MAT02025 - Amostragem 1

AAS: proporções e porcentagens por amostragem

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2021

Características qualitativas

Características qualitativas

- Às vezes, desejamos estimar o número total, a proporção ou a porcentagem de unidades na população que possuem alguma **característica** ou **atributo** ou se enquadram em alguma classe definida.
  - Muitos dos resultados regularmente publicados de censos ou pesquisas são desta forma, por exemplo, o **número de pessoas desempregadas**, a **porcentagem da população nativa**.
- A classificação pode ser introduzida diretamente no questionário, como nas perguntas que são respondidas com um simples **"sim"** ou **"não"**.
- Em outros casos, as medidas originais são mais ou menos contínuas e a classificação é introduzida na tabulação dos resultados.
  - Assim, podemos registrar as idades dos respondentes até o ano mais próximo, mas publicar a porcentagem da população com 60 anos ou mais.

Características qualitativas

Notação

Supomos que cada unidade na população cai em uma das duas classes C e C'. A notação é a seguinte:

Número de unidades da categoria C na		Proporção de unidades de C na	
População	Amostra	População	Amostra
A	a	$P = A/N$	$p = a/n$

Características qualitativas

- A estimativa amostral de P é p, e a estimativa amostral de A é Np ou Na/n.
- No trabalho estatístico, a **distribuição binomial** é frequentemente aplicada a estimativas como a e p.
- Como será visto, a distribuição correta para populações finitas é a **hipergeométrica**, embora o binomial seja geralmente uma aproximação satisfatória.

Variâncias das estimativas amostrais

Variâncias das estimativas amostrais

- Por meio de um artifício simples, é possível aplicar os teoremas estabelecidos nas aulas anteriores a essa situação.
- Para qualquer unidade na amostra ou população, atribui-se valor 1 a Y_i se a unidade estiver em C, e 0 se estiver em C'.
- Para esta população de valores Y_i, é evidente que

$$Y_T = \sum_{i=1}^{N} Y_i = A,$$

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i = \frac{A}{N} = P.$$

Variâncias das estimativas amostrais

- E também, para a amostra,

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{a}{n} = p.$$

Variâncias das estimativas amostrais

- Consequentemente, o problema de estimar A e P pode ser considerado como o de estimar o total e a média de uma população em que cada Y_i é 1 ou 0.
- Para usar os teoremas das aulas anteriores, primeiro expressamos S^2 e s^2 em termos de P e p.
- Observe que

$$\sum_{i=1}^{N} Y_i^2 = A = NP, \quad \sum_{i=1}^{n} Y_i^2 = a = np.$$

Variâncias das estimativas amostrais

- Portanto,

$$S^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1} = \frac{\sum_{i=1}^{N} Y_i^2 - N\overline{Y}^2}{N-1}$$
$$= \frac{1}{N-1}(NP - NP^2) = \frac{N}{N-1}PQ,$$

em que $Q = 1 - P$. Semelhantemente,

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{y})^2}{n-1} = \frac{n}{n-1}pq. \qquad (1)$$

Variâncias das estimativas amostrais

- A aplicação dos **teoremas das aulas 9, 10 e 11** a essa população fornece os seguintes resultados para uma amostragem aleatória simples das unidades que estão sendo classificadas.

Teorema

A proporção da amostra $p = a/n$ é uma estimativa não enviesada da proporção da população $P = A/N$.

Teorema

A variância de p é

$$\operatorname{Var}(p) = \mathrm{E}(p - P)^2 = \frac{S^2}{n}\left(\frac{N-n}{N}\right) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right). \qquad (2)$$

Variâncias das estimativas amostrais

Corolário

Se p e P são as porcentagens da amostra e da população, respectivamente, caindo na classe C, (2) continua valendo para a variância de p.

Corolário

A variância de $\widehat{A} = Np$, o número total estimado de unidades na classe C, é

$$\operatorname{Var}(\widehat{A}) = \frac{N^2 PQ}{n}\left(\frac{N-n}{N-1}\right).$$

Variâncias das estimativas amostrais

Teorema

Uma estimativa imparcial da variância de p, derivada da amostra, é

$$\widehat{Var}(p) = s_p^2 = \frac{N-n}{(n-1)N}pq.$$

Variâncias das estimativas amostrais

Demonstração. No corolário do teorema da aula 11 foi mostrado que para uma variável Y_i uma estimativa não enviesada da variância da média amostral $\overline{y}$ é

$$\widehat{\text{Var}}(\overline{y}) = \frac{s^2}{n}\frac{(N-n)}{N}.$$

- Para proporções, p toma o lugar de $\overline{y}$, e em (1) mostramos que

$$s^2 = \frac{n}{n-1}pq.$$

Variâncias das estimativas amostrais

- Portanto,

$$\widehat{Var}(p) = s_p^2 = \frac{N-n}{(n-1)N} pq.$$

- Segue-se que se N é muito grande em relação a n, de modo que a **fpc** é desprezível, uma estimativa não enviesada da variância de p é

$$\frac{pq}{n-1}.$$

- O resultado pode parecer intrigante, uma vez que a expressão pq/n é quase invariavelmente usada na prática para a variância estimada.
  - O fato é que pq/n não é imparcial, mesmo com uma população infinita.

Variâncias das estimativas amostrais

Corolário

Uma estimativa não enviesada da variância de $\widehat{A} = Np$, o número total estimado de unidades da classe C na população, é

$$\widehat{\text{Var}}\,(\widehat{A}) = s^2_{N_p} = \frac{N(N-1)}{n-1}pq.$$

Exemplo

Exemplo

- De uma lista de 3042 nomes e endereços, uma amostra aleatória simples de 200 nomes mostrou na investigação 38 endereços errados.
- **Problema:** estimar o número total de endereços que precisam de correção na lista e encontrar o erro padrão dessa estimativa.
- Nós temos

$$N = 3042; \quad n = 200; \quad a = 38; \quad p = 0,19.$$

Exemplo

- O número total estimado de endereços errados é

$$\widehat{A} = Np = 3042 \times 0,19 \approx 578.$$

- O erro padrão será

$$s_{\widehat{A}} = \sqrt{[(3042 \times 2842 \times 0,19 \times 0,81)/199]} \approx 81,8.$$

Exemplo

- Como a fração de amostragem está abaixo de 7%, a fpc faz pouca diferença.
- Para removê-lo, substitua o termo $N - n$ por N.
- Se, além disso, substituirmos $n - 1$ por n, temos a fórmula mais simples

$$s_{N_p} = N\sqrt{pq/n} = (3042)\sqrt{[(0,19 \times 0,81)/200]} = 84,4.$$

- Isso está bastante de acordo com o resultado anterior, 81,8.

Considerações finais

Considerações finais

- As expressões anteriores para a variância e a variância estimada de p são válidas apenas se as unidades forem classificadas em C ou C', de modo que p seja a razão entre o número de unidades em C na amostra e o número total de unidades na amostra.
- Em muitos levantamentos por amostragem, cada unidade é composta por um **grupo de elementos**, e são os elementos que são classificados. Alguns exemplos são os seguintes:

Unidade de amostragem	Elementos componentes
Família/domicílio	Membros da família/domicílio
Restaurante	Funcionários
Engradados de ovos	Cada ovo
Pessegueiro	Cada pêssego

Considerações finais

- Se uma amostra aleatória simples de unidades for delineada para estimar a proporção P dos **elementos** na população que pertencem à classe C, as fórmulas anteriores **não se aplicam**.
  - Os métodos apropriados são fornecidos em aulas futuras.

Para casa

- Revisar os tópicos discutidos nesta aula.
- Atividade de avaliação 2.

Próxima aula

- A influência de P nos erros padrões.
- As distribuições binomial e hipergeométrica.

Por hoje é só!

Bons estudos!