Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Models de VA i simulació
Annex

Bloc B – Probabilitat i Estadística

2023

Índex

Mostra Aleatòria Simple (MAS)

Sigui la VA: $X: \Omega \rightarrow R$

$\omega_i \rightarrow X(\omega_i) = x_i$

Direm que M.A.S. de grandària **n** de la v.a. X

és una funció vectorial $M = (X_1, X_2, \ldots, X_n)$ tal que

$M: \Omega^n \rightarrow R^n$

$\omega = (\omega_1, \omega_2, \ldots \omega_n) \rightarrow M(\omega_i) = (X_1, X_2, \ldots, X_n)$

Direm que és una MAS si i només si es compleixen les dues condicions següents:

(1) **Tots els elements** de la població tenen la **mateixa probabilitat** de pertànyer a la mostra.

(2) **Qualsevol combinació** de *n* elements té la **mateixa probabilitat** de pertànyer a la mostra.

La informació aportada per les diferents unitats ha de ser **independent** entre sí:

- les X_i han de ser VA independents i idènticament distribuïdes (i.i.d.)

R té funcions per generar mostres aleatòries simples seguint un model. Per exemple,

R: rbinom(), rpois(), rexp(), rnorm()

Descriptiva (gràfica)

L'**Estadística Descriptiva** permet resumir dades gràficament (i també numèricament tal com es veurà relacionant-ho amb inferència estadística)

En la següent taula hi ha algunes funcions (bàsiques) en R per **Estadística Descriptiva gràfica** en dades discretes i continues de forma univariant o bivariant:

	UNIVARIANT VAC	UNIVARIANT VAD	BIVARIANT
GRÀFIQUES	hist( ) boxplot( )	barplot(table( ))	plot(,)

(més funcions gràfiques en R: https://www.r-graph-gallery.com/)

Simulació de VA amb R

R té diverses funcions per a generar valors aleatòriament:

`sample(x, n, replace=y)` **n** valors del conjunt **x**, amb reemplaçament (**y**=T) o no (**y**=F).
També podem assignar probabilitats particulars a cada element de **x**.
Exemple: `sample(1:10, 25, replace=TRUE)`

Generalment, volem reproduir valors d'un model probabilístic particular: Poisson, Normal, uniforme, ... Tots els models vistos (tots els que té implementats R, en realitat, apart de les funcions `d...`, `p...` i `q...` tenen la funció `r...` per simular aquests models.

Exemples:

`rnorm(1):` 1 valor a l'atzar d'un model N(0,1)

`rpois(5, 2.5):` 5 valors a l'atzar d'un model P(λ=2.5)

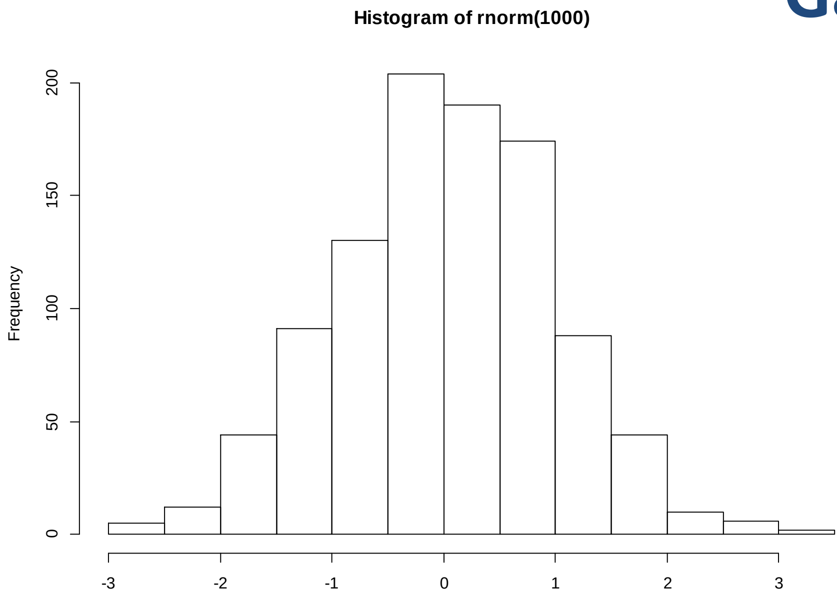
`rbinom(50, 30, 0.1):` 50 valors a l'atzar d'un model B(n=30, p=0.1)

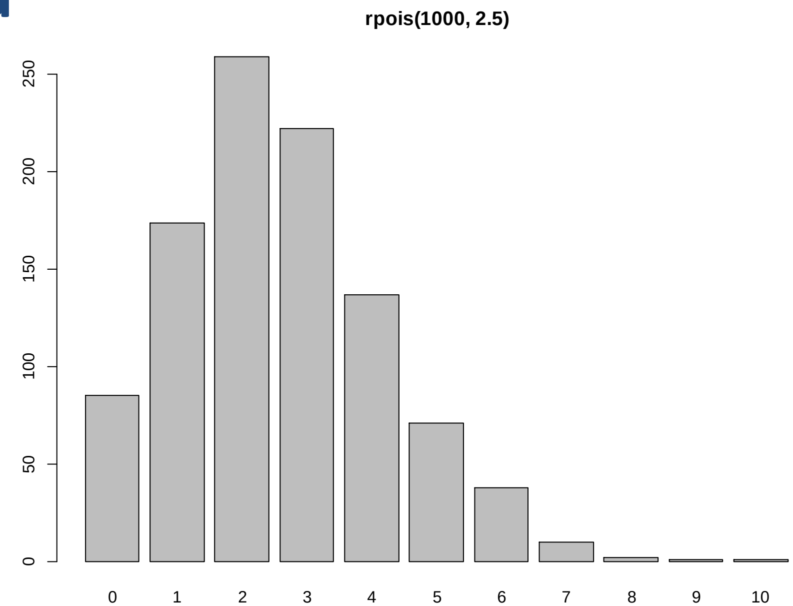
`runif(3, -1, 1):` 3 valors a l'atzar d'un model U(-1,1)

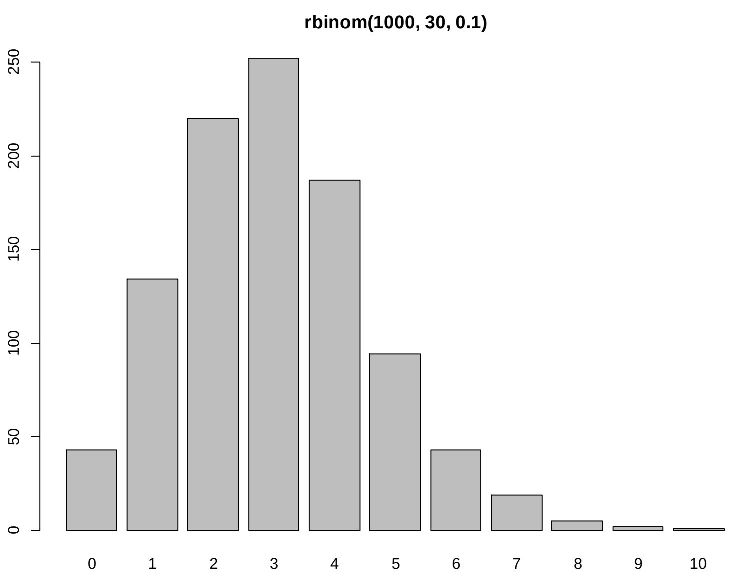
`rnbinom(1, 5, 0.7):` 1 valor a l'atzar d'un model BN(r=5, p=0.7)

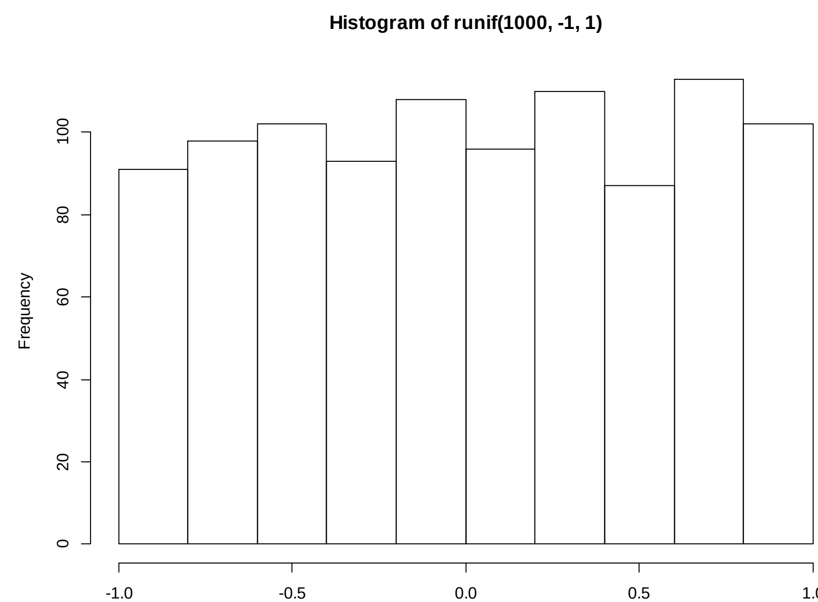
compte: R contempla el nombre de fracassos abans de r èxits, no el nombre total d'intents

Galeria


Histogram of rnorm(1000)
Frequency
0
50
100
150
200
-3
-2
-1
0
1
2
3


rpois(1000, 2.5)
0
50
100
150
200
250
0
1
2
3
4
5
6
7
8
9
10


rbinom(1000, 30, 0.1)
0
50
100
150
200
250
0
1
2
3
4
5
6
7
8
9
10


Histogram of runif(1000, -1, 1)
Frequency
0
20
40
60
80
100
-1.0
-0.5
0.0
0.5
1.0

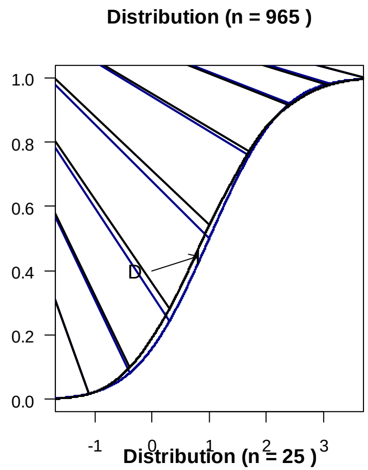
Anàlisis gràfica de Normalitat

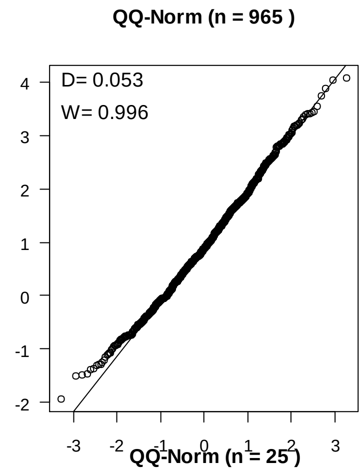
- Hi ha dues mesures que ajuden a valorar el grau d'ajustament, afinitat o similitud a una certa distribució de referència.

Kolmogorov-Smirnov (Estadístic D)	Shapiro-Wilk (Estadístic W)
Distància màxima entre la funció de distribució empírica i la teòrica. Valors elevats indiquen No Normalitat Entre 0 i 1 (usualment, prop de 0). Valors alts indiquen des-ajustament	Mesura la correlació entre els quantils observats i els teòrics. Valors elevats indiquen Normalitat Entre 0 i 1 (usualment, prop de 1). Valors alts indiquen bon ajustament

- Totes dues mesures fluctuen a les mostres i han de interpretar-se amb prudència. En farem només una anàlisi descriptiva i visual (qqplot), especialment respecte la distribució Normal (QQnorm)
- A continuació mostrem alguns exemples generats a partir de distribucions conegudes i amb diferents mides mostrals.

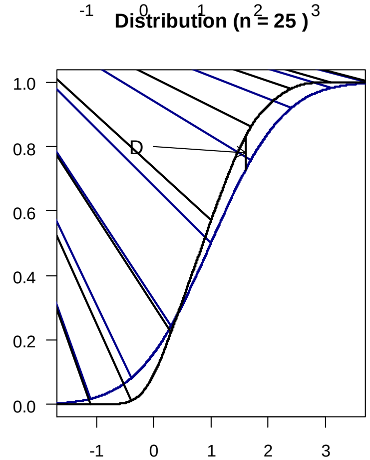
Anàlisis gràfica de Normalitat

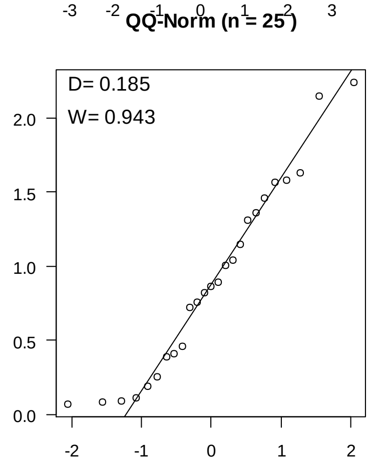




- Aquestes 965 observacions van estar generades seguint una distribució Normal
- D, W i el QQ-Norm mostren que les dades s’ajusten prou bé a una Normal

```
x <- rnorm(965)
qqnorm(x)
qqline(x)
```



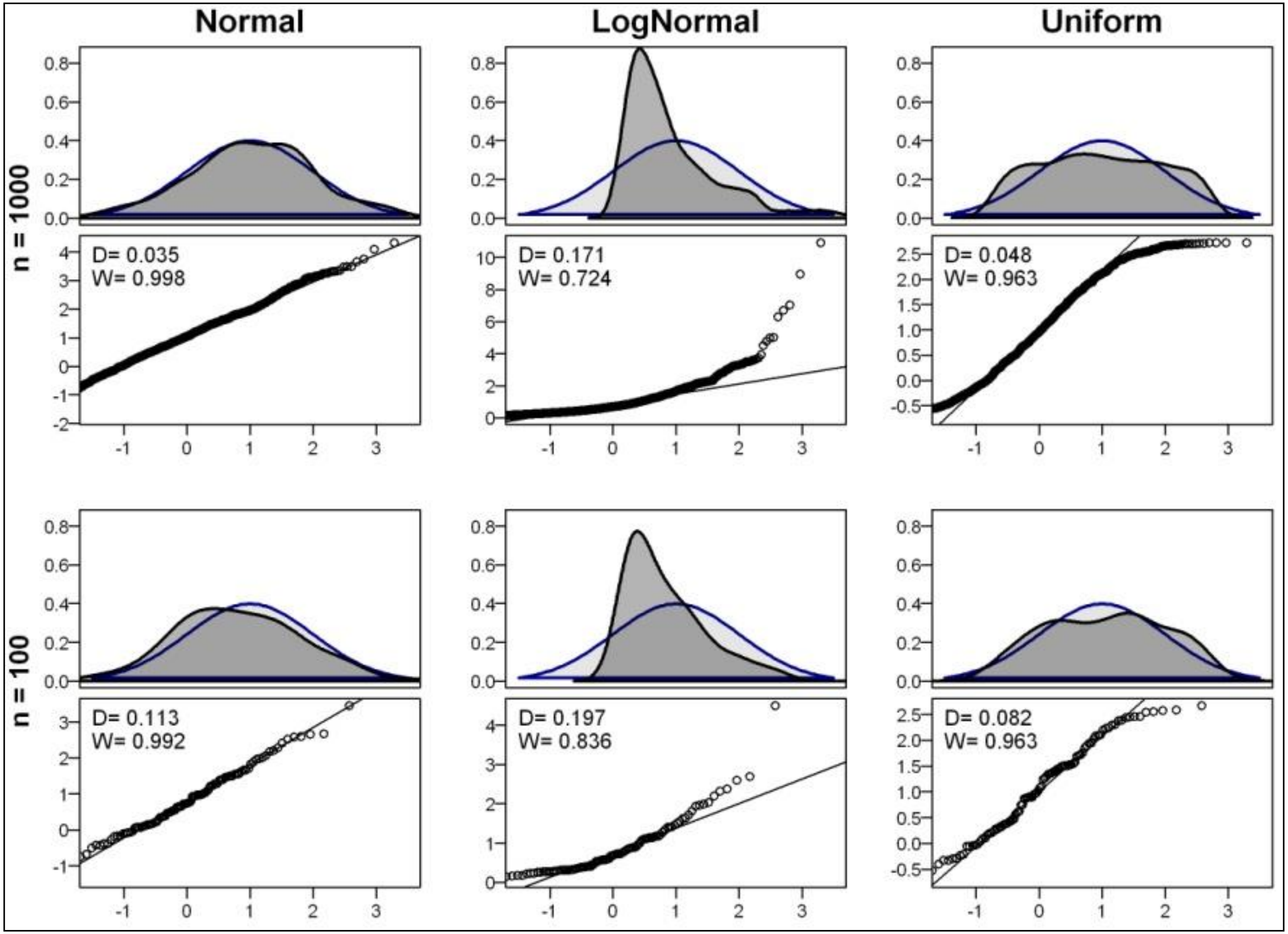


- Aquestes 25 observacions van estar generades seguint una distribució Exponencial
- F_X i D mostren la distància amb la teòrica, encara que W té una bona correlació.

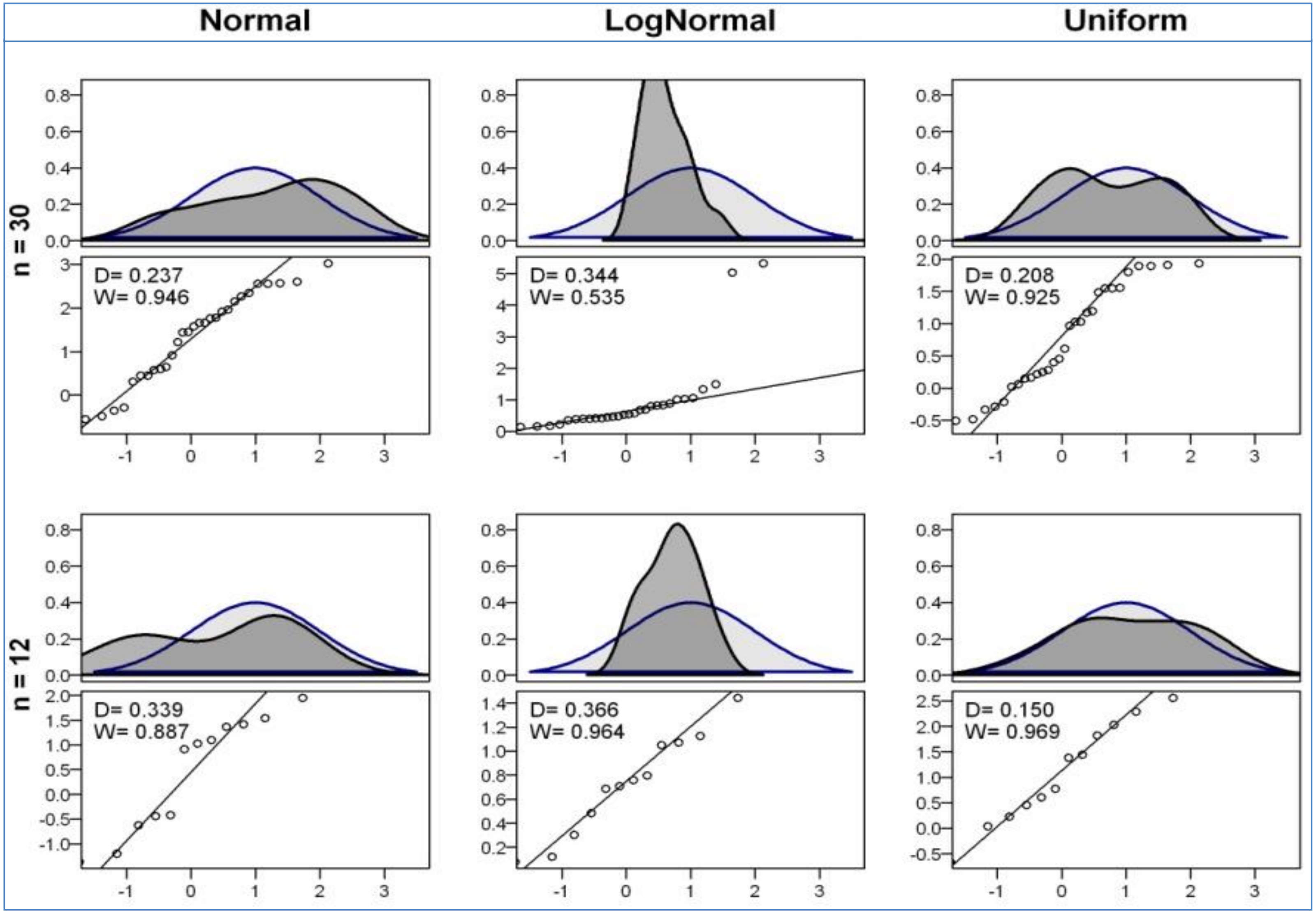
```
x <- rexp(25)
qqnorm(x)
qqline(x)
```

Noteu la paradoxa: quan la mostra és petita, és important detectar la No Normalitat. En canvi, quan la mostra és gran, és poc important detectar-la (pel bon comportament asimptòtic)

Anàlisis gràfica de Normalitat (n gran)



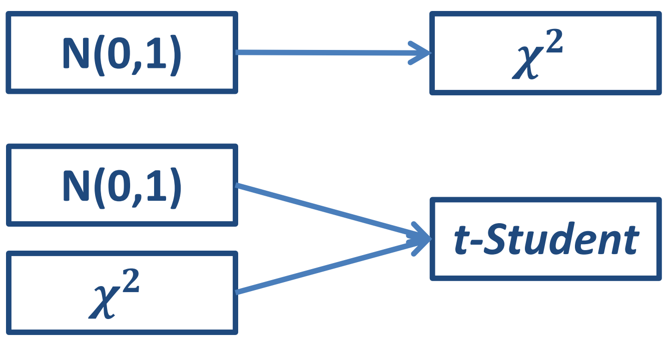
Anàlisis gràfica de Normalitat (n petita)



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

- Hi ha un parell de distribucions noves que s'usaran a inferència: χ^2 i **t-Student**

- 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 entre les possibles mostres.

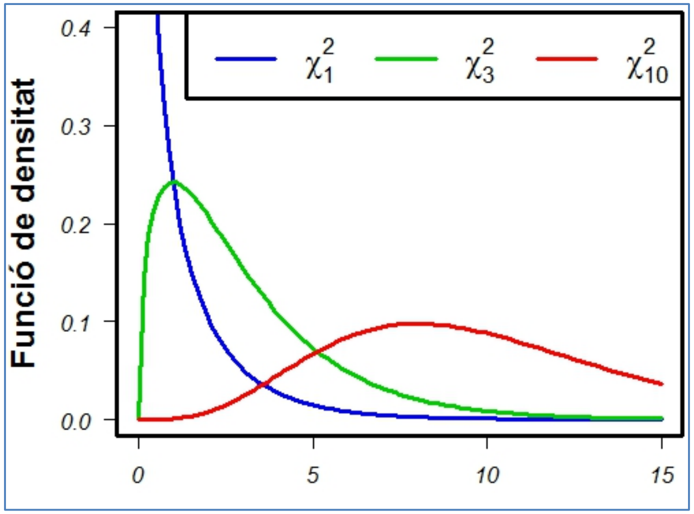
Model derivats de la Normal: χ^2

- **Definició**: Siguin $X_i \sim N(0,1)$. 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**:



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

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

Γ : funció Gamma
γ : funció Gamma incompleta
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:

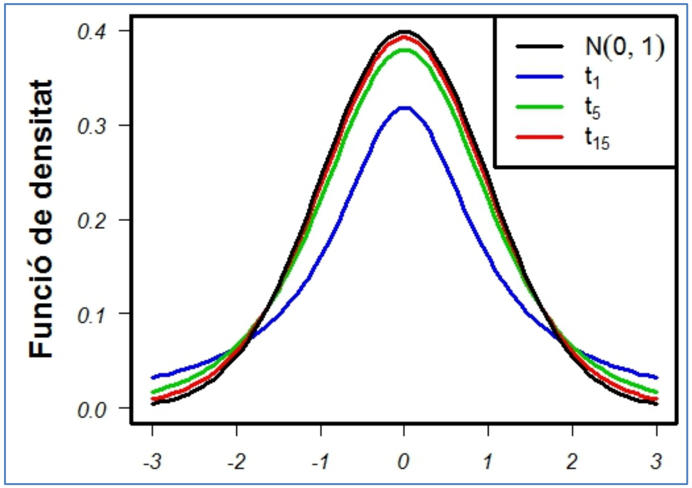
$$\frac{Z}{\sqrt{Y_n/n}} \sim \mathrm{t_n}$$

[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 $X_1 \sim \chi^2_n$ i $X_2 \sim \chi^2_m$. Llavors:

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

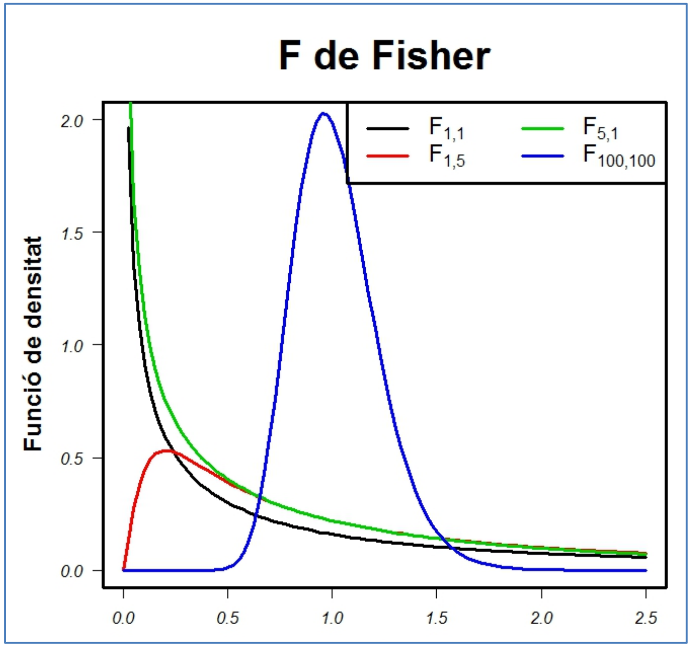
$1/Y \sim F_{m,n}$

- **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)
pbinom(5,10,0.4)       # P(X<=5)
qbinom(0.5,10,0.4)     # P(X<=?)=0.5
# Y es Poi(lambda=4)
dpois(5,4)             # P(Y=5)
ppois(5,4)             # P(Y<=5)
qpois(0.5,4)           # P(Y<=?)=0.5
# E es exp(lambda=4)
pexp(1,4)              # P(E<=1)
qexp(0.5,4)            # P(E<=?)=0.5
# Z es N(0,1)
pnorm(1.96)            # P(Z<=1.96)
qnorm(0.95)            # P(Z<=?)=0.95 (?= z_0.95)

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

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