Aproximação normal para as distribuições binomial e Poisson

Distribuição normal: aproximação para uma variável aleatória com um grande número de amostras.

$$\text{Distribuição binomial} \xrightarrow{\; n \longrightarrow \infty \;} \text{Distribuição normal}$$

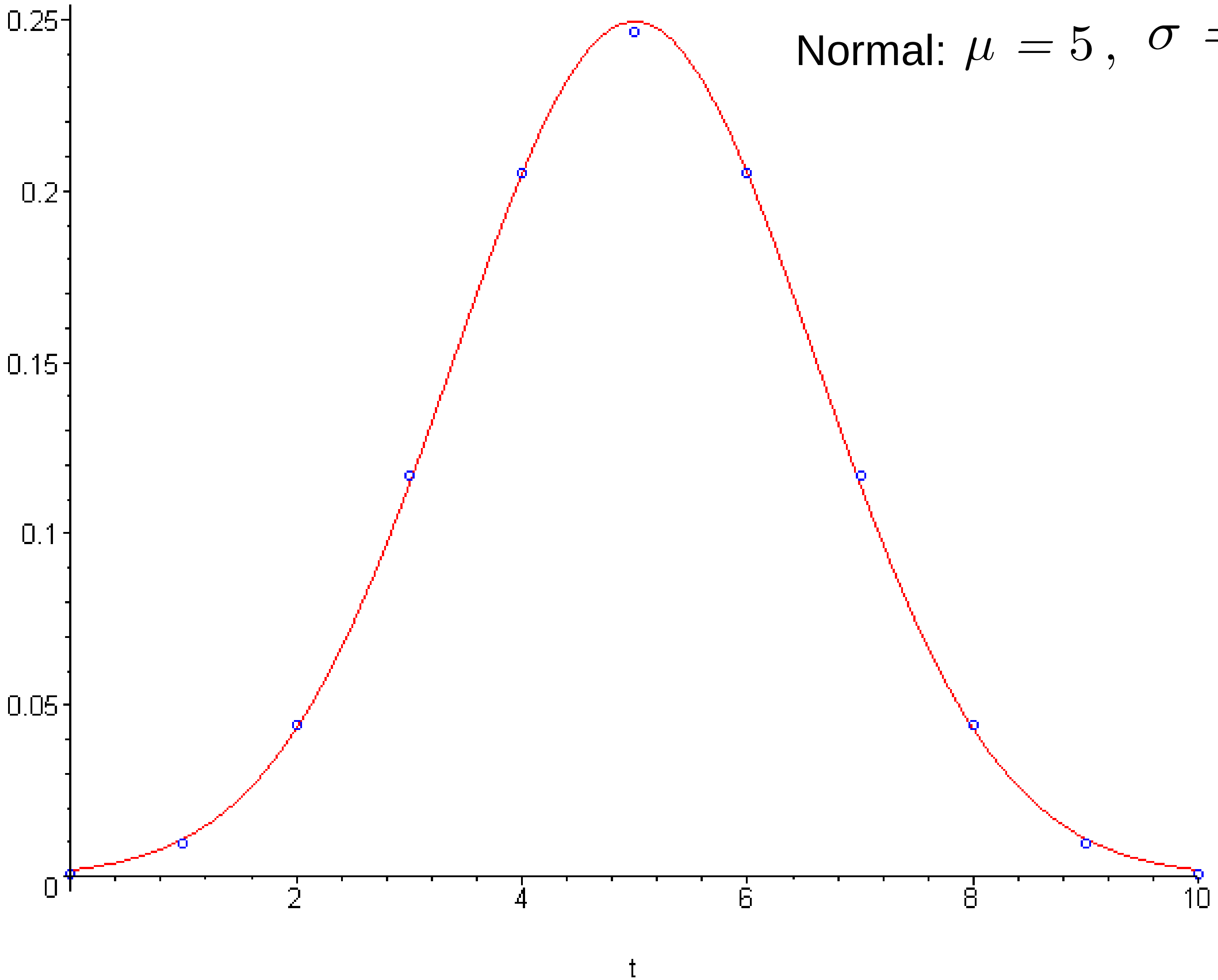
Difícil de calcular quando $n \longrightarrow \infty$

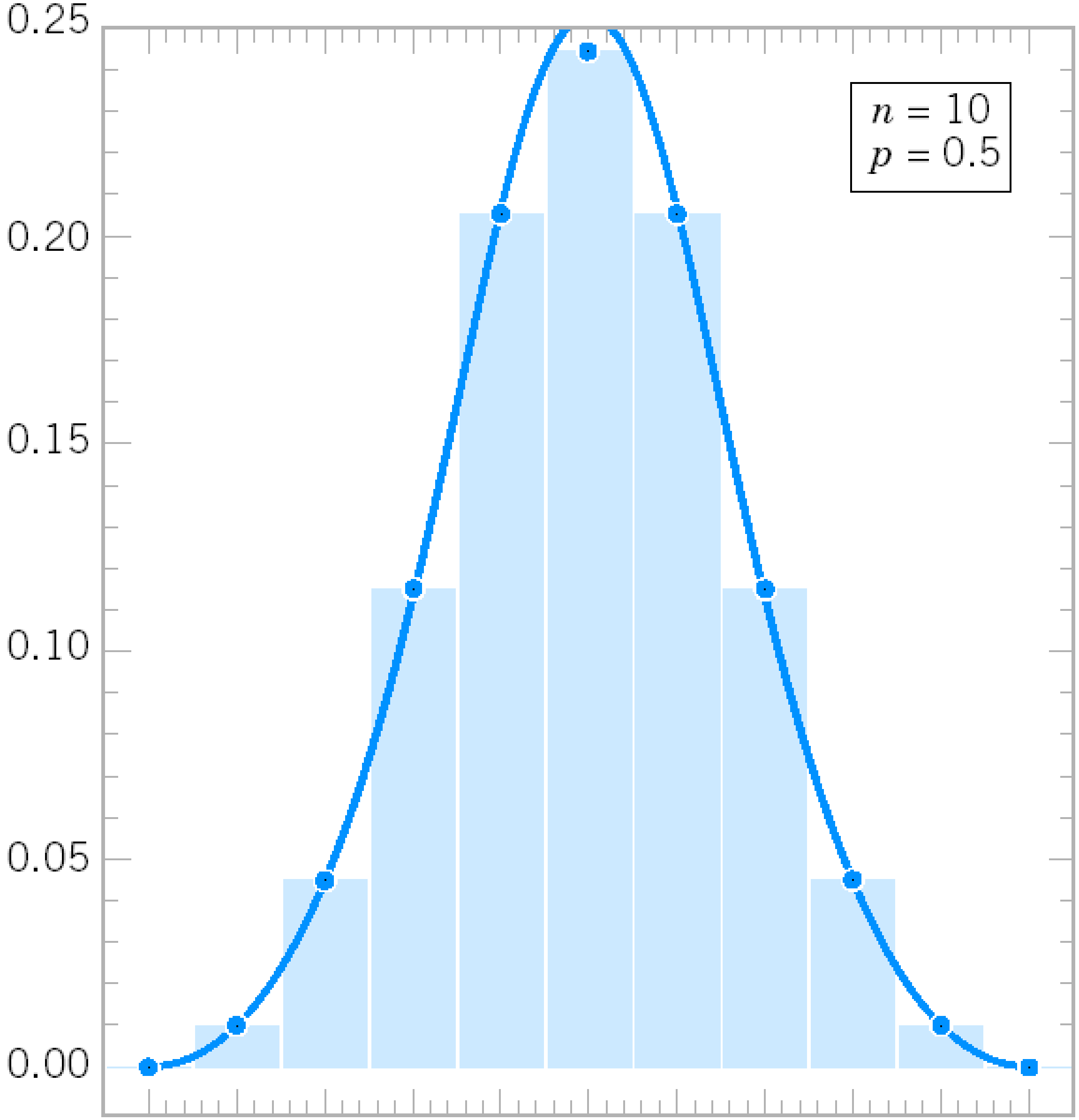
Exemplo com Maple

```
> with(Statistics):with(plots):

> n:=10:p:=0.5:

> for x from 0 to n do

    P[x]:=binomial(n,x)*p^x*(1-p)^(n-x);

od:

> xdata:=[seq(x,x=0..n)]:

> ydata:=[seq(P[x],x=0..n)]:

> PL1:=PointPlot(ydata,xcoords=xdata, color=blue,
symbol=circle):

> f:=(mu,sigma)->1/(sqrt(2*Pi)*sigma)*exp(-(t-mu)^2/(2*sigma^2));

> PL2:=plot(f(5,1.6),t=0..10):

> display([PL1,PL2]);
```

Binomial : n=10, p=0.5

Normal: $\mu = 5$, $\sigma = 1.6$




0.25
0.20
0.15
0.10
0.05
0.00
$n = 10$
$p = 0.5$

Se X é uma variável aleatória binomial, então

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

é aproximadamente uma variável aleatória padrão se

$$np > 5 \text{ e } n(1-p) > 5$$

Note que $E(X) = np$ $\quad V(X) = np(1-p)$

Portanto,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}}$$

Exemplo. Em um canal de comunicação digital, suponha que o número de bits em erro possa ser modelado por uma variável aleatória binomial. Suponha que a probabilidade de que um bit seja recebido com erro seja 1E-5. Se 16E6 bits são recebidos, qual é a probabilidade de que 150 cheguem com erro ?

$$P(X > 150) = 1 - P(x \leq 150)$$

$$= 1 - \sum_{x=0}^{150} \frac{16\times 10^6\,!}{x!\;\; 16\times 10^6 - x\;\;!}(10^{-5})^x(1-10^{-5})^{16\times 10^6 - x}$$

Tal expressão pode ser difícil de ser calculada.

Notemos que

$$np = 160, \quad n(1-p) \simeq 10^7$$

Padronização

$$P(X > 150) = P\left(\frac{X - 160}{\sqrt{160(1 - 10^{-5})}} > \frac{150 - 160}{\sqrt{160(1 - 10^{-5})}}\right)$$
$$= P(Z > -0.79) = P(Z < 0.79) = 0.785$$

```
> with(stats):
> statevalf[cdf,normald[0,1]](0.79);
```

0.7852361158

Lembremos que a distribuição binomial é uma aproximação satisfatória para a distribuição hipergeométrica quando o tamanho da amostra n é pequeno com relação ao tamanho da população N.

Regra geral: a aproximação binomial é efetiva quando

$$\frac{n}{N} < 0.1$$

Portanto, quando

$$\boxed{\frac{n}{N} < 0.1, \quad np > 5 \;\text{ e }\; n(1-p) > 5}$$

a distribuição normal é uma boa aproximação para a distribuição hipergeométrica.

Distribuição de Poisson: aproximação para a distribuição binomial quando o número de amostragens tende a infinito.

De fato,

Se X é uma variável aleatória de Poisson com

$$E(X) = \lambda, \quad V(X) = \lambda$$

então

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

é uma boa aproximação para uma variável aleatória normal padrão se

$$\lambda > 5$$

Exemplo: Suponha que o número de partículas de asbestos em um metro quadrado de poeira sobre uma superfície segue uma distribuição de Poisson com uma média de 1000. Se 1 m^2 de poeira é analisada, qual é a probabilidade de que no máximo 950 partículas sejam encontradas ?

A distribuição de Poisson é dada por

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

de modo que

$$P(X \leq 950) = \sum_{0}^{950} \frac{e^{-1000}1000^{x}}{x!}$$

Embora o Maple realize tal soma sem dificuldades,

```
> f:=x->exp(-1000)*1000^x/x!;
```

$$f := x \to \frac{\mathbf{e}^{(-1000)}\, 1000^{x}}{x!}$$

```
> evalf(sum(f(x),x=0..950));
```

$$0.05783629296$$

Podemos também usar a aproximação normal:

$$P(X < 950) = P\left(Z < \frac{950 - 1000}{\sqrt{1000}}\right)$$

$$= P(Z < 1.58) = 0.05692314901$$

Cálculo:

```
> z:=(950-1000)/sqrt(1000.);
```

$$z := -1.58113883$$

```
> statevalf[cdf,normald[0,1]](z);
```

$$0.0569231490$$

Distribuição Exponencial

Exemplo Falhas ocorrem aleatoriamente ao longo de um fio de cobre.

Variável aleatória de Poisson: número de falhas em um comprimento L de fio.

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Variável aleatória de interesse agora : distância entre as falhas.

X : comprimento desde qualquer ponto de partida at;e que uma falha seja detectada.

A distância até a primeira falha excede x mm se e somente não há qualquer falha em um comprimento de x mm.

N - número de falhas no em um comprimento x do fio.

Média de falhas por mm : λ

$$\Downarrow$$

N tem distribuição de Poisson com média

$$E(N) = \lambda x$$

Então, se o fio tem comprimento maior que x ,

$$P(X > x) = P(N = 0) = \frac{e^{-\lambda x}(\lambda x)^0}{0!} = e^{-\lambda x}$$

$$F(x) = P(X \leq x) = 1 - P(X > x) = 1 - e^{-\lambda x}$$

Função distribuição cumulativa de X

$$F(x) = 1 - e^{-\lambda x}$$

Função densidade de probabilidade de X

$$f(x) = \frac{dF(x)}{dx} = \lambda e^{-\lambda x}$$

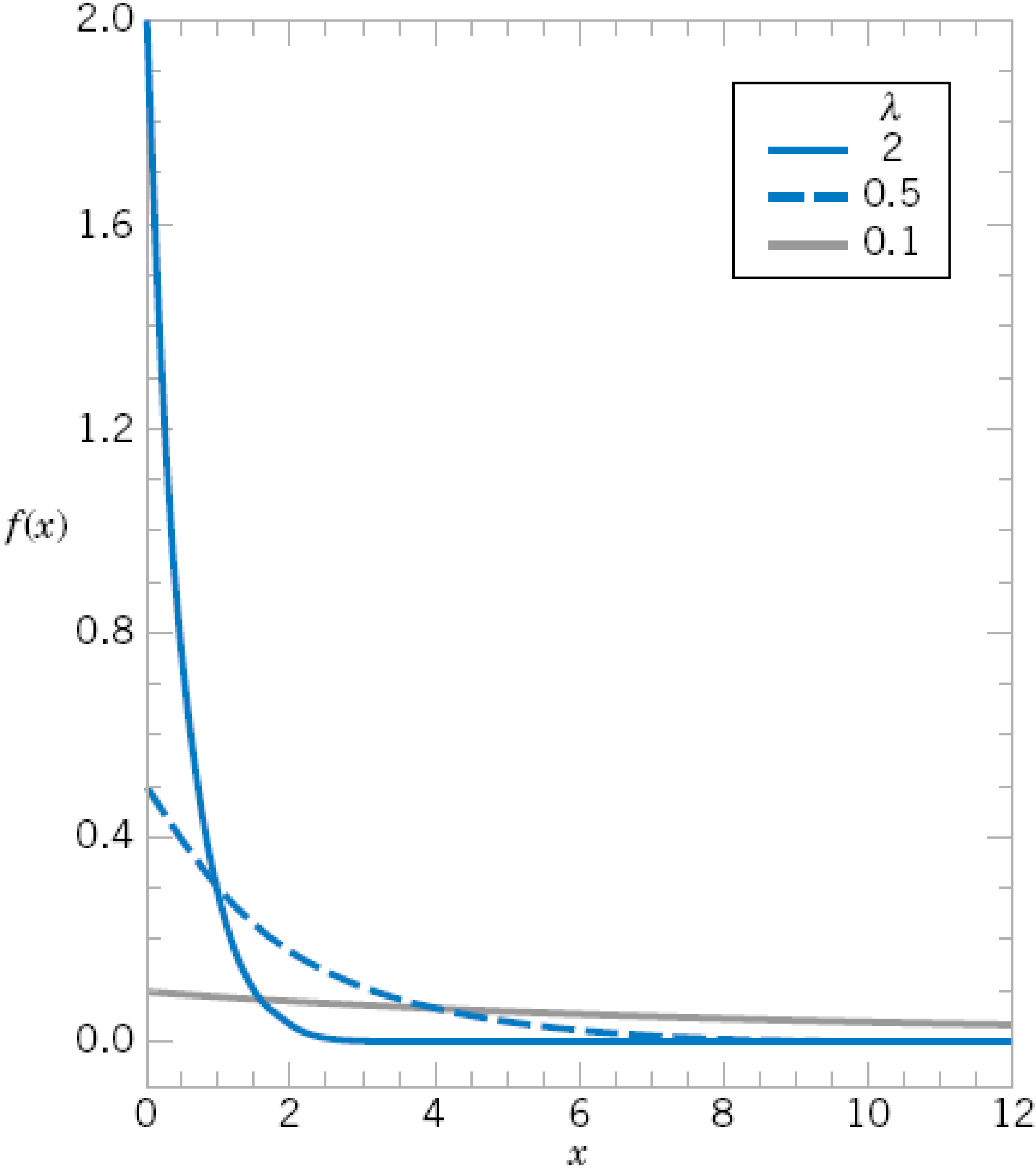
Suposição essencial: falhas no fio seguem uma distribuição de Poisson

Definição. Uma variável aleatória X que é igual à distância entre sucessivas contagens de um processo de Poisson com média

$$\lambda > 0$$

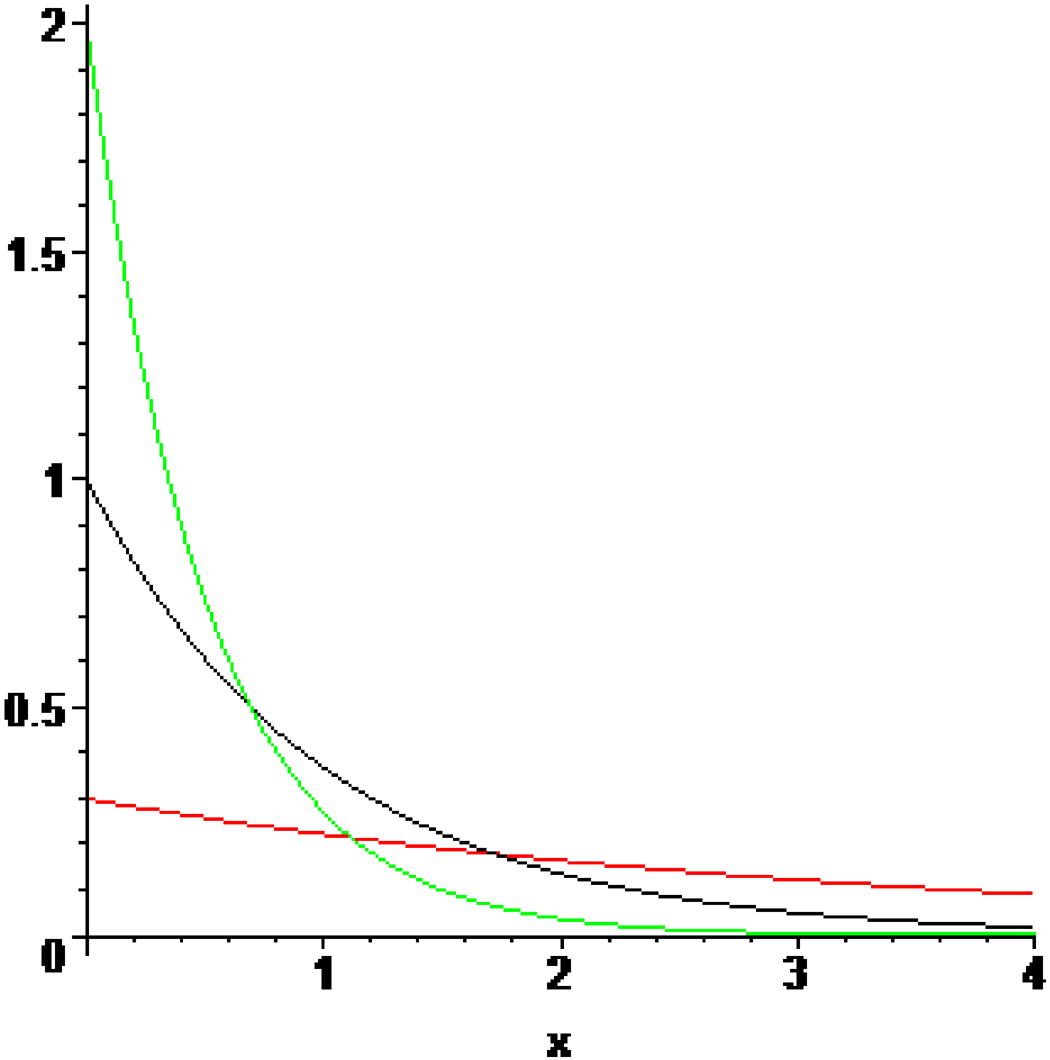
é uma variável aleatória exponencial com parâmetro λ e função densidade de de probabilidade dada por

$$\boxed{f(x) = \lambda e^{-\lambda x}, \quad 0 < x < \infty}$$


2.0
1.6
1.2
0.8
0.4
0.0
f(x)
0
2
4
6
8
10
12
x
λ
2
0.5
0.1

Gráficos no Maple

```
> f:=(lambda,x)->lambda*exp(-lambda*x):

> plot([f(2,x),f(1,x),f(0.3,x)],x=0..4,color=[green,black,red]);
```



Valor esperado e variância

$$\mu = E(X) = \frac{1}{\lambda} \qquad \sigma^2 = V(X) = \frac{1}{\lambda^2}$$

```
> assume(lambda>0):

> int(f(lambda,x),x=0..infinity);
```

$$1$$

```
> E:=int(x*f(lambda,x),x=0..infinity);
```

$$E := \frac{1}{\lambda\sim}$$

```
> V:=int((x-E)^2*f(lambda,x),x=0..infinity);
```

$$V := \frac{1}{\lambda\sim^2}$$

Exemplo. Em uma rede de computadores, conexões de usuários podem ser modeladas como um processo de Poisson, com uma média de 25 logons/hora. Qual é a probabilidade de que não haja logons num intervalo de 6 min. ?

X – tempo em horas a partir do início do intervalo até o primeiro logon.

$$\Downarrow$$

X tem uma distribuição exponencial com $\lambda = 25$ logons/hora.

Como 6 min = 0.1 h, desejamos calcular

$$P(X > 0.1)$$

Ou seja,

$$P(X > 0.1) = \int_{0.1}^{\infty} 25e^{-25x}\,dx = -e^{-25x}\Big|_{0.1}^{\infty} = e^{-2.5}$$

$$= 0.082$$

Qual é a probabilidade de que o tempo até o próximo logon esteja entre 2 e 3 min. ?

$$P(0.033 < X < 0.05) = \int_{0.033}^{0.05} 25e^{-25x}\,dx = -e^{-25x}\Big|_{0.033}^{0.005}$$

$$= 0.148$$

No Maple,

```
> a:=2/60: b:=3/60:
> evalf(int(f(25,x),x=a..b));
```

$$0.1480934116$$

Determine o intervalo de tempo tal que a probabilidade de que nenhum logon ocorra no intervalo é 0.9.

Devemos achar x tal que

$$P(X > x) = \int_x^{\infty} 25e^{-25t}\, dt = -e^{-25t}\Big|_x^{\infty} = e^{-25x} = 0.9$$

Portanto,

$$-25x = \ln 0.9$$

$$x = 0.00421 \text{ h} \;\; ; \quad 0.25 \text{ min}$$

Notas

A probabilidade de que não haja logons em um intervalo de 6 min é 0.082, independentemente do ponto de partida do intervalo.

Um processo de Poisson supõe que eventos ocorrem uniformemente ao longo do intervalo de observação - não há acumulação de eventos.

Nosso ponto de partida para observar o sistema não é importante.

Se há horários de alta e baixa demanda de logons, o processo de Poisson não é apropriado. Processo pode ser aplicado em intervalos.

Propriedade de falta de memória

Exemplo. Seja X o tempo entre detecções de uma partícula com um contador geiger. Suponhamos que X tem uma distribuição exponencial com

$$\lambda = 1.4 \min$$

A probabilidade se detectarmos uma partícula dentro de 30s a partir do momento em que o contador entra em funcionamento é:

$$P(X < 0.5) = \int_0^{0.5} 1.4e^{-(1.4)x}\,dx = 0.5034146962$$

Suponhamos agora que ligamos o contador geiger e 3 min se passam sem que a partícula seja detectada. Qual é a probabilidade de que uma partícula seja detectada nos próximos 30 s ?

Como já esperamos 3 min poderíamos pensar que a probabilidade de uma detecção nos próximos 30s é maior que 0.5. No entanto, os 3 min iniciais não têm qualquer influência. De fato,

$$P(X < 3.5 \mid X > 3.0) = \frac{P(3.0 < X < 3.5)}{P(X > 3.0)} = 0.503414$$

```
> f:=1.4*exp(-1.4*x):
> I1:=int(f,x=3..3.5);
```

$$I1 := 0.00754899375$$

```
> I2:=int(f,x=3..infinity);
```

$$I2 := 0.0149955768$$

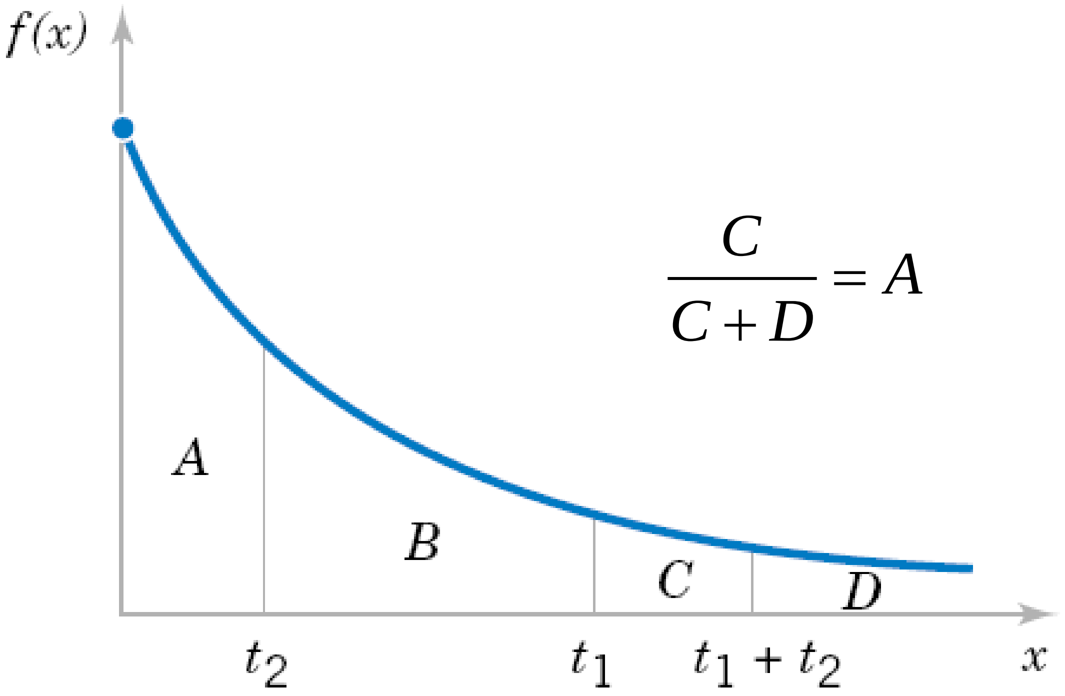
```
> I1/I2;
```

$$0.5034146963$$

Propriedade da falta de memória

$$P(X < t_1 + t_2 \mid X > t_1) = P(X < t_2)$$

$$= P(t_1 < X < t_1 + t_2) / P(X > t_1) = P(X < t_2)$$



$$\frac{C}{C + D} = A$$

Uso da distribuição exponencial:

Estudos de confiabilidade para o tempo até a falha de um dispositivo, por exemplo, vida de um semicondutor.

Falta de memória → Dispositivo não se desgasta

↓

Falha ocorre por choques mecânicos aleatórios, sobrecargas aleatórias de corrente.

Desgaste lento → Distribuição de Weibull