Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-03-28

Sumário

Prefácio

Capítulo 1

Introdução

Capítulo 2

Referencial para Inferência

Capítulo 3

Estimação Baseada no Plano Amostral

Capítulo 4

Efeitos do Plano Amostral

4.1 Introdução

O cálculo de desvio padrão e o uso de testes de hipóteses desempenham papel fundamental em estudos analíticos. Além de estimativas pontuais, na inferência analítica é necessário transmitir a idéia de precisão associada a essas estimativas e construir intervalos de confiança associados. Valores de desvios padrões, ou alternativamente comprimentos de intervalos de confiança, permitem avaliar a precisão da estimação. O cálculo do desvio padrão também possibilita a construção de estatísticas para testar hipóteses relativas a parâmetros do modelo (tradição de modelagem) ou de parâmetros da população ão finita (tradição de amostragem). Testes de hipóteses são também usados na fase de seleção de modelos.

Os pacotes mais comuns de análise estatística incluem em suas saídas valores de estimativas pontuais e seus desvios padrões, além de valores-p relativos a hipóteses de interesse. Contudo, as fórmulas usadas nestes pacotes para o cálculo dos desvios padrões e obtenção de testes são, em geral, baseadas nas hipóteses de independência e de igualdade de distribuição (IID) das observações, ou equivalentemente, de amostragem aleatória simples com reposição (AASC). Tais hipóteses quase nunca valem para dados obtidos através de pesquisas por amostragem, como as que realizam o IBGE e outras agências produtoras de estatísticas.

Este capítulo trata de avaliar o impacto sobre desvios padrões, intervalos de confiança e níveis de significância de testes usuais quando há afastamentos das hipóteses IID mencionadas, devidos ao uso de planos amostrais complexos para obter os dados. Como veremos, o impacto pode ser muito grande em algumas situações, justificando os cuidados que devem ser tomados na análise de dados deste tipo. Neste capítulo, usaremos como referência básica (Skinner, 989a).

4.2 Efeito do Plano Amostral (EPA) de Kish

Para medir o efeito do plano amostral sobre a variância de um estimador, Kish(1965) propôs uma medida que denominou `Efeito do Plano Amostral` (**EPA**) (em inglês, design effect ou, abreviadamente, deff). O objetivo desta medida é comparar planos amostrais no estágio de planejamento da pesquisa. O **EPA** de Kish é uma razão entre variâncias (de aleatorização) de um estimador, calculadas para dois planos amostrais alternativos. Vamos considerar um estimador $\hat{\theta}$ e calcular a variância de sua distribuição induzida pelo plano amostral complexo (verdadeiro) $V_{VERD}\left(\hat{\theta}\right)$ e a variância da distribuição do estimador induzida pelo plano de amostragem aleatória simples $V_{AAS}\left(\hat{\theta}\right)$.

Definição 4.1 *O **Efeito do Plano Amostral** (EPA) **de Kish** para um estimador $\hat{\theta}$ é*

$$\mathbf{EPA}_{Kish}\left(\hat{\theta}\right) = \frac{V_{VERD}\left(\hat{\theta}\right)}{V_{AAS}\left(\hat{\theta}\right)}. \tag{4.1}$$

Para ilustrar o conceito do $\mathbf{EPA}_{Kish}\left(\hat{\theta}\right)$, vamos considerar um exemplo.

Exemplo 4.1 *Efeitos de plano amostral de Kish para estimadores de totais com amostragem conglomerada em dois estágios.*

(Nascimento Silva and Moura, 1990) estimaram o $\mathbf{EPA}_{Kish}$ para estimadores de totais de várias variáveis sócio-econômicas no nível das Regiões Metropolitanas (RMs) utilizando dados do questionário de amostra do Censo Demográfico de 1980. Essas medidas estimadas do efeito do plano amostral foram calculadas para três esquemas amostrais alternativos, todos considerando amostragem conglomerada de domicílios em dois estágios, tendo o setor censitário como unidade primária e o domicílio como unidade secundária de amostragem. Duas das alternativas consideraram seleção de setores com equiprobabilidade via amostragem aleatória simples sem reposição (AC2AAS) e fração amostral constante de domicílios no segundo estágio (uma usando o estimador simples ou π-ponderado do total, e outra usando o estimador de razão para o total calibrando no número total de domicílios da população), e uma terceira alternativa considerou a seleção de setores com probabilidades proporcionais ao tamanho (número de domicílios por setor), denominada AC2PPT, e a seleção de 15 domicílios em cada setor da amostra, e empregando o correspondente estimador π-ponderado. Os resultados referentes à Região Metropolitana do Rio de Janeiro para algumas variáveis são apresentados na Tabela 4.1 a título de ilustração. Note que a população alvo considera apenas moradores em domicílios particulares permanentes na Região Metropolitana do Rio de Janeiro.

Os valores apresentados na Tabela 4.1 para a RM do Rio de Janeiro são similares aos observados para as demais RMs, se consideradas as mesmas variáveis. Nota-se grande variação dos valores do EPA, cujos valores mínimo e máximo são de 1,28 e 111,27 respectivamente. Para algumas variáveis (1,2,4,5 e 9), o EPA varia consideravelmente entre as diferentes alternativas de plano amostral, enquanto para outras variáveis (3,6,7 e 8) as variações entre os planos amostrais é mínima.

Os valores elevados do EPA observados para algumas variáveis realçam a importância de considerar o plano amostral verdadeiro ao estimar variâncias e desvios padrões associados às estimativas pontuais. Isso ocorre porque estimativas ingênuas de variância baseadas na hipótese de AAS subestimam substancialmente as variâncias corretas.

Outra regularidade encontrada nesse valores é que o EPA para o plano amostral AC2AAS com estimador simples apresenta sempre os valores mais elevados, revelando que este esquema é menos eficiente que os competidores considerados. Em geral, o EPA é menor para o esquema AC2PPT, com valores próximos aos do esquema AC2AAS com estimador de razão.

Os valores dos EPAs calculados por (Nascimento Silva and Moura, 1990) podem ser usados para planejar pesquisas amostrais (ao menos nas regiões metropolitanas), pois permitem comparar e antecipar o impacto do uso de alguns esquemas amostrais alternativos sobre a precisão de estimadores de totais de várias variáveis relevantes. Permitem também calcular tamanhos amostrais para garantir determinado nível de precisão, sem emprego de fórmulas complicadas. Portanto, tais valores seriam úteis como informação de apoio ao planejamento de novas pesquisas por amostragem, antes que as respectivas amostras sejam efetivamente selecionadas.

Entretanto, esses valores têm pouca utilidade em termos de usos analíticos dos dados da amostra do Censo Demográfico 80. é que tais valores, embora tendo sido estimados com essa amostra, foram calculados para planos amostrais distintos do que foi efetivamente adotado para seleção da amostra do censo. A amostra de domicílios usada no censo é estratificada por setor censitário com seleção sistemática de uma fração fixa (25% no Censo 80) dos domicílios de cada setor. Já os planos amostrais considerados na tabulação dos EPAs eram planos amostrais em dois estágios, com seleção de setores no primeiro estágio, os quais foram considerados

Tabela 4.1: Efeitos de plano amostral de Kish para variáveis selecionadas - Região Metropolitana do Rio de Janeiro

Plano amostral →	AC2AAS		AC2PPT
Variável ↓	Estimador Simples	Estimador de Razão	Estimador π-ponderado
1) Número total de moradores	10,74	2,00	1,90
2) Número de moradores ocupados	5,78	1,33	1,28
3) Rendimento monetário mensal I	5,22	4,92	4,49
4) Número total de filhos nascidos vivos de mulheres com 15 anos ou mais	4,59	2,02	1,89
5) Número de domicílios que têm fogão	111,27	1,58	1,55
6) Número de domicílios que têm telefone	7,11	7,13	6,41
7) Valor do aluguel ou prestação mensal	7,22	7,02	6,45
8) Número de domicílios que têm automóvel e renda $<$ 5SM	1,80	1,67	1,55
9) Número de domicílios que têm geladeira e renda $\geq$ 5SM	46,58	2,26	2,08

por sua similaridade com os esquemas adotados nas principais pesquisas domiciliares do IBGE tais como a PNAD e a PME (Pesquisa Mensal de Emprego). Portanto, a utilidade maior dos valores tabulados dos EPAs seria a comparação de planos amostrais alternativos para planejamento de pesquisas futuras, e não a análise dos resultados da amostra do censo 80.

4.3 Efeito do Plano Amostral Ampliado

O que se observou no Exemplo 4.1 com respeito à dificuldade de uso dos EPAs de Kish calculados para fins analíticos também se aplica para outras situações e é uma deficiência estrutural do conceito de EPA proposto por Kish. Para tentar contornar essa dificuldade, é necessário considerar um conceito ampliado de EPA, correspondente ao conceito de misspecification effect `meff` proposto por p. 24, (Skinner et al., 1989), que apresentamos e discutimos nesta seção.

Para introduzir este conceito ampliado de EPA, que tem utilidade também para fins de inferência analítica, vamos agora considerar um modelo subjacente às observações usadas para o cálculo do estimador pontual $\hat{\theta}$. Designemos por $v_0 = \widehat{V}_{IID}\left(\hat{\theta}\right)$ um estimador usual (consistente) da variância de $\hat{\theta}$ calculado sob a hipótese (ingênua) de que as observações são IID. A inadequação da hipótese de IID poderia ser consequência ou de estrutura da população ou de efeito de plano amostral complexo. Em qualquer dos casos, a estimativa v_0 da variância de $\hat{\theta}$ calculada sob a hipótese de observações IID se afastaria da variância de $\hat{\theta}$ sob o plano amostral (ou modelo) verdadeiro, denotada $V_{VERD}\left(\hat{\theta}\right)$. Note que $V_{VERD}\left(\hat{\theta}\right) = V_M\left(\hat{\theta}\right)$ na abordagem baseada em modelos e $V_{VERD}\left(\hat{\theta}\right) = V_p\left(\hat{\theta}\right)$ na abordagem de aleatorização.

Para avaliar se este afastamento tende a ser grande ou pequeno, vamos considerar a distribuição de v_0 com relação à distribuição de aleatorização verdadeira (ou do modelo verdadeiro) e localizar $V_{VERD}\left(\hat{\theta}\right)$ com relação a esta distribuição de referência. Como em geral seria complicado obter esta distribuição, vamos tomar uma medida de centro ou locação da mesma e compará-la a $V_{VERD}\left(\hat{\theta}\right)$.

Podemos desta forma introduzir uma medida de efeito da especificação incorreta do plano amostral (ou do modelo) sobre a estimativa v_0 da variância do estimador $\hat{\theta}$.

Definição 4.2 *O efeito da especificação incorreta do plano amostral (ou do modelo) sobre a estimativa v_0 da variância do estimador $\hat{\theta}$ é*

$$\mathbf{EPA}\left(\hat{\theta}, v_0\right) = \frac{V_{VERD}\left(\hat{\theta}\right)}{E_{VERD}\left(v_0\right)}\,. \tag{4.2}$$

Desta forma, o $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$ mede a tendência de v_0 a subestimar ou superestimar $V_{VERD}\left(\hat{\theta}\right)$, variância `verdadeira` de $\hat{\theta}$. Quanto mais afastado de 1 for o valor de $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$, mais incorreta será considerada a especificação do plano amostral ou do modelo.

Enquanto a medida proposta por Kish baseia-se nas distribuições induzidas pela aleatorização dos planos amostrais comparados, o $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$ pode ser calculado com respeito a distribuições de aleatorização ou do modelo envolvido, bastando calcular V_{VERD} e E_{VERD} da Definição (4.2) com relação à distribuição correspondente.

Em geral, são esperadas as seguintes consequências sobre o **EPA** ao ignorar o plano amostral efetivamente adotado e admitir que a seleção da amostra foi AAS:

1. Ignorar os pesos em v_0 pode inflacionar o **EPA**;
2. Ignorar conglomeração em v_0 pode inflacionar o **EPA**;

Tabela 4.2: Definição da estratificação da população de empresas

Estrato	Condição	Tamanho
1	empresas com PO > 21	161 empresas
2	empresas com PO ≤ 21	588 empresas

3. Ignorar estratificação em v_0 pode reduzir o **EPA**.

Combinações destes aspectos num mesmo plano amostral, resultando na especificação incorreta do plano amostral subjacente a v_0, podem inflacionar ou reduzir o **EPA**. Nesses casos é difícil prever o impacto de ignorar o plano amostral (ou modelo) verdadeiro sobre a análise baseada em hipóteses IID. Por essa razão, é recomendável ao menos estimar os EPAs antes de concluir a análise padrão, para poder então avaliar se há impactos importantes a considerar.

Exemplo 4.2 *Efeitos de plano amostral para estimação de médias na amostragem estratificada simples com alocação desproporcional*

Neste exemplo consideramos uma população de $N = 749$ empresas, para as quais foram observadas as seguintes variáveis:

1. pessoal ocupado em 31/12/94 (PO);
2. total de salários pagos no ano de 94 (SAL);
3. receita total no ano de 94 (REC).

A idéia é considerar o problema de estimar as médias populacionais das variáveis SAL e REC (variáveis de pesquisa, nesse exemplo), usando amostras estratificadas simples com alocação desproporcional, implicando em unidades amostrais com pesos desiguais numa situação bastante simples. A variável PO é a variável de estratificação. As médias populacionais das variáveis de pesquisa (SAL e REC) são conhecidas, porém supostas desconhecidas para efeitos do presente exercício, em que se supõe que amostragem seria usada para sua estimação.

Para estimar estas médias, as empresas da população foram divididas em dois estratos, definidos a partir da variável PO, conforme indicado na Tabela 4.2.

Foram então selecionadas de cada um dos estratos amostras aleatórias simples sem reposição de 30 empresas, implicando em uso de alocação igual e em frações amostrais desiguais, em vista dos diferentes tamanhos populacionais dos estratos. Como o estrato 1 contém cerca de 21% das observações da população, a proporção de 50% das observações da amostra no estrato 1 (das maiores empresas) na amostra é bem maior do que seria esperado sob amostragem aleatória simples da população em geral. Desta forma, a média amostral de uma variável de pesquisa y qualquer (SAL ou REC) dada por

$$\bar{y} = \frac{1}{n} \sum_{h=1}^{2} \sum_{i \in s_h} y_{hi}$$

tenderia a superestimar a média populacional $\overline{Y}$ dada por $\overline{Y} = \frac{1}{N} \sum\limits_{h=1}^{2} \sum_{i \in U_h} y_{hi}$, onde y_{hi} é o valor da variável de pesquisa y para a $i-$ésima observação do estrato h, $(h = 1, 2)$. Neste caso, um estimador não-viciado da média populacional $\bar{Y}$ seria dado por

$$\bar{y}_w = \sum_{h=1}^{2} W_h \bar{y}_h$$

Tabela 4.3: Propriedades dos estimadores da média das variáveis de pesquisa

Quantidade de interesse	Salários	Receitas
1) Média populacional $\overline{Y}$	78,328	2,107
2) Média de $\bar{y}$ sobre 500 amostras	160,750	4,191
3) Média de $\bar{y}_w$ sobre 500 amostras	76,700	2,054

onde $W_h = \frac{N_h}{N}$ é a proporção de observações da população no estrato h e $\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_{hi}$ é a média amostral dos $y^{'}s$ no estrato h, $(h = 1, 2)$.

Com a finalidade de ilustrar o cálculo do **EPA**, vamos considerar o estimador não-viciado $\bar{y}_w$ e calcular sua variância sob o plano amostral realmente utilizado (amostra estratificada simples - AES com alocação igual). Essa variância poderá então ser comparada com o valor esperado (sob a distribuição induzida pelo plano amostral estratificado) do estimador da variância obtido sob a hipótese de amostragem aleatória simples.

No presente exercício, a variância do estimador $\bar{y}_w$ pode ser obtida de duas formas: calculando a expressão da variância utilizando os dados de todas as unidades da população (que são conhecidos, mas admitidos desconhecidos para fins do exercício de estimação de médias via amostragem) e por simulação.

A variância de $\bar{y}_w$ sob a distribuição de aleatorização verdadeira é dada por

$$V_p\left(\bar{y}_w\right) = \sum_{h=1}^{2} W_h^2 \left(1 - f_h\right) \frac{S_h^2}{n_h} \tag{4.3}$$

onde $f_h = n_h/N_h$, n_h é o número de observações na amostra no estrato h, e $S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} \left(y_{hi} - \overline{Y}_h\right)^2$ é a variância populacional da variável de pesquisa y dentro do estrato h, com $\overline{Y}_h = \frac{1}{N_h} \sum_{i \in U_h} y_{hi}$ representando a média populacional da variável y dentro do estrato h.

Um estimador usual da variância de $\bar{y}_w$ sob amostragem aleatória simples é $v_0 = (1 - f)\frac{s^2}{n}$ onde $s^2 = \frac{1}{n-1} \sum_{h=1}^{2} \sum_{i \in s_h} \left(y_{hi} - \bar{y}\right)^2$ e $f = \sum_{h=1}^{2} n_h / \sum_{h=1}^{2} N_h = n/N$.

O cálculo do **EPA** foi feito também por meio de simulação. Geramos 500 amostras de tamanho 60, segundo o plano amostral estratificado considerado. Para cada uma das 500 amostras e cada uma das duas variáveis de pesquisa (SAL e REC) foram calculados:

1. média amostral $(\bar{y})$;
2. estimativa ponderada da média $(\bar{y}_w)$;
3. estimativa da variância da estimativa ponderada da média $(\bar{y}_w)$ considerando observações IID (v_0);
4. estimativa da variância da estimativa ponderada da média $(\bar{y}_w)$ considerando o plano amostral verdadeiro $\left(\hat{V}_{AES}\ \left(\bar{y}_w\right)\right)$.

Note que na apresentação dos resultados os valores dos salários foram expressos em milhares de Reais (R$ 1.000,00) e os valores das receitas em milhões de Reais (R$ 1.000.000,00). Como a população é conhecida, os parâmetros populacionais de interesse podem ser calculados, obtendo-se os valores na primeira linha da Tabela 4.3.

Tabela 4.4: Propriedades dos estimadores de variância do estimador $\bar{y}_w$

Quantidade de interesse	Salários	Receitas
1) Variância populacional $V_{AES}(\bar{y}_w)$	$244,18$	$0,43500$
2) Média de $\widehat{V}_{AES}(\bar{y}_w)$ usando 500 amostras	$231,84$	$0,32569$
3) Valor esperado de v_0 usando população	$1.613,3$	$1,1880$
4)Média de v_0 usando 500 amostras	$1.636,1$	$1,2121$

Em contraste com os valores dos parâmetros populacionais, calculamos a média das médias amostrais não ponderadas ($\bar{y}$) dos salários e das receitas obtidas nas 500 amostras simuladas, obtendo os valores na segunda linha da Tabela 4.3. Como previsto, observamos um vício para cima na estimativa destas médias, da ordem de 105% para os salários e de $98,9\%$ para as receitas.

Usamos também o estimador $\bar{y}_w$ para estimar a média dos salários e das receitas na população, obtendo para esse estimador as médias apresentadas na terceira linha da Tabela 4.3. Observamos ainda um pequeno vício da ordem de $-1,95\%$ e $-2,51\%$ para os salários e receitas, respectivamente. Note que o estimador $\bar{y}_w$ é não-viciado sob o plano amostral adotado, entretanto o pequeno vício observado na simulação não pode ser ignorado pois é significantemente diferente de 0 ao nível de significância de 5%, apesar do tamanho razoável da simulação (500 replicações).

Além dos estimadores pontuais, o interesse maior da simulação foi comparar valores de estimadores de variância, e consequentemente de medidas do efeito do plano amostral. Como o estimador pontual dado pela média amostral não ponderada ($\bar{y}$) é grosseiramente viciado, não consideramos estimativas de variância para esse estimador, mas tão somente para o estimador não-viciado dado pela média ponderada $\bar{y}_w$. Para esse último, consideramos dois estimadores de variância, a saber o estimador ingênuo sob a hipótese de AAS (dado por v_0) e um estimador não viciado da variância sob o plano amostral $\hat{V}_{AES}(\bar{y}_w)$, que foi obtido substituindo as variâncias dentro dos estratos S_h^2 por estimativas amostrais não viciadas dadas por $s_h^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\overline{y}_h)^2$, $h=1,2$, na fórmula de $V_{AES}(\bar{y}_w)$ conforme definida em (4.3).

Como neste exercício a população é conhecida, podemos calcular $V_{AES}(\bar{y}_w)$ através das variâncias de y dentro dos estratos $h=1,2$ ou através da simulação. Esses valores são apresentados respectivamente na primeira e segunda linhas da Tabela 4.4, para as duas variáveis de pesquisa consideradas.

Os valores de $E_{VERD}\left(v_0\left(\overline{SAL}_w\right)\right)$ e de $E_{VERD}\left(v_0\left(\overline{REC}_w\right)\right)$ foram também calculados a partir das variâncias dentro e entre estratos na população, resultando nos valores na linha 3 da Tabela @ref{tab44}, e estimativas desses valores baseadas nas 500 amostras da simulação são apresentadas na linha 4 da Tabela @ref{tab44}. Os valores para o **EPA** foram calculados tanto com base nas estimativas de simulação como nos valores populacionais das variâncias, cujos cálculos estão ilustrados a seguir:

$$\begin{aligned}
\textbf{EPA}\left(\overline{SAL}_w, v_0\left(\overline{SAL}_w\right)\right) &= \frac{231,84}{1.636,1} = 0,142 \\
\textbf{EPA}\left(\overline{REC}_w, v_0\left(\overline{REC}_w\right)\right) &= \frac{0,32569}{1,2121} = 0,269
\end{aligned}$$

Tabela 4.5: Valores dos Efeitos de Plano Amostral (EPA) para as médias de Salário e de Receita

Variável	Estimativa	Simulação	População
Salário	Variância	$231,84$	$244,18$
	EPA	$0,142$	$0,151$
Receita	Variância	$0,32569$	$0,43500$
	EPA	$0,269$	$0,366$

$$\begin{aligned} \mathbf{EPA}\left(\overline{SAL}_w, v_0\left(\overline{SAL}_w\right)\right) &= \frac{244,18}{1.613,3} = 0,151 \text{ e} \\ \mathbf{EPA}\left(\overline{REC}_w, v_0\left(\overline{REC}_w\right)\right) &= \frac{0,43500}{1,1880} = 0,366. \end{aligned}$$

A Tabela 4.5 resume os principais resultados deste exercício, para o estimador ponderado da média $\bar{y}_w$. Apesar das diferenças entre os resultados da simulação e suas contrapartidas calculadas considerando conhecidos os valores da população, as conclusões da análise são similares:

1. ignorar os pesos na estimação da média provoca vícios substanciais, que não podem ser ignorados; portanto, o uso do estimador simples de média ($\bar{y}$) é desaconselhado;

2. ignorar os pesos na estimação da variância do estimador ponderado $\bar{y}_w$ também provoca vícios substanciais, neste caso, superestimando a variância por ignorar o efeito de estratificação; os efeitos de plano amostral são substancialmente menores que 1 para as duas variáveis de pesquisa consideradas (salários e receita); portanto o uso do estimador ingênuo de variância v_0 é desaconselhado.

Essas conclusões são largamente aceitas pelos amostristas e produtores de dados baseados em pesquisas amostrais para o caso da estimação de médias e totais, e respectivas variâncias. Entretanto ainda há exemplos de usos indevidos de dados amostrais nos quais os pesos são ignorados, em particular para a estimação de variâncias associadas a estimativas pontuais de médias e totais. Tal situação se deve ao uso ingênuo de pacotes estatísticos padrões desenvolvidos para analisar amostras IID, sem a devida consideração dos pesos e plano amostral.

Observação 4.1 *Neste exemplo não foi feito uso analítico dos dados e sim descritivo, onde é usual incorporar os pesos no cálculo de estimativas e variâncias. Não seria esperado usar um estimador ponderado para a média e não considerar os pesos no cálculo de variâncias, como fizemos neste exemplo.*

Observação 4.2 *O exemplo mostra que ignorar a estratificação ao calcular v_0 diminui o* **EPA**.

Um outro exemplo relevante é utilizado a seguir para ilustrar o fato de que o conceito do **EPA** adotado aqui é mais abrangente do que o definido por Kish, em particular porque a origem do efeito pode estar na estrutura da população e não no plano amostral usado para obter os dados.

Exemplo 4.3 *População conglomerada com conglomerados de tamanho 2 (Skinner et al., 1989), p. 25*

Considere uma população de conglomerados de tamanho 2, isto é, onde as unidades (elementares ou de referência) estão grupadas em pares (exemplos de tais populações incluem pares de irmãos gêmeos, casais, jogadores numa dupla de vôlei de praia ou tênis, etc.). Suponha que os valores de uma variável de pesquisa medida nessas unidades têm média θ e variância σ^2, além de uma correlação ρ entre os valores dentro de cada par (correlação intraclasse, veja (Nascimento Silva and Moura, 1990), cap. 2 e (Haggard, 1958). Suponha que um único par é sorteado ao acaso da população e que os valores y_1 e y_2 são observados para as duas

unidades do par selecionado. O modelo assumido pode então ser representado como

$$\begin{cases} E_M(Y_i) = \theta \\ V_M(Y_i) = \sigma^2 \\ CORR_M(Y_1; Y_2) = \rho \end{cases} \qquad i = 1, 2.$$

Um estimador não viciado para θ é dado por $\widehat{\theta} = (y_1 + y_2)/2$, a média amostral. Assumindo a (falsa) hipótese de que o esquema amostral é AASC de unidades individuais e não de pares, ou equivalentemente, que y_1 e y_2 são observações de variáveis aleatórias IID, a variância de $\widehat{\theta}$ é dada por

$$V_{AAS}\left(\widehat{\theta}\right) = \sigma^2/2$$

com um estimador não viciado dado por

$$v_0\left(\widehat{\theta}\right) = (y_1 - y_2)^2/4 \, .$$

Entretanto, na realidade a variância de $\widehat{\theta}$ é dada por

$$V_{VERD}\left(\widehat{\theta}\right) = V_M\left(\widehat{\theta}\right) = \sigma^2(1+\rho)/2$$

e o valor esperado do estimador de variância $v_0\left(\widehat{\theta}\right)$ é dado por

$$E_{VERD}\left[v_0\left(\widehat{\theta}\right)\right] = \sigma^2(1-\rho)/2 \, .$$

Consequentemente, considerando as equações (4.1) e (4.2), tem-se que

$$\mathbf{EPA}_{Kish}\left(\hat{\theta}\right) = 1 + \rho$$

e

$$\mathbf{EPA}\left(\hat{\theta}, v_0\right) = (1+\rho)/(1-\rho) \, .$$

A Figura 4.1 plota os valores de $\mathbf{EPA}_{Kish}\left(\hat{\theta}\right)$ e $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$ para valores de ρ entre 0 e $0,8$. Como se pode notar, o efeito da especificação inadequada do plano amostral ou da estrutura populacional pode ser severo, com valores de $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$ chegando a 9. Um aspecto importante a notar é que o $\mathbf{EPA}_{Kish}\left(\hat{\theta}\right)$ tem variação muito mais modesta que o $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$.

Este exemplo ilustra bem dois aspectos distintos do uso de medidas como o efeito de plano amostral. O primeiro é que as duas medidas são distintas, `embora os respectivos estimadores baseados numa particular amostra coincidam`. No caso particular deste exemplo, o $\mathbf{EPA}_{Kish}\left(\hat{\theta}\right)$ cresce pouco com o valor do coeficiente de correlação intraclasse ρ, o que implica que um plano amostral conglomerado como o adotado (seleção ao acaso de um par da população) seria menos eficiente que um plano amostral aleatório simples (seleção de duas unidades ao acaso da população), mas a perda de eficiência seria modesta. Já se o interesse é medir, a posteriori, o efeito da má especificação do plano amostral no estimador de variância, o impacto, medido pelo $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$, seria muito maior.

Vale ainda notar que o $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$ mede o impacto da má especificação do plano amostral ou do modelo para a estrutura populacional. Neste exemplo, ignorar a estrutura da população (o fato de que as observações são pareadas) poderia provocar subestimação da variância do estimador de média, que seria tanto maior quanto maior fosse o coeficiente de correlação intraclasse ρ. Efeitos como esse são comuns também devido ao planejamento amostral, mesmo em populações onde a conglomeração é imposta artificialmente pelo amostrista.

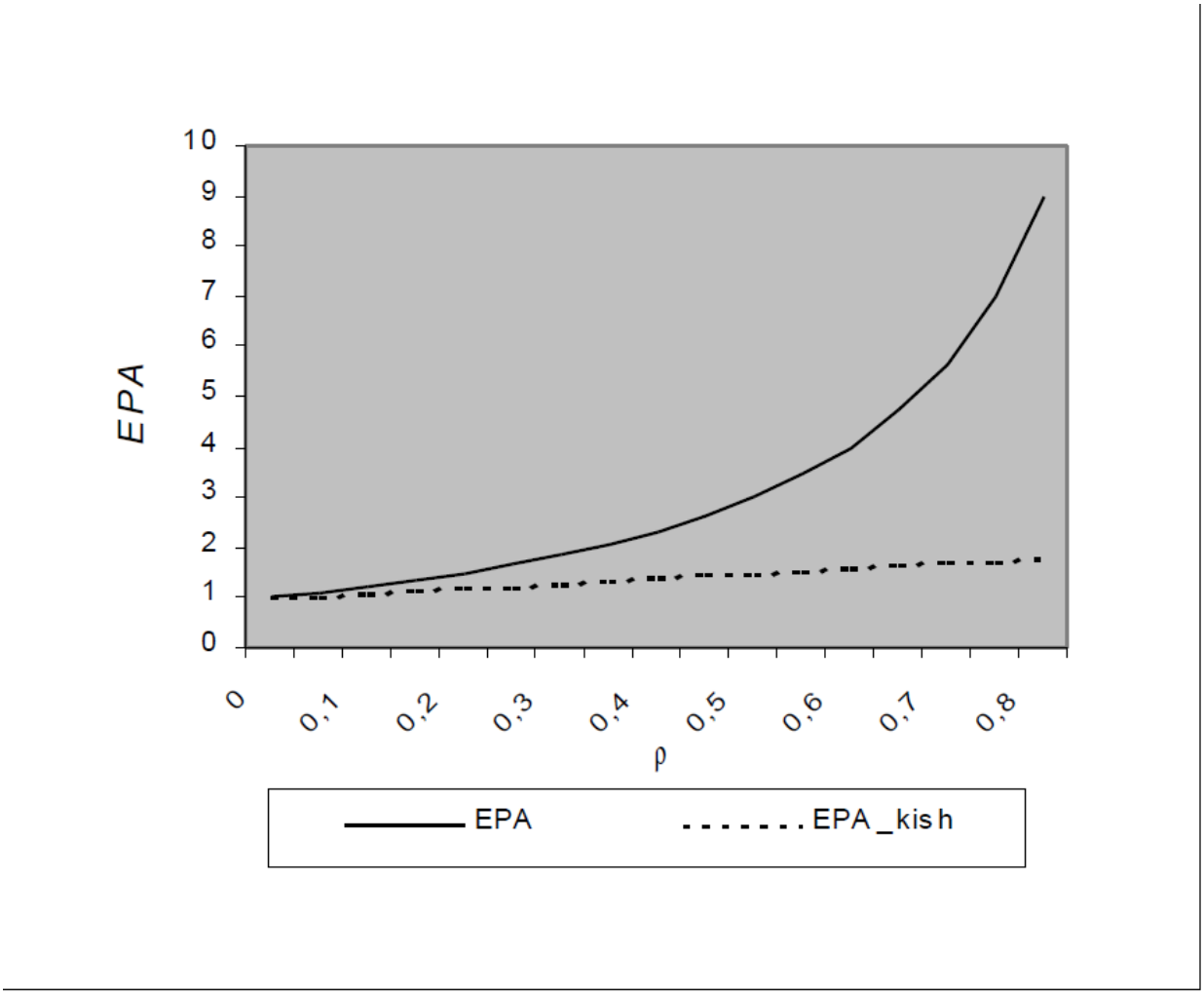

Figura 4.1: Valores de EPA e EPA de Kish para conglomeração

4.4 Intervalos de Confiança e Testes de Hipóteses

A partir da estimativa pontual $\hat{\theta}$ de um parâmetro θ (da população finita ou do modelo de superpopulação) é possível construir um intervalo de confiança de nível de confiança aproximado $(1-\alpha)$ a partir da distribuição assintótica de

$$t_0 = \frac{\hat{\theta} - \theta}{v_0^{1/2}}$$

que, sob a hipótese de que as observações são IID, frequentemente é $N(0;1)$.

Neste caso, um intervalo de confiança de nível de confiança aproximado $(1-\alpha)$ é dado por $\left[\hat{\theta} - z_{\alpha/2} v_0^{1/2}, \hat{\theta} + z_{\alpha/2} v_0^{1/2}\right]$, onde z_α é definido por $\int_{z_\alpha}^{+\infty} \varphi(t)\, dt = \alpha$, onde φ é a função de densidade da distribuição normal padrão.

Vamos analisar o efeito de um plano amostral complexo sobre o intervalo de confiança. No caso de um plano amostral complexo, a distribuição que é aproximadamente normal é a de

$$\frac{\hat{\theta} - \theta}{\left[\widehat{V}_{VERD}\left(\hat{\theta}\right)\right]^{1/2}}.$$

Por outro lado, para obter a variância da distribuição assintótica de t_0 note que

$$\frac{\hat{\theta} - \theta}{v_0^{1/2}} = \frac{\hat{\theta} - \theta}{\left[\widehat{V}_{VERD}\left(\hat{\theta}\right)\right]^{1/2}} \times \frac{\left[\widehat{V}_{VERD}\left(\hat{\theta}\right)\right]^{1/2}}{v_0^{1/2}} \ .$$

Como o primeiro fator tende para uma $N(0;1)$, a variância assintótica de t_0 é aproximadamente igual ao quadrado do segundo fator, isto é, a $\frac{\widehat{V}_{VERD}(\hat{\theta})}{v_0}$ que é um estimador para $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$. Porém quando a amostra é grande esse valor aproxima o $\mathbf{EPA}\left(\hat{\theta}, v_0\right) = \frac{V_{VERD}(\hat{\theta})}{E_{VERD}(v_0)}$, pois v_0 é aproximadamente igual a $E_{VERD}(v_0)$ e $\widehat{V}_{VERD}\left(\hat{\theta}\right)$ é aproximadamente igual a $V_{VERD}\left(\hat{\theta}\right)$. Logo temos que a distribuição assintótica verdadeira de t_0 é dada por

$$t_0 \sim N\left[0; \mathbf{EPA}\left(\hat{\theta}, v_0\right)\right] \ .$$

Dependendo do valor de $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$, o intervalo de confiança baseado na distribuição assintótica verdadeira de t_0 pode ser bem distinto daquele baseado na distribuição assintótica obtida sob a hipótese de observações IID. Em geral, a probabilidade de cobertura assintótica do intervalo $\left[\hat{\theta} - z_{\alpha/2} v_0^{1/2}, \hat{\theta} + z_{\alpha/2} v_0^{1/2}\right]$ será aproximadamente igual a

$$2\Phi\left(z_{\alpha/2} / \left[\mathbf{EPA}\left(\hat{\theta}, v_0\right)\right]^{1/2}\right) - 1 \ ,$$

onde Φ é a função de distribuição acumulada de uma $N(0;1)$. Calculamos esta probabilidade para alguns valores do **EPA**, que apresentamos na Tabela @ref{tab46}.

À medida que o valor do $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$ aumenta, a probabilidade real de cobertura diminui, sendo menor que o valor nominal para valores de $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$ maiores que 1.

Utilizando a correspondência existente entre intervalos de confiança e testes de hipóteses, podemos derivar os níveis de significância nominais e reais subtraindo de 1 os valores da Tabela 4.6. Por exemplo, para $\alpha = 0,05$ e $\mathbf{EPA}\left(\hat{\theta}, v_0\right) = 2$, o nível de significância real seria aproximadamente $1 - 0,83 = 0,17$.

Tabela 4.6: Probabilidades de cobertura para níveis nominais de 95% e 99%

EPA $\left(\hat{\theta}, v_0\right)$	$1-\alpha = 0,95$	$1-\alpha = 0,99$
$0,90$	$0,96$	$0,99$
$0,95$	$0,96$	$0,99$
$1,0$	$0,95$	$0,99$
$1,5$	$0,89$	$0,96$
$2,0$	$0,83$	$0,93$
$2,5$	$0,78$	$0,90$
$3,0$	$0,74$	$0,86$
$3,5$	$0,71$	$0,83$
$4,0$	$0,67$	$0,80$

Exemplo 4.4 *Teste de hipótese sobre proporção*

Vamos considerar um exemplo hipotético de teste de hipótese sobre uma proporção, semelhante ao de (Sudman, 1976), apresentado em p. 196, (Lehtonen and Pahkinen, 1995). Uma amostra de $m = 50$ conglomerados é extraída de uma grande população de empresas industriais (conglomerados). Suponhamos que cada empresa $i = 1, \ldots, 50$ da amostra tenha $n_i = 20$ empregados. O tamanho total da amostra de empregados (unidades elementares) é $n = \sum_i n_i = 1.000$. Queremos estudar o acesso dos trabalhadores das empresas a planos de saúde.

Usando-se conhecimento do ano anterior, foi estabelecida a hipótese de que a proporção de trabalhadores cobertos por planos de saúde é 80%, ou seja $H_0 : p = p_0 = 0,8$. Vamos adotar o nível de significância $\alpha = 5\%$.

A estimativa obtida na pesquisa foi $\widehat{p} = n_A/n = 0,84$, onde $n_A = 840$ é o número de trabalhadores na amostra com acesso a planos de saúde. Ignorando o plano amostral e a conglomeração das unidades elementares na população, podemos considerar um teste binomial e usar a aproximação normal $N(0;1)$ para a estatística de teste

$$Z = |\widehat{p} - p_0| / \sqrt{p_0 \left(1 - p_0\right) / n}, \tag{4.4}$$

onde o denominador é o desvio padrão da estimativa $\widehat{p}$ sob a hipótese nula.

Vamos calcular o valor da estatística Z, supondo que tenha sido usada amostragem aleatória simples com reposição (AASC) de empregados. Vamos também considerar uma abordagem baseada no plano amostral de conglomerados. O desvio padrão de $\widehat{p}$, no denominador de Z, será baseado na hipótese de distribuição binomial, com tamanhos amostrais diferentes para as duas abordagens.

Para o teste baseado na amostragem aleatória simples, ignoramos a conglomeração e usamos na fórmula do desvio padrão o tamanho total da amostra de unidades elementares (empregados), isto é, $n = 1.000$. O valor da estatística de teste Z definida em (4.4) é, portanto,

$$Z_{bin} = |0,84 - 0,8| / \sqrt{0,8 \left(1 - 0,8\right) / 1.000} = 3,162 > Z_{0,025} = 1,96 \tag{4.5}$$

onde $\sqrt{0,8 \left(1 - 0,8\right) / 1.000} = 0,0126$ é o desvio padrão de $\widehat{p}$ sob a hipótese nula. Este resultado sugere a rejeição da hipótese H_0.

Por outro lado, é razoável admitir que se uma empresa for coberta por plano de saúde, cada empregado dessa empresa terá acesso ao plano. Essa é uma informação importante que foi ignorada no teste anterior. De fato, selecionar mais de uma pessoa numa empresa não aumenta nosso conhecimento sobre a cobertura por plano de saúde no local. Portanto, o `tamanho efetivo` da amostra é $\overline{n} = 50$, em contraste com o valor

1.000 usado no teste anterior. O termo `tamanho efetivo` foi introduzido em (Kish, 1965) para designar o tamanho de uma amostra aleatória simples necessário para estimar p com a mesma precisão obtida por uma amostra conglomerada de tamanho n (neste caso, igual a 1.000) unidades elementares.

Usando o tamanho efetivo de amostra, temos a estatística de teste baseada no plano amostral verdadeiro

$$Z_p = |\widehat{p} - p_0| / \sqrt{p_0 \left(1 - p_0\right) / 50} = 0,707,$$

onde o valor $\sqrt{0,8 \left(1 - 0,8\right) / 50} = 0,0566$ é muito maior que o valor do desvio padrão obtido no teste anterior. Portanto, o valor observado de Z_p é menor que o de Z_{bin}, e o novo teste sugere a não rejeição da mesma hipótese nula.

Neste exemplo, portanto, se verifica que ignorar a conglomeração pode induzir a uma decisão incorreta de rejeitar a hipótese nula, quando a mesma não seria rejeitada se o plano amostral fosse corretamente incorporado na análise. Efeitos desse tipo são mais difíceis de antecipar para inferência analítica, particularmente quando os planos amostrais empregados envolvem combinação de estratificação, conglomeração e probabilidades desiguais de seleção. Por essa razão, a recomendação é procurar sempre considerar o plano amostral na análise, ao menos como forma de verificar se as conclusões obtidas por formas ingênuas de análise ignorando os pesos e plano amostral são as mesmas.

4.5 Efeitos Multivariados de Plano Amostral

O conceito de efeito de plano amostral introduzido em (4.2) é relativo a inferências sobre um parâmetro univariado θ. Consideremos agora o problema de estimação de um vetor θ de K parâmetros. Seja $\hat{\theta}$ um estimador de θ e seja $\mathbf{V}_0$ um estimador da matriz $K \times K$ de covariância de $\hat{\theta}$, baseado nas hipóteses de independência e igualdade de distribuição das observações (IID), ou equivalentemente, de amostragem aleatória simples com reposição (AASC). é possível generalizar a equação (4.2), definindo o `efeito multivariado do plano amostral de` $\hat{\theta}$ `e` $\mathbf{V}_0$ como

$$\mathbf{EMPA}(\hat{\theta}, \mathbf{V}_0) = \mathbf{\Delta} = \mathbf{E}_{VERD} \left(\mathbf{V}_0\right)^{-1} \mathbf{V}_{VERD}(\hat{\theta}), \tag{4.6}$$

onde $\mathbf{E}_{VERD} \left(\mathbf{V}_0\right)$ é o valor esperado de $\mathbf{V}_0$ e, $\mathbf{V}_{VERD}(\hat{\theta})$ é a matriz de covariância de $\hat{\theta}$, ambas calculadas com respeito `{a} distribuição de aleatorização induzida pelo plano amostral efetivamente utilizado, ou alternativamente sob o `modelo correto`.

Os autovalores $\delta_1 \geq \ldots \geq \delta_K$ da matriz $\mathbf{\Delta}$ são denominados `efeitos generalizados do plano amostral`. A partir deles, e utilizando resultados padrões de teoria das matrizes (p.64, (Johnson and Wichern, 1988)) é possível definir limitantes para os efeitos (univariados) do plano amostral para combinações lineares $\mathbf{c}^{'}\widehat{\theta}$ das componentes de $\widehat{\theta}$. Temos os seguintes resultados:

$$\begin{aligned} \delta_1 &= \max \mathbf{EPA}(\mathbf{c}^{'}\widehat{\theta}, \mathbf{c}^{'}\mathbf{V}_0\mathbf{c}), \\ \delta_K &= \min \mathbf{EPA}(\mathbf{c}^{'}\widehat{\theta}, \mathbf{c}^{'}\mathbf{V}_0\mathbf{c}). \end{aligned}$$

No caso particular onde $\mathbf{\Delta} = \mathbf{I}_{K \times K}$, temos $\delta_1 = \ldots = \delta_K = 1$ e os efeitos (univariados) do plano amostral das combinações lineares para componentes de $\hat{\theta}$ são todos iguais a 1. Para ilustrar esse conceito, vamos reconsiderar o Exemplo 4.2 de estimação de médias com amostragem estratificada desproporcional anteriormente apresentado, mas agora considerando a natureza multivariada do problema (há duas variáveis de pesquisa).

Exemplo 4.5 *Efeitos Multivariados do Plano Amostral para as médias de Salários e de Receitas*

Vamos considerar as variáveis Salário (em R$ 1.000) e Receita (em R$ 1.000.000) definidas na população de empresas do Exemplo 4.2 e calcular a matriz $\mathbf{EMPA}\left(\hat{\theta}, \mathbf{V}_0\right)$, onde $\hat{\theta} = \left(\overline{SAL}_w, \overline{REC}_w\right)'$. Neste exemplo, os dados populacionais são conhecidos, e portanto podemos calcular a covariância dos estimadores $\left(\overline{SAL}_w, \overline{REC}_w\right)$. Usando a mesma notação do Exemplo 4.2, temos que

$$COV_{AES}(\overline{SAL}_w, \overline{REC}_w) = \sum_{h=1}^{2} W_h^2 \frac{(1-f_h)}{n_h} S_{SAL,REC}^{(h)}$$

onde

$$S_{SAL,REC}^{(h)} = \frac{1}{N_h - 1} \sum_{i \in U_h} \left(SAL_{hi} - \overline{SAL}_h\right)\left(REC_{hi} - \overline{REC}_h\right) \ .$$

Substituindo os valores conhecidos na população das variáveis SAL_{hi} e REC_{hi}, obtemos para esta covariância o valor

$$COV_{AES}(\overline{SAL}_w, \overline{REC}_w) = 3,2358$$

e portanto a matriz de variância dos estimadores ponderados da média fica igual a

$$\mathbf{V}_{AES}(\overline{SAL}_w, \overline{REC}_w) = \begin{bmatrix} 244,18 & 3,2358 \\ 3,2358 & 0,4350 \end{bmatrix} \tag{4.7}$$

onde os valores das variâncias em (4.7) foram os calculados no Exemplo 4.2 e coincidem, respectivamente, com os valores usados nos numeradores de $\mathbf{EPA}\left(\overline{SAL}_w\right)$ e de $\mathbf{EPA}\left(\overline{REC}_w\right)$ lá apresentados. Para calcular o $\mathbf{EMPA}(\hat{\theta}, \mathbf{V}_0)$ é preciso agora obter $\mathbf{E}_{VERD}\left(\mathbf{V}_0\right)$.

Neste exemplo, a matriz de efeito do plano amostral $\mathbf{EMPA}(\hat{\theta}, \mathbf{V}_0) = \mathbf{\Delta}$ pode também ser calculada através de simulação, de modo análogo ao que foi feito no Exemplo 4.2. Para isto, foram utilizadas outras 500 amostras de tamanho 60 segundo o plano amostral descrito no Exemplo 4.2. Para cada uma das 500 amostras foram calculadas estimativas:

1. da variância da média amostral ponderada do salário e da receita assumindo observações IID;
2. da covariância entre médias ponderadas do salário e da receita assumindo observações IID;
3. da variância da média amostral ponderada do salário e da receita considerando o plano amostral verdadeiro;
4. da covariância entre médias ponderadas do salário e da receita considerando o plano amostral verdadeiro.

A partir da simulação foram obtidos os seguintes resultados:

$$\mathbf{E}_{AES}\left(\mathbf{V}_0\right) = \begin{bmatrix} 1785,3 & 27,734 \\ 27,734 & 1,2852 \end{bmatrix}, \tag{4.8}$$

$$\mathbf{V}_{AES}(\hat{\theta}) = \begin{bmatrix} 250,41 & 3,2683 \\ 3,2683 & 0,42267 \end{bmatrix} \text{ e} \tag{4.9}$$

$$\mathbf{\Delta} = \left[\mathbf{E}_{AES}\left(\mathbf{V}_0\right)\right]^{-1} \mathbf{V}_{AES}(\hat{\theta}) = \begin{bmatrix} 0,1516 & -4,931 \\ -0,0007277 & 0,4353 \end{bmatrix} \ . \tag{4.10}$$

Os autovalores $\delta_1 = 0,447$ e $\delta_2 = 0,139$ de $\mathbf{\Delta}$ fornecem os efeitos generalizados do plano amostral.

Da mesma forma que o $\mathbf{EPA}\left(\hat{\theta}, v_0\right)$ definido em (4.2) para o caso uniparamétrico foi utilizado para corrigir níveis de confiança de intervalos e níveis de significância de testes, o $\mathbf{EMPA}(\hat{\theta}, \mathbf{V}_0)$ definido em (4.6) pode

ser utilizado para corrigir níveis de confiança de regiões de confiança e níveis de significância de testes de hipóteses no caso multiparamétrico. Para ilustrar, vamos considerar o problema de testar a hipótese $H_0 : \mu = \mu_0$, onde μ é o vetor de médias de um vetor de variáveis de pesquisa $\mathbf{y}$. A estatística de teste usualmente adotada para este caso é a T^2 de Hottelling dada por

$$T^2 = n\left(\bar{\mathbf{y}} - \mu_0\right)' \mathbf{S}_y^{-1}\left(\bar{\mathbf{y}} - \mu_0\right), \tag{4.11}$$

onde

$$\begin{aligned} \bar{\mathbf{y}} &= \frac{1}{n}\sum_{i\in s}\mathbf{y}_i, \quad \mathbf{S}_y = \frac{1}{n-1}\sum_{i\in s}\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)', \text{ e} \\ \mu_0 &= \left(\mu_{10}, \mu_{20}, \ldots, \mu_{K0}\right)' . \end{aligned}$$

Se as observações $\mathbf{y}_i$ são IID normais, a estatística T^2 tem a distribuição $\frac{(n-1)}{(n-K)}\mathbf{F}\left(K; n-K\right)$ sob H_0, onde $\mathbf{F}\left(K; n-K\right)$ denota uma variável aleatória com distribuição $\mathbf{F}$ com K e $(n-K)$ graus de liberdade. Mesmo se as observações $\mathbf{y}_i$ não forem normais, T^2 tem distribuição assintótica $\chi^2\left(K\right)$ quando $n \to \infty$, (Johnson and Wichern, 1988), p.191.

Contudo, se for utilizado um plano amostral complexo, T^2 tem aproximadamente a distribuição da variável $\sum_{i=1}^{K}\delta_i Z_i^2$, onde $Z_1, \ldots, Z_K$ são variáveis aleatórias independentes com distribuição normal padrão e os δ_i são os autovalores da matriz $\mathbf{\Delta} = \Sigma_{AAS}^{-1}\Sigma$, onde $\Sigma_{AAS} = E_p(\mathbf{S}_y/n)$ e $\Sigma = V_p(\bar{\mathbf{y}})$.

Vamos analisar o efeito do plano amostral sobre o nível de significância deste teste. Para simplificar, consideremos o caso em que $\delta_1 = \ldots = \delta_K = \delta$. Neste caso, o nível de significância real é dado aproximadamente por

$$P\left(\chi^2\left(K\right) > \chi_\alpha^2\left(K\right)/\delta\right) \tag{4.12}$$

onde $\chi_\alpha^2\left(K\right)$ é o quantil superior α de uma distribuição χ^2 com K graus de liberdade, isto é, o valor tal que $P\left[\chi^2\left(K\right) > \chi_\alpha^2\left(K\right)\right] = \alpha$.

A Tabela 4.7 apresenta os níveis de significância reais para $\alpha = 5\%$ para vários valores de K e δ. Mesmo quando os valores dos δ_i são distintos, os valores da Tabela 4.7 podem ser devidamente interpretados. Para isso, consideremos o $p$valor do teste da hipótese $H_0 : \mu = \mu_0$, sob a hipótese de amostragem aleatória simples com reposição e sob o plano amostral efetivamente utilizado. Por definição este valor é dado por

$$p\text{valor}_{AAS}\left(\bar{\mathbf{y}}\right) = P\left[\chi^2\left(K\right) > \left(\bar{\mathbf{y}} - \mu_0\right)' \mathbf{\Sigma}_{AAS}^{-1}\left(\bar{\mathbf{y}} - \mu_0\right)\right]$$

e H_0 é rejeitada com nível de significância α se valor-p $_{AAS} < \alpha$.

O verdadeiro valor-p pode ser definido analogamente como

$$p\text{valor}_{VERD}\left(\bar{\mathbf{y}}\right) = P\left[\chi^2\left(K\right) > \left(\bar{\mathbf{y}} - \mu_0\right)' \mathbf{\Sigma}_{VERD}^{-1}\left(\bar{\mathbf{y}} - \mu_0\right)\right] . \tag{4.13}$$

Os valores na Tabela 4.7 podem ser usados para quantificar a diferença entre estes valores-p. Consideremos a região crítica do teste de nível α baseado na hipótese de AAS:

$$\begin{aligned} RC_{AAS}\left(\bar{\mathbf{y}}\right) &= \left\{\bar{\mathbf{y}} : \left(\bar{\mathbf{y}} - \mu_0\right)' \mathbf{\Sigma}_{AAS}^{-1}\left(\bar{\mathbf{y}} - \mu_0\right) > \chi_\alpha^2\left(K\right)\right\} \\ &= \left\{\bar{\mathbf{y}} : p\text{valor}_{AAS}\left(\bar{\mathbf{y}}\right) < \alpha\right\}. \end{aligned} \tag{4.14}$$

Tabela 4.7: Níveis de significância (%) verdadeiros do teste T^2 para o nível nominal de 5% assumindo autovalores iguais para $\mathbf{\Delta}$

	K			
δ	1	2	3	4
0,9	4	4	3	3
1,0	5	5	5	5
1,5	11	14	16	19
2,0	17	22	27	32
2,5	22	30	37	44
3,0	26	37	46	53

Pode-se mostrar que o máximo de $p\text{valor}_{VERD}\left(\bar{\mathbf{y}}\right)$ quando $\bar{\mathbf{y}}$ pertence à $RC_{AAS}\left(\bar{\mathbf{y}}\right)$ é dado por:

$$\max_{\bar{\mathbf{y}} \in RC_{AAS}(\bar{\mathbf{y}})} p\text{valor}_{VERD}\left(\bar{\mathbf{y}}\right) = P\left(\chi^2\left(K\right) > \chi^2_{\alpha}\left(K\right)/\delta_1\right). \tag{4.15}$$

Observe que o segundo membro de (4.15) é da mesma forma que o segundo membro de (4.12). Logo, os valores da Tabela 4.7 podem ser interpretados como valores máximos de $p\text{valor}_{VERD}\left(\bar{\mathbf{y}}\right)$ para $\bar{\mathbf{y}}$ na região $RC_{AAS}\left(\bar{\mathbf{y}}\right)$, considerando-se δ_1 no lugar de δ.

4.6 Laboratório de R

Vale a pena incluir as contas que não dependem da simulação?

Exemplo 4.2 Parte de simulação do exemplo para salario

```
library(survey)
```

```
## Loading required package: grid
```

```
## Loading required package: Matrix
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survey'
```

```
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
source("~\\GitHub\\adac\\data\\popul.R")
names(popul)<- c("ID","PO", "SAL","REC","ESTRAT")
```

```
N<-nrow(popul)
n1<-30; n2<-30
nh=c(n1,n2)
n<-sum(nh)
Nh<-table(popul$ESTRAT)
fh<-nh/Nh ; Wh<-Nh/N ; f<- n/N
```

Simulação (incluir?)

```
SAL.mat.res<-matrix(NA,500,4)
REC.mat.res<-matrix(NA,500,4)
popul$SAL <- popul$SAL/1000
popul$REC <- popul$REC/1000000
library(sampling)
```

```
##
## Attaching package: 'sampling'
```

```
## The following objects are masked from 'package:survival':
##
##     cluster, strata
```

```
#Geracão dos dados
for(i in 1:500){
s<-strata(popul, "ESTRAT",c(30,30),method= "srswor")
dados<-getdata(popul,s)
SAL_med_y<-mean(dados$SAL)
REC_med_y<-mean(dados$REC)
# estimador v0
SAL_var_AAS<-(1-f)*var(dados$SAL)/n
REC_var_AAS<-(1-f)*var(dados$REC)/n
# Vhat_AES estimador não-viciado
desenho<-svydesign(~1,strata=~ESTRAT,data=dados,fpc=~Prob)
SAL_med_yw<-svymean(~SAL,desenho)
REC_med_yw<-svymean(~REC,desenho)
SAL.mat.res[i,]<-c(SAL_med_y,coef(SAL_med_yw),SAL_var_AAS,SE(SAL_med_yw)^2)
REC.mat.res[i,]<-c(REC_med_y,coef(REC_med_yw),REC_var_AAS,SE(REC_med_yw)^2)
}
# 1-media de ybar; 2- média de ybar_w; 3- média de v0; 4- média de Vhat_AES
SAL_res_mean <-colMeans(SAL.mat.res)
REC_res_mean <- colMeans(REC.mat.res)
# EPA = média de Vhat_AES/ média de v0

EPA_SAL1 <- SAL_res_mean[4]/SAL_res_mean[3]
EPA_SAL1
```

```
## [1] 0.1407099
```

```
EPA_REC1 <- REC_res_mean[4]/REC_res_mean[3]
EPA_REC1
```

```
## [1] 0.3457303
```

Usando a fórmula (4.3)

```
# variância sob distr. de aleatorização verdadeira
Vp_SAL_h <- aggregate(SAL~ESTRAT, popul, var)
Vp_REC_h <- aggregate(REC~ESTRAT, popul, var)
Vp_SAL <- sum(Wh^2*(1-fh)*Vp_SAL_h[,"SAL"]/nh)
Vp_SAL <- sum(Wh^2*(1-fh)*Vp_REC_h[,"REC"]/nh)
```

```
# valor esperado v0 usando população
```

Capítulo 5

Ajuste de Modelos Paramétricos

Capítulo 6

Modelos de Regressão

Capítulo 7

Testes de Qualidade de Ajuste

Capítulo 8

Testes em Tabelas de Duas Entradas

Capítulo 9

Estimação de densidades

Capítulo 10

Modelos Hierárquicos

Capítulo 11

Não-Resposta

Capítulo 12

Diagnóstico de ajuste de modelo

Capítulo 13

Agregação vs. Desagregação

Capítulo 14

Pacotes para Analisar Dados Amostrais

Capítulo 15

Placeholder

Referências Bibliográficas

Haggard, E. A. (1958). Intraclass Correlation and the Analysis of Variance. Dryden Press, Nova Iorque.

Johnson, R. A. and Wichern, D. W. (1988). Applied Multivariate Statistical Analysis. Prentice Hall, Englewood Cliffs, New Jersey.

Kish, L. (1965). Survey Sampling. John Wiley and Sons, Nova Iorque.

Lehtonen, R. and Pahkinen, E. J. (1995). Practical Methods for Design and Analysis of Complex Surveys. John Wiley and Sons, Chichester.

Nascimento Silva, P. L. D. and Moura, F. A. S. (1990). Efeitos de conglomeração da malha setorial do censo demográfico 80. Série Textos para Discussão 32, IBGE, Diretoria de Pesquisas, Rio de Janeiro.

Skinner, C. J. (1989a). Introduction to Part A. In Analysis of Complex Surveys, pages 23--57. John Wiley and Sons, Chichester.

Skinner, C. J., Holt, D., and Smith, T. M. F., editors (1989). Analysis of Complex Surveys. John Wiley and Sons, Chichester.

Sudman, S. (1976). Applied Sampling. Academic Press, Nova Iorque.