Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Introdução à Inferência Estatística -
Intervalo de Confiança para Média, Proporção e Variância

Edna A. Reis e Ilka A. Reis

Apostila Didática
01-2020

SUMÁRIO

1. REVISÃO DAS TABELAS NORMAL, T-STUDENT E QUI-QUADRADO

Nesta seção, vamos relembrar como ler as três tabelas mais usadas na Estatística: Normal Padrão, t-Student e Qui-quadrado. Isto será muito importante para compreender as seções seguintes.

1.1. Tabela Normal Padrão

A Tabela Normal Padrão (ou, simplesmente, *Tabela Z),* apresentada na Tabela 1, é formada por conjuntos de duas colunas: a do percentil *z* e a da área (probabilidade) acima de *z* na curva Normal com média=0 e desvio-padrão=1 (Figura 1.1-a). Note que a *curva*[1] Normal Padrão é simétrica em torno do zero, o que significa que metade da área (probabilidade=0,5) está abaixo do zero (e a outra metade está acima do zero).

A Tabela Normal Padrão foi construída para responder a duas perguntas:

1) Qual é o valor de *z* que deixa *acima* dele uma *área (probabilidade)* igual a *p*, por exemplo, p=0,0054? Ou seja, qual é o valor de *z* tal que P(*Z*>*z*) = *p*? (Figura 1.1-b)
2) Qual é a *área (probabilidade)* na curva Normal Padrão *acima* de um dado valor de z, por exemplo, *z*=1,64? Ou seja, qual é o valor de P(*Z>1,64*)? (Figura 1.1-c)

Para responder à pergunta 2, consultamos a Tabela 1 na coluna referente a *z* e buscamos o valor *z*=1,64. Este valor está no cruzamento da terceira coluna de *z* com a linha 14. Ao lado direito do valor 1,64, temos a resposta à pergunta: P(Z>1,64)=0,0505 (representação visual na Figura 1.1-c).

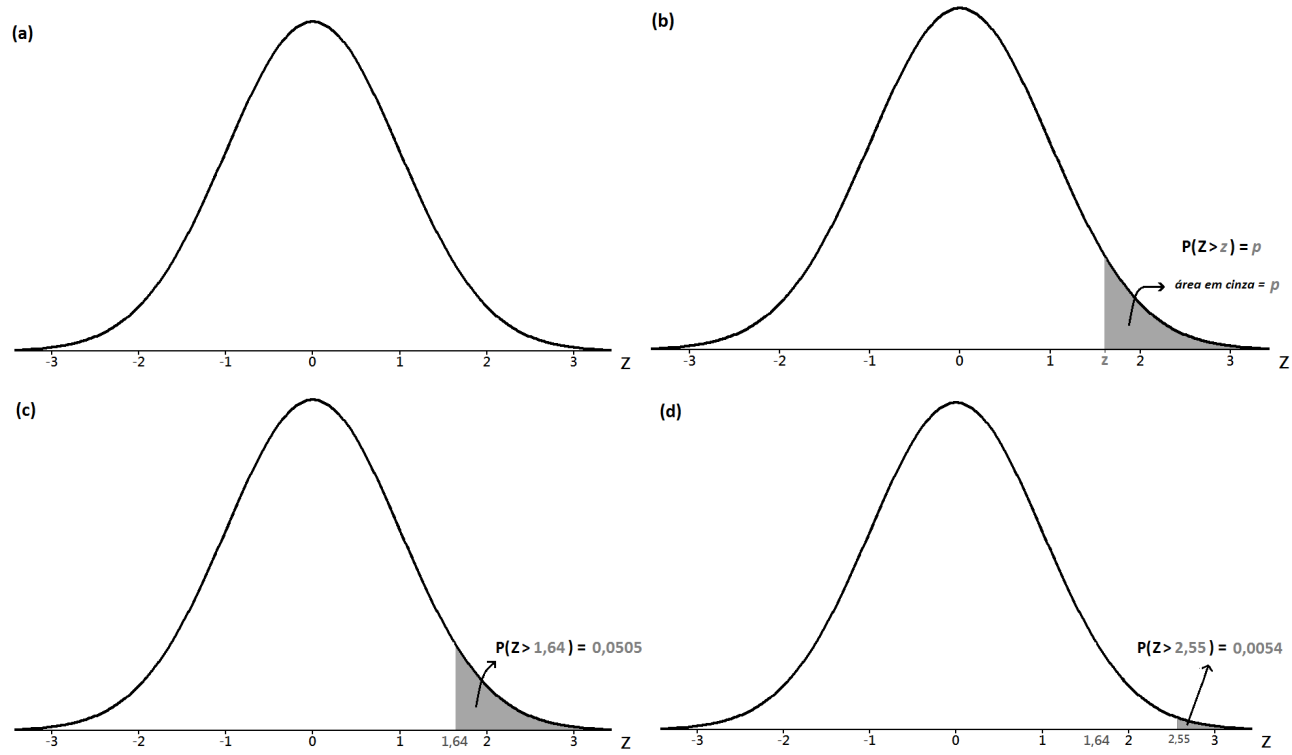

Figura 1.1. (a) Curva Normal padronizada (Curva Z), com área total abaixo dela (entre –∞ e +∞) igual a 1 (um).
(b) Representação do valor *z* (genérico) que deixa uma probabilidade igual a *p* acima dele.
(c) Representação do valor z=1,64, que deixa uma probabilidade (aprox.) igual a 0,0505 acima dele.
(d) Representação do valor z=2,55, que deixa uma probabilidade (aprox.) igual a 0,0054 acima dele.

[1] O nome formal é *função de densidade de probabilidade*, mas neste texto usamos o termo informal *curva*.

Tabela 1: Valores na distribuição Normal Padronizada segundo a probabilidade deixada acima.

z	P(Z ≥ z)	z	P(Z ≥ z)	z	P(Z ≥ z)	z	P(Z ≥ z)	z	P(Z ≥ z)	z	P(Z ≥ z)
0,01	0,4960	0,51	0,3050	1,01	0,1562	1,51	0,0655	2,01	0,0222	2,51	0,0060
0,02	0,4920	0,52	0,3015	1,02	0,1539	1,52	0,0643	2,02	0,0217	2,52	0,0059
0,03	0,4880	0,53	0,2981	1,03	0,1515	1,53	0,0630	2,03	0,0212	2,53	0,0057
0,04	0,4840	0,54	0,2946	1,04	0,1492	1,54	0,0618	2,04	0,0207	2,54	0,0055
0,05	0,4801	0,55	0,2912	1,05	0,1469	1,55	0,0606	2,05	0,0202	2,55	0,0054
0,06	0,4761	0,56	0,2877	1,06	0,1446	1,56	0,0594	2,06	0,0197	2,56	0,0052
0,07	0,4721	0,57	0,2843	1,07	0,1423	1,57	0,0582	2,07	0,0192	2,57	0,0051
0,08	0,4681	0,58	0,2810	1,08	0,1401	1,58	0,0571	2,08	0,0188	2,58	0,0049
0,09	0,4641	0,59	0,2776	1,09	0,1379	1,59	0,0559	2,09	0,0183	2,59	0,0048
0,10	0,4602	0,60	0,2743	1,10	0,1357	1,60	0,0548	2,10	0,0179	2,60	0,0047
0,11	0,4562	0,61	0,2709	1,11	0,1335	1,61	0,0537	2,11	0,0174	2,61	0,0045
0,12	0,4522	0,62	0,2676	1,12	0,1314	1,62	0,0526	2,12	0,0170	2,62	0,0044
0,13	0,4483	0,63	0,2643	1,13	0,1292	1,63	0,0516	2,13	0,0166	2,63	0,0043
0,14	0,4443	0,64	0,2611	1,14	0,1271	1,64	0,0505	2,14	0,0162	2,64	0,0041
0,15	0,4404	0,65	0,2578	1,15	0,1251	1,65	0,0495	2,15	0,0158	2,65	0,0040
0,16	0,4364	0,66	0,2546	1,16	0,1230	1,66	0,0485	2,16	0,0154	2,66	0,0039
0,17	0,4325	0,67	0,2514	1,17	0,1210	1,67	0,0475	2,17	0,0150	2,67	0,0038
0,18	0,4286	0,68	0,2483	1,18	0,1190	1,68	0,0465	2,18	0,0146	2,68	0,0037
0,19	0,4247	0,69	0,2451	1,19	0,1170	1,69	0,0455	2,19	0,0143	2,69	0,0036
0,20	0,4207	0,70	0,2420	1,20	0,1151	1,70	0,0446	2,20	0,0139	2,70	0,0035
0,21	0,4168	0,71	0,2389	1,21	0,1131	1,71	0,0436	2,21	0,0136	2,71	0,0034
0,22	0,4129	0,72	0,2358	1,22	0,1112	1,72	0,0427	2,22	0,0132	2,72	0,0033
0,23	0,4090	0,73	0,2327	1,23	0,1093	1,73	0,0418	2,23	0,0129	2,73	0,0032
0,24	0,4052	0,74	0,2296	1,24	0,1075	1,74	0,0409	2,24	0,0125	2,74	0,0031
0,25	0,4013	0,75	0,2266	1,25	0,1056	1,75	0,0401	2,25	0,0122	2,75	0,0030
0,26	0,3974	0,76	0,2236	1,26	0,1038	1,76	0,0392	2,26	0,0119	2,76	0,0029
0,27	0,3936	0,77	0,2206	1,27	0,1020	1,77	0,0384	2,27	0,0116	2,77	0,0028
0,28	0,3897	0,78	0,2177	1,28	0,1003	1,78	0,0375	2,28	0,0113	2,78	0,0027
0,29	0,3859	0,79	0,2148	1,29	0,0985	1,79	0,0367	2,29	0,0110	2,79	0,0026
0,30	0,3821	0,80	0,2119	1,30	0,0968	1,80	0,0359	2,30	0,0107	2,80	0,0026
0,31	0,3783	0,81	0,2090	1,31	0,0951	1,81	0,0351	2,31	0,0104	2,81	0,0025
0,32	0,3745	0,82	0,2061	1,32	0,0934	1,82	0,0344	2,32	0,0102	2,82	0,0024
0,33	0,3707	0,83	0,2033	1,33	0,0918	1,83	0,0336	2,33	0,0099	2,83	0,0023
0,34	0,3669	0,84	0,2005	1,34	0,0901	1,84	0,0329	2,34	0,0096	2,84	0,0023
0,35	0,3632	0,85	0,1977	1,35	0,0885	1,85	0,0322	2,35	0,0094	2,85	0,0022
0,36	0,3594	0,86	0,1949	1,36	0,0869	1,86	0,0314	2,36	0,0091	2,86	0,0021
0,37	0,3557	0,87	0,1922	1,37	0,0853	1,87	0,0307	2,37	0,0089	2,87	0,0021
0,38	0,3520	0,88	0,1894	1,38	0,0838	1,88	0,0301	2,38	0,0087	2,88	0,0020
0,39	0,3483	0,89	0,1867	1,39	0,0823	1,89	0,0294	2,39	0,0084	2,89	0,0019
0,40	0,3446	0,90	0,1841	1,40	0,0808	1,90	0,0287	2,40	0,0082	2,90	0,0019
0,41	0,3409	0,91	0,1814	1,41	0,0793	1,91	0,0281	2,41	0,0080	2,91	0,0018
0,42	0,3372	0,92	0,1788	1,42	0,0778	1,92	0,0274	2,42	0,0078	2,92	0,0018
0,43	0,3336	0,93	0,1762	1,43	0,0764	1,93	0,0268	2,43	0,0075	2,93	0,0017
0,44	0,3300	0,94	0,1736	1,44	0,0749	1,94	0,0262	2,44	0,0073	2,94	0,0016
0,45	0,3264	0,95	0,1711	1,45	0,0735	1,95	0,0256	2,45	0,0071	2,95	0,0016
0,46	0,3228	0,96	0,1685	1,46	0,0721	1,96	0,0250	2,46	0,0069	2,96	0,0015
0,47	0,3192	0,97	0,1660	1,47	0,0708	1,97	0,0244	2,47	0,0068	2,97	0,0015
0,48	0,3156	0,98	0,1635	1,48	0,0694	1,98	0,0239	2,48	0,0066	2,98	0,0014
0,49	0,3121	0,99	0,1611	1,49	0,0681	1,99	0,0233	2,49	0,0064	2,99	0,0014
0,50	0,3085	1,00	0,1587	1,50	0,0668	2,00	0,0228	2,50	0,0062	3,00	0,0013

Fonte: Gerada no programa R através da função `pnorm`.

Exemplo: $Z_{[0,0505]} = 1{,}64$ ou seja, $P(Z > 1{,}64) = 0{,}0505$.

Quanto à pergunta 1, suponha que *p*=0,0054. Assim, para responder à pergunta 1, consultamos a Tabela 1 na coluna referente a P(*Z*>*z*) e buscamos o valor 0,0054. Este valor está no cruzamento da última coluna de P(*Z*>*z*) com a linha 5. Ao lado esquerdo do valor 0,0054, temos que *z*=2,55 (representação visual na Figura 1.1-d).

Vamos relembrar a notação que será útil nas seções seguintes. Temos, por exemplo, que $Z_{[0,0505]} = 1{,}64$ ou seja, $P(Z > 1{,}64) = 0{,}0505$, é a resposta à pergunta 1 se p=0.0505.
Se p=0,0250, então $Z_{[0,0250]} = 1{,}96$ (quarta coluna de P(Z>z), linha 46).

Como veremos a seguir, os quatro valores mais usados na construção de intervalos de confiança são *p*=0,0500, *p*=0,0250, *p*=0,0100 e *p*=0,0050. Já sabemos que $Z_{[0,0250]} = 1{,}96$. Assim, quais seriam os valores de $Z_{[0,0050]}$, $Z_{[0,0100]}$ e $Z_{[0,0500]}$? Para descobrir esses valores, basta consultar o valor entre colchetes [] na coluna P[*Z*>*z*] e a resposta será o valor de *z* na coluna à esquerda. Vejamos:

Para encontrar o valor de $Z_{[0,0050]}$, devemos procurar o valor 0,0050 na coluna de P[*Z*>*z*]. Como a tabela Normal Padrão não comporta todos os valores da distribuição Normal (eles são infinitos), é comum que não encontremos exatamente o valor procurado. É o caso do valor 0,0050. Na última coluna de P[*Z*>*z*], encontramos os valores 0,0051 e 0,0049 (linhas 7 e 8, respectivamente). Como os dois valores estão igualmente próximos do valor desejado (0,0050), podemos tomar qualquer um dos dois valores de *z* (2,57 e 2,58, respectivamente) para ser o valor de $Z_{[0,0050]}$. O valor mais comumente usado é $Z_{[0,0050]} = 2{,}58$.

No caso de $Z_{[0,0500]}$, o valor 0,0500 também não se encontra na Tabela 1, mas, na quarta coluna de P[*Z*>*z*], encontram-se dois valores muito próximos, 0,0505 e 0,0495 (linhas 14 e 15, respectivamente). Como os dois valores estão igualmente próximos do valor desejado (0,0500), podemos tomar qualquer um dos dois valores de z (1,64 ou 1,65, respectivamente) para ser o valor de $Z_{[0,0500]}$. O valor mais comumente usado é $Z_{[0,0500]} = 1{,}64$.

Usando o mesmo raciocínio dos parágrafos anteriores, encontramos que $Z_{[0,0100]}$=2,33.
Recapitulando, temos que os valores mais usados de $Z_{[p]}$ são

$$Z_{[0,0050]} = 2{,}58,\ Z_{[0,0100]} = 2{,}33,\ Z_{[0,0250]} = 1{,}96 \text{ e } Z_{[0,0500]} = 1{,}64.$$

1.2. Tabela T-Student

Enquanto a tabela Normal Padrão representa a distribuição de apenas um modelo probabilístico (a única curva Normal Padrão que existe), a *tabela da distribuição T-Student centralizada*, ou simplesmente *Tabela T*, foi construída para representar uma família de distribuições, ou seja, várias curvas T. Cada curva T é gerada com um valor diferente de um parâmetro chamado *graus de liberdade* (*g.l.*) (Figura 1.2-a). Note que as curva T também são simétricas em torno do zero.

Na Tabela 2, que apresenta a tabela t-Student, cada linha representa um valor diferente dos *graus de liberdade* e, portanto, uma curva t-Student diferente. Cada coluna da Tabela 2 representa o valor de *p*, a área (probabilidade) acima dos percentis dentro da tabela (Figura 1.2-b). Assim, na notação usada para representar os percentis da Tabela t-Student, temos que nos referir tanto ao valor dos graus de liberdade da distribuição quanto à área que o percentil deixa acima dele. Por exemplo,

$$T_{[5;0,10]} = 1{,}476 \quad \text{ou seja,} \quad P\left(T_{[5]} > 1{,}476\right) = 0{,}10 \text{ (Figura 1.2-c).}$$

O primeiro valor dentro dos colchetes [] se refere ao valor dos *graus de liberdade* (g.l.=5) e o segundo valor se refere à área (probabilidade) que o percentil deixará acima dele (0,10). Consultando a linha 5 da tabela na coluna p=0,10, encontramos o valor 1,476. Do mesmo modo, $T_{[5;0,025]} = 2{,}571$ (linha 5 e coluna p=0,025, Figura 1.2-d).

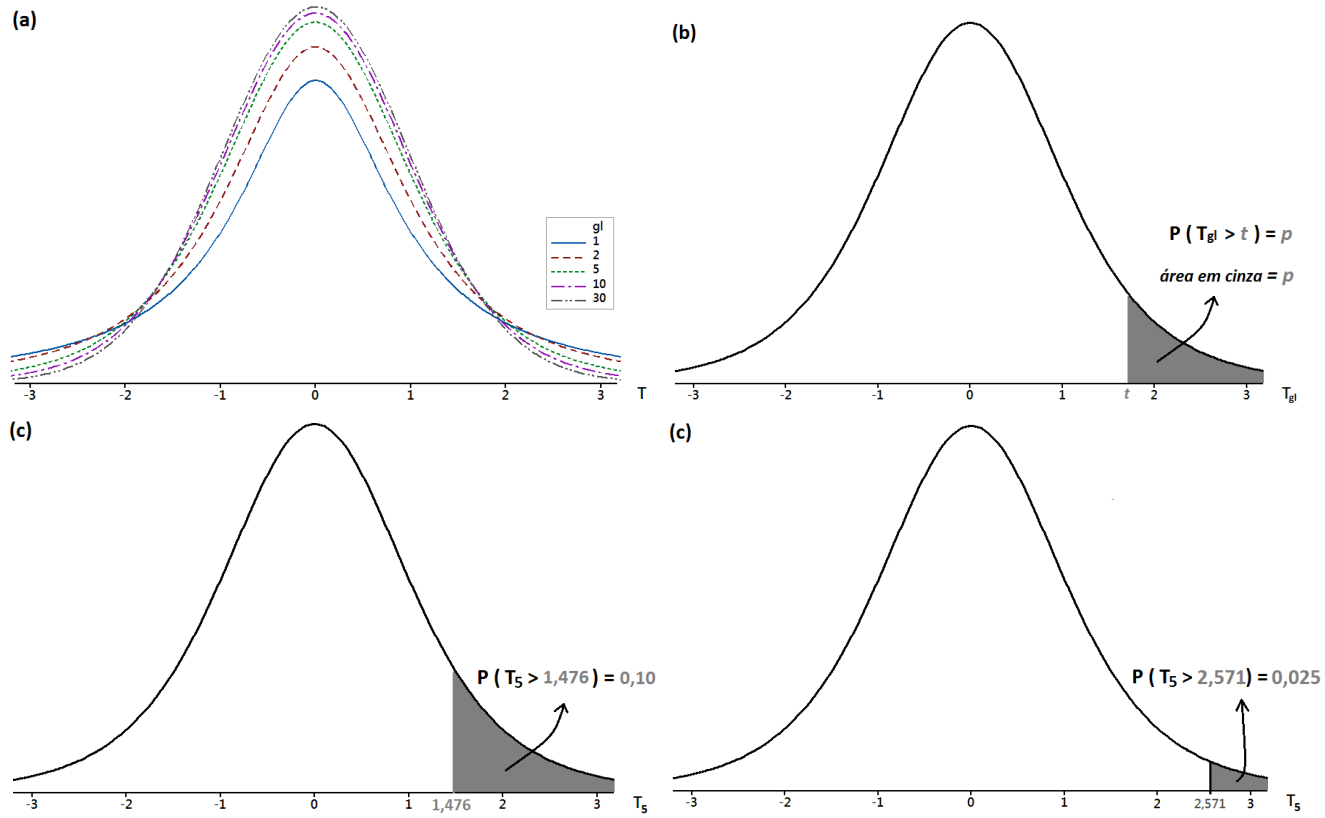

Figura 1.2. (a) Curvas T-Student (Curvas T) para diferentes valores de graus de liberdade (gl). A área total abaixo de cada curva (entre –∞ e +∞) é igual a 1 (um).
(b) Representação, em uma curva T genérica, do valor *t* que deixa uma probabilidade *p* acima dele.
(c) Representação, na curva T com gl=5, do valor 1,476, que deixa uma probabilidade (aprox.) igual a 0,10 acima dele.
(d) Representação, na curva T com gl=5, do valor 2,571, que deixa uma probabilidade (aprox.) igual a 0,025 acima dele.

Tabela 2: Valores na distribuição T de Student segundo os graus de liberdade (g.l.) e a probabilidade deixada acima (p).

	Probabilidade Acima (p)									
g.l.	**0,20**	**0,15**	**0,10**	**0,05**	**0,025**	**0,010**	**0,005**	**0,0025**	**0,001**	**0,0005**
1	1,376	1,963	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	1,061	1,386	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,978	1,250	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,941	1,190	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,920	1,156	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,906	1,134	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,889	1,108	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,883	1,100	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,879	1,093	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,876	1,088	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,768
23	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,792
24	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
31	0,853	1,054	1,309	1,696	2,040	2,453	2,744	3,022	3,375	3,633
32	0,853	1,054	1,309	1,694	2,037	2,449	2,738	3,015	3,365	3,622
33	0,853	1,053	1,308	1,692	2,035	2,445	2,733	3,008	3,356	3,611
34	0,852	1,052	1,307	1,691	2,032	2,441	2,728	3,002	3,348	3,601
35	0,852	1,052	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
36	0,852	1,052	1,306	1,688	2,028	2,434	2,719	2,990	3,333	3,582
37	0,851	1,051	1,305	1,687	2,026	2,431	2,715	2,985	3,326	3,574
38	0,851	1,051	1,304	1,686	2,024	2,429	2,712	2,980	3,319	3,566
39	0,851	1,050	1,304	1,685	2,023	2,426	2,708	2,976	3,313	3,558
40	0,851	1,050	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
50	0,849	1,047	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
60	0,848	1,045	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
70	0,847	1,044	1,294	1,667	1,994	2,381	2,648	2,899	3,211	3,435
80	0,846	1,043	1,292	1,664	1,990	2,374	2,639	2,887	3,195	3,416
90	0,846	1,042	1,291	1,662	1,987	2,368	2,632	2,878	3,183	3,402
100	0,845	1,042	1,290	1,660	1,984	2,364	2,626	2,871	3,174	3,390

Fonte: Gerada no programa R através da função `qt`.

Exemplo: $T_{[5;0,10]} = 1,476$ ou seja, $P\left(T_{[5]} > 1,476\right) = 0,10$.

Como a tabela t-Student não comporta todos os valores possíveis para os graus de liberdade (eles são infinitos), pode ocorrer que o valor de g.l. desejado não esteja na tabela. Quando isto acontecer (por exemplo, g.l.=44), devemos usar a linha dos graus de liberdade (g.l.) imediatamente menor (no exemplo, seria g.l.=40).

Note que, para uma probabilidade fixa, o valor do percentil na distribuição t-Student vai variando cada vez menos (nas casas decimais) à medida que aumentam os graus de liberdade e se aproximando do percentil da Normal Padrão. Por exemplo, para uma probabilidade de 0,025 (para cima), o percentil da t-Student começa em 12,71 (g.l.=1) e rapidamente cai para valores próximos de 2, chegando a 1,98 (valor bem próximo do valor 1,96 da Normal) quando n=100. Para valores acima de n=100, podemos usar a Tabela Normal Padrão para encontrar os percentis desejados. De fato, a curva T se aproxima da curva Normal Padrão à medida que os graus de liberdade aumentam.

1.3. Tabela Qui-Quadrado

Assim como a tabela da distribuição *t-Student*, a Tabela Qui-quadrado foi construída para representar uma família de distribuições, pois a distribuição Qui-quadrado também depende de uma quantidade chamada *graus de liberdade* (*g.l.*) (Figura 1.3-a). Porém, diferentemente das distribuições Normal e t-*student*, a distribuição Qui-quadrado não é simétrica e só assume valores maiores do que 0 (zero).

A Tabela 3 apresenta a tabela Qui-quadrado, ou simplesmente Tabela Q. Cada linha da Tabela Q representa um valor diferentes dos *graus de liberdade* e, portanto, uma curva Qui-quadrado diferente. Cada coluna da Tabela Q representa o valor de *p*, a *área (probabilidade)* acima dos percentis dentro da tabela (Figura 1.3-b). Assim, na notação usada para representar os percentis da Tabela Q, temos que nos referir tanto ao valor dos graus de liberdade da distribuição quanto à área que o percentil deixa acima dele. Por exemplo,

$$Q_{[10;0,10]} = 15{,}987 \quad \text{ou seja,} \quad P\left(Q_{[10]} > 15{,}987\right) = 0{,}10 \qquad \text{(Figura 1.3-c)}$$

O primeiro valor dentro dos colchetes [] se refere ao grau de liberdade (g.l.=10) e o segundo valor se refere à área que o percentil deixará acima dele (0,10). Consultando a linha 10 da tabela na coluna p=0,10, encontramos o valor 15,987. Do mesmo modo, $Q_{[10;0,05]}$ = 18,307 (linha 10 e coluna p=0,05, Figura 3.2-d)

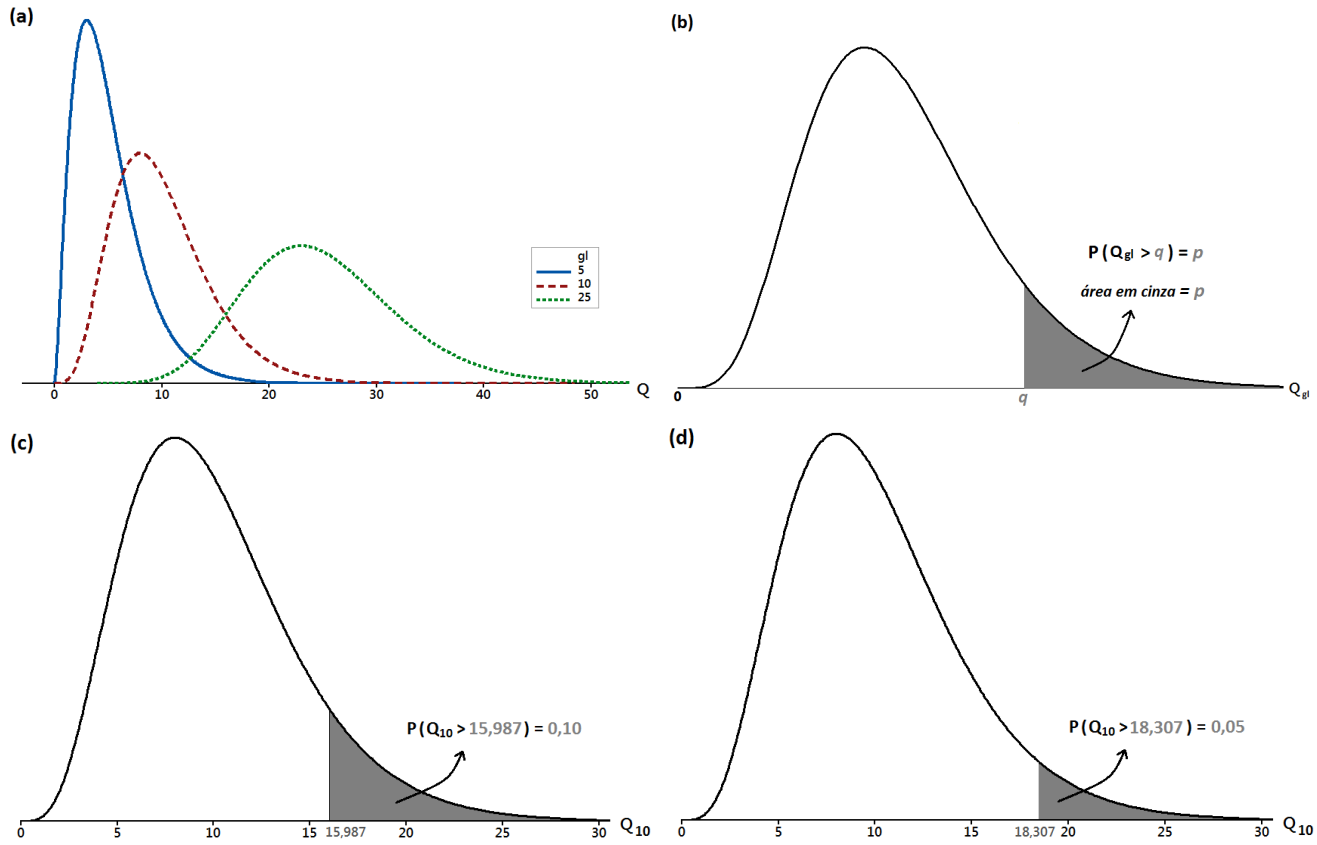

Figura 1.3. (a) Curvas Qui-Quadrado (Curvas Q) para diferentes valores de graus de liberdade (g.l.). A área total abaixo de cada curva (entre 0 e +∞) é igual a 1 (um).
(b) Representação, em uma curva Q genérica, do valor *q* que deixa uma probabilidade *p* acima dele.
(c) Representação, na curva Q com gl=10, do valor 15,987, que deixa uma probabilidade (aprox.) igual a 0,10 acima dele.
(d) Representação, na curva Q com gl=10, do valor 18,307, que deixa uma probabilidade (aprox.) igual a 0,05 acima dele.

Tabela 3: Valores na distribuição Qui-Quadrado segundo os graus de liberdade (g.l.) e a probabilidade deixada acima (p).

	Probabilidade Acima (p)									
g.l.	**0,995**	**0,990**	**0,975**	**0,95**	**0,90**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
1	0,000	0,000	0,001	0,004	0,016	2,706	3,814	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,071	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,299
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,042	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,335	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,194	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,257	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,954	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807	140,169

Fonte: Gerada no programa R através da função `qchisq`.

Exemplo: $Q_{[5;0,10]} = 9,236$ ou seja, $P\left(Q_{[5]} > 9,236\right) = 0,10$.

Como a tabela t-*Student*, a Tabela Qui-quadrado também não comporta todos os valores possíveis para os graus de liberdade (eles são infinitos). Assim, pode ocorrer que o valor do g.l. desejado não esteja na tabela. Quando isto acontecer (por exemplo, g.l.=56), devemos usar a linha dos graus de liberdade (g.l.) imediatamente menor (no exemplo, seria g.l.=50).

2. CONCEITOS E DEFINIÇÕES

A *Inferência Estatística* pode ser definida como o conjunto de métodos de análise estatística que permitem tirar conclusões sobre uma população com base em somente uma parte dela, a amostra.

A **população** é o conjunto de todos os elementos, indivíduos ou objetos cujas características (variáveis) estão sendo estudadas. Por exemplo, uma pesquisadora deseja estudar as características físicas dos peixes de determinada espécie que habita um lago. Ela pode estar interessada nas variáveis *peso* e *comprimento*, por exemplo.

Em geral, devido a restrições financeiras ou de tempo, não é possível observar ou medir as variáveis de interesse em todos os elementos da população. Neste caso, podemos ter acesso a apenas uma parte dos elementos da população, ou seja, uma amostra. Portanto, a **amostra** é o subconjunto da população no qual a variável é realmente observada ou medida.

Uma amostra é representativa da população se ela tem o mesmo tipo de variação que a população, porém, em uma escala menor.

Há muitas maneiras de selecionar os elementos de uma amostra de modo que ela seja representativa da população. Cada método de amostragem é o mais adequado dependendo do contexto do problema a ser resolvido. A área da Estatística que estuda esses esquemas de seleção chama-se *Amostragem*. Para mais detalhes, consulte Bolfarine e Bussab (2005).

Neste texto, consideraremos que as amostras foram selecionadas pelo esquema de *Amostragem Aleatória Simples (com reposição)*, no qual todos os elementos da população têm a mesma probabilidade de participar da amostra.

2.1. Parâmetro, Estimador e Estimativa

Um **parâmetro** é um valor relacionado à variável de estudo considerando-se todos os elementos da população no cálculo. Pode ser uma medida simples, como as de tendência central (a *média* ou a *mediana*) ou as de variabilidade (como a *variância* ou o *desvio-padrão*) de uma variável quantitativa, ou a *proporção* de indivíduos em uma categoria de uma variável qualitativa. Também pode ser uma medida de interpretação mais complexa, como uma razão de chances ou um risco relativo.

Como raramente podemos observar ou medir a variável em todos os elementos da população, o valor do parâmetro não pode ser calculado, sendo, portanto, *desconhecido*. Entretanto, podemos inferir sobre o valor do parâmetro, ou seja, estimá-lo, por meio de uma amostra representativa da população.

Neste texto, abordaremos a inferência estatística para os parâmetros: **média** (cujo valor na população será denotado por μ), **variância** (σ^2) ou o **desvio-padrão** (σ), de uma variável quantitativa Normal; e para a **proporção** (***p***) de elementos na categoria ***sucesso*** de uma variável qualitativa com duas categorias (chamadas genericamente de *sucesso* e *fracasso*).

Considere os seguintes exemplos de variáveis:
Variável X: tempo gasto na locomoção de casa até o trabalho/escola (em minutos) dos habitantes de uma cidade;
Variável Y: *status* dos empregados de uma empresa quanto ao nível de atividade física diária (sedentário ou ativo).

No caso da Variável X, podemos usar os parâmetros média (**μ**) e desvio-padrão (**σ**) para caracterizar os habitantes de uma cidade quanto ao tempo gasto na locomoção diária. No caso da Variável Y, podemos escolher uma das categorias da variável como sendo o evento sucesso (ex: sedentário) e caracterizar os empregados por meio da proporção de pessoas sedentárias (**p**).

Um **estimador** é uma função matemática que é calculada usando-se os valores da variável obtidos na amostra. O resultado desse cálculo é chamado **estimativa**.

No caso das variáveis quantitativas, os estimadores da sua *média* μ e da sua *variância* σ^2 na população são, respectivamente, a ***média amostral*** ($\overline{X}$) e a ***variância amostral*** (S^2) calculadas, como

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{e} \quad S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1},$$

na qual n é o tamanho da amostra e X_1, X_2, X_3, ..., X_n são os valores da variável X medidos, respectivamente, no *primeiro*, *segundo*, *terceiro*, ..., *n-ésimo* indivíduo da amostra.

Se a variável estudada é do tipo qualitativa com duas categorias (sucesso ou fracasso), o estimador pontual da ***proporção de sucessos p*** na população é dado pela ***proporção de sucessos na amostra*** ($\hat{p}$), ou seja,

$$\hat{p} = \frac{número\ de\ sucessos\ na\ amostra}{n}.$$

(Naturalmente, a *proporção de fracassos* será estimada por $1 - \hat{p}$).

O resultado da aplicação dos estimadores aos valores amostrais da variável de interesse é apenas um valor (ponto). Por este motivo, as estimativas resultantes desse processo são chamadas **estimativas pontuais**.

Apesar de ser simples e de fácil interpretação, o uso da estimativa pontual não leva em conta que a amostra utilizada é somente uma das muitas possíveis amostras que poderiam ser retiradas da população e que, portanto, o valor da estimativa pontual *varia* de amostra para amostra. Isto ficará mais claro na próxima subseção, na qual estudaremos o comportamento dos valores de um estimador considerando todas as amostras possíveis, ou seja, a *distribuição amostral* desse estimador.

2.2. Distribuição Amostral do Estimador

Cada amostra retirada de uma mesma população gera um valor diferente da estimativa do parâmetro estudado. Para ilustrar esta afirmação, consideremos o experimento ilustrado pela Figura 2.1. O experimento consiste em retirar várias amostras de mesmo tamanho da população de valores apresentada na Figura 2.1-a. A média populacional é igual a μ=57.16 (representada pela linha pontilhada na figura). Para construir a Figura 2.1-b, foram retiradas 10000 amostras de tamanho *n*=5.

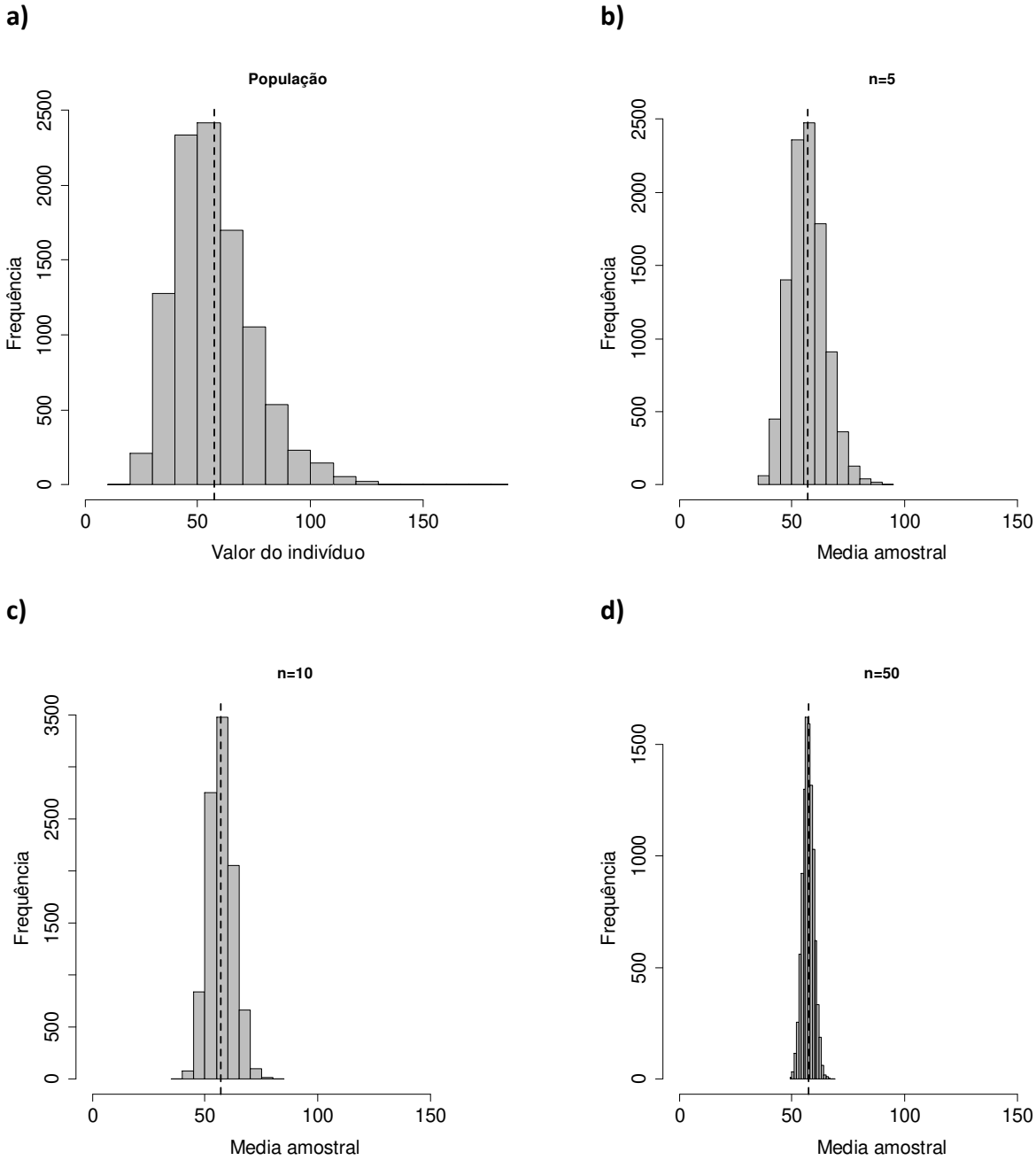

Figura 2.1. (a) Distribuição de frequências dos valores da população (μ=57.16 representada pela linha pontilhada).
(b) Distribuição de frequências dos valores da média de 10000 amostras de tamanho *n*=5 retiradas da população em a). (média dos valores igual a 56.89 representada pela linha pontilhada)
(c) Distribuição de frequências dos valores da média de 10000 amostras de tamanho *n*=10 retiradas da população em a). (média dos valores igual a 57.06 representada pela linha pontilhada)
(d) Distribuição de frequências dos valores da média de 10000 amostras de tamanho *n*=50 retiradas da população em a). (média dos valores igual a 57.17 representada pela linha pontilhada)

Para cada amostra, foi calculado o valor da média amostral. Assim, o histograma da Figura 2.1-b representa a distribuição de frequências dos 10000 valores da média amostral. Note que a média desses 10000 valores (linha pontilhada) é muito próxima do valor populacional (μ=57.16). Note também que a variabilidade dos valores das médias amostrais é *menor* do que a variabilidade dos valores populacionais (os gráficos foram construídos na mesma escala horizontal para facilitar essa comparação). Em outras palavras,

> o valor esperado (médio) do estimador média amostral coincide com o valor que ele pretende estimar (média populacional) e a variabilidade desses valores em torno da média populacional é menor do que a variabilidade da característica na população de interesse.

Esse comportamento dos valores do estimador média amostral se repete quando aumentamos o tamanho das amostras retiradas no experimento para *n*=10 e *n*=50 (figuras 2.1-c e 2.1-d, respectivamente). Ou seja, a média dos 10000 valores da média amostral (média das médias) coincide com o valor populacional e a variabilidade desses valores em torno do seu valor médio diminui cada vez mais com o aumento do tamanho da amostra *n*.

Além disso, a forma da distribuição de frequências se torna mais simétrica e mais próxima da distribuição Normal à medida que o tamanho das amostras aumenta. Esta aproximação da curva Normal por parte da distribuição dos valores das médias amostrais é mais fácil de ver na Figura 2.2, que reproduz a figura anterior e acrescenta a distribuição de frequências das médias para amostra de tamanho *n*=100.

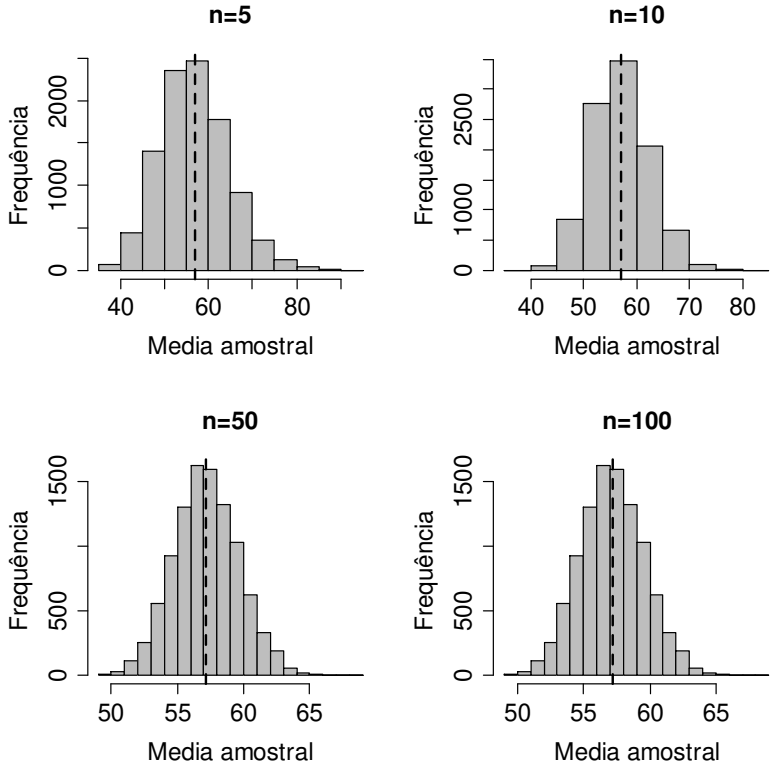

Figura 2.2. Distribuição de frequências dos valores da média de 10000 amostras de tamanho *n*=5, *n*=10, *n*=50 e *n*=100 retiradas da população representada na Figura 2.1-a. Nota-se que, à medida que o tamanho das amostras cresce, o valor esperado (média) das médias amostrais (linha pontilhada) permanece igual ao valor populacional (μ=57.16). Além disso, a variabilidade em torno do valor médio diminui e a forma da distribuição de frequências se torna mais simétrica e mais próxima do Modelo Normal.

O experimento ilustrado nas figuras 2.1 e 2.2 é uma maneira de entender o conceito de **Distribuição Amostral do Estimador**. De maneira geral, a distribuição amostral de estimador nos permite conhecer o comportamento dos valores de um estimador como se tivéssemos acesso a todas as amostras possíveis de uma população. Com este comportamento em mente, saberemos o quanto podemos confiar no único valor que temos em mãos como sendo uma boa estimativa do valor populacional desconhecido.

No caso do estimador **média amostral**, vimos que a maioria das amostras gera valores de média próximos ao valor populacional e que essa proximidade aumenta à medida que amostra fica maior (figuras 2.1 e 2.2). Esse resultado também vale para os estimadores **proporção amostral** e **variância amostral**.

Na próxima seção, veremos como levar em conta a variabilidade dos valores de estimador no momento de fornecer uma estimativa para o valor populacional desconhecido com base em uma *única amostra* da população. Os resultados apresentados nas próximas seções são obtidos no conhecimento teórico sobre a distribuição amostral dos estimadores.

3. ESTIMAÇÃO INTERVALAR

Como vimos anteriormente, o valor das estimativas pontuais varia de amostra para amostra. Assim, divulgar somente um único valor como estimativa de um parâmetro (**estimativa pontual**) deixa de lado toda a **incerteza** envolvida no processo de estimação desse parâmetro.

Para nos lembrar da incerteza envolvida no resultado amostral, vamos associar uma **margem de erro** à estimativa pontual, gerando o que é denominado **estimativa intervalar** que, nos casos mais simples[2], é calculada por

$$\textit{Estimativa Intervalar} = \textit{Estimativa Pontual} \pm \textit{Margem de Erro.}$$

Exemplo:
Um estudo pretende estimar o valor de μ, a escolaridade média das mães dos alunos da UFMG, medida em anos de estudo formal. Em uma amostra de 40 alunos da universidade, encontrou-se um número médio de 10 anos de estudo (estimativa pontual) com desvio-padrão amostral s=3,5 anos. A estimativa pontual para escolaridade média das mães dos alunos da UFMG é de 10 anos de estudo.

Suponha que a **margem de erro** foi calculada em 1,1 anos de estudo (vamos ver como foi calculada na seção 2.23). Assim, a **estimativa intervalar** para a escolaridade média das mães dos alunos da UFMG é de

$$[10 \pm 1{,}1] = [8{,}9 \ ; \ 11{,}1] \text{ anos de estudo.}$$

Toda estimativa intervalar tem associada a ela um **nível de confiança**, geralmente expresso em porcentagem e que está entre 0 e 100%. **Ex:** nível de confiança de 95%. Então, falamos em **Intervalo de Confiança.**

No exemplo, o intervalo de 95% de confiança para a escolaridade média das mães dos alunos da UFMG vai de 8,9 a 11,1 anos de estudo.

---

[2] Nem todas as estimativas intervalares são obtidas somando-se e subtraindo-se a margem de erro da estimativa pontual, como é o caso da estimativa intervalar da variância. No entanto, por ora, esse modo de cálculo nos servirá para entender as ideias da estimação intervalar.

Interpretação: Estima-se, com 95% de confiança, que o valor *desconhecido* da escolaridade média das mães dos alunos da UFMG esteja 8,9 e 11,1 anos de estudo.

O nível de confiança é um valor arbitrário e é escolhido por quem está fazendo a Inferência Estatística. Em geral, os valores do nível de confiança são altos, como 90%, 95%, 98% e, muito raramente, 99%.

Como veremos na seção a seguir, o valor do nível de confiança tem influência no tamanho da margem de erro do intervalo: quanto maior o nível de confiança, maior a margem de erro. Por isto, a escolha do nível de confiança deve ser feita com cuidado e o valor mais usado é o de 95%.

Por causa dessa influência do nível de confiança de um intervalo em sua largura, um intervalo com 100% de confiança seria tão largo[3] que não teria nenhuma utilidade. Assim, para conseguirmos um intervalo de confiança que tenha alguma utilidade prática, reduzimos essa confiança total em alguns pontos percentuais. Por exemplo, para um nível de 95% de confiança, teremos uma redução de 5 pontos percentuais. Desse modo, é usual expressar o nível de confiança em função dessa redução. Uma notação comum é usar 100(1-α)%, no qual α é a redução. Para um nível de 98% de confiança, por exemplo, teremos 100(1-α)%=98%, o que leva a (1-α)=0,98 e ao valor de α=0,02. Se o nível de confiança for de 95%, então α=0,05.

Como veremos nas próximas seções, identificar o valor de α é importante tanto para o cálculo do intervalo de confiança como para o seu uso.

3.1. Intervalo de Confiança para a Média

Considere que a variável quantitativa de interesse (X) tenha distribuição Normal, com média cujo valor μ queremos estimar, e com variância desconhecida. A margem de erro na estimação de μ com 100(1-α)% de confiança é dada por:

$$me_{\mu}^{100(1-\alpha)\%} = t_{[n-1;\alpha/2]}\frac{s}{\sqrt{n}} \qquad [1.1]$$

e, desse modo, o intervalo de confiança de 100(1-α)% para a média populacional μ é dado por

$$IC_{\mu}^{100(1-\alpha)\%} = \left[\bar{x} - t_{[n-1;\alpha/2]}\frac{s}{\sqrt{n}}\ ;\ \bar{x} + t_{[n-1;\alpha/2]}\frac{s}{\sqrt{n}}\right], \qquad [1.2]$$

no qual $\bar{x}$ e *s* e são, respectivamente, a média e o desvio-padrão na amostra de tamanho *n*, α é a redução na confiança e $t_{[n-1;\alpha/2]}$ é o valor na distribuição t-*Student* com *n-1* graus de liberdade que deixa uma probabilidade $\alpha/2$ acima dele (Tabela 2).

Exemplo 2.3.1: Em um experimento com uma amostra de *n*=20 crianças, a idade média ao falar foi de $\bar{x}$ = 10 meses com desvio-padrão de *s* = 1,5 meses.
Vamos construir um intervalo de 95% de confiança para a média μ da idade ao falar na população de crianças de onde esta amostra foi retirada. Como 100(1-α)=95, então α=0,05 e $\alpha/2$=0,025.
Na tabela T, tem-se que $t_{[19;\ 0,025]}$ =2,093. A margem de erro é igual a $(2,093)\left(1,5/\sqrt{20}\right)=0,7$ meses. Assim,

[3] Um intervalo de 100% de confiança teria limites infinitos, sendo, portanto, completamente inútil.

$$IC_{\mu}^{95\%} = [10 - 0{,}7 \ ; \ 10 + 0{,}7] = [9{,}3 ; 10{,}7] \text{meses.}$$

Estima-se, com 95% de confiança, que a média da idade ao falar das crianças desta população esteja entre 9,3 e 10,7 meses.

Exemplo 2.3.2: Suponha que a amostra do exemplo anterior tenha sido maior, digamos, 41 crianças. Neste caso, $t_{[40;\ 0{,}025]}$=2,021 e a margem de erro seria menor, pois $(2{,}021)\left(1{,}5/\sqrt{41}\right) = 0{,}5$ meses.

O efeito do tamanho da amostra aparece tanto na divisão por valor maior em $\sqrt{n}$, quanto pelo menor valor na tabela *t-Student* para um número maior de graus de liberdade (*g.l.*=40).

Isto acontece porque, na distribuição t-*Student*, as caudas ficam mais leves com o aumento dos graus de liberdade que, no intervalo de confiança em [1.2], crescem com o tamanho da amostra *n*. Isto significa que o valor de $t_{[n-1;\ \alpha/2]}$ e, consequentemente, a margem de erro, fica cada vez menor com o aumento do tamanho da amostra.

O gráfico da Figura 3.1 mostra o efeito do tamanho da amostra (de 10 a 100 crianças) na margem de erro deste exemplo: quanto maior o tamanho da amostra, menor a margem de erro (*me*). Mas note que a relação entre *n* e *me* não é linear: a curva que descreve a relação entre *n* e *me* "cai" mais rapidamente no início (*n* menores) e mais lentamente no final (*n* maiores). Veja que, quando passamos de *n*=10 para *n*=20, a redução na margem de erro (de 0,70 para 0,56 anos) é maior do que de *n*=80 para *n*=90 (de 0,33 para 0,31 anos), mesmo que o aumento absoluto na amostra tenha sido de 10 crianças nos dois casos.

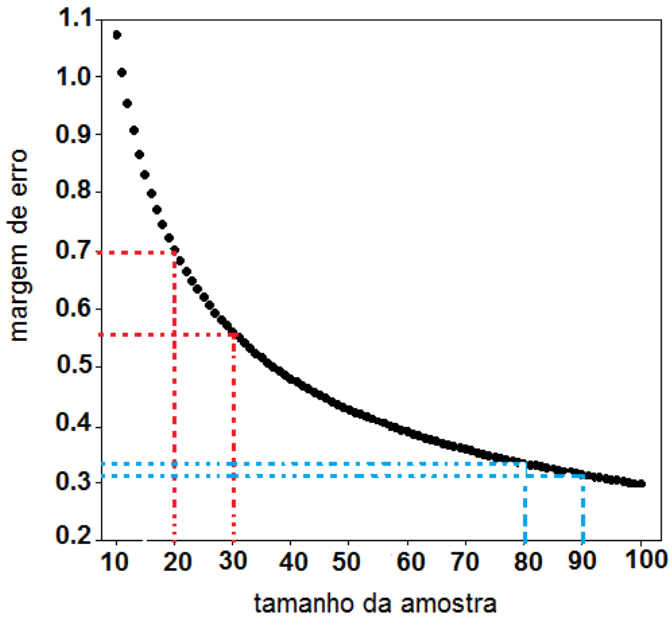

Figura 3.1. Margem de erro (em anos) em função do tamanho da amostra no IC95% para a média do Exemplo 2.3.1 (*s*=1.5 anos).

3.2. Intervalo de Confiança para a Proporção

O intervalo de confiança de 100(1-α)% para a proporção populacional *p* é dado por

$$IC_p^{100(1-\alpha)\%} = \left[\hat{p} - z_{[\alpha/2]}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \;;\; \hat{p} + z_{[\alpha/2]}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \qquad [2]$$

na qual $\hat{p}$ é a proporção observada na amostra de tamanho *n*, e $z_{[\alpha/2]}$ é o valor na distribuição Normal Padronizada) que deixa uma probabilidade $\alpha/2$ acima dele (Tabela 1).

Exemplo 2.4.1: Deseja-se saber a eficácia de um novo tratamento contra micose em adultos, ou seja, deseja-se estimar *proporção p de pessoas que seriam curadas* com o novo tratamento. Uma amostra de 50 pessoas doentes foi tratada com o novo tratamento e 40 deles foram curadas, gerando uma proporção amostral $\hat{p}$ = 40/50 = 0,80. Como 100(1-α)=95, então α=0,05, α/2=0,025 e $z_{[\alpha/2]} = z_{[0,025]}$ =1,96. O intervalo de confiança de 95% é dado por

$$IC_p^{95\%} = \left[0.80 \pm 1{,}96\sqrt{\frac{0.80(0.20)}{50}}\right] = \left[0.80 \pm 0.11\right] = \left[0.69\,;0.91\right].$$

Assim, com base nesta amostra, estimamos que a proporção de cura com o novo tratamento está entre 69% e 91%, com 95% de confiança.

Note que a margem de erro do IC95% é de 11 p.p. (pontos percentuais). Se a amostra fosse dez vezes maior (n=500), a margem de erro seria de 3,5 p.p. Para um valor fixo de $\hat{p}$, a relação entre o tamanho da amostra e a margem de erro é do mesmo tipo daquela mostrada na Figura 3.1.

Para um tamanho de amostra fixo, a margem de erro também depende da estimativa $\hat{p}$ através da expressão $\hat{p}(1-\hat{p})$, que atinge seu valor máximo de 0,25 em $\hat{p}$ = 0,5 (Figura 3.2).

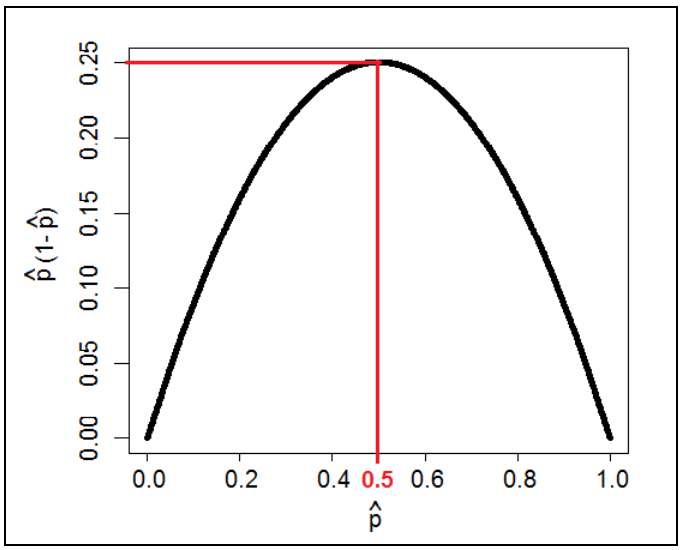

Figura 3.2. Valor do termo $\hat{p}(1-\hat{p})$ da margem de erro para uma proporção em função do valor da proporção amostral $\hat{p}$, atingindo o máximo quando $\hat{p} = 0{,}5$.

Para valores fixos de n e $\hat{p}$, a margem de erro diminui se reduzimos o nível de confiança do intervalo. No exemplo, fazendo 100(1-α)=90, então α=0,10, α/2=0,05 e $z_{[\alpha/2]}=z_{[0,05]}$=1,64; desse modo, a margem de erro do IC90% é dada por (1,64)(0.0566)=0,09, menor que a margem de erro do IC95%.

3.3. Intervalo de Confiança para a Variância (e Desvio-Padrão)

Diferentemente da média e da proporção, no caso da variância os limites do intervalo de confiança não são calculados como estimativa *pontual* $\pm$ *margem de erro*, não sendo, portanto, equidistantes da estimativa pontual. Isto acontece porque a distribuição amostral da variância amostral não é simétrica, como vimos na Seção 1.2.

O intervalo de confiança de 100(1-α)% para a variância σ^2 de uma variável com distribuição Normal é dado por

$$IC_{\sigma^2}^{100(1-\alpha)\%} = \left[\frac{(n-1)s^2}{Q_{[n-1;\alpha/2]}} \; ; \; \frac{(n-1)s^2}{Q_{[n-1;1-\alpha/2]}} \right] \qquad [3]$$

na qual s^2 é a variância amostral (ou seja, s é o desvio-padrão amostral), n é o tamanho da amostra e $\chi^2_{[n-1;\ \alpha/2]}$ é o valor na tabela Qui-quadrado com *n-1* graus de liberdade que deixa uma probabilidade igual α/2 acima dele e $Q_{[n-1;\ 1-\alpha/2]}$ é o valor na tabela Qui-quadrado com *n-1* graus de liberdade que deixa uma probabilidade igual 1-α/2 acima dele (Tabela 3).

Se quisermos o **intervalo de confiança para o desvio-padrão σ**, basta tomar a raiz quadrada nos limites inferior e superior do intervalo para a variância σ^2.

Exemplo 2.5.1: Deseja-se estimar o desvio padrão das medidas de uma balança de precisão. Um objeto com peso conhecido de 1 grama foi pesado 30 vezes na balança, obtendo-se um desvio-padrão das medidas igual a 0,02 gramas (variância igual a 0,0004). Supondo que as medidas do peso sigam uma distribuição Normal, o intervalo de 95% de confiança para a variância dos pesos é dados por

$$IC_{\sigma^2}^{95\%} = \left[\frac{(30-1)(0,02)^2}{Q_{[29;0,025]}} \; ; \; \frac{(30-1)(0,02)^2}{Q_{[29;0,975]}} \right] = \left[\frac{0.0116}{45,772} \; ; \; \frac{0.0116}{16,047} \right] = [0,00025;0,00072].$$

Assim, o intervalo de 95% de confiança para desvio-padrão dos pesos é calculado como

$$IC_{\sigma}^{95\%} = \left[\sqrt{0,00025}; \sqrt{0,00072} \right] = [0,016;0,027] \text{ gramas.}$$

3.4. Interpretação do Intervalo de Confiança

Um erro comum na interpretação do IC é dizer que "o valor de μ está no IC100(1-α)% com probabilidade (1-α)". Mas como μ é uma constante (desconhecida, mas fixa) e não uma variável aleatória, não faz sentido falar em "probabilidade de que μ assuma este ou aquele valor".

O que podemos dizer é que, se fizermos um grande número de intervalos nestas condições, aproximadamente 100(1-α)% desses intervalos conterão de fato o verdadeiro valor de μ (que permanece desconhecido), é esta ideia que é traduzida por "confiança". O gráfico da Figura 3.3

ilustra esta interpretação: das 10 amostras retiradas, 9 (ou seja, 90%) produziram um IC que contém o valor (desconhecido) de μ.

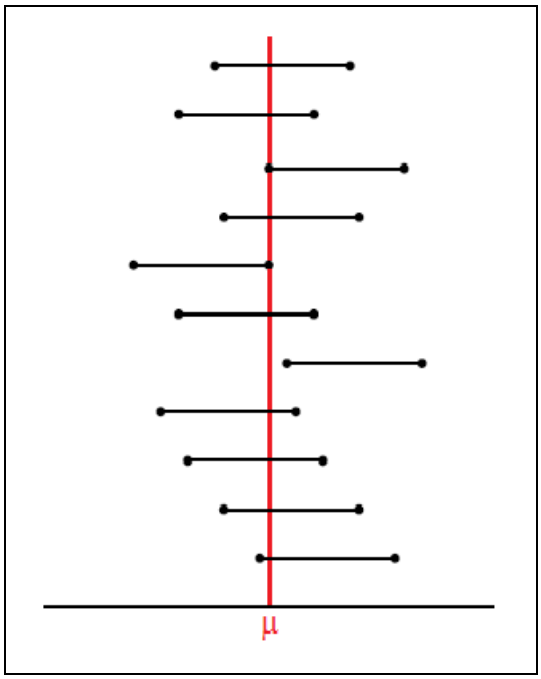

Figura 3.3. Representação de intervalos de 90% confiança para a média populacional construídos a partir de diferentes amostras de mesmo tamanho obtidas da população. Note que 9 dos 10 intervalos contém o valor μ, reprentando o nível de confiança de 90%.

4. CÁLCULO DO TAMANHO DA AMOSTRA EM ESTIMAÇÃO

Como vimos anteriormente, a ligação entre a margem de erro nos intervalos de confiança e o tamanho da amostra é bem clara. De fato, em muitos estudos a coleta da amostra é planejada tal que seu tamanho é calculado em função de um valor máximo *d* para a margem de erro.

4.1. Média

No caso da estimação da média populacional, a margem de erro para 100(1-α)% de confiança é igualada ao valor desejado *d*:

$$t_{[n-1;\alpha/2]}\frac{s}{\sqrt{n}} = d. \qquad [4.1]$$

Note que o termo *n* não pode ser "isolado" na equação, pois ele está "preso" nos graus de liberdade. Deste modo, não conseguimos uma "fórmula fechada" para o tamanho da amostra, mas podemos aumentar o tamanho da amostra até conseguir a margem de erro desejada.

Exemplo 2.6.1: No Exemplo 2.3.1 (idade ao falar de crianças), deseja-se estimar a média da idade ao falar com margem de erro *d*=0,4 anos (considerando s=1,5) em um intervalo de confiança de 95%. Segundo o gráfico na Figura 3.4 (uma réplica da Figura 3.2), o tamanho da amostra deve ser de 58 crianças.

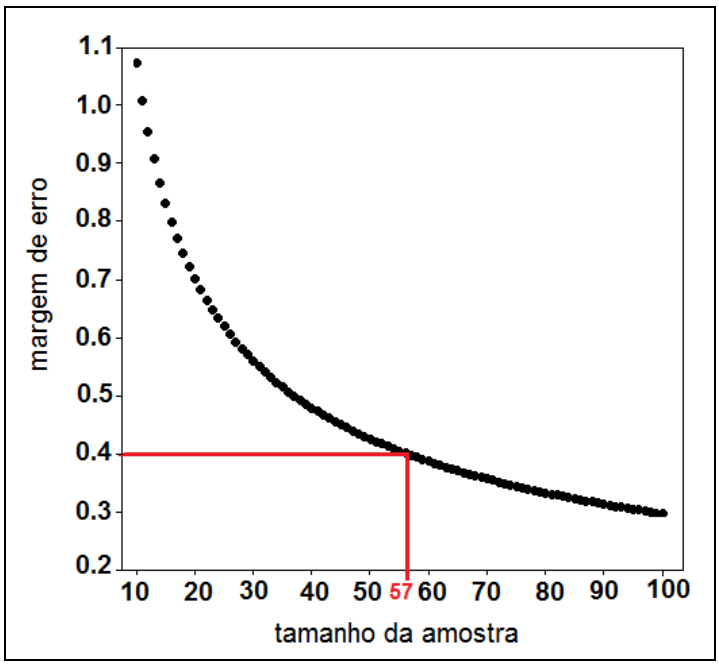

Figura 3.4. Margem de erro (em anos) em função do tamanho da amostra no IC95% para a média do Exemplo 2.3.1 (*s*=1,5 anos), usando a distribuição T. Tamanho de amostra mínimo (n = 57,02 ≈ 58) para margem de erro máxima de 0,4 anos.

Para amostras grandes, podemos substituir o percentil $t_{[n-1;\alpha/2]}$ por $z_{[\alpha/2]}$ (da tabela Normal Padrão), obtendo a seguinte expressão para o tamanho da amostra:

$$n = \left(\frac{z_{[\alpha/2]} \cdot s}{d} \right)^2 . \quad , \qquad [4.2]$$

sempre arredondando para cima.

No exemplo, para uma margem de erro *d*=0,4 no IC95% (considerando s=1,5), o tamanho da amostra pela fórmula anterior (com o percentil da Normal) é dado por $[(1,96)(1,5)/(0,4)]^2 = 54.02 \approx 55$ crianças, valor muito próximo do cálculo n=58 encontrado no gráfico da Figura 3.4 (no qual foi usada a fórmula com o percentil da distribuição T).

Tanto em [4.1] quanto em [4.2], temos que escolher um valor para s. Podemos usar o valor encontrado em uma amostra piloto (como foi feito no exemplo da Figura 3.4) ou "chutar" um valor. Uma alternativa é definir o tamanho da amostra em função da razão *s/d*. Por exemplo, queremos um tamanho de amostra tal que a margem de erro *d* seja um quarto do desvio-padrão *s*, ou seja, que d=s/4 o que equivale a dizer que s/d=4. Em geral, desejamos valores da margem de erro que sejam menores do que o desvio-padrão amostral, ou seja, d/s < 1, o que implica em s/d > 1.

No exemplo, o tamanho da amostra pela fórmula [4.2] seria dado diretamente por $n = [(1,96)*(4)]^2 = 61.5 \approx 62$. Pela fórmula iterativa [4.1] (cálculos não mostrados), chegaríamos a *n* = 64. De fato, podemos mostrar que, para um mesmo valor s/d, o tamanho da amostra requerido pelo cálculo exato do IC95% (e de 90%) via distribuição t-*Student* é sempre três unidades a mais do que o valor calculado (arredondado para cima) pela aproximação Normal.

4.2. Proporção

Na fórmula [2.1], a margem de erro pode ser igualada ao valor desejado *d* e reescrita em função do tamanho da amostra mínimo como

$$n = \hat{p}(1-\hat{p})\left(\frac{z[\alpha/2]}{d}\right)^2 . \quad , \qquad [5.1]$$

Note que esta fórmula depende de uma estimativa $\hat{p}$. Você pode tomar $\hat{p}$ como uma estimativa anterior ou mesmo um "chute". Caso não tenha um bom palpite, você pode fazer $\hat{p}$ =0,5, que gera o valor máximo de 0,25 para $\hat{p}(1-\hat{p})$ (Figura 3.3), tornando a expressão para o tamanho mínimo da amostra igual a:

$$n = (0.25)\left(\frac{z_{[\alpha/2]}}{d}\right)^2 .$$

No cálculo de n para 95% de confiança (o mais usual), se arredondamos 1,96 para 2, a fórmula simplifica para $n = (1/d^2)$, ou seja, o tamanho da amostra é inversamente proporcional ao quadrado da margem de erro desejada.

Exemplo 2.6.2: Considerando novamente o exemplo 2.4.1, suponha que se queira estimar a eficácia do novo tratamento com margem de erro máxima de 1 p.p. (0,01) com 95% de confiança. Neste caso, usando a estimativa $\hat{p}$ = 0,9, o tamanho mínimo da amostra deve ser

$$n \geq 0{,}9(1-0.9)\left(\frac{1.96}{0.01}\right)^2 = (0{,}09)(38416) = 3458.$$

Se fosse usado o valor $\hat{p}$ = 0,5, o tamanho mínimo requerido para a amostra seria muito maior,

$$n^* \geq (0.25)\left(\frac{1{,}96}{0{,}01}\right)^2 = 9604 .$$

4.3. Variância (e Desvio-padrão)

Assim como no caso da média, na equação do intervalo de confiança para a variância não conseguimos isolar o valor de n:

$$IC_{\sigma^2}^{100(1-\alpha)\%} = \left[\frac{(n-1)s^2}{Q_{[n-1;\alpha/2]}} \; ; \; \frac{(n-1)s^2}{Q_{[n-1;1-\alpha/2]}}\right] = [LI; LS].$$

Desse modo, sugerimos duas alternativas de abordagem:

A) Como a margem de erro inferior deste intervalo (s^2 - LI) é menor que a margem de erro superior (LS-s^2), podemos nos concentrar em encontrar o tamanho de amostra para limitar a margem de erro superior ao valor escolhido *d*. Ou seja, calcular o *n* mínimo tal que (LS - s^2) ≤ *d*, com um "chute" para o valor de s^2 ou seu valor estimado em uma amostra piloto.

B) Podemos escolher o tamanho da amostra que limite o valor da largura do intervalo de confiança (LS-LI) em relação à magnitude da variância (s^2). Ou seja, que escolher o valor de *n* tal que (LS-LI)/s^2 < R, onde R é um valor escolhido pelo pesquisador, como 1, 1.5, 2, 10, etc. Isto nos leva a encontrar o *n* tal que (LS-LI)/s^2 = (n-1)*[(1/QS)-(1/QI)] < R, onde QI = $Q_{[n-1;\alpha/2]}$ e QS=$Q_{[n-1;1-\alpha/2]}$. Note que não precisamos mais "chutar" o valor de s^2, pois ele "sumiu" da expressão.

Exemplo 2.6.3: No Exemplo 2.5.1, a amostra de tamanho n=30 e com variância amostral s^2= 0,0004 nos levou ao IC95% = [0,00025;0,00072]. Note que a margem de erro inferior deste intervalo (=0,00040-0,00025=0,00015) é menor que a margem de erro superior (=0,00072-0,0004=0,00032) e que a largura do intervalo é igual a LS-LI=0,00072-0,00025=0,00047.

A) Usando valor de s^2=0,0004, suponha que desejamos calcular o tamanho de amostra n tal que a margem de erro superior do IC95% não ultrapasse 0,00015 (cerca de metade do valor de 0,00032 no estudo piloto). Assim, queremos que

$$[(n-1)(0,0004)/Q_{[n-1;0,975]}] - 0,0004 \leq 0,00015$$

$$(n-1)(0,0004)/Q_{[n-1;0,975]} \leq 0,00055$$

$$(n-1)\ /Q_{[n-1;0,975]} \leq 1.375.$$

Os cálculos a seguir (com valores de Q retirados da Tabela 3) nos levam a n=91.

n	g.l.=n-1	$Q_{0,975}$	$(n-1)/Q_{0,975}$
30	29	16,047	1,81
31	30	16,791	1,79
41	40	24,433	1,64
51	50	32,357	1,55
61	60	40,482	1,48
71	70	48,758	1,44
81	80	57,153	1,40
91	**90**	65,647	**1,37**

B) Na amostra piloto (n=30), temos, no IC95% construído, que (LS-LI)/s^2=0,00047/0,0004=1,75. Suponha que desejamos calcular o tamanho de amostra n tal que (LS-LI)/$s^2 \leq 0,7$, ou seja, que

$$(n-1)*[(1/QS)-(1/QI)] \leq 0,7.$$

Os cálculos a seguir (com valores de Q retirados da Tabela 3) nos levam a n=71.

n	g.l.=n-1	$QI_{[0,025]}$	$QS_{[0,975]}$	(n-1)*[(1/QS)-(1/QI)]
30	29	45,722	16,047	1.173
31	30	46,979	16,791	1.148
41	40	59,342	24,433	0.963
51	50	71,420	32,357	0.845
61	60	83,298	40,482	0.762
71	70	95,023	48,758	**0.699**

4. Uso do Intervalo de Confiança para Testar Hipóteses

Além de estimar o valor de parâmetro da população, podemos estar interessados em compará-lo com um valor de referência, geralmente retirado da literatura sobre o problema ou que pertence a outra população. Por exemplo, podemos fazer as seguintes perguntas:

1) A média da renda familiar per capita dos alunos de uma universidade é igual a 2000 reais?
2) A proporção de alunos da universidade que trabalham (p) é igual a 0,2?
3) O desvio-padrão das medidas de uma balança (σ) é igual a 0,010 gramas ?

Nesses três casos temos uma *hipótese* acerca do valor do parâmetro de interesse: $\mu=2000$; $p=0,2$ e $\sigma=0,010$, respectivamente. Quando essa hipótese é confrontada com sua hipótese contrária ($\mu\neq2000$; $p\neq0,2$ e $\sigma\neq0,010$, respectivamente), temos o que é chamado de **Teste de Hipóteses**.

De modo geral, para um parâmetro θ qualquer, testamos uma *hipotese nula* (denotada por H_0) e uma *hipótese alternativa*[4] (denotada por H_1): $H_0: \theta = \theta_0$ é testada contra $H_1: \theta \neq \theta_0$. Nos exemplos,

1) Testaremos $H_0: \mu=2000$ contra $H_1: \mu\neq2000$.
2) Testaremos $H_0: p=0,2$ contra $H_1: p\neq0,2$.
3) Testaremos $H_0: \sigma = 0,010$ contra $H_1: \sigma\neq0,010$

Podemos usar o resultado do intervalo de confiança para o parâmetro θ para decidir entre H_0 e H_1. Basta verificar se o valor de está ou não dentro do intervalo: **se θ_0 (valor de θ sob H_0) estiver "dentro do intervalo de confiança, "não rejeitamos com H_0"; caso contrário, "rejeitamos H_0".**

No Exemplo 2.3.1, o parâmetro de interesse é a média da idade ao falar (em meses) das crianças de uma população (μ). Suponha que um pesquisador queira saber se $H_0: \mu=9$ ou $H_1: \mu\neq9$. Como o IC95%=[9,3 ; 10,7] não contém o valor 9, podemos rejeitar a hipótese nula e concluir que $\mu\neq9$.

No Exemplo 2.4.1, suponha que o pesquisador afirme que a proporção de cura com o novo tratamento contra micose em adultos (p) seja de 0,75. O IC90%=[0,71 ; 0,89] nos mostra que esta hipótese não pode ser rejeitada, pois o valor 0,75 está dentro do intervalo.

No Exemplo 2.5.1, a hipótese nula de que desvio-padrão dos pesos da balança é igual a 0,010 gramas é rejeitada, pois o IC95%=[0,016 ; 0,027] gramas não contém o valor 0,010.

O nível de confiança do intervalo é denotado por $100(1-\alpha)\%$, com $0\leq\alpha\leq1$. Na nomenclatura de Teste de Hipóteses, este valor de $\boldsymbol{\alpha}$ é chamado de **nível de significância do teste**. Desse modo, para um IC90%, o nível de significância é de $\alpha=0,10$ (10%); para um IC95%, $\alpha=0,05$ (5%) e assim por diante.

Quando escolhemos um valor para α, estamos escolhendo o valor máximo tolerado para a probabilidade de cometermos o **Erro do Tipo I**, que é o erro de *rejeitar H_0 quando H_0 é verdadeira*. No caso do Exemplo 2, o Erro do Tipo I seria dizer que a proporção de estudantes que trabalham diferente de 0,2 quando ela é, na verdade, igual a 0,2.

O Teste de Hipóteses também é uma técnica de *Inferência Estatística*, assim como os Intervalos de Confiança. Para saber mais, consulte Triola (2013).

[4] A hipótese alternativa de um teste também pode levar os sinais < ou > (ex: $\mu < 2000$). São os chamados testes unilaterais. No entanto, no caso desse tipo de teste de hipóteses, o IC não pode ser usado.

REFERÊNCIAS

- Bolfarine, H.; Bussab, W. (2005) ***Elementos de Amostragem*** (1ª Edição). Editora Blucher, 290 pp.

- Triola, M.F. (2013) ***Introdução à Estatística*** (11ª edição). Editora LTC.

- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.