# Comparació de distribucions

*Antoni Gomà Nasarre*
Programa d'Informàtica Educativa, 1991.



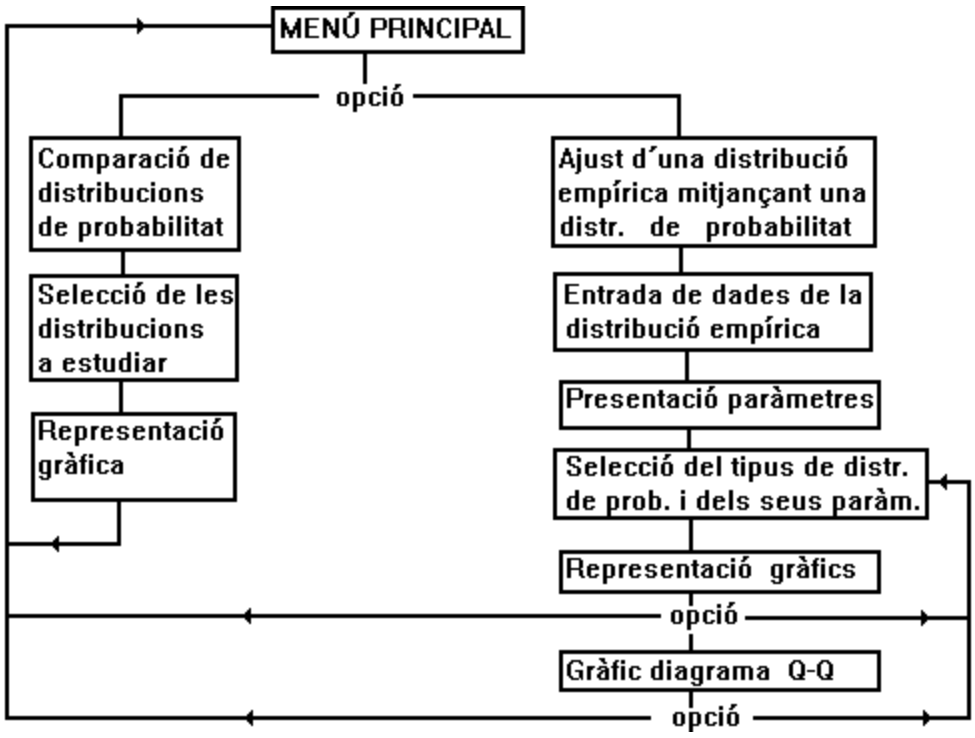
## 1. DESCRIPCIÓ TÈCNICA

L'arxiu que conté el programa es denomina DISTRIB:EXE i està escrit en llenguatge GW-BASIC i compilat posteriorment. Requereix la presència dels arxius BRUN20G.EXE i PROTADA:BIN en el mateix disquet o subdirectori des d'on es cridi el programa, que caldrà executar amb la comanda DISTRIB.

## 2. INSTRUCCIONS DE FUNCIONAMENT

### 2.1. Estructura del programa

Aquest diagrama explica com està estructurat el programa, amb dues opcions fonamentals de treball, cadascuna d'elles relacionada amb un dels objectius del programa.




## 2.2. Opcions del programa i dinàmica de funcionament

D'acord amb el menú principal, al qual s'arriba després d'unes planes de presentació, ofereix tres opcions:

1. **Comparar els gràfics de diverses distribucions.**
2. **Comparar la distribució empírica amb un model teòric.**
0. **Acabar.**

Per escollir la part del programa que es vol treballar, cal prémer la tecla que indica l'opció.

### 2.2.1 Com se seleccionen les distribucions a estudiar

Per a la primera de les opcions, el programa ens ofereix la possibilitat d'escollir d'entre cinc distribucions, tres de tipus discret -binomial, Poisson i geomètrica- i dues de probabilitat contínua -la normal i l'exponencial-.

La selecció de tipus de distribucions que es vol estudiar no ofereix cap dificultat. Quan s'escau, en pantalla es pot veure quin número correspon a cada tipus de distribució i la pregunta

```
Quin tipus de distribució voleu? (1, 2, 3, 4, 5 o 0 per
acabar)
```

a la qual cal respondre prement la tecla numèrica adequada.
Seguidament, el programa demanarà els paràmetres que defineixen la distribució concreta que volem estudiar. Caldrà entrar els corresponents valors numèrics. Els exemples que es donen al final d'aquesta documentació poden ajudar a entendre quins són, en cada cas, els paràmetres que cal entrar.

Convé comentar que aquesta entrada de paràmetres es pot fer amb les operacions indicades. Això pot ser útil per exemple quan vulgueu estudiar l'aproximacióde la distribució binomial B(p=0.4,n=100) mitjançant la distribució normal de mitjana μ=npp i desviació tipus σ=[np(1-p)12. Per fer aquest exemple podeu actuar així des de la pantalla de selecció de les distribucions:

- Polseu després la tecla *4* per fer la tria de la distribució normal. Entreu 0.4*100 com a valor de la mitjana i, també amb l'operació indicada, sqr(0.4*0.6*100) com a valor de la desviació tipus.

### 2.2.2. Comparació de diverses distribucions

Els gràfics que es presenten en aquesta opció són:

* la funció de probabilitat o funció de masses si la distribució és de tipus discret.

* la funció de densitat de probabilitat si la distribució és de tipus continu.

Es poden obtenir els gràfics d'un màxim de tres distribucions en la mateixa pantalla. Si se'n volen comparar només dues, o visualitzar-ne una sola, caldrà indicar-ho polsant 0 enlloc de la tecla que indicaria el tipus de distribució a representar.

### 2.2.3. Entrada de dades de la distribució empírica

La primera matisació que cal fer per al treball amb aquesta opció és la d'indicar si les dades empíriques corresponen a una variable estadística discreta o bé contínua. Aquesta matisació es fa prement una tecla (*D* o *C*) i té importància per als tipus de gràfics que es presenten i també per a la forma de càlcul de les quantil.les. Seguidament, caldrà entrar les dades empíriques amb les següents indicacions:

* Si es tracta de dades que corresponen a una variable estadística discreta, caldrà indicar prèviament els extrems del rang de valors observats i la separació entre els possibles valors, que haurà de ser constant.

* Si volem treballar amb dades corresponents a una variable contínua, caldrà que les tinguem tabulades en intervals de classe, tots ells de la mateixa longitud. El programa ens demanarà la marca de classe del primer i de l'últim interval de classe i la longitud comuna de tots els intervals de classe.

* Llavors, com a idea fonamental, cal saber que les dades s'entraran en forma de

dada-freqüència absoluta. Per cada valor observat o cada marca de classe caldrà entrar-ne la freqüència absoluta qun el programa ho demani.

En acabar aquesta entrada de dades el programa ens facilitarà alguns paràmetres que ens poden ser d'utilitat a l'hora de triar la distribució de probabilitat amb què volem ajustar les nostres dades empíriques.

### 2.2.4. Ajust de la distribució empírica mitjançant el model teòric

Després de la selecció de la distribució de probabilitat que serà el nostre model teòric de treball, el programa mostrarà el gràfic més senzill per a la comparació del conjunt de dades empíriques amb el model teòric.

A la vista del diagrama comparat ja es pot tenir una idea intuïtiva de la bondat de l'ajust i, llavors, es pot optar, polsant S o N a les diverses preguntes es formulin:

* per representar el diagrama de les quantil.les (que és una altra manera gràfica d'estudiar l'ajust).

* per seleccionar una altra distribució de probabilitat com a model teòric amb què assajar l'ajust d'aquella mateixa distribució de dades empíriques.

* per tornar al menú principal.

### 2.2.5. Tecles preprogramades

*F10* A part de la possibilitat d'acabar l'execució del programa des del menú principal, la tecla *F10* està pre-programada amb aquesta mateixa probabilitat.
*F9* Aquesta tecla porta al menú principal en qualsevol moment de l'execució del programa.

## 3. ASPECTES PEDAGÒGICS

### 3.1. Objectius del programa

* Facilitar l'estudi i la comparació de les distribucions de probabilitat des d'un punt de vista gràfic i poder-ho fer d'una forma àgil.
* Disposar d'eines gràfiques per analitzar l'ajust d'una distribució estadística de dades empíriques mitjançant una distribució de probabilitat.

### 3.2. Coneixements previs

És del tot necessari tenir assolit el concepte de distribució de probabilitat i saber distingir el tractament del cas discret d'allò que cal fer en el cas continu.

Convé conèixer, per altra banda, tant l'extraordinària importància de les distribucions de probabilitat com el model teòric que dóna les bases per a l'estudi de les nocions

estadístiques.

En aquest sentit convé tenir perfectament compresa la significació del diagrama de barres o l'histograma asociats a un recull empíric de dades i el concepte de freqüència esperada d'un valor o d'una classe pel que fa a una distribució de probabilitat.

## 3.3. Fonamentació teòrica

### 3.3.1. Distribucions de probabilitat

Convé que una eina gràfica per estudiar les distribucions de probabilitat distingeixi ben clarament els gràfics que s'empren en el cas de distribucions de tipus continu.

Aquesta distinció, que seguidamnent recordarem, és crucial des del punt de vista teòric i, doncs, cal que també es faci visual.

Pel que fa a les distribucions de probabilitat associades a una variable aleatòria discreta, X, es treballa usualment amb:

* la funció de probabilitat o funció de masses que és, essencialment, discreta, perquè només els valors x i d'un conjunt discret verifiquen p[X=xi]<>0. (Aquest és el gràfic que presenta el programa).

* la funció de distribució (probabilitat acumulada) que resulta ser una funció gradual, creixent, discontínua pels valors xi que tenen probabilitat diferent de 0.

Pel que fa a les distribucions de probabilitat contínues es considera:

* la funció de densitat de probabilitat que és contínua. ( I el gràfic de la qual és el que mostra el programa).

* la funció de distribució de probabilitat (probabilitat acumulada) que és creixent i contínua.

El paquet estadístic STATGRAPHICS no fa aquesta distinció; podem dir que fa un tractament incorrecte del cas discret. Això ens va impulsar a dissenyar aquesta eina informàtica de suport per a l'ensenyament de l'Estadística.

Al mateix temps vàrem procurar agilitzar una tasca que, si bé és possible amb STATGRAPHICS, és una mica farragosa: poder veure els gràfics de distribucions de probabilitat de tipus diferents en el mateix gràfic, cosa que té clares aplicacions des d'un punt de vista teòric: aproximació de la binomial mitjançant la normal; la distribució de Poisson com a cas límit de la binomial, etc.

### 3.3.2. Ajust d'una distribució

Quan s'han recollit dades i, per tant, s'ha obtingut una distribució estadística

experimental sobre la que s'especula si s'ajusta o no a una determinada distribució de probabilitat, és del tot necessari- abans de passar a quantificar la bondat de l'ajust- adquirir-ne una idea gràfica.

Una primera aproximació es pot obtenir comparant:

* en el cas discret, el diagrama de barres que dóna les freqüències relatives de la distribució experimetnal amb el gràfic de la funció de probabilitat o funció de masses de la distribució teòrica.

* en el cas de distribucions contínues, l'histograma de la distribució estadística, que haurem hagut de tabular adequadament, confrontat amb un diagrama que mostra la freqüència relativa esperada en cada classe per al model teòric escollit (que hem representat com un diagrama de barres, col.locades sobre les respectives marques de classe).

No hi ha "teories" que ens diguin quin ha de ser el resultat de la comparació dels dos gràfics per tal que poguem dir que l'ajust "sigui bo". Cal que sigui la nostra intuïció la que ens ho digui!

Tanmateix, si constatem que els dos gràfics comparats que se'ns presenten són ben diferents, podem assegurar que l'ajust no és bo: caldrà cercar un altre model teòric adequat al nostre conjunt de dades.

Ara bé, en cas de pensar que l'ajust pot ser correcte i, abans de quantificar la bondat de l'ajust amb algun test numèric, convé conèixer un altre mètode propi de l'anàlisis exploratòria de dades: els diagrames Q-Q (quantil.la-quantil.la).

Aquests diagrames consisteixen en representar en un sitema cartesià d'eixos el valor de cadascuna de le quantil.les de la distribució experimental confrontats amb el valor dels corresponents quantil.les de la distribució teòrica.

Així, per exemple, si una distribució experimental consta de 50 dades, aquestes , una vegada ordenades de menor a major -y1, y2, ..., y50 correspondran a les quantil.les (1-0.5)/50, (2-0.5)/50, ..., (49-0.5)/50 i (50-0.5)/50, és a dir, la quantil.la corresponent a pi=(i-0.5)/50 és la dada yi.

Si això mateix es fa amb la distribució teòrica, a la qual s'intenta ajustar la distribució experimental, obtindrem les quantil.les x1, x2, ..., x50.

Seguidament, representarem els punts (xi, yi) per i=1, ...,50 i ja tindrem el diagrama Q_Q.

S'hauria de fer una cosa semblant si el nombre de dades fos un altre. Això sí, si el conjunt de dades és molt nombrós ben segur que serà suficient calcular les 100 quantil.les corresponents a cada percentatge enter i mostrar-ne el gràfic per poder intuir sobre la bondat de l'ajust. Així ho fa el programa: el que no fa sinó que ho haurà de fer la intuïció de l'usuari, és especular sobre el núvol de punts resultant per deduir

quina és la bondat de l'ajust. Ara bé, la idea important següent: si l'ajust fos "perfecte" els punts (xi, yi) haurien de quedar alineats sobre la bisectriu perquè els valors de les quantil.les coincidirien per la distribució empírica i la distribució teòrica. Seria xi=yi.

Al final d'aquesta documentació es poden veure els diagrames Q-Q corresponents a tres exemples: un bon ajust i dos clars desajustaments.

**Per ampliar la informació sobre els diagrames Q-Q podeu consultar:**
**- J.M.Chambers, Graphical Methods for Data Analysis, pp. 191-242.**

### 3.4. Metodologia d'ús

Creiem que el programa és d'ús individual però amb una"tasca a fer" prèviament assenyalada i dirigida pel professor. Tanmateix si es pot disposar de sortida a la pantalla gran, aquest programa pot fer molt bé el paper de "pissarra electrònica" per acompanyar les explicacions del professor.
Assenyalem seguidament alguns exemples de "treballs a fer" que creiem interessants.

#### 3.4.1. Amb l'opció 1 del programa

a) Podem aconseguir que l'alumne visualitzi la influència que tenen els paràmetres de les principals distribucions discretes en les gràfiques de les respectives funcions de probablilitat o funcions de massa.
Amb aquesta finalitat podem comparar, per exemple:

* tres distribucions binomials amb el mateix nombre n de repeticions i variant la probabilitat p d'èxit.

* repetir el procés anterior, però ara, mantenint constant igual a 0.5 el valor p de la probabilitat d'èxit i variant el nombre n de repeticions.

* tres distribucions de Poisson amb diferent paràmetre p.

b) Podem aconseguir que l'alumne s'adoni de l'equivalència entre les distribucions binomial i de Poisson quan la probabilitat d'èxit és molt petita i el nombre de repeticions és suficientment gran.

c) Amb un treball semblant al que hem assenyalat a l'apartat a) podem fer que l'alumne copsi la transcendència dels paràmetres de les principals distribucions contínues. Sobre tot:

* Comparar tres distribucions normals amb la mateixa $\mu$ i diferent s.

* Comparar tres distribucions normals amb la mateixa s i diferent $\mu$.

d) Aproximació de la binomial mitjançant la normal. Anteriorment en aquesta mateixa documentació ja n'hem comentat un bon exemple. Repetiu-ho, amb p=0.5 i n=10 i, després, amb p=0.2 i n=10. Veureu com, a mesura que es va baixant el valor de np l'aproximació és cada vegada més dolenta i, en aquest darrer cas, ja no es pot pas dir que la normal doni un bon model per la B(0.2, 0). De fet, veureu que aquesta és ben poc simètrica.

e) Com a suggeriment de treball final en l'exercitació d'aquesta opció del programa DISTRIB donem el de comparar una distribució exponencial amb la distribució geomètrica que té la mateixa mitjana. Traieu-ne conclusions.

### 3.4.2. Amb l'opció 2 del programa

Suggerim a continuació alguns possibles exemples. Val a dir que seria bo refer aquells exemples per als quals es donen recollides de dades fetes anteriorment per grups d'alumnes, és a dir, perquè la idea didàctica d'aprenentatge de l'Estadística fos més acurada seria convenient que el grup-classe reculli, ell mateix, les dades.

**Exemples amb distribucions discretes:**

- Distribució experimental:  dóna les freqüències del nombre de cares obtingudes en sèries de 10 llançaments d'una moneda no trucada.
Distribució de probabilitat: Binomial B(n=10;p=0.5).

- Distribució experimental: del nombre d'1s obinguts en sèries de 10 llançaments d'un dau esbiaixat.
Distribució de probabilitat: Binomial B(n=10;p) on el valor de p cal estimar prèviament el valor de p.

- Distribució experimental: es van comptar els nombres de vehicles que, a intervals de 30 s, vàren passar per l'autopista A-7 en direcció a Barcelona a l'alçada del municipi de Salt.
Es va obtenir la taula següent:

| Nombre vehicles cada 30 s | Freqüència |
|---|---|
| 0 | 2 |
| 1 | 9 |
| 2 | 12 |
| 3 | 11 |
| 4 | 11 |
| 5 | 8 |
| 6 | 3 |
| 7 | 3 |
| 8 | 1 |

Podem estudiar si s'ajusta significativament a un model de llei de Poisson amb la mitjana $\mu$ estimada a partir de les dades i que ens serà facilitada pel programa

- Distribució experimental: la que resulta d'haver realitzar 314 vegades l'experiment de llançar repetidament un dau fins que surti un 6, prenent nota del nombre de tirades realitzades en cadascuna de les proves. Els resultats obinguts es resumeixen en la següent taula:

| Nombre de llançaments | Freqüència |
|---|---|
| 1 | 53 |
| 2 | 51 |
| 3 | 25 |
| 4 | 36 |
| 5 | 19 |
| 6 | 31 |
| 7 | 24 |
| 8 | 14 |
| 9 | 7 |
| 10 | 14 |
| 11 | 10 |
| 12 | 9 |
| 13 | 3 |
| 14 | 2 |
| 15 | 4 |
| 16 | 2 |

| Nombre de llançaments | Freqüència |
|---|---|
| 17 | 1 |
| 18 | 0 |
| 19 | 1 |
| 20 | 1 |
| 21 | 0 |
| 22 | 1 |
| 23 | 1 |
| 24 | 0 |
| 25 | 1 |
| 26 | 1 |
| 27 | 0 |
| 28 | 0 |
| 29 | 1 |
| 30 | 0 |
| 31 | 1 |
| 32 | 1 |

Estudieu el grau d'ajustament d'aquesta taula de freqüències empíriques amb una distribució geomètrica de paràmetre p=1/6.

**Exemples amb distribucions contínues:**

- Emprarem unes dades extretes de l'Anuari Estadístic Nacional de 1987 - lleugerament modificades- que corresponen als pesos dels nois que vàren ser cridats a fer el servei l'any 1985.

| Pes | Percentatges |
|---|---|
| Menys de 40 | 0.0 |
| 40 - 44 kg | 0.1 |
| 45 - 49 kg | 1.0 |
| 50 - 54 kg | 5.4 |
| 55 - 59 kg | 13.0 |
| 60 - 64 kg | 23.2 |
| 65 - 69 kg | 21.9 |
| 70 - 74 kg | 16.5 |
| 75 - 79 kg | 10.3 |
| 80 - 84 kg | 5.0 |
| 85 - 89 kg | 2.2 |
| 90 - 94 kg | 1.0 |
| 95 - 99 kg | 0.4 |
| Més de 100 | 0.0 |

Es tracta de veure si aquestes dades poden ser enteses com una mostra presa d'una població en la qual el pes es distribueixi seguint un model teòric normal. Els paràmetres $\mu$ i $\sigma$ es podràn dedïr de les dades que facilita el mateix programa.
Quan el programa pregunti "Marca de classe inferior", entreu-li el valor 42. A la pregunta de "Longitut de cada classe" entreu-li el valor 5. Quan es demanin les freqüències observades de cada classe ( que serà referenciada per la marca de classe), entreu-hi els percentatges multiplicats per 10, per tal que esdevinguin nombres enters.

- Podeu estudiar si amb algun altre valor de la mitjana $\mu$ i per a la desviació tipus s per a la distribució normal teòrica aconseguiu que millori l'ajust.

- Distribució empírica: els temps de vida de 500 bateries elèctriques han estat estudiats i classificats així:

| Temps (hores) | Freqüència |
|---|---|
| 0 - 50 | 208 |
| 50 - 100 | 112 |
| 100 - 150 | 75 |
| 150 - 200 | 40 |
| 200 - 250 | 30 |
| 250 - 300 | 18 |
| 300 - 350 | 11 |
| 350 - 400 | 6 |

Haureu de prendre la marca de classe inferior: 25; la marca de classe superior: 375; la longitud de les classes: 50.
Es tracta de veure si s'ajusten a un adistribució exponencial. Com que la mitjana és 95, el valor del paràmetre haurà de ser $k= 1/95 = 0.0105$.

- Com a darrer exemple suggerim la recollida d'alçades dels alumnes i les alumnes del grup-classe en què estiguem treballant; la posterior tabulació en intervals de classe i, finalment l'ús del programa per estudiar el seu possible ajust mitjançant una distribució normal. És més que possible que no aconseguim de cap manera que l'ajust sigui bo pel fet que es presenti una distribució pràcticament bimodal, lligada a les subpoblacions nois-noies.