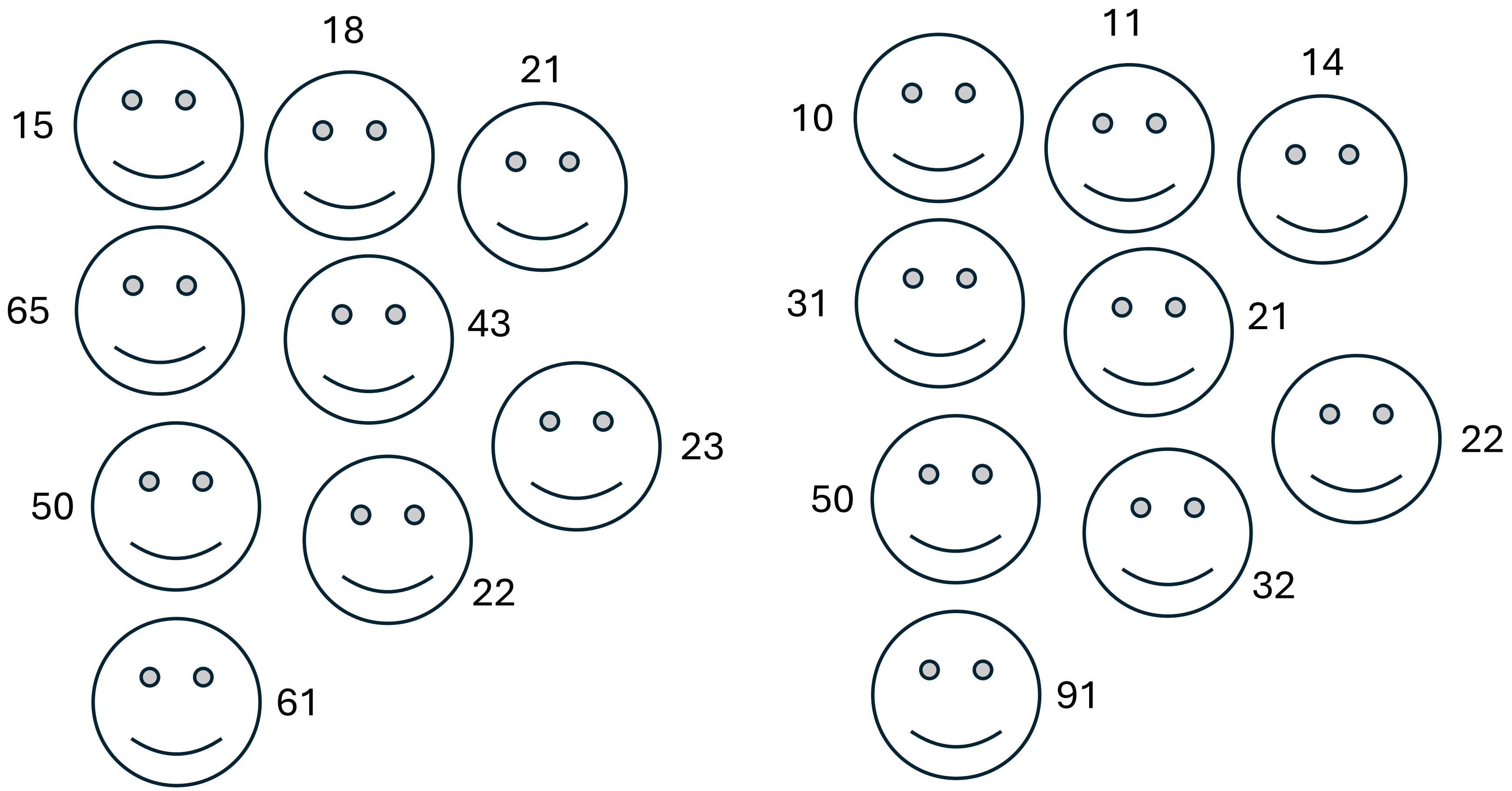
Final CLT e teste de proporções

População



Média : 33,33

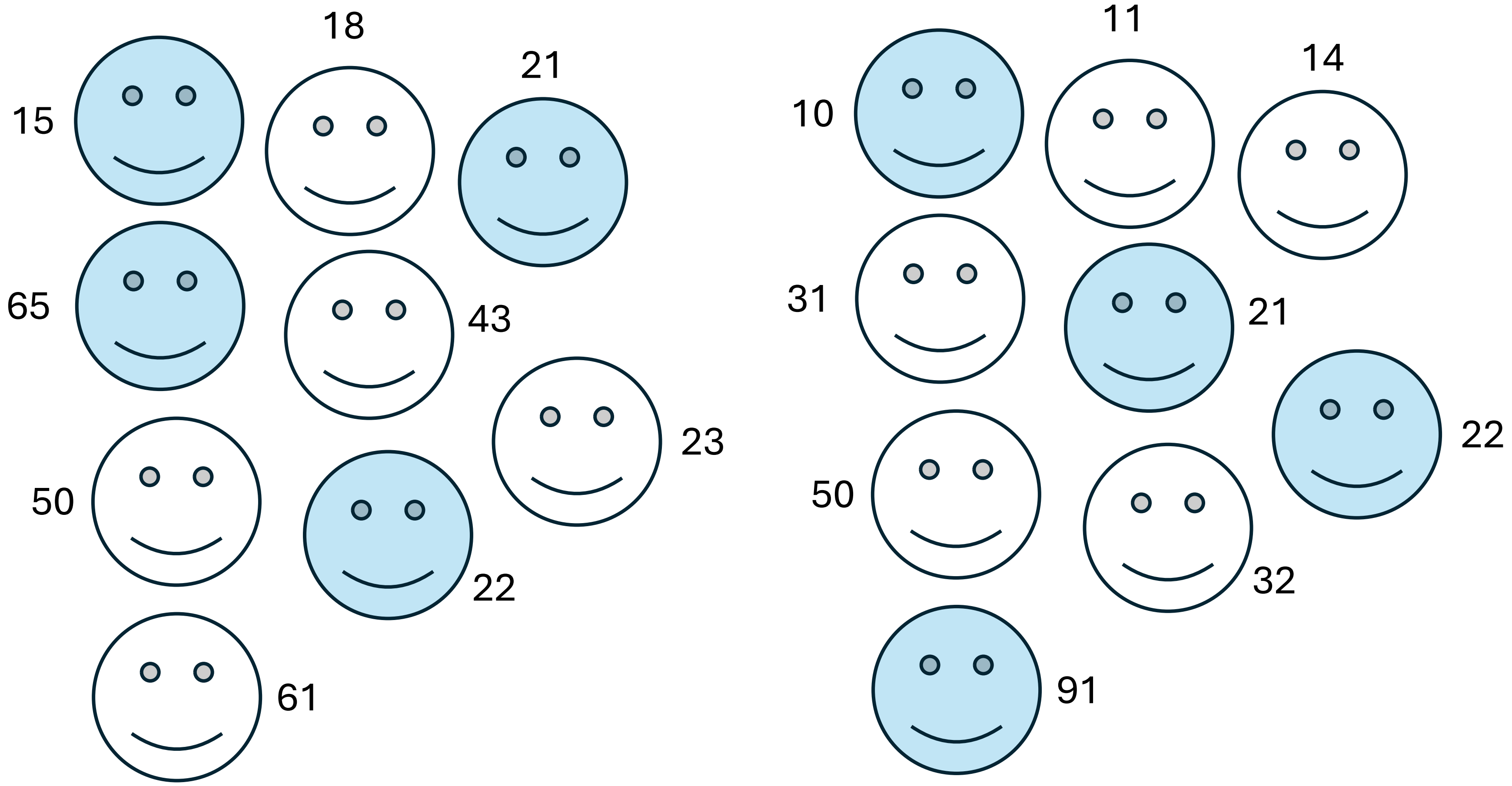
Máximo: 91

Mínimo: 10

Populacional

Desvio Padrão:
21

Uma amostra possível (tamanho amostral n = 8)



Média: 33,3375

Máximo: 91

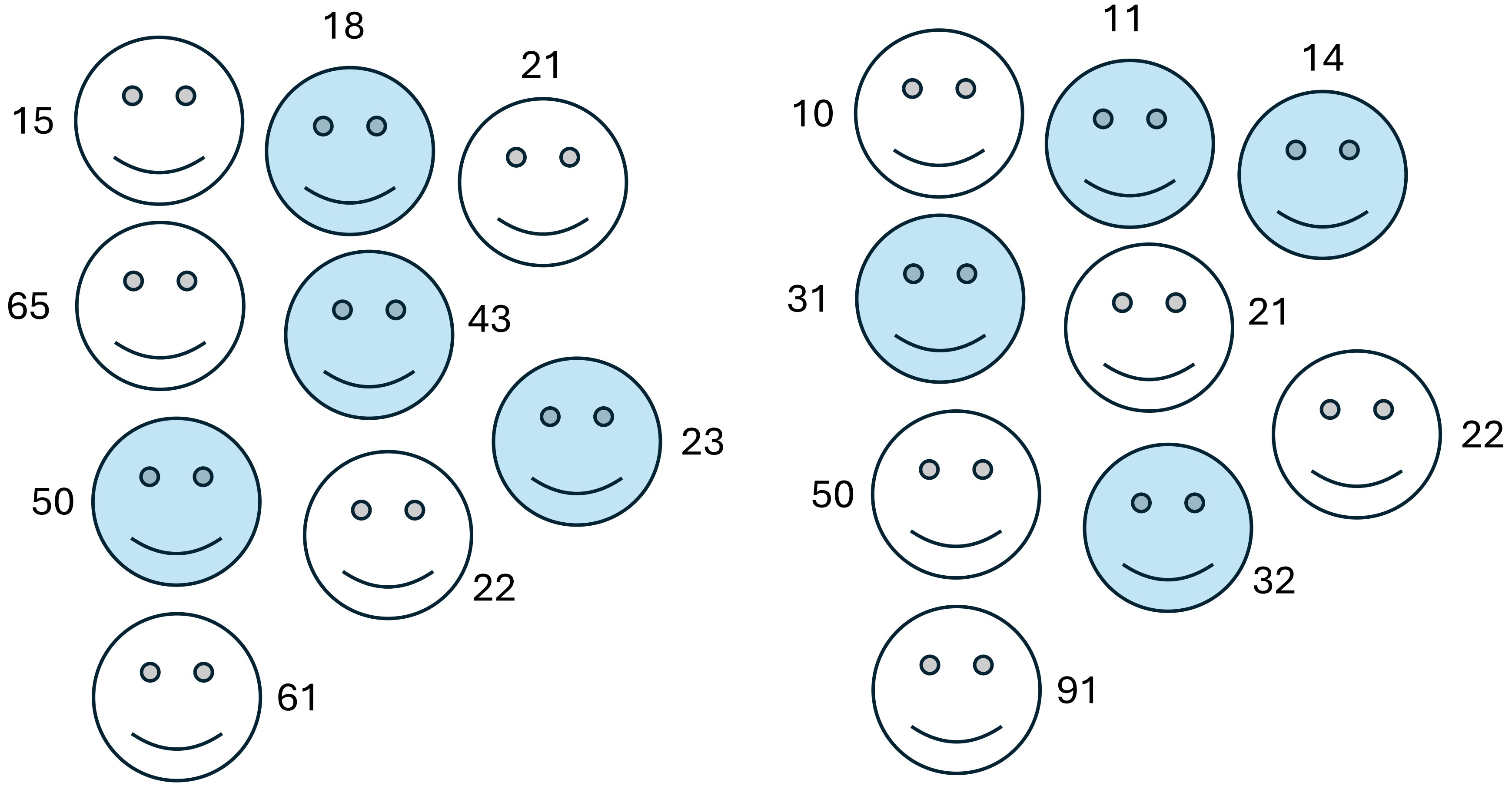
Amostral

Mínimo: 10

Desvio Padrão: 26,88

Amostra 1

Outra amostra possível (tamanho amostral n = 8)



Média: 27,75

Máximo: 50

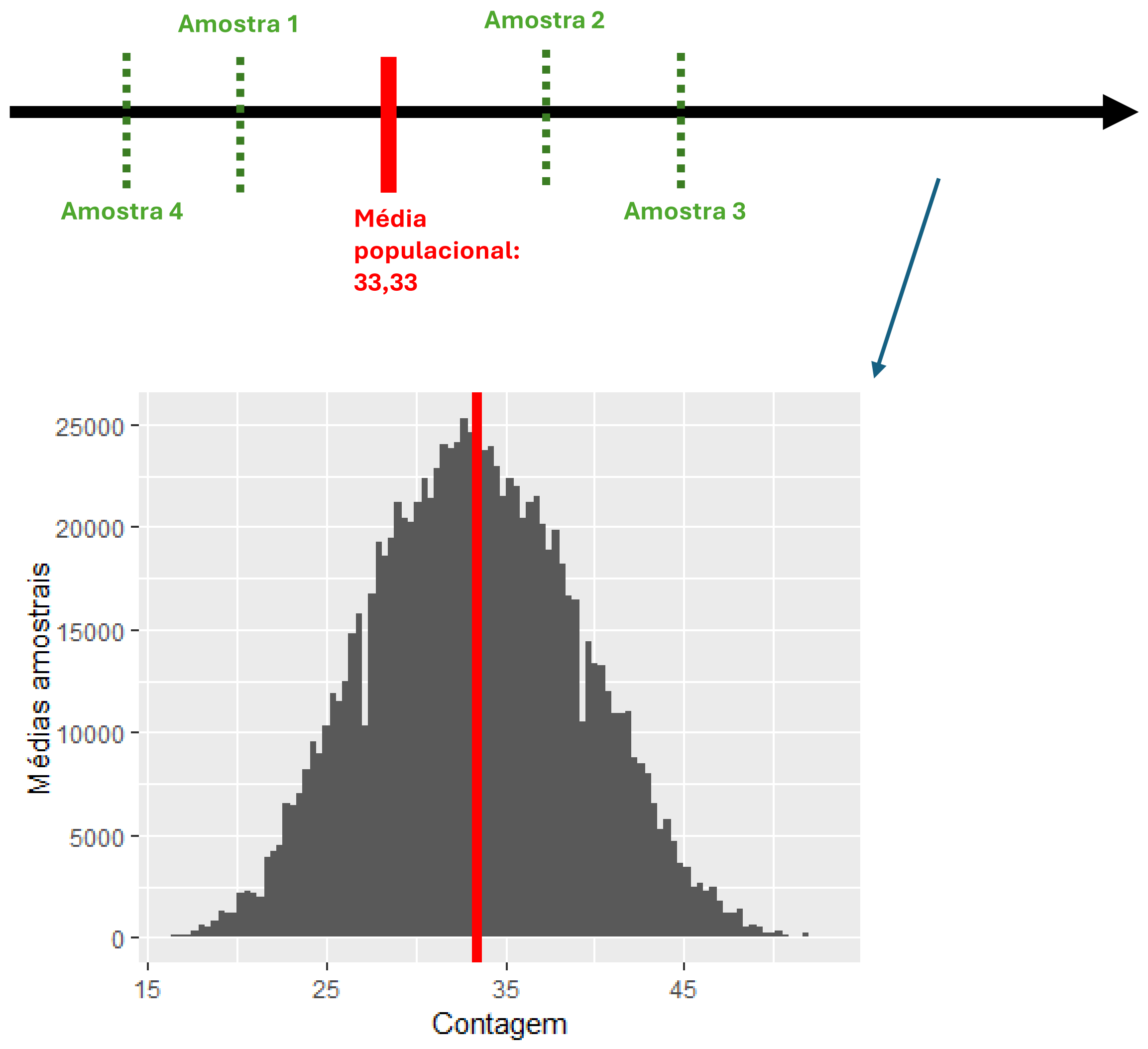
Amostral

Mínimo: 11

Desvio Padrão: 12,95

Amostra 2

Histograma **da média amostral de várias amostras**



O Teorema Central do Limite nos diz que:

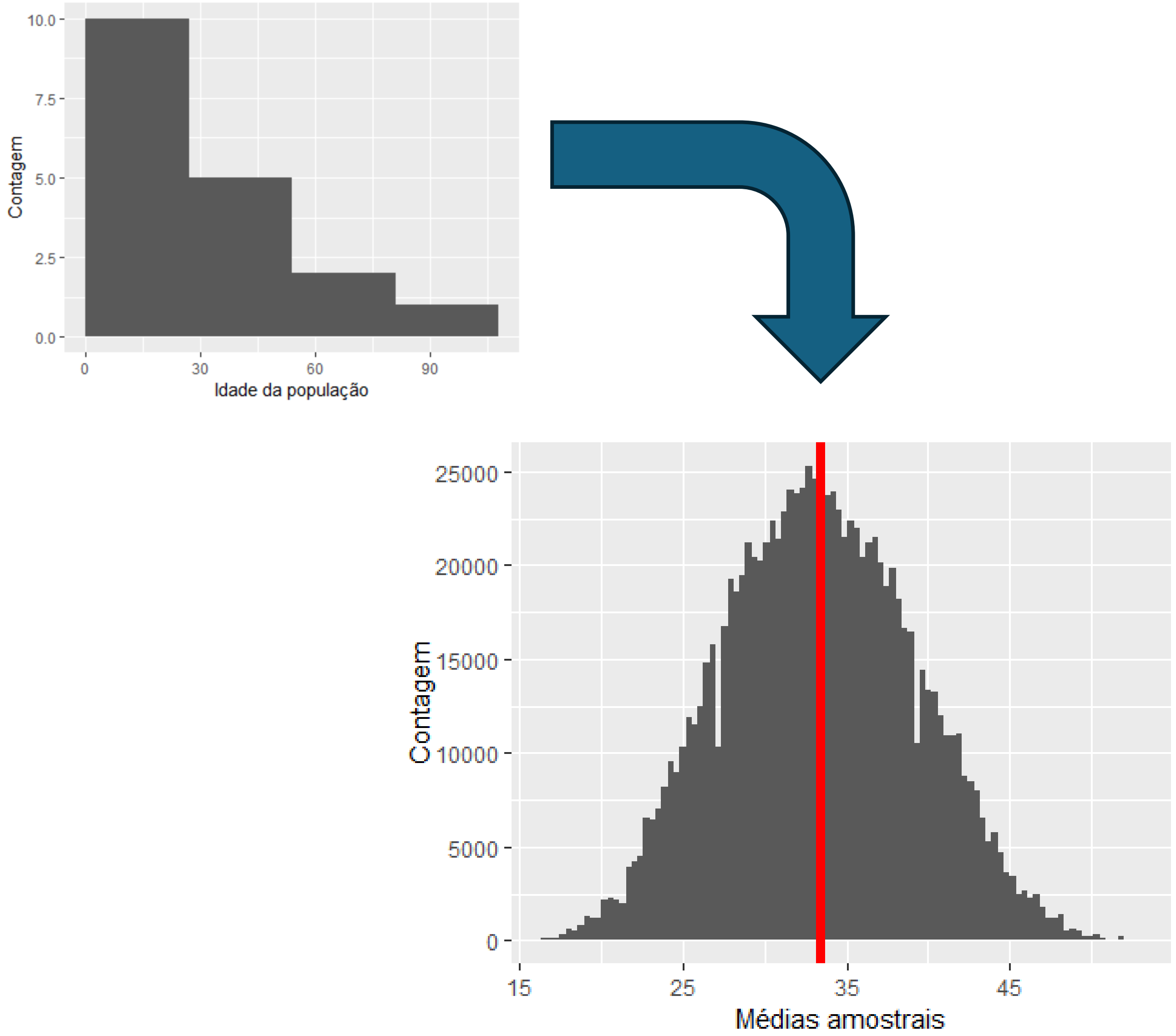
(se os elementos da amostra forem escolhidos aleatoriamente dentro da população)

1. A média amostral vai ficar perto da média populacional

2. A **probabilidade** de observarmos diferenças com relação à média populacional pode ser calculada usando a curva normal

3. Quão maior for o tamanho amostral, menor a diferença entre a média amostral e a média populacional

Histograma **da média amostral de várias amostras**



O Teorema Central do Limite nos diz que:

(se os elementos da amostra forem escolhidos aleatoriamente dentro da população)
Não importa a distribuição da população!

1. A média amostral vai ficar perto da média populacional

2. A **probabilidade** de observarmos diferenças com relação à média populacional pode ser calculada usando a curva normal

3. Quão maior for o tamanho amostral, menor a diferença entre a média amostral e a média populacional

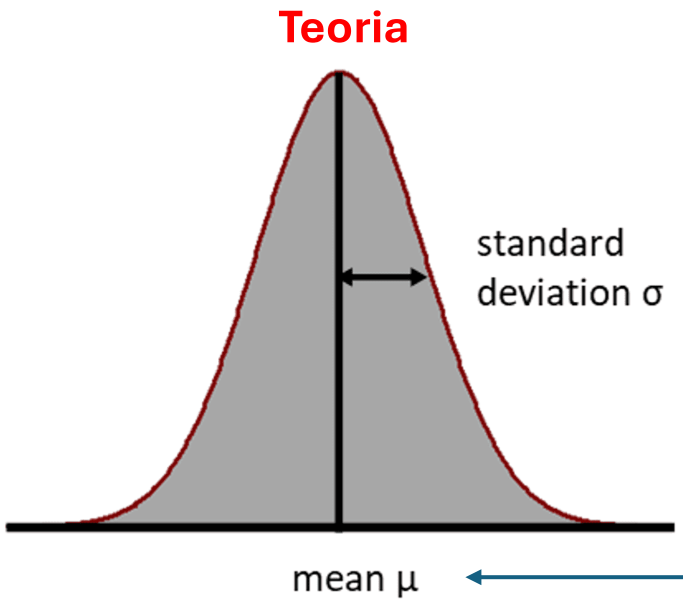
1. A média amostral vai ficar perto da média populacional

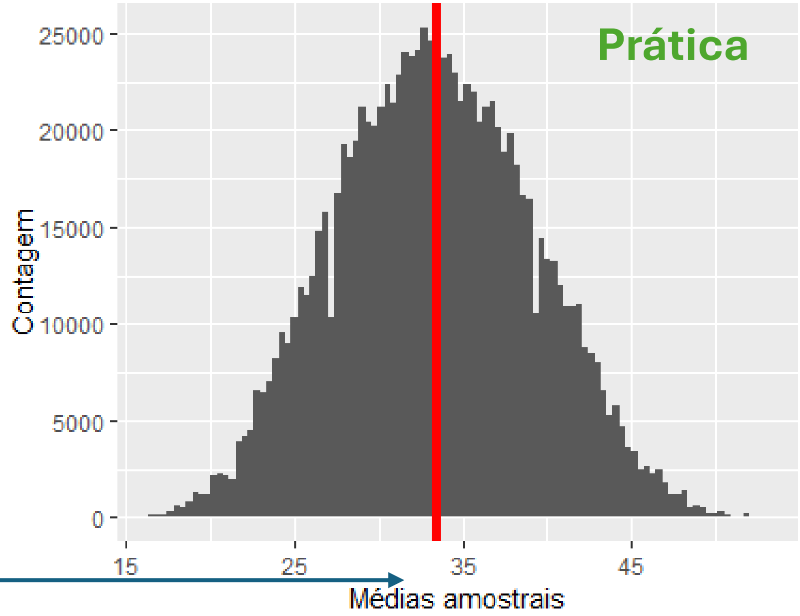
Quando o teorema diz “a média amostral vai ficar perto da média populacional” matematicamente o que está sendo dito é que NÃO IMPORTA QUAL FOR A MÉDIA POPULACIONAL

Vamos representa-la por uma incógnita μ

A distribuição amostral de $\bar{X}$ é uma normal com média μ

O desvio padrão σ vem depois, o primeiro fato é que o pico da curva vai ser na média populacional!





2. A **probabilidade** de observarmos diferenças com relação à média populacional pode ser calculada usando a curva normal

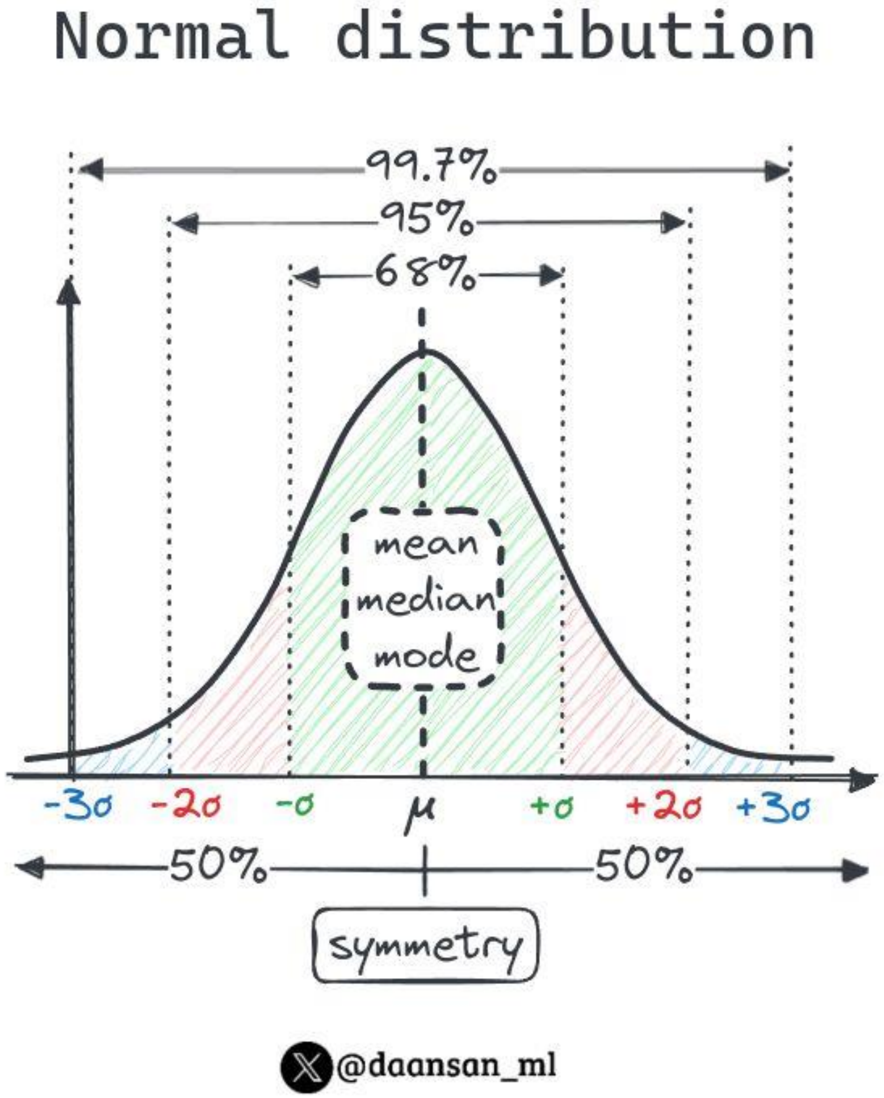
A curva normal é uma curva parametrizada por dois números:

O pico da curva μ, que já sabemos que no caso da distribuição de $\bar{X}$ é a média populacional

O desvio padrão σ, que resta definir no nosso caso

O segundo ponto do teorema nos diz que, para algum desvio padrão que vamos definir no próximo slide, podemos calcular probabilidades sobre $\bar{X}$ a partir das probabilidades conhecidas da curva normal

Isso quer dizer, por exemplo, que é **muito improvável** que um $\bar{X}$ caia a mais do que 3 σ pra mais ou pra menos da média populacional (menos de 99,7% de chance)



3. Quão maior for o tamanho amostral, menor a diferença entre a média amostral e a média populacional

O terceiro ponto do teorema central do limite é dizer quanto vale o σ da normal que dá a distribuição amostral de $\bar{X}$:

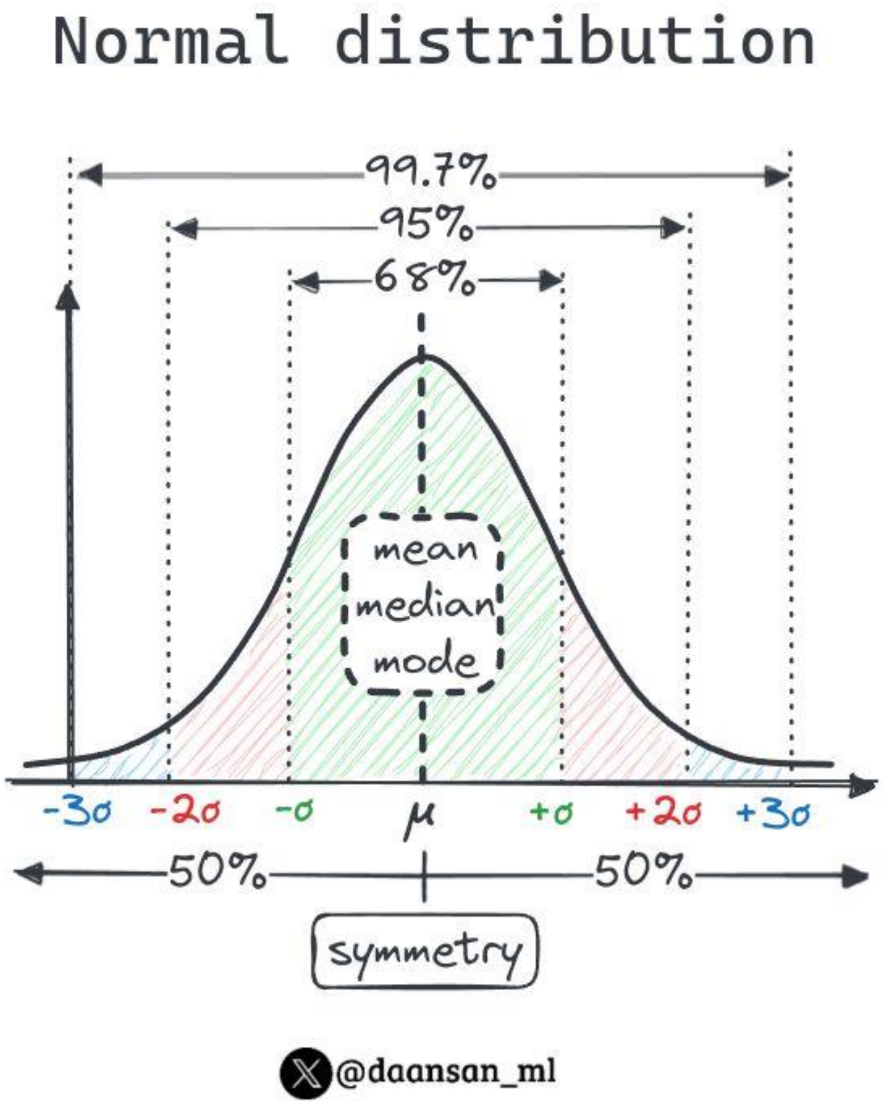
(Caso particular do Teorema Central do Limite para amostragem)

Se os elementos de uma população são selecionados aleatoriamente para formar uma amostra de tamanho n $X_1, X_2, \ldots, X_n$ então a média amostral $\bar{X}$ segue aproximadamente uma distribuição normal com parâmetros:

$$\mu = Média\ Populacional$$

$$\sigma = \frac{Desvio\ padrão\ populacional}{Raiz\ de\ n}$$

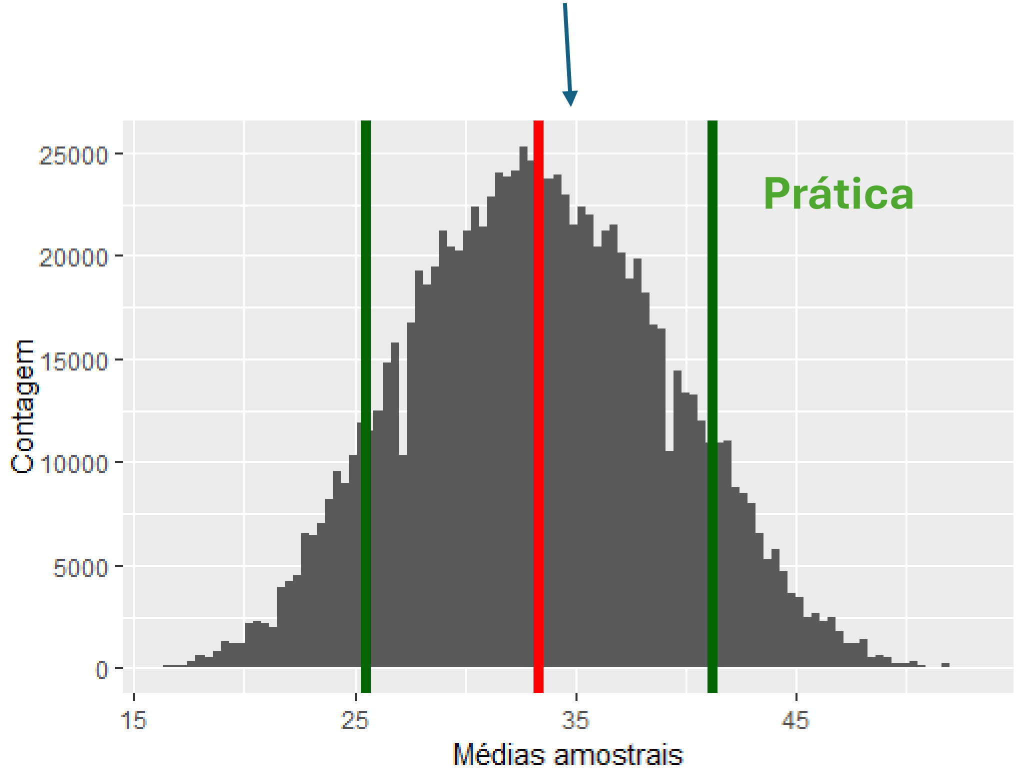
$$= \frac{Desvio\ padrão\ populacional}{\sqrt{n}} =$$

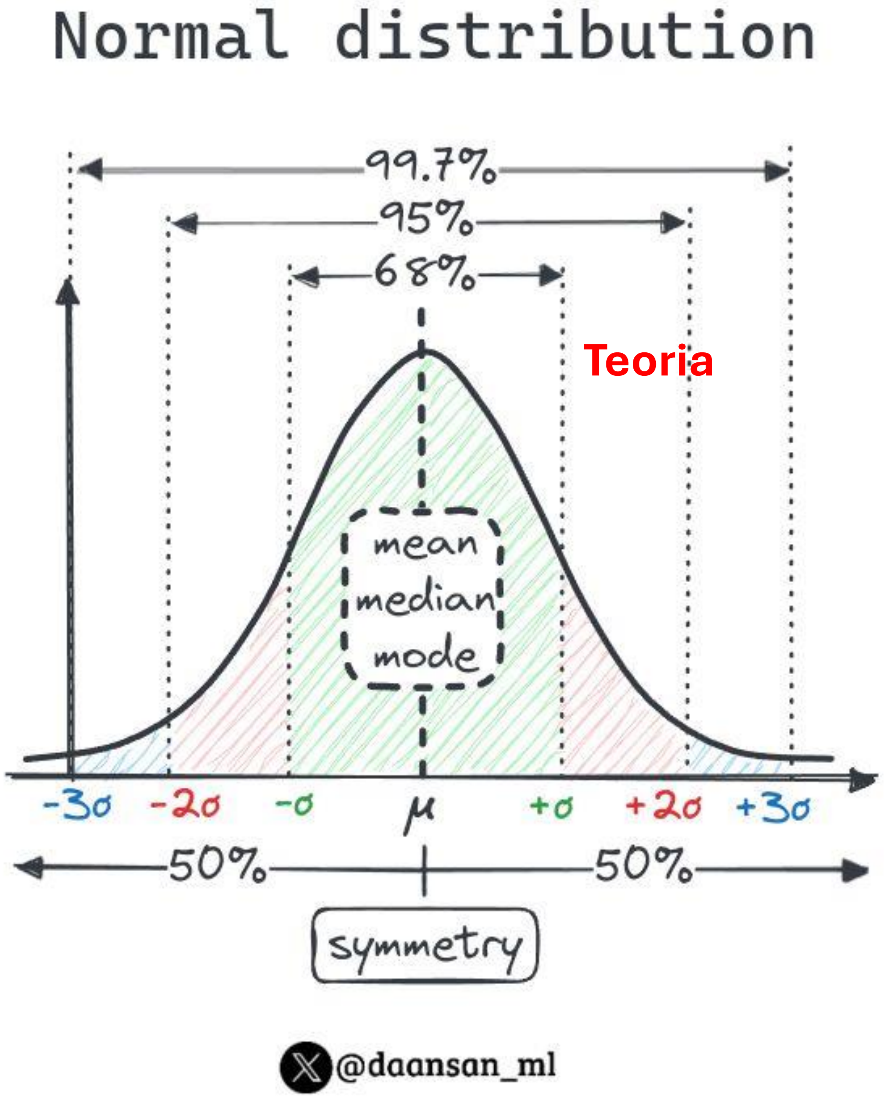


3. Quão maior for o tamanho amostral, menor a diferença entre a média amostral e a média populacional

$$\mu = Média\ Populacional = 33{,}33$$

$$\sigma = \frac{Desvio\ padrão\ populacional}{\sqrt{n}} = \frac{22{,}23}{2{,}82} = 7{,}86$$





3. Quão maior for o tamanho amostral, menor a diferença entre a média amostral e a média populacional

A dispersão é sempre ao redor do mesmo número
a média populacional

O valor da dispersão, por outro lado, vai caindo porque estamos dividindo por raiz de n ($\sqrt{n}$)

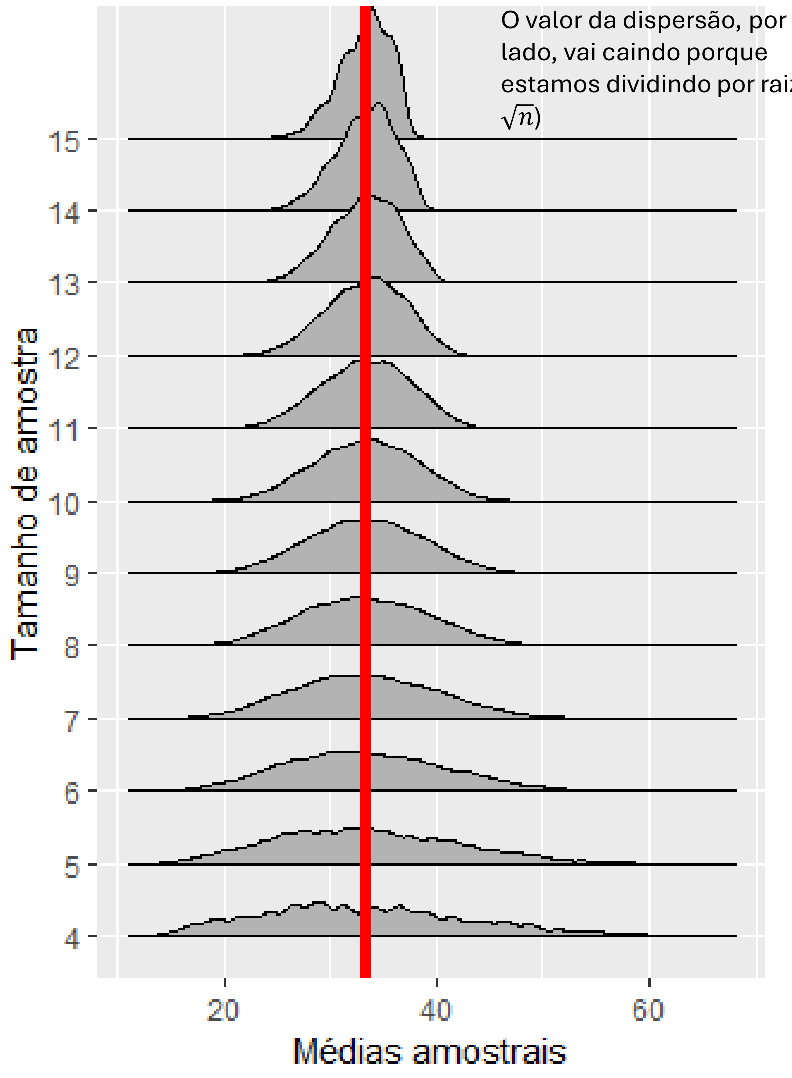
Caso particular com n = 8:

$$\mu = Média\ Populacional = 33{,}33$$

$$\sigma = \frac{Desvio\ padrão\ populacional}{\sqrt{n}} = \frac{22{,}23}{2{,}82} = 7{,}86$$

Em cima, o desvio padrão populacional, é fixo e desconhecido

Dividi-lo por $\sqrt{n}$ quer dizer que, se aumentarmos suficientemente a amostra, o "espalhamento" da distribuição amostral de $\bar{X}$ fica tão pequeno quanto possível. Nós mandamos no n



O que é probabilidade?

O Teorema Central do Limite fala sobre probabilidades

O que é probabilidade?

Tem muitas definições e existe muita discussão filosófica e científica sobre o que é probabilidade

Para compreender o fundamento das ferramentas estatísticas desse curso probabilidade para nós será:

A probabilidade de uma coisa é a percentagem do tempo em que é esperado que ela aconteça, quando ela pode ser repetida várias e várias vezes de maneira independente e sob as mesmas condições

O conceito de probabilidade tem a ver com **repetições**, nas **mesmas condições**, de **maneira independente**

Vamos esmiuçar os conceitos, mas o exemplo principal dessas repetições são **experimentos científicos**

O que é probabilidade?

A probabilidade de uma coisa é a percentagem do tempo em que é esperado que ela aconteça, quando ela pode ser repetida várias e várias vezes de maneira independente e sob as mesmas condições

Vamos pensar em um experimento científico da área média: oferecer um remédio e verificar se diminui sintomas/cura doenças etc:

Mesmas condições = Exatamente o mesmo remédio é administrado para pacientes praticamente idênticos, de mesma idade, mesmo histórico etc

Repetida várias e várias vezes = tamanho amostral grande, conceitualmente a probabilidade deveria descrever perfeitamente a realidade de tivéssemos **infinitas realizações** (nunca vamos ter)

De maneira independente = “independência” tem dois significados. O significado coloquial quer dizer “sem relação alguma”, “um experimento não impacta no outro”, “os pacientes nem os médicos nem aplicadores se conhecem” etc. No sentido estatístico essa noção é formalizada como “a probabilidade da intersecção é o produto das probabilidades”

Para nós, importa mais o significado coloquial

Consequências do Teorema Central do Limite

O Teorema Central do Limite tem muitas consequências práticas

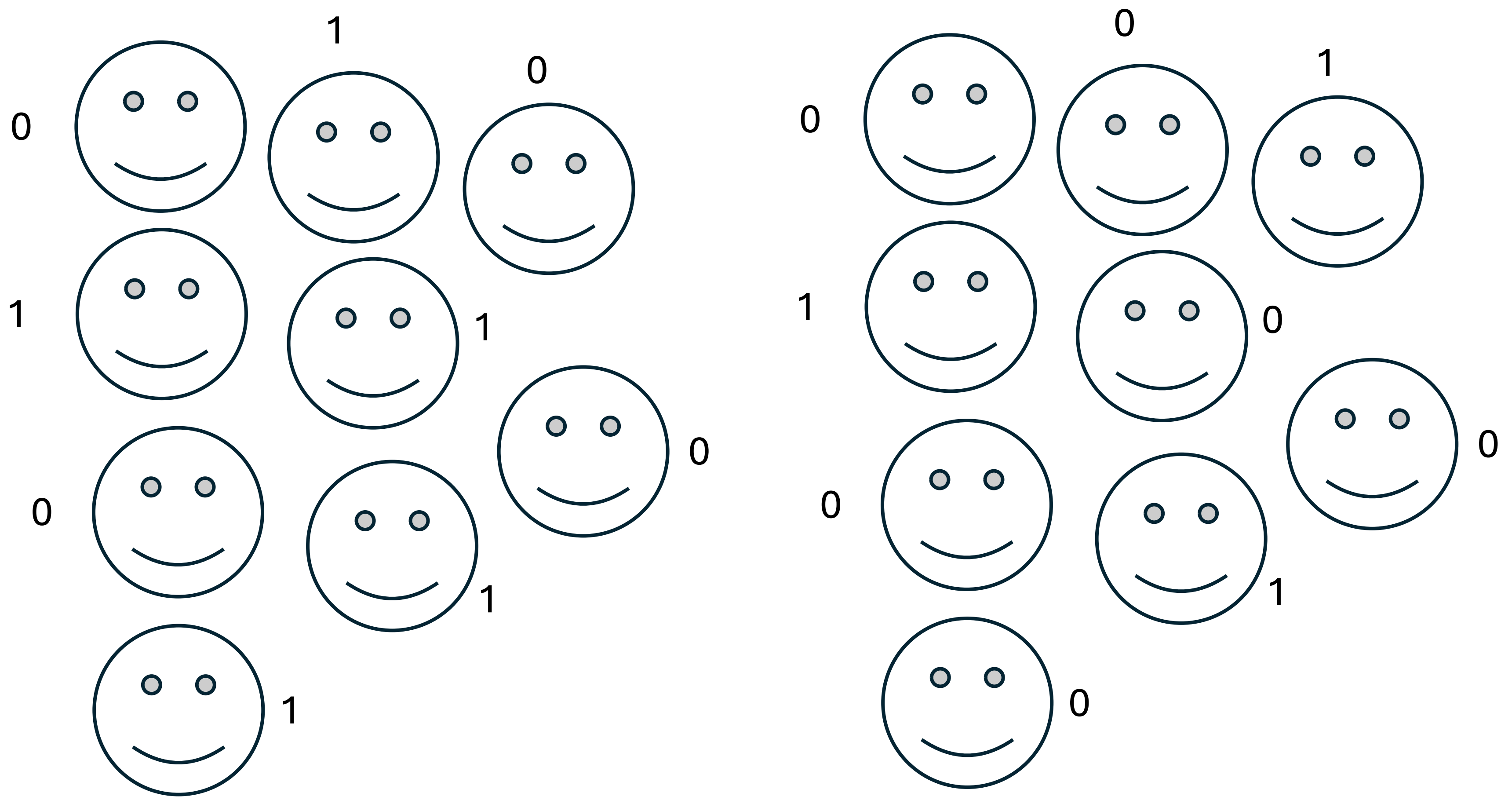
Consequência número 1:

Podemos testar se um determinado valor de interesse é compatível com a media amostral observada

Isso é exatamente o que fizemos para decidir se o dado era honesto:

1. Observamos amostra
2. Escolhemos uma estatística
3. Encontramos a distribuição amostral dessa estatística no cenário que nos interessa
4. Posicionamos o valor que encontramos dentro dessa escala

Valores populacionais formados apenas por 0 e 1



Média : 42,10%

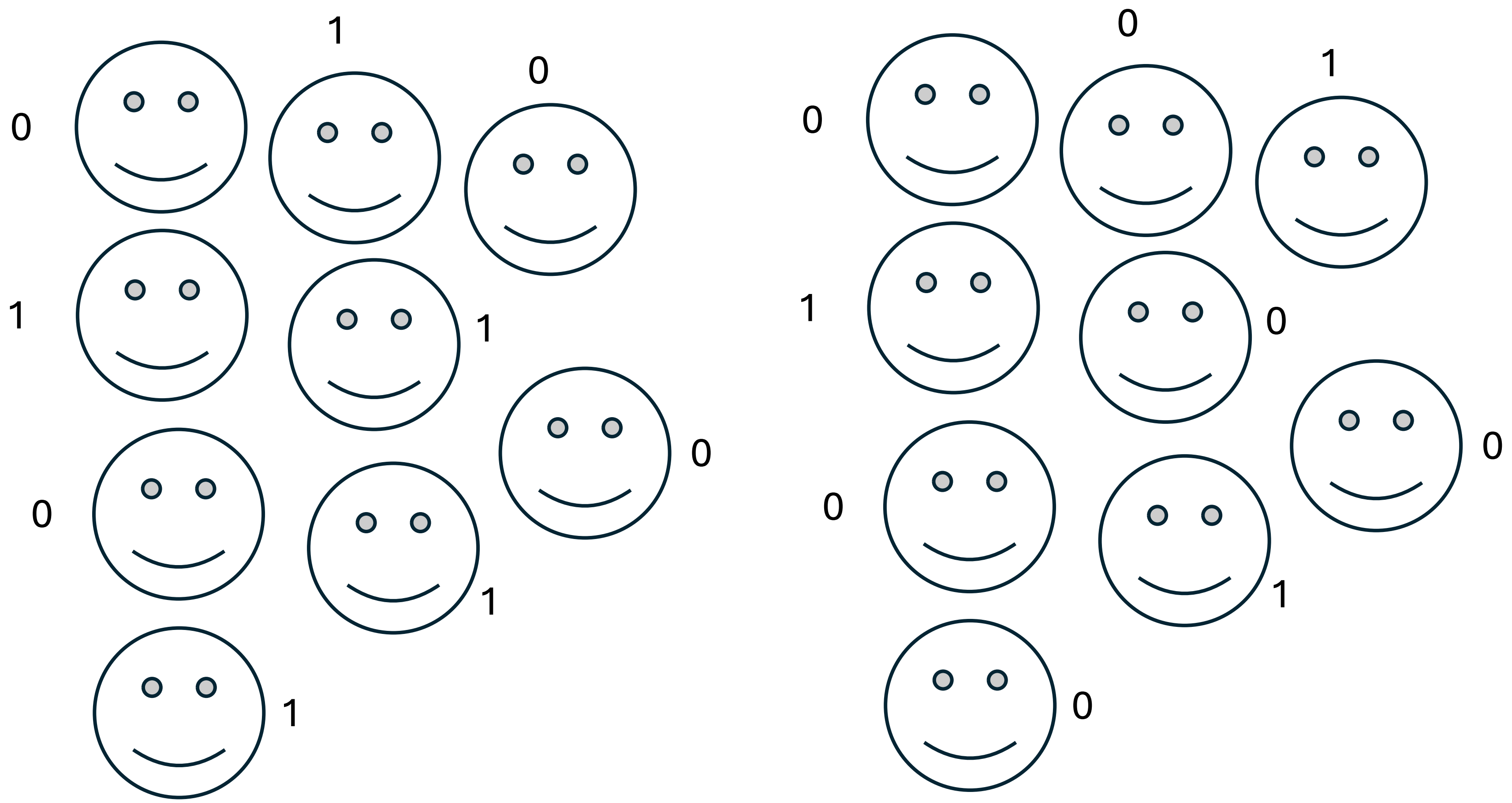
Máximo: 1

Populacional

Mínimo: 0

Desvio Padrão: 51,13%

Valores populacionais formados apenas por 0 e 1



Média (proporção): 42,10%

Máximo: 1

Mínimo: 0

Desvio Padrão: 51,13%

Populacional

Esse caso particular é prático porque se a população é grande o desvio padrão populacional segue aproximadamente uma fórmula que só depende da média:

Desvio padrão = raiz(proporção * (1 – proporção))

Na população ao lado:

Raiz(42,1% * (100% - 42,1%)) = 49,37%

Caso particular importante: proporções

Podemos testar se um certo valor é compatível com uma media amostral

No caso de proporção, podemos testar se uma proporção é compatível com o $\bar{X}$ observado, que escrevemos $\bar{x}$

O processo é o mesmo de antes:

1. Observamos amostra
2. Escolhemos uma estatística
3. Encontramos a distribuição amostral dessa estatística no cenário que nos interessa
4. Posicionamos o valor que encontramos dentro dessa escala

1. Observamos uma amostra

Amostras em que a hipótese de que os elementos da amostra foram sorteados aleatoriamente dentre toda a população são relativamente comuns, mas são muito caras

Quem conduz essas pesquisas normalmente são institutos de pesquisa grandes como Datafolha, IPESP (antigo IBOPE) e o próprio IBGE

Para esse primeiro exemplo, vamos considerar os dados da PNAD-continua de Janeiro de 2024

Essa pesquisa tem abrangência nacional e tem um delineamento amostral bem cuidadoso, vamos considerer que todos da população tem igual chance de serem selecionados

O tamanho de amostra de domicílios é de cerca de 180.000 normalmente

Em Janeiro o desemprego nessa amostra ficou em 7,8%

2. Escolhemos uma estatística

Digamos que o nosso interesse é verificar se o desemprego subiu de dezembro pra Janeiro. O dado divulgado em dezembro foi de 7,4%

Será que esse 7,8% amostral é compatível com o 7,4% observado anteriormente?

Considerando que toda proporção é uma media de 0 e 1, $\bar{X}$ é estatística interessante

O tamanho de amostra de domicílios é de cerca de 180.000 normalmente

Em Janeiro o desemprego nessa amostra ficou em 7,8%, ou seja, $\bar{X}$ observado, que escrevemos $\bar{x} = 7{,}8\%$

3. Encontramos a distribuição amostral dessa estatística no cenário que nos interessa

O tamanho de amostra de domicílios é de cerca de 180.000 normalmente

Em Janeiro o desemprego nessa amostra ficou em 7,8%, ou seja, $\bar{X}$ observado, que escrevemos $\bar{x} = 7{,}8\%$

Será que esse 7,8% amostral é compatível com o 7,4% observado anteriormente?

Pelo teorema central do limite, **se fosse verdade que a media populacional é 7,4%** então a distribuição amostral de $\bar{X}$ seria:

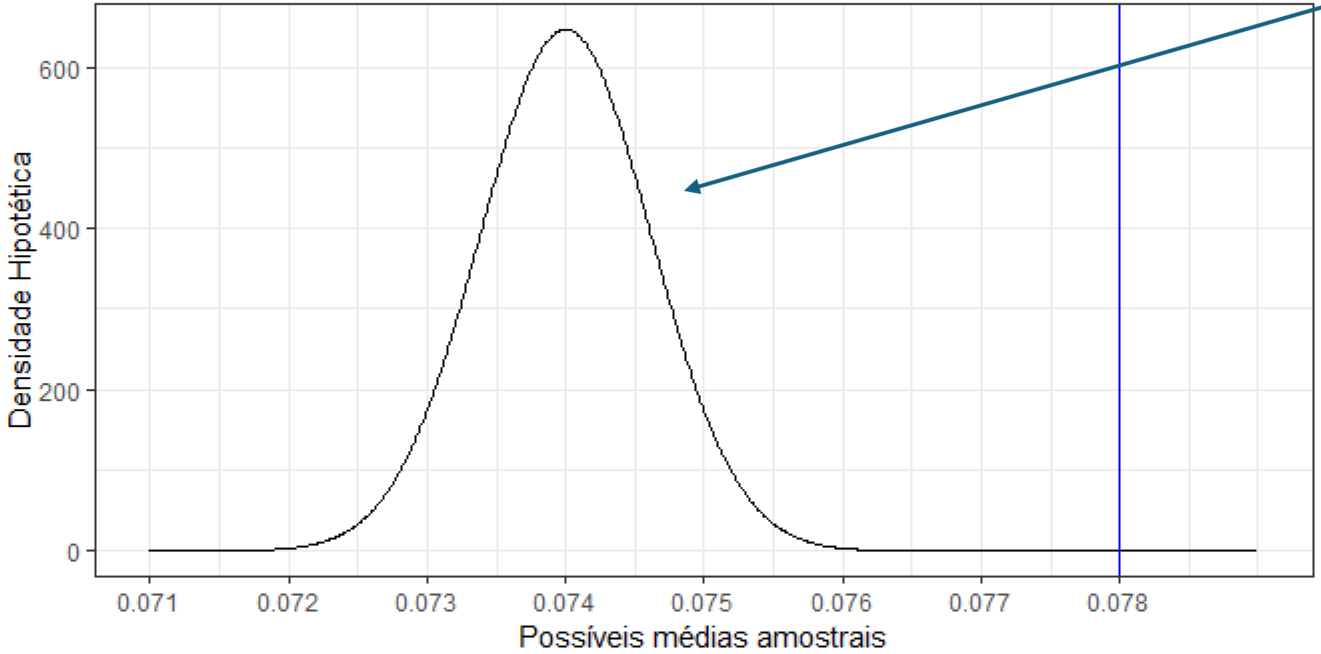
$$\bar{X} \sim N\left(média\ populaciona, \frac{desvio\ padrão\ populacional}{raiz\ de\ n}\right) = N\left(7{,}4\%, \frac{\sqrt{7{,}4\% \times (1 - 7{,}4\%)}}{\sqrt{18000}}\right)$$

4. Posicionamos o valor que encontramos dentro dessa escala

Será que esse 7,8% amostral é compatível com o 7,4% observado anteriormente?

Pelo teorema central do limite, **se fosse verdade que a media populacional é 7,4%** então a distribuição amostral de $\bar{X}$ seria:

$$\bar{X} \sim N\left(média\ populaciona, \frac{desvio\ padrão\ populacional}{raiz\ de\ n}\right) = N\left(7{,}4\%, \frac{\sqrt{7{,}4\% \times (1 - 7{,}4\%)}}{\sqrt{18000}}\right)$$



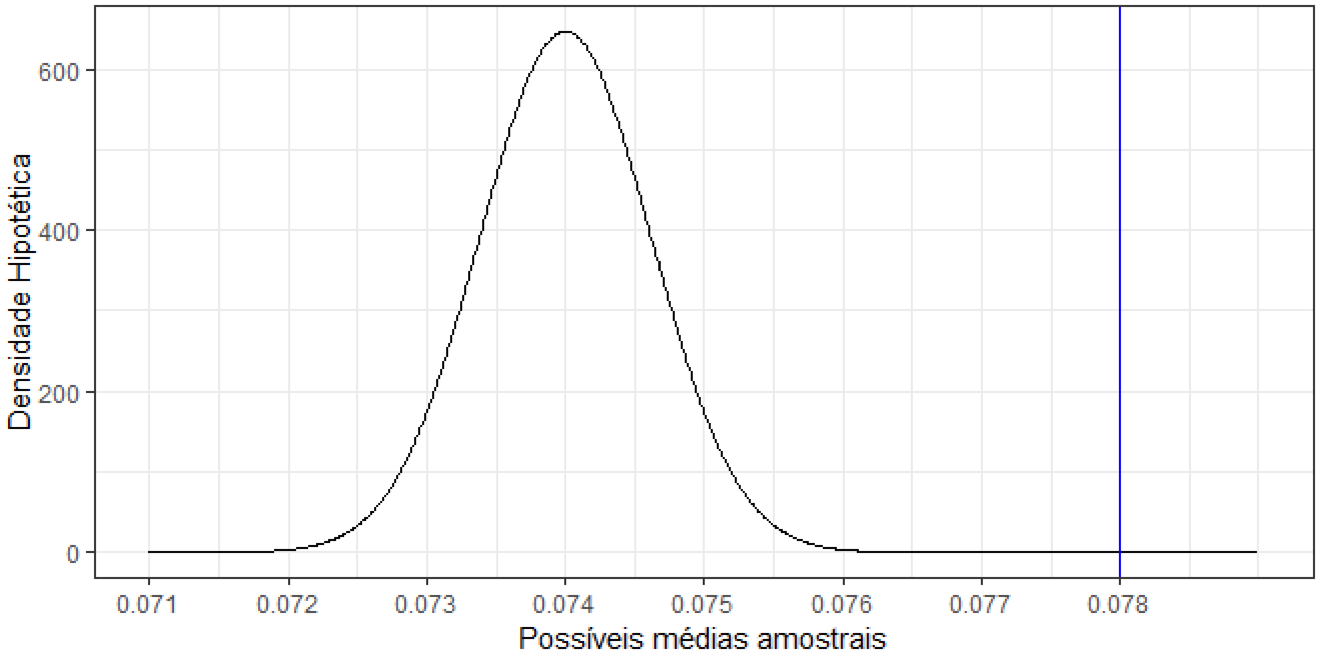
7,8% é bem improvável, o esperado seria Ver x barra perto de 7,4%, com bastante precisão

Como decidir se o posicionamento na escala está “ruim o suficiente?”

Será que esse 7,8% amostral é compatível com o 7,4% observado anteriormente?

Pelo teorema central do limite, **se fosse verdade que a media populacional é 7,4%** então a distribuição amostral de $\bar{X}$ seria:

$$\bar{X} \sim N\left(média\ populaciona, \frac{desvio\ padrão\ populacional}{raiz\ de\ n}\right) = N\left(7{,}4\%, \frac{\sqrt{7{,}4\% \times (1 - 7{,}4\%)}}{\sqrt{18000}}\right)$$



O critério globalmente mais adotado pra esse tipo de comparação é o que chamamos de “valor-p” ou “p-value”

O “valor-p” na imagem ao lado corresponde a área delimitada pela linha azul e a curva preta

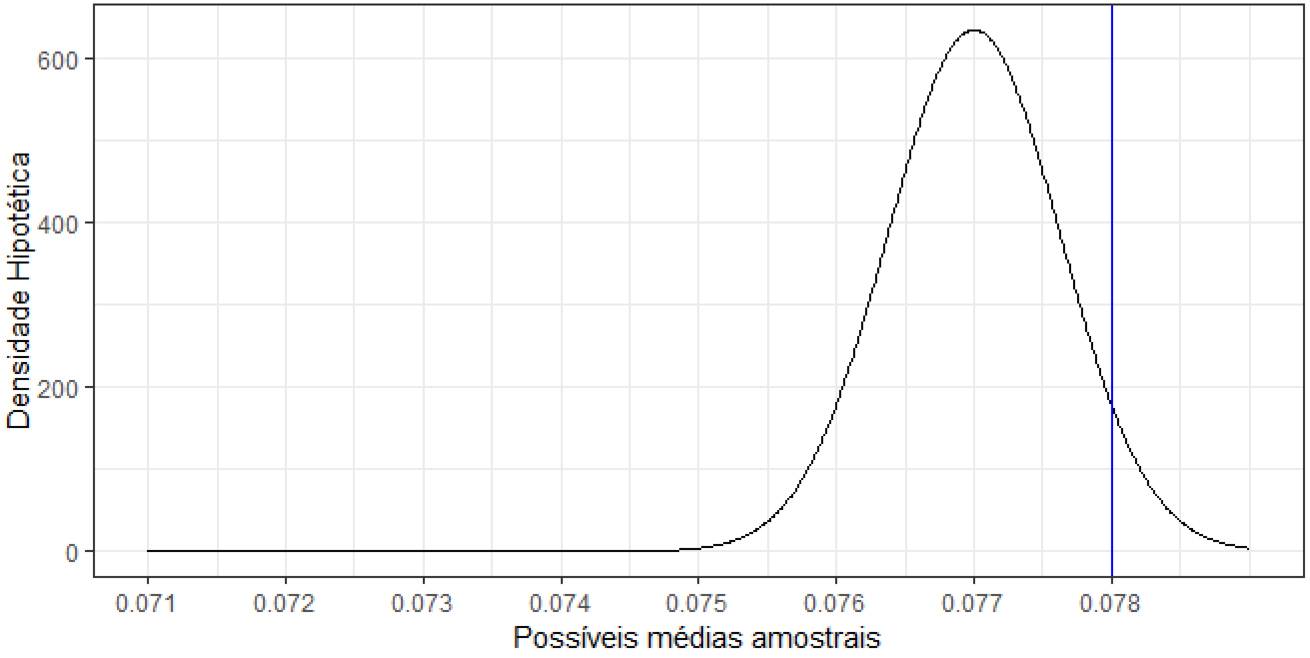
Como essa curva preta é uma normal 99% da área está lá perto do 7,4%. Acima de 7,8% a área é muito muito pequena

Como decidir se o posicionamento na escala está “ruim o suficiente?”

Será que esse 7,8% amostral é compatível com 7,7% (outro valor de interesse, por exemplo)?

Pelo teorema central do limite, **se fosse verdade que a media populacional é 7,7%** então a distribuição amostral de $\bar{X}$ seria:

$$\bar{X} \sim N\left(média\ populaciona, \frac{desvio\ padrão\ populacional}{raiz\ de\ n}\right) = N\left(7{,}7\%, \frac{\sqrt{7{,}7\% \times (1 - 7{,}7\%)}}{\sqrt{18000}}\right)$$



O “valor-p” na imagem ao lado é bem maior! O número fica em 11% (calculado no R)

11% é grande ou é pequeno? Na “estatística das ruas” só consideramos um resultado pra rejeitar uma hipótese, se o valor-p for menor que 5%

Isso não necessariamente é uma estratégia boa, mas é a prática usual e as nuances dessa escolha fogem do escopo do curso

Teste de significância

Esse procedimento que acabamos de executar é o que chamamos de **teste de hipóteses baseados em significância**

O fluxo é sempre o mesmo:

1. Observamos amostra
2. Escolhemos uma estatística
3. Encontramos ou definimos a distribuição amostral dessa estatística no cenário que nos interessa
4. Posicionamos o valor que encontramos dentro dessa escala (calculando o valor-p)

O que vai variar é a estatística escolhida, a distribuição amostral dessa estatística (e o método pra encontrar ela) e o critério que vamos adotar

Teste de significância

Testes de hipóteses estatísticas falam em **parâmetros populacionais**. Em casos de amostragem esses parâmetros sempre são **estatísticas da população**, como por exemplo a média, o desvio padrão, o máximo e o mínimo. A média populacional μ se a população é formada apenas por 0s e 1s é uma proporção. Sendo assim, hipóteses que queremos testar costumam ser escritas assim:

$$\mathrm{H}_0\colon \mu = \text{valor fixo ou } \mathrm{H}_0\colon \mu \geq \text{valor fixo ou } \mathrm{H}_0\colon \mu \leq \text{valor fixo}$$

No exemplo anterior, vamos ver que seguindo os métodos usuais a primeira hipótese testada poderia ser:

$$\mathrm{H}_0\colon \mu \leq 7{,}4\%$$

Ou talvez:

$$\mathrm{H}_0\colon \mu \leq 7{,}7\%$$

Teste de significância

A lógica do teste de significância é sempre a mesma:

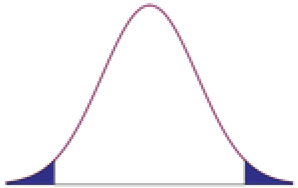
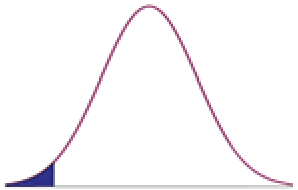
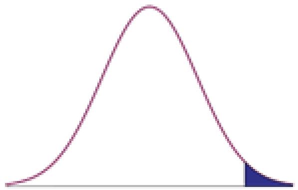
1. Definimos a distribuição amostral para o parâmetro que fica na bordinha do que estamos tentando testar

2. Calculamos o valor-p, que é **a probabilidade de uma amostra sorteada sob a hipótese nula produzir uma estatística pior para a hipótese nula do que a que foi observada**

3. Se o valor-p for muito pequeno, isso quer dizer que a amostra observada estava em uma região muito improvável do eixo X, seria improvável ver algo pior. Se for grande, quer dizer que aconteceu o que de fato deveria acontecer: um $\bar{X}$ perto da media populacional, por exemplo

Valor p

O valor-p sempre vai ser uma área sob a distribuição amostral da estatística escolhida. Dependendo se estamos avaliando se o parâmetro é maior/menor/igual a um valor fixo as áreas calculadas vão mudar, mas sempre são parecidas

z-Test	Null hypothesis H_0	Alternative hypothesis H_1	p-value
Two-tailed	$\boldsymbol{\mu = \mu_0}$	$\boldsymbol{\mu \neq \mu_0}$	
Lower-tailed	$\mu \geq \mu_0$	$\mu < \mu_0$	
Upper-tailed	$\mu \leq \mu_0$	$\mu > \mu_0$	

Aqui para calcular se o $\bar{x}$ é ruim mesmo precisamos pensar que seria possível observar resultados piores abaixo da média

No caso anterior, observamos $\bar{x}$ e usamos ela pra calcular um valor-p bilateral, só interessava a chance de ver percentuais de desemprego maiores

Erros em testes de significância

Qual é a chance de termos decidido incorretamente no caso do desemprego?

Se o desemprego realmente não tiver saído do patamar de 7,4%, qual é a chance de termos errado decidindo que ele mudou?

Quais decisões são possíveis?

Existem muitos jeitos de usar estatística, especificamente nos testes de significância a ideia é sempre **encontrar evidência para rejeitar** hipóteses

Na estatística clássica nunca vamos aceitar hipóteses com base em valor-p. Isso é, o valor-p não ajuda a confirmar uma conclusão, ou ele pelo menos não foi inventado pra isso

Sendo assim existem duas possibilidades de “erro”: rejeitar quando não devia (controlado pelo valor de corte do valor-p) e não rejeitar quando deveria, mas vamos discutir esse caso depois

O que o valor-p não é

“Se P = 5%, a hipótese nula só tem 5% de chance de ser verdade”

Não. 5% é a chance de sob a hipótese nula observamos uma amostra com menos aderência à hipótese nula. Ou seja, isso não diz nada sobre se a hipótese nula é verdade ou mentira, só diz que, dentro dos parâmetros da hipótese nula, o resultado que observamos é mais estranho que 95% das possibilidades

“Um resultado estatisticamente significante é importante do ponto de vista prático”

Não. Muitas hipóteses que podemos testar são inúteis do ponto de vista prático. Um aumento de 1% pode ser grande ou pequeno a depender do context, não necessariamente do valor-p.

Existem vários papers e muita polêmica sobre se é bom usar valor-p, o que ele quer dizer etc. Essas duas observações são importantes e tem muitas outras interessantes.