

Webscraping Básico

Benilton Carvalho & Guilherme Ludwig

Webscraping

- É possível "raspar" (*scrape*) informações de páginas da internet e guardá-las em um banco de dados para análise posterior;
- Esta prática é chamada de *webscraping*;
- Utilizaremos os pacotes:
 - tidyverse: que disponibiliza o operador %>%;
 - rvest: que simplifica algumas operações dos pacotes xml2 e httr;
 - RSQLite: que permite o uso do SQLite dentro do R.

```
library(tidyverse)
library(rvest)
library(RSQLite)
```

Idéias

Uma página da web é um documento que pode ser exibido por um navegador. Estes documentos normalmente exibem resultados de consultas à bancos de dados, que são nosso principal interesse nesta disciplina. Em geral:

- Páginas simples podem ser acessadas através do R com o pacote `rvest`.
- Páginas dinâmicas que exijam autenticação do usuário, na forma de *cookies*, podem ser acessadas por meio do pacote `httr`.
- Nosso objetivo é coletar dados com o `rvest` e armazená-los em um banco de dados.

HTML

É preciso o conhecimento de HTML! Em geral, páginas HTML são texto estruturado, interpretado pelo navegador. Veja exemplos em:

https://www.w3schools.com/html/html_basic.asp

```
<!DOCTYPE html>
<html>
<body>

<h1>Um título</h1>
<p>Um parágrafo</p>

<a href="http://www.uol.com.br">Link para o site do UOL</a>

</body>
</html>
```

Para o rvest, os itens `<h1>`, `<p>`, `` e `<a>` são nós (em inglês, *node*). O nó tipo `h1` é um cabeçalho, o `p` é parágrafo, enquanto o `img` é imagem e o `a` indica um link.

Ainda sobre HTML

Um nó `table` (rvest) define tabelas em HTML.

```
<!DOCTYPE html>
<html>
<body>

  <table>
    <tr>
      <th>Curso</th>
      <th>Código</th>
    </tr>
    <tr>
      <td>Estatística</td>
      <td>02</td>
    </tr>
    <tr>
      <td>Matemática</td>
      <td>01</td>
    </tr>
  </table>

</body>
</html>
```

Ainda sobre HTML

Um nó `li` (rvest) define listas em HTML.

```
<!DOCTYPE html>
<html>
<body>

<ul>
  <li>Café</li>
  <li>Chá</li>
  <li>Leite</li>
</ul>

</body>
</html>
```

Exemplo: wikipedia

A *wikipedia* é particularmente interessante para scraping, pois ela possui muitas páginas com listas, de onde podemos começar nossas buscas. Por exemplo,

https://en.wikipedia.org/wiki/List_of_statisticians

Podemos estar interessados em compilar uma lista com nome, *alma mater*, data de nascimento (e local), e data de falecimento (caso já tenha falecido) de estatísticos famosos.


Lista de Estatísticos

W List of statisticians - Wikipedia

← → 🔒 https://en.wikipedia.org/wiki/List_of_statisticians

⋮ 🔒 ⭐ ↺ 🏠 ⬇ 🔍 Pesquisar

👤 Not logged in Talk Contributions Create account Log in



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

[Interaction](#)
[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

[Tools](#)
[What links here](#)
[Related changes](#)
[Upload file](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Wikidata item](#)
[Cite this page](#)

[Print/export](#)
[Create a book](#)
[Download as PDF](#)
[Printable version](#)

[Languages](#)

Article Talk

Read Edit View history

List of statisticians

From Wikipedia, the free encyclopedia

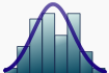
This list of **statisticians** lists people who have made notable contributions to the theories or application of **statistics**, or to the related fields of probability or machine learning. Also included are **actuaries** and **demographers**.

Contents: A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V · W · Y · Z · See also · External links

A [\[edit\]](#)


- Aalen, Odd Olai (1947–1987)
- Abbott, Edith (1876–1957)
- Abelson, Robert P. (1928–2005)
- Abramovitz, Moses (1912–2000)
- Achenwall, Gottfried (1719–1772)
- Adelstein, Abraham Manie (1916–1992)
- Ahsan, Riaz (1951–2008)
- Aitchison, Beatrice (1908–1997)
- Aitchison, John (1926–)
- Aitken, Alexander (1895–1967)
- Akaike, Hirotugu (1927–2009)
- Ali, Mir Masoom (1937–)
- Allen, R. G. D. (1906–1983)
- Allison, David B.
- Altman, Doug (1948–)
- Amemiya, Takeshi (1938–)
- Anderson, Oskar (1887–1960)
- Anderson, Theodore Wilbur
- Anscombe, Francis (1918–2001)

Statistics



[Outline](#) · [Statisticians](#) · [Glossary](#) · [Notation](#) · [Journals](#) · [Lists of topics](#) · [Articles](#) · [Portal](#) · [Category](#)

V · T · E



22:46

POR

21/10/2018

[illegible]

Tabela de Interesse

Wikipedia - George E. P. Box

George E. P. Box

From Wikipedia, the free encyclopedia

For the ice hockey player, see [George Box \(ice hockey\)](#).

George Edward Pelham Box FRS⁽¹⁾ (16 October 1919 – 26 March 2013) was a British statistician; who worked in the areas of **quality control**, **time-series analysis**, **design of experiments**, and **Bayesian inference**. He has been called "one of the great statistical minds of the 20th century".^{[3][4][5][6]}

Contents [hide]

- Education and early life
- Career and research
- Awards and honours
- Personal life
- References
- External links

Education and early life [edit]

George Box

HTML Inspector:

```
<a class="mw-jump-link" href="#p-search">Jump to search</a>
<div id="mw-content-text" class="mw-content-ltr" dir="ltr" lang="en">
  <div class="mw-parser-output">
    <div class="hatnote navigation-not-searchable" role="note"></div>
    <p class="mw-empty-elt"></p>
    <table class="infobox biography vcard" style="width:22em">
      <tbody>
        <tr></tr>
        <tr>
          <td colspan="2" style="text-align:center"></td>
        </tr>
      </tbody>
    </table>
  </div>
</div>
```

Styles:

```
element {
  text-align: center;
}
.infobox td, .infobox th {
  vertical-align: top;
  text-align: left;
}
Inherited from table
.infobox {
  border-spacing: 3px;
  color: black;
}
```

Layout:

```
border-spacing: 3px 3px;
color: rgb(0, 0, 0);
direction: ltr;
font-family: sans-serif;
font-size: 12.3167px;
```

SelectorGadget

Uma ferramenta recomendada pelo rvest é o chamado SelectorGadget (<https://selectorgadget.com/>), que mostra o nome de um "selector" em CSS. Há uma extensão para o navegador Chrome que permite que você use o SelectorGadget em qualquer página.

Com o selector correto, você pode acessá-lo usando `html_nodes()`. Selectors interessantes incluem `"table.<nome>"` e `"li"`. É preciso inspecionar as páginas de interesse caso a caso.

Usando SelectorGadget (Chrome)

W George E. P. Box - Wikipedia x +

← → ↻ ⓘ https://en.wikipedia.org/wiki/George_E._P._Box

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

George E. P. Box

From Wikipedia, the free encyclopedia

For the ice hockey player, see [George Box \(ice hockey\)](#).

George Edward Pelham Box FRS^[1] (18 October 1919 – 28 March 2013) was a British statistician, who worked in the areas of [quality control](#), [time-series analysis](#), [design of experiments](#), and [Bayesian inference](#). He has been called "one of the great statistical minds of the 20th century".^{[3][4][5][6]}

Contents [hide]

- [1 Education and early life](#)
- [2 Career and research](#)
- [3 Awards and honours](#)
- [4 Personal life](#)
- [5 References](#)
- [6 External links](#)

Education and early life [edit]

He was born in [Gravesend, Kent](#), England. Upon entering university he began to study [chemistry](#), but was called up for service before finishing. During [World War II](#), he performed experiments for the [British Army](#) exposing small animals to poison gas. To analyze the results of his experiments, he taught himself statistics from available texts. After the war, he enrolled at [University College London](#) and obtained a bachelor's degree in mathematics and statistics. He received a [PhD](#) from the [University of London](#) in 1953, under the supervision of [Egon Pearson](#).^{[2][7]}

Career and research [edit]

From 1948 to 1956, Box worked as a statistician for [Imperial Chemical Industries](#) (ICI). While at ICI, he took a leave of absence for a year and served as a visiting professor at the North Carolina State University at Raleigh (now [North Carolina State University](#)). He later went to [Princeton University](#) where he served as Director of the Statistical Research Group.

In 1960, Box moved to the [University of Wisconsin–Madison](#) to create the [Center for Quality and Productivity Improvement](#) at the University of Wisconsin–Madison in 1994. Box officially retired in 1999, becoming an emeritus professor.

George Box



Born 18 October 1919
[Gravesend, Kent, England](#)

Died 28 March 2013 (aged 93)
[Madison, Wisconsin](#)

Residence United Kingdom, United States

Alma mater [University College London](#)

Known for "All models are wrong"
[Response-surface methodology](#)
[Box–Jenkins method](#)
[Box–Cox transformation](#)

.vcard Clear (1) Toggle Position XPath ? X

23:31 21/10/2018

Tabela de interesse

```
url = "https://en.wikipedia.org/wiki/George_E._P._Box"
webpage <- read_html(url)
```

```
table <- webpage %>%
  html_node(".vcard") %>%
  html_table(header = FALSE) %>%
  as_tibble()
```

```
table
```

```
## # A tibble: 17 x 2
```

```
##       X1                X2
```

```
##       <chr>            <chr>
```

```
## 1 George Box          George Box
```

```
## 2 ""                  ""
```

```
## 3 Born                (1919-10-18)18 October 1919Gravesend, Kent, England
```

```
## 4 Died                28 March 2013(2013-03-28) (aged&nbsp;93)Madison, Wisconsin
```

```
## 5 Residence           United Kingdom, United States
```

```
## 6 Alma&nbsp;mater         University College London
```

```
## 7 Known&nbsp;for          "“All models are wrong”\nResponse-surface methodol
```

```
## 8 Awards              "Shewhart Medal (1968)\nWilks Memorial Award (1972)\nR.
```

```
## 9 Scientific car... Scientific career
```

Conteúdo (limpeza com regex)

```
table %>% mutate(X1 = str_replace_all(X1, "\\s", " "),
                 X2 = str_replace_all(X2, "\\s", " "),
                 X2 = str_replace_all(X2, "\\[[[:digit:]]\\]", " "),
                 X2 = str_replace_all(X2, "&nbsp;", " "))
```

```
## # A tibble: 17 x 2
```

```
##       X1                X2
```

```
##       <chr>             <chr>
```

```
## 1 George Box          George Box
```

```
## 2 ""                  ""
```

```
## 3 Born                (1919-10-18)18 October 1919Gravesend, Kent, England
```

```
## 4 Died                28 March 2013(2013-03-28) (aged 93)Madison, Wisconsin
```

```
## 5 Residence           United Kingdom, United States
```

```
## 6 Alma mater          University College London
```

```
## 7 Known for           "All models are wrong" Response-surface methodology EVO
```

```
## 8 Awards              Shewhart Medal (1968) Wilks Memorial Award (1972) R. A.
```

```
## 9 Scientific car... Scientific career
```

```
## 10 Fields             Statistics Design of experiments Bayesian statistics Ti
```

```
## 11 Institutions       ICI Princeton University University of Wisconsin-Madison
```

```
## 12 Thesis             Departures from Independence and Homoskedasticity in th
```

```
## 13 Doctoral advis... "Egon Pearson H. O. Hartley "
```

```
## 14 Doctoral stude... John F. MacGregor Greta M. Ljung
```

Procurando Links

Inspecionando a página no navegador, é possível observar que, dentro de `body #content` (o conteúdo da página), os links estão guardados no node `"li"`.

```
url = "https://en.wikipedia.org/wiki/List_of_statisticians"
listPages <- read_html(url)
links <- listPages %>%
  html_nodes("li")
```

Procurando Links

links

```
## {xml_nodeset (742)}
## [1] <li><a href="/wiki/Outline_of_statistics" title="Outline of statistico
## [2] <li><a class="mw-selflink selflink">Statisticians</a></li>
## [3] <li><a href="/wiki/Glossary_of_probability_and_statistics" title="Glo
## [4] <li><a href="/wiki/Notation_in_probability_and_statistics" title="Not
## [5] <li><a href="/wiki/List_of_statistics_journals" title="List of statis
## [6] <li><a href="/wiki/Lists_of_statistics_topics" title="Lists of statis
## [7] <li><a href="/wiki/List_of_statistics_articles" title="List of statis
## [8] <li>\n<a href="/wiki/File:Nuvola_apps_edu_mathematics_blue-p.svg" cla
## [9] <li><a href="/wiki/Category:Statistics" title="Category:Statistics">C
## [10] <li class="nv-view"><a href="/wiki/Template:Statistics_topics_sidebar
## [11] <li class="nv-talk"><a href="/wiki/Template_talk:Statistics_topics_si
## [12] <li class="nv-edit"><a class="external text" href="https://en.wikiped
## [13] <li><a href="#A">A</a></li>
## [14] <li><a href="#B">B</a></li>
## [15] <li><a href="#C">C</a></li>
## [16] <li><a href="#D">D</a></li>
## [17] <li><a href="#E">E</a></li>
## [18] <li><a href="#F">F</a></li>
## [19] <li><a href="#G">G</a></li>
```


"Sajid Ali Khan, Rawalakot" até "Zipf, George Kingsley"

```
estat1 = links %>%  
  as.character %>%  
  grep("Sajid Ali Khan, Rawalakot", .)  
estatN = links %>%  
  as.character %>%  
  grep("Zipf, George Kingsley", .)  
estat1
```

```
## [1] 40
```

```
estatN
```

```
## [1] 679
```

```
links <- links[estat1:estatN]
```

Páginas individuais

O objeto `links` possui os endereços no formato XML e não se restringem apenas aos endereços.

```
links
```

```
## {xml_nodeset (640)}
## [1] <li>\n<a href="/w/index.php?title=Sajid_Ali_Khan&action=edit&";
## [2] <li>\n<a href="/wiki/Odd_Aalen" title="Odd Aalen">Aalen, Odd Olai</a>
## [3] <li>\n<a href="/wiki/Edith_Abbott" title="Edith Abbott">Abbott, Edith
## [4] <li>\n<a href="/wiki/Robert_P._Abelson" class="mw-redirect" title="Ro
## [5] <li>\n<a href="/wiki/Moses_Abramovitz" title="Moses Abramovitz">Abram
## [6] <li>\n<a href="/wiki/Gottfried_Achenwall" title="Gottfried Achenwall"
## [7] <li>\n<a href="/wiki/Abraham_Manie_Adelstein" title="Abraham Manie Ac
## [8] <li>\n<a href="/wiki/Riaz_Ahsan" title="Riaz Ahsan">Ahsan, Riaz</a> (
## [9] <li>\n<a href="/wiki/Beatrice_Aitchison" title="Beatrice Aitchison">A
## [10] <li>\n<a href="/wiki/John_Aitchison" title="John Aitchison">Aitchison
## [11] <li>\n<a href="/wiki/Alexander_Aitken" title="Alexander Aitken">Aitke
## [12] <li>\n<a href="/wiki/Hirotsugu_Akaike" class="mw-redirect" title="Hir
## [13] <li>\n<a href="/wiki/Mir_Masoom_Ali" title="Mir Masoom Ali">Ali, Mir
## [14] <li>\n<a href="/wiki/R._G._D._Allen" title="R. G. D. Allen">Allen, R.
## [15] <li><a href="/wiki/David_B._Allison" title="David B. Allison">Allison
## [16] <li>\n<a href="/wiki/Doug_Altman" title="Doug Altman">Altman, Doug</a>
```

Páginas individuais

Devemos lembrar que os endereços nos links são armazenados no nó `<a>`.

```
links %>%  
  html_nodes("a")
```

```
## {xml_node_set (640)}  
## [1] <a href="/w/index.php?title=Sajid_Ali_Khan&action=edit&redlin  
## [2] <a href="/wiki/Odd_Aalen" title="Odd Aalen">Aalen, Odd Olai</a>  
## [3] <a href="/wiki/Edith_Abbott" title="Edith Abbott">Abbott, Edith</a>  
## [4] <a href="/wiki/Robert_P._Abelson" class="mw-redirect" title="Robert P  
## [5] <a href="/wiki/Moses_Abramovitz" title="Moses Abramovitz">Abramovitz,  
## [6] <a href="/wiki/Gottfried_Achenwall" title="Gottfried Achenwall">Acher  
## [7] <a href="/wiki/Abraham_Manie_Adelstein" title="Abraham Manie Adelstei  
## [8] <a href="/wiki/Riaz_Ahsan" title="Riaz Ahsan">Ahsan, Riaz</a>  
## [9] <a href="/wiki/Beatrice_Aitchison" title="Beatrice Aitchison">Aitchis  
## [10] <a href="/wiki/John_Aitchison" title="John Aitchison">Aitchison, Joh  
## [11] <a href="/wiki/Alexander_Aitken" title="Alexander Aitken">Aitken, Ale  
## [12] <a href="/wiki/Hirotsugu_Akaike" class="mw-redirect" title="Hirotsugu  
## [13] <a href="/wiki/Mir_Masoom_Ali" title="Mir Masoom Ali">Ali, Mir Masoom  
## [14] <a href="/wiki/R._G._D._Allen" title="R. G. D. Allen">Allen, R. G. D.  
## [15] <a href="/wiki/David_B._Allison" title="David B. Allison">Allison, Da  
## [16] <a href="/wiki/Doug_Altman" title="Doug Altman">Altman, Doug</a>
```

Páginas Individuais

No slide anterior, você deve observar que, dentro do nó <a>, existe um atributo href, que possui o link relativo (dentro da página da Wikipedia) para cada uma das páginas.

```
links %>%  
  html_nodes("a") %>%  
  html_attr("href") # Salvar title também!
```

```
## [1] "/w/index.php?title=Sajid_Ali_Khan&action=edit&redlink=1"  
## [2] "/wiki/Odd_Aalen"  
## [3] "/wiki/Edith_Abbott"  
## [4] "/wiki/Robert_P._Abelson"  
## [5] "/wiki/Moses_Abramovitz"  
## [6] "/wiki/Gottfried_Achenwall"  
## [7] "/wiki/Abraham_Manie_Adelstein"  
## [8] "/wiki/Riaz_Ahsan"  
## [9] "/wiki/Beatrice_Aitchison"  
## [10] "/wiki/John_Aitchison"  
## [11] "/wiki/Alexander_Aitken"  
## [12] "/wiki/Hirotsugu_Akaike"  
## [13] "/wiki/Mir_Masoom_Ali"  
## [14] "/wiki/R._G._D._Allen"
```

Criando os links completos

Como visto anteriormente, o atributo href refere-se ao endereço **relativo** ao endereço padrão da Wikipedia. Para montar o endereço completo, é preciso adicionar a expressão `https://en.wikipedia.org` antes do endereço relativo.

```
li <- links %>% html_nodes("a") %>% html_attr("href")
li <- paste0("https://en.wikipedia.org", li)
li %>% head()
```

```
## [1] "https://en.wikipedia.org/w/index.php?title=Sajid_Ali_Khan&action=edit"
## [2] "https://en.wikipedia.org/wiki/Odd_Aalen"
## [3] "https://en.wikipedia.org/wiki/Edith_Abbott"
## [4] "https://en.wikipedia.org/wiki/Robert_P._Abelson"
## [5] "https://en.wikipedia.org/wiki/Moses_Abramovitz"
## [6] "https://en.wikipedia.org/wiki/Gottfried_Achenwall"
```

```
names <- links %>% html_nodes("a") %>% html_attr("title")
names %>% head()
```

```
## [1] "Sajid Ali Khan (page does not exist)"
## [2] "Odd Aalen"
```

Curadoria Manual

```
bad = c("page does not exist", "Florence Nightingale",  
        "Harold Wilson", "Robert P. Abelson")  
bad1 = unlist(sapply(bad, grep, names))  
bad2 = unlist(sapply(c("mshkhan", "redlink", "orghttp"), grep, li))  
(remove = c(bad1, bad2))
```

```
## page does not exist1 page does not exist2 Florence Nightingale1  
## 1 382 131  
## Florence Nightingale2 Harold Wilson Robert P. Abelson  
## 428 617 4  
## mshkhan redlink1 redlink2  
## 469 1 382  
## orghttp1 orghttp2  
## 469 567
```

```
rm(bad, bad1, bad2)
```

Curadoria Manual

```
names = names[-remove]
li = li[-remove]
pessoas = tibble(nome=names, link=li)
pessoas %>% head()
```

```
## # A tibble: 6 x 2
##   nome                                link
##   <chr>                             <chr>
## 1 Odd Aalen                         https://en.wikipedia.org/wiki/Odd_Aalen
## 2 Edith Abbott                     https://en.wikipedia.org/wiki/Edith_Abbott
## 3 Moses Abramovitz                 https://en.wikipedia.org/wiki/Moses_Abramovitz
## 4 Gottfried Achenwall              https://en.wikipedia.org/wiki/Gottfried_Achenwal
## 5 Abraham Manie Adelstein          https://en.wikipedia.org/wiki/Abraham_Manie_Adel
## 6 Riaz Ahsan                       https://en.wikipedia.org/wiki/Riaz_Ahsan
```

Função para Extração de Tabela

```
f <- function(x){  
  ifelse(length(x) == 0, NA_character_, x)  
}  
  
extraiTabela = function(mylink){  
  table <- read_html(mylink) %>% html_node(".vcard")  
  if (is.na(html_name(table))) return(NULL)  
  if (html_name(table) != "table") return(NULL)  
  table = table %>% html_table(header = FALSE) %>%  
    mutate(X1 = str_replace_all(X1, "\\s", " "),  
           X2 = str_replace_all(X2, "\\s", " "),  
           X2 = str_replace_all(X2, "\\[[[:digit:]]\\]", " "),  
           X2 = str_replace_all(X2, "&nbsp;", " "))  
  tibble(link = mylink,  
         Born = f(table[grep("Born", table$X1), 2]),  
         Died = f(table[grep("Died", table$X1), 2]),  
         AlmaMater = f(table[grep("Alma", table$X1), 2]))  
}
```


Extraindo tabelas (demora alguns minutos...)

```
library(doMC) ## se windows library(doParallel)
registerDoMC(4) ## se windows registerDoParallel(nproc)
out = foreach(thislink=li, .combine=rbind) %dopar% {
  extraiTabela(thislink)
}
final = pessoas %>% inner_join(out, by='link')
final[, 1:3] %>% head
```

```
## # A tibble: 6 x 3
```

##	nome	link	Born
##	<chr>	<chr>	<chr>
## 1	Odd Aalen	https://en.wikipedia.org/wiki...	(1947-05-06) May 6, 1947
## 2	Edith Abbott	https://en.wikipedia.org/wiki...	(1876-09-26) September 26,
## 3	Moses Abramov...	https://en.wikipedia.org/wiki...	(1912-01-01) January 1, 1912
## 4	Gottfried Ach...	https://en.wikipedia.org/wiki...	(1719-10-20) 20 October 1719
## 5	Riaz Ahsan	https://en.wikipedia.org/wiki...	1951, December 25 Karachi,
## 6	Beatrice Aitc...	https://en.wikipedia.org/wiki...	(1908-07-18) 18 July 1908 Po

```
final[, -(1:3)] %>% head
```

```
## # A tibble: 6 x 2
```

```
##   Died                      AlmaMater
```

```
##   <chr>                     <chr>
```

```
## 1 <NA>                     University of Oslo
```

```
## 2 July 28, 1957(1957-07-28) (aged ... <NA>
```

```
## 3 December 1, 2000(2000-12-01) (ag... Harvard University and Columbia Univer
```

```
## 4 1 May 1772(1772-05-01) (aged 52)... <NA>
```

```
## 5 November 8, 2008(2008-11-08) (ag... University of KarachiAdamjee Governmen
```

```
## 6 22 September 1997(1997-09-22) (a... Goucher College (BA)Johns Hopkins Univ
```

Obrigado!