Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Models de VA i simulació
Conceptes

Bloc B – Probabilitat i Estadística

2024

Índex

Models de VAD i VAC

- **A Wikipedia** (https://en.wikipedia.org/wiki/List_of_probability_distributions): *"Many probability distributions that are important in theory or applications 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 → $E(X) = \mu_X$
  - Variància → $V(X) = \sigma_X^2$
- A partir de les **funcions de probabilitat i distribució** de probabilitat es calculen probabilitats:

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

- 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: $F_X(x_\alpha) = \alpha \quad (0 \leq \alpha \leq 1)$

Model de Bernoulli

- **Definició**: Número d'èxits en la realització d'un únic experiment amb 2 possibles resultats*: **0** ("no èxit") i **1** ("èxit")

* Parlem de respostes binàries o dicotòmiques

- **Notació**: X~Bern(p)
- **Paràmetres**: p (probabilitat d'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
  - **Geomètrica:** número de repeticions fins observar el primer "èxit"
  - **Binomial negativa:** número de repeticions fins observar el r-èssim "èxit"

- 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_point_process)

Model Binomial

- **Definició**: Número d'èxits en la repetició de *n* proves de *Bernoulli* independents amb probabilitat constant *p*
- **Notació**: X~B(n,p)

  Notació de funcions en **R** pel model: dbinom(), pbinom(), qbinom()

- **Paràmetres**: n (nombre de repeticions), p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat**:

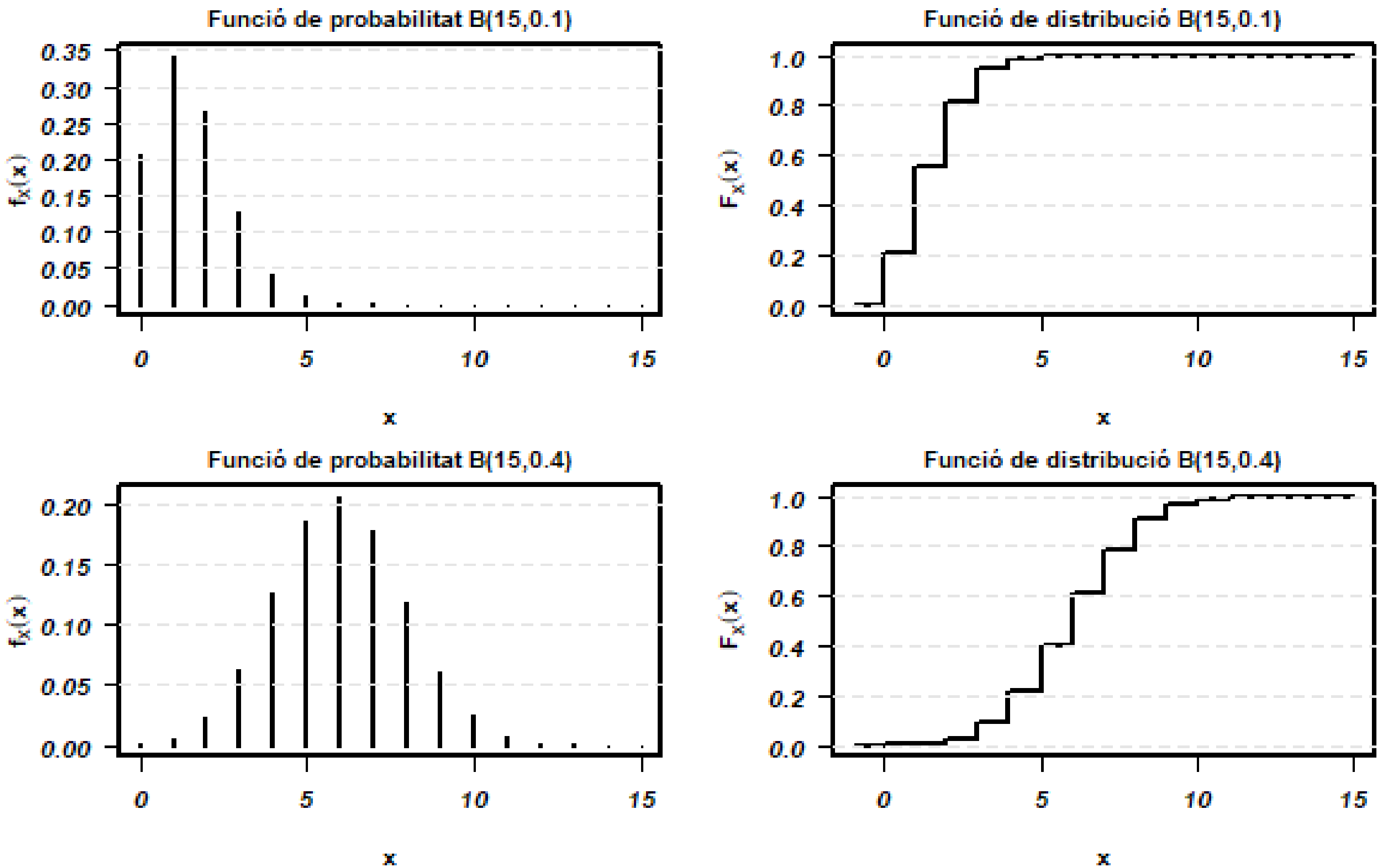
$$P(X = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k} \quad amb \quad k = 0, 1, \ldots, n$$

on $q = 1 - p$ *i* $\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$ *(o choose(n,k) en R,* https://rdrr.io/snippets/ per R online*)*

- No té **funció de distribució** analítica explícita → És el sumatori de probabilitats puntuals
- **Indicadors:**
  - *E(X) = n·p*
  - *V(X) = n·p·q*

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 $X \sim B(n=20, p=0.5)$

- **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 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 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 R →$P(X \leq x_{0.95}) = 0.95 \rightarrow qbinom(p=0.95, size=20, prob=0.5) = \mathbf{14}$

dbinom, pbinom, qbinom són funcions en R (https://rdrr.io/snippets/ per R online)

El model Binomial permet relacionar probabilitats de variables que segueixen models amb paràmetre p o bé 1-p:
$P(X \leq k) = 1 - P(Y \leq n-k-1)$ on $X \sim B(n, p)$ i $Y \sim B(n, 1-p)$ (o bé pbinom(k,n,p)=1-pbinom(n-k-1,n,1-p))

Model Geomètric

- **Definició**: nombre d'intents (k) d'un experiment de Bernoulli fins observar el **primer èxit**
- **Notació**: X~Geom(p)

Notació de funcions en **R** pel model: dgeom(), pgeom(), qgeom()

Les funcions en **R** són pel número de fracassos enlloc d'intents *(k intents són k-1 fracassos)*

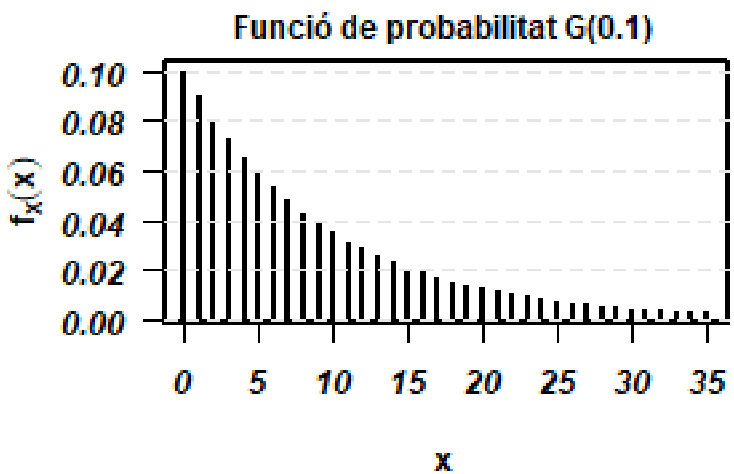
- **Paràmetres**: p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat** :

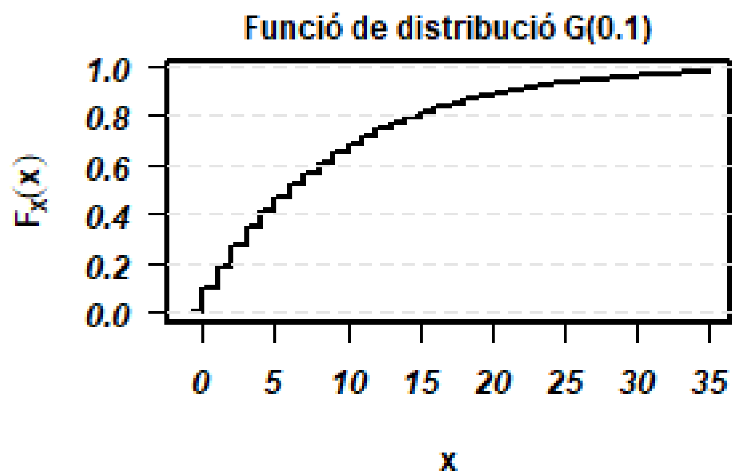
$$P(X = k) = q^{k-1} \cdot p \qquad k \geq 1$$

Indicadors:

E(X) = 1/p

V(X) = (1-p)/p²





Qualsevol valor enter k > 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.

Model Binomial negativa

- **Definició**: nombre d'intents (k) d'un experiment de Bernoulli fins observar ***r* èxits**
- **Notació**: X~BN(r, p)

Notació de funcions en **R** pel model: dnbinom(), pnbinom(), qnbinom()

Les funcions en **R** són pel número de fracassos enlloc d'intents *(k intents són k-r fracassos)*

- **Paràmetres**: r (nombre d'èxits), p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat**:

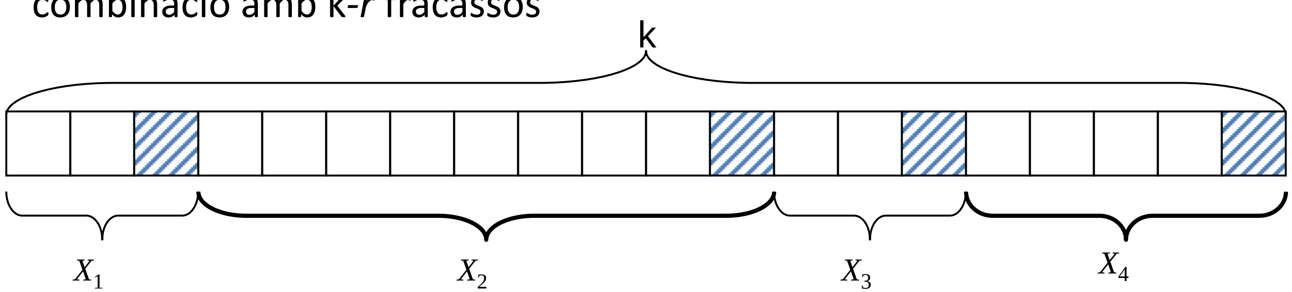
$$P(X = k) = \binom{k-1}{r-1} \cdot p^r \cdot q^{k-r}$$

Indicadors:

E(X) = r/p

V(X) = r(1-p)/p²

Sense comptar l'últim intent (que <u>ha de</u> ser un èxit), són *r* -1 èxits barrejats en qualsevol combinació amb k-*r* fracassos



Nota: En general, podem pensar que una BN és una suma de *r* geomètriques independents
Si *r*=1 tenim la distribució geomètrica

Model Poisson

- **Definició**: Número d'ocurrències en un determinat interval de temps o espai

  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: trucades a una centraleta es poden representar amb una variable Poisson de "nombre de trucades per hora", sabent la mitjana de trucades rebudes per hora)

- **Notació**: X~P(λ)

  Notació de funcions en **R** pel model: dpois(), ppois(), qpois()

- **Paràmetres**: λ (taxa de l'esdeveniment)
- **Funció de probabilitat**:

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

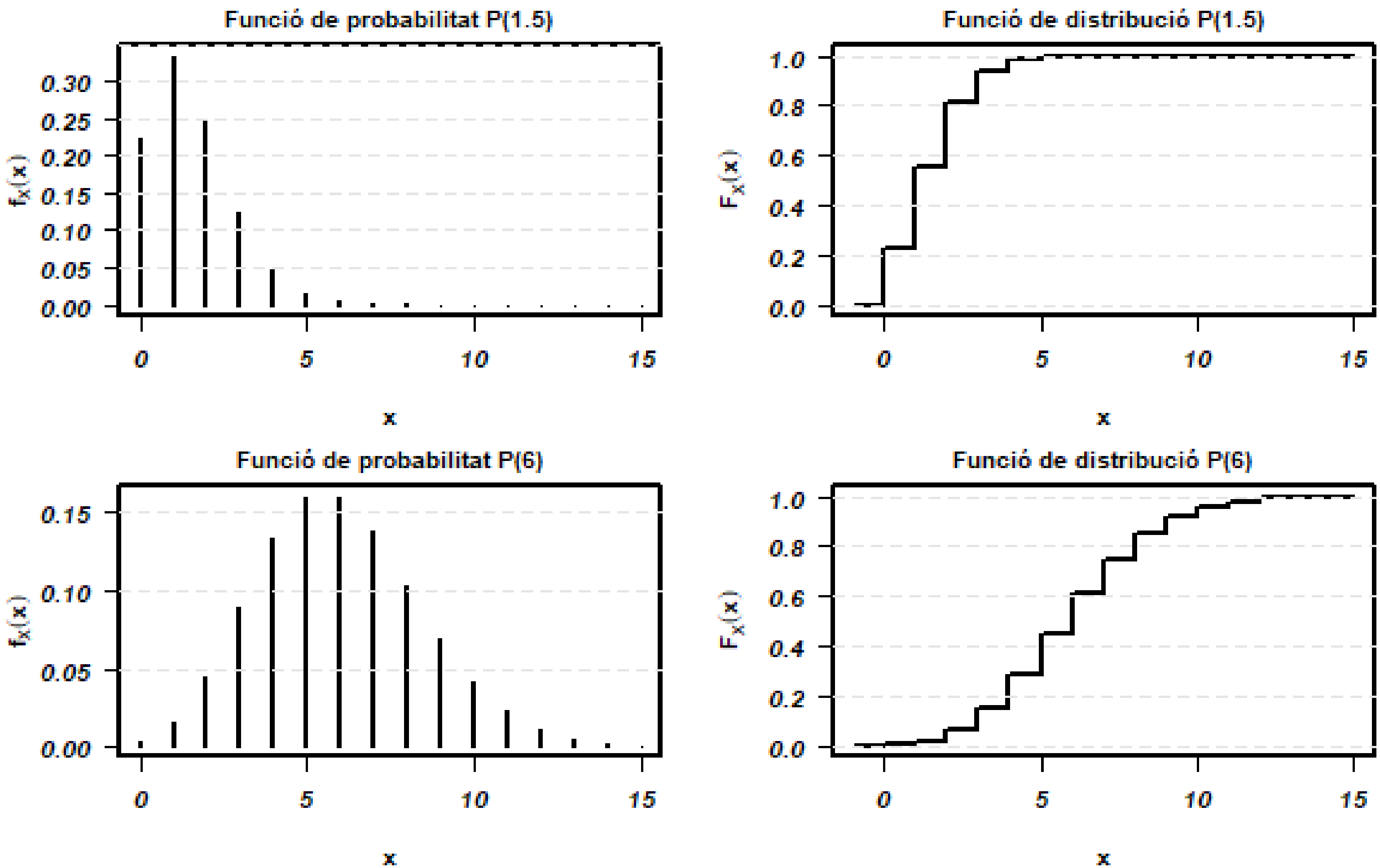
- No té **funció de distribució** analítica → És el sumatori de probabilitats puntuals
- **Indicadors:**
  - E(X) = λ
  - V(X) = λ

  (**λ** és un número real positiu que representa la taxa mitjana d'ocurrències per unitat considerada)

Nota: 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

Model Poisson. Representació gràfica

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



Model Poisson. Ex. de càlcul de probabilitats

Sigui $X \sim P(\lambda=2)$:

- **Probabilitat puntual**. Quina és la probabilitat de 3?
  - Amb fórmules → $P(X=3)=\frac{e^{-2}\cdot 2^3}{3!}=\mathbf{0.1804}$
  - 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 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 R →$P(X\leq x_{0.95})=0.95\rightarrow qpois(p=0.95,lambda=2)=\mathbf{5}$

Nota: dpois, ppois, qpois són funcions en R (https://rdrr.io/snippets/ per R online)

Nota: La suma de VAD Poisson és també una VAD Poisson amb paràmetre λ igual a la suma dels paràmetres:

X ~ P(λ1) Y ~ P(λ2) → X+Y ~ P(λ1+ λ2)

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) sent k>=0

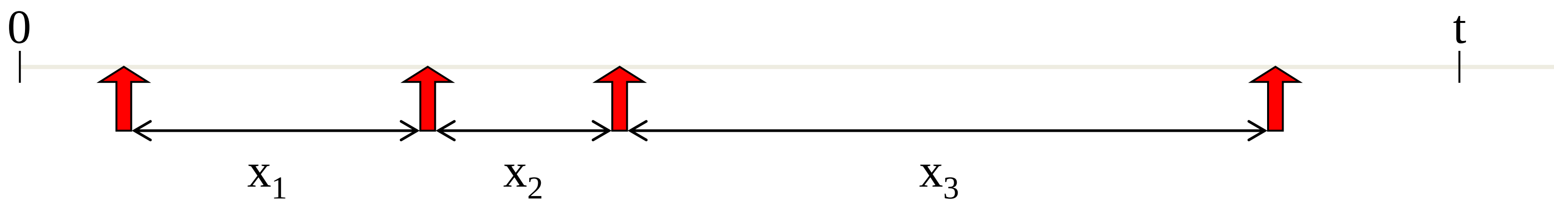
TAULA resum de models de VAD

Distribució	Declaració	Domini	Esperança $E(X) = \mu_X$	Variància $V(X) = \sigma_X{}^2$
Bernoulli	Bern(p)	0, 1	p	$p \cdot q$
Binomial	B(n,p)	0,1,…,n	$n \cdot p$	$n \cdot p \cdot q$
Geomètrica	Geom(p)	1,2,3,…	$1/p$	q/p^2
Binomial negativa	BN(r,p)	r, r+1,…	r/p	$q \cdot r/p^2$
Poisson	P(λ)	0, 1, 2,…	λ	λ

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

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~Exp(λ)

Notació de funcions en **R** pel model: dexp(), pexp(), qexp()

- **Paràmetres**: λ (taxa d'aparició de l'esdeveniment per unitat de temps)
- **Funció de densitat i de distribució**:

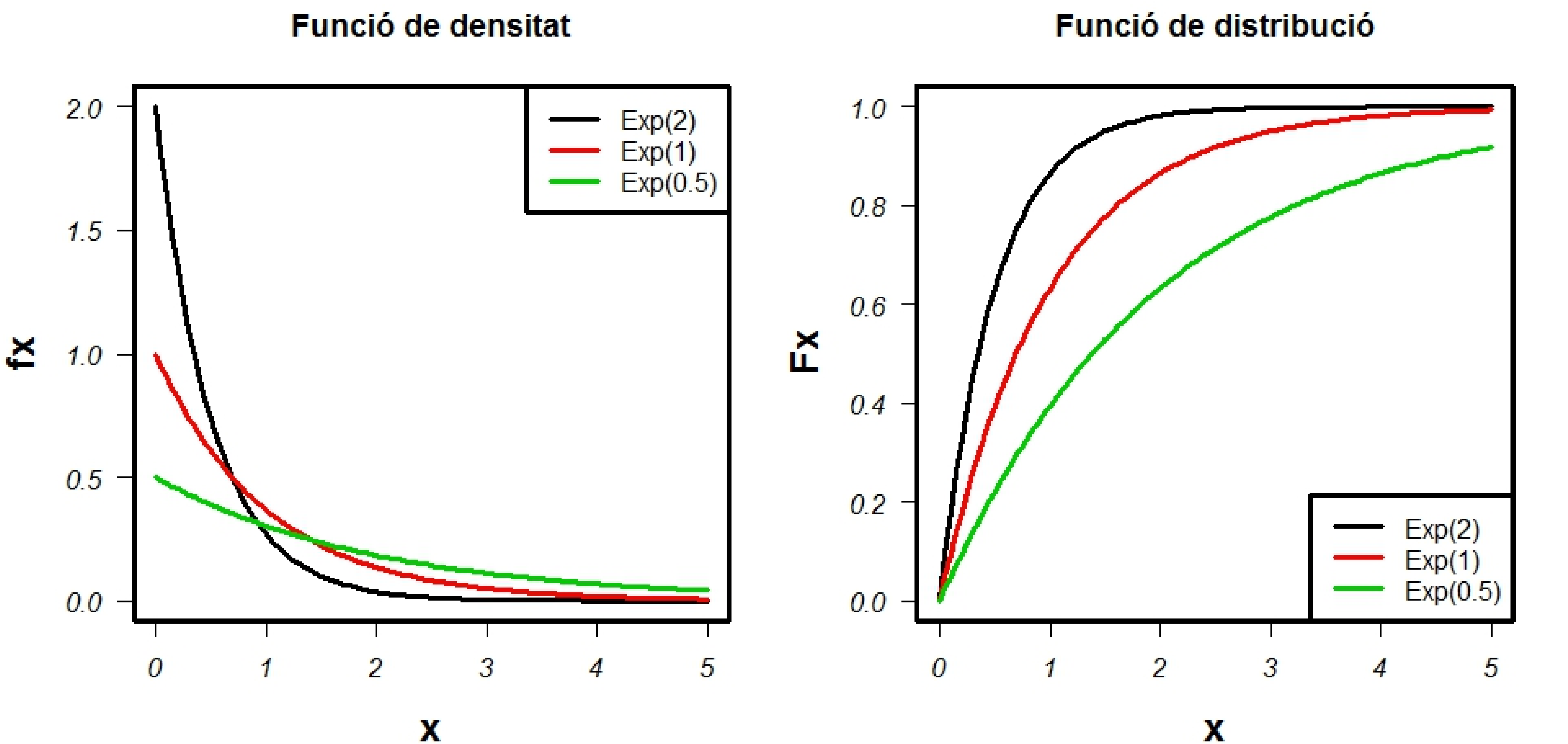
$$f_X(x) = \lambda \cdot e^{-\lambda x} \qquad amb \quad x > 0$$

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

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

Model Exponencial. Representació gràfica

Ex: Com es distribueix el <u>temps entre correus *spam* </u>si rebo una mitjana 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 0.95?
  - Amb fórmules → $P(X \leq x_{0.95}) = 0.95 \rightarrow 1 - e^{-2\cdot x_{0.95}} = 0.95 \rightarrow x_{0.95} = \mathbf{1.4979}$
  - Amb R → $P(X \leq x_{0.95}) = 0.95 \rightarrow qexp(p = 0.95, rate = 2) = \mathbf{1.497866}$

Nota: pexp, qexp són funcions en R (https://rdrr.io/snippets/ per R online)

Nota: La distribució Exponencial té funció de distribució amb expressió analítica

Model Exponencial (algunes propietats)

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

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

Recordem també que en el model exponencial les probabilitats acumulades 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:

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

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. Què és més probable: (A) rebre en en els 10'' següents, o (B) rebre 10'' desprès d'una arribada?

  Solució: Igual

Model Uniforme (continu)

- **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)

  Notació de funcions en **R** pel model: dunif(), punif(), qunif()
- **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 $x > b$

$$F_X(x) = \frac{x-a}{b-a} \qquad amb \quad a < x < b$$

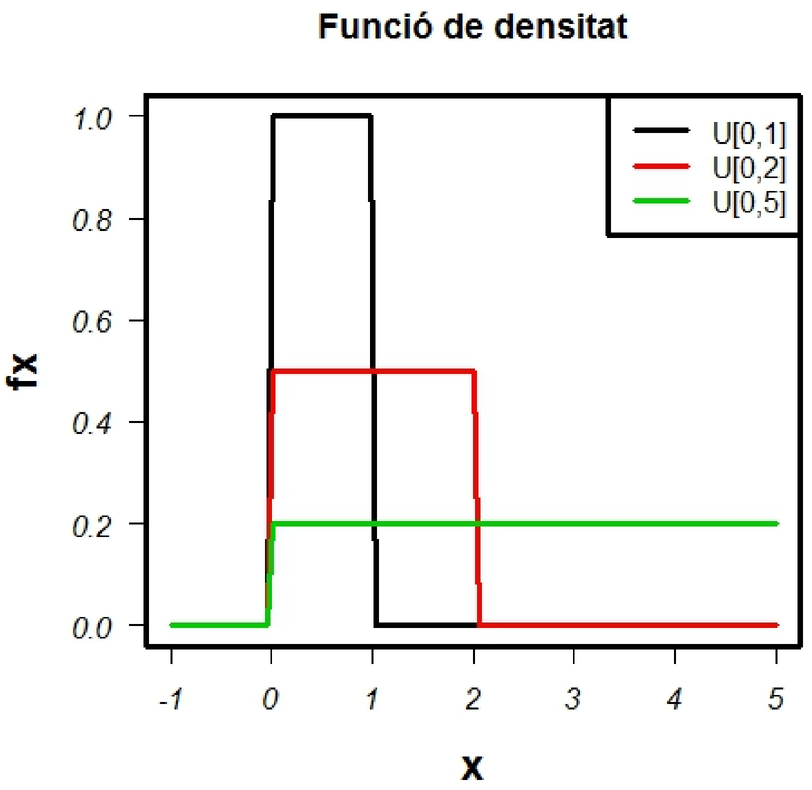
- **Indicadors:**
  - E(X) = (b + a)/2
  - V(X) = $(b - a)^2/12$

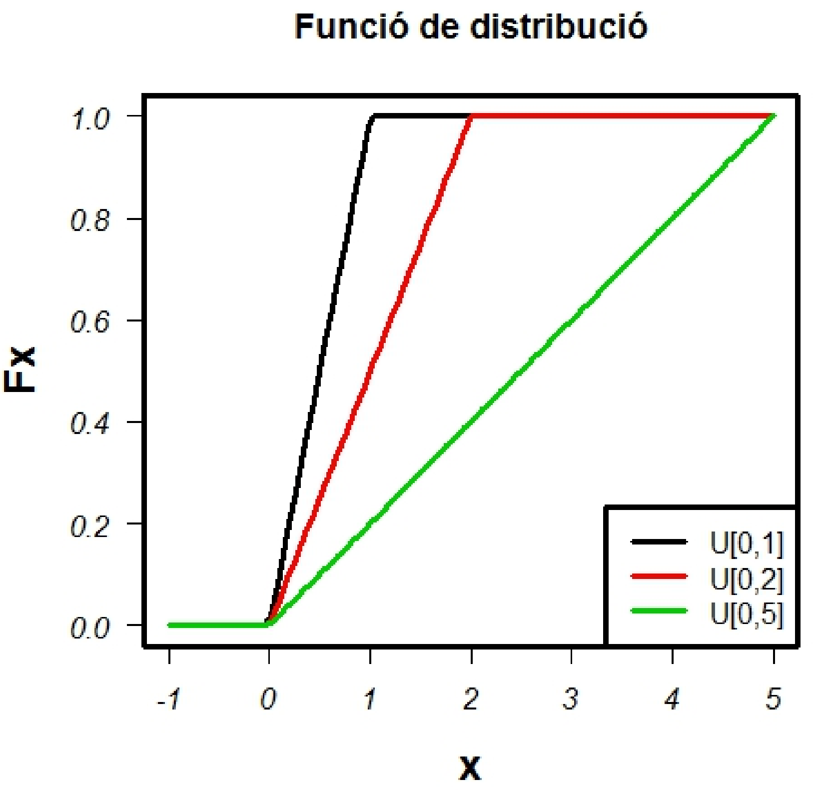
$$E(X) = \mu_X = \int_{-\infty}^{+\infty} x \cdot f_X(x)dx = \int_a^b x \cdot \frac{1}{b-a} dx =$$
$$= \frac{1}{b-a} \cdot \frac{x^2}{2}\Big]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

(intuïtivament ja es veu que la mitjana ha de ser el centre de l'interval)

Model Uniforme. Representació gràfica

Ex: Com es distribueixen el <u>nombres reals aleatoris entre 0 i 1</u>? I entre 0 i 2? I entre 0 i 5?



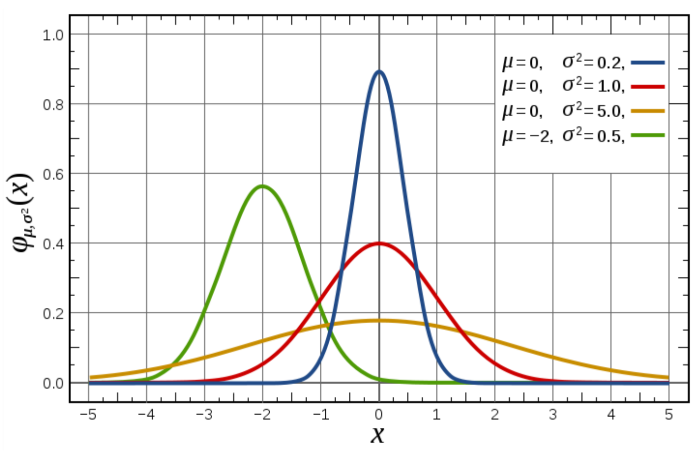


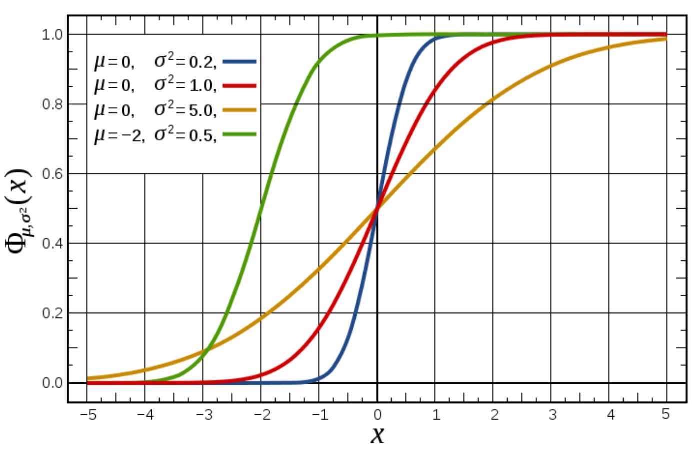
Model Normal (o de Gauss). Introducció

(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 serveix per representar els 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)

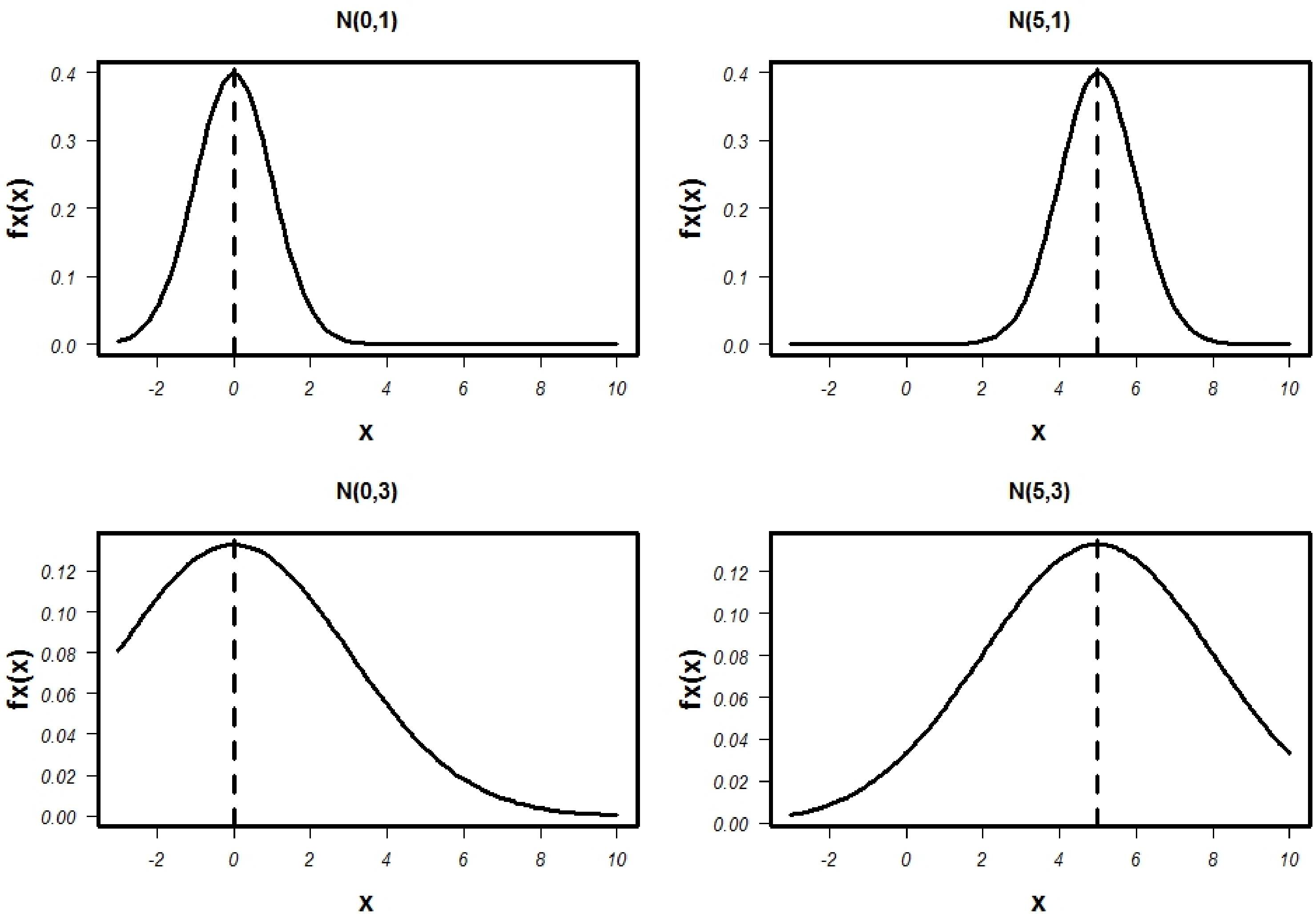
Notació de funcions en **R** pel model: dnorm(), pnorm(), qnorm()

- **Paràmetres**: μ (esperança), σ (desviació estàndard) [vigilar si s'usa σ^2 i no σ com a paràmetre]
- **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 explícita → R
- **Indicadors:**
  - E(X) = μ
  - V(X) = σ^2

Nota: la Normal amb paràmetres μ = 0 i σ=1 s'anomena **Normal estàndard**

Model Normal. Representació gràfica

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



Model Normal. Exemple 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 R → $P(X \leq 2) = pnorm(q = 2, \text{mean} = 0, sd = 1) = \mathbf{0.9772499}$

    $$(pnorm(2) = 0.9772499)$$
- **Quantils**. Quin és el valor tq la probabilitat de quedar per sota d'ell és 0.95?
  - Amb fórmules →**No es pot fer!!!**
  - Amb R →$P(X \leq x_{0.95}) = 0.95 \rightarrow qnorm(p = 0.95, mean = 0, sd = 1) = \mathbf{1.645}$

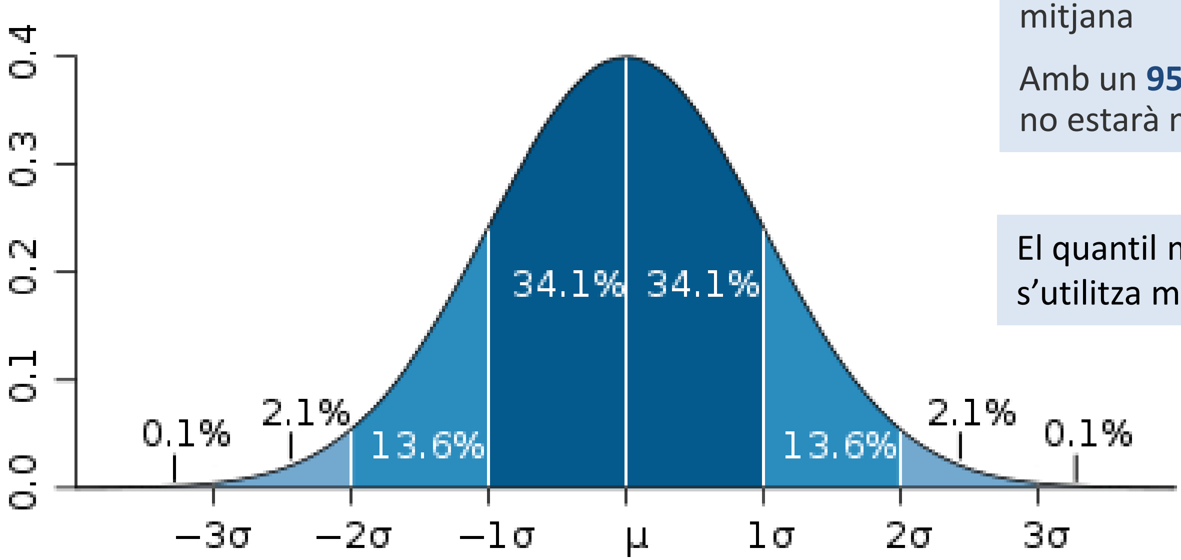
    $$(qnorm(0.95) = 1.64)$$

Nota: pnorm, qnorm són funcions en R (https://rdrr.io/snippets/ per R online)

Nota: les anteriors funcions en R serveixen per a qualsevol Normal, però és habitual calcular-ho amb l'estàndard (**N(0,1)**) amb la transformació que veurem d'**estandarització**. De fet, mean=0 i sd=1 són els valors per defecte en les funcions

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.
- En qualsevol Normal, la probabilitat d'allunyar-se de μ una quantitat inferior a σ ($\mu-\sigma < x < \mu+\sigma$) és aproximadament 0.682
- Els quantils de la Normal estàndard $Z \sim N(0,1)$, 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 $\mathbf{Z_{0.975} = 1.96}$ ja que s'utilitza molt en inferència estadística

Model Normal. Estandardització

- La **combinació lineal de variables Normals** és Normal:
  - Sigui a i b, dos escalars i $X \sim N(\mu_X , \sigma_X)$ → $Y = a \cdot X + b \sim N(\boldsymbol{\mu_Y} = a \cdot \mu_X + b, \boldsymbol{\sigma_Y} = a \cdot \sigma_X)$
  - Sigui a i b, dos escalars, $X1 \sim N(\mu_1 , \sigma_1)$ i $X2 \sim N(\mu_2 , \sigma_2)$ →

    → $X = a \cdot X1 + b \cdot X2 \sim N\left(\boldsymbol{\mu_X} = a \cdot \mu_1 + b \cdot \mu_2 , \boldsymbol{\sigma_X} = \sqrt{a^2\sigma_1{}^2 + b^2\sigma_2{}^2 + 2 \cdot ab \cdot \rho_{X1X2} \cdot \sigma_1 \cdot \sigma_2}\right)$

- 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:

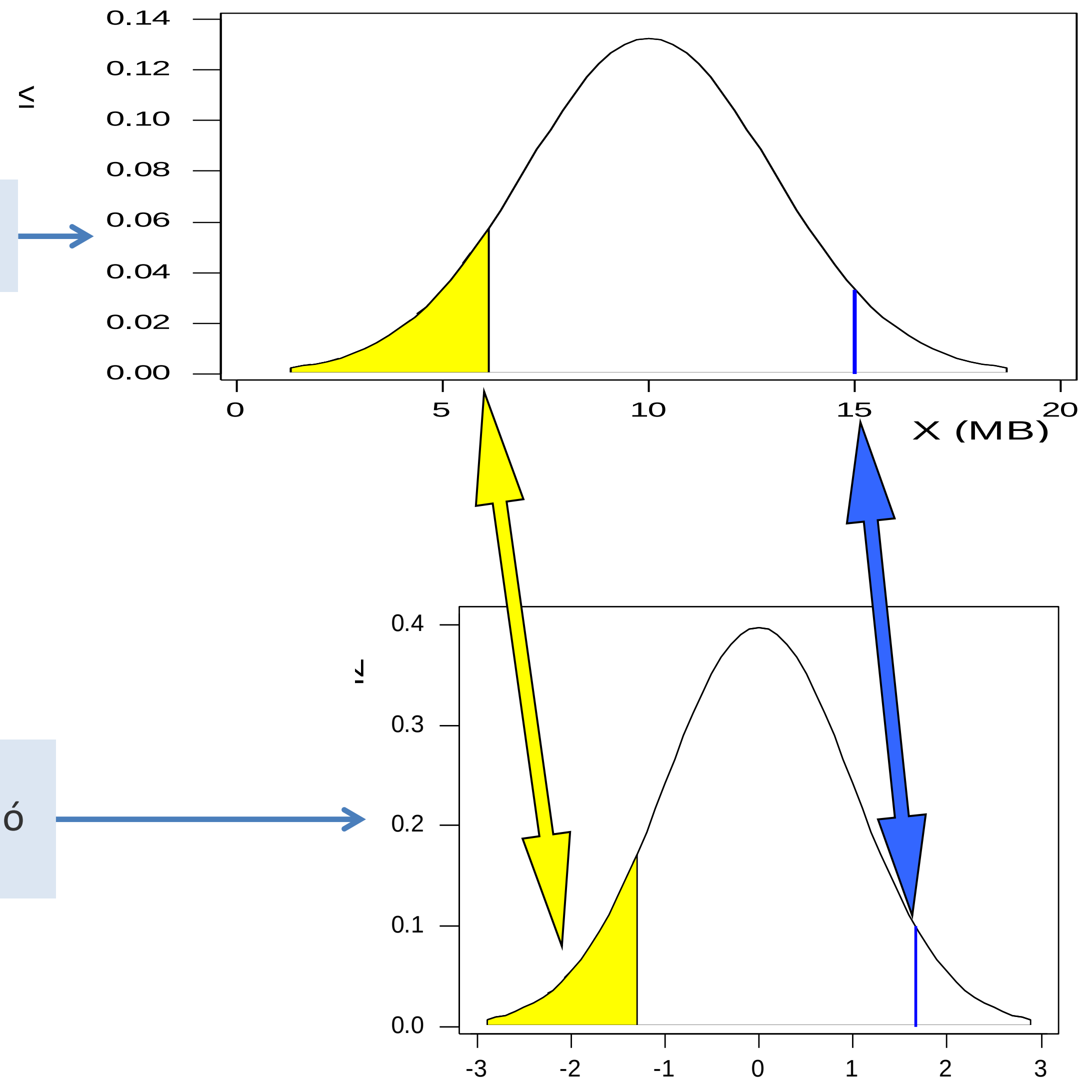
$$\boldsymbol{Z = \frac{X-\mu}{\sigma} \sim N(0, 1) \rightarrow X = \mu + \sigma \cdot Z}$$

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)

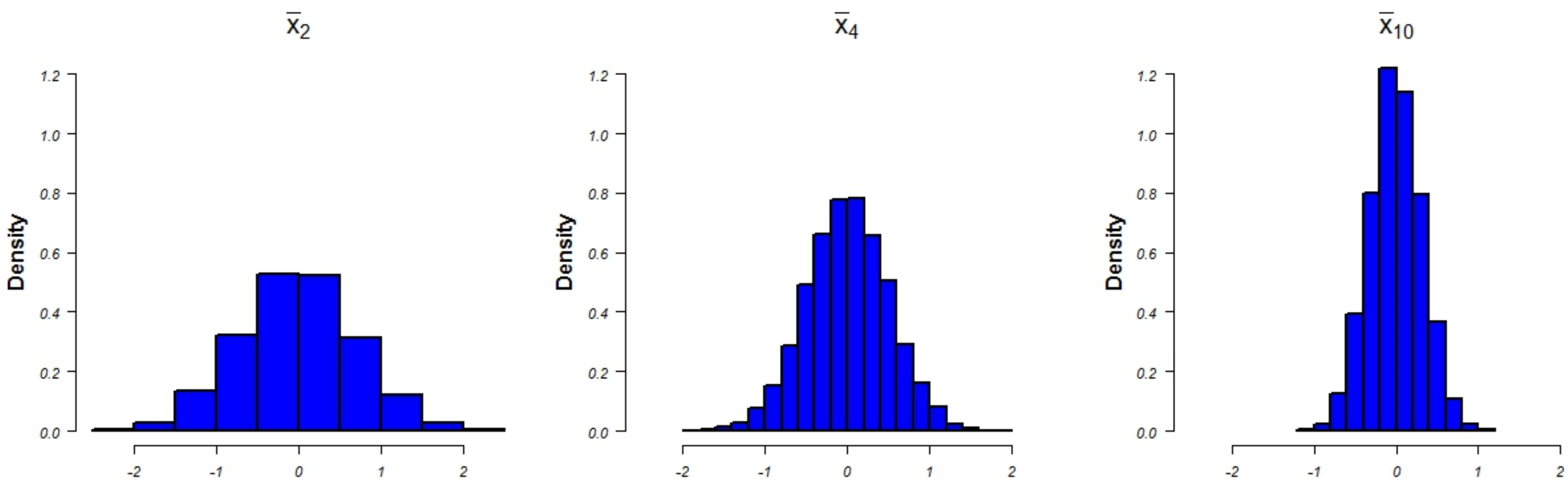


Distribució de la mitjana de v.a.

- 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)

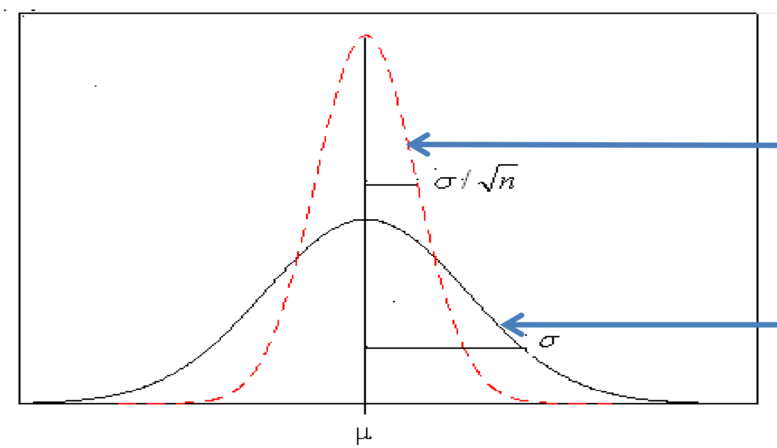
- Siguin X_1, X_2, …, X_n independents i idènticament distribuïdes amb esperança μ i desviació típica σ. Llavors:

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

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

- 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!**
- Veure [app]

Teorema Central del Límit (TCL)

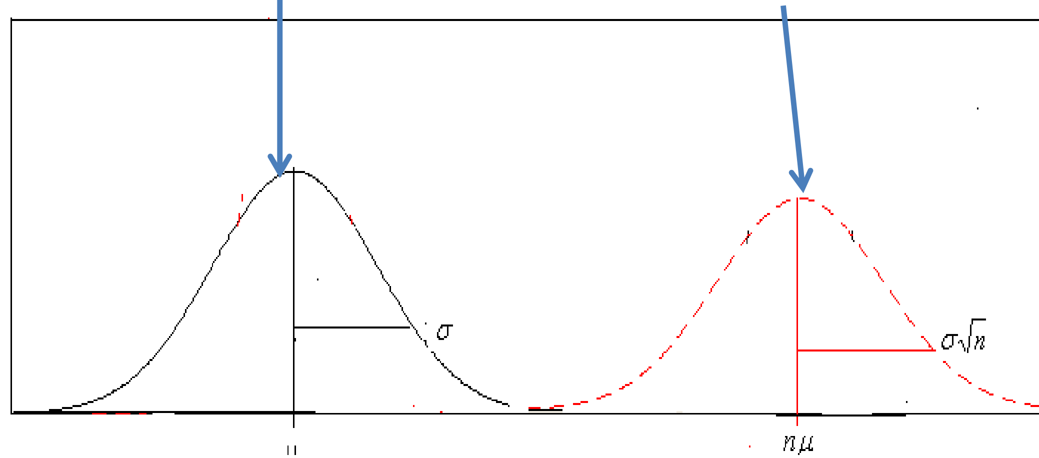


$$\bar{X}_n = \frac{\sum X_i}{n} \xrightarrow[n\ gran]{} N(\mu, \sigma/\sqrt{n})$$

$$X_i$$

$$X_i$$

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



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

però han de tenir la mateixa esperança i variància:

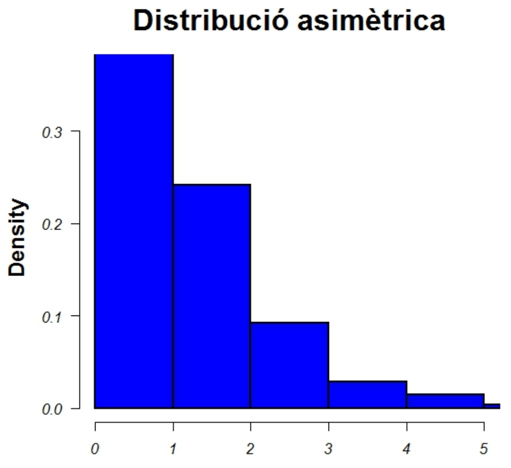
E(X) = μ

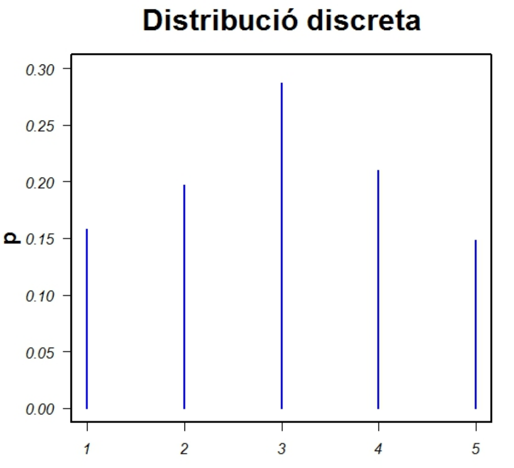
V(X) = σ²

Teorema Central del Límit (TCL). ”n”

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

- Depèn de com sigui la distribució original i de què 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]

TCL. Relacions entre distribucions

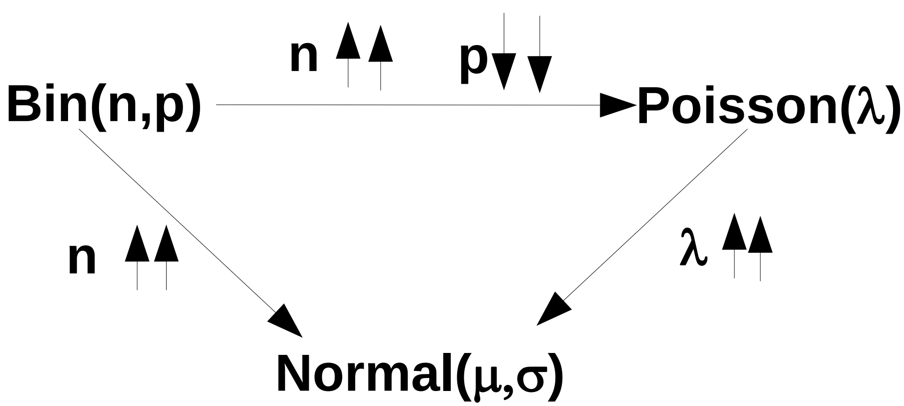
Una de les aplicacions pràctiques del TCL és que la distribució Normal es pot emprar com a aproximació d'altres distribucions:

- La **distribució Binomial** (suma de Bernoullis) 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

  $\mu = n \cdot p$ i $\sigma^2 = n \cdot p \cdot (1-p)$

- 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:

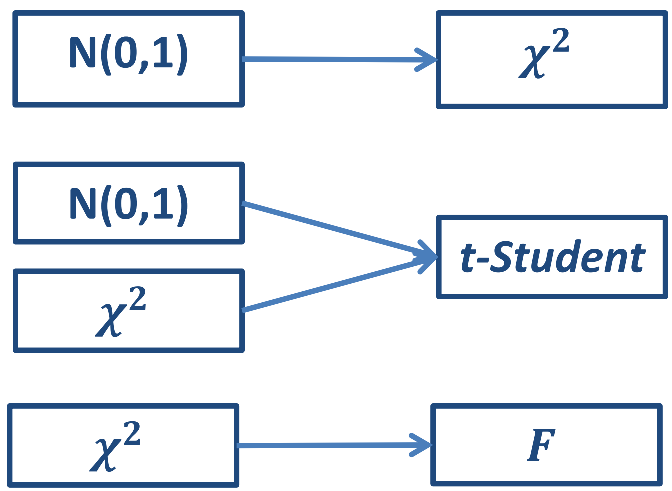
  $\mu = \lambda$ i $\sigma^2 = \lambda$



Nota: la Binomial es pot aproximar a una Poisson quan la *n* és prou gran i la *p* prou petita. Llavors $\lambda = np$

Models derivats de la Normal: χ^2 i *t-Student* i *F*

- Veurem tres distribucions noves que s'usaran a inferència: χ^2 i **t-Student** i ***F***
- Aquestes distribucions provenen de fer operacions amb VA provinents d'altres distribucions, entre elles la Normal estàndard



- A diferència de les distribucions vistes prèviament NO modelen fenòmens de la vida real, sinó el comportament dels estadístics (que veurem que són indicadors calculats a partir d'unes dades) entre les possibles mostres

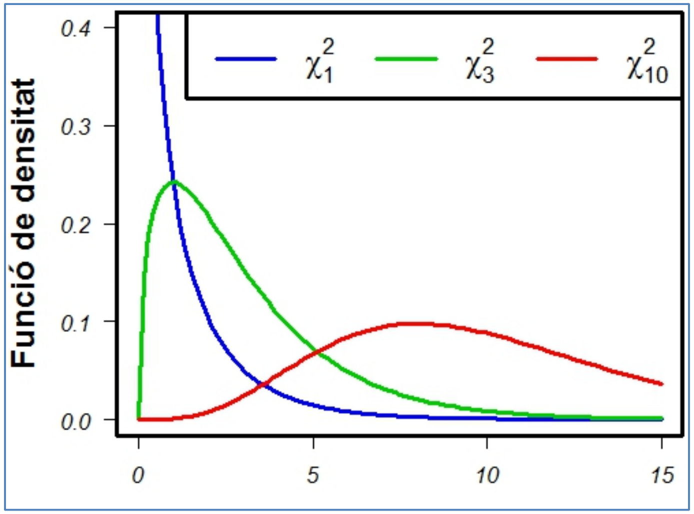
Model derivats de la Normal: χ^2

- **Definició**: Siguin X_i ~ N(0,1). Llavors:

$$X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi_n^2$$

[Concretament, per n = 1 → $X_1^2 \sim \chi_1^2$]

- **Notació**: $X \sim \chi_n^2$
- **Paràmetres**: *n* (graus de llibertat)
- **Funció de probabilitat i distribució**:



$$f(x) = \frac{x^{k/2-1} \cdot e^{-x/2}}{2^{k/2} \cdot \Gamma(k/2)} \quad per\ x > 0$$

$$F(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)} \quad per\ x > 0$$

Γ : funció Gamma
γ : funció Gamma incomplerta
n: graus de llibertat

R: dchisq, pchisq, qchisq

Script per veure que la suma de Normals estàndard al quadrat és una χ^2

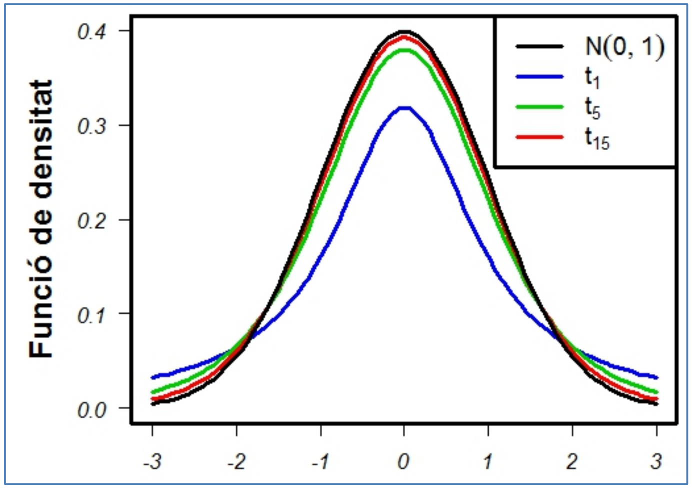
```
M = 500                                  # Mostres de normals
n = 7                                    # Graus de llibertat
sample = array(rnorm(M*n),dim=c(M,n))    # n mostres de N(0,1)
sample2 = sample*sample                  # n mostres de (N(0,1))^2
sum = apply(sample2, 1, sum)             # Suma de les mostres al^2
hist(sum, breaks="Scott", freq=FALSE)    # Distribució empírica sumant Normals
curve(dchisq(x, n),add=TRUE,col=2,lwd=2)# Distribució teòrica de la chi-quadrat
quantile(sum,c(0.25,0.50,0.75))          # Q1, Mediana i Q3 de la suma de Normals
qchisq(c(0.25, 0.50, 0.75),n)            # Q1, Mediana i Q3 de la chi-quadrat
```

Model derivats de la Normal: *t-Student*

- **Definició**: Siguin dues VA independents, $Z \sim N(0,1)$ i $Y_n \sim \chi^2_n$. Llavors:

$$\frac{Z}{\sqrt{Y_n/n}} \sim \mathrm{t_n}$$

[Quan n→∞ (n>30), llavors t_n → N(0,1)]



- **Notació**: $X \sim t_n$
- **Paràmetres**: *n* (graus de llibertat)
- **Funció de probabilitat i distribució**:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad per\ x>0$$

$$F(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n}\cdot\mathrm{B}(1/2,n/2)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad per\ x>0$$

Γ : funció Gamma
B: funció Beta
n: graus de llibertat

R: dt, pt, qt

Script per veure que a partir de una Z i una Y_n s'obté una t

```
M = 500; n = 7                              # Número de mostres i graus de llibertat
samplez = rnorm(M, 0, 1)                    # Mostra de normals
samplechi2 = rchisq(M,n)                    # Mostra de chi-quadrats
samplechi2n = sqrt(samplechi2/n)            # Càlcul dels denominadors
t = samplez / samplechi2n                   # Càlcul de la t-student
hist(t, breaks="Scott", freq=FALSE)         # Distribució empírica
curve(dt(x, n),add=TRUE,col=2,lwd=2)        # Distribució teòrica d'una t-Student
quantile(t, c(0.25, 0.50, 0.75))            # Q1, Mediana i Q3 de Z/sqrt(Yn/n)
qt(c(0.25, 0.50, 0.75),n)                   # Q1, Mediana i Q3 de la chi-quadrat
```

Distribució F de Fisher-Snedecor

- **Definició**: Siguin $X_1 \sim \chi^2_n \quad i \quad X_2 \sim \chi^2_m$. Llavors:

$$Y = \frac{X_1/n}{X_2/m} \sim F_{n,m}$$

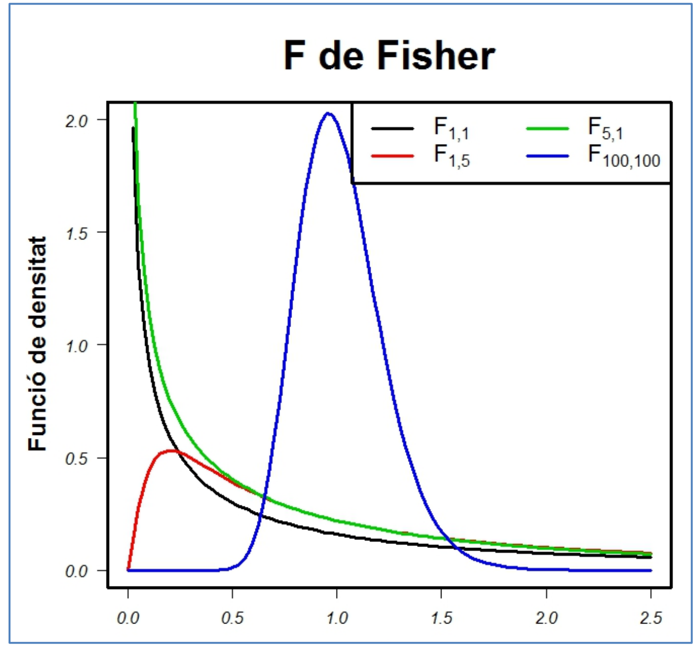
$1/Y \sim F_{m,n}$

- **Notació**: $F \sim F_{n,m}$
- **Paràmetres**: *n* (graus de llibertat numerador)

  m (graus de llibertat denominador)

- **Funció de probabilitat i distribució**:

  F distribution at Wikipedia



NOTA: La distribució F de Fisher, la farem servir per comparar variàncies de 2 poblacions

R: df, pf, qf

Script per veure que el quocient de χ^2 dividits pels g.ll és una F

```
M=500 ; n=5; m=7
samplechi2n = rchisq(M,n)
samplechi2m = rchisq(M,m)
F = (samplechi2n/n) / (samplechi2m/m)
hist(F, breaks="Scott", freq=FALSE)
curve(df(x, n, m),add=TRUE,col=2,lwd=2)
quantile(F, c(0.25, 0.50, 0.75))
qf(c(0.25, 0.50, 0.75), n, m)
```

Probabilitats i quantils de models de VA usant R

```
# X es Bin(n=10,p=0.4)
dbinom(5,10,0.4)      # P(X=5)          → 0.2006581
pbinom(5,10,0.4)      # P(X<=5)         → 0.8337614
qbinom(0.5,10,0.4)    # P(X<=?)=0.5     → 4
# Y es Poi(lambda=4)
dpois(5,4)             # P(Y=5)         → 0.1562935
ppois(5,4)             # P(Y<=5)        → 0.7851304
qpois(0.5,4)           # P(Y<=?)=0.5    → 4
# E es exp(lambda=4)
pexp(1,4)              # P(E<=1)        → 0.9816844
qexp(0.5,4)           # P(E<=?)=0.5     → 0.1732868
# Z es N(0,1)
pnorm(1.96)            # P(Z<=1.96)     → 0.975
qnorm(0.95)           # P(Z<=?)=0.95 → ? = 1.645 = z_0.95
```

Funcions en R,
o bé instruccions en R online:
https://rdrr.io/snippets/

```
# T es t15
pt(1.96,15)           # P(T<=1.96)      → 0.9655779
qt(0.95,15)          # P(T<=?)=0.95     → ? = 1.753 = t_15,0.95
# X es Chi10
pchisq(5,10)          # P(X<=5)         → 0.108822
qchisq(0.95,10)      # P(X<=?)=0.95     → ? = 18.307
# F es F1,5
pf(1,1,5)             # P(F<=1)         → 0.6367825
qf(0.95,1,5)         # P(F<=?)=0.95     → ? = 6.607891
```

```
# anàlisis gràfica de normalitat
#(per ex 30 valors generats de Normal
# o d'Exponencial)
x = rnorm(100)        # o x = rexp(100)
qqnorm(x)             # i afegir qqline(x)
```