

# Introdução à estimadores regressão

Prof. Caio Azevedo

# Motivação

- **Vimos que**, quando há informações (em nível populacional) sobre uma variável auxiliar ( $x_i$ ) e que a relação desta com a variável de interesse ( $y_i$ ) puder ser expressa como uma reta passando pela origem (além de outros fatores: magnitude dos coeficientes de variação e da correlação entre as duas variáveis), estimadores razão podem ser mais apropriados do que os estimadores usuais.
- Tal abordagem também pode ser útil quando não conhecemos, por exemplo, o tamanho da população ( $N$ ).

# Cenário

- Exploraremos, agora, estimadores que são mais apropriados quando a relação entre tais variáveis puder ser expressa como uma reta que não passa pela origem, ou seja:  $y_i = \alpha + \beta x_i + e_i$  (nível populacional)  $Y_i = \alpha + \beta X_i + E_i$  (nível amostral) ([link](#) sobre regressão).
- Consideraremos as mesmas notações e suposições (acerca dos parâmetros populacionais da variável  $x$ ) [feitas para o estimador razão](#).

# Cenário

- O estimador regressão para a média é definido como:

$$\hat{\mu}_{Reg} = \hat{\mu} + b(\mu_x - \bar{X})$$

em que  $b$  é uma estimativa apropriada de  $\beta$  (slide anterior).

Consideraremos, portanto, em princípio, que  $b$  é não-aleatório e

$$\hat{\mu} = \bar{Y}.$$

- Note que se  $b \approx 0$  e/ou  $\bar{x}$  for pequeno com relação à  $\mu_x$  então, devido à linearidade entre  $y$  e  $x$ , a diferença entre  $\hat{\mu}_{Reg}$  e  $\hat{\mu}$  também é pequena.

# Propriedades dos estimadores

- Note ainda que o estimador  $\hat{\mu}_{Reg}$  faz uma “correção” em  $\hat{\mu}$ , isto é, adiciona a  $\hat{\mu}$  uma quantidade proporcional a  $\mu_x - \bar{X}$ , ou seja,  $b(\mu_x - \bar{X})$ .
- O estimador regressão para o total populacional é dado por  $\hat{\tau}_{Reg} = N\hat{\mu}_{Reg}$ .
- Consideraremos, inicialmente, que  $b = b_0$  (é um valor conhecido).
- Note que, nesse caso, temos duas variáveis aleatórias no estimador  $\hat{\mu}_{Reg}$ , nomeadamente  $\hat{\mu} (= \bar{Y})$  e  $\bar{X}$ .

# Propriedades dos estimadores

- Tem-se

$$\begin{aligned}\mathcal{E}(\hat{\mu}_{Reg}) &= \mathcal{E}[\hat{\mu} + b_0(\mu_x - \bar{X})] = \mathcal{E}(\hat{\mu}) + b_0[\mu_x - \mathcal{E}(\bar{X})] \\ &= \mu_y + b_0(\mu_x - \mu_x) = \mu_y\end{aligned}$$

- Para o cálculo da variância, defina  $D_i = Y_i - b_0(X_i - \mu_x)$ ,  $i = 1, 2, \dots, n$  e note que

$$\begin{aligned}\hat{\mu}_{Reg} &= \frac{1}{n} \sum_{i=1}^n Y_i - b_0 \left( \frac{1}{n} \sum_{i=1}^n X_i - \mu_x \right) = \frac{1}{n} \sum_{i=1}^n [Y_i - b_0(X_i - \mu_x)] \\ &= \frac{1}{n} \sum_{i=1}^n D_i = \bar{D}\end{aligned}$$

# Propriedades dos estimadores

- Defina agora  $d_i = y_i - b_0(x_i - \mu_x)$ ,  $i=1,2,\dots,N$ . Note que

$$\begin{aligned}\mu_d &= \frac{1}{N} \sum_{i=1}^N d_i = \frac{1}{N} \sum_{i=1}^N [y_i - b_0(x_i - \mu_x)] \\ &= \frac{1}{N} \sum_{i=1}^N y_i - b_0 \left( \frac{1}{N} \sum_{i=1}^N x_i - \mu_x \right) = \mu_y - b_0(\mu_x - \mu_x) = \mu_y\end{aligned}$$

## Cont.

- Além disso:

$$\begin{aligned}\sigma_d^2 &= \frac{1}{N} \sum_{i=1}^N [d_i - \mu_d]^2 = \frac{1}{N} \sum_{i=1}^N [y_i - b_0(x_i - \mu_x) - \mu_y]^2 \\ &= \frac{1}{N} \sum_{i=1}^N [(y_i - \mu_y) - b_0(x_i - \mu_x)]^2 \\ &= \frac{1}{N} \sum_{i=1}^N \left[ (y_i - \mu_y)^2 - 2b_0(x_i - \mu_x)(y_i - \mu_y) + b_0^2(x_i - \mu_x)^2 \right]^2 \\ &= \frac{1}{N} \sum_{i=1}^N (y_i - \mu_y)^2 - 2b_0 \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y) \\ &+ b_0^2 \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)^2 = \sigma_y^2 - 2b_0\sigma_{xy} + b_0^2\sigma_x^2\end{aligned}$$



# Propriedades dos estimadores

- (Cont.) Em que  $\sigma_y^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu_y)^2$  (analogamente para  $s_x^2$ ),  
 $\sigma_{xy} = \text{Cov}(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)$ .
- Portanto, temos que

$$\mathcal{V}_{A_1}(\hat{\mu}_{Reg}) = \mathcal{V}_{A_1}(\bar{D}) = \frac{\sigma_d^2}{n}$$

$$\mathcal{V}_{A_2}(\hat{\mu}_{Reg}) = \mathcal{V}_{A_2}(\bar{D}) = (1 - f) \frac{s_d^2}{n}$$

em que  $f = \frac{n}{N}$ ,  $s_d^2 = s_y^2 - 2b_0 s_{xy} + b_0^2 s_x^2$ ,  $s_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu_y)^2$   
(analogamente para  $s_x^2$ ) e  $s_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)$ .

# Propriedades dos estimadores

- Estimadores não viesados para as variâncias anteriores são dados, respectivamente, por:

$$\begin{aligned}\widehat{V}_{A_1}(\widehat{\mu}_{Reg}) &= \frac{\widehat{\sigma}_d^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^n [(Y_i - \bar{Y}) - b_0(X_i - \bar{X})]^2 \\ &= \frac{1}{n} (\widehat{\sigma}_y^2 - 2b_0\widehat{\sigma}_{xy} + b_0^2\widehat{\sigma}_x^2)\end{aligned}$$

$$\begin{aligned}\widehat{V}_{A_2}(\widehat{\mu}_{Reg}) &= (1-f)\frac{\widehat{S}_d^2}{n} = \frac{1-f}{n(n-1)} \sum_{i=1}^n [(Y_i - \bar{Y}) - b_0(X_i - \bar{X})]^2 \\ &= \frac{1-f}{n} (\widehat{S}_y^2 - 2b_0\widehat{S}_{xy} + b_0^2\widehat{S}_x^2)\end{aligned}$$

# Propriedades dos estimadores

- (Cont.) Em que

$$\hat{\sigma}_d^2 = \hat{s}_d^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \hat{\mu}_{Reg})^2, \hat{\sigma}_y^2 = \hat{s}_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

$$\hat{\sigma}_x^2 = \hat{s}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

$$\hat{\sigma}_{xy} = \hat{s}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}).$$

- Prova: Basta utilizar algumas propriedades do valor esperado e lembrar que  $\mathcal{E}_{A_1}(\hat{\sigma}_y^2) = \sigma_y^2$ ,  $\mathcal{E}_{A_1}(\hat{\sigma}_x^2) = \sigma_x^2$ ,  $\mathcal{E}_{A_2}(\hat{s}_y^2) = \sigma_y^2$ ,  $\mathcal{E}_{A_2}(\hat{s}_x^2) = \sigma_x^2$ ,  $\mathcal{E}_{A_1}(\hat{\sigma}_{xy}) = \sigma_{xy}$ ,  $\mathcal{E}_{A_1}(\hat{s}_{xy}) = s_{xy}$ .

# Otimidade do estimador regressão

- O valor de  $b_0$  que minimiza  $\mathcal{V}_{A_1}(\hat{\mu}_{Reg})$  é dado por

$$B_0 = \frac{\sum_{i=1}^N (y_i - \mu_y)(x_i - \mu_x)}{\sum_{i=1}^N (x_i - \mu_x)^2} = \frac{\sigma_{xy}}{\sigma_x^2} \quad (1)$$

- Além disso, para o  $b_0$  acima,  $\mathcal{V}_{A_1(\min)}(\hat{\mu}_{Reg}) = \frac{\sigma_y^2}{n} (1 - \rho^2[x, y])$ ,  
 $\rho[x, y] = \sigma_{xy}/(\sigma_x\sigma_y)$ .

## Cont;

- Prova: Seja  $b_0 = B_0 + c$ ,  $c \in \Re$ . Tem-se, então, para este  $b_0$  que

$$\begin{aligned}\mathcal{V}_{A_1}(\hat{\mu}_{Reg}) &= \frac{1}{n} [\sigma_y^2 - 2(B_0 + c)\sigma_{xy} + (B_0 + c)^2\sigma_x^2] \\ &= \frac{1}{n} \left( \sigma_y^2 - 2\frac{\sigma_{xy}^2}{\sigma_x^2} - 2c\sigma_{xy} + \frac{\sigma_{xy}^2}{\sigma_x^2} + 2c\sigma_{xy} + c^2\sigma_x^2 \right) \\ &= \frac{1}{n} \left\{ \left( \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} \right) + c^2\sigma_x^2 \right\}\end{aligned}$$

que é mínimo quando  $c = 0$ , lembrando que

$$\rho_{xy} = \rho[x, y] = \sigma_{xy} / \sigma_x \sigma_y.$$

## Cont.

- Como na prática não é possível utilizar  $B_0$  (parâmetro), devemos considerar um estimador apropriado, ou seja:

$$\hat{B}_0 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \quad (2)$$

- Como estimador para  $\mathcal{V}_{A_1}(\hat{\mu}_{Reg})$  consideramos

$$\hat{\mathcal{V}}_{A_1}(\hat{\mu}_{Reg}) = \frac{1}{n} \left( \hat{\sigma}_y^2 - 2\hat{B}_0\hat{\sigma}_{xy} + \hat{B}_0^2\hat{\sigma}_x^2 \right)$$

- Note que, neste caso, tal estimador não é, necessariamente, não viciado.

## Exemplo

- Considere uma população formada por três domicílios,  $\mathcal{U} = \{1, 2, 3\}$  e que se observam as seguintes variáveis: nome (do chefe), sexo, idade, fumante ou não, renda bruta (mensal em salários mínimos) familiar e número de trabalhadores.
- Considere que o objetivo é estimar  $\bar{f}$  a média da renda bruta familiar.

## Cont.

Variável	Valores			Notação
Unidade	1	2	3	$i$
nome do chefe	Ada	Beto	Ema	$a_i$
sexo (0: F, 1:M)	0	1	0	$x_i$
idade	20	30	40	$y_i$
fumante (0: N, 1:S)	0	1	1	$g_i$
renda bruta familiar	12	30	18	$f_i$
nº de trabalhadores	1	3	2	$t_i$



## Cont.

- Considere um plano AASc e  $n = 2$ . Além disso, defina  $\bar{F}_R = \mu_t \left( \frac{\bar{F}}{\bar{T}} \right)$  e  $\bar{F}_{Reg} = \bar{F} + B_0(\mu_t - \bar{T})$  (para  $B_0$ , veja slide seguinte)
- Dessa forma, temos que:

<b>s</b>	11	12	13	21	22	23	31	32	33
$\bar{F}$	12	21	15	21	30	24	15	24	18
$\bar{T}$	1,0	2,0	1,5	2,0	3,0	2,5	1,5	2,5	2,0
$\bar{F}_{Reg}$	21,0	21,0	19,5	21	21,0	19,5	19,5	19,5	18,0
$P(\mathbf{s}) :$	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

## Cont.

- (Cont.) De onde obtemos que  $\mathcal{E}(\bar{F}_{Reg}) = 20$ ,  $\mathcal{V}(\bar{F}_{Reg}) = 1$ ,  
 $\mathcal{E}(\bar{F}_R) \approx 20,27$ ,  $\mathcal{V}(\bar{F}_R) \approx 2,52$ .
- Neste caso temos que  $EQM(\bar{F}_{Reg}) = 1 < EQM(\bar{F}_R) = 2,59$  (este último obtido em [aqui](#), páginas 9 e 10. Em geral, tal resultado é válido.

## Cont.

- Podemos ainda obter a variância do estimador regressão utilizando a fórmula obtida anteriormente. Note que  $\mu_t = 2$ ,  $\mu_f = 20$ ,  $\sum_{i=1}^3 f_i t_i = 138$ ,  $\sum_{i=1}^3 f_i^2 = 1368$ ,  $\sum_{i=1}^3 t_i^2 = 14$ . Assim,

$$B_0 = \frac{\sum_{i=1}^N f_i t_i - N \bar{f} \bar{t}}{\sum_{i=1}^N t_i^2 - N \bar{t}^2} = \frac{18}{2} = 9$$

- Também, com  $b_0 = B_0 = 9$ ,  $\sigma_f^2 = 56$ ,  $\sigma_t^2 = 2/3$  e  $\sigma_{ft} = 6$ , tem-se, para  $n = 2$ , que

$$\mathcal{V}(\bar{F}_{Reg}) = \frac{1}{n} (\sigma_f^2 - 2b_0 \sigma_{ft} + b_0^2 \sigma_t^2) = 1.$$

# Comparação entre os estimadores razão e regressão

- Considerando  $B_0$  dado em (1) e sob  $AAS_c$ , temos que

1  $\mathcal{V}(\hat{\mu}_{Reg}) \leq \mathcal{V}(\hat{\mu})$ .

2  $\mathcal{V}(\hat{\mu}_{Reg}) \leq \mathcal{V}(\hat{\mu}_R)$ .

- Prova: Temos que  $\mathcal{V}(\hat{\mu}) = \frac{\sigma_y^2}{n}$  e

$$\mathcal{V}(\hat{\mu}_{Reg}) = \frac{\sigma_y^2}{n} (1 - \rho[x, y]^2) = \mathcal{V}(\hat{\mu}) (1 - \rho[x, y]^2),$$

de onde segue o item 1), pois  $0 \leq \rho[x, y]^2 \leq 1$ .

# Comparação entre os estimadores razão e regressão

- Por outro lado, temos que

$$\mathcal{V}(\hat{\mu}_R) \approx \frac{1}{n} (\sigma_y^2 - 2r\rho[x, y]\sigma_x\sigma_y + r^2\sigma_x^2).$$

- Portanto, vem que

$$\begin{aligned}\mathcal{V}(\hat{\mu}_R) - \mathcal{V}(\hat{\mu}_{Reg}) &\approx \frac{\sigma_y^2}{n} - \frac{2}{n}\rho[x, y]\sigma_x\sigma_y + \frac{1}{n}r^2\sigma_x^2 - \frac{\sigma_y^2}{n} + \frac{\sigma_y^2}{n}\rho[x, y]^2 \\ &= \frac{1}{n} (\rho^2[x, y]\sigma_y^2 - 2r\rho[x, y]\sigma_x\sigma_y + r^2\sigma_x^2) \\ &= \frac{1}{n} (\rho[x, y]\sigma_y - r\sigma_x)^2 \geq 0\end{aligned}$$

De onde se conclui o item 2).

# Comparação entre os estimadores razão e regressão (para o total populacional)

- Temos que  $\mathcal{E}(\widehat{\tau}_{Reg}) = \mathcal{E}(N\widehat{\mu}_{Reg}) = N\mathcal{E}(\widehat{\mu}_{Reg}) = N\mu_y = \tau_y$ .
- Variâncias do estimador:  $\mathcal{V}_{A_i}(\widehat{\tau}_{Reg}) = N^2\mathcal{V}_{A_i}(\widehat{\mu}_{Reg})$ ,  
 $\mathcal{V}_{A_1}(\widehat{\tau}_{Reg}) = N^2\frac{\sigma_d^2}{n}$  e  $\mathcal{V}_{A_2}(\widehat{\tau}_{Reg}) = (1-f)N^2\frac{s_d^2}{n}$ .
- Estimativas da variância do estimador:  $\widehat{\mathcal{V}}_{A_1}(\widehat{\tau}_{Reg}) = N^2\frac{\widehat{\sigma}_d^2}{n}$  e  
 $\widehat{\mathcal{V}}_{A_2}(\widehat{\tau}_{Reg}) = (1-f)N^2\frac{\widehat{s}_d^2}{n}$ .
- Intervalos de confiança, testes de hipótese, e tamanho de amostra podem ser obtidos de forma semelhante aos estimadores usuais (para a média e total) sob  $AAS_c$  e  $AAS_s$ .

# Comentários

- À rigor, portanto, o estimador regressão (para a média) é dado por:

$$\hat{\mu}_{Reg}^* = \hat{\mu} + \hat{B}_0 (\mu_x - \bar{X})$$

- Note que as maioria das expressões (esperança e variância dos estimadores regressão) foram obtidas considerando-se  $b_0(B_0)$  fixo (não aleatório). Como  $\hat{B}_0 \xrightarrow[n \rightarrow \infty, N-n \rightarrow \infty]{P} B_0$  ([link 1](#), [link 2](#)), então os resultados são válidos, para  $n$  ( $N-n$ ) suficientemente grandes.
- Ou seja,  $|\hat{\mu}_{Reg}^* - \hat{\mu}_{Reg}| \xrightarrow[n \rightarrow \infty, N-n \rightarrow \infty]{P} 0$ .
- Assim,  $\mathcal{E}_{PA}(\hat{\mu}_{Reg}^*) \approx \mathcal{E}_{PA}(\hat{\mu}_{Reg})$  e  $\mathcal{V}_{PA}(\hat{\mu}_{Reg}^*) \approx \mathcal{V}_{PA}(\hat{\mu}_{Reg})$

## Resumo para $AAS_s$ (para $AAS_c$ é análogo)

- Se  $b = b_0$  for conhecido, então:

$$\begin{aligned}\hat{\mu}_{Reg} &= \hat{\mu} + b_0 (\mu_x - \bar{X}), \\ \mathcal{V}_{A_2}(\hat{\mu}_{Reg}) &= \frac{1-f}{n} (s_y^2 - 2b_0 s_{xy} + b_0^2 s_x^2), \\ \hat{\mathcal{V}}_{A_2}(\hat{\mu}_{Reg}) &= \frac{1-f}{n} (\hat{s}_y^2 - 2b_0 \hat{s}_{xy} + b_0^2 \hat{s}_x^2), \\ EP_{A_2}(\hat{\mu}_{Reg}) &= \sqrt{\mathcal{V}_{A_2}(\hat{\mu}_{Reg})}, \\ \hat{EP}_{A_2}(\hat{\mu}_{Reg}) &= \sqrt{\hat{\mathcal{V}}_{A_2}(\hat{\mu}_{Reg})}\end{aligned}$$



## Resumo para $AAS_s$ (para $AAS_c$ é análogo)

- Se  $b = b_0$  for desconhecido, usa-se  $\widehat{B}_0$  dado em (2), então:

$$\widehat{\mu}_{Reg}^* = \widehat{\mu} + \widehat{B}_0 (\mu_x - \bar{X}),$$

$$\mathcal{V}_{A_2}(\widehat{\mu}_{Reg}^*) = \frac{1-f}{n} (s_y^2 - 2B_0 s_{xy} + B_0^2 s_x^2) = \frac{1-f}{n} (1 - \rho_{xy}^2),$$

$$\widehat{\mathcal{V}}_{A_2}(\widehat{\mu}_{Reg}^*) = \frac{1-f}{n} (\widehat{s}_y^2 - 2\widehat{B}_0 \widehat{s}_{xy} + \widehat{B}_0^2 \widehat{s}_x^2) = \frac{1-f}{n} (1 - \widehat{\rho}_{xy}^2),$$

$$EP_{A_2}(\widehat{\mu}_{Reg}^*) = \sqrt{\mathcal{V}_{A_2}(\widehat{\mu}_{Reg}^*)},$$

$$\widehat{EP}_{A_2}(\widehat{\mu}_{Reg}^*) = \sqrt{\widehat{\mathcal{V}}_{A_2}(\widehat{\mu}_{Reg}^*)}$$

- Em que  $\widehat{\rho}_{xy} = \frac{\widehat{s}_{xy}}{\widehat{s}_x \widehat{s}_y}$ ,  $\widehat{s}_x = \sqrt{\widehat{s}_x^2}$  e  $\widehat{s}_y = \sqrt{\widehat{s}_y^2}$ . O análogo vale para

$$\widehat{\tau}_{Reg} = N \widehat{\mu}_{Reg}.$$

# Teoria assintótica

- Para  $n$  e  $N$  suficientemente grandes, a convergência em distribuição para a normal padrão é válida, de modo semelhante ao caso dos estimadores usuais sob  $AAS_c$  e  $AAS_s$  (PA: indica um plano amostral apropriado).
- Portanto, sob certas condições, em relação à estimação da média, temos que (válido, de forma análoga, para  $\hat{\mu}_{Reg}^*$ )

$$\frac{\hat{\mu}_{Reg} - \mu}{EP_{PA}(\hat{\mu}_{Reg})} \xrightarrow[n \rightarrow \infty, N - n \rightarrow \infty]{D} N(0, 1)$$
$$\frac{\hat{\mu}_{Reg} - \mu}{\widehat{EP}_{PA}(\hat{\mu}_{Reg})} \xrightarrow[n \rightarrow \infty, N - n \rightarrow \infty]{D} N(0, 1)$$

# Intervalos de confiança

- Analogamente aos casos anteriores, temos que dois intervalos assintóticos, com coeficiente de confiança de aproximadamente  $\gamma$ , são dados por:

$$IC(\mu, \gamma) \approx \left[ \hat{\mu}_{Reg} - z_\gamma \widehat{EP}_{PA}(\hat{\mu}_{Reg}); \hat{\mu}_{Reg} + z_\gamma \widehat{EP}_{PA}(\hat{\mu}_{Reg}) \right]$$
$$IC^*(\mu, \gamma) \approx \left[ \hat{\mu}_{Reg}^* - z_\gamma \widehat{EP}_{PA}(\hat{\mu}_{Reg}^*); \hat{\mu}_{Reg}^* + z_\gamma \widehat{EP}_{PA}(\hat{\mu}_{Reg}^*) \right]$$

- Erro da estimativa:  $z_\gamma \widehat{EP}_{PA}(\hat{\mu}_{Reg}) \left( z_\gamma \widehat{EP}_{PA}(\hat{\mu}_{Reg}^*) \right)$ .

# Testes de Hipótese

- Hipóteses usuais ( $\mu_0$  conhecido)

- 1  $H_0 : \mu = \mu_0$  vs  $H_1 : \mu < \mu_0$ .

- 2  $H_0 : \mu = \mu_0$  vs  $H_1 : \mu > \mu_0$ .

- 3  $H_0 : \mu = \mu_0$  vs  $H_1 : \mu \neq \mu_0$ .

- Estatística do teste  $Z_t = \frac{\hat{\mu}_{Reg} - \mu_0}{\text{EP}_{PA}(\hat{\mu}_{Reg})}$ .

- Sob  $H_0$ , vimos que  $Z_t \approx N(0, 1)$ , para  $n$  e  $N-n$  suficientemente grandes.

# Testes de Hipótese

- Defina  $z_t = \frac{\tilde{\mu}_{Reg} - \mu_0}{EP_{PA}(\hat{\mu}_{Reg})}$  o valor calculado da estatística do teste e  $z_c$  o(s) valor(es) crítico(s).
- Defina ainda  $Z \sim N(0, 1)$ . Os mecanismos de tomada de decisão são basicamente aqueles descritos nas páginas 20 a 23 dos [slides](#)
- O análogo pode ser feito considerando-se  $\hat{\mu}_{Reg}^*$  ao invés de  $\hat{\mu}_{Reg}$ .

# Determinação do tamanho amostral

- Estabelece-se algum critério de interesse acerca da acurácia/precisão na estimativa da média populacional.
- Sob o estimador proposto, calcula-se o tamanho da amostra, com base em sua distribuição assintótica obtida e critério estabelecido.
- Erro de estimativa:  $z_\gamma \widehat{EP}_{PA}(\widehat{\mu}_{Reg})$ . Fixa-se um erro de estimativa de interesse.
- Probabilidade do módulo da diferença  $P(|\widehat{\mu}_{Reg} - \mu| < \delta) > \gamma$ ,  
 $\delta > 0$ ,  $\gamma \in (0, 1)$ .

# Determinação do tamanho amostral: erro da estimativa

- $AAS_c$

$$\delta = z_\gamma \sqrt{\frac{\sigma_d^2}{n}} \rightarrow n = \frac{z_\gamma^2 \sigma_d^2}{\delta^2}$$

- Em geral, o (um) valor de  $\sigma_R^2$  é obtido através de pesquisas anteriores ou de uma amostra piloto, de tamanho apropriado.

# Determinação do tamanho amostral: erro da estimativa

## ■ $AAS_s$

$$\begin{aligned}\delta &= z_\gamma \sqrt{\frac{(1-f)s_d^2}{n}} \rightarrow \left(\frac{1}{n} - \frac{1}{N}\right) = \frac{\delta^2}{z_\gamma^2 s_d^2} \rightarrow \frac{1}{n} = \frac{\delta^2}{z_\gamma^2 s_d^2} + \frac{1}{N} \\ \rightarrow \frac{1}{n} &= \frac{\delta^2 N + z_\gamma^2 s_d^2}{N z_\gamma^2 s_d^2} \rightarrow n = \frac{N z_\gamma^2 s_d^2}{\delta^2 N + z_\gamma^2 s_d^2} = \frac{1}{\frac{\delta^2}{s_d^2 z_\gamma^2} + \frac{1}{N}}\end{aligned}$$

- Em geral, o (um) valor de  $s_d^2$  é obtido através de pesquisas anteriores ou de uma amostra piloto, de tamanho apropriado.
- Exercício: Construir IC's e testes de hipótese para o total populacional.



# Sob amostragem estratificada

- Se a população está estratificada, podemos considerar uma combinação dos resultados obtidos, anteriormente, sob AE, com os resultados obtidos para o estimador razão.
- A estrutura é a mesma daquela apresentada [aqui](#).
- Defina:  $\bar{Y}_h$ ,  $\bar{X}_h$ ,  $\mu_{xh}$ , as médias amostrais das variáveis  $y$  e  $x$  e a média populacional da variável  $x$ , respectivamente, no estrato  $h$ .

# Sob amostragem estratificada

- Como estimadores para a média e o total populacionais, podemos considerar, respectivamente:

$$\hat{\mu}_{Reges} = \sum_{h=1}^H W_h [\hat{\mu}_h + b_{0h} (\mu_{xh} - \bar{X}_h)] = \sum_{h=1}^H W_h \hat{\mu}_{Regh},$$

$$\hat{T}_{Reges} = N \hat{\mu}_{Reges} = \sum_{h=1}^H N_h \hat{\mu}_{Regh},$$

em que  $\hat{\mu}_{Regh} = \hat{\mu}_h + b_{0h} (\mu_{xh} - \bar{X}_h)$ .

# Sob amostragem estratificada

- Usando resultados anteriores, temos que:

$$\mathcal{V}_{AE_1}(\hat{\mu}_{Reges}) = \sum_{h=1}^H W_h^2 \frac{\sigma_{dh}^2}{n_h}$$

$$VM_{AE_2}(\hat{\mu}_{Reges}) = \sum_{h=1}^H W_h^2 (1 - f_h) \frac{s_{dh}^2}{n_h}$$

em que  $f_h = \frac{n_h}{N_h}$ ,  $\sigma_{dh}^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} \left( \sigma_{yh}^2 - 2b_0 \sigma_{xyh} + b_0^2 \sigma_x^2 \right)$  e  
 $s_{dh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left( s_{yh}^2 - 2b_0 s_{xyh} + b_0^2 \sigma_{xh}^2 \right)$ .

# Sob amostragem estratificada

- Estimadores para as variâncias são obtidos substituindo-se as quantidades desconhecidas por estimadores apropriados, ou seja

$$\hat{V}_{AE_1}(\hat{\mu}_{Reges}) = \sum_{h=1}^H W_h^2 \frac{\hat{\sigma}_{dh}^2}{n_h}$$

$$\hat{V}_{AE_2}(\hat{\mu}_{Reges}) = \sum_{h=1}^H W_h^2 (1 - f_h) \frac{\hat{S}_{dh}^2}{n_h}$$

em que  $\hat{\sigma}_{dh}^2 = \hat{S}_{dh}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left( \hat{s}_{yh}^2 - 2b_0 \hat{s}_{xyh} + b_0^2 \hat{s}_{xh}^2 \right)$ .

# Sob amostragem estratificada

- Exercício: desenvolver as fórmulas das variâncias e dos estimadores das variâncias para o estimador do total populacional  $\hat{T}_{Reges}$ .
- O comportamento assintótico dos estimadores são semelhantes àquele relativos aos **estimadores usuais sob AE**.
- Intervalos de confiança e testes de hipótese podem ser desenvolvidos de **modo semelhante ao exposto anteriormente**.

# Alocação ótima

- Sob alocação ótima e  $AAS_c$  dentro de cada estrato, como visto [aqui](#), temos que:

$$n_h = n \frac{N_h \sigma_{dh} / \sqrt{c_h}}{\sum_{h=1}^H N_h \sigma_{dh} / \sqrt{c_h}}$$

- Raciocínio análogo pode ser considerado sob  $AAS_s$  dentro de cada estrato.

# Comparaçãõ entre os estimadores razão e regressãõ

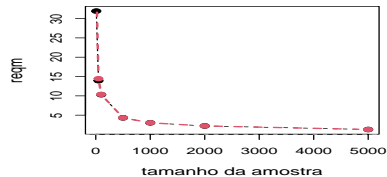
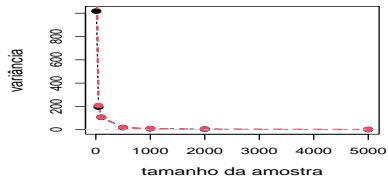
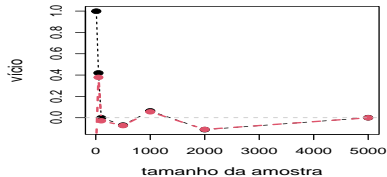
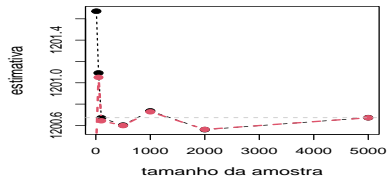
- Estudo de simulaçãõ sob o PA  $AAS_s$ .
- Duas variáveis (y-resposta, x-auxiliar) foram simuladas considerando  $N=100.000$  em dois cenários:  $Y_i = 1,2X_i + E_i$  (cenário 1),  
 $Y_i = 100 + 1,2X_i + E_i$  (cenário 2),  $X_i \overset{i.i.d.}{\sim} N(1.000, 10.000)$  e  
 $E_i \overset{i.i.d.}{\sim} N(0, 10.000)$ .
- Objetivo: estimar  $\mu_y = \mathcal{E}(Y)$ .

# Comparaçãõ entre os estimadores razãõ e regressãõ

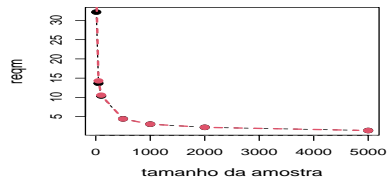
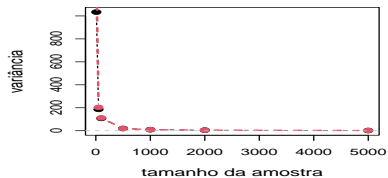
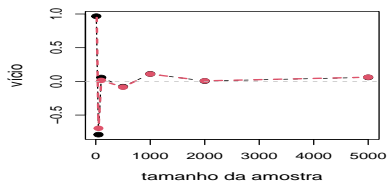
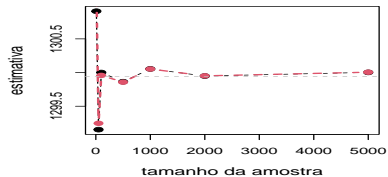
- R= 1.000 amostras foram selecionadas, sob diversos tamanhos  $n = (10, 50, 100, 500, 1.000, 2.000, 5.000)'$ .
- Foram calculados a média, vício, variância e raiz quadrática do erro quadrático médio dos estimadores razãõ e regressãõ (respectivamente em preto e vermelho nos dois grãficos a seguir).
- As linhas em cinza representam o valor verdadeiro de  $\mu_y$  (primeiro grãfico) ou o valor zero (grãficos restantes).



# Cenário 1



## Cenário 2



## Aplicação: Exercício 6.11, Bolfarine & Bussab (2005)

- Uma fábrica de suco de laranja quer estimar quanto um caminhão com ( $\tau_x =$ ) 1.000 kg de laranja produzirá de suco natural. Para isso, selecionaram-se 10 laranjas, com os seguintes resultados:

Laranja	1	2	3	4	5	6	7	8	9	10
peso (g) ( $x$ )	150	130	140	120	160	160	130	170	140	150
suco (ml) ( $y$ )	60	55	50	40	70	60	45	65	55	65

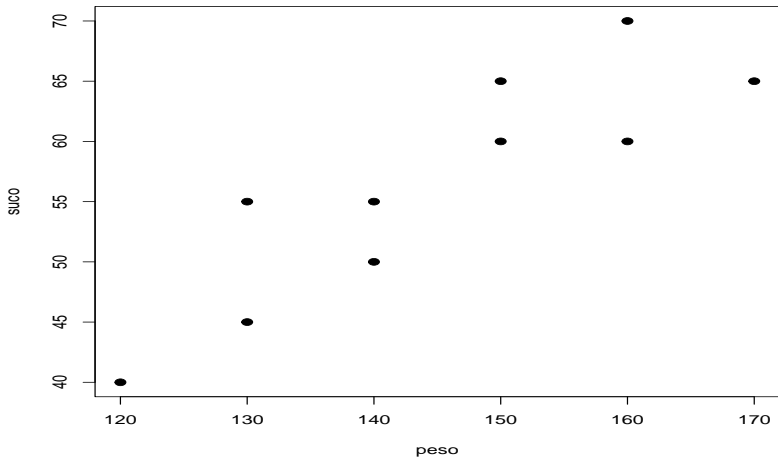
# Comentários

- Consideraremos que a seleção fora feita segundo um PA:  $AAS_s$ .
- Como não dispomos de  $N$  (tamanho da população), e com os recursos vistos até o momento, só poderíamos usar o **estimador razão**.
- Contudo, vamos “estimar”  $N$  através de  $\tilde{N} = \frac{T_x}{\bar{x}} \approx 6897$ , a fim de poder utilizar os estimadores regressão (estes slides) e **expansão**.

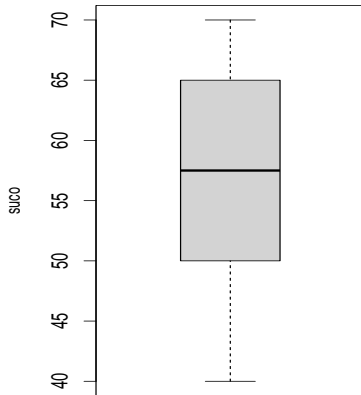
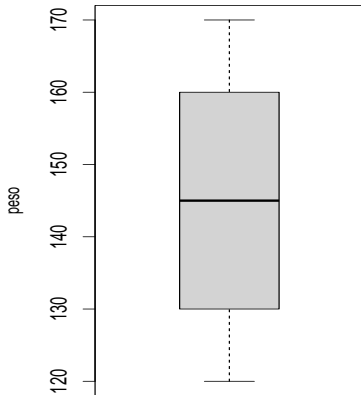
# Medidas resumo

medida resumo	peso	suco
média	145,00	56,50
dp	15,81	9,44
variância	250,00	89,17
cv(%)	10,90	16,71
mínimo	120,00	40,00
mediana	145,00	57,50
máximo	170,00	70,00
curtose	-1,42	-1,27
ca	0,00	-0,30

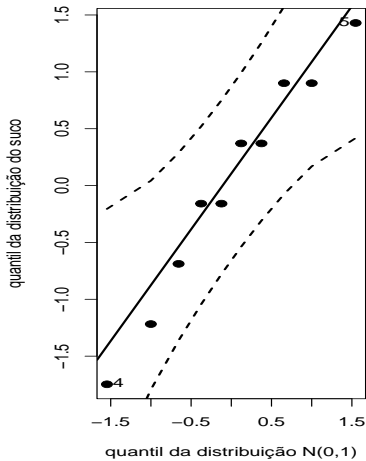
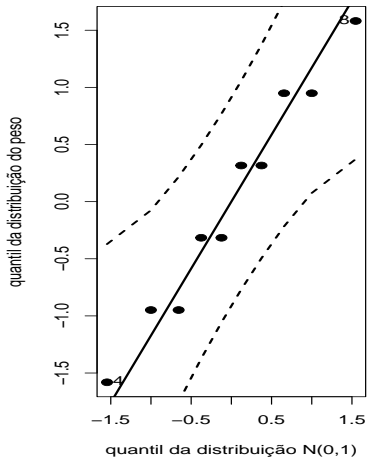
# Gráfico de dispersão entre as variáveis



# Boxplot das variáveis



# QQ plot das variáveis





# Resultados

- Os testes de normalidade de **Kolmogorov-Smirnov (KS)** e **Shapiro - Wilks (SW)** não rejeitaram a normalidade: KS - peso : 0,13 (0,9964); suco : 0,15 (0,9850); SK - peso : 0,97 (0,8486); suco : 0,97 (0,8353).
- A correlação amostral de Pearson (estimativa (erro-padrão)) resultou em 0,874 (0,172). O teste para nulidade da correlação de Pearson resultou em  $t = 5,10$  (0,0009). Estes resultados indicam que as variáveis são significativa (e positivamente) correlacionadas.

# Resultados

- Usando **modelos de regressão apropriados** temos que a relação entre as duas variáveis parecer ser “adequadamente” representável por uma reta passando pela origem, ( $p=0,2340$ ), (modelo de regressão normais lineares homocedásticos). Isso pode indicar um melhor desempenho do estimador razão, em relação aos outros dois.
- Estimativas dos estimadores **razão** e regressão (estes slides) podem ser obtidos usando as funções “**calibrate**” e “**predict**” ([link](#)).

# Resultados

- Em geral, no pacote “[survey](#)” os erros-padrão são estimados usando o método de [Horvitz-Thompson \(HT\)](#) ([link 1](#), [link 2](#), [link 3](#)) que, em geral, são mais robustos que os estimadores usuais (presentes em vários livros, como aqueles que apresentamos ao longo do curso). Isso se aplica, em particular, ao estimador regressão.
- De uma certa forma, as estimativas dos erros-padrão via HT levam em consideração, de uma forma mais ampla, as incertezas associadas à estimação dos parâmetros.
- Vamos apresentar as duas estimativas do erro-padrão para tal estimador.

# Comentários

- Se se fosse utilizar a **modelagem clássica** (distribuição dos dados) uma escolha usual (em nível de Graduação) seria a distribuição normal.
- Com efeito, as análises indicam que tal suposição parece ser razoável para as duas variáveis, (QQ-plots e testes de normalidade).

# Comentários

- Neste caso, análises inferenciais exatas (sob normalidade) são possíveis e, como (também) a presente abordagem (**modelagem probabilística**), utiliza resultados assintóticos, aquela pode ser melhor do que esta.
- Em geral, quanto mais as suposições de uma determinada abordagem forem válidas (para um dados problema) melhor é seu desempenho em relação à abordagens que demandam menos suposições.

# Resultados (estimativas)

- Estimativas dos totais (em ml):

Estimador	Estimativa	EP	IC(95%)
Expansão	389.680,50	20.580,05	[349.344,34 ; 430.016,66]
Razão	389.655,17	10.979,25	[368.136,24 ; 411.174,11]
Regressão	389.646,56	9.984,52	[370.077,26 ; 409.215,85]

- A estimativa do EP de HT (estimador regressão) foi igual a 9.983,91 (próximo do estimador usual, acima).
- As estimativas pontuais são próximas. Contudo, a precisão relativa ao estimador regressão é a maior (o que coaduna com o resultado destes slides, pag. 20).