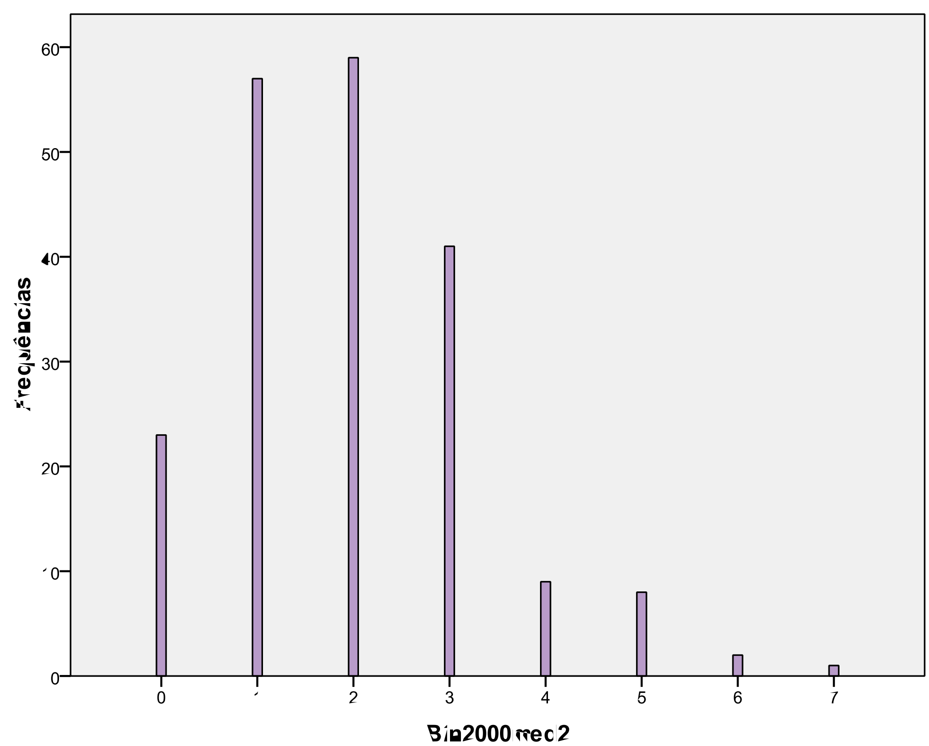
Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Departamento de Matemática

Escola Delfos

Fenómenos raros, pequenas amostras e lei de Poisson

Maria de Nazaré Mendes Lopes



Setembro 2009

Resumo

A ocorrência de fenómenos (aleatórios) raros será discutida, e avaliada probabilisticamente, através da introdução da distribuição de Poisson. Esta distribuição surgirá como limite da lei binomial que, como é já do conhecimento dos estudantes, descreve o número de ocorrências de um acontecimento em n (fixo a priori) realizações idênticas e independentes de uma experiência ou fenómeno aleatório. Este estudo permitirá introduzir de forma intuitiva, mas rigorosa, uma probabilidade discreta de suporte não finito.

Algumas propriedades da distribuição de Poisson serão analisadas, sendo dado particular relevo, pelo seu interesse prático, à propriedade de estabilidade para somas.

A distribuição de Poisson, dita também dos fenómenos raros, será então usada para ilustrar a viabilidade das conclusões estatísticas em amostras de baixa dimensão. Este estudo será acompanhado e motivado por uma situação prática, ligada às Ciências da Saúde, onde a recolha de grandes amostras não é, em geral, possível.

Sumário

1 Da lei Binomial à lei de Poisson

1.1 Lei binomial

Seja $\mathcal{E}$ uma experiência aleatória e A um acontecimento associado a $\mathcal{E}$ tal que $P(A) = p$ ($p \in]0,1[$).

Realiza-se a experiência, $\mathcal{E}$, n vezes sempre nas mesmas condições e tal que as realizações são independentes umas das outras.

Considere-se uma variável aleatória, X, tal que

> X descreve **"o número de vezes que o acontecimento A ocorre nas n realizações da experiência"**.

Nestas condições, a variável X toma os valores $0, 1, 2, ..., n$ com probabilidades:

$$P(\{X = k\}) = \binom{n}{k} p^k (1-p)^{n-k}, \ k \in \{0, 1, ..., n\}.$$

<u>*Prova*</u>

Uma realização possível do acontecimento $\{X = k\}$ é a ocorrência da intersecção

$$\textbf{(1.1)} \qquad A_1 \cap A_2 \cap ... \cap A_k \cap B_1 \cap ... \cap B_{n-k}$$

onde A_i, $i = 1, ..., k$, representa a ocorrência de A na realização i, e B_j, $j = 1, ..., n-k$, representa a ocorrência de $\overline{A}$ na realização $k + j$.

As n realizações da experiência ocorrem sempre nas mesmas condições. Assim, a probabilidade de qualquer acontecimento A_i é $P(A)$ e de qualquer acontecimento B_j é $P(\overline{A})$. Além disso, tratando-se o acontecimento em **(1.1)** de uma intersecção de n acontecimentos independentes, obter-se-á para a sua probabilidade

$$\underbrace{P(A)...P(A)}_{k \ vezes}\underbrace{P(\overline{A})...P(\overline{A})}_{n-k \ \ vezes} = p^k (1-p)^{n-k}$$

Por outro lado, o acontecimento $\{X = k\}$ pode escrever-se como a reunião de todos os acontecimentos que se obtêm daquele considerando todas as trocas possíveis dos A_i com os

B_j. O número de trocas é $\dfrac{n!}{k!(n-k)!} = \left(\begin{array}{c} n \\ k \end{array}\right)$ (número de permutações de n elementos em que k são de um tipo e $(n-k)$ de outro tipo).

Como os acontecimentos resultantes são dois a dois incompatíveis e equiprováveis obtém-se

$$(\mathbf{1.2}) \qquad P\left(\{X = k\}\right) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k \left(1-p\right)^{n-k}, \ k = 0, 1, .., n. \ \blacksquare$$

Definição 1. Uma variável aleatória X verificando **(1.2)** diz-se *binomial*, ou que segue a *lei (distribuição) binomial* de parâmetros n e p. Escreve-se, de modo abreviado,

$$X \sim B\left(n, p\right).$$

Exercícios de aplicação

1. Prove que $S = \{0, 1, ..., n\}$ é o suporte da variável X, isto é, verifica as duas condições seguintes:

i) $\forall k \in S, \ P\left(\{X = k\}\right) > 0$;
ii) $P\left(X \in S\right) = 1$.

2. Sabe-se que com certo tratamento se alcançam 90% de curas para determinada doença quando o mesmo é aplicado a pacientes em condições bem definidas. Supondo que o tratamento é aplicado a 20 pacientes nessas condições, calcule

a) a probabilidade de obter pelo menos 5 curas.

b) o número máximo de curas que, com uma probabilidade de cerca de 98%, podem ocorrer naquele conjunto de doentes.

1.2 Lei de Poisson

1.2.1 Construção da lei de Poisson

Considere-se a situação aleatória descrita na definição da lei binomial e suponha-se que

- n é muito grande (isto é, a experiência é realizada um número muito grande de vezes);
- p é muito pequeno (isto é, o fenómeno é raro);
- o produto np tem uma ordem de grandeza razoável (isto é, nem n domina p, nem p domina n).

As condições anteriores são matemática e rigorosamente verificadas considerando uma sucessão $\left(X_n\right)_{n \in \mathbb{N}}$ de variáveis aleatórias binomiais tal que

$$X_n \sim B\left(n, p\left(n\right)\right)$$

com

$$\lim_{n\to+\infty} np\,(n) = \lambda, \quad \lambda \in \left]0, + + \infty\right[.$$

É óbvio que, nestas condições, $p\,(n)$ converge necessariamente para zero.

Tem-se (verificação ao cuidado do aluno)

$$\forall k \in \mathbb{N}, \forall n \geq k,\ \frac{P\,(X_n = k)}{P\,(X_n = k-1)} = \frac{n-k+1}{k}\frac{p\,(n)}{1-p\,(n)}.$$

Assim,

$$\forall k \in \mathbb{N},\ \lim_{n\to+\infty} \frac{P\,(X_n = k)}{P\,(X_n = k-1)} = \frac{\lambda}{k}.$$

Esta igualdade admite a seguinte interpretação para n suficientemente elevado:

X_n tem quase o mesmo comportamento que uma variável aleatória Y com suporte $\mathbb{N}_0$ e tal que

$$\forall k \in \mathbb{N},\ P\,(Y = k) = \frac{\lambda}{k} P\,(Y = k-1)\,.$$

Se existir uma tal probabilidade de suporte $\mathbb{N}_0$, ter-se-á (verificação ao cuidado do aluno)

$$P\,(Y = k) = \frac{\lambda^k}{k!} P\,(Y = 0)\,, \forall k \in \mathbb{N};$$

e([1])

$$\lim_{N\to+\infty} \sum_{k=0}^{N} P\,(Y = k) = \sum_{k=0}^{++\infty} P\,(Y = k) = e^{\lambda} P\,(Y = 0) = 1\ ,$$

isto é,

$$P\,(Y = 0) = e^{-\lambda}.\ \blacksquare$$

Definição 2. Uma variável aleatória X é *Poissoniana*, ou segue a *lei (distribuição) de Poisson de parâmetro* λ, se tem por suporte $\mathbb{N}_0$ e

$$P\,(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k \in \mathbb{N}_0.$$

Escreve-se, de modo abreviado,

$$X \sim P(\lambda).$$

[1] Tenha em conta o desenvolvimento da função exponencial em série de potências (aproximação da função exponencial por uma função polinomial):
$e^x = \lim\limits_{N\to\infty} \sum\limits_{k=0}^{N} \frac{x^k}{k!} = \sum\limits_{k=0}^{+\infty} \frac{x^k}{k!}.$

Propriedade 1. Se uma variável aleatória X é distribuída de acorddo com a lei de Poisson de parâmetro λ, então a sua média (valor médio, valor esperado ou esperança matemática) é λ.

<u>*Prova*</u>
Designando por $E(X)$ a média de X, vem

$$\begin{aligned} E(X) &= \lim_{N\to+\infty}\sum_{k=0}^{N} k\frac{\lambda^k}{k!}e^{-\lambda} = \lim_{N\to+\infty}\lambda e^{-\lambda}\sum_{k=1}^{N}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda}\lim_{N\to+\infty}\sum_{j=0}^{N-1}\frac{\lambda^j}{j!} = \lambda e^{-\lambda}e^{\lambda} = \lambda. \ \blacksquare \end{aligned}$$

1.2.2 Ilustração deste resultado com o recurso à geração de números aleatórios de acordo com uma lei binomial e de acordo com uma lei de Poisson

A/ Foi gerada uma amostra de dimensão 200 de uma variável aleatória, Bin2000med2, seguindo a lei binomial B(2000, 0.001)([2]).

Análise descritiva da amostra

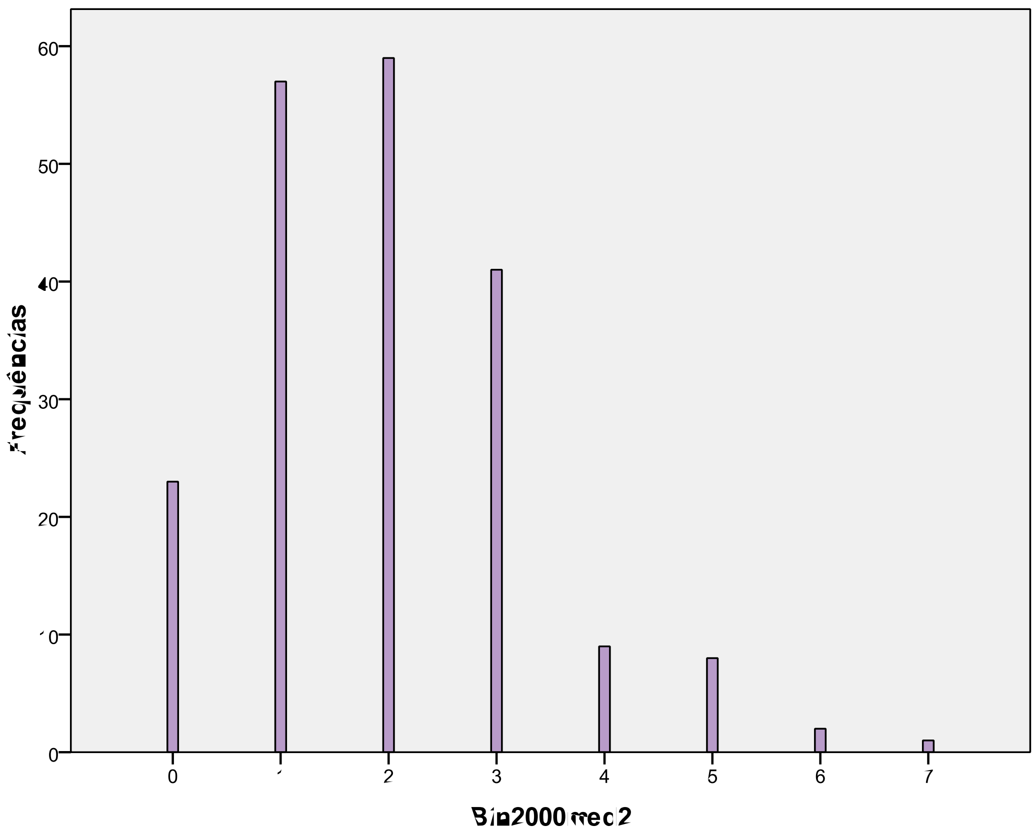
Os **resumos descritivos** básicos associados à amostra gerada estão resumidos no quadro seguinte:

Descriptives

			Statistic
Bin2000med2	Mean		1,96
	95% Confidence Interval for Mean	Lower Bound	1,78
		Upper Bound	2,15
	Median		2,00
	Variance		1,783
	Std. Deviation		1,335
	Minimum		0
	Maximum		7

[2] Utilização do software estatístico SPSS.

O gráfico abaixo representa o **diagrama de barras** associado à amostra considerada.



Comparação com a distribuição de Poisson de parâmetro igual a 2 - teste de Kolmogorov-Smirnov

Testou-se, relativamente à sua alternativa, a hipótese

$$\mathbf{H_0}\textbf{: A amostra é Poissoniana.}$$

One-Sample Kolmogorov-Smirnov Test

N		200
Poisson Parameter[a]	Mean	1,96
Kolmogorov-Smirnov Z		,518
Asymp. Sig. (2-tailed)		,951
		Bin2000med2

a. Test distribution is Poisson.

Sendo a *significância* ou *p-valor*[3] do teste igual a 0.951 dever-se-á aceitar H_0. Poder-se-á

[3] Chama-se *p-valor* ou *significância* de um teste de uma hipótese H_0 contra a sua alternativa H_1 (contrário de H_0) ao menor valor da probabilidade do erro que se comete ao rejeitar H_0 sendo esta verdadeira.

assim afirmar que a variável Bin2000med2 segue aproximadamente $P(2)$, isto é,

$$\textbf{Bin2000med2} \overset{\bullet}{\sim} P(2).$$

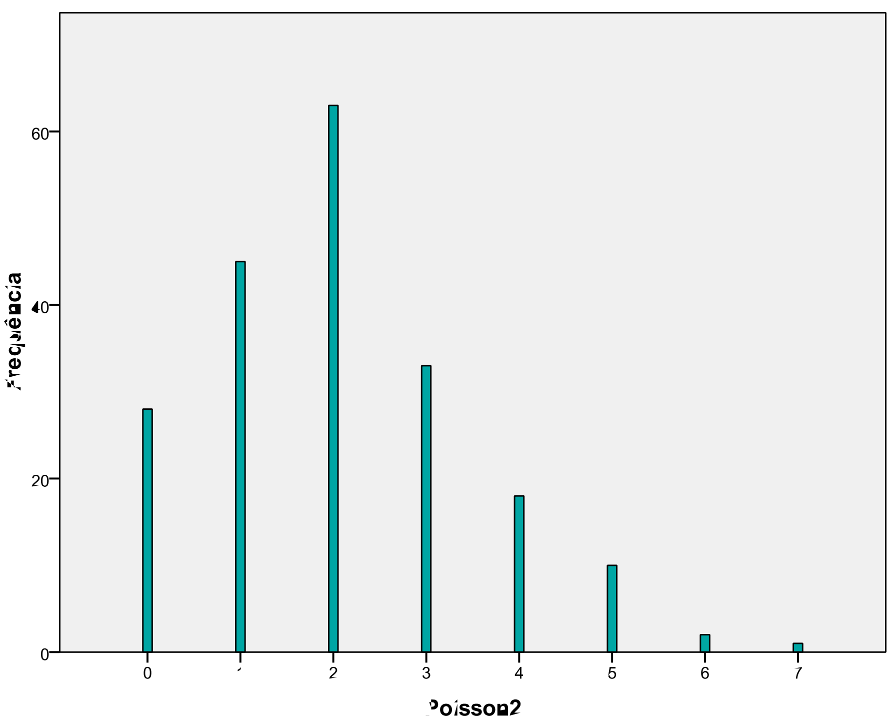
B/ Foi também gerada uma amostra de dimensão 200 de uma variável aleatória, Poisson2, seguindo a lei de Poisson $P(2)$.

Os **resumos descritivos** básicos associados à amostra gerada estão resumidos no quadro seguinte:

Descriptives

			Statistic
Poisson2	Mean		2,06
	95% Confidence Interval for Mean	Lower Bound	1,86
		Upper Bound	2,25
	Median		2,00
	Variance		2,052
	Std. Deviation		1,433
	Minimum		0
	Maximum		7

Apresenta-se de seguida o **diagrama de barras** associado à amostra considerada.



Teste de Kolmogorov-Smirnov relativamente à distribuição teórica de Poisson

Testando a hipótese H_0 anteriormente considerada obteve-se agora o p-valor 0.999.

One-Sample Kolmogorov-Smirnov Test 2

Most Extreme Differences		
N		200
Poisson Parameter[a]	Mean	2,06
Kolmogorov-Smirnov Z		,372
Asymp. Sig. (2-tailed)		,999
		Poisson2

a. Test distribution is Poisson.

Exercício de aplicação

Determinado estudo estatístico permitiu concluir que a probabilidade de, em períodos de fortes chuvas, um veículo entrar em *aquaplaning* numa determinada zona da auto-estrada é de 2×10^{-5}. Suponha que numa semana de fortes chuvas se regista a passagem nessa zona de 150000 veículos. Defina a variável aleatória real (v.a.r.) X: "número de veículos que nessa semana entram em *aquaplaning* naquela zona".

a) Calcule a probabilidade de, nessa semana, pelo menos um veículo entrar em *aquaplaning* naquela zona.

b) Determine m de modo que a probabilidade de que ocorram, durante essa semana, mais do que m situações de *aquaplaning* naquela zona seja no máximo de 5%.

1.3 Estabilidade para somas da lei de Poisson

1.3.1 Propriedade teórica

Sejam X e Y variáveis aleatórias independentes, isto é, os acontecimentos que tais variáveis definem são entre si independentes. Além disso, são variáveis Poissonianas tais que

$$X \sim P(\lambda) \text{ e } Y \sim P(\mu),\ \ \lambda > 0 \text{ e } \mu > 0.$$

Propriedade 2. Sob as hipóteses anteriores a variável aleatória $Z = X + Y$ é Poissoniana, tendo-se que

$$Z \sim P(\lambda + \mu).$$

Prova

De facto, o suporte da variável aleatória Z é $\mathbb{N}_0$. Além disso,

$$
\begin{aligned}
P(Z &= k) = \sum_{i=0}^{k} P(X=i, Y=k-i) \\
&= \sum_{i=0}^{k} P(X=i)P(Y=k-i) = \sum_{i=0}^{k} \frac{\lambda^i}{i!} e^{-\lambda} \frac{\mu^{k-i}}{(k-i)!} e^{-\mu} \\
&= \frac{e^{-(\lambda+\mu)}}{k!} \sum_{i=0}^{k} \frac{k!}{i!\,(k-i)!} \lambda^i \mu^{k-i} = \frac{e^{-(\lambda+\mu)}}{k!} (\lambda+\mu)^k,
\end{aligned}
$$

isto é,

$$
P(Z=k) = \frac{e^{-(\lambda+\mu)}}{k!} (\lambda+\mu)^k, \; k \in \mathbb{N}_0. \; \blacksquare
$$

Mais geralmente tem-se a propriedade seguinte:

Propriedade 3. Sejam $X_1, X_2, ..., X_n$ n variáveis aleatórias independentes tais que X_j segue a lei de Poisson de parâmetro $\lambda_j, \; j = 1, 2, ..., n$. A variável aleatória

$$
Y = \sum_{j=1}^{n} X_j
$$

segue a lei de Poisson de parâmetro $\sum_{j=1}^{n} \lambda_j$.

1.3.2 Ilustração deste resultado com o recurso à geração de números aleatórios de acordo com leis de Poisson

Foi considerada uma amostra de dimensão 200 de uma variável aleatória resultante da soma de amostras da mesma dimensão de variáveis aleatórias seguindo as leis de Poisson $P(1), \; P(2)$ e $P(2.5)$ ([4]). A variável resultante chamou-se SomaPoisson.

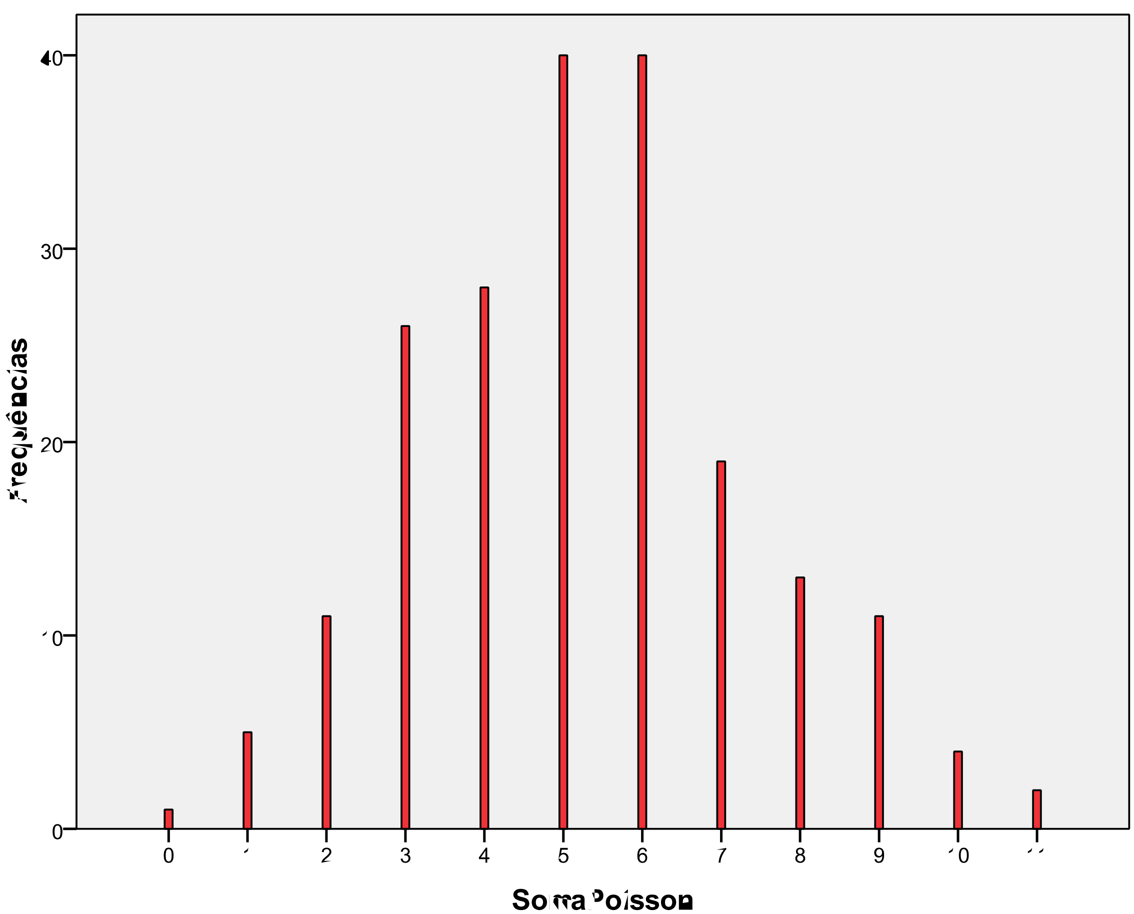
Os **resumos descritivos** básicos associados à amostra da variável SomaPoisson estão resumidos no quadro seguinte:

[4] Utilização do software estatístico SPSS.

Descriptives

			Statistic
SomaPoisson	Mean		5,28
	95% Confidence Interval for Mean	Lower Bound	4,98
		Upper Bound	5,57
	Median		5,00
	Variance		4,532
	Std. Deviation		2,129
	Minimum		0
	Maximum		11

O gráfico abaixo é o **diagrama de barras** associado a tal amostra.



Comparação entre a distribuição da variável SomaPoisson e a distribuição de Poisson - teste de Kolmogorov- Smirnov

Testou-se a hipótese

H_0: A variável SomaPoisson é Poissoniana.

One-Sample Kolmogorov-Smirnov Test

N			200
Poisson Parameter[a]	Mean		5,28
Kolmogorov-Smirnov Z		,546	
Asymp. Sig. (2-tailed)		,926	
		SomaPoisson	

a. Test distribution is Poisson.

Sendo o *p-valor* obtido 0.926, aceita-se a hipótese em teste.

Exercício de aplicação

Suponha que X e Y são v.a.r. descrevendo, respectivamente, o número de veículos a gasolina e a gasóleo que chegam, num determinado período de tempo, a uma estação de serviço para abastecer de combustível. Sabe-se que X e Y seguem leis de Poisson de parâmetros, respectivamente, iguais a 10 e a 6 e que as chegadas dos veículos são independentes entre si.

a) Determine a probabilidade de que nesse período de tempo cheguem pelo menos 5 veículos para abastecer de gasóleo e no máximo 8 para abastecer de gasolina.

b) Determine o número mínimo de veículos que, com probabilidade de 95%, chegam à respectiva estação de serviço, durante aquele período, para abastecer.

2 Aplicação da lei de Poisson a ensaios estatísticos

2.1 Exemplo de motivação

Determinado estudo estatístico permitiu concluir que o número de doentes com insuficiência renal que chegam de urgência, diariamente, ao serviço de hemodiálise de um certo hospital é bem descrito por uma lei de Poisson de parâmetro λ ($\lambda > 0$, desconhecido).

O equipamento técnico existente no referido hospital foi adquirido supondo que a unidade de hemodiálise atende, de urgência, em média um doente por dia. O médico responsável pela unidade afirma que o equipamento existente não é suficiente pois, segundo ele, tal média já terá aumentado.

Com o objectivo de averiguar a veracidade da afirmação feita pelo médico, observou-se durante 15 dias o número de doentes que chegaram àquela unidade. Os resultados obtidos estão resumidos no seguinte quadro:

n^o de doentes	0	1	2	3	4	5	>5
n^o de dias	1	4	6	2	1	1	0

Com base nesta amostra pretende-se concluir se a referida unidade deve adquirir mais equipamento, isto é, se o número médio de doentes que são atendidos de urgência na referida unidade de hemodiálise é agora superior um.

Uma forma de analisar este problema é, com base nesta amostra, testar se o valor médio (desconhecido) da população é 1 ou maior do que 1, isto é, testar as hipóteses

$$\mathbf{H}_0\text{: } \boldsymbol{\lambda} = \mathbf{1} \;\; \textbf{contra} \;\; \mathbf{H}_1\text{: } \boldsymbol{\lambda} > \mathbf{1}.$$

2.2 Teste para a média de uma população de Poisson

Seja $(X_1, X_2, ..., X_n)$ uma amostra aleatória de dimensão n de uma variável aleatória X, que mede uma certa característica de uma população, distribuída de acordo com uma lei de Poisson de parâmetro λ ($\lambda > 0$, desconhecido).

Pretende-se, com base numa qualquer amostra observada, testar

$$\mathbf{H}_0\text{: } \boldsymbol{\lambda} = \boldsymbol{\lambda}_0 \;\; \textbf{contra} \;\; \mathbf{H}_1\text{: } \boldsymbol{\lambda} > \boldsymbol{\lambda}_0, \quad \boldsymbol{\lambda}_0 \text{ fixo},$$

isto é, decidir se a amostra nos leva, ou não, à rejeição de $\mathbf{H}_0$.

Sabe-se que a média da amostra,

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

é um bom estimador da média da população λ.

Assim, se $\mathbf{H}_0$ é verdadeira, a média de uma qualquer amostra (representativa da população) deverá ser próxima de λ_0. Então é natural que se rejeite $\mathbf{H}_0$ quando a amostra pertence a um conjunto, $\mathcal{C}$, de amostras de dimensão n, $(x_1, x_2, ..., x_n)$, cuja média, $\overline{x}_n$, é tal que

$$\overline{x}_n - \lambda_0 > k, \;\; k > 0 \text{ a determinar.}$$

A determinação de k envolve o erro que se comete na decisão de rejeitar erradamente H_0, isto é, rejeitar H_0 sendo H_0 verdadeira.

A probabilidade de tal erro será

$$P\left(\left\{\overline{X}_n - \lambda_0 > k\right\}/\lambda = \lambda_0\right).$$

Ora, tendo em conta a estabilidade da lei de Poisson, tal probabilidade é igual a

$$\begin{aligned}
P\left(\left\{\overline{X}_n - \lambda_0 > k\right\}/\lambda = \lambda_0\right) &= P\left(\left\{\sum_{i=1}^{n} X_i - n\lambda_0 > nk\right\}/\lambda = \lambda_0\right) \\
&= P\left(\left\{\sum_{i=1}^{n} X_i > n\left(\lambda_0 + k\right)\right\}/\lambda = \lambda_0\right) \\
&= P\left(n\lambda_0\right)\left(\left]n\left(\lambda_0 + k\right), + + \infty\right[\right),
\end{aligned}$$

pois, sob $\mathbf{H}_0$, $\sum\limits_{i=1}^{n} X_i \backsim P\left(n\lambda_0\right).$

Escolhendo o valor desta probabilidade de erro (que se admite cometer ao tomar a decisão de rejeitar $\mathbf{H}_0$), determina-se k.

A este valor chama-se **nível de significância** do teste.

Observando uma amostra particular $(x_1, x_2, ..., x_n)$, rejeitar-se-á $\mathbf{H}_0$ se $n\left(\lambda_0 + k\right) < \sum\limits_{i=1}^{n} x_i$.

2.3 Aplicação prática

Na prática pretende-se tomar uma das decisões em teste ($\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ **ou** $\boldsymbol{\lambda} > \boldsymbol{\lambda}_0$) com base numa amostra observada $(x_1, x_2, ..., x_n)$.

Ora a amostra observada pertence ao referido conjunto $\mathcal{C}$ (isto é, conduz à rejeição de $\mathbf{H}_0$) se $\sum\limits_{i=1}^{n} x_i > n\left(\lambda_0 + k\right)$. Assim,

$$P\left(\left\{\sum_{i=1}^{n} X_i > \sum_{i=1}^{n} x_i\right\}/\lambda = \lambda_0\right) \leq P\left(\left\{\sum_{i=1}^{n} X_i > n\left(\lambda_0 + k\right)\right\}/\lambda = \lambda_0\right).$$

Determina-se então

$$\begin{aligned}
\alpha &= P\left(\left\{\sum_{i=1}^{n} X_i > \sum_{i=1}^{n} x_i\right\}/\lambda = \lambda_0\right) \\
&= P\left(n\lambda_0\right)\left(\left]\sum_{i=1}^{n} x_i, + + \infty\right[\right),
\end{aligned}$$

sendo portanto α o menor valor da probabilidade do erro que, com base na amostra observada, leva à rejeição de $\mathbf{H}_0$.

Este valor é designado por **p-valor** ou **significância** do teste para a amostra observada.

Tomar-se-á a decisão de rejeitar a hipótese $\mathbf{H}_0$ em teste se este valor α for pequeno (em geral inferior ou igual ao nível de significância mais usual de 5%). Em contrapartida, aceitar-se-á $\mathbf{H}_0$ se α for grande (em geral superior ou igual a 90%).

Aplicação ao exemplo de motivação

Pretende-se testar

$$\mathbf{H}_0\text{: } \boldsymbol{\lambda = 1} \ \ \textbf{contra} \ \ \mathbf{H}_1\text{: } \boldsymbol{\lambda > 1}.$$

A amostra observada é tal que

$$\sum_{i=1}^{15} x_i = 31.$$

Por outro lado, da estabilidade da lei de Poisson vem que, sob H_0,

$$\sum_{i=1}^{15} X_i \backsim P(15).$$

Então o p-valor do teste associado à amostra observada é

$$\begin{aligned} \alpha &= P\left(\left\{\sum_{i=1}^{15} X_i > 31\right\} / \lambda = 1\right) \\ &= P\left(15\right)\left(]31, + + \infty[\right) = 0,000090311608411, \end{aligned}$$

pelo que a hipótese $\mathbf{H}_0$ deve ser rejeitada.

Assim, tal como afirma o médico responsável, a média de doentes que acede à referida unidade aumentou. A amostra observada alerta, de facto, para a necessidade de reequipamento da referida unidade de hemodiálise.