MODELS de VAD i VAC

- Models de VAD i VAC
- El model BERNOULLI
- El procés de Bernoulli i el procés de Poisson
- El model BINOMIAL
- Models: Geomètrica i Binomial Negativa
- El model POISSON
- El model EXPONENCIAL
- El model Uniforme
- El model NORMAL
- Teorema Central del Límit

Models de VAD i VAC

(Wikipedia.org) *List of probability distributions:* "Many probability distributions are so important in theory or applications that they have been given specific names"

En PE destacarem:
Binomial (VAD), **Poisson** (VAD), **Exponencial** (VAC), **Normal** (VAC)

i comentarem: Bernoulli(VAD), Geomètrica(VAD), Binomial Negativa(VAD), Uniforme(VAC)

Recordem que, en funció dels paràmetres de cada model s'expressen: $E(X) = \mu_X = \ldots$ $V(X) = \sigma_X^{\,2} = \ldots$

Recordem que calcularem probabilitats directes:

- En VAD
  - $P(X=k) = p_X(k)$
  - $P(X \leq k) = F_X(k) = \Sigma_{j \leq k}\, p_X(j)$
  - $P(X<k) = P(X \leq k-1) = F_X(k-1)$
  - $P(a < X \leq b) = F_X(b) - F_X(a)$
- En VAC
  - $P(X=k) = 0$
  - $P(X \leq k) = F_X(k)$
  - $P(X<k) = P(X \leq k) = F_X(k)$
  - $P(a \leq X \leq b) = F_X(b) - F_X(a)$

i probabilitats inverses: donada una probabilitat α calcular el quantil α (o percentil α en %) és el problema invers al càlcul de probabilitats acumulades:
x_α és el quantil α de X si es compleix: $F_X(x_\alpha) = \alpha$ $\quad (0 \leq \alpha \leq 1)$ (en VAD el càlcul és aproximat)

i també probabilitats condicionades: P(valors en un model | valors al mateix o altre model)

El model BERNOULLI

El model teòric general més senzill aplicable a una variable aleatòria és el model de Bernoulli:

k	$P_X(k)$
0	$1 - p = q$
1	p

Els valors "0" i "1" tenen un sentit ampli:

"1" significa "èxit" en l'opció d'interès. En un sentit ampli pot significar *encert, positiu,…*;

"0" significa "no èxit" en l'opció d'interès. Representa el complementari: *error*, *fracàs*, *negatiu*,...

El paràmetre p és la probabilitat d'observar un "èxit".

El cas més habitual són experiències aleatòries que impliquen repeticions de proves Bernoulli, que donen lloc al Procés de Bernoulli i el de Poisson.
En aquests és necessari que unes proves siguin *independents* d'altres i que la probabilitat d'èxit sigui constant i igual a p

Procés de BERNOULLI i procés de POISSON

En una experiència aleatòria que implica repetició de proves Bernoulli independents, ens situa davant d'un **Procés de Bernoulli** en el que poden plantejar-ne com distribucions interessants:

- sobre *n* repeticions, número "d'èxits" totals (dist. binomial)
- número de repeticions fins observar el primer "èxit" (dist. geomètrica)
- número de repeticions fins observar el *r*-èssim "èxit" (dist. binomial negativa)

En experiències aleatòries on el número de repeticions n té un valor gran i p un valor petit (fenòmens estranys), pot ser més fàcil identificar el promig (np) d'"èxits" (en un interval) que explícitament el valor de n i p. Estem davant d'un **Procés de Poisson,** on es plantegen com distribucions interessants:

- número "d'èxits" en el interval (dist. de Poisson)
- temps entre "èxits" (dist. Exponencial)

El model BINOMIAL: $X \sim B(n, p)$

L'experiència suposa un número fix *n* de proves realitzades, dins d'un procés de Bernoulli.

La variable *X* que conta el número (de 0 a n) d'encerts en les n repeticions es denota $X \sim B(n, p)$ i té com a funció de probabilitat:

$$P(X=k) = \binom{n}{k} p^k q^{n-k} \qquad k=0, 1, \ldots, n$$

resultat				X	probabilitat			
0	0	0	0	0	q	q	q	q
0	0	0	1	1	q	q	q	p
0	0	1	0	1	q	q	p	q
0	0	1	1	2	q	q	p	p
0	1	0	0	1	q	p	q	q
0	1	0	1	2	q	p	q	p
0	1	1	0	2	q	p	p	q
0	1	1	1	3	q	p	p	p
1	0	0	0	1	p	q	q	q
1	0	0	1	2	p	q	q	p
1	0	1	0	2	p	q	p	q
1	0	1	1	3	p	q	p	p
1	1	0	0	2	p	p	q	q
1	1	0	1	3	p	p	q	p
1	1	1	0	3	p	p	p	q
1	1	1	1	4	p	p	p	p

Exemple: 4 tirades del dau (0 és no surt "1", 1 és sí surt "1", i X és el nombre de "1")

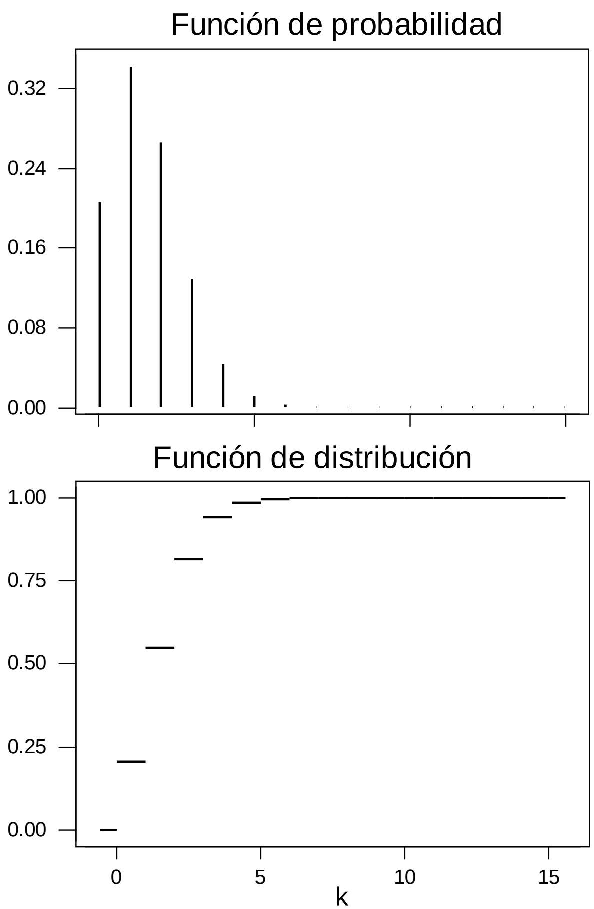
Contarem 1 vegada 0 i 4,
4 vegades 1 i 3,
6 vegades 2 i 3.

I a més, els resultats no són equiprobables (excepte si *p*=1/2):

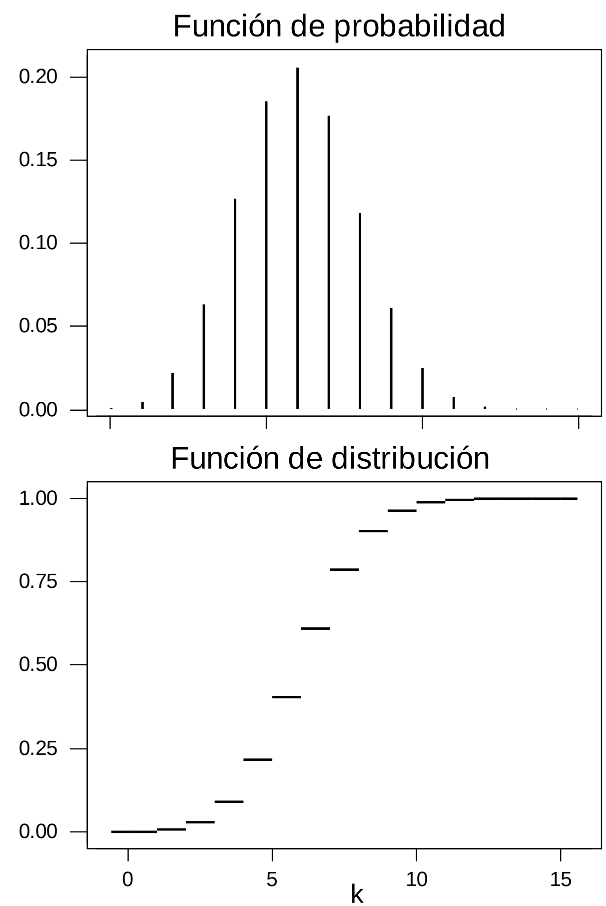
- obtenir 4 cares diferents de "1" té probab. de $(5/6)^4=0.48$;
- i obtenir 4 "1" té una probab. de $(1/6)^4=0.00077$.

EXEMPLE. Un tècnic agrupa per quinzenes les seves notes sobre la paginació del sistema. Com es distribueix el <u>número de dies de cada 15 amb paginació alta</u> segons si la probabilitat de paginació alta diària és...?

p=0.1



p=0.4



EXERCICI. Càlcul de probabilitats en el model Binomial:

Sigui $X \sim B(20, 0.5)$ i alguna de les probabilitats tabulades:

…	
12	0.8684
13	0.9423
14	**0.9793**
15	0.9941
16	0.9987
17	0.9998
…	

Probabilitat puntual

Quina és la probabilitat de 14?:

. . .

Probabilitats acumulades

Quina és la probabilitat de 14 o menys?

. . .

Quin és el quantil de 0,95?

(o, quin és el percentil 95 (95%)?)

. . .

EXERCICI. Binomial

La cabina de discos d'un servidor conté 18 discs idèntics. Un disc de cada 5 necessita ser substituït al cap d'un any per ser reparat.

- Una variable convenient:
- Llei de probabilitat:
- Número esperat de reparacions anuals:
- Variància i desviació típica:
- Prob. d'observar 4 casos:
- Prob. sols una averia:
- Prob. de tenir menys de 3:
- Prob. de que al menys 6 discs han de ser reparats:

- R: "número de discs que han de ser reparats al cap de l'any"
- $R \sim$
- $E(R) =$
- $V(R) =$ $\sigma_R =$
- $P(R=4) =$
- $P(R=1) =$
- $P(R<3)=$
- $P(R\geq 6) =$

Models: GEOMÈTRICA i BINOMIAL NEGATIVA

Número d'intents (k) fins observar el primer èxit (distribució geomètrica)

$k=1 \Leftrightarrow (1)$: $P_X(1) = p$ $k=2 \Leftrightarrow (0,1)$: $P_X(2) = q \cdot p$ $k=3 \Leftrightarrow (0,0,1)$: $P_X(3) = q^2 \cdot p$ …

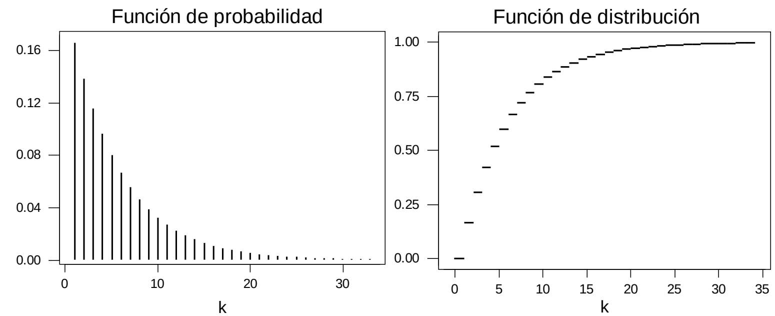
k intents $\Leftrightarrow$ $(0,\ldots^{(k-1)}\ldots,0,1)$: $P_X(k) = q^{k-1} \cdot p$, per tant $P(X=k) = q^{k-1} \cdot p$

Observeu en els gràfics: -

no hi ha un màxim, qualsevol valor enter > 0 és possible

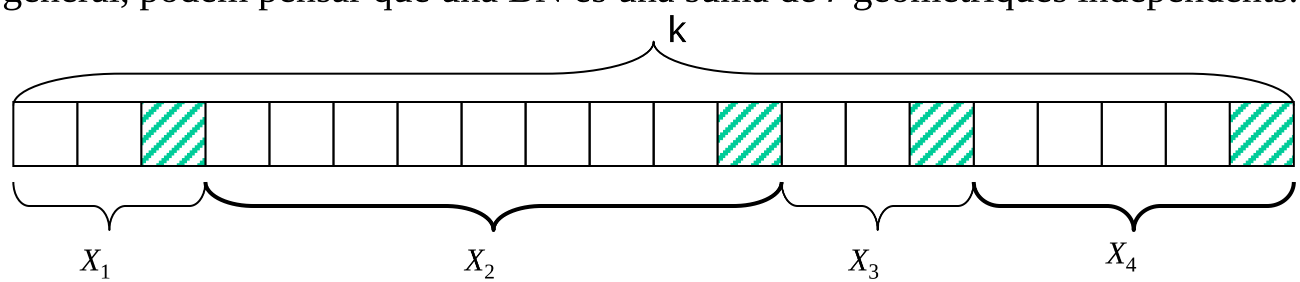
(Ex. "tirar el dau moltes vegades fins que surti el primer 1")

- el més probable és que el número d'intents no sigui molt alt
- si p augmenta, $P_X(k)$ es traslladada a valors més baixos, i $F_X(k)$ creix més ràpid.



Número d'intents (k) fins observar el *r*-èssim èxit (distribució binomial negativa)

Quantes repeticions fan falta per aconseguir r èxits (r >1)? Si r =1, tenim la distribució geomètrica. En general, podem pensar que una BN és una suma de r geomètriques independents.



Sense comptar l'últim intent (que ha de ser un èxit), són r -1 èxits barrejats en qualsevol combinació amb k-r fracassos. $P_X(k)$, amb $k \geq r$ és

$$P(X = k) = \binom{k-1}{r-1} p^r q^{k-r}$$

El model POISSON: $X \sim P(\lambda)$

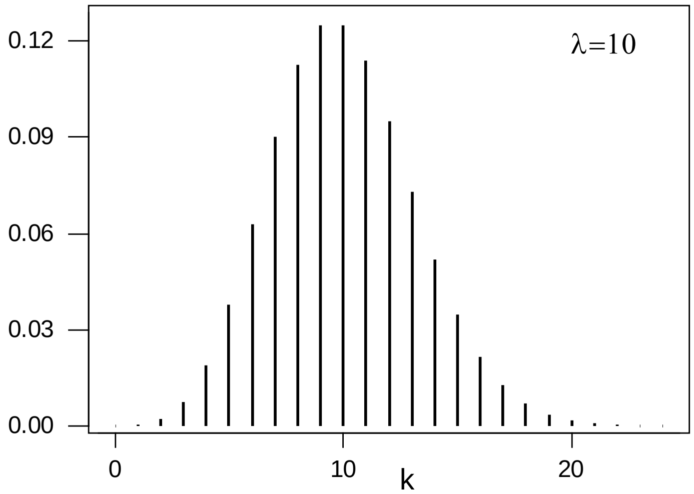
Al igual que la binomial, compta el número d'ocurrències favorables en un determinat entorn. No hi ha pròpiament una repetició d'experiències idèntiques tipus Bernoulli, sinó un procés continu en el que inesperadament pot ocórrer el fenomen en qüestió (per ex., una trucada a una centraleta)

La variable X que compta el número d'ocurrències favorables (en un interval)
es denota $X \sim \mathcal{P}(\lambda)$ i té com a funció de probabilitat:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

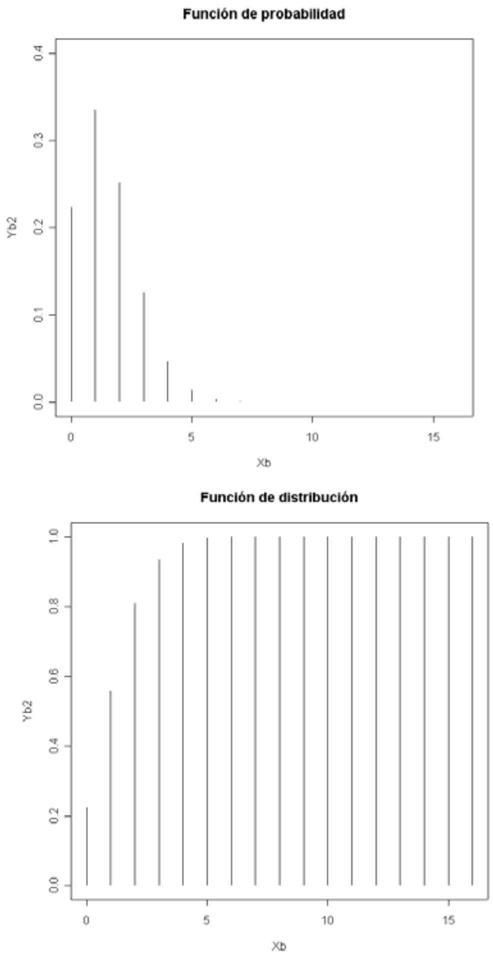
λ és un número real positiu que representa la taxa mitja d'ocurrències per unitat considerada (10 trucades/hora, p. ex.).

Una variable de Poisson pot agafar qualsevol valor enter $k \geq 0$, encara que en la pràctica sols els que estan relativament a prop de λ tenen probabilitats rellevants.

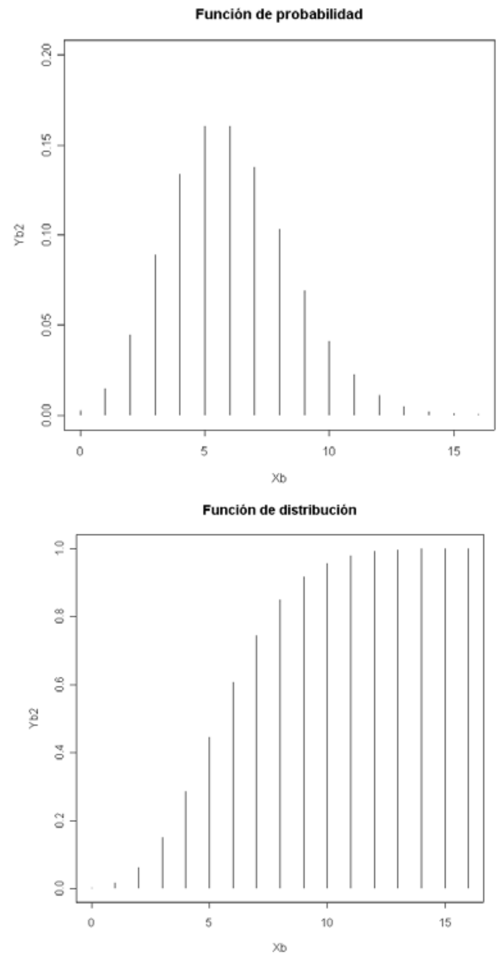


EXEMPLE. Un tècnic anota per períodes la mitjana de dies amb paginació alta. Com es distribueix el número de dies amb paginació alta segons el valor d'aquesta mitjana …?

λ=1.5



λ=6



Algunes característiques dels anteriors models

Característiques d'una variable aleatòria X segons el model de distribució

$(E(X) = \mu_X)$ $(V(X) = \sigma_X^2)$

Distribució	Declaració	Imatge	Esperança	Variància
Bernoulli	Bern(p)	0,1	p	$p*q$
Geomètrica	Geom(p)	1,2,3,...	$1/p$	q/p^2
Binomial	B(n,p)	0,1,...,n	$n*p$	$n*p*q$
Binomial negativa	BN(r,p)	r,r+1,...	r/p	$q*r/p^2$
Poisson	P(λ)	0,1,2,...	λ	λ

$0 < p < 1$; $q = 1 - p$; n enter > 0; r enter > 0; λ real > 0.0

EXERCICI: Poisson

El centre de càlcul d'una important empresa atén els problemes (o incidències) que els hi sorgeix als treballadors. S'ha observat que les incidències apareixen esporàdicament, encara que l'elevat número d'usuaris suposa que el volum de problemes a tractar diàriament sigui considerable (s'ha suposat un promig de 2.35 incid./dia).

1. Probabilitat de que en un dia es produeixi 3 incidències:

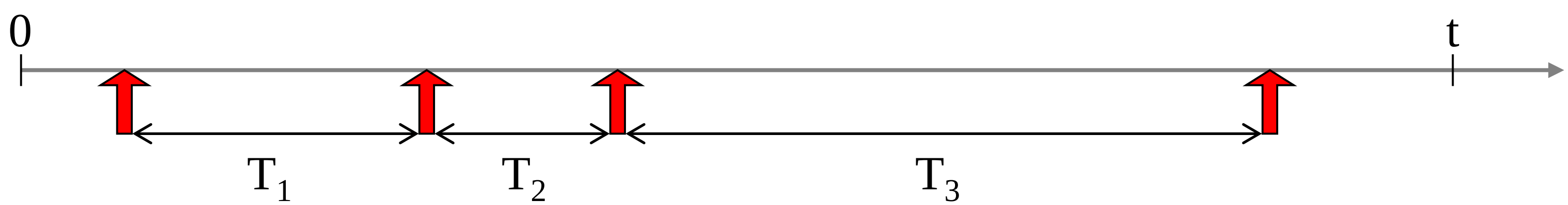
2. Probabilitat d'observar menys de 3 incidències en un dia:

3. Entre el dilluns i el dimarts han rebut sis incidències.
Quina és la probabilitat que de dilluns a divendres es tractin no més de 15?

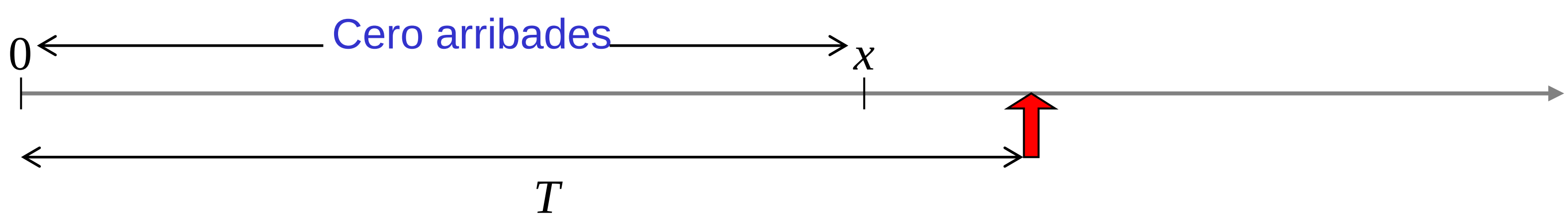
4. Quina és la probabilitat de que cap dia de la setmana laboral presenti incidències?

El model EXPONENCIAL: $X \sim \text{Exp}(\lambda)$

Distribució del temps entre arribades (ocurrències) en un procés de Poisson:
En l'interval [0, t] les arribades al sistema (N_t) segueixen una distribució de Poisson, amb taxa $\lambda \cdot t$ (la taxa per unitat de temps és λ). El temps entre dos arribades consecutives és una magnitud continua i és indeterminista. Com es distribueix?



Els valors possibles són $x > 0$. Sigui T la variable "temps entre dos arribades consecutives".
Busquem, per qualsevol valor x, $P(T \leq x)$.
És més simple buscar $P(T > x)$: quins esdeveniments reuneix?



$$F_T(x) = P(T \leq x) = 1- P(T > x) = 1- P(N_x = 0) = 1- e^{-\lambda \cdot x}\, \lambda^0/0! = 1 - e^{-\lambda \cdot x}$$

La distribució que acabem de deduir es coneix com distribució exponencial

La variable temps entre dos arribades s'expressa com $T \sim \text{Exp}(\lambda)$ $(\lambda > 0)$ i té com a funció de distribució de probabilitat:

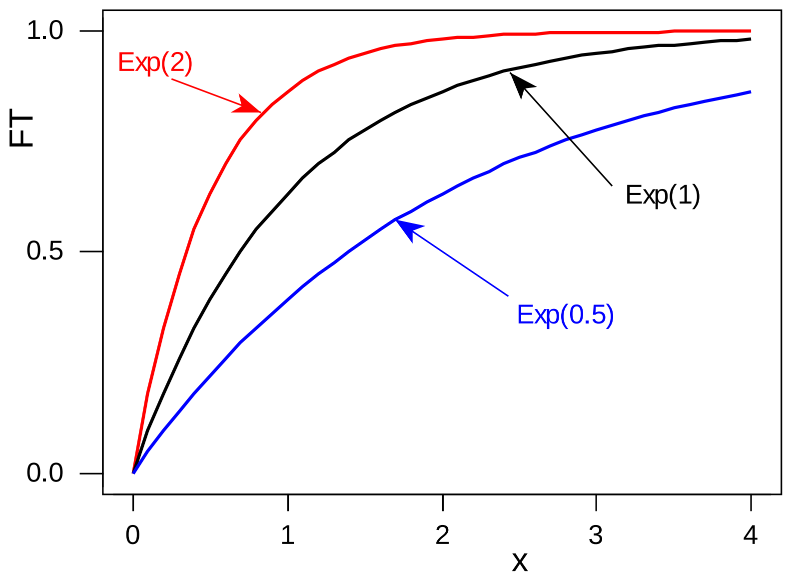
$$P(T \leq x) = 1 - e^{-\lambda \cdot x}$$

Exemple:

- N: número de peticions/seg a un servidor de BBDD, $N \sim P(\lambda)$
- X: temps (seg.) transcorregut entre dos peticions consecutives

$P(X \leq 0.5\text{s}) =$

- 0.632 $(\lambda = 2$ arr/s)
- 0.393 $(\lambda = 1$ arr/s)
- 0.221 $(\lambda = 0.5$ arr/s)



Si les arribades són més freqüents (λ alt), el temps entre arribades són més curts. Per tant, és més probable trobar temps per sota de mig segon quan λ és major (*F_X creix més ràpid*).

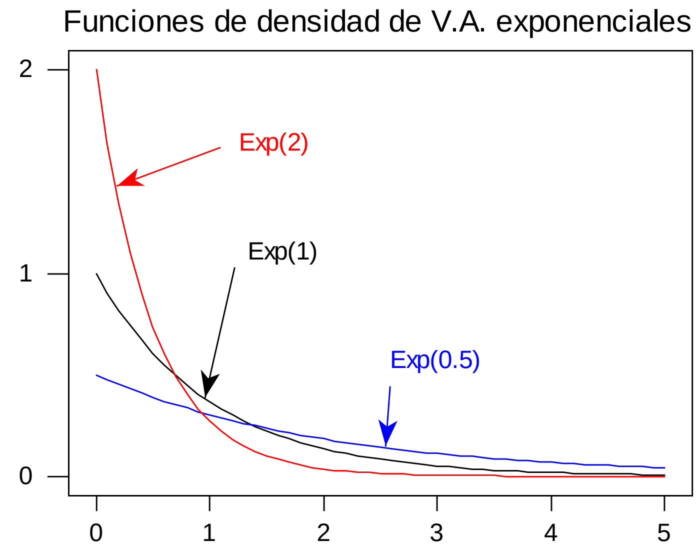
Més intuïtiva, és la funció de densitat de probabilitat: $f_X(x)$.

Si $X \sim \mathrm{Exp}(\lambda)$

amb $F_X(x) = 1 - e^{-\lambda \cdot x}$

llavors $$f_X(x) = \frac{dF_X(x)}{dx}$$

és a dir $\mathbf{f_X(x) = \lambda \cdot e^{-\lambda \cdot x}}$ **(x>0)**



- $f_X(x)$ no és $P(X=x)$ (que és igual a 0 per definició)

  ($f_X(x)$ no és una probabilitat, a diferencia de la $p_X(x)$ de les discretes)

- Recordem que en VAC $P(a \leq X) = P(a < X)$

  i $P(a \leq X \leq b) = P(a < X < b) = F_X(b) - F_X(a)$

Altre exemple habitual: vida útil d'un component electrónic. La seva duració pot ser molt llarga, però no són infreqüents els errors primerencs.

Integrant per parts, obtenim que $E(X) = 1/\lambda$ i $V(X) = 1/\lambda^2$ $(\sigma_X = 1/\lambda)$

(*en l'exemple del servidor de BBDD, si les peticions arriben a un promig de 1/2 per segon, el temps mig entre dos peticions és de 2 segons*)

Propietat de Markov. En l'ex. del servidor de BBDD, en un instant donat fa 10" que no arriben peticions. Que és més probable: (A) rebre en aquest moment, o (B) rebre justament desprès d'una arribada?

(Solució: igual)

$\mathbf{P(X > t+s \mid X>t)} =$

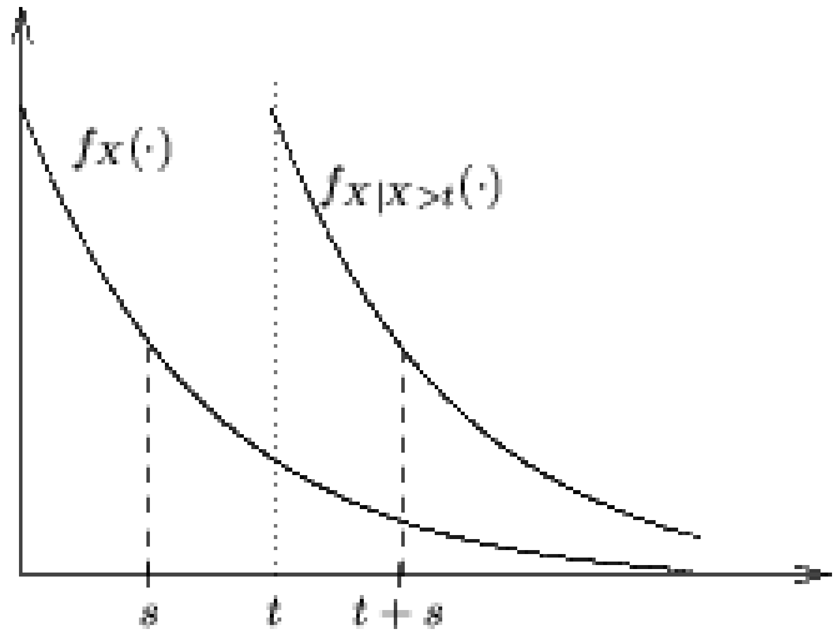
$P(X > t+s \cap X>t) / P(X>t) =$

$P(X > t+s) / P(X>t) =$

$(1-F_X(t+s)) / (1-F_X(t)) =$

$e^{-\lambda(t+s)} / e^{-\lambda t} = e^{-\lambda s} =$

$1-F_X(s) = \mathbf{P(X > s)}$



EXERCICI: Poisson i Exponencial

El centre de càlcul d’una important empresa atén els problemes (o incidències) que els hi sorgeix als treballadors. Es suposa un promig de 4 incid./dia).

1. Probabilitat de rebre 0 incidències en un dia:

   i 0 incidències en una hora:

2. Quina és l’esperança de la variable temps (en hores) entre incidències:

3. Probabilitat d’estar 8 o més hores sense rebre incidències:

Altres aplicacions del model exponencial:

"Failure Rate and Reliability" (Jane Horgan lecture 17) *(es veurà ex. més endavant)*

$X \sim Exp(\lambda)$ $F_X(x)=P(X<=x)=1 - e^{-\lambda \cdot x}$ $R_X(x)= P(X>x) = e^{-\lambda \cdot x}$

R és la funció de "reliability" o de fiabilitat (probabilitat de durar més de..)

λ és "failure rate" o taxa d'error

$E(X) = 1/\lambda$ és "MTTF" o "Mean Time To Failure"

"Modelling Response Times: M/M/1" (Jane Horgan lecture 17) (*es veurà ex. al laboratori)*

La teoria de cues permet estudiar sistemes en que interaccionen més d'una variable (per exemple temps de resposta en un sistema d'espera amb cues).

El model M/M/1 indica: 1 per contemplar una sola cua no finita, i dos M's pels dos temps d'arribada i servei exponencials (*Exp*(λ) i *Exp*(μ) o número d'arribades $\mathcal{P}(\lambda)$ i serveis $\mathcal{P}(\mu)$), que compleixen la propietat de Markov de no tenir memòria.

Amb λ i μ es defineix un indicador del sistema com factor de càrrega o "traffic intensity" (λ/μ):

- si $\lambda/\mu>1$ ($\lambda>\mu$) la taxa d'arribades és superior a la de sortides (sobreutilització del sistema)
- si $\lambda/\mu=1$ ($\lambda=\mu$) la taxa d'arribades s'iguala amb la taxa de sortides
- si $\lambda/\mu<1$ ($\lambda<\mu$) la taxa d'arribades és inferior a la de sortides (infrautilització del sistema)

(el model M/M/1 permet un tractament per teoria de cues o per simulació. Models més complexos poden permetre sols el tractament per simulació)

EXERCICI: Poisson i Exponencial (fiabilitat)

El centre de càlcul d’una important empresa garanteix treballar amb un promig de 2 hores entre incidències

1. Quina és la funció de distribució de probabilitat:

i la funció de fiabilitat (“reliability”):

i la “taxa d’errors”:

i el “MTTF”:

2. Quin és el valor d’hores entre incidències que podem garantir que es superarà amb una fiabilitat del 78%?

Algunes característiques dels anteriors models

Distribució	*Declaració*	**Funció de probabilitat o de densitat**	**Funció de distribució**	**Esperança** ($E(X) = \mu_X$)	**Variància** ($V(X) = \sigma_X^2$)
Bernoulli	X~Bern(p)			p	p q
Geomètrica	X~Geom(p)				
Binomial	**X~B(n,p)**	$P_X(k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}$ **o `R:dbinom(k,n,p)`**	$F_X(k) = \sum_{i<=k} P_X(i)$ **o taules estadístiques** **o `R:pbinom(k,n,p)`**	**n p**	**n p q**
Binomial negativa	X~BN(r,p)				
Poisson	**X~$\mathcal{P}$(λ)**	$P_X(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$ **o `R:dpois(k,λ)`**	$F_X(k) = \sum_{i<=k} P_X(i)$ **o taules estadístiques** **o `R:ppois(k,λ)`**	λ	λ
Exponencial	**X~Exp(λ)**	$f_X(x) = \lambda \cdot e^{-\lambda \cdot x}$ **o `R:dexp(x,λ)`**	$F_X(x) = 1 - e^{-\lambda \cdot x}$ **o `R:pexp(x,λ)`**	$1/\lambda$	$1/\lambda^2$
Uniforme	X~U[a,b]				
Normal	X~N(μ,σ)				

$0 < p < 1$; $q = 1 - p$; n enter > 0; r enter > 0; λ real > 0.0 (paràmetre procés de Poisson)

EXERCICI: Airport III (e-status)

Continuem analitzant el cas del aeroport, donant importància al procés d'arribades de passatgers als punts de facturació.
Respon a les següents preguntes. Els valors de probabilitat han de ser correctes fins al tercer decimal, com a mínim.

1. Un determinat punt de facturació es caracteritza perquè el número de viatgers que arriben per minut es distribueix segons una llei Poisson amb promig=9.5. Calcula la probabilitat de que aquest número sigui menor que 7.

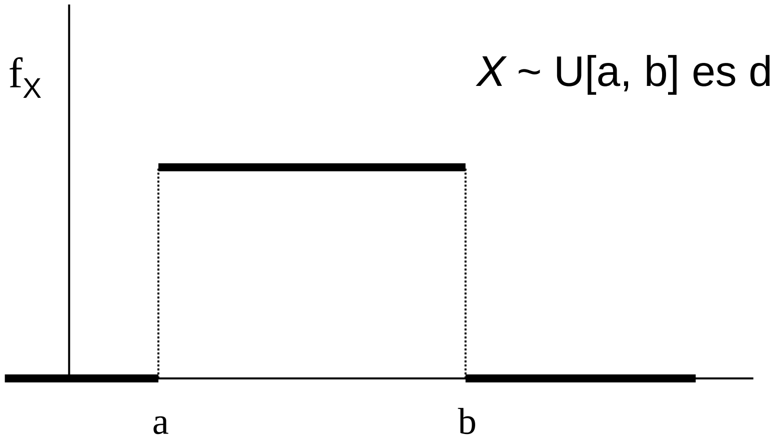
2. En aquest punt de facturació, quina és la probabilitat d'observar exactament 10 arribades en un minut?

3. En el mateix punt de facturació, quina és l'esperança de la variable temps (*en segons*) entre dos arribades?
Introduir al menys dos decimals correctes.

4. Al punt de facturació, quina és la probabilitat d'estar menys de 4 segons sense arribades?

5. Considerant 18 punts de facturació caracteritzats per una probabilitat 0.8 d'observar exactament 0 arribades en un minut, quina és la probabilitat de tenir més de 14 punts amb 0 arribades?

El model UNIFORME: $X \sim U[a, b]$



$X \sim U[a, b]$ es defineix per $F_X(u) = (u-a)/(b-a)$, $a \le u \le b$;

$F_X(u) = 0$, $u < a$;

$F_X(u) = 1$, $u > b$.

$$f_X(x) = \begin{cases} 1/(b-a) & a < x < b \\ 0 & x < a \cup x > b \end{cases}$$

Intuïtivament, el valor promig o esperat és el centre, en qüestió de simetria. Comprovem-ho $\Rightarrow$

$$E(X) = \mu_X = \int_{-\infty}^{+\infty} x \, f_X(x) dx = \int_a^b x \frac{1}{b-a} dx =$$

$$= \frac{1}{b-a} \left. \frac{x^2}{2} \right|_a^b = \frac{b^2 - a^2}{2(b-a)} = \boldsymbol{\frac{b+a}{2}}$$

En quant a la variància, arribaríem al següent resultat: $\sigma_X^2 = (b-a)^2/12$ (proveu-ho). Observeu que la desviació típica és proporcional a l'amplitud del interval (a, b).

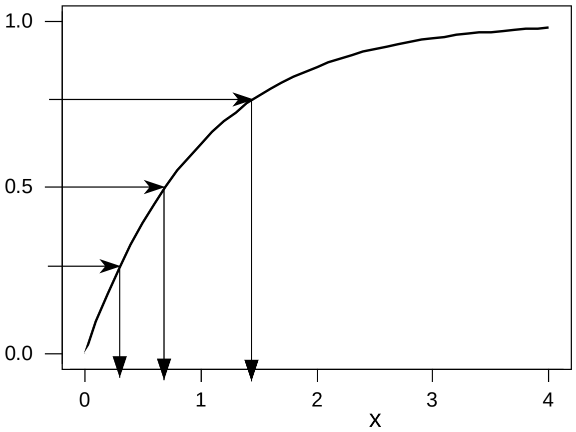
Un generador de números pseudo-aleatoris U segueix una distr. U[0, 1]. La relació X=a+(b-a)·U és U[a, b]. És més fàcil trobar E(U) i V(U) i aplicar propietats.

Aplicació del model uniforme i exponencial en simulació:

Generació de números aleatoris amb distribució exponencial.

Els programes de simulació solen requerir valors que segueixen una llei Exp(λ) per reproduir determinades situacions. Llavors s'utilitza el procediment de la transformació inversa. Es parteix d'un valor *u* entre 0 i 1 agafant d'un generador de números pseudo-aleatoris que segueixen una distribució U[0, 1].

Partint de $F_X(x) = 1 - e^{-\lambda \cdot x}$;
agafem un valor "*u*", uniforme en [0, 1], busquem a que "x" li correspon segons $F_X(x)$: $u = 1 - e^{-\lambda \cdot x}$;
$e^{-\lambda \cdot x} = 1 - u$; $-\lambda \cdot x = \ln(1-u)$ $\quad x = -\ln(1-u)/\lambda$
(o bé $-\ln(u)/\lambda$)



EXEMPLE							
u:	0,442	0,4248	0,9177	0,0317	0,5943	0,1738	
x:	0,8165	0,8562	0,0859	3,4513	0,5203	1,7497	
							$\lambda=1$
u:	0,8344	0,2286	0,8542	0,0947	0,4288	1,1036	
x:	0,7244	5,9039	0,6305	9,4274	3,3874	9,0682	$\lambda=1/4$

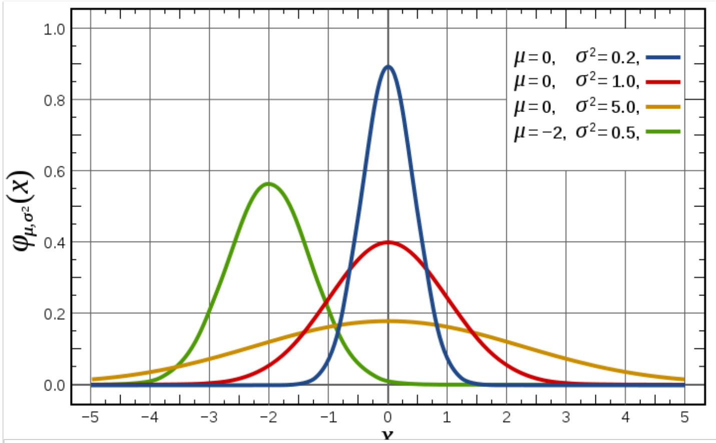
El model NORMAL (de Gauss-Laplace): $X \sim N(\mu, \sigma)$

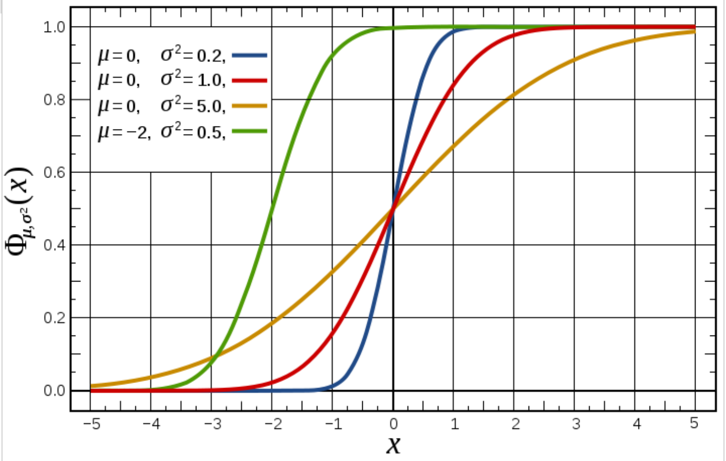
(Wikipedia.org) *Normal distribution:*

"the **normal** (or **Gaussian**) **distribution**, is a continuous probability distribution that is often used as a first approximation to describe real-valued random variables that tend to cluster around a single mean value"

"the normal distribution is commonly encountered in practice, and is used throughout statistics, natural sciences, and social sciences"

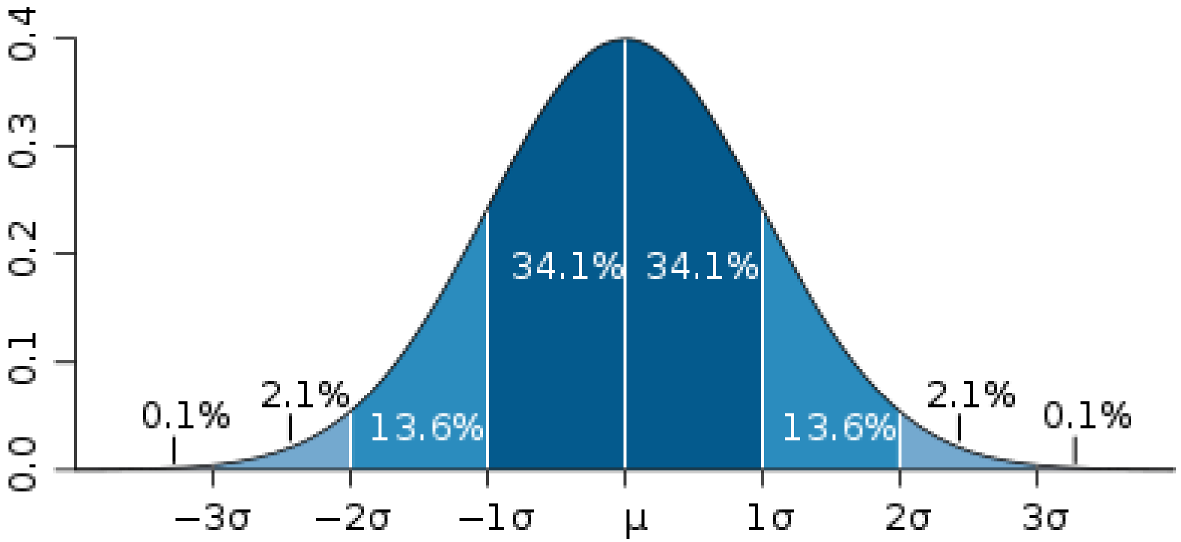
"The normal distribution is usually denoted by $N(\mu, \sigma^2)$"
"The distribution with $\mu = 0$ and $\sigma^2 = 1$ is called the **standard normal**", $Z \sim N(0,1)$





Veurem característiques i propietats del model Normal que fan que sigui molt utilitzat per modelar, i com a base de les tècniques d'inferència (amb la implicació de la necessitat de suposar i/o analitzar la normalitat com a mínim gràficament)

“Function $f(x)$ is symmetric around the point $x = \mu$, which is at the same time the mode, the median and the mean of the distribution”
“The inflection points of the curve occur one standard deviation away from the mean (i.e., at $x = \mu - \sigma$ and $x = \mu + \sigma$)”



“Quantiles of the standard normal distribution (Z) are commonly denoted as z_p.
The quantile z_p represents such a value that a standard normal random variable Z has the probability of exactly p to fall inside the $(-\infty, z_p]$ interval” (se usará en inferencia)

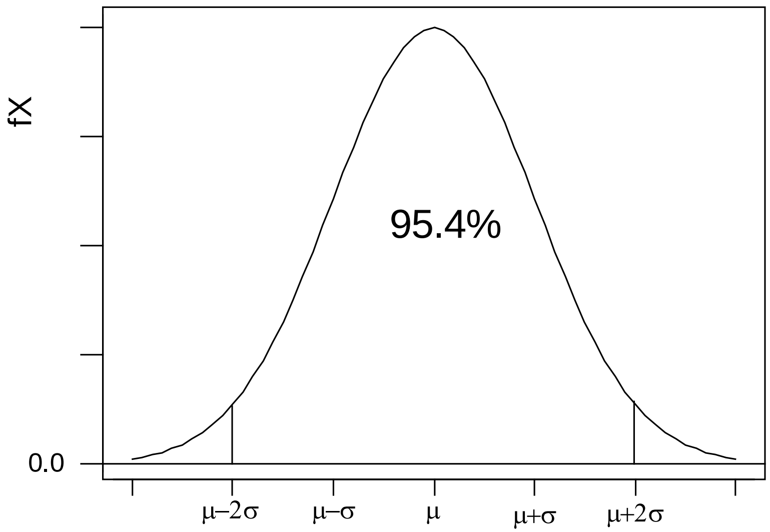
“The most “famous” standard normal quantile is $1.96 = z_{0.975}$. A standard normal random variable Z is greater than 1.96 in absolute value in only 5% of cases”

La funció de densitat d'una variable *X*, normal, amb mitjana μ i desviació típica σ és:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}$$

estenent-se per tota la recta real.

El seu aspecte és:



A la pràctica, *X* es concentra molt a prop de la mitjana: amb un 95.4% de probabilitat, un valor al atzar no estarà més lluny de 2σ de la mitjana μ. No és estrany utilitzar la llei Normal per modelar variables estrictament positives (per exemple, en biologia).

La funció de distribució de *X* no té forma algebraica.

Per probabilitats acumulades serà necessari utilitzar taules o R. Les taules recullen $F_Z(z)$, amb $Z \sim N(0,1)$ (*Normal estandarditzada = centrada i reduïda)*

Els indicadors d'esperança i variància són respectivament:

$$E(X)=\mu_X=\mu \quad \text{i} \quad V(X)= \sigma_X^2= \sigma_X^2$$

Algunes característiques dels anteriors models

Distribució	*Declaració*	**Funció de probabilitat o de densitat**	**Funció de distribució**	**Esperança** ($E(X) = \mu_X$)	**Variància** ($V(X) = \sigma_X^{\,2}$)
Bernoulli	X~Bern(p)			p	p q
Geomètrica	X~Geom(p)				
Binomial	**X~B(n,p)**	$P_X(k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}$ **o R:dbinom(k,n,p)**	$F_X(k) = \sum_{i<=k} P_X(i)$ o taules estadístiques **o R:pbinom(k,n,p)**	**n p**	**n p q**
Binomial negativa	X~BN(r,p)				
Poisson	**X~P(λ)**	$P_X(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$ **o R:dpois(k,λ)**	$F_X(k) = \sum_{i<=k} P_X(i)$ o taules estadístiques **o R:ppois(k,λ)**	**λ**	**λ**
Exponencial	**X~Exp(λ)**	$f_X(x) = \lambda \cdot e^{-\lambda \cdot x}$ **o R:dexp(x,λ)**	$F_X(x) = 1 - e^{-\lambda \cdot x}$ **o R:pexp(x,λ)**	**1/λ**	**1/λ²**
Uniforme	X~U[a,b]				
Normal	**X~N(μ,σ)**	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)2 / 2\sigma^2)$ **o R:dnorm(x,μ,σ)**	$F_X(x) = ?$ **taules estadístiques N(0,1)** **o R:pnorm(x,μ,σ)**	**μ**	**σ**

$0 < p < 1$; $q = 1 - p$; n enter > 0; r enter > 0; λ real > 0.0 (paràmetre procés de Poisson)

Propietat. Combinació lineal de variables normals

- Si $X \sim N(\mu_1, \sigma_1^2)$, a, b escalars: $aX + b \sim N(a\mu_1 + b, a^2\sigma_1^2)$
- Si $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, a, b escalars:

$$aX + bY \sim N(a\mu_1 + b\mu_2, \underbrace{a^2\sigma_1^2 + b^2\sigma_2^2 + 2\,ab\,\rho_{XY}\,\sigma_1\,\sigma_2}_{Variància})\ ^{\dagger}$$

Aquesta propietat assegura que la combinació de Normals segueixen sent Normal, i permet relacionar distribucions Normals a base de translacions i escalars.
En particular transformar a la Normal estàndard Z~N(0,1), "estandarditzar", permet buscar en les taules de Z, probabilitats de qualsevol llei Normal

Amb $X \sim N(\mu, \sigma)$ i $Z \sim N(0, 1)$ podem relacionar:

- $Z = X/\sigma - \mu/\sigma = (X - \mu) / \sigma \sim N(0, 1)$ {a=1/σ, b= $-\mu/\sigma$ són escalars}

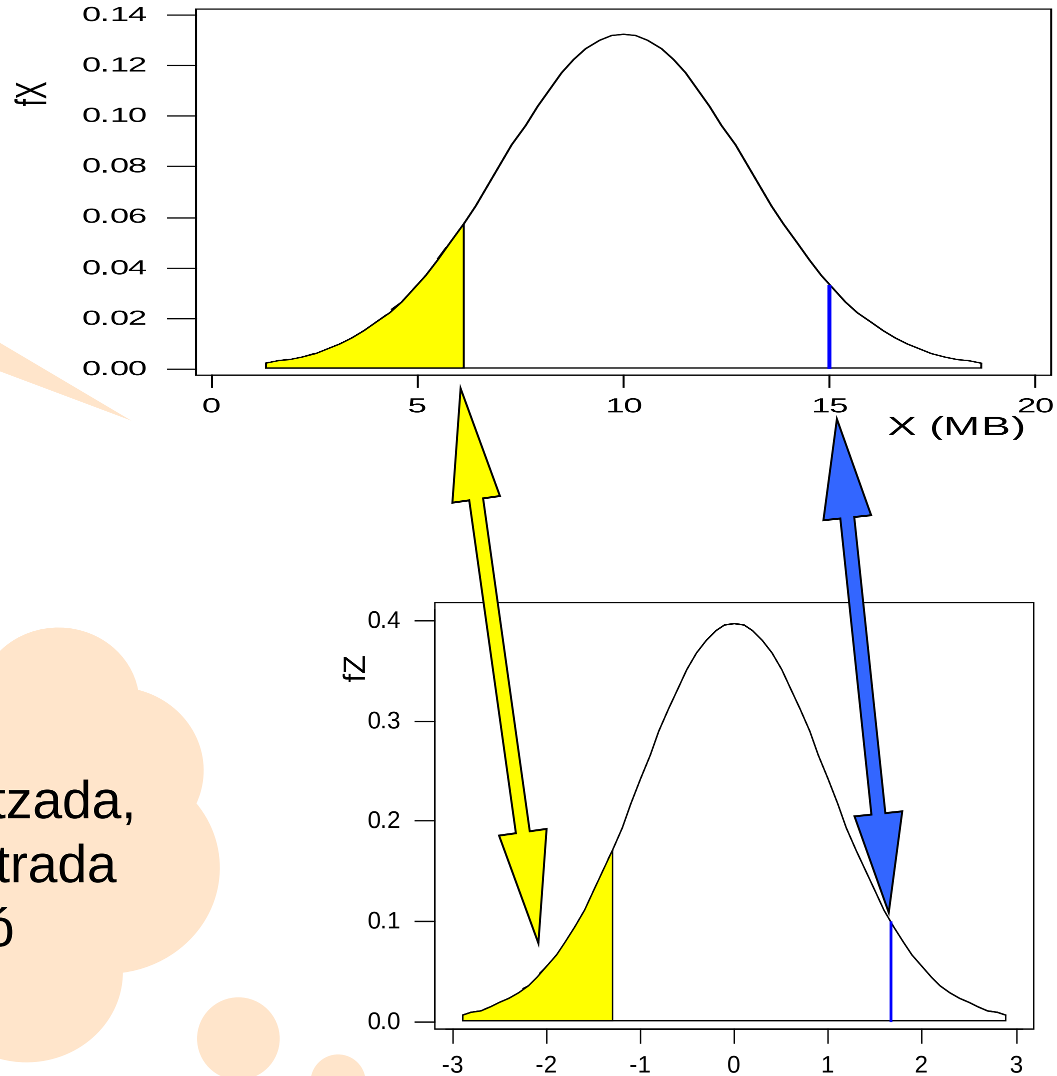
és a dir $Z = \dfrac{X - \mu}{\sigma}$

- O bé, $X = \mu + \sigma \cdot Z$

Estandaritzar.

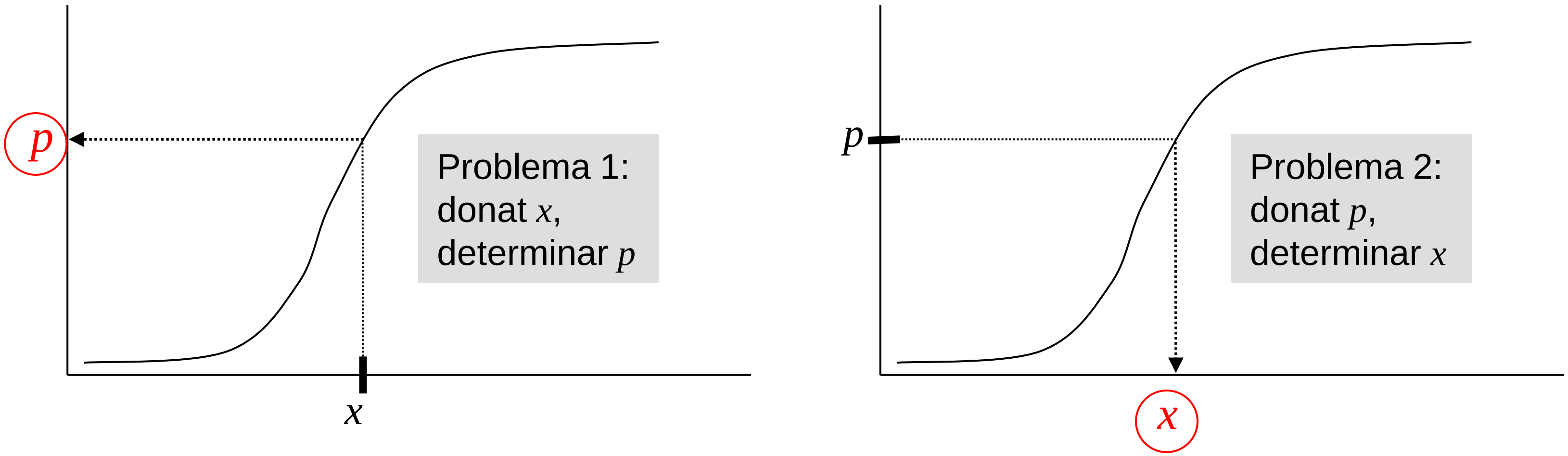
Variable *X*:
situació real
(unitats reals,
per exemple MB)

$$Z = \frac{X-\mu}{\sigma}$$



Variable *Z*:
situació estandarditzada,
sense unitats, centrada
en 0, dispersió
tipificada

Problemes típics:



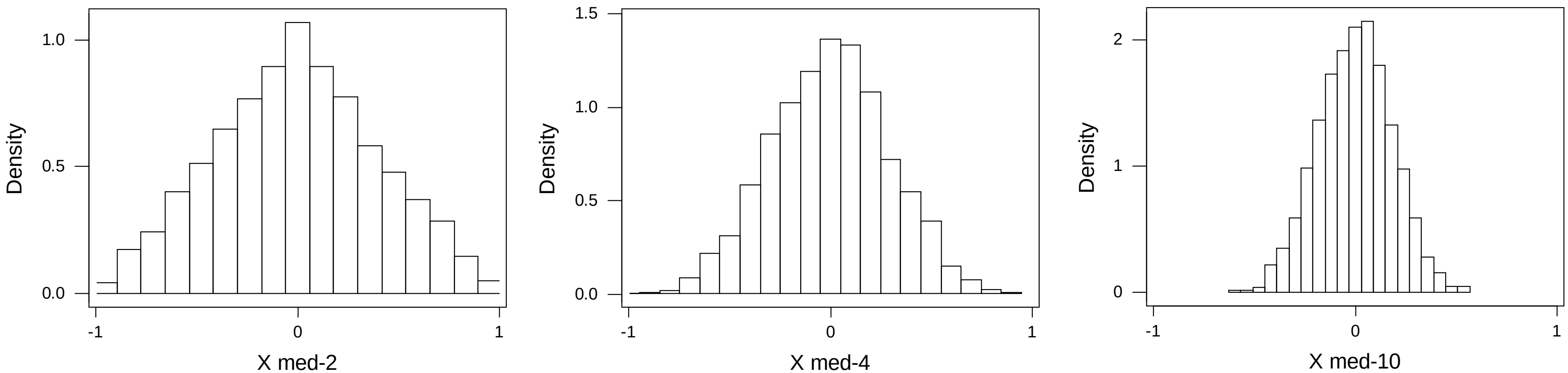
EXEMPLE 1. W ~ N(0, 1). w per p=0.25?
Com 0.25 < 0.5, no ve en taules, utilitzar simetries:
P(Z < -|z|) =

EXEMPLE 2. *X*: "Increment diari espai disc" ~ N(10 MB, 3 MB)

♣ P(*X*>15)? Estandaritzant, *Z* = (*X* – 10)/3 ~ N(0, 1), i P(*X*>15) =. . .

♣ 1 de cada 10 dies, l'augment és inferior a, quant? (*t* : P(*X* < *t*) = 0.1?)
t =

La llei de probabilitat de la *mitjana* de distribucions aleatòries



Hem simulat $(X_1 + X_2 + \ldots + X_n)/n$ (X_i, i.i.d.), i observem que:

- ♦ tendeix a concentrar-se al voltant de μ quan n augmenta, i
- ♦ tendeix a assemblar-se a una normal a mesura que n es fa gran.

$$E\left(\overline{X}_n\right) = \frac{E\left(\sum X_i\right)}{n} = \frac{\sum E(X_i)}{n} = \frac{n\mu}{n} = \mu \qquad V\left(\overline{X}_n\right) = \frac{V\left(\sum X_i\right)}{n^2} \underset{¡!}{=} \frac{\sum V(X_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Per qualsevol n, l’esperança de la mitjana és μ; la variància decreix amb n: amb una mostra gran, utilitzant la mitjana mostral ens aproximem més a μ.

$$\boxed{\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) = 0}$$

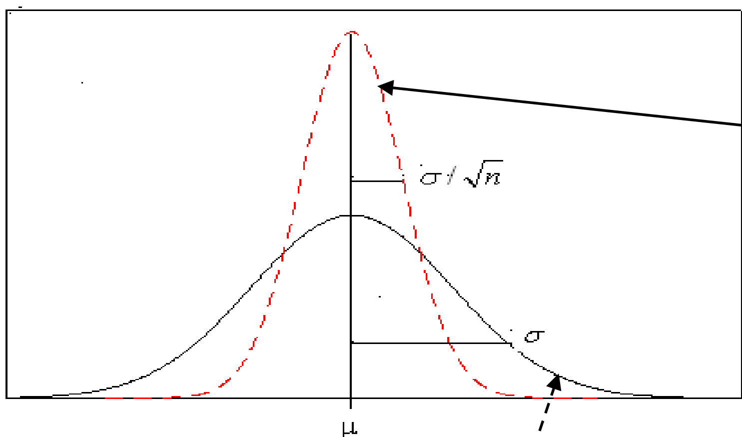
TEOREMA CENTRAL DEL LÍMIT: TCL

$X_1, X_2, \ldots, X_n$ indep., amb esperança μ i desviació típica σ.

En el límit , $n \to \infty$, la funció de distribució de la variable:

$$\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} = \frac{\sum X_i - n\mu}{\sigma\sqrt{n}}$$

tendeix a la funció de distribució d'una N(0,1).

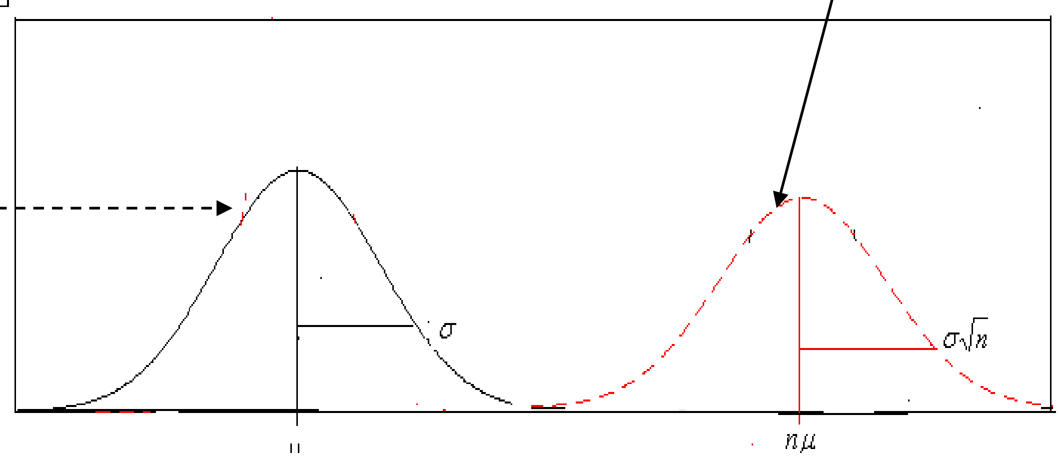


$$\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}_n \approx N(\mu, \sigma/\sqrt{n})$$

$$\sum_{i=1}^{n} X_i \approx N(n\mu, \sigma\sqrt{n})$$

X_i

(X_i no té perquè ser normal, sols complir: $E(X_i) = \mu, V(X_i) = \sigma^2$)



EXERCICI**.** El treball de CPU per un procés backup presenta unes característiques diàries de mitjana 30’/dia i una desviació de 15’/dia.
Si volem calcular probabilitats sobre el consum de CPU total mensual (suposant independència entre els 30 dies del mes), haurem de plantejar-nos la variable:
$X = X_1+X_2+ \ldots +X_{30}$. (no ens donen la distrib. de X_i, sols μ i σ)

Quina és la llei de X?:

I la probabilitat d’un consum total mensual de més de 18h de CPU?:

$P(X > 18) =$

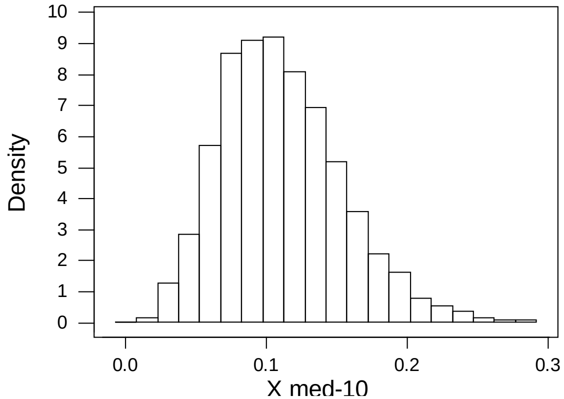
I la probabilitat de, en un mes, un consum mitjà diari inferior a 36’?:

Quan n és *suficientment gran* per aplicar el TCL?

Depèn de com sigui la llei original i de que es desitgi calcular.

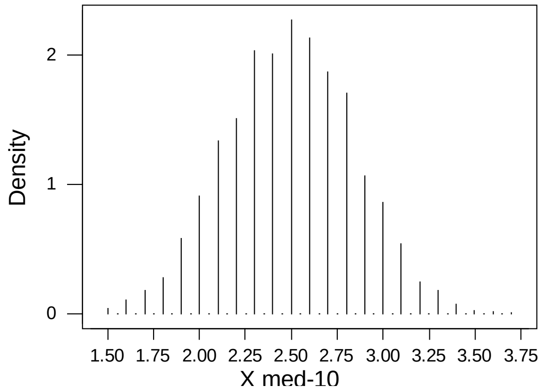
La convergència a la normal és més lenta si:

la distribució de les X_i és poc simètrica.



Distribució continua asimètrica, n=10

les X_i són variables discretes (especialment si agafen pocs valors)



Uniforme discreta {1, 2, 3, 4}, n=10

Altres aplicacions: la normal aproxima bé certes distribucions.
Exemple: variable de Poisson, si λ és gran (>5, segons Peña)

[la t de Student, i la χ^2, que es veuran més endavant, també]

EXERCICI. L'error de mesura del temps d'un procés és normal amb σ=1/4 seg. i mitjana 0. Es considera repetir les mesures de forma independent.

- Una variable convenient:
- Llei de probabilitat:
- Prob. error menor de 0.1 s.:
- Error màxim (prob 0.95) en una mesura:
- Idem amb promig de 10 mesures:
- Número mínim de mesures tal que l'error màxim (prob 0.95) sigui inferior a 0.1 seg.:

- E_n: "error en la n-èssima mesura"
- $E_n \sim N(\mu= \quad , \sigma = \quad)$
- $P(| E_1 | \leq 0.1) = P(-0.1 \leq E_1 \leq 0.1) = . .$
- $P(-e \leq E_1 \leq e) = 0.95;$

  $e =$

$X_n = (\Sigma_{i=1,n} E_i)/n \sim N(\mu= \quad , \sigma = \qquad)$

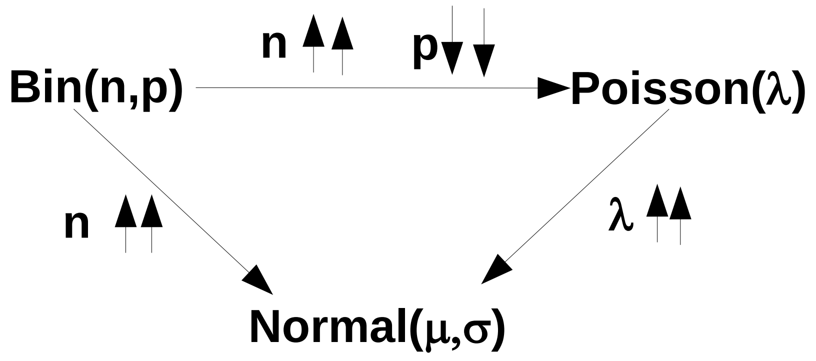
  » $P(-x \leq X_{10} \leq x) = 0.95$

  » $x =$

- n? tq $P(-0.1 \leq X_n \leq 0.1) \geq 0.95$

  »

TCL: aproximació entre Binomial, Poisson i Normal



http://www.wikipedia.org/ **"Normal distribution" (4.4) :**

The practical importance of the central limit theorem is that the normal distribution can be used as an approximation to some other distributions:

A binomial distribution with parameters n and p is approximately normal for large n and p not too close to 1 or 0 (the approximating normal distribution has mean $\mu = np$ and variance $\sigma^2 = np(1 - p)$)

A Poisson distribution with parameter λ is approximately normal for large λ. The approximating normal distribution has mean $\mu = \lambda$ and variance $\sigma^2 = \lambda$.

EXERCICI: Bin, P i Normal

El control de qualitat del temps de resposta d'una determinada pàgina web ha comprovat que la probabilitat que el servei al llarg d'un dia sigui inadequat és d'un 5%. Per calcular probabilitats del número de dies en que el servei és inadequat durant una setmana, un mes (30 dies) o cinc anys (1825 dies), necessitem les següents variables aleatòries per les quals podem identificar el model més adequat:

Xdia="nombre de dies en 1 setmana amb servei inadequat"   Xdia ~

Xmes="nombre de dies en 1 mes amb servei inadequat "   Xmes ~

X5anys="nombre de dies en 5 anys amb servei inadequat   X5anys ~

EXERCICI: Airport IV (e-status)

Per una altra part, anem a suposar que s’ha establert que el pes del equipatge d’un viatger segueix una distribució Normal amb mitjana 19.7 Kg. i desviació 2.85 Kg.

Habitualment, si l’equipatge d’un viatger sobrepassa els 20 Kg., llavors té un sobrepreu que depèn de l'excés de pes. En les següents preguntes anem a explorar temes relacionats amb la distribució d’un equipatge al atzar, d’un grup d’equipatges conduït per un carro de transport, i del pes promig d’un equipatge del carro.

Formulari : **Distribucions de variables discretes i contínues**

Distribució	Declaració	Funció de probabilitat o de densitat	Funció distribució $F_X(k)=\sum_{i<=k} P_X(i)$ o $\int_{-\infty}^{k} f_X(x)dx$	Esperança $\mathrm{E}(X)$	Variància $\mathrm{V}(X)$
Bernoulli	X~Bern(p)	$P_X(k)=\begin{cases} q & k=0 \\ p & k=1 \end{cases}$	$F_X(k)=\sum_{i<=k} P_X(i)$	p	$p\cdot q$
Binomial	X~B(n,p)	$P_X(k)=\binom{n}{k}\cdot p^k\cdot q^{n-k}$ $k=0,1,\ldots,n$ **(R:dbinom(k,n,p))**	$F_X(k)=\sum_{i<=k} P_X(i)$ (taules estadístiques) **(R:pbinom(k,n,p))**	$p\cdot n$	$p\cdot q\cdot n$
Poisson	X~P(λ) *	$P_X(k)=\frac{\lambda^k\cdot e^{-\lambda}}{k!}$ $k=0,1,2,\ldots$ **(R:dpois(k,λ))**	$F_X(k)=\sum_{i<=k} P_X(i)$ (taules estadístiques) **(R:ppois(k,λ))**	λ	λ
Exponencial	X~Exp(λ) *	$f_X(x)=\lambda\cdot e^{-\lambda\cdot x}$ $x>0$ **(R:dexp(x,λ))**	$F_X(x)=1-e^{-\lambda\cdot x}$ **(R:pexp(x,λ))**	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Uniforme	X~U[a,b]	$f_X(x)=\frac{1}{b-a}$ $a<x<b$ **(R:dunif(k,a,b))**	$F_X(x)=\frac{x-a}{b-a}$ **(R:punif(k,a,b))**	$\frac{(a+b)}{2}$	$(b-a)^2/12$
Normal	X~N(μ,σ)	$f_X(x)=\frac{1}{\sigma\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$ **(R:dnorm(k,μ,σ))**	$F_X(x)=?$ (taules estadístiques N(0,1)) **(R:pnorm(k,μ,σ))**	μ	σ^2

0 < p < 1; q = 1 - p; n, r enter > 0; λ, a, b, μ, σ real > 0.0
*λ paràmetre del procés Poisson: variables Poisson i Exponencial