

# MAT02025 - Amostragem 1

AAS: o pacote survey

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, 2023

## Exemplo

## Exemplo



- ▶ As assinaturas de uma petição foram coletadas em 676 folhas.
- ▶ Cada folha tinha espaço suficiente para 42 assinaturas, mas em muitas folhas um número menor de assinaturas foi coletado.
- ▶ O número de assinaturas por folha foi contado em uma amostra aleatória de 50 folhas (cerca de 7% da amostra), com os resultados apresentados na tabela a seguir.

$Y_i$	$n_i$	$Y_i$	$n_i$
42	23	14	1
41	4	11	1
36	1	10	1
32	1	9	1
29	1	7	1
27	2	6	3
23	1	5	2
19	1	4	1
16	2	3	1
15	2	-	-

## Exemplo

- ▶ O objetivo do levantamento amostral é estimar o **número total de assinaturas** para a petição e os limites de confiança de 80%.
- ▶ A **unidade de amostragem é a folha** e as **observações  $Y_i$  são os números de assinaturas por folha**.
- ▶ Observe que a distribuição original parece estar longe da normal, a maior frequência estando na extremidade superior.
- ▶ No entanto, há razões para acreditar, por experiência, que as médias das amostras de 50 unidades são aproximadamente normalmente distribuídas.

## Exemplo

```
N <- 676
n <- 50
id <- 1:n
yi <- c(3, 4, 5, 6, 7, 9, 10, 11, 14, 15,
        16, 19, 23, 27, 29, 32, 36, 41, 42)
ni <- c(1, 1, 2, 3, 1, 1, 1, 1, 1, 2, 2,
        1, 1, 2, 1, 1, 1, 4, 23)
y <- rep(x = yi, times = ni)
df.assinatura <- data.frame(id, y, cpf = N)
head(df.assinatura)
```

```
##   id y  cpf
## 1  1 3 676
## 2  2 4 676
## 3  3 5 676
## 4  4 5 676
## 5  5 6 676
## 6  6 6 676
```

## Exemplo

```
# Estimativa da média  
(Y.barra <- mean(df.assinatura$y))
```

```
## [1] 29.42
```

```
# Estimativa do total  
(YT.chapeu <- N * Y.barra)
```

```
## [1] 19887.92
```

```
# Estimativa da variância de Y  
(s2 <- var(x = df.assinatura$y)) # ?var
```

```
## [1] 228.9833
```

```
# Estimativa do desvio padrão de Y  
(s <- sd(x = df.assinatura$y)) # ?sd
```

```
## [1] 15.13219
```

```
# s <- sqrt(s2)
```

```
(f <- n/N) # fração de amostragem
```

## Exemplo

```
## [1] 0.0739645
```

```
# Estimativa do erro padrão do total  
(s.YT.chapeu <- ((N * s)/sqrt(n)) * sqrt(1 - f))
```

```
## [1] 1392.122
```

```
# IC 80% para o total de assinaturas  
# (utilizando a distribuição normal)  
YT.chapeu + c(-1, 1) * qnorm(p = 0.9) * s.YT.chapeu
```

```
## [1] 18103.84 21672.00
```

```
round(YT.chapeu + c(-1, 1) * qnorm(p = 0.9) * s.YT.chapeu)
```

```
## [1] 18104 21672
```

```
# IC 80% para o total de assinaturas  
# (utilizando a distribuição t)  
round(YT.chapeu + c(-1, 1) * qt(p = 0.9, df = n - 1) * s.YT.chapeu)
```

```
## [1] 18079 21696
```

# Levantamentos amostrais complexos no R



## Why are surveys special?

---

- Probability samples come with a lot of meta-data that has to be linked correctly to the observations, so software was specialized
- Interesting data sets tend to be large, so hardware was specialized
- Design-based inference literature is largely separate from rest of statistics, doesn't seem to have a unifying concept such as likelihood to rationalize arbitrary choices.

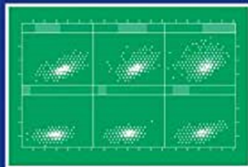
In part, too, the specialized terminology has formed barriers around the territory. At one recent conference, I heard a speaker describe methods used in his survey with the sentence, We used BRR on a PPS sample of PSUs, with MI after NRFU

—Sharon Lohr, The American Statistician, 5/2004



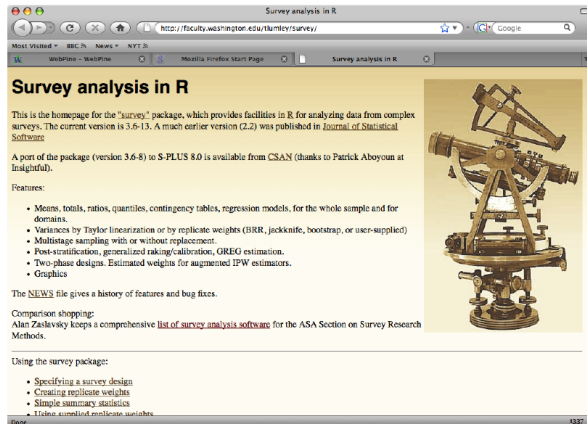
# Complex Surveys

A Guide to Analysis Using R



Thomas Lumley

# R Survey package



The screenshot shows a web browser window titled "Survey analysis in R". The address bar displays the URL <http://faculty.washington.edu/tlumley/survey/>. The page content includes the title "Survey analysis in R", a description of the package, its features, and a list of links for comparison shopping and using the package. An image of a surveying instrument is also visible on the right side of the page.

**Survey analysis in R**

This is the homepage for the "survey" package, which provides facilities in R for analyzing data from complex surveys. The current version is 3.6-13. A much earlier version (2.2) was published in *Journal of Statistical Software*.

A port of the package (version 3.6-8) to S-PLUS 8.0 is available from CSAN (thanks to Patrick Aboyoun at Insightful).

Features:

- Means, totals, ratios, quantiles, contingency tables, regression models, for the whole sample and for domains.
- Variances by Taylor linearization or by replicate weights (BRR, jackknife, bootstrap, or user-supplied)
- Multistage sampling with or without replacement.
- Post-stratification, generalized raking/calibration, GREG estimation.
- Two-phase designs. Estimated weights for augmented IPW estimators.
- Graphics

The [NEWS](#) file gives a history of features and bug fixes.

Comparison shopping:  
Alan Zaslavsky keeps a comprehensive [list of survey analysis software](#) for the ASA Section on Survey Research Methods.

Using the survey package:

- [Specifying a survey design](#)
- [Creating replicate weights](#)
- [Simple summary statistics](#)
- [Using raked and replicate weights](#)

<http://faculty.washington.edu/tlumley/survey/>

## Design principles

---

- Ease of maintenance and debugging by code reuse
- Speed and memory use not initially a priority: don't optimize until there are real use-cases to profile.
- Rapid release, so that bugs and other infelicities can be found and fixed.
- Emphasize features that look like biostatistics (regression, calibration, survival analysis, exploratory graphics)

## Intended market

---

- Methods research (because of programming features of R)
- Teaching (because of cost, use of R in other courses)
- Secondary analysis of national surveys (regression features, R is familiar to non-survey statisticians)
- Two-phase designs in epidemiology

## Outline

---

- Describing survey designs: `svydesign()`
- Database-backed designs
- Summary statistics: mean, total, quantiles, design effect
- Tables of summary statistics, domain estimation.
- Graphics: histograms, hexbin scatterplots, smoothers.
- Regression modelling: `svyglm()`
- Multiply-imputed data
- Calibration of weights: `postStratify()`, `calibrate()`

## Objects and Formulas

---

Collections of related information should be kept together in an object. For surveys this means the data and the survey meta-data.

The way to specify variables from a data frame or object in R is a formula

$$\sim a + b + I(c < 5*d)$$

The survey package **always** uses formulas to specify variables in a survey data set.

## Basic estimation ideas

---

Individuals are sampled with known probabilities  $\pi_i$  from a population of size  $N$  to end up with a sample of size  $n$ . The population can be a full population or an existing sample such as a cohort.

We write  $R_i = 1$  if individual  $i$  is sampled,  $R_i = 0$  otherwise

The design-based inference problem is to estimate what any statistic of interest would be if data from the whole population were available.



## Basic estimation ideas

---

For a population total this is easy: an unbiased estimator of

$$T_X = \sum_{i=1}^N x_i$$

is

$$\hat{T}_X = \sum_{i: R_i=1} \frac{1}{\pi_i} X_i$$

Standard errors follow from formulas for the variance of a sum:  
main complication is that we do need to know  $\text{cov}[R_i, R_j]$ .

## Retornando ao exemplo das assinaturas

- ▶ Embora amostras aleatórias simples não exijam *software* especializado para análise, elas fornecem um exemplo simples para a introdução do pacote *survey*.
- ▶ A **primeira etapa** na análise de dados é descrever o delineamento (desenho, *design*) para o R.
- ▶ A função `svydesign` pega essa descrição e a adiciona ao conjunto de dados para produzir um **objeto de delineamento** do levantamento por amostragem.
- ▶ O objeto de delineamento do levantamento é então usado em **todas as análises**.
- ▶ O envelopamento das informações do delineamento e dos dados em um único objeto garante que as informações do delineamento não possam ser inadvertidamente separadas ou usadas com o conjunto de dados incorreto.

## Retornando ao exemplo das assinaturas

```
# install.packages("survey")
library(survey)

# O objeto design
(ass.des <- svydesign(ids = ~1,
                    fpc = ~cpf,
                    data = df.assinatura))
```

```
## Independent Sampling design
## svydesign(ids = ~1, fpc = ~cpf, data = df.assinatura)
summary(ass.des)
```

```
## Independent Sampling design
## svydesign(ids = ~1, fpc = ~cpf, data = df.assinatura)
## Probabilities:
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07396 0.07396 0.07396 0.07396 0.07396 0.07396
## Population size (PSUs): 676
## Data variables:
## [1] "id"  "y"   "cpf"
```

## Retornando ao exemplo das assinaturas

```
# Estimativa da média  
svymean(x = ~y, design = ass.des)
```

```
##      mean      SE  
## y 29.42 2.0594
```

```
# Estimativa do total  
svytotal(x = ~y, design = ass.des)
```

```
##      total      SE  
## y 19888 1392.1
```

```
# IC 80% para o total de assinaturas  
# (utilizando a distribuição normal)  
confint(svytotal(x = ~y, design = ass.des),  
        level = 0.8)
```

```
##          10 %   90 %  
## y 18103.84 21672
```

## Retornando ao exemplo das assinaturas

```
# IC 80% para o total de assinaturas  
# (utilizando a distribuição tl)  
confint(svytotal(x = ~y, design = ass.des),  
        level = 0.8, df = survey::degf(ass.des))
```

```
##          10 %      90 %  
## y 18079.46 21696.38
```

```
# IC 95% para o total de assinaturas  
confint(svytotal(x = ~y, design = ass.des),  
        level = 0.95)
```

```
##          2.5 %    97.5 %  
## y 17159.41 22616.43
```

**Um novo exemplo**

# Índice de Desempenho Acadêmico

- ▶ O **Índice de Desempenho Acadêmico da Califórnia** (*Academic Performance Index*, API) é calculado a partir de testes padronizados administrados a alunos em escolas da Califórnia.
- ▶ Além dos dados de desempenho acadêmico das escolas, há uma ampla gama de variáveis socioeconômicas disponíveis.

```
data(api)

# View/apisrs
# ?apisrs

# O objeto design
api.des <- svydesign(id = ~1,
                    fpc = ~fpc,
                    data = apisrs)

# Estimativa do total de alunos
# matriculados
svytotal(x = ~enroll, design = api.des)
```

# Índice de Desempenho Acadêmico

```
##           total      SE
## enroll 3621074 169520
# Estimativa da média de alunos
# matriculados por escola
svymean(x = ~enroll, design = api.des)
```

```
##           mean      SE
## enroll 584.61 27.368
```



# Índice de Desempenho Acadêmico

- ▶ A população total estimada é de 3,6 milhões de alunos matriculados, com erro padrão de 169.000; o tamanho médio estimado da escola é cerca de 585, com um erro padrão de 27.
- ▶ Os **valores verdadeiros (na população)** são 3,8 milhões e 619, respectivamente, portanto, os erros padrões fornecem uma representação precisa da incerteza nas estimativas.
- ▶ A correção da população finita tem muito pouco impacto sobre essas estimativas e podemos seguramente tê-la ignorado.

# Índice de Desempenho Acadêmico

- ▶ Se o tamanho da população não for especificado, é necessário especificar as **probabilidades de amostragem** ou **pesos de amostragem**.
- ▶ A variável  $p_w$  no conjunto de dados contém o peso de amostragem,  $6194/200 = 30,97$ .
- ▶ A seguir, é mostrado o impacto da omissão do tamanho da população.
- ▶ Quando o objeto de *design* é impresso, a falta de informações sobre o tamanho da população é indicada por **“com reposição”** na saída.
- ▶ A média e o total estimados são iguais, mas os erros padrões são ligeiramente maiores.

# Índice de Desempenho Acadêmico

```
# O objeto design
api.des2 <- svydesign(id = ~1,
                    weights = ~pw,
                    data = apisrs)

# Estimativa do total de alunos
# matriculados
svytotal(x = ~enroll, design = api.des2)
```

```
##           total      SE
## enroll 3621074 172325
```

```
# Estimativa da média de alunos
# matriculados por escola
svymean(x = ~enroll, design = api.des2)
```

```
##           mean      SE
## enroll 584.61 27.821
```

## Para casa

- ▶ Utilize os dados do objeto `api` e, utilizando um esquema de amostragem simples sem reposição, construa intervalos de confiança de 90%, 95% e 99% para o número total e médio de alunos matriculados.
  - ▶ Interprete os seus resultados.

## Próxima aula

- ▶ Amostragem aleatória simples com reposição.

# Por hoje é só!

Bons estudos!

