Departament d'Estadística
i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Models de variable aleatòria (VAC i VAD)

Bloc 3 – Probabilitat i Estadística

Setembre 2019

Índex

Models de VAD i VAC

- **Definició Wikipedia**: *List of probability distributions: "Many probability distributions are so important in theory or applications that they have been given specific names".*
- **VAD – VA Discretes**: Binomial, Poisson, Bernoulli, Geomètrica, Binomial Negativa
- **VAC – VA Contínues**: Exponencial, Normal, Uniforme
- A partir dels paràmetres de cada model es calculen **indicadors**
  - Esperança → E(X) = μ_X
  - Variància → V(X) = σ_X^2
- A partir de les **funcions de probabilitat i distribució** de probabilitat es calculen probabilitats:

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

- A més, es calcularan inverses (donada una probabilitat α, calcular el **quantil α**, o **percentil α en %**):
  - x_α és el quantil α de X si es compleix: $F_X(x_\alpha) = \alpha$ $(0 \leq \alpha \leq 1)$

Model de Bernoulli

- **Definició**: Número d'èxits en la realització d'un únic experiment amb 2 possibles resultats: **0** ("no èxit") i **1** ("èxit")
- **Notació**: X~Bern(p)
- **Paràmetres**: p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat**:

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

Els valors "0" i "1" poden tenir 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*,...

- És el model teòric general més senzill aplicable a una variable aleatòria. El cas més habitual són experiències aleatòries que impliquen repeticions de proves Bernoulli. És necessari que unes proves siguin ***independents*** d'altres i que la **probabilitat d'èxit sigui constant** i igual a *p*

Models associats a la Bernoulli

- En una experiència aleatòria que implica repetició de proves Bernoulli independents, es plantegen com a distribucions interessants les següents **VAD**:
  - **Binomial:** sobre n repeticions, número "d'èxits" totals → **VAD**
  - **Geomètrica:** número de repeticions fins observar el primer "èxit" → **VAD**
  - **Binomial negativa:** número de repeticions fins observar el r-èssim "èxit" → **VAD**

- En experiències aleatòries on el número de repeticions *n* és molt gran i *p* és un valor petit (fenòmens *estranys*), pot ser més fàcil identificar la mitjana (*np*) d'"èxits" (en l'interval-unitat) que explícitament el valor de *n* i *p*. En aquest cas es plantegen com distribucions interessants:
  - **Poisson:** número "d'èxits" en l'interval → **VAD**
  - **Exponencial:** temps entre "èxits" → **VAC**

- En aquests darrers casos parlem d'un procés de Poisson en el qual la VAD Poisson i la VAC Exponencial comparteixen un paràmetre o taxa que relaciona la mitjana d'èxits i el temps entre "èxits" (http://en.wikipedia.org/wiki/Poisson_process)

Model Binomial

- **Definició**: Número de èxits en la repetició de *n* proves de Bernoulli independents amb probabilitat constant *p*
- **Notació**: X~B(n,p)
- **Paràmetres**: n (nombre de repeticions), p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat**:

$$P(X=k)=\binom{n}{k}\cdot p^k\cdot q^{n-k} \quad amb \quad k=0,1,\dots,n$$

on [q = 1 – p]

- No té **funció de distribució** analítica → S'utilitzen taules o R
- **Indicadors:**
  - $E(X) = n \cdot p$
  - $V(X) = n \cdot p \cdot q$

R: dbinom, pbinom, qbinom

Model Binomial. Explicació funció de probabilitat

Suposem la VAD X: "número de uns en 4 tirades d'un dau" ~ B(n=4, p=1/6)

$$P(X=k)=\binom{n}{k}\cdot p^k\cdot q^{n-k} \quad amb \quad k=0,1,\dots,n$$

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

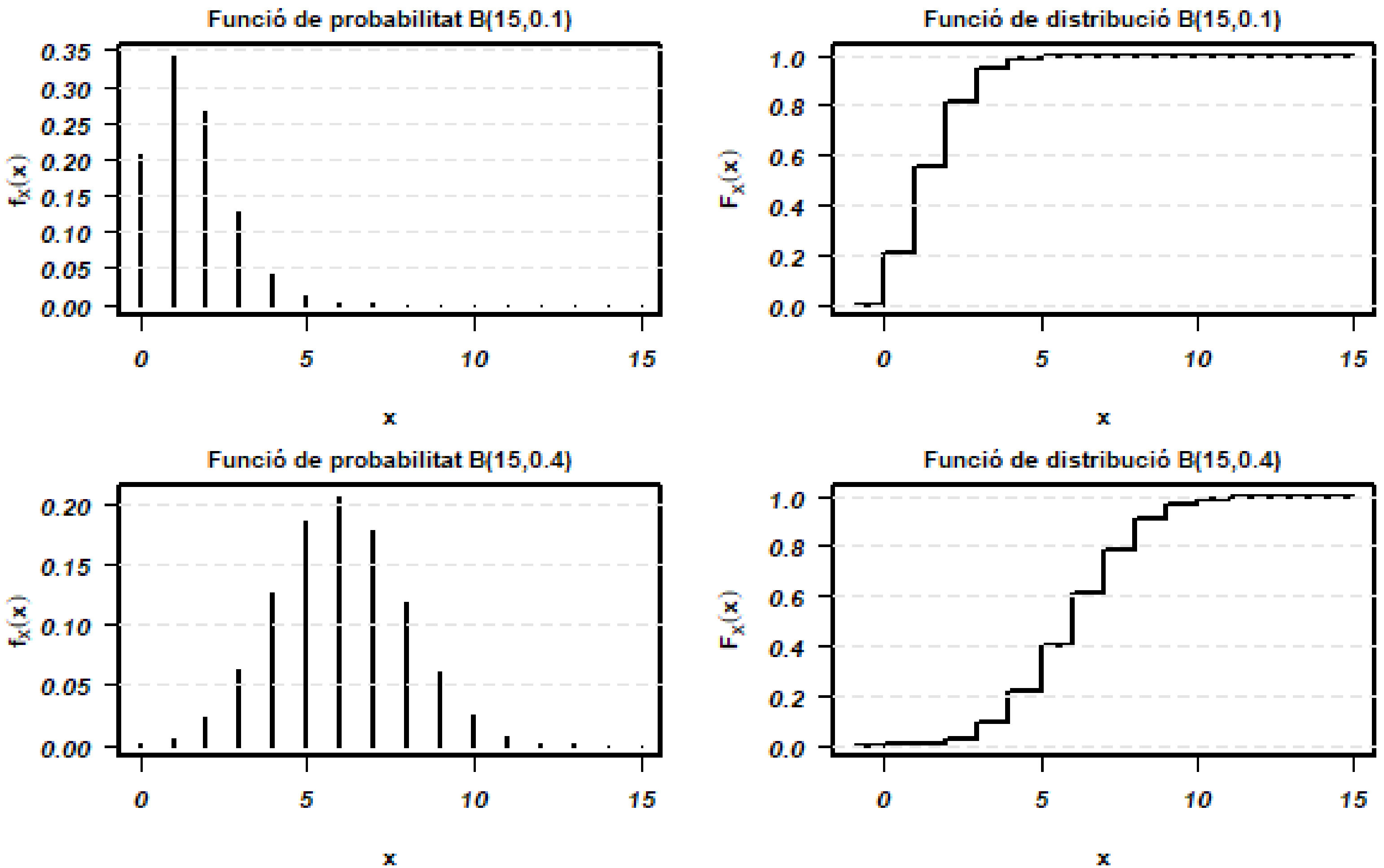
- Contarem 0 i 4, 1 vegada → $\binom{4}{0}=\binom{4}{4}=1$
- Contarem 1 i 3 → 4 vegades → $\binom{4}{1}=\binom{4}{3}=4$
- Contarem 2 → 6 vegades → $\binom{4}{2}=6$

- Els resultats no són equiprobables excepte si p=0.5:

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

Model Binomial. Representació gràfica

Ex: Com es distribueix el número de correus *spam* entre els 15 primers rebuts al dia segons si la probabilitat de que un correu sigui *spam* és 0.1 o 0.4?



Model Binomial. Ex. de càlcul de probabilitats

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

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

Veure vídeo sobre taules

- **Probabilitat puntual**. Quina és la probabilitat de 14?
  - Amb fórmules → $P(X=14) = \binom{20}{14} \cdot 0.5^{14} \cdot 0.5^{6} = \mathbf{0.037}$
  - Amb taules → $P(X=14) = P(X \leq 14) - P(X \leq 13) = \mathbf{0.037}$
  - Amb R → $P(X=14) = dbinom(x=14, size=20, prob=0.5) = \mathbf{0.03696442}$
- **Probabilitat acumulada**. Quina és la probabilitat de 14 o menys?
  - Amb fórmules → $P(X \leq 14) = \binom{20}{0} \cdot 0.5^{0} \cdot 0.5^{20} + \cdots + \binom{20}{14} \cdot 0.5^{14} \cdot 0.5^{6} = \mathbf{0.979}$
  - Amb taules → $P(X \leq 14) = \mathbf{0.979}$
  - Amb R → $P(X \leq 14) = pbinom(q=14, size=20, prob=0.5) = \mathbf{0.9793053}$
- **Quantils**. Quin és el valor tq la probabilitat de quedar per sota d'ell és almenys 0.95?
  - Amb fórmules → $P(X \leq x_{0.95}) = 0.95 \rightarrow Molt\ complicat!!!$
  - Amb taules → $P(X \leq x_{0.95}) = 0.95 \rightarrow x_{0.95} = \mathbf{14}\ ja\ que\ P(X \leq 14) = 0.9793 > 0.95$
  - Amb R → $P(X \leq x_{0.95}) = 0.95 \rightarrow qbinom(p=0.95, size=20, prob=0.5) = \mathbf{14}$

Nota: Les taules estadístiques contenen les probabilitats acumulades per a alguns valors dels paràmetres n ($2 \leq n \leq 20$) i p ($0.05 \leq p \leq 0.5$). Per valors de $p > 0.5$ es poden obtenir probabilitats acumulades amb una Binomial complementària:

$$P(X \leq k) = 1 - P(Y \leq n-k-1) \quad \text{on} \quad X \sim B(n, p) \quad \text{i} \quad Y \sim B(n, 1-p)$$

Model Binomial. Exercici

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

VAD que compta substitucions:

R: "número de discs que han de ser substituïts al cap de l'any"

Model per a la VAD:	R ~ B(n=18, p=0.2)
Número esperat de substitucions anuals:	E(R)=3.6
Variància i desviació típica:	V(R)=2.88 → $\sigma_x = 1.697$
Prob. d'observar 4 casos:	P(R=4)=0.2153
Prob. sols una substitució:	P(R=1) =0.0811
Prob. de tenir-ne menys de 3:	P(R<3)=0.2713
Prob. que almenys 6 discs siguin substituïts:	P(R≥6)=0.1329

Model Geomètric

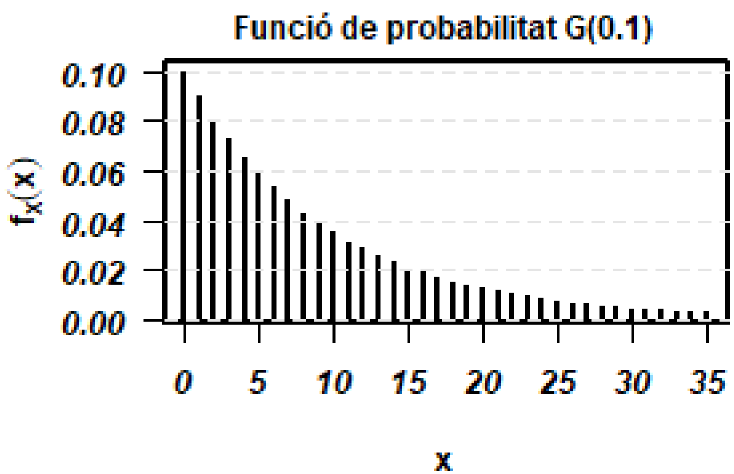
- **Definició**: Número d'intents (k) d'un experiment de Bernoulli fins observar el primer èxit
- **Notació**: X~Geom(p)
- **Paràmetres**: p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat** : $P(X=k)=q^{k-1}\cdot p$

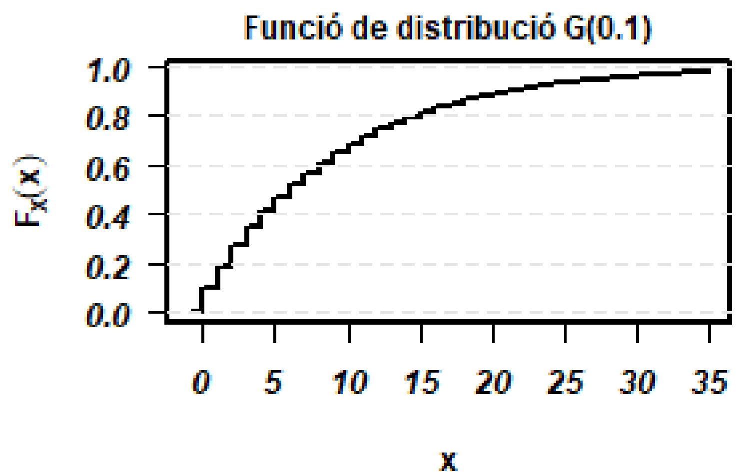
R: dgeom, pgeom, qgeom

Indicadors:

$E(X) = 1/p$

$V(X) = (1-p)/p^2$





Nota: no està acotada, qualsevol valor enter > 0 és possible [Ex. "tirar el dau moltes vegades fins que surti el primer 1"]. Encara que el més probable és que el número d'intents no sigui molt alt: quan *p* augmenta, $P_X(k)$ es trasllada a valors més baixos, i $F_X(k)$ creix més ràpid.

Model Geomètric. Exercici

La probabilitat d'un disc dur defectuós d'una determinada marca és 0.05. Es durà a terme una inspecció sobre un lot de discs durs.

Quina distribució segueix la VAD X: "Nombre de discs durs a revisar fins a trobar el primer defectuós"?:

X~Geom(p=0.05)

Quina és la probabilitat d'haver-ne de revisar només 1? P(X=1) = 0.05

Quina és la probabilitat d'haver-ne de revisar 10? P(X=10) = 0.959 · 0.05 = 0.032

Quina és la probabilitat d'haver-ne de revisar més de 2?

P(X>2) = 1 – P(X=1) – P(X=2) = 1 – 0.05 – 0-95*0.05 = 0.9025

Quin és el nombre esperat de discos que s'hauran de revisar? E(X) = 1/0.05 = 20

Amb quin nombre de discos en el lot podrem estar segurs que trobarem algun defectuós amb una probabilitat almenys del 50% P(X<x) > 0.50 → x= 13

```
qgeom(0.50,0.05)
```

Model Binomial negativa

- **Definició**: Número de repeticions (k) d'un experiment de Bernoulli fins observar *r* èxits
- **Notació**: X~BN(r , p)
- **Paràmetres**: r (nombre d'èxits), p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat**:

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

Indicadors:

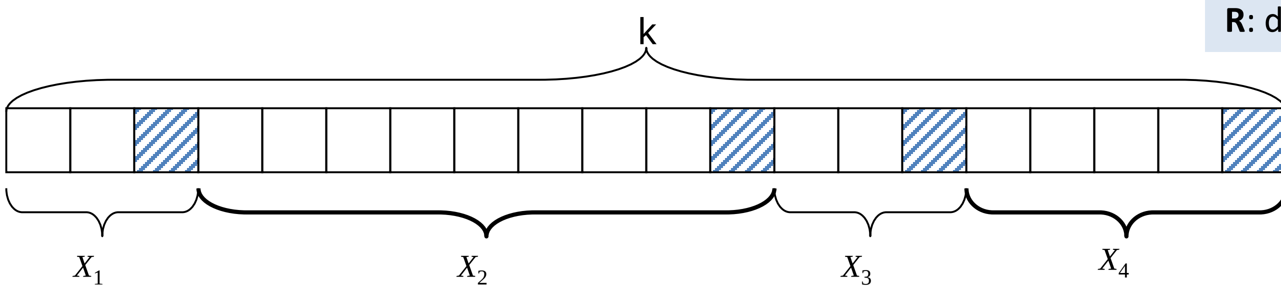
E(X) = r/p

V(X) = r(1-p)/p²

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 ≥ r és

- En general, podem pensar que una BN és una suma de r geomètriques independents.

R: dnbinom, pnbinom, qnbinom



- Ex: Quantes repeticions fan falta per aconseguir r èxits (r>1)? Si r=1, tenim la distribució geomètrica

Model Binomial negatiu. Exercici

Per aprovar un determinada assignatura, es requereix treure més d'un 5 en un problema d'estatus almenys 3 vegades. Un alumne concret té una probabilitat invariable de 0.6 de treure més d' un 5.

Quina distribució segueix la VAD X: "Nombre d'execucions per treure 3 notes superiors a 5"?:

X~BN(r = 3, p=0.6)

Quina és la probabilitat d'haver-ne de fer 1 execució?

P(X=1) = 0

Quina és la probabilitat d'haver-ne de fer 10 execucions?

$$P(X=10) = \binom{9}{2} \cdot 0.6^3 \cdot 0.4^7 = 0.013$$

Quin és el nombre esperat de execucions que s'hauran de fer per aprovar?

E(X) = 3/0.6 = 5

Model Poisson

- **Definició**: Número d'ocurrències favorables en un determinat interval de temps o espai
- **Notació**: X~P(λ)
- **Paràmetres**: λ (taxa de l'esdeveniment)
- **Funció de probabilitat**:

$$P(X=k)=\frac{e^{-\lambda}\cdot\lambda^{k}}{k!}\quad amb\quad k=0,1,\ldots$$

- **Indicadores:**
  - E(X) = λ
  - V(X) = λ
- **λ** és un número real positiu que representa la taxa mitjana d'ocurrències per unitat considerada

  [Ex: 10 trucades/hora]

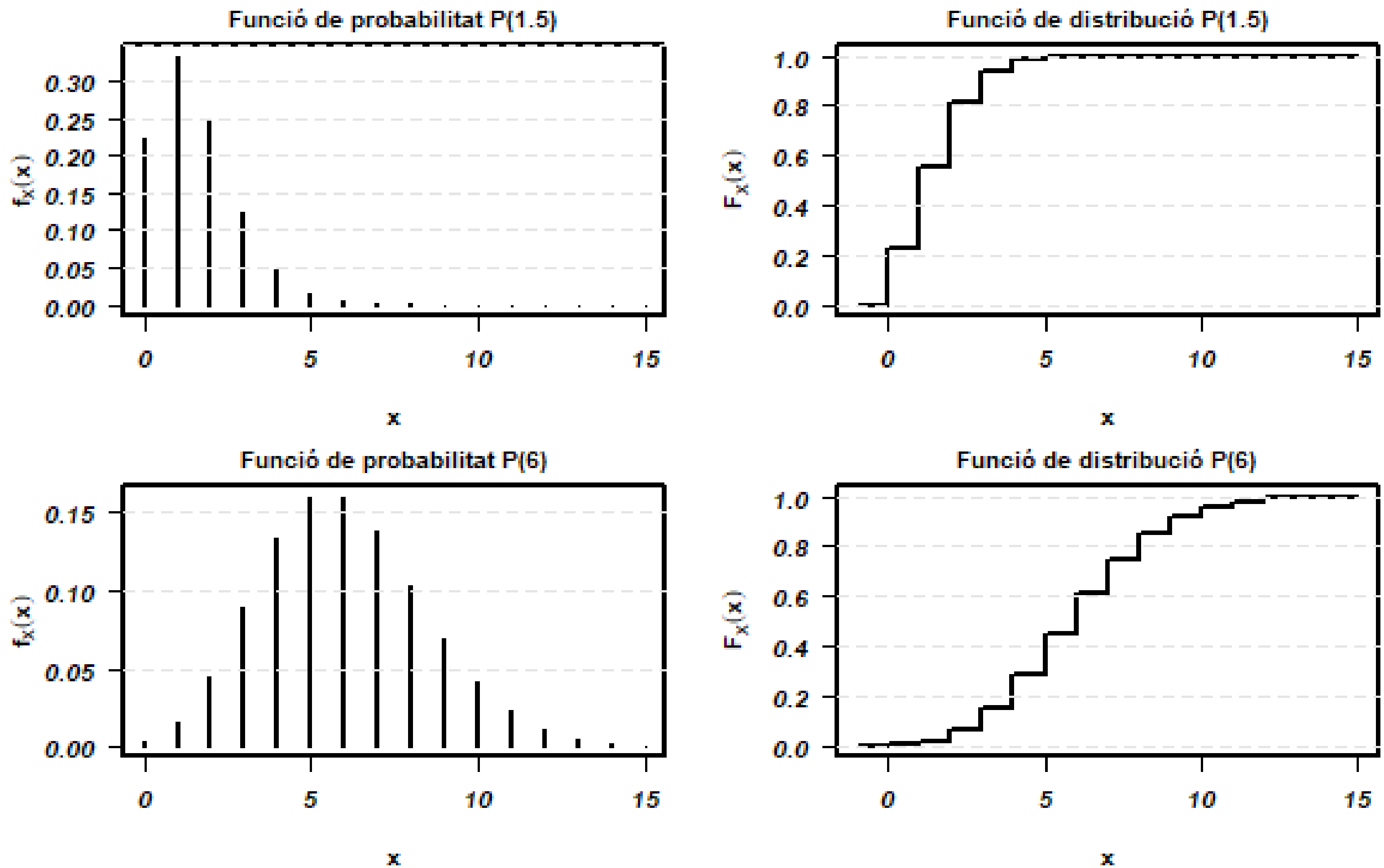
Al igual que en el model binomial pot haver-hi pròpiament una repetició d'experiències idèntiques tipus Bernoulli, però també pot correspondre a fenòmens que ocorren inesperadament (Ex: una trucada a una centraleta)

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

R: dpois, ppois, qpois

Model Poisson. Representació gràfica

EXEMPLE. Com es distribueix el número de correus *spam* rebuts al dia segons si el valor de la mitjana és 1.5 o 6?



Model Poisson. Ex. de càlcul de probabilitats

Sigui X ~ P(λ =2) i alguna de les prob. tabulades:

λ	0	1	2	3	4	5	6	7	8
2	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000

- **Probabilitat puntual**. Quina és la probabilitat de 3?
  - Amb fórmules → $P(X=3)=\frac{e^{-2}\cdot 2^3}{3!}=\mathbf{0.1804}$
  - Amb taules → $P(X=3)=P(X\leq 3)-P(X\leq 2)=\mathbf{0.18}$
  - Amb R → $P(X=3)=dpois(x=3,\text{lambda}=2)=\mathbf{0.1804470}$

- **Probabilitat acumulada**. Quina és la probabilitat de 3 o menys?
  - Amb fórmules → $P(X\leq 3)=\frac{e^{-2}\cdot 2^0}{0!}+\cdots+\frac{e^{-2}\cdot 2^3}{3!}=\mathbf{0.857}$
  - Amb taules → $P(X\leq 3)=\mathbf{0.857}$
  - Amb R → $P(X\leq 3)=ppois(q=3,\text{lambda}=2)=\mathbf{0.8571235}$

- **Quantils**. Quin és el valor tq la probabilitat de quedar per sota d'ell és almenys 0.95?
  - Amb fórmules → $P(X\leq x_{0.95})=0.95\rightarrow Molt\ complicat!!!$
  - Amb taules → $P(X\leq x_{0.95})=0.95\rightarrow x_{0.95}=\mathbf{5}\ ja\ que\ P(X\leq 5)=0.983>0.95$
  - Amb R → $P(X\leq x_{0.95})=0.95\rightarrow qpois(p=0.95, lambda=2)=\mathbf{5}$

Nota: A les taules estadístiques només trobarem les probabilitats acumulades per a alguns valors del paràmetre λ
Nota: La suma de VAD Poisson és una VAD Poisson amb paràmetre λ igual a la suma dels paràmetres de cadascuna: a partir d'una VAD Poisson, podem definir altres VAD aplicant proporcionalment al paràmetre el canvi en l'interval:

X="...en interval t" ~ P(λ) → *Y="...en interval kt"* ~ P(kλ)

Model Poisson. Exercici

El centre de càlcul d'una empresa atén les incidències que sorgeixen als treballadors. S'ha observat que aquestes apareixen esporàdicament, encara que l'elevat número d'usuaris implica que el volum de problemes a tractar diàriament sigui considerable (s'ha suposat una mitjana de 2.35 incidències/dia).

Quina és la distribució de la v.a "número d'incidències diàries"?

X: "número d'incidències al dia" → $X \sim P(\lambda=2.35)$

Quina és la seva Esperança? I la seva desviació típus?

$E(X) = 2.35$ incidències i $\sigma_X=1.53$ incidències

Quina és la distribució de la v.a "número d'incidències en 5 dies"? I en 7?

X_5: "número d'incidències 5 dies" → $X_5 \sim P(\lambda=11.75)$; $X_7 \sim P(\lambda=16.45)$

Probabilitat que en un dia es produeixi 3 incidències

$P(X=3) = 0.2063$

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

$P(X<3) = P(X\leq 2) = 0.5828$

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

$P(X_3\leq 15-6) = P(X_3\leq 9) = 0.8254$

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

$P(X_7=0) = 0.0000079$

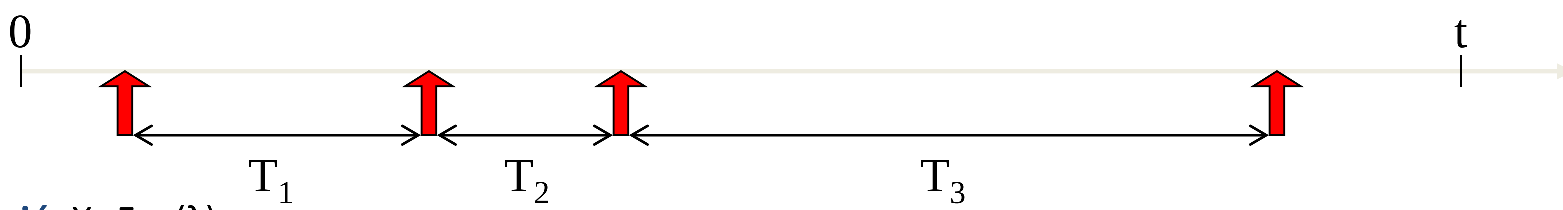
TAULA resum de models de VAD

Distribució	Declaració	Domini	Esperança $E(X) = \mu_X$	Variància $V(X) = \sigma_X{}^2$
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	$q \cdot r/p^2$
Poisson	P(λ)	0, 1, 2,...	λ	λ

$0 < p < 1$ $q = 1 - p$ $n \in N$ $r \in N$ $\lambda \in R^+$

Model Exponencial

- **Definició**: Distribució del temps entre arribades (ocurrències) en un procés de Poisson. És a dir, si a 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), llavors el temps entre dues arribades consecutives és una magnitud continua i indeterminista que es distribueix exponencialment. [Ex. d'aplicació: vida útil d'un component electrónic]



- **Notació**: $X \sim Exp(\lambda)$
- **Paràmetres**: λ (taxa d'aparició de l'esdeveniment per unitat de temps)
- **Funció de densitat i de distribució**:

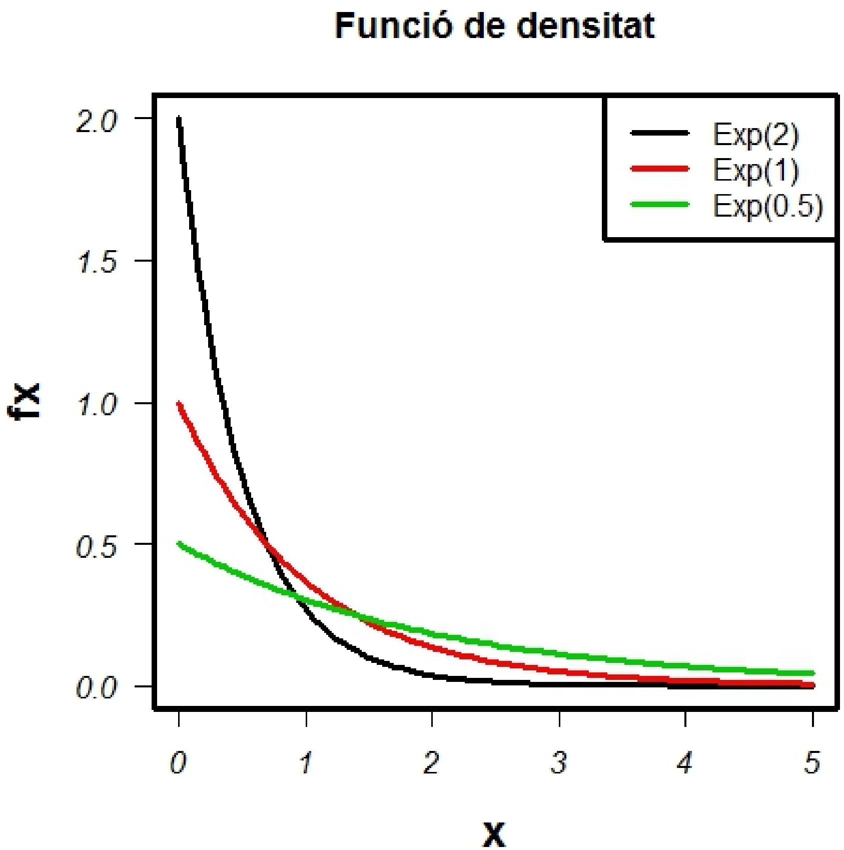
$$f_X(x) = \lambda \cdot e^{-\lambda x} \qquad amb \quad x > 0$$

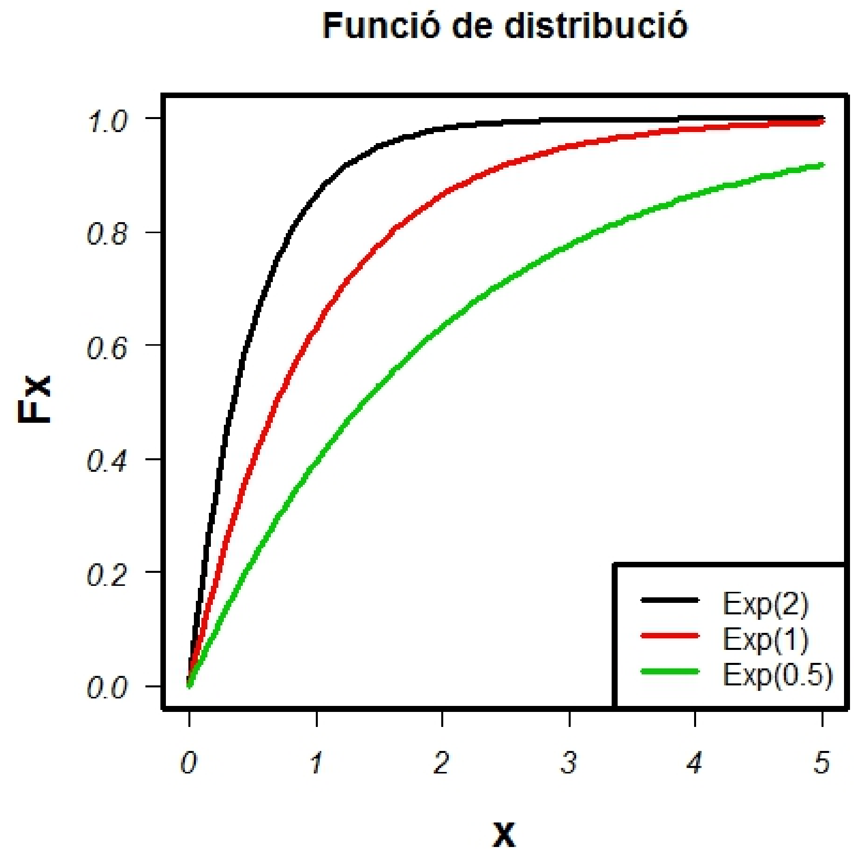
$$F_X(x) = 1 - e^{-\lambda x} \qquad amb \quad x > 0$$

- **Indicadors:**
  - $E(X) = 1/\lambda$
  - $V(X) = 1/\lambda^2$

Model Exponencial. Representació gràfica

Ex: Com es distribueix el temps entre correus *spam* si rebo un promig de 2 per hora? I si rebo 1 per hora? I si rebo 1 cada dues hores?





Model Exponencial. Ex. de càlcul de probabilitats

Sigui $X \sim \text{Exp}(\lambda = 2)$:

- **Probabilitat puntual**. → Recordeu que P(X=x) = 0 per qualsevol x ja que és una VAC

- **Probabilitat acumulada**. Quina és la probabilitat de 2 o menys?
  - Amb fórmules → $P(X \leq 2) = 1 - e^{-2\cdot 2} = 1 - e^{-4} = \mathbf{0.9817}$
  - Amb R → $P(X \leq 2) = pexp(q = 2, \text{rate} = 2) = \mathbf{0.9816844}$

- **Quantils**. Quin és el valor tq la probabilitat de quedar per sota d'ell és almenys 0.95?
  - Amb fórmules → $P(X \leq x_{0.95}) = 0.95 \rightarrow 1 - e^{-2\cdot x_{0.95}} = 0.95 \rightarrow x_{0.95} = \mathbf{1.497866}$
  - Amb R → $P(X \leq x_{0.95}) = 0.95 \rightarrow qexp(p = 0.95, rate = 2) = \mathbf{1.497866}$

La distribució Exponencial no té taules, perquè la funció de distribució té expressió analítica

Model Exponencial. Exemples

Siguin les següents variables aleatòries:

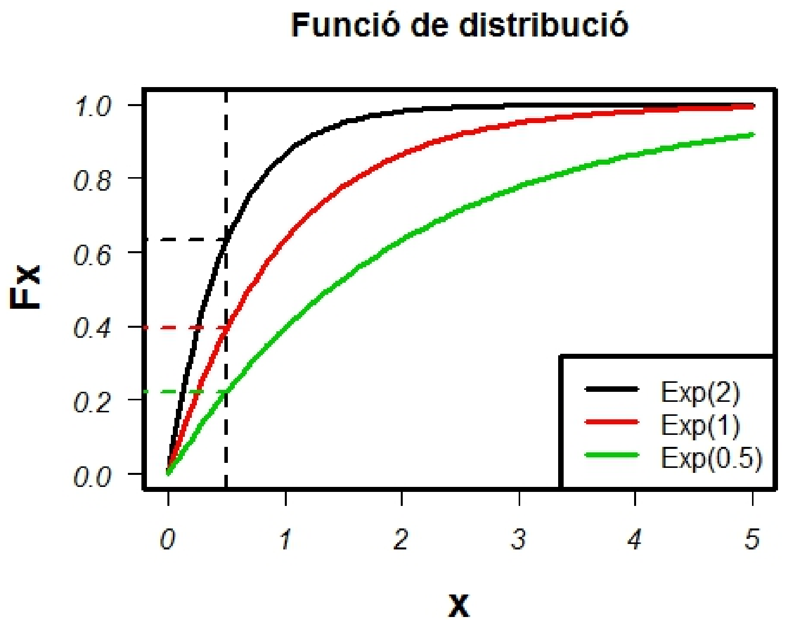
N: "número de peticions/seg a un servidor de BBDD" → N ~ P(λ)

T: temps (seg.) transcorregut entre dos peticions consecutives

Si λ = 2 arr./s. → P(T<0.5) = 0.632

Si λ = 1 arr./s. → P(T<0.5) = 0.393

Si λ = 0.5 arr./s. → P(T<0.5) = 0.221



Nota: 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*).

Model Exponencial. Observacions

- $f_X(x)$ no és $P(X=x)$ [$P(X=x) = 0$ per definició] → $f_X(x)$ no és una probabilitat, a diferencia de la $p_X(x)$ de les VAD
- Recordem que en una VAC:

$$P(a \leq X) = P(a < X) \quad \text{i} \quad P(a \leq X \leq b) = P(a < X < b) = F_X(b) - F_X(a)$$

- No tindrem taules estadístiques per probabilitats acumulades en el model Exp. Es calculen directament amb la funció de distribució de probabilitat:

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

- **Propietat de Markov (o de NO memòria):** La distribució de probabilitat d'una variable aleatòria Exponencial no depèn del que hagi passat amb anterioritat al moment present. Matemàticament:

$$P(T > t_1 | T > t_0) = P(T > t_1 - t_0) \quad per\ t_1 > t_0$$

Atenció:
$P(T>t_1 | T>t_0) \neq P(T>t_1)$

- Ex: En el servidor de BBDD, en un instant donat fa 10" que no arriben peticions. Que és més probable: (A) rebre en en els 10'' següents, o (B) rebre 10'' desprès d'una arribada?

Solució: Igual

Model Exponencial (i Poisson). Exercici

El centre de càlcul d'una important empresa atén les incidències que els sorgeixen als treballadors. Se suposa una mitjana de 4 incidències/dia, i 8 hores laborables al dia.

Quina és la distribució de la v.a "número d'incidències diàries"?

Quina és la distribució de la v.a "número d'incidències per hora"?

Probabilitat de rebre 0 incidències en un dia:

Probabilitat de rebre 0 incidències en una hora:

Quina és l'esperança de la variable temps (en hores) entre incidències?

I la desviació tipus?

Probabilitat d'estar 8 o més hores sense rebre incidències:

Model Exponencial. Aplicacions

- **"Failure Rate and Reliability"** (Jane Horgan lecture 17)

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

  - 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/λ és "MTTF" o "Mean Time To Failure"

- **"Modelling Response Times: M/M/1"** (Jane Horgan lecture 17)
  - 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 **P**(λ) i serveis **P**(μ)), que compleixen la propietat de Markov de no tenir memòria.

Amb l i m es defineix un indicador del sistema com factor de càrrega o "traffic intensity" (λ / μ):
- si λ /μ > 1 (λ > μ) → la taxa d'arribades és superior a la de sortides (sobreutilització del sistema)
- si λ /μ = 1 (λ = μ) →la taxa d'arribades s'iguala amb la taxa de sortides
- si λ /μ < 1 (λ < μ) → 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)

Model Exponencial (i Poisson). Exercici fiabilitat

El centre de càlcul d'una important empresa garanteix treballar amb una mitjana de 2 hores entre incidències.

Quina és la distribució de la v.a "temps en hores entre incidències"?

Quina és la funció de distribució?

Quina és la funció de fiabilitat?

Quina és la taxa d'errors?

Quin és el MTTF?

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

Model Exponencial (i Poisson). Exercici

Continuem analitzant el cas de l'aeroport, donant importància al procés d'arribades de passatgers als punts de facturació. Respon a les següents preguntes.

1. Un determinat punt de facturació es caracteritza perquè el número de viatgers que arriben per minut es distribueix segons una Poisson amb mitjana de 9.5. Calcula la probabilitat que aquest número sigui menor que 7 [0.165]
2. En aquest punt de facturació, quina és la probabilitat d'observar exactament 10 arribades en un minut? [0.123]
3. En el mateix punt de facturació, quina és l'esperança de la variable temps (*en segons*) entre dues arribades? [6.316 s.]
4. Al punt de facturació, quina és la probabilitat d'estar menys de 4 segons sense arribades? [0.469]
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? [0.501]

Model Uniforme

- **Definició**: VAC amb funció de densitat constant en un determinat rang [la probabilitat de pertànyer a un interval concret en aquest rang només depèn de la longitud de l'interval]
- **Notació**: X~U(a,b)
- **Paràmetres**: a (valor mínim del rang de X), b (valor màxim del rang de X)
- **Funció de densitat i distribució**:

Constant!!! → $$f_X(x) = \frac{1}{b-a} \qquad amb \quad a < x < b$$

$F_X(x) = 0$ si $x < a$
$F_X(x) = 1$ si $x > b$ → $$F_X(x) = \frac{x-a}{b-a} \qquad amb \quad a < x < b$$

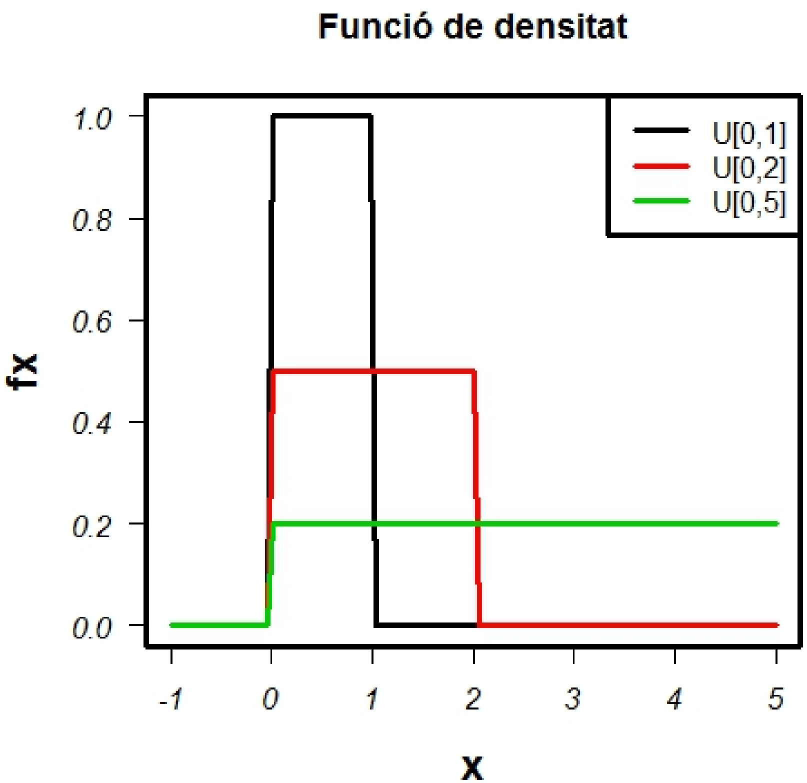
- **Indicadors:**
  - E(X) = (b + a)/2
  - V(X) = (b – a)²/12

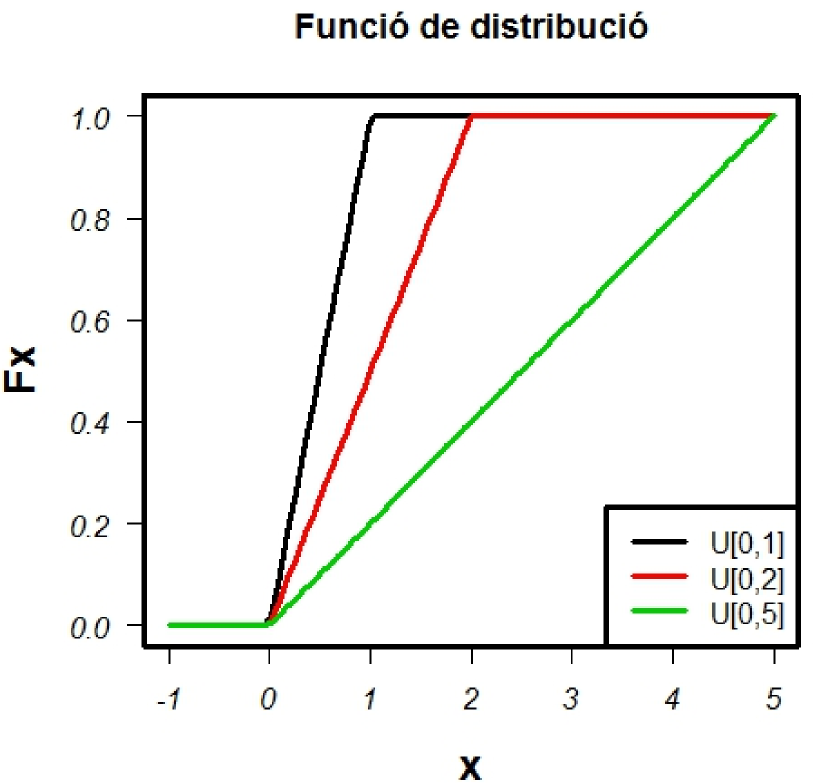
$$E(X) = \mu_X = \int_{-\infty}^{+\infty} x \cdot f_X(x)dx = \int_a^b x \cdot \frac{1}{b-a} dx =$$
$$= \frac{1}{b-a} \cdot \left.\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

Intuïtivament ja es veu que la mitjana ha de ser el centre de l'interval

Model Uniforme. Representació gràfica

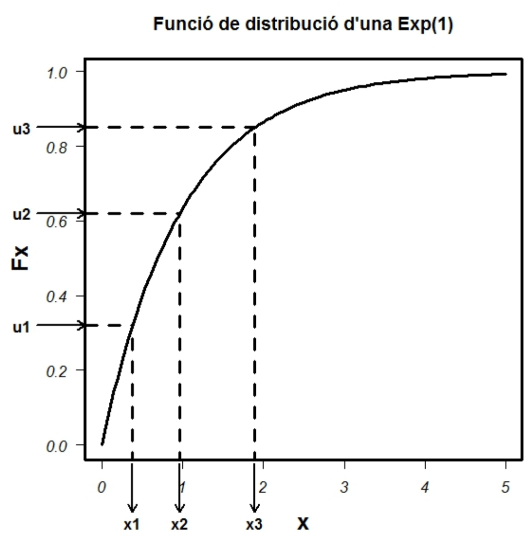
Ex: Com es distribueixen el nombres aleatoris entre 0 i 1? I entre 0 i 2? I entre 0 i 5?





Model Uniforme i Exponencial. Aplicació

- Generar números pseudo-aleatoris amb distribució U[0,1] és relativament complicat
- No obstant, generar números aleatoris de qualsevol distribució, un cop tenim els nombres pseudo-aleatoris amb distribució U[0,1] és senzill:
  - Per generar Y ~ U[a,b] → Y = a + (b-a)·u ~U[a,b] on u són valors d'una U[0,1]
  - Per generar T ~ exp(λ) → T = $F^{-1}(u) = -\ln(1-u)/\lambda$ ~exp(λ) on u són valors d'una U[0,1]
- En general, el mètode de la transformació inversa (que empra F^{-1}) permet generar valors de qualsevol distribució (no cal que F tingui expressió analítica)



Generar valos d'una exponencial:
- Es parteix de $F_X(x) = 1 - e^{-\lambda \cdot x}$
- S'agafen valors u_i~ U[0, 1]
- Es troben les x_i tals que $F_X(x_i) = u_i = 1 - e^{-\lambda \cdot x_i}$
- Els x_i segueixen una distribució Exponencial(λ)

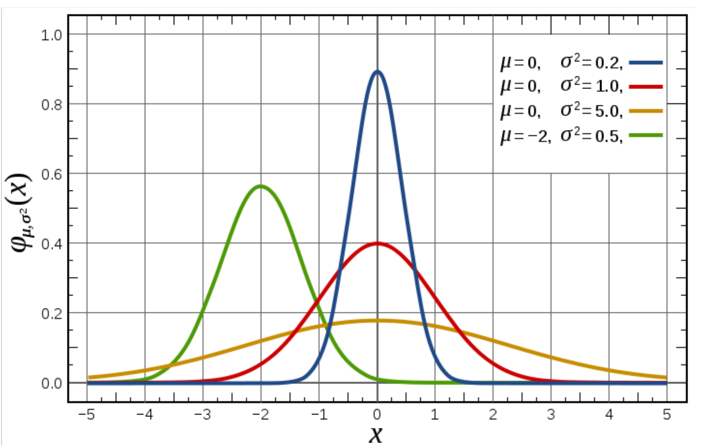
R: Per generar valors psudoaleatoris tenim funcions per a cada distribució: rbinom, rpois, runif, rexp ….

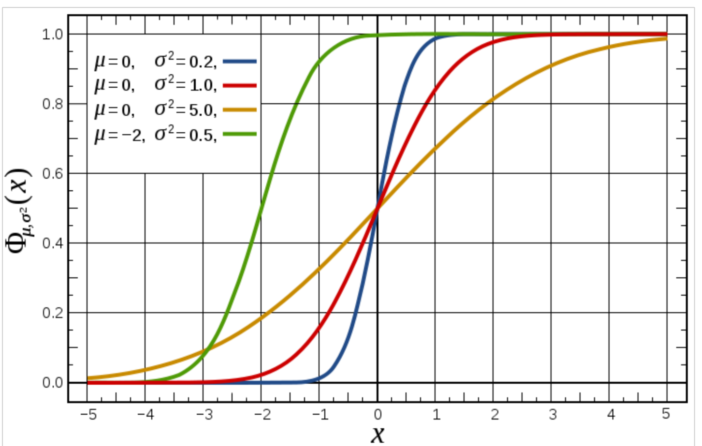
Model Normal (o de Gauss). Introducció

(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”





Model Normal

- **Definició**: Model que s'ajusta a valors provinents de múltiples fenòmens trobats en diferents disciplines científiques [Ex: alçades de persones, efecte d'un fàrmac, soroll en telecomunicacions...]
- **Notació**: X~N(μ , σ) [a vegades s'usa σ^2 en comptes de σ com a paràmetre]
- **Paràmetres**: μ (esperança), σ (desviació estàndard)
- **Funció de densitat**:

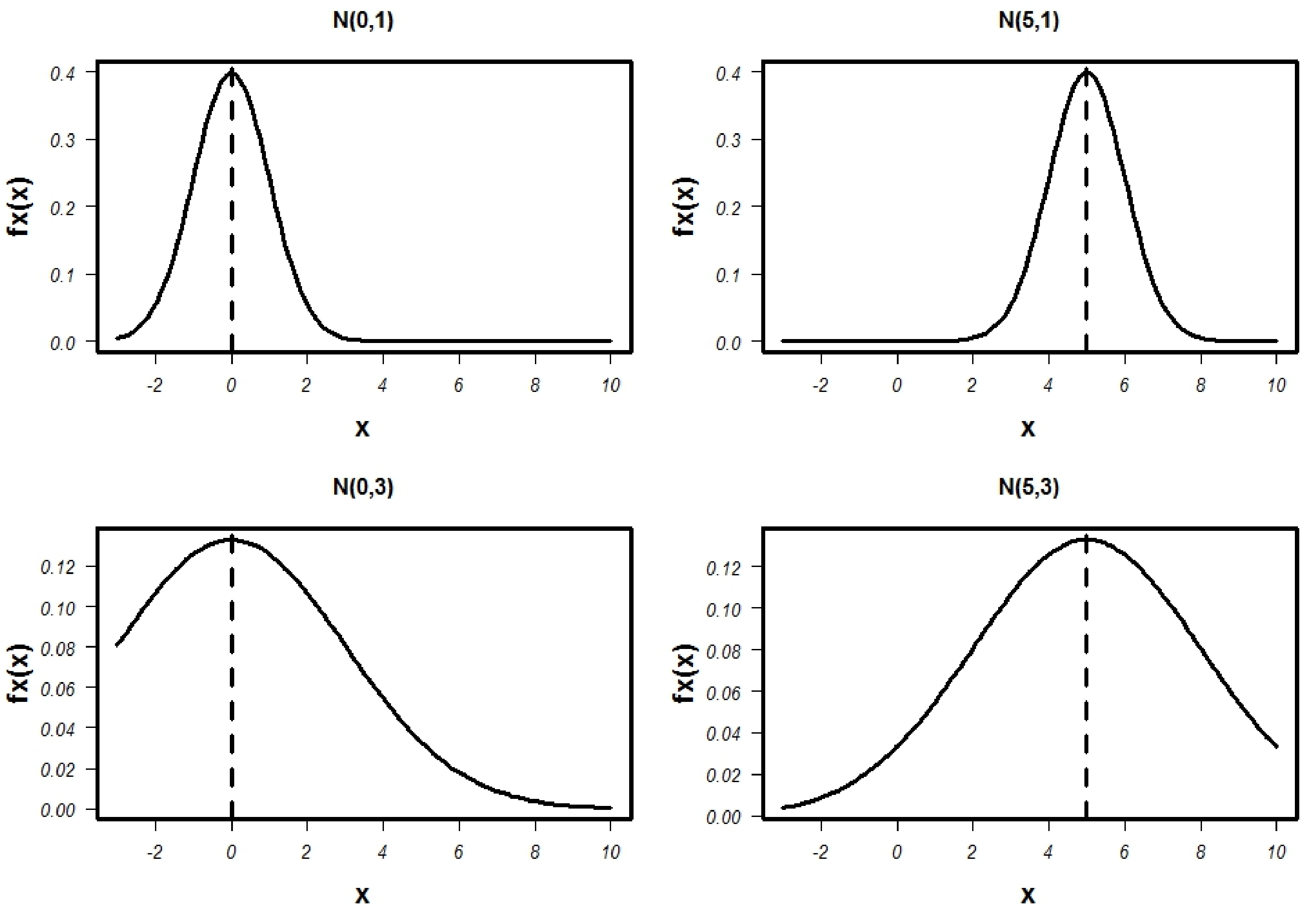
$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad amb \quad x \in R$$

- **La funció de distribució** no té expressió analítica → Taules
- **Indicadors:**
  – E(X) = μ
  – V(X) = σ^2

La Normal amb paràmetres μ = 0 i σ=1 s'anomena **Normal estàndard** i és la que apareix a les taules

Model Normal. Representació gràfica

Ex: Com són les funcions de densitat de diferents Normals segons els valors de μ i σ?



Model Normal. Ex. de càlcul de probabilitats

Sigui $X \sim N(\mu=0, \sigma=1)$:

- **Probabilitat puntual**. → Recordeu que P(X=x) = 0 per qualsevol x ja que és una VAC

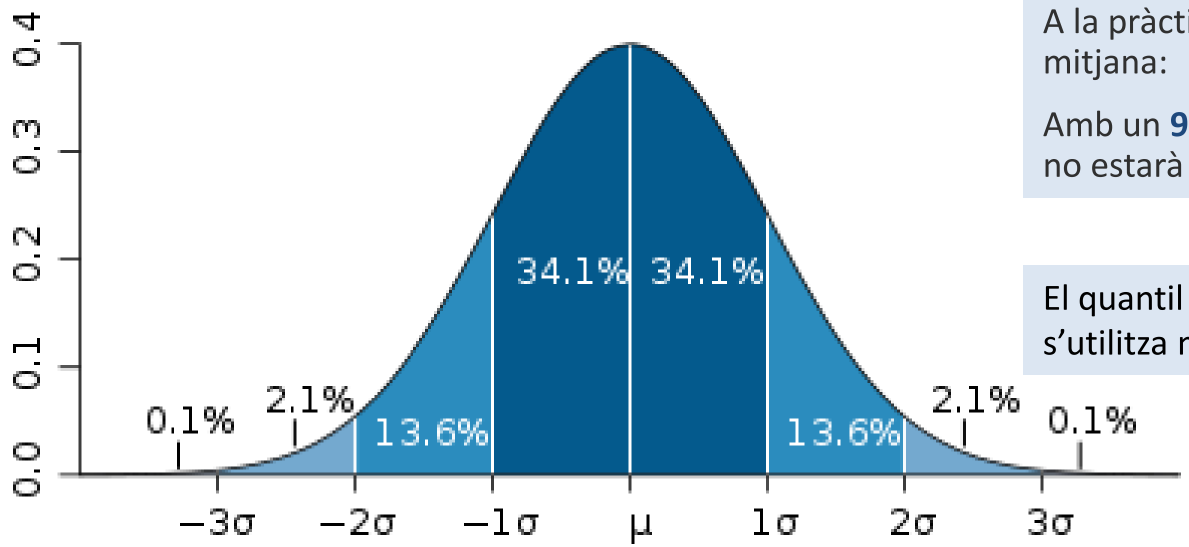
- **Probabilitat acumulada**. Quina és la probabilitat de 2 o menys?
  - Amb fórmules →**No es pot fer!!!**
  - Amb taules →$P(X \leq 2) = \mathbf{0.977}$
  - Amb R → $P(X \leq 2) = pnorm(q = 2, \text{mean} = 0, sd = 1) = \mathbf{0.9772499}$

- **Quantils**. Quin és el valor tq la probabilitat de quedar per sota d'ell és almenys 0.95?
  - Amb fórmules →**No es pot fer!!!**
  - Amb taules → $P(X \leq x_{0.95}) = 0.95 \rightarrow x_{0.95} = \mathbf{1.645}$
  - Amb R →$P(X \leq x_{0.95}) = 0.95 \rightarrow qnorm(p = 0.95, mean = 0, sd = 1) = \mathbf{1.645}$

Nota: Només hi ha taules de la Normal estàndar, **N(0,1)**. En la resta de casos, s'ha **d'estandarditzar** per fer els càlculs amb taules

Model Normal. Propietats i quantils

- La funció de densitat *f*(*x*) és simètrica respecte al punt $x = \mu$, que és a la vegada, la mitjana i la mediana de la distribució.
- Els punts d'inflexió es troben a una desviació estàndard de la μ ($x=\mu-\sigma$ i $x=\mu+\sigma$)
- Els quantils de la Normal estàndard Z~N(0,1), normalment, es denoten amb z_p. El quantil z_p representa aquell valor tal que en una Normal estàndard té una probabilitat p de caure en l'interval $(-\infty, z_p]$



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.

El quantil més emprat és el $Z_{0.975} = 1.96$ ja que s'utilitza molt en la inferència estadística

Model Normal. Estandardització

- La **combinació lineal de variables Normals** és Normal:
  - Sigui a i b, dos escalars i $X \sim N(\mu_X, \sigma_X) \rightarrow Y = a \cdot X + b \sim N(\boldsymbol{\mu_Y} = a \cdot \mu_X + b, \boldsymbol{\sigma_Y} = a \cdot \sigma_1)$
  - Sigui a i b, dos escalars, $X \sim N(\mu_X, \sigma_X)$ i $Y \sim N(\mu_Y, \sigma_Y) \rightarrow$

    $\rightarrow W = a \cdot X + b \cdot Y \sim N\left(\boldsymbol{\mu_W} = a \cdot \mu_X + b \cdot \mu_Y, \boldsymbol{\sigma_W} = \sqrt{a^2 {\sigma_1}^2 + b^2 {\sigma_2}^2 + 2 \cdot ab \cdot \rho_{XY} \cdot \sigma_1 \cdot \sigma_2}\right)$

- Aquesta propietat permet relacionar distribucions Normals a base de translacions i escalars. En particular, transformar a la Normal estàndard $Z \sim N(0,1)$, **estandarditzar**, permet buscar en les taules de Z, probabilitats de qualsevol Normal

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

  $Z = X/\sigma - \mu/\sigma = (X - \mu) / \sigma \sim N(0, 1)$ ($a = 1/\sigma$, $b = -\mu/\sigma$ són escalars). És a dir:

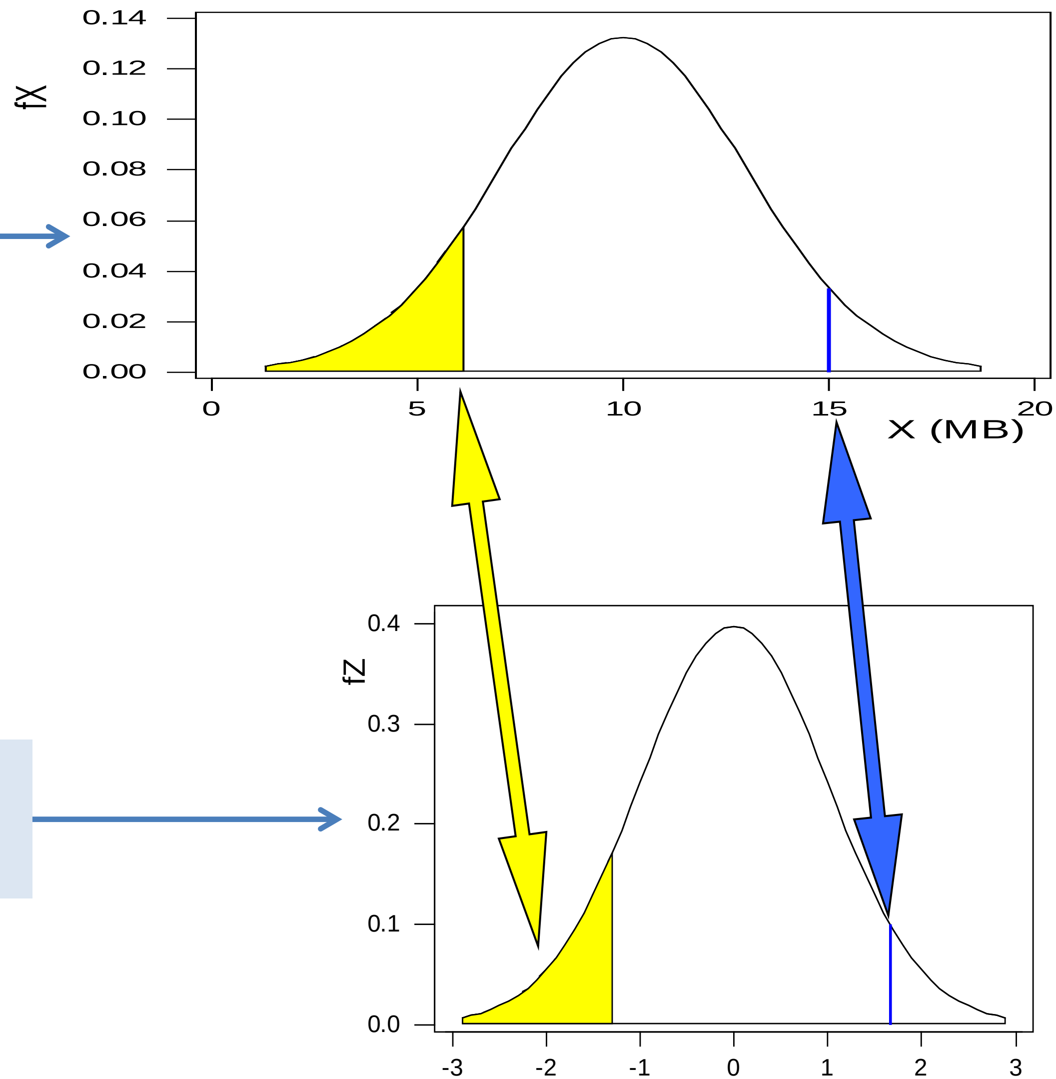
$$\boldsymbol{Z = \frac{X-\mu}{\sigma} \sim N(0,1) \rightarrow X = \mu + \sigma \cdot Z}$$

Model Normal. Estandardització

Variable X: situació real (per exemple, MB d'un fitxer)

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

Variable Z: situació estandarditzada, sense unitats, centrada en 0, dispersió tipificada (igual a la unitat)



Model Normal. Exemple d'estandardització

Exemple 1: [**probabilitats inferiors a 0.5**] . Sigui Z ~ N(0, 1). Quin és el valor *z* que deixa una probabilitat per sota de p = 0.25?

Com 0.25 < 0.5, no surt a les taules; s'ha d'utilitzar simetries:

$$P(Z < -|z|) = P(Z > |z|) \rightarrow F_Z(0.6745) = 0.75 \rightarrow z = -0.6745$$

Exemple 2: [**Estandardització**] Sigui X: "Increment diari espai disc" ~ N(10 MB, 3 MB)

1) Quant val P(*X*>15)?

$$\mathrm{Z} = \frac{\mathrm{X} - 10}{3} \sim \mathrm{N}(0,1) \rightarrow \mathrm{P}(\mathrm{X} > 15) = \mathrm{P}\left(\mathrm{Z} > \frac{15 - 10}{3}\right) = \mathrm{P}\left(\mathrm{Z} > \frac{5}{3}\right) = 1 - \mathrm{P}\left(\mathrm{Z} \leq \frac{5}{3}\right)$$

$$= 1 - \mathrm{F_Z}\left(\frac{5}{3}\right) = 1 - 0.9522 = 0.0478$$

2) 1 de cada 10 dies, l'augment és inferior a quant? (Quant val *t* tal que P(*X* < *t*) = 0.1?)

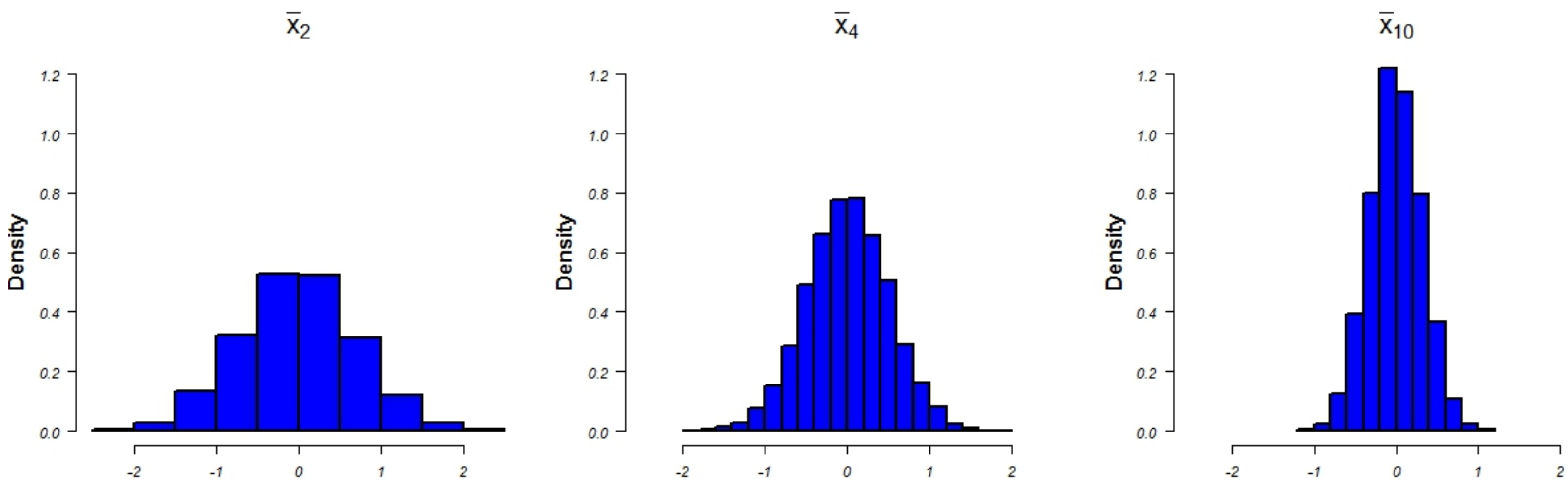
$$P(Z < t') = 0.1 \rightarrow t' = F_Z^{-1}(0.1) = -1.2816 \rightarrow t = 10 + 3 \cdot t' = 6.15\ MB$$

Distribució de la mitjana de v.a. Model Normal

- Hem simulat $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ tal que X_i siguin i.i.d. Observem que:
  - **tendeix a concentrar-se al voltant de μ** quan *n* augmenta
  - **tendeix a assemblar-se a una Normal** a mesura que n es fa gran.

$$E(\bar{X}_n) = \frac{E(\sum X_i)}{n} = \frac{\sum E(X_i)}{n} = \frac{n\mu}{n} = \mu \qquad V(\bar{X}_n) = \frac{V(\sum X_i)}{n^2} = \frac{\sum V(X_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

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



Teorema Central del Límit (TCL)

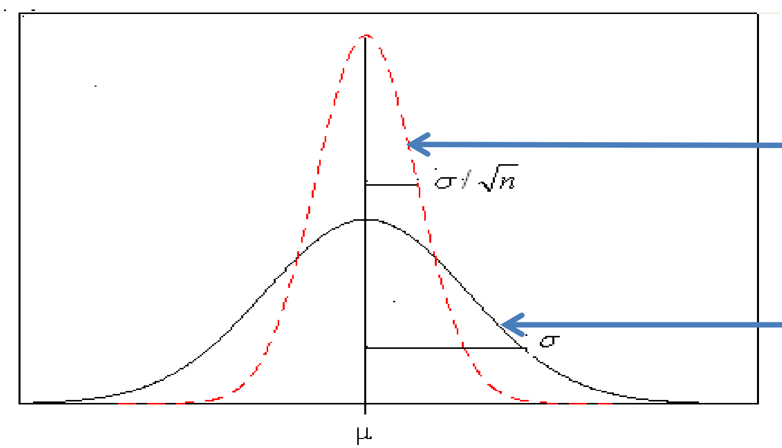
- Siguin X_1, X_2, …, X_n independents i idènticament distribuïdes amb esperança μ i desviació típica σ. Llavors:

$$S_n = \sum X_i \xrightarrow[n\ gran]{} N(n\mu, \sigma\sqrt{n}) \Rightarrow \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\ gran]{} N(0,1)$$

$$\bar{X}_n = \frac{\sum X_i}{n} \xrightarrow[n\ gran]{} N(\mu, \sigma/\sqrt{n}) \Rightarrow \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\ gran]{} N(0,1)$$

- És a dir, amb n gran, la funció de distribució de la variable Suma (S_n) i mitjana ($\bar{X}_n$) tendeix a una Normal amb uns determinats paràmetres **independentment de la distribució de les X_i!**

Teorema Central del Límit (TCL)

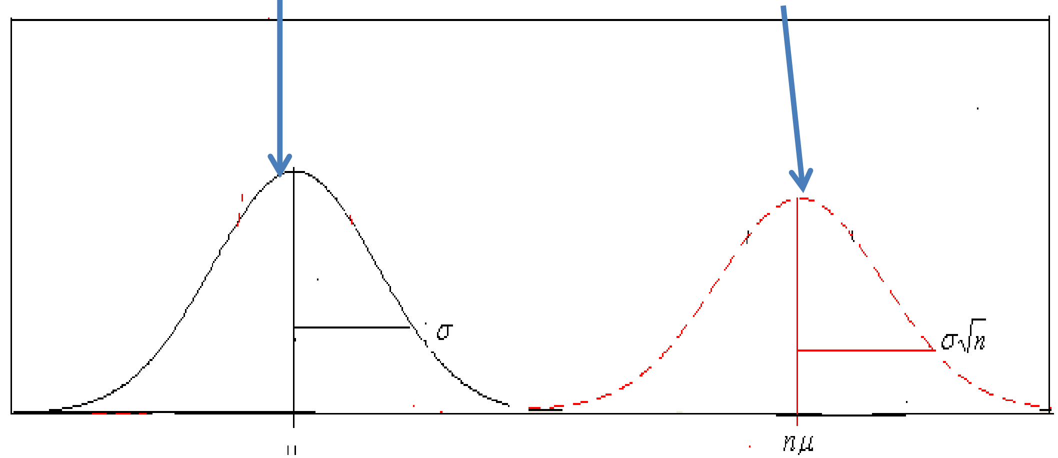


$$\bar{X}_n = \frac{\sum X_i}{n} \xrightarrow[n\ gran]{} N(\mu, \sigma/\sqrt{n})$$

X_i

X_i

$$S_n = \sum X_i \xrightarrow[n\ gran]{} N(n\mu, \sigma\sqrt{n})$$



Els X_i no necessàriament han de ser Normals!!!!

Només han de complir:

E(X) = μ

V(X) = σ^2

Teorema Central del Límit (TCL). Exemple

El treball de CPU per fer un *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:

$S_n = X_1 + X_2 + \ldots + X_{30}$ (no ens donen la distribució de X_i, sols μ i σ)

1) Quina és la distribució de S_n?

$$S_n = \sum X_i \xrightarrow[n\ gran]{} N(n\mu, \sigma\sqrt{n}) = N(900\,, 82.15)\ min = N(15{,}1.37)\ hores$$

2) Quina és la probabilitat d'un consum total mensual de més de 18 hores de CPU?

$$P(S_n > 18) = P\left(\frac{S_n - 15}{1.37} > \frac{18-15}{1.37}\right) = P(Z > 2.19) = 1 - p(Z < 2.19) = 1 - 0.9857 = 0.0143$$

[un de cada 70 mesos]

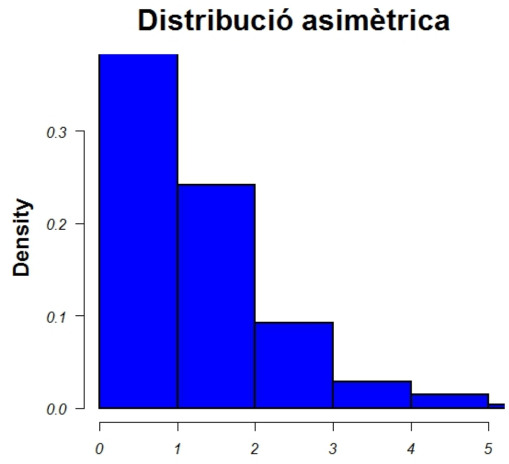
3) I la probabilitat de, en un mes, un consum mitjà diari inferior a 36'?

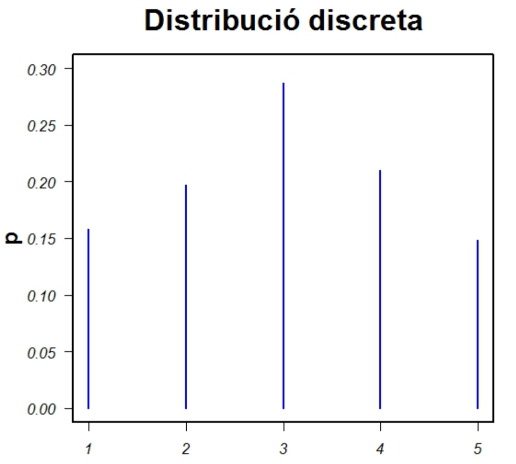
$$\bar{X}_{30} \sim N\left(\mu, \frac{\sigma}{\sqrt{30}}\right) = N(30\,, 2.74)\ min \rightarrow P(\bar{X}_{30} < 36) = P\left(Z < \frac{36-30}{2.74}\right) = P(Z < 2.19) = 0.9857$$

Teorema Central del Límit (TCL). ”n”

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

- Depèn de com sigui la distribució 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** o són **variables discretes** (especialment si només pot prendre pocs valors):





- Aplicacions del TCL: la normal aproxima bé certes distribucions. [Exemple: variable de Poisson, si λ és gran. La t-Student, i la χ^2 son derivades de la Normal que es veuran més endavant]

Teorema Central del Límit (TCL). Exercici

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

1) VAC d'una mesura de l'error: E_n: "error en la n-èssima mesura"
2) Model de la VAC: $E_n \sim N(\mu = 0, \sigma = 0.25)$
3) Prob. error menor de 0.1 s.:

$$P(|E_1| \leq 0.1) = P(-0.1 \leq E_1 \leq 0.1) = F_{E_1}(0.1) - F_{E_1}(-0.1) = 0.655 - 0.435 = 0.311$$

4) Error màxim (amb prob. 0.95) en una mesura:

$$P(-e \leq E_1 \leq e) = 0.95 \;\rightarrow P(E_1 \leq e) = 0.975 \rightarrow e = 0.49\ s.$$

5) Ídem per VAC de mitjana de 10 mesures:

$$X_{10} \sim N\left(\mu = 0, \sigma = \frac{0.25}{\sqrt{10}}\right) \rightarrow P(-f \leq X_{10} \leq f) = 0.95 \rightarrow f = 0.156\ \mathrm{s.}$$

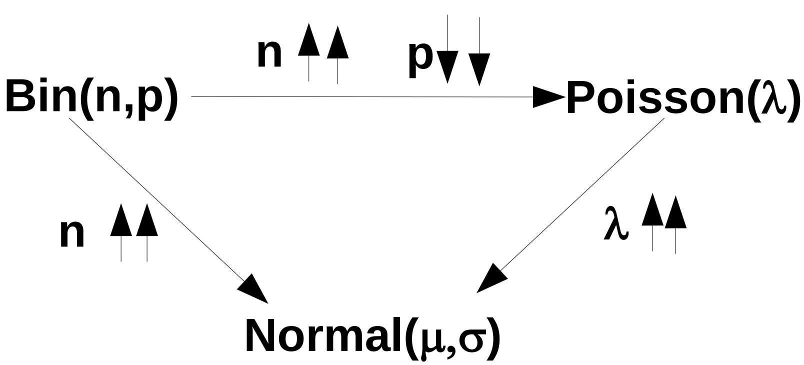
6) Número n mínim de mesures per tal que l'error màxim de la mitjana de les *n* mesures (amb prob. 0.95) sigui inferior a 0.1 s. És a dir n tal que

$$P(-0.1 \leq X_n \leq 0.1) \geq 0.95 \rightarrow P\left(\frac{-0.1-0}{\frac{0.25}{\sqrt{n}}} \leq Z \leq \frac{0.1-0}{\frac{0.25}{\sqrt{n}}}\right) \rightarrow \frac{0.1}{\frac{0.25}{\sqrt{n}}} = 1.956 \rightarrow n = 24.01 \rightarrow n = 25$$

TCL. Relacions entre distribucions

Una de les aplicacions pràctiques del TCL és que la distribució Normal es pot emprar com a aproximació d'altres distribucions:

- La **distribució Binomial** amb paràmetres *n* i *p* es pot aproximar per una Normal quan *n* és gran i la *p* no massa extrema (ni molt a prop de 0 ni de 1). Llavors, els paràmetres de la Normal són
  - $\mu = n \cdot p$
  - $\sigma^2 = n \cdot p \cdot (1-p)$
- La **distribució de Poisson** amb paràmetre λ es pot aproximar per una Normal quan la λ és prou gran. Llavors els paràmetres de la Normal són:
  - $\mu = \lambda$
  - $\sigma^2 = \lambda$



Nota: la Binomial es pot aproximar a una Poisson quan la *n* és prou gran i la *p* prou petita

TCL. Relacions entre distribucions. Exemple

S’ha comprovat que la probabilitat que el temps de resposta d’una determinada pàgina web 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 (7 dies), un mes (30 dies) o cinc anys (1825 dies), necessitem les següents variables aleatòries per les quals podem identificar el model més adequat:

X_{set}=“nombre de dies en 1 setmana amb servei inadequat”

X_{set}~Bin(n=7,p=0.05)

X_{mes}=“nombre de dies en 1 mes amb servei inadequat“

X_{mes} ~ Bin(n=30,p=0.05)

X_{mes} ~ P(λ=1.5)

X_{5anys}=“nombre de dies en 5 anys amb servei inadequat”

X_{5anys} ~ Bin(n=1825, p=0.05)

X_{5anys} ~ P(λ=91.25)

X_{5anys} ~ N(μ=91.25, σ=9.55) [Encara que p és petita, n és molt gran]

TCL. Exercici

Suposem que s'ha establert que el pes del equipatge d'un viatger segueix una distribució Normal amb mitjana 18.9 Kg. i desviació 4 Kg. Habitualment, si el equipatge d'un viatger sobrepassa els 20 Kg., llavors té un sobrepreu que depèn de l'excés de pes. Contesta les següents qüestions:

1. Troba la probabilitat que un viatger hagi de pagar sobrepreu per excedir el seu equipatge els 20 Kg. de pes. 0.392 [`1-pnorm(20,18.9,4)`]

2. Quin és el pes que podem assegurar , amb un 95% de probabilitat , que un equipatge no superarà? 25.479 [`qnorm(0.95,18.9,4)`]

3. Indica els paràmetres (μ , σ) de la variable pes <u>total</u> de 10 equipatges

$N(\mu = 189, \sigma = 4\sqrt{10} = 12.649)$

4. Indica els paràmetres (μ , σ) de la variable pes <u>promig</u> de 10 equipatges

$N(\mu = 18.9, \sigma = 4/\sqrt{10} = 1.265)$

TAULA resum de tots els models rellevants

Distribució	Declaració	Funció de probabilitat o de densitat	Funció distribució	Esperança E(X)	Variància 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) *	$P_X(k)=\frac{\lambda^k \cdot e^{-\lambda}}{k!} \quad k=0,1,2,\ldots$ **(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) *	$f_X(x)=\lambda \cdot e^{-\lambda \cdot x} \quad x>0$ **(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]	$f_X(x)=\frac{1}{b-a} \quad a<x<b$ **(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

$0 < p < 1 \qquad q = 1 - p \qquad n \in \mathrm{N} \qquad r \in \mathrm{N} \qquad \lambda \in \Re^+$

*λ paràmetre del procés Poisson: variables Poisson i Exponencial