproximadamente (no jargão estatístico, assintoticamente) Normal padrão. Logo, podemos tomar na formulação geral $\theta=\mu$, $\hat{\theta}=\hat{\mu}=\bar{X}$, $\hat{\sigma}_{\bar{X}}=S/\sqrt{n}$,

$$\frac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \sqrt{n}\,\frac{\bar{X}-\mu}{S} \underset{\text{aprox.}}{\sim} \ \mathrm{Normal}(0,1)\,,$$

$a_I = -z_{\alpha/2}$, $a_S = z_{\alpha/2}$ e $a_C = z_\alpha$. Assim, o IC acima (agora aproximado, ou assintótico) fica

$$\bar{x} - z_{\alpha/2}\,\frac{S}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\,\frac{S}{\sqrt{n}}\,,$$

enquanto a região crítica para testar $H_a : \mu > \mu_0$ é

$$z_{obs} = \sqrt{n}\,\frac{\bar{x} - \mu_0}{S} > z_\alpha$$

ou, equivalentemente, $\bar{x} \ge \mu_0 + z_\alpha\,\frac{S}{\sqrt{n}}$.

- **Distribuição Normal, variância desconhecida**: Suponha que $X_1, \ldots, X_n$ é uma amostra aleatória de uma distribuição Normal com média μ e variância σ^2 desconhecidas. Então a quantidade $\sqrt{n}\,\frac{\bar{X}-\mu}{S}$ segue a (assim chamada) distribuição t de Student com $(n-1)$ graus de liberdade (gl). Portanto fazemos $\theta = \mu$, $\hat{\theta} = \hat{\mu} = \bar{X}$, $\hat{\sigma}_{\bar{X}} = S/\sqrt{n}$,

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \sqrt{n}\,\frac{\bar{X} - \mu}{S} \sim\; t_{n-1}$$

  $a_I = -t_{n-1;\alpha/2}$, $a_S = t_{n-1;\alpha/2}$ e $a_C = t_{n-1;\alpha}$, onde $t_{n-1;\alpha}$ e o valor que deixa área ou probabilidade α a sua direita na distribuição t com $(n-1)$ gl (para achar $t_{n-1;\alpha}$, usamos a tabela da distribuição t de Student ou, numa sessão do **R**, digite '`> qt(1-`α`,`$n-1$`)`'). Assim, o IC acima fica

$$\bar{x} - t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}}\,,$$

  enquanto a região crítica para testar $H_a : \mu > \mu_0$ é

$$t_{obs} = \sqrt{n}\,\frac{\bar{x} - \mu_0}{S} > t_{n-1;\alpha}$$

  ou, equivalentemente, $\bar{x} \ge \mu_0 + t_{n-1;\alpha}\,\frac{S}{\sqrt{n}}$.

- **Inferência para proporções, n grande**: Se $X \sim$ Binomial(n,p) e $\hat{p} = X/n$, segue pelo Teorema de DeMoivre-Laplace (que é um caso particular do TCL) que, quando $n \to \infty$, a distribuição de

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} = \frac{\hat{p} - p}{\sqrt{\hat{p}\,(1-\hat{p})/n}} = \sqrt{n}\,\frac{\hat{p} - p}{\sqrt{\hat{p}\,(1-\hat{p})}} \underset{\text{aprox.}}{\sim} \text{Normal}(0,1) \tag{1}$$

  ou, quando $p = p_0$,

$$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} = \frac{\hat{p} - p_0}{\sqrt{p_0\,(1-p_0)/n}} = \sqrt{n}\,\frac{\hat{p} - p_0}{\sqrt{p_0\,(1-p_0)}} \underset{\text{aprox.}}{\sim} \text{Normal}(0,1)\,,$$

  Assim por um argumento semelhante ao feito nos outros casos, o IC com confiança aproximada $100(1-\alpha)\%$ para p é

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

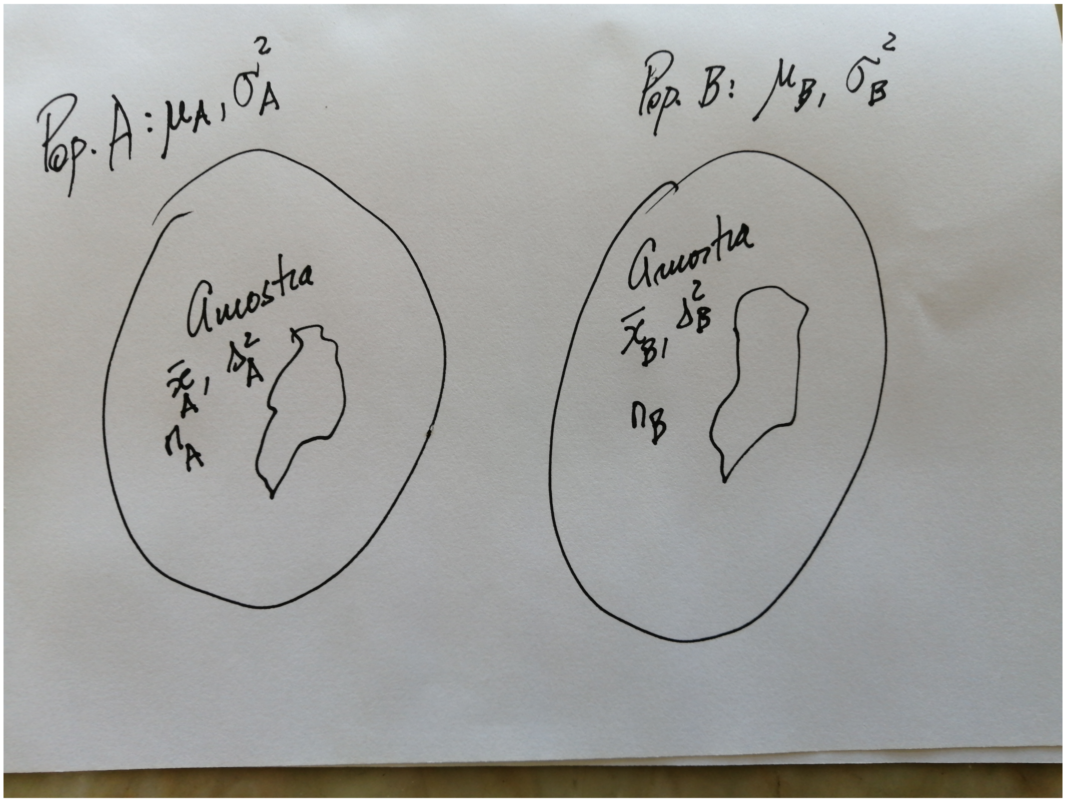

Figura 1: Amostras independentes.

e a região crítica para testar a $H_a : p > p_0$ é

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} > z_\alpha \,.$$

2 Comparação de duas médias: Amostras independentes

Considere duas populações A e B, cada uma delas com médias μ_A e μ_B e desvios padrões σ_A e σ_B respectivamente (veja a Figura 1). O problema é fazer inferência (estimar pontualmente, construir um IC ou testar se é maior, menor ou diferente do que um valor dado) sobre a diferença das médias, digamos $\mu_B - \mu_A$.

Como as médias populacionais μ_A e μ_B são desconhecidas, sabemos da Unidade 0 que elas podem ser estimadas a partir de amostras $X_{A,1}, \ldots, X_{A,n_A}$ da população A e $X_{B,1}, \ldots, X_{B,n_B}$ da população B. Usualmente, é fácil fazer com que o mecanismo amostral garanta que os valores observados na primeira amostra não influenciem os da segunda nem vice-versa e, nesse caso, dizemos que as amostras são *independentes* (eis o título desta Seção). Outra observação importante é que algumas vezes os dois tamanhos amostrais são iguais mas, como será visto mais adiante, pode ser conveniente que eles sejam diferentes, de forma que usamos símbolos diferentes n_A e n_B.

Uma vez que temos as duas amostras, sabemos da Unidade 0 que $\bar{X}_A = n_A^{-1}\sum_{i=1}^{n_A} X_{A,i}$ é um estimador não-viesado para μ_A (i.e. que $\mathrm{E}(\bar{X}_A) = \mu_A$) e que a sua variância é dada por $\mathrm{Var}(\bar{X}_A) = \sigma_A^2/n_A$. De forma semelhante temos também que $\mathrm{E}(\bar{X}_B) = \mu_B$ e $\mathrm{Var}(\bar{X}_B) = \sigma_B^2/n_B$. Devido a linearidade do operador *valor esperado*, segue que

$$\mathrm{E}(\bar{X}_B - \bar{X}_A) = \mu_B - \mu_A\,.$$

Em outras palavras, se na notação usada na introdução definimos $\theta = \mu_B - \mu_A$, um bom candidato para estimá-lo é $\hat{\theta} = \bar{X}_B - \bar{X}_A$. Para avaliar o erro desse estimador precisamos calcular a sua variância, que é fácil lembrando que para quaisquer duas variáveis aleatórias X e Y com variâncias finitas tem-se que $\mathrm{Var}(aX+bY) = a^2\ \mathrm{Var}(X)+ b^2\ \mathrm{Var}(Y) + 2ab\ \mathrm{cov}(X,Y)$, de forma que

$$\mathrm{Var}(\bar{X}_B - \bar{X}_A) = \mathrm{Var}(\bar{X}_B) + \mathrm{Var}(\bar{X}_A) - 2\,\mathrm{cov}(\bar{X}_B, \bar{X}_A) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\,, \tag{2}$$

onde temos usado o fato que $\bar{X}_A$ e $\bar{X}_B$ são independentes e, portanto, a sua covariância é nula.

A partir desse resultado, o maior problema que vai diferenciar as diferentes situações é como estimar a variância (2).

2.1 Populações normalmente distribuidas e variâncias conhecidas

O caso em que σ_A^2 e σ_B^2 são conhecidas é muito raro na prática, mas o seu estudo vai ajudar a entender melhor o resto da Seção. Veja que nesse caso não existe necessidade de estimar a variância (2). Se for ainda possível assumir que a distribuição das duas populações é Normal, sabemos da Unidade 0 que tanto $\bar{X}_A$ quanto $\bar{X}_B$ seguem distribuições normais e, como combinações lineares de variáveis aleatórias normais também são normais, segue que

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} = \frac{\bar{X}_B - \bar{X}_A - (\mu_B - \mu_A)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim\ \mathrm{Normal}(0,1)\,. \tag{3}$$

Com base nesse resultado, um IC de nível $100(1-\alpha)\%$ para a diferença $\mu_B - \mu_A$ é

$$\bar{x}_B - \bar{x}_A \pm z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}\,,$$

onde $z_{\alpha/2}$ é obtido da tabela da distribuição Normal Padrão. De forma semelhante, os testes de hipóteses sobre a diferença $\mu_B - \mu_A$ são baseados no estatístico

$$z_{obs} = \frac{(\bar{x}_B - \bar{x}_A) - d_0}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

onde d_0 é o valor da diferença $\mu_B - \mu_A$ especificado pela H_0. Mais precisamente,

- Para testar $H_0: \mu_B - \mu_A \leq d_0$ contra a $H_a: \mu_B - \mu_A > d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $z_{obs} > z_\alpha$. O p-valor é $P[N(0,1) > z_{obs}]$;

- Para testar $H_0: \mu_B - \mu_A \geq d_0$ contra a $H_a: \mu_B - \mu_A < d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $z_{obs} < -z_\alpha$. O p-valor é $P[N(0,1) < z_{obs}]$ e

- Para testar $H_0: \mu_B - \mu_A = d_0$ contra a $H_a: \mu_B - \mu_A \neq d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $|z_{obs}| > z_{\alpha/2}$. O p-valor é $P[|N(0,1)| > |z_{obs}|]$.

Exemplo 1 (Continuação). A locadora abasteceu $n_A = 20$ carros escolhidos aleatoriamente da sua frota com o combustível A e outros $n_B = 20$ carros também escolhidos aleatoriamente com o combustível B. Todos os 40 carros realizaram depois um percurso predeterminado e ao final do percurso foi aferido o consumo para cada carro. Suponha que os carros grupo A tiveram consumo médio $\bar{x}_A = 13.4$km/l, enquanto a média do grupo B foi $\bar{x}_A = 14.3$. As refinarias que produzem os dois tipos de combustível informam que o desvio padrão do consumo para esse tipo de modelo é de 1.5km/l para o combustível A e 2.4km/l para o combustível B. Ao nível de significância de 5%, existe evidência que o combustível B é mais eficiente que o A para esse tipo de carros? Construa um IC com nível 95% para a diferença de consumo entre o combustível B e o A.

Para resolver, assumimos que a distribuição do consumo nos dois grupos é Normal. O desvio padrão (2) de $\bar{X}_B - \bar{X}_A$ é $\sigma_{\bar{X}_B - \bar{X}_A} = \sqrt{\frac{1.5^2}{20} + \frac{2.4^2}{20}} \doteq 0.633$. Logo o IC para $\mu_B - \mu_A$ é $\bar{x}_B - \bar{x}_A \pm z_{\alpha/2}\, \sigma_{\bar{X}_B - \bar{X}_A} \doteq 14.3 - 13.4 \pm (1.96)\, 0.633 = (-0.340; 2.140)$km/l. Para testar a hipótese alternativa $H_a: \mu_B > \mu_A$ contra a nula $H_0: \mu_B \leq \mu_A$, calculamos o estatístico do teste $z_{obs} = (\bar{x}_B - \bar{x}_A)/\sigma_{\bar{X}_B - \bar{X}_A} \doteq (14.3 - 13.4)/0.633 \doteq 1.422$. Como ele é menor que o valor crítico $z_\alpha = z_{0.05} = 1.645$, não rejeitamos a H_0, i.e. ao nível de 5% não existe evidência que o combustível B tem menor consumo (ou maior rendimento) do que o A. Para ter ideia da quantidade da evidência em favor da H_a, podemos calcular o p-valor do teste. Neste caso, como a $H_a: \mu_B - \mu_A > 0$ é unilateral à direita, o p-valor é a área à direita do $z_{obs} \doteq 1.422$ na curva Normal Padrão, isto é $P[N(0,1) > 1.422) \doteq 0.077$ ou 7.7%, que corresponde a evidência moderada contra H_0 (por exemplo, a H_0 sería rejeitada ao nível de significância de 10% mas, como vimos, não ao de 5%).

Para refletir: No caso do exemplo, quais seriam as duas populações da Figura 1? □

2.2 Quaisquer populações e tamanhos amostrais grandes

Quando **os dois** tamanhos amostrais n_A e n_B são grandes, semelhante ao que ocorre para uma única população, tanto a distribuição de $\bar{X}_A$ quanto a de $\bar{X}_B$ são aproximadamente Normais e, portanto, também a da diferença $\bar{X}_B - \bar{X}_A$. A variância 2 pode ser estimada por

$$\hat{\sigma}_{\bar{X}_B - \bar{X}_A} = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}},$$

onde $S_A^2 = (n_A-1)^{-1}\sum_{i=1}^{n_A}(X_{A,1}-\bar{X}_A)^2$ e $S_B^2 = (n_B-1)^{-1}\sum_{i=1}^{n_B}(X_{B,1}-\bar{X}_B)^2$ são as respectivas variâncias amostrais. Assim, temos nesse caso que

$$\frac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} = \frac{\bar{X}_B - \bar{X}_A - (\mu_B - \mu_A)}{\sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}} \underset{\text{aprox.}}{\sim} \text{Normal}(0,1)\,. \tag{4}$$

Com base nessa distribuição assintótica um IC aproximado de nível $100(1-\alpha)\%$ para $\mu_B - \mu_A$ será

$$\bar{x}_B - \bar{x}_A \pm z_{\alpha/2}\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}\,,$$

onde $z_{\alpha/2}$ é obtido da tabela da distribuição Normal Padrão. De forma semelhante, os testes de hipóteses são baseados no estatístico

$$z_{obs} = \frac{(\bar{x}_B - \bar{x}_A) - d_0}{\sqrt{\frac{s_A^2}{n_A}+\frac{S_B^2}{n_B}}}\,,$$

onde d_0 é o valor da diferença $\mu_B - \mu_A$ especificado pela H_0. Mais precisamente,

- Para testar $H_0: \mu_B - \mu_A \leq d_0$ contra a $H_a: \mu_B - \mu_A > d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $z_{obs} > z_\alpha$. O p-valor é $P[N(0,1) > z_{obs}]$;
- Para testar $H_0: \mu_B - \mu_A \geq d_0$ contra a $H_a: \mu_B - \mu_A < d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $z_{obs} < -z_\alpha$. O p-valor é $P[N(0,1) < z_{obs}]$ e
- Para testar $H_0: \mu_B - \mu_A = d_0$ contra a $H_a: \mu_B - \mu_A \neq d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $|z_{obs}| > z_{\alpha/2}$. O p-valor é $P[|N(0,1)| > |z_{obs}|]$.

Exemplo 1 (Continuação). Suponha que agora a locadora deseja comparar o consumo dos dois combustíveis num outro modelo/marca, para o qual as refinarias não fornecem informação sobre o desvio padrão do consumo (i.e. não conhecemos os valores de σ_A nem o de σ_B). Como agora existe maior incerteza, a locadora abasteceu $n_A = 40$ carros desse novo modelo, escolhidos aleatoriamente da sua frota, com o combustível A e outros $n_B = 40$ carros, também escolhidos aleatoriamente com o combustível B. Todos os 40 carros realizaram depois um percurso predeterminado e ao final do percurso foi aferido o consumo para cada carro. Os carros do grupo A tiveram consumo médio $\bar{x}_A = 9.4$km/l e desvio padrão $s_A = 1,8$, enquanto para os do grupo B obteve-se $\bar{x}_A = 10.5$ e $s_B = 2.6$. Construa um IC com nível 95% para a diferença de consumo entre o combustível B e o A e teste se o combustível B oferece menor consumo do que o A ao nível de significância de 5%

Veja que se aceitamos que $n_A = 40$ e $n_B = 40$ são suficientemente grandes para que, *via* o TCL, a aproximação 4 seja válida, não precisamos assumir mais nada para construir o IC aproximado. Esse IC para $\mu_B - \mu_A$ é $\bar{x}_B - \bar{x}_A \pm z_{\alpha/2}\sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}} \doteq 10.5 - 9.4 \pm (1.96)\,(0.5) \doteq (0.120; 2.080)$km/l.

Para testar a $H_a : \mu_B > \mu_A$, o estatístico do teste é $z_{obs} = \frac{10.5-9.4}{\sqrt{\frac{1.8^2}{40}+\frac{2.6^2}{40}}} \doteq 2.200$. Como ele é maior que o valor crítico $z_{0.05} = 1.645$, rejeitamos a H_0, i.e. ao nível de 5% existe evidência que o combustível B tem menor consumo do que o A. $P[N(0,1) > 2.200] \doteq 0.014$ ou 1.4%, que corresponde a evidência forte contra H_0. □

2.3 Variâncias desconhecidas e tamanho amostral pequeno

Quando um ou os dois tamanhos amostrais n_A e/ou n_B são pequenos e as variâncias σ_A^2 e σ_B^2 são desconhecidas não podemos usar nenhum dos dois métodos discutidos acima. Nesse caso é comum assumir que (i) as duas populações são normalmente distribuídas [i.e. $X_{A,1}, \ldots, X_{A,n_A} \overset{\text{iid}}{\sim} \text{Normal}(\mu_A,\sigma_A^2)$ e $X_{B,1}, \ldots, X_{B,n_B} \overset{\text{iid}}{\sim} \text{Normal}(\mu_B,\sigma_B^2)$] e que as duas variâncias σ_A^2 e σ_B^2 são iguais, digamos $\sigma_A^2 = \sigma_B^2 = \sigma^2$. É claro que esses dois supostos restringem a aplicabilidade da metodologia que vamos a discutir. Embora uma ou as duas hipóteses podem ser retiradas, a comparação de médias no caso de tamanhos amostrais pequenos sem assumir normalidade e/ou igualdade das variâncias requer ferramental estatístico mais sofisticado que fica fora do escopo da disciplina, embora mais na frente nesta unidade vamos estudar o problema realizar testes de hipóteses para verificar se duas variâncias são iguais.

Como foi discutido acima, um problema fundamental fundamental é como estimar a variância (2) de $\bar{X}_B - \bar{X}_A$. Veja que como $\sigma_A^2 = \sigma_B^2 = \sigma^2$, a equação (2) fica agora

$$\text{Var}(\bar{X}_B - \bar{X}_A) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \sigma^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right) , \tag{5}$$

de forma que estimar $\text{Var}(\bar{X}_B - \bar{X}_A)$ é equivalente a estimar σ^2. Nesse sentido, poderíamos pensar em muitos estimadores de σ^2. Por exemplo, como a população A tem agora variância σ^2, um possível estimador é $S_A^2 = (n_A - 1)^{-1} \sum_{i=1}^{n_A} (X_{A,1} - \bar{X}_A)^2$. Da mesma forma, $S_B^2 = (n_B - 1)^{-1} \sum_{i=1}^{n_B} (X_{B,1} - \bar{X}_B)^2$ também estima σ^2. Se tivéssemos que escolher entre um deles, escolheríamos possivelmente o da amostra com maior tamanho amostral, embora de certa forma estaríamos jogando fora a informação da outra amostra, pelo menos para estimar a variância comum σ^2. É possível mostrar então que o melhor estimador de σ^2 não é nem S_A^2 nem S_B^2, mas uma *média ponderada* deles de acordo ao número de graus de liberdade, isto é

$$S_c^2 = \frac{(n_A - 1)\, S_A^2 + (n_B - 1)\, S_B^2}{n_A + n_B - 2}) \tag{6}$$

(o subíndice c é usado aquí no sentido de "estimador **c**ombinado"). Uma vez que temos esse estimador podemos substituir σ^2 por S_c^2 na equação (5) para obter um estimador de $\text{Var}(\bar{X}_B - \bar{X}_A)$. Depois disso, o resultado fundamental para construir ICs e regiões críticas de testes de hipóteses é que, sob o suposto de normalidade,

$$\frac{\bar{X}_B - \bar{X}_A - (\mu_B - \mu_A)}{\hat{\sigma}_{\bar{X}_B - \bar{X}_A}} = \frac{\bar{X}_B - \bar{X}_A - (\mu_B - \mu_A)}{S_c \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim \; t_{n_A + n_B - 2} \tag{7}$$

Consequentemente, o IC com nível de confiança $100(1-\alpha)\%$ será

$$\bar{x}_B - \bar{x}_A \pm t_{\alpha/2}\, s_c \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}\,,$$

onde $t_{\alpha/2}$ é obtido da tabela da distribuição t de Student com $(n_A + n_B - 2)$ graus de liberdade. De forma semelhante, os testes de hipóteses são baseados no estatístico

$$t_{obs} = \frac{(\bar{x}_B - \bar{x}_A) - d_0}{s_c \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}\,,$$

onde d_0 é o valor da diferença $\mu_B - \mu_A$ especificado pela H_0. Mais precisamente,

- Para testar $H_0 : \mu_B - \mu_A \leq d_0$ contra a $H_a : \mu_B - \mu_A > d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $t_{obs} > t_\alpha$. O p-valor é $P[t_{n_A+n_B-2} > t_{obs}]$;
- Para testar $H_0 : \mu_B - \mu_A \geq d_0$ contra a $H_a : \mu_B - \mu_A < d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $t_{obs} < -t_\alpha$. O p-valor é $P[t_{n_A+n_B-2} < t_{obs}]$ e
- Para testar $H_0 : \mu_B - \mu_A = d_0$ contra a $H_a : \mu_B - \mu_A \neq d_0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $|t_{obs}| > t_{\alpha/2}$. O p-valor é $P[|t_{n_A+n_B-2}| > |t_{obs}|]$.

Exemplo 1 (Continuação). Suponha que a locadora deseja comparar os dois combustíveis num outro modelo de carro com menor disponibilidade de unidades. Assim, foram alocados $20 carros$ para cada combustível. Durante o experimento um dos carros alocados ao grupo A quebrou e foi portanto retirado do estudo, de forma que temos efetivamente que $n_A = 19$ e $n_B = 20$.

Suponha que os carros do grupo A tiveram consumo médio $\bar{x}_A = 13.4$km/l e desvio padrão $s_A = 1.5$, enquanto para os do grupo B obteve-se $\bar{x}_A = 14.3$ e $s_B = 2.4$. Construa um IC com nível 95% para a diferença de consumo entre o combustível B e o A.

Para resolver, assumimos normalidade e variâncias iguais e calculamos primeiro $s_c^2 = \frac{(19-1)\,(1.5)^2 + (20-1)\,(2.4)^2}{19+20-2} \doteq (2.013)^2$. Logo, o IC para $\mu_B - \mu_A$ é $14.3 - 13.4 \pm (2.026)\,(2.013)\sqrt{\frac{1}{19} + \frac{1}{20}} = (-0.407; 2.207)$km/l.

Veja que, embora os números são muito parecidos aos apresentados na Subseção 2.4, o IC aqui é mais comprido, o que é devido à maior incerteza devida ao desconhecimento das variâncias.

2.4 Uma nota sobre alocação amostral

Como foi notado no começo da Seção, é conveniente ter a flexibilidade de poder ter tamanhos amostrais diferentes nos dois grupos, i.e. $n_A \neq n_B$. Para fixar ideias, suponha como na Subseção 2.4 que as duas variâncias σ_A^2 e σ_B^2 são conhecidas. Como gostaríamos que a variância (2) seja pequena, de forma a obter, por exemplo, intervalos de confiança mais precisos, faria sentido alocar mais observações ao grupo onde a variância da população é maior. Está ideia pode ser formalizada como segue. Suponha que orçamento do experimento permite observar um total de n observações que podem ser alocadas a qualquer um dos dois grupos, de forma que temos uma restrição $n_A + n_B = n$. O objetivo é então minimizar a função $f(n_A, n_B) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$ sujeito a essa restrição. Falando estritamente, esse é um problema de programação inteira, pois obviamente as variáveis n_A e n_B podem tomar somente valores inteiros. Porém, podemos ter uma boa ideia da solução ótima procedendo como se n_A e n_B fossem variáveis reais. Nesse caso usando o método dos multiplicadores de Lagrange construímos a função $f(n_A, n_B, \lambda) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} + \lambda\,(n - n_A - n_B)$ e igualamos a zero as suas derivadas parciais. É bastante fácil verificar que a solução ótima deve satisfazer que $\frac{n_B}{n_A} = \frac{\sigma_B}{\sigma_A}$. Se acrescentamos a restrição $n_A + n_B = n$, devemos ter então que

$$n_A = \frac{n}{1 + \sigma_B/\sigma_A} \quad \text{e} \quad \mathrm{n_B} = \frac{\mathrm{n}}{1 + \sigma_\mathrm{A}/\sigma_\mathrm{B}} \tag{8}$$

(verifique!) Note que esse resultado confere com a nossa intuição inicial: quanto maior é a variância de um grupo dado, com a outra fixa, maior será o tamanho alocado a esse grupo.

Os valores n_A e n_B na (8) não são necessariamente inteiros, de forma que é necessário arredondar esse resultado. Também, como na prática σ_A^2 e σ_B^2 não são conhecidas, para fazer a alocação amostral usam-se muitas vezes valores conjeturados por exemplo a partir de experimentos semelhantes realizados anteriormente ou, quando for possível, obtidos a partir de amostras piloto colhidas previamente do experimento principal.

Exemplo 1 (Continuação). Na situação descrita na Subseção 2.4 ($\sigma_A = 1.5$ e $\sigma_B = 2.4$), quais seriam os tamanhos amostrais ótimos caso fosse possível observar somente 40 veículos?

Os tamanhos amostrais n_B e n_A devem estar na mesma relação que os desvios $\sigma_B/\sigma_A = 2.4/1.5 = 1.6$. Logo a equação (8) fornece $n_A = 40/(1 + 1.6) \doteq 15.38$ e $n_B = 40/(1 + 1/1.6) \doteq 24.615$, que arredondamos para $n_A = 15$ e $n_B = 25$. Verifique que, se tivessem sido usados esses valores de n_A e n_B, a variância (2) seria 0.617^2, ligeiramente menor que o valor 0.633^2 que obtivemos com $n_A = n_B = 20$. □

Terminamos está seção notando que uma formulação um pouco mais geral deste problema permite introduzir custos por observação diferentes para os dois grupos. Mais precisamente, suponha que o experimento tem um orçamento $\mathcal{O}$ fixo e custos por observação c_A e c_B para cada grupo, de forma que o problema é minimizar a variância (2) sujeito à restrição $c_A\,n_A + c_B\,n_B = \mathcal{O}$. A solução fica como exercício.

3 Comparação de duas médias: Amostras pareadas

No Exemplo 1 acima podemos nos perguntar o motivo da locadora escolher carros da mesma marca e modelo para realizar o experimento, ou determinar que cada carro realize o mesmo percurso predeterminado. A resposta tem a ver com as variabilidades medidas pelas variâncias σ_A^2 e $sigma_B^2$. Como tipicamente o rendimento de um tipo de combustível depende do tipo de veículo, se fossem usados veículos diferentes observaríamos a variabilidade inerente ao combustível mas aquela devida ao tipo de veículo e/ou do percurso. Observando as fórmulas que usamos para construir os IC, vemos que maior variabilidade implica IC menos precisos (i.e. maior comprimento). Assim, o interesse do experimentador é controlar por todas aquelas variáveis que não dizem respeito ao fator que é o objetivo do experimento, que no caso da locadora tem a ver com os dois tipos de combustível.

Voltando agora ao Exemplo 2, a fábrica de sapatos poderia realizar o experimento de forma que as crianças "testadas" usassem cada uma um par de tênis fabricados com o material A ou com o material B. No final do experimento o desgaste poderia ser medido como a diferença média (para os dois pés, direito e esquerdo) entre a espessura das solas antes e após o uso, em cujo caso usaríamos as ferramentas apresentadas na Seção 3. Porém, realizar o experimento como *amostras independentes* faria com que a variabilidade inerente a cada material seria acumulada com a variabilidade devida por exemplo aos padrões de atividade das crianças (por exemplo, algumas podem praticar mais esporte enquanto outras podem ser mais sedentárias). Por isso, na situação do Exemplo 2, é tipicamente mais eficiente realizar o experimento de forma que cada criança usa num pé um tênis fabricado com o material A e no outro um fabricado com o material B e, ao final do experimento medir para cada criança a diferença de desgaste D_i entre o tênis do material B e o do material A. Dessa forma, boa parte da variabilidade devida às crianças será eliminada ao se fazer a diferença.

Observe que embora o título desta Seção é "amostra**s** pareada**s**", existe neste caso **somente** uma amostra, a das crianças, de forma que uma vez que as diferenças d_i são calculadas, usamos os mesmos métodos discutidos para uma única amostra na Seção 1. Em consequência, neste caso temos um único tamanho amostral n, ao invés da situação discutida na Seção anterior, onde tinha n_A e n_B que podiam ou não ser iguais.

Se no caso do Exemplo denotarmos por $X_{A,i}$ e $X_{B,i}$ os desgastes dos respectivos tênis e $D_i = X_{B,i} - X_{A,i}$, é verdade por causa da linearidade do operador *valor esperado* que $\mathrm{E}(D_i) = \mathrm{E}(X_{B,i}) - \mathrm{E}(X_{A,i}) = \mu_B - \mu_A$ (digamos, usando a notação da Seção 3), mas tipicamente esperamos que a variabilidade das diferenças D_i seja muito menor que a de $X_{B,i}$ e que a de $X_{A,i}$, devido a que o padrão de atividade das crianças, por exemplo, fará que $X_{A,i}$ e $X_{B,i}$ sejam positivamente correlacionadas.

É importante apontar que certos experimentos são impossíveis ou muito difíceis de *parear*. Por exemplo, num estudo para comparar o efeito de dois tratamentos médicos, pode ser que o fator importante a ser controlado, além daquele inerente aos dois tratamentos ou drogas, seja a variabilidade genética dos pacientes. Uma forma de fazer esse controle seria parear os tratamentos para gêmeos idênticos. Porém, como esse tipo de gêmeos é relativamente raro, realizar um experimento pareado pode ser mais difícil do que realizá-lo em amostras independentes. Nesse caso, a opção é entre realizar um experimento pareado com n relativamente pequeno ou realizá-lo em amostras independentes com tamanhos n_A e n_B maiores, de forma que a comparação da eficiência das duas opções pode ser na prática mais complicada.

Exemplo 2 (Continuação) A fábrica de sapatos escolheu $n = 9$ crianças entre os filhos dos seus empregados e para cada um deles fabricou um par de tênis de forma que um dos pés tinha a sola do material A e o outro do material B (qual pé recebeu qual material foi decidido ao acaso). Após 3 meses, os nove pares de tênis foram recolhidos e foi medido o desgaste em cada uma das 18 solas. As medidas do desgaste são dadas na tabela abaixo:

	Criança										
	1	2	3	4	5	6	7	8	9	**Média**	**DP**
A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	10.33	2.40
B	14.0	9.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	10.86	2.42
B - A	0.8	1.6	0.3	-0.1	1.1	-0.2	0.3	0.5	0.5	0.53	0.57

Construa um intervalo com confiança 90% para $\mu_B - \mu_A$. Ao nível de significância de 10%, existe evidência que o material A é mais resistente do que o B? Calcule o p-valor.

Para resolver, usamos os resultados da Seção 1. Como o tamanho amostral é pequeno, precisaremos assumir que a distribuição das diferenças é Normal. Nesse caso, o IC é $\bar{d} \pm t_{n-1;\alpha/2} \frac{s_d}{\sqrt{n}} = 0.53 \pm (1.860)(0.53/3) = (0.177; 0.883)$, onde o valor $t_{8,0.95} = 1.860$ é obtido da tabela da Distribuição t de Student com 8gl (ou, numa sessão do **R**, digitando `> qt(0.95,8)`). Para testar a $H_a : \mu_d = \mu_B - \mu_A > 0$, o estatístico do teste é $t_{obs} = \sqrt{n}\,\bar{d}/s_d = 3(0.53)/(0.57) \doteq 2.789$. Como ele é maior do que $t_{8,0.90} = 1.397$, rejeitamos a H_0, i.e. podemos afirmar que esses dados fornecem evidência suficiente para concluir ao nível de 10% que o material A é mais resistente em média do que o B. Finalmente, o p-valor desse teste é $P(t_8 > t.obs) = P(t_8 > 2.789) = 0.012$ ou 1.2%. □

Para refletir: Seria possível realizar um experimento pareado no caso do Exemplo 1? Como?

4 Comparação de duas proporções

O interesse e fazer inferência sobre a diferença $p_B - p_A$ das proporções ou probabilidades de uma certa característica para duas populações A e B. Se obtemos amostras aleatórias independentes com reposição de tamanhos n_A e n_B respectivamente, o número de indivíduos com a característica em estudo em cada amostra, X_A e X_B, seguem distribuições Binomiais com parâmetros n_A e p_A e n_B e p_B respectivamente.

Das propriedades da Distribuição Binomial sabemos que $E(X_A) = n_A\, p_A$, de forma que $\hat{p}_A = \frac{X_A}{n_A}$ é um estimador não viesado para p_A e, analogamente, $\hat{p}_B = \frac{X_B}{n_B}$ é um estimador não viesado para p_B. Logo, para estimar a diferença $p_B - p_A$ usamos o estimador $\hat{p}_B - \hat{p}_A$. A variância desse estimador é

$$\mathrm{Var}(\hat{p}_B - \hat{p}_A) = \mathrm{Var}(\hat{p}_A) + \mathrm{Var}(\hat{p}_B) = \frac{p_A\,(1-p_A)}{n_A} + \frac{p_B\,(1-p_B)}{n_B}\,, \tag{9}$$

onde usamos o fato que $\hat{p_A}$ e $\hat{p}_B$ são independentes e portanto a sua covariância é nula. Finalmente, quando $n_A \to \infty$ e $n_B \to \infty$, pode-se mostrar que a distribuição limite do estimador é Normal ou, mais precisamente, que

$$\frac{\hat{p}_B - \hat{p}_A - (p_B - p_A)}{\sqrt{\frac{\hat{p}_A\,(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B\,(1-\hat{p}_B)}{n_B}}} \underset{\text{aprox.}}{\sim} \mathrm{Normal}(0,1)\,.$$

Com esse resultado, quando n_A e n_B são grandes, o IC aproximado com nível $100(1-\alpha)\%$ para $p_B - p_A$ é

$$\hat{p}_B - \hat{p}_A \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_A\,(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B\,(1-\hat{p}_B)}{n_B}}\,. \tag{10}$$

O problema para realizar testes de hipóteses para verificar se a diferença $p_B - p_A$ é maior (menor, diferente) que um certo valor especificado d_0 é mais complicado que no caso de diferença de médias. Abordamos aqui tão somente o caso que $d_0 = 0$, quando a hipótese nula específica que as duas proporções são iguais, digamos $p_A = p_B = p$. Veja que nesse caso a variância (9) fica

$$\text{Var}(\hat{p}_B - \hat{p}_A) = p\,(1-p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \tag{11}$$

e, para estimá-la, devemos estimar o valor comum p de p_A e de p_B. É claro que p poderia ser estimado tanto por $\hat{p}_A$ quanto por $\hat{p}_B$ mas, de forma semelhante à discussão sobre a estimação de σ^2 na Subseção 2.3, é possível obter um estimador mais preciso usando informação das duas amostras. A idéia principal é simples: quando $p_A = p_B = p$, podemos misturar as duas populações para ter uma única onde a proporção da característica em estudo. Portanto, as duas amostras de tamanho n_A e n_B podem ser pensadas como uma única amostra de tamanho $n = n_A + n_B$ onde foram observados $X = X_A + X_B$ indivíduos com a característica em estudo. Dessa forma, o estimador $\hat{p}_c$ de p é

$$\hat{p}_c = \frac{X}{n} = \frac{X_A + X_B}{n_A + n_B} = \frac{n_A}{n_A + n_B}\,\frac{X_A}{n_A} + \frac{n_B}{n_A + n_B}\,\frac{X_B}{n_B} = \frac{n_A}{n_A + n_B}\,\hat{p}_A + \frac{n_B}{n_A + n_B}\,\hat{p}_B\,.$$

que é uma média ponderada dos estimadores originais $\hat{p}_A$ e $\hat{p}_B$ com pesos proporcionais aos respectivos tamanhos amostrais. Substituindo $\hat{p}_c$ em (11) obtemos o estatístico do teste

$$z_{obs} = \frac{\hat{p}_B - \hat{p}_A}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Usando esse estatístico,

- Para testar $H_0 : p_B - p_A \leq 0$ contra a $H_a : p_B - p_A > 0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $z_{obs} > z_\alpha$. O p-valor é $P[N(0,1) > z_{obs}]$;
- Para testar $H_0 : p_B - p_A \geq 0$ contra a $H_a : p_B - p_A < 0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $z_{obs} < -z_\alpha$. O p-valor é $P[N(0,1) < z_{obs}]$ e
- Para testar $H_0 : p_B - p_A = 0$ contra a $H_a : p_B - p_A \neq 0$ ao nível de significância $100\alpha\%$, rejeite H_0 se $|z_{obs}| > z_{\alpha/2}$. O p-valor é $P[|N(0,1)| > |z_{obs}|]$;

Exemplo 3. Numa pesquisa eleitoral foram entrevistados 500 eleitores da cidade A e 400 da cidade B. Entre os eleitores da cidade A, 37% declararam a sua preferência pelo candidato NN, enquanto essa proporção foi de 42% na cidade B. Teste, ao nível de significância de 1%, se existe evidência no sentido que a preferência dos eleitores por NN é diferente nas duas cidades. Calcule o p-valor do teste e construa um IC de nível 99% para a diferença.

Os tamanhos amostrais são suficientemente grandes para poder usar os métodos descritos acima. Para testar $H_0 : p_B - p_A = 0$ contra a $H_a : p_B - p_A \neq 0$, calculamos primeiro $\hat{p}_c = [(0.37)(500)+(0.42)(400)]/(500+400) \doteq 0.392$. Logo, o estatístico do teste é $z.obs \doteq (0.42 - 0.37)/\sqrt{(0.392)(1 - 0.392)(500^{-1} + 400^{-1})} \doteq 1.527$. Como esse valor é (em valor absoluto) menor do que $z_{0.005} \doteq 2.576$, não rejeitamos a H_0 (i.e. com esses dados não existe evidência suficiente para concluir que $p_A \neq p_B$). O p-valor do teste é $P[|N(0,1)| > 1.527] \doteq 0.127$, sinalizando que existe no máximo evidência fraca no sentido que as preferências nas duas cidades são diferentes. Finalmente, calculamos primeiro a variância estimada a partir da equação 9, $\hat{\sigma}_{\hat{p}_B - \hat{p}_B} = \frac{(0.37)(1-0.37)}{500} + \frac{(0.42)(1-0.42)}{400} \doteq (0.033)^2$, de forma que o IC (10) é $0.42 - 0.37 \pm (2.576)(0.033) \doteq (-0.034; 0,134)$, i.e. com 99% de confiança podemos afirmar que a diferença entre a preferência por NN na cidade A e na cidade B está entre -3.4 e $+13.4\%$. □