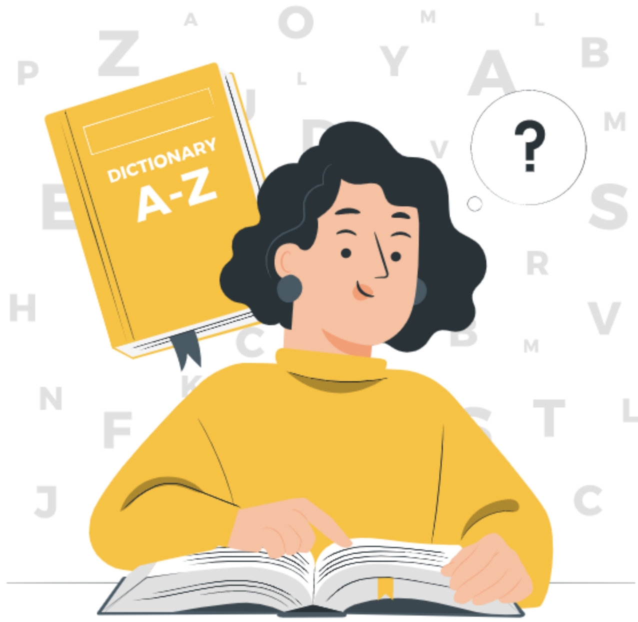
GLOSSÁRIO DE SÍMBOLOS



ITEM DE UMA SEQUÊNCIA

Se temos um conjunto de dados, cada item pode ser representado como sendo x1, x2, x3, etc. Por exemplo, temos o conjunto 10, 21, 32, 43, 58, então:

x1 = 10, x2 = 21, x3 = 32, x4= 43, x5= 58

SOMATÓRIO - ∑

Em matemática, somatório ou somatória é a adição de uma sequência de quaisquer tipos de números. O resultado é sua soma ou total.

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

O índice i significa a partir de qual elemento devemos começar a soma. O símbolo n indica que, nesse caso, vamos somar a sequencia toda - que tem um total de n elementos.

Vamos a alguns exemplos. Vamos supor que queremos somar os números 1, 2 e 3, ou seja, temos um total de 3 elementos para somar:

$$\sum_{i=1}^{3} xi = x1 + x2 + x3 = 1 + 2 + 3 = 6$$

Muitas vezes vemos o somatório em expressões como:

$$\sum_{n=1}^{3} 2n - 1$$

Esse somatório indica que devemos somar a expressão 2*n - 1, substituindo o n por cada um dos valores da série. Logo, para a série 1, 2 e 3 temos

$$\sum_{n=1}^{3} 2n - 1$$

$$= \underbrace{[2(1) - 1]}_{n=1} + \underbrace{[2(2) - 1]}_{n=2} + \underbrace{[2(3) - 1]}_{n=3}$$

$$= 1 + 3 + 5$$

$$= 9$$

PRODUTÓRIO - Π

O produtório é a multiplicação de uma sequência de objetos matemáticos (números, funções, vetores, matrizes, etc.), chamados fatores, que tem como resultado o produto. É uma operação análoga ao somatório, embora seja menos utilizada quanto esse último. É representado pela letra grega pi maiúscula (Π).

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots \times x_n$$

OUTROS SÍMBOLOS IMPORTANTES

Símbolo	**Parâmetro**
μ	Média populacional
σ	Desvio-padrão populacional
σ^2	Variância populacional
p	Proporção populacional
$\overline{x}$	Média amostral
s	Desvio-padrão amostral
s^2	Variância amostral
$\hat{p}$	Proporção amostral
α	Nível de significância

1. COMO USAR ESSA APOSTILA

Bem-vindos(as) ao nosso curso de Teste de hipótese - Estatística do básico ao avançado! Sou a professora Renata Biaggi e estarei com vocês de perto nessa jornada em busca de mais conhecimento técnico e de tomadas de decisões bem fundamentadas.

Como vocês bem sabem, durante nosso curso teremos aulas online com bastante teoria e bastante prática. Teremos também exercícios extra-aulas e essa apostila, que escrevi pensando em guiar e complementar os estudos de vocês.

A apostila é composta por todos os tópicos que veremos em sala de aula de uma forma muito detalhada, teórica e exemplificada. Minha sugestão é que vocês leiam o tópico referente a aula **antes** de assistirem as aulas referentes. Isso vai possibilitar que vocês tenham um desempenho muito maior.

Ao longo da apostila vocês vão encontrar várias citações de autores, bem como indicações de livros e blogs para que vocês complementem os estudos - especialmente para os tópicos nos quais não vamos nos aprofundar durante o curso. Novamente, recomendo fortemente que todo esse material seja lido antes de cada aula, beleza?

Antes de partirmos para o conteúdo de fato, vamos alinhar algumas expectativas. O intuito do nosso curso não é aprender como usar linguagens de programação ou ferramentas de uma forma geral. Queremos aqui focar em **estatística e teste de hipótese**, entretanto para isso precisamos do auxílio de alguma interface para fazermos cálculos de uma forma mais rápida e para lidar com todos os nossos dados. Por isso, os dois próximos tópicos vão se dedicar a abordar de forma bastante rápida as principais ferramentas que guiarão esse curso: Python e Excel. Depois desses dois tópicos, entraremos no conteúdo que aborda a nossa tão amada ~~matemágica~~ matemática.

Outro ponto importantíssimo de ressaltarmos é que nesta apostila vamos detalhar os cálculos de todos os exemplos para demonstrarmos como usar cada fórmula. Na prática, dificilmente calcularemos qualquer coisa "na mão". Softwares como Excel ou linguagens como Python já tem esses cálculos intrínsecos, salvando bastante nosso tempo. Daí a importância dos capítulos introdutórios, uma vez que precisaremos manipular tais ferramentas para que elas façam todos os cálculos por nós.

Então Renata, pra quê eu preciso ver todas as fórmulas e entender os exemplos de cálculo? Porque aqui nós **não seremos ferramenteiros**! Nosso intuito é sair do curso entendendo de fato cada conceito que dá suporte às suas análises e modelos preditivos para que, quando vocês se depararem com situações complexas reais, vocês saibam agir sozinhos e possam pensar criticamente no que fazer.

Preparados? Bora lá!

7. INTRODUÇÃO A PROBABILIDADE

A **teoria da probabilidade** é o campo da matemática que estuda experimentos ou fenômenos **aleatórios** e através dela é possível analisar as chances de um determinado evento ocorrer.

Muitos eventos não podem ser previstos com total certeza. Podemos prever apenas a chance de um evento ocorrer, ou seja, qual a probabilidade de acontecer, usando-o. A probabilidade pode variar de 0 a 1 (ou 0 a 100% dependendo da denotação), onde 0 significa que o evento é impossível e 1 indica que definitivamente aquilo vai acontecer.

Nas seções anteriores, falamos bastante sobre a distribuição de frequências e sua importância para avaliarmos a variabilidade das observações. A partir dessas frequentes podemos calcular medidas de posição e variabilidade, como média, mediana, desvio padrão etc. Essas frequências (relativas) são estimativas de probabilidades de ocorrências de certos eventos de interesse.

De maneira geral, a fórmula da probabilidade é:

$$P(A) = \frac{\text{Nº DE RESULTADOS FAVORÁVEIS}}{\text{Nº DE RESULTADOS POSSÍVEIS}}$$

Sendo A o evento que queremos prever. O numerador da equação acima representa a quantidade de vezes que acontece o evento A (resultados favoráveis) e o numerador representa a quantidade de espaços amostrais que temos (todos os resultados possíveis). Vamos a um exemplo

Há 6 travesseiros em uma cama, 3 são vermelhos, 2 são amarelos e 1 é azul. Qual é a probabilidade de escolher um travesseiro amarelo?

Resposta:

A probabilidade é igual ao número de travesseiros amarelos na cama dividido pelo número total de travesseiros, ou seja, 2/6 = 1/3.

EXPERIMENTO ALEATÓRIO

Um **experimento aleatório** é aquele que não é possível conhecer qual resultado será encontrado antes de realizá-lo. Os acontecimentos deste tipo quando repetidos nas mesmas condições, podem dar resultados diferentes e essa inconstância é atribuída ao acaso. Um exemplo de experimento aleatório é jogar um dado não viciado (dado que apresenta uma distribuição homogênea de massa) para o alto. Ao cair, não é possível prever com total certeza qual das 6 faces estará voltada para cima.

ESPAÇO AMOSTRAL

Representado pela letra Ω (ômega), o **espaço amostral** corresponde ao conjunto de todos os pontos amostrais, ou, resultados possíveis obtidos a partir de um experimento aleatório.

Por exemplo, ao retirar ao acaso uma carta de um baralho, o espaço amostral corresponde às 52 cartas que compõem este baralho.

No caso do dado, se o lançarmos para cima, qualquer uma das 6 faces podem ocorrer. Ou seja, nosso espaço amostral seria as 6 faces desse dado (face 1, face 2, ..., face 6). Nesse caso, a denotação do espaço amostral é:

$\Omega = \{1, 2, 3, 4, 5, 6\}$

Vamos a mais um exemplo para fixar bem esse conceito. Se lançarmos uma moeda para cima, podemos obter cara ou coroa. Nesse caso, nosso espaço amostral é:

Ω = {cara, coroa}

Para facilitar, vou apelidar "cara" com a letra H e "coroa" com a letra T. Ou seja, teriamos:

Ω = {H, T}

E se lançarmos 2 moedas simultaneamente para cima, qual seria nosso espaço amostral? Nesse caso, podemos obter cara em uma coroa em outra, cara em uma, cara em outra... e assim por diante. Logo, denotamos o espaço amostral como sendo:

Ω = {HT, HH, TH, TT}

Agora vamos a um exemplo um pouco diferente. Considere o experimento que consiste em retirar uma lâmpada de um lote e medir seu "tempo de vida" antes de se queimar. Um espaço amostral conveniente é

$\Omega = \{t \in \mathbb{R} : t \geqslant 0\}$

Estranho né? Vou explicar. Primeiro, o símbolo $\in$ significa "pertence" e o símbolo $\mathbb{R}$ significa "reais" - ou seja, números reais. Quando falamos $t \in \mathbb{R}$ dizemos que t (nossa variável de tempo) pertence aos reais, portanto, t pode ser qualquer valor desde que esteja no conjunto dos números reais. Também dizemos que $t \geqslant 0$, uma vez que aqui estamos interessados somente nos números reais não negativos (não faz sentido ter um tempo de vida negativo).

Se tivermos um evento A que indica "o tempo de vida da lâmpada é inferior a 20 horas", então

$A = \{t : 0 \leqslant t \leqslant 20\}$.

Ou seja, de todo o espaço amostral possível (qualquer valor de tempo possível), estaríamos interessados em encontrar quando o tempo é menor que 20 horas. Como o tempo nunca é negativo, t então tem que estar entre 0 a 20 horas. Esse é um exemplo de um espaço amostral contínuo, contrastado com os anteriores, que são discretos.

TIPOS DE ENVENTOS

Um evento, como já explicado, é qualquer subconjunto do espaço amostral de um experimento aleatório.

EVENTO IMPOSSÍVEL

O conjunto do evento é vazio.

Exemplo: jogar um dado para o alto e tirar um número 8

EVENTO COMPLEMENTAR

Os conjuntos de dois eventos formam todo o espaço amostral, sendo um evento complementar ao outro.

Exemplo: No experimento de lançar uma moeda, o espaço amostral é Ω = {cara, coroa}.

Seja o evento A sair cara, A={cara}, o evento B sair coroa é complementar ao evento A, pois, B={coroa}. Juntos formam o próprio espaço amostral.

EVENTO MUTUAMENTE EXCLUSIVO

Os conjuntos dos eventos não possuem elementos em comum. A intersecção entre os dois conjuntos é vazia.

Exemplo: Seja o experimento lançar um dado, os seguintes eventos são mutuamente exclusivos

A: ocorrer um número menor que 5, A={1, 2, 3, 4}
B: ocorrer um número maior que 5, A={6}

EVENTOS DEPENDENTES E INDEPENDENTES

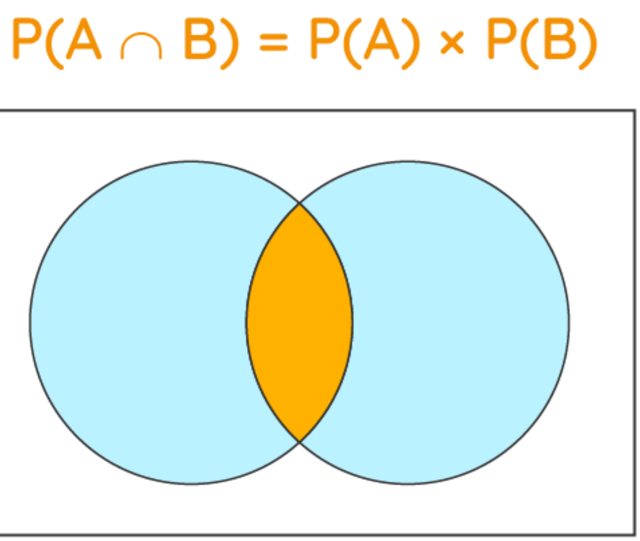
Em probabilidade, dizemos que dois **eventos são independentes** quando o fato de saber que um evento ocorreu não altera a probabilidade do outro evento.
Eventos dependentes são eventos que, caso um aconteça, a probabilidade do outro evento acontecer muda. Vamos pegar um exemplo bem simples.

Quando você vai pegar o seu ônibus na parada, você precisa sair de casa horário pré determinado por você (evento A) para chegar no trabalho às 8h da manhã (evento B). Ou seja, caso você saia de casa em outro determinado horário, a probabilidade de você chegar no trabalho às 8h da manhã muda.

PROBABILIDADE DA INTERSECÇÃO DE EVENTOS

Para dois eventos **independentes** a probabilidade da intersecção de dois eventos envolve a chance de o evento A **E** de o evento B ocorrer. O cálculo é feito por:

$$P(A \cap B) = P(A) \times P(B)$$



Para **eventos dependentes**, a probabilidade passa a ser:

$$P(A \cap B) = P(A) \times P(B|A)$$

Em que $P(B|A)$ é a probabilidade de um evento B acontecer dado que A já aconteceu - ou seja, a probabilidade de você chegar ao trabalho às 8h da manhã dado que você saiu de casa em um outro horário. Esse tipo de probabilidade é chamada de **probabilidade condicional** e vamos falar sobre falar sobre ela mais a frente

<u>Vamos a um exemplo:</u>

Qual é a probabilidade de selecionar uma carta vermelha **e** um 6 quando uma carta é selecionada aleatoriamente de um baralho de 52 cartas?

Resposta:

Sejam A e B as probabilidades individuais de obter uma carta vermelha e um 6, respectivamente.

Sabemos que o número de cartas vermelhos é 26 e que o número de cartas número 6 é 4

A probabilidade de obter uma carta vermelha de um baralho de 52 cartas, P(A) = 26/52

A probabilidade de obter um 6 de um baralho de 52 cartas, P(B) = 4/52

Usando a fórmula P(A∩B),

$P(A \cap B) = P(A) \times P(B)$

$P(A \cap B) = 26/52 \times 4/52 = 1/26 = 0{,}038$

Ou seja, a probabilidade de retirarmos uma carta vermelha que seja 6 é de 3,8%..

PROBABILIDADE DA UNIÃO DE DOIS EVENTOS

A probabilidade da união de dois eventos envolve a chance de o evento A **OU** de o evento B ocorrer. O cálculo é feito por:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$P(A \cap B)$ é a probabilidade da interseção entre A e B.

Podemos dizer, então, que a probabilidade da união de dois eventos é calculada pela probabilidade do primeiro evento ocorrer mais a probabilidade do segundo evento ocorrer menos a probabilidade da intersecção de ambos, sendo que a probabilidade da intersecção de dois eventos é igual à probabilidade do primeiro e do segundo evento ocorrerem simultaneamente.

Vamos a um exemplo

Uma urna contém 20 bolas numeradas de 1 a 20. Quando uma bola é retirada ao acaso, qual é a probabilidade do número ser múltiplo de 3 **ou** de 5?

Resposta:

- Evento A: Múltiplos de 3 no espaço amostral de 1 a 20: {3, 6, 9, 12, 15, 20}

  Ou seja, de um total de 20 bolinhas, 6 são múltiplas de 3

  P(A) = probabilidade de ser múltiplo de 3 = 6/20 = 0.3

- Evento B: Múltiplos de 5 no espaço amostral de 1 a 20: {5, 10, 15, 20}

  Ou seja, de um total de 20 bolinhas, 4 são múltiplas de 5

  P(B) = probabilidade de ser múltiplo de 5 = 4/20 = 0.2

- Evento A ∩ B: Múltiplos de 3 e 5 simultaneamente no espaço amostral de 1 a 20: {15, 20}

  P(A ∩ B) = probabilidade de ser múltiplo de 3 e 5 simultaneamente = 2/20 = 0.1

Logo, a probabilidade de ser múltiplo de 3 ou de 5 é de:

P (A U B) = 0.3+ 0.2 - 0.1 = 0.4

Logo, há 40% de chance de tirarmos um múltiplo de 3 ou de 5.

Vamos a um segundo exemplo

Considere o experimento: lançamento de um dado. Qual a probabilidade de sair um número maior que 5 **ou** um número ímpar?

Resposta:

- Evento A: Número maior que 5: {6}

  Ou seja, de um total de 6 números (dado tem 6 faces), 1 número é maior que 5.

P(A) = probabilidade de ser maior que 6 = 1/6 = 0.166

- Evento B: Número ímpar: {1, 3, 5}

  Ou seja, de um total de 6 números, 3 são ímpares

  P(B) = probabilidade de ser ímpar= 3/6 = 0.5

- Evento A ∩ B: Maior que 5 **e** ímpar: não existe

  Nenhum valor é maior que 5 e é ímpar simultaneamente pois, o único valor maior que 5 é o número 6

  P(A ∩ B) = 0

Logo, a probabilidade de ser múltiplo de 3 ou de 5 é de:

P (A U B) = 0.166 + 0.5 - 0 = 0.666

Chamamos esse tipo de evento em que a probabilidade de intersecção dos eventos é 0 de **eventos mutuamente exclusivos**

PROBABILIDADE CONDICIONAL

Se a probabilidade de ocorrência de um evento B interfere na probabilidade de ocorrência de um evento A, então dizemos que a probabilidade de A está **condicionada** à probabilidade de B e representamos por P(B/A). Lê-se: probabilidade de A dado B. A probabilidade condicional pode ser calculada por:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Onde,

$P(A \cap B)$ é a probabilidade da interseção entre A e B.

$P(B)$ é a probabilidade do evento B.

<u>Vamos a um exemplo</u>

Dois dados são lançados ao acaso. Qual a probabilidade da soma ser igual a 6, dado que o primeiro dado saiu número menor que 3.

Resposta:

Aqui temos um exemplo clássico de probabilidade condicional, em que:

Evento A: “a soma ser igual a 6”
Evento B: “o primeiro dado é menor 3”

- A = {soma igual a 6} = {(1, 5),(2, 4),(3, 3),(4, 2),(5, 1)} -> 5 combinações

  P(A) - de um total de 36 combinações (6 faces de 2 dados), a soma pode ser 6 em apenas 5 casos.

  Apesar de não ser importante ter a probabilidade de A na fórmula de P(A|B), saber o espaço amostral será útil para calcular P(A ∩ B). Veremos abaixo.

- B = {primeiro dado < 3} = {(1, 1),(1, 2),(1, 3),(1, 4),(1, 5),(1, 6), (2, 1),(2, 2),(2, 3),(2, 4),(2, 5),(2, 6)} -> 12 combinações

  P(B) - de um total de 36 combinações, ele é menor que 3 em apenas 12 casos. Ou seja, P(B) = 12/36 = 0.3

- P(A ∩ B) = probabilidade da soma ser 6 e o primeiro dado ter um valor menor que 3.

  Observando o espaço amostral de A, vemos que o primeiro dado é menor que 3 apenas para os casos (1,5) e (2,4). Ou seja, para todas as 36 combinações possíveis, somente em 2 casos essa condição será satisfeita. Ou seja, P(A ∩ B) = 2/36 = 0.0555

Logo, P(A|B) = 0.055/0.3 = 0.183

Portanto, a probabilidade da soma ser 6 ao lançarmos dois dados uma vez que o primeiro dado tem valor menor que 3 é de 0.183.

Vamos a um segundo exemplo

Duas cartas são selecionadas, sem reposição da primeira carta, de um baralho normal de 52 cartas. Encontre a probabilidade de selecionar um rei e depois uma rainha.

Resposta:

Queremos a intersecção de dois eventos (A e B acontecerem) e são eventos dependentes (seleção uma carta em seguida da outra sem reposição)

A: Selecionar um rei

B: Selecionar uma rainha

$P(A \cap B) = P(A)xP(B|A)$

Existem 4 reis e 4 rainhas no baralho. Na primeira retirada, temos 52 cartas. Logo, a probabilidade do evento A é:

$P(A) = 4/52$

Depois de retirar uma carta, restam 51 cartas. Dado que o primeiro evento aconteceu (retirar rei), continuam sobrando 4 rainhas.

$P(B|A) = 4/51$

Logo, $P(A \cap B) = 4/52 \times 4/51 = 0{,}006$

TEOREMA DA PROBABILIDADE TOTAL

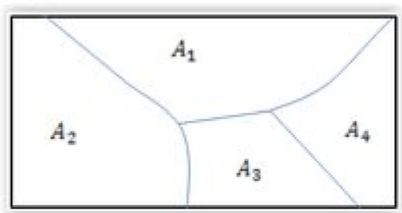
A definição formal do teorema da probabilidade total é a seguinte.

Seja um evento do espaço amostral Ω, e {Ai: i = 1, 2, 3,..., n} um conjunto de eventos distintos cuja união é todo o espaço amostral, ou seja Ai é uma partição do espaço amostral Ω. A probabilidade do evento pode ser calculada como segue:

$$P(B) = \sum_{i=1}^{n} P(B|A_i).P(A_i)$$

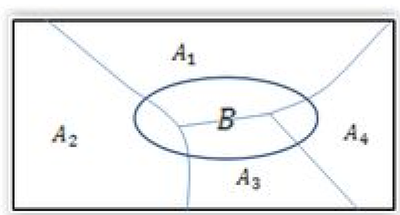
Não entendeu nada? Vamos traduzir!

Primeira coisa que precisamos entender é que temos um espaço Ω que é formado por vários eventos A1, A2, A3...., todos mutuamente exclusivos.



Ω

Agora, temos um evento B nesse espaço amostral que depende dos eventos anteriores



Ω

Podemos reescrever esse evento B de forma que:

B = (A1 ∩ B) + (A2 ∩ B) + (A3 ∩ B) + (A4 ∩ B)

Logo,

P(B) = P(A1 ∩ B) + P(A2 ∩ B) + P(A3 ∩ B) + P(A4 ∩ B)

Usando a probabilidade condicional P(Ai ∩ B) = P(Ai)xP(B|Ai), reescrevemos a fórmula acima de forma que:

$$\sum_{i=1}^{4} P(B|A_i) \cdot P(A_i)$$

Generalizando, temos que:

$$P(B) = \sum_{i=1}^{n} P(B|A_i).\, P(A_i)$$

<u>Agora vamos a um exemplo de uso:</u>

Considere um processo de fabricação de semicondutores. Nesses processos, quando um chip está sujeito a altos níveis de contaminação, a probabilidade de que ele cause defeito na produção é de 0,1. Se o chip não está sujeito a altos níveis de contaminação, a probabilidade dele causar defeito na produção é 0,005. Sabemos ainda que a probabilidade de um chip estar sob altos níveis de contaminação é 0,2. Estamos interessados no evento: o chip causa uma falha na produção (geral, estando ou não sob efeito de contaminação).
As condições a que esse evento está sujeito são: está sujeito a altos níveis de contaminação; não está sujeito a altos níveis de contaminação.

Resposta:

Nesse nosso caso, Ω será composto por A1 = está sujeito a altos níveis de contaminação e A2 = não está sujeito a altos níveis de contaminação. Nosso evento de interesse é B = o chip causa uma falha na produção.

Usando a fórmula anterior, temos que:

$P(B) = P(B|A1)*P(A1) + P(B|A2)*P(A2)$

Sabemos que:

$P(B|A1) = 0,1$ - Probabilidade do chip causar falha na produção dado que está sujeito a altos níveis de contaminação

$P(B|A2) = 0,005$ - Probabilidade do chip causar falha na produção dado que não está sujeito a altos níveis de contaminação

$P(A1) = 0,2$ - Probabilidade de um chip estar sob altos níveis de contaminação

Agora vamos usar a lógica. O chip está ou não está sujeito a altos níveis de contaminação - não há um intermediário. Ou seja, a probabilidade de ele estar **OU** não estar sujeito a altos níveis de contaminação é 100%, uma vez que não existe outra possibilidade. Usamos a letra Ω para indicar todo nosso espaço amostral possível. Logo

$\Omega = A1 \cup A2$

Usando a regra da união de dois eventos:

$P(\Omega) = P(A1) + P(A2) - P(A1 \cap A2)$

$P(A1 \cap A2)$ é zero, uma vez que não existe uma opção de estar **E** não estar sob efeitos de alta contaminação. $P(\Omega)$ é 100% (ou seja, 1), uma vez que é nosso espaço amostral inteiro. Como já dito, $P(A1) = 0,2$. Então temos:

$1 = 0,2 + P(A2) \rightarrow P(A2) = 1-0,2 = 0,8$

Portanto, há 80% chance do chip não estar sujeito a altos níveis de contaminação. Agora, podemos voltar a nossa fórmula:

P(B) = P(B|A1)*P(A1) + P(B|A2)*P(A2) = 0,1*0,2 + 0,005*0,8 = 0,0235 = 2,35%

Isso quer dizer que existe 2,35% de probabilidade do chip causar uma falha na produção no geral, considerando todos os cenários (contaminação ou não).

TEOREMA DE BAYES

Uma das relações mais importantes envolvendo probabilidades condicionais é dada pelo Teorema de Bayes. A versão mais simples desse teorema é dada pela fórmula:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Vimos que a informação muitas vezes é apresentada em forma de probabilidade condicional. Elas nos fornecem a probabilidade de um evento (no caso do chip, uma falha no processo) dada uma condição (estar contaminado). Pode ser que estejamos interessados em investigar: depois que o evento deu um resultado (falha no processo); qual a probabilidade de uma certa condição estar presente (alta contaminação)?

P(A1|B) = Probabilidade do chip estar contaminado dado que aconteceu uma falha na produção.

***IMPORTANTE**: não confunda com P(B|A1) = 0,1 - Probabilidade do chip causar falha na produção dado que está sujeito a altos níveis de contaminação*

Usando o teorema de Bayes temos que:

P(A1|B) = P(B|A1)*P(A1)/P(B) = 0,1*0,2/0,0235 = 0,85 = 85%.

Logo, existe 85% de probabilidade do chip estar contaminado dado que aconteceu uma falha na produção.

Vamos a um outro exemplo:

Uma empresa oferece aos candidatos a uma vaga um curso de treinamento durante uma semana. No final do curso, eles são submetidos a uma prova e 25% são classificados como bons (B), 50% como médios (M) e os restantes 25% como fracos (F). A empresa pretende substituir o treinamento por um teste contendo questões referentes a conhecimentos gerais e específicos. Para isso, gostaria de conhecer qual a probabilidade de um indivíduo aprovado no teste ser considerado fraco, caso tivesse feito o curso. Assim, neste ano, fizeram um experimento. Antes do início do curso, os candidatos foram submetidos ao teste e receberam o conceito aprovado (A) ou reprovado (R). Posteriormente, fizeram o curso e, ao final, obtiveram-se as seguintes probabilidades condicionais:

P(A|B) = 0,80 - Probabilidade de ser "bom" dado que foi aprovado (são as pessoas aprovadas que receberam critério "bom")

P (A|M) = 0,50 - Probabilidade de ser "médio" dado que foi aprovado (são as pessoas aprovadas que receberam critério "médio")

P (A|F) = 0,20 - Probabilidade de ser "fraco" dado que foi aprovado (são as pessoas aprovadas que receberam critério "fraco")

Agora, a empresa quer entender qual seria a probabilidade de ter uma pessoa aprovada dado que era "fraco". Ou seja, P (F|A).

Evento A: Probabilidade de ser aprovado

Evento F: Probabilidade de ser "fraco"

Resposta:

Usando teorema de Bayes temos que:

P (F|A) = P(A|F)*P(F)/P(A)

A probabilidade de ser "fraco" já foi dita no próprio enunciado -> P(F) = 0,25. Também temos P(A|F) = 0,2. Porém, qual seria P(A)?

Note que a probabilidade de ser aprovado é uma probabilidade que ainda não temos. Porém, temos a probabilidade de ser aprovado dado um determinado cenário anterior: Aprovado dado que era bom, aprovado dado que era médio e aprovado dado que era fraco. Sabendo que só existem essas 3 possibilidades anteriores (ser bom, médio ou fraco), podemos usar aqui o teorema da probabilidade total.

$$P(B) = \sum_{i=1}^{n} P(B|A_i).\,P(A_i)$$

P(A) = P(A|B)*P(B) + P(A|M)*P(M) + P(A|F)*P(F) = 0,8*0,25 + 0,5*0,5 + 0,2*0,25 = 0,5.

Ou seja, a probabilidade de ser aprovado é de 50%.

Voltando no teorema de Bayes:

P (F|A) = P(A|F)*P(F)/P(A) = 0,2*0,25/0,5 = 0,1 = 10%

Portanto, existe 10% de probabilidade do candidato ser fraco dado que ele foi aprovado.

8. DISTRIBUIÇÕES DISCRETAS E CONTÍNUAS DE PROBABILIDADE

No capítulo anterior vimos o que é um espaço amostral e como calcular a probabilidade de um evento. Neste capítulo vamos ampliar esses conceitos associando espaço amostral e eventos a valores numéricos.

Antes de iniciarmos, precisamos introduzir um conceito importante. Uma **variável aleatória** (abreviadamente, v.a.) é uma função que associa a cada elemento de um espaço amostral um número real.

Na prática, usualmente não existe a preocupação de se explicitar qual é o espaço amostral no qual está definida a variável aleatória. O que importa é definir o conjunto de valores reais que a variável pode admitir e explicitar como se calcula a probabilidade de que ela admita tais valores.

O conceito de variável aleatória é particularmente útil em situações nas quais se dispõe de um nível de conhecimento parcial ou incompleto do comportamento da grandeza que está sendo estudada. Essa incerteza pode ser então introduzida sob a forma de um modelo probabilístico.

Por exemplo, um engenheiro encarregado de realizar estudos em uma empresa mede o tempo que os operários gastam em executar certas tarefas. Naturalmente, para cada tarefa o tempo gasto depende da experiência e da destreza do operário. Suponha que, para uma particular tarefa, o tempo médio gasto é de 285 segundos. Aqui a variável aleatória é X = “tempo em segundos gasto na execução da tarefa”, e tudo indica que, para um operário novato, pouco treinado, é alta a probabilidade P (X >285).

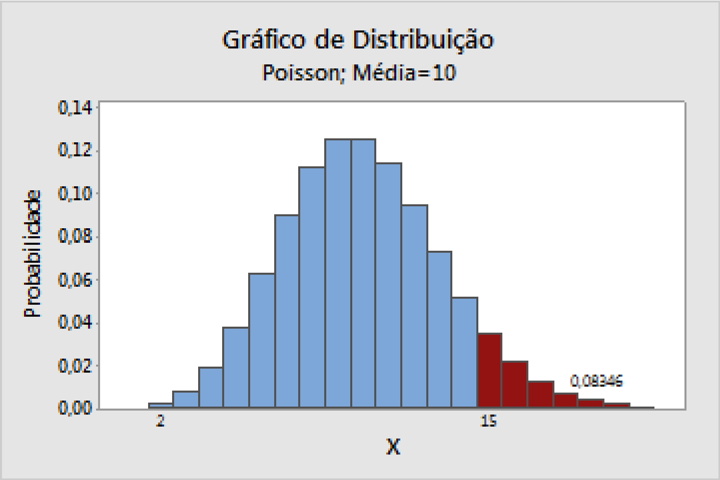
FUNÇÃO DE PROBABILIDADE - VARIÁVEIS DISCRETAS

Dizemos que X é uma **v.a. discreta** se o número de valores que ela pode admitir é finito. Por exemplo, em uma linha de produção as peças produzidas são examinadas até que sejam encontradas 10 peças defeituosas, e então o número total de peças examinadas é anotado. Nesse caso, a v.a. X é o número total de peças examinadas. Nesse caso, X pode admitir os valores 10, 11, 12, 13, 14, ... Assim, claramente X é uma v.a. aleatória discreta.

Uma distribuição discreta descreve a probabilidade de ocorrência de cada valor de uma variável aleatória discreta. Com uma distribuição de probabilidade discreta, cada valor possível da variável aleatória discreta pode ser associado a uma probabilidade diferente de zero. Deste modo, uma distribuição de probabilidade discreta é, por vezes, apresentada em forma de tabela. Com uma distribuição discreta é possível calcular a probabilidade de que X é exatamente igual a algum valor.

Vamos a um exemplo:

O gráfico abaixo mostra a frequência de quantidade de reclamações por dia em um call center.



A probabilidade do call center receber 15 reclamações por dia é exatamente a frequência (ou seja, em torno de 0,04 = 4% de acordo com o gráfico).

As barras sombreadas neste exemplo representam o número de ocorrências quando as reclamações de clientes diárias são 15 ou mais. A altura das barras somam 0,08346; por conseguinte, a probabilidade de que o número de chamadas por dia seja de 15 ou mais é 8,35%.

Apresentamos a seguir alguns dos modelos probabilísticos discretos que costumam ser mais utilizados nas aplicações práticas da Estatística. Existem diversas modelagens para prever a probabilidade de variáveis discretas, como os modelos que envolvem ensaios de Bernoulli (Bernoulli, Binomial, Geométrico) e o modelo de Poisson. Aqui, vamos abordar rapidamente o modelo de Bernoulli e o modelo Binomial apenas, mas deixaremos dicas de onde vocês podem encontrar outras informações de outros tipos de modelos.

MODELO DE BERNOULLI

Num experimento aleatório é comum que estejamos interessados apenas na ocorrência de um resultado particular. Por exemplo:

1. Na seleção de um chip extraído de um lote, podemos querer saber somente se ele é perfeito ou não;
2. Na seleção de uma peça fabricada, queremos saber somente se ela satisfaz ou não às especificações exigidas pelo consumidor;

Em todos esses casos, o experimento realizado admite somente dois resultados possíveis.

Um experimento dessa natureza é chamado de "experimento de Bernoulli" ou, mais popularmente, "ensaio de Bernoulli". Os dois resultados de um ensaio de Bernoulli são comumente chamados de "sucesso" e "fracasso", sendo P a probabilidade de sucesso e (1-P) a probabilidade de fracasso.

Por exemplo, numa turma com 50 alunos, dos quais 30 são homens e 20 mulheres, escolhe-se um aluno ao acaso. Se levarmos em consideração apenas o sexo do aluno selecionado, isso trata-se de um ensaio de Bernoulli. Se considerarmos como sucesso a escolha de uma mulher, teremos p = 20/50 = 0,4 = e 1 – p = 0,6.

MODELO BINOMIAL

No modelo Binomial um mesmo experimento de Bernoulli é repetido n vezes, independentemente, e a v.a. de interesse representa o número de sucessos a serem obtidos nos n ensaios.

Sejam p e (1–p), respectivamente, as probabilidades de sucesso e de fracasso em cada ensaio de Bernoulli. Se os resultados de cada ensaio são denotados por S (sucesso) e F (fracasso) teremos, para cada ensaio, P(S) = p e P(F) = 1– p.

O espaço amostral do experimento resultante dos n ensaios de Bernoulli será composto por resultados que podem ser escritos como uma sequência de letras S e F. Em particular, um resultado com k sucessos e (n – k) fracassos pode ser descrito, sem perda de generalidade, como uma sequência de k S's, seguida de (n–k) F's, como a seguinte:

SSSSSS...SFFF...FF.

Como os n ensaios são independentes, a probabilidade de ocorrência desse resultado particular é:

$$p^{k} * (1 - p)^{n-k}$$

O evento “k sucessos e (n–k) fracassos” pode ocorrer de diversas outras maneiras. O cálculo do número de maneiras de se obter “k sucessos e (n – k) fracassos” é o número de combinações de n objetos tomados de k em k

$$P_x = \binom{n}{x} p^x q^{n-x}$$

O símbolo $\binom{n}{x}$ indica combinação de n objetos tomados x a x. Por exemplo, vamos supor que temos um total de 50 elementos e o particionaremos em grupos de 5. Temos inúmeras possibilidades de formar de forma diferente esse grupo de 5. O cálculo de combinatória nos mostra quantas exatamente existem, de forma que:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Na fórmula acima representamos x como sendo r, que também é uma denotação possível que vocês devem encontrar em vários livros por aí. O símbolo "!" indica " fatorial. A fórmula de um fatorial r!, por exemplo, é dada por:

r! = 1*2*3*4...*r

Essa fórmula acima pode ser simplificada por:

$$\binom{n}{r} = \frac{n(n-1)(n-2).....(n-r+1)}{1 \cdot 2 \cdot 3......r}$$

Logo, voltando ao exemplo de termos 50 elementos e queremos formar 5 grupos, temos que podemos formá-lo:

$$\binom{50}{5} = \frac{50*(50-1)*(50-2)*...*(50-5+1)}{5!} = \frac{50*49*48*47*46}{1*2*3*4*5}$$

= 2.118.760 maneiras distintas

NOTA IMPORTANTE: É CONVENCIONADO QUE 0! = 1! =1. OU SEJA, NUNCA TEREMOS UMA DIVISÃO POR ZERO

Quando temos as combinações n tomado a n ou n tomado a 0, ficamos com:

$$\binom{n}{n} = \binom{n}{0} = \frac{n!}{0!n!} = 1$$

Vamos ver em um exemplo como o modelo binomial funciona:

Geralmente, em cerca de 80% dos chamados que um certo técnico em computação recebe para resolver panes nos computadores de clientes ele constata que o problema decorreu da presença de algum vírus. Suponha que, em um determinado dia, esse técnico vai visitar seis desses clientes cujos computadores necessitam de conserto, e admita também que os seis clientes não se comunicam por meio de computador (o que garante a independência da existência de vírus em cada computador). Calcule a probabilidade de que:

a) Pelo menos quatro entre os seis computadores estejam com vírus.
b) No máximo dois dentre eles estejam com vírus.
c) Todos os seis estejam com vírus.

Resposta:

Considere:
Sucesso = "o defeito no computador é devido a presença de vírus" (p = P(sucesso) = 0,80)
X = número de computadores com vírus entre os 6 a serem consertados.

a) $P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6) =$

$$=\binom{6}{4}0,8^4\times 0,2^2+\binom{6}{5}0,8^5\times 0,2+\binom{6}{6}0,8^6=0,90112\,.$$

Isso significa que é bem alta a probabilidade de pelo menos quatro entre os seis computadores estarem com vírus (90,112%)

b) P(X ≤ 2) = P(X = 0) + P(X = 1) + P(X = 2)

$$=\binom{6}{0}0,2^6+\binom{6}{1}0,8\times 0,2^5+\binom{6}{2}0,8^2\times 0,2^4=0,01696\,.$$

Este valor indica que é baixíssima a probabilidade de que no máximo dois deles estejam com vírus.

NOTA IMPORTANTE: NESSE CASO, A COMBINATÓRIA DE 6 ELEMENTOS TOMADOS 0 A 0 É IGUAL A 1

c) $P(X=6)=0,8^6=0,26214.$

Como dito anteriormente, existem diversos modelos probabilísticos para variáveis discretas. O modelo de **Poisson** prevê a distribuição do número de indivíduos ou outro dados de contagem por unidade de tempo ou espaço. Por exemplo

- Se os organismos distribuem-se independentemente no espaço, espera-se que a contagem de indivíduos por m² se ajuste a uma distribuição de Poisson.
- Número de chamadas telefônicas que chegam a uma Central em um dado intervalo de tempo
- Número de navios que chegam ao cais de um porto em um dia
- Número de defeitos encontrados em uma geladeira recém-fabricada

O modelo **Geométrico** resulta de experimentos similares aos experimentos binomiais. Suponha novamente uma variável resposta pode assumir somente dois valores. Enquanto a distribuição binomial se preocupa com a probabilidade do número de sucessos em n tentativas, o modelo Geométrico conta o número de fracassos até a observação do primeiro sucesso.Alguns exemplos de quando podemos usá-lo:

- Suponha que um ambiente tenha N manchas de hábitats que pode ser ou não ocupado por uma espécie. Queremos saber: quantas manchas de habitat devem ser avaliadas até que a espécie seja detectada?
- O engenheiro responsável pelo Controle da Qualidade de uma linha de produção examina, uma após a outra, as peças fabricadas. Se achar uma defeituosa, ele para a produção para detectar e corrigir as causas do defeito. Se após examinar 10 peças verificar que nenhuma é defeituosa, ele mantém a linha funcionando. Se a probabilidade de se achar uma peça defeituosa em cada exame é 0,05, qual é a probabilidade de: a) a produção ser parada antes que a quinta peça seja examinada? b) a produção não precisar ser parada?

Você pode encontrar mais detalhes sobre esses modelos na referência Pinheiro, J., Cunha, S., S. Santiago, Gomes, G. - Probabilidade e Estatística: Quantificando a incerteza.

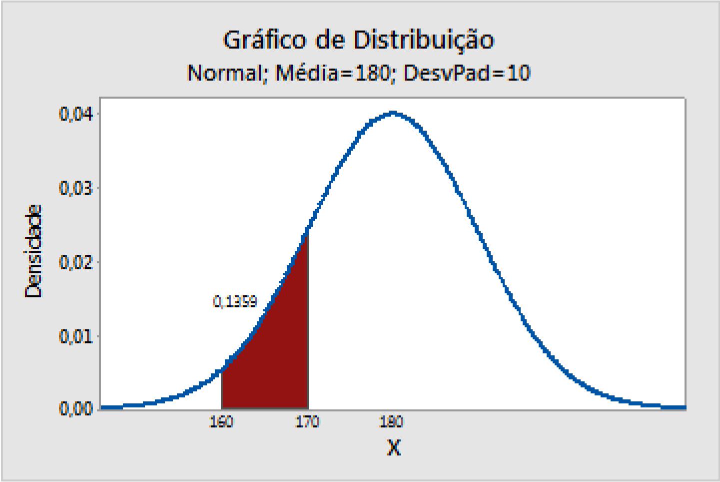
FUNÇÃO DE PROBABILIDADE - VARIÁVEIS CONTÍNUAS

A distribuição contínua descreve as probabilidades dos possíveis valores de uma variável aleatória contínua. Uma variável aleatória contínua é uma variável aleatória com um conjunto de valores possíveis (conhecidos como intervalos) que é infinito e incontável.

As probabilidades de variáveis aleatórias contínuas (X) são definidas como a **área sob a curva da sua distribuição**. Assim, apenas as faixas de valores podem ter uma probabilidade diferente de zero. A probabilidade de que uma variável aleatória contínua seja igual a algum valor é sempre zero.

Vamos a um exemplo:

A distribuição normal contínua pode descrever a distribuição de peso de indivíduos do sexo masculino adultos. Por exemplo, você pode calcular a probabilidade de que um homem pesa entre 160 e 170 libras



A região sombreada sob a curva, neste exemplo, representa o intervalo entre 160 e 170 libras (72 a 77 kg, aproximadamente). A área deste intervalo é 0,136; por conseguinte, a probabilidade de um homem ser selecionado aleatoriamente pesar entre 160 e 170 libras é de 13,6%. Toda a área sob a curva equivale a 1,0.

No entanto, a probabilidade de que X seja exatamente igual a algum valor é sempre zero porque a área sob a curva em um único ponto, que não tem nenhuma largura, é zero. Por exemplo, a probabilidade de um homem pesar exatamente 190 libras para a precisão infinita é zero. É possível calcular uma probabilidade não nula de que um homem pese mais do que 190 libras, ou menos do que 190 libras, ou entre 189,9 e 190,1 libras, mas a probabilidade de que ele pesa exatamente 190 libras é zero.

Como dito anteriormente, agora vamos definit uma probabilidade para um dado intervalo, ou seja, $P(a \leq X \leq b)$ para dois números reais a e b. Isso é obtido ao substituir-se a função de probabilidade p por uma função f, chamada **função de densidade de X,** ou simplesmente, **densidade de X.**

Definindo formalmente esse conceito:

Dizemos que X é uma variável aleatória contínua se existe uma função f, chamada função de densidade de X, satisfazendo as seguintes condições:

1. $f(x) \geq 0$ para todo x real
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. Para quaisquer a, b reais (a < b), $P(a \leq X \leq b) = \int_a^b (f(x)dx)$

Vamos traduzir esse trem! O primeiro ponto diz que essa função de densidade assume valor maior ou igual a zero para toda variável X existente. Ou seja, não há valores negativos para uma função de densidade.

O ponto dois mostra que a integral (símbolo ∫) dessa função deve ser igual a 1. Quando falamos de integral estamos automaticamente falando de "área embaixo da curva". Ou seja, se pegarmos a área inteirinha embaixo da curva, ela vai ser igual a 1. Não entraremos aqui no detalhe de cálculo de integrais, até porque muitas dessas fórmulas já são pré-calculadas na forma de tabela para nós (ufa!). Esse ponto 2 também é chamado de **Função de Distribuição Acumulada (FDA).**

E os pontos 1 e 2 você já sabia! Afinal, uma probabilidade sempre está entre 0 e 1.

O ponto 3 diz que a probabilidade de uma faixa será exatamente a área embaixo da curva dentro dessa faixa. É exatamente o que falamos no exemplo acima sobre os pesos de homens.

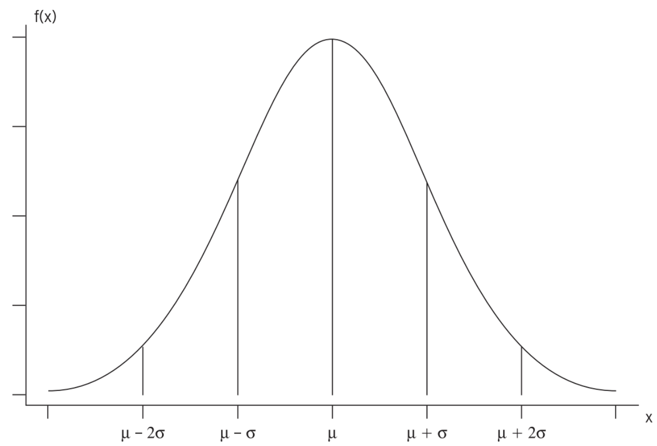
Apresentaremos a seguir alguns dos modelos probabilísticos contínuos que costumam ser mais utilizados nas aplicações práticas da Estatística. Existem várias, como o modelo Uniforme, Exponencial, Normal, t-student, chi-quadrado, F, etc. Nesse capítulo vamos abordar apenas a Normal, mas nos capítulos de teste de hipótese ainda abordaremos a t-student, chi-quadrado e F.

MODELO NORMAL

É, de longe, a distribuição mais importante da estatística. A curva Normal (também chamada de Gaussiana ou de paramétrica) descreve de forma muito adequada o comportamento de uma variável que se distribui de forma simétrica em relação a um valor central. Os dois parâmetros que a caracterizam são μ, que especifica a média, e **σ**, que define seu desvio-padrão. Tendo esses dois parâmetros definidos, a função é escrita como:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\}, \text{ para todo x real.}$$

Dessa forma, se tivermos a média e o desvio-padrão, conseguiremos traçar uma curva com o formato:



Lá no início dos anos 1800, supunha-se que todos os fenômenos da vida real devessem ajustar-se a uma curva em forma de sino; caso contrário, suspeitava-se de alguma anormalidade no processo de coleta de dados. Daí a designação de curva normal.

Ainda que hoje saibamos que isso não é verdade, a distribuição de probabilidade normal é importante na inferência estatística por três razões distintas:

a) as medidas produzidas em diversos processos aleatórios seguem essa distribuição;

b) as probabilidades normais podem ser usadas frequentemente como aproximações de outras distribuições de probabilidade, tais como a binomial e a de Poisson;

c) as distribuições de estatísticas da amostra, tais como a média e a proporção, frequentemente seguem a distribuição normal independentemente da distribuição da população. Veremos isso mais pra frente, quando falarmos sobre o Teorema do Limite Central.

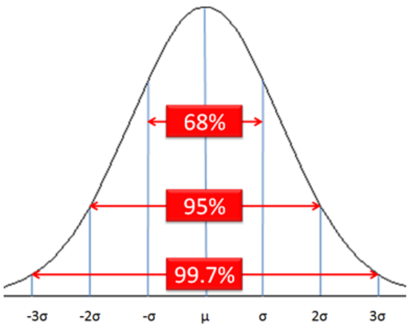
As características da distribuição normal são:

1. A média, a mediana e a moda são iguais - ou seja, não tem outliers!

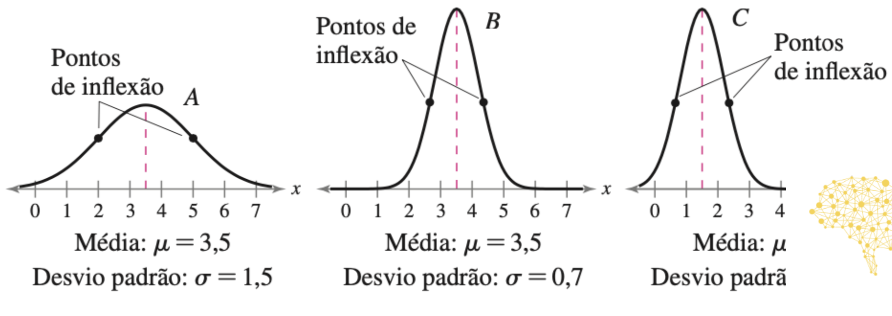
2. Tem forma de sino e é simétrica em torno da média.

3. À medida que a curva normal se distancia da média, ela se aproxima do eixo x, mas sem tocá-lo.

Outra grande característica dela é poder afirmar que cerca de 68% dos dados encontram-se dentro do intervalo de ± 1 desvios padrão da média, 95% dos dados encontram-se dentro do intervalo de ± 2 desvios padrão da média e assim por diante, conforme mostra a figura abaixo.



Uma **distribuição normal** pode ter qualquer média e qualquer desvio padrão positivo. Esses dois parâmetros, determinam o formato da curva normal. A média dá a localização da linha de simetria e o desvio padrão descreve o quanto os dados estão dispersos. Logo, observando a curva abaixo, sabendo que ela respeita todas as características de distribuição normal vistas acima, conseguimos ver que todas elas são distribuições normais!



Renata Biaggi

Como para qualquer distribuição contínua de probabilidade, o valor da probabilidade pode somente ser determinado para um intervalo de valores da variável.

Dentro dessa classe de funções, temos também um tipo específico de curva normal. Se uma v.a. tem distribuição Normal com média igual a 0 (zero) e variância igual a 1 (um), diremos que ela tem distribuição **Normal Padrão** ou distribuição **Normal Reduzida**.

Qualquer curva normalmente distribuída pode se tornar uma Normal Padrão. Para isso, pegamos cada valor de X e o padronizamos pela seguinte fórmula:

$$Z = \frac{X - \mu}{\sigma}$$

Em que μ é a média da sua distribuição atual **σ** é o desvio-padrão da distribuição atual. Nesse caso, Z será seu valor de X padronizado. E por que isso é legal? Pois temos uma tabela que indica a probabilidade para cada valor de Z, ou seja, não precisaremos fazer nenhum cálculo de probabilidade!

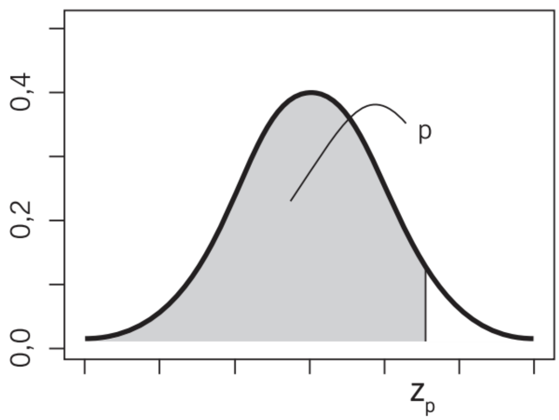
Um pedaço da tabela pode ser visto abaixo.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298

Nessa tabela, caso obtenhamos um z = 0.56, podemos dizer que a probabilidade de termos um **z menor ou igual a** 0.56 é de 0.71226 - ou seja, a área embaixo da curva até o z = 0.56 é de 0.71226.

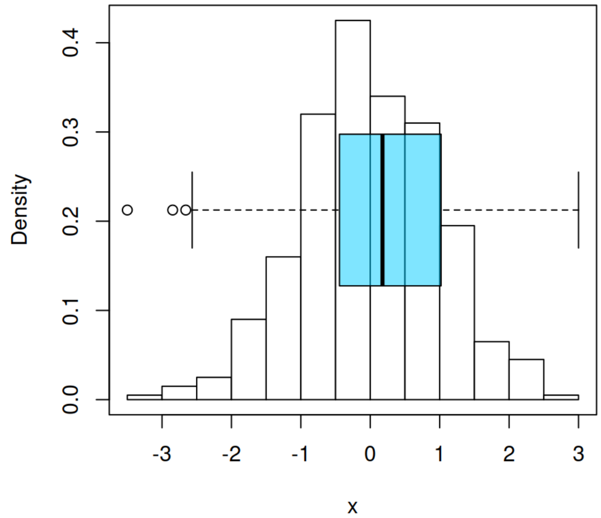
A notação que usamos para ler isso é: P(z<0.56) = 0.71226

Antes de partirmos para um exemplo de uso, note que acima falei "z menor ou igual a". É importante ressaltar aqui que essa tabela nos mostra somente a probabilidade **acumulada** até o valor de Z, como mostra a figura abaixo.



Isso é importante pois se quisermos calcular a probabilidade de uma **faixa** e não a acumulada, precisamos fazer algumas transformações. Para fazermos essas transformações, lembrem-se que a curva de probabilidade varia de 0 a 1, sendo que 0 é uma probabilidade nula de um valor acontecer e 1 uma probabilidade de 100% de que todos os valores estejam naquela faixa. Vocês entenderão mais sobre isso no exemplo 2.

Outra propriedade importante de lembrar é que em uma distribuição de probabilidade, a probabilidade acumulada P(z<zc) para um determinado valor corresponde ao **percentil** em que aquele valor se encontra. Existe um paralelo muito grande entre o histograma e o boxplot que ajuda lembrar desse conceito.



Notem que se tivéssemos um boxplot traçado em cima do histograma, saberíamos exatamente qual é o percentil 25, 50, etc.

Mesmo sem ter o boxplot em cima da nossa distribuição, usando a tabela de Z conseguimos saber qual é o percentual correspondente a um Z calculado e, consequentemente, a um X que corresponde a esse Z. Por exemplo, considerando que temos uma distribuição normal e temos um X = 20 que corresponde a um Z = 0.56. Sabendo que a probabilidade P(z<zc) nesse zc é de 0.71226 (de acordo com a tabela Z), podemos dizer que o x = 20 está no percentil 71,226.

Por exemplo, considerando que temos uma distribuição normal e que nosso x = 20 e, o z correspondente a esse x nessa distribuição é 0.56 (ou seja, zc = 0.56). Sabendo que a probabilidade P(z<zc) nesse zc é de 0.71226 (de acordo com a tabela z), podemos dizer que o x = 20 está no percentil 71,226.

Agora vamos a outros exemplos de uso

<u>Exemplo 1.</u>

Em uma população de homens com IMC médio = 29 e desvio-padrão = 6, qual é o IMC que representa o percentil 90?

Resposta:

Lembrando da fórmula:

$$z = \frac{x - \mu}{\sigma}$$

Queremos o X que representa o percentil 90. Sabemos que a média é 29 e desvio-padrão é 6. Logo, reescrevendo a fórmula acima e isolando o valor que queremos (x), temos:

X = Z*6+29

Para encontrarmos o valor de Z, precisamos lembrar que queremos o Z que representaria o percentil 90, ou seja, a área embaixo da curva deverá ser 0,900. Como os valores de z são tabelados, voltamos a tabela e procuramos o valor mais próximo a 0,900.

Z_i	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

Esse valor está entre 0.8997 e 0.9015. O mais próximo seria 0.8997, que corresponde a um z de 1.28. Logo,

x = 1,28*6 + 29 = 36,7

Portanto, o IMC que corresponde ao pctl 90 é de 36,7.

<u>Exemplo 2.</u>

Suponha que o tempo X, em minutos, corresponde ao tempo que uma pessoa leva para executar determinada tarefa e varia conforme uma distribuição Normal com parâmetros μ (média) e σ (desvio padrão). Suponha também que a probabilidade de que a tarefa seja executada em 70 minutos no máximo é 0,75, e a probabilidade de que a tarefa seja executada em no máximo 50 minutos é 0,25.

a) Determine os valores da média e desvio-padrão .

b) Qual a porcentagem das pessoas que precisarão de mais de 85 minutos?

Resposta:

a) Sabemos que X tem distribuição normal, então podemos padronizar sua distribuição para o Z normal padrão. A probabilidade acumulada quando X = 70 minutos é 0,75. Então:

   P(Z ≤ 70) = 0,75. Consultando a tabela, para essa probabilidade acumulada, Z = 0,67.

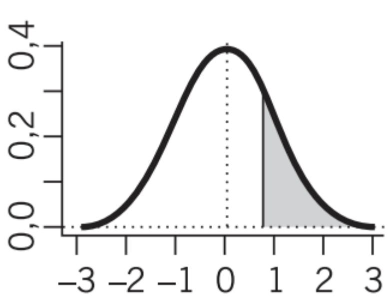
   Logo, Z = $\frac{70 - \mu}{\sigma}$ = 0,67

   Da mesma forma, sabemos que a probabilidade acumulada quando X = 50 minutos é 0,25. Da tabela, a probabilidade 0,25 corresponde a um Z = -0,67. Logo,

   Z = $\frac{50 - \mu}{\sigma}$ = -0,67

Temos agora 2 equações e 2 incógnitas. Resolvendo essas equações, temos μ = 60 e σ = 14,9, ambos em minutos.

b) Aqui temos um problemas: Não estamos em busca de uma probabilidade acumulada até certo valor - queremos saber a probabilidade a partir de um certo valor, ou seja, teríamos algo como:



Esse problema é fácil de resolver quando lembramos que a área total embaixo da curva é 1. Ou seja, ainda que a tabela Z apenas seja capaz de nos dar a probabilidade acumulada (área branca desse gráfico), sabemos que a área cinza deverá ser 1 - área branca. Ou seja, conseguimos fazer o cálculo mesmo sem o valor exatamente tabelado.

Dessa forma, $Z = (85-60)/14,9 = 1,677$

$P(z \geq 1,677) = 1 - P(z \leq 1,677) = 0,9535$

Você pode encontrar mais detalhes sobre esses e outros modelos probabilísticos na referência Pinheiro, J., Cunha, S., S. Santiago, Gomes, G. - Probabilidade e Estatística: Quantificando a incerteza.

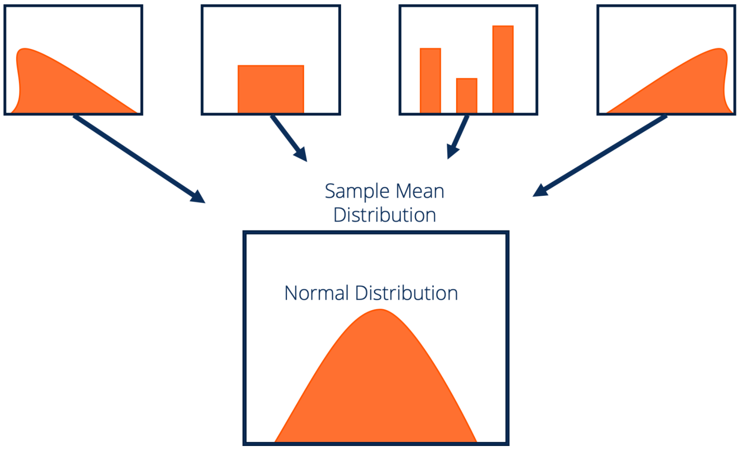
TEOREMA DO LIMITE CENTRAL

Vocês viram até agora que distribuições normais são ótimas para se trabalhar: temos suas propriedades, tabelas referências para encontrar probabilidades e um comportamento muito bem conhecido. Por conta disso, diversas estatísticas são descritas especialmente para distribuições normais.

Porém, em vários casos do dia a dia vamos nos deparar com **distribuições não normais**. Quando isso acontece, torna-se mais complexo e, em alguns casos, menos precisos, os métodos disponíveis para cálculo de probabilidade, teste de hipótese, entre outros.

Dessa forma, uma alternativa para garantir a normalidade dos dados é proveniente do **teorema do limite central**. O teorema nos diz que, para uma distribuição com **pelo menos 30 dados**, a **distribuição da média de subamostras** dessa distribuição seguirá uma normal.

Simplificando esse raciocínio: tendo os dados da distribuição, podemos selecionar várias subamostras dessa distribuição e calcular a média de cada uma dessas subamostras. Plotando em um gráfico a média de cada uma dessas subamostras, veremos uma distribuição normal, conforme exemplifica a imagem abaixo:



Para exemplificar o que que isso quer dizer

Supondo que temos uma distribuição com esses dados:

Distrib = [1, 1, 2, 3, 4, 5, 6, 7, 8,8,8,8,8,8,8,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9]

1) A amostra tem pelo menos 30 dados? Sim
2) Vamos selecionar várias subamostras aleatórias (podendo ter repetição) dessa distribuição:
   S1 = [1, 1, 9]
   S2 = [3, 4, 9]
   S3 = [4,6, 9]
   .... até Sn
3) Vamos calcular a média para cada uma dessas subamostras
   Média S1 = 3,5
   Média S2 = 5,3
   Média S3 = 6,3
   .... Calcular para todas as subamostras (Sn)
4) Plotando Média S1, Média S2, Média S3..., Média Sn, de acordo com o teorema do limite central, nós teremos uma normal

O aparecimento de uma distribuição normal proveniente da distribuição populacional que é distorcida tem algumas aplicações muito importantes na prática estatística. Muitas práticas em estatística, como aquelas que envolvem testes de hipóteses ou intervalos de confiança, fazem algumas suposições sobre a população da qual os dados foram obtidos.

Outro fator importante do teorema do limite central é o cálculo da nova média e do novo desvio-padrão:

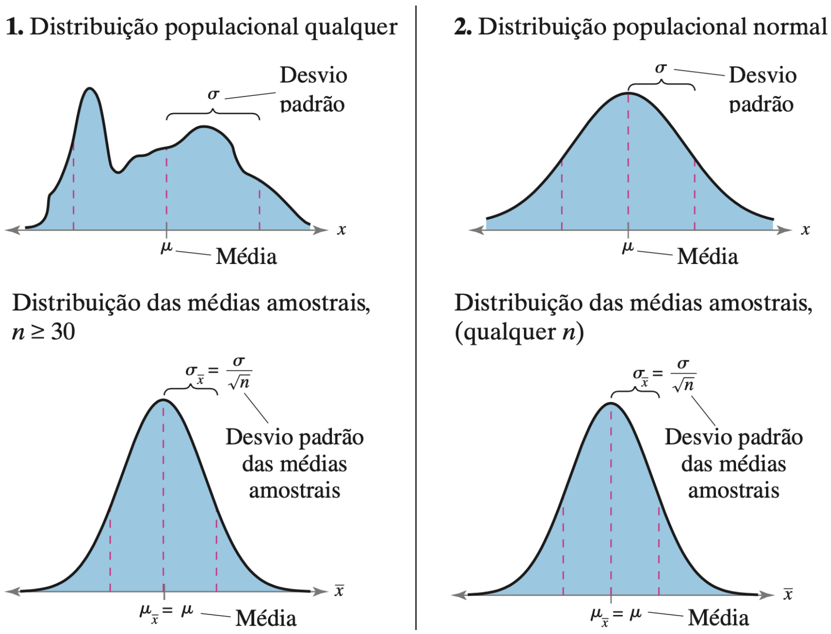
1) A média das médias amostrais será igual à média da distribuição original

$$\mu_{\bar{x}} = \mu$$

2) A distribuição amostral das médias tem uma variância igual a 1/n vezes a variância da população e um desvio padrão igual ao desvio padrão da população dividido pela raiz quadrada de n

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Ou seja, a distribuição das médias amostrais tem a mesma média que a população, mas o seu desvio padrão é menor que o desvio padrão da população. Isso nos diz que a distribuição das médias amostrais tem o mesmo centro que a população, porém é mais concentrada. Além disso, a distribuição das médias amostrais torna-se cada vez menos dispersa (maior concentração em relação à média) conforme o tamanho **n** da amostra aumenta.



<u>Exemplo</u>

O gasto médio com alojamento e refeição, por ano, em faculdades de quatro anos é de US$ 9126 (supondo conjunto com mais de 1000 faculdades). Você seleciona aleatoriamente 40 dessas faculdades. Qual é a probabilidade de que a média de gastos com alojamento e refeição seja menor que US$ 9400? Suponha que os gastos tem desvio padrão de US$ 1500 na amostra original

Resposta:

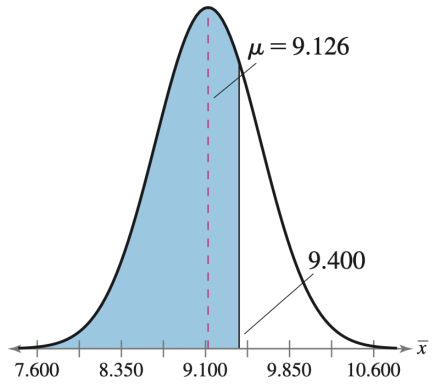
Selecionamos 40 amostras das 1000 originais. Como n>30 e como estamos querendo calcular a probabilidade da **média**, podemos usar o teorema do limite central.

De acordo com o teorema, a média das médias deve ser a mesma da população original. Logo, a média das médias amostrais deve ser US$ 9126.

O desvio-padrão das médias amostrais deve ser 1500/√40 = US$ 237.17.

Logo, agora temos uma distribuição normalizada com média US$ 9126 e desvio-padrão US$ 237.17. A partir dessa distribuição, podemos aplicar tudo que vimos sobre escore-z, pois essa é uma distribuição normalizada.

Nossa nova distribuição:



Queremos encontrar o valor de z que corresponde a 9400 e, posteriormente, na tabela de z, queremos encontrar a probabilidade acumulada até esse valor. Logo:
Z = (9400 - 9126)/ 237.17 = 1,15.

De acordo com a tabela de z, P(z<1,15) = 0.8739.

TRANSFORMAÇÕES

Como já dissemos, vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma distribuição normal (em forma de sino) ou então mais ou menos simétrica. Mas, em muitas situações de interesse prático, a distribuição dos dados da amostra é assimétrica e pode conter valores atípicos, como vimos em exemplos anteriores.

Se estamos interessados em entender probabilidades de médias amostrais, conseguimos superar esse problema com o teorema do limite central. Entretanto, se estivermos interessados em qualquer outra medida, precisamos pensar em outras alternativas. O que se propõe é efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. Aqui vamos ver algumas transformações mais usuais.

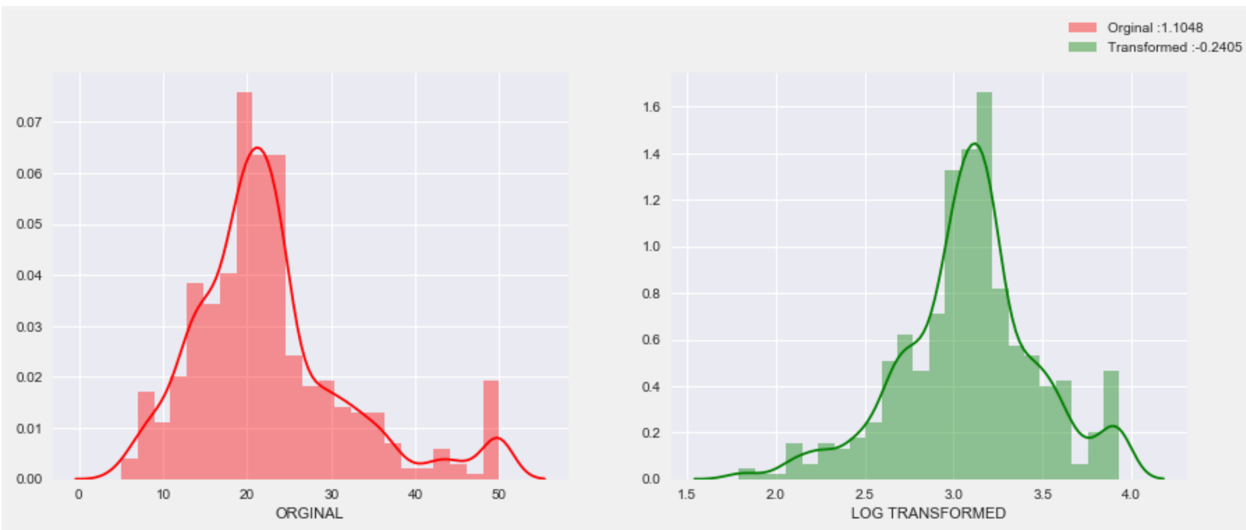
1. Transformação de Log

Variáveis numéricas podem ter distribuição altamente assimétrica e não normal (Distribuição Gaussiana) causada por outliers, distribuições altamente exponenciais, etc. Portanto, optamos pela transformação de dados.

Na transformação Log cada variável de x será substituída por log(x) com base 10, base 2 ou log natural.

Em Python teríamos, caso quisessemos aplicar log em uma variável chamada "Target", teríamos

```
import numpy as np
log_target = np.log1p(df["Target"])
```



O gráfico acima é a comparação dos dados originais e transformados em Log. Aqui vemos que a assimetria é reduzida nos dados transformados (melhor valor de inclinação deve ser quase zero)

2. Box-cox

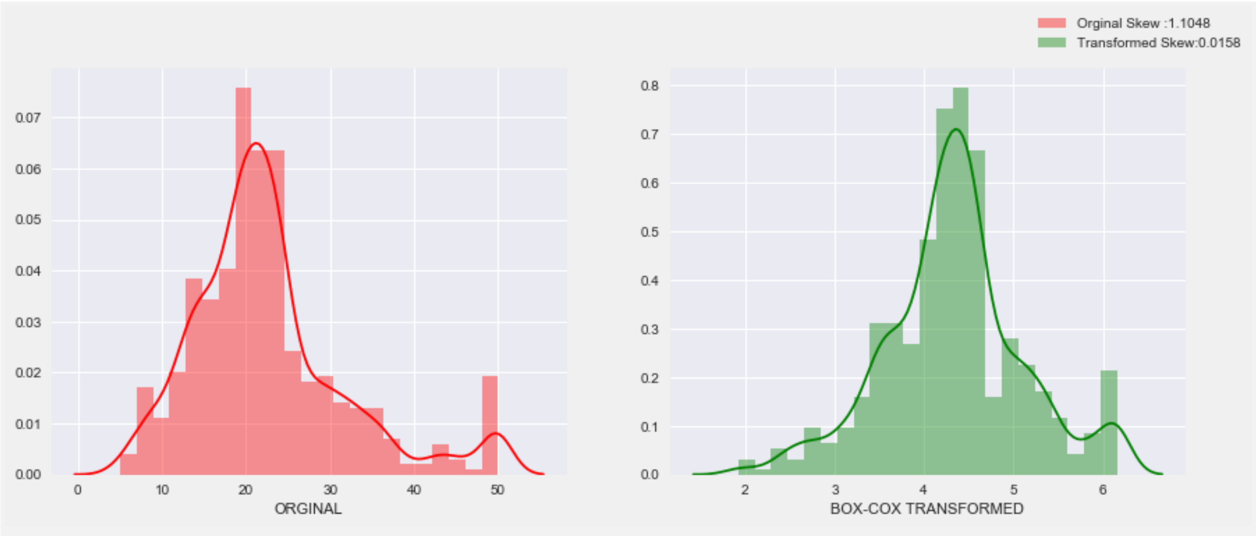
É uma das técnicas de transformação mais usadas. A transformação Box-cox funciona muito bem para muitas naturezas de dados. A imagem abaixo é a fórmula matemática para a transformação Box-cox.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Valores de lambda de -5 a 5 vão ser testados e o melhor valor para os dados é selecionado. O valor “Melhor” é aquele que resulta na menor assimetria da distribuição. A transformação de log ocorrerá quando tivermos lambda zero.

Em Python, faríamos:

```
from scipy.stats import boxcox
bcx_target, lam = boxcox(df["Target"])
#lam vai ser o lambda que der a menor assimetria
```



Aqui, notamos que a função Box-cox reduziu a assimetria e é quase igual a zero. Funcionou bem!

Vocês verão nos capítulos de teste de hipótese que existem diversos tipos de testes que podem ser aplicados **apenas** se a distribuição dos dados forem normais. Entretanto, se nossos dados não forem normalmente distribuídos, podemos usar essas transformações para que, após a transformação, o dado se torne normalmente distribuído.

REFERÊNCIAS BIBLIOGRÁFICAS

Bussan, W., Morettin, P. - Estatística Básica - Editora Saraiva - 2010 - 6th ed

Frost, J. - Hypothesis testing - An intuitive guide for making data driven decisions - Jim Frost - 2020 - 1st ed

Huyen, C. - Designing machine learning systems - Editora O'Reilly - 2022 - 1st ed

Knaflic, C.N - Storytelling com dados: Um guia sobre visualização - Editora Alta Books - 2019

Larson, R., Farber B. - Estatística Aplicada - Editora Pearson - 2016 - 6th ed

Pinheiro, J., Cunha, S., S. Santiago, Gomes, G. - Probabilidade e Estatística: Quantificando a incerteza - Elsevier Editora Ltda - 2012