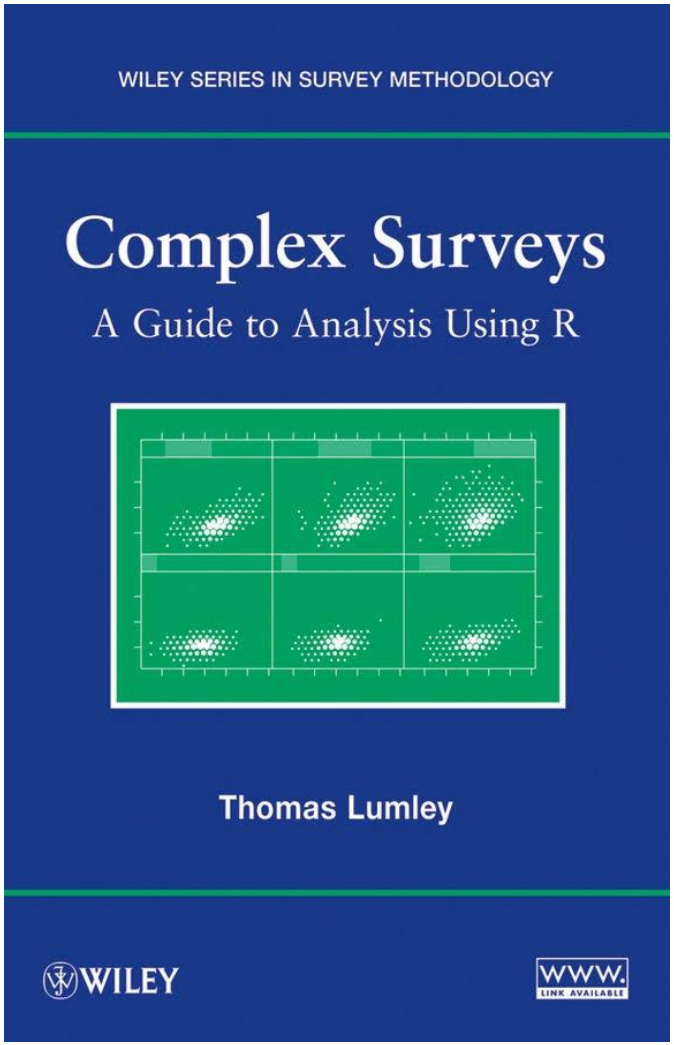
Introdução a análise de dados com pesos amostrais usando o pacote survey

Vanessa Bielefeldt Leotti

Referência principal



Outras referências

- https://djalmapessoa.github.io/adac/index.html
- https://rpubs.com/BragaDouglas/335574

- Outros pacotes
  - PNADcIBGE
  - AmostraBrasil

O que é o survey?

- Um pacote para análise de amostras probabilísticas, com abordagem baseada no **desenho**
- Em geral, trabalhamos em estatística com a abordagem baseada no **modelo**
- Assume-se uma distribuição de probabilidades para os dados
- Na abordagem baseada no desenho, assume-se apenas a aleatoriedade de amostra
- População finita X infinita (modelos teóricos)

Amostras probabilísticas

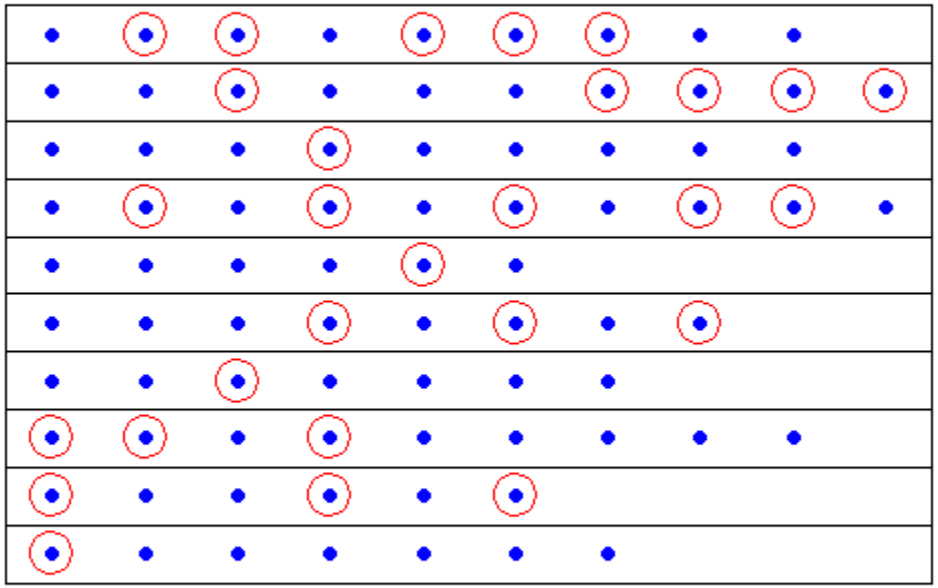
- Cada unidade deve ter uma probabilidade conhecida e não nula de pertencer a amostra
- Essas probabilidades podem ser desiguais
- Um mecanismo aleatório é usado para sortear as unidades respeitando as probabilidades

Alguns tipos de amostras probabilísticas

- Amostragem aleatória simples
  - Todas as unidades tem a mesma probabilidade
- Amostragem estratificada
  - Divide-se a população em grupos, em cada grupo faz-se uma AAS
- Amostragem por conglomerados
  - Divide-se a população em grupos, faz-se uma AAS em grupos
- Amostragens complexas ou em estágios múltiplos
  - Combinam técnicas de amostragem

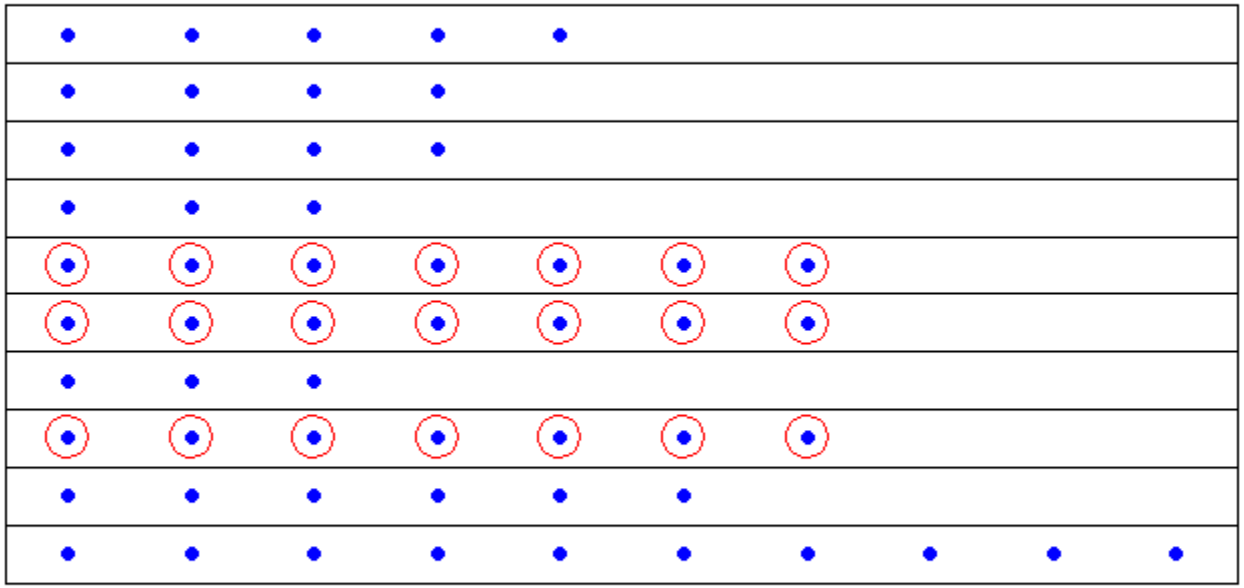
Estratificada

```
library(animation)
sample.strat()
```



Conglomerados

```
library(animation)
sample.cluster()
```



Pesos amostrais

- Se pegarmos uma amostra aleatória de 10% dos habitantes de POA, então cada pessoa teria chance de 1 em cada 10 de ser selecionada.
- Cada pessoa amostrada em POA representaria 10 porto alegrenses.
- Se por acaso, 5000 pessoas amostradas tiverem hipertensão, significa que POA tem 50000 hipertensos (5000 * 10).
- Por definição o peso amostral é o inverso da probabilidade de sorteio de uma unidade.

Porque é interessante saber analisar dados com pesos amostrais?

- Grandes pesquisas nacionais usam estratégias de amostragem com pesos desiguais
  - PNAD
  - VIGITEL
  - PME
  - PNS
  - ...

Vamos então ao survey!

- Há bancos de dados exemplo dentro do pacote

```
library(survey)
data(api)
?api
```

- API = Índice de Performance Acadêmica para escolas da califórnia
- apipop é o banco do censo,
- apisrs é uma AAS,
- apiclus1 é uma amostragem por conglomerados de um estágio, onde os conglomerados são os distritos,
- apistrat é uma amostragem estratificada por tipo de escola,
- apiclus2 é uma amostragem por conglomerados de dois estágios, com sorteio de escolas dentro de distritos.

API

```
> # Suponha que o interesse é inferir sobre
  o número de estudantes matriculados
>
> # Parâmetros
> total = sum(apipop$enroll, na.rm=T)
> total
[1] 3811472
>
> media = mean(apipop$enroll, na.rm=T)
> media
[1] 619.0469
```

API

- Agora vamos analisar a AAS
- O primeiro passo é descrever o desenho

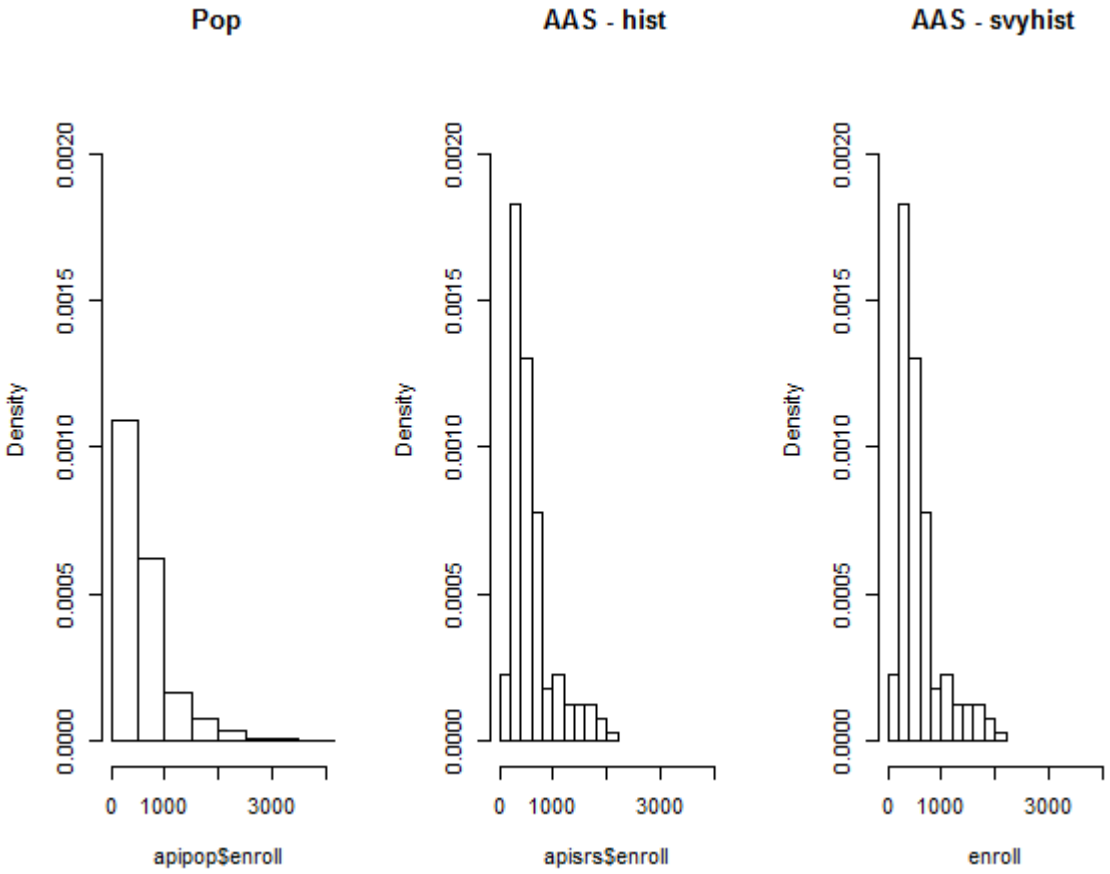
```
aas = svydesign(id=~1, fpc=~fpc,
  data=apisrs)
```

- id=~1 significa que não há clusters (pode ser 0 também)
- fpc=~fpc aplica a correção de população finita, mostrando qual variável do banco tem o tamanho da população
- data define o banco de dados

API

```
> # Após isso aplica-se qualquer função de análise
> svytotal(~enroll, aas)
          total     SE
enroll 3621074 169520
> svymean(~enroll, aas)
          mean     SE
enroll 584.61 27.368
> confint(svytotal(~enroll, aas))
          2.5 %  97.5 %
enroll 3288822 3953327
> confint(svymean(~enroll, aas))
          2.5 %  97.5 %
enroll 530.969 638.251
```

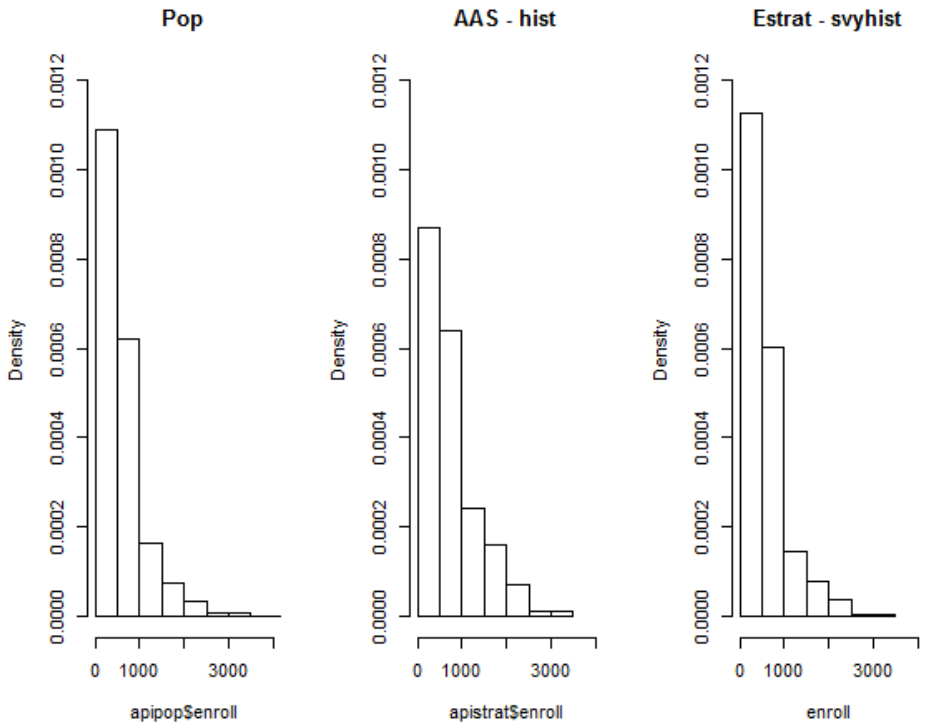
```
> # Há funções de gráficos
> par(mfrow=c(1,3))
> hist(apipop$enroll, main = "Pop", freq=F, ylim=c(0,0.0022),
xlim=c(0,4000))
> hist(apisrs$enroll, main = "AAS - hist", freq=F, ylim=c(0,0.0022),
xlim=c(0,4000))
> svyhist(~enroll, aas, main="AAS - svyhist", ylim=c(0,0.0022),
xlim=c(0,4000))
```



```
> # Agora a análise da amostra estratificada
> # A variável estratificadora foi tipo de escola (E, H, M)
> # Distribuição dos estratos na população
> table(apipop$stype)

   E    H    M
4421  755 1018
>
> # 200 escolas foram selecionadas, 100 do tipo E, 50 do tipo H e 50
do tipo M
> # Então os pesos são
> table(apipop$stype)/c(100,50,50)

    E     H     M
44.21 15.10 20.36
>
> # São os mesmos presentes no banco apistrat
> table(apistrat$stype, apistrat$pw)

  15.1000003814697 20.3600006103516 44.2099990844727
E                0                0              100
H               50                0                0
M                0               50                0
```

```
> # Vejamos se são os mesmos que o survey calcula
> estrat = svydesign(id=~1, fpc=~fpc, strata=~stype, data=apistrat)
> table(1/estrat$prob)

 15.1 20.36 44.21
   50    50   100
> # Aqui fpc é o número de escolas por estrato
> table(apistrat$stype, apistrat$fpc)

    755 1018 4421
  E   0    0  100
  H  50    0    0
  M   0   50    0
```

Histograma e efeito de delineamento




```
> svymean(~enroll, estrat, deff=T)
             mean     SE  DEff
enroll 595.282  18.509 0.362
```

Análise de dados da PNAD

```
> ##### Agora exemplos mais interessantes
> # Pacote para obter dados da PNAD
> library(PNADcIBGE)
>
> pnad181 = get_pnadc(year=2018, quarter=1)
trying URL
'ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Trimestral/Microdados/Documentacao/Dicionario_e_input_20170517.zip'
downloaded 50 KB

trying URL
'ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Trimestral/Microdados/2018/PNADC_012018.zip'
downloaded 20.5 MB

|======================================================================| 100%  248 MB
Warning message:
In factor(as.numeric(unlist(data_pnadc[varsf[i]])), levels = suppressWarnings(as.numeric(unlist(dictionary %>%  :
  NAs introduced by coercion
>
> # Veja que é importado um objeto pronto para o survey
> class(pnad181)
[1] "survey.design2" "survey.design"
>
> # Então basta analisar!
```

Renda (Média geral e por sexo)

```
> # Média de renda
> svymean(~VD4020, pnad181, na.rm=T)
          mean     SE
VD4020 2374.3 28.595
>
> # Renda X Sexo
> svyby(~VD4020, ~V2007, pnad181, svymean, na.rm=T)
        V2007   VD4020       se
Homem   Homem 2617.758 35.75846
Mulher Mulher 2051.593 23.66315
```

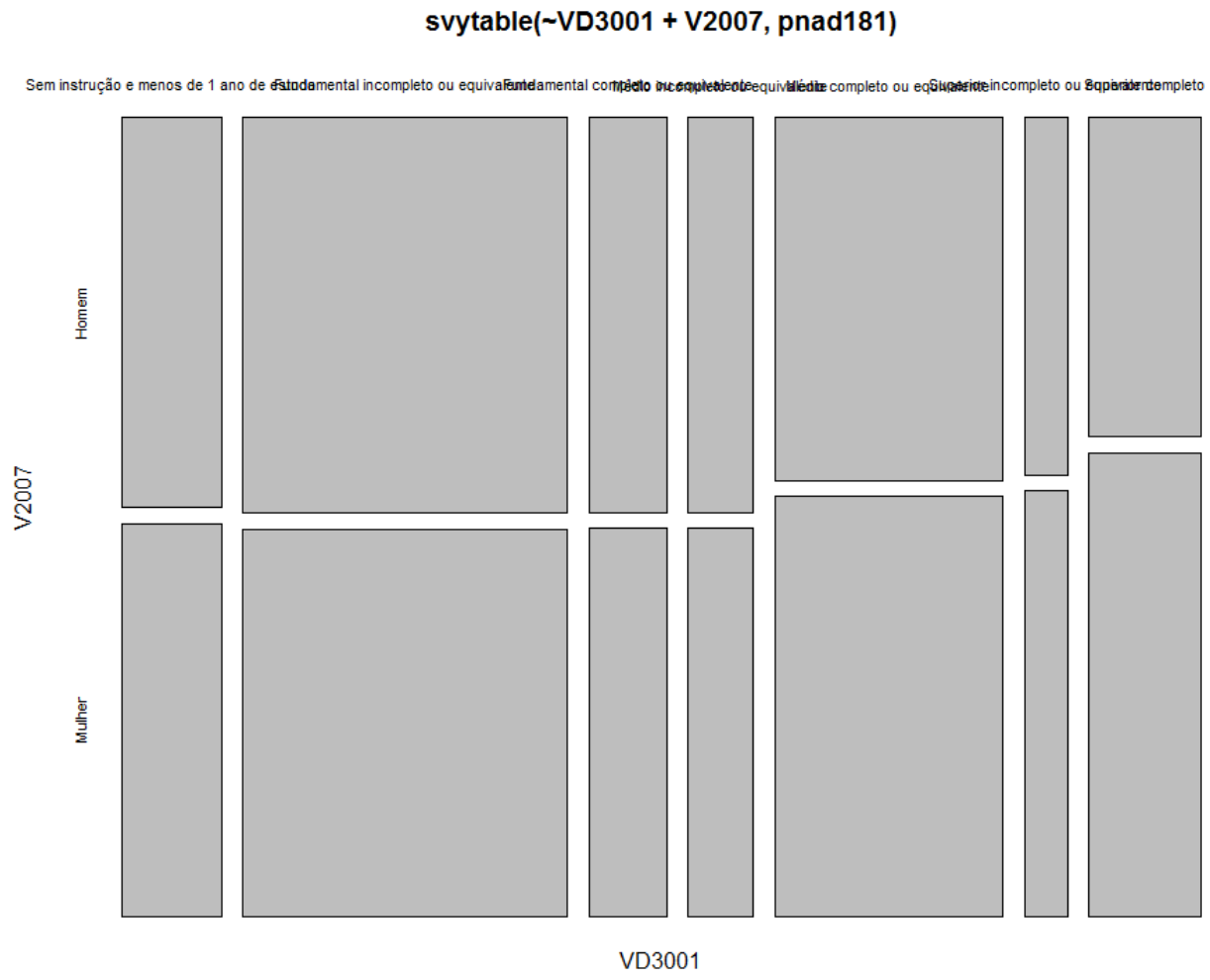
Renda (Média por sexo e raça)

```
> # Estimando médias por sexo e raça
> svyby(~VD4020, ~V2007 + V2010, pnad181, svymean, na.rm=T)
                   V2007    V2010   VD4020          se
Homem.Branca       Homem   Branca 3480.572   68.10212
Mulher.Branca     Mulher   Branca 2589.140   40.70402
Homem.Preta        Homem    Preta 1883.017   28.82112
Mulher.Preta      Mulher    Preta 1531.636   27.79774
Homem.Amarela      Homem  Amarela 5535.998  592.46351
Mulher.Amarela    Mulher  Amarela 3850.126  360.85029
Homem.Parda        Homem    Parda 1899.957   21.04574
Mulher.Parda      Mulher    Parda 1542.320   15.42830
Homem.Indígena     Homem Indígena 2216.592  282.63411
Mulher.Indígena   Mulher Indígena 1829.717  206.48531
Homem.Ignorado     Homem Ignorado 7678.056 1817.50348
Mulher.Ignorado   Mulher Ignorado 3136.912  919.49404
```

Instrução X Sexo

```
> # Table cruzada de instrução por sexo
> svytable(~VD3001 + V2007, pnad181)
                                                V2007
VD3001                                             Homem   Mulher
  Sem instrução e menos de 1 ano de estudo      10240018 10303346
  Fundamental incompleto ou equivalente         33674608 32983898
  Fundamental completo ou equivalente            8014319  7865681
  Médio incompleto ou equivalente                6711329  6594215
  Médio completo ou equivalente                 21724307 25120547
  Superior incompleto ou equivalente             4036454  4804029
  Superior completo                              9349114 13585442
> par(mfrow=c(1,1))
> plot(svytable(~VD3001 + V2007, pnad181))
> # Teste qui-quadrado Instrução X sexo
> svychisq(~VD3001 + V2007, pnad181, statistic="Chisq")

        Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~VD3001 + V2007, pnad181, statistic = "Chisq")
X-squared = 2350.4, df = 6, p-value < 2.2e-16
```

Instrução X Sexo



Instrução X Sexo

```
> svyby(~V2007, ~VD3001, pnad181, svymean, na.rm = T)
                                                 V2007Homem V2007Mulher
Sem instrução e menos de 1 ano de estudo          0.4984587   0.5015413
Fundamental incompleto ou equivalente             0.5051810   0.4948190
Fundamental completo ou equivalente               0.5046800   0.4953200
Médio incompleto ou equivalente                   0.5044010   0.4955990
Médio completo ou equivalente                     0.4637501   0.5362499
Superior incompleto ou equivalente                0.4565875   0.5434125
Superior completo                                 0.4076431   0.5923569
```

Outras inferências

```
> # Teste-t Renda X Sexo
> svyttest(VD4020 ~ V2007, pnad181)

        Design-based t-test

data:  VD4020 ~ V2007
t = -22.924, df = 14503, p-value < 2.2e-16
alternative hypothesis: true difference in mean is not equal to 0
95 percent confidence interval:
 -614.5712 -517.7576
sample estimates:
difference in mean
         -566.1644
```

Outras inferências

```
> # Modelo de regressão para renda ajustada por escolaridade, sexo e idade
> reg = svyglm(VD4020 ~ V2007 + VD3001 + V2009, pnad181)
> summary(reg)

Call:
svyglm(formula = VD4020 ~ V2007 + VD3001 + V2009, pnad181)

Survey design:
survey::postStratify(design = data_pre, strata = ~posest, population = popc.types)

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              -1056.389     66.541  -15.88   <2e-16 ***
V2007Mulher                              -1081.440     32.216  -33.57   <2e-16 ***
VD3001Fundamental incompleto ou equivalente  703.674     25.387   27.72   <2e-16 ***
VD3001Fundamental completo ou equivalente   1209.803     31.023   39.00   <2e-16 ***
VD3001Médio incompleto ou equivalente       1446.482     48.862   29.60   <2e-16 ***
VD3001Médio completo ou equivalente         1808.770     36.956   48.94   <2e-16 ***
VD3001Superior incompleto ou equivalente    2490.626     55.850   44.59   <2e-16 ***
VD3001Superior completo                     5366.911    114.891   46.71   <2e-16 ***
V2009                                         43.489      1.318   32.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 6197161)

Number of Fisher Scoring iterations: 2
```

Outras funções

- svyciprop para intervalos de confiança de proporções
- svyquantile para quantis
- svysurvreg ou svycoxph para modelos de sobrevida