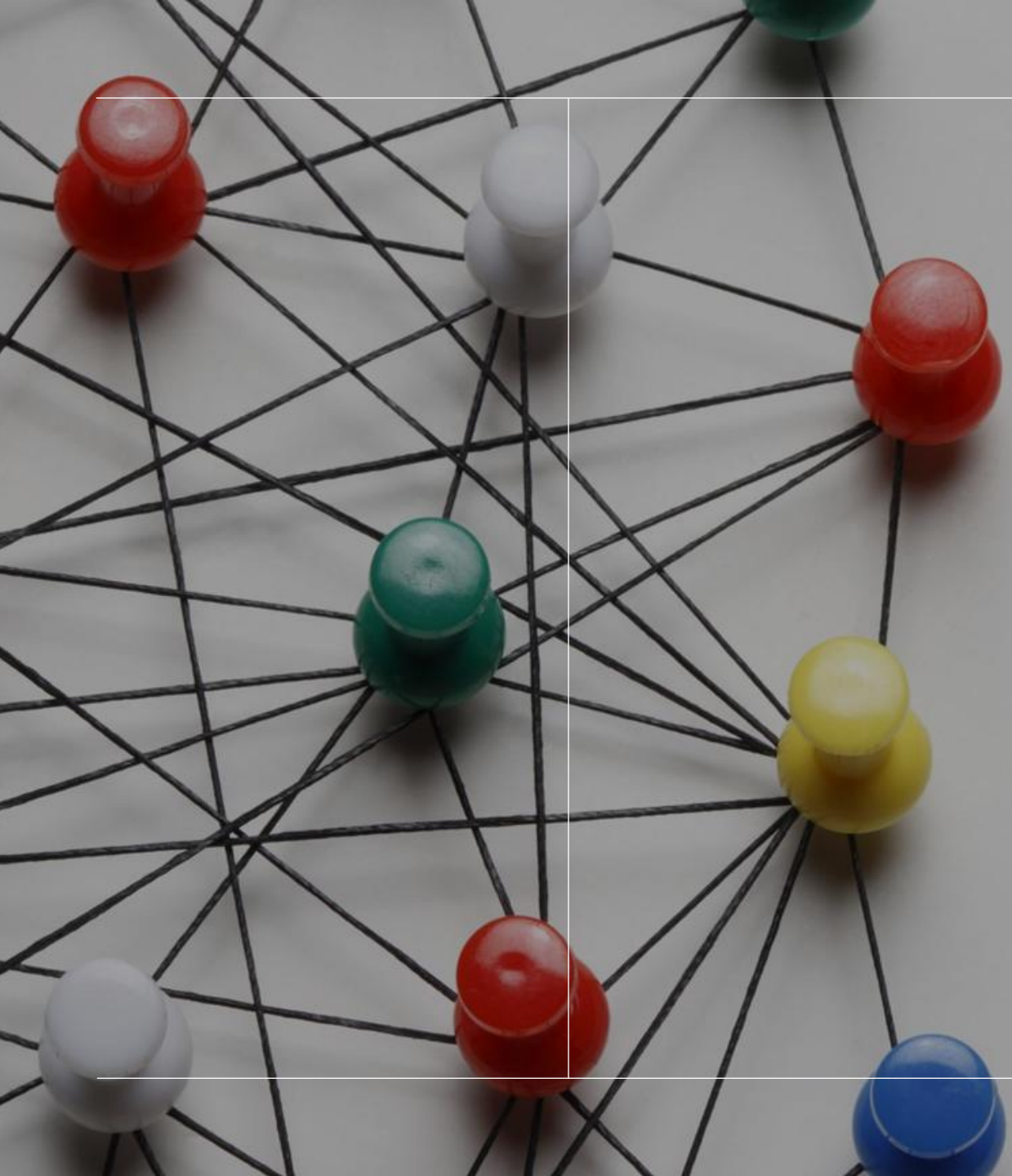
Aulas 6, 7 e 8
Inferência Estatística

PROF. ME. NATAN KLEIN
IA & ANÁLISE DE DADOS
CIÊNCIA DA COMPUTAÇÃO
ATITUS EDUCAÇÃO

Conteúdos

- Inferência
- Amostras e populações
- Representatividade
- Amostragem
- Distribuições probabilísticas
- Distribuição normal
- Desvios da normalidade
- Distribuição amostral
- Teorema do limite central
- Erro padrão
- Intervalos de confiança
- Testagem de hipóteses
- Bootstrap



Inferência

Tipos de raciocínio

- Existem três tipos básicos de raciocínio:

Característica	Raciocínio Dedutivo	Raciocínio Indutivo	Raciocínio Abdutivo
Ponto de Partida	Premissas gerais	Observações específicas	Observação de um fato surpreendente ou incompleto
Conclusão	Necessariamente verdadeira se as premissas forem verdadeiras	Provavelmente verdadeira, generalização das observações	Melhor explicação possível para o fato observado (hipótese)
Direção do Raciocínio	Do geral para o específico	Do específico para o geral	Do efeito para a possível causa
Certeza da Conclusão	Conclusiva	Probabilística	Plausível, mas não garantida
Objetivo Principal	Provar uma conclusão com base em fatos conhecidos	Descobrir padrões e formular generalizações	Explicar um fenômeno ou encontrar a melhor hipótese
Exemplo Simplificado	Todo A é B. C é A. Logo, C é B.	Cisne 1 é branco, cisne 2 é branco... Logo, todos os cisnes são brancos.	A grama está molhada. Pode ter chovido.

Indução, inferência e hipóteses

- A inferência é um exemplo de raciocínio indutivo ... as conclusões que inferimos são, sempre, provavelemente verdadeiras.

- Uma hipótese é uma inferência que fazemos para explicar algo.

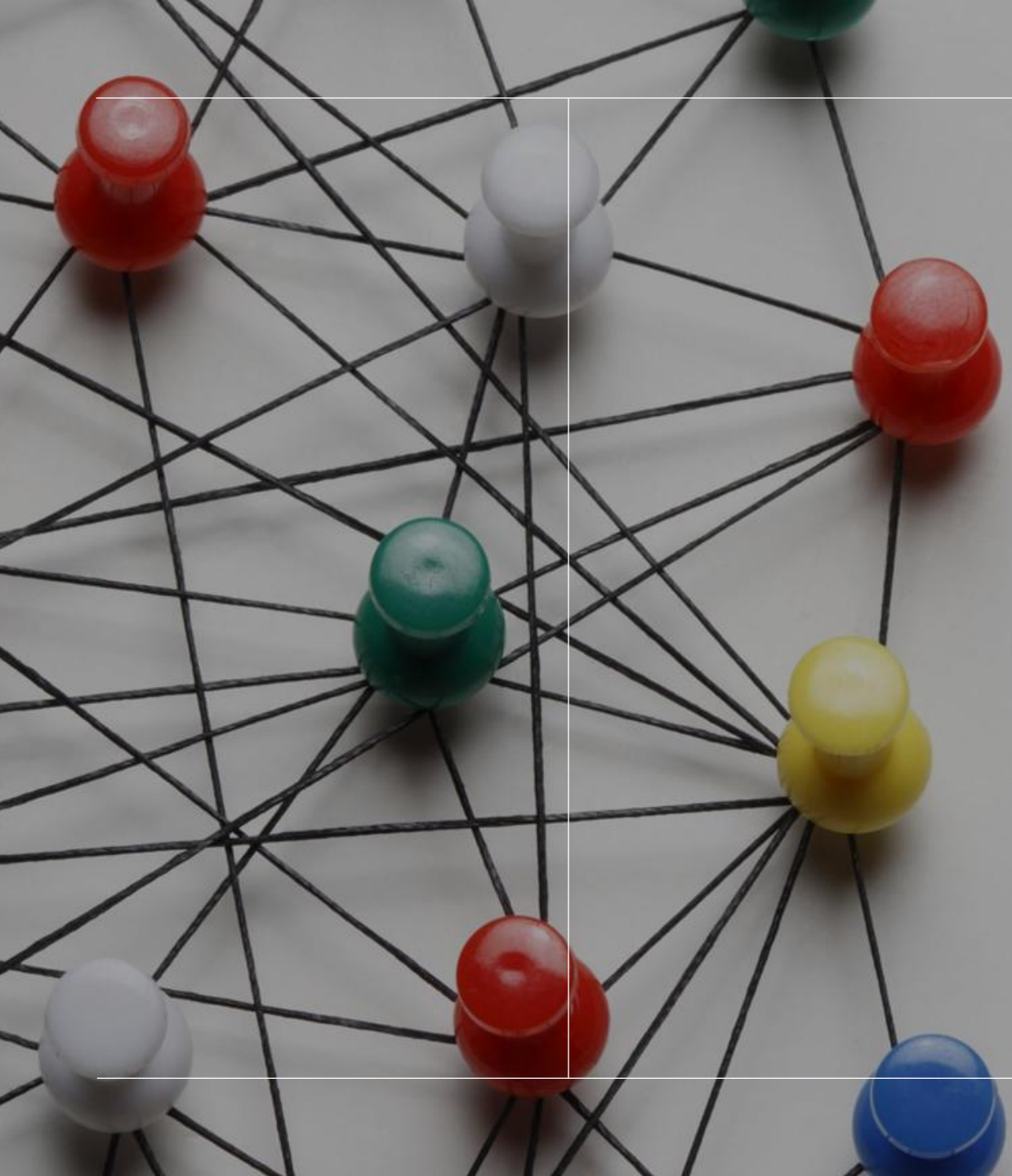
- Podemos usar estatística para formular hipóteses (análise exploratória de dados) ou testar hipóteses (estatística inferencial).

Abordagens científicas

Abordagem confirmatória, guiada pela teoria.

Abordagem exploratória, guiada pelos dados.

Vamos praticar!



Amostras e populações

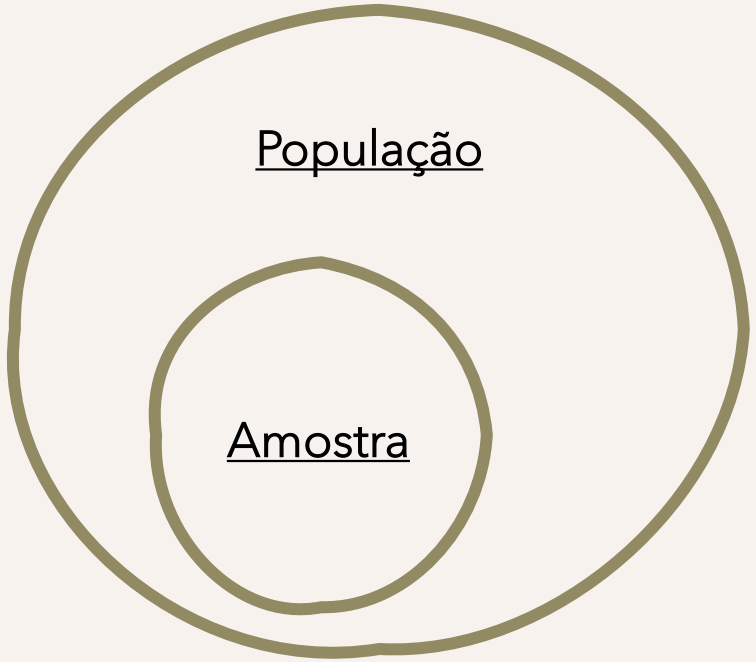
Exemplos

População	Amostra
População Brasileira	Subconjunto de residentes do Brasil
Alunos de uma faculdade	% de alunos de cada curso
Sangue de uma pessoa	Amostra de sangue em jejum
Tempo	Períodos aleatórios
Linha de produção (xícaras)	Peças em períodos aleatórios de tempo
Características psicológicas (nível de disciplina)	Comportamentos específicos

Definições

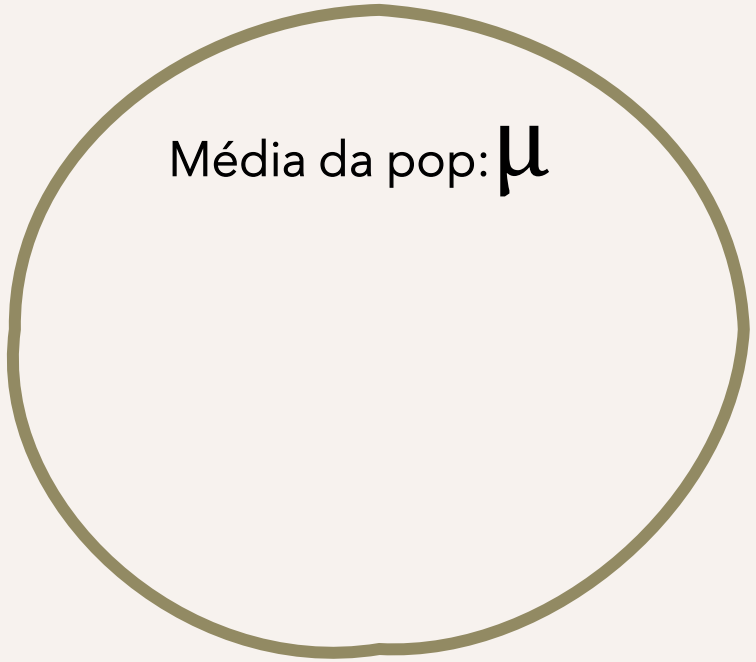
- **População:** um grupo completo de referência sobre o qual queremos tirar conclusões.

- **Amostra:** subconjunto da população que usamos para fazer inferências.



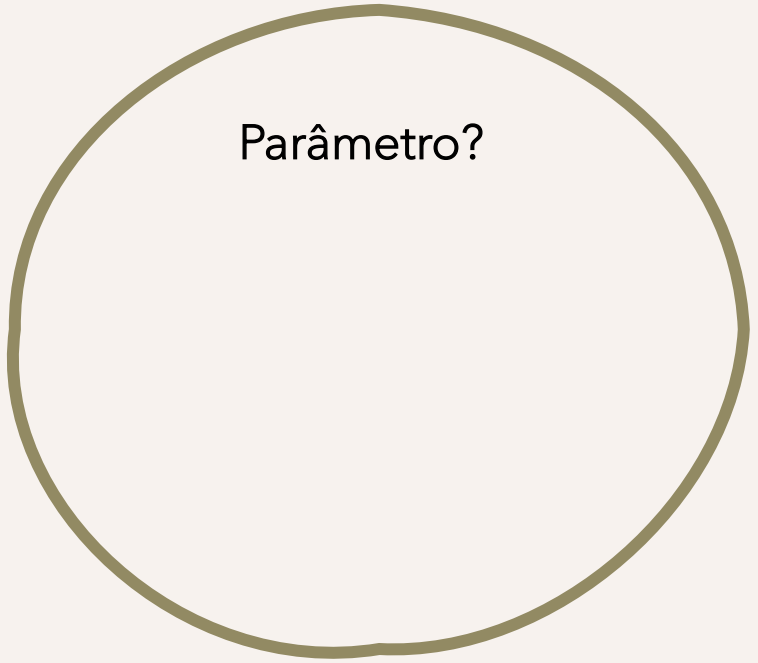
Parâmetros

- Queremos calcular métricas que digam algo sobre uma característica da população (variável).
- Chamamos essas métricas da população de parâmetros.
- **Exemplo de parâmetro:** consumo médio de refrigerantes por dia dos brasileiros.



Parâmetros

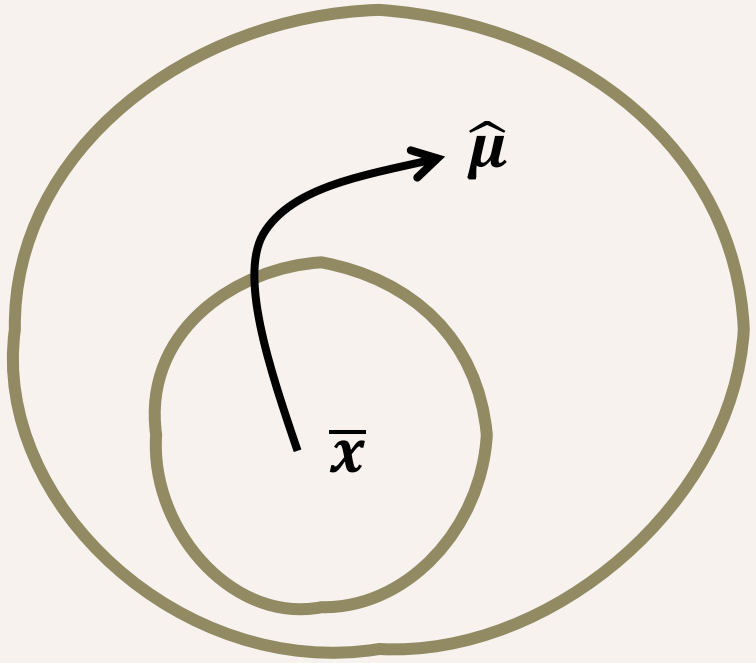
- Mas não temos acesso direto a muitos parâmetros:
  - Nível de depressão de estudantes de graduação
  - Taxa de suicídios por pessoa no Brasil
  - Nível de felicidade das pessoas de um país
  - Número de erros em uma linha de produção ao longo de um ano
  - ...



Estimativa de parâmetros

- O que fazemos para descobrir um valor aproximado para o parâmetro é estimá-lo ou inferi-lo a partir de uma amostra que represente a população!

- Nessa amostra, calculamos uma estatística e usamos ela como estimativa do parâmetro (melhor chute possível).



Exemplo de estimativa de parâmetros

- Exemplo: Pesquisa Nacional de Saúde

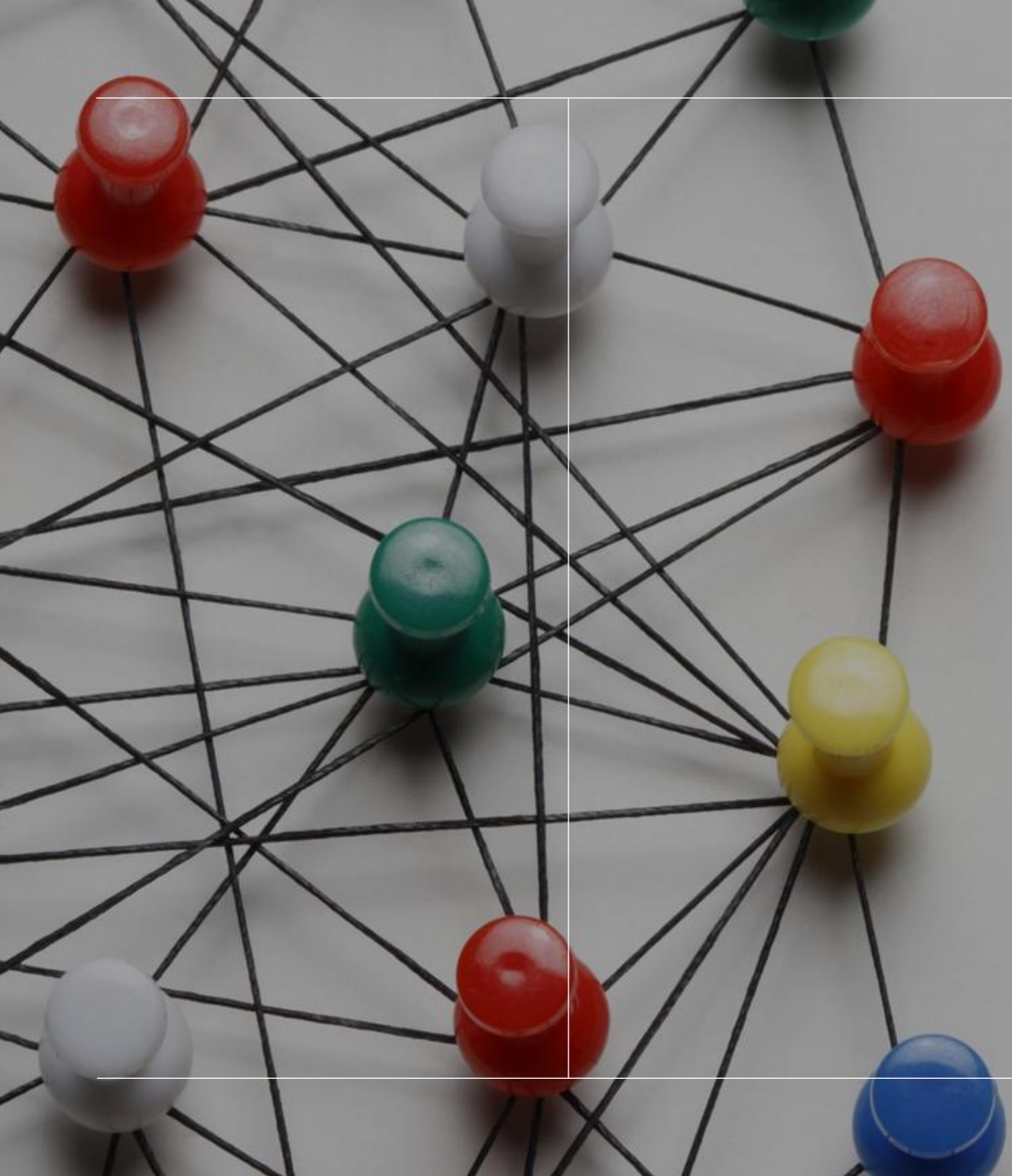
Convenções

- Existem muitos tipos de parâmetros e estatísticas. Uma convenção é usar:
  - Parâmetros → letras gregas: $\alpha\ \beta\ \gamma\ \delta\ \varepsilon\ \epsilon\ \zeta\ \eta\ \theta\ \vartheta\ \iota\ \kappa\ \lambda\ \mu\ \nu\ \xi\ o\ \pi\ \varpi\ \rho\ \varphi\ \phi\ \chi\ \psi\ \omega$ ...
  - Estatísticas → letras latinas: $a\ b\ c\ d\ e\ f\ g\ h\ i\ j\ k\ l$ ...

- Existem estatísticas/parâmetros:
  - **Univariados:** calculados com apenas uma variável.
  - **Bivariados:** calculados com duas variáveis.
  - **Multivariados:** calculados com mais de duas variáveis.

Exemplos de parâmetros/estatísticas

- Ver tabela!

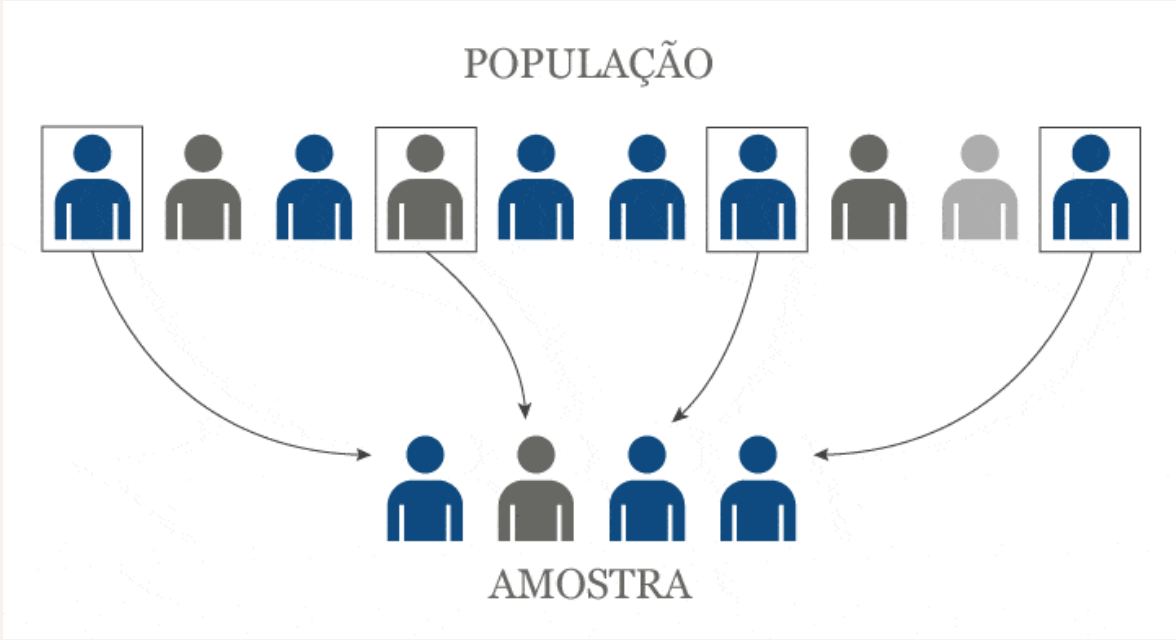


Representatividade

Definição

- Queremos extrair uma amostra que seja representativa da população.

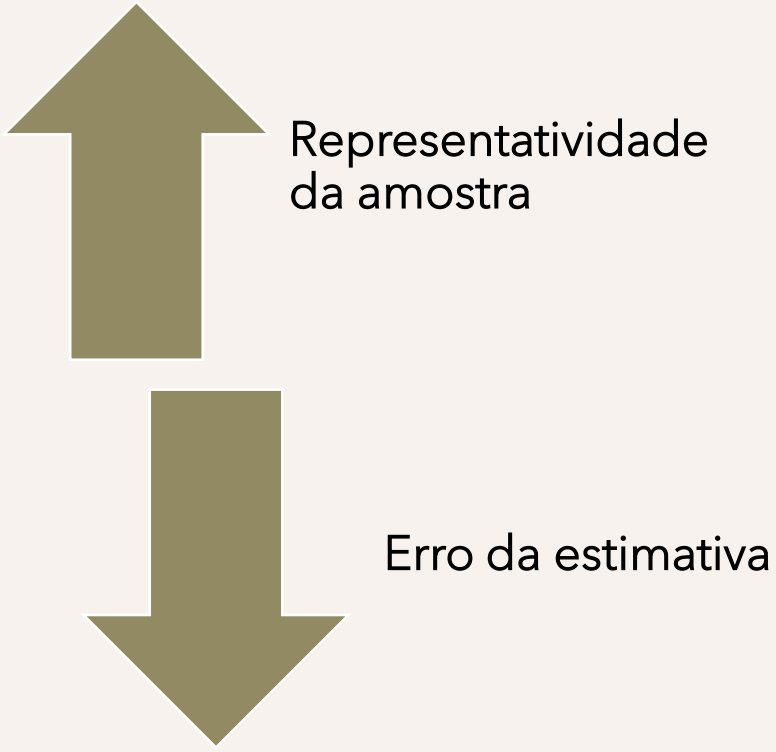
- Uma amostra representativa da população é aquela que reproduz as características fundamentais da população e nos permite calcular estatísticas que estimam parâmetros com o mínimo de erro.



Confiança

- Como posso saber que minhas estatísticas amostrais são medidas confiáveis, fidedignas, dos parâmetros da população?

- As estimativas serão tão boas quanto a amostra for representativa da população, i.e., amostra representativa = boas estimativas.



Erros e vieses

- O erro de estimativa de um parâmetro pode ser expresso pela equação:

$$erro = parâmetro - estatística$$

- Uma amostra não-representativa é fundamentalmente diferente da população em algum aspecto importante, o que enviesa (introduz erro) nas estimativas.

Erros e vieses

Tipo de Viés	Definição	Exemplo
Viés de seleção	Ocorre quando a amostra não é selecionada aleatoriamente, favorecendo certos grupos.	Pesquisa sobre hábitos alimentares em academias, favorecendo pessoas preocupadas com a saúde.
Viés de não resposta	Ocorre quando certos grupos têm maior probabilidade de não responder à pesquisa.	Pesquisa online sobre privacidade com baixa resposta de idosos, menos familiarizados com a internet.
Viés de sobrevivência	Amostra composta apenas por "sobreviventes" de um processo, ignorando os que não sobreviveram.	Estudo de sucesso de empresas iniciantes, incluindo apenas as que ainda operam.
Viés de conveniência	Amostra selecionada pela facilidade de acesso aos participantes.	Pesquisa em shopping center, representando apenas frequentadores de shoppings.
Viés de subcobertura	Alguns membros da população são representados inadequadamente na amostra.	Pesquisa por telefone excluindo pessoas sem telefone fixo ou que não atendem desconhecidos.
Viés de resposta voluntária	Amostra composta apenas por voluntários, geralmente com opiniões fortes.	Pesquisa online sobre produto controverso, atraindo pessoas com opiniões extremas.

Garantias da representatividade

- Existem duas formas de garantir a representatividade de uma amostra:
  - Tamanho da amostra
  - Método de seleção da amostra (amostragem) → **mais importante!**

Tamanho da amostra

- Estatísticas calculadas em amostras muito pequenas dificilmente serão representatividas da população.

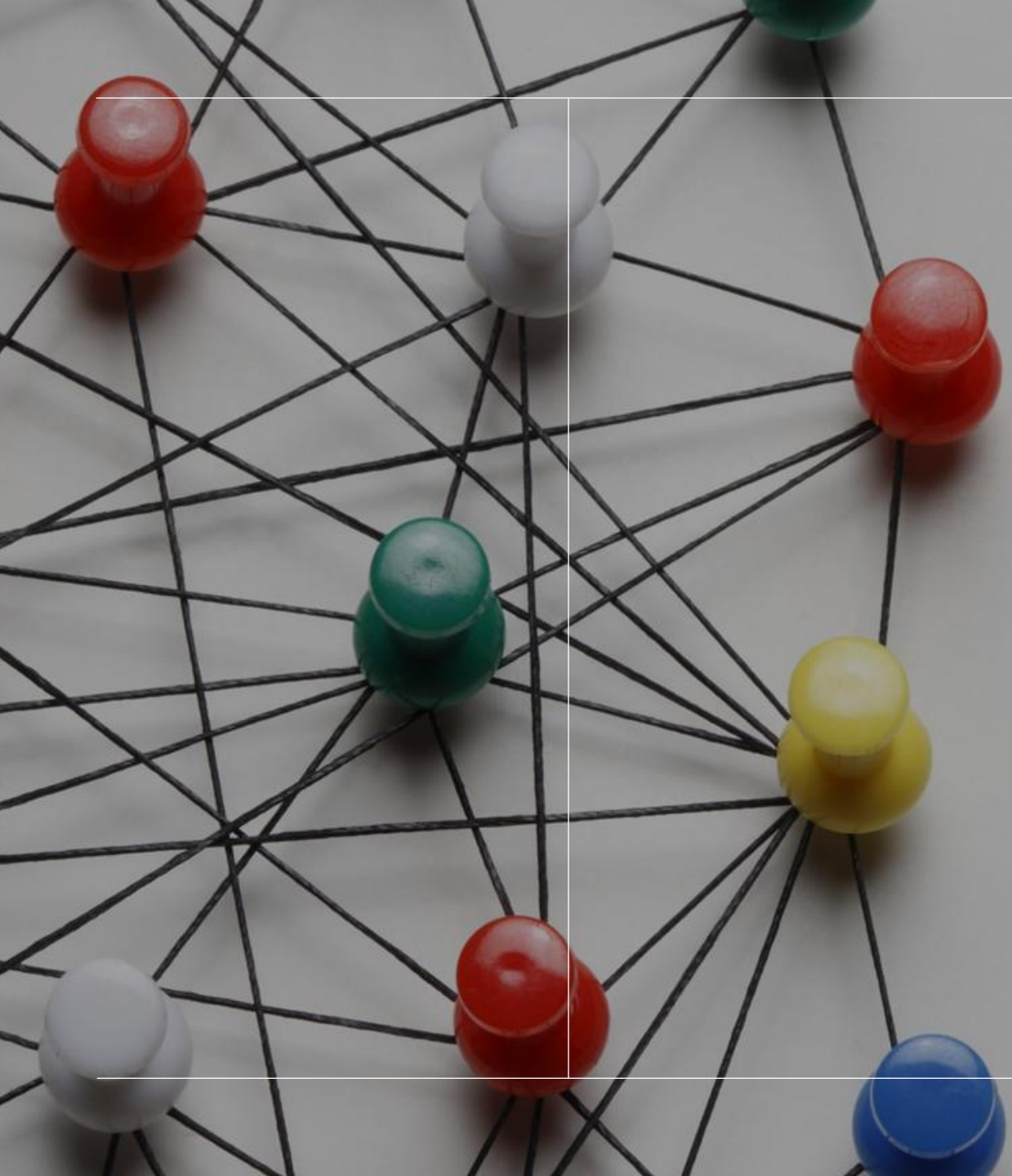
- Não há um critério para o tamanho mínimo de uma amostra porque isso varia de um contexto para outro, mas $n > 30$ é aconselhável.

- Existem formas de descobrir o tamanho ideal de uma amostra por meio de cálculos de tamanho amostral (estatísticas de poder amostral).

Tamanho da amostra

- Vamos usar o banco de dados dos passageiros do Titanic para calcular a média de idade dos passageiros com diferentes tamanhos de amostra selecionadas aleatóriamente!

Amostra	Média	Erro
Total (N)	29,70	-
n = 10	37,28	+7,58
n = 30	34,73	+5,03
n = 60	31,44	+1,74
n = 100	30,38	+0,68



Amostragem

Amostragem aleatória simples

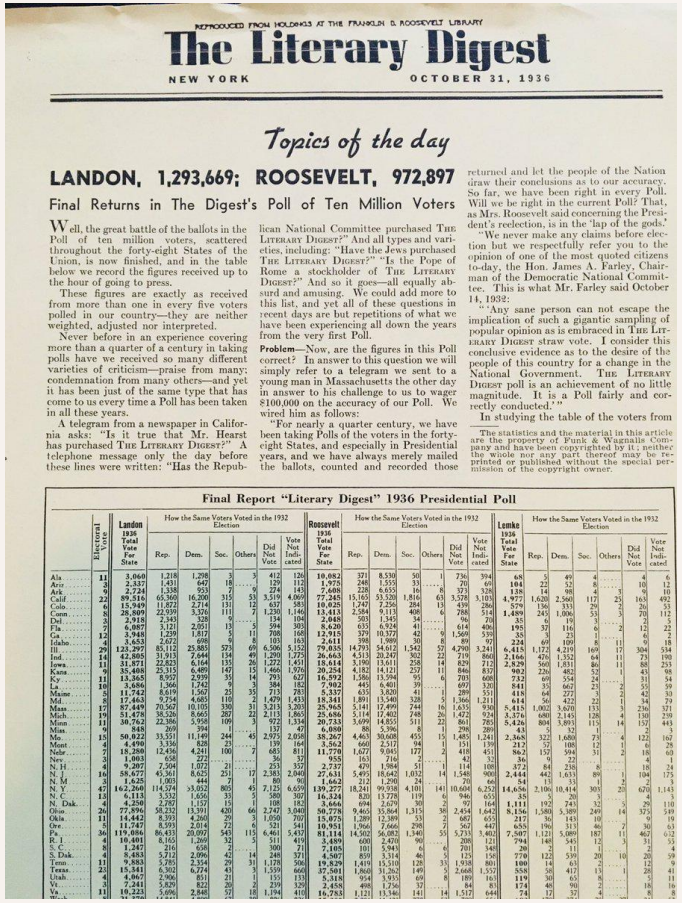
- O melhor método para selecionar objetos da população para nossa amostra é a seleção aleatória:
  - Garante a máxima representatividade.
  - Cada elemento da população é conhecido e é selecionado aleatoriamente para entrar na amostra.
  - Todos tem a mesma chance de entrar na amostra.

Amostragem aleatória simples

- Problemas:
  - Necessita uma base amostral com todos os elementos da população.
  - Processo custoso.
  - Vieses ainda podem atrapalhar o processo.

- Exemplos:
  - Estudantes da Atitus
  - Pacientes de um hospital
  - População geral

Literary Digest vs Gallup



REPRODUCED FROM HOLDINGS AT THE FRANKLIN D. ROOSEVELT LIBRARY

The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased THE LITERARY DIGEST?" A telephone message only the day before these lines were written: "Has the Republican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased THE LITERARY DIGEST?" "Is the Pope of Rome a stockholder of THE LITERARY DIGEST?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the lap of the gods."

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in THE LITERARY DIGEST straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. THE LITERARY DIGEST poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of the voters from

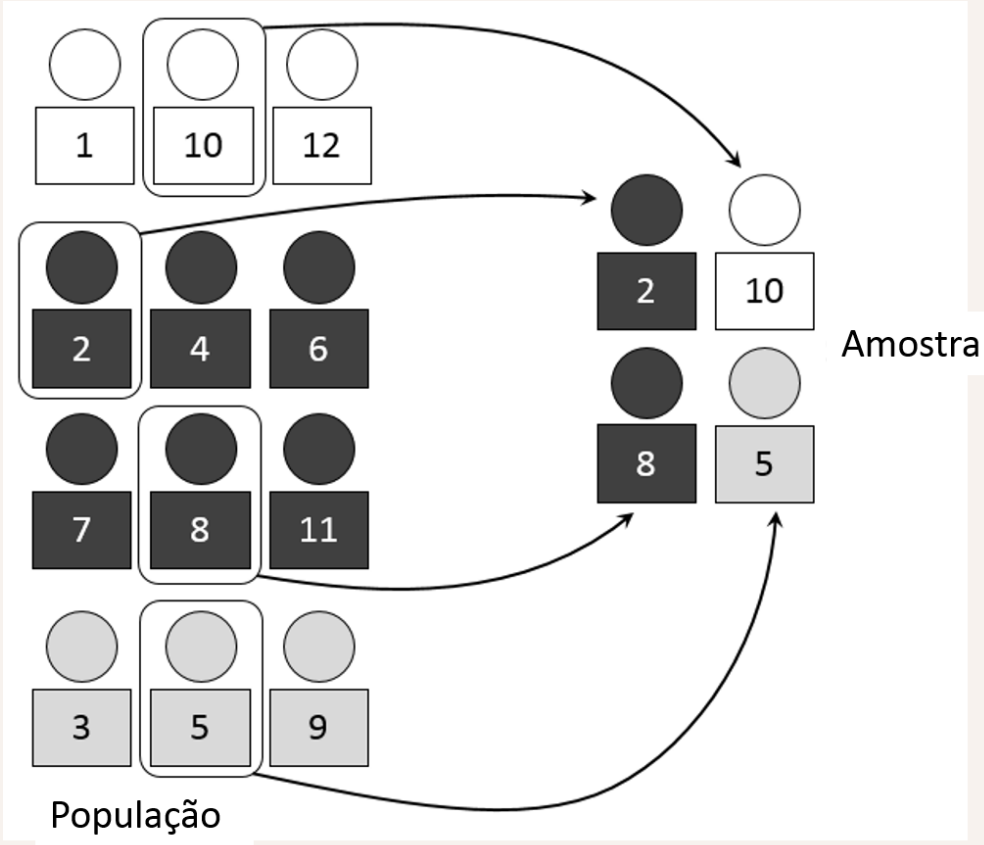
Final Report "Literary Digest" 1936 Presidential Poll

	Electoral Votes	Landon 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated	Roosevelt 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated	Lemke 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated
Ala.	11	3,060	1,218	1,208	3	3	412	126	10,082	321	8,538	56	....	736	394	68	5	49	4	....	4	6
Ariz.	3	2,337	1,431	647	18	....	129	112	1,975	248	1,555	33	....	70	69	104	22	52	8	....	10	12
Ark.	9	2,724	1,338	955	7	....	274	145	7,608	228	6,633	16	6	323	328	138	14	68	5	5	9	10
Calif.	22	89,516	65,360	16,200	315	53	3,519	4,069	77,245	15,165	53,520	1,816	63	3,528	3,103	4,977	1,620	2,560	112	25	163	492
Colo.	6	15,949	11,872	2,714	131	12	637	583	10,025	1,747	7,256	284	13	439	286	579	136	333	29	2	26	53
Conn.	8	28,809	22,939	5,376	113	7	1,230	1,146	13,413	2,594	9,113	400	6	788	514	1,489	245	1,006	53	5	70	112
Del.	3	2,918	2,343	328	9	....	134	104	2,048	503	1,345	34	....	96	70	35	6	19	3	....	2	5
Fla.	7	6,007	3,121	2,051	13	5	594	383	8,420	635	6,924	41	....	614	406	195	37	116	6	2	12	22
Ga.	12	3,948	1,239	1,817	5	11	708	168	12,915	379	10,377	42	9	1,569	539	35	3	23	1	....	6	2
Idaho	4	3,653	2,672	698	9	8	103	163	2,611	398	1,989	50	8	89	97	224	69	109	8	11	9	18
Ill.	29	123,297	85,112	25,885	573	69	6,506	5,152	79,035	14,793	54,612	1,542	57	4,790	3,241	6,415	1,172	4,219	169	17	304	534
Ind.	14	42,805	31,913	7,644	134	49	1,200	1,775	26,663	4,513	20,247	302	22	719	860	2,166	476	1,352	64	11	73	190
Iowa	11	31,871	22,823	6,164	133	26	1,272	1,451	18,614	3,190	13,611	258	14	829	712	2,829	560	1,831	80	11	88	253
Kans.	9	35,408	25,315	6,489	147	15	1,466	1,976	20,254	4,182	14,121	257	11	846	837	902	226	482	52	1	43	98
Ky.	11	13,365	8,957	2,939	35	14	793	627	16,592	1,586	13,594	93	6	703	608	732	69	554	24	....	31	54
La.	10	3,686	1,366	1,742	9	3	384	182	7,902	445	6,401	39	....	697	320	541	35	462	23	....	55	59
Maine	5	11,742	8,619	1,567	25	35	715	781	5,337	635	3,820	41	....	289	551	418	64	277	3	2	42	30
Md.	8	17,462	9,754	4,685	110	2	1,479	1,433	15,341	1,891	13,540	328	5	1,366	1,211	414	54	422	25	3	34	72
Mass.	17	87,449	70,567	10,105	330	31	3,213	3,203	25,965	5,141	17,499	744	16	1,635	930	5,415	1,002	3,670	135	3	236	371
Mich.	19	51,479	38,526	8,665	207	22	2,113	1,865	25,686	5,114	17,402	748	26	1,472	924	3,376	680	2,145	128	4	130	239
Minn.	11	30,762	22,386	5,958	109	3	972	1,336	20,733	3,699	14,855	511	23	861	785	5,426	894	3,893	115	14	157	443
Miss.	9	662	269	394	1	....	137	61	6,000	88	5,396	8	....	298	209	43	7	32	1	....	2	3
Mo.	15	50,022	33,551	11,149	244	45	2,975	2,058	38,267	4,463	30,608	455	15	1,485	1,241	2,363	322	1,608	73	2	122	107
Mont.	4	4,490	3,336	828	23	....	139	164	3,562	660	2,517	94	1	151	139	212	57	105	12	1	6	29
Nebr.	7	15,200	12,456	4,241	100	7	665	811	11,770	1,627	9,045	177	2	418	451	86	157	594	31	2	18	63
Nev.	3	1,003	658	272	....	....	58	37	955	163	716	7	....	42	32	36	9	22	....	....	4	1
N. H.	4	9,207	7,504	1,072	21	....	253	357	2,737	479	1,904	51	1	114	108	372	84	238	8	....	18	24
N. J.	16	58,677	45,361	8,625	251	17	2,383	2,040	27,631	5,495	18,642	1,032	14	1,548	900	2,444	442	1,633	89	1	104	175
N. M.	3	1,625	1,003	444	7	1	89	90	1,662	212	1,290	24	....	70	66	54	13	33	1	2	2	3
N. Y.	47	162,260	114,574	33,052	805	45	7,125	6,659	139,277	18,241	99,938	4,101	141	10,604	6,252	14,636	2,106	10,414	303	20	670	1,143
N. C.	13	6,113	3,532	1,656	33	5	580	302	16,224	820	13,728	119	6	946	653	35	5	20	3	....	4	3
N. Dak.	4	4,250	2,787	1,157	15	....	108	182	3,666	694	2,679	30	2	97	164	1,111	192	743	32	5	29	110
Ohio	26	77,596	58,232	13,391	420	66	2,747	3,040	50,778	9,465	35,864	1,315	38	2,434	1,642	8,136	1,581	5,389	249	14	375	549
Okla.	11	14,442	8,393	4,260	29	3	1,050	707	15,875	1,289	12,389	53	2	682	655	217	36	143	10	....	9	19
Ore.	5	11,747	8,593	2,016	73	6	521	541	10,951	1,964	7,666	298	7	562	447	653	196	313	46	7	30	63
Pa.	36	119,086	86,433	20,097	543	115	6,461	5,437	81,114	14,502	56,082	1,340	55	5,731	3,402	7,507	1,121	5,089	187	11	467	632
R. I.	4	10,401	8,165	1,269	32	5	511	419	3,489	600	2,470	90	....	200	121	794	140	545	12	3	31	53
S. C.	8	1,247	216	658	2	....	300	71	7,105	181	5,943	6	6	731	145	20	2	11	1	....	2	4
S. Dak.	4	8,683	5,712	2,096	42	14	240	371	4,507	859	3,314	46	5	123	150	770	122	539	20	10	20	59
Tenn.	11	9,583	5,785	2,354	29	31	1,178	906	19,829	1,419	15,510	128	33	1,938	801	180	14	63	2	....	12	9
Texas	23	15,341	6,302	6,774	45	3	1,559	660	37,501	1,860	31,262	149	5	2,668	1,557	558	58	417	13	1	28	41
Utah	4	4,067	2,906	851	21	1	155	133	5,318	954	3,935	69	8	189	163	119	30	65	8	....	5	11
Vt.	3	7,241	5,829	822	20	2	239	329	2,458	498	1,736	37	....	84	83	174	46	90	2	....	18	16
Va.	11	10,228	5,696	2,848	57	18	1,194	415	16,783	1,121	13,346	141	14	1,517	644	74	12	37	4	....	8	8

Amostragem estratificada

- Além de ser uma amostragem aleatória, a estratificada busca manter proporções importantes da população

- A amostragem estratificada faz isso ao retirar amostras aleatórias em subconjuntos da população.



Amostragem sistemática

- Todos os elementos da população são conhecidos e uma lógica é usada para selecionar os elementos, por exemplo, selecionar todo décimo elemento.

Amostragem por conveniência

- Usada quando não é possível utilizar a amostragem aleatória, logo, os casos são selecionados pela conveniência do pesquisador.

- Embora seja o método menos rigoroso, é o mais usado e menos custoso.

- Tenta-se minimizar os vieses ao buscar assemelhar as características da amostra às da população, quando essas características são conhecidas.

- **Exemplo:** amostra de estudantes universitários

Amostragem por conglomerados

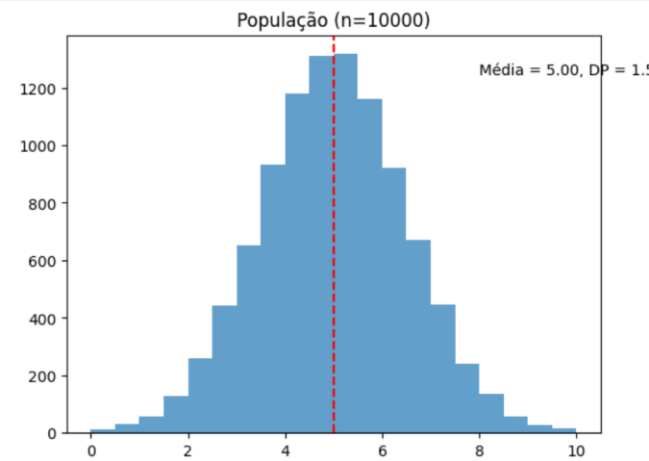
- A população é dividida em grupos chamados de conglomerados e uma amostragem aleatória é realizada entre os grupos.

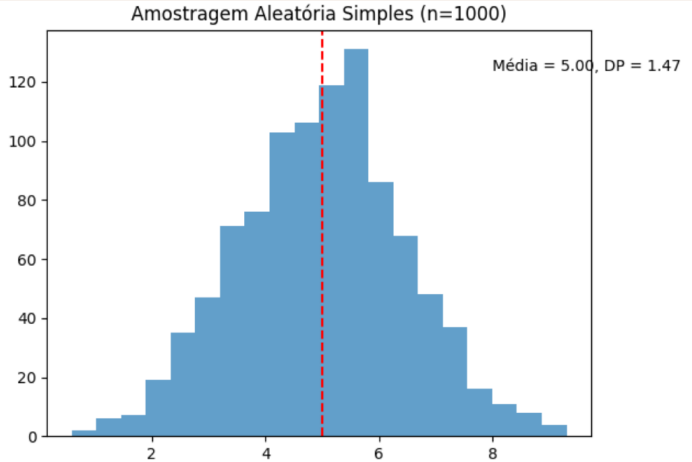
- Usada em pesquisas nacionais importantes:
  - PNAD
  - SAEB
  - ...

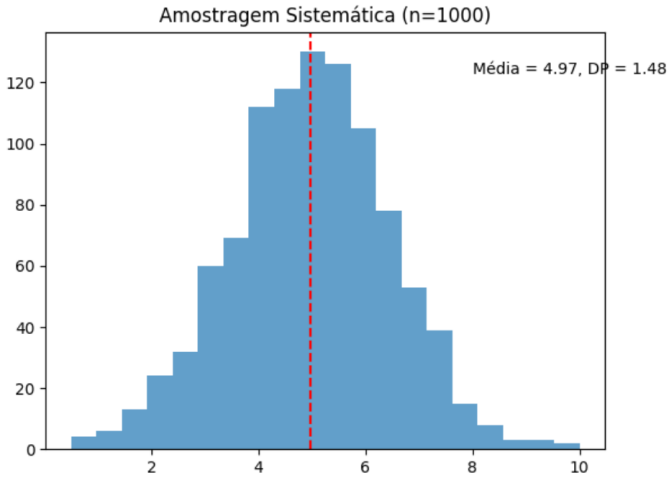
Resumo

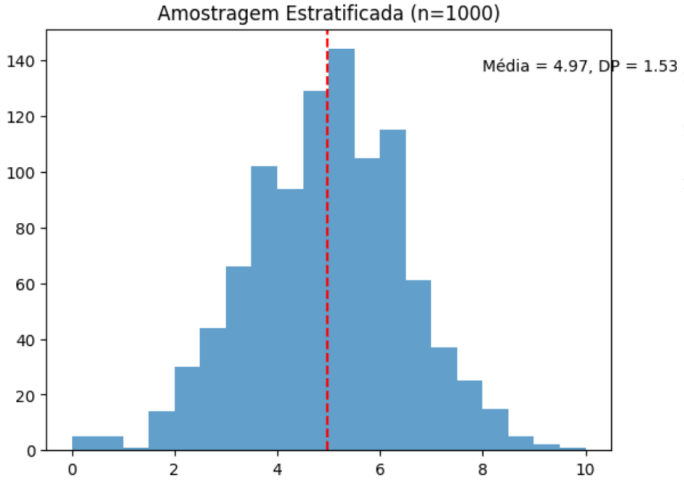
Método de Amostragem	Descrição	Exemplo
Amostragem Aleatória Simples	Cada elemento da população tem a mesma probabilidade de ser selecionado.	Sortear números aleatórios para escolher 50 pessoas de um grupo de 500.
Amostragem Sistemática	Os elementos são selecionados a partir de intervalos fixos.	Escolher a cada 10ª pessoa em uma lista ordenada de 1.000 indivíduos.
Amostragem Estratificada	A população é dividida em subgrupos (estratos), e elementos são selecionados de cada grupo.	Selecionar proporcionalmente alunos de diferentes faixas etárias de uma escola.
Amostragem por Conglomerados	A população é dividida em grupos (conglomerados), e alguns grupos são selecionados.	Escolher algumas salas de aula de uma escola e entrevistar todos os alunos dessas salas.
Amostragem por Conveniência	Os elementos são selecionados com base na facilidade de acesso.	Entrevistar as pessoas que estão disponíveis em um shopping.
Amostragem por Quotas	A seleção é baseada em características específicas da população.	Escolher 20 homens e 20 mulheres para uma pesquisa.

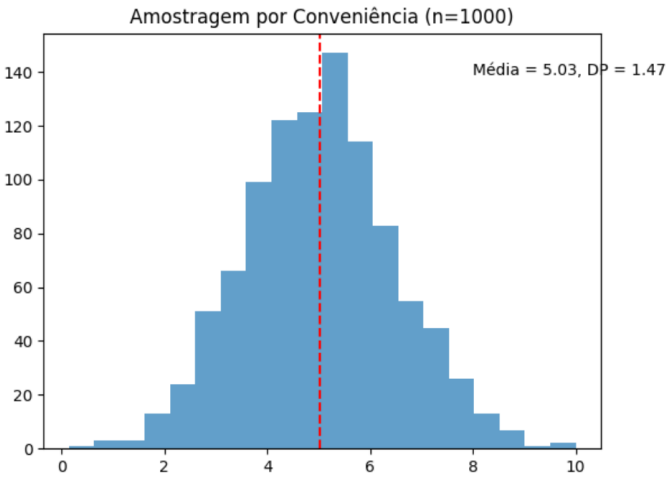
Comparação dos métodos







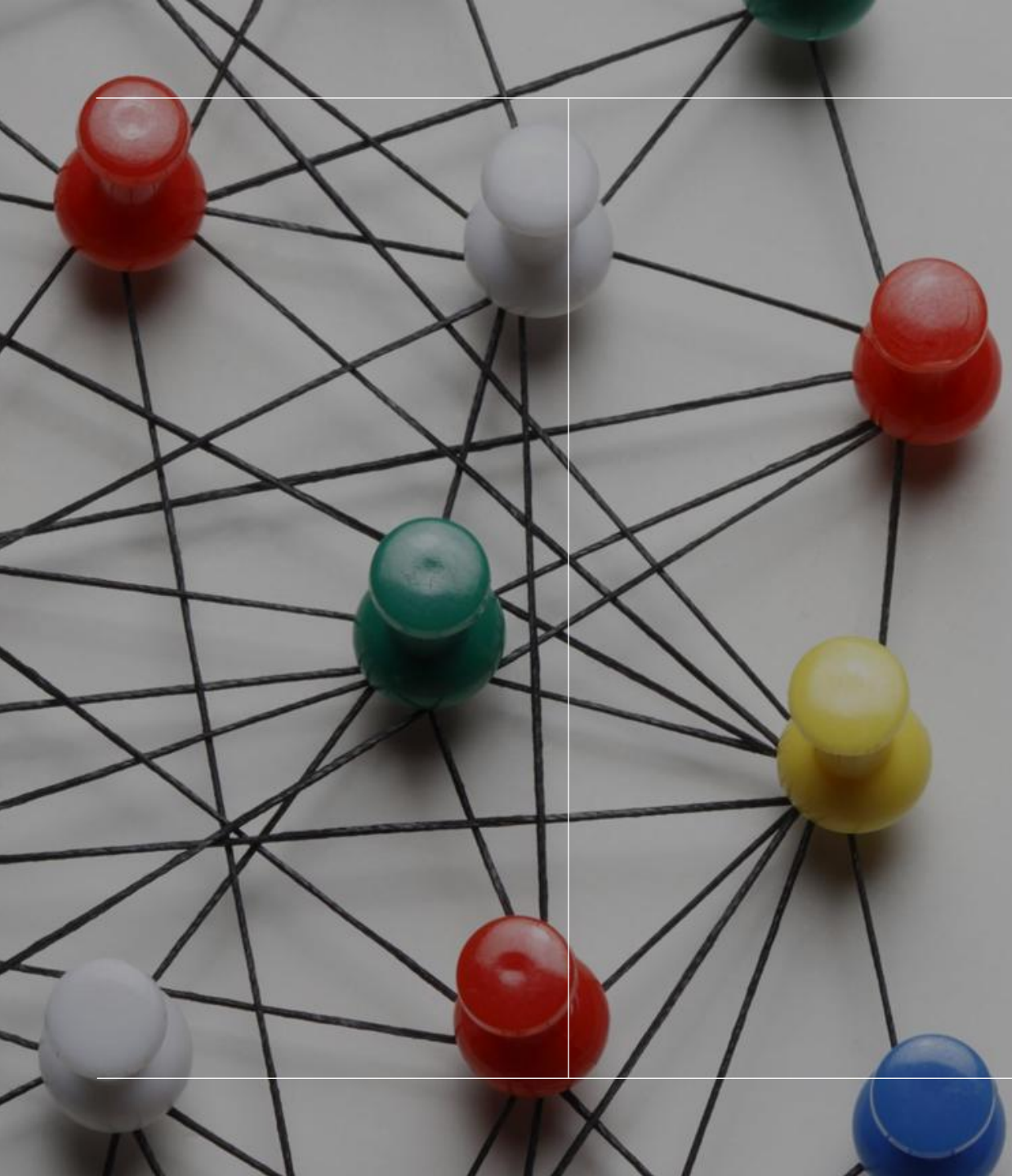




E aí?

- O que é mais preciso?

- Uma amostra de 200 casos ou 10 amostras de 20 casos?



Distribuições probabilíticas

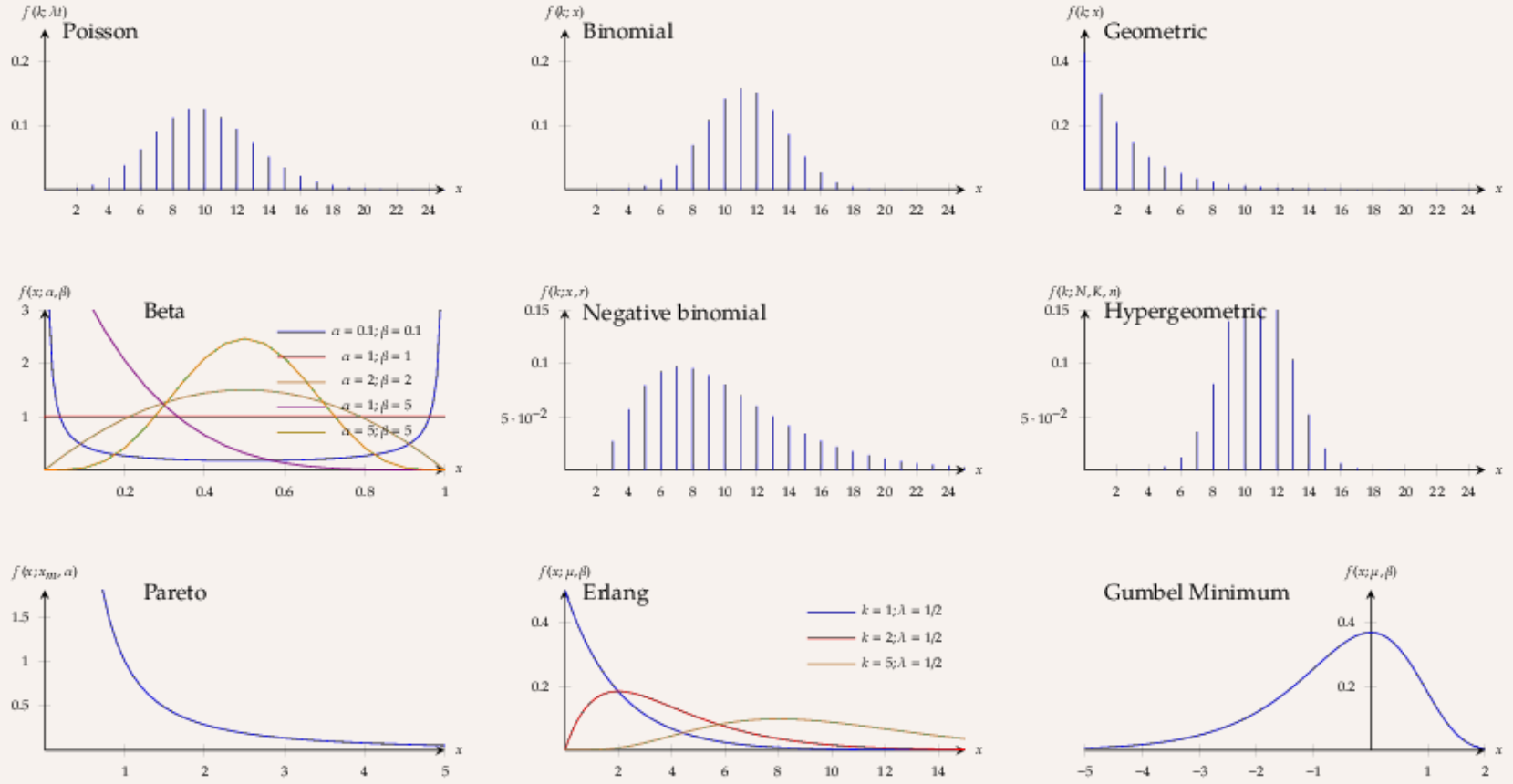
Dois tipos de distribuições

- Existem dois tipos de distribuições:
  - **Teóricas:** usadas para representar padrões de distribuição que, teoricamente, podem existir.
  - **Empíricas:** as distribuições dos dados reais coletados.

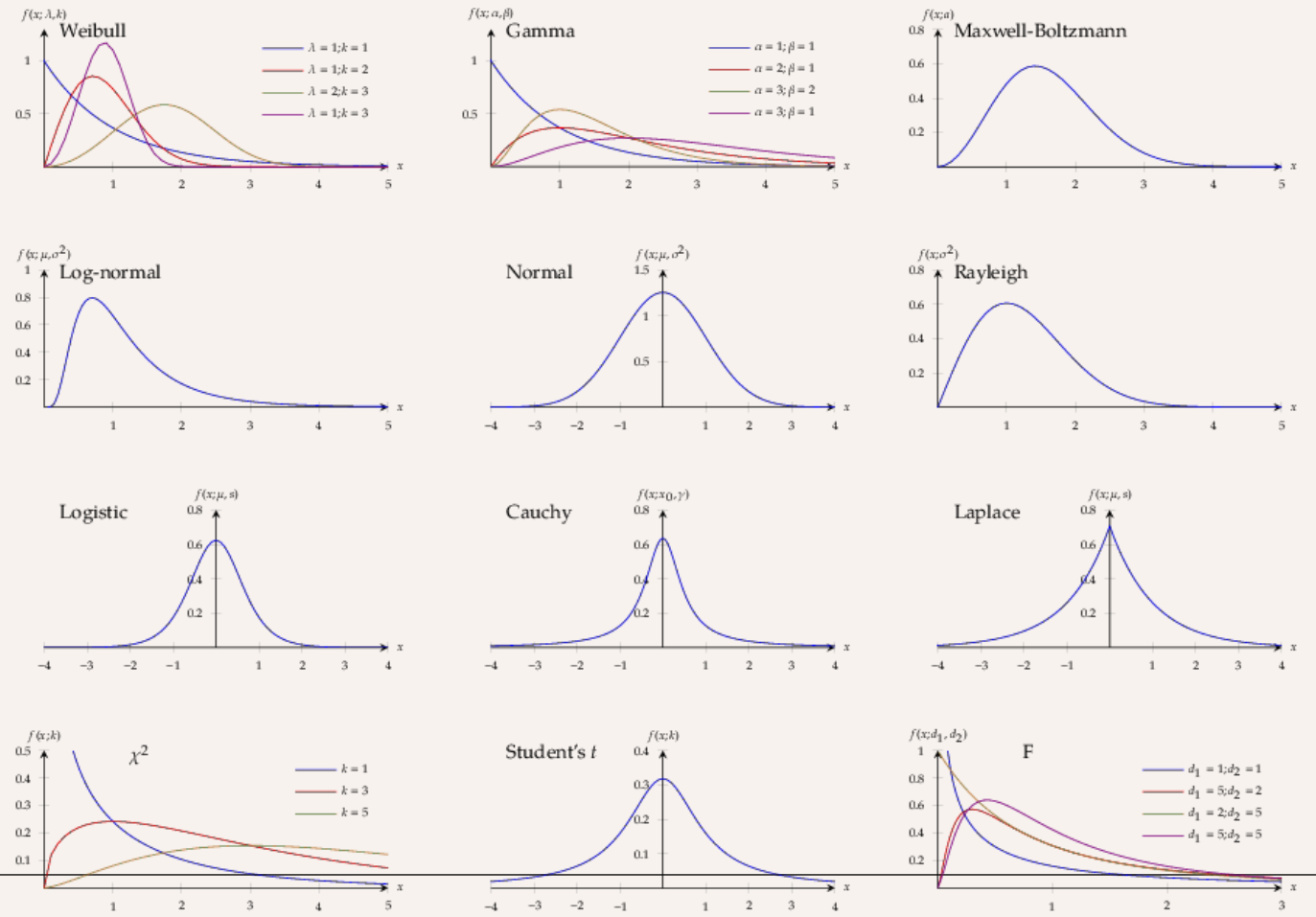
- Podemos visualizar uma distribuição empírica plotando um histograma ou um gráfico de densidade.

- As distribuições teóricas são úteis para modelar eventos probabilíticos.
  - **Eixo-x** = valores da distribuição
  - **Eixo-y** = probabilidade
  - **Área sob a curva** = 1

Várias distribuições



Várias distribuições

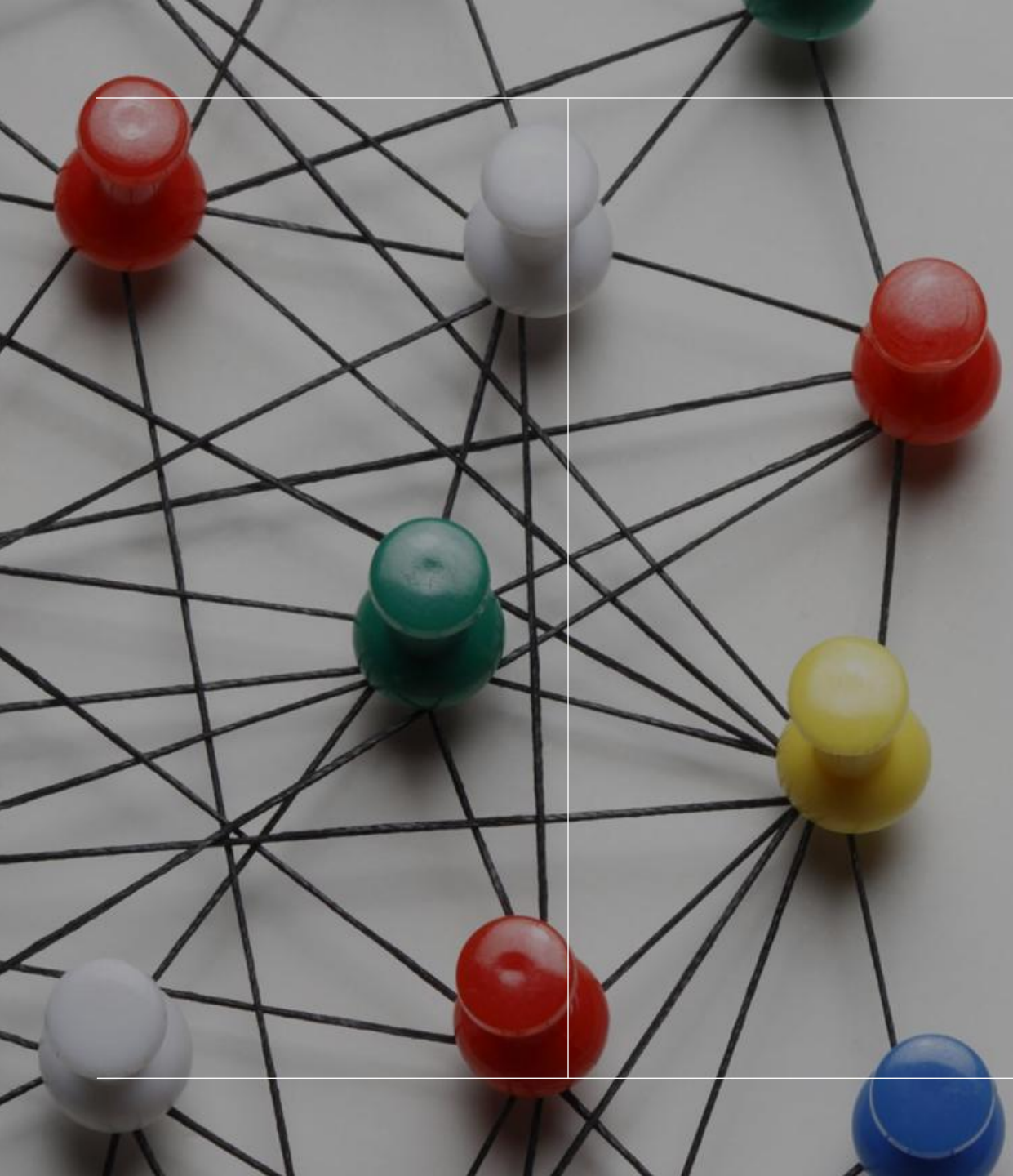


Distribuições mais usadas

- As distribuições mais usadas na prática são:
  - Normal
  - Exponencial
  - Poisson
  - Logística
  - qui-quadrado
  - *t* de Student
  - *F*

Calma!

- Essas distribuições tendem a ser mais importantes para a teoria do que para a prática da estatística e análise de dados. Por isso, não vamos nos aprofundar nelas.



Distribuição normal

Tábua de Galton

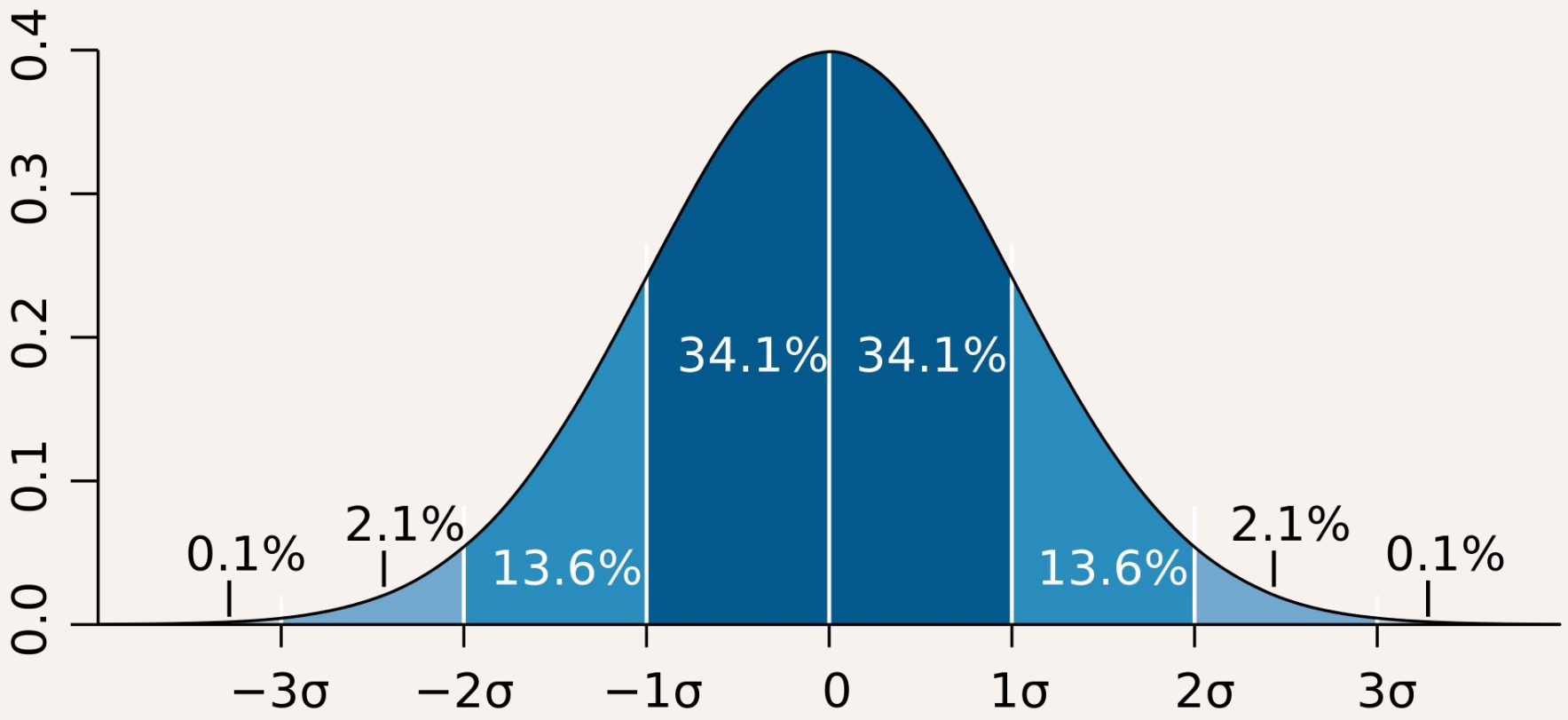
Teórica e empírica

- O nome distribuição normal (Gaussiana) pode se referir a duas coisas:

  - A distribuição de probabilidades teórica chamada distribuição normal padrão (DNP);

  - Dizemos que uma variável no nosso conjunto de dados tem distribuição normal quando sua forma é semelhante à DNP (uma distribuição normal empírica).

DNP



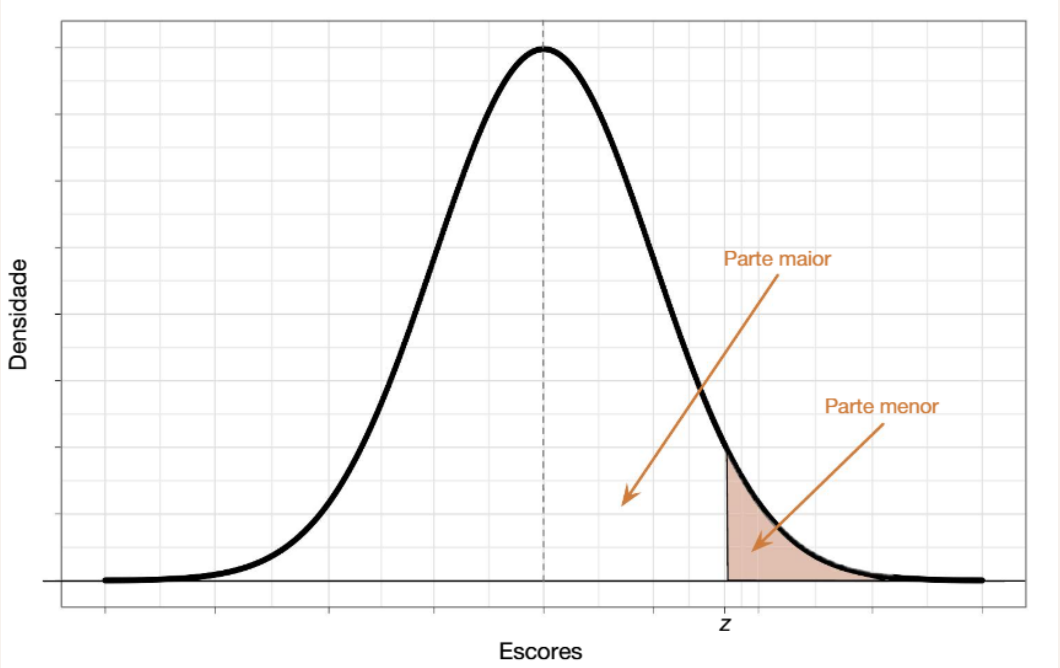
Teórica e empírica

- Para cada valor da DNP há uma probabilidade correspondente, i.e.

$$x_i \rightarrow y_i = p(x_i)$$

Características

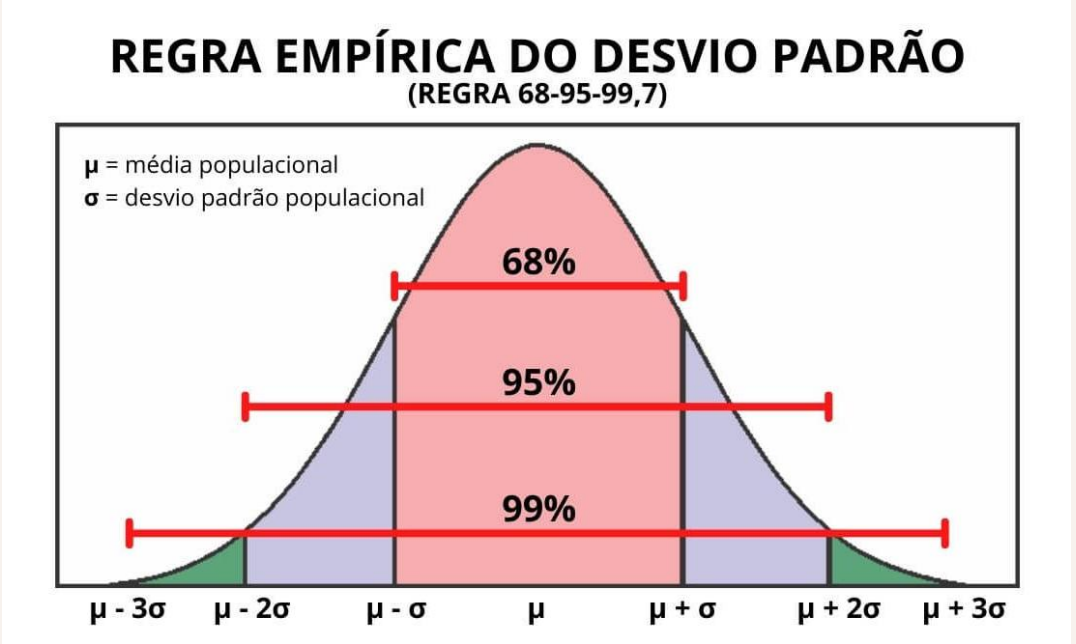
- Forma de sino
- Centrada na média e simétrica em torno dela
- **Média = 0 e desvio-padrão = 1**
- Cada escore no eixo-x é chamado de escore Z, uma medida de desvios-padrão da média.
- Para cada escore Z existe uma região abaixo e acima, que é a probabilidade de um escore Z menor que ou maior que aquele.



Regra empírica

- Uma característica da DNP é a regra empírica, que especifica proporções da curva para intervalos de desvio-padrão acima e abaixo da média:
  - $\pm 1\ DP\ da\ média = 68\%\ da\ curva$
  - $\pm 2\ DP\ da\ média = 95\%\ da\ curva$
  - $\pm 3\ DP\ da\ média = 99\%\ da\ curva$

- Na realidade, qualquer intervalo pode ser computado usando os escores z. As probabilidades acima e abaixo da curva para cada escore z já estão tabeladas.



Escores Z

- Para usar as propriedades da distribuição normal com os valores de uma variável, posso transformar os escores da minha variável em escores z:

$$z_i = \frac{x_i - \bar{x}}{DP}$$

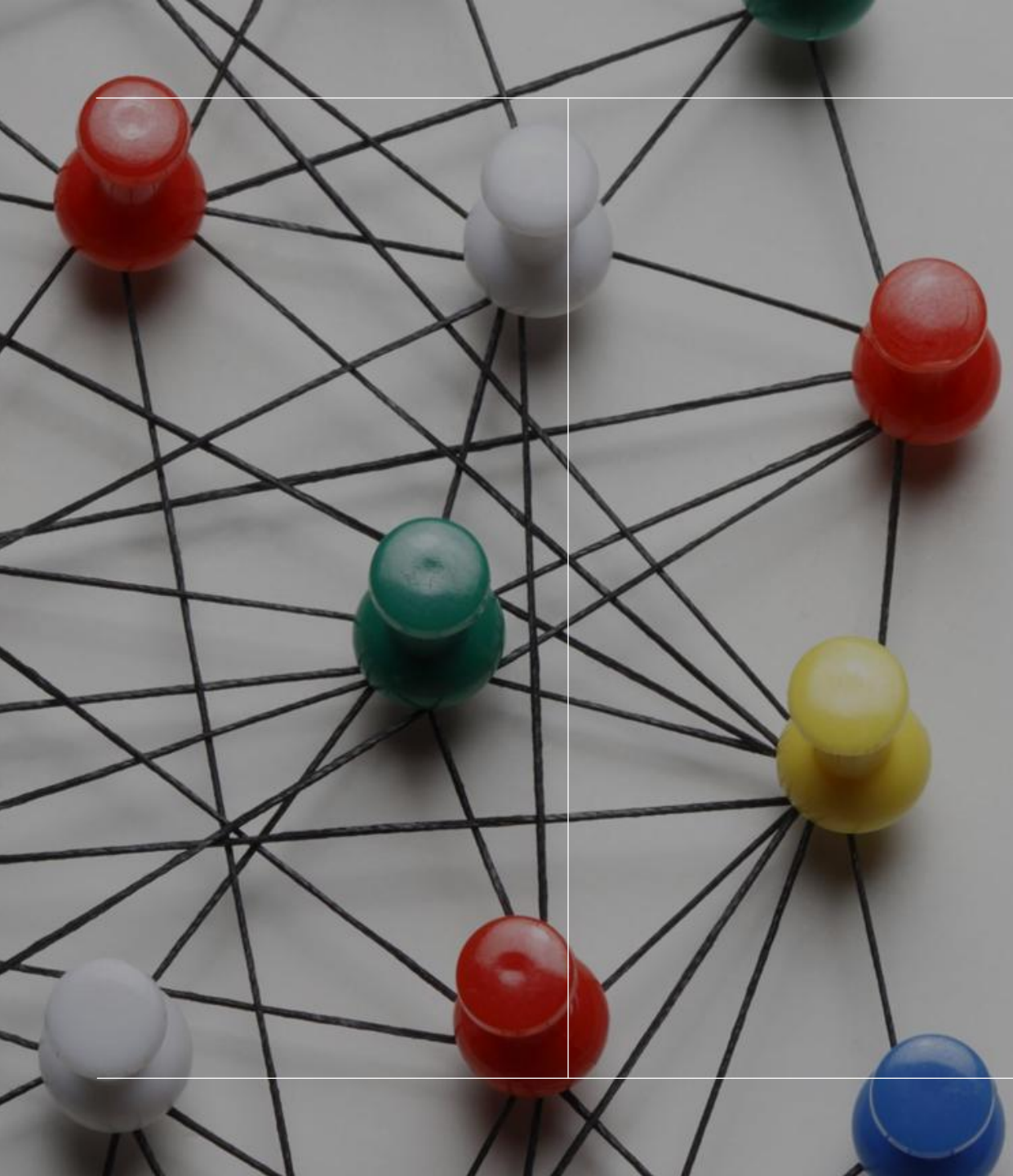
- Assim, os escores z representam o afastamento da média na DNP, cuja média é 0 e desvio-padrão é 1.

	Variável	Desvios	Escore Z
	1	-3,86	-1,33
	4	-0,86	-0,30
	7	2,14	0,74
	8	3,14	1,08
	2	-2,86	-0,99
	3	-1,86	-0,64
	9	4,14	1,43
Média	4,86		0
DP	2,90		1

Exercício

- Use a tabela de escores Z para encontrar as probabilidades dos escores Z correspondentes!

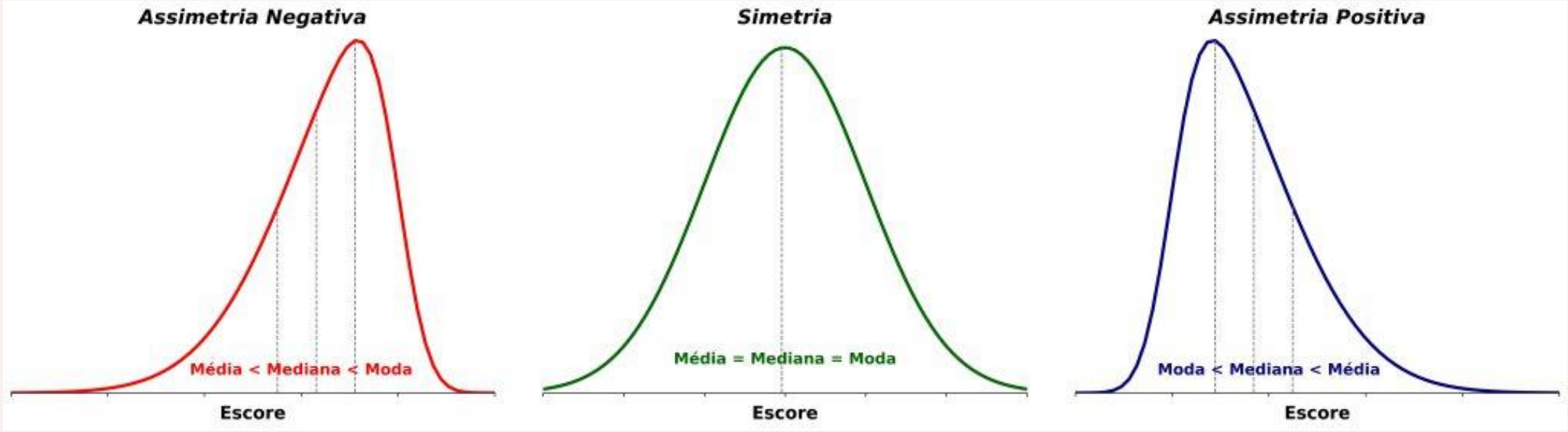
	Variável	Desvios	Escore Z
	1	-3,86	-1,33
	4	-0,86	-0,30
	7	2,14	0,74
	8	3,14	1,08
	2	-2,86	-0,99
	3	-1,86	-0,64
	9	4,14	1,43
Média	4,86		0
DP	2,90		1



Desvios da normalidade

Assimetria

- A distribuição normal é simétrica em torno da média, *i.e.*, tem a mesma proporção de dados acima e abaixo da média (50-50%). Uma distribuição assimétrica viola essa característica.

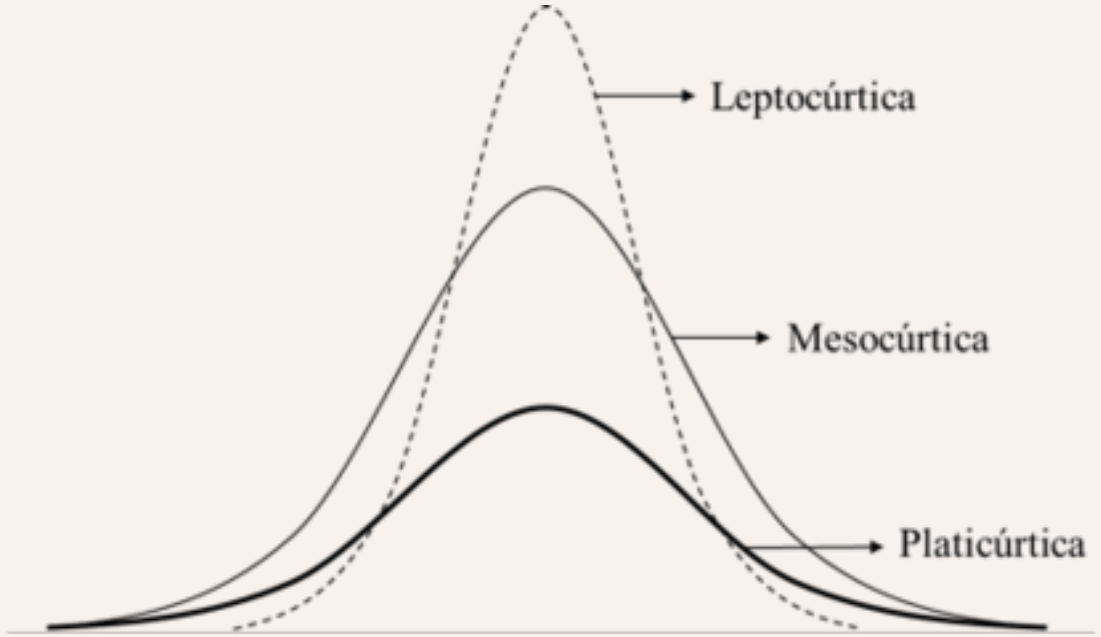


Métrica

- A assimetria (desvio da simetria, *skewness*) pode ser calculada e assumir valores positivos, negativos ou zero:
  - Assimetria = 0: distribuição simétrica;
  - Assimetria > 0: assimetria positiva;
  - Assimetria < 0: assimetria negativa.

Curtose

- A distribuição normal é mesocúrtica, *i.e.:*
  - **Nem pontuda demais:** dados muito concentrados na média, desvio-padrão muito pequeno;
  - **Nem larga demais:** dados muito espalhados da média, desvio-padrão grande.



Métrica

- A curtose (*kurtosis*) pode ser calculada e assumirá valores positivos, negativos ou zero:
  - Curtose = 3: distribuição mesocúrtica;
  - Curtose > 3: distribuição leptocúrtica;
  - Curtose < 3: distribuição platicúrtica.

- Em algumas implementações, calcula-se o excesso de curtose:
  - Curtose = 0: distribuição mesocúrtica;
  - Curtose > 0: distribuição leptocúrtica;
  - Curtose < 0: distribuição platicúrtica.

Como saber se a minha distribuição é normal?

- Podemos:
  - Avaliar o histograma
  - Verificar os valores de assimetria e curtose
  - Plotar um gráfico-QQ
  - Realizar o teste de normalidade

Gráfico-QQ

- Gráfico Quantil-Quantil:
  - Converte os valores da minha variável para percentis (eixo-x)
  - Converte os escores Z da DNP para percentis (eixo-y)
  - Plota os eixos em um gráfico
  - Se os pontos recaem sob a linha diagonal, a distribuição é normal

- Exemplo no JASP!

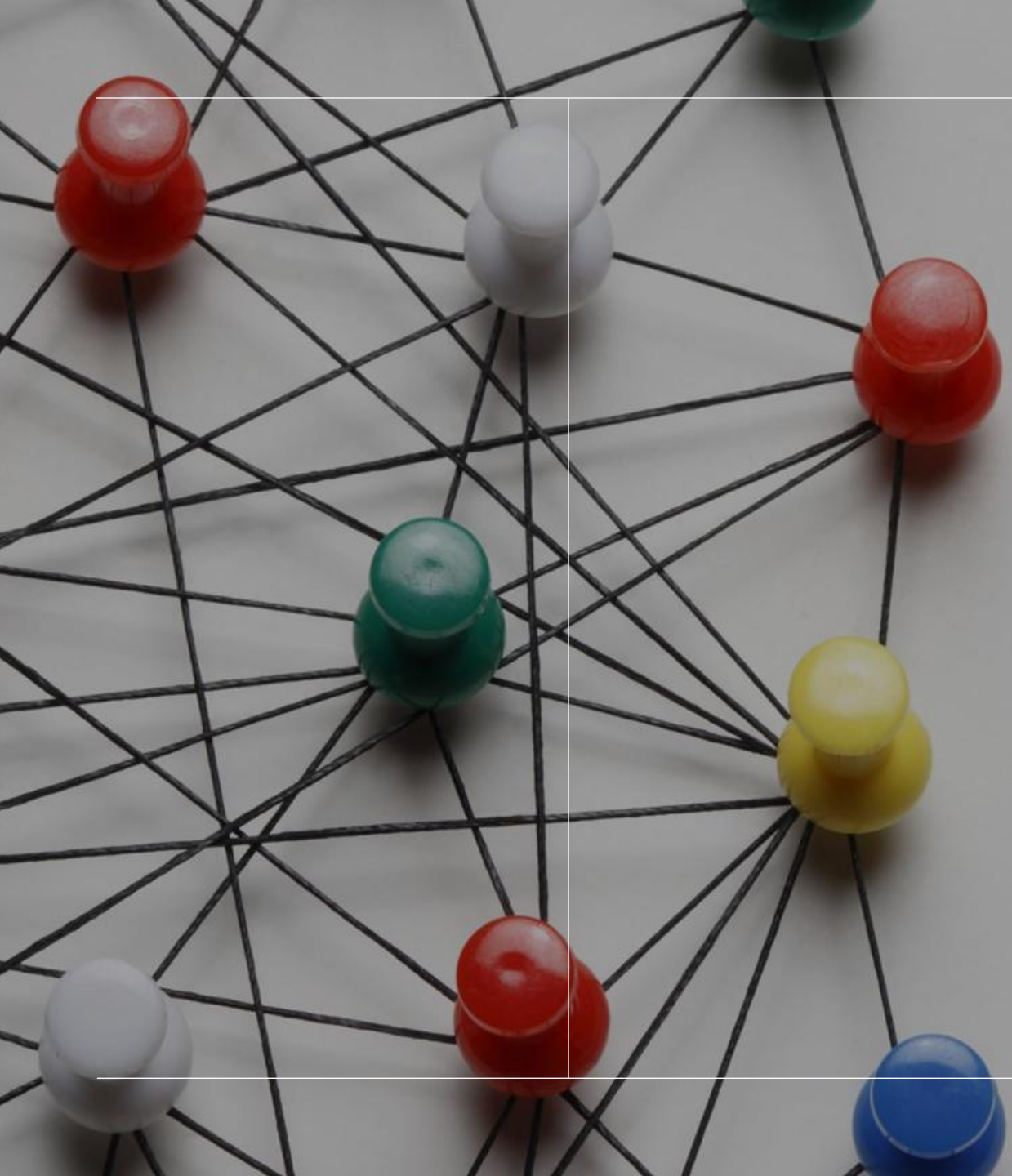
Testes de normalidade

- Avaliam quão diferente a minha distribuição é de uma distribuição normal:
  - Teste de Kolmogorov-Smirnov ou Shapiro-Wilk
  - São muito usados, mas não deveriam
  - Se $p < 0,05$, então a distribuição não é normal
  - Esses testes tem muitos problemas ... amostras grandes sempre dão $p < 0,05$

- Exemplo no JASP!

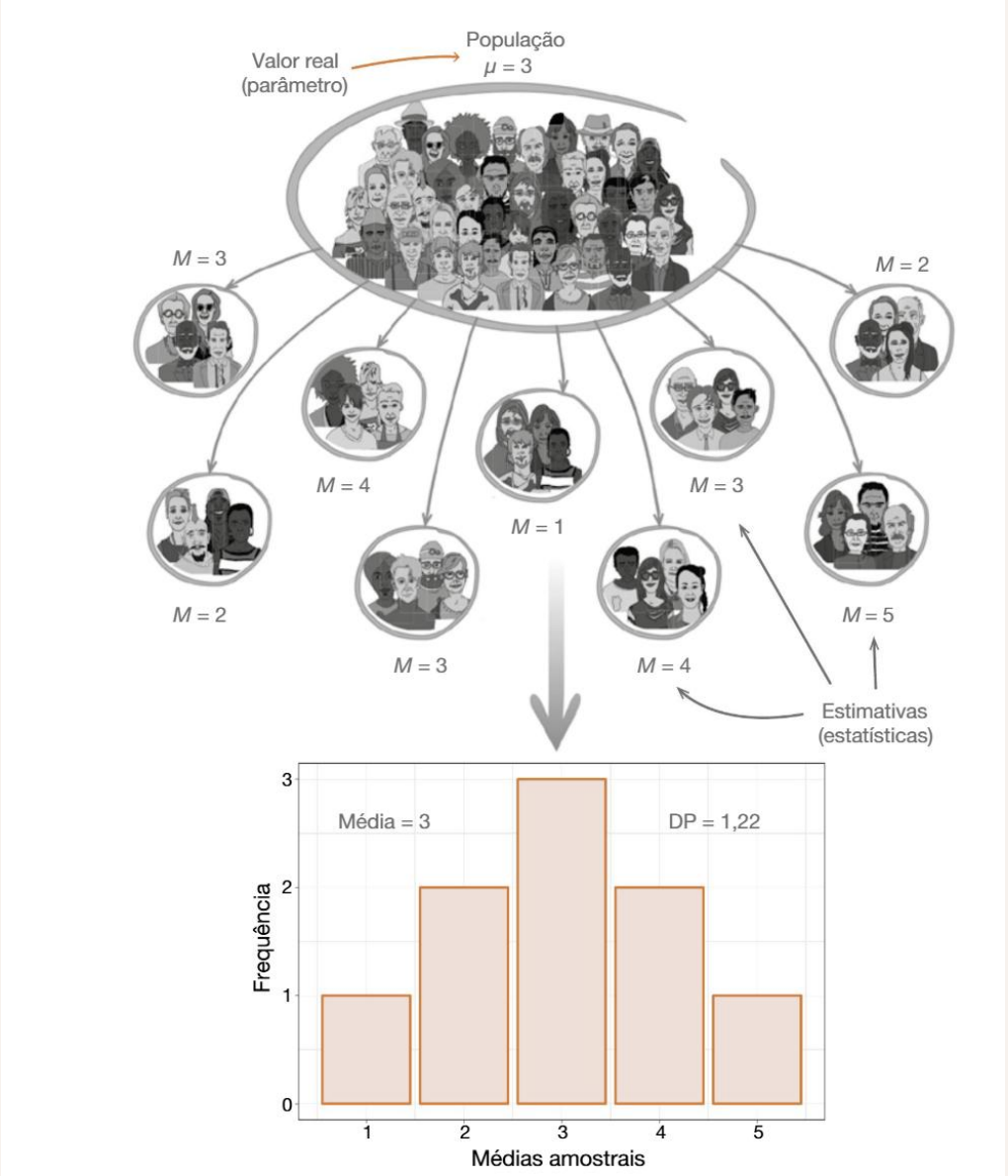
É importante saber se a minha dist. é normal?

- Uma vez era, hoje nem tanto!
- É importante saber se existem grandes desvios da normalidade, mas isso não vai nos impedir de calcular estatísticas.
- Não é a normalidade dos dados que importa, mas da distribuição amostral (em breve).
- Métodos computacionais como *bootstrap* podem contornar a não-normalidade;



Distribuição amostral

Experimento mental



Experimento mental

- Vamos fazer um experimento mental:
  1. Retiramos uma amostra aleatória de **30 casos (n = 30)** de uma certa população
  2. Calculamos uma estimativa nessa amostra (*e.g.*, média)
  3. Repetimos os passos 1 e 2 **muitas vezes (k)**
  4. Montamos um histograma com as estimativas
  5. A distribuição representada pelo histograma com as estimativas é a distribuição amostral

Exemplo

- Vamos ver isso!

Normalidade

- Na medida em que as amostras retiradas se tornam maiores, a distribuição amostral da estatística calculada nessa amostra tende a ter uma forma de sino (normal);

- Isso acontece mesmo se tirarmos as amostras de uma distribuição que não seja normal;

Erro

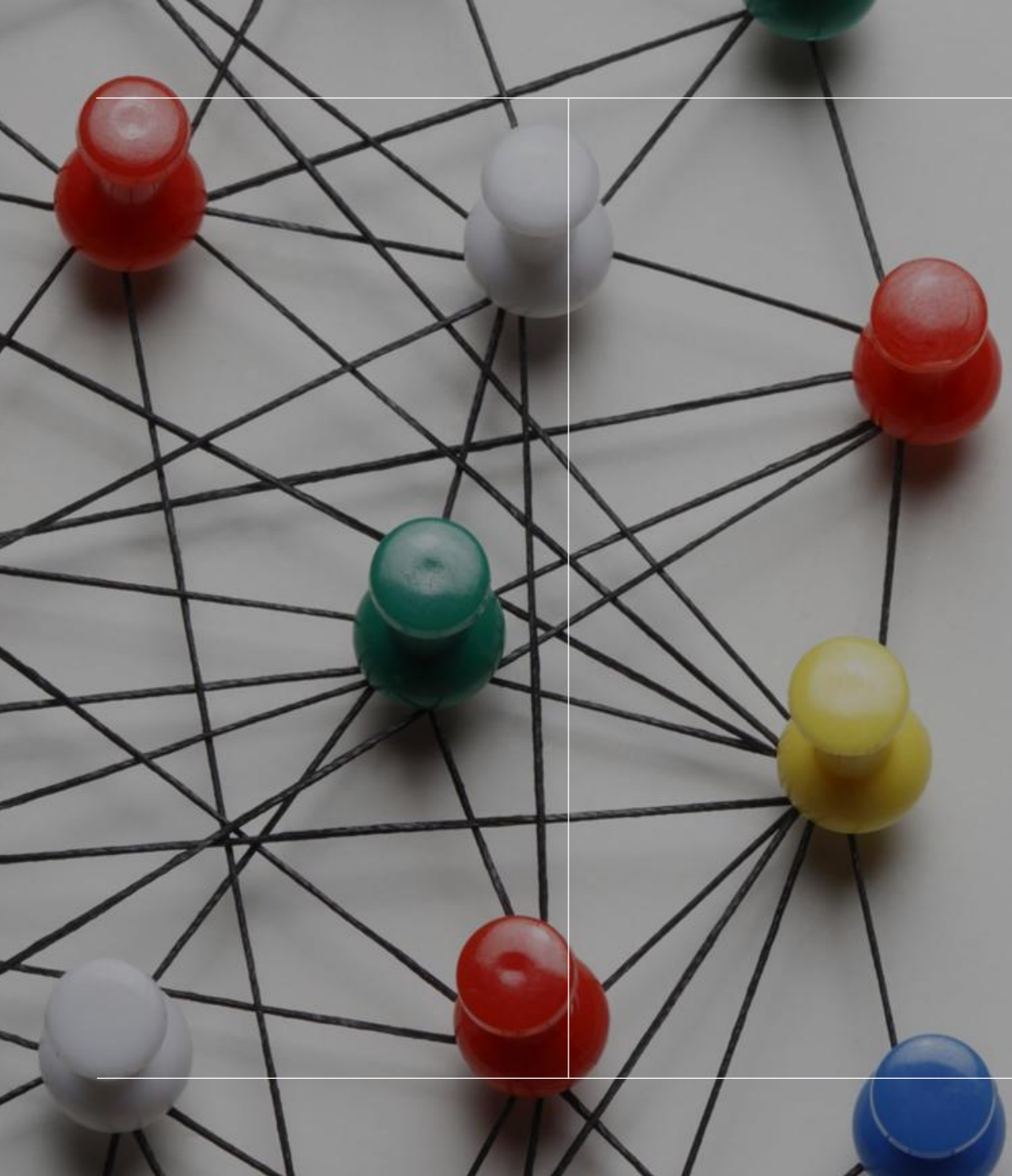
- A média das estatísticas usadas para a distribuição amostral é a melhor estimativa possível do parâmetro populacional.

- O desvio-padrão da distribuição amostral indica a variabilidade entre amostras na estimativa do parâmetro, i.e., o grau médio de erro, chamado de erro-padrão.

É pura Teoria!

- A distribuição amostral é uma ideia, uma teoria. Nunca vamos montar uma de fato (exceto no processo de *bootstrap*).

- Usamos essa ideia para entender que nossa estimativa é uma dentre muitas possíveis (podemos tirar diferentes amostras da mesma população) e, com amostras maiores, sabemos que estamos mais perto do valor verdadeiro do parâmetro.



Teorema do Limite Central

Fundamental!

- O teorema do limite central (TLC) é um dos resultados matemáticos mais importantes na estatística.

- Ele diz que se retirarmos infinitas amostras aleatórias de uma população, cada um de tamanho $n \geq 30$, e calcularmos uma estatística em cada amostra, reunindo essas estatísticas em uma distribuição amostral, então teremos uma distribuição amostral com forma normal, cuja média será uma boa estimativa do parâmetro na população.

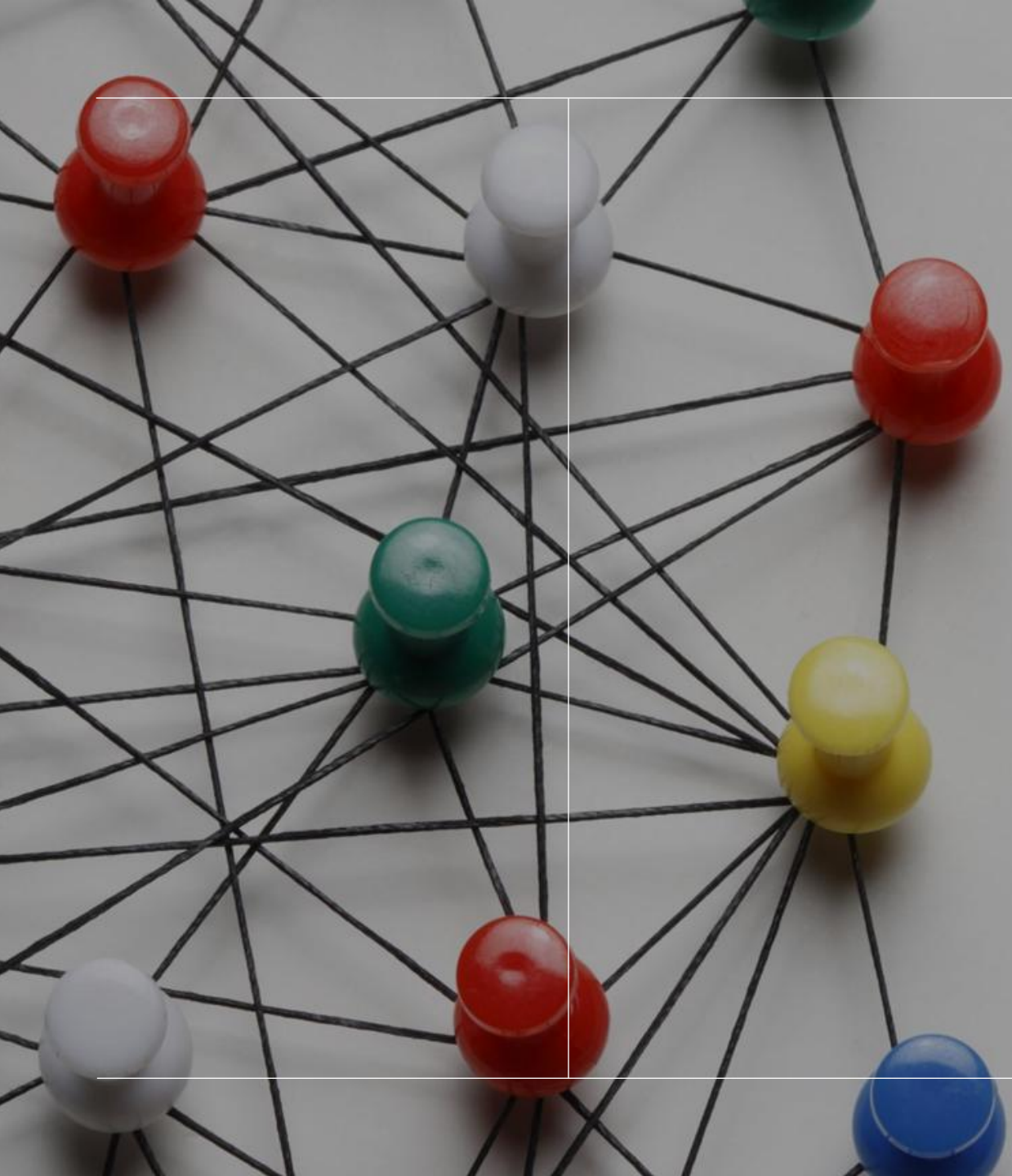
Por partes ...

- Retirar infinitas amostras aleatórias de uma população ($n \geq 30$) **+**
- Calcular uma estatística em cada amostra (*e.g.*, uma média) **+**
- Reunir essas estatísticas em uma distribuição **+**

- **=** distribuição amostral com <u>forma normal</u>
- **=** média da distribuição amostral <u>é igual ao</u> parâmetro na população

Implicação

- Se tivermos uma amostra grande o suficiente, não vamos precisar nos preocupar com a normalidade da distribuição amostral!



Erro-padrão

Definição

- O erro-padrão é o desvio-padrão da distribuição amostral.

- Também é chamado de erro amostral, *i.e.*, a diferença média entre diferentes amostras na estatística calculada. Em outras palavras, a métrica que calculamos varia de amostra para amostra.

- Se fizéssemos o processo de formar a distribuição amostral e calculássemos o desvio-padrão dessa distribuição, teríamos o erro-padrão.

Cálculo

- Para evitar o processo de formar uma distribuição amostral (que seria muito custoso), existe um cálculo direto do erro-padrão, onde:
- EP = Erro-Padrão
- DP = Desvio-padrão da amostra
- N = tamanho da amostra

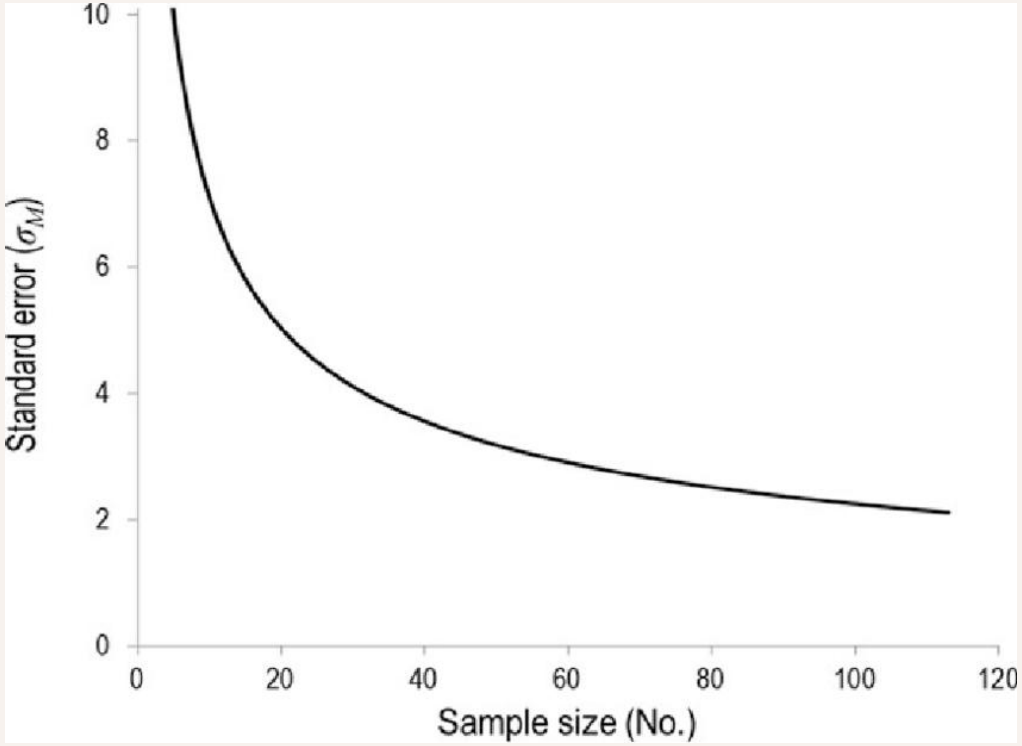
$$EP = \frac{DP}{\sqrt{N}}$$

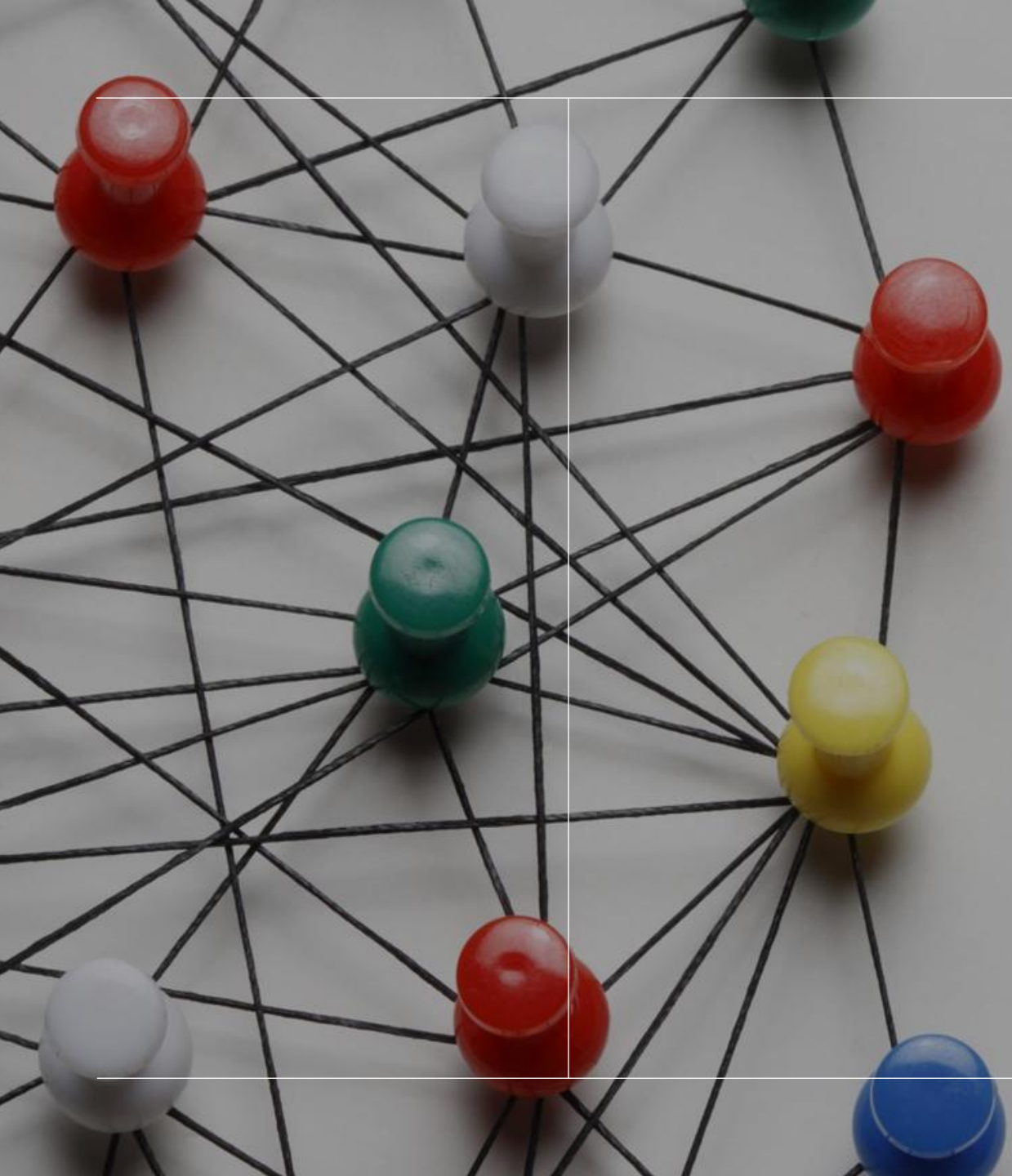
Característica fundamental do EP

- O erro-padrão caí na medida em que o tamanho da amostra aumenta.

- Tamanhos de amostra grande produzirão erros-padrão muito pequenos.

$$EP = \frac{DP}{\sqrt{N}}$$





Intervalo de confiança

Estimativa pontual e intervalar

- A estatística calculado em nossa amostra é uma estimativa pontual do parâmetro da população ... mas, e se tivéssemos como estimativa um intervalo de valores dentro do qual o parâmetro pode estar?

- Essa estimativa intervalar chama-se intervalo de confiança!

Exemplo:

- Vamos imaginar o seguinte caso:
  - Temos uma amostra de 498 estudantes brasileiros que se inscreveram para cursar uma pós-graduação em uma universidade americana, cada um com um escore do TOEFL, com escores que podem variar de 0 a 120.
  - Queremos saber o escore médio do TOEFL na população de estudantes brasileiros que se inscrevem para uma universidade americana
  - Calculamos a média na amostra: $M = 107{,}20$

Estatísticas

- Calculamos a média na amostra: $M = 107{,}20$
- Calculamos o desvio-padrão: $DP = 6{,}09$
- Calculamos o erro-padrão: $\frac{6{,}09}{\sqrt{498}} = 0{,}273$

Sabemos ...

- Que entre -1,96 e 1,96 escores-Z da distribuição normal padrão temos 95% da área sob a curva, ou seja, temos uma probabilidade de 95% de encontrar um escore entre -1,96 e 1,96 se retirarmos um valor qualquer de uma distribuição normal padrão.

- Que a distribuição amostral é uma distribuição normal.

- Logo, podemos usar o erro-padrão para calcular um intervalo de 95% na distribuição amostral em torno da estatística calculada na amostra.

Cálculo do intervalo para média

- Calculamos esse intervalo da seguinte forma:

$$Limite_{Superior} = \bar{x} + 1{,}96 \times EP$$

$$Limite_{inferior} = \bar{x} - 1{,}96 \times EP$$

- No exemplo:
  - **Média:** 107,20
  - **LS:** 107,73
  - **LI:** 106,66

Interpretação frequentista

- Um intervalo de confiança de 95% significa que em 95 de 100 vezes que retiramos uma amostra, calcularmos uma estatística e calcularmos um intervalo, o verdadeiro valor do parâmetro estará dentro do intervalo. Em outras palavras, somente em 5 vezes o valor do parâmetro não estará dentro do intervalo.

- Interpretações erradas:
  - O intervalo de confiança indica uma probabilidade de 95% do intervalo conter o valor do parâmetro
  - O intervalo de confiança indica que só há 5% de chance de erro aleatório

Interpretação frequentista

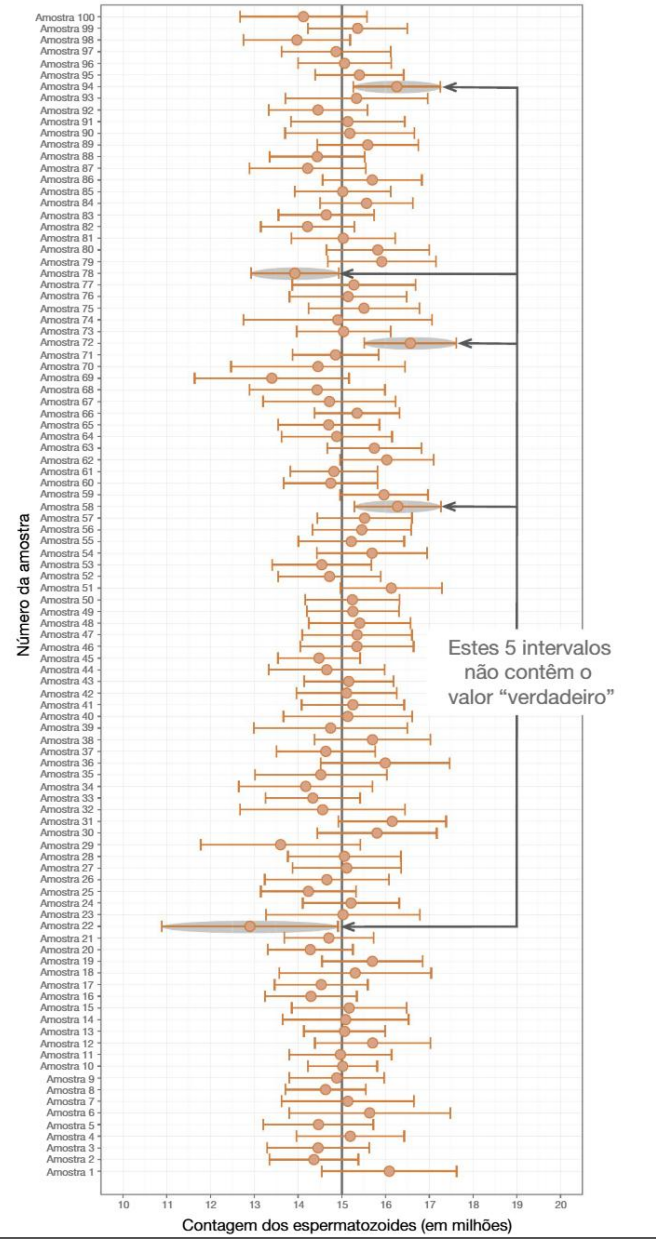


Figura 2.9 Intervalos de confiança das contagens de espermatozoides da codorna japonesa (eixo horizontal) para 100 amostras diferentes (eixo vertical).

Características

- Podemos calcular intervalos de confiança para qualquer estatística. Versão generalizada da fórmula: $IC = b \pm z \times EP$

- Os intervalos são afetados pelo valor de Z. Quanto maior, mais amplo fica o intervalo:
  - 90%: z = 1,68
  - 95%: z = 1,98
  - 99%: z = 2,57
  - 99,9%: z = 3,29

- Os intervalos são afetados pelo erro padrão. Quanto mais erro, mais amplo é o intervalo.

Características

- Quanto mais estreito o IC, mais precisa é nossa estatística.

- Intervalos que cruzam o zero sugerem um grau de erro muito elevado na estimativa, exceto em alguns poucos casos.

Exemplo de uso dos intervalos

- Existe diferença nos escores do TOEFL entre um grupo de estudantes que foi aprovado para uma universidade americana e um grupo que não foi?
  - Média Aprovados = 110, 32
  - Média Não-Aprovados = 102,46

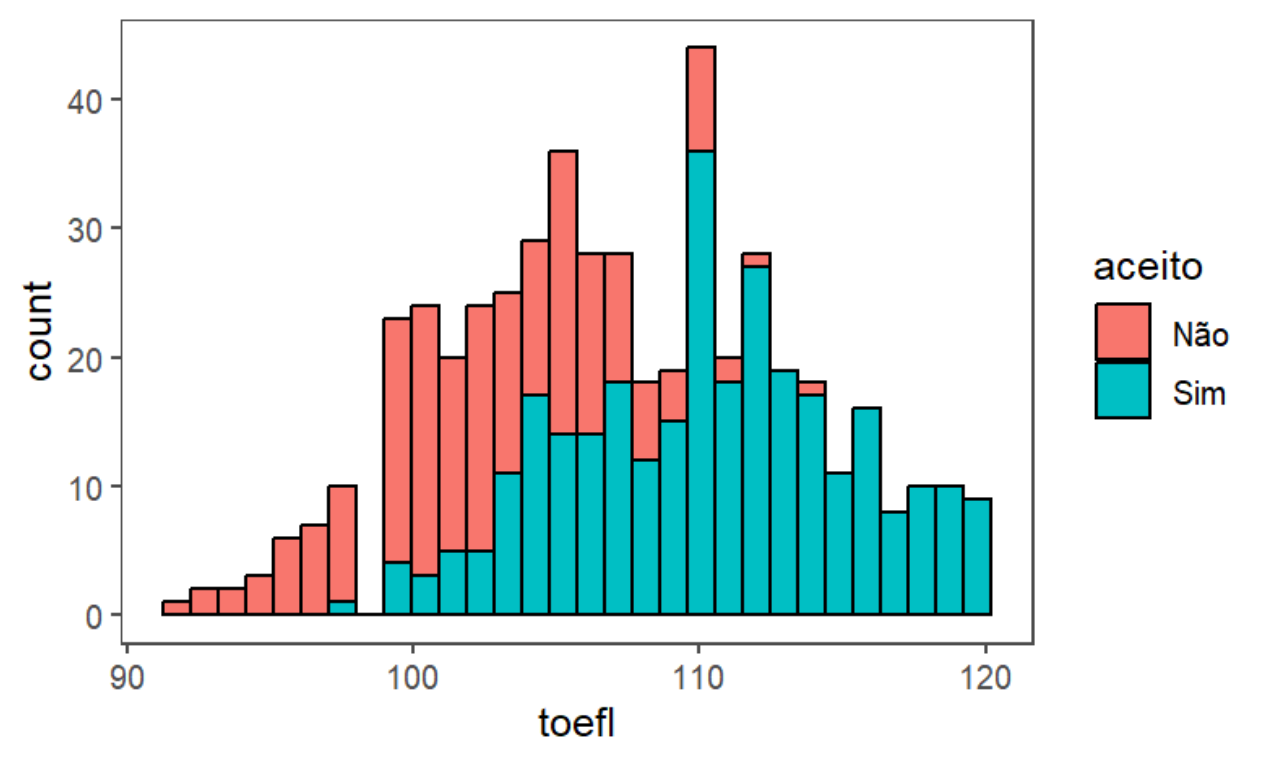
- Existe uma diferença matemática, mas essa é uma diferença estatística? As distribuições dos dois grupos são mais diferentes do que semelhantes entre si?

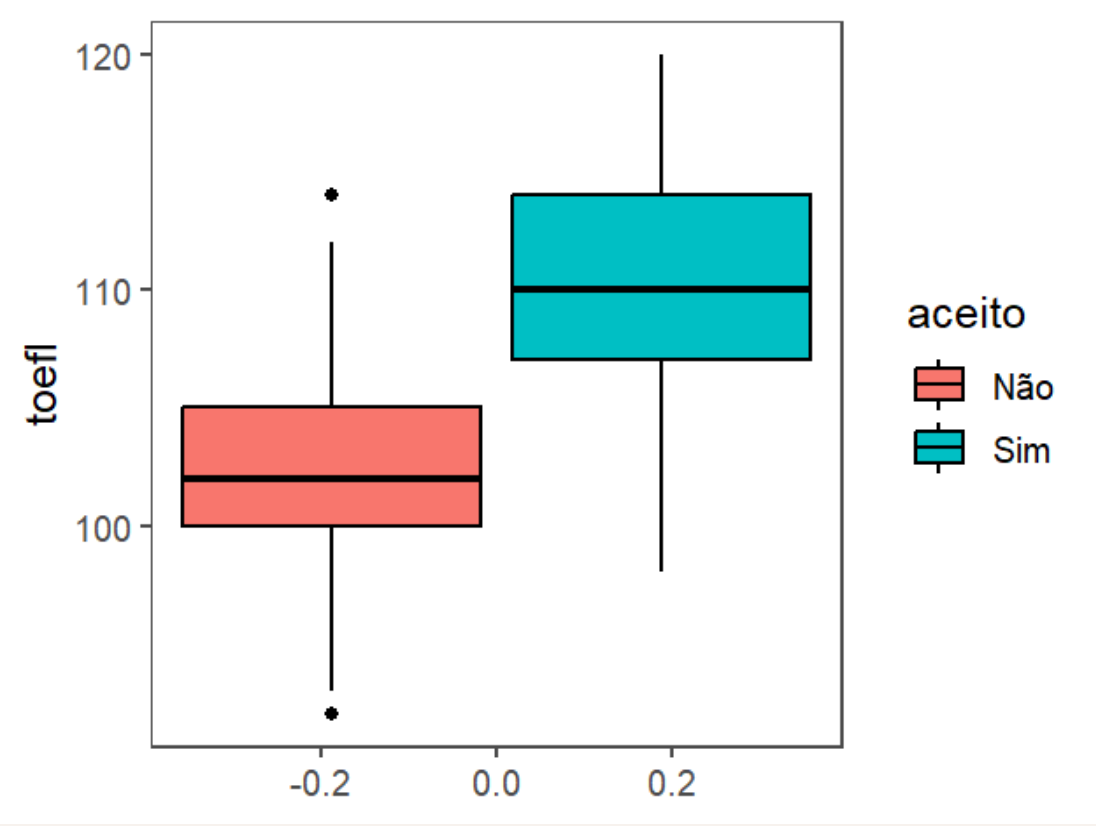
Diferenças

- **Diferença matemática:** dois números, quando subtraídos, tem resultado diferente de 0, i.e., $x - y \neq 0$

- **Diferença estatística:** uma diferença entre dois números que é suficientemente grande ao ponto de não poder ter ocorrido ao acaso, *i.e.*, os números são métricas calculadas de distribuições distintas, vindas de populações distintas. Muitas diferenças podem ocorrer ao acaso quando lidamos com amostras.

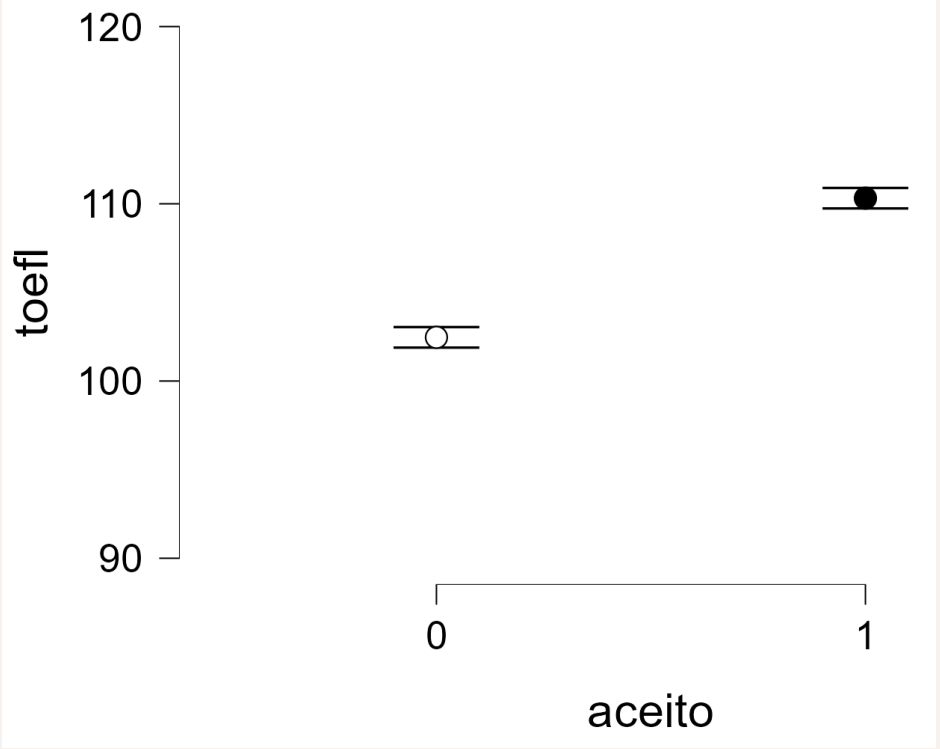
Comparações

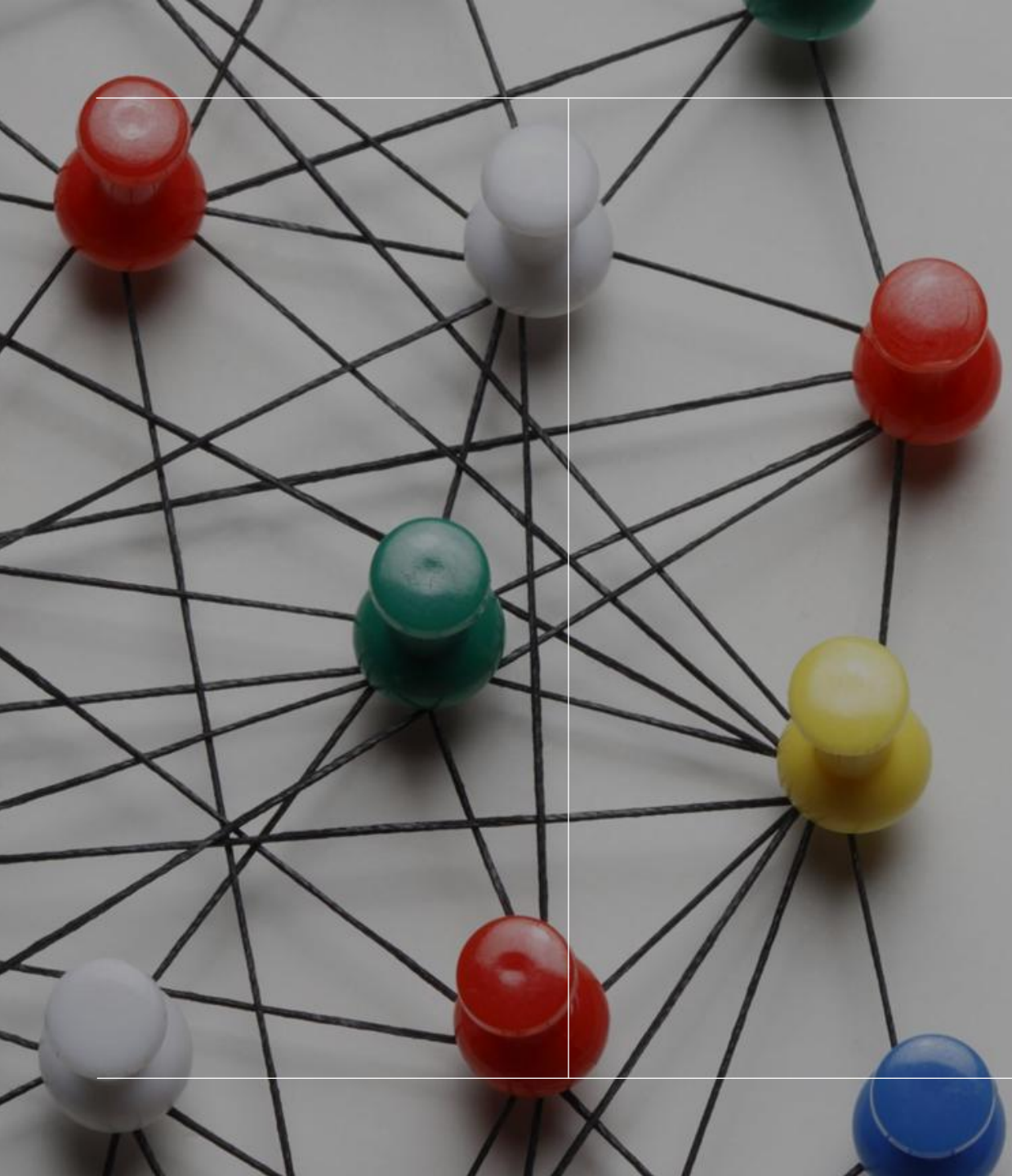




Intervalos

Grupo	Média	IC 95% [LI; LS]
Aceito (n =300)	102,47	101,89; 103,04
Não-aceito (n = 198)	110,32	109,74; 110,89





Testagem de hipóteses

Propósito

- O Teste de Significância da Hipótese Nula (TSHN ou NHST) é um processo desenvolvido para testar hipóteses verificando se nossas estatísticas representam algo real ou se devem ao acaso (variação amostral).

- As estatísticas operacionalizam nossas hipóteses, i.e., trazem elas do mundo teórico para a realidade física/empírica.

Hipóteses

- No TSHN sempre vão existir dois tipos de hipóteses:

  - Hipótese nula $\boldsymbol{h_0}$: representa a nulidade, ausência de efeito, acaso.
  - Exemplo: tomar ômega 3 não está associado ao desempenho cognitivo.

  - Hipótese alternativa $\boldsymbol{h_1}$: representa um efeito.
  - Exemplo: tomar ômega 3 está associado ao desempenho cognitivo.

Hipóteses

- As hipóteses podem ser:

  - **Unilaterais/unicaudais:** especificam uma direção para o efeito, por exemplo, tomar ômega 3 aumenta o desempenho cognitivo;

  - **Bilaterais/bicaudais:** não especificam uma direção para o efeito, por exemplo, tomar ômega 3 está associado ao desempenho cogntivo.

Tipos de hipóteses

*Os testes de hipóteses, em geral, se referem às relações entre as variáveis. Cada "estatística" na tabela é um só valor que representa a relação entre duas variáveis.

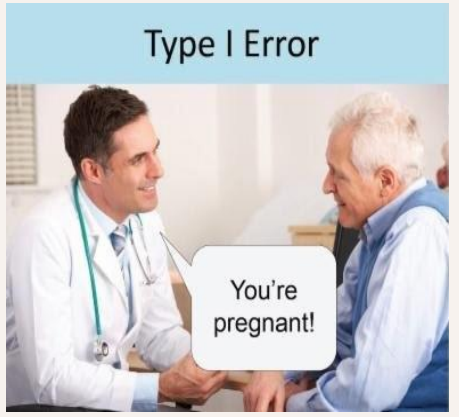
A questão é: esse valor pode ter acontecido ao acaso porque nossa amostra era diferente da população ou ele indica algo real, uma padrão que acontece na população?

Hipótese	Variáveis	Teste	Estatística
Associação de 2 variáveis	Numéricas	Correlação	r
Associação de 2 variáveis	Categóricas	Qui-Quadrado	χ^2
Diferença de médias	1 numérica e 1 categórica	Teste t	t
Predição de valores	n variáveis de qualquer tipo e 1 alvo numérico	Regressão linear	b
Classificação	n variáveis de qualquer tipo e 1 alvo categórico	Regressão logística e classificadores	Odds-ratio etc

Erros

- Existem dois erros que queremos evitar no TSHN:
  - **Erro Tipo I (falso positivo):** dizer que um efeito existe quando ele não existe.
  - **Erro Tipo II (falso negativo):** dizer que um efeito não existe quando ele existe.

- Existem taxas máximas <u>probabilísticas</u> aceitáveis para esses erros:
  - Taxa de erro Tipo I: $\alpha = 0{,}05$ ou $\alpha = 0{,}01$
  - Taxa de erro Tipo II: $\beta = 0{,}20$ ou $\beta = 0{,}05$



Processo do TSHN

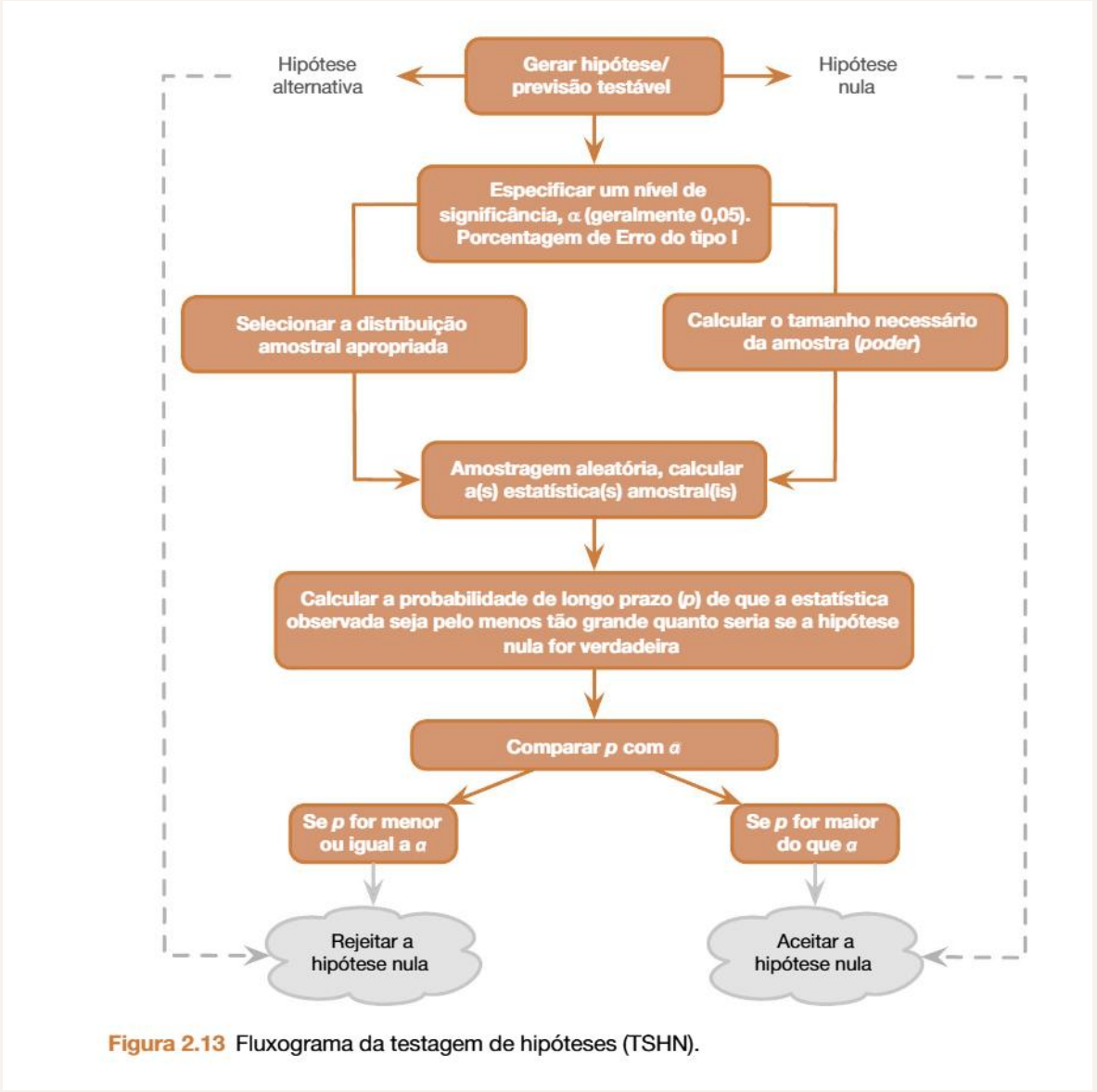


Figura 2.13 Fluxograma da testagem de hipóteses (TSHN).

Exemplo TSHN

- Vamos fazer um teste Z para verificiar se estudantes brasileiros aceitos em universidades americanas tem escores significativamente mais elevados no TOEFL em comparação aos não aceitos.

Hipóteses

- h_0: a diferença nos escores do TOEFL entre estudantes brasileiros que foram aceitos ou não em uma universidade americana não é estatisticamente significativa.

- h_1: a diferença nos escores do TOEFL entre estudantes brasileiros que foram aceitos ou não em uma universidade americana é estatisticamente significativa.
- Vamos assumir que h_0 é verdadeira!

Alfa e distribuição amostral

- Vamos usar uma taxa de erro de tipo II de 0,05, o que significa que admitiremos uma chance de erro de 5 em 100, *i.e.*, em 5 de 100 estudos exatamente iguais a este estaremos errados!

- Vamos supor que nossa estatística, uma diferença entre médias, vem de uma distribuição normal padrão, composta de escores Z.

Estatística

Grupo	Média	Variância	Diferença entre médias	Estatística Z
Aceito (*n* =300)	102,47	25,87	102,47 – 110,32 = -7,85	-18,96
Não-aceito (*n* = 198)	110,32	17,07		

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Valor de p

- A estatística Z é um escore Z da DNP, logo, podemos usar a distribuição normal para saber a probabilidade dela. Usamos a distribuição normal padrão, com média = 0 e DP = 1 porque ela representa a hipótese nula, que supomos ser verdadeira.

- A probabilidade de um valor menor que ou igual a -18,96 é inferior a 0,001, assim, dizemos:

$$P(z \leq -18{,}96) < 0{,}001$$

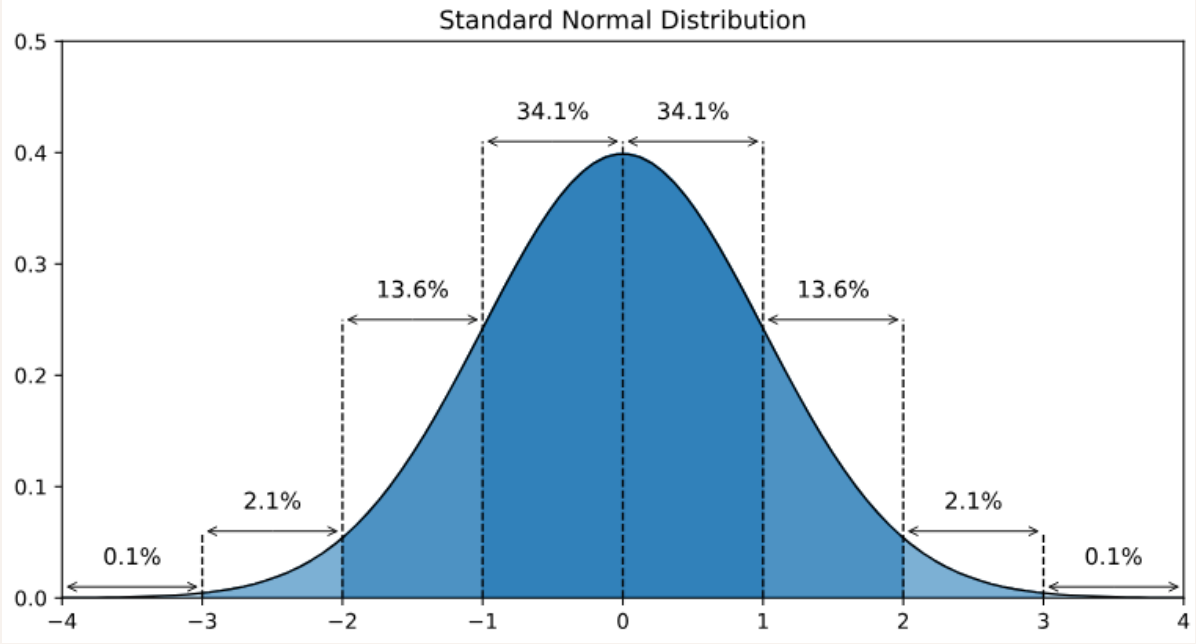
ou,

$$P(z \geq 18{,}96) < 0{,}001$$

Valor de p

$$p < 0{,}001$$

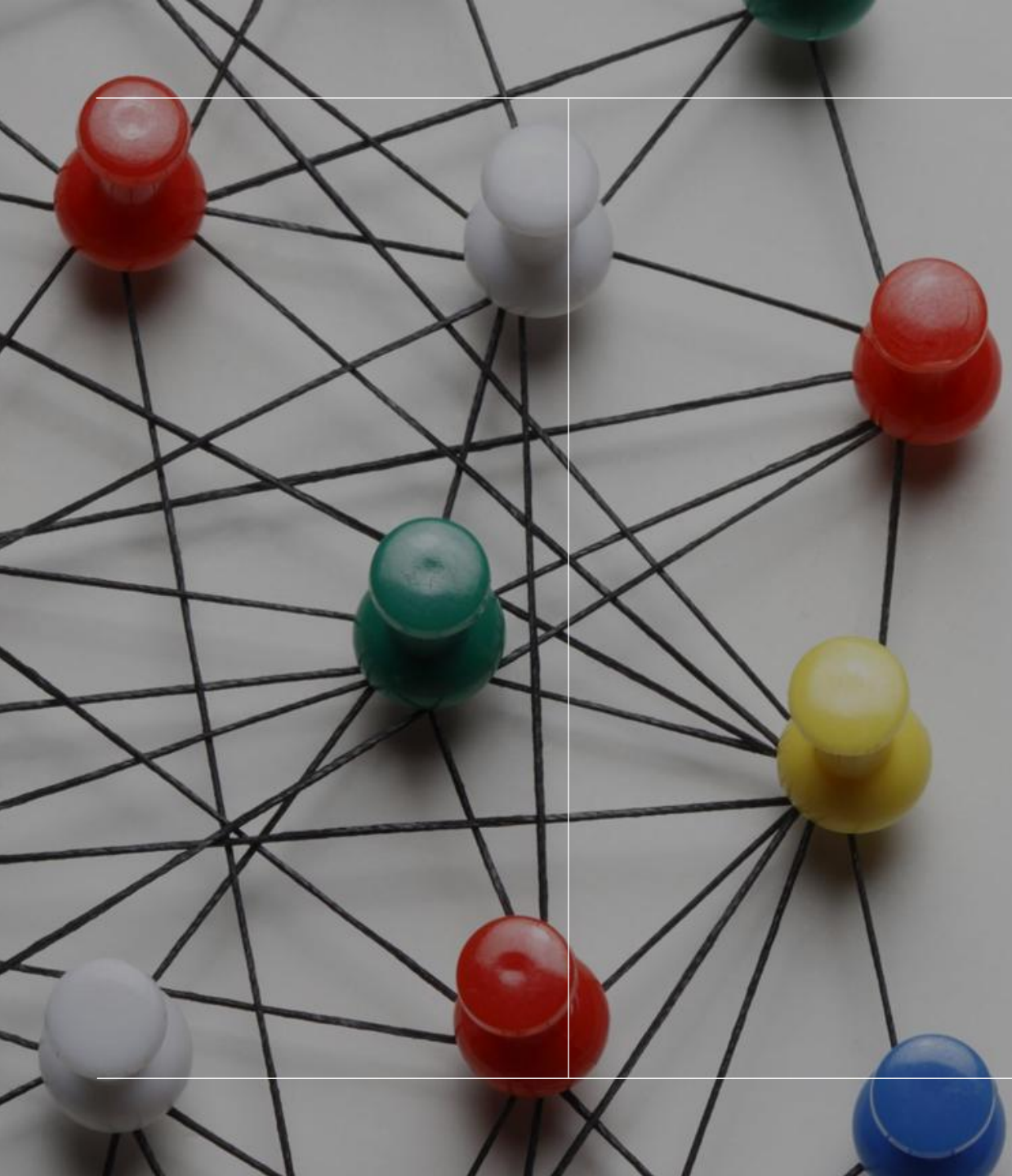
- Significa que a probabilidade de encontrar uma estatística tão grande quanto essa ou maior é muito pequena.
- Ou seja, nosso resultado é muito <u>improvável</u> se supormos que a hipótese nula é verdadeira.



Interpretação

- Como nossa estatística é muito improvável sob a suposição de que a hipótese nula é verdadeira, rejeitamos a hipótese nula e assumimos a hipótese alternativa.

- O TSHN funciona como um juri onde o suspeito é considerado inocente até que as evidências e a ideia de que ele seja inocente se torne ridícula.



bootstrap

Definição

- *Bootstrap* (alça de bota) é um termo usado para descrever uma expressão idiomática no inglês *"By one's own bootstraps"*, que sugere a tarefa impossível de alguém se levantar pelas alças das próprias botas. O termo se refere a um processo independente, autossustentável.
- Um procedimento de *bootstrap* ou reamostragem em estatística consiste em uma técnica de simulação que produz uma distribuição amostral.

Execução

1. Retirar uma amostra de tamanho n de um banco de dados com n casos, com reposição
2. Calcular uma estatística de interesse
3. Repetir os passos 1 e 2 k vezes (k = 500 ou k = 1000)
4. Unir as k estatísticas calculadas em uma distribuição
5. Usar a distribuição das estatísticas como uma distribuição amostral, calculando intervalos de confiança a partir dos percentis

Uso

- Quase todos os testes estatísticos podem ser usados com procedimentos de reamostragem.

- Se os intervalos de confiança não cruzarem o 0, concluímos que as estatísticas são significativas.

- Ver exemplo no JASP!

Para se aprofundar...