Qual a proporção de casos de Covid-19 no RS?

Prevalência de Covid-19 no RS com amostra aleatória e informação populacional.

Idéia geral

Considere duas quantidades de interesse:

- N_y: o número de casos já confirmados no RS.
- N_x: o número real de casos no RS.

Autoridades e pesquisadores gostariam de conhecer o número verdadeiro de casos da doença N_x, para uma população de tamanho N. Geralmente governos e órgãos responsáveis por testes conseguem divulgar somente o valor de N_y. Suponha que $N_y \leq N_x$ e que N_y seja conhecido, algumas razões para isso são discutidas em (ref.).

Se todas as quantidades definidas acima fossem observadas poderíamos apresentar os dados na forma de uma tabela 2×2 como essa abaixo.

Casos Reais	Não confirmados	Casos confirmados	$\sum$
Não	???	???	$N - N_x$???
Sim	???	???	N_x???
$\sum$	$N - N_y$	N_y	N

Proporções de casos em populações finitas podem ser representadas por probabilidades. Primeiramente defina as variáveis aleatórias:

- Y: indicadora de caso já confirmado no RS ($Y = 1$).
- X: indicadora de casos real no RS ($X = 1$).

Assim a proporção de casos reais da doença, N_x/N, pode ser escrita como a probabilidade de uma pessoa na população ser um caso real, $\pi = P(X = 1)$. A proporção de casos já confirmados N_y/N, representa a probabilidade de uma pessoa ser um caso confirmado, $\rho = P(Y = 1)$. Suponha que $\rho \leq \pi$ (ref.).

Como calcular π?

1. Estatísticas oficiais?
   Devido a quantidade limitada de testes (em situações de emergência), geralmente os casos são detectados apenas em unidades da população em estado grave que procuram atendimento, pessoas que pertencam a grupos de risco, profissionais de saude, ou algum outro critério de prioridade. O que provavelmente não representará a proporcao real de casos na regiao de interesse.
2. Testar toda a população do RS?
   Inviável! (Impossível seja pelo custo ou tempo, não eficiente, ...)
3. Estimar π através de uma amostra aleatória?
   Se coletarmos uma amostra aleatória de tamanho n da variável X, $\boldsymbol{x} = (x_1, \ldots, x_n)$, podemos calular a proporção amostral

$$\hat{\pi} = \frac{\sum_{u=1}^{n} x_u}{n} = \frac{n_x}{n}.$$

E se também observamos Y para cada unidade?

- Utilizando informações amostrais que representem as contagens no interior da tabela acima conseguimos estimar π com maior eficiência (ref.). Como?

Na amostra teríamos a tabela

Casos Reais	Não confirmados	Casos confirmados	$\sum$
Não	$n - n_x$	0	$n - n_x$
Sim	$n_x - n_y$	n_y	n_x
$\sum$	$n - n_y$	n_y	n

Obs.: Pesquisadores ainda podem ter interesse em quantificar a proporção da diferença entre as duas probabilidades, ou seja, desejam estimar uma quantidade μ tal que $\pi = \mu \times \rho$.

Obs. 2: Nesse caso precisaramos conhecer ρ. Sabemos o valore de ρ? Se assumirmos ρ conhecido (quando o número de casos detectados são precisamente considerados corretos...

Obs. 3: Conhecendo ρ, um estimador para μ pode ser dado por $\hat{\mu} = \frac{\hat{\pi}}{\rho} = \frac{n_x}{n\rho}$, se estimarmos $\hat{\pi} = n_x/n$, a proporção amostral de casos reais.

No caso do RS

Estimativas do governo indicam $\pi =?$. QUal o estimador utilizado? (media amostral?) Assumindo uma popuação de $N = 113.000.00$ habitantes no RS (ref.)... Informacoes oficiais dizem que existem $Y = 348$ casos confirmados na população do RS...

Fase 1 Populacao

Casos Reais	Não confirmados	Casos confirmados	$\sum$
Não	???	0	
Sim	???	???	
$\sum$	111.299.616	384	11.300.000

Amostra

Casos Reais	Não confirmados	Casos confirmados	$\sum$
Não	4.187	0	4.187
Sim	2	0	2
$\sum$	4.189	0	4.189

Amostragem + Dados populacionais

Ao invés de utilizar apenas o número de casos na amostra podemos tornar mais eficiente...

Além do teste, se soubermos o número de unidades na amostra que já foram consideradas caso anteriormente...

Função de verossimilhança

Para o problema acima temos $X \sim Binomial(n, \pi)$ e $Y|X \sim Binomial\left(X, \frac{q}{p}\right)$, assim a distribuição conjunta das variáveis é dada por

$$p(Y, X) = \pi^X (1-\pi)^{1-X} \times (\rho/\pi)^Y \left[1 - (\rho/\pi)\right]^{X-Y}$$

Agora sponha que observamos uma amostra aleatória (simples) de tamanho n da população de interesse, $(\boldsymbol{y}, \boldsymbol{x}) = \{(y_1, x_1), \ldots, (y_n, x_n)\}$. A distribuição conjunta de probabilidade pode ser dada por

$$L(\pi; \boldsymbol{y}, \boldsymbol{x}) = \binom{n}{n_x} \pi^{n_x} (1-\pi)^{n-n_x} \times \binom{n_x}{n_y} (\rho/\pi)^{n_y} \left[1 - (\rho/\pi)\right]^{n_x - n_y}$$

Verossimilhanca em termos de μ...

$$L(\mu; \boldsymbol{y}, \boldsymbol{x}) = \binom{n}{n_x} \pi^{n_x} (1-\pi)^{n-n_x} \times \binom{n_x}{n_y} (\rho/\pi)^{n_y} \left[1 - (\rho/\pi)\right]^{n_x - n_y}$$

- Isso é o mesmo que pós estratificação, raking e calibragem...

Referências

Estimating the proportion of Corona cases with a random sample

- https://grazeconomics.wordpress.com/2020/04/22/estimating-the-proportion-of-corona-cases-with-a-random-sample/