Universidade de São Paulo

Escola Superior de Agricultura "Luiz de Queiroz"

Seção Técnica de Informática

Proc Univariate:Testando a normalidade

Marcelo Corrêa Alves

Proc Step

– Piracicaba / 2016 –

_____Proc Univariate: Testando a normalidade

Sumário

1 Introdução

Uma preocupação bastante recorrente na aplicação de técnicas de análise de dados se refere à avaliação da aderência à distribuição normal, ou Gaussiana[1]. Independentemente do objetivo da avaliação da aderência a esta distribuição específica, este capítulo se dedica ao uso do SAS, e mais especificamente do procedimento **Univariate**, na avaliação da aderência de dados à essa distribuição.

A distribuição gaussiana apresenta diversas características que a fazem interessante como embasamento para o desenvolvimento de testes estatísticos e, em certos casos, o rigor exigido requer que se avalie a aderência de um conjunto de dados à essa distribuição.

Conhecendo-se as características da aderência e a eventual existência de desvios pode-se ter mais certeza da validade da aplicação de uma técnica ou das limitações que essa técnica terá em função dos desvios encontrados e, com isso, pode-se justificar ou refutar a sua aplicação.

2 Objetivos

Nesse capítulo objetiva-se que você:

- Conheça a opção normal do procedimento **univariate**
- Conheça a opção **plot** do procedimento **univariate**
- Interprete os resultados dos testes formais para normalidade
- Interprete gráficos de avaliação da aderência à distribuição normal
- Interprete coeficientes que mensuram características da distribuição

3 A distribuição normal

A distribuição gaussiana é muito frequentemente associada à figura de um sino e os dados que se distribuem de acordo com a distribuição normal, se plotados em um histograma apresentam a aparência dessa curva (figura 1).

[1] Homenagem a Carl Friedrich Gauss, matemático, astrônomo e físico alemão.

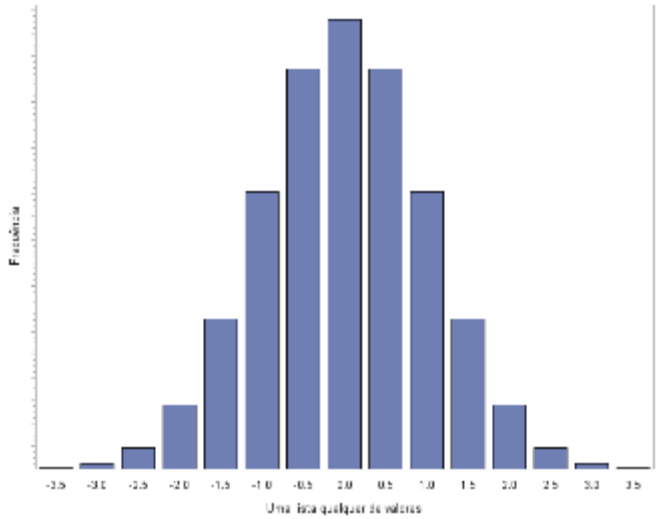


Figura 1. Histograma com dados distribuídos de maneira aderentes a uma distribuição normal.

Ao observarmos o histograma apresentado na figura 1, algumas características devem ser especialmente percebidas para que possamos avaliar a questão da normalidade. **Os dados usados na construção do gráfico são contínuos**, ou seja, temos valores que se distribuem deste -3,5 até mais +3,5 e quaisquer valores dentro desse intervalo são válidos. Na figura 1 esses valores contínuos são agrupados em classes.

A ideia de continuidade é importante já que a distribuição normal é tida como contínua, sendo assim, quando analisamos a barra que representa a frequência de dados na classe 3,0; devemos imaginar que há diversos números diferentes representados nessa barra que deve incluir valores tais como 2,8; 2,9; 3,0; 3,01; 3,13 e assim por diante de forma que ao imaginarmos dados com distribuição aderente à normal, imaginaríamos um gráfico de linha, conforme apresentado na figura 2.

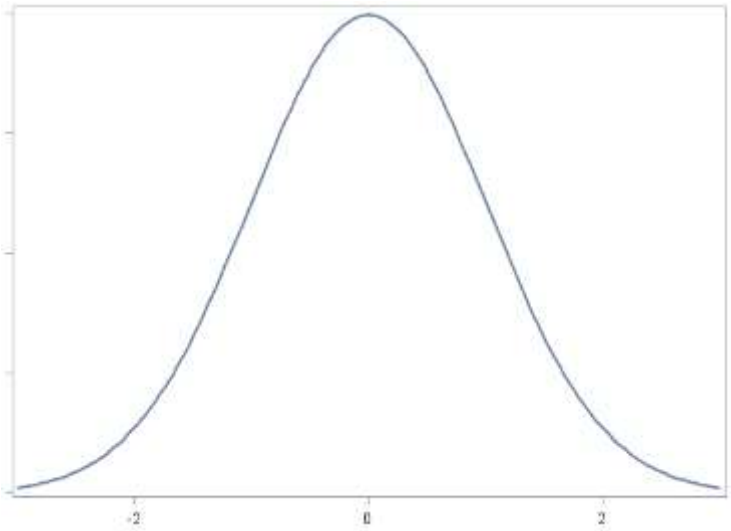


Figura 2. Representação da continuidade da distribuição normal.

Esse é o formato de sino, referido no primeiro parágrafo do tópico e se refere a uma possível distribuição normal dentro de uma **família de distribuições** que seguem esse comportamento.

Há dois parâmetros que caracterizam distribuições normais diferentes, dessa forma, quando falamos de distribuição normal nos referimos a uma família de distribuições com parâmetros diferentes, mas cada uma delas, ainda chamada de distribuição normal. **Os dois parâmetros que permitem identificar uma única distribuição normal são: a média e o desvio padrão**.

3.1 Distribuições normais com médias diferentes

Podemos, por exemplo, variar a distribuição em relação à média e teríamos duas distribuições normais diferentes, conforme ilustra a figura 3.

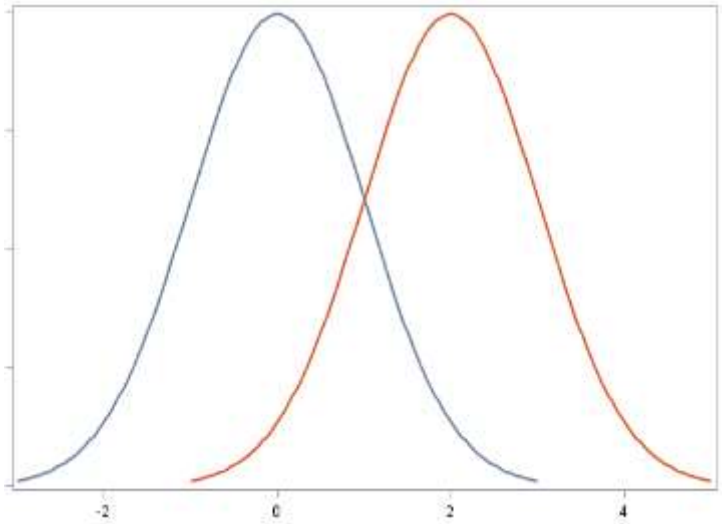


Figura 3. Representação de duas distribuições normais com médias diferentes.

Note que as duas distribuições representadas acima podem ser consideradas aderentes à distribuição normal, entretanto, há uma diferença entre as duas, uma delas tem o ponto mais alto no valor 0 da abcissa enquanto que na outra, o pico fica no valor 2 da abcissa.

Na distribuição "normal", esse ponto mais alto **representa a média dos valores, também representa a mediana (já que metade dos dados são menores que ele e metade dos dados são maiores) e ainda a moda** já que este é o dado que mais se repete e por isso é o pico da distribuição de frequências representada no gráfico.

Uma característica que faz com que exista uma família de distribuições normais é a variação em relação à média, na figura 3, temos a representação de uma distribuição normal com média 0 e uma segunda distribuição normal com média 2.

3.2 Desvios padrão diferentes

Há um segundo parâmetro que modifica a aparência da distribuição normal: o desvio padrão.

Na figura 4, observamos duas representações de dados normalmente distribuídos, ambas com média 0, medida tomada apenas para se observar as diferenças causadas pela variação de um parâmetro, no caso o desvio padrão que, para uma distribuição foi fixado em 1,0 e em outra curva foi fixado em 1,5.

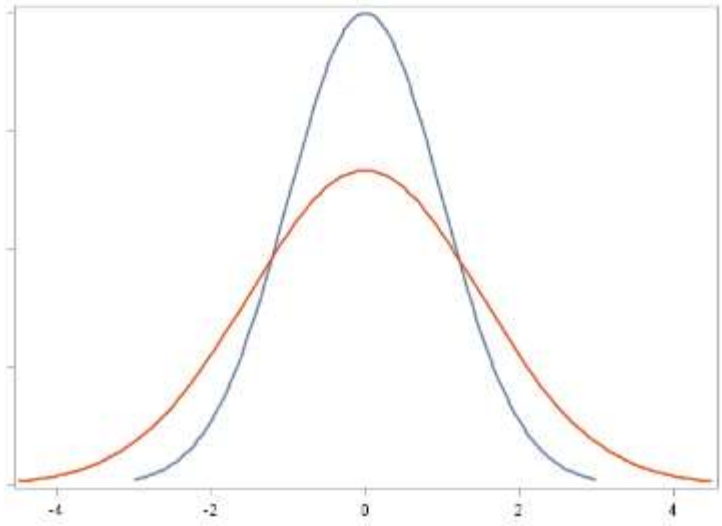


Figura 4. Sobreposição de distribuições normais diferentes em relação ao desvio padrão.

Mas note que há uma diferença em relação ao desvio padrão o que implica na diferença do achatamento. A mais alta foi construída com base em dados com desvio padrão igual a 1 enquanto que a mais baixa representa uma distribuição normal com desvio padrão igual a 1,5. Quanto maior o desvio padrão, maior a amplitude, note que que com desvio padrão 1, os valores da abcissa se situam entre os valores -3 e +3, aproximadamente, enquanto que quando o desvio padrão é 1,5 os valores representados na abcissa variam desde -4,5 até +4,5, também aproximadamente.

3.3 Regra empírica da distribuição normal

No final das contas, o que têm em comum todas as distribuições normais como as representadas na figura 5.

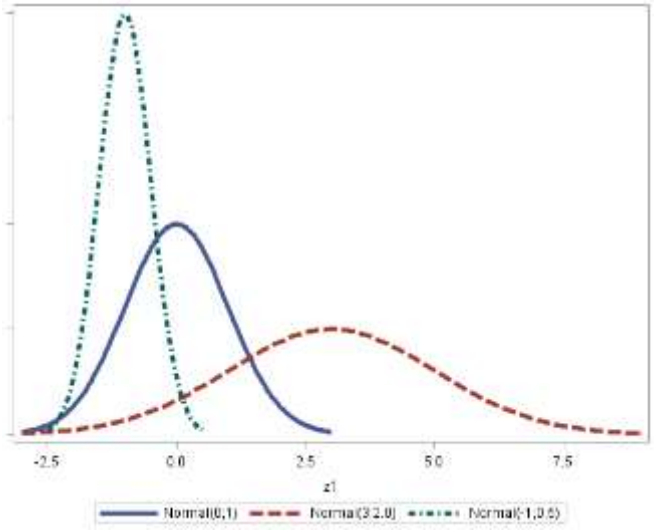


Figura 5. Diferentes distribuições normais, construídas de acordo variação da média e do desvio padrão.

As três curvas são bastante diferentes entre si, visualmente, apenas se assemelhando pelo formato de sino, mas todas elas foram construídas com base na distribuição normal e, apesar da aparente diferença, todas elas seguem uma regra empírica.

A regra empírica da distribuição normal nos permite distribuir a frequência de valores em termos da média e do desvio padrão de acordo com as porcentagens ilustradas na figura 6.

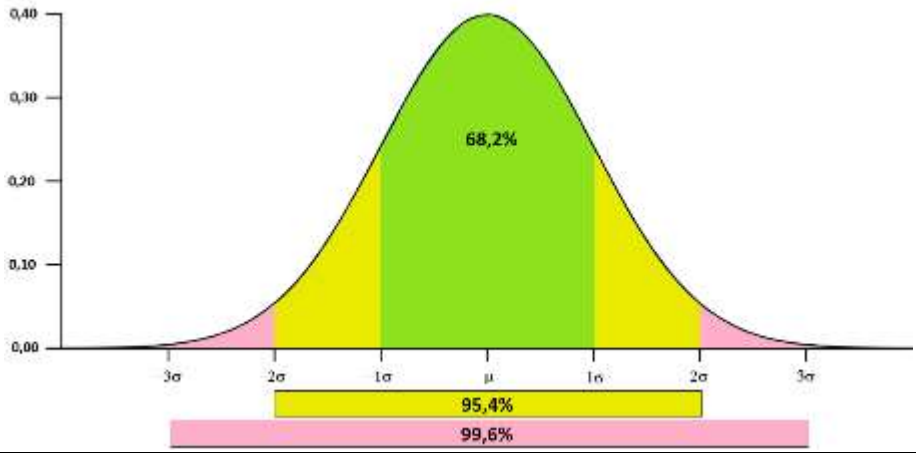


Figura 6. Representação da regra empírica da distribuição normal.

Pela regra empírica, 68,2% dos dados se situam na faixa entre mais ou menos um desvio padrão em torno da média; 95,4% dos dados se situam entre mais ou menos dois desvios padrão em torno da média e mais de 99,6% dos dados se situam entre mais ou menos três desvios padrão.

Essas proporções são importantes na definição da avaliação da aderência de dados à normalidade e também são importantes ao determinar propriedades da distribuição normal.

3.4 Síntese das características da distribuição normal

Convém sintetizar as principais características da distribuição normal, algumas previamente citadas nos parágrafos anteriores. Essas características são importantes para identificarmos a aderência de uma variável aleatória contínua a distribuição normal e, com isso, podermos nos valer de suas propriedades:

- A distribuição normal é contínua e desde o mínimo até o máximo de desenvolve de maneira suave;
- É definida por meio da média e do desvio padrão;
- A distribuição normal é simétrica;
- A distribuição normal tem um grau de achatamento pertinente à regra empírica;
- A média a mediana e a moda são iguais;

4 O procedimento Univariate

O procedimento **univariate** do SAS é usado quando se deseja efetuar análises descritivas de variáveis numéricas e é especialmente útil quando se necessita estudar a distribuição dos dados.

A quantidade de estatísticas por ele geradas é bastante grande e uma fração delas é interessante em relação ao estudo da aderência dos dados à distribuição normal e esses recursos serão tratados de forma detalhada no presente material.

A sintaxe completa do procedimento **univariate** é ilustrada no quadro 1.

```
proc univariate <opções> ;
        by lista de variáveis ;
        cdfplot < lista de variáveis > < / opções > ;
        class variável_1 <(opções)> < variável_2 <( opções)>> </ keylevel= valor1 | ( valor1 valor2 )> ;
        freq variável;
        histogram < lista de variáveis > < / opções> ;
        id lista de variáveis; inset lista de palavras chave </ opções> ;
        output <out=SAS-data-set> <estatística1=nomes ...estatístican=nomes> <opções de percentis> ;
        ppplot < lista de variáveis > < / opções> ;
        probplot < lista de variáveis > < / opções> ;
        qqplot < lista de variáveis > < / opções> ;
        var lista de variáveis;
        weight variável ;
```

Quadro 1. Sintaxe do procedimento **univariate**.

Nesse capítulo serão estudadas apenas as opções que se relacionem ao estudo da aderência dos dados à distribuição normal, o que envolve um subconjunto relativamente grande das opções elencada no quadro 1.

4.1 Comando Proc Univariate

Para ativar o procedimento, usamos o comando **proc univariate** o qual permite a especificação de duas opções interessantes para quem deseja avaliar a normalidade de uma variável: as opções **normal** e **plot**, conforme ilustrado no programa 1.

Programa 1. Geração de dados aleatórios e avaliação da normalidade por opções do procedimento univariate.

```
ods html close;
ods html;
data amostra;
   do x = 1 to 300;
      y1 = rand('Normal', -2,1.5);
      output;
   end;
run;
proc univariate data=amostra normal plot;
   var y1;
run;
```

O programa 1 conta com um *data step* que gera um SAS data set denominado amostra e que contém 300 observações (valores da variável x se iniciando em 1 e terminando em 300) determinando a execução de 300 iterações e a cada uma delas um valor aleatório é gerado e armazenado na variável y1. O resultado do processamento desse *data step*, quando analisado pelo conteúdo da janela **log** revela:

```
NOTE: The data set WORK.AMOSTRA has 300 observations and 2 variables.
NOTE: DATA statement used (Total process time):
real time           0.11 seconds
cpu time            0.00 seconds
```

A segunda parte do programa é um proc step que ativa o procedimento univariate e três opções foram especificadas: a opção **data=**, a opção **normal** e a opção **plot**.

A opção **data=** é desnecessária nesse programa pois ela somente especifica que o *SAS data set* a ser analisado é o denominado "amostra". Na ausência da opção **data=**, o SAS utiliza em qualquer *proc step*, o SAS data set mais recentemente criado que nesse caso seria o "amostra".

Por meio do comando **var** foi especificado que somente deveria ser analisada a variável y1. Se esse comando não fosse especificado, também seria avaliada a variável x.

As opções **normal** e **plot** já são interessantes para quem deseja avaliar a aderência dos dados à distribuição normal, vamos avaliar os resultados emitidos pelo procedimento univariate. Os resultados são fragmentados em tabelas e algumas contém informações interessantes, iniciando pela tabela de momentos (*Moments*) representada na figura 7.

Moments			
N	300	Sum Weights	300
Mean	-2.0338549	Sum Observations	-610.15646
Std Deviation	1.58364134	Variance	2.50791989
Skewness	0.0686856	Kurtosis	0.81537246
Uncorrected SS	1990.83772	Corrected SS	749.868048
Coeff Variation	-77.864029	Std Error Mean	0.09143158

Figura 7. Tabela de momentos gerada pelo procedimento **univariate**.

Para avaliação da normalidade, a tabela de momentos traz duas estatísticas interessantes: o coeficiente de assimetria (**Skewness**) e o coeficiente de curtose (**Kurtosis**), cuja interpretação permite identificar problemas em relação à aderência à distribuição gaussiana. Começando pelo coeficiente de assimetria que considera a similaridade das caudas, conforme ilustrado na figura 8.

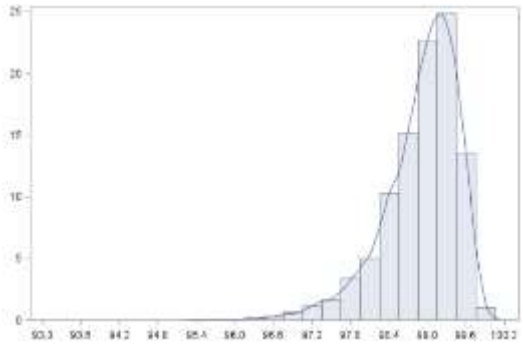
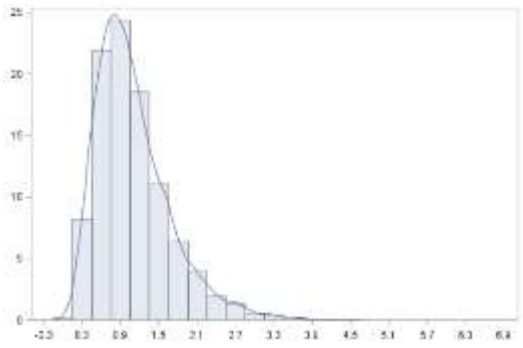
Coeficiente de assimetria	
< -2	> 2
	

Figura 8. Aparência da distribuição dos dados de acordo com a avaliação dos coeficientes de assimetria.

Quanto maior, em termos absolutos, o coeficiente de assimetria, mais desiguais são as caudas da distribuição e o sinal nos dá indicação de qual é a cauda mais longa, se negativo: aquela que indica a variabilidade dos dados pequenos e se positivo, a maior variabilidade ocorre nos valores grandes.

Em geral, assimetrias entre -2 e +2 não são indicadoras de severas discrepâncias em relação a uma distribuição simétrica e, entre as distribuições simétricas encontramos a distribuição normal.

Também o coeficiente de curtose pode ser avaliado, mas ele se refere ao achatamento da curva, conforme ilustra a figura 9.

Coeficiente de curtose	
< -2	> 2

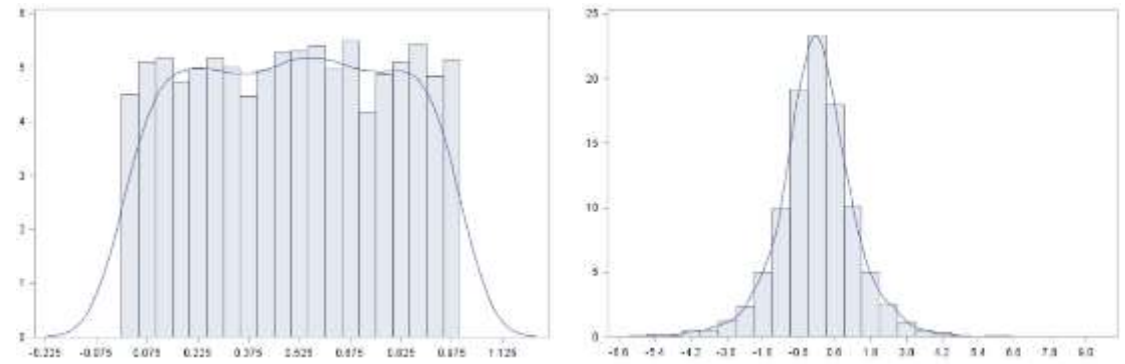


Figura 9. Aparência da distribuição dos dados de acordo com a avaliação dos coeficientes de curtose.

Em relação ao que se espera de uma distribuição aderente à normalidade, nenhuma das condições acima são razoáveis, daí a primeira avaliação da aderência dos dados à distribuição normal sendo feita por meio dos dois coeficientes.

Além dos coeficientes de assimetria e curtose, o procedimento **univariate** efetua testes estatísticos para a hipótese de que os dados provém de população normalmente distribuída e esses testes são apresentados no quadro "*Tests for Normality*", representado na figura 10.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.994183	Pr < W	0.3073
Kolmogorov-Smirnov	D	0.041272	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.084098	Pr > W-Sq	0.1898
Anderson-Darling	A-Sq	0.461804	Pr > A-Sq	>0.2500

Figura 10. Testes de normalidade efetuados pelo procedimento **univariate**.

Observe que o procedimento *univariate* calcula valores de 4 testes de normalidade, o de *Shapiro-Wilk*, o de *Kolmogorov-Smirnov*, *Cramer-von Mises* e o de *Anderson-Darling*.

A interpretação dos testes estatísticos parte de uma hipótese de nulidade (H_0) e, no caso dos quatro testes apresentados, parte-se da seguinte:

H_0: Amostra provém de população normalmente distribuída

Partimos então da premissa de que estamos trabalhando com uma amostra e que esta amostra permite que se tome uma decisão válida para a população. Este processo denominado inferência é válido, entretanto, sempre que o fazemos, podemos cometer um erro.

Isso quer dizer que a essa amostra pode nos conduzir a uma conclusão diferente da verdade, sendo assim, o teste estatístico nos auxilia ao permitir que estimemos qual é a probabilidade de errarmos ao tomar, sempre a mesma decisão: a de rejeitar H_0.

Os valores-p (*p Value*) representam a probabilidade de erro envolvido na rejeição da hipótese de nulidade. Note que os 4 testes apresentam valores-p diferentes. Se adotado o teste de Shapiro-Wilk, se rejeitarmos H_0 temos 30,73% de probabilidade de estarmos tomando a decisão errada. Se observamos o teste de Kolmogorov-Smirnov, por sua vez, temos 15,00% de probabilidade de estarmos tomando a decisão errada de rejeitar a hipótese de nulidade.

A interpretação dos testes estatísticos requer que seja arbitrado uma probabilidade de erro tolerável que conduza à rejeição da hipótese de nulidade. Esse valor arbitrário recebe o nome de Nível de Significância e é representado pela letra grega alfa (α).

Se estabelecermos um nível de significância de 10% (α=0,10) estamos assumindo que essa é uma probabilidade de erro razoavelmente pequena e que, caso tenhamos até essa probabilidade de erro, ainda assim devemos rejeitar a hipótese de nulidade.

Rejeitar a hipótese de nulidade implica em assumir como verdadeira uma segunda hipótese, a chamada de hipótese alternativa, da seguinte forma enunciada para estes testes:

H_a: Amostra provém de população não distribuída de forma aderente à distribuição normal

4.2 Atividade prática

Analise a aderência dos dados da tabela 1 à distribuição gaussiana, observe que são 30 observações, ou seja, que as variáveis se repetem para aproveitar a largura da página.

Tabela 1. Dados de DAP, Altura e Produção de observados em 30 árvores.

Árvore	DAP	Altura	Produção	Árvore	DAP	Altura	Produção
1	0,11	6,91	13,37	16	0,30	6,65	13,29
2	0,32	5,10	12,52	17	0,26	6,49	13,21
3	0,35	8,18	13,47	18	0,12	5,30	12,51
4	0,43	8,59	13,43	19	0,25	6,20	13,33
5	0,37	6,62	13,18	20	0,26	10,73	12,79
6	0,29	6,47	13,46	21	0,10	5,85	13,21
7	0,32	6,13	12,74	22	0,32	7,09	12,72
8	0,11	8,49	12,72	23	0,25	5,30	13,22
9	0,35	7,00	13,43	24	0,35	5,62	12,73
10	0,10	5,19	13,11	25	0,07	7,66	12,75
11	0,25	6,33	13,30	26	0,11	5,29	13,30
12	0,16	5,36	12,69	27	0,23	5,90	12,50
13	0,23	6,52	12,90	28	0,34	5,95	12,93
14	0,21	5,69	12,81	29	0,22	6,18	13,37
15	0,27	6,74	13,01	30	0,06	6,92	13,40

Fonte: Dados fictícios gerados por meio do SAS.