UM CASO DE TRANSFORMAÇÃO ATRAVÉS DA RAIZ QUADRADA NAS ANÁLISES ESTATÍSTICAS

IARACILDA DE ANDRADE COCENZA

Orientador: Dr. Clovis P. de Abreu

Dissertação apresentada à Escola Superior de Agricultura "Luiz de Queiroz", da Universidade de São Paulo, para obtenção do título de Mestre em Estatística e Experimentação Agronômica.

P I R A C I C A B A
Estado de São Paulo - Brasil
Novembro, 1979

ÍNDICE

Pág.

1. RESUMO

A justificativa do uso das transformações, em suas formas atualmente mais usadas, é feita considerando-se relações entre médias e variâncias e a estabilidade da variância é a condição imposta para se chegar a forma final:

$$y = \sqrt{x}$$

$$y = \log x$$

$$y = \text{arc sen } \sqrt{x}$$

Assim, entre as três razoes para se aplicar transformações, tendo em vista a validade da análise de variância:

- não-normalidade na distribuição dos erros;
- não-aditividade dos diversos efeitos previstos no modelo matemático;

- a terceira e considerada a mais importante pela literatura e a mais frequentemente usada como motivo do uso da transformaçao de dados. Porém, quando se deseja corrigir problemas causados pelas duas outras razões, as mesmas transformações são usadas.

Este trabalho tem como objetivo a verificação da capacidade da transformação da raiz quadrada em corrigir não-normalidade.

A transformação é considerada em três formas:

$$y = \sqrt{x} \qquad y = \sqrt{x + 1/2} \qquad e \qquad y = \sqrt{x + 3/8}$$

através de suas inversas:

$$x = y^2 \qquad x = y^2 - 1/2 \qquad e \qquad x = y^2 - 3/8 \,.$$

O material é simulado. Dados normais foram gerados, transformados, e a anormalidade foi medida pelos coeficientes dos momentos de assimetria e curtose, testados pelo teste t.

Pelos resultados pode-se inferir que, de modo geral, a transformação da raiz quadrada é útil para normalizar distribuiçoes desde que as médias sejam pequenas.

O resultado das transformações sobre a homogeneidade das variâncias e das médias foi mais efetivo e se tornou evidente, mesmo nao constando dos objetivos iniciais.

A comparação das três formas da mesma transformação, nao permitiu que se considerasse a superioridade de uma delas sobre as demais.

2. INTRODUÇÃO

A transformação de dados é um processo auxiliar, utilizado junto a análise de variância, com o objetivo de torná-la mais eficiente.

A análise de variância é uma técnica estatística desenvolvida há mais de meio século por R.A. Fisher, para sistematizar a análise e a interpretação de dados obtidos através de ensaios de campo e experimentos laboratoriais.

Para ser utilizada corretamente, é necessário que as hipóteses básicas em que se fundamenta (EISENHART, 1947), sejam satisfeitas, sob pena de serem invalidadas as conclusões que, o método e os testes de significância a ele associados, permitem tirar.

A transformação de dados tem sido usada como um dos métodos para corrigir as falhas nas hipóteses básicas, garantindo a validade das conclusões.

As falhas corrigíveis por transformação de dados sao:

- não-normalidade na distribuição dos erros;
- não-aditividade dos diversos efeitos previstos no modelo matemático adotado;
- heterogeneidade da variância do erro.

Neste trabalho consideraremos a não-normalidade como o motivo da transformação.

A não-normalidade ocorre quando os erros nao se adaptam à distribuição normal de probabilidades. Sendo os experimentos realizados em tamanhos limitados, os dados resultantes sao uns poucos valores isolados que dificilmente se adaptariam perfeitamente a uma distribuição contínua. No entanto, pode-se supor que uma função destes dados leve o intervalo entre eles a ser tão pequeno que torne possível uma suposição de continuidade.

As transformações são, pois, funções que aproximam os elementos de um conjunto diminuindo a amplitude do intervalo que os contém, como se pode inferir pelo exame das formas mais usadas:

$$y = \sqrt{x}$$

$$y = \log x$$

$$y = \text{arc sen } \sqrt{x}$$

Quando ocorre, a não-normalidade provoca perda de eficiência na estimação dos efeitos dos tratamentos, levando a média aritmética dos dados, a perder sua característica de medida de tendência central. Esta situação é medida pela *assimetria*, caracterizada pelo desvio da curva da distribuição de frequência para a direita ou para a esquerda, perdendo sua forma tradicional, que ocorre sempre que os dados têm distribuição normal. O excesso de dados concentrados próximo a média, ou o excesso de dados moderadamente afastados da média, é indicado pela medida de *curtose*.

A curtose exerce influência sobre a variância e consequentemente, sobre os testes de homogeneidade de variâncias.

O teste básico para comparar estimativas de variâncias, é o teste "F", de Snedecor. Este parece ser o mais robusto a não-normalidade, mas os testes utilizados para comparar estimativas das médias dos efeitos são afetados, com perda de potência, em vista da perda de eficiência na estimação destes parâmetros.

A estimação da variância, numa distribuição com razoável afastamento da normalidade, pode conter sério viés. Se a não-normalidade estiver acompanhada de heterogeneidade dos erros, o quadrado médio do resíduo não terá a representatividade requerida para ser utilizado na aplicação dos testes.

Embora as demonstrações que justificam as formas de transformações mais usadas se baseiem em homogeneização de variâncias, elas têm sido usadas na correção da não-normalidade e não-aditividade, com bons resultados. Tal fato pode ser decorrente do relacionamento entre os três problemas, apontados por vários autores en tre eles BARTLETT (1947), COCHRAN (1947) e DOLBY (1963).

Para estudar a não-normalidade usaremos a transformaçao

$$y = \sqrt{x}$$

através de sua inversa

$$x = y^2$$

aplicada a dados normais.

O enfoque do assunto nos foi sugerido pelo Dr. F. Pimentel Gomes.

Ao utilizar a transformação x^2, em dados normais, o objetivo pode ser assim descrito: verificar se a transformação é capaz de efetuar aproximação à normalidade, através de sua inversa, que quando aplicada a dados reconhecidamente normais, deverá efetuar afastamentos da normalidade, tanto mais efetivos quanto for a capacidade da transformação direta de normalizá-los.

As transformações

$$y = \sqrt{x + 1/2} \quad \text{(BARTLETT, 1936)},$$

e

$$y = \sqrt{x + 3/8} \quad \text{(ANSCOMBE, 1948)}$$

através de suas inversas:

$$x = y^2 - 1/2$$

e

$$x = y^2 - 3/8$$

também serao testadas com o objetivo de verificar se, neste trabalho, sua eficiência em relação a $\sqrt{x}$ pode ser comprovada.

As transformações serão testadas em dados simulados, obtidos por computação, devido à dificuldade de se conseguir dados experimentais suficientemente próximos da distribuição normal, como requerido aqui.

3. REVISÃO DE LITERATURA

A literatura a respeito de transformações de dados é vasta, porem neste trabalho procurou-se relacionar apenas as que diziam respeito ao aspecto do assunto aqui tratado, ou seja,a transformação da raiz quadrada, a aproximação à normalidade e os testes de assimetria e curtose.

TIPPETT (1934), em trabalho referente à pesquisa na indústria têxtil, utilizou as transformações da raiz quadrada e angular, relacionando o tipo de dados obtidos em pesquisa com algodão, com os dados das distribuições de Poisson e binomial. Seu objetivo era tornar as variâncias independentes da média.

Para chegar às formas utilizadas nas transformações, o autor baseou-se em método utilizado por Fisher para deduzir a transformação para coeficientes de correlação:

$$Z = tg\ h^{-1}\ r$$

TIPPETT partiu do princípio que r fosse uma variável aleatória com distribuição binomial, cuja média fosse p e variância σ_r^2. Supondo uma função

$$u = f(r)\ ,$$

com média u e variância σ_u^2, expandiu o 2º membro da função:

$$u = f(p + dr)$$

obtendo:

$$\sigma_u^2 = \left[f'(p)\right]^2\ \sigma_r^2$$

Considerando que a variância de u fosse constante, e fazendo

$$\sigma_r^2 = \frac{p(1 - p)}{}$$

a resolução da equaçao diferencial forneceu

$$u = f(r) = \text{arc sen}\ \sqrt{r}$$

dispensadas as constantes.

Neste artigo nao consta a obtenção da transformação da raiz quadrada, mas o autor informa que foi obtida pelo mesmo processo.

BARTLETT (1936) utiliza a transformação da raiz quadrada com o objetivo de corrigir heterogeneidade de variâncias em experimentos cujos dados se adaptem a uma distribuição de Poisson.

A razao de tal indicação é explicada pela propriedade da distribuição, que relaciona a média com a variância. Segundo o autor, tal relacionamento provoca, em dados experimentais, heterogeneidade dos erros e assimetrias.

O tipo de experimento que Bartlett identifica com a distribuição de Poisson, é do tipo de campo, com dados inteiros, advindos de contagens; de natureza não limitada ou que ocorram num dos extremos do intervalo.

Partindo do pressuposto de que a média dos dados observados é uma estimativa eficiente do parâmetro m de uma série de Poisson e considerando que, para médias pequenas entre 0 e 2 ou 3 a descontinuidade de tal curva é muito acentuada, o autor sugere que se use neste caso, a transformação $\sqrt{x + 1/2}$, por anologia com a correção de continuidade, no teste de χ^2 com 1 grau de liberdade, devida a F. Yates, afirmando que a variância, nestas condições, seria mais estável que a de $\sqrt{x}$.

COCHRAN (1940) discute bases teóricas para o uso das transformações da raiz quadrada e angular quando os dados seguem, respectivamente, as leis de Poisson e binomial. Desenvolve equaçoes de estimação para cada caso, usando o método da máxima verossimilhança, mas admite que o método é muito complicado para uso frequen te. Nas aplicações numéricas, conclui que a transformação da raiz quadrada e do arco seno sao apropriadas mesmo para dados derivados de pequenos números, exceto quando há parcelas nulas, no caso de Poisson, e porcentagens próximas de zero ou 100 no caso da binomial.

BEALL (1942), partindo de gráficos obtidos em seus experimentos, considerou que as relações entre médias e variâncias eram mais curvilíneas do que lineares, como havia sido sugerido por BARTLETT (1936). Utilizou a forma:

$$\sigma^2 = m + km^2 \ ,$$

para expressar o relacionamento. Através do método já utilizado por TIPPETT (1934) e BARTLETT (1936) e desta relação, o autor chegou a transformação:

$$y = k^{-1/2} \ \text{senh}^{-1} \ (kx)^{1/2}$$

como a forma mais apropriada, sempre que necessário estabilizar variâncias. O autor desenvolve, em série, a expressão acima

$$y = x^{1/2} - \frac{1}{6} k x^{3/2} + \frac{3}{40} k^2 \qquad \frac{5}{112} k^3 x^{7/2} + \ldots$$

para mostrar que a mesma se reduziria a

$$y = \sqrt{x} \qquad \text{se} \quad k = 0 \ ,$$

e, para valores suficientemente grandes de kx a série tenderia a

$$y = \log x$$

Beall também considera, neste artigo, a transformaçao $\sqrt{x + 1/2}$, proposta por Bartlett em 1936, comparando-a com a sua forma que considerou mais eficiente, se comparadas quanto à homogeneização de variâncias e de médias.

CURTISS (1943) procurou bases teóricas para o uso das transformações através da busca de soluções assintóticas para os problemas de normalização e estabilização de variâncias. Considerando a transformação $\sqrt{x + \alpha}$,o autor cita a transformação de Bartlett (α = 1/2) e seus resultados quanto à convergência da variância; no entanto, considera que a questão de convergência à normalidade e a possibilidade de selecionar um valor ótimo para α, permanece aberto.

BARTLETT (1947) define os objetivos das transformaçoes de dados, como sendo:

1) Eliminar o relacionamento da média com a variância, causadores de heterogeneidade do erro e não-normalidade (assimetria acentuada);

2) Normalizar a distribuição da variável;

3) Levar a média aritmética a ser uma estimativa eficiente da verdadeira média, para qualquer grupo particular de medidas;

4) Tornar os efeitos lineares e aditivos.

Segundo o autor, num mesmo experimento pode-se ter uma ou mais razoes para aplicar transformações. Embora seja difícil conseguir todos os objetivos numa só transformação, se há uma causa comum para a falha de diversas hipóteses, ou se elas estão de algum modo relacionadas, a transformação que vise à obtenção de um dos ítens pode frequentemente aproximar os outros. Como exemplo, cita: a obtenção de estabilidade para a variância, normalidade e aditividade implicam a obtenção de uma média eficiente.

No mesmo artigo, a suposição de que a variância fosse uma função da média:

$$\sigma_x^2 = f(m)$$

e usada para chegar a relação

$$\sigma_g^2 = \left(\frac{dg}{dm}\right)^2 f(m)$$

onde g é uma função de x. O processo e o mesmo já utilizado por Tippet em 1934.

Fazendo

$$\sigma_x^2 = m$$

como na distribuição de Poisson, a solução da equaçao diferencial e imediata:

$$g(x) = \sqrt{x}$$

dispensando-se as constantes.

É também sugerido que em se fazendo:

$$\sigma_x^2 = m^2$$

ou

$$\sigma_x = m$$

a transformação obtida e a logarítmica.

A transformação $\sqrt{x + 1/2}$, seria sugerida no lugar de $\sqrt{x}$ quando ocorressem zeros entre os valores observados. Ambas as formas são indicadas, mesmo que os dados não se adaptem a uma distribuição de Poisson, mas desde que a heterogeneidade das variâncias esteja presente.

COCHRAN (1947) enumera os objetivos da análise de variância e as hipóteses básicas do método, descrevendo os efeitos das falhas de cada uma e as correçoes possíveis. Prescreve transformações de dados para a não-normalidade, para a não aditividade e a heterogeneidade quando surge da não-normalidade, decorrente, frequentemente, do relacionamento da média com a variância, o que ocorre quando a variância oscila com a média a ponto de nao se poder atribuir um único valor dela a todos os tratamentos.

ANSCOMBE (1948) comenta, entre outras, a transformaçao

$$y = \sqrt{x + 3/8}$$

deduzida por A.H.L. Johnson.

Em comparação com

$$y = \sqrt{x + 1/2}$$

Anscombe considera que a forma de Johnson é mais eficaz na convergência da variância. Ao ser testada, usando-se os coeficientes de assimetria e curtose, mostrou-se bastante eficiente, tanto mais quanto as médias aumentaram.

KLECZKOWSKI (1949), estudando o numero de lesões locais provocadas por vírus, usou as medidas de assimetria e curtose deduzidas por Fisher, adaptando-as a um experimento com repetições. Usou a transformação logarítmica para obter normalidade e variância constante na distribuição dos dados.

TUKEY (1949) expõe o processo associado à análise de variância, apropriado para detectar a não-aditividade dos efeitos, com o isolamento de 1 grau de liberdade do resíduo. Considera os casos em que a transformação de dados deve ser empregada e apresenta um quadro onde o sinal da soma final de produtos, obtida da análise, serve como indicador da transformação a ser usada.

FISHER (1954) analisa várias transformações. Enfatiza a importância da aditividade do modelo, quando se analisa efeitos de vários fatores, controláveis ou não, e estes são muito variáveis. Adverte quanto à excessiva importância dada à homogeneidade de variâncias, como objetivo das transformações. Ao comentar a transformação introduzida por BARTLETT (1936),

$$y = \sqrt{x + 1/2}$$

o autor diz: "é difícil prever o que influenciou Bartlett a levar adiante sua proposta de ajustar a variável de Poisson". Fisher considerou a comparaçao com a correção de F. Yates, para continuidade, como muito fraca e acrescentou que Bartlett não havia demonstrado que tal mudança tivesse resultado em melhoramentos efetivos.

ARRUDA (1959) fez uso da transformação da raiz quadrada buscando maior homogeneidade das variâncias e aditividade dos efeitos de tratamentos e blocos. Após a transformação, usou o teste de TUKEY (1949) para a não-aditividade, com resultado não significativo. O exame das variâncias permitiu considerá-las homogêneas e o coeficiente de variação baixou de 42,8% para 19,1%.

AQUINO (1969) utilizou-se da transformação logarítmica para normalizar dados pluviométricos. Com o auxílio da computaçao, o autor usou o princípio de que o valor da assimetria, medido pelo coeficiente do momento de assimetria, não deveria se afastar de 0,8. Com esta condição, calculou a constante da transformação

$$y = \log (x + a).$$

Assimetrias e curtoses foram testadas pelo teste t, usando as variâncias teóricas deduzidas por Fisher.

ARRUDA (1971) apresentou o desenvolvimento detalhado do processo utilizado por TIPPETT (1934) para justificar o uso da transformação angular, para dados de porcentagem, admitidos como tendo distribuição binomial. Mostrou que apos a transformação, a variância deveria se estabilizar em torno de

$$\frac{1}{4N} \quad \text{ou} \quad \frac{821}{N}$$

conforme o arco da transformação estivesse medido em radianos ou graus. Embora isto ocorra quando testado com dados da distribuição binomial, o autor observa que, em 35 experimentos, somente em 9 deles, a variância residual concordou com a teórica.

DEMÉTRIO (1978) aplicou as transformações da raiz quadrada, logarítmica e angular a dados experimentais, testando os resultados quando à não-aditividade, heterogeneidade de variâncias e não normalidade. Apresentou a justificativa das transformações, pelo método descrito por TIPPETT (1934), mostrando a convergência das variâncias, nos três casos utilizados. Usou o coeficiente do momento de assimetria para teste de normalidade.

4. MATERIAL E MÉTODOS

A fim de alcançar o objetivo a que se propoe o presente trabalho, foram utilizados dados simulados, gerados por computação eletrônica, em linguagem FORTRAN, com equipamento IBM-1130,de configuração de 16K.

Os seguintes passos foram seguidos:

1) Inicialmente 1000 dados, com distribuição normal foram gerados. Foram utilizadas duas sub-rotinas do Manual da IBM.

A primeira sub-rotina utilizada foi a RANDU, sendo que sua chamada é feita por:

CALL RANDU (IX, IY, YFL) ,

onde IX é um numero inteiro positivo, diferente de zero, que precisa ser fornecido no momento em que se chama, pela primeira vez, a sub-rotina RANDU, constando do programa original.

Após a primeira chamada, IX deve ter valor igual ao prévio valor de IY. Este IY é o número inteiro aleatório gerado.

Outra sub-rotina auxiliar é a GAUSS. Sua chamada e feita por:

CALL GAUSS (IX, S, AM, V)

onde, AM = média da amostra: zero;

S = desvio padrão: 0,25.

Os dados foram subdivididos em 10 amostras, consideradas como repetições.

2) Foram calculados, para cada amostra, a média ($\bar{X}$), o desvio padrão (s), o 3º e 4º momentos (m_3 e m_4) em relação à média:

$$m_3 = \frac{\Sigma(X - \bar{X})^3}{n} \quad ,$$

$$m_4 = \frac{(X - \bar{X})^4}{n} \quad ;$$

os coeficientes dos momentos de assimetria (g_1) e curtose (g_2):

$$g_1 = \frac{m_3}{s^3} \quad ,$$

$$g_2 = \frac{m_4}{s^4} - 3 \quad ;$$

as variâncias teóricas dos coeficientes:

$$V(g_1) = \frac{6n(n - 1)}{(n - 2)(n + 1)(n + 3)}$$

$$V(g_2) = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}$$

os valores do teste t:

$$t_1 = \frac{g_1}{\sqrt{V(g_1)}} ,$$

$$t_2 = \frac{g}{\sqrt{V(g_2)}} ;$$

e o coeficiente de variação:

$$C.V. = \frac{100s}{\bar{X}} .$$

3) Em seguida, os dados, amostra por amostra, foram sendo sucessivamente incrementados por constantes, de modo a obter-se novas emostras. As constantes somadas foram: 1,0; 2,0; 10,0 e 15,0.

4) A cada nova amostra obtida, os procedimentos descritos no passo 2 foram repetidos. Até este estágio, contava-se com 5 amostras, cada uma com 10 repetições.

5) A partir daí, as transformações foram aplicadas em sua forma inversa, conforme ficou esclarecido na Introdução.

6) Aos dados transformados foi novamente aplicado o procedimento indicado no passo 2.

Ao término da computação, o material disponível para estudo constava de 20 amostras, cada uma com 10 repetições de 100 dados cada.

Os afastamentos da normalidade foram medidos em função da assimetria e curtose, através dos coeficientes dos momentos de assimetria e curtose. Foram testados pelo teste $\underline{t}$ como foi feito por SNEDECOR (1967) e a variância utilizada no teste foi a variância teórica para uma amostra do mesmo tamanho: V(g) deduzida por FISHER (1958).

Foram utilizadas amostras de 100 dados, por julgar-se que uma amostra desta ordem é bastante representativa. Cada experimento foi repetido 10 vezes, para garantir a validade das conclusões.

As médias teóricas utilizadas foram: 0 (zero), 1,0; 2,0; 10,0 e 15,0, procurando assim cobrir os campos mais polêmicos do assunto.

As transformações utilizadas foram:

<u>Diretas</u>		<u>Inversas</u>
$y = \sqrt{x}$		$x = y^2$
$y = \sqrt{x + 1/2}$	(BARTLETT, 1936)	$x = y^2 \quad 1/2$
$y = \sqrt{x} + 3/8$	(ANSCOMBE, 1948)	$x = y^2 \quad 3/8$

Para verificar a significância do teste t, os valores t_1 e t_2 foram comparados com:

Nível 5%: 1,96 ,

Nível 1%: 2,57 .

O teste t é bilateral, com ∞ g.l.

5. RESULTADOS E DISCUSSÃO

Ao apresentar os resultados obtidos, uma parte da pesquisa foi omitida. A primeira amostra, gerada com média teórica μ = 0 apresentou dados negativos. A nosso ver, estes dados não poderiam ser comparados com as demais amostras, devido à característica do trabalho. Durante a aplicação da transformação inversa, estes dados seriam elevados ao quadrado e estaria evidentemente eliminado o problema de sinal. No entanto, ao considerarmos a transformação direta (extração da raiz quadrada do dado já positivo), não teríamos um critério para estabelecer o sinal da raiz. Felizmente, nas demais amostras, obtidas por soma de constantes aos dados originais, os numeros resultantes foram positivos.

Tabela 1 - Médias das amostras obtidas com média teórica $\mu = 1$.

REPETIÇÃO	SEM TRANSFORMAÇÃO	TRANSFORMAÇÃO INVERSA		
		y^2	$y^2 - 1/2$	$y^2 - 3/8$
1	1,02	1,12	0,62	0,74
2	1,02	1,13	0,63	0,75
3	1,01	1,08	0,58	0,70
4	0,99	1,05	0,55	0,68
5	0,99	1,03	0,53	0,65
6	1,00	1,08	0,58	0,70
7	0,99	1,05	0,55	0,67
8	1,04	1,15	0,65	0,77
9	1,01	1,08	0,58	0,71
10	0,99	1,03	0,53	0,65

Tabela 2 - Médias das amostras obtidas com média teórica $\mu = 2$.

REPETIÇÃO	SEM TRANSFORMAÇÃO	TRANSFORMAÇÃO INVERSA		
		y^2	$y^2 - 1/2$	$y^2 - 3/8$
1	2,02	4,17	3,67	3,79
2	2,02	4,17	3,67	3,80
3	2,01	4,11	3,61	3,73
4	1,99	4,04	3,54	3,67
5	1,98	4,00	3,50	3,62
6	2,00	4,09	3,59	3,71
7	1,99	4,03	3,53	3,65
8	2,04	4,25	3,75	3,86
9	2,01	4,11	3,61	3,73
10	1,99	4,01	3,51	3,64

Tabela 3 - Médias das amostras obtidas com média teórica μ = 10.

REPETIÇÃO	SEM TRANSFORMAÇÃO	TRANSFORMAÇÃO INVERSA		
		y^2	y^2 - 1/2	y^2 - 3/8
1	10,02	100,55	100,05	100,18
2	10,02	100,56	100,06	100,18
3	10,01	100,35	99,84	99,97
4	9,99	99,96	99,46	99,59
5	9,99	99,78	99,28	99,40
6	10,00	100,18	99,68	99,80
7	9,99	99,88	99,38	99,51
8	10,04	100,91	100,41	100,54
9	10,01	100,31	99,81	99,93
10	9,99	99,89	99,39	99,51

Tabela 4 - Médias das amostras obtidas com média teórica μ = 15.

REPETIÇÃO	SEM TRANSFORMAÇÃO	TRANSFORMAÇÃO INVERSA		
		y^2	y^2 - 1/2	y^2 - 3/8
1	15,02	225,8	225,3	225,4
2	15,02	225,8	225,3	225,4
3	15,01	225,5	225,0	225,1
4	14,99	224,9	224,4	224,5
5	14,99	224,6	224,1	224,3
6	15,01	225,2	224,7	224,9
7	14,99	224,8	224,3	224,4
8	15,04	226,3	225,8	226,0
9	15,01	225,4	224,9	225,1
10	14,99	224,8	224,3	224,4

Tabela 5 - Desvios padrões.

REPETIÇÃO	SEM TRANSFORMAÇÃO	TRANSFORMAÇÃO INVERSA			
		$\mu = 1$	$\mu = 2$	$\mu = 10$	$\mu = 15$
1	0,27	0,53	1,07	5,36	8,04
2	0,28	0,57	1,13	5,64	8,46
3	0,22	0,44	0,88	4,44	6,66
4	0,25	0,51	1,02	5,08	7,62
5	0,23	0,48	0,95	4,69	7,03
6	0,26	0,51	1,03	5,15	7,73
7	0,26	0,52	1,03	5,17	7,75
8	0,25	0,52	1,03	5,07	5,59
9	0,25	0,50	0,98	4,91	7,37
10	0,21	0,43	0,85	4,22	6,32

Tabela 6 - Variâncias.

REPETIÇÃO	SEM TRANSFORMAÇÃO	TRANSFORMAÇÃO INVERSA			
		$\mu = 1$	$\mu = 2$	$\mu = 10$	$\mu = 15$
1	0,0720	0,284	1,136	28,69	64,63
2	0,0797	0,321	1,267	31,80	71,61
3	0,0493	0,197	0,783	19,67	44,30
4	0,0645	0,261	1,032	25,78	58,03
5	0,0548	0,230	0,893	21,97	49,38
6	0,0665	0,265	1,057	26,55	59,77
7	0,0668	0,271	1,068	26,69	60,07
8	0,0639	0,274	1,052	25,68	57,68
9	0,0603	0,246	0,970	24,13	54,28
10	0,0444	0,184	0,719	17,78	39,97

Tabela 7 - Resultados do teste t para assimetria.

REPETIÇÃO	SEM TRANSFORMAÇÃO	TRANSFORMAÇÃO INVERSA			
		$\mu = 1$	$\mu = 2$	$\mu = 10$	$\mu = 15$
1	-1,38	1,20	-0,06	-1,12	-1,21
2	-1,19	2,04*	0,51	-0,84	-0,95
3	-1,09	1,74	0,35	-0,80	-0,90
4	-0,06	2,58**	1,28	0,21	0,12
5	0,93	3,47**	2,26*	1,21	1,16
6	-0,58	1,50	0,48	-0,37	-0,44
7	0,01	3,10**	1,57	0,32	0,22
8	-0,63	1,81	0,60	-0,39	-0,47
9	-0,41	1,48	0,56	-0,22	-0,28
10	0,60	2,90**	1,76	0,84	0,76

Tabela 8 - Resultados do teste t para curtose.

REPETIÇÃO	SEM TRANSFORMAÇÃO	TRANSFORMAÇÃO INVERSA			
		$\mu = 1$	$\mu = 2$	$\mu = 10$	$\mu = 15$
1	-0,67	-0,84	-0,99	-0,77	-0,74
2	0,15	-0,06	-0,43	-0,06	0,01
3	0,33	0,20	0,04	0,24	0,27
4	-0,69	0,41	-0,39	-0,67	-0,68
5	-0,20	0,67	0,06	-0,18	-0,19
6	-1,38	-1,26	-1,48	-1,43	-1,41
7	-0,21	1,49	0,33	-0,15	-0,17
8	-0,93	0,02	-0,62	-0,89	-0,91
9	-1,51	-1,37	-1,57	-1,54	-1,53
10	-0,50	0,84	0,01	-0,43	-0,46

Tabela 9 Coeficientes de variação (%) dos dados sem transformação.

REPETIÇÃO	AMOSTRAS			
	μ = 1	μ = 2	μ = 10	μ = 15
1	26,2	13,2	2,67	1,78
2	27,6	13,9	2,81	1,88
3	21,9	11,0	2,22	1,48
4	25,5	12,7	2,54	1,69
5	23,7	11,8	2,34	1,56
6	25,6	12,9	2,58	1,72
7	26,1	13,0	2,58	1,72
8	24,2	12,4	2,51	1,68
9	24,2	12,2	2,45	1,63
10	21,2	10,6	2,10	1,40

Tabela 10 - Coeficientes de variação (%) após transformação y^2.

REPETIÇÃO	AMOSTRAS			
	μ = 1	μ = 2	μ = 10	μ = 15
1	47,6	25,6	5,32	3,56
2	50,3	27,0	5,60	3,74
3	41,1	21,5	4,42	2,95
4	48,5	25,1	5,08	3,38
5	46,7	23,6	4,70	3,12
6	47,8	25,1	5,14	3,43
7	49,7	25,6	5,17	3,44
8	45,5	24,2	5,02	3,35
9	45,7	24,0	4,90	3,27
10	41,7	21,1	4,22	2,81

Tabela 11 - Coeficientes de variação (%) após transformação y^2-1/2.

REPETIÇÃO	AMOSTRAS			
	μ = 1	μ = 2	μ = 3	μ = 4
1	86,0	29,0	5,35	3,56
2	90,3	30,6	5,63	3,76
3	76,6	24,5	4,44	2,95
4	92,4	28,7	5,10	3,40
5	91,1	27,0	4,72	3,13
6	89,2	28,6	5,17	3,44
7	95,0	29,3	5,19	3,45
8	80,5	27,5	5,04	3,36
9	84,8	27,3	4,92	3,28
10	81,1	24,1	4,24	2,81

Tabela 12 - Coeficientes de variação (%) apos transformação y^2- 3/8.

REPETIÇÕES	AMOSTRAS			
	μ = 1	μ = 2	μ = 10	μ = 15
1	71,6	28,1	5,34	3,36
2	75,3	29,6	5,62	3,75
3	63,0	23,7	4,43	2,95
4	75,4	27,7	5,09	3,39
5	73,6	26,1	4,71	3,13
6	73,4	27,7	5,16	3,44
7	77,4	28,3	5,19	3,45
8	67,5	26,6	5,04	3,36
9	69,9	26,4	4,91	3,27
10	65,6	23,3	4,23	2,81

5.1 - Homogeneidade das Médias

A transformação inversa provocou aumento da amplitude do intervalo de variação das médias em cada amostra, como se pode notar na Tabela 13.

Quando considerada a transformação direta (dados sem transformação), nota-se que a amplitude do intervalo é menor, mantendo-se estável para as várias amostras utilizadas.

Tabela 13 - Amplitude do intervalo de variação das médias.

μ	Sem Transformação	Com Transformação Inversa
1	0,04	0,12
2	0,06	0,24
10	0,05	1,13
15	0,05	1,70

5.2 - Homogeneidade das Variâncias

Pela Tabela 14, podemos observar que, apos submetidos à transformação inversa, as variâncias perderam homogeneidade. Nesta tabela observamos o crescimento da amplitude do intervalo,com o aumento da média. Entre os dados não transformados não houve variação, uma vez que a soma de constantes aos dados não altera a variância.

Tabela 14 - Amplitude do intervalo de variação da variância em cada amostra.

	Sem Transformação	Transformação Inversa	F_{max}
1	0,0353	0,137	1,75
2	0,0353	0,548	1,76
10	0,0353	14,020	1,79
15	0,0353	31,640	1,79

Uma vez que a heterogeneidade das variância precisava ser comprovada, o teste F_{max} foi utilizado, por sua simplicidade de aplicação, embora não estivesse planejado.

Pela tabela encontrada em HARTLEY (1950), pode-se considerar o resultado do teste como significativo ao nível de 5% de probabilidade, para todas as amostras.

5.3 - Comparação das Três Formas de Transformação

Não houve diferenças entre as três formas de transformação aplicadas, de modo a indicar uma delas como a mais eficiente.

Os resultados das três sao iguais, quanto às variâncias e coeficientes dos momentos de assimetria e curtose.

A variação das médias pode ser tomada como previsível, uma vez que a soma ou subtração de uma constante aos dados, afeta a média em valor igual ao da constante:

$$\frac{\Sigma(x_i \pm k)}{n} = \bar{x} \pm k \quad .$$

A variância e o 3º e 4º momentos não sao afetados pela constante:

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{\Sigma\left[(x_i \quad k) - (\bar{x} \quad)\right]^2}{n - 1}$$

$$= \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \quad ,$$

dentro de cada amostra com a mesma média teórica (μ).

5.4 - Resultados do Teste t Para Assimetria e Curtose

Conforme se pode ver pela Tabela 7, apenas nas faixas de médias mais baixas a transformação inversa provocou assimetria significativa. Quando μ = 1 houve 5 casos em 10 de assimetrias detectadas pelo teste t, todas positivas. Os valores da tabela apresentam tendência decrescente com o aumento da média dos dados.

Na Tabela 8, nota-se que a transformação inversa nao teve efeito sobre a curtose dos dados. Não houve nenhum valor de t significativo.

5.5 - Coeficientes de Variação

Os coeficientes de variação para os dados normais (não transformados) foram razoavelmente altos para médias pequenas ($\mu = 1$), baixando sensivelmente para as médias mais elevadas.

Quanto aos dados transformados, a transformação $y^2 - 1/2$ apresentou os maiores coeficientes de variação, para dados com médias pequenas, devido à subtração da constante que ocasionou as menores médias.

Não houve indicação de que o coeficiente de variação tivesse alguma outra função, senão a de indicar o relacionamento entre a magnitude da média e do desvio padrão. Alguns trabalhos usam o coeficiente de variação como indicador de não-aditividade. Neste trabalho, porém, não foi possível comprovar-se isto.

6. CONCLUSÕES

Embora tenhamos testado as transformações inversa, e possível, em alguns casos, inferir resultados que se obteriam se a transformação direta fosse usada.

Não se pode concluir que a transformação da raiz quadrada fosse realmente eficiente para remover não-normalidade, exceto para dados com médias pequenas. Mesmo assim, somente a assimetria parece sensível a esta transformação. A transformação não parece apropriada para remover curtose.

Embora não tenha sido planejado nenhum teste para verificar a homogeneidade das variâncias, o teste F_{max} foi aplicado com resultado significativo, indicando a perda de homogeneidade, após a transformação inversa. Considerando a transformação direta, a raiz quadrada parece mais indicada para homogeneizar variâncias,con cordando com a teoria das transformações.

A comparação das três formas de transformação

$$y = \sqrt{x} \quad ,$$

$$y = \sqrt{x + 1/2} \quad ,$$

$$y = \sqrt{x + 3/8} \quad ,$$

nao demonstrou superioridade de nenhuma delas sobre as outras, quando analisadas em relação à normalização de dados e heterogeneização de variâncias.

7. SUMMARY

The use of the square root transformation is studied by means of the application of its reciprocal to simulated data.

Four samples with 10 replications of 100 values each are used. The theoretical values of the samples are: 1, 2, 10 and 15.

Skewness and kurtosis are measured in order to test non-normality.

The maximum F-ratio is used as a short-cut test for heterogeneity of variance.

The tests demonstrated that the transformation can be efficient in the correction of skewness in data with low averages.

With regard to kurtosis, there was no reaction.

The resultant heterogeneity of the variances demonstrated that the transformation is better used to obtain homogeneity of data than normality.

8. BIBLIOGRAFIA

ANSCOMBE, F.J., 1948. The Transformation of Poisson, Binomial and Negative-Binomial Data. Biometrika, Cambridge, 35: 246-254.

AQUINO, L.H., 1969. Análise Estatística de Dados Pluviométricos com Auxílio da Programação Fortran. Piracicaba, ESALQ/USP. (Dissertação de Mestrado).

ARRUDA, H.V. de, 1959. Aplicação da Transformação da Raiz Quadrada na Análise de Variância de Dados Experimentais. Bragantia, Campinas, 18: XV-XIX.

ARRUDA, H.V. de, 1971. Transformação Angular de Dados de Porcentagem em Face da Distribuição Binomial. Piracicaba, ESALQ/USP. (Dissertação de Mestrado).

BARTLETT, M.S., 1936. The Square Root Transformations in Analysis of Variance. Jour. Royal Stat. Soc. Suppl., Londres, 3: 68-78.

BARTLETT, M.S., 1947. The Uses of Transformations. Biometrics, Raleigh, 3: 39-52.

BEALL, G., 1942. The Transformation of Data from Entomological Field Experiments. Biometrika, Cambridge, 32: 243-262.

COCHRAN, W.G., 1940. The Analysis of Variance when Experimental Errors follow the Poisson or Binomial Laws. Ann. Math. Stat., Baltimore, 11: 335-347.

COCHRAN, W.G., 1947. Some Consequences when the Assumptions for the Analysis of Variance Are Not Satisfied. Biometrics, Raleigh, 3: 22-38.

CURTISS, J.H., 1943. On Transformation Used in the Analysis of Variance. Ann. Math. Stat., Baltimore, 14: 107-122.

DEMÉTRIO, C.G.B., 1978. Transformação de Dados. Efeitos sobre a Análise de Variância. Piracicaba, ESALQ/USP. (Dissertação de Mestrado).

DOLBY, J.L., 1963. A Quick Method for Choosing a Transformation. Technometrics, 5: 317-325.

EISENHART, C., 1947. The Assumptions Underlying the Analysis of Variance. Biometrics, Raleigh, 3: 1-21.

FISHER, R.A., 1954. The Analysis of Variance with Various Binomial Transformations. Biometrics, Raleigh, 10: 130-139.

FISHER, R.A., 1958. Statistical Methods for Research Workers. 13a. ed. Londres, Oliver and Boyd.

HARTLEY, H.O., 1950. The Maximum F-Ratio as a Short-Cut Test for Heterogeneity of Variance. Biometrika, Cambridge, 37: 308-312.

KLECZKOWSKI, A., 1949. The Transformation of Local Lesion Counts for Statistical Analysis. Ann. App. Biol., 36: 139-152.

SNEDECOR, G.W. e W.G. COCHRAN, 1967. Statistical Methods. 6a. ed. Ames, Iowa, The Iowa State University Press.

TIPPETT, L.H.C., 1934. Statistical Methods in Textile Research. Part 2. Uses of the Binomial and Poisson Distributions. Shirley Inst. Mem., 13: 35-72.

TUKEY, J.W., 1949. One Degree of Freedom for Non-Additivity. Biometrics, Raleigh, 5: 232-242.