Licap

Formação Cientista de Dados

PUC Minas

Formação do Cientista de Dados

Estatística Inferêncial – Módulo Básico

Luis Enrique Zárate

Licap

Conteúdo do Curso

Licap

Teoria Amostral

1. População finita e Infinita
2. Intervalos de confiança
3. Tamanho da amostra
4. Tipo de amostragem
5. Processo de amostragem

Teoria Amostral

Licap

N = Tamanho da população
$\mu(x)$ = média da população
$\sigma(x)$ = desvio padrão da população

n = Tamanho da Amostra
X = média da amostra
S(x) = desvio padrão da amostra

Para trabalhar com amostra e definir o tamanho desta, devemos fixar o nível de confiança desejado e o erro máximo tolerável. Esta análise é feita para cada variável.

População Finita e Infinita

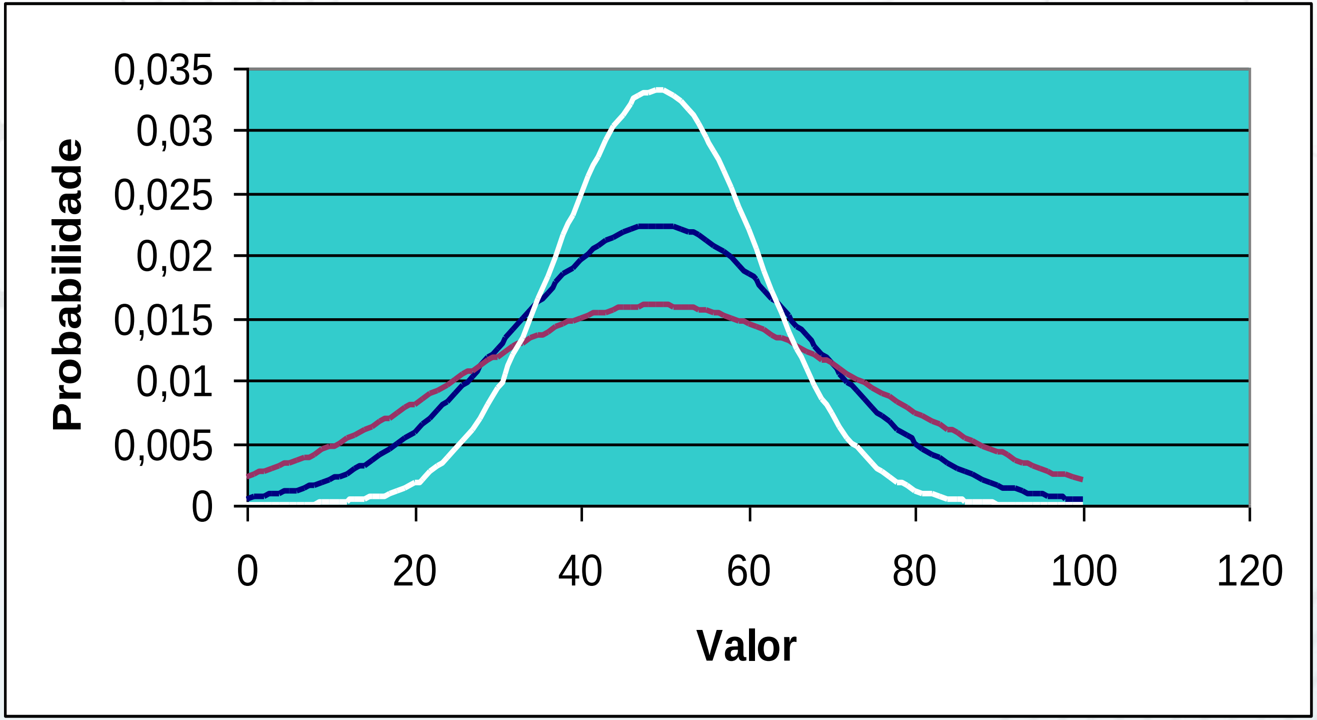
Licap

Caso não seja conhecido o tamanho da população (**N**), a mesma é considerada infinita, onde (**n**) representa o tamanho da amostra.

Caso N seja conhecido:

Se $(20*n)<N$ a população é considerada infinita

Se $(20*n)\geq N$ a população é considerada finita



Qual amostra representa melhor a população ??

Amostragem

A obtenção de uma AMOSTRA é uma tarefa a ser executada cuidadosamente, pois através dela serão tomadas decisões sobre a população.

A representatividade envolve a determinação do número “n” de itens e da técnica a ser usada na sua obtenção.

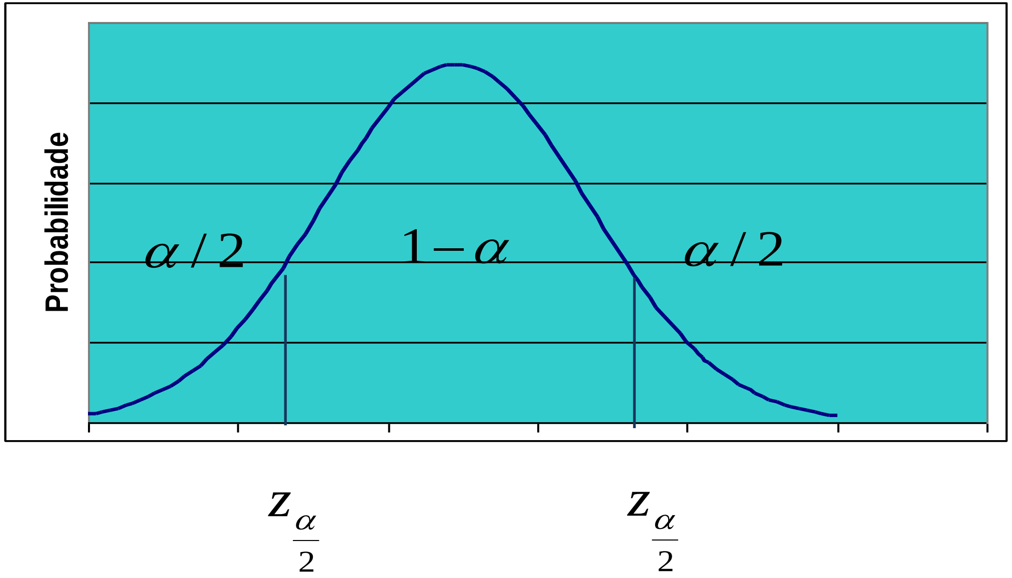
$$n=f(\sigma,\alpha,\beta,\delta)$$

onde:

σ = dispersão da população
α = risco por rejeitar informação válida
β = risco por aceitar informação inválida
δ = distância entre o parâmetro populacional e o amostral (por exemplo na média)

Intervalos de Confiança

A partir de uma amostra, extraída **aleatoriamente** de uma população com distribuição Normal N(μ,σ) é possível construir um **intervalo de confiança** para o parâmetro populacional (μ ou σ) desconhecido, com probabilidade 1 - α, chamado de **nível de confiança**, de que o **intervalo** contenha o verdadeiro parâmetro. Onde α é chamado de **nível de significância**.



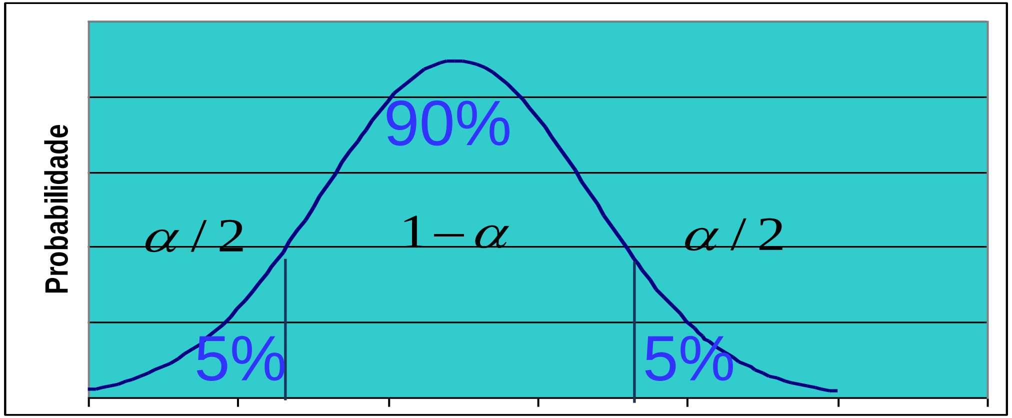
Intervalos de Confiança

Por exemplo:

Para o nível de significância α = 10%

O nível de confiança é: 1 - α = 90%

Intervalo de confiança: $\theta 1 \leq \theta \leq \theta 2$ 90% de probabilidade que o parâmetro θ se encontre dentro dos limites do intervalo.



$-z_{\frac{\alpha}{2}}$ $z_{\frac{\alpha}{2}}$

IC- para a média populacional $\mu(x)$

Licap

Com $\sigma(x)$ conhecido

Exemplo 1:

amostra: 20,25,27,30,32,35,38,40,37,23 com X=30,70; $\sigma(x)=6$;

para $\alpha=6\%$ (0,06); z=1,881

Solução:

$$P(X-e \leq \mu(x) \leq X+e) = 1-\alpha$$

onde $e=z.[\sigma(x)/(n)^{½}]=3,57$

$$P(30,70 - 3,57 \leq \mu(x) \leq 30,70 + 3,57)=94\%$$

$$P(27,13 \leq \mu(x) \leq 34,27) = 94\%$$

IC- para a média populacional $\mu(x)$

Licap

Com $\sigma(x)$ desconhecido

Exemplo 3:

amostra: 20,25,27,30,32,35,38,40,37,23 com X=30,70; S(x)=6,83; para α=5% (0,05); t=2,2622 (Dist. "t" de Student para n<30)

Solução:

$$P(X - e \leq \mu(x) \leq X + e) = 1 - \alpha$$

onde $e = t.[s(x)/(n)^{½}] = 4,89$

$$P(30,70 - 4,89 \leq \mu(x) \leq 30,70 + 4,89) = 95\%$$

$$P(25,81 \leq \mu(x) \leq 35,59) = 95\%$$

IC- para a média populacional $\mu(x)$

Licap

Com N e $\sigma(x)$ conhecidos

Exemplo 2:

amostra: 20,25,27,30,32,35,38,40,37,23 com X=30,70;

N=150; $\sigma(x)=6$; para $\alpha=6\%$ (0,06); z=1,881

Solução:

$$P(X-e \leq \mu(x) \leq X+e) = 1-\alpha$$

onde $e=z\{\sigma(x)/[(N-n)/(N-1)]^{1/2}\}=3{,}46$

$$P(30{,}70 - 3{,}46 \leq \mu(x) \leq 30{,}70 + 3{,}46)=94\%$$

$$P(27{,}24 \leq \mu(x) \leq 34{,}16) = 94\%$$

Tamanho da amostra univariável

Licap

Com $\sigma(x)$ e N desconhecidos

Exemplo:

Quanto deve ser "n" para nível α=4% com erro máximo tolerável de e=0,45 e S(x)=3,87 obtido de uma amostra piloto.

Solução:

para α=0,04, z=2,054

$$n = [z.(S(x)/e)]^2$$

$$n = [2{,}054.(3{,}87/0{,}45)]^2 = 312{,}51$$

$$\Rightarrow n = 313$$

Tamanho da amostra univariável

Licap

Com $\sigma(x)$ desconhecida e N conhecido

Exemplo:

Quanto deve ser "n" se N=3000, para nível α=4% com erro máximo tolerável de e=0,45 e S(x)=3,87 obtido de uma amostra piloto.

Solução:

para α=0,04, z=2,054

$$n = [N.(z.S)^2] / [(N-1)e^2+ (z.S)^2]$$

$$n = 283,11$$

$$\Rightarrow n = 283$$

IC- para uma proporção populacional

Licap

Tipo de Filme	N° de pessoas
Aventura	180
Comédia	240
Erótico	320
Ficção	300
Musical	80
Policial	260
Romance	240
Terror	**380**
Outros	40
Total	2040

Exemplo 5:

Construir Intervalo de Confiança para: Pessoas (p) que gostam de "terror" ao nível $\alpha=5\%$

Solução:

$x=380$; $n=2040$; $f=x/n=0{,}1863$

Para $\alpha=0{,}05$; $z=1{,}96$

$$P(f-e \leq p \leq f+e) = 1-\alpha$$

onde $e=z[(f(1-f)/n)^{½}]=0{,}0169$

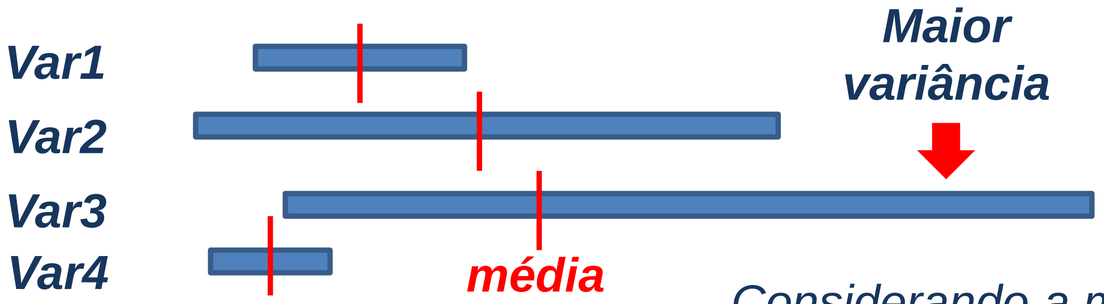
$$P(0{,}1694 \leq p \leq 0{,}2032) = 95\%$$

$$P(16{,}94\% \leq p(\text{terror}) \leq 20{,}32\%) = 95\%$$

Tamanho da amostra para o problema multivariável

Data Mining lida com problemas multivariáveis. Daí pode ser **difícil** encontrar o tamanho de amostra ideal que atenda ao problema multivariável.

Utilizando as técnicas estatísticas apresentadas, para lidar com o problema multivariável deverá ser escolhida a variável (atributo) que apresente maior variabilidade, é dizer maior variância ou desvio padrão.



Considerando a mesma escala para as variáveis

Tipos de Amostragem

Licap

Com reposição (over-sampling):

Onde o elemento sorteado volta a ser parte da população.

O número total de amostras de tamanho “n” que pode-se retirar, com reposição, de uma população de tamanho “N” é dado pela expressão:

$$(ACR)_{N,n} = N^n$$

Exemplo:
N=30
n=6
ACR=729000000 *amostras*

Tipos de Amostragem

Licap

<u>Sem reposição (under-sampling):</u>

Onde o elemento sorteado não volta ser parte da população (testes destrutivos)

O número total de amostras de tamanho “n” que pode-se retirar, sem reposição, de uma população de tamanho “N” é dado pela expressão:

$$(ASR)_{N,n} = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Exemplo:
N=30
n=6
ASR=593775 *amostras*

Processo de Amostragem

Licap

Técnica de Amostragem Aleatória:

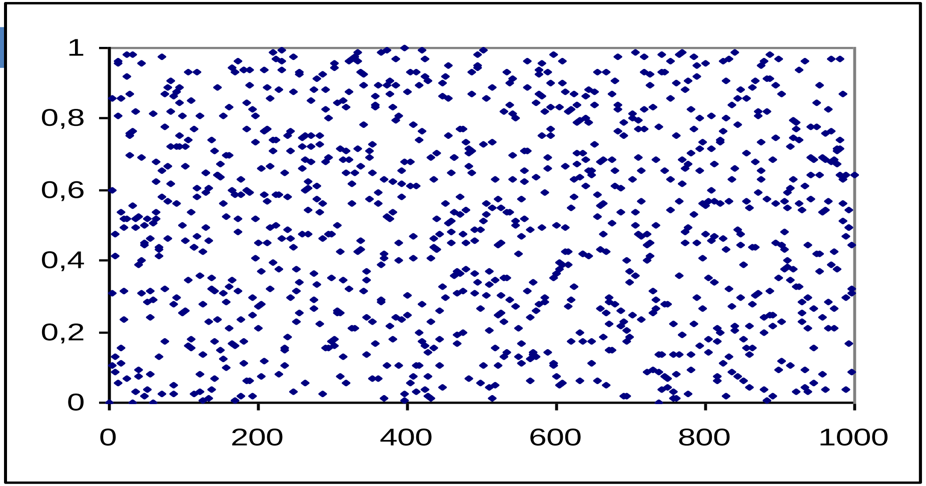
A idéia básica é que os itens sejam obtidos por sorteio, cuidando que se anule toda tendência possível (bias).

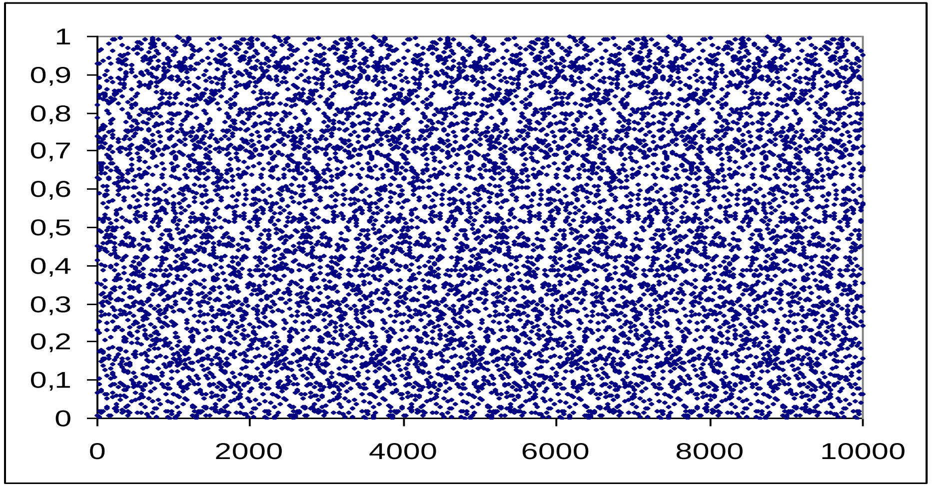
Geração de Números aleatórios :

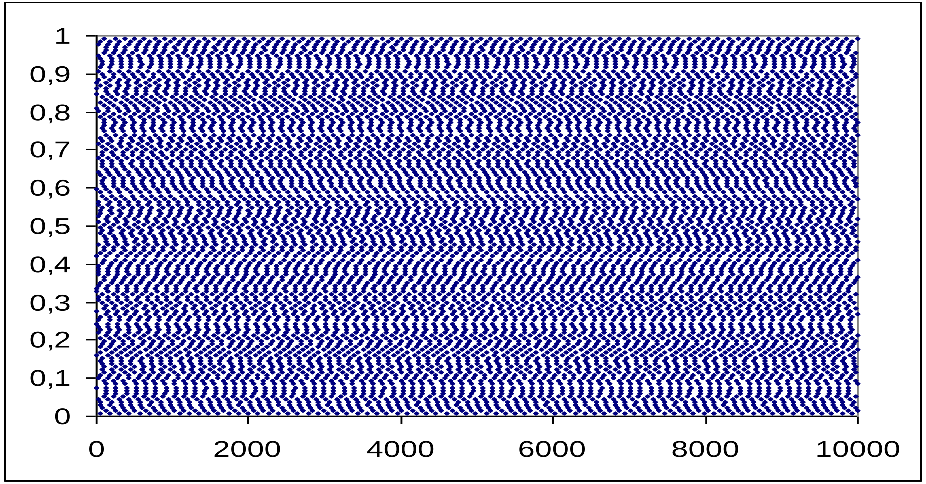
Método congruente: n(i+1)=(n(i)*b) MOD m

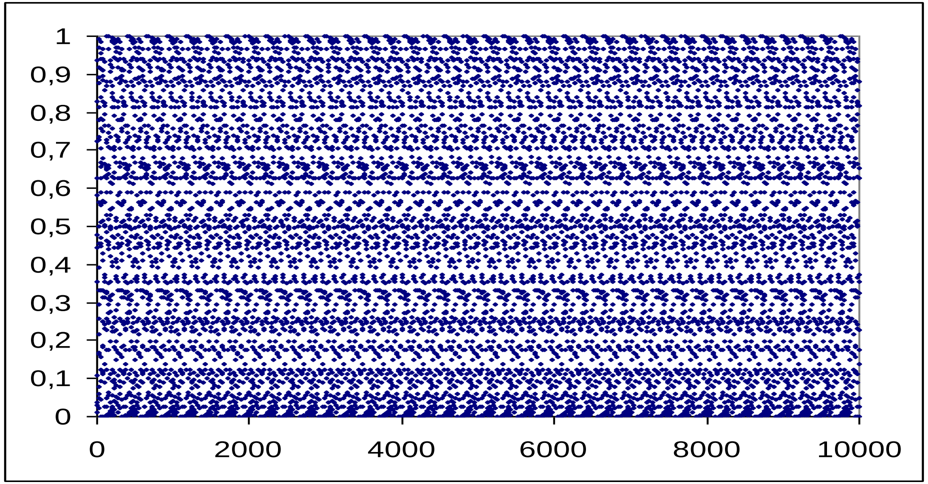
onde: n(i), b e m são constantes positivas escolhidas por conveniência, com **n(i) < m.**

n(0) é chamado de semente


1
0,8
0,6
0,4
0,2
0
0
200
400
600
800
1000


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
2000
4000
6000
8000
10000


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
2000
4000
6000
8000
10000


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
2000
4000
6000
8000
10000

Amostragem

Incremento exponencial da amostra

Propriedades para dados de alta dimensionalidade

1) O tamanho do conjunto de dados com a mesma densidade de uma amostra de tamanho "n" para um espaço d-dimensão cresce exponencialmente.

$$tamanho = n^d$$

Exemplo:

Se o tamanho da amostra do conjunto 1-dim for n=100, para 5-dim, é necessário:

n^5=10^10 amostras para manter a mesma densidade nos dados.

Amostragem

Licap

Propriedades para dados de alta dimensionalidade

2) Para capturar uma uniforme e pequena porção de dados num espaço de alta-dimensão é necessária uma grande vizinhança. O cálculo desta expansão é dado por:

$$expansão(p,d) = p^{1/d}$$

Exemplo:

Se a porção de amostras é p = 10%

para 1-dim: expansão(0,1;1)=0,1

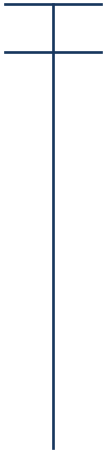
para 2-dim: expansão(0,1;2)=0,32

para 3-dim: expansão(0,1;3)=0,46

para 10-dim: expansão(0,1;3)=0,80

Amostragem

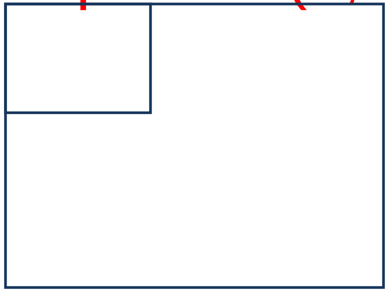
para 1-dim:
expansão(0,1;1)=0,1



0,10

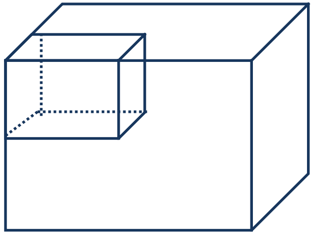
para 2-dim:
expansão(0,1;2)=0,32

0,32



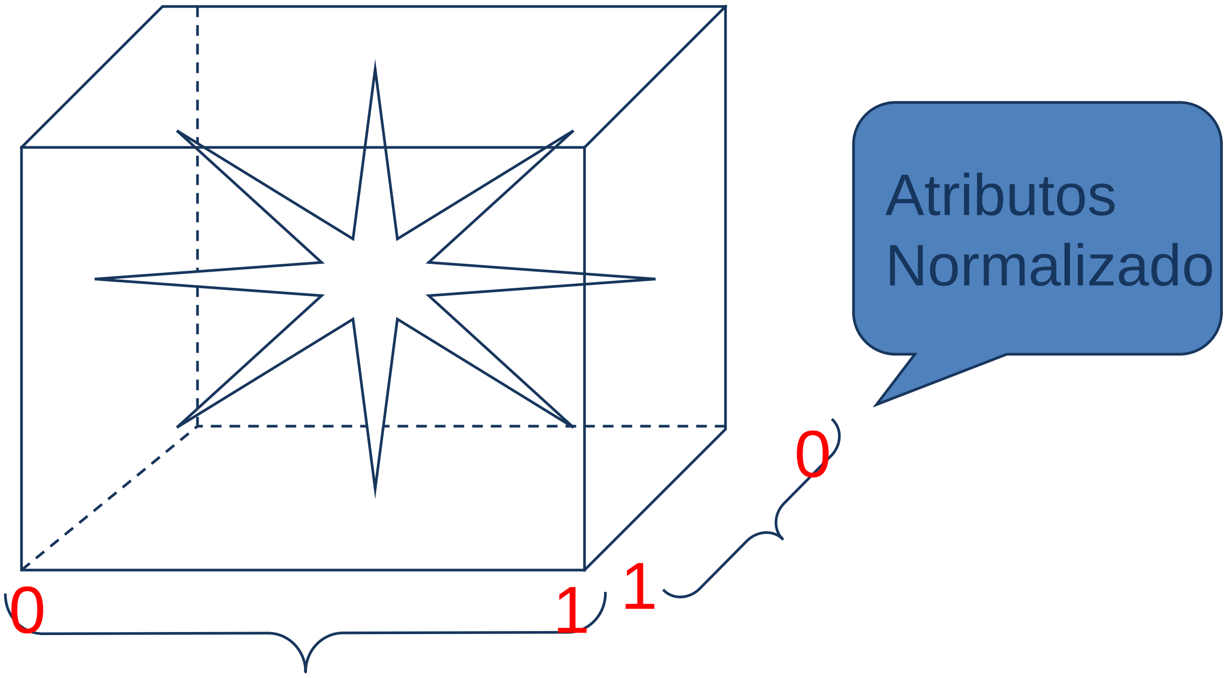
para 3-dim: expansão(0,1;3)=0,46

0,46



$$expansão(p,d) = p^{1/d}$$

Amostragem



Amostragem

Licap

Propriedades para dados de alta dimensionalidade

3) Cada objeto (exemplo) deve estar mais próximo da fronteira do sub-hipercubo, que de outro objeto da amostra. Por tanto a distância esperada "D" entre objetos num espaço d-dim é dado por:

$$D(d,n) = \frac{1}{2}\left(\frac{1}{n}\right)^{\frac{1}{d}}$$

<u>Exemplo:</u>

Se n=10000

para 1-dim: D(1,10000)=0,00005

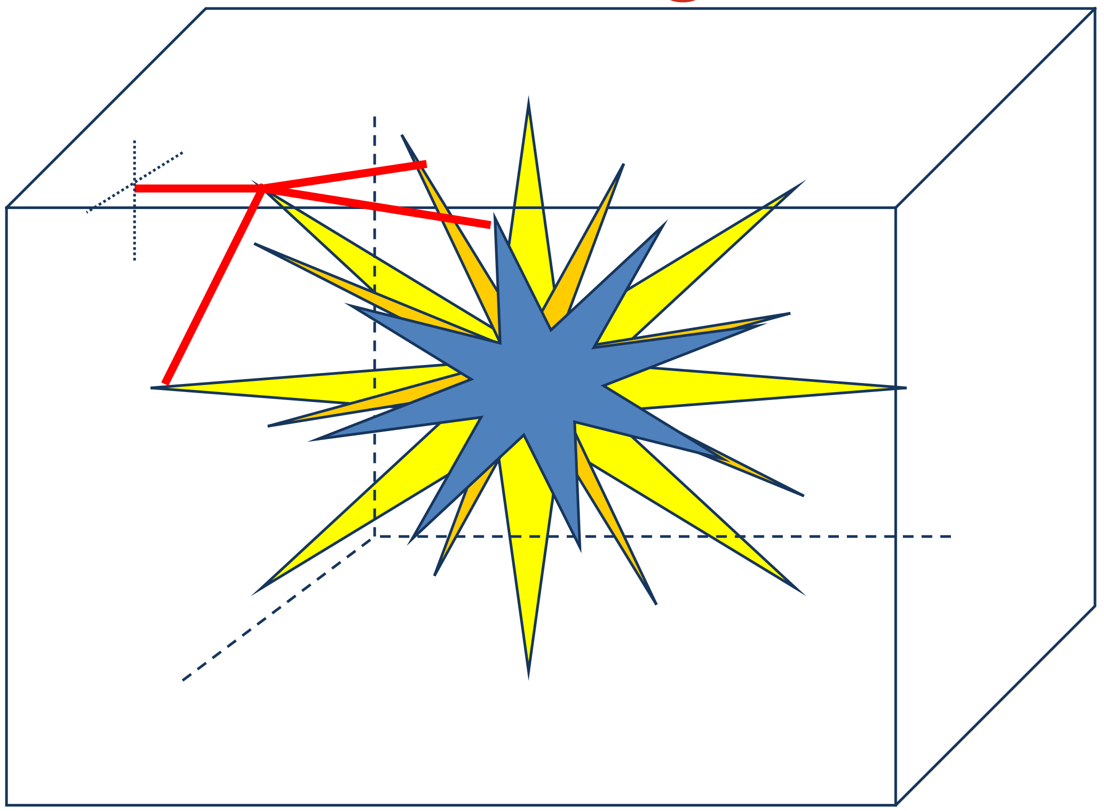
para 2-dim: D(2,10000)=0,005

para 3-dim: D(3,10000)=0,023

para 10-dim: D(10,10000)=0,20

Amostragem

Licap



Amostragem

Licap

Propriedades para dados de alta dimensionalidade

4) Cada objeto é *quase* um outlier. Quanto maior a dimensão maior será o desvio padrão entre os objetos e o centro do dados.

Lidar com dados com alta-dimensionalidade exige condições a serem satisfeitas. Um número inadequado de exemplos leva a um conhecimento falho ou a um conhecimento com severas restrições.

Aplicação em R

Licap

```
#Programa R para análise estatística e Inferêncial
dados = read.table(file="DadosExemploCurso4.txt", header=T)
View(dados)              #Mostra os dados
dim(dados)               #Mostra dimensão do objeto
head(dados)              #Mostra nome das variáveis

#Extrai variáveis do conjunto de dados
Peso<-dados$Peso
Target<-dados$Target

#Estatistica descritiva
summary(Peso)
summary(Target)
hist(Peso)
plot(function(y)dnorm(y,mean=mean(Peso),sd=sd(Peso)),65,95, ylab="f(Peso)")
```

Aplicação em R

Licap

```
#Ordena a variável Peso deacordo com o Target A ou B
dad<-dados[order(dados$Target,na.last=TRUE,decreasing=FALSE),]
View(dad)

#contabiliza número de entradas por cada classe A e B
nA<-length(Target[Target=="A"])
nB<-length(Target[Target=="B"])

#Extrai os valores pelo tipo de classe A e B
dadA<-dad[1:nA,1:2]
dadB<-dad[1:nB,1:2]

#Constroi fdp
plot(function(y)dnorm(y,mean=mean(dadA$Peso),sd=sd(dadA$Peso)),65,95, ylab="f(Peso1)")
plot(function(z)dnorm(z,mean=mean(dadB$Peso),sd=sd(dadB$Peso)),65,95, ylab="f(Peso2)")
hist(dadA$Peso,nclass=10,col="red")
hist(dadB$Peso,add=T,col=rgb(0,1,0,0.5))

#boxplot
boxplot(dadA$Peso)
boxplot(dadB$Peso)
```

Aplicação em R

```
#Dados padronizados
meanA<-mean(dadA$Peso)
varA<-var(dadA$Peso)
sdA<-sd(dadA$Peso)

ZA<-(dadA$Peso-rep(meanA,nA))/sdA
meanZ<-mean(ZA)
varZ<-var(ZA)
sdZ<-sd(ZA)

#Intervalo de confiança para a Média com Variância conhecida para classe A
conf.level<-0.94
z<-1.881
LI <- meanA - z*(sdA/sqrt(nA))
LS <- meanA + z*(sdA/sqrt(nA))
View(LI)
View(LS)
```

Licap

Formação Cientista de Dados

Obrigado!

PUC Minas