Licap

Formação Cientista de Dados

PUC Minas

Licap

Formação do Cientista de Dados

Estatística Descritiva – Módulo Básico

Luis Enrique Zárate

Conteúdo do Curso

Licap

Estatística descritiva

1. Medidas de tendência central
2. Histograma de frequências
3. Distribuição normal
4. Distribuição normal padronizada
5. Análise de correlação

Estatística descritiva - Medidas de tendência central

Licap

Media: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mediana: $$\tilde{x} = x_{([n+1]/2)}$$

para "n" impar

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

para "n" par

Desvio Padrão: $$S(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

n-1: para amostra; n: para população

Variância: $$Var(x) = S^2(x)$$

Seja o conjunto de dados

81,80	87,10	82,70	79,80	81,30	79,50	88,50	75,90
81,60	73,90	84,50	87,10	82,00	79,30	82,50	87,10
83,00	87,30	79,70	82,00	83,60	84,50	80,40	78,10
86,40	76,70	83,70	78,40	76,00	80,90	80,20	78,90
77,40	78,50	82,90	81,90	80,70	78,40	78,00	81,40
84,60	79,50	82,30	80,50	80,70	79,00	90,00	79,90
86,80	80,10	83,20	78,20	80,40	85,50	85,50	79,30
83,00	78,10	83,40	83,60	85,70	86,80	86,50	83,80
86,80	83,50	79,90	76,60	84,30	78,50	74,40	71,80
79,10	82,10	84,50	78,40	80,70	70,70	78,50	85,20

Medidas de Tendência Central	
Média	81,44
Mediana	81,35
D. Pad	3,79
Variância	14,36

Estatística descritiva - Histograma de frequências

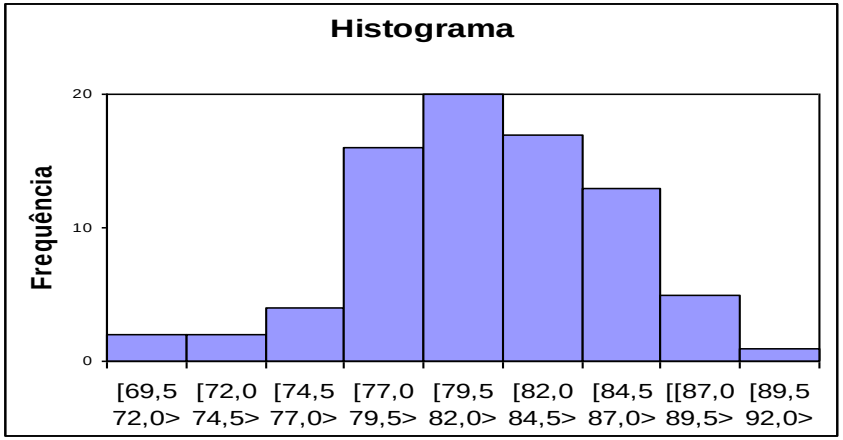
Licap

Seja o conjunto de dados

81,80	87,10	82,70	79,80	81,30	79,50	88,50	75,90
81,60	73,90	84,50	87,10	82,00	79,30	82,50	87,10
83,00	87,30	79,70	82,00	83,60	84,50	80,40	78,10
86,40	76,70	83,70	78,40	76,00	80,90	80,20	78,90
77,40	78,50	82,90	81,90	80,70	78,40	78,00	81,40
84,60	79,50	82,30	80,50	80,70	79,00	90,00	79,90
86,80	80,10	83,20	78,20	80,40	85,50	85,50	79,30
83,00	78,10	83,40	83,60	85,70	86,80	86,50	83,80
86,80	83,50	79,90	76,60	84,30	78,50	74,40	71,80
79,10	82,10	84,50	78,40	80,70	70,70	78,50	85,20

Procedimentos		
Número de Dados =	80	
Num. Intervalos K =	9	
Mínimo (MIN) =	70,70	
Máximo (MAX) =	90,00	
Amplitude R =	19,30	
Comp. Interv h=R/(K-1)	2,41	
Arredondamento h =	2,50	
Li1 = MIN - h/2	69,45	69,50
Ls1 = Li1+h	72,00	
Li2 = Ls1	72,00	
Ls2 = Li2+h	74,50	

Intervalo i	Limites	Ponto Médio	Freq. Simples	Freq. Relativa
1	[69,5 72,0>	70,75	2,00	0,0250
2	[72,0 74,5>	73,25	2,00	0,0250
3	[74,5 77,0>	75,75	4,00	0,0500
4	[77,0 79,5>	78,25	16,00	0,2000
5	[79,5 82,0>	80,75	20,00	0,2500
6	[82,0 84,5>	83,25	17,00	0,2125
7	[84,5 87,0>	85,75	13,00	0,1625
8	[[87,0 89,5>	88,25	5,00	0,0625
9	[89,5 92,0>	90,75	1,00	0,0125
		Total:	80,00	1,0000



Distribuição Normal

Licap

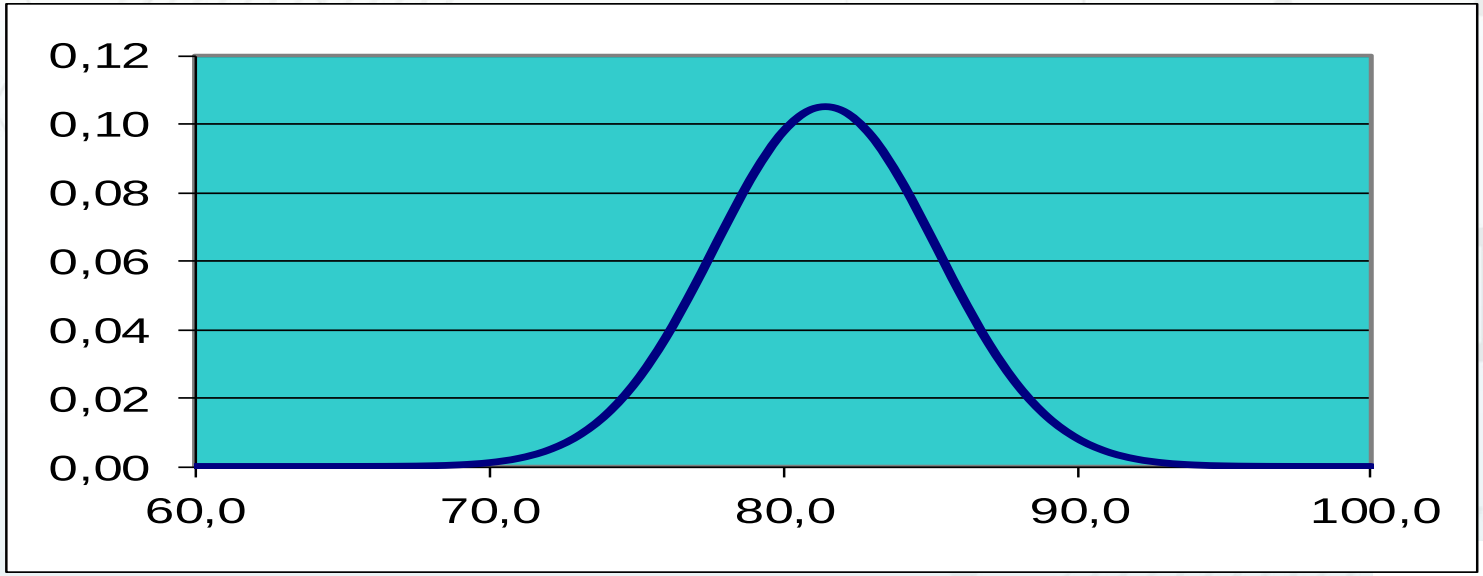
A distribuição normal é um modelo estatístico que fornece uma base teórica para o estudo do padrão de ocorrência dos elementos de uma população.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

$\mu(x)$ média da população (ou da amostra)

$\sigma(x)$ desvio padrão da população (ou da amostra)

Distribuição Normal



Medidas de Têndencia Central	
Média	81,44
Mediana	81,35
D. Pad	3,79
Variância	14,36

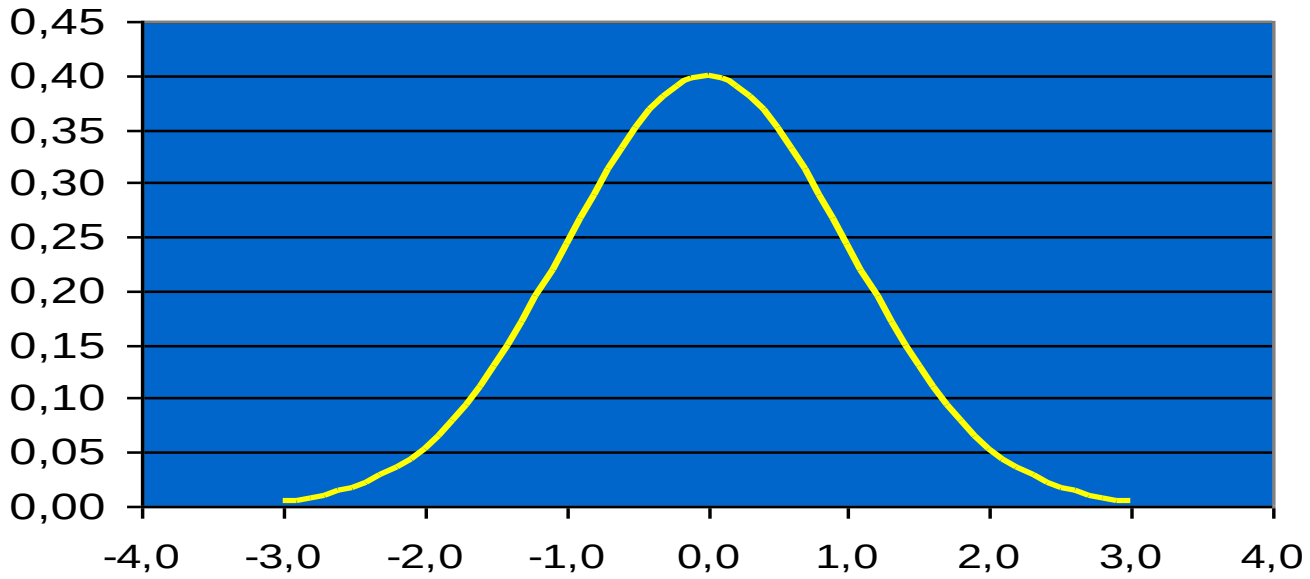
Intervalo	Probabilidade	
	Interna	Externa
$\mu \pm 1\sigma$	68,2%	31,74%
$\mu \pm 2\sigma$	95,46%	4,54%
$\mu \pm 3\sigma$	99,73%	0,27%

$N(\mu,\sigma)$

PUC Minas

Distribuição Normal Padronizada

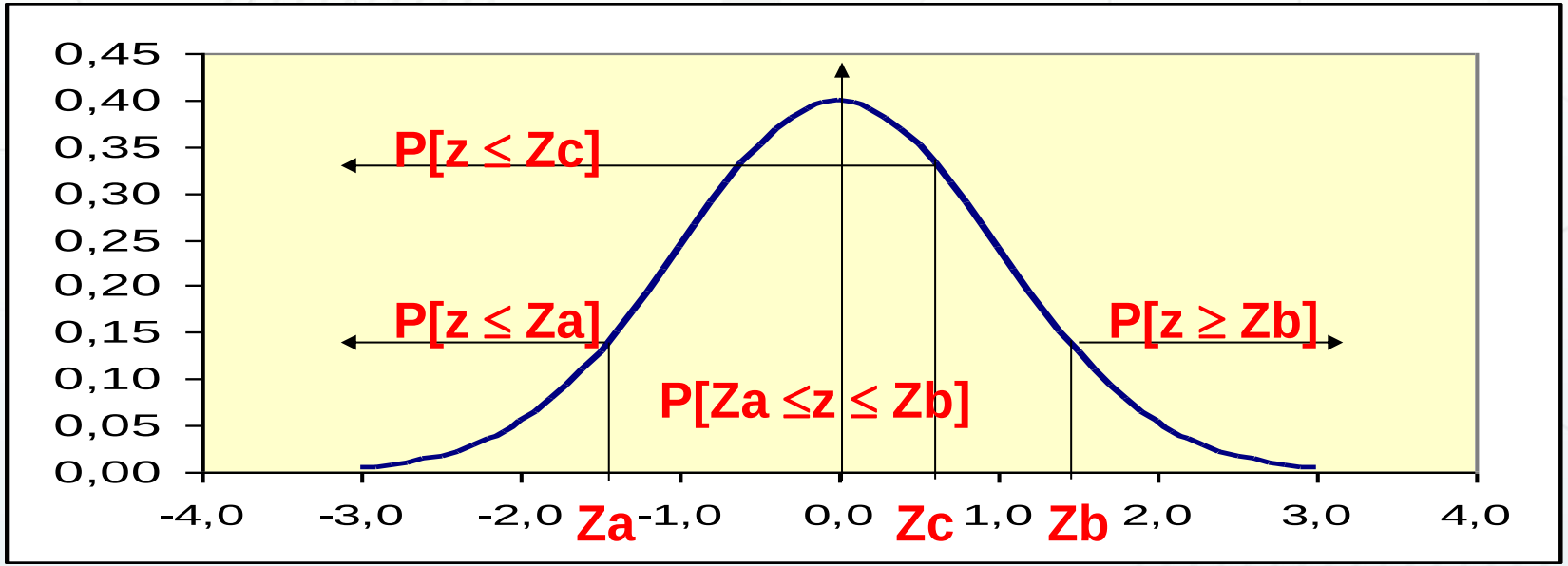
Para calcular probabilidades associadas a uma variável $x \sim N(\mu,\sigma)$ é comúm aplicar uma transformação sobre a variável x para obter a variável normal padronizada z. Onde $z \sim N(0,1)$.



$$z = \frac{x - \mu}{\sigma}$$

N(0,1)

Entendendo o cálculo de Probabilidades



P[z ≤ Za]: probabilidade de z ser inferior ou igual a Za

P[Za ≤ z ≤ Zb] = P[z ≤ Zb] - P[z < Za]: probabilidade de z estar entre Za e Zb

Entendendo o cálculo de Probabilidades

Exemplos:

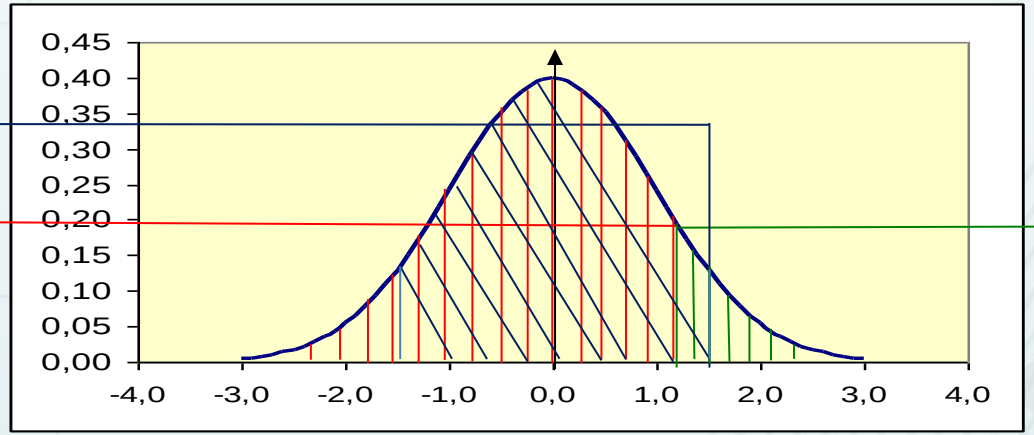
$P[z \leq 1{,}18] = 0{,}8810$

$P[z > 1{,}18] = 1 - P[z \leq 1{,}18] = 1 - 0{,}8810 = 0{,}1190$

$P[-1{,}40 \leq z \leq 1{,}82] = P[z \leq 1{,}82] - P[z < -1{,}40] = 0{,}9656 - 0{,}0808 = 0{,}8848$

P[-1,40 ≤ z ≤ 1,82] = 88,40%

P[z ≤ 1,18] = 88,10%



P[z > 1,18] =11,90%

Distribuição de Poisson

Licap

É uma distribuição discreta de probabilidades utilizada para eventos periódicos ou intervalares, tais como:

a) Número de chamadas telefônicas durante um dia;

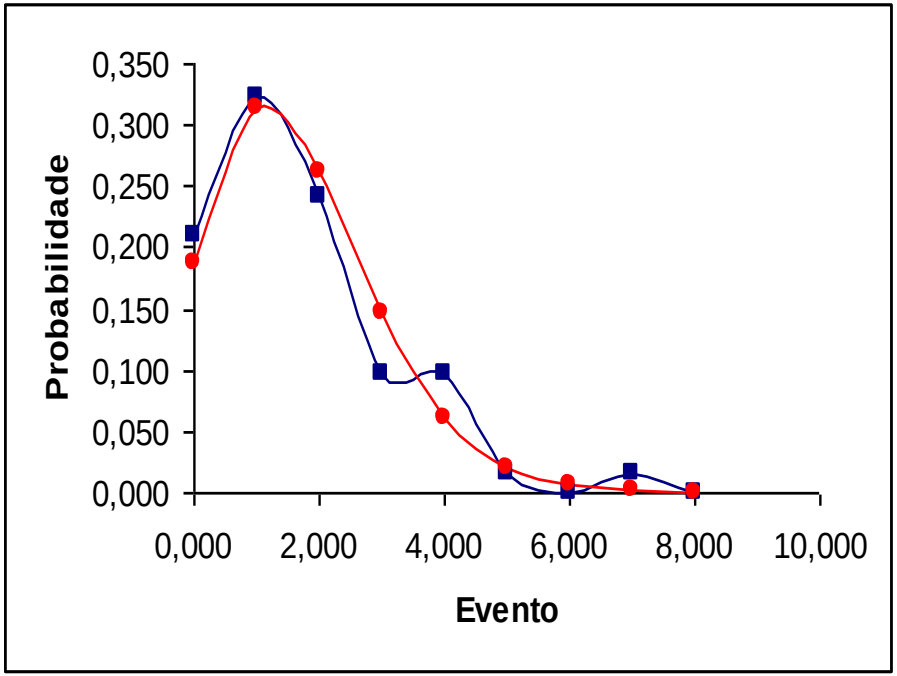
b) Número de acidentes de transito, numa cidade, durante um período do dia;

c) Número de consultas a uma página Web durante uma semana; etc.

Distribuição de Poisson

Licap

evento (xi)	Freq. Simples (fi)	Freq. Relativa (hi)	xi * fi	Dist. Poisson
0,000	13,000	0,210	0,000	0,187
1,000	20,000	0,323	20,000	0,313
2,000	15,000	0,242	30,000	0,263
3,000	6,000	0,097	18,000	0,147
4,000	6,000	0,097	24,000	0,062
5,000	1,000	0,016	5,000	0,021
6,000	0,000	0,000	0,000	0,006
7,000	1,000	0,016	7,000	0,001
8,000	0,000	0,000	0,000	0,000
	62,000	1,000	104,000	
		Média=	1,677	



$$f(x) = \frac{\alpha^x}{x!}\exp(-\alpha) \qquad \alpha = \frac{\sum x_i f_i}{n}$$

Análise de Correlação de Pearson

Na prática, é muitas vezes essencial estudar a relação entre duas variáveis associadas.

x	y	x	y	x	y
8,6	0,889	8,4	0,894	8,7	0,896
8,9	0,884	8,2	0,864	9,3	0,928
8,8	0,874	9,2	0,922	8,9	0,886
8,8	0,891	8,7	0,909	8,9	0,908
8,4	0,874	9,4	0,905	8,3	0,881
8,7	0,886	8,7	0,892	8,7	0,882
9,2	0,911	8,5	0,877	8,9	0,904
8,6	0,912	9,2	0,885	8,7	0,912
9,2	0,895	8,5	0,866	9,1	0,925
8,7	0,896	8,3	0,896	8,7	0,872

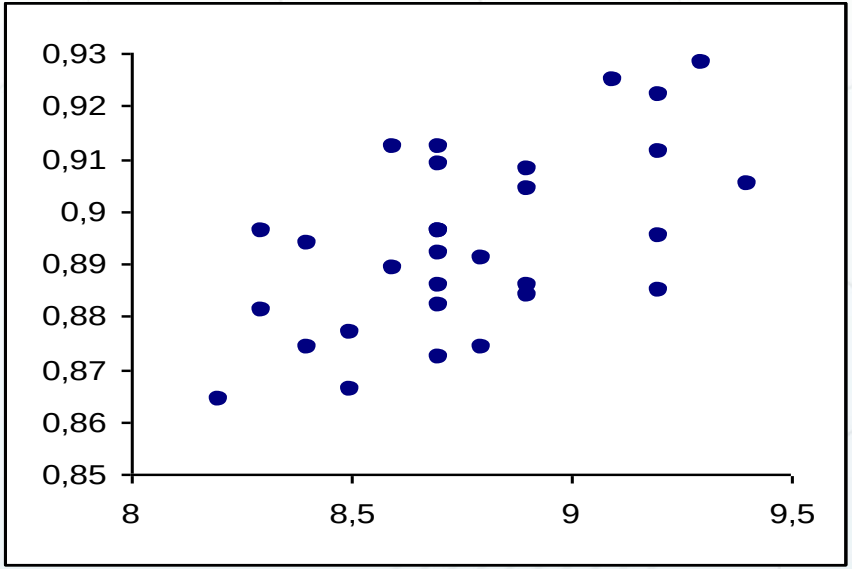
$$r = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}} \qquad -1 \leq r \leq +1$$

Análise de Correlação

$$Sxx = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$Syy = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$Sxy = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$



Sxx = 2,88 Syy = 0,00840 Sxy = 0,59 r = 0,59

0.9 correlação muito forte.
0.7 a 0.9 correlação forte.
0.5 a 0.7 correlação moderada.
0.3 a 0.5 correlação fraca.
0 a 0.3 correlação desprezível.

Licap

Formação Cientista de Dados

Obrigado!

PUC Minas