Um Curso Intuitivo de Probabilidade (e Estatística)

para a área de dados

Caio Velasco

Engenheiro Mecânico, UFRJ

Master in Public Policy, University of California Los Angeles

Junho 2023

> *“When you explain a why, you have to be in some framework that allows something to be true. Otherwise, you're perpetually asking why.”*
>
> Richard Feynman, Theoretical Physicist

Contents

List of Figures

A Real Motivação

Apesar de ter estudado *Probabilidade e Estatística* na Engenharia (UFRJ) e de ter feito alguns cursos no mestrado em UCLA, percebi que não dominava o assunto. Até mesmo quando iniciei o PhD em Economia na pandemia na Holanda (que infelizmente não pude continuar), onde as disciplinas eram altamente matemáticas (por exemplo, *Análise Real* e noções de *Teoria da Medida* eram pré-requisitos), percebi que algo estava faltando.
Eu tinha um bom entendimento dos conceitos, mas quanto as coisas ficaram complexas, comecei a ter dificuldades. Porém, ainda não sabia o porquê. Então, resolvi dar uns passos atrás e aumentar o nível matemático.

Foi com base nessa experiência que percebi o problema e criei o **Um Curso Intuitivo de Probabilidade e Estatística**, *para a área de dados.* Eu acredito no seguinte:

- Não importa qual seja seu objetivo no mundo dos dados, **é crucial** entender a estrutura matemática da teoria da probabilidade.

Convido você a se fazer a seguinte pergunta: *Será que eu realmente sei o que está descrito abaixo?*

O **Modelo Estatístico** foi desenvolvido para formalizar *padrões de regularidade estatística* (ou seja, as *informações sistemáticas*) presentes nos dados observados. A formalização matemática deste modelo é baseada na teoria da probabilidade e busca capturar tal informação sistemática gerada por *mecanismos estocásticos*. Portanto, para tomar uma decisão sob incerteza de maneira lógica e consistente, precisamos estabelecer um framework (de modelagem) para tais *fenômenos estocásticos*.

Bem, se sua resposta for não, faça o curso :)
Acredito que você romperá barreiras se dedicar um tempo para este curso.

Chapter 1

Probability

Nesse capítulo, o objetivo principal é aprender o conceito de **Probabilidade**.

Começaremos com o entendimento sobre o que é algo *Estocástico*, conceito que fundamenta o mundo da estatística e o mundo dos dados. Em seguida, aprenderemos sobre a ideia de *Regularidade Estatística*, uma característica de **crucial** que é **observada** em situações que acreditamos ser do tipo *"fenômeno estocástico"* (a palavra *observada* é bem importante e muito usada). Ao aprender sobre esses conceitos, você perceberá que Probabilidade nada mais é do que uma tentativa de "adestrar" (matematicamente) a tal regularidade estatística.

Você usará esse conceitos para construir um **Modelo Estatístico** e para isso precisará saber construir um **Modelo Probabilístico**.

Com isso, você estará pronto para entender criteriosamente sobre o que é um *Experimento* na Estatística.

Este capítulo foi baseado, em sua maioria, no livro *Probability Theory and Statistical Inference*, do Aris Spanos [1], um dos meus preferidos.

Vamos lá.

1.1 O que é um fenômeno "Determinístico"?

Um fenômeno é dito *determinístico* quando seu futuro **não é** incerto.

Um exemplo é a função matemática, que é uma regra com certas propriedades (*todos os elementos do domínio* ***devem*** *ser cobertos e você* ***não pode*** *associar o mesmo elemento do domínio com vários elementos do contradomínio* - você não deve esquecer disso, mas a maioria esquece). Portanto, se você conhece a regra, assim que conhece a entrada (o *argumento* da função), também conhece a saída (o *valor* que essa função *assume*).

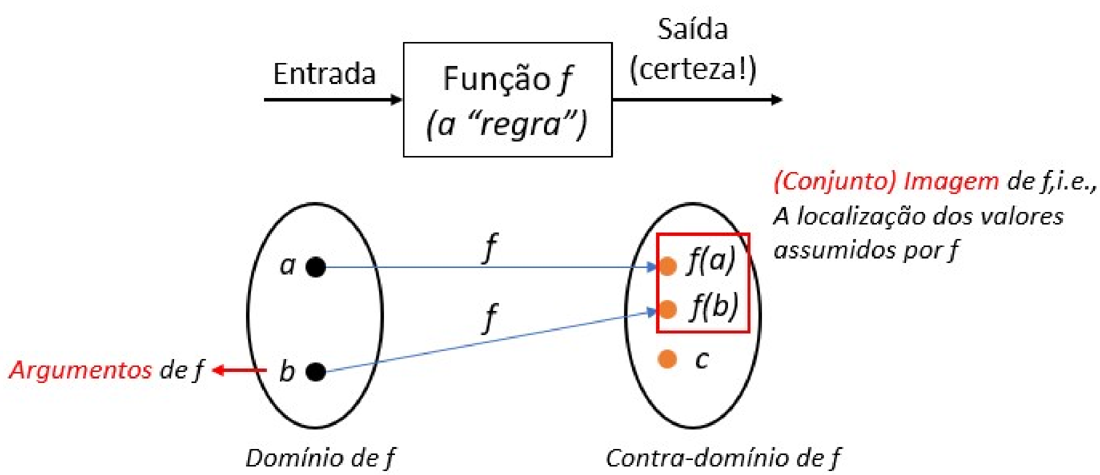

Figure 1.1: A Caixa Preta de uma Função

Se pegarmos uma função f cuja regra é $f(x) = 2x$ e avaliarmos o valor dessa função para o $x = 3$, sabemos *com certeza* que a função assumirá o valor $f(3) = 6$. Não há incerteza sobre o futuro nesse processo. (note que a terminologia aqui é importante!)

Outra maneira de colocar é: uma situação é *determinística* quando o estado anterior pode prever *com certeza* o próximo estado. Em outras palavras, dado que temos um estado inicial, sempre podemos prever qualquer estado futuro de interesse porque o caminho para chegar lá é determinístico.

Não se esqueça: Determinismo diz respeito à **certeza** em prever o futuro.

1.2 O que é um fenômeno "Estocástico"?

Vamos definir o que entendemos por *fenômenos estocásticos* e introduzir o conceito de *regularidade estatística*.

Definição - Fenômeno Estocástico e Regularidade Estatística

Fenômeno Estocástico é aquela situação onde os dados apresentam padrões de *Regularidade Estatística*. Em inglês, também é chamado de *chance regularity*. Okay, mas o que isso significa?

Os dados que nós observamos no mundo real **exibem certos padrões**, chamados de *padrões de regularidade estatística*, e o que queremos é justamente tentar matematizar (ou modelar matematicamente) esses padrões.

Construir esse o modelo é justamente o que usaremos para tirar conclusões sobre o que está acontecendo.

Preste bastante atenção agora, pois toda a construção do conceito de probabilidade é baseada na seguinte ideia:

> O conceito de probabilidade é construído sobre a propriedade de que os dados que observamos **exibem certos padrões regulares**.

Ótimo, mas como algo incerto pode exibir um comportamento regular?

Para responder a essa pergunta, precisamos dividir o conceito em duas partes:

$$chance + regularity$$

ou, no português, $regularidade + da\ chance$.

Ao dividi-lo em duas partes, podemos diferenciar a ideia por trás de cada uma e entender como que a ideia de *incerteza* (a chance de algo acontecer) está conectada a um certo **padrão**:

O mundo real apresenta incertezas e, claro, não sabemos sobre o futuro. No entanto, essa noção está intimamente ligada ao fato de que no nível **individual** temos incerteza, mas no nível **agregado** notamos um certo padrão. (Isto é **muito importante**. Veja abaixo).

Intuição - Regularidade Estatística

Considere o seguinte experimento: Jogar dois dados e observar a soma.
Você sabe (com certeza) os resultados: $2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$
Mas onde está a parte *incerta* e onde está a parte *regular*?
Nível Individual Se você jogar os dois dados pela primeira vez, você não saberá qual será o resultado. Se jogar na segunda vez, também não saberá. Se jogar pela milésima vez (sem memorizar os valores anteriores), também não saberá.
Então, no nível *individual*, observamos nos dados a ideia de **chance** ou **incerteza**.

Nível Agregado
Agora, se você jogar os dados várias vezes, **acompanhar** todos os resultados e plotá-los em um gráfico do tipo histograma (aquele que agrega a **frequência** de resultados iguais em um experimento - veja mais abaixo), você poderá observar um certo padrão.
Portanto, no nível *agregado*, observamos **uma certa regularidade** nos dados.

Abaixo, extraída do livro [1] e traduzida para o português, coloco um frase muito forte que resume tudo isso:

"De fato, a essência da regularidade estatística decorre do fato de que a desordem no nível individual cria (de alguma forma) ordem no o nível agregado."

Note que esta figura possui uma propriedade interessante, a *agregação* dos

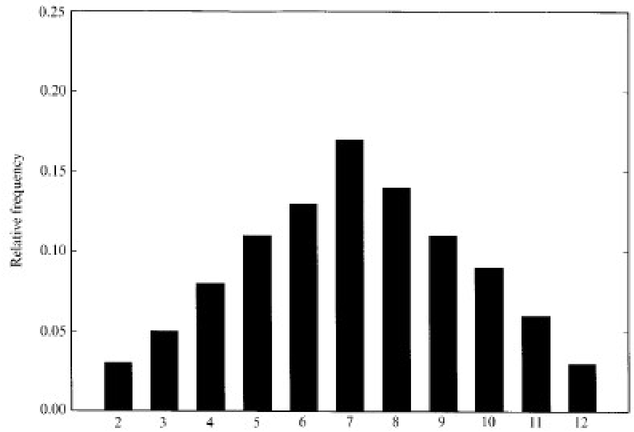

Figure 1.2: Histograma da soma de dois dados jogados 100 vezes

dados.

Neste caso particular, a regra de agregação foi uma contagem simples de cada resultado igual que apareceu nas 100 tentativas. Por exemplo, se $soma = 7$ apareceu 17 vezes, este histograma de *frequências relativas* tem uma altura (eixo y) de 17%, ou $0,17$.

Obs.: o **Histograma** pode ser apresentado de duas formas:

- mostrando a *frequência absoluta* no eixo y, que representa a contagem de cada resultado
- ou mostrando a *frequência relativa* no eixo y, que representa a porcentagem de cada resultado no todo, neste caso, nas 100 tentativas.

  De qualquer forma, o histograma apresenta a ideia de **agregação**.

Nota: O histograma foi apenas um exemplo para nos ajudar a visualizar os padrões de *regularidade estatística* e justificar toda a história da Probabilidade. Há outras formas de visualizar tais padrões e é justamente nessas outras formas que surgem conceitos **extremamente** importantes, como:

- o conceito de **Distribuição** dos dados, que se refere a como os resultados do experimento estão distribuídos após várias tentativas, que nos faz acreditar que há uma lei que seja estável o suficiente para chegar a esse formato (o que pode ser visualizado pelo histograma).

- o conceito de **Independência** se refere à (não) influência que um resultado anterior (ou resultados anteriores) exerce sobre os seguintes. Nesse caso, a forma de visualizar é perceber que se você tapar **um pedacinho** da parte de dentro da janela vermelha da figura abaixo e tentar prever o próximo resultado, não conseguirá prever. Se você tapar um outro pedacinho, ainda não será capaz de prever. Portanto, saber as informações que não foram tapadas não ajudam a *prever* o resultado seguinte. Neste caso, dizemos que os resultados do experimento são *independentes*.

- e o conceito de **Homogeneidade**, que se refere a uma propriedade exibida pelos dados que nos faz acreditar que, ao deslizarmos a janela vermelha para a direita, a variação vertical observada nos resultados é praticamente constante. Quando não é, dizemos que os dados exibem *heterogeneidade*.

Você notou que tocamos no assunto *frequência* com a introdução do *histograma*, certo? Notou que extraímos conceitos muito importantes como *Distribuição* e *Independência* do mesmo assunto? Além disso, notou que focamos na ideia de frequência **relativa**?

Bem, é aí que começamos a desvendar a **intuição** por trás da Probabilidade.

Considere o mesmo experimento: lançar dois dados e observar a soma.

No entanto, vamos nos concentrar na estrutura dos dados por trás da soma, ou seja, vamos observar o que aparece nas duas faces dos dados (também conhecido como espaço amostral, sobre o qual aprenderemos mais adiante).

Observe na figura abaixo que é mais *provável* que a soma seja 4 do que 12,

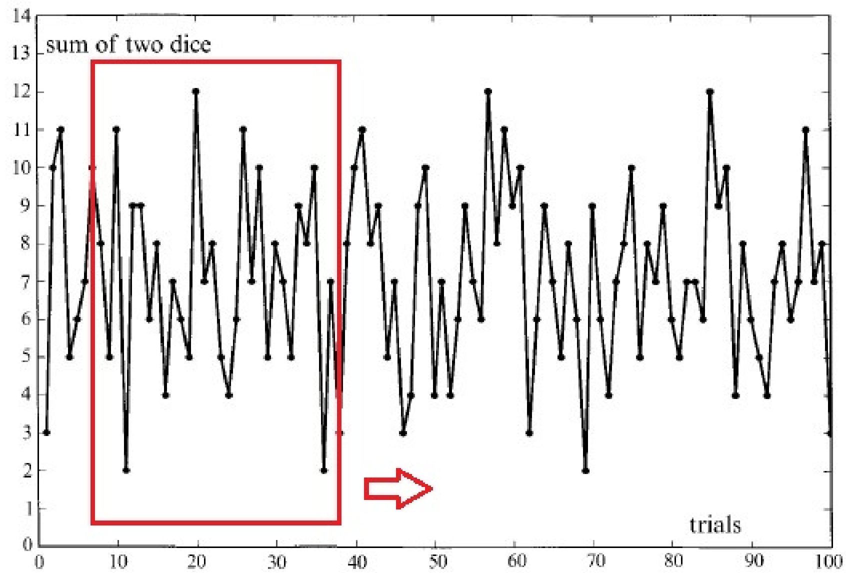

Figure 1.3: Gráfico da sequência cronológica da soma das 100 jogadas dos dados

pois há uma propriedade inerente a esse experimento de que alguns resultados são mais **frequentes** do que outros. Neste caso, poderíamos dizer que é *três vezes mais provável* obter uma soma de 4 do que uma soma de 12.

Se voltarmos ao histograma, você verá que essa ideia de frequência, na forma de frequência *relativa*, está intimamente relacionada ao significado intuitivo de probabilidade no sentido de ser *proporcional ao todo* .

O histograma, neste caso, transmite a ideia da *frequência relativa* com que aparecem os resultados da experiência e indica que **quanto mais lançamos os dados, mais fácil se torna perceber o padrão da regularidade estatística** que parece estar por trás da explicação do conceito de probabilidade. Isso é o **coração** da definição de Probabilidade, como veremos ma seguir.

No entanto, embora o histograma nos ajude a entender o fenômeno estocástico observado de regularidade estatística, ele não formaliza o conceito de probabilidade. Afinal, vemos apenas as contagens de resultados iguais quando repetimos um experimentos. Para entender, vamos começar com a tabela abaixo.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Figure 1.4: Tabela com todos os possíveis resultados da face dos dois dados jogados ao mesmo tempo

Nesta tabela, fica claro que existe uma possível conexão entre o histograma e a Probabilidade. É justamente nessa tentativa de capturar a chance de cada resultado dentre todos os resultados possíveis que surge a ideia de avaliar o quão provável (**ou quão frequente é**) observar um determinado resultado após **muitas** tentativas. A razão pela qual *muitas* está em negrito é porque isso está no centro da definição. Nós chegaremos lá.

1.3 O que é "Probabilidade"?

Observe que, por exemplo, se a possibilidade de obter uma soma de 4 aparece com mais frequência (ou seja, com maior frequência) do que uma soma de 12, é muito provável que a probabilidade de a soma ser 4 seja menor do que a da soma sendo 12, conforme mostrado em vermelho na figura abaixo. Essa comparação é feita para todos os resultados possíveis quando lançamos os dados várias vezes e observamos o padrão que surge quando contamos os resultados obtidos e os plotamos em um gráfico semelhante a um histograma.

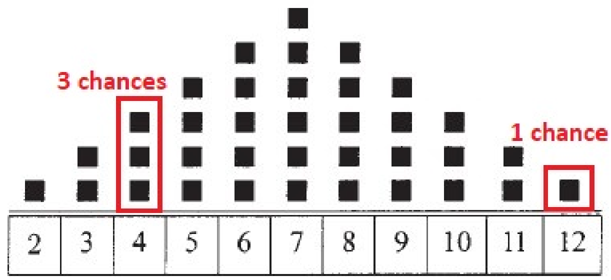

Figure 1.5: Gráfico que tenta capturar a ideia da distribuição da probabilidade do nosso experimento

Intuição - Probabilidade e seu comportamento a longo prazo

O conceito de **Probabilidade** está intimamente ligado à ideia da **frequência relativa** com que ocorrem os resultados ao repertirmos um experimento **muitas** vezes. Nós agrupamos os resultados idênticos, os contamos e, finalmente, comparamos suas proporções em todas as tentativas do experimento.

Para ser mais preciso, a realidade é que a essência da Probabilidade não está apenas embutida na *frequência* com que resultados idênticos aparecem durante várias tentativas de um experimento, mas também está relacionada à uma certa estrutura física por trás do experimento, que poderíamos apontar como a *simetria* ou *geometria* natural do problema em questão.

No caso de jogarmos dois dados, a geometria dos mesmos é um bom indicador da razão pela qual observamos os resultados da tabela acima, ou seja, é uma boa razão para argumentar que algumas somas ocorrem com mais frequência do que outros.

O último ponto importante nesta seção é que nosso objetivo é apresentar o caminho que a modelagem matemática percorreu para modelar tais situações estocásticas, partindo da observação de fenômenos estocásticos, capturando sua *regularidade estatística*) com o conceito de probabilidade (e a tal distribuição de probabilidade) e formalizando todo esse meio de campo com o tal **Modelo Estatístico**. Tudo isso baseado na Teoria da Probabilidade.

É dentro desta formalização que surgem os conceitos de *Probabilidade, Distribuição de Probabilidade, Espaço de Probabilidade, Variável Aleatória, Independência, Modelo Probabilístico, Parâmetros, Estimativa, Especificação, Identificação*, entre outros.

Esta formalização via modelos estatísticos (incluindo uma de suas partes, os modelos probabilísticos) é feita dentro de um *Modeling Framework*.

Este *framework* (de modelagem) é conhecido como **Teoria da Probabilidade**, uma disciplina ensinada em matemática, estatística e física, mas que geralmente não é abordada em profundidade nem mesmo na Engenharia, onde geralmente é chamada de Probabilidade e Estatística que é apresentada de uma forma mais condensada, o que acaba deixando de lado a essência que apresentamos aqui e contribuindo para a motivação que eu descrevi no início: tive que voltar do zero e me aprofundar matematicamente para realmente dominar o assunto.

> A teoria da probabilidade é um framework de modelagem que visa formalizar matematicamente a *regularidade estatística*, característica presente nos mecanismos que constituem a "essência" dos *fenômenos estocásticos*. Chegaremos lá.

E é isso que vamos aprender neste curso :)

Nota: Frequentistas vs. Bayesianos

Não vou entrar em detalhes sobre as diferentes **interpretações de probabilidade**[1] no mundo dos estudiosos da estatística. Porém, vale ressaltar que é justamente pela relação entre *Probabilidade* e *Frequência* que dizemos que estamos no reino da *Probabilidade Frequentista*, também conhecida como *Probabilidade Clássica*. Esta última parece estar até mais ligada à tal simetria (ou geometria) natural de um problema, uma ideia que veio primeiro na história. O outro lado da história são os Bayesianos, mas é muito cedo para aprofundar nisso. Por enquanto, vamos **finalmente** nos ater à definição do conceito de Probabilidade. Veja abaixo.

[1] Veja Wikipedia

Para finalize, vamos finalmente definir Probabilidade, dentro do mundo dos Frequentistas/Clássicos.

Definição: Frequência Relativa de Longo Prazo ou Probabilidade!)

A **Probabilidade** de um evento é definida como o *limite* da frequência relativa de seus resultados em um número muito grande de tentativas hipotéticas.

A ideia **limitante** refere-se a um contexto onde tentamos responder sobre a probabilidade de um evento numa ideia hipotética pautada em **"COMO SE"** fôssemos repetir o experimento muitas vezes. Esta é a Interpretação Frequentista (ou Clássica) da Probabilidade.

Apenas para ver um exemplo, veja a figura abaixo, onde uma moeda foi lançada várias vezes e a frequência relativa do aparecimento de *Caras* foi avaliada para que pudéssemos perceber qual poderia ser sua probabilidade de ocorrência. Veja a foto abaixo. [2]

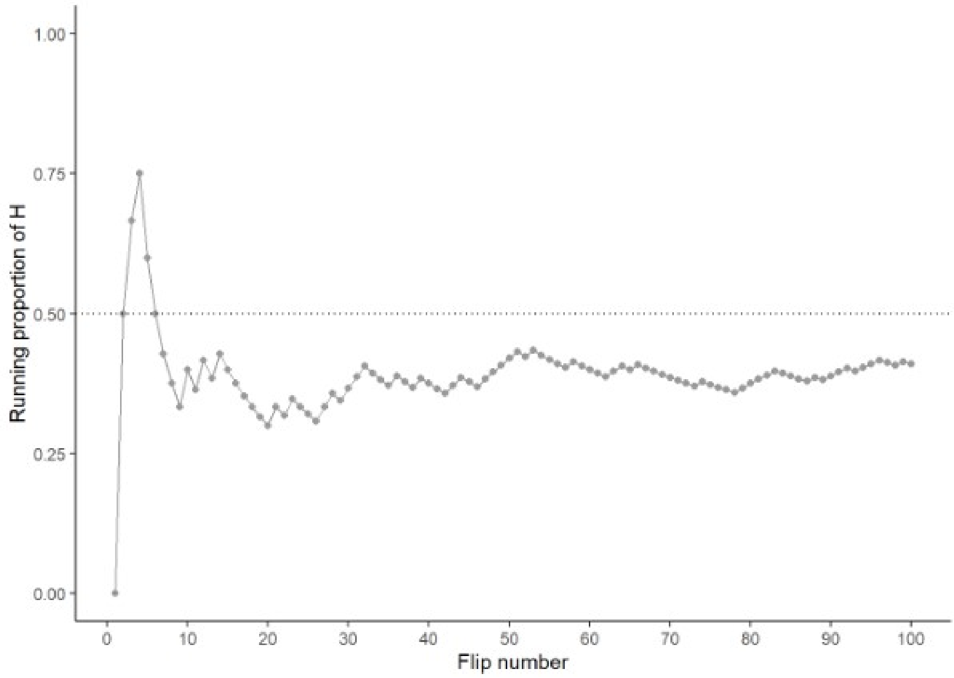

Figure 1.6: Moeda lançada várias vezes e sua proporção atual de Caras

Na figura acima, nota-se que conforme jogamos a moeda cada vez mais, o comportamento *no limite* de obter Cara se aproxima de $0,5$ ou 50% ou $\frac{1}{2}$.

Pronto, agora você aprendeu sobre o que é o conceito de Probabilidade e percebeu que a tal Distribuição de Probabilidade saiu da regularidade estatística presente em fenômenos estocásticos que o estatístico quer tentar adestrar.

A ideia agora é entender como que o estatístico vai adestrar esses fenômenos e para isso daremos mais um passo em direção ao **Modelo Estatístico**. Este próximo passo será entender o que é um **Modelo de Probabilidade**.

[2]Veja Fonte

Bibliography

[1] Aris Spanos. *Probability Theory and Statistical Inference*. Wiley, 2003.