Analytics e Inteligência Artificial Data Science

Tema da aula
Inferência Estatística

FIA BUSINESS SCHOOL lab.data

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseadas em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

FIA BUSINESS SCHOOL lab.data

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil.

Os diretores foram professores de grandes especialistas do mercado.

+10 anos de atuação.

+9.000 alunos formados.

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria;
- Larga experiência de mercado na resolução de *cases;*
- Participação em congressos nacionais e internacionais;
- Professor assistente que acompanha o aluno durante todo o curso.

Estrutura

- 100% das aulas realizadas em laboratórios;
- Computadores para uso individual durante as aulas;
- 5 laboratórios de alta qualidade (investimento +R$2MM);
- 2 unidades próximas à estação de metrô (com estacionamento).

FIA BUSINESS SCHOOL lab.data

PROFA. DRA. ALESSANDRA DE ÁVILA MONTINI

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Tem muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e Inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em Estatística Aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Membro do Conselho Curador da FIA, coordenadora de grupos de pesquisa no CNPQ, parecerista da FAPESP e colunista de grandes portais de tecnologia.

Currículo do Professor

PROF. ÂNGELO CHIODE, MSc

Bacharel, mestre e candidato ao PhD em Estatística (IME-USP), atua como professor de Estatística Aplicada para turmas de especialização, pós-graduação e MBA na FIA. Trabalha como consultor nas áreas de Analytics e Ciência de Dados há 13 anos, apoiando empresas na resolução de desafios de negócio nos contextos de finanças, adquirência, seguros, varejo, tecnologia, aviação, telecomunicações, entretenimento e saúde. Nos últimos 5 anos, tem atuado na gestão corporativa de times de Analytics, conduzindo projetos que envolviam análise estatística, modelagem preditiva e *machine learning*. É especializado em técnicas de visualização de dados e design da informação (Harvard) e foi indicado ao prêmio de Profissional do Ano na categoria Business Intelligence, em 2019, pela Associação Brasileira de Agentes Digitais (ABRADi).

linkedin.com/in/achiode

Conteúdo Programático

DISCIPLINAS

- IA E TRANSFORMAÇÃO DIGITAL
- ANALYTICS
- INTELIGÊNCIA ARTIFICIAL: MACHINE LEARNING
- INTELIGÊNCIA ARTIFICIAL: DEEP LEARNING
- EMPREENDEDORISMO E INOVAÇÃO
- COMPORTAMENTO HUMANO E SOFT SKILLS

TEMAS: ANALYTICS E MACHINE LEARNING

- ANÁLISE EXPLORATÓRIA DE DADOS
- INFERÊNCIA ESTATÍSTICA
- TÉCNICAS DE PROJEÇÃO
- TÉCNICAS DE CLASSIFICAÇÃO
- TÓPICOS DE MODELAGEM
- TÉCNICAS DE SEGMENTAÇÃO
- TÓPICOS DE ANALYTICS
- MANIPULAÇÃO DE BASE DE DADOS
- AUTO ML

TEMAS: DEEP LEARNING

- REDES DENSAS
- REDES CONVOLUCIONAIS
- REDES RECORRENTES
- MODELOS GENERATIVOS

FERRAMENTAS

- LINGUAGEM R
- LINGUAGEM PYTHON
- DATABRICKS

FIA BUSINESS SCHOOL lab.data

Conteúdo da Aula

1. Introdução
2. Objetivo
3. Probabilidade
4. Tarefa da Inferência Estatística
5. Intervalo de Confiança
6. Tópico Extra: Tamanho Amostral

Referências Bibliográficas

1. Introdução

Conceitos Preliminares

A seguir, vamos introduzir alguns conceitos importantes de **estatística**:

- **população** *versus* **amostra**
- **estatística descritiva** *versus* **estatística inferencial**
- **tipos de amostra**

População e Amostra

População	Amostra
Todas as observações de um universo de referência.	**Parte das observações** de uma população.
O tamanho da população é denotado pela letra **N**.	O tamanho da amostra é denotado pela letra **n**.

Todos os eleitores do Brasil com idades entre 18 e 70 anos

3.000 respondentes de uma pesquisa de intenção de voto para eleições nacionais

Todos os veículos fabricados em uma montadora em 1 semana

30 veículos fabricados em 1 semana e inspecionados pelo controle de qualidade

Todos os indivíduos que serão clientes de um banco daqui a 6 meses

1 milhão de clientes atuais do banco, com seu devido histórico transacional

FIA BUSINESS SCHOOL lab.data

População e Amostra

Em muitas situações, **não é possível** coletar dados acerca de todos os elementos da população de interesse, pois:

- Pode ser **inviável operacionalmente.**

  Exemplo: entrevistar todos os habitantes do Brasil com idades entre 18 e 70 anos para questioná-los sobre sua intenção de voto.

- Pode envolver um **custo financeiro muito elevado.**

  Exemplo: examinar de forma minuciosa 100% dos veículos fabricados por uma montadora, em busca de defeitos.

- Pode se tratar de **elementos futuros**, com dados ainda não observados.

  Exemplo: analisar dados de todos os indivíduos que serão clientes do banco daqui a 6 meses ***(impossível)****.*

Dessa forma, a amostra consiste em uma **pequena parte** da população, cujos dados podem ser coletados de forma operacionalmente viável e sem custo demasiadamente elevado.

Estatística Descritiva *versus* Estatística Inferencial

Estatística Descritiva (ou Exploratória)

Tem como objetivo descrever comportamentos observados a partir de **dados populacionais** (quando conhecidos) ou **dados amostrais**.

Entre todos os eleitores que votaram nas eleições, **25%** escolheram o candidato A

Entre os 3.000 eleitores que abordados na pesquisa, **23%** escolheram o candidato A

Entre todos os veículos fabricados, **2%** têm algum defeito

Entre os 30 veículos inspecionados, **3%** têm algum defeito

Entre todos os clientes do banco daqui a 6 meses, **15%** possuem o produto A

Entre os 1 milhão de clientes atuais do banco, **10%** possuem o produto A

Estatística Descritiva *versus* Estatística Inferencial

Estatística Inferencial

Tem como objetivo inferir (ou prever, ou predizer, ou estimar) comportamentos **populacionais desconhecidos** a partir de **dados amostrais conhecidos**.

Entre todos os eleitores que votaram nas eleições, **X%** escolheram o candidato A

Entre os 3.000 eleitores que abordados na pesquisa, **23%** escolheram o candidato A

Estimativa

Entre todos os veículos fabricados, **X%** têm algum defeito

Entre os 30 veículos inspecionados, **3%** têm algum defeito

Estimativa

Entre todos os clientes do banco daqui a 6 meses, **X%** possuem o produto A

Entre os 1 milhão de clientes atuais do banco, **10%** possuem o produto A

Estimativa

FIA BUSINESS SCHOOL lab.data

Tipos de Amostra

Para garantir que os resultados obtidos a partir de uma amostra representem **boas estimativas** acerca do comportamento populacional, é necessário que a amostra seja coletada de forma que reflita, aproximadamente, as **mesmas características da população**.

As formas mais comuns de **coleta de amostras** são:

- **Amostragem aleatória simples (AAS)**
  Exemplo: sortear aleatoriamente 5% de todos os veículos produzidos e destiná-los para o controle de qualidade.
  Funciona como jogar um dado não viciado, no qual todas as 6 faces possuem a mesma probabilidade de caírem para cima (1/6 ≈ 16,7%).

- **Amostragem estratificada (AE)**
  Exemplo: entrevistar aleatoriamente 3.000 eleitores no Brasil, respeitando proporções específicas de gênero, faixa etária, região e renda.
  Funciona como um conjunto de amostras aleatórias simples, uma para cada segmento (“estrato”) populacional que se deseja controlar.

- **Amostragem temporal (AT)**
  Exemplo: analisar todos os dados disponíveis e devidamente registrados em um determinado “recorte” ou ”foto” temporal.
  Esse tipo de amostragem funciona bem para projetar resultados futuros desde que as características da população permaneçam estáveis.

2. Objetivo

Objetivo

Como vimos, o objetivo da **inferência estatística** consiste em predizer aspectos desconhecidos de uma população a partir de dados amostrais.

Nesta aula, vamos aprender conceitos introdutórios de inferência estatística que nos permitirão realizar predições a respeito de **médias** e **proporções** populacionais, utilizando o conceito de **distribuição normal**.

3. Probabilidade

O que é Probabilidade?

Utilizamos a palavra **probabilidade** em diversas situações do cotidiano.

> *A previsão do tempo indicou que há 70% de probabilidade de chover hoje à tarde; é melhor levar meu guarda-chuva.*

> *Com esse trânsito, a probabilidade de eu chegar a tempo para a primeira reunião do dia é zero.*

> *Jogamos cara ou coroa para decidir qual time começaria o jogo, para que ambos tivessem a mesma probabilidade.*

> *Estou muito cansado, a probabilidade de eu ficar em casa nesse próximo feriado é de 100%.*

O que essas frases realmente significam? Qual é o conceito de **probabilidade**?

Definição de Probabilidade e Evento

A **probabilidade** é uma medida que quantifica o grau de **certeza/incerteza** que temos a respeito da ocorrência de um evento. É uma função matemática representada pela letra P.

Eventos são acontecimentos quaisquer observáveis em um experimento da realidade. São representados por meio de letras maiúsculas do alfabeto.

Voltando aos exemplos:

Evento **A**: *Chover hoje à tarde*	P(**A**) = **70%**
Evento **B**: *Chegar a tempo para a primeira reunião do dia*	P(**B**) = **0%**
Evento **C**: *O time adversário iniciar o jogo, após cara ou coroa*	P(**C**) = **50%**
Evento **D**: *Ficar em casa no próximo feriado*	P(**D**) = **100%**

Quais valores uma probabilidade pode assumir?

Ela assume valores de **0%** (certeza de não ocorrência do evento) até **100%** (certeza de ocorrência do evento). Quanto mais próxima de **50%**, maior a incerteza.

FIA BUSINESS SCHOOL lab.data

Probabilidade como Medida de Frequência

A probabilidade também pode ser interpretada como medida de **frequência relativa** de ocorrência de eventos, tal como estudamos em análise exploratória.

Voltando aos exemplos:

Evento **A**: *Chover hoje à tarde* P(**A**) = **70%**

Entre todas as possibilidades de configuração de clima avaliadas pelo instituto de meteorologia, **70%** delas resultam em chuva hoje à tarde.

Evento **B**: *Chegar a tempo para a primeira reunião do dia* P(**B**) = **0%**

Entre todas as possibilidades de rearranjo de trânsito de veículos, abertura de semáforos, incidentes etc., **0%** delas me levam a chegar a tempo para a reunião.

Evento **C**: *O time adversário iniciar o jogo, após cara ou coroa* P(**C**) = **50%**

Entre todas as possibilidades de lançamento da moeda, ignorando fatores como força, velocidade e direção do lançamento, **50%** delas resultam em cara (ou coroa).

Evento **D**: *Ficar em casa no próximo feriado* P(**D**) = **100%**

Entre todas as possibilidades de planos pessoais que venham à minha cabeça para o próximo feriado, **100%** delas incluem ficar em casa para descansar.

Exercício

Calculemos a **probabilidade** de ocorrência de alguns eventos:

Experimento 1: Sorteio aleatório de 1 cliente para responder uma pesquisa de satisfação, em uma base de dados de 20 clientes, identificados por números sequenciais de 1 a 20.

Evento A: Sortear o cliente de número 17.

Probabilidade: P(A) = 1/20 = 5%

Experimento 2: Decisão entre 6 jogadores sobre quem iniciará um jogo de tabuleiro, a partir do lançamento de um dado de 6 lados. Cada jogador escolheu um número.

Evento A: Iniciar com o jogador de número 1.

Probabilidade: P(A) = 1/6 ≈ 17%

Exercício

Calculemos a **probabilidade** de ocorrência de alguns eventos:

Experimento 3: Sorteio aleatório de 2 entre 10 colaboradores de um time, que ganharão um *voucher* promocional de um aplicativo de restaurantes. O time é composto por 4 analistas de dados e 6 cientistas de dados.

Evento A: Ambos os colaboradores sorteados serem analistas de dados.

Probabilidade: P(A) = (4/10) * (3/9) = 12/90 = 4/30 ≈ 13%

Experimento 4: Avaliação de se determinado cliente de uma empresa de telefonia irá ou não manter a sua conta ativa nos próximos 12 meses, a depender do tempo de casa, histórico transacional, chamados abertos etc.

Evento A: O cliente cancelar a conta nos próximos 12 meses.

Probabilidade: P(A) = ?

Exercício

Calculemos a **probabilidade** de ocorrência de alguns eventos:

Experimento 3: Sorteio aleatório de 2 entre 10 colaboradores de um time, que ganharão um *voucher* promocional de um aplicativo de restaurantes. O time é composto

Evento A: Ambos

Probabilidade: P

Note que pode ser bem difícil calcular a probabilidade de **eventos complexos**, especialmente quando eles podem ser influenciados por múltiplos fatores concomitantes.

Experimento 4: Avaliação de se determinado cliente de uma empresa de telefonia irá ou não manter a sua conta ativa nos próximos 12 meses, a depender do tempo de casa, histórico transacional, chamados abertos etc.

Evento A: O cliente cancelar a conta nos próximos 12 meses.

Probabilidade: P(A) = ?

Distribuições de Probabilidade

Quando analisamos variáveis em uma base de dados, podemos calcular as **probabilidades** de observar os seus valores, mediante **amostragem aleatória**.

Vejamos o exemplo a seguir.

Case: Farmácias

Uma rede de farmácias atua em 19 cidades do estado de São Paulo. Deseja-se analisar as informações a seguir, disponíveis em uma base de dados consolidada na visão **cidade**.

Variável	Descrição
CIDADE	Nome da cidade
FLAG_SHOPPING	Indica se a rede possui ao menos uma loja localizada em *shopping centers* na cidade
SEGUNDO_PRINCIPAL_SETOR	Setor com a segunda maior quantidade de produtos vendidos na cidade, após medicamentos
FATURAMENTO_TRI	Faturamento total das lojas da cidade no último trimestre, em milhões de R$

Arquivo: Farmacias.xlsx

Case: Farmácias

Uma rede de farmácias atua em 19 cidades do estado de São Paulo. Deseja-se analisar as informações a seguir, disponíveis em uma base de dados consolidada na visão **cidade**.

CIDADE	FLAG_SHOPPING	SEGUNDO_PRINCIPAL_SETOR	FATURAMENTO_TRI
São Paulo	Sim	Cosméticos	34,8
Campinas	Sim	Cosméticos	28,6
Guarulhos	Sim	Higiene pessoal	27,3
Osasco	Sim	Perfumaria	24,5
São Bernardo do Campo	Sim	Cosméticos	24,1
Ribeirão Preto	Sim	Cosméticos	23,6
Diadema	Não	Perfumaria	21,8
Sorocaba	Sim	Cosméticos	19,8
Praia Grande	Não	Higiene pessoal	19,4
Jundiaí	Não	Perfumaria	18,5
Santo André	Sim	Higiene pessoal	17,6
Santos	Não	Higiene pessoal	16,2
Piracicaba	Sim	Perfumaria	14,9
São José do Rio Preto	Não	Cosméticos	13,4
Limeira	Sim	Perfumaria	13,2
Mogi das Cruzes	Sim	Higiene pessoal	12,6
Suzano	Não	Perfumaria	8,9
Americana	Não	Higiene pessoal	6,3
Franca	Não	Perfumaria	2,1

Arquivo: Farmacias.xlsx

Temos dados disponíveis de **todas as cidades de atuação**. Supondo que não haja interesse em realizar inferência (predizer resultados para outras cidades, ou mesmo para o futuro), consideramos que a base de dados é **populacional**.

Case: Farmácias

Uma rede de farmácias atua em 19 cidades do estado de São Paulo. Deseja-se analisar as informações a seguir, disponíveis em uma base de dados consolidada na visão **cidade**.

CIDADE	FLAG_SHOPPING	SEGUNDO_PRINCIPAL_SETOR	FATURAMENTO_TRI
São Paulo	Sim	Cosméticos	34,8
Campinas	Sim	Cosméticos	28,6
Guarulhos	Sim	Higiene pessoal	27,3
Osasco	Sim	Perfumaria	24,5
São Bernardo do Campo	Sim	Cosméticos	24,1
Ribeirão Preto	Sim	Cosméticos	23,6
Diadema	Não	Perfumaria	21,8
Sorocaba	Sim	Cosméticos	19,8
Praia Grande	Não	Higiene pessoal	19,4
Jundiaí	Não	Perfumaria	18,5
Santo André	Sim	Higiene pessoal	17,6
Santos	Não	Higiene pessoal	16,2
Piracicaba	Sim	Perfumaria	14,9
São José do Rio Preto	Não	Cosméticos	13,4
Limeira	Sim	Perfumaria	13,2
Mogi das Cruzes	Sim	Higiene pessoal	12,6
Suzano	Não	Perfumaria	8,9
Americana	Não	Higiene pessoal	6,3
Franca	Não	Perfumaria	2,1

Arquivo: Farmacias.xlsx

Se selecionarmos uma **amostra aleatória** de 1 cidade, qual a probabilidade de que nela haja atuação em *shopping center*?

- Resposta: 11/19 ≈ 0,58 = **58%**

Representando a variável FLAG_SHOPPING pela letra X, podemos representar a sua **distribuição de probabilidades** como:

$$\begin{cases} P(X = \text{Sim}) = 58\% \\ P(X = \text{Não}) = 42\% \end{cases}$$

Case: Farmácias

Uma rede de farmácias atua em 19 cidades do estado de São Paulo. Deseja-se analisar as informações a seguir, disponíveis em uma base de dados consolidada na visão **cidade**.

CIDADE	FLAG_SHOPPING	SEGUNDO_PRINCIPAL_SETOR	FATURAMENTO_TRI
São Paulo	Sim	Cosméticos	34,8
Campinas	Sim	Cosméticos	28,6
Guarulhos	Sim	Higiene pessoal	27,3
Osasco	Sim	Perfumaria	24,5
São Bernardo do Campo	Sim	Cosméticos	24,1
Ribeirão Preto	Sim	Cosméticos	23,6
Diadema	Não	Perfumaria	21,8
Sorocaba	Sim	Cosméticos	19,8
Praia Grande	Não	Higiene pessoal	19,4
Jundiaí	Não	Perfumaria	18,5
Santo André	Sim	Higiene pessoal	17,6
Santos	Não	Higiene pessoal	16,2
Piracicaba	Sim	Perfumaria	14,9
São José do Rio Preto	Não	Cosméticos	13,4
Limeira	Sim	Perfumaria	13,2
Mogi das Cruzes	Sim	Higiene pessoal	12,6
Suzano	Não	Perfumaria	8,9
Americana	Não	Higiene pessoal	6,3
Franca	Não	Perfumaria	2,1

Arquivo: Farmacias.xlsx

Se selecionarmos uma **amostra aleatória** de 1 cidade, qual a probabilidade de que nela o segundo principal setor seja cosméticos?

- Resposta: 6/19 ≈ 0,32 = **32%**

Representando a variável SEGUNDO_PRINCIPAL_SETOR pela letra Y, podemos representar a sua **distribuição de probabilidades** como:

$$\begin{cases} P(Y = \text{Cosméticos}) = 32\% \\ P(Y = \text{Perfumaria}) = 37\% \\ P(Y = \text{Higiene}) \quad\;\; = 32\% \end{cases}$$

FIA BUSINESS SCHOOL lab.data

Case: Farmácias

Uma rede de farmácias atua em 19 cidades do estado de São Paulo. Deseja-se analisar as informações a seguir, disponíveis em uma base de dados consolidada na visão **cidade**.

CIDADE	FLAG_SHOPPING	SEGUNDO_PRINCIPAL_SETOR	FATURAMENTO_TRI
São Paulo	Sim	Cosméticos	34,8
Campinas	Sim	Cosméticos	28,6
Guarulhos	Sim	Higiene pessoal	27,3
Osasco	Sim	Perfumaria	24,5
São Bernardo do Campo	Sim	Cosméticos	24,1
Ribeirão Preto	Sim	Cosméticos	23,6
Diadema	Não	Perfumaria	21,8
Sorocaba	Sim	Cosméticos	19,8
Praia Grande	Não	Higiene pessoal	19,4
Jundiaí	Não	Perfumaria	18,5
Santo André	Sim	Higiene pessoal	17,6
Santos	Não	Higiene pessoal	16,2
Piracicaba	Sim	Perfumaria	14,9
São José do Rio Preto	Não	Cosméticos	13,4
Limeira	Sim	Perfumaria	13,2
Mogi das Cruzes	Sim	Higiene pessoal	12,6
Suzano	Não	Perfumaria	8,9
Americana	Não	Higiene pessoal	6,3
Franca	Não	Perfumaria	2,1

Arquivo: Farmacias.xlsx

Se selecionarmos uma **amostra aleatória** de 1 cidade, qual a probabilidade de que ela tenha faturamento de 10M a 15M no último tri?

- Resposta: 4/19 ≈ 0,21 = **21%**

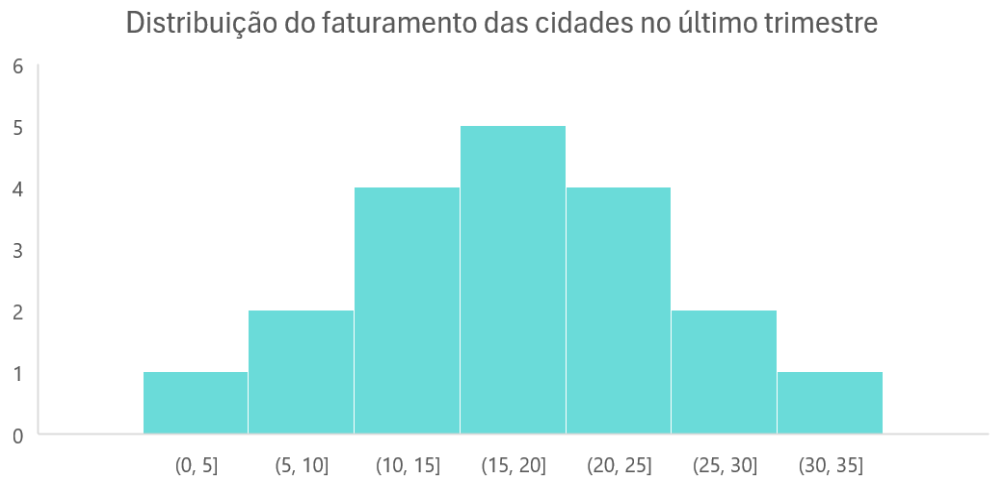
Como podemos representar a **distribuição de probabilidades** da variável FATURAMENTO_TRI?

FIA BUSINESS SCHOOL lab.data

Case: Farmácias

Uma rede de farmácias atua em 19 cidades do estado de São Paulo. Deseja-se analisar as informações a seguir, disponíveis em uma base de dados consolidada na visão **cidade**.

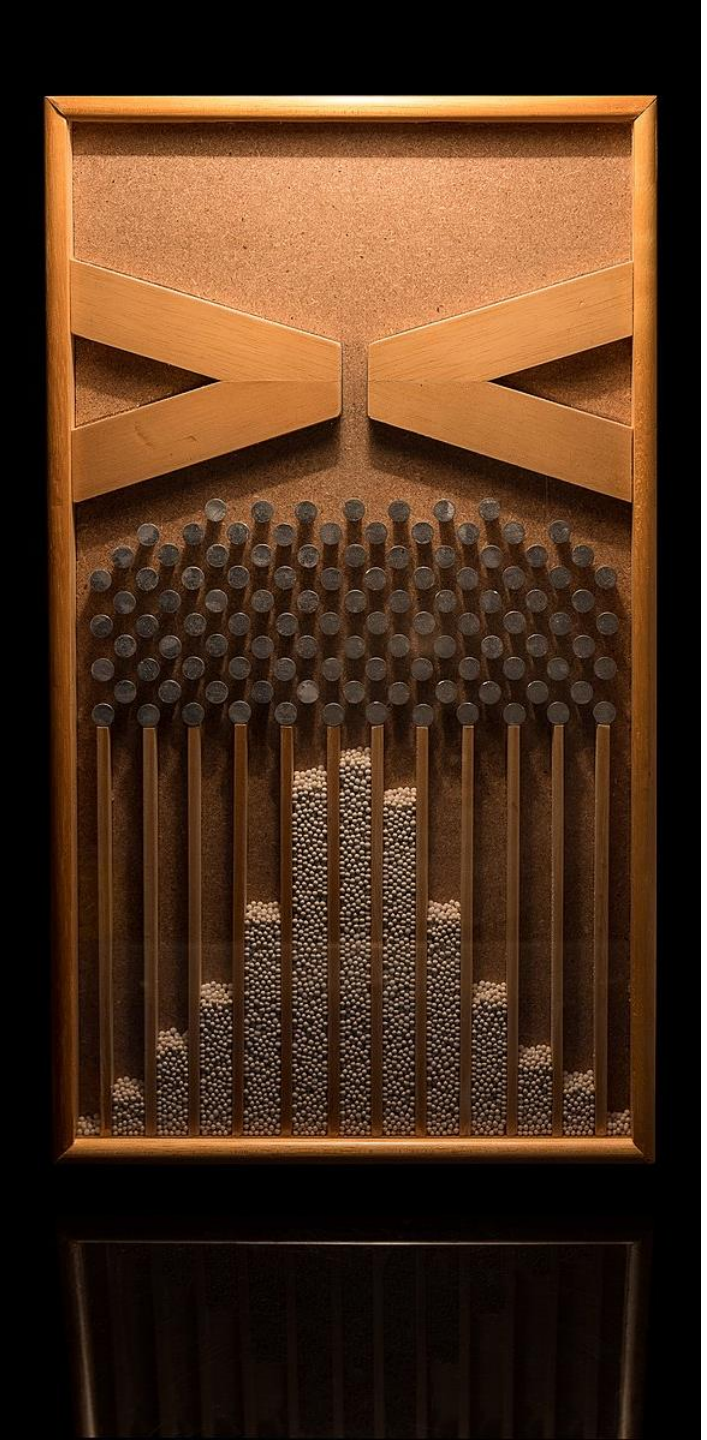
CIDADE	FLAG_SHOPPING	SEGUNDO_PRINCIPAL_SETOR	FATURAMENTO_TRI
São Paulo	Sim	Cosméticos	34,8
Campinas	Sim	Cosméticos	28,6
Guarulhos	Sim	Higiene pessoal	27,3
Osasco	Sim	Perfumaria	24,5
São Bernardo do Campo	Sim	Cosméticos	24,1
Ribeirão Preto	Sim	Cosméticos	23,6
Diadema	Não	Perfumaria	21,8
Sorocaba	Sim	Cosméticos	19,8
Praia Grande	Não	Higiene pessoal	19,4
Jundiaí	Não	Perfumaria	18,5
Santo André	Sim	Higiene pessoal	17,6
Santos	Não	Higiene pessoal	16,2
Piracicaba	Sim	Perfumaria	14,9
São José do Rio Preto	Não	Cosméticos	13,4
Limeira	Sim	Perfumaria	13,2
Mogi das Cruzes	Sim	Higiene pessoal	12,6
Suzano	Não	Perfumaria	8,9
Americana	Não	Higiene pessoal	6,3
Franca	Não	Perfumaria	2,1



Média = R$ **18,3 M**
Desvio padrão = R$ **7,8 M**

Arquivo: Farmacias.xlsx

FIA BUSINESS SCHOOL lab.data

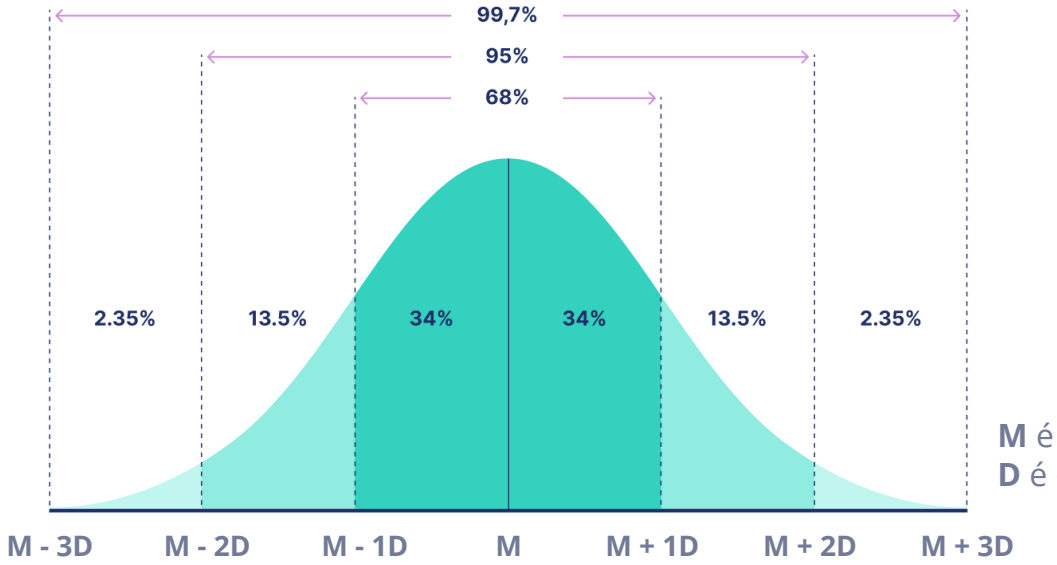


Distribuição Normal

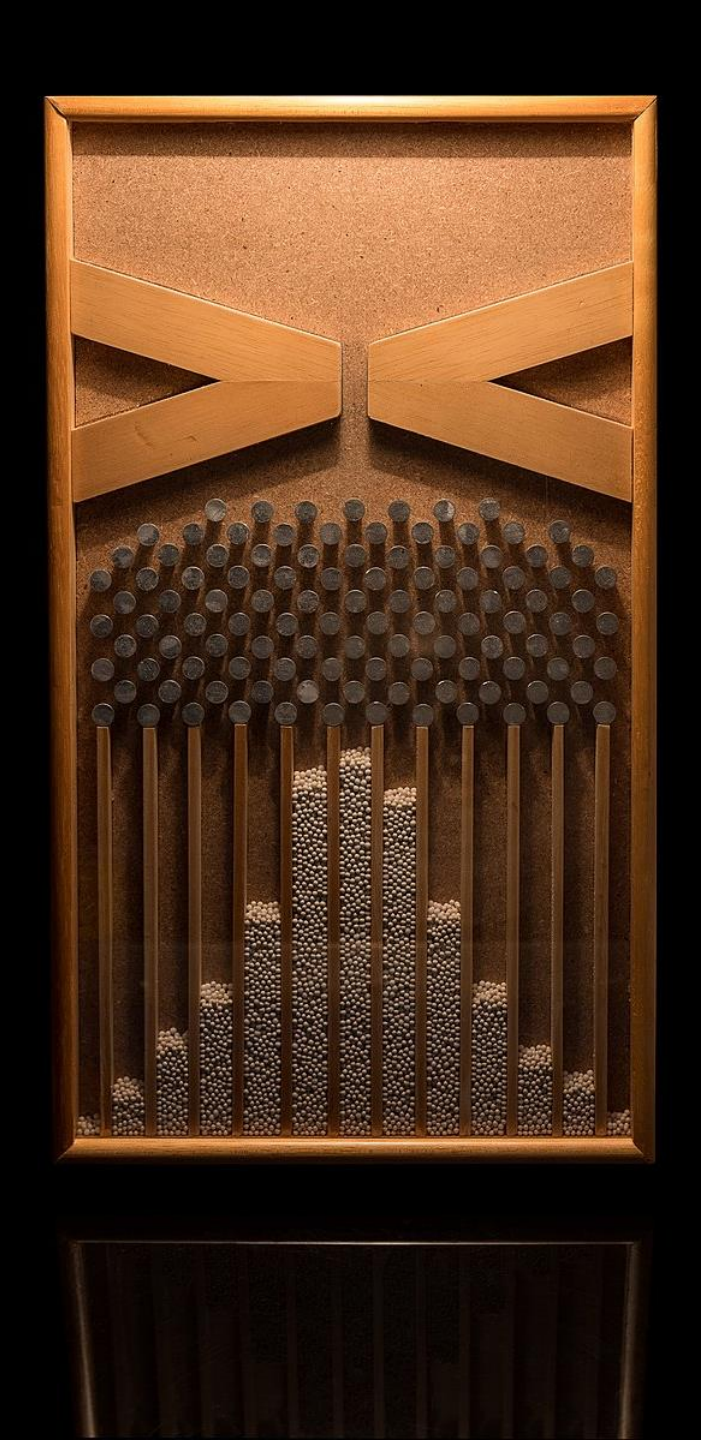
Quando uma variável é quantitativa contínua, uma das possíveis distribuições de probabilidade é a distribuição **normal**.

A distribuição normal é caracterizada por:

- Um formato de **curva** que cresce e decresce (como um "sino"), com alta concentração de probabilidades em torno da **média**.
- Um decaimento **simétrico** de probabilidades para valores que se distanciam da média, sendo que a intensidade do decaimento depende do **desvio padrão**.



M é a **média**
D é o **desvio padrão**



Distribuição Normal

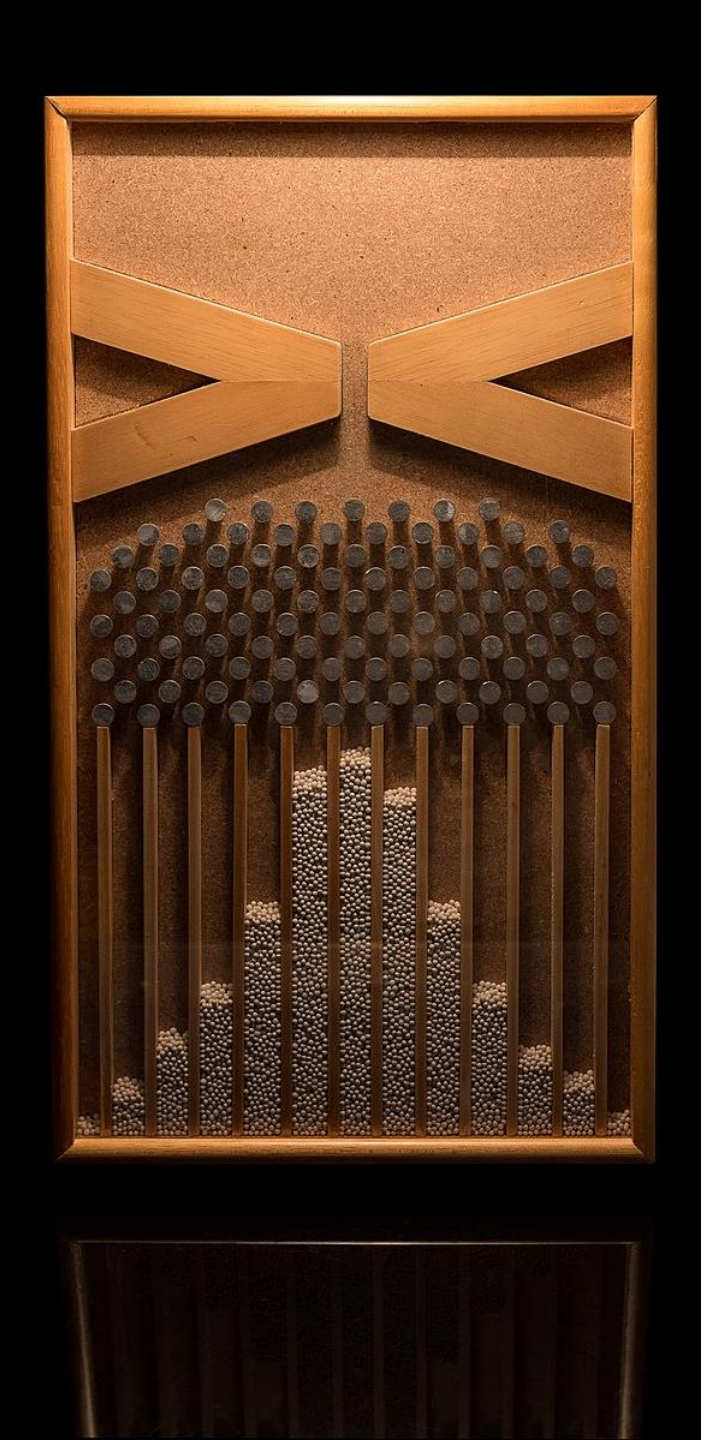
A distribuição normal possui esta denominação pois ela aparece de forma **natural** (ou “normal”) em diversos contextos ao nosso redor.

Exemplos:

- ✓ Distribuição de **aspectos biológicos** de seres humanos (ex.: peso, altura).
- ✓ Distribuição de **aspectos biológicos** de animais e vegetais.
- ✓ Distribuição de **características** de minerais (ex.: concentração de elementos).
- ✓ Distribuição de **temperaturas diárias** de uma região.
- ✓ Distribuição de **características de produtos industriais** (ex.: dimensões).
- ✓ Distribuição de **erros de medida** em experimentos físicos (ex.: gravidade).

Ou seja, a distribuição normal tende a surgir naturalmente em fenômenos da natureza que possuem uma **tendência** (definida pela média), mas envolvem **aleatoriedade** (definida pelo desvio padrão).

Este comportamento é representado a partir do ***Galton Board*** (imagem ao lado).



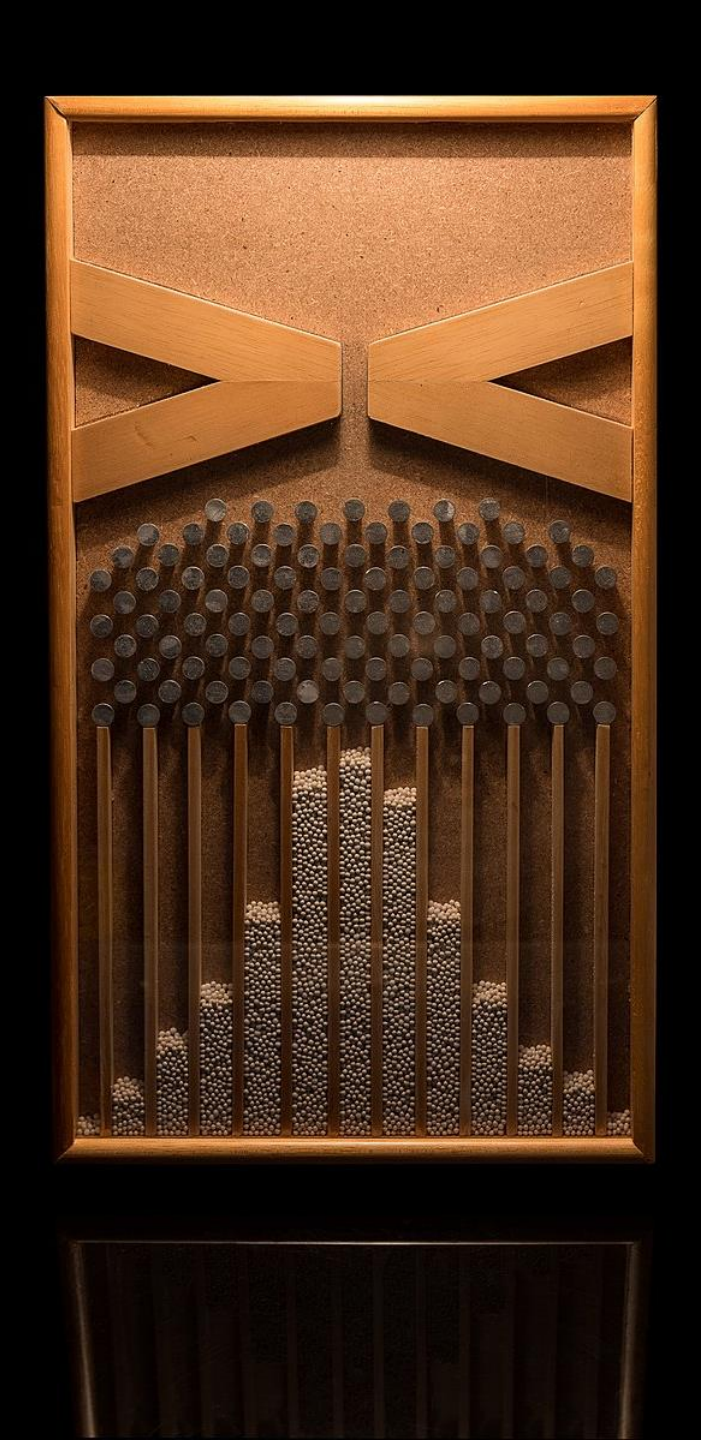
Distribuição Normal

Consideremos o seguinte exemplo:

Experimento: Selecionar aleatoriamente um paciente, após 1 mês em internação hospitalar para tratamento de câncer, e avaliar a sua oscilação de peso desde o início da internação.

Suposição: A distribuição populacional das oscilações de peso após 1 mês de tratamento deste câncer é **normal**, com média -4kg e desvio padrão 3kg.

Variável: Diferença de peso entre a data da internação e a atual (em kg).



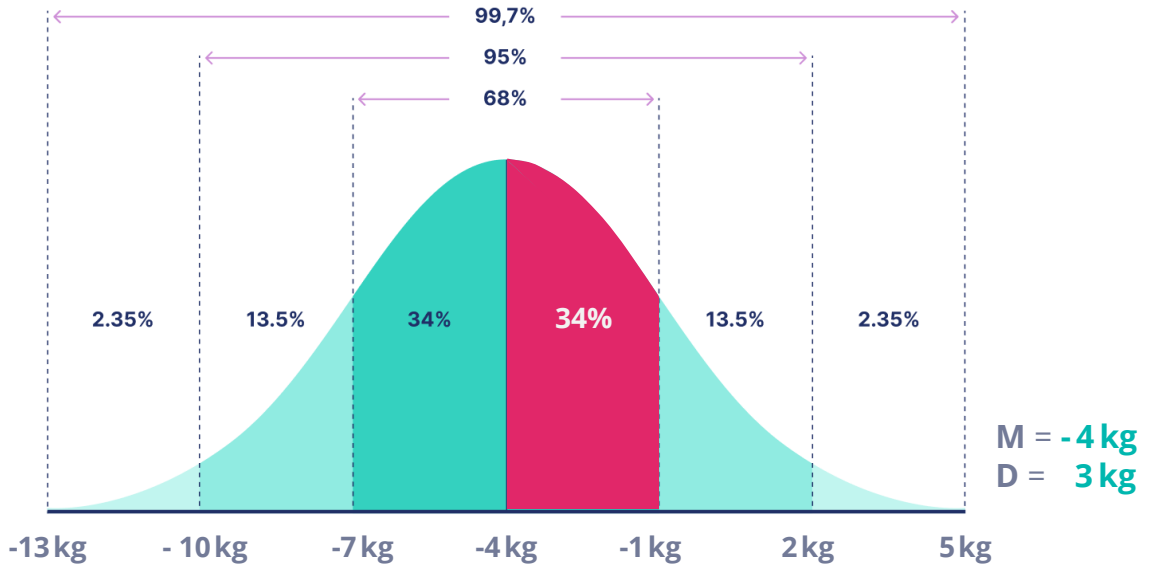
Distribuição Normal

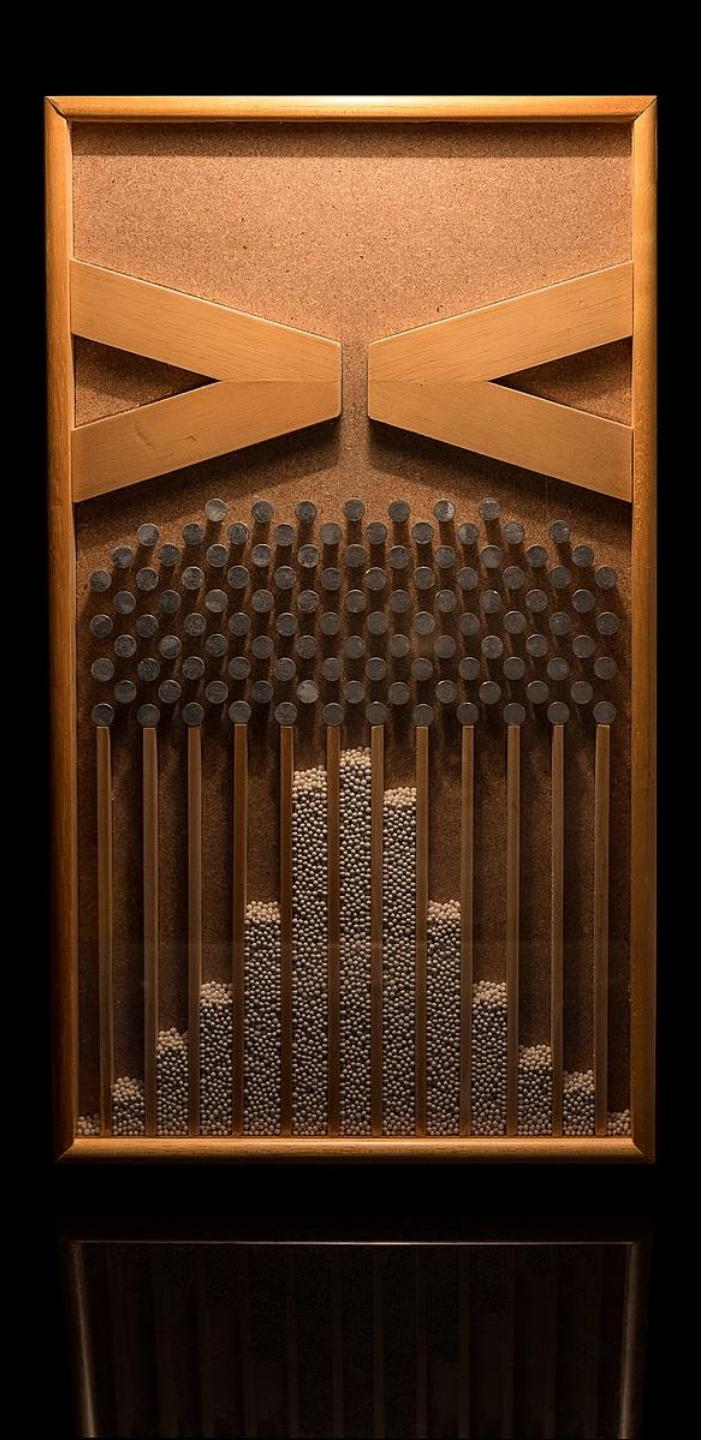
Como a variável é **contínua**, não é possível estabelecer uma probabilidade individual para cada um dos seus infinitos valores possíveis.

Em uma distribuição contínua, tal como a normal, só podemos calcular probabilidades com base em **intervalos**.

Exemplo: Qual a probabilidade de observar um paciente cuja perda de peso tenha sido de 1 kg a 4kg no último mês?

Resposta: 34%



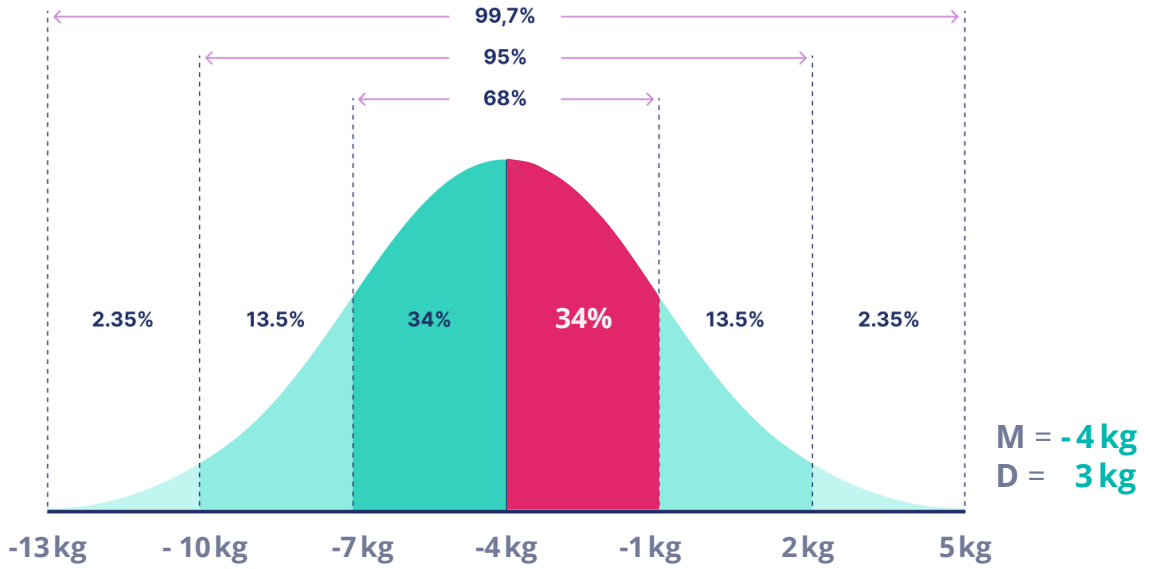


Distribuição Normal

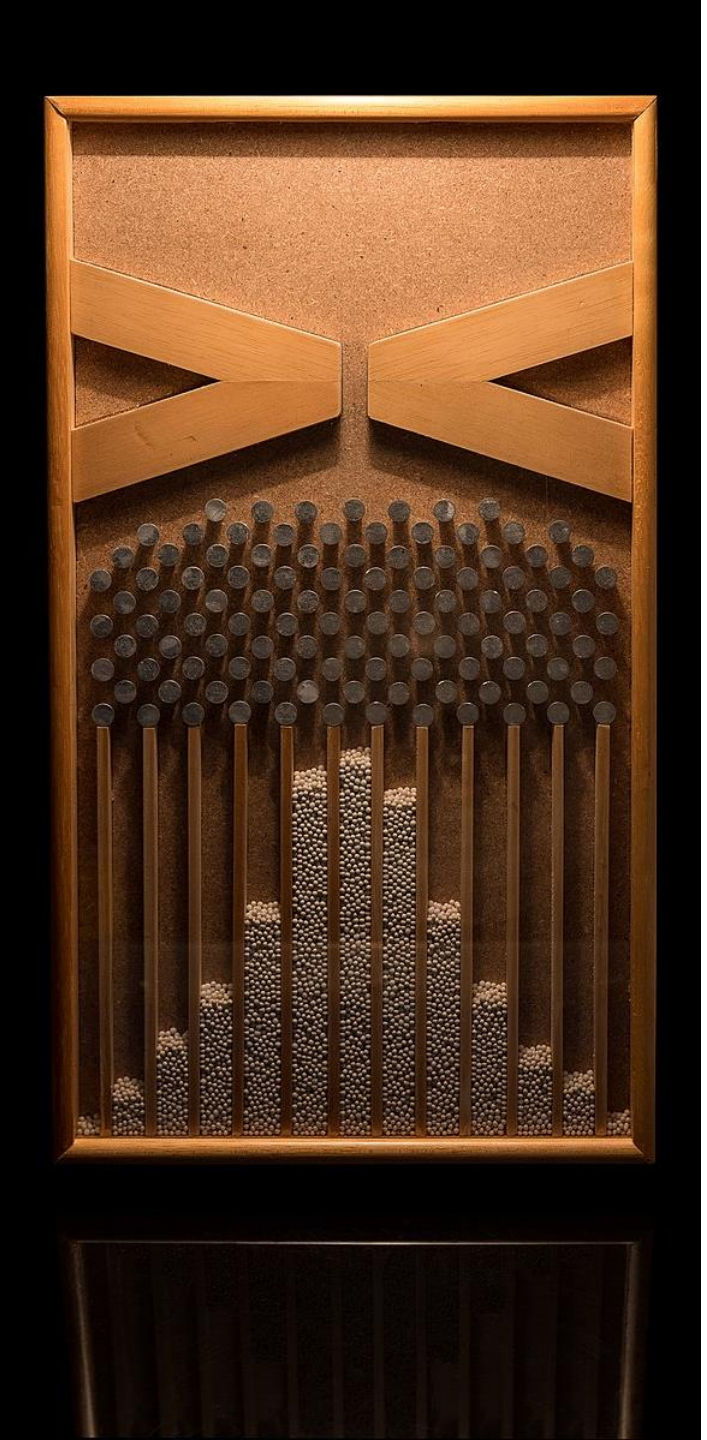
A probabilidade de 34% corresponde ao tamanho da **área abaixo da curva** da distribuição normal. Ela é passível de ser calculada matematicamente a partir de **integrais**. Este é um tópico bastante complexo, que não vamos aprofundar em nosso curso.

É possível calcular facilmente as probabilidades a partir do **Excel**, pela função DIST.NORM.N, que retorna a probabilidade acumulada **abaixo** de algum valor.

Neste exemplo: **= DIST.NORM.N(-1; -4; 3; 1) – DIST.NORM.N(-4; -4; 3; 1)** = 34%.



FIA BUSINESS SCHOOL lab.data



Distribuição Normal

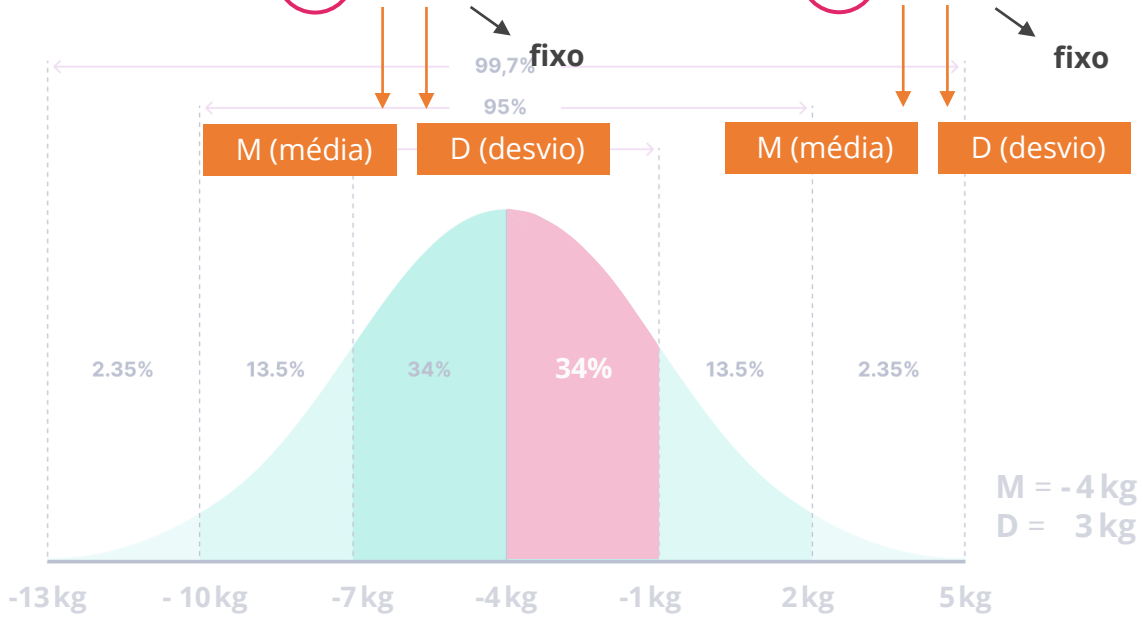
A probabilidade de 34% corresponde ao tamanho da **área abaixo da curva** da distribuição normal. Ela é passível de ser calculada matematicamente a partir de **integrais**. Este é um tópico bastante complexo, que não vamos aprofundar em nosso curso.

É possível calcu des
DIST.NORM.N, que retorna a probabilidade acumulada **abaixo** de algum valor.

Probabilidade de observar um paciente com oscilação de peso ≤ -1 kg.

Probabilidade de observar um paciente com oscilação de peso ≤ -4 kg.

Neste exemplo: **= DIST.NORM.N(-1; -4; 3; 1) – DIST.NORM.N(-4; -4; 3; 1)** = 34%.



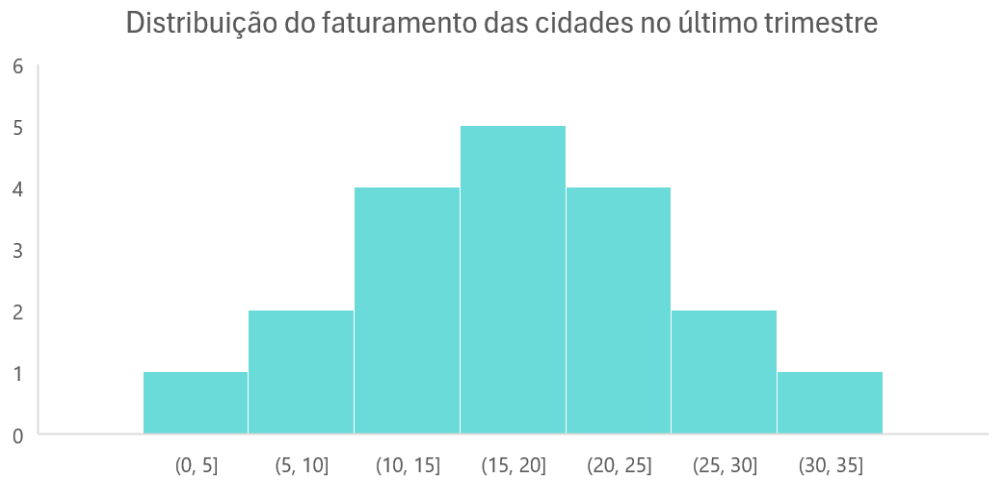
FIA BUSINESS SCHOOL lab.data

Case: Farmácias

Uma rede de farmácias atua em 19 cidades do estado de São Paulo. Deseja-se analisar as informações a seguir, disponíveis em uma base de dados consolidada na visão **cidade**.

CIDADE	FLAG_SHOPPING	SEGUNDO_PRINCIPAL_SETOR	FATURAMENTO_TRI
São Paulo			
Campinas			
Guarulhos			
Osasco			
São Bernardo do Campo			
Ribeirão Preto			
Diadema			
Sorocaba	Sim	Cosméticos	18
Praia Grande	Não	Higiene pessoal	
Jundiaí	Não	Perfumaria	18,5
Santo André	Sim	Higiene pessoal	17,6
Santos	Não	Higiene pessoal	16,2
Piracicaba	Sim	Perfumaria	14,9
São José do Rio Preto	Não	Cosméticos	13,4
Limeira	Sim	Perfumaria	13,2
Mogi das Cruzes	Sim	Higiene pessoal	12,6
Suzano	Não	Perfumaria	8,9
Americana	Não	Higiene pessoal	6,3
Franca	Não	Perfumaria	2,1

Portanto, a variável FATURAMENTO_TRI segue uma distribuição de probabilidades **aproximadamente normal**, com média 18,3M e desvio padrão 7,8M.



Média = R$ **18,3 M**
Desvio padrão = R$ **7,8 M**

Arquivo: Farmacias.xlsx

FIA BUSINESS SCHOOL lab.data

Outras Distribuições

Existem inúmeras **distribuições de probabilidade** utilizadas em estatística, que vão além do nosso escopo neste curso.

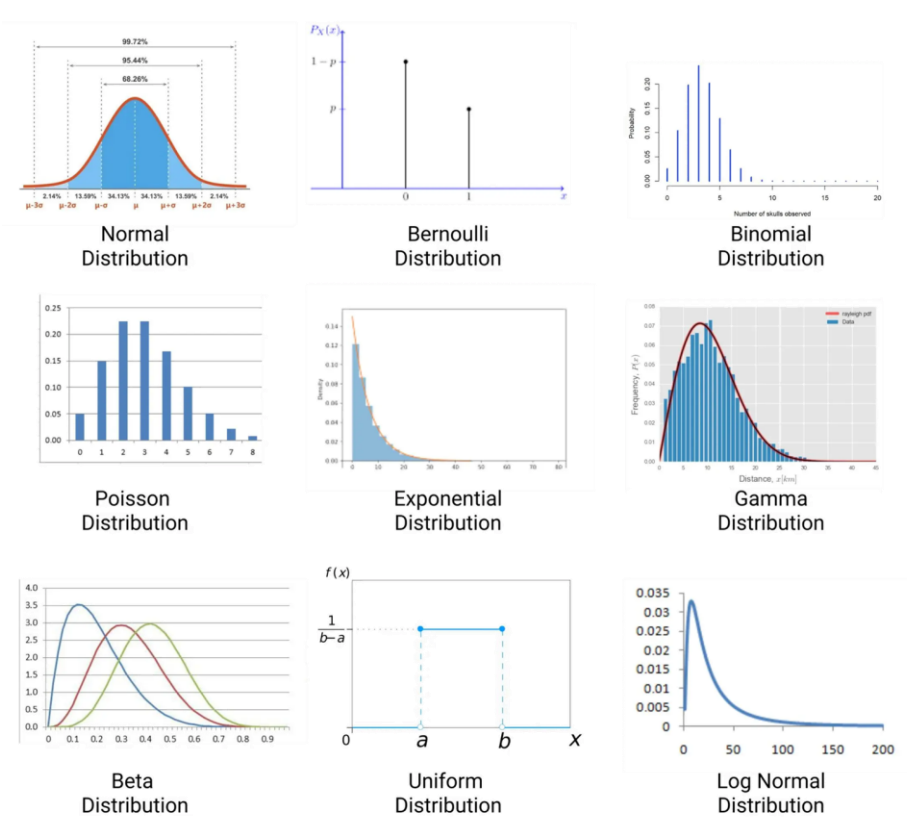
As principais distribuições são:

Distribuições discretas

- ✓ Uniforme discreta
- ✓ Bernoulli
- ✓ Binomial
- ✓ Poisson
- ✓ Geométrica
- ✓ Hipergeométrica

Distribuições contínuas

- ✓ Normal
- ✓ Log normal
- ✓ t de Student
- ✓ Uniforme contínua
- ✓ Exponencial
- ✓ Gama
- ✓ Beta



***Fonte**: https://datasciencedojo.com/wp-content/uploads/9-data-science-distributions-scaled.jpg*

FIA BUSINESS SCHOOL lab.data

Exercícios: Probabilidade em Distribuição Normal

1. Um fabricante produz pacotes de café de 500g em uma linha de produção automatizada. O peso dos pacotes segue uma distribuição normal, com média de 500g e desvio padrão de 10g. Qual a probabilidade de que um consumidor compre um pacote de café que contenha menos de 485g?

2. Em uma agência bancária, o tempo de resposta dos caixas eletrônicos para realizar a leitura de cartões segue uma distribuição normal, com média de 4 segundos e desvio padrão de 1,1 segundos. Qual a probabilidade de que um cliente tenha que esperar mais de 6 segundos para ler seu cartão?

3. O tempo de vida de lâmpadas LED produzidas por um fabricante segue uma distribuição normal, com média de 20.000 horas e desvio padrão de 2.000 horas. Qual a probabilidade de que um consumidor adquira uma lâmpada que dure entre 17.000 horas e 21.000 horas?

4. Tarefa da Inferência Estatística

Tarefa da Inferência Estatística

Em todos os exemplos iniciais sobre os quais falamos de **probabilidade**, bem como no *case* de farmácias, realizamos a suposição (direta ou indireta) de que conhecíamos toda a **população**.

Evento **A**: *Chover hoje à tarde* P(**A**) = **70%**

Entre **todas** as possibilidades de configuração de clima avaliadas pelo instituto de meteorologia, **70%** delas resultam em chuva hoje à tarde.

Evento **B**: *Chegar a tempo para a primeira reunião do dia* P(**B**) = **0%**

Entre **todas** as possibilidades de rearranjo de trânsito de veículos, abertura de semáforos, incidentes etc., **0%** delas me levam a chegar a tempo para a reunião.

Evento **C**: *O time adversário iniciar o jogo, após cara ou coroa* P(**C**) = **50%**

Entre **todas** as possibilidades de lançamento da moeda, ignorando fatores como força, velocidade e direção do lançamento, **50%** delas resultam em cara (ou coroa).

Evento **D**: *Ficar em casa no próximo feriado* P(**D**) = **100%**

Entre **todas** as possibilidades de planos pessoais que venham à minha cabeça para o próximo feriado, **100%** delas incluem ficar em casa para descansar.

Tarefa da Inferência Estatística

Quando a população é **conhecida**, já sabemos o que esperar de eventuais amostras aleatórias provenientes dela, antes mesmo de observá-las.

Já quando a população é **desconhecida**, nosso objetivo passa a ser o de **inferir** características relevantes da população a partir de uma amostra, conforme definimos anteriormente.

Para desenvolver um conhecimento científico abrangente e/ou tomar decisões de negócio acuradas, o nosso interesse nunca está nas amostras, mas na **população**!

Proporções e Médias

Nesta aula, vamos estudar o problema de estimar **dois tipos** de características populacionais desconhecidas, mediante amostragem.

Proporções (para variáveis qualitativas)

$$p \rightarrow \hat{p}$$

Vamos estimar uma **proporção populacional** p
a partir da **proporção amostral** $\hat{p}$

Médias (para variáveis quantitativas)

$$\mu \rightarrow \hat{\mu}$$

Vamos estimar uma **média populacional** μ
a partir da **média amostral** $\hat{\mu}$

FIA BUSINESS SCHOOL lab.data

Obs.: A letra grega μ se lê como “mu”.

Proporções e Médias

Nesta aula, vamos estudar o problema de estimar **dois tipos** de características populacionais desconhecidas, mediante amostragem.

Proporções (para variáveis qualitativas)

- ✓ Qual a proporção de clientes que estão satisfeitos?
- ✓ Qual a proporção de consumidores que preferem o produto A ao B?
- ✓ Qual a proporção de pessoas que possuem determinada doença?
- ✓ Qual a proporção de lares que assistem a determinado programa de TV?

Médias (para variáveis quantitativas)

- ✓ Qual a média mensal de vezes que os clientes vão a um supermercado?
- ✓ Qual a média de idade dos veículos que circulam em uma cidade?
- ✓ Qual a média diária de calorias ingeridas pelos habitantes de uma região?
- ✓ Qual a média de gastos mensais (R$) das famílias brasileiras?

5. Intervalo de Confiança

Teorema do Limite Central (TLC)

O **Teorema do Limite Central** (TLC) é uma formulação matemática a partir da qual é possível chegar a um importante resultado para a inferência estatística.

Tanto a **proporção amostral** quanto a **média amostral** seguem uma distribuição **aproximadamente normal**, independentemente da distribuição populacional da característica em questão, quando a amostra é **aleatória** e grande.

Além disso, essa distribuição está centrada no real valor da proporção ou média **populacional**.

Note que existem inúmeras amostras aleatórias possíveis de serem retiradas de uma população, e cada uma fornecerá um **valor diferente** de proporção amostral ($\hat{p}$) e/ou de média amostral ($\hat{\mu}$).

O que o TLC afirma é que esses possíveis valores de $\hat{p}$ (ou $\hat{\mu}$) estão distribuídos de forma simétrica, de acordo com uma normal, em torno do **real valor populacional** (p ou μ) que está sendo estimado, desde que se trate de uma amostra aleatória grande – com, pelo menos, **30 observações**.

FIA BUSINESS SCHOOL lab.data

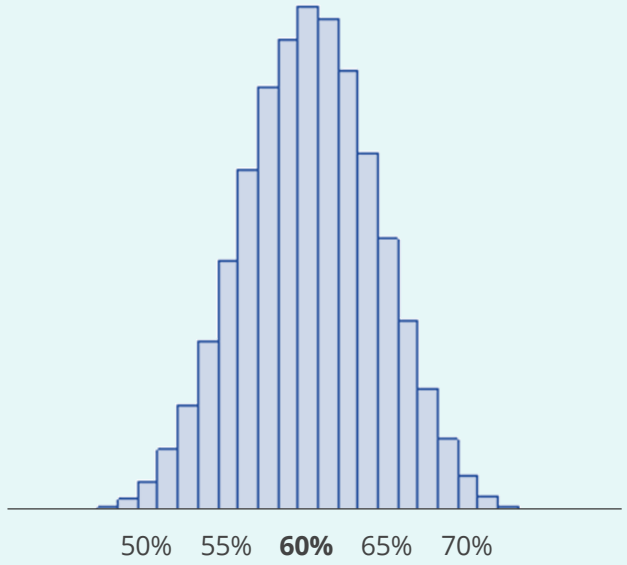
Teorema do Limite Central (TLC)

Exemplo 1: Qual a proporção (p) de consumidores que preferem o produto A ao B?

Não conhecemos o valor de p.

De forma hipotética, suponha que extraíssemos 1.000 amostras aleatórias (cada uma com pelo menos 30 consumidores). Consequentemente, obteríamos 1.000 estimativas amostrais $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{1000}$.

Histograma das proporções amostrais ($\hat{p}$) de consumidores que preferem o produto A ao B



- Note que a distribuição normal não é exata. Em teoria, a aproximação é melhor conforme os tamanhos amostrais *(n)* aumentam.
- Qual seria uma boa estimativa para a proporção populacional (p)?

  Resposta: **60%**

FIA BUSINESS SCHOOL lab.data

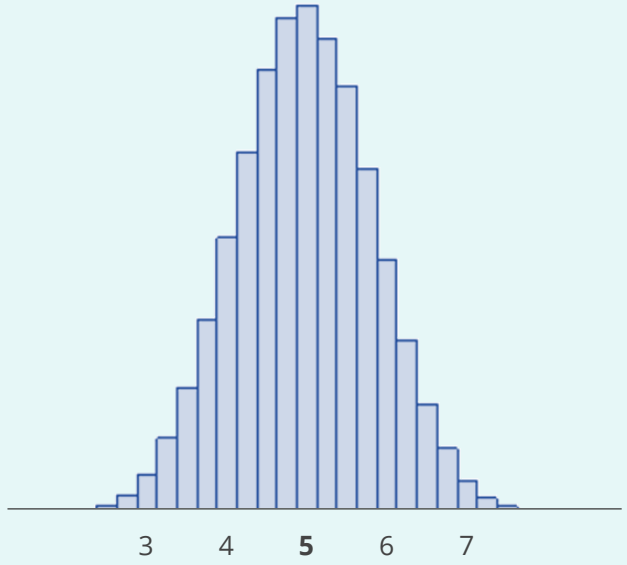
Teorema do Limite Central (TLC)

Exemplo 2: Qual a média (μ) de idade dos veículos que circulam em uma cidade?

Não conhecemos o valor de μ.

De forma hipotética, suponha que extraíssemos 500 amostras aleatórias (cada uma com pelo menos 30 veículos). Consequentemente, obteríamos 500 estimativas amostrais $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_{500}$.

Histograma das médias amostrais ($\hat{\mu}$) de idade dos veículos que circulam na cidade, em anos



- Note que a distribuição normal não é exata. Em teoria, a aproximação é melhor conforme os tamanhos amostrais *(n)* aumentam.
- Qual seria uma boa estimativa para a média populacional (μ)?

  Resposta: **5 anos**

FIA BUSINESS SCHOOL lab.data

Estimação Intervalar: Proporções

Tendo em vista que a **estimativa pontual** ($\hat{p}$ ou $\hat{\mu}$) de uma amostra específica não é uma verdade absoluta, por conta da aleatoriedade da amostra, podemos estabelecer um **intervalo** de valores plausíveis para o parâmetro populacional.

Para isso, utilizamos o fato de $\hat{p}$ (ou $\hat{\mu}$) seguir uma distribuição aproximadamente normal, cujas probabilidades são **conhecidas**, bem como o fato de essa distribuição estar centrada no real valor de p (ou μ).

Intervalo de confiança para a proporção p

$$IC(p; 95\%) = \left[\hat{p} \pm 1{,}96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

onde n é o tamanho amostral

FIA BUSINESS SCHOOL lab.data

Estimação Intervalar: Proporções

Tendo em vista que a **estimativa pontual** ($\hat{p}$ ou $\hat{\mu}$) de uma amostra específica não é uma verdade absoluta, por conta da aleatoriedade da amostra, podemos estabelecer um **intervalo** de valores plausíveis para o parâmetro populacional.

Para isso, utilizamos o fato de $\hat{p}$ (ou $\hat{\mu}$) seguir uma distribuição aproximadamente normal, cujas probabilidades são **conhecidas**, bem como o fato de essa distribuição estar centrada no real valor de p (ou μ).

Intervalo de confiança para a proporção p

$$IC(p;95\%) = \left[\hat{p} \pm 1{,}96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Estimativa amostral pontual

Margem de erro

Estimação Intervalar: Proporções

Tendo em vista que a **estimativa pontual** ($\hat{p}$ ou $\hat{\mu}$) de uma amostra específica não é uma verdade absoluta, por conta da aleatoriedade da amostra, podemos estabelecer um **intervalo** de valores plausíveis para o parâmetro populacional.

Para isso, utilizamos o fato de $\hat{p}$ (ou $\hat{\mu}$) seguir uma distribuição aproximadamente normal, cujas probabilidades são **conhecidas**, bem como o fato de essa distribuição estar centrada no real valor de p (ou μ).

Intervalo de confiança para a proporção p

$$IC(p;95\%) = \left[\hat{p} \pm 1{,}96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Interpretação: Temos **95% de probabilidade** de que o intervalo acima, centrado na estimativa amostral $\hat{p}$, acrescida/subtraída de uma margem de erro, contenha o verdadeiro valor (populacional e desconhecido) de p.

FIA BUSINESS SCHOOL lab.data

Estimação Intervalar: Médias

Tendo em vista que a **estimativa pontual** ($\hat{p}$ ou $\hat{\mu}$) de uma amostra específica não é uma verdade absoluta, por conta da aleatoriedade da amostra, podemos estabelecer um **intervalo** de valores plausíveis para o parâmetro populacional.

Para isso, utilizamos o fato de $\hat{p}$ (ou $\hat{\mu}$) seguir uma distribuição aproximadamente normal, cujas probabilidades são **conhecidas**, bem como o fato de essa distribuição estar centrada no real valor de p (ou μ).

Intervalo de confiança para a média μ

$$IC(\mu; 95\%) = \left[\hat{\mu} \pm 1{,}96 \cdot \frac{s}{\sqrt{n}} \right]$$

onde s é o desvio padrão amostral e n é o tamanho amostral

Estimação Intervalar: Médias

Tendo em vista que a **estimativa pontual** ($\hat{p}$ ou $\hat{\mu}$) de uma amostra específica não é uma verdade absoluta, por conta da aleatoriedade da amostra, podemos estabelecer um **intervalo** de valores plausíveis para o parâmetro populacional.

Para isso, utilizamos o fato de $\hat{p}$ (ou $\hat{\mu}$) seguir uma distribuição aproximadamente normal, cujas probabilidades são **conhecidas**, bem como o fato de essa distribuição estar centrada no real valor de p (ou μ).

Intervalo de confiança para a média μ

$$IC(\mu; 95\%) = \left[\hat{\mu} \pm 1{,}96 \cdot \frac{s}{\sqrt{n}} \right]$$

Estimativa amostral pontual

Margem de erro

FIA BUSINESS SCHOOL lab.data

Estimação Intervalar: Médias

Tendo em vista que a **estimativa pontual** ($\hat{p}$ ou $\hat{\mu}$) de uma amostra específica não é uma verdade absoluta, por conta da aleatoriedade da amostra, podemos estabelecer um **intervalo** de valores plausíveis para o parâmetro populacional.

Para isso, utilizamos o fato de $\hat{p}$ (ou $\hat{\mu}$) seguir uma distribuição aproximadamente normal, cujas probabilidades são **conhecidas**, bem como o fato de essa distribuição estar centrada no real valor de p (ou μ).

Intervalo de confiança para a média μ

$$IC(\mu; 95\%) = \left[\hat{\mu} \pm 1{,}96 \cdot \frac{s}{\sqrt{n}} \right]$$

Interpretação: Temos **95% de probabilidade** de que o intervalo acima, centrado na estimativa amostral $\hat{\mu}$, acrescida/subtraída de uma margem de erro, contenha o verdadeiro valor (populacional e desconhecido) de μ.

FIA BUSINESS SCHOOL lab.data

Alterando o Nível de Confiança

Quanto **maior a confiança/probabilidade** desejada de que o intervalo contenha o verdadeiro valor do parâmetro populacional p (ou μ), **maior precisa ser o tamanho do intervalo**, e vice-versa.

É possível controlar isso alterando o multiplicador **1,96** nas fórmulas, conforme referências da tabela a seguir.

Nível de confiança desejado	Multiplicador (associado à distribuição normal)
80%	1,28
90%	1,64
95%	1,96
99%	2,58
99,5%	2,81

Case: Internet Banking

Uma empresa bancária realizou uma pesquisa com 2.400 clientes que acessaram o internet banking no último mês, com a seguinte pergunta: "*Qual dos seguintes conteúdos é mais relevante para você?*". As opções de resposta foram:

- Dicas de educação financeira
- Dicas de investimentos
- Dicas de segurança
- Promoções e ofertas especiais

O banco complementou o resultado da pesquisa com a informação de renda mensal declarada (R$) dos clientes respondentes. Com isso, chegou à seguinte base de dados:

Variável	Descrição
ID_CLIENTE	Código de identificação do cliente
CONTEUDO_PREFERIDO	Categoria de conteúdo preferido pelo cliente, com base na pesquisa
RENDA_MENSAL	Renda mensal declarada pelo cliente, em R$

Nosso objetivo é **estimar a proporção** (%) **de clientes** que preferem cada um dos quatro tipos de conteúdo, na carteira completa de 1 milhão de clientes do banco; bem como **estimar a renda média** (R$) associada a cada um desses quatro grupos.

Arquivo: Internet_Banking.xlsx

Case: Internet Banking

Denominando os parâmetros de interesse:

- p_1 é a **proporção** de clientes que preferem receber **dicas de educação financeira**
- μ_1 é a **média** de renda mensal declarada (em R$) entre clientes que preferem receber **dicas de educação financeira**
- p_2 é a **proporção** de clientes que preferem receber **dicas de investimentos**
- μ_2 é a **média** de renda mensal declarada (em R$) entre clientes que preferem receber **dicas de investimentos**
- p_3 é a **proporção** de clientes que preferem receber **dicas de segurança**
- μ_3 é a **média** de renda mensal declarada (em R$) entre clientes que preferem receber **dicas de segurança**
- p_4 é a **proporção** de clientes que preferem receber **promoções e ofertas especiais**
- μ_4 é a **média** de renda mensal declarada (em R$) entre clientes que preferem receber **promoções e ofertas especiais**

Utilizando o Microsoft Excel, podemos chegar aos seguintes tamanhos amostrais (n_1, ..., n_4), bem como estimativas pontuais das proporções ($\hat{p}_1$, ..., $\hat{p}_4$), médias ($\hat{\mu}_1$, ..., $\hat{\mu}_4$) e desvios padrão (s_1, ..., s_4):

- $n_1 = 178$
- $n_2 = 525$
- $n_3 = 708$
- $n_4 = 989$
- $\hat{p}_1 = 7{,}4\%$
- $\hat{p}_2 = 21{,}9\%$
- $\hat{p}_3 = 29{,}5\%$
- $\hat{p}_4 = 41{,}2\%$
- $\hat{\mu}_1 = 5.317$
- $\hat{\mu}_2 = 10.976$
- $\hat{\mu}_3 = 5.942$
- $\hat{\mu}_4 = 4.110$
- $s_1 = 2.914$
- $s_2 = 4.340$
- $s_3 = 3.111$
- $s_4 = 1.462$

Case: Internet Banking

Cálculo da estimativa intervalar (ou intervalo de confiança) para p_1, que é a **proporção** de clientes que preferem receber **dicas de educação financeira**, dado que $\hat{p}_1 = 7{,}4\%$:

$$IC(p_1; 95\%) = \left[\hat{p}_1 \pm 1{,}96 \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n}}\right] = \left[0{,}074 \pm 1{,}96 \cdot \sqrt{\frac{0{,}074 \cdot (1-0{,}074)}{2.400}}\right] = [0{,}064; 0{,}085]$$

Portanto, temos 95% de confiança para afirmar que a **proporção populacional** de clientes que preferem receber dicas de educação financeira está entre 6,4% e 8,5%.

Cálculo da estimativa intervalar (ou intervalo de confiança) para μ_1, que é a **média** de renda mensal declarada (em R$) entre clientes que preferem receber **dicas de educação financeira**, dado que $n_1 = 178$ e $\hat{\mu}_1 = 5.317$:

$$IC(\mu_1; 95\%) = \left[\hat{\mu}_1 \pm 1{,}96 \cdot \frac{s}{\sqrt{n_1}}\right] = \left[5.317 \pm 1{,}96 \cdot \frac{2.914}{\sqrt{178}}\right] = [4.889; 5.745]$$

Portanto, temos 95% de confiança para afirmar que a **média populacional** de renda mensal dos clientes que preferem receber dicas de educação financeira está entre R$ 4.889 e R$ 5.745.

Case: Internet Banking

Cálculo de todos os intervalos de confiança (arquivo: *Internet_Banking.xlsx).*

INTERVALOS DE CONFIANÇA PARA AS PROPORÇÕES

Tamanho amostral	**2.400**
Nível de confiança	**95%**

Conteúdo	Estimativa pontual	Margem de erro	Limite inferior	Limite superior
Dicas de educação financeira	7,4%	1,0 pp	6,4%	8,5%
Dicas de investimentos	21,9%	1,7 pp	20,2%	23,5%
Dicas de segurança	29,5%	1,8 pp	27,7%	31,3%
Promoções e ofertas especiais	41,2%	2,0 pp	39,2%	43,2%

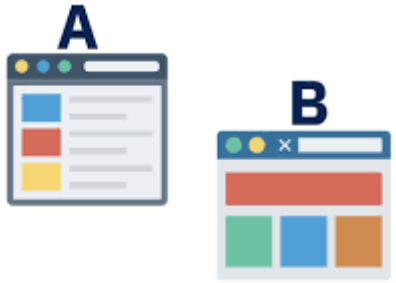
INTERVALOS DE CONFIANÇA PARA AS MÉDIAS

Nível de confiança	**95%**

Conteúdo	Tamanho amostral	Estimativa pontual	Margem de erro	Limite inferior	Limite superior
Dicas de educação financeira	178	R$ 5.317	R$ 428	R$ 4.889	R$ 5.745
Dicas de investimentos	525	R$ 10.976	R$ 371	R$ 10.604	R$ 11.347
Dicas de segurança	708	R$ 5.942	R$ 229	R$ 5.713	R$ 6.171
Promoções e ofertas especiais	989	R$ 4.110	R$ 91	R$ 4.019	R$ 4.201

Exercícios: Intervalo de Confiança

1. Uma empresa de *streaming* de conteúdo musical deseja saber se os usuários que possuem o plano básico já conhecem os benefícios do plano *premium* recém lançado. Para isso, realizaram uma pesquisa junto a uma amostra aleatória de 500 usuários ativos do plano básico, entre os quais 28% responderam que não conheciam o plano *premium*. Construa um intervalo, com 95% de confiança, para a proporção populacional de usuários que ainda não conhecem o plano *premium*. Diminua o nível de confiança para 90% e reavalie o resultado.



2. A área de marketing de um *e-commerce* deseja determinar se existe diferença no tempo médio que os visitantes passam na *home page* do site, a depender de se a diagramação da página é do tipo A ou B. Para isso, realizaram um teste, por meio do qual uma amostra aleatória de 200 visitantes visualizou a diagramação A, e passaram 37 segundos na *home page*, em média; e outra amostra aleatória de 200 visitantes visualizou a diagramação B, tendo ficado 41 segundos na *home page*, em média. O desvio padrão do tempo de ambos os grupos foi de 15 segundos. Construa dois intervalos, com 95% de confiança, para o tempo médio populacional de cada grupo de visitantes. Com base nos resultados, você diria que existe diferença entre as duas estratégias de diagramação?

6. Tópico Extra: Tamanho Amostral

Case: Internet Banking

Cálculo de todos os intervalos de confiança (arquivo: *Internet_Banking.xlsx*).

E se quiséssemos que nossas estimativas possuíssem uma margem de erro de no máximo **1,5 pp**?

INTERVALOS DE C

Tamanho amostral	**2.400**
Nível de confiança	**95%**

Conteúdo	Estimativa pontual	Margem de erro	Limite inferior	Limite superior
Dicas de educação financeira	7,4%	1,0 pp	6,4%	8,5%
Dicas de investimentos	21,9%	1,7 pp	20,2%	23,5%
Dicas de segurança	29,5%	1,8 pp	27,7%	31,3%
Promoções e ofertas especiais	41,2%	2,0 pp	39,2%	43,2%

INTERVALOS DE CONFIANÇA PARA AS MÉDIAS

Nível de confiança	**95%**

Conteúdo	Tamanho amostral	Estimativa pontual	Margem de erro	Limite inferior	Limite superior
Dicas de educação financeira	178	R$ 5.317	R$ 428	R$ 4.889	R$ 5.745
Dicas de investimentos	525	R$ 10.976	R$ 371	R$ 10.604	R$ 11.347
Dicas de segurança	708	R$ 5.942	R$ 229	R$ 5.713	R$ 6.171
Promoções e ofertas especiais	989	R$ 4.110	R$ 91	R$ 4.019	R$ 4.201

FIA BUSINESS SCHOOL lab.data

Definição de Tamanho Amostral

Anteriormente, calculamos margens de erro e intervalos de confiança a partir de uma determinada amostra aleatória, com n observações. Porém, podemos ter o objetivo inverso: como calcular n, o tamanho da amostra, para garantir uma margem de erro de interesse?

Para isso, basta **reescrever as fórmulas anteriores**, isolando o valor de n e deixando-o em função da margem de erro.

Tamanho amostral para estimar uma proporção p

$$n = \frac{1{,}96^2 \cdot \hat{p}(1-\hat{p})}{ME^2}$$

Tamanho amostral para estimar uma média μ

$$n = \frac{1{,}96^2 \cdot s^2}{ME^2}$$

onde ME é a margem de erro amostral e s é o desvio padrão amostral

FIA BUSINESS SCHOOL lab.data

Definição de Tamanho Amostral

As fórmulas anteriores consideram **95% de confiança**. Para aumentar ou diminuir o nível de confiança, tal como antes, podemos alterar o multiplicador **1,96** por outros valores de referência associados à distribuição normal.

Nível de confiança desejado	Multiplicador (associado à distribuição normal)
80%	1,28
90%	1,64
95%	1,96
99%	2,58
99,5%	2,81

Case: Internet Banking

Cálculos de tamanho amostral necessário para reduzir as margens de erro das **proporções** para **1,5 pp**.

TAMANHO AMOSTRAL NECESSÁRIO PARA REDUZIR O ERRO PARA AS PROPORÇÕES

Tamanho amostral inicial	**2.400**
Nível de confiança	**95%**

Conteúdo	Estimativa pontual	Margem de erro	Margem de erro desejada	Tamanho amostral ideal
Dicas de educação financeira	7,4%	1,0 pp	-	-
Dicas de investimentos	21,9%	1,7 pp	1,5 pp	**2.918**
Dicas de segurança	29,5%	1,8 pp	1,5 pp	**3.551**
Promoções e ofertas especiais	41,2%	2,0 pp	1,5 pp	**4.136**

Exemplo: proporção de clientes que preferem receber **dicas de investimentos** (p_2):

$$n = \frac{1{,}96^2 \cdot \hat{p}_2(1-\hat{p}_2)}{ME^2} = \frac{1{,}96^2 \cdot 0{,}219 \cdot (1-0{,}219)}{0{,}015^2} = 2.918$$

Interpretação: A partir dos cenários calculados, é necessário que o tamanho amostral seja aumentado de 2.400 para **4.136 respondentes**, a fim de que tenhamos 95% de confiança de que as proporções estimadas na amostra tenham uma margem de erro de, no máximo, 1,5 pp.

Cálculos de tamanho amostral necessário para reduzir as margens de erro das **médias** para **300 reais**.

TAMANHO AMOSTRAL NECESSÁRIO PARA REDUZIR O ERRO PARA AS MÉDIAS	
Nível de confiança	**95%**

Conteúdo	Tamanho amostral inicial	Desvio padrão	Margem de erro	Margem de erro desejada	Tamanho amostral ideal
Dicas de educação financeira	178	R$ 2.914	R$ 428	R$ 300	363
Dicas de investimentos	525	R$ 4.340	R$ 371	R$ 300	804
Dicas de segurança	708	R$ 3.111	R$ 229	-	-
Promoções e ofertas especiais	989	R$ 1.462	R$ 91	-	-

Exemplo: renda média mensal de clientes que preferem receber **dicas de educação financeira** (μ_1):

$$n = \frac{1{,}96^2 \cdot s^2}{ME^2} = \frac{1{,}96^2 \cdot 2.914^2}{300^2} = 363$$

Interpretação: É necessário que o tamanho amostral seja aumentado de tal forma que se observe ao menos 363 clientes que prefiram receber dicas de educação financeira, o que representa um aumento de **104%** (363/178 - 1) no tamanho amostral desse grupo e, consequentemente, dos demais. Isso resulta num aumento de 2.400 para **4.894 respondentes** na amostra total, a fim de que tenhamos 95% de confiança de que todas as médias estimadas tenham uma margem de erro de, no máximo, **300 reais**.

Exercícios: Tamanho Amostral

1. Com base em levantamentos anteriores, um fabricante de sorvetes acredita que cerca de metade dos consumidores preferem o sabor chocolate em vez do sabor morango. Para estimar melhor esse percentual, a empresa entrevistará uma amostra de clientes a respeito de sua preferência. Calcule o tamanho amostral ideal para que a proporção estimada tenha uma margem de erro de 2 pp para mais ou para menos, com 95% de confiança. Considere diferentes cenários de proporção amostral, de 40% a 60%.

2. O departamento governamental de saúde de uma região conduzirá um estudo clínico para avaliar o IMC da população em internação cardíaca, dividindo-a entre portadores e não portadores de hipertensão arterial. Supondo que o desvio padrão amostral do IMC seja de 4 unidades entre não hipertensos e de 6 unidades entre hipertensos, calcule os tamanhos amostrais necessários para estimar o IMC médio em cada um dos dois grupos, com uma margem de erro de 1 unidade para mais ou para menos, com 90% de confiança.

Referências Bibliográficas

INFERÊNCIA ESTATÍSTICA

- Anderson, R. A. et al. *Estatística Aplicada a Administração e Economia.* 5ª edição. Cengage, 2021.
- Bussab, W. O., Morettin, P. A. *Estatística Básica*. 9ª edição. Saraiva Uni, 2017.
- Illowsky, B., Dean, S. *Introductory Statistics*. Open Stax, 2018.

  Download gratuito em *https://openstax.org/details/books/introductory-statistics*

lab.data

http://labdata.fia.com.br
Instagram: @labdatafia
Facebook: @LabdataFIA