T

Marcelo Colonno
Cientista de Dados
Bayer Crop Science

Formulação e
Levantamento de
Hipóteses

AGENDA

- **Bloco 1: Cálculo de Amostragem**
- **Bloco 2: Teste de Hipótese**
  + Intervalo - 10 min
- **Bloco 3: Teste Z-Score e t-Student**
- **Bloco 4: Aplicações Práticas**
- **Dúvidas e reflexões finais**
- **Como foi?**

RECAPITULANDO...

Como definir uma amostra

- A **amostra** pode ser definida dentro do universo como "**os indivíduos que responderam à pesquisa**".

- É através desses indivíduos selecionados que iremos tirar conclusões válidas para todo o grupo em estuado. O critério mais importante é o da **REPRESENTATIVIDADE**.

- **Como definir o tamanho da amostra? (Gancho c/ teste de hipóteses)**

1. **margem de erro**

2. **desvio-padrão**

3. **nível de confiança**

Calculo de Amostragem

Como definir o tamanho da amostra

- Para uma população desconhecida

$$n = \frac{z^2 \,.\, \sigma(1 \, - \sigma)}{(e)^2}$$

- Exemplo: Determine o tamanho da amostra necessário para uma população desconhecida considerando um nível de confiança de 90%, um desvio padrão de 50% e uma margem de erro de 3%

$z = 1{,}645$

$e = 0{,}03$

$\sigma = 0{,}5$

$$n = \frac{z^2 \,.\, \sigma(1 \, - \sigma)}{(e)^2} = \frac{1{,}645^2 \,.\, 0{,}5(1 \, - 0{,}5)}{(0{,}03)^2} = 756{,}22$$

Como definir o tamanho da amostra

- Para uma população de tamanho conhecido

$$n = \frac{\frac{z^2 . \sigma(1 - \sigma)}{(e)^2}}{1 + \left(\frac{z^2 . \sigma(1 - \sigma)}{(e)^2 . N}\right)}$$

- Exemplo*:* Determine o tamanho ideal de amostra para uma população de 425 pessoas. Utilize um intervalo de confiança de 99%, um desvio padrão de 50% e uma margem de erro de 5%.

$N = 425$

$z = 2{,}58$

$e = 0{,}05$

$\sigma = 0{,}5$

$$n = \frac{\frac{z^2 . \sigma(1 - \sigma)}{(e)^2}}{1 + \left(\frac{z^2 . \sigma(1 - \sigma)}{(e)^2 . N}\right)} = 259{,}39$$

Fórmula de Slovin

- é uma equação geral que é utilizada quando precisamos estimar uma população mas não temos ideia do comportamento dela. A formula é descrita como:

$$n = \frac{N}{1 + Ne^2}$$

Obs: essa é a formula mais imprecisa e menos recomendável de todas. Ela só deve ser utilizada em casos onde é impossível determinar um desvio padrão e um nível de confiança apropriados (o que também impede a definição de um escore z)

- Exemplo*:* Calcule o tamanho necessário de amostra para uma população de 240 considerando uma margem de erro de 4%:

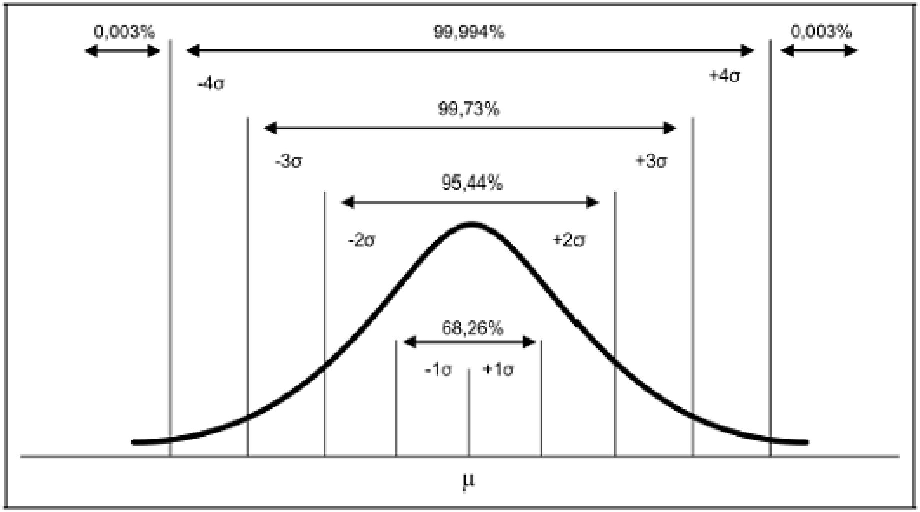
$$\mathrm{N} = 240$$

$$e = 0{,}04$$

$$n = \frac{N}{1 + Ne^2} = 173{,}41$$

Margem de Erro

- A margem de erro é uma porcentagem que indica a proximidade dos resultados obtidos da amostra do valor real para a população total do estudo



- Margens de erro menores oferecem resultados mais precisos, mas também exigem amostras maiores

- Na apresentação dos resultados da pesquisa, a margem de erro geralmente é mostrada em pontos percentuais. Por exemplo: "35% das pessoas concordam com a *opção A*, com uma margem de erro de dois pontos percentuais para mais ou para menos".

Como definir o tamanho da amostra

- Regra de bolso que fornece o número de questionários para a margem de erro desejada considerando X% de margem de erro.

$$n = \frac{1}{(e)^2}$$

- Qual o número de pessoas entrevistadas para a pesquisa abaixo?

$$n = \frac{1}{(0{,}02)^2} = 2500$$

Desvio-Padrão

- Padrão de desvio de uma série de números em relação à sua média

Desvio-padrão de uma população

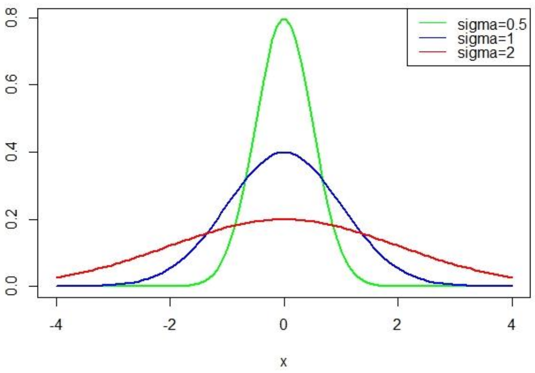
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Desvio-padrão de uma amostra

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- O desvio padrão é o quadrado da variância

- Variância mede a dispersão de um conjunto de pontos de dados ao redor da media.



Calculando o Z-Score

- Z-Score: "valor padronizado“ constante que indica o número de desvios padrão acima ou abaixo da média da população.
- Como os níveis de confiança relativamente padronizados, a maioria dos pesquisadores simplesmente memorizará o escore Z a ser utilizado para os principais níveis de confiança:

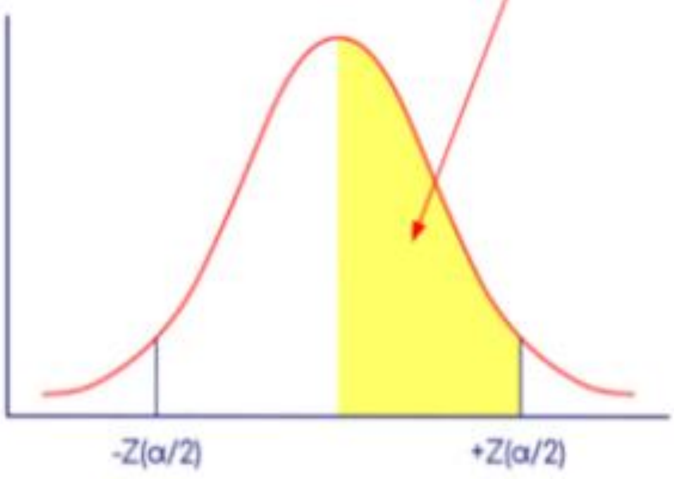
Como calcular o Z-Score

$$Z = \frac{x_h - \bar{x}}{\left({}^{\sigma}/_{\sqrt{n}} \right)}$$

Margem de Erro (α)		Nível de confiança $(1-\alpha)$		Z
	em %		em %	
0,01	1%	0,99	99%	2,576
0,02	2%	0,98	98%	2,33
0,05	5%	0,95	95%	1,96
0,1	10%	0,9	90%	1,645

Encontrando o Z-Score

$$(1-\alpha)/2 = 0{,}95/2 = 0{,}475$$



$$\boxed{Z_{0,025} = 1{,}96}$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803

Teste de Hipótese

Teste de Hipótese

Definição: Os teste de hipótese é um método científico para a tomada de decisão baseado em dados.

Como conduzir um Teste de Hipótese

1. **Formular** as hipóteses.
2. **Identificar** o teste apropriado
3. Escolha um **nível de significância**, α
4. Estabelecer a **regra de decisão**
5. Calcular o valor da **estatística de prova**
6. **Tomar a decisão** com base no resultado

1. Formular a Hipótese

- Uma hipótese é **uma ideia que pode ser testada**
- Existem duas hipóteses: a **hipótese nula**, H_0 e a **hipótese alternativa**, H_1 ou H_A.
- A hipótese nula é a única a ser **testada** e a alternativa é todo o resto.
- Nas estatísticas, a hipótese nula é a afirmação que estamos tentando **rejeitar**.
  - ✓ *Pense na **hipótese nula** como o status quo e a alternativa como a mudança ou inovação que desafia esse status quo.*
  - ✓ *A hipótese nula é o **estado atual das coisas**, enquanto a alternativa é a nossa opinião pessoal.*
- O conceito da **hipótese nula** é semelhante a: inocente até que se prove o contrário.

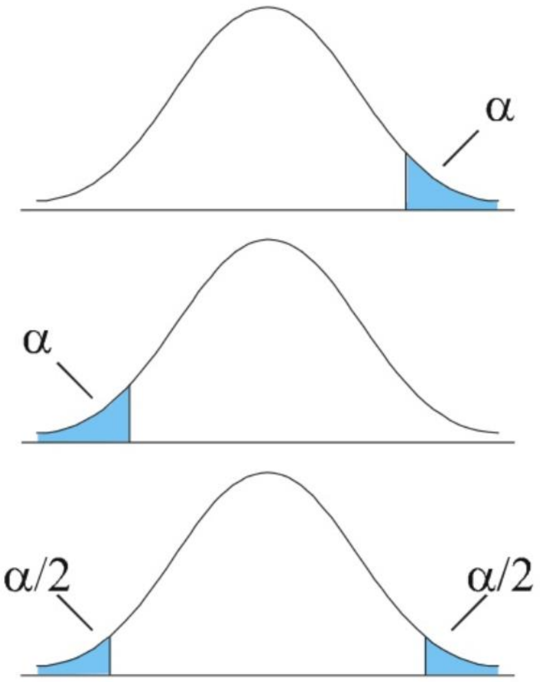
1. Formular a Hipótese

Hipótese **unilateral ou uni-caudal**

- Direita
  - ✓ $H_0: \bar{x} = \mu_0$
  - ✓ $H_1: \bar{x} > \mu_0$
- Esquerda
  - ✓ $H_0: \bar{x} = \mu_0$
  - ✓ $H_0: \bar{x} < \mu_0$

Hipótese **bilateral ou bi-caudal**

- $H_0: \bar{x} = \mu_0$
- $H_0: \bar{x} \neq \mu_0$



2. Identificar o Teste Apropriado

- **Exemplos dos testes**
  - ✓ **Teste Z** (teste para taxa, proporção)
  - ✓ **Teste t** de Student (teste para media)
  - ✓ **Teste qui-quadrado** de Pearson (teste para proporção)
  - ✓ **Teste F** (teste para variância)
  - ✓ **Teste ANOVA** (teste para variância)

3. Escolher o nível de Significância

- O nível de significância, α é a **probabilidade de rejeitar a hipótese nula** quando ela é verdadeira.
- A **escolha do alfa**, α depende da situação e do campo de estudo, mas o valor mais usado é 0,05.
- α mais baixos indicam que você precisa de evidências mais fortes antes de rejeitar a **hipótese nula**.

Conclusão do teste (baseada na amostra)	"Realidade"	
	H_0 verdadeira	H_0 falsa
Rejeitar H_0	erro tipo I (α) False Positive (FP)	decisão correta True Negative(TN)
Não rejeitar H_0	decisão correta True Positive (TP)	erro tipo II (β) False Negative (FN)

3. Escolher o nível de Significância

- O nível de significância, α é a **probabilidade de rejeitar a hipótese nula** quando ela é verdadeira.

- A **escolha do alfa**, α depende da situação e do campo de estudo, mas o valor mais usado é 0,05.

- α mais baixos indicam que você precisa de evidências mais fortes antes de rejeitar a **hipótese nula**.

Conclusão do teste (baseado na amostra)	Realidade	
	H_o *verdadeira*	H_o *falsa*
Rejeitar H_o	erro tipo I False Positive (FP)	Decisão Correta True Negative (TN)
Não Rejeitar H_o	Decisão Correta True Positive (TP)	erro tipo II False Negative (FN)

4. Estabelecer a Regra de Decisão

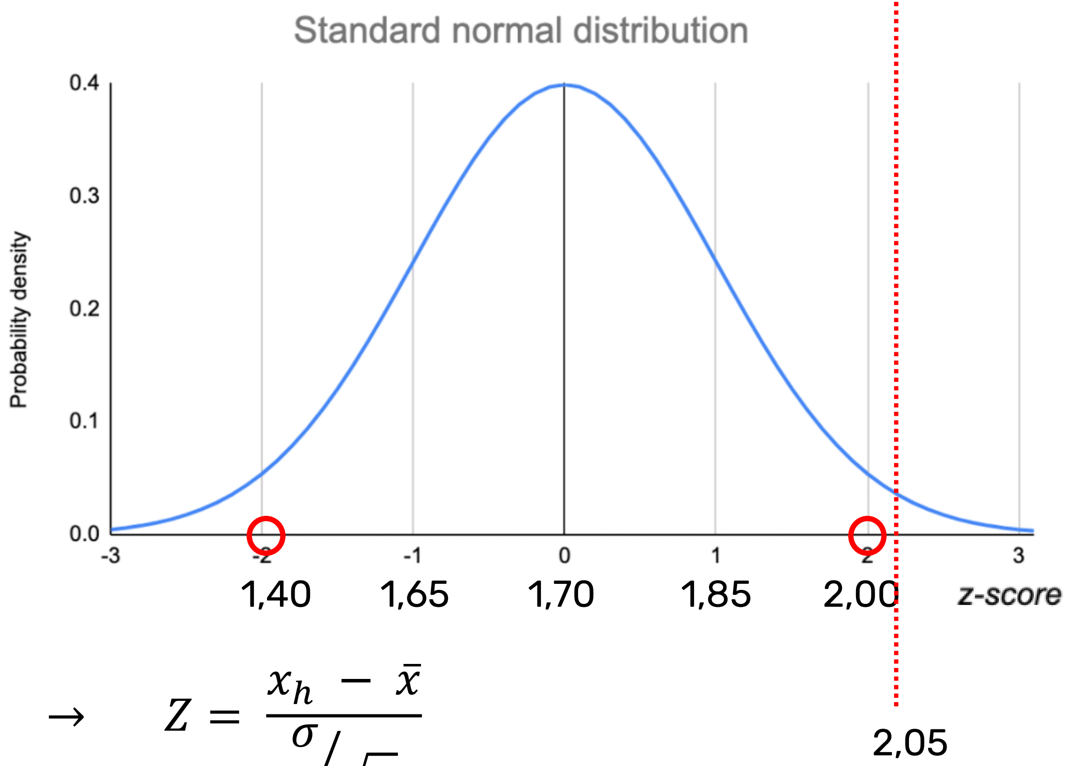
- Dado o **nível de significância**, teremos valores críticos que separam a região de não rejeição da região de rejeição.
- Determinar os **valores críticos** baseados na α escolhida e o teste.

5. Calcular a estatística de prova

- Calcular o valor amostral da estatística do teste escolhido
  - ✓ Teste Z: valor **Z-Score**
  - ✓ Teste t: de Student; valor **t-Statistic**
  - ✓ Teste qui-quadrado: **Valor Qui-quadrado** de Pearson
  - ✓ Teste F: **valor F**

6. Tomar a Decisão

- A altura média da população é 1,70 m
- **Margem de erro = 0,15**
- Hipótese nula: uma pessoa de 2,05 metros tem **altura acima da média**, com 95% de significância
- Aceita ou rejeita a hipótese?
- **Qual é o intervalo de confiança?**



$$margem\ erro = x_h - \bar{x} = \frac{Z\,.\,\sigma}{\sqrt{n}} \quad \rightarrow \quad Z = \frac{x_h - \bar{x}}{\sigma/\sqrt{n}}$$

INTERVALO 10 MIN

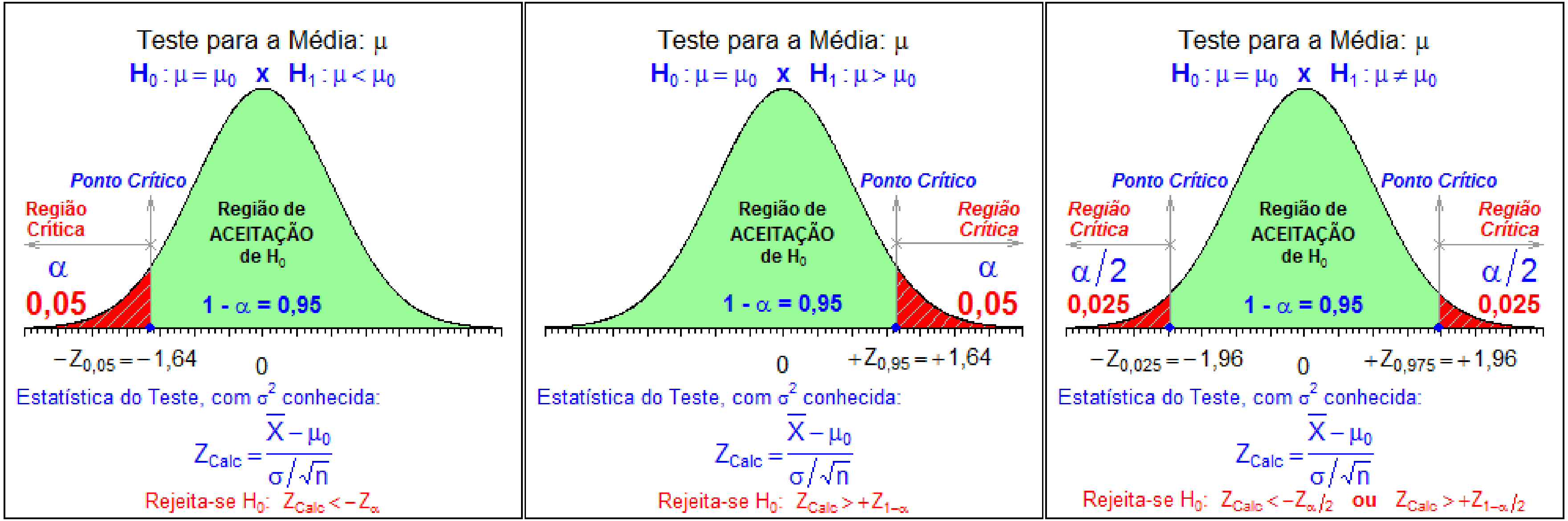
APROVEITE PARA:

- Fazer anotações do que viu até agora (aprendizados, insights, dúvidas)
- Levantar-se, esticar os braços e as pernas, relaxar por mais tempo
- Comer algo para voltar com energia renovada
- Ir ao toalete

Teste Z-Score

Teste Z-Score

- **Z-Score é o valor crítico** porque ele é o número que divide a distribuição em duas regiões: a região de falha ao rejeitar a hipótese nula e a região de rejeição da hipótese nula



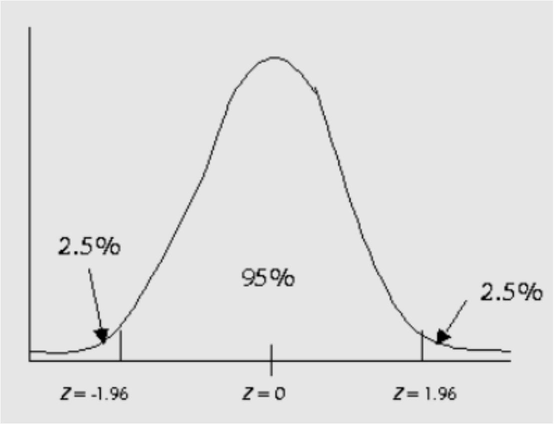
Teste Z-Score

- **Z-Score** é definido matematicamente como a quantidade de desvios-padrão médio um valor está distante da sua média (expectativa matemática)

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- **p-valor** é a de que ele é a probabilidade de obter um efeito tão extremo quanto o observado, considerando que a hipótese nula é verdadeira

- O valor de **alfa** é uma constante que representa uma área dentro da distribuição normal. Essa área vai de uma das extremidades até o valor definido



Teste Z-Score: Exemplo

Uma indústria produz discos de metal, segundo o vendedor, os diâmetros dos discos são de 10 cm, com desvio padrão de 0,13 cm. O comprador selecionou 30 discos aleatoriamente para confirmar os diâmetros e obteve média 9,95 cm. O comprador deseja confirmar os diâmetros para uma α=0,05.

$$H_0: \bar{x} = 10\ cm\ (os\ discos\ tem\ diâmetro\ igual\ a\ 10\ cm)$$

$$H_1: \bar{x} \neq 10\ cm\ (os\ discos\ tem\ diämetro\ diferente\ de\ 10\ cm)$$

$$\sigma = 0{,}13$$

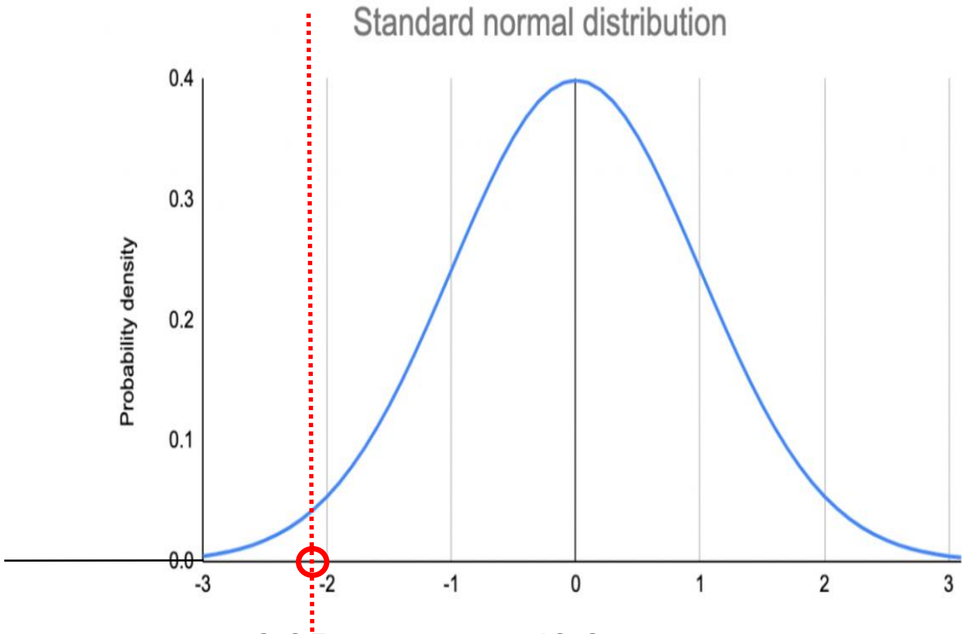
$$n = 30$$

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{9{,}95 - 10{,}0}{0{,}13 / \sqrt{30}}$$

$$\boldsymbol{Z = -2,10}$$

Teste Z-Score: Resultado

Nível de significância (α)		Nível de confiança $(1-\alpha)$		Z
	em %		em %	
0,01	1%	0,99	99%	2,576
0,02	2%	0,98	98%	2,33
0,05	5%	0,95	95%	1,96
0,1	10%	0,9	90%	1,645



9,95 cm
Z = -2,10

10,0 cm

Intervalo de Confiança: α = 95%

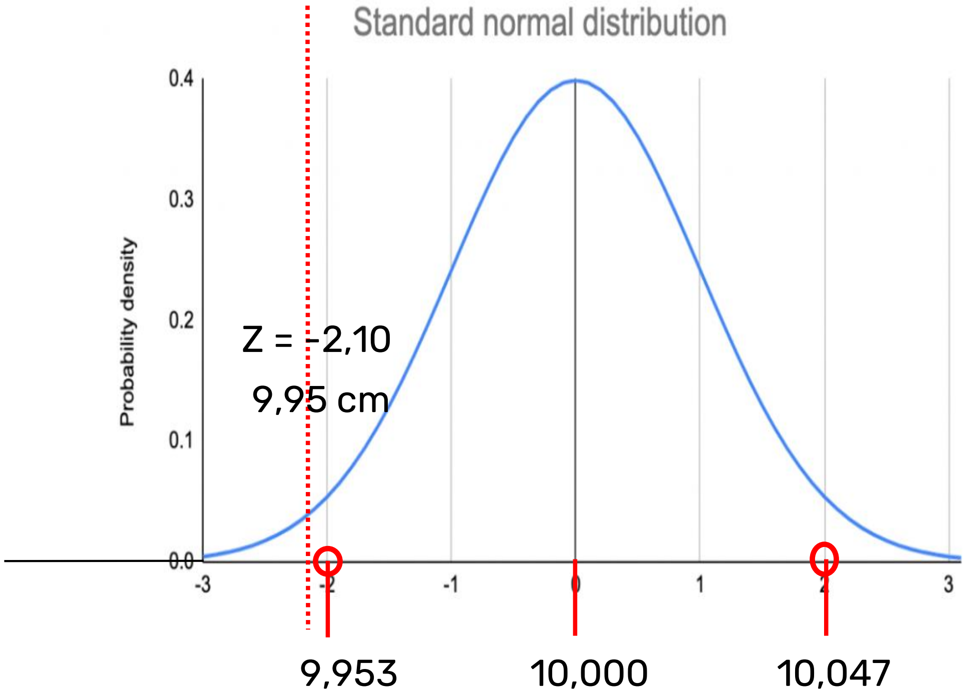
$$margem\ erro = \frac{Z\,.\,\sigma}{\sqrt{n}}$$

$\alpha = 0{,}95 \ \rightarrow \ Z = 1{,}96$

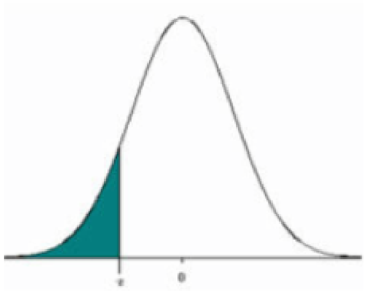
$$margem\ erro = \frac{1{,}96\,.\,0{,}13}{\sqrt{30}}$$

$$margem\ erro = 0{,}0465\ cm$$

$$IC = (9{,}953,\ 10{,}047)$$



p-valor: Tabela Z-Score



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

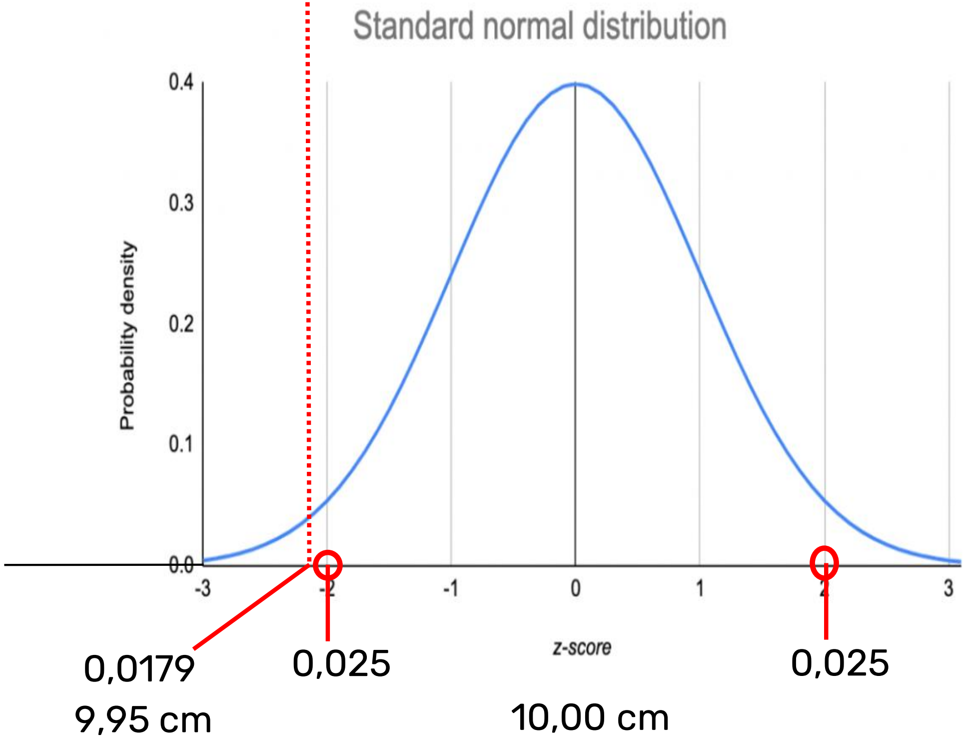
Intervalo de Confiança: α = 95%

$$p_{value} = 0{,}0179 \ \ (unicaudal)$$

$$p_{value} = 0{,}00358 \ \ (bicaudal)$$

$$p_{value} < 0{,}05$$

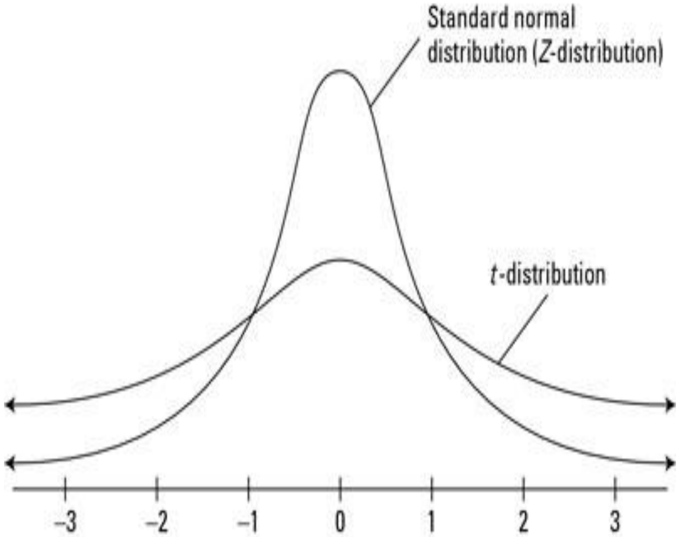
$$9{,}95 \ e \ 10{,}00 \ são$$
$$estatisticamente \ distintos$$



Teste t-Student

Teste t-Student

- O teste t mostra a importância das **diferenças entre os grupos**; Em outras palavras, permite que você saiba se essas diferenças (medidas em médias / médias) poderiam ter acontecido por acaso.
- O teste mais comum para **comparar amostras**
- Todo valor p tem um **teste t** respectivo, só que **um pouco maior**



Teste t-Student

- **Quando utilizar o teste Z-Score ou t-Student**:
  - ✓ **Z-Score**: quando **o desvio-padrão é conhecido** (população)
  - ✓ **t-Student**: quando o **desvio-padrão é desconhecido** (calcula-se o desvio-padrão para a amostra) Teste t de uma amostra: testa a média de um único grupo em relação a uma média conhecida.
- Existem **três tipos** de teste t:
  - ✓ Teste t de **uma amostra**: testa a média de um **único grupo** em relação a uma média conhecida.
  - ✓ Teste t de **amostras independentes**: compara as médias de **dois grupos**
  - ✓ Teste t de **amostra pareada**: compara médias do **mesmo grupo em momentos diferentes** (por exemplo, um ano de intervalo).

Teste t-Student

- Teste t de **uma amostra** $t = \frac{x_h - \bar{x}}{\left(s/\sqrt{n}\right)}$

- Teste t de **amostra pareada** $t = \frac{x_h - \bar{x}}{\left(s_d/\sqrt{n}\right)}$

- Teste t de **amostras independentes** $t = \frac{\bar{x}_2 - \bar{x}_1}{\left(\frac{s_1}{\sqrt{n_1}}\right) + \left(\frac{s_2}{\sqrt{n_2}}\right)}$

Teste t-Student: Exemplo

Um engenheiro de produção quer testar se a altura média de uma haste está próxima do valor nominal de **1055mm**. Uma amostra de **20 hastes** foi analisada as medidas obtidas são dadas a seguir.

903,88	**915,38**	**993,45**	**941,83**	**1098,04**	**1097,79**	**860,41**	**1086,98**	**1020,7**	**1214,08**
1036,92	**1014,53**	**1120,19**	**936,78**	**1011,26**	**934,52**	**1039,19**	**1144,94**	**950,38**	**1066,12**

Estabelecer as hipóteses:

$H_0 \colon x_h = \bar{x}$ Vs. $H_1 \colon x_h \neq \bar{x}$

Fixar o nível de significância:

$\alpha = 0{,}05$

Calcular a média:

$\bar{x} = 1019{,}37$

Calcular o Desvio-Padrão:

$\sigma = 89{,}06$

Calcular o t-Student

$$t = \frac{x_h - \bar{x}}{\left({}^{s}/_{\sqrt{n}}\right)}$$

$$\boxed{t = 1,744}$$

Mas qual o valor de referência do t-Student?

df/alpha	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.706200	31.820520	63.656740	636.619200
2	0.288675	0.816497	1.885618	2.919986	4.302650	6.964560	9.924840	31.599100
3	0.276671	0.764892	1.637744	2.353363	3.182450	4.540700	5.840910	12.924000
4	0.270722	0.740697	1.533206	2.131847	2.776450	3.746950	4.604090	8.610300
5	0.267181	0.726687	1.475884	2.015048	2.570580	3.364930	4.032140	6.868800
6	0.264835	0.717558	1.439756	1.943180	2.446910	3.142670	3.707430	5.958800
7	0.263167	0.711142	1.414924	1.894579	2.364620	2.997950	3.499480	5.407900
8	0.261921	0.706387	1.396815	1.859548	2.306000	2.896460	3.355390	5.041300
9	0.260955	0.702722	1.383029	1.833113	2.262160	2.821440	3.249840	4.780900
10	0.260185	0.699812	1.372184	1.812461	2.228140	2.763770	3.169270	4.586900
11	0.259556	0.697445	1.363430	1.795885	2.200990	2.718080	3.105810	4.437000
12	0.259033	0.695483	1.356217	1.782288	2.178810	2.681000	3.054540	4.317800
13	0.258591	0.693829	1.350171	1.770933	2.160370	2.650310	3.012280	4.220800
14	0.258213	0.692417	1.345030	1.761310	2.144790	2.624490	2.976840	4.140500
15	0.257885	0.691197	1.340606	1.753050	2.131450	2.602480	2.946710	4.072800
16	0.257599	0.690132	1.336757	1.745884	2.119910	2.583490	2.920780	4.015000
17	0.257347	0.689195	1.333379	1.739607	2.109820	2.566930	2.898230	3.965100
18	0.257123	0.688364	1.330391	1.734064	2.100920	2.552380	2.878440	3.921600
19	0.256923	0.687621	1.327728	1.729133	2.093020	2.539480	2.860930	3.883400
20	0.256743	0.686954	1.325341	1.724718	2.085960	2.527980	2.845340	3.849500
21	0.256580	0.686352	1.323188	1.720743	2.079610	2.517650	2.831360	3.819300
22	0.256432	0.685805	1.321237	1.717144	2.073870	2.508320	2.818760	3.792100
23	0.256297	0.685306	1.319460	1.713872	2.068660	2.499870	2.807340	3.767600
24	0.256173	0.684850	1.317836	1.710882	2.063900	2.492160	2.796940	3.745400
25	0.256060	0.684430	1.316345	1.708141	2.059540	2.485110	2.787440	3.725100
26	0.255955	0.684043	1.314972	1.705618	2.055530	2.478630	2.778710	3.706600
27	0.255858	0.683685	1.313703	1.703288	2.051830	2.472660	2.770680	3.689600
28	0.255768	0.683353	1.312527	1.701131	2.048410	2.467140	2.763260	3.673900
29	0.255684	0.683044	1.311434	1.699127	2.045230	2.462020	2.756390	3.659400
30	0.255605	0.682756	1.310415	1.697261	2.042270	2.457260	2.750000	3.646000

$t = 1,744$

O teste t-Student depende dos graus de liberdade

t-Student

$t = 1,744$

$t_{crítico} = 2,093$

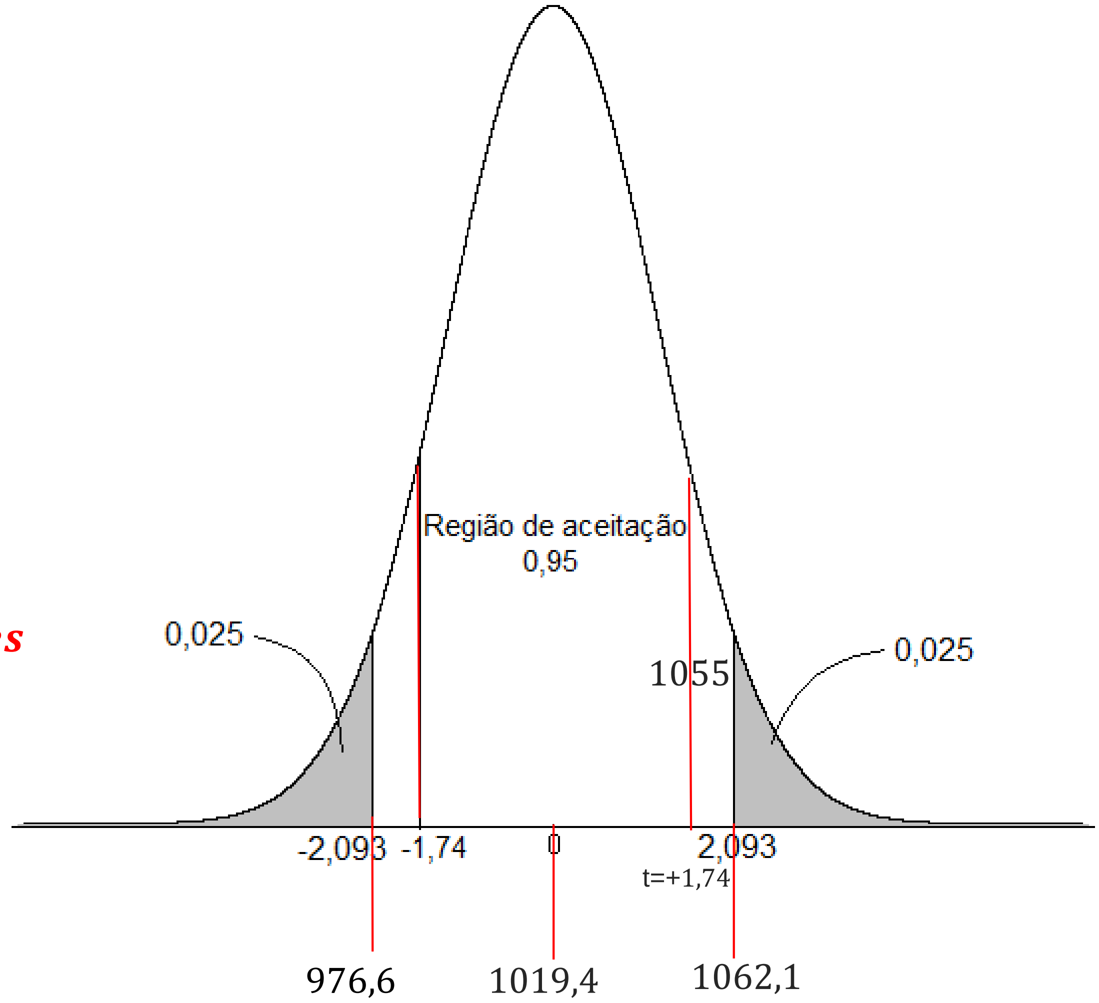
$t < t_{crítico}$

Os valores são estatisticamente equivalentes

$$mg.erro = \frac{t_{crítico}.\sigma}{\sqrt{n}}$$

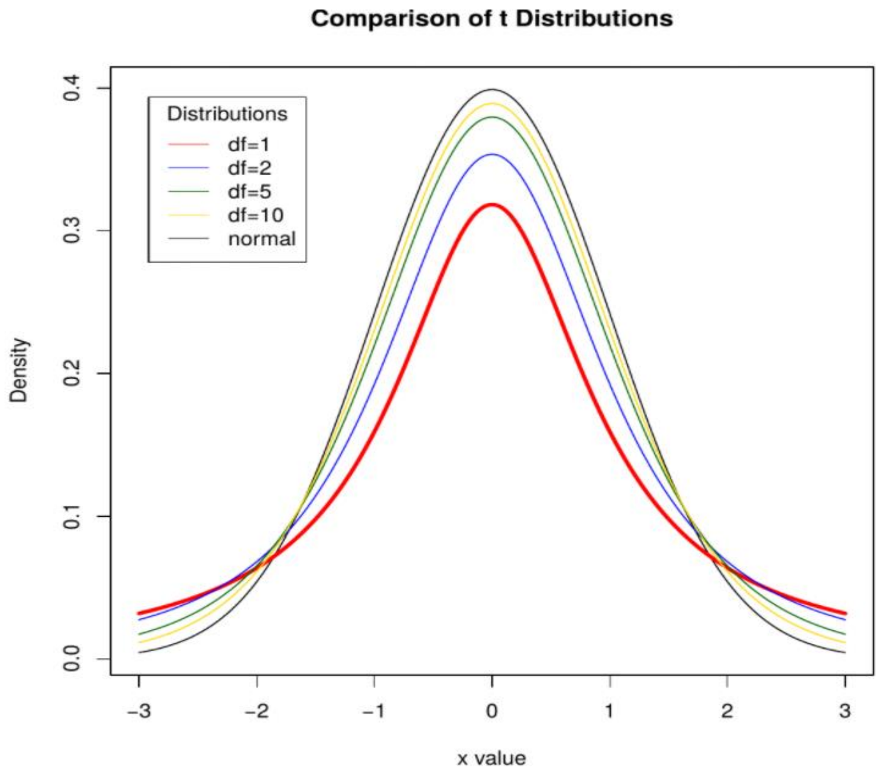
$$mg.erro = \frac{2,093 \,.\, 89,0}{\sqrt{19}}$$

$$mg.erro = 71,3$$



Teste Z-Score Vs. t-Student

- Quanto mais graus de liberdade par um ***t-Distribution***, mas achatada será a curva
- Isto indica **mais variância** nos dados causada por um baixo número de observações
- Portanto tratam-se de dados **menos confiáveis**



Going Deeper: Teste Qui-Quadrado

Teste qui-quadrado

- **Qui-quadrado**: Suponha que temos duas variáveis qualitativas X classificadas em r categorias e Y classificadas em s categorias
- O teste χ^2 (***chi square***) é, essencialmente, um mecanismo pelo qual os desvios de uma proporção hipotética são reduzidos a um único valor, que permite determinar uma probabilidade a respeito da casualidade ou não dos desvios entre as proporções observadas e esperadas

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

De forma simplificada

$$\frac{(o_i - e_i)^2}{e_i}$$

Teste qui-quadrado

- **Qui-quadrado**: Suponha que temos duas variáveis qualitativas X classificadas em r categorias e Y classificadas em s categorias
- O teste χ^2 (***chi square***) é, essencialmente, um mecanismo pelo qual os desvios de uma proporção hipotética são reduzidos a um único valor, que permite determinar uma probabilidade a respeito da casualidade ou não dos desvios entre as proporções observadas e esperadas

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

De forma simplificada

$$\frac{(o_i - e_i)^2}{e_i}$$

Exemplo: Teste qui-quadrado

Queremos verificar se a criação de determinado tipo de cooperativa está associada com algum fator regional.

Dados: Cooperativas autorizadas a funcionar por tipo e estado (1974)

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	214	237	78	119	648
Paraná	51	102	126	22	301
Rio G do Sul	111	304	139	48	602
Total	376	643	343	189	1551

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	33%	37%	12%	18%	100%
Paraná	17%	34%	42%	7%	100%
Rio G do Sul	18%	50%	23%	8%	100%
Total	24%	41%	22%	12%	100%

Calculado: Valores esperados assumindo a independência entre as variáveis

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157	269	143	79	648
Paraná	73	125	67	37	301
Rio G do Sul	146	250	133	73	602
Total	376	643	343	189	1551

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	24%	41%	22%	12%	100%
Paraná	24%	41%	22%	12%	100%
Rio G do Sul	24%	41%	22%	12%	100%
Total	24%	41%	22%	12%	100%

Exemplo: Teste qui-quadrado

Desvios entre observados e esperados.

$$o_i - e_i$$

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57	-32	-65	40
Paraná	-22	-23	59	-15
Rio G do Sul	-35	54	6	-25

- *A célula Escola-São Paulo é aquela que apresenta o maior desvio da suposição de não-associação (−65). Nessa célula esperávamos 143 casos.*
- *A célula Escola-Paraná também tem um desvio alto (59), mas o valor esperado é bem menor (67).*

$$\frac{(o_i - e_i)^2}{e_i}$$

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	20,62	3,73	29,55	20,30
Paraná	6,61	4,16	51,96	5,87
Rio G do Sul	8,36	11,87	0,26	8,77

- *A célula Escola-São Paulo obtemos* $(-65)^2/_{143} = 29.55$
- *A célula Escola-Paraná obtemos* $(59)^2/_{67} = 51.96$*, o que é uma*
- *Indicado que o desvio devido a essa última célula é "maior" do que aquele da primeira.*

- Uma medida do afastamento global pode ser dada pela soma de todas as medidas na segunda tabela.
- Essa medida é denominada χ^2 (qui-quadrado) de Pearson,
- $\chi^2 = 20.69 + 3.81 + 29.55 + 20.25 + 6.63 + 3.90 + 51.96 + 6.08 + 8.39 + 11.66 + 0.27 + 8.56 = 171.76$
- Um valor grande de χ^2 indica **independência** entre as variáveis, o que parece ser o caso.

Exemplo: Teste qui-quadrado

Desvios entre observados e esperados.

$$o_i - e_i$$

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57	-32	-65	40
Paraná	-22	-23	59	-15
Rio G do Sul	-35	54	6	-25

- *A célula Escola-São Paulo é aquela que apresenta o maior desvio da suposição de não-associação (−65). Nessa célula esperávamos 143 casos.*
- *A célula Escola-Paraná também tem um desvio alto (59), mas o valor esperado é bem menor (67).*

$$\frac{(o_i - e_i)^2}{e_i}$$

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	20,62	3,73	29,55	20,30
Paraná	6,61	4,16	51,96	5,87
Rio G do Sul	8,36	11,87	0,26	8,77

- *A célula Escola-São Paulo obtemos* $^{(-65)^2}/_{143} = 29.55$
- *A célula Escola-Paraná obtemos* $^{(59)^2}/_{67} = 51.96$*, o que é uma*
- *Indicado que o desvio devido a essa última célula é "maior" do que aquele da primeira.*

- Uma medida do afastamento global pode ser dada pela soma de todas as medidas na segunda tabela.
- Essa medida é denominada χ^2 (qui-quadrado) de Pearson,
- $\chi^2 = 20.69 + 3.81 + 29.55 + 20.25 + 6.63 + 3.90 + 51.96 + 6.08 + 8.39 + 11.66 + 0.27 + 8.56 = 171.76$
- Um valor grande de χ^2 indica **independência** entre as variáveis, o que parece ser o caso.

Going Deeper: Teste F

Teste ANOVA

- Análise de variância é a técnica estatística que permite avaliar afirmações sobre as médias de populações. A análise visa, fundamentalmente, verificar se existe uma diferença significativa entre as médias

- A análise de variância compara médias de diferentes populações para verificar se essas populações possuem médias iguais ou não. Assim, essa técnica permite que vários grupos sejam comparados a um só tempo.

Fonte de variação	Soma dos quadrados	Graus de liberdade	Variância	Valor de F
Entre amostras	$SQ_{Ent} = \sum_{i=1}^{27} n_i \cdot (\overline{x}_i - \overline{\overline{x}})^2$	$k - 1 = 26$	$S_e^2 = \frac{SQ_{Ent}}{k-1}$	$F = \frac{S_e^2}{S_d^2}$
Dentro das amostras	$SQ_{Den} = SQ_{Tot} - SQ_{Ent}$	$\sum_{i=1}^{27} n_i - k = \sum_{i=1}^{27} n_i - 27$	$S_d^2 = \frac{SQ_{Den}}{\sum_{i=1}^{27} n_i - k}$	
Total	$SQ_{Tot} = \sum_{i=1}^{27}\sum_{j=1}^{Ni} \left(x_i^j - \overline{\overline{x}}\right)^2$	$\sum_{i=1}^{27} n_i - 1$[3]		

Exemplo: Teste ANOVA

- Vamos analisar quatro amostras de plástico para determinar se elas têm diferentes forças médias
- A ANOVA com um fator calculou uma média para cada uma das quatro amostras de plástico. As médias do grupo são: 11,203, 8,938, 10,683 e 8,838 Essas médias de grupo estão distribuídas em torno da média global para todas as 40 observações, que é 9,915. Se as médias dos grupos estão aglomeradas próximas à média global, suas variâncias é baixa. No entanto, se as médias do grupo estiverem mais afastadas da média global, a variância delas será maior.

Médias

Amostra	N	Média	DesvPad	IC de 95%
1	10	11.203	1.995	(9.857, 12.548)
2	10	8.938	2.980	(7.592, 10.283)
3	10	10.683	1.102	(9.337, 12.028)
4	10	8.838	1.879	(7.492, 10.184)

Exemplo: Teste ANOVA

- Qual valor usamos para medir a variância entre as médias amostrais para o exemplo da resistência do plástico? Na saída da ANOVA com um fator, usaremos o quadrado médio ajustado (QM Aj) para o Fator, que é 14.540**. É a soma dos desvios quadrados divididos pelo GL do fator**.

- Também precisamos de uma estimativa da variabilidade dentro de cada amostra. Para calcular essa variância, precisamos calcular o quão distante cada observação está em relação à média do grupo, para todas as 40 observações. Tecnicamente é **a soma dos desvios ao quadrado da diferença de cada observação em relação à média do grupo, dividido pelo s GL do erro**, que resulta em 4,402.

Análise de Variância

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Amostra	3	43.62	14.540	$\boxed{3.30}$	0.031
Erro	36	158.47	4.402		
Total	39	202.09			

Para este exemplo de ANOVA com um fator, o valor que usaremos para a variância dentro das amostras é o **QM Aj** para o erro, que é 4,402. É chamado de "erro" porque é a variabilidade que não é explicada pelo fator.

Exemplo: Teste ANOVA

- Qual valor usamos para medir a variância entre as médias amostrais para o exemplo da resistência do plástico? Na saída da ANOVA com um fator, usaremos o quadrado médio ajustado (QM Aj) para o Fator, que é 14.540**. É a soma dos desvios quadrados divididos pelo GL do fator**.

- Também precisamos de uma estimativa da variabilidade dentro de cada amostra. Para calcular essa variância, precisamos calcular o quão distante cada observação está em relação à média do grupo, para todas as 40 observações. Tecnicamente é **a soma dos desvios ao quadrado da diferença de cada observação em relação à média do grupo, dividido pelo s GL do erro**, que resulta em 4,402.

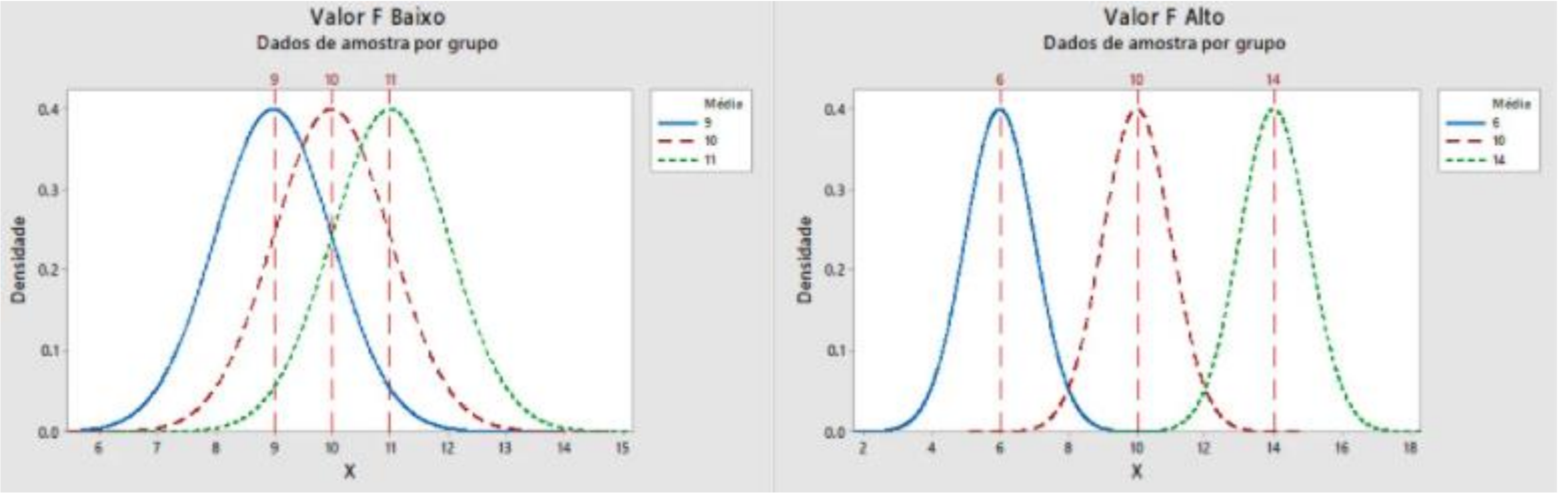
Análise de Variância

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Amostra	3	43.62	14.540	$\boxed{3.30}$	0.031
Erro	36	158.47	4.402		
Total	39	202.09			

Para este exemplo de ANOVA com um fator, o valor que usaremos para a variância dentro das amostras é o **QM Aj** para o erro, que é 4,402. É chamado de “erro” porque é a variabilidade que não é explicada pelo fator.

Teste F

- O gráfico com o baixo valor-F mostra um caso em que as médias dos grupos estão próximas (baixa variabilidade) em relação à variabilidade dentro de cada grupo. O gráfico com o alto valor-F mostra um caso em que a variabilidade das médias dos grupos é grande em relação à variabilidade intragrupo. Para rejeitar a hipótese nula de que as médias do grupo são iguais, precisamos de um valor F alto



ALGUMA DÚVIDA ATÉ AQUI?

Aplicações Práticas: Jupyter

DÚVIDAS FINAIS

COMO FOI?

T