



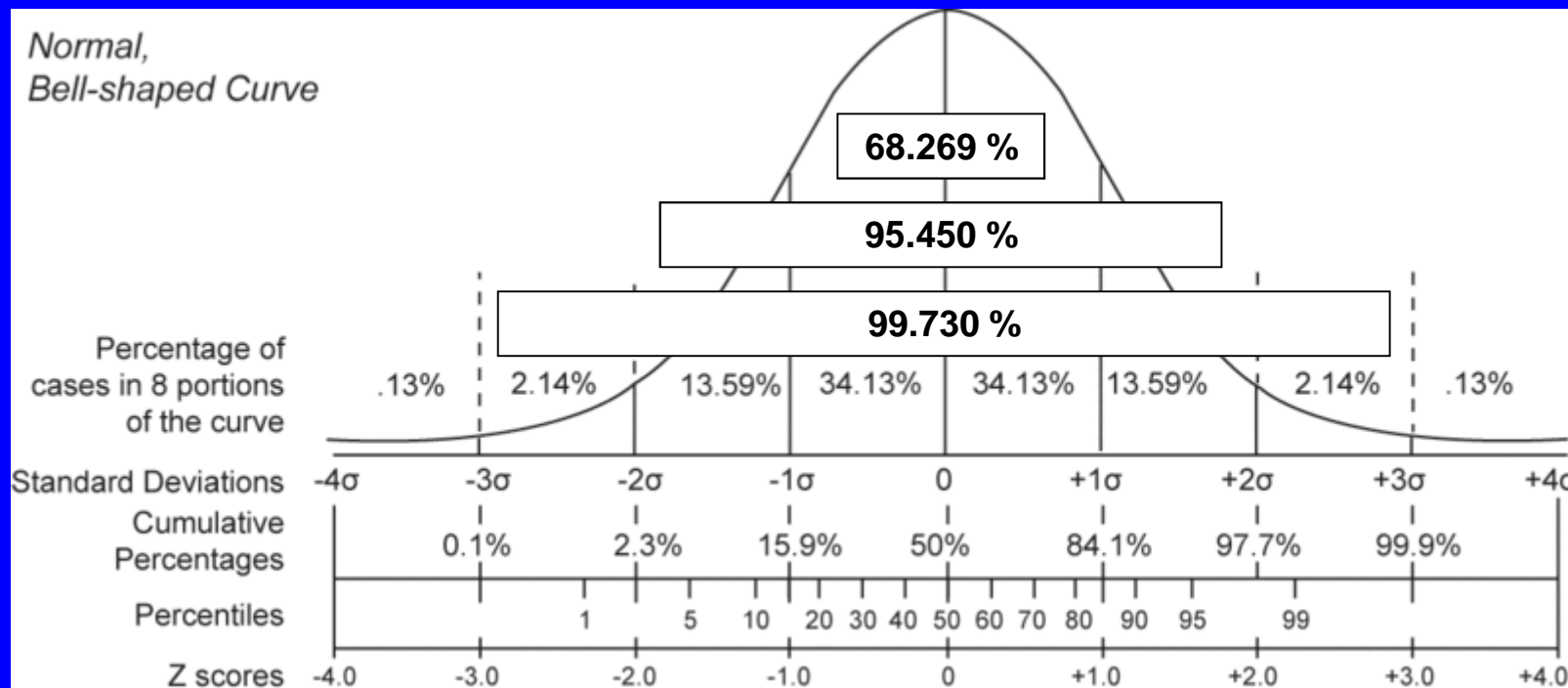
# Introducció

En l'àmbit de la Teledetecció i els SIG és freqüent realitzar anàlisis estadístiques. En aquest context hi ha alguns temes que apareixen de manera freqüent:

- En moltes anàlisis cal assumir que les variables es distribueixen de forma normal; tanmateix, convé fer algunes reflexions al respecte: **Caracterització de distribucions i verificació de normalitat.**
- En moltes anàlisis cal assumir que no existeix una correlació elevada entre "variables independents": **Correlació en imatges de teledetecció.**
- En moltes anàlisis pot ser útil reduir el nombre de variables independents i treballar amb un subconjunt menor però amb similar poder informatiu: **Anàlisi de Components Principals.**
- En moltes anàlisis cal fer **càlculs de Distàncies estadístiques.**
- En moltes anàlisis cal tenir present l'**Autocorrelació espacial.**



# Caracteritzar distribucions. Recordatori de la normal univariant



Font: [http://en.wikipedia.org/wiki/Standard\\_score](http://en.wikipedia.org/wiki/Standard_score), modificada



## Caracteritzar distribucions

- Obriu un full MS-Excel en blanc i genereu una columna de 30000 valors aleatoris en el rang  $[-5,5)$ . Mostreu-los amb 4 decimals

**Com?** Invoqueu la funció d'Excel que genera un valor aleatori: ALEATORIO(). Com que F9 (o reobrir el fitxer, etc) refà la sèrie, convé convertir les fórmules en valors (copiar+enganxat especial).

- Comproveu visualment, mitjançant un histograma, que la distribució és realment aleatòria i no segueix, per exemple, una normal.

**Com?** Escriviu, en una columna nova, els rangs de dades: des de -5.5 a 4.5 en salts de valor 1.0. Després aneu a "Herramientas | Análisis de datos | Histograma" (si no hi fos, afegiu-la amb "Herramientas | Complementos..."). Segons la versió d'Excel, indiqueu "Crear gráfico", "En una hoja nueva"



## Caracteritzar distribucions

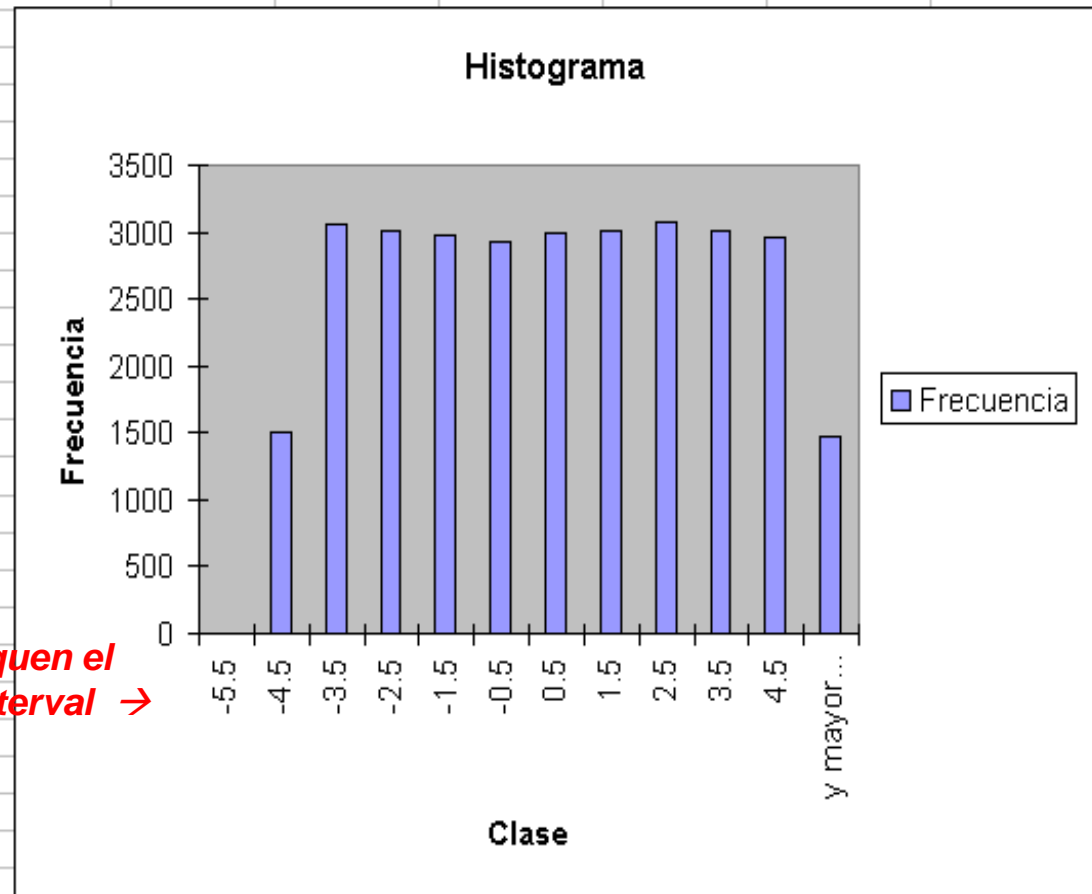
- A títol informatiu, calculeu N, mínim, màxim, mitjana i desviació estàndard d'aquesta població.

**Com?** Contar(), Min(), Max(), Promedio() i DesvEstP() [divideix entre n].



	A	B	C	D	E	F	G	H	I	J	K
1	<b>Sèrie aleatòria entre -5 i 5</b>										
2	<b>=10*(ALEATORIO()-0.5)</b>										
3	-4.0752										
4	-0.5357	<b>N:</b>	<b>30000</b>								
5	0.7661	<b>Mínim:</b>	<b>-4.9987</b>								
6	-3.5881	<b>Màxim:</b>	<b>4.9999</b>								
7	3.2710	<b>Mitjana:</b>	<b>-0.0095</b>								
8	-3.4843	<b>DesvEstàndard:</b>	<b>2.8874</b>								
9	2.4089										
10	0.3657	<b>Classes proposa</b>	<i>Clase</i>	<i>Frecuencia</i>							
11	3.4172	-5.5	-5.5	0							
12	2.4067	-4.5	-4.5	1508							
13	-1.8225	-3.5	-3.5	3056							
14	3.1081	-2.5	-2.5	3012							
15	-3.1085	-1.5	-1.5	2979							
16	3.3092	-0.5	-0.5	2930							
17	2.0153	0.5	0.5	2997							
18	4.2483	1.5	1.5	3002							
19	4.6105	2.5	2.5	3071							
20	1.8187	3.5	3.5	3012							
21	3.6233	4.5	4.5	2967							
22	-3.5957		y mayor...	1466							
23	-1.7173										
24	-3.0595										
25	4.7687										
26	-4.8182										
27	2.7295										
28	-1.9265										
29	2.4162										
30	1.2890										
31	1.4818										
32	4.4222										
33	2.3365										
34	-0.9518										

Noteu que els dos extrems són ~1/2 perquè l'interval només té una amplitada 0.5



Aquests valors indiquen el límit superior de l'interval →



- A través de l'aplicació `AleatNormal_AmbParametres.exe`, genereu 30000 valors normalment distribuïts al voltant d'una mitjana 0 i una desviació estàndard 1.5. Inclogueu-los en un full MS-Excel en blanc com heu fet en l'exercici anterior.

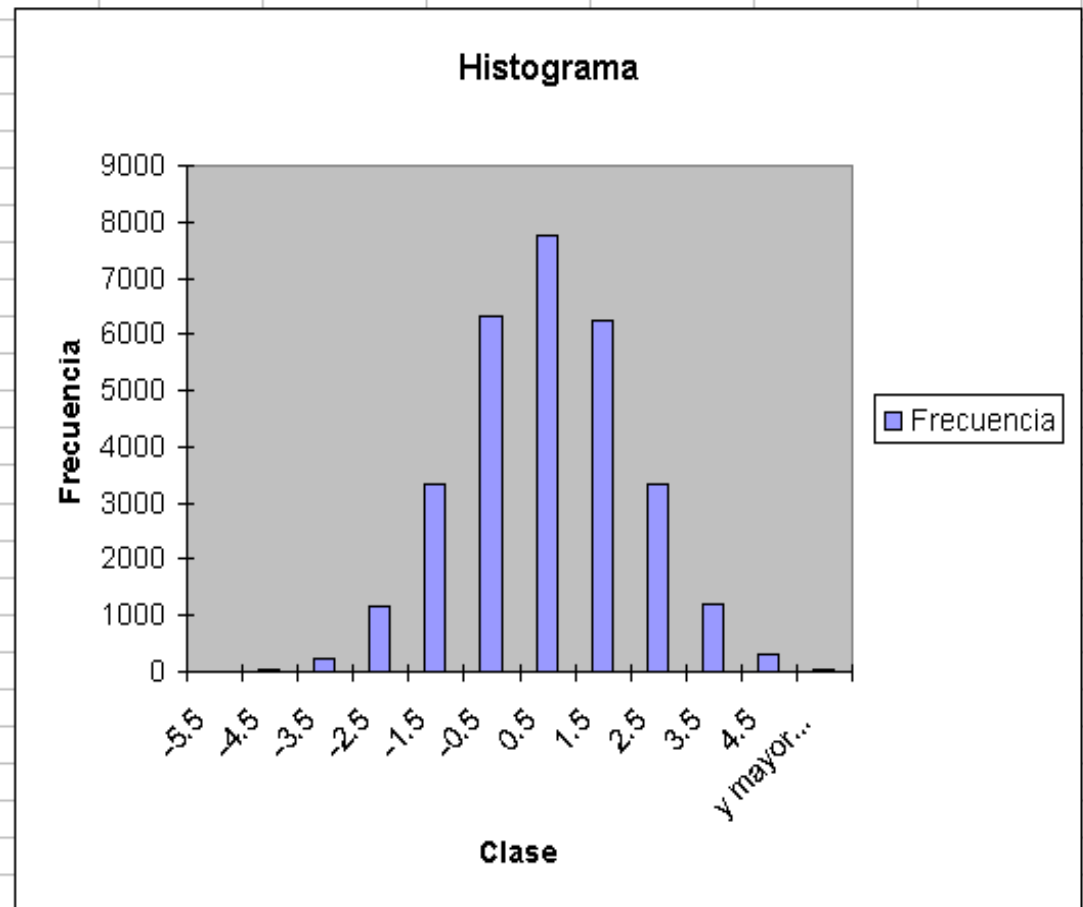
**Com?** Redirigeu la sortida del programa a un fitxer de text, obriu-lo i, a través del portapapers, copieu els valors a Excel:

```
AleatNormal_AmbParametres 0 1.5 30000 4 > resu.txt
```

- Comproveu visualment, mitjançant un histograma, que la distribució segueix realment una normal.
- Calculeu N, mínim, màxim, mitjana i desviació estàndard d'aquesta mostra imaginària.



	A	B	C	D	E	F	G	H	I	J	K
1	<b>Sèrie normal de mitjana 0 i desviació estàndard 1.5</b>										
2	<b>AleatNormal_AmbParametres 0 1.5 30000 4 &gt; resu.txt</b>										
3	-1.2553										
4	-0.2584	<b>N:</b>	<b>30000</b>								
5	0.2807	<b>Mínim:</b>	<b>-5.5962</b>								
6	2.4232	<b>Màxim:</b>	<b>6.9732</b>								
7	-0.2652	<b>Mitjana:</b>	<b>0.0079</b>								
8	0.9797	<b>DesvEstàndard:</b>	<b>1.5107</b>								
9	-0.8195										
10	0.2912	<b>Classes proposa</b>	<i>Clase</i>	<i>Frecuencia</i>							
11	1.3886	-5.5	-5.5	1							
12	1.8065	-4.5	-4.5	43							
13	2.2958	-3.5	-3.5	246							
14	-2.0333	-2.5	-2.5	1156							
15	0.0772	-1.5	-1.5	3325							
16	1.5303	-0.5	-0.5	6331							
17	-1.8392	0.5	0.5	7765							
18	1.0627	1.5	1.5	6253							
19	1.3075	2.5	2.5	3352							
20	-1.1846	3.5	3.5	1198							
21	0.4981	4.5	4.5	296							
22	0.3084	y mayor...		34							
23	-0.2541										
24	-0.4776										
25	-0.4435										
26	0.7834										
27	-3.3466										
28	0.3874										
29	-0.1317										
30	-2.4703										
31	0.4302										
32	0.4500										
33	1.6559										
34	1.1141										





## Verificar la normalitat de les dues distribucions anteriors

- Coneixem les  $k$  freqüències dels histogrames calculats (12 en els nostres exemples).
- Ens basarem en els problemes 7.33 i 12.13 del llibre d'Spiegel (1991) "Estadística" [McGraw Hill, 2<sup>a</sup> Ed.]. La idea és comparar el grau de semblança entre la freqüència ABSOLUTA de cada columna de l'histograma i la que tindria una normal teòrica amb la mitjana i desviació estàndard que es dedueix de la mostra que tenim.  
**Com?** Invoqueu la funció d'Excel que calcula l'àrea acumulada des de  $-\infty$  fins al valor desitjat: `DISTR.NORM(Valor; Mitjana; DesvEst.;1)`.





La decisió final sobre si acceptem o no que s'ajusta a una normal la prendrem fent un test de  $X^2$ . A tal efecte, recordeu que:

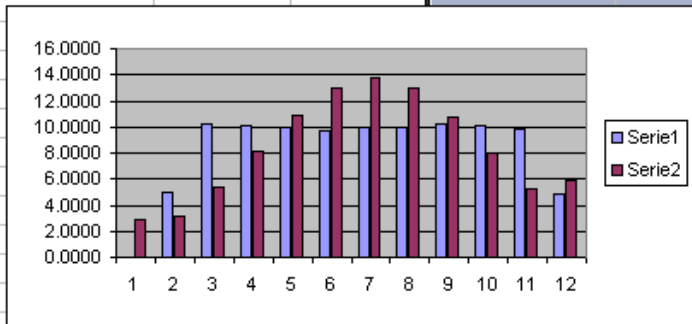
1. Es planteja que la hipòtesi nul·la,  $H_0$ , és que les freqüències observades i esperades no difereixen significativament.
2. Si  $X^2=0$  vol dir que les freqüències observades i les teòriques coincideixen completament
3. El nombre de graus de llibertat,  $\mu$ , val  $k-1-m$ , on  $m$  és el nombre de paràmetres que hem estimat de la població (2 en una normal).  $k$  és el nombre de mostres.
4. El valors crítics (de tall) depenen dels graus de llibertat i del nivell de significació<sup>(1)</sup>,  $\alpha$ , desitjat (compte perquè hi ha taules on els típics nivells de significació 0.05 i 0.01 cal buscar-los a  $X^2_{.95}$  i  $X^2_{.99}$ ).

**Com?**  $k=\text{CONTAR}()$ ;  $X^2_{.95}=\text{PRUEBA.CHI.INV}(0.05; k-1-2)$

<sup>(1)</sup>  $\alpha$ : Probabilitat de cometre un error de Tipus I (rebutjar  $H_0$  quan era certa).  
 $\downarrow \alpha \rightarrow \downarrow \text{prob. eT1}$  però  $\text{prob. eT2}$  (acceptar  $H_0$  falsa)



	A	B	C	D	E	F	G	H
1	Sèrie aleatòria entre -5 i 5							
2	=10*(ALEATORIO(0.5))							
3	-4.0752							
4	-0.5357	N:	30000					
5	0.7661	Mínim:	-4.9987					
6	-3.5881	Màxim:	4.9999					
7	3.2710	Mitjana:	-0.0095					
8	-3.4843	DesvEstàndard:	2.8874					
9	2.4089							
10	0.3657	Classes proposa	Clase	Frecuencia	Freq. observada (%)	Freq. esperada (%)	(Obs-Esp)^2/Esp	
11	3.4172	-5.5	-5.5	0	0.0000	2.8617	2.8617	
12	2.4067	-4.5	-4.5	1508	5.0267	3.1334	1.1440	
13	-1.8225	-3.5	-3.5	3056	10.1867	5.3408	4.3968	
14	3.1081	-2.5	-2.5	3012	10.0400	8.0840	0.4733	
15	-3.1085	-1.5	-1.5	2979	9.9300	10.8659	0.0806	
16	3.3092	-0.5	-0.5	2930	9.7667	12.9699	0.7911	
17	2.0153	0.5	0.5	2997	9.9900	13.7478	1.0271	
18	4.2483	1.5	1.5	3002	10.0067	12.9406	0.6652	
19	4.6105	2.5	2.5	3071	10.2367	10.8169	0.0311	
20	1.8187	3.5	3.5	3012	10.0400	8.0293	0.5035	
21	3.6233	4.5	4.5	2967	9.8900	5.2927	3.9932	
22	-3.5957	y mayor...		1466	4.8867	5.9170	0.1794	
23	-1.7173						Suma=	16.1471
24	-3.0595						X2.90=	14.6837
25	4.7687						X2.95=	16.9190
26	-4.8182						X2.99=	21.6660
27	2.7295							
28	-1.9265							
29	2.4162							
30	1.2890							
31	1.4818							
32	4.4222							
33	2.3365							
34	-0.9518							



=100\*D11/\$C\$4

=100\*DISTR.NORM(C11;\$C\$7;\$C\$8;1)

=100\*(DISTR.NORM(C12;\$C\$7;\$C\$8;1) - DISTR.NORM(C11;\$C\$7;\$C\$8;1))

=100\*(DISTR.NORM(C20;\$C\$7;\$C\$8;1) - DISTR.NORM(C21;\$C\$7;\$C\$8;1))

o bé

=100\*(1 - DISTR.NORM(C21;\$C\$7;\$C\$8;1))

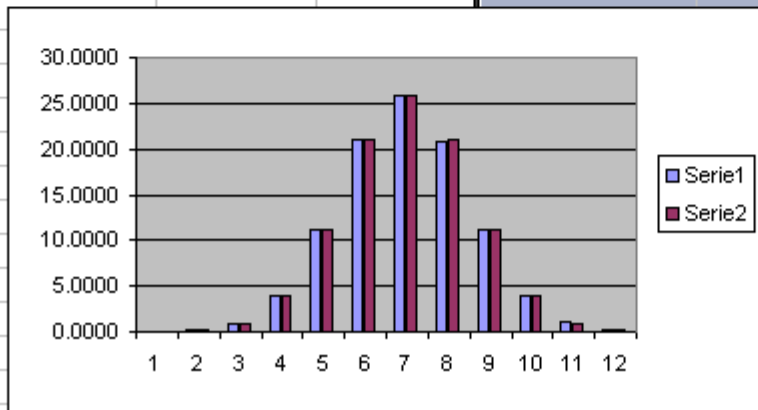
=PRUEBA.CHI.INV(0.01;CONTAR(\$G\$11:\$G\$22)-1-2)

Observeu que si exigim un 5% o un 1% de probabilitat de rebutjar  $H_0$  i que fos certa, no ens atrevim a dir que no és normal. En canvi, podem dir que no és normal acceptant un 10% de probabilitats que ho sigui.

A aquest artefacte hi contribueixen les dues freqüències "tallades" a 1/2 als dos extrems (que simulen les cues de la normal) i que fins i tot es comparen amb la cua acumulada dreta, i el fet que no hi hagi cap dada en les probabilitats més extremes (a les dues cues).

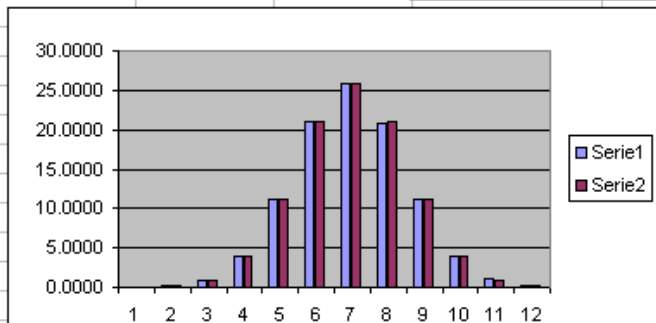


	A	B	C	D	E	F	G	H
1	<b>Sèrie normal de mitjana 0 i desviació estàndard 1.5</b>							
2	<b>AleatNormal_AmbParametres 0 1.5 30000 4 &gt; resu.txt</b>							
3	-1.2553							
4	-0.2584	<b>N:</b>	<b>30000</b>					
5	0.2807	<b>Mínim:</b>	<b>-5.5962</b>					
6	2.4232	<b>Màxim:</b>	<b>6.9732</b>					
7	-0.2652	<b>Mitjana:</b>	<b>0.0079</b>					
8	0.9797	<b>DesvEstàndard:</b>	<b>1.5107</b>					
9	-0.8195							
10	0.2912	<b>Classes proposa</b>	<i>Clase</i>	<i>Frecuencia</i>	<b>Freq. observada (%)</b>	<b>Freq. esperada (%)</b>	<b>(Obs-Esp)^2/Esp</b>	
11	1.3886	-5.5	-5.5	1	0.0033	0.0133	0.0075	
12	1.8065	-4.5	-4.5	43	0.1433	0.1289	0.0016	
13	2.2958	-3.5	-3.5	246	0.8200	0.8692	0.0028	
14	-2.0333	-2.5	-2.5	1156	3.8533	3.8329	0.0001	
15	0.0772	-1.5	-1.5	3325	11.0833	11.0652	0.0000	
16	1.5303	-0.5	-0.5	6331	21.1033	20.9255	0.0015	
17	-1.8392	0.5	0.5	7765	25.8833	25.9335	0.0001	
18	1.0627	1.5	1.5	6253	20.8433	21.0661	0.0024	
19	1.3075	2.5	2.5	3352	11.1733	11.2145	0.0002	
20	-1.1846	3.5	3.5	1198	3.9933	3.9108	0.0017	
21	0.4981	4.5	4.5	296	0.9867	0.8928	0.0099	
22	0.3084	y mayor...		34	0.1133	0.1472	0.0078	
23	-0.2541					<b>Suma=</b>	0.0355	
24	-0.4776					<b>X2.90=</b>	14.6837 <b>normal</b>	
25	-0.4435					<b>X2.95=</b>	16.9190 <b>normal</b>	
26	0.7834					<b>X2.99=</b>	21.6660 <b>normal</b>	
27	-3.3466							
28	0.3874							
29	-0.1317							
30	-2.4703							
31	0.4302							
32	0.4500							
33	1.6559							
34	1.1141							
35	0.3364							



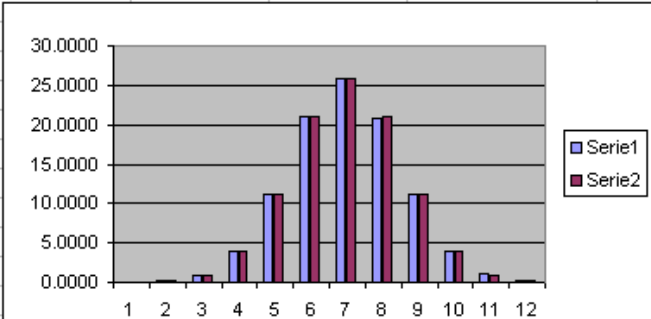


	A	B	C	D	E	F	G	H	I
1	Sèrie normal de mitjana 0 i desviació estàndard 1.5								
2	AleatNormal_AmbParametres 0 1.5 30000 4 > resu.txt								
3	-1.2553								
4	-0.2584	<b>N:</b>	<b>30000</b>						
5	0.2807	<b>Mínim:</b>	<b>-5.5962</b>						
6	2.4232	<b>Màxim:</b>	<b>6.9732</b>						
7	-0.2652	<b>Mitjana:</b>	<b>0.0079</b>						
8	0.9797	<b>DesvEstàndard:</b>	<b>1.5107</b>						
9	-0.8195				<b>Freq. observada</b>	<b>Freq. esperada</b>	<b>Freq. esperada</b>	<b>(Obs-Esp)^2/Esp</b>	
10	0.2912	<b>Classes proposa</b>	<b>Clase</b>	<b>Frecuencia</b>	<b>(%)</b>	<b>(%)</b>	<b>(absoluta)</b>		
11	1.3886	-5.5	-5.5	1	0.0033	0.0133	4	2.2465	
12	1.8065	-4.5	-4.5	43	0.1433	0.1289	39	0.4837	
13	2.2958	-3.5	-3.5	246	0.8200	0.8692	261	0.8340	
14	-2.0333	-2.5	-2.5	1156	3.8533	3.8329	1150	0.0325	
15	0.0772	-1.5	-1.5	3325	11.0833	11.0652	3320	0.0089	
16	1.5303	-0.5	-0.5	6331	21.1033	20.9255	6278	0.4536	
17	-1.8392	0.5	0.5	7765	25.8833	25.9335	7780	0.0291	
18	1.0627	1.5	1.5	6253	20.8433	21.0661	6320	0.7068	
19	1.3075	2.5	2.5	3352	11.1733	11.2145	3364	0.0454	
20	-1.1846	3.5	3.5	1198	3.9933	3.9108	1173	0.5222	
21	0.4981	4.5	4.5	296	0.9867	0.8928	268	2.9606	
22	0.3084	y mayor...		34	0.1133	0.1472	44	2.3368	
23	-0.2541						<b>Suma=</b>	10.6602	
24	-0.4776						<b>X2.90=</b>	14.6837	<b>normal</b>
25	-0.4435						<b>X2.95=</b>	16.9190	<b>normal</b>
26	0.7834						<b>X2.99=</b>	21.6660	<b>normal</b>
27	-3.3466								
28	0.3874								
29	-0.1317								
30	-2.4703								
31	0.4302								
32	0.4500								
33	1.6559								





	A	B	C	D	E	F	G	H	I
1	Sèrie normal de mitjana 0 i desviació estàndard 1.5								
2	AleatNormal_AmbParametres 0 1.5 30000 4 > resu.txt								
3	-1.2553								
4	-0.2584	N:	30000						
5	0.2807	Mínim:	-5.5962						
6	2.4232	Màxim:	6.9732						
7	-0.2652	Mitjana:	0.0079						
8	0.9797	DesvEstàndard:	1.5107						
9	-0.8195				Freq. observada	Freq. esperada	Freq. esperada	(Obs-Esp)^2/Esp	
10	0.2912	Classes proposa	Clase	Frecuencia	(%)	(%)	(absoluta)		
11	1.3886	-5.5	-5.5	1	0.0033	0.0133	4	2.2465	
12	1.8065	-4.5	-4.5	43	0.1433	0.1289	39	0.1837	
13	2.2958	-3.5	-3.5	246	0.8200	0.8692	261	0.8340	
14	-2.0333	-2.5	-2.5	1156	3.8533	3.8329	1150	0.0325	
15	0.0772	-1.5	-1.5	3325	11.0833	11.0652	3320	0.0089	
16	1.5303	-0.5	-0.5	6331	21.1033	20.9255	6278	0.4536	
17	-1.8392	0.5	0.5	7765	25.8833	25.9335	7780	0.0291	
18	1.0627	1.5	1.5	6253	20.8433	21.0661	6320	0.1701	
19	1.3075	2.5	2.5	3352	11.1733	11.2145	3364	0.0454	
20	-1.1846	3.5	3.5	1198	3.9933	3.9108	1173	0.5222	
21	0.4981	4.5	4.5	296	0.9867	0.8928	268	2.9606	
22	0.3084	y mayor...		34	0.1133	0.1472	44	2.3368	
23	-0.2541						Suma=	10.6602	
24	-0.4776						X2.90=	21.6137	normal
25	-0.4435						X2.95=	16.9190	normal
26	0.7834						X2.99=	21.8860	normal
27	-3.3466								
28	0.3874								
29	-0.1317								
30	-2.4703								
31	0.4302								
32	0.4500								
33	1.6559								



=100\*D11/\$C\$4

=100\*DISTR.NORM(C11;\$C\$7;\$C\$8;1)

=100\*(DISTR.NORM(C12;\$C\$7;\$C\$8;1)

-DISTR.NORM(C11;\$C\$7;\$C\$8;1))

=100\*(DISTR.NORM(C20;\$C\$7;\$C\$8;1)

-DISTR.NORM(C21;\$C\$7;\$C\$8;1))

o bé

=100\*(DISTR.NORM(C21;\$C\$7;\$C\$8;1)

-DISTR.NORM(C21;\$C\$7;\$C\$8;1))

=PRUEBA.CHI.INV(0.01;CONTAR(\$G\$11:\$G\$22)-1-2)

En aquest cas surt que sempre s'accepta normalitat. Tanmateix (EXERCICI) proveu què passa amb mostres petites i observareu que si exigim un 5% o un 1% de probabilitat de rebutjar  $H_0$  i que fos certa, no ens atrevim a dir que no és normal. En canvi, podem dir que no és normal acceptant un 10% de probabilitats que ho sigui. A aquest artefacte hi contribueixen les dues freqüències "tallades" a 1/2 als dos extrems (que simulen les cues de la normal) i que fins i tot es comparen amb la cua acumulada dreta, i el fet que no hi hagi cap dada en les probabilitats més extremes (a les dues cues).



**Exercici 1:** A partir del Mapa d'Usos del Sòl de Catalunya 1997-98, disponible al *web* del Departament de Medi Ambient de la Generalitat i del Model Digital d'Elevacions de Catalunya "cata200.img", extret del *web* de l'Institut Cartogràfic de Catalunya, esbrineu si es pot afirmar que la categoria "Bosquines i prats" segueix una distribució d'altituds normal a un nivell de significació 0.10, 0.05 i 0.01.

**Notes:**

1. Podeu obtenir Cata200.img a partir del *web* de l'ICC o a \\joanma\MaterialsMaster\MEII\GerardMore\cata200.IMG
2. Mostregeu les altituds a les quals se situa la categoria seguint una malla de 1000 m x 1000 m.



## Correlació en imatges de teledetecció

- La proximitat espectral de les bandes de Teledetecció, i la seva inespecificitat en el cas de bandes que abasten un rang ampli de longituds d'ona fa que sovint la correlació entre bandes sigui elevada.
- L'histograma bivariant és una excel·lent forma d'apreciar l'eventual existència d'aquest fenomen.
  - Com?** Invoqueu el mòdul REGRESS d'Idrisi. A partir de la imatge Landsat-7 ETM+ de 13-06-2002, compareu la informació de les bandes 3 i 4, 3 i 5, 4 i 5, 5 i 6 i 4 i 6. Comenteu els resultats.  
Podeu trobar el fitxer a \\joanma\materialsMASTER\MEII\XavierPons
- L'histograma bivariant també permet mostrar la presència de grans grups naturals (classes) a la imatge. Identifiqueu quins veieu en els histogrames anteriors.



**Exercici 2:** Tradicionalment es pensa que l'ús d'imatges de diferents dates sobre una mateixa zona geogràfica aporta informació útil per a la classificació. Tanmateix, ens podem plantejar fins a quin punt la fenologia dels conreus i de la vegetació, la presència de neu i núvols pot aportar més informació que les diferències espectrals. A partir de les imatges de 26-04-2002 i 13-06-2002, compareu la informació de les bandes 3, 4, 5 i 6 entre dates. Comenteu els resultats.

**Nota:**

Material a \\joanma\materialsMASTER\MEII\XavierPons





## Anàlisi de Components Principals (*Pearson 1901*)

- En moltes anàlisis pot ser útil reduir el nombre de variables independents i treballar amb un subconjunt menor però amb similar poder informatiu. Això permet, per exemple, dur a terme processos de classificació d'imatges de Teledetecció en base a un nombre menor de bandes i, per tant, en menys temps de càlcul.
- A més, ja hem vist que, particularment en les imatges de Teledetecció hi ha una alta redundància entre bandes. Tanmateix, aquest fet és comú a molts estudis en què s'utilitzen diverses dades en forma multivariant: són ben conegudes les correlacions entre altitud topogràfica i temperatura de l'aire (positiva), o entre superfície urbanitzada i pendent topogràfic (negativa).



**L'anàlisi de components principals (ACP o PCA)** permet obtenir un nou conjunt de variables (noves bandes o canals en el cas de la Teledetecció) en nombre igual al del conjunt original, però amb les següents característiques:

1 Cada nova variable (component principal, CP) s'obté per combinació lineal de totes les  $p$  variables originals:

$$\begin{aligned} \text{CP}_1 &= a_{1,1}X_1 + a_{1,2}X_2 + \dots + a_{1,p}X_p \\ \text{CP}_2 &= a_{2,1}X_1 + a_{2,2}X_2 + \dots + a_{2,p}X_p \\ &\vdots \\ \text{CP}_p &= a_{p,1}X_1 + a_{p,2}X_2 + \dots + a_{p,p}X_p \end{aligned}$$

2 Cada nova variable està descorrelacionada amb l'anterior.

3 Cada nova variable explica una quantitat menor (o igual) de la variació original total.



- Si les variables en joc estan molt poc o gens correlacionades, l'ACP no serveix de res.
- Habitualment no s'obtenen totes les CP (aquesta és la gràcia).
- La suma de les variàncies de totes les CP és igual a la suma de les variàncies de totes les variables originals ( $X$ ).
- Per tal d'evitar que alguna de les variables tingui un pes inadequat en l'anàlisi és habitual normalitzar (estandarditzar) prèviament les variables (mitjana 0 i desviació estàndard 1). Això equival a efectuar l'anàlisi en base a la matriu de correlacions en comptes d'en base a la matriu de variàncies/covariàncies.



## Normalització (estandardització)

(Font: [http://en.wikipedia.org/wiki/Standard\\_score](http://en.wikipedia.org/wiki/Standard_score)). El llibre d'Espiegel (1991, p. 160), tradueix **standard score** per unitat estàndard, però jo prefereixo valor estàndard perquè permet usar **score** com a valor en tots els casos

- In statistics, a **standard score** (also called **z-score** or **normal score**) is a dimensionless quantity derived by subtracting the population mean from an individual (raw) score and then dividing the difference by the population standard deviation.  
$$z = \frac{x - \mu}{\sigma}$$
- The z score reveals how many units of the standard deviation a case is above or below the mean. The z score allows us to compare the results of different normal distributions, something done frequently in research.
- The conversion process is called **standardizing**. The quantity z represents the distance between the raw score and the population mean in units of the standard deviation. z is negative when the raw score is below the mean, positive when above.



- En el procés es calculen els vectors  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p$ , que reben el nom de vectors propis (*eigenvectors*) i contenen els coeficients de cada CP.
- També es calculen uns valors  $\lambda_1, \lambda_2, \dots, \lambda_p$ , que reben el nom de valors propis (*eigenvalues*) i són la variància de cada CP.
- L'ACP també és interessant perquè dóna idea de quina és la variabilitat intrínseca de les nostres dades (quantes fonts de variació hi ha).
- Les noves variables perden les seves unitats i sentit físic. Per exemple, en el cas de bandes de Teledetecció, les noves bandes ja no tenen el sentit radiomètric que tenien.



- Efectueu una ACP de la imatge Landsat-7 ETM+ de 13-06-2002 i obtingueu un nombre de CP igual al nombre de bandes originals.

*Com?* Invoqueu el mòdul PCA d'Idrisi.

- A partir de quina nova CP sembla, visualment, que no hi ha aportació d'informació rellevant? Com lliga això amb els valors propis?

- Quin és el paper dels elements singulars però poc representats a la imatge? Queden preservats?

- Comproveu la descorrelació que s'ha aconseguit en les noves variables.

*Com?* Useu el mòdul REGRESS d'Idrisi.



**Exercici 3:** Comproveu, en la mateixa imatge, si el fet de multiplicar per un factor 3 una banda mentre dividim entre 3 les altres canvia substancialment l'ACP en mode estandarditzat i en mode no estandarditzat.

**Exercici 4:** Obtingueu de nou les bandes originals a partir de les CP però obviant les darreres components (menys informatives). Aprecieu alguna millora?