Estatística e Informática

Aula 10 - Estimação e Intervalo de Confiança

Alan Rodrigo Panosso alan.panosso@unesp.br

Departamento de Ciências Exatas FCAV/UNESP

(09-05-2024)

Parâmetro e Estatística

Parâmetro: é uma medida usada para descrever uma característica da população.

Estatística ou **Estimador**: é qualquer função de uma amostra aleatória (fórmula ou expressão), construída com o propósito de servir como instrumento para descrever alguma característica da amostra e para fazer *inferência* a respeito da característica na população.

Resumo	Parâmetro	Estatística
Média	μ	$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Variância	σ^2	$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2$
Proporção	π	$\hat{p} = \frac{X}{n}$

O valor numérico da estatística ou estimador de um parâmetro, calculado para uma amostra observada, é chamado de **estimativa desse parâmetro**.

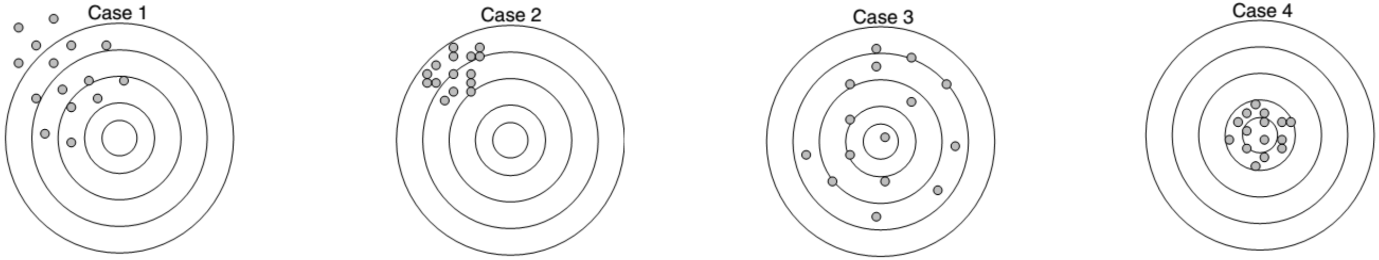
A diferença entre estatística e estimativa é que a **estatística** é uma variável aleatória, e a estimativa é um particular valor dessa variável aleatória.

Acurácia

A acurácia mede quão próximo o valor estimado está do valor real, ou seja, é a habilidade do estimador de estimar o valor real.

Precisão

A Precisão mede quão próximas estimativas individuais estão umas das outras, ou seja é a habilidade do estimador de estimar valores similares de maneira consistente.



Baixa Acurácia e Baixa Precisão | Baixa Acurácia e Alta Precisão | Alta Acurácia e Baixa Precisão | Alta Acurácia e Alta Precisão

Propriedades de um bom estimador

1) Consistência: é uma propriedade por meio da qual a acurácia de uma estimativa aumenta quando o tamanho da amostra aumenta. Assim dado um parâmetro populacional θ e sendo $\hat{\theta}$ o estimador desse parâmetro. As condições suficientes para um estimador ser consistente são:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta$$

$$\lim_{n\to\infty} Var(\hat{\theta}) = 0$$

Exemplo

$$E(\bar{X}) = \mu \text{ e } Var(\bar{X}) = \frac{Var(X)}{n}$$

Propriedades de um bom estimador

2) Não viciado ou não viesado: O estimador $\hat{\theta}$ como uma variável aleatória, tem uma certa distribuição em repetidas amostras de tamanho n. Não viciado é uma propriedade que assegura que, em média, o estimador é correto:

O **estimador** é chamado **não viciado** ou **imparcial** se seu valor esperado ou médio for igual ao verdadeiro valor do parâmetro, ou seja:

$$E(\hat{\theta}) = \theta$$

Entretanto, se

$E(\hat{\theta}) = \theta + b(\theta)$ com $b(\theta) \neq 0$,

o estimador é **viciado** e a quantidade $b(\theta)$ é chamada vício ou viés.

Exemplos de Estimadores

Estes estimadores nada mais são do que as próprias definições dos respectivos parâmetros, mas aplicadas à amostra:

$E(\bar{X}) = \mu$ e $E(\hat{p}) = p$

Por sua vez, para a variância o estimador populacional $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^2$ é viciado, pois, podemos demonstrar que:

$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{1}{n}\sigma^2$

onde o viés $b(\sigma^2) = -\frac{1}{n}\sigma^2$.

Abaixo segue o estimador não viciado para variância:

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

No entanto, para $n \to \infty$, têm-se para ambos os estimadores convergem para σ^2, ou seja, $\hat{\sigma}^2$ e s^2 são assintoticamente não viciados.

Acesse o Link para estudarmos essas Propriedades

https://arpanosso.shinyapps.io/estatinfo/

Estimativa por Ponto e por Intervalo

Estimativa por ponto

É a estimativa de um parâmetro populacional dada por um único valor para a estatística, exemplo:

$$\hat{X} = \mu$$

esse procedimento não permite julgar qual a possível magnitude do erro que se está cometendo.

Exemplo: O diâmetro a altura do peito de árvores de Eucalipto tem uma média de $105\ cm$,

Estimativa por intervalo

É a estimativa de um parâmetro populacional baseada na distribuição amostral do estimador pontual, dada por dois valores a e b $(a < b)$, entre os quais se considera que o parâmetro esteja contido.

Essas estimativas indicam a sua precisão ou acurácia, por isto são preferíveis às estimativas por ponto.

A declaração da precisão de uma estimativa por intervalo denomina-se grau de confiança ou **nível de confiança**, daí a denominação de **Intervalo de Confiança**.

Exemplo: O diâmetro a altura do peito de árvores de Eucalipto tem uma média de $105 \pm 0,05\ cm$,

Estimativa por Intervalo de Confiança

Estimativa por intervalo de confiança

Um intervalo de confiança para θ é um intervalo construído a partir das observações da amostra, de modo que ele inclui o verdadeiro e desconhecido valor de θ, **com uma específica e alta probabilidade** denotada por $1-\alpha$, é tipicamente tomada como:

$$NC = P(a \leq \theta \leq b) = 1 - \alpha$$

Então, o intervalo $]a, b[$ é chamado intervalo com $100 \cdot (1-\alpha)\%$ de confiança para o parâmetro θ, onde: 1- α é o **nível de confiança** associado ao intervalo a e b são os **limites de confiança**, inferior e superior, respectivamente, do intervalo.

Onde temos a seguinte relação:

Nível de Confiança (NC)	**Nível de significância** (α)
$0,90$	$0,10$
$0,95$	$0,05$
$0,99$	$0,01$

Intervalo de Confiança para a Média Populacional (μ)

Precisamos definir 4 casos:

(a) Caso em que amostras são grandes $(n \geq 30)$ e σ conhecido;

(b) Caso em que amostras são grandes $(n \geq 30)$ e σ desconhecido;

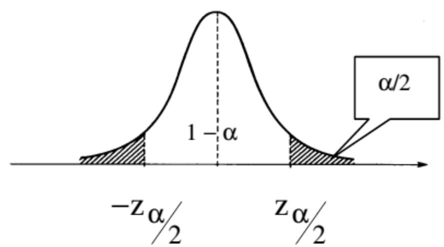
(c) Caso em que as amostras são pequenas $(n < 30)$ σ conhecido;

(d) Caso em que as amostras são pequenas $(n < 30)$ e σ desconhecido.

(a) Caso em que amostras são grandes $(n \geq 30)$ e σ conhecido.

O desenvolvimento de intervalos de confiança para μ é baseado na distribuição amostral de $\bar{X}$ se o tamanho da amostra (n) é grande:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \cong N(0,1)$$



onde:
$a = \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ e $b = \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

$$\begin{cases} \frac{\sigma}{\sqrt{n}} = \sigma_{\bar{X}} = \text{erro padrão da média} \\ z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = \text{erro da estimativa da média} \end{cases}$$

Exemplo

Se $1 - \alpha = 0,95$ nesse caso $\alpha = 0,05$

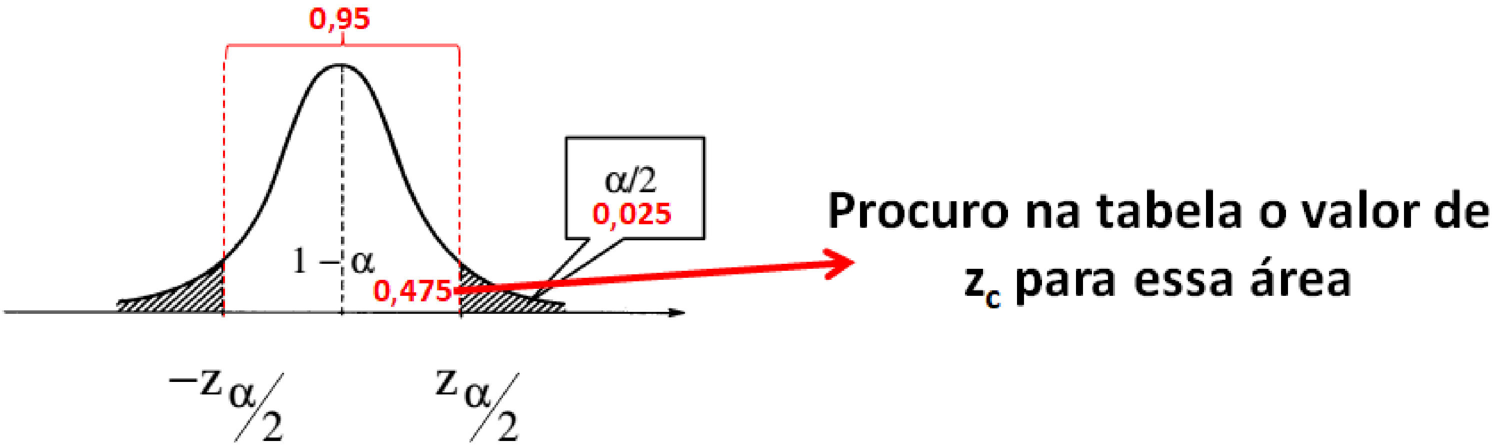


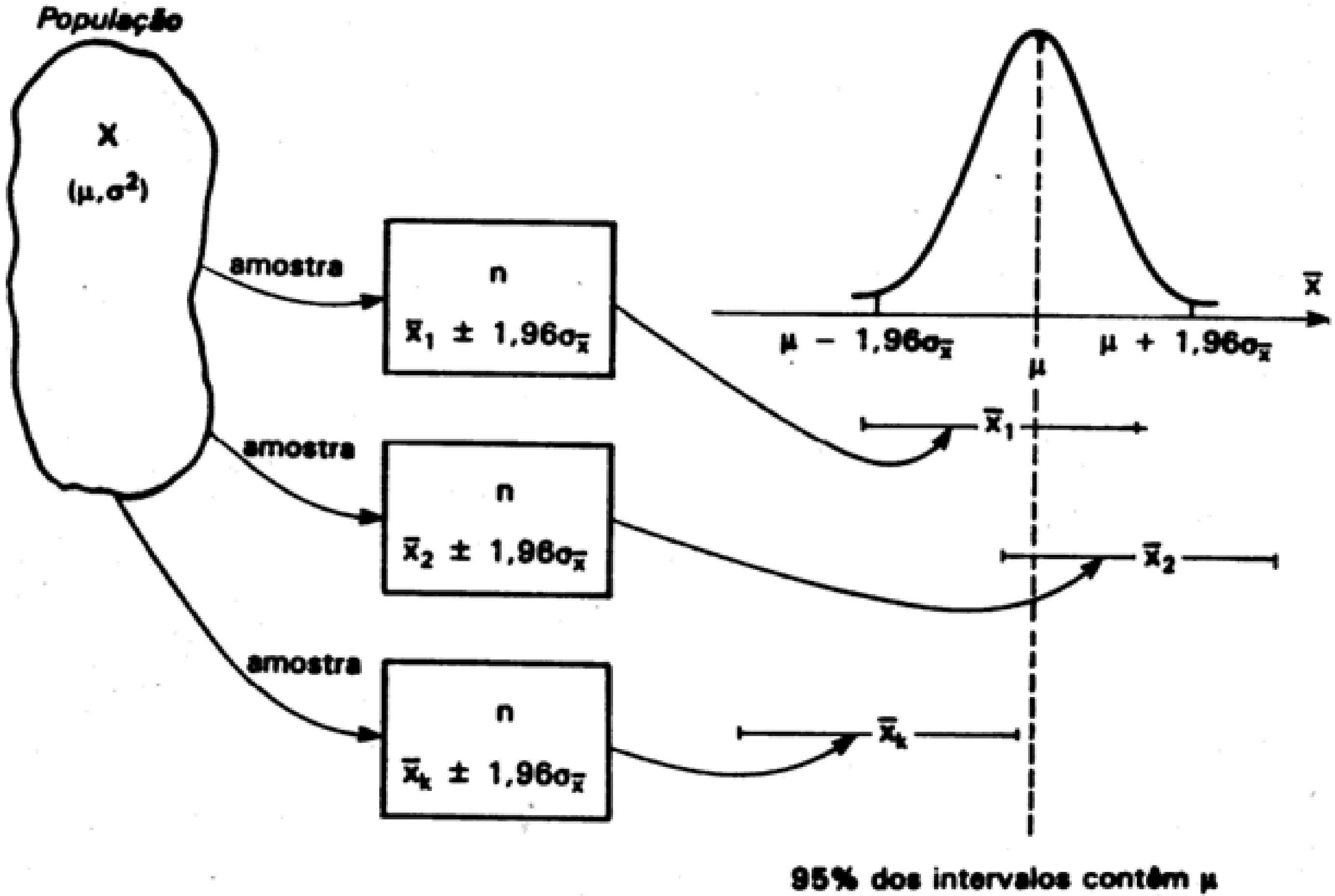
Tabela - Normal Padrão

primeira decimal de Z_c	0	1	2	3	4	5	6	7	8	9	primeira decimal de Z_c
	p = 0										
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0

$$z_{\alpha/2} = 1{,}96$$

$$P(\bar{X} - 1{,}96\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1{,}96\,\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

Esta expressão, deve ser interpretada do seguinte modo: construídos todos os intervalos da forma $\bar{X} \pm 1,96\sigma_{\bar{X}}$, 95% deles conterão μ. Lembrando que μ não é uma variável aleatória, mas um parâmetro, isto é, não é o mesmo que dizer que μ tem 95% de probabilidade de estar entre os limites indicados.



Então, denotamos o intervalo de confiança como:

$$IC(\mu; 1-\alpha) =]\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}[$$

Exemplo

Considerando uma amostra de 100 animais da raça Nelore, onde o peso médio a desmama é $171,70\ kg$, encontre um IC de 95% para μ, supondo que o desvio padrão da população (σ) seja igual a $7,79\ kg$.

$$IC(\mu; 95\%) =]171,70 \pm 1,96 \frac{7,79}{\sqrt{100}}$$

$$IC(\mu; 95\%) =]170,17\ kg; 173,23\ kg[$$

(b) Caso em que amostras são grandes $(n \geq 30)$ e σ desconhecido.

Como n é grande, a substituição de σ pelo desvio padrão amostral (s) não afeta apreciavelmente a estimativa de IC, assim, temos que:

$$IC(\mu; 1-\alpha) =]\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}; \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}[$$

(c) Caso em que as amostras são pequenas $(n < 30)$ e σ conhecido.

Se $X_1, X_2, \cdots, X_n$ é uma amostra aleatória de uma população com distribuição normal $N(\mu, \sigma^2)$, a média amostral $\bar{X}$ é exatamente distribuída como $N(\mu, \frac{\sigma^2}{n})$. Sendo σ conhecido, o $IC(\mu : 1 - \alpha)$ é dado por:

$$IC(\mu; 1-\alpha) =]\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}[$$

(d) Caso em que as amostras são pequenas $(n < 30)$ e σ desconhecido.

Fato que ocorre na maioria dos casos, uma aproximação intuitiva é substituir σ por s considerar a razão:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Essa substituição causa uma considerável diferença se a amostra for pequena. A notação t é requerida porque s aumenta a variância de t para um valor maior do que um (1), de modo que a razão não é padronizada.

A distribuição da razão t é conhecida como **distribuição** t **- Student** com parâmetro $r = n - 1$ graus de liberdade.

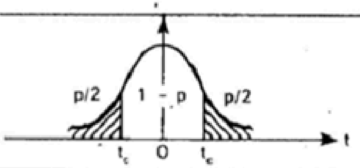
Distribuição t de Student

Tabela - Distribuição t-Student

TÁBUA V

Distribuição de Student: St(n)

Valores críticos de t tais que $P(-t_c < t < t_c) = 1 - p$



Graus de liberdade	p = 90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%	Graus de liberdade
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	15,894	31,821	63,657	318,309	636,619	1
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	4,849	6,965	9,925	22,327	31,598	2
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	3,482	4,541	5,841	10,214	12,924	3
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	2,998	3,747	4,604	7,173	8,610	4
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	2,756	3,365	4,032	5,893	6,869	5
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	2,612	3,143	3,707	5,208	5,959	6
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,517	2,998	3,499	4,785	5,408	7
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,449	2,896	3,355	4,501	5,041	8
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,398	2,821	3,250	4,297	4,781	9
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,359	2,764	3,169	4,144	4,587	10
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,328	2,718	3,106	3,025	4,437	11
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,303	2,681	3,055	3,930	4,318	12
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,282	2,650	3,012	3,852	4,221	13
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,264	2,624	2,977	3,787	4,140	14
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,248	2,602	2,947	3,733	4,073	15
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,235	2,583	2,921	3,686	4,015	16
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,224	2,567	2,898	3,646	3,965	17
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,214	2,552	2,878	3,610	3,922	18
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,205	2,539	2,861	3,579	3,883	19
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,197	2,528	2,845	3,552	3,850	20
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,189	2,518	2,831	3,527	3,819	21
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,183	2,508	2,819	3,505	3,792	22
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,177	2,500	2,807	3,485	3,768	23
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,172	2,492	2,797	3,467	3,745	24
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,166	2,485	2,787	3,450	3,725	25
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,162	2,479	2,779	3,435	3,707	26
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,158	2,473	2,771	3,421	3,690	27
28	0,127	0,256	0,389	0,530	0,684	0,855	1,056	1,313	1,701	2,048	2,154	2,467	2,763	3,408	3,674	28
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,150	2,462	2,756	3,396	3,659	29
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,147	2,457	2,750	3,385	3,646	30
35	0,126	0,255	0,388	0,529	0,682	0,852	1,052	1,306	1,690	2,030	2,133	2,438	2,724	3,340	3,591	35
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,123	2,423	2,704	3,307	3,551	40
50	0,126	0,254	0,387	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,109	2,403	2,678	3,261	3,496	50
60	0,126	0,254	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,099	2,390	2,660	3,232	3,460	60
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,076	2,358	2,617	3,160	3,373	120
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,054	2,326	2,576	3,090	3,291	∞
	p = 90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%	

Distribuição t de Student

As distribuições t são simétricas em torno de zero mas têm caudas mais espalhadas do que a distribuição $N(0,1)$. Entretanto, com o aumento de r, a distribuição t se aproxima da distribuição $N(0,1)$, pois $Var(t)$ tende à unidade (1).

$$\begin{cases} E(t) = 0 \\ Var(t) = \frac{r}{r-2} = \frac{n-1}{n-3} \end{cases}$$

Pode-se concluir da distribuição t, que

$$P(-t_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \leq +t_{\frac{\alpha}{2}}) = 1-\alpha$$

em que $t_{\frac{\alpha}{2}}$ é obtido na tabela da distribuição t com $r = n-1$ graus de liberdade, ou seja:

$$IC(\mu; 1-\alpha) =]\bar{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}; \bar{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}[$$

Exercício

Uma amostra de 10 cães sofrendo de uma determinada doença apresentou um tempo de sobrevivência médio de $46,9$ meses e o desvio padrão de $43,3$ meses. Determinar os limites de confiança de 90% para μ.

Solução: $\bar{x}_a = 46,9$ meses $\quad s = 43,3$ meses

$1 - \alpha = 0,90 \qquad n - 1 = 9 \qquad t_{\alpha/2} = 1,833$

Limites de confiança para μ: $\bar{x} \pm t_{\alpha/2} \dfrac{s}{\sqrt{n}} = 46,9 \pm 1,833 \dfrac{43,3}{\sqrt{10}} = 21,8 \;\; e \;\; 72,0 \; meses$

Portanto, $\quad IC(\mu : 90\%) =]21,8; \; 72,0[$

Intervalo de Confiança para a Proporção

Fazendo uso do fato que, para n grande, a distribuição binomial pode ser aproximada com a normal:

$$Z = \frac{x - n.p}{\sqrt{n.p.q}} = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

Temos:

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{x - np}{\sqrt{np(1-p)}} \leq +z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Substituindo p, visto que é desconhecido, por seu estimador $\hat{p}$ dentro das raízes, obtêm-se:

$$IC(\hat{p}; 1-\alpha) = \left]\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}; \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right[$$

Exemplo

Suponha que em $n = 400$ animais são administrados uma droga, obtendo $X = 320$ sucessos, ou seja, 80% dos animais melhoraram. A partir destes dados, obtenha um IC para p, com $1 - \alpha = 0,90$.

primeira decimal de Z_c	0	1	2	3	4	5	6	7	8	9	primeira decimal de Z_c
	p = 0										
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0

$$z_{\alpha/2} = 1{,}64$$

$$P\left(\bar{X} - 1{,}64\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1{,}64\frac{\sigma}{\sqrt{n}}\right) = 0{,}90$$

Exemplo

Suponha que em $n = 400$ animais doentes são administrados uma droga, obtendo $X = 320$ sucessos, ou seja, 80% dos animais melhoraram. A partir destes dados, obtenha um IC para p, com $1 - \alpha = 0,90$.

Solução: $\hat{p} = 320/400 = 0,80 \qquad \hat{q} = 0,20$

$$IC = 0,80 \pm 1,64\sqrt{\frac{0,80.0,2}{400}} =]0,767;0,833[$$

Portanto, IC(p : 90%) =]0,767 ; 0,833[