# Análise Estatística da Distribuição de Poisson (*)

**J. T. A. Gurgel**
*Docente-livre*
*Seção de Genética*

*Escola Superior de Agricultura "Luiz de Queiroz", Universidade de S. Paulo*

ÍNDICE



(*) Entregue para publicação em 2 de Agosto de 1945.

## INTRODUÇÃO

A distribuição ou série de Poisson ou ainda a Lei dos pequenos números é de há muito conhecida na matemática, tendo sido publicada originalmente pelo seu autor em 1837.

Na estatística analítica, a série de Poisson tem sido bastante aplicada, tanto teórica como pràticamente. Assim, seguindo esta lei, encontramos na literatura exemplos sôbre a distribuição da frequência em vários tipos de acidentes humanos( na indústria, no tráfego, etc.), a diluição de microorganismos e posterior contagem no hematímetro, a frequência de emissão das partículas $\alpha$ do polônio em pequenos intervalos, a frequência de algumas doenças em plantas e animais, etc.

Todavia, a-pesar-de ter sido já bem estudada, os livros de estatística trazem-na sob uma forma bastante complexa, principalmente no que se refere à sua análise. Isto dificulta a sua utilização e tentando contornar esta dificuldade, apresentamos aqui com detalhes os testes de análise e a sua aplicação a exemplos práticos encontrados no nosso trabalho sôbre Citrus.

Existem vários modos de definir a distribuição de Poisson. Assim, por exemplo, a definição dada por ARNE FISHER no seu livro "Frequency Curves" (2), pag. 95, baseia-se na teoria dos semi-invariantes, enquanto que YULE, em "Introduction to the Theory of Statistics" (11), pag. 187, chega à distribuição de Poisson partindo da distribuição binomial; há ainda outros autores que a definem baseando-se em considerações sôbre a probabilidade de tirar bolas brancas e pretas de urnas, onde o número de bolas permanece relativamente inconstante.

Todavia, podemos reunir em 3 pontos os principais caracteres da distribuição de Poisson: a) trata-se de uma série descontínua, na qual as frequências das classes só podem ser números inteiros e positivos; b) a primeira classe teórica tem sempre o valor zero, e em consequência disto, a pressão do limite zero geralmente se manifesta, tornando a distribuição assimétrica; c) a frequência dos acontecimentos esperados é tão pequena em relação ao número total de acontecimentos possíveis, que a frequência observada é pràticamente independente do número total de observações.

O teste mais utilizado para determinar se as frequências observadas não diferem daquelas esperadas na série de Poisson é o $X^2$ teste e ùltimamente BRIEGER recomenda um novo teste de $\vartheta$, o qual já foi recentemente aplicado em dois trabalhos publicados em Bragantia, sob os títulos de "Comportamento de variedades e progênies de fumo na resistência ao vira-cabeça" (1) e "Poliembrionia em Citrus" (6).

Reunindo assim a exposição sôbre o modo de calcular as frequências esperadas e a aplicação de $X^2$ teste, o processo de calcular o novo teste de $\vartheta$ e os exemplos por nós encontrados em Citrus, esperamos que o presente trabalho represente algum valor aos estudiosos da estatística.

Queremos deixar aqui consignados os nossos agradecimentos ao Prof. BRIEGER pela crítica feita no desenvolvimento de todo êste trabalho.

## DERIVAÇÃO TEÓRICA

Como já foi mencionado, poderíamos derivar a distribuição de POISSON de várias maneiras, mas aqui preferimos derivá-la partindo ùnicamente da distribuição binomial.

Na **distribuição binomial** $(p+q)n$, as probabilidades p e q podem tomar valores iguais ou diferentes, porém não excessivamente diferentes, contanto que se mantenha a igualdade $p+q = 1$; o expoente n pode variar de 1 até $\infty$. Se $p = q$, a curva é simétrica e no caso do expoente n crescer até tornar-se infinitamente grande, a distribuição, que era descontínua, passa a ser contínua e temos, então, a **distribuição normal,** ou de GAUSS, que é uma curva "lisa".

Se $p \lesseqgtr q$ e o expoente n é baixo, a distribuição é assimétrica e essa assimetria será tanto mais acentuada quanto maior fôr a diferença entre as probabilidades p e q. Se porém o expoente n crescer, a assimetria vai pràticamente desaparecendo e no caso de n tomar valores muito altos, poderemos ainda aceitar a distribuição como aproximadamente simétrica.

Se $p << q$, consequentemente a probabilidade q aproxima-se de 1, e mesmo para valores de **n** pequenos ou grandes, a distribuição fica assimétrica e descontínua, com uma frequência média **m** pràticamente independente do valor n. Êste é o limite entre a distribuição binomial e a **distribuição de Poisson.**

Para fins práticos, aceitamos que na distribuição binomial a probabilidade p pode chegar até 0,05 ou 5%. Aqui já notamos que o produto das probabilidades $p.q = 0{,}05.0{,}95 = 0{,}0475$ aproxima-se bastante de $p = 0{,}05$.

Se a probabilidade p cai ainda mais, chegando por exemplo até 0,01 ou 1%, e o produto $p.q = 0{,}01.0{,}99 = 0{,}0099$ torna-se igual a p ou 0,01 e portanto independente de q, que fica pràticamente igual a 1.

Como já frisámos no capítulo anterior, sendo na distribuição de POISSON a frequência esperada muito pequena em relação ao número de acontecimentos, as frequências ficam pràticamente independentes do número total de observações.

Para provarmos isto, suponhamos que numa cidade a probabilidade diária de mortes em acidentes do trânsito é de 0,1% ou seja 1 pessoa em 1.000 habitantes. Numa cidade de 5.000 almas, a probabilidade seria portanto de 5 mortes, e numa outra de 6.000 habitantes seria de 6 mortes. A menor variação possível de mortes é um indivíduo inteiro e só o aumento da população da ordem de um ou mais milhares de habitantes terá efeito. Uma vez que devemos comparar séries baseadas sôbre o número esperado de mortes, isto é, no nosso experimento que varia em centenas de habitantes, o número esperado ou frequência esperada fica independente do total.

As principais fórmulas das duas distribuições discutidas são:

Distribuição binomial: $\bar{v} = pn \qquad \sigma = \pm \sqrt{pqn}$

Distribuição de Poisson: $m = f\ esp = pn \qquad \sigma i = \pm \sqrt{pn} = \pm \sqrt{m}$

Por essas fórmulas é evidente que para podermos desenvolver qualquer binômio, precisamos saber o valor de **p** e de **n**; o valor de q é automàticamente definido pela fórmula $p+q=1$.

Na distribuição de POISSON, porém, precisamos saber apenas o valor de **m**. Todavia, sabemos por definição que $m = pn$ e não só é desnecessário saber os valores individuais de **p** e **n**, como em muitos casos nem se pode determinar êsses valores.

Para explicarmos isso, podemos citar da literatura alguns exemplos:

1) No caso clássico, tantas vezes citado, de QUETELET e VON BORTKIEWICZ, (11) sôbre o número de acidentes por coice de cavalo em alguns corpos do exército prussiano, durante 20 anos (1875-94), além do valor **m**, também o valor **n** é acidentalmente conhecido. O valor **m** representa o número de acidentes e o **n** é o número de soldados por corpo de exército. Daí pode-se calcular o valor de $p = m : n$

2) No exemplo da resistência do fumo a "vira-cabeça", de BRIEGER e colaboradores, (1) **m** representa o número médio de plantas doentes por canteiro e o **n** o número total de plantas por canteiro.

3) De outro lado, quando determinamos o número de bactérias em um quadradinho do hematímetro, exemplo dado por STUDENT (11), podemos sempre calcular a média **m** de bacté-

rias encontradas em tôdas as contagens, mas o número **n** de bactérias possíveis é desconhecido; êsse número é que corresponde, no primeiro exemplo, ao número de soldados do exército que poderiam sofrer o acidente, e no segundo exemplo, ao número total de plantas por canteiro que poderiam se tornar doentes.

Até agora usámos sempre o símbolo **n** como constituindo o número total de casos possíveis, mas devemos ainda distinguir claramente o número de amostras investigadas em cada caso, que nos exemplos acima citados é respectivamente o número de corpos do exército, o número de canteiros estudados e o número de quadradinhos contados no hematímetro. Para indicar o número de amostras usaremos o símbolo **a**.

Os exemplos que iremos dar sôbre o número de sementes em Citrus, cai no caso do 3.o exemplo que acabamos de discutir, isto é, onde se conhece apenas o valor de **m**, que é o número médio de sementes encontradas por fruto; mas o valor **n**, número total de sementes possíveis, quer dizer, incluindo o número de óvulos abortivos, é desconhecido. Além disso, conhecemos também o valor de **a**, que é o número total de frutos contados.

---

Terminando êste capítulo, queremos frisar que não existe uma só distribuição de acaso, mas que o resultado do jôgo do acaso dependerá evidentemente das condições concretas nas quais cada ensaio é realizado. Assim, os velhos estatísticos idealizaram diferentes modalidades de tirar bolas pretas e brancas de uma urna, creando desta forma condições variadas de uma tiragem sempre de acaso. De acôrdo com essa probabilidade e condições pre-estabelecidas, o jôgo do acaso seguirá as séries binomial ou de BERNOULLI, a de POISSON ou enfim a de LEXIS.

A curva normal ou de GAUSS-LAPLACE, como comumente é sabido, é apenas o caso extremo da distribuição binomial, onde as probabilidades favoráveis e desfavoráveis são iguais, e o número de tiragem de bolas sendo considerado infinito.

A comparação de uma distribuição observada com a série de POISSON e a determinação se êsses desvios encontrados podem ser atribuidos ao jôgo de acaso, torna-se um pouco mais difícil de que no caso da comparação com uma distribuição normal, precisando, por isso, aplicar métodos especiais. Em vez de podermos usar diretamente as tábuas existentes, vai ser geralmente necessário calcular as frequências esperadas das séries ideais de POISSON, a-pesar-de que para algumas séries, com valores especiais de m e n, essas frequências já estão calculadas, de acôrdo com uma citação de YULE (11), pag. 190.

## CÁLCULO DAS FREQUÊNCIAS ESPERADAS

A distribuição de POISSON, com n amostras e média m por amostra, tem por limite a expressão:

$$n\,e^{-m}\left(1+m+\frac{m^2}{1.2}+\frac{m^3}{1.2.3}+\cdots\cdots+\frac{m^a}{1.2.3\ldots a}\right)=$$

$$n\cdot e^{-m}+n\cdot e^{-m}\cdot m+n\cdot e^{-m}\cdot\frac{m^2}{1.2}+n\cdot e^{-m}\cdot\frac{m^3}{1.2.3}+\cdots\cdots\cdots+n\cdot e^{-m}\cdot\frac{m^a}{1.2.3\ldots a}$$

onde os têrmos seguidos indicam o valor algébrico das frequências das classes sucessivas.

Para o cálculo dessas frequências, precisamos fazer uma transformação logarítmica, quer usando os logarítmos naturais, quer os decimais. Assim, teremos em forma matemática as seguintes frequências em logarítmos:

### Logarítmos naturais

1.º têrmo: [(—m) + log nat n]
2.º têrmo: [(—m) + log nat n] + log nat m
3.º têrmo: [(—m) + log nat n] + 2 log nat m — log nat 2
4.º têrmo: [(—m) + log nat n] + log nat m — (log nat 2 + log nat 3)
a.º têrmo: [(—m) + log nat n] + a log nat m — (log nat 2 + log nat 3 + ...... + log a)

### Logarítmos decimais

1.º têrmo (—m. log e) + log n
2.º têrmo (—m. log e) + log n + log m
3.º têrmo (—m. log e) + log n + 2 log m — log 2
4.º têrmo (—m. log e) + log n + 3 log m — (log 2 + log 3)
a.º têrmo (—m. log e) + log n + a log m — [(log 2 + log3 + ..... + log a)]

O cálculo dos têrmos dessa série é muito facilitado pela organização de quadros, dos quais daremos oportunamente um exemplo. As frequências esperadas no cálculo devem estender-se até que o antilogarítmo da última frequência da série atinja a ordem de centésimos.

## TESTES DE ANALISE

**$X^2$ teste** — Êste teste tem a vantagem de nos permitir analisar a variação de cada classe e do conjunto das classes. Porém, pode acontecer, como adiante veremos, de alguns X2 parciais serem um pouco grandes, mas não ainda significantes, porém uma vez tomados em conjunto, podem dar um resultado significante. Isto dificulta o nosso julgamento e esperamos que combinando êsse teste com o $\vartheta$ teste, poderemos resolver satisfatòriamente a questão.

Como sabemos, o $X^2$ teste só é seguro para as classes onde o valor numérico da frequência esperada é igual ou maior do que 5. Acontece que sempre as frequências em uma ou nas duas extremidades da distribuição de POISSON são menores do que êstes limites. Neste caso, é necessário reunir as classes, somando as suas frequências da extremidade para dentro, até que a soma das frequências esperadas atinja ou ultrapasse o valor 5.

Uma vez calculadas as frequências esperadas, fazemos para cada classe da série um $X^2$, definido pela fórmula (f.obs — f. esp)$^2$: f. esp. e finalmente somamos todos os $X^2$ parciais.

Se na comparação dos $X^2$ parciais, com 1 grau de liberdade, resultar um valor menor do que 3,84 ou 5% de probabilidade, dizemos que o $X^2$ é **insignificante**; se fôr maior do que 6,64 ou 1% de probabilidade, dizemos que é **significante**. Se ficar compreendido entre êstes limites de 5% e 1%, dizemos que está na **região de dúvida**. Para o $X^2$ total, o grau de liberdade é igual ao número de $X^2$ somados menos um, quando se refere a média calculada, e n quando se refere a variação em volta de um valor ideal.

Finalmente, para cada caso temos que verificar na tabela se o valor obtido é ou não significante.

**$\vartheta$ teste** — Êste teste, tendo a vantagem de mostrar se a variação total da amostra observada não difere daquela esperada teòricamente, nos dá uma melhor visão do conjunto. Todavia, acontecendo do $\vartheta$ ser significante, torna-se difícil precisar qual a classe responsável pela variação excessiva.

Para realizarmos êste teste, temos que calcular a média **m**, o êrro "standard" $\sigma$ calculado pela fórmula comum da variação descontínua com classes e o êrro ideal $\sigma_i$, que ,como acima dissemos, para a distribuição de POISSON é igual a $\sqrt{m}$ Embora no cálculo dêstes valores tenhamos que usar os dados obtidos da série observada, todavia não poderemos nos esquecer dos requisitos teóricos sôbre a distribuição de POISSON, como, por exemplo, que o valor da primeira classe teórica tem que ser zero.

Para melhor compreendermos êste ponto, seria bom explicá-lo com alguns exemplos. Nos três casos da literatura citados, o valor da primeira classe foi de fato zero: nenhum soldado morto por coice de cavalo, nenhuma planta de fumo com "vira-cabeça" e nenhuma bactéria encontrada em um quadradinho do hematímetro.

O exemplo a ser aqui discutido do número de sementes por fruto, em Citrus, também cai na mesma categoria. Contràriamente a muitas outras plantas, onde os frutos só se desenvolvem quando têm no mínimo algumas sementes, o que aliás constitui a regra, os Citrus e outras plantas frutíferas como a macieira, pereira, etc., apresentam o fenômeno da partenocarpia, isto é, os frutos podem se desenvolver sem possuirem nenhuma semente. Assim, neste caso, a primeira classe da distribuição de POISSON tem o valor **zero** sementes por fruto.

Mencionaremos agora alguns exemplos de uma situação bem diferente. Num trabalho publicado em Bragantia (6), MOREIRA e GURGEL estudaram a questão da poliembrionia em várias espécies cítricas. Aqui, o material experimental é representado pelas sementes, e não haverá sementes que não contenham no mínimo um embrião. O problema é então analisar a **poliembrionia**, isto é, a ocorrência de **mais** de um embrião. Dêste modo, a primeira classe da distribuição de POISSON será formada pelas sementes de um só embrião, quer dizer, com **zero** embriões adicionais.

Neste exemplo, poderemos indicar a média de duas formas diferentes: a média geral do número de embriões por semente, e a média do número de embriões adicionais, também por semente; para o Calamondin, essas médias foram respectivamente de 3,7 e de 2,7. No caso de aplicarmos a distribuição de POISSON, sòmente o último valor deve ser utilizado.

Em um outro exemplo, tirado de dados de BRIEGER, ainda não publicados, sôbre a variação do número de fileiras em espigas de milho, a situação é semelhante à da poliembrionia. Nas linhagens de milho estudadas, o número mínimo de fileiras encontradas nas espigas foi de 8 e a variação não inclui um aumento de fileiras simples, mas sim de pares de fileiras.

Para aplicarmos a distribuição de POISSON, temos que raciocinar da seguinte forma: o que desejamos saber é a variação do aparecimento de pares de fileiras adicionais, e a primeira classe será formada pelas espigas de 8 fileiras, ou com zero par de fileiras adicionais; a segunda classe terá 10 fileiras ou um par de fileiras adicionais, etc..

Em virtude do que expuzemos, no cálculo da média e do êrro "standard" da distribuição de POISSON, usando as fórmulas da variação descontinua com classes, o primeiro centro de classe Vk é zero e os demais seguem a ordem numérica. As fórmulas a serem empregadas são:

$$m = f\,(esp) = \frac{\Sigma[f\,(obs) \cdot Vk]}{a} \qquad \sigma = \pm\sqrt{\frac{\Sigma[f\,(obs) \cdot Vk^2] - C}{a-1}}$$

$$C\,(\text{Correção}) = \frac{[\Sigma f\,(obs) \cdot Vk]^2}{a} \qquad \sigma_i = \pm\sqrt{m}$$

onde f (obs) e f (esp) são as frequências observadas e esperadas em cada classe Vk da série de POISSON e a o número de amostras.

Finalmente, fazemos o quociente $\vartheta = \sigma : \sigma i$ sendo o grau de liberdade do dividendo nf1 = a-1 e do divisor nf2 = $\infty$.

Sendo quase sempre o grau de liberdade do dividendo elevado, as tábus comuns do $\vartheta$ não se prestam bem e precisamos procurar outras, mais completas. Para os nossos exemplos utilizámos uma tábua de BRIEGER, ainda inédita.

Se na comparação do $X^2$ total ou do quociente $\vartheta$ resultar um valor insignificante, concluiremos que se trata de uma série de POISSON, ou que a variação original, ao redor da média m, é de acaso.

## APLICAÇAO PRATICA

**Problema** — Foi determinado no Calamondin e nas laranjas Natal e Côco o número de sementes por fruto e foram achadas as frequências dadas no quadro I. Desejamos saber se essas frequências por fruto variam apenas com o jôgo de acaso em volta do número médio calculado sôbre todos os frutos.

Parece-nos evidente que a variação de acaso, que nós devemos aqui aceitar como variação teòricamente esperada, deverá ser a da série de POISSON. Isto porque referindo-nos aos critérios anteriormente estabelecidos, verificamos que: 1) trata-se de uma variação descontínua, sendo presentes 0, 1, 2.... sementes por fruto, números êstes sempre inteiros e positivos; 2) as frequências sempre são baixas, de modo que devemos esperar o efeito do limite zero; 3) o número total de sementes possíveis, o qual corresponde ao número total de óvulos, inclusive os abortivos, é desconhecido, porém, pode ser aceito como muito grande em relação ao número de sementes presentes por fruto.

Podemos então proceder à análise estatística, supondo tratar-se de uma distribuição de POISSON.

## CALCULO DAS FREQUÊNCIAS ESPERADAS

**Laranja Natal** — (Logarítmos naturais) — Para a determinação das frequências esperadas pela série de POISSON, temos que calcular primeiramente algumas constantes e depois pela lei da formação da série facilmente obteremos todos os têrmos.

As constantes necessárias são:

$\log \text{nat}\ m = \log \text{nat}\ 3{,}53 = 1{,}26130$

$(\log \text{nat}\ n) - m = \log \text{nat}\ 150 - 3{,}53 = [\log \text{nat}\ 1{,}5 + (\log 10^2) - 3{,}53] = (0{,}40547 + 4{,}60517) - 3{,}53] = 1{,}48064$

Como já frizámos anteriormente, a organização de uma tábua facilita muito o cálculo dos têrmos da série e para isto organizámos o quadro II.

**Laranja Côco** — (Logarítmos decimais) — O cálculo das frequências esperadas pelos logarítmos decimais não oferece dificuldade e só temos que fazer algumas transformações logarítmicas, nas constantes a usar na tabela.

$\log m = \log 3{,}42 = 0{,}53403$

$(\log n) - m = \log n - \log 10^{e}.m = \log 144 - (0{,}43429\ .\ 3{,}42) = 2{,}5836 - 1{,}48527 = 0{,}67309$

Para facilitar o cálculo de tôdas as frequências esperadas, organizámos o quadro III.

**Calamondin** — (Logarítmos naturais). — As constantes que precisamos são as seguintes:

log nat m = log nat 3,64 = 1,29198
(log nat n) — m = log nat 157 — 3,64 = 1,41625

Sendo o modo de calcular as frequências esperadas para êste caso uma repetição do primeiro exemplo, limitar-nos-emos a dar as frequências já calculadas (quadro IV).

$X^2$ TESTE (QUADRO IV)

**Laranjas Natal e Côco** — A comparação de cada $x^2$ parcial com nf = 1 e do total com nf = 7, nos dois casos com médias respectivamente de 3,51 e 3,42 sementes por fruto, mostra-nos que todos os valores são insignificantes, isto é, abaixo de 5% limite de probabilidade. Disto concluimos que a frequência de cada classe observada não difere daquela esperada e que o conjunto também forma uma amostra da distribuição de POISSON.

**Calamondin** — Fazendo-se um $x^2$ teste entre a frequência observada e aquela esperada da distribuição de POISSON com média de 3,64 sementes por fruto, nota-se que o $x^2$ total está justamente no limite de 5% de probabilidade. Estudando-se a distribuição dos sinais, podemos constatar que êles não se distribuem ao acaso, mas que a classe do centro, que corresponde aos frutos de 2 a 5 sementes, tem uma frequência excessiva, enquanto que as classes extremas, com 0 a 1 semente ou com 1 a 10 sementes, mostram deficiências.

Ainda mais, o $x^2$ teste para a classe sem e **uma** semente, dá um valor bem perto do limite de 1% de probabilidade. Daí podemos concluir que se trata talvez de uma distribuição de POISSON, com um certo constrangimento da variação.

Outra prova do que acabamos de dizer acima, **que os sinais não se distribuem ao acaso**, é a seguinte: fazendo-se uma reunião das classes seguidas, o que aliás é justificado, e calculando-se finalmente os $x^2$ (última coluna do quadro IV), deveríamos esperar, se a distribuição dos sinais fôsse ao acaso, que a redução dos $x^2$ parciais, que em número caiu de 8 a 5, fôsse também acompanhada pela redução dos próprios valores, dando finalmente um $x^2$ menor ou pelo menos da mesma dimensão que o primitivo. Porém, tal não aconteceu e o $x^2$ total obtido, depois da reunião das classes, com um valor de 13,93 e 4 graus de liberdade é bem significante, fora do limite de 1% de probabilidade.

## $\vartheta$ TESTE (QUADRO IV)

**Laranjas Natal e Côco** — A comparação do êrro calculado diretamente das séries observadas com o da distribuição de POISSON com médias respectivamente de 3,51 e 3,42 para cada caso, dá um quociente $\vartheta$ insignificante. Daí concluimos que as duas séries em aprêço são uma variação ao acaso da distribuição de POISSON.

**Calamondin** — A comparação do êrro calculado diretamente com o $\sigma$ i da distribuição de POISSON, com média de 3,64 demonstra que o primeiro é significantemente menor que o segundo. Assim, fica mais uma vez acentuada a existência de uma redução da variação, já mencionada no capítulo anterior.

Nas laranjas Natal e Côco, onde o $x^2$ teste foi insignificante, também o $\vartheta$ teste nada mostrou de anormal. Todavia, já para o Calamondin, tomando os dois testes em conjunto, podemos dizer que a anormalidade da variação constatada é provocada pela ausência de frutos sem sementes; a falta de tais frutos talvez possa ser explicada por razões fisiológicas.

## RESULTADO E COMPARAÇÃO DO $X^2$ E DO $\vartheta$ TESTE

Comparando o $x^2$ teste com o $\vartheta$ teste, devemos esperar que de um modo geral os resultados coincidam. Assim aconteceu com os dois testes na análise das laranjas Natal e Côco, onde não houve nenhuma razão para supor que a variação do número de sementes por fruto não fôsse de acaso, segundo a série de POISSON.

De outro lado, para o Calamondin, o $\vartheta$ teste demonstrou que a variação evidentemente foi mais restrita do que a variação de uma série ideal de POISSON correspondente, não se podendo, porém, dizer em detalhe, no que consiste essa restrição. Neste ponto, o $x^2$ teste mostra-se mais eficiente, porque, como aliás já foi mencionado anteriormente, as frequências das classes com poucas e muitas sementes foram deficientes, enquanto que as classes de 4 e 5 sementes foram excessivas, sendo especialmente notável e estatisticamnte significante a falta relativa de frutos com zero e 1 semente. Deveríamos então tentar procurar a causa dêsses desvios da distribuição ideal de POISSON. Talvez haja, para o Calamondin, uma impossibilidade fisiológica de produzir frutos sem sementes, o que poderia explicar a ausência de tais frutos.

Para provar estatisticamente esta hipótese provisória, teremos que formular o problema do seguinte modo : é necessário fisiològicamente que os frutos tenham no mínimo uma semente e o que é variável é o número adicional de sementes por fruto. Consequentemente, a média m para a distribuição de POISSON torna-se igual a 2,64 sementes adicionais por fruto e o êrro ideal $\sigma$ i será igual a $\pm \sqrt{2,64} = \pm 1,62$ O $\vartheta$ entre o $\sigma$ calculado diretamente da frequência observada com o $\sigma$ i é de 1,48 : 1,62 = 0,91. Sendo êste quociente $\vartheta$ insignificante, poderemos dizer que agora encontrámos de fato uma variação de acaso, segundo a série de POISSON.

Poderemos ainda fazer o $x^2$ teste para comprovar melhor o que acima dissemos, e como podemos ver no quadro V, tanto os $x^2$ parciais como o $x^2$ total foram insignificantes, mostrando mais uma vez que o nosso raciocínio sôbre a necessidade de haver pelo menos uma semente nos frutos de Calamondin, foi verdadeiro.

Em geral, o $\vartheta$ teste é mais fácil de calcular, não necessita de usar logarítmos e prova de um modo sumario se temos ou não uma distribuição de acaso, segundo a série de POISSON. Se desejarmos um teste mais exato e detalhado, temos então de recorrer ao $x^2$ teste, e cremos que neste caso, o trabalho adicional de cálculo será largamente compensado pelos resultados.

## ABSTRACT

The general properties of POISSON distributions and their relations to the binomial distribuitions are discussed. Two methods of statistical analysis are dealt with in detail :

$X^2$-test. In order to carry out the $X^2$-test, the mean frequency and the theoretical frequencies for all classes are calculated. Than the observed and the calculated frequencies are compared, using the well nown formula: f(obs) — f(esp) 2: f(esp). When the expected frequencies are small, one must not forget that the value of $X^2$ may only be calculated, if the expected frequencies are biger than 5. If smaller values should occur, the frequencies of neighboroughing classes must ge pooled.

As a second test reintroduced by BRIEGER, consists in comparing the observed and expected error standard of the series. The observed error is calculated by the general formula:

$$\sigma = \pm \sqrt{\frac{\Sigma f.Vk}{n-1}}$$

where n represents the number of cases.

The theoretical error of a POISSON series with mean frequency m is always $\pm \sqrt{m}$.

These two values may be compared either by dividing the observed by the theoretical error and using BRIEGER's tables for $\vartheta$ or by dividing the respective variances and using SNEDECOR's tables for F. The degree of freedom for the observed error is one less the number of cases studied, and that of the theoretical error is always infinite.

In carrying out these tests, one important point must never be overlloked. The values for the first class, even if no concrete cases of the type were observed, must always be zero, an dthe value of the subsequent classes must be 1, 2, 3, etc..

This is easily seen in some of the classical experiments. For instance in BORKEWITZ example of accidents in prussian armee corps, the classes are: no, one, two, etc., accidents. When counting the frequency of bacteria, these values are: no, one, two, etc., bacteria or cultures of bacteria. Ins studies of plant diseases equally the frequencies are : no, one, two, etc., plants deseased.

Howewer more complicated cases may occur. For instance, when analising the degree of polyembriony, frequently the case of "no polyembryony" corresponds to the occurrence of one embryo per each seed. Thus the classes are not: no, one, etc., embryo per seed, but they are: no additional embryo, one additional embryo, etc., per seed with at least one embryo.

Another interestin case was found by BRIEGER in genetic studies on the number os rows in maize. Here the minimum number is of course not : no rows, but : no additional beyond eight rows. The next class is not: nine rows, but: 10 rows, since the row number varies always in pairs of rows. Thus the value of successive classes are: no additional pair of rows beyond 8, one additional pair (or 10 rows), two additional pairs (or 12 rows) etc..

The application of the methods is finally shown on the hand of three examples : the number of seeds per fruit in the oranges "Natal" and "Côco" and in "Calamondin". As shown in the text and the tables, the agreement with a POISSON series is very satisfactory in the first two cases. In the third case BRIEGER's error test indicated a significant reduction of variability, and the $X^2$ test showed that there were two many fruits with 4 or 5 seeds and too few with more or with less seeds. Howewer the fact that no fruit was found without seed, may be taken to indicate that in Calamondin fruits are not fully parthenocarpic and may develop only with one seed at the least. Thus a new analysis was carried out, on another class basis. As value for the first class the following value was accepted: no additional seed beyond the indispensable minimum number of one seed, and for the later classes the values were: one, two, etc., additional seeds. Using this new basis for all calculations, a complete agreement of the observed and expected frequencies, of the correspondig POISSON series was obtained, thus proving that our hypothesis of the impossibility of obtaining fruits without any seed was correct for Calamondin while the other two oranges were completely parthenocarpic and fruits without seeds did occur.

## BIBLIOGRAFIA

1 — BRIEGER, F. G., A. RODRIGUES LIMA e R. FORSTER — Comportamento de variedades e progênies de fumo na resistência ao "vira-cabeça". Bragantia 2: 275-294.— 1942.

2 — FISHER, Arne — Frequency curves. 1.a edition — The Macmillan Company — New York — 1922.

3 — FISHER, Arne — The Mathematical Theory of Probabilities — 1.a edition — The Macmillan Company — New York — 1915.

4 — FISHER, R. A. — Statistical Methods of Research Workers — 5.a edition — Oliver and Boyd — London — 1934.

5 — MOREIRA, S. e J. T. A. GURGEL — A fertilidade do pólen e sua correlação com o número de sementes, em espécies e formas do gênero Citrus — Bragantia 1: 669-712 — 1941.

6 — MOREIRA, S., e J. T. A. GURGEL — Poliembrionia em Citrus — Bragantia (em impressão) — 1945.

7 — RIDER, P. R. — An Introduction to Modern Statistical Methods — 1.a edition — John Wiley and Sons, Inc. — London — 1939.

8 — SNEDECOR, G. W. — Statistical Methods — Collegiate Press, Inc., Ames — Iowa — 1938.

9 — TIPPET, L. H. C. — The Methods of Statistics — 2.a edition — Williams and Norgate — London — 1939.

10 — TRELOAR, A. E. — Elements of Statistical Reasoning — 1.a edition — John Wiley and Sons, Inc. — London — 1939.

11 — YULE, E. U. and M. E. KENDALL — An Introduction to the Theory of Statistics — 11.a edition — Charles Griffin and Company — London — 1937.

QUADRO I

**Número de sementes por fruto**

| Número de sementes por fruto | Frequência observada | | |
|---|---|---|---|
| | Laranja Natal | Laranja Côco | Calamondin |
| 0 | 4 | 7 | 0 |
| 1 | 16 | 15 | 8 |
| 2 | 30 | 20 | 29 |
| 3 | 26 | 36 | 40 |
| 4 | 31 | 30 | 38 |
| 5 | 24 | 20 | 26 |
| 6 | 9 | 9 | 10 |
| 7 | 7 | 2 | 4 |
| 8 | 1 | 3 | 2 |
| 9 | 2 | 2 | — |
| **Total n** | 150 | 144 | 157 |
| **Média m** | 3,51 | 3,42 | 3,64 |

## QUADRO II

**Laranja Natal**

Logarítmos naturais

| I | II | III | IV | V | VI | VII | VIII |
|---|---|---|---|---|---|---|---|
| Centro das classes Vk | Log. nat. Fatorial n! | Log. nat. m | (Log. nat n) — m | -II+III+IV | | Logarítmo de f. esp. | Antilogarítno de f. esp. |
| 0 | — | — | 1,48064 | 1,48064 | ——— | 1,48064 | 4,40 |
| 1 | — | 1,26130 | 1,48064 | 2,74194 | — 2,30258 | 0,43936 | 15,52 |
| 2 | 0,69315 | 2,52260 | 1,48064 | 3,31009 | — 2,30258 | 1,00751 | 27,39 |
| 3 | 1,79176 | 3,78390 | 1,48064 | 3,47278 | — 2,30258 | 1,17020 | 32,22 |
| 4 | 3,17805 | 5,04520 | 1,48064 | 3,34779 | — 2,30258 | 1,04521 | 28,44 |
| 5 | 4,78749 | 6,20650 | 1,48064 | 2,99965 | — 2,30258 | 0,69707 | 20,08 |
| 6 | 6,57925 | 7,56780 | 1,48064 | 2,46919 | — 2,30258 | 0,16661 | 11,81 |
| 7 | 8,52516 | 8,82910 | 1,48064 | 1,78458 | ——— | 1,78458 | 5.96 |
| 8 | 10,60460 | 10,09040 | 1,48064 | 0,96644 | ——— | 0,96644 | 2,63 |
| 9 | 12,80182 | 11,35170 | 1,48064 | 0,03052 | ——— | 0,03052 | 1,03 |
| 10 | 15,10440 | 12,61300 | 1,48064 | — 1,01076 | 2,30258 | 1,29182 | 0,36 |
| 11 | 17,45229 | 13.87430 | 1,48064 | — 2,09735 | 2,30258 | 0,20523 | 0,12 |
| 12 | 19,93719 | 15.13560 | 1,48064 | — 3,32095 | 4,60517 | 1,28422 | 0,04 |

**Coluna VI** — Uma vez que a tábua usada continha apenas os log. nat. de 0 até 2,30258, corespondendo aos antilogarítmos de 1 a 100, foi necessário fazer uma correção nos log. calculados, subtraindo-se ou somando-se ao log. da coluna V os log. das potências de 10, dado na coluna VI; correspondentemente, deve-se multiplicar ou dividir os antilog. tirados da tábua com as respectivas potências de 10, para se obter os valores da coluna VIII.

## QUADRO III

**Laranja Côco**

Logarítmos decimais

| I | II | III | IV | V | VI | VII |
|---|---|---|---|---|---|---|
| Centro das classes Vk | Log. Fatorial n! | Log. m | (Log. n.) — m | −II+III+IV | Logarítmo de f. esp. | Antilogarítmo de f. esp. |
| 0 | — | — | 0,67309 | 0,67309 | 0,67309 | 4,71 |
| 1 | — | 0,53403 | 0,67309 | 1,20712 | 1,20712 | 16,11 |
| 2 | 0,30103 | 1,06806 | 0,67309 | 1,44012 | 1,44012 | 27,55 |
| 3 | 0,77815 | 1,60209 | 0,67309 | 1,49703 | 1,49703 | 31,41 |
| 4 | 1,38021 | 2,13612 | 0,67309 | 1,42900 | 1,42900 | 26,85 |
| 5 | 2,07918 | 2,67015 | 0,67309 | 1,26406 | 1,26406 | 18,37 |
| 6 | 2,85733 | 3,20418 | 0,67309 | 1,01994 | 1,01994 | 10,47 |
| 7 | 3,70243 | 3,73821 | 0,67309 | 0,70887 | 0,70887 | 5,12 |
| 8 | 4,60552 | 4,27224 | 0,67309 | 0,33981 | 0,33981 | 2,19 |
| 9 | 5,55976 | 4,80627 | 0,67309 | — 0,08040 | $\bar{1}$,91960 | 0,83 |
| 10 | 6,55976 | 5,34030 | 0,67309 | — 0,54637 | $\bar{1}$,45363 | 0,28 |
| 11 | 7,60115 | 5,87433 | 0,67309 | — 1,05373 | $\bar{2}$,94627 | 0,09 |
| 12 | 8,68033 | 6,4036 | 0,67309 | — 1,59888 | $\bar{2}$,40112 | 0,03 |

QUADRO IV — **Número de sementes por fruto**

| Número de sementes por fruto | LARANJA NATAL Frequência obs. | LARANJA NATAL Frequência esp. | LARANJA NATAL $X^2$ | LARANJA CÔCO Frequência obs. | LARANJA CÔCO Frequência esp. | LARANJA CÔCO $X^2$ | CALAMONDIN Frequência obs. | CALAMONDIN Frequência esp. | CALAMONDIN $X^2$ | CALAMONDIN $X^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 4 | 4,40 | 0,04 | 7 | 4,71 | 1,11 | 0 } 8 | 4,12 } 19,12 | (6,47) | 6,47 |
| 1 | 16 | 15,52 | 0,01 | 15 | 16,11 | 0,08 | 8 | 15,00 | | |
| 2 | 30 | 27,39 | 0,25 | 20 | 27,55 | 2,07 | 29 | 27,30 | 0,06 | 1,22 |
| 3 | 26 | 32,23 | 1,20 | 36 | 31,41 | 0,67 | 40 | 33,10 | 1,44 | |
| 4 | 31 | 28,44 | 0,23 | 30 | 28,85 | 0,05 | 38 | 30,10 | 2,07 | 2,87 |
| 5 | 24 | 20,08 | 0,77 | 20 | 18,37 | 0,14 | 26 | 21,70 | 0,85 | |
| 6 | 9 | 11,81 | 0,67 | 9 | 10,47 | 0,21 | 10 | 13,30 | 0,82 | 1,91 |
| 7 | 7 | 5,96 | | 2 | 5,12 | | 4 | 6,[illegible]2 | 1,23 | |
| 8 | 1 | 2,63 | | 3 | 2,19 | | 2 | 3,15 | | 1,46 |
| 9 | 2 } 10 | 1,03 } 10,14 | 0,002 | 2 } 7 | 0,83 } 8,54 | 0,28 | — } 2 | 1,28 } 5,07 | 1,46 | |
| 10 | — | 0,36 | | — | 0,28 | | — | 0,46 | | |
| 11 | — | 0,12 | | — | 0,09 | | — | 0,18 | | |
| 12 | — | 0,04 | | — | 0,03 | | | | | |
| Soma | 150 | 150,01 | 3,17 nf($x^2$) = 7 | 144 | 144,01 | 5,34 nf($x^2$) = 7 | 157 | 156,61 | 14,40 nf($x^2$) = 7 | 13,93 nf($x^2$) = 4 |
| Média **m** | 3,51 | | | 3,42 | | | 3,64 | | | |
| Êrro «standard» $\sigma$ | ± 2,12 | | | ± 1,85 | | | ± 1,48 | | | |
| Êrro ideal $\sigma i$ | ± 1,87 | | | ± 1,85 | | | ± 1,91 | | | |
| $\vartheta = \frac{\sigma}{\sigma i}$ | 1,13 | $\frac{nf1 = 149}{nf^2 = \infty}$ | | 1,00 | $\frac{nf1 = 143}{nf^2 = \infty}$ | | **0,77** | $\frac{nf1 = 156}{nf^2 = \infty}$ | | |

## QUADRO V

| Número de sementes por fruto | CALAMONDIN | | |
|---|---|---|---|
| | Frequência obs. | Frequência esp. | $\chi^2$ |
| 0 | 8 | 11,20 | 0,91 |
| 1 | 29 | 29,58 | 0,01 |
| 2 | 40 | 39,04 | 0,02 |
| 3 | 38 | 34,36 | 0,39 |
| 4 | 26 | 22,68 | 0,49 |
| 5 | 10 | 11,98 | 0,33 |
| 6 | 4) | 5,27) | |
| 7 | 2) | 1,99) | |
| 8 | —) 6 | 0,66) 8,17 | 0,58 |
| 9 | —) | 0,19) | |
| 10 | —) | 0,05) | |
| 11 | —) | 0,01) | |
| Soma | 157 | 157,01 | 2,73 nf $(\chi^2)$ = 6 |
| Média m | 2,64 | | |
| Êrro "standard" $\sigma$ | $\pm$ 1,48 | | |
| Êrro ideal $\sigma$i | $\pm$ 1,62 | | |
| $\vartheta = \frac{\sigma}{\sigma i}$ | 0,91 | | nfl = 156 / nf2 = $\infty$ |