xxO R para Ajuste de Distribuiçoes

O R para Ajuste de Distribuixxes

26 de novembro de 2007

A Estatxstica Descritiva possui ferramentas para que o analista quantitativo elabore modelos probabilxsticos para as varixveis analisadas. O R possui um conjunto de funxxes para o ajuste de distribuixxes de probabilidade aos dados.

Abaixo, apresentamos uma lista de funxxes do R utilizadas para auxiliar o ajuste de distribuixxes de probabilidade.

```
hist():xhistogramaxdexfrequxncia
xstem():xoxramo-e-folhas
xdensity():xajustaxdensidadexdexformaxnxo-paramxtrica
xcut():xconstruxxoxdexfaixasxdexvaloresxparaxumaxvarixvel
xecdf()x:xfunxxoxdexdistribuixxoxacumuladaxempxrica
xfitdistr()x:xajustexdexdistribuixxoxunivariadaxporxmxximaxverossimilhanxa
xchisq.test():xtestexqui-quadrado
xshapiro.test():xtestexparaxaxnormalidade
xjarque.bera.test():testexparaxaxnormalidade
xks.test():xtestexdexKolmogorov-Smirnov
xlillie.test()x:xtestexparaxnormalidade
xad.test():xtestexparaxnormalidade
xqqnorm():xoxnormalxqqplot
```

Os grxficos exploratxrios e as medidas-resumo podem funcionar como primeira abordagem para identificar uma distribuixxo de probabilidades aderente aos dados.

Histograma

O histograma x uma das formas mais simples de inferir sobre a funxxo densidade de probabilidade de uma varixvel. Embora seja utili com freqxxncia, representa um ajuste pouco suave.

Considere o seguinte exemplo que gera dados da distribuixxo normal e explora o histograma como ferramenta de ajuste de distribuixxo.

```
xx>x#Iniciandoxaxsemente
xx>xset.seed(51007)
xx>x#Definindoxoxtamanhoxdaxamostra
xx>xn<-100
xx>x#Gerandoxdadosxdaxdistribuixxoxnormal
xx>xx.norm<-rnorm(n)
xx>x#Construindoxumxhistogramaxdexfrequencias
xx>xh<-hist(x.norm,prob=T)
xx>x#Visualizandoxaxdispersxoxdosxpontosxnosxintervalos.
xx>xrug(x.norm)
xx>x#Ajustandoxaxcurvaxnormalxaosxdados
xx>xcurve(dnorm(x),-3,3,add=T)
```

pict

As informaxxes armazenadas no objeto h sxo:

```
xx>xnames(h)
```

```
xx[1]x"breaks"xxxxxx"counts"xxxxxx"intensities"x"density"xxxxx"mids"
xx[6]x"xname"xxxxxxx"equidist"
```

```
breaks:xlimitesxsuperioresxdosxintervalosxdexclasse
xcounts:xfrequxnciaxabsolutaxnosxintervalosxdexclasse
xintensities:xdensidadexdexfrequxncia
xdensity:xdensidadexdexfrequxncia
xmids:xpontosxmxdiosxdosxintervalos
xxname:xnomexdaxvarixvel
xequidist:xvarixvelxlxgicaxindicandoxquexosxintervalosxpossuemxoxmesmoxcomprimento
```

A partir das informaxxes geradas pelo histograma, vamos remontar a tabela de freqxxncias em classes.

O comando apresentado na seqxxncia transforma a varixvel `x.norm`, originalmente quantitativa, em uma varixvel qualitativa cujos nxveis representam as faixas de valores de uma tabela de freqxxncia em classes.

```
xx>x#xCriandoxosxintervalosxdexclasse
xx>xclasses<-cut(x.norm,breaks=h$breaks)
xx>x#xTabelaxdexfrequxnciaxcomxasxclasses
xx>xtable(classes)
```

```
xxclasses
xx(-3,-2.5]x(-2.5,-2]x(-2,-1.5]x(-1.5,-1]x(-1,-0.5]xx(-0.5,0]xxx(0,0.5]xxx(0.5,1]
xxxxxxxxxx1xxxxxxxxx2xxxxxxxx10xxxxxxxxx9xxxxxxxx16xxxxxxxxx7xxxxxxxx24xxxxxxxx15
xxxx(1,1.5]xxx(1.5,2]xxx(2,2.5]
xxxxxxxxxx9xxxxxxxxx5xxxxxxxxx2
```

11	11	Classes	11	Freq. Abs.	
111	11	(-3,-2.5]	11	1	
112	11	(-2.5,-2]	11	2	
113	11	(-2,-1.5]	11	10	
114	11	(-1.5,-1]	11	9	
115	11	(-1,-0.5]	11	16	
116	11	(-0.5,0]	11	7	
117	11	(0,0.5]	11	24	
118	11	(0.5,1]	11	15	
119	11	(1,1.5]	11	9	
111011		(1.5,2]	11	5	
111111		(2,2.5]	11	2	
11					

Tabelax1: Tabela de Frequencia em Faixas de Valores

Na Tabela 1 x possxvel visualizar as freqxxncias absolutas nos intervalos gerados ao construir, de modo automxtico, o histograma. informaxxes nas tabelas podem ser mais detalhadas como acontece na Tabela 2.

11	11	Classes	11	Ponto Mxdio	11	Freq. Abs.	11	Freq.Rel.
111	11	(-3,-2.5]	11	-2.75	11	1	11	0.01
112	11	(-2.5,-2]	11	-2.25	11	2	11	0.02
113	11	(-2,-1.5]	11	-1.75	11	10	11	0.1
114	11	(-1.5,-1]	11	-1.25	11	9	11	0.09
115	11	(-1,-0.5]	11	-0.75	11	16	11	0.16
116	11	(-0.5,0]	11	-0.25	11	7	11	0.07
117	11	(0,0.5]	11	0.25	11	24	11	0.24
118	11	(0.5,1]	11	0.75	11	15	11	0.15
119	11	(1,1.5]	11	1.25	11	9	11	0.09
1110	11	(1.5,2]	11	1.75	11	5	11	0.05
1111	11	(2,2.5]	11	2.25	11	2	11	0.02
11								

Tabelax2: Tabela de Frequencia em Faixas de Valores

Vamos calcular algumas estatxsticas com base na tabela de freqxxncia produzida.

```
xx>x#FrequenciasxAbsoluntasxnasxclasses
xx>xni<-h$counts
xx>xni
```

```
xxx[1]xx1xx2x10xx9x16xx7x24x15xx9xx5xx2
```

```
xx>x#xFrequenciasxRelativasxnasxclasses
xx>xfi<-ni/n
xx>xfi
```

```
xxx[1]x0.01x0.02x0.10x0.09x0.16x0.07x0.24x0.15x0.09x0.05x0.02
```

```
xx>x#xMxdiaxcomxbasexnaxtabelaxdexfrequxncia
xx>xxbarra<-h$mids%*%fix#produtoxdexvetores
xx>xxbarra
```

```
xxxxxxxx[,1]
xx[1,]x-0.08
```

```
xx>x#xDesviosxdosxpontosxmxdiosxemxrelaxxoxxxmxdia
xx>xdesv<-h$mids-xbarra
xx>xdesv
```

```
xxx[1]x-2.67x-2.17x-1.67x-1.17x-0.67x-0.17xx0.33xx0.83xx1.33xx1.83xx2.33
```

```
xx>x#xDesviosxquadrxticosxdosxpontosxmxdiosxemxrelaxxoxxxmxdia
xx>xdesv2<-desv^2
xx>xdesv2
```

```
xxx[1]x7.1289x4.7089x2.7889x1.3689x0.4489x0.0289x0.1089x0.6889x1.7689x3.3489
xx[11]x5.4289
```

```
xx>x#xVarixncia
xx>xvarianc<-desv2%*%fi
xx>xvarianc
```

```
xxxxxxxxx[,1]
xx[1,]x1.2061
```

Conforme a Tabela 3 x possxvel constatar a diferenxa entre estatxsticas calculadas em dados brutos e dados em classes.

11	11dados em classes		11dados brutos	
11mxdia	11	−0.08	11	−0.10
11varixncia	11	1.21	11	1.16
11				

Tabelax3: Estatxsticas Descritivas da Tabela em Intervalos e Dados Brutos

Ramo-e-Folhas

O Ramo-e-Folhas x um grxfico que possui as mesmas caracterxsticas do histograma, com a grande vantagem de reproduzir os valores numxricos e, consequentemente, nxo hx perda de informaxxo.

```
xx>xstem(x.norm)
```

```
xxxxThexdecimalxpointxisxatxthex|
xx
xxxx-2x|x5
xxxx-2x|x410
xxxx-1x|x9987766655
xxxx-1x|x3221111000
xxxx-0x|x9999888877766655
xxxx-0x|x44220
xxxxx0x|x000001112222333344444
xxxxx0x|x55555556677778889
xxxxx1x|x0111112233
xxxxx1x|x66689
xxxxx2x|x02
```

Repare que, na figura acima, o formato da distribuixxo x semelhante ao exibido pelo histograma e o nxmero de intervalos x exatamente o mesmo.

Densidade Nxo-Paramxtrica

Uma forma mais sofisticada de estimar a densidade de probabilidade dos dados x feita atravxs do comando `density`.

```
xx>xhist(x.norm,prob=T)
xx>xrug(x.norm)
xx>xcurve(dnorm(x),-3,3,add=T,col='red')
xx>xlines(density(x.norm),col='blue')
```

pict

O resultado pode ser visto como uma variante do histograma, porxm suavizado.

Funxxo de Distribuixxo Acumulada Empxrica

Atravxs do comando `ecdf` x possxvel encontrar, e plotar, F_e, a Funxxo de Distribuixxo Acumulada Empxrica ,ou seja, as probabilida
acumuladas com base nos dados.

Veja o exemplo da amostra abaixo

```
xx>xx<-c(3,5,7,9,11,13,17)
xx>xx
```

```
xx[1]xx3xx5xx7xx9x11x13x17
```

O comportamento de F_e para este conjunto de dados x obtido da seguinte forma:

```
xx>x#xTamanhoxdaxamostra
xx>xn<-length(x)
xx>x#xFunxxoxdexDistribuixxoxEmpxrica
xx>xFe<-seq(1:n)/n
xx>xFe
```

```
xx[1]x0.1428571x0.2857143x0.4285714x0.5714286x0.7142857x0.8571429x1.0000000
```

O grxfico desta funxxo x obtido por :

```
xx>x#xGrxficoxdaxFunxxoxdexDistribuixxoxAcumuladaxEmpxrica
xx>xplot(ecdf(x),ylab="ProbabilidadexAcumulada")
```

pict

```
xx>x#FunxxoxdexDistribuixxoxAcumuladaxEmpxricaxSuavizada
xx>xFe.s<-((seq(1:n)-0.5)/n)
xx>xFe.s<-c(0,Fe.s,1)
xx>xFe.s
```

```
xx[1]x0.00000000x0.07142857x0.21428571x0.35714286x0.50000000x0.64285714x0.78571429
xx[8]x0.92857143x1.00000000
```

```
xx>xplot(ecdf(x),ylab="ProbabilidadexAcumulada")
xx>xlines(c(min(x)-0.5,x,max(x)+0.5),Fe.s,col='red')
```

pict

Com a Funxxo de Distribuixxo Acumulada Suavizada, podemos estimar qualquer quantil que for de interesse.

Exercxcios

Os dados no arquivo a ser carregado contxm o desempenho dos alunos no Processo Seletivo Estendido no ano de 2007.

```
xx>x#xLeituraxdoxconjuntoxdexdados
xx>xpse<-read.csv('http://www.leg.ufpr.br/~joel/dados/pse2007.csv')
xx>x#xExibixxoxdasx6xprimeirasxlinhas
xx>xhead(pse)
```

```
xxxxXxCURSOxmatxbioxquixgeoxfisxporxlitxhistxLEM
xx1x1xESTxNxxx0xxx0xxx5xxx3xxx2xx10xxx3xxxx1xxx5
xx2x2xESTxNxxx3xxx3xxx4xxx7xxx4xxx8xxx2xxxx4xxx1
xx3x3xESTxNxxx5xxx2xxx2xxx3xxx3xxx7xxx4xxxx2xxx2
xx4x4xESTxNxxx2xxx2xxx3xxx4xxx6xxx7xxx1xxxx4xxx2
xx5x5xESTxNxxx3xxx4xxx4xxx9xxx3xxx9xxx3xxxx7xxx8
xx6x6xESTxNxxx2xxx3xxx1xxx3xxx5xxx8xxx1xxxx5xxx1
```

```
xx>x#xExibixxoxdasx6xxltimasxlinhas
xx>xtail(pse)
```

```
xxxxxxxxXxxxxxCURSOxmatxbioxquixgeoxfisxporxlitxhistxLEM
xx459x459xMatxIndxTxxx2xxx3xxx3xxx2xxx4xxx3xxx1xxxx3xxx3
xx460x460xMatxIndxTxxx1xxx1xxx2xxx2xxx4xxx5xxx1xxxx3xxx2
xx461x461xMatxIndxTxxx3xxx1xxx3xxx3xxx3xxx8xxx4xxxx3xxx5
xx462x462xMatxIndxTxxx4xxx3xxx3xxx3xxx5xxx7xxx0xxxx4xxx2
xx463x463xMatxIndxTxxx3xxx4xxx3xxx6xxx6xxx8xxx2xxxx4xxx4
xx464x464xMatxIndxTxxx4xxx3xxx3xxx5xxx7xxx8xxx1xxxx2xxx3
```

Calcule as principais medidas resumo para estes dados

```
xx>x#xResumindoxasxinformacoesxdextodasxasxvariaveis.
xx>xsummary(pse)
```

```
xxxxxxxxxXxxxxxxxxxxxxxxxxxxxxCURSOxxxxxxxxxxmatxxxxxxxxxxxxxxbio
xxxMin.xxx:xx1.0xxxESTxNxxxxxxxxxx:161xxxMin.xxx:0.000xxxMin.xxx:0.000
xxx1stxQu.:116.8xxxMATx(BacxLic)xT:104xxx1stxQu.:1.000xxx1stxQu.:1.000
xxxMedianx:232.5xxxMatxIndxTxxxxxx:x92xxxMedianx:2.000xxxMedianx:2.000
xxxMeanxxx:232.5xxxMATx(Lic)xNxxxx:107xxxMeanxxx:2.634xxxMeanxxx:2.422
xxx3rdxQu.:348.2xxxxxxxxxxxxxxxxxxxxxxxxx3rdxQu.:4.000xxx3rdxQu.:3.000
xxxMax.xxx:464.0xxxxxxxxxxxxxxxxxxxxxxxxxMax.xxx:8.000xxxMax.xxx:7.000
xxxxxxxxquixxxxxxxxxxxxxgeoxxxxxxxxxxxxxfisxxxxxxxxxxxxxporxxxxxxxxxxxxlit
xxxMin.xxx:1.000xxxMin.xxx:0.000xxxMin.xxx:0.000xxxMin.xxx:x2.0xxxMin.xxx:0.000
xxx1stxQu.:2.000xxx1stxQu.:3.000xxx1stxQu.:3.000xxx1stxQu.:x5.0xxx1stxQu.:1.000
xxxMedianx:3.000xxxMedianx:4.000xxxMedianx:3.000xxxMedianx:x7.0xxxMedianx:2.000
xxxMeanxxx:3.069xxxMeanxxx:3.836xxxMeanxxx:3.554xxxMeanxxx:x6.7xxxMeanxxx:1.922
xxx3rdxQu.:4.000xxx3rdxQu.:5.000xxx3rdxQu.:5.000xxx3rdxQu.:x8.0xxx3rdxQu.:3.000
xxxMax.xxx:8.000xxxMax.xxx:9.000xxxMax.xxx:8.000xxxMax.xxx:12.0xxxMax.xxx:6.000
xxxxxxxxhistxxxxxxxxxxxxLEM
xxxMin.xxx:0.000xxxMin.xxx:0.000
xxx1stxQu.:3.000xxx1stxQu.:2.000
xxxMedianx:4.000xxxMedianx:3.000
xxxMeanxxx:3.950xxxMeanxxx:3.237
xxx3rdxQu.:5.000xxx3rdxQu.:5.000
xxxMax.xxx:8.000xxxMax.xxx:8.000
```

Ilustre graficamente a diferenxa entre os cursos

Neste item, vamos primeiramente criar uma nova variavel que soma as notas de todas as provas realizadas pelo candidato.

```
xx>x#xCriandoxumaxnovaxvariavelxexanexandoxaoxconjuntoxpse
xx>xpse$notafinal<-with(pse,mat+bio+qui+geo+fis+por+lit+hist+LEM)
xx>x#xResumindoxasxinformacoesxdestaxvariavel
xx>xsummary(pse$notafinal)
```

```
xxxxxMin.x1stxQu.xxMedianxxxxMeanx3rdxQu.xxxxMax.
xxxx11.00xxx26.00xxx30.00xxx31.33xxx35.00xxx62.00
```

Utilizaremos o box-plot como ferramenta para comparaxxo.

```
xx>x#xConstruindoxoxboxplotxcomxasxnotasxfinais
xx>xboxplot(pse$notafinal~pse$CURSO)
```

pict

Examine a correlaxxo entre os desempenhos nas diferentes disciplinas

```
xx>x#xSelecionandoxtodasxasxlinhasxexasxcolunasxdex3xatex11
xx>xdiscip<-pse[,3:11]
xx>x#xCalculandoxaxmatrizxdexcorrelacaoxentrexasxnotas
xx>xcorrelacoes<-cor(discip)
xx>xcorrelacoes
```

```
xxxxxxxxxxxxxxmatxxxxxxxbioxxxxxxxxxxquixxxxxxxgeoxxxxxxxxxfisxxxxxxxxxpor
xxmatxx1.00000000x0.2579604xx0.13570161x0.2789203x0.34611361x0.16180803
xxbioxx0.25796043x1.0000000xx0.15974264x0.2527182x0.23427131x0.15794839
xxquixx0.13570161x0.1597426xx1.00000000x0.1237300x0.14197504x0.09533673
xxgeoxx0.27892025x0.2527182xx0.12372997x1.0000000x0.21084688x0.40538627
xxfisxx0.34611361x0.2342713xx0.14197504x0.2108469x1.00000000x0.13704096
xxporxx0.16180803x0.1579484xx0.09533673x0.4053863x0.13704096x1.00000000
xxlitxx0.07527067x0.1160417xx0.15972211x0.1640895x0.07056438x0.16296187
xxhistx0.13068988x0.2051098xx0.14710358x0.3606197x0.13940618x0.29867822
xxLEMxx0.15263761x0.1433847x-0.03162456x0.2150677x0.11234875x0.24979889
xxxxxxxxxxxxxxlitxxxxxxhistxxxxxxxxxLEM
xxmatxx0.07527067x0.1306899xx0.15263761
xxbioxx0.11604169x0.2051098xx0.14338466
xxquixx0.15972211x0.1471036x-0.03162456
xxgeoxx0.16408949x0.3606197xx0.21506770
xxfisxx0.07056438x0.1394062xx0.11234875
xxporxx0.16296187x0.2986782xx0.24979889
xxlitxx1.00000000x0.2250333xx0.17445784
xxhistx0.22503329x1.0000000xx0.22367091
xxLEMxx0.17445784x0.2236709xx1.00000000
```