Bootcamp Data Analytics

Estatística
Probabilidade & Amostragem

WOMAKERS® CODE

Inferencia estatística & Probabilidade

Estatística

Estatística descritiva
"Coletar, Sumarizar e Apresentar os dados"

Medidas de tendencia central

Medidas de disperção

Análises de tipos de variáveis

Inferência estatística
"A partir dos dados de uma amostra assumir hipóteses / inferir resultados"

Probabilidade & Amostragem

Testes de Hipótese

Modelos

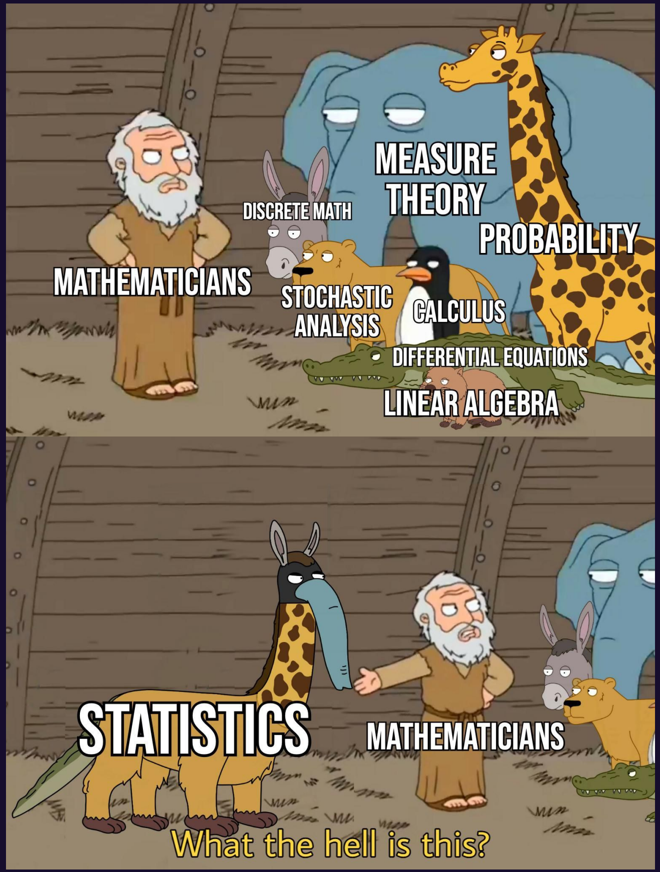
Aula de Hoje

Probabilidade

Probabilidade e Análise de dados

- A probabilidade estuda a "chance" de eventos ocorrerem e portanto é importante para conseguirmos como analistas ou cientistas de dados elaborar hipóteses, sendo a base da inferencia estatistica

- **Inferencia Estatística vs Probabilidade**: A probabilidade quantifica a incerteza dos eventos enquanto que a estatística é a ciência que nos permite a partir de amostras, inferir relações entre variáveis e hipóteses.



Conceitos de Probabilidade

Dados, Populacao vs Amostra

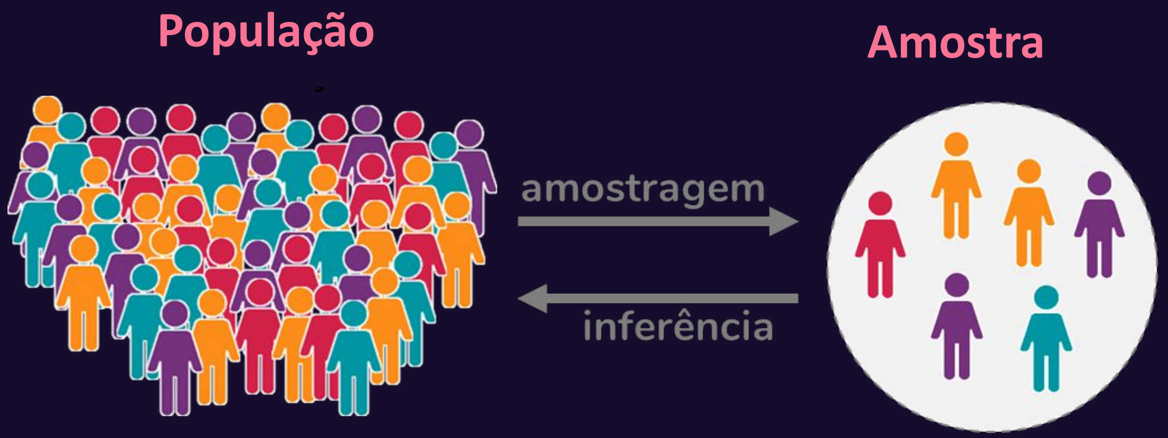
- **Dados:** observações documentadas ou resultados da medição. Os dados podem ser obtidos pela percepção através dos sentidos (por exemplo observação) ou pela execução de um processo de medição.

- **População**: É a população total de interesse sobre a qual desejamos obter informações.

  Ex: todas as transações bancárias de um banco

- **Amostra:** Conjunto formado por um subconjunto da população.

  Ex: transações bancárias de um determinado período de tempo (Safra) e uma região



Experimento, Espaço Amostral e Evento

- **Experimento**: Um procedimento performado com o objetivo de verificar e validar hipóteses, sendo realizado em condições controladas. Ex: jogar uma moeda não viesada 2 vezes seguidas.

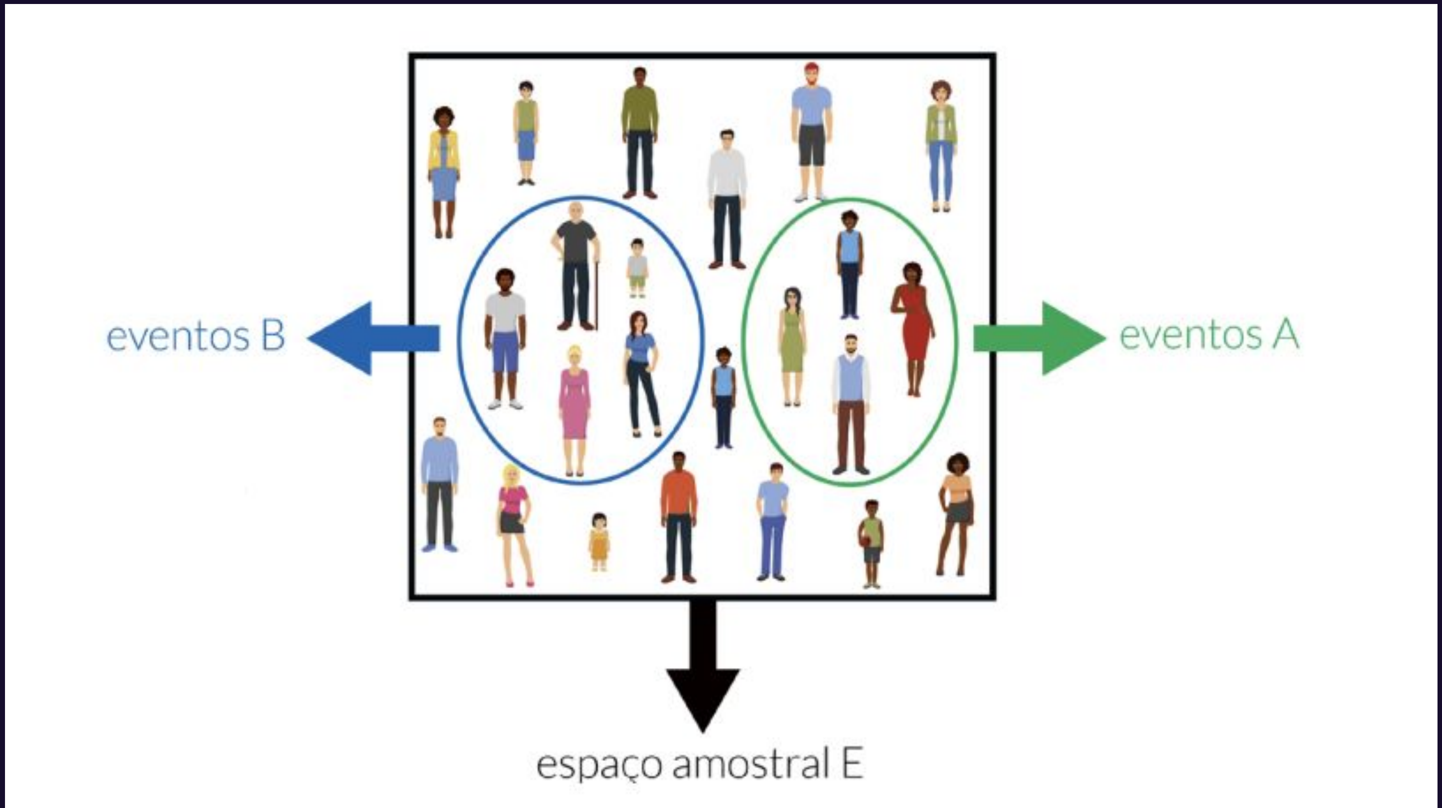
- **Espaco Amostral:** é o conjunto de todos os resultados possíveis de um experimento. Exemplo : Ω = {(Cara, Coroa), (Cara, Cara), (Coroa, Cara), (Coroa , Coroa)}

- **Evento:** É um subconjunto do espaço amostral. Exemplo: Evento "Sair pelo menos 1 vez cara" é representado por {(Cara, Coroa), (Cara, Cara), (Coroa, Cara)}

- **Evento complementar**: É aquele em que somado ao evento resulta no espaço amostral completo. No caso anterior, o evento complementar é {(Coroa, Coroa)}

Experimento, Espaço Amostral e Evento

- **Experimento:** Pesquisa Eleitoral
- **Espaco Amostral Ω:** contém todos os entrevistados, 10 mil pessoas
- **Evento A** : eleitores do candidato 1
- **Evento B**: eleitores do candidato 2



A Probabilidade de um Evento

A probabilidade de um evento A = "Sair pelo menos 1 vez cara" no experimento: jogar uma moeda não viesada 2 vezes seguidas.

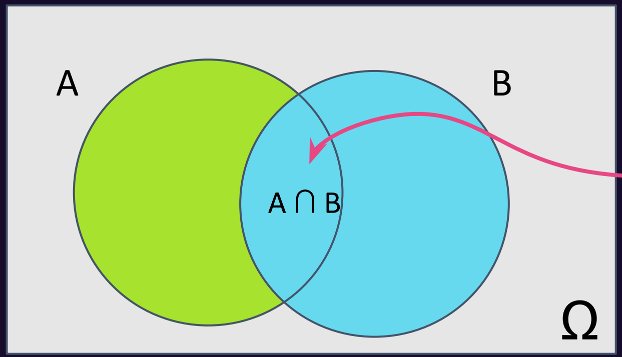
É denotada por:

$$P(A) = \frac{\#\text{Número de casos possíveis no evento A}}{\#\text{Número de casos possíveis no espaço amostral } (\Omega)}$$

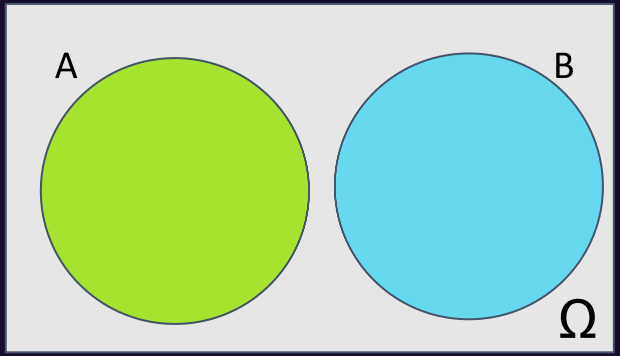
No exemplo: N(A) = 3 e N(Ω) = 4, logo P(A) = ¾ = 75%

Regras Básicas de Probabilidade

- 1. $0 \leq P(A) \leq 1$: A probabilidade de um evento A é no mínimo zero e máximo 1
- 2. $P(\Omega) = 1$: A probabilidade do espaço amostral é a máxima, 1.
- 3. $P(\varnothing) = 0$: A probabilidade do conjunto / event vazio é 0
- 4. Probabilidade da União de dois eventos:
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, sendo $P(A \cap B)$ a probabilidade da intersecção de dois eventos



Eventos não disjuntos (possuem interseccao)



Eventos disjuntos (intersecção é nula / não existe)

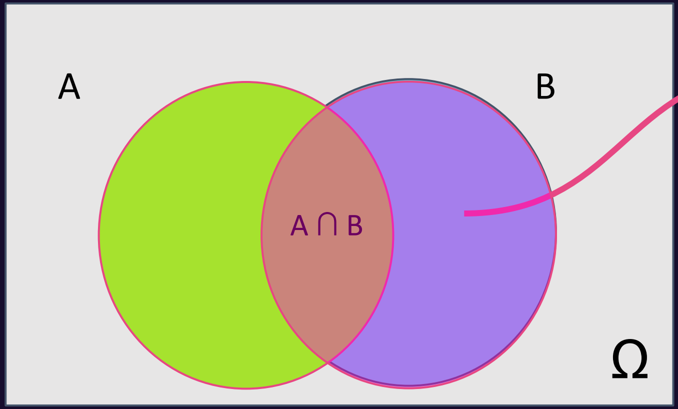
Regras Básicas de Probabilidade

- 5. P(A c) = 1 - P(A): A probabilidade de um evento A complementar é 1 - P(A)

- 6. P(A/B) é A probabilidade do evento A, quando se sabe que o evento B ocorreu, ´e chamada probabilidade condicional de A dado B
  - Na probabilidade condicional, a ocorrencia de um evento altera a probabilidade de ocorrência do outro.

  - $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ se $P(B) > 0$

Reescrevendo: $P(A \cap B) = P(B)*P(A|B)$



Dado que B ocorreu , A probabilidade condicional se restringe ao espaço amostral do evento ocorrido

Eventos Independentes ou Mutuamente Exclusivos

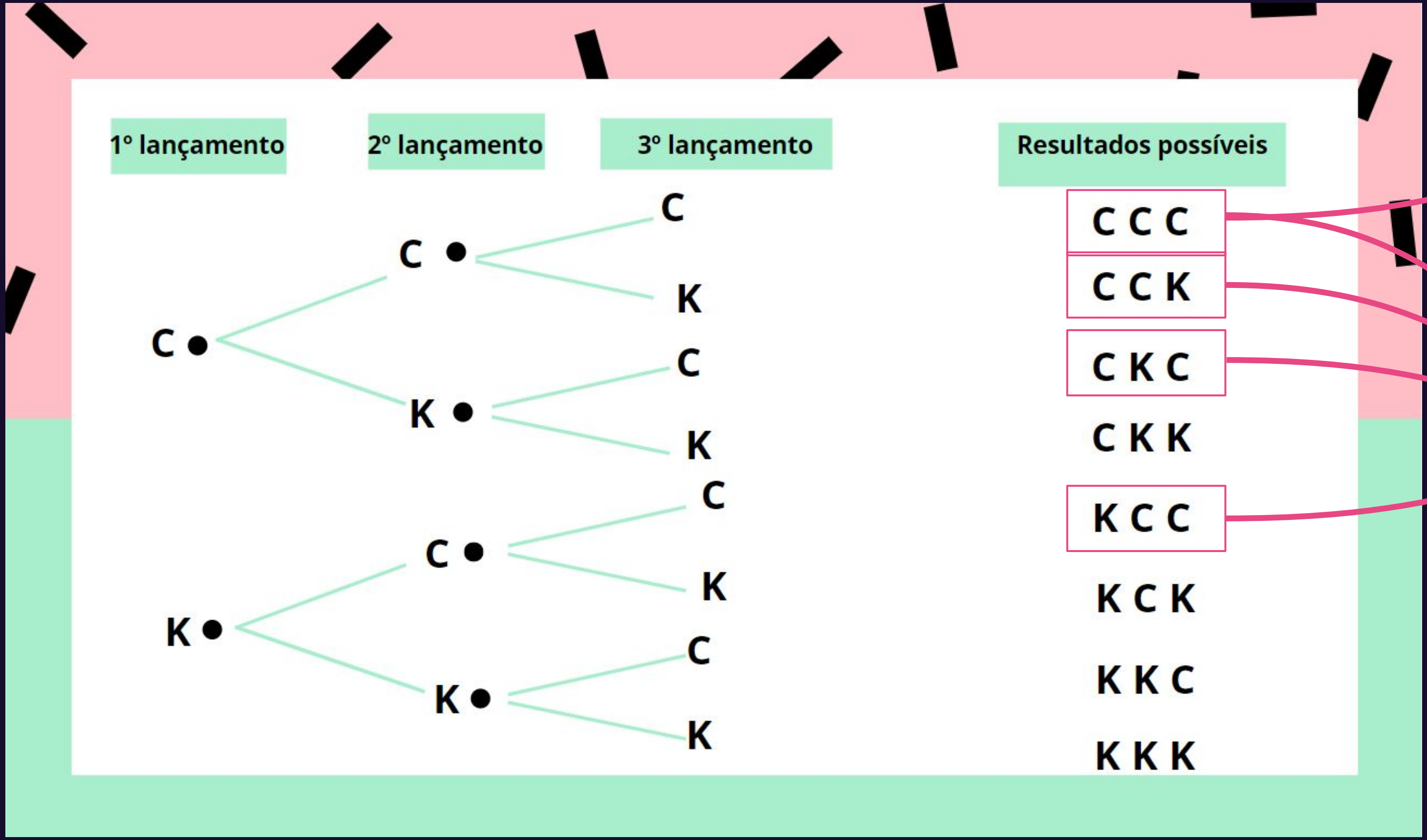
- 7. Dois eventos são independentes se a ocorrência de um deles não afeta a ocorrência do outro. Sendo assim :
    - $P(A \cap B) = P(B)*P(A|B) = P(B)*P(A)$, pois a ocorrência de A não depende de B.

- 8. Dois eventos são mutuamente exclusivos se: não podem ocorrer simultaneamente, ou seja $P(A \cap B) = \varnothing$,
  Assim : $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A)+P(B)$

Ao contrario de eventos mutuamente exclusivos, dois eventos independentes podem ocorrer simultaneamente, a diferença é que a ocorrência de um deles não afeta o outro.

Exemplo: Cara ou Coroa

Supondo o experimento de lançar 3 moedas seguidas. Qual a probabilidade de se obter 3 caras ? E a probabilidade de se obter pelo menos duas caras? Sendo cara representado por C e coroa K.

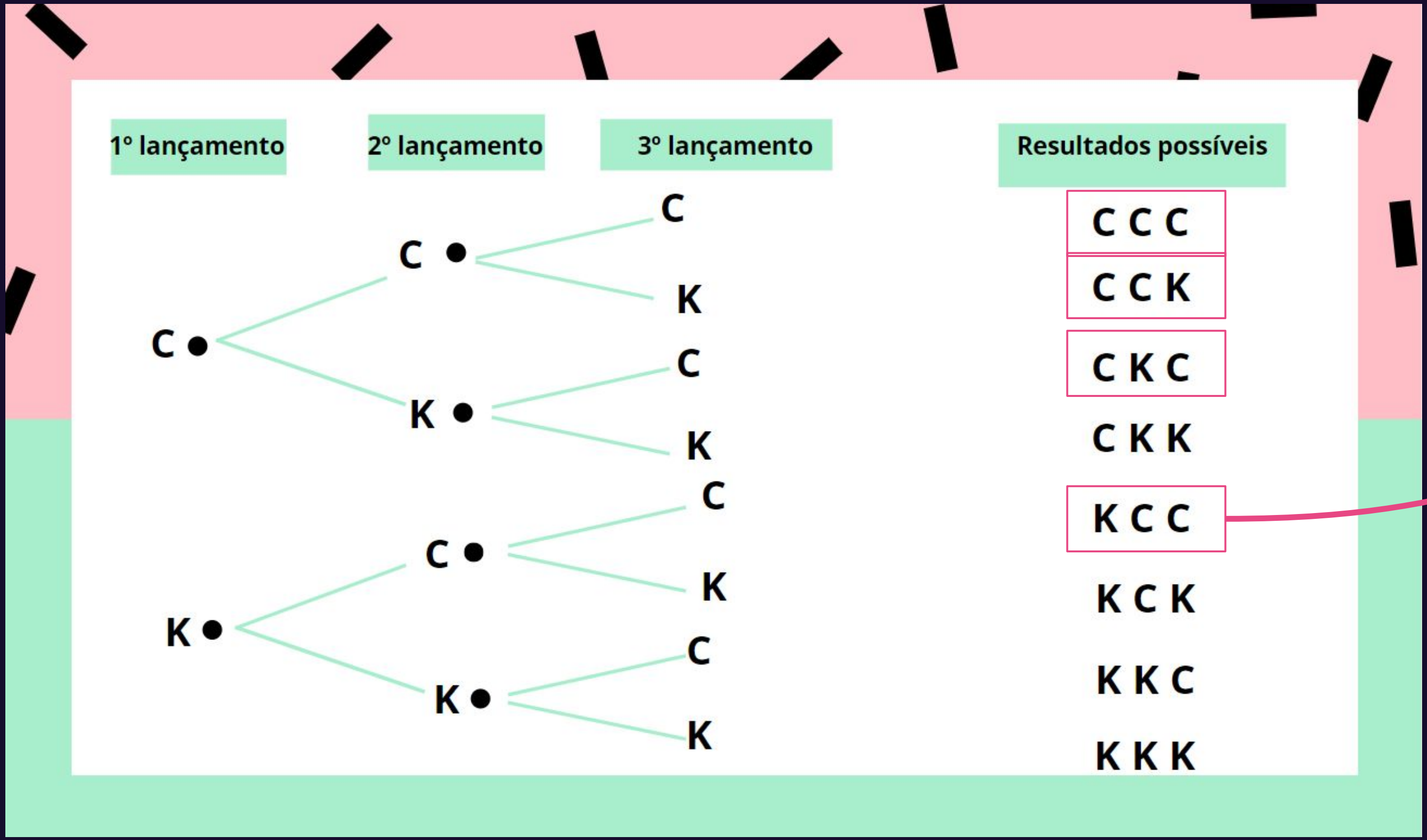


P(3 caras) = 1/8

P(Pelo Menos 2 Caras) = 4/8

Exemplo: Cara ou Coroa

Qual a probabilidade condicional de se obter pelo menos 2 caras dado que o primeiro lançamento foi Coroa, K?



P(Pelo Menos 2 Caras dado que o primeiro foi Coroa, K,) = 1/4

Distribuições de Probabilidade:

Variável Aleatória

"Uma variável aleatória é uma função associa cada evento do espaço amostral à um número real."

Simplificando: Supondo o experimento de jogarmos 2 moedas

Ω = {CaraCara; CaraCoroa; CoroaCara; CoroaCoroa}

Fenômeno Probabilístico / evento do espaço amostral

Podemos definir uma variável aleatória X como o número de Caras em dois lançamentos. Desse modo, os valores possíveis de X são:

$X = \{0, 1, 2\}$

Número real

Variável Aleatória

Uma variável aleatória pode assumir valores discretos ou contínuos e ela é escrita como convenção com uma letras maiúscula. Ex: X = {0, 1, 2}, enquanto uma observação dessa variável é escrita com letra minúscula, x = 0.

- No caso da cara ou coroa que vimos uma variável ***aleatória discreta***, pois ela tem ***observações finitas e enumeráveis***: 0, 1 ,2)

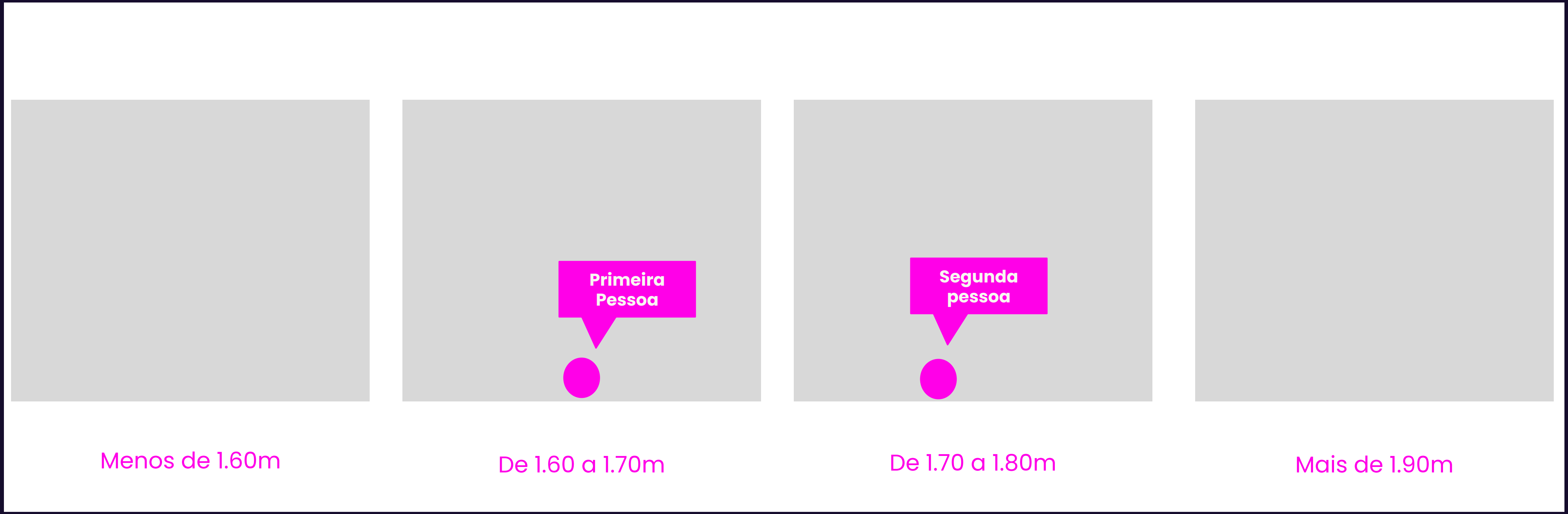
- Uma ***variável aleatória contínua***, diferentemente, pode assumir ***qualquer valor dentro de um intervalo e portanto é não enumerável.***

Ex V. A. contínua: Quantidade de açúcar em um café ; O Tempo para finalizar uma prova;

Note que : O tempo para realizar uma prova poderia ser 2 minutos , mas se melhorar a precisão poderia ser 2.05 , ou 2.049482 e assim por diante, portanto não é um valor numerável

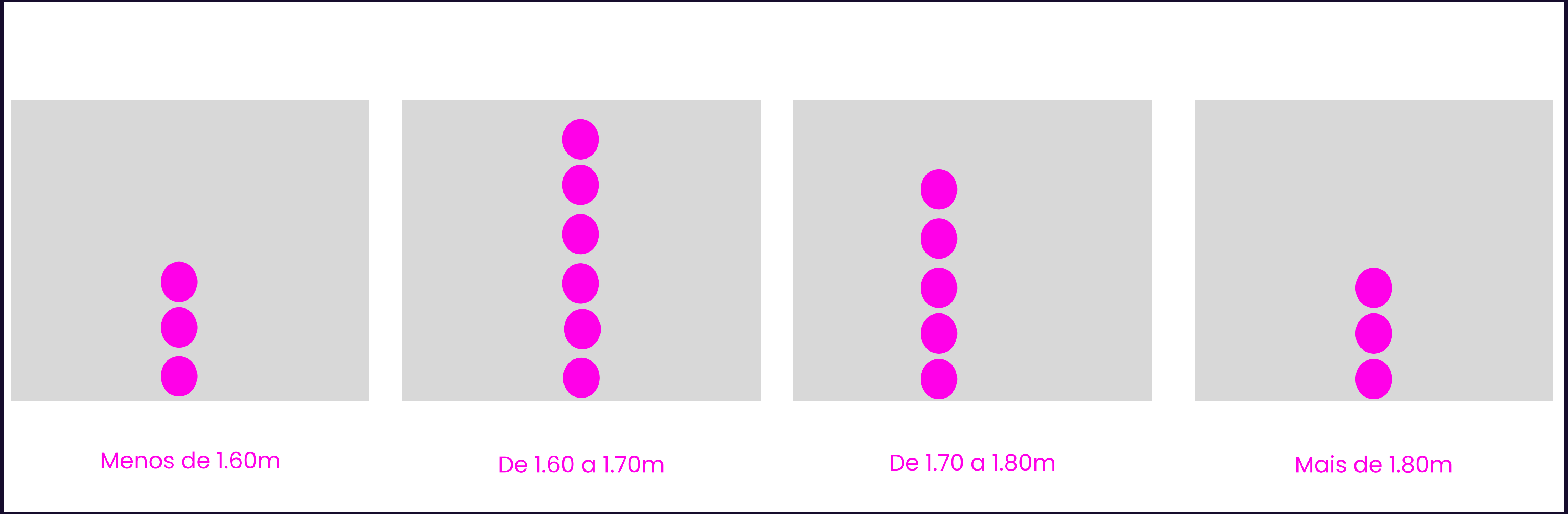
Função Densidade de Probabilidade

Imagine que estamos medindo a altura de várias pessoas.



Função Densidade de Probabilidade

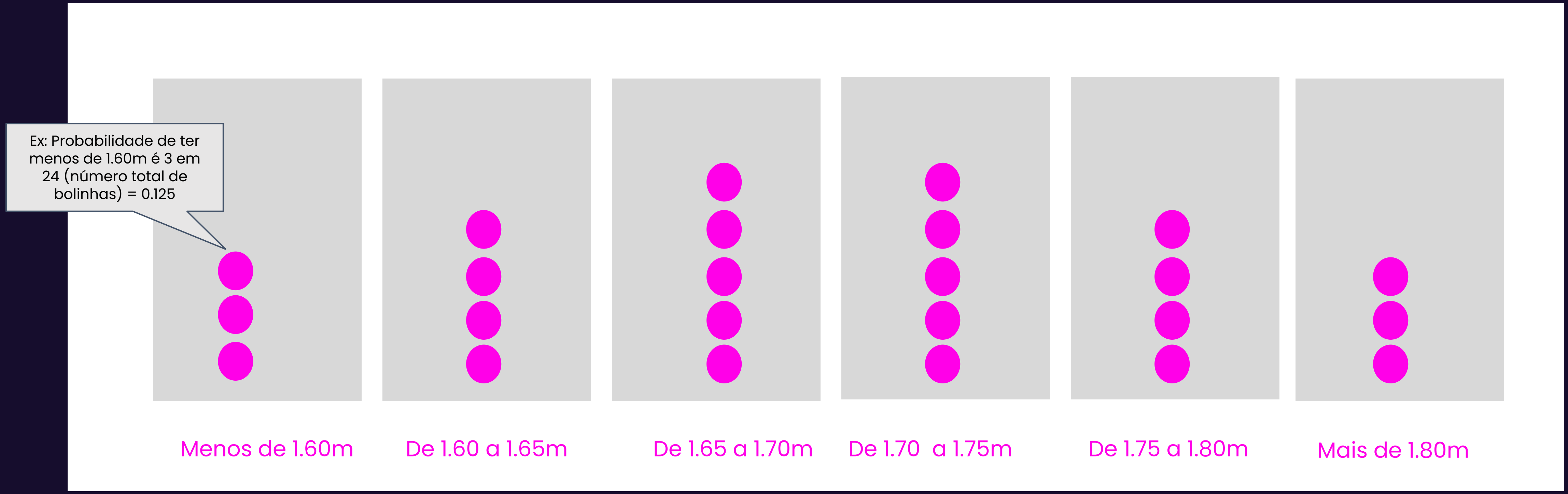
A medida que vamos mensurando as alturas e vamos categorizando esses indivíduos obtemos um histograma.



- A maioria das pessoas têm alturas de 1.60 a 1.70m de modo que as alturas maiores e menores são mais raras.
- Assim, se selecionarmos uma pessoa aleatoriamente ela teria maior probabilidade de estar no grupo de 1.60 a 1.70m

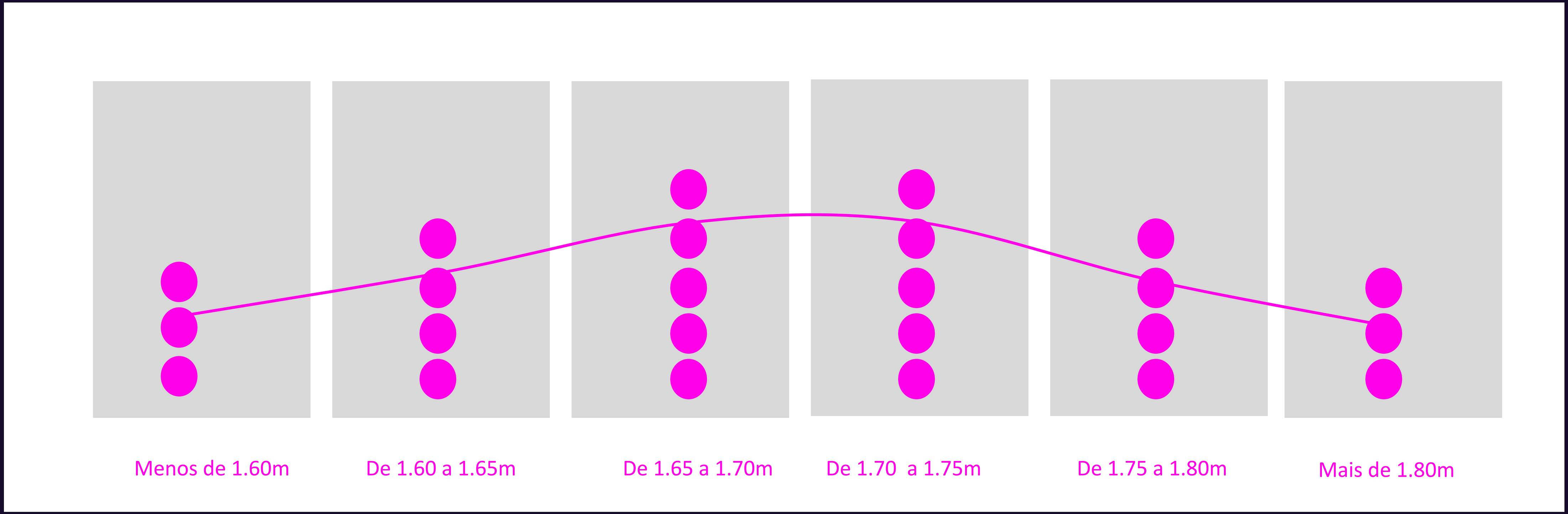
Função Densidade de Probabilidade

Poderíamos usar mais grupos ou seja, usarmos intervalos mais estreitos e assim temos maior precisão



Função Densidade de Probabilidade

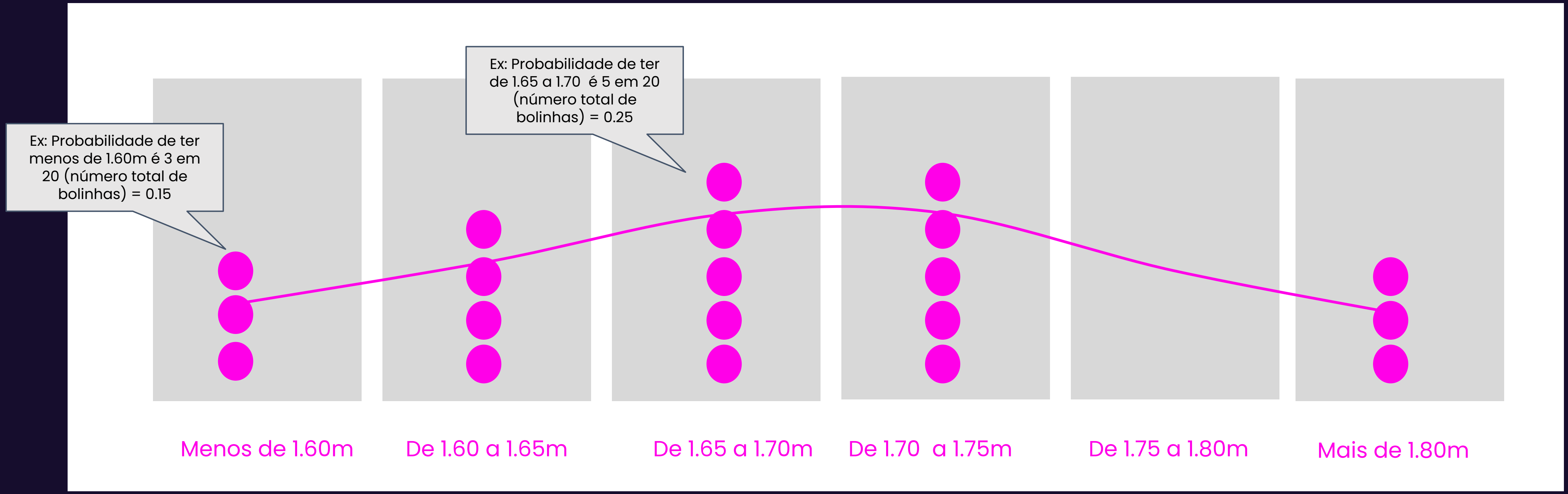
Com maiores bins, podemos usar uma curva para aproximarmos esses dados



Essa curva nos dá a mesma intuição do histograma. As caldas apresentam probabilidades menores e o centro maior probabilidade.

Função Densidade de Probabilidade

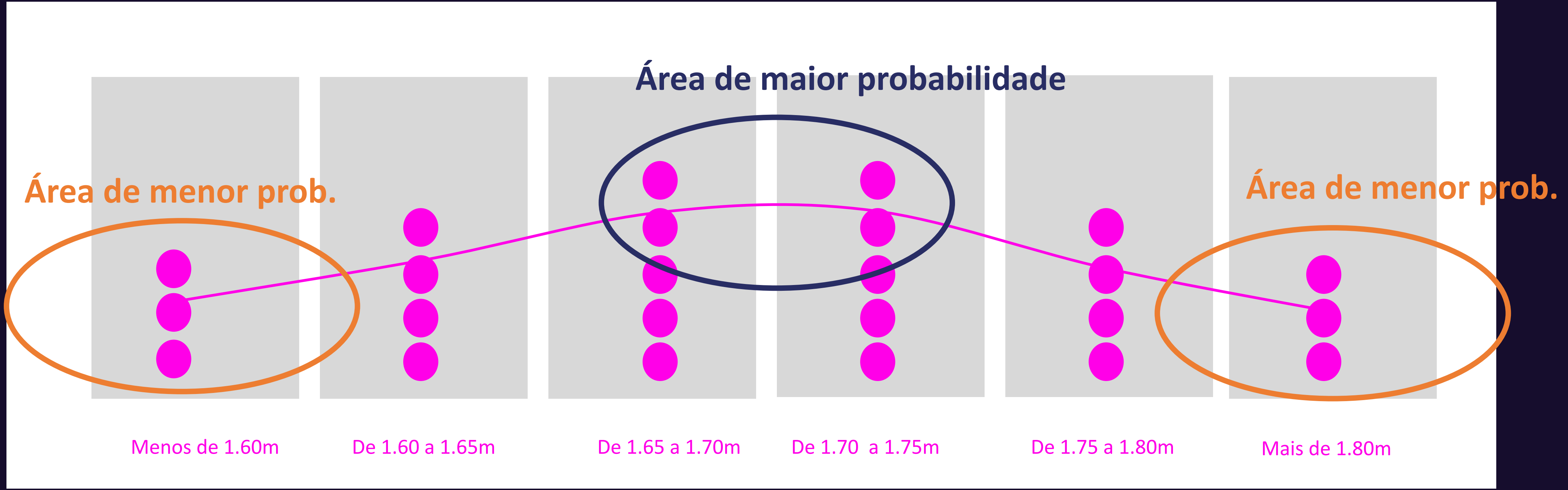
A curva entretanto nos dá uma grande vantagem. Imagine que não tenhamos obtido em nossa amostra nenhum indivíduo com altura de 1.75 a 1.80m



Com a curva podemos estimar a probabilidade de um indivíduo ter 1.75 a 1.80m de altura!!
E com a curva poderiamos calcular a probabilidade de um indivíduo conter intervalos de altura!! ex: Probabilidade de ter altura entre 1.60 a 1.75.

Função Densidade de Probabilidade

A curva estimada e o histograma são portanto distribuições de probabilidade



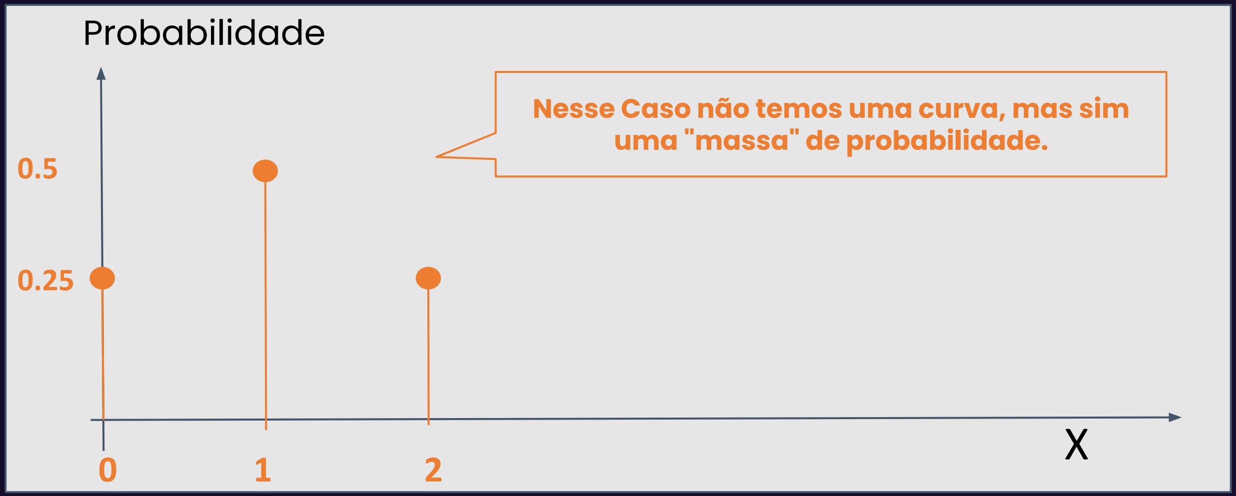
A função que descreve a curva pode ser descrita como: $f(x) = P(X = x)$

Função Massa de Probabilidade

Dado esse contexto vimos que a **função densidade de probabilidade** descreve a probabilidade de ocorrência de cada valor da minha variável aleatória contínua, no caso a altura dos indivíduos.

Mas e no caso das variáveis discretas?
Nesse caso não teremos uma curva ou seja uma função densidade de probabilidade. Teremos o que chamamos de **função massa de probabilidade**.

Exemplo: Variável aleatória X = Número de Caras em dois lançamentos de uma moeda



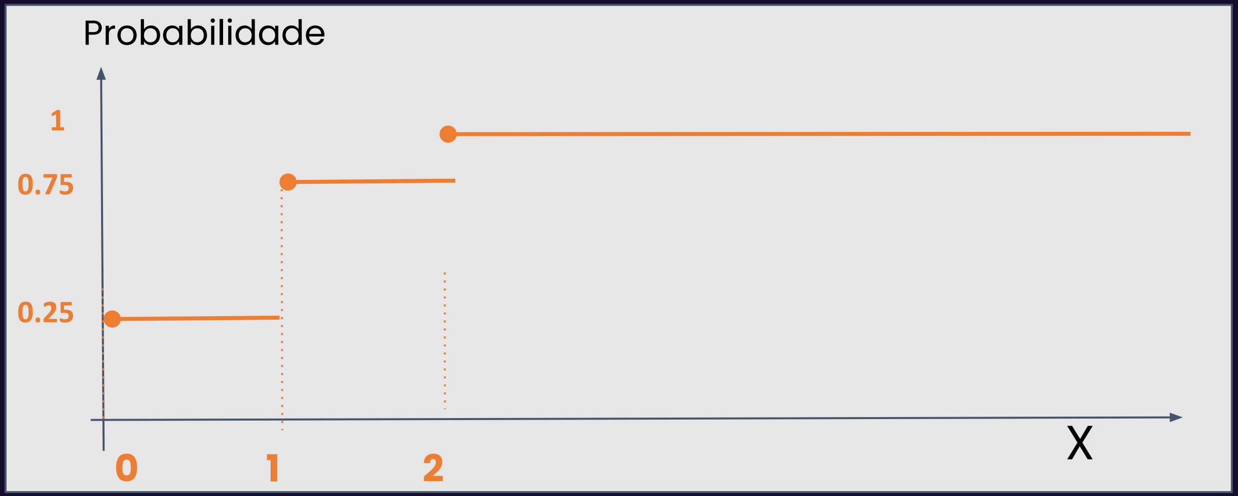
Ω = {CaraCara; CaraCoroa; CoroaCara;CoroaCoroa}

X	f(X)
0	0.25
1	0.5
2	0.25

Função de Densidade Acumulada.

- Vimos a função densidade e massa de probabilidade, expressada em cada evento possível $f(x) = P(X = x)$.
- Mas uma outra função muito utilizada por analistas e cientistas de dados é a função densidade de probabilidade acumulada.
- A densidade acumulada calcula a probabilidade acumulada para um determinado valor de x. E é expressada por $F(x) = P(X<=x)$

Ex: Vamos calcular a probabilidade acumulada do exemplo anterior. Temos a V.A X = Número de Caras em dois lançamentos de uma moeda



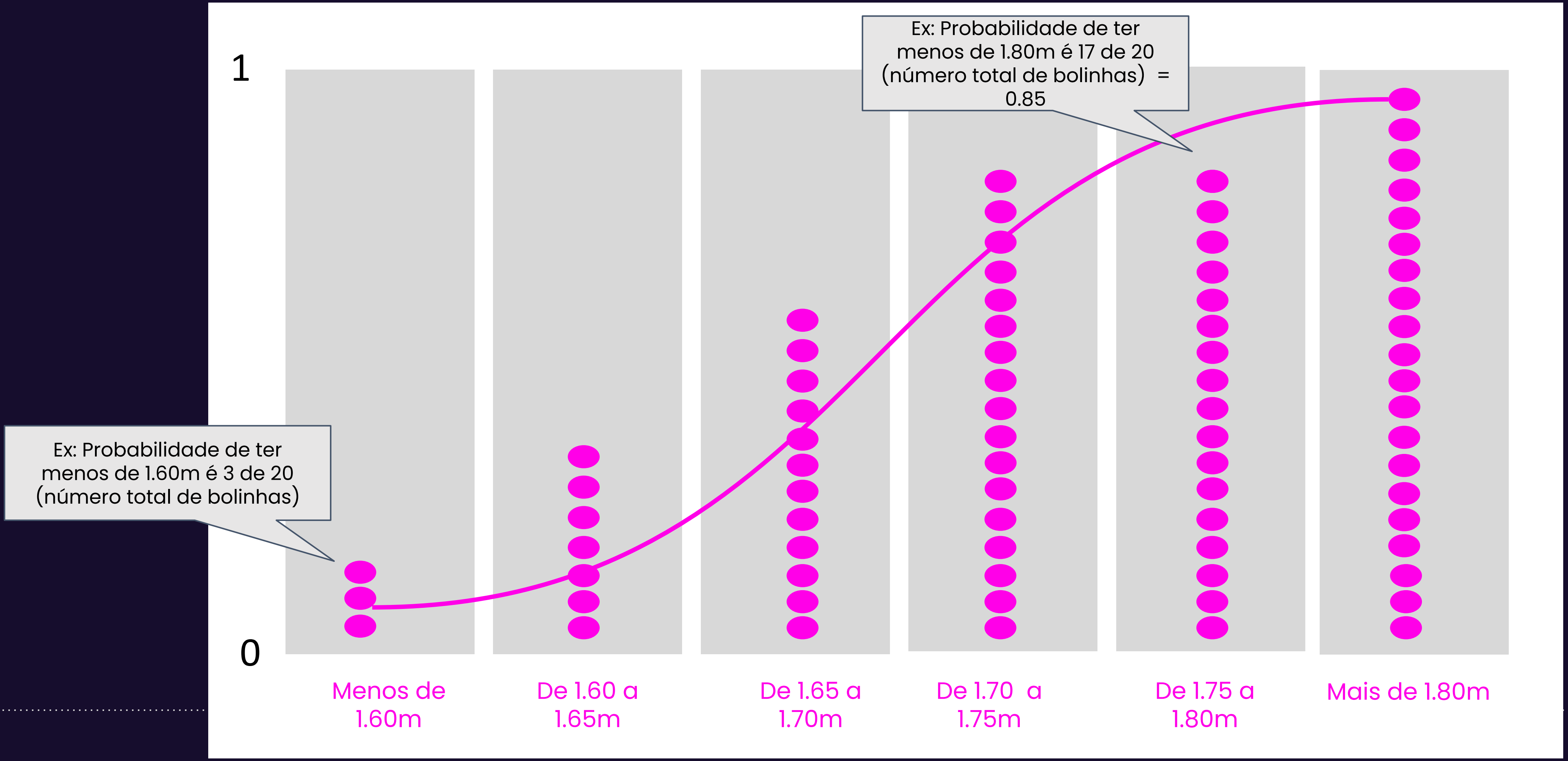
Ω = {CaraCara; CaraCoroa; CoroaCara;CoroaCoroa}

X	f(X)	F(X)
0	0.25	0.25
1	0.5	0.75
2	0.25	1

Função de Densidade Acumulada.

Ex2: No caso da altura, para calcular a FDA vamos "somando" as probabilidades possíveis de cada evento de X.

Note que: TODA FDA vai de 0 a 1



Distribuições Discretas vs Contínuas

A distribuição de probabilidade é o processo que descreve o comportamento aleatório de fenômenos. E de acordo com as características dos processos aleatórios podemos classificá-las em 2 grandes grupos.

Discretas

- Binomial
- Geométrica
- Poisson
- Uniforme discreta

Contínuas

- Normal
- Uniforme contínua
- Exponencial
- Gamma
- Valores Extremos

Distribuições de Probabilidade: Discretas

Distribuição Binomial

A distribuição binomial descreve situações em que os resultados de uma variável podem ser agrupados em duas categorias mutuamente excludentes (Ex: sucesso ou falha)

Características da distribuição binomial

Uma distribuição de probabilidade binomial resulta de um experimento que satisfaz os seguintes requisitos:

1. O experimento tem um número finito de tentativas.
2. As tentativas devem ser independentes (o resultado de qualquer tentativa individual não afeta as probabilidades nas outras tentativas).
3. Cada tentativa deve ter todos os resultados classificados em duas categorias (em geral, chamadas de sucesso e fracasso).
4. A probabilidade de sucesso permanece constante em todas as tentativas.

Ex: Número de vezes que sai coroa no lançamento de 3 moedas não viciadas..

(3 tentativas ; cada tentativa com probabilidade ½ constante; cada tentativa pode ter 1 sucesso e 1 fracasso)

Probabilidade de sucessos de n tentativas

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

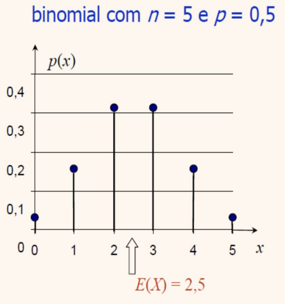
p= a probabilidade de sucesso em uma tentativa;
q= a probabilidade de fracasso em uma tentativa;
n= o número de tentativas;
x= a quantidade de sucesso nas n tentativas.
sendo q = 1-p

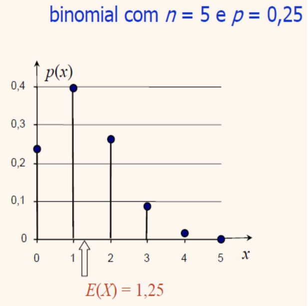
Medidas

média (u)	n*p
variancia	n*p*q

Lembrando que a média amostral pode ser expressada por E(x), vs a populacional por u

Exemplos:





Distribuição Geométrica

A distribuição geométrica é responsável por representar eventos ou tarefas repetidas até que um sucesso ocorra. Por exemplo, a probabilidade de uma vendedora de telemarketing realizar uma venda, na sexta ligação.

Características da distribuição Geométrica

Uma distribuição de probabilidade geométrica resulta de um experimento que satisfaz os seguintes requisitos:

1. A tentativa deve ser repetida até que um sucesso ocorra;
2. Cada tentativa é independente;
3. A probabilidade de sucesso é a mesma em cada tentativa;
4. A variável aleatória X representa o número de tentativas até o primeiro sucesso.

Probabilidade em que o primeiro sucesso ocorra

$$P(X = k) = (1 - p)^k p$$

p= a probabilidade de um sucesso
k = número de tentativas

Medidas

média (u)	1/p
variancia	(1-p)/p^2

Exemplos:

Ex: Suponha que uma vendedora de telemarketing tem a probabilidade de vender em uma ligação p = 15%.

Qual a probabilidade dela vender somente na terceira ligação do dia?

P(X = 3) = ((1-0.15)^3)*0.15 = 9,21%

Distribuição Poisson

A distribuição de Poisson descreve resultados de experiências nos quais contamos acontecimentos que ocorrem aleatoriamente a uma taxa média definida. Por exemplo: o nº de bebés que nasce por mês num determinado hospital; o número de peças fabricadas por dia

Características da distribuição Poisson

Uma distribuição de probabilidade poisson resulta de um experimento que satisfaz os seguintes requisitos:

1. Dois eventos não podem ocorrer simultaneamente.
2. ***A taxa média entre a ocorrência do evento é constante.***
3. Os eventos são independentes um do outro (se um acontecer, isso não tem nenhuma influência sobre a probabilidade de outro evento ocorrer).
4. Os eventos podem ocorrer em qualquer número de vezes

Ex: Em uma indústria qual a probabilidade de 10 mil peças serem produzidas em 1 dia.

Probabilidade de ocorrência de x vezes em um intervalo de tempo:

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!}$$

p= a probabilidade de ocorrência de x em um intervalo de tempo;
λ = a taxa de ocorrência do evento.
e = 2,7182 aproximadamente

Medidas

média (u)	λ
variancia	λ

No caso de poisson, a média e igual a variância que será igual a taxa média de ocorrência

Exemplos:

Em uma loja estima-se que entram 5 clientes a cada 10 minutos. Qual a probabilidade de entrarem 4 pessoas em um período qualquer de 10 minutos?

Para responder a essa pergunta, considere uma distribuição de Poisson com média igual a:
$\lambda=5$;

Substituindo na fórmula P(4) = 0.1755

Distribuição Uniforme Discreta

A distribuição uniforme descreve eventos equiprováveis; Ex: Considere a variável aleatória X o valor obtido em um lançamento de um dado. Podemos obter 1 com probabilidade ⅙ ; até 6 com mesma probabilidade.

Características da distribuição Uniforme discreta

Uma distribuição de probabilidade uniforme resulta de um experimento que satisfaz os seguintes requisitos:

1. A variável aleatória X assume valores de 1 até N.
2. Cada valor de x tem uma igual probabilidade de ocorrência
3. Os eventos são independentes um do outro (se um acontecer, isso não tem nenhuma influência sobre a probabilidade de outro evento ocorrer).

Probabilidade

$$P(X = x) = \frac{1}{N}$$

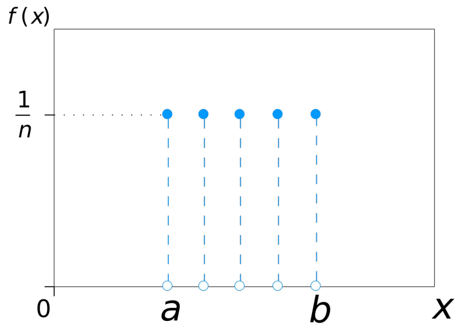
p= a probabilidade de ocorrência de x
N = valor máximo que a variável pode assumir

Medidas

média (u)	(N + 1)/2
variancia	(N^2 -1)/12

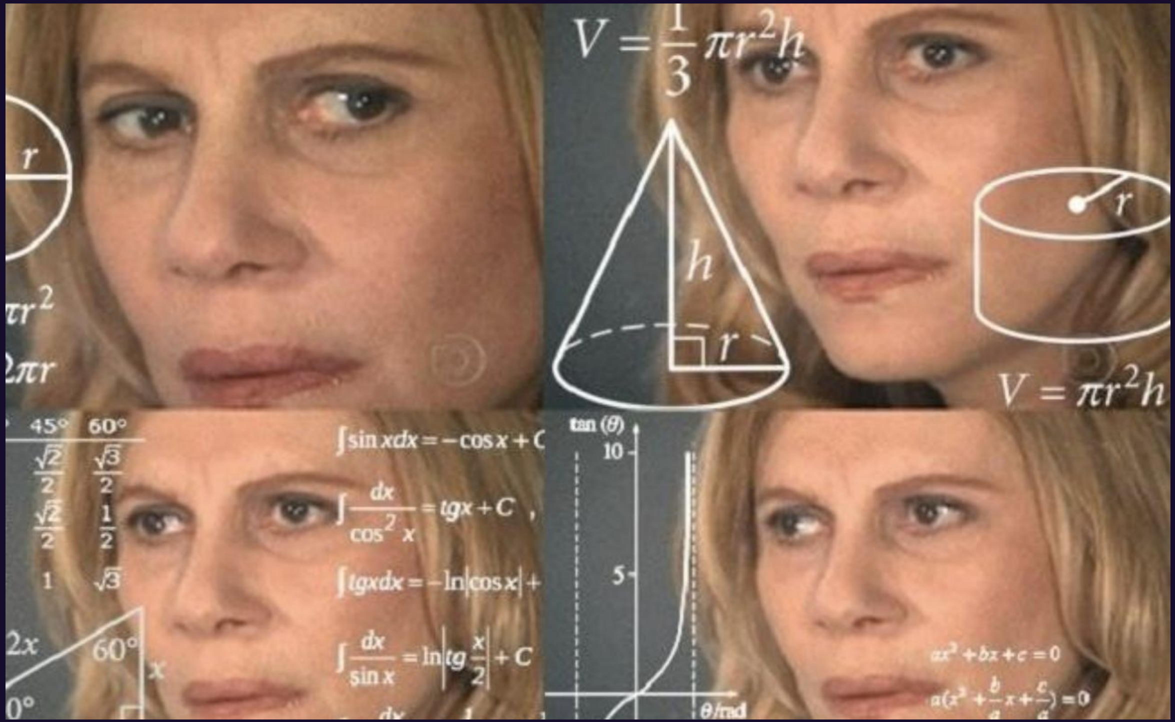
Podemos encontrar fórmulas para intervalos de [a, b] ao inves de 1 até N. Nesse caso: P(X =x) = 1 / [b - a] Média = (a + b)/2 ; Variancia = (b-a)^2/12

Visualmente:



- Agora que vimos distribuições discretas, podemos nos perguntar: Como mensurar P(X = 1,75)?

- E se diminuíssemos cada vez mais os intervalos entre os valores possíveis de X; Ou seja e se x assumisse qualquer número Real, e não mais discreto?



Distribuições de Probabilidade: Contínuas

Distribuição Uniforme Contínua

Assim como no caso discreto, a distribuição uniforme contínua descreve eventos equiprováveis agora em um universo real. Por exemplo: Imagine que você chegou no ponto de ônibus agora e seu ônibus passa de 1 em 1 hora. Você não sabe a hora em que o último ônibus passou. Sendo assim o horário de chegada do próximo ônibus segue uma distribuição uniforme contínua no intervalo de horas de 0 a 1.

Características da distribuição Uniforme contínua

Assim como no caso discreto, uma distribuição de probabilidade uniforme resulta de um experimento que satisfaz os seguintes requisitos:

1. A variável aleatória X assume valores de A até B.
2. Cada valor de x tem uma igual probabilidade de ocorrência
3. Os eventos são independentes um do outro (se um acontecer, isso não tem nenhuma influência sobre a probabilidade de outro evento ocorrer).

Probabilidade

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \leq x \leq b \\ 0 & \text{caso contrário} \end{cases}$$

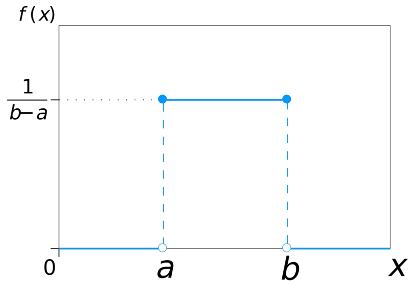
a = valor mínimo do intervalo que a variável pode assumir
b = valor máximo do intervalo que a variável pode assumir

Note que: No caso contínuo P(x) = f(x) pois a probabilidade é uma função contínua no intervalo

Medidas

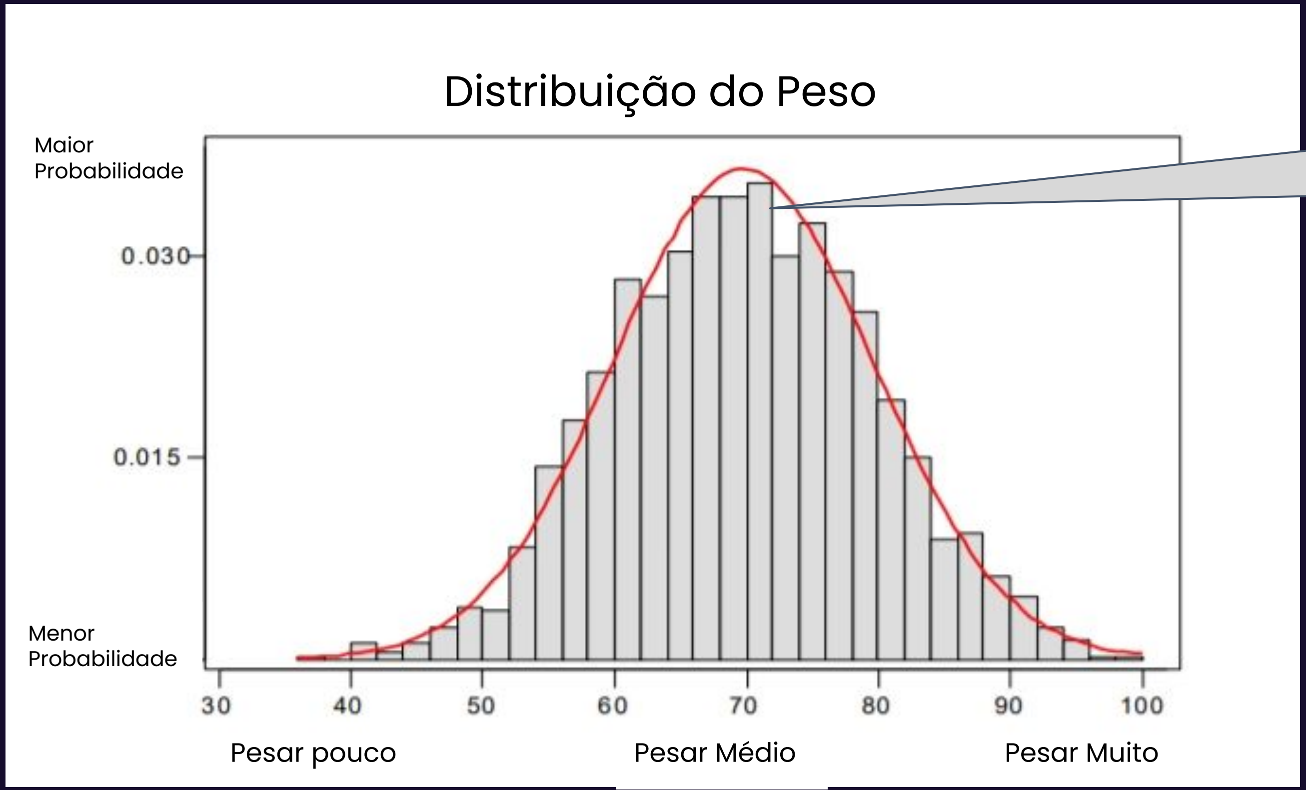
média (u)	(a + b)/2
variancia	(b -a)^2/12

Visualmente:



Distribuição Normal

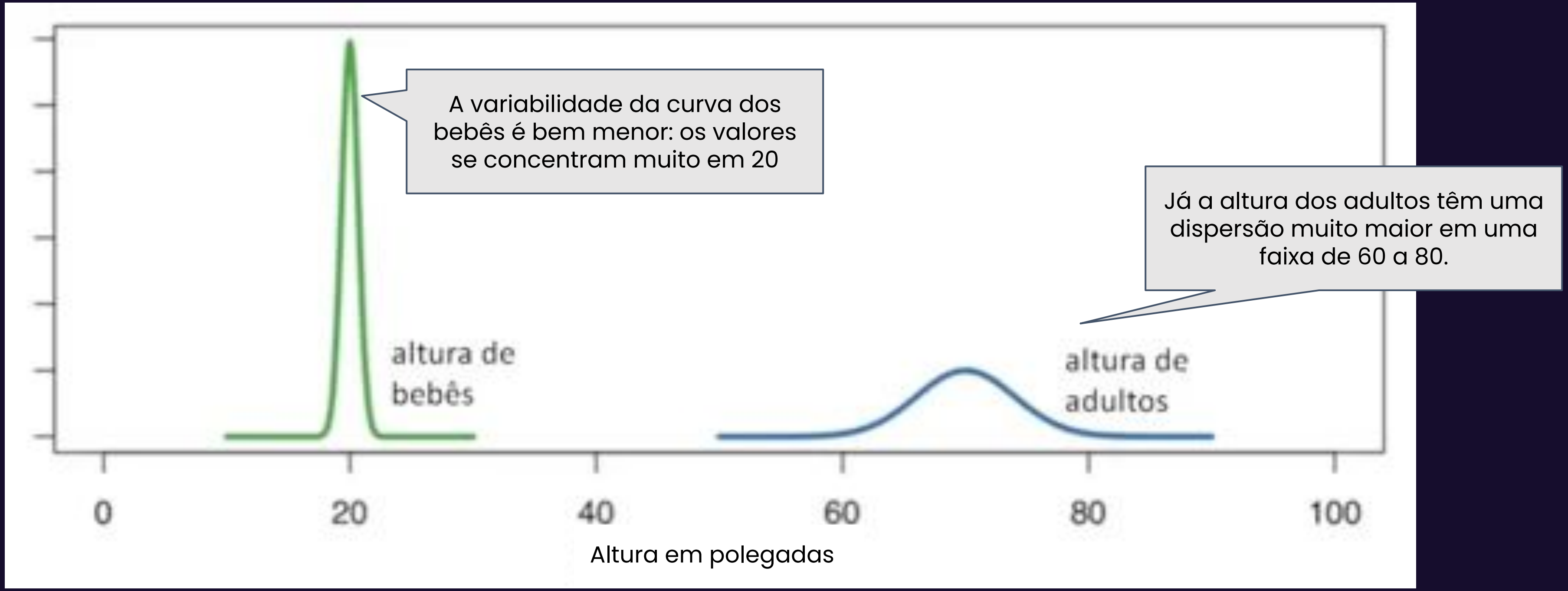
A distribuição normal é uma das mais importantes distribuições da estatística. Ela é tão importante pois representa a probabilidade de muitos eventos na natureza.



Distribuição normal é centrada em valores médios. A Área de maior probabilidade é no centro de 60 a 80 kg

Distribuição Normal

O Formato da distribuição Normal também nos dá uma intuição à respeito da variabilidade dos dados.



Distribuição Normal

A distribuição normal é uma das mais importantes distribuições da estatística. Ela é tão importante pois representa a probabilidade de muitos eventos na natureza.

Características da distribuição Normal

Uma distribuição de probabilidade Normal resulta de um experimento que satisfaz os seguintes requisitos:

1. Curva em formato de sino.
2. Distribuição simétrica em torno da média
3. Não chega a tocar o eixo x, vai se aproximando no infinito
4. É delimitada pelo seu grau de dispersão (desvio padrão) e medida central, sua média
5. O pico da curva , valor de maior probabilidade está na média
6. A área embaixo da curva é um percentual

Probabilidade

$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}$$

u = média populacional

σ = desvio padrão populacional

e = 2,71828 aproximadamente

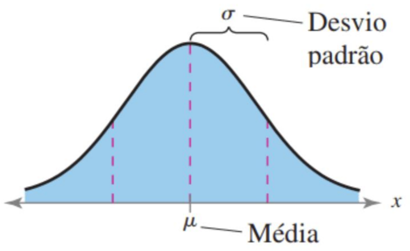
Note que: No caso contínuo P(x) = f(x) pois a probabilidade é uma função contínua no intervalo

Medidas

média	u
variancia	σ^2

Visualmente:

Distribuição populacional normal



Distribuição Normal Padrão

Uma distribuição normal muito conhecida é a Normal Padrão. Cuja média é 0 e desvio padrão é 1. Ela é chamada de distribuição Z.

Qualquer distribuição normal pode ser "padronizada" convertendo os valores de x em z-scores.

Características da distribuição Normal

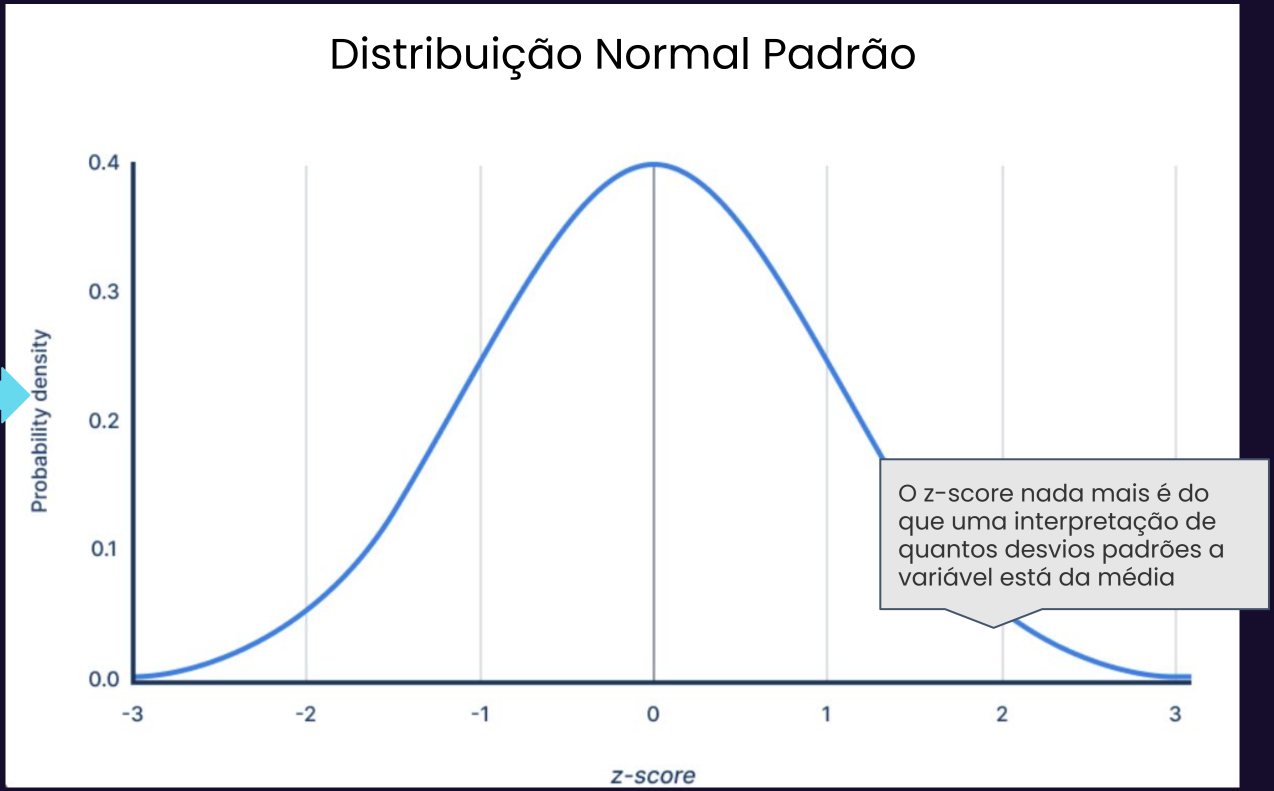
Uma distribuição de probabilidade Normal resulta de um experimento que satisfaz os seguintes requisitos:

1. Média 0 e desvio padrão 1
2. O seu eixo x pode ser chamado de Z, e cada valor de z pertencente a Z pode ser interpretado como a quantos desvios padrões z está da média.

Para padronizar uma distribuição normal podemos aplicar a fórmula:

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

X = variável aleatória com dist. Normal
u = média populacional
σ = desvio padrão populacional



Teorema do limite central

Muitas vezes no nosso dia a dia, temos acesso somente a uma distribuição amostral e não populacional.

Por exemplo: Suponha que você trabalha no departamento de marketing de uma empresa. E você deseja entender a distribuição de vendas de um determinado produto para publicos de 15 a 30 anos versus de 30 a 60 anos. Agora suponha que a sua área não tem acesso a todos os dados de vendas, mas somente a uma amostra de 100 mil clientes. Como saber com como essas distribuições se comparam?

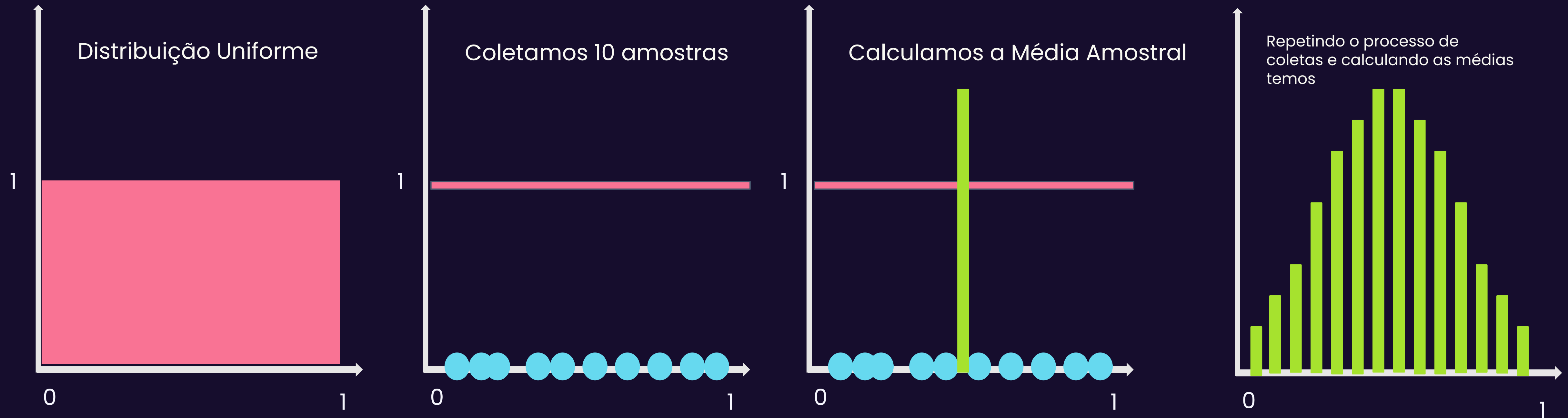
O Teorema do limite central nos explicará que se a distribuição da população de origem for desconhecida, ao retirarmos amostras suficientemente grandes (acima de 30 elementos) a distribuição amostral das médias dos dados se aproxima de uma normal.

Mas ...

1. O que é a distribuição amostral das médias dos dados?
2. Qual a intuição desse teorema?

Teorema do limite central

Entendendo na prática: Suponha que temos um distribuição uniforme e coletamos uma amostra dos dado se repetirmos esse processo independente da distribuição original, a distribuição das médias amostrais segue uma distribuição Normal



Implicaçoes: Se não sabemos a distribuição original na prática sabemos que se coletarmos uma amostra suficientemente grande a **distribuição das médias de vendas será Normal**. E utilizaremos a distribuição das médias para construir intervalos de confiança e fazer testes sobre a população

Intervalo de Confiança

Intervalos de Confiança

1. Suponha que anotamos peso de 20 pacientes em um laboratório. (temos uma amostra com N = 20).
2. Em seguida calculamos a média dessa amostra = 63kg;
3. Em seguida vamos fazer um **processo chamado de bootstrap**:
   - vamos selecionar uma amostra aleatória com repetição com 20 dados e calcular a sua média = 69kg (Img 2)
   - vamos repetir esse processo 10000 vezes.

4. Como sabemos que a distribuição das médias pelo TCL é normal e portanto simétrica; intervalo de confiança de x% conterá x% de todas as médias de uma distribuição.

Intervalos de Confiança

Entendemos que o intervalo de confiança por exemplo de 95%. nos diz que com 95% de confiança o valor populacional , ex: média do peso, estará em um determinado intervalo.

Mas como podemos calcular os valores mínimos e máximos desse intervalo?

Fórmula do IC

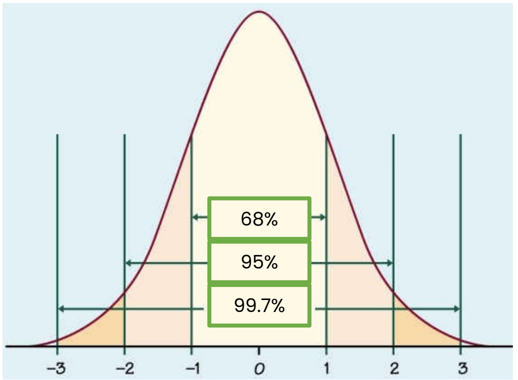
$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

x: média da amostra.
Z: valor crítico da distribuição normal padrão correspondente ao nível de confiança desejado
s: é o desvio padrão da amostra.
n: é o tamanho da amostra.

Valor Crítico Z

Vimos que a distribuição Normal Padrão (média é 0 e desvio padrão é 1) é chamada de distribuição Z. E é dessa distribuição que vemos encontrar os valores críticos z.

O valor crítico z nos indicará quantos desvios padrões da média desejamos o nosso nível de confiança.



Ex:

Escolhendo 95% de confianca, vamos encontrar valores críticos de z aproximadamente entre (-2 e 2)

Podemos nos referir a z como z(a/2), sendo a o nível de significância a área abaixo da curva colorida ao lado.
Para 95% de confianca a = 5% ; para 99%, a = 1%

Amostragem

Conceitos de Amostragem

- Em geral no processo de coleta de dados, não se tem acesso a toda a população do interesse de estudo. Com isso recorremos a uma amostra dessa população para a análise dos dados.

- Amostragem é uma técnica/processo da área de estatística em que estudamos formas de selecionar subconjuntos da população para fazermos posteriormente inferências estatísticas sobre a população de interesse.

- Erro Amostral: É a diferença entre o resultado amostral e o populacional. Ex: média da altura de uma amostra é 1.65m na população do estudo 1.74m

- Amostragem vs Censo:
  O censo é o estudo de TODOS os elementos da população enquanto a amostra é um subconjunto o censo visa minimizar ao máximo o erro amostral.

Métodos de Amostragem

Na Amostragem Probabilística os critérios de seleção da amostra são definidos a partir da probabilidade de modo que sabemos a probabilidade de cada indivíduo na amostra.

Amostragem Probabilística

Amostragem Não Probabilística

Amostra Aleatória Simples

A seleção dos indivíduos se dá por um sorteio.

Todos os indivíduos têm a mesma probabilidade de seleção

Amostra Aleatória Estratificada

Seleção de proporções de diferentes "estratos" ou subgrupos diferentes .
Ex: 50% homens e 50% mulheres

Amostra Sistemática

População ordenada e se seleciona elementos na ordem.
Ex: Fila do banco selecionar a cada 5 pessoas.

Definindo o Tamanho da Amostra e a Margem de erro.

Quando selecionamos uma amostra buscamos que ela seja o mais representativa possível da população, ou seja procuramos um erro amostral pequeno.

A partir do erro amostral podemos calcular o erro padrão, $e = \sigma/\sqrt{n}$. Sendo n o tamanho da amostra.

Sendo assim, para definirmos o tamanho da amostra precisaremos definir a Margem de Erro desejada.

Quanto Maior o Tamanho da Amostra

Erro Amostral

Mas o que é margem de erro?

A margem de erro é uma porcentagem que mede a proximidade dos resultados obtidos da amostra do valor real para a população total do estudo. Ela é uma métrica de precisão do seu estudo.

Exemplo: Margem de erro de uma pesquisa eleitoral é de 2 p.p.. Isso significa que, se 60% dos entrevistados disseram que irão votar no candidato A, você deve considerar que a porcentagem real de votos fica entre 58% e 62%.

$$\text{Margem de Erro} = z(a/2) * \frac{\sigma}{\sqrt{n}}$$

Z-score (a/2)	1.96 (apx 2) para a = 5%
σ	desvio padrão populacional
n	tamanho da amostra

Mas e se não sabemos o desvio padrão populacional?

1. Podemos calcular a ***margem de erro amostral da Proporção populacional.***

   Supondo por exemplo para pesquisas em que gostariamos de saber uma resposta binária (Sim ou não).

Ex: pesquisa realizada com 100 pessoas em que 45 disseram votariam no candidato A. A proporção = 45%.
Qual a margem de erro com 95% de confiança?

$$ME = \frac{z(a/2) * \sqrt{(p*(1-p))}}{\sqrt{n}} = \frac{1.96*\sqrt{(0.45*(0.55))}}{\sqrt{100}} = 9{,}75\%$$

Para 95% de confiança, a = 5% e assim, a/2 = 2.5% ; Z(2.5%) como vimos anteriormente na normal padrão é 1.96

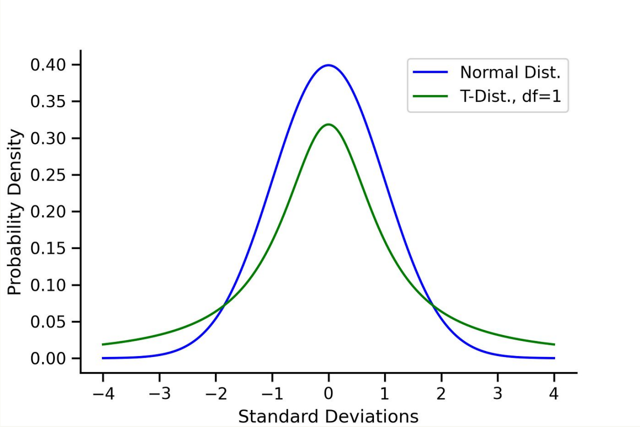
Mas e se não sabemos o desvio padrão populacional e temos amostra pequena?

2. Podemos calcular a margem de erro amostral utilizando uma distribuição diferente

No caso de não só sabermos o desvio padrão da amostra e não o da população podemos usar a distribuição t. A distribuição t é muito semelhante a Normal padrão sua principal diferença é: ela é obtida a partir da padronização utilizando valores da amostra a partir da seguinte fórmula:

$$T = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

onde:
u = média populacional
s = desvio padrão da amostra
n = tamanho da amostra
X = média amostral



Cálculo da margem de erro amostral:

$$\text{Margem de Erro} = \frac{t(a/2) * s}{\sqrt{n}}$$

onde:
s = desvio padrão da amostra
n = tamanho da amostra
t = valor crítico na distribuição t para o nível de confiança desejado
a = nível de significancia

E Se não sabemos o desvio padrão populacional e temos amostra grande?

Nesse caso podemos estimar o desvio padrão populacional a partir do desvio padrão amostral, ao contrário do caso em que temos amostra pequena

Estimação de sigma:

1. Primeira Opção : $\sigma = \frac{(max - min)}{4}$

2. Segunda Opção : Utilizar o desvio padrao S no lugar de sigma

onde:
max e min se referem aos valores máximos e mínimos

Cálculo da margem de erro amostral:

$$\text{Margem de Erro} = \frac{z(a/2) * \sigma'}{\sqrt{n}}$$

onde:
σ' = desvio padrão estimado
n = tamanho da amostra
t = valor crítico na distribuição t para o nível de confiança desejado
a = nível de significancia

Resumindo:

- MARGEM DE ERRO
  - Margem de erro para média
    Ex: margem de erro da média da altura
    - Sabendo desvio padrão populacional
      $z(a/2) * \frac{\sigma}{\sqrt{n}}$
    - Desvio padrão populacional desconhecido e amostra > 30
      $\frac{z(a/2) * \sigma'}{\sqrt{n}}$
    - Desvio padrão populacional desconhecido e amostra < 30*
      $\frac{t(a/2) * s}{\sqrt{n}}$
  - Margem de erro para proporção
    Ex: margem de erro da pesquisa eleitoral
    $z(a/2) * \frac{\sqrt{(p*(1-p))}}{\sqrt{n}}$

Como determinar o tamanho da Amostra?

Ao inves de usar as fórmulas anteriores para calcular a margem de erro também podemos utilizá-la para calcular qual o tamanho de amostra N necessário para uma determinada margem de erro.

Por exemplo: Supondo por exemplo para pesquisas em que gostariamos de saber uma resposta binária (Sim ou não). Ex: Qual a quantidade de pessoas necessárias em que 45 disseram votariam no candidato A. A proporção seria de 45% e a margem de erro de 5%?

Vamos Praticar!

Teste A/B

É muito comum no dia a dia de analista de dados o uso de testes A/B.

Por exemplo:

- Equipes de Design e experiência do usuário (UX) lançam novos botões de adicionar ao carrinho no site e analisam o impacto do novo botão nas vendas dos produtos finais.

- Equipes de marketing analisam o impacto da campanha A vs campanha B nas vendas.

- Cientistas de dados comparam efeitos de um modelo A vs modelo B de previsão

Teste A/B

Nesses casos em que queremos analisar o impacto de uma medida ou então comparar uma nova funcionalidade é muito comum estruturarmos o que chamamos de um teste A/B para decidirmos qual a melhor medida.

Basicamente a finalidade desse teste é, por exemplo: Comparar as vendas (a conversao do cliente) com dois cenários:
- Cenário A : com o botão atual (como funciona hoje) também chamado de grupo controle
- Cenário B : com o novo botão (nova ferramenta) também chamado de grupo teste ou tratamento

Podemos realizar essas comparações com base na média. Ex: comparar a conversão média do grupo A vs grupo B ou comparar proporções, Ex: queremos verificar se no teste A/B se clicar no botão aumentou a proporção de homens ou mulheres.

Teste A/B

Para determinar o tamanho da amostra em um teste A/B de diferentes médias, desse modo vamos utilizar a seguinte fórmula.

Valor crítico do nível de significância alpha

Valor crítico do poder do teste

Tamanho da amostra por grupo

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 \sigma^2}{\Delta^2}$$

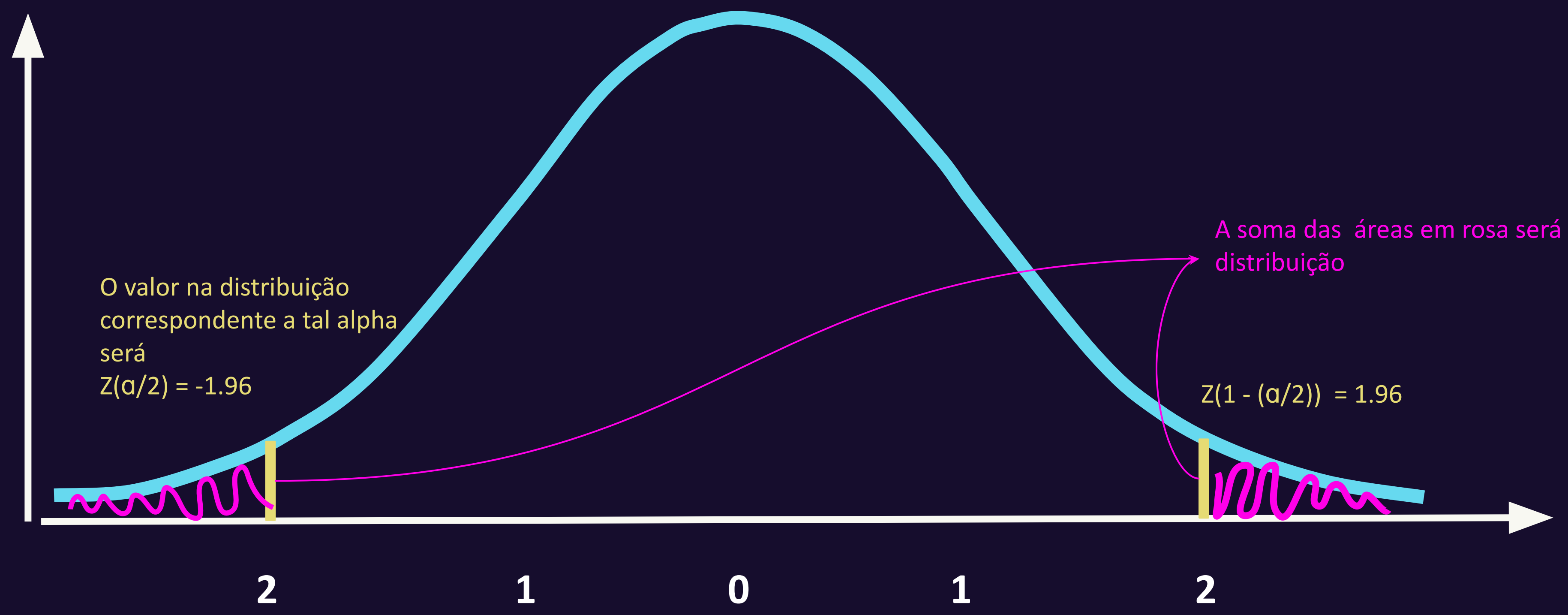
Variância

Delta (Diferença Absoluta)

Teste A/B

Revisando os conceitos:

1. Nível de significancia (α):
   - Nos dá a Probabilidade de um evento ocorrer de modo aleatório.
   - Ou seja se selecionarmos 5%, significa que. queremos que com 95% de certeza o grupo A difere do B e 5% de chance essa diferença é aleatória.
   - Encontramos na Distribuição normal padrão os valores críticos Z para alpha. Para 5% eles serão :-1.96 e 1.96



Distribuição Normal Padrão:

Teste A/B

2. Poder do teste (1 - β):
   - β nos dá a probabilidade de o teste não encontrar diferença nas campanhas mesmo que ela exista
   - 1- β Nos dá a probabilidade de o meu teste A/B encontrar de fato uma diferença entre teste e controle, dado que elas são diferentes.
   - Selecionado de 0 a 100%, geralmente selecionamos β = 20%

3. Delta (Δ):

   - Será a diferença entre médias ou proporções das métricas dos dois grupos.
   - Exemplo:
     - Grupo de controle A (utilizando o botão atual) tem taxa de conversão de 15%
     - Suponha que queremos um aumento de 20% de conversão para o grupo B (tratamento). CTR2 = 0.15*(1+0.2) = 0.18
     - Delta = 0.18 - 0.15

4. Variância:

   - Vamos calcular a variância amostral do grupo controle (atual) e chamá-la de S1
   - A variância estimada para colocarmos na fórmula será: 2* a variancia do grupo controle (valor obtido por aproximação quando fazemos testes com duas amostras)

Teste A/B

EXEMPLO:

Suponha que queremos comparar o impacto de uma mudança de botão na conversão dos usuários de um e-commerce utilizando um teste A/B. Para isso vamos utilizar como métrica o CTR (click through rate) das pessoas que estiveram na página quantos % apertaram no botão de adicionar ao carrinho.

Suponha que desejamos com o novo botão obter um aumento de conversão de 10%

Queremos ter 95% de certeza que o efeito na conversão não foi aleatório (alpha = 5%)
Queremos também com 80% de certeza conseguir capturar o efeito do novo botão.

Com o botão atual temos uma média de CTR de 10 e variância de 20.

Determine o tamanho da amostra necessária para o teste A/B

Teste A/B

<u>SOLUÇÃO:</u>

1. Nível de significância (α) = 5% ; Assim olhando na distribuição normal padrão $Z(1-\alpha/2) = 1.96$
2. Poder do teste ($1 - \beta$) = 80% ; Assim olhando na distribuição normal padrão $Z(1-\beta) = 0.84$ vamos aproximar por 0.8
3. Variância (σ^2) = 2*(20) = 40
4. Delta (Δ) =

    CTR baseline = 10

    CTR desejado = 10*(1.1) = 11

    Δ = (11- 10)^2 = 1

Substituindo na fórmula original:

Valor crítico do nível de significância alpha

Valor crítico do poder do teste

Tamanho da amostra por grupo

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 \sigma^2}{\Delta^2}$$

Variância

Delta (Diferença Absoluta)

$$n = \frac{((1.96 + 0.8)\text{^}2)*40}{1} \qquad n = 314$$

Como n é o tamanho da amostra por grupo, vamos precisar de 2*n = 628 dados