```
Enter cell code...
```

Por que estudar métodos quantitativos?

Aperfeiçoamento da forma de pensar sobre o mundo

- Qual é a probabilidade de uma mulher que testa positivo para câncer em um exame de mamografia realmente estar com a doença?
- Por que os gráficos sobre a evolução dos casos e mortos por Covid-19 são apresentados em escala logarítimica?
- O que quer dizer a "margem de erro com intervalo de confiança de 95%" nas pesquisas de opinião?
- Por que a taxa de homicídios é calculada por 100 mil?
- Como podemos ter segurança para dizer que as mulheres apoiam majoritariamente Lula na corrida eleitoral de 2022?
- A desigualdade está crescendo ou diminuindo no Brasil?

Quantificação da ciência, das ciências sociais e do jornalismo

- Máquina de aprendizagem, big dada e algorítimos
- Ciência de dados e massificação de programação (R, Python, Julia)
- Jornalismo de dados
- Novas fronteiras da pesquisa quantitativa em ciências sociais: análise de textos, redes sociais, documentos históricos
- Causalidade e estatística bayesiana

Ampliação das oportunidades profissionais

- Estágio e oportunidades profissionais para alunos que têm familiaridade com programação e estatística
- Predileção das revistas por artigos baseados em pesquisa quantitativa

💡**Hans Roling: um mestre da arte de apresentar dados**

Hans Roling foi um médico sueco que se notabilizou nos anos 2010 por suas apresentações fabulosas sobre estateisticas sobre demografia e pobreza. Faleceu em 2017, com 68 anos.

Hans Rosling
Global population growth, box by box

Jairo Nicolau: minha história com os dados

Há exatos 40 anos atrás eu comecei o curso de ciências sociais na Universidade Federal Fluminense (UFF), em Nitéroi. Quase todas as discplinas do curso tinham um formato semelhante. O professor indicava um texto para ser lido na aula seguinte; a aula consistia em diversas formas em torno do texto selecionado: exposição do professor, exposição de um aluno (ou grupo de alunos) ou uma discussão livre. Nas primeiras semanas de aula, estranhei esse formato, pois estava acostumado com o modelo do ensino médio, de fazer a leitura dos livros após a aula.

Um dos poucos cursos que fugiu desse modelo foi o de Estatística. Na época, uma aula de matemática, com uma avaliação final consistindo em uma série de exercícios feitos em casa e entregues ao professor, um argentino boa praça, do Departamento de Matemática, que acabava aprovando todo mundo.

Imagino que o professor soubesse que aquele curso não fazia nenhum sentido para os alunos de Ciências Sociais. Afinal, fomos para a grande área de humanas justamente porque erámos péssimos

alunos de matemática e no fim do segundo grau tínhamos uma certeza: nunca mais precisaríamos de usar matemática na vida. E de fato, ao longo da graduação, foram poucos os textos em que havia uma menção a números. Li alguns livros e muitos artigos de antropologia, teoria política, teoria sociológica, filosofia e história política. Professores que faziam trabalhos empíricos usavam entrevistas ou etnografia.

O primeiro texto com o qual eu tive contato que usava extensivamente estatística foi o livro *Discriminação e Desigualdades Raciais no Brasil*, de Carlos Hasenbalg (1979). O livro fazia parte da bibliografia da prova de mestrado que tive que ler em 1987. Era um texto cheio de tabelas e de leitura árida e um trabalho praticamente único da nascente sociologia da estratificação social brasileira. Uma das questões da prova consistiu em analisar um das tabelas do livro.

No mestrado, a disciplina de métodos quantitativos também era odiada pelos alunos, já que quase todos vinham da cultura anti-matemática que sempre dominou as humanidades no Brasil. Nos anos 1980, o IUPERJ era uma das centros de ciências sociais que exigiam que os alunos cursassem uma disciplina de métodos quantitativos e tinha em seus quadro professores que usavam estatatística. As aulas eram completadas por exercicios em um computador que tinha uma versão instalada do SPSS, o que era uma raridade.

Nos anos 1990 houve uma ampla massificação do uso do computador pessoal. Comprávamos as peças separadas e alguém montava o computador. Não havia computador pronto para usar. Meu primeiro computador foi comprado em 1992. No começo, ele funcionava praticamente como uma máquina de escrever a qual acrescentávamos alguns jogos. Na segunda metade dos 1990 o e-mail se espandiu e a internet foi inventada, o que revolucionou a forma que os cientistas sociais passaram a trabalhar.

À medida que os computadores foram ficando mais potentes eles tornaram-se aptos a rodar programas estatísticos. No começo dos anos 2000 já era possivel instalar o SPSS no computador. Como a licença era carrísima, todos recorreriam aos ambulantes do Edifício Central, onde versões atualizadas (que quase sempre funcionavam) eram vendidas por uma bagatela.

Na tese de doutorado, escrita na primeira metade dos anos 1990, usei estatística descritiva; todos os dados foram feitos por uma modesta calculadora pessoal. Nos anos seguintes, já como professor do IUPERJ, continuei a usar amplamente estatística em meu trabalho (basicamente índices e estatística básica). Nesse período, fazia o que a maioria dos professores fazia: contava com a ajuda de alunos e pesquisadores versados no usos do EXCEL e do SPPS para me ajudar. Comecei a ganhar alguma autonomia, quando fiz, em meados dos anos 2000, alguns cursos sobre o SPSS.

No meu pós-doutorado em 2005/2006 investi em aprender um pouco mais de estatística. Li muitos livros, fiz alguns cursos e passei a usar o SPSS razoavelmente bem. Na volta ao Brasil, tive confiança para ensinar a temida disciplina de metodologia quantitativa. Tentei inovar usando slides e

integrando um pouco mais as aulas ao laboratório de SPSS.

Com a descoberta do Tableau (um software amigável que faz gráficos bem bonitos) em 2009 passei a me interessar por visualização de dados. Na década seguinte, minha vida mudou com o uso do Stata (2013) e posteriormente com a descoberta da programação. Aos poucos, comecei a utilizar o R, o Python e até o Julia para analisar os meu dados e fazer os meus gráficos.

Uma das melhores coisas que aconteceram na minha carreira foi ter aprendido a usar a programaação para fazer análises estatísticas e gráficos. Todos os passos da minha atividade ficam registrados, o que é fundamental na comunidade científica atual. Gosto de dizer que isso aconteceu quando eu me aproximava dos 50 anos. Quase tudo que sei aprendi sozinho. E como todo autodidata sei coisas muito avançadas e não sei outras bem básicas. Falo isso porque creio que posso incentivar os mais jovens. Sobretudo, os que acham que já decidiram que não querem aprender estatística e consideram que programação é coisa de nerd.

População e amostra

População

É o conjunto de todas as observações de interesse: os moradores do Complexo da Maré, os estudantes do Pedro II, os jogadores que participam da Série A do brasileiro

Amostra

É um segmento da população cujos os dados estão disponíveis: os ndivíduos que responderam à PNAD; os alunos do Pedro II sorteados para realizar um teste de desempenho escolar; os respondentes de uma pesquisa de opinião do Instituto Quaest.

O objetivo da maioria das análises de dados é aprender sobre a população. Mas quase sempre é necessário, e mais prático, observar apenas amostras dessa população. Por exemplo: o Datafolha ouv cerca de 1.000 brasileiros para coletar informações a respeito das opiniões e crenças da população.

A **estatística inferencial** fornece avaliações sobre uma população com base em dados de uma amostra. Por exemplo, uma pesquisa realizada nos EUA em 2018 perguntou: “Você acredita no céu?” A população de interesse era composta por todos os adultos dos Estados Unidos. Dos 1.141 indivíduos da amostra, 81% responderam que sim. Estamos interessados, no entanto, não apenas nessas 1.141 pessoas, mas na população de mais de 250 milhões de adultos que residem nos Estados Unidos.

Casos e Variáveis

Nas pesquisas quantitativas é fundamental enteder o que é um caso, o que é uma variável.

Caso

Um caso é um indivíduo/objeto de uma determinada população.

Imagine, por exemplo, a população de municípios brasileiros (5570 municípios). O municipio do Rio de Janeiro é um caso da população.

Um pesquisador que estuda o PSDB, dedica-se a entender um caso de uma população de partidos brasileiros. Atualmente, existem 34 partidos regsitrados no Brasil.

Variável

Uma variável é um atributo ou característica dos indivíduos/objetos.

O termo variável dá ênfase ao fato de que os atributos dos indivíduos variam. Numa população de mulheres, o sexo não é uma variável.

Em uma pesquisa eleitoral, a idade e a escolaridade são variáveis, já que que observamos pessoas de diferentes idades e escolaridades na população.

O que é significa a letra N nas pesquisas quantitativas?

O **n** faz referência ao número de casos de uma população. A população de municípios brasileiros é de 5570, portanto o **n = 5570**.

Tipos de variáveis

Variável quantitativa (numérica)

Porta algum valor numérico que é passível de operação matemática (soma, divisão, multiplicação): idade, número de filhos, renda per capita, anos de estudo

Variável categórica

Cada observação pertence a uma categoria, em um conjunto de categorias:

- Gênero: masculino, feminino
- Religião: católica, evangélica, espírita, budista, outras
- Tipo de moradia: apartamento, casa
- Crença na vida após a morte: sim, não

💡Quizz 1

Identifique se cada uma da variáveis é categórica (factor) ou quantitativa:

- número de crianças em uma creche
- sessão eleitoral em uma pesquisa sobre geografia do voto
- estado civil
- distância (em quilômetros) de deslocamento até a escola
- CPF

Resposta do Quizz 1

💡Nate Silver: o mais conhecido analista de dados

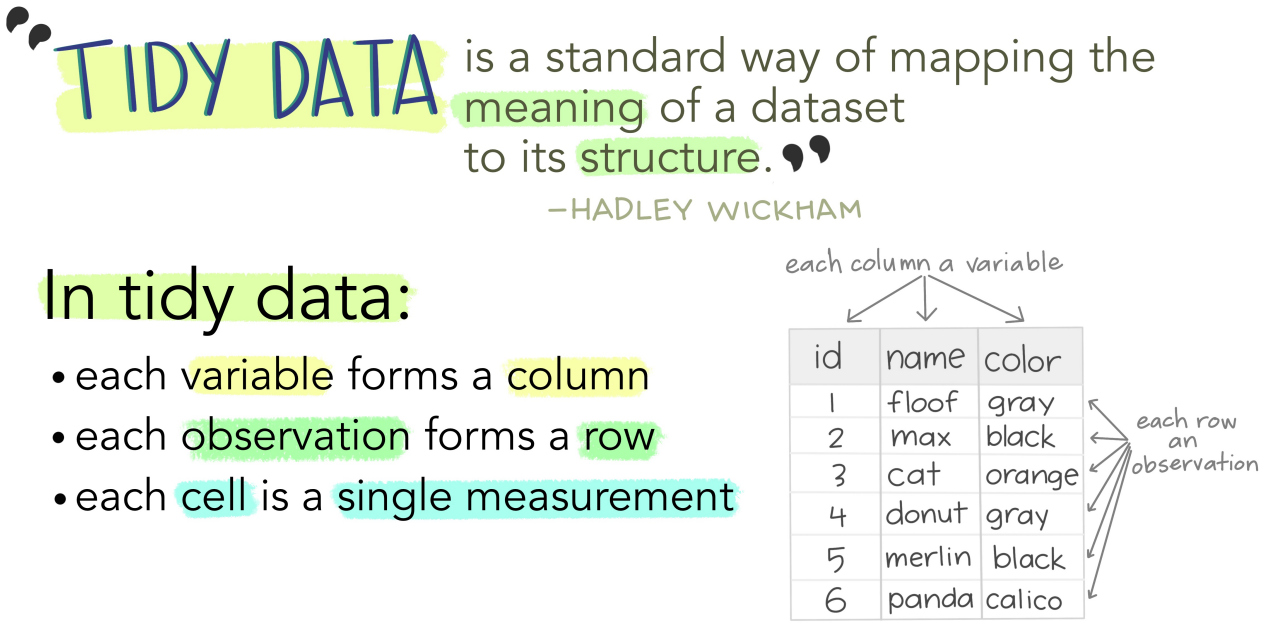
Nate Silver é fundador do site 538, especializado em jornalismo de dados e autor do livro *O Sinal e o Ruído* (recomendo a leitura).

Dados Tabulares

Quase todos os dados que os pesquisadore lidam estão no formato tabular

Onde:

- Cada **coluna** é uma variável
- Cada **linha** é uma observação
- Cada **célula** é uma mensuração única



Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

Figura com licença creative commons de `@allisonhorst`

Exemplo de um banco em formato tabular

2000 American National Election Studies: Dataframe sobre o nível de informação dos cidadãos

Um banco de dados com 1807 observações e 8 variáveis:

- **y** classificação do entrevistador com os levels: Very Low, Fairly Low, Average, Fairly High, Very

High

- **collegeDegree** categóricas com levels: No, Yes
- **female** categórica com os levels: No, Yes
- **agea** numérica com a idade dos respondente em anos
- **homeOwn** categórica com levels: No, Yes
- **govt** categórica com levels: No, Yes
- **length** numérica, mostra a duração da entrevista em minutos
- **id** um número que identifica cada respondente

`Informação_Política` =

	Y	CollegeDegree	Female	Age	HomeOwn
1	"Fairly High"	"Yes"	"No"	49	"Yes"
2	"Average"	"No"	"Yes"	35	"Yes"
3	"Very High"	"No"	"Yes"	57	"Yes"
4	"Average"	"No"	"No"	63	"Yes"
5	"Fairly High"	"Yes"	"Yes"	40	"Yes"
6	"Average"	"No"	"No"	77	"Yes"
7	"Average"	"No"	"No"	43	"Yes"
8	"Fairly High"	"Yes"	"Yes"	47	"Yes"
9	"Average"	"Yes"	"Yes"	26	"Yes"
10	"Very High"	"No"	"Yes"	48	"No"
	more				
1807	"Average"	"No"	"Yes"	62	"Yes"