

PROBABILITAT I ESTADÍSTICA PER A CIÈNCIES II

Toni Monleón-Getino
Clara Rodríguez Casado

Departament de Genètica, Microbiologia i Estadística
Secció d'Estadística

PROBABILITAT I ESTADÍSTICA PER A CIÈNCIES II

Toni Monleón-Getino
Clara Rodríguez Casado

Departament de Genètica, Microbiologia i Estadística
Secció d'Estadística



TEXTOS DOCENTS



UNIVERSITAT DE
BARCELONA

Edicions

Índex

PRESENTACIÓ	9
1. TEORIA DEL MOSTRATGE I BASES DE L'ESTADÍSTICA	11
1.1. Aproximació de distribucions de probabilitat	12
1.1.1. Suma de variables aleatòries amb distribució binomial	12
1.1.2. Reproductibilitat de famílies de distribucions	13
1.1.3. Aproximació de funcions de distribució	14
1.1.4. Aproximació de la distribució binomial a la distribució de Poisson	15
1.1.5. Aproximació de la distribució binomial a la distribució normal	16
1.1.6. Aproximació de la distribució de Poisson a la distribució normal	18
1.1.7. Resum de relacions	20
1.1.8. La llei dels grans nombres	21
1.1.9. Teorema central del límit	23
1.1.9.1. Distribució uniforme discreta	25
1.1.9.2. Els errors de mesura	25
1.2. Mostres i poblacions	26
1.2.1. Inferència estadística: de la mostra a la població	26
1.2.1.1. Cas 1: exemple en poblacions finites	27
1.2.1.2. Cas 2: exemple en poblacions infinites	28
1.2.2. Representativitat de la mostra	29
1.2.3. Mostratge	29
1.2.4. Tècniques de mostratge	29
1.2.5. Mostratge aleatori simple	30
1.2.6. Mostratge sistemàtic	31
1.2.7. Mostratge per estrats	31
1.2.8. Mostratge per conglomerats	31
1.2.9. Distribució empírica de la mostra i teorema de Glivenko-Cantelli	32
1.2.9.1. La representativitat de la mostra	32
1.3. Estadístics mostrals i estimació puntual	35
1.3.1. Concepte d'estadístic	35
1.3.2. Principals estadístics mostrals	35
1.3.3. Un estadístic és una variable aleatòria	35
1.3.4. El teorema central del límit explica la distribució de la mitjana mostral	37
1.3.5. Estimadors i estimació	38
1.3.5.1. Estadístics mostrals com a estimadors	39
1.3.5.2. Propietats dels estimadors puntuals: consistència, no esbiaixat (centrat) i eficiència	39
1.4. Interval de confiança	44
1.4.1. Definició d'interval de confiança	44
1.4.2. Com s'interpreta la «confiança»	45

1.4.3.	Interval de confiança per l'estimació de μ amb σ coneguda	45
1.4.4.	Amplada de l'interval de confiança: confiança <i>versus</i> precisió	47
1.4.5.	Interval de confiança per l'estimació d'una proporció	48
1.4.6.	Distribucions de probabilitat en el mostratge	50
1.4.6.1.	Distribució χ^2 (khi quadrat)	50
1.4.6.2.	Distribució t de Student	52
1.4.6.3.	Importància de les distribucions khi quadrat i t de Student	53
1.4.6.4.	Interval de confiança per a μ amb variància desconeguda	54
1.4.6.5.	Construcció de l'interval de confiança per a la variància	55
	Problemes proposats	56
2.	INFERÈNCIA ESTADÍSTICA I REGRESSIÓ	61
2.1.	Introducció al contrast d'hipòtesis	61
2.1.1.	Les hipòtesis en estadística	61
2.1.2.	Llindar de decisió de les hipòtesis: probabilitats i taula d'errors	65
2.1.2.1.	La regió crítica (W) i el nivell de significació (α)	66
2.1.2.2.	Error de tipus II i la potència del test ($1 - \beta$)	68
2.1.3.	Tipus de test d'hipòtesis	69
2.1.3.1.	Tests unilaterals i tests bilaterals	70
2.1.4.	Passos que cal seguir en un contrast	72
2.1.5.	El valor de probabilitat o valor p	72
2.2.	Contrast d'hipòtesis per a la mitjana μ i la variància σ^2 d'una població.	
	Cas normal	75
2.2.1.	Contrast per a la mitjana ($\mu = E(X)$) amb σ coneguda	75
2.2.2.	Contrast per a la mitjana ($\mu = E(X)$) amb σ desconeguda	79
2.2.3.	Contrast per a la variància ($\sigma = \text{Var}(X)$)	81
2.3.	Contrast d'hipòtesis per a les proporcions p i λ d'una població.	
	Cas no normal	84
2.3.1.	Contrast d'hipòtesis per a una proporció p	84
2.3.2.	Contrast per a la proporció λ d'una distribució de Poisson	87
2.4.	Comparació de mitjanes de dues poblacions normals independents	90
2.4.1.	Contrast de comparació de mitjanes: hipòtesis	91
2.4.1.1.	Comparació de mitjanes: cas normal amb σ_1 i σ_2 conegudes	91
2.4.1.2.	Comparació de mitjanes: cas normal amb σ_1 i σ_2 desconegudes	92
2.5.	Test d'homogeneïtat de variàncies de dues poblacions normals independents ...	95
2.5.1.	Comparació de variàncies	95
2.6.	Contrast d'hipòtesis per a la proporció p de dues poblacions. Cas no normal....	98
2.6.1.	Cas particular: contrast de proporcions de dues poblacions	98
2.7.	Comparació de mitjanes de dades aparellades	101
2.8.	El model multinomial	105
2.8.1.	Distribució de probabilitats multinomial	105
2.8.2.	Càlcul de probabilitats en el model multinomial	106
2.8.3.	Propietats i aplicacions del model multinomial en estadística (tests de χ^2)	107
2.8.3.1.	Test de bondat d'ajust	107
2.8.3.2.	Test χ^2 d'independència (taules de contingència)	112
2.8.3.3.	Test χ^2 d'homogeneïtat de proporcions	115
	Problemes proposats	118

3.1. La relació entre dues variables: covariància, correlació i regressió	122
3.2. Covariància entre dues variables aleatòries.....	122
3.3. Correlació lineal.....	124
3.3.1. Correlació lineal: propietats	125
3.3.2. Criteri per interpretar el coeficient de correlació	126
3.4. Regressió lineal en dues variables	128
3.4.1. Concepte de regressió	128
3.4.2. Classificació de la regressió: simple i múltiple.....	129
3.4.3. Regressió lineal simple.....	129
3.4.4. Error residual.....	131
3.4.5. Solució per mínims quadrats	131
3.4.6. Prediccions	134
3.4.7. Bondat de la regressió: coeficient de determinació.....	135
3.4.8. Variància residual i error típic.....	136
3.4.9. Intervals de confiança en regressió	136
Problemes proposats.....	137
BIBLIOGRAFIA RECOMANADA	143

Presentació

Aquest segon volum de l'obra general *Probabilitat i estadística per a ciències* se centra en l'estadística i les seves bases matemàtiques, com ara la llei dels grans nombres, el teorema de Glivenko-Cantelli i el teorema central del límit. Es treballa especialment la part d'estadística (teoria del mostratge, estimació de paràmetres, càlcul d'interval de confiança, inferència estadística, test d'hipòtesi, correlació i regressió simple) i s'aporten nombrosos exemples desenvolupats i problemes aplicats al camp de la ciència.

D'altra banda, aquest segon volum es basa en material procedent de recursos propis, per exemple, problemes, esquemes i apunts docents, com també apunts i explicacions de diferents assignatures d'estadística que s'imparteixen en els graus de Biologia, Geologia, Biotecnologia, Bioquímica..., i apunts i taules recopilats pels professors del Departament d'Estadística de la Universitat de Barcelona (Pedro Sánchez, Jordi Ocaña, Antoni Arcas, Miquel Salicrú, Miquel Calvo, Miquel de Cáceres, Ángel Vilarroya, entre d'altres).

La novetat principal és que els gràfics, els esquemes i els problemes s'han realitzat íntegrament amb R package, la qual cosa ha significat un enorme esforç d'aprenentatge i de temps per part dels autors.

Volem fer arribar el nostre agraïment als professors Marta Cubedo, Toni Miñarro i Martín Ríos, del Departament d'Estadística de la UB, per revisar el llibre i donar-nos el seu suport. I volem dedicar el llibre a tots els professors del Departament d'Estadística, especialment a Carles Cuadras i a Pedro Sánchez, els nostres millors mestres.

1. TEORIA DEL MOSTRATGE I BASES DE L'ESTADÍSTICA

Introducció

Una variable aleatòria X es caracteritza per la seva distribució de probabilitat, és a dir, pel repartiment de les probabilitats entre tots els valors que prengui. Aquesta distribució es pot representar mitjançant dues funcions que assignen probabilitat (de diferent forma) als valors que pren la variable: 1) Funció de distribució i 2) Funció de probabilitat o funció de densitat.

Recordem que la distribució de probabilitat d'una variable aleatòria és una funció que assigna a cada succés definit sobre la variable aleatòria X la probabilitat que aquest succés ocorri (vegeu *Probabilitat i estadística per a ciències I*). La funció de distribució de probabilitat, $F(X)$, està definida sobre el conjunt de tots els successos, i cada un dels successos és el rang de valors de la variable aleatòria X . La distribució de probabilitat està completament especificada per la funció de distribució $F_X(X = x)$, el valor en cada x real és la probabilitat que la variable aleatòria X sigui menor o igual que x . Tanmateix, la distribució de probabilitat es pot referir a la funció de densitat de probabilitat, $f(x)$ o funció de densitat, o, simplement, densitat d'una variable aleatòria contínua, que descriu la probabilitat relativa segons la qual aquesta variable aleatòria prendrà determinat valor x .

Recordem també que les distribucions de probabilitat distingeixen entre variables discretes i variables contínues, distinció que es basa en el tipus de valors que pot prendre la variable: numerable (normalment finit) o no numerable. Entre les primeres, la més important és la distribució binomial (un cas particular de la distribució multinomial), amb un bon nombre d'aplicacions de caràcter pràctic. Entre les segones, la més important és la distribució normal, a la qual s'ajusten fenòmens de caràcter biològic, geològic, químic, físic, etc. Les distribucions utilitzades més freqüentment en la investigació científica, a més de la distribució normal i la binomial, són F de Snedecor, t de Student i khi quadrat, entre d'altres. La majoria de valors observats sobre variables contínues al nostre voltant solen aproximar-se a una distribució normal. Aquesta és una funció de probabilitat que ofereix un gran interès per les múltiples aplicacions que presenta. Per exemple, l'àrea sota la corba normal està tabulada i s'interpreta en termes de probabilitat, proporció o percentatge.

El teorema central del límit, un dels fonamentals de l'estadística, estudia el comportament de la suma de variables aleatòries quan creix el nombre de sumands, i assegura la seva convergència cap a una distribució normal en condicions molt generals. Aquest teorema, del qual hi ha diferents versions que s'han anat desenvolupant al llarg de la història, té una gran aplicació en inferència estadística, ja que molts paràmetres de diferents distribucions de probabilitat, com la mitjana, poden expressar-se en funció d'una suma de variables. Permet també apropar moltes distribucions d'ús freqüent: Bernoulli, binomial, Poisson, khi quadrat, t de Student, gamma, etc., quan els seus paràmetres creixen i el càlcul es fa difícil. D'altra banda, la suma de variables aleatòries apareix de manera natural en moltes aplicacions de la ciència: determinació de biomassa, concentració d'un compost, recompte d'organismes o tipus, temps d'espera de serveis, càrrega suportada per una estructura, etc. Tot això explica per què molts mètodes estadístics requereixen la condició de normalitat per a la seva aplicació correcta. Aquest teorema dóna suport i relaciona entre si conceptes i pro-

cediments bàsics en estadística, com els de variable aleatòria i les seves transformacions, distribució mostral, convergència, tipificació, càlcul de probabilitats, etc.

1.1. Aproximació de distribucions de probabilitat

Recordem que en la teoria de la probabilitat i en estadística, una funció de distribució acumulada $F_X(X = x)$ és la probabilitat que una variable aleatòria real X subjecta a certa llei de distribució de probabilitat se situï a la zona de valors menors o iguals a x .

Intuïtivament, assumint la funció de densitat de probabilitat $f(x)$ com la llei de distribució de probabilitat, $F(X)$ seria la funció amb la recta real com a domini, amb imatge de l'àrea fins aquí de la funció de densitat $f(x)$, i aquí el valor x és per a la variable aleatòria real X .

Com ja sabem, les variables aleatòries es classifiquen en discretes i contínues:

- *Variables aleatòries discretes*: la variable aleatòria X és discreta si els nombres assignats als esdeveniments elementals de l'espai mostral Ω són punts aïllats. Els seus possibles valors constitueixen un conjunt finit o infinit numerable. Per exemple, suposem l'experiment consistent a llançar tres vegades una moneda no trucada; si considerem la variable aleatòria X = nombre de cares obtingudes en els dos llançaments, els valors que pot prendre aquesta variable aleatòria són finits (0, 1, 2).
- *Variable aleatòria contínua*: la variable aleatòria X és contínua si els valors assignats poden ser qualssevol dins de certs intervals, és a dir, poden prendre qualsevol valor de R . Per exemple, si considerem l'experiment aleatori consistent a mesurar el nivell d'aigua freàtica en un terreny argilós i prenem la variable aleatòria X = nivell d'aigua, aquesta variable pot prendre valors entre 0 i més infinit.

A continuació veurem operacions senzilles amb variables aleatòries conegudes, com binomials, Poisson i normal, i què succeeix quan les sumem.

1.1.1. Suma de variables aleatòries amb distribució binomial

Considerem que la probabilitat que té una moneda no trucada d'obtenir cara és igual que la d'obtenir creu ($p = 0,5$).

Si llancem 7 cops una moneda i X_1 és el nombre de cares, la distribució que porta associada és $X_1 \sim \text{Bin}(7, 0,5)$.

Si llancem 7 cops més la moneda i X_2 és el nombre de cares, la distribució és també $X_2 \sim \text{Bin}(7, 0,5)$.

Com que les dues seqüències de llançament són independents, la probabilitat d'obtenir cara és sempre la mateixa i és equivalent a considerar el nombre de cares sumant X_1 i X_2 .

Resulta immediat: $(X_1 + X_2) \sim \text{Bin}(14, 0,5)$.

Així, doncs, donades dues variables amb distribucions binomials independents amb el mateix paràmetre p : $B_1 \sim \text{Bin}(n_1, p)$ i $B_2 \sim \text{Bin}(n_2, p)$, la variable suma de les anteriors també segueix una distribució binomial:

$$(B_1 + B_2) \sim \text{Bin}(n_1 + n_2, p)$$

És a dir, és reproducible.

I això també és cert per a una suma finita de variables aleatòries d'aquest tipus.

1.1.2. Reproductibilitat de famílies de distribucions

Aquesta propietat es coneix com a reproductibilitat: en sumar dues variables aleatòries independents d'una mateixa família de funcions de distribució, s'obté una nova variable aleatòria de la mateixa família.

Les tres famílies que vam presentar en el volum anterior tenen aquesta propietat:

- Si $B_1 \sim \text{Bin}(n_1, p)$ i $B_2 \sim \text{Bin}(n_2, p)$, llavors $(B_1 + B_2) \sim \text{Bin}(n_1 + n_2, p)$.
- Si $X_1 \sim \text{Poisson}(\lambda_1)$ i $X_2 \sim \text{Poisson}(\lambda_2)$, llavors $(X_1 + X_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$.
- Si $Y_1 \sim N(\mu_1, \sigma_1)$ i $Y_2 \sim N(\mu_2, \sigma_2)$, llavors $(Y_1 + Y_2) \sim N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

Però aquesta propietat no és general i no és certa en totes les famílies de distribucions de variables aleatòries, com veurem en els exemples següents:

Penseu en la tirada d'un dau: $X = \text{puntuació d'un dau}$.

Imaginem que el dau és regular i no trucat. La funció de densitat és:

$$f_X(x) = P(X = x) = \frac{1}{6} \quad x = 1, 2, \dots, 6$$

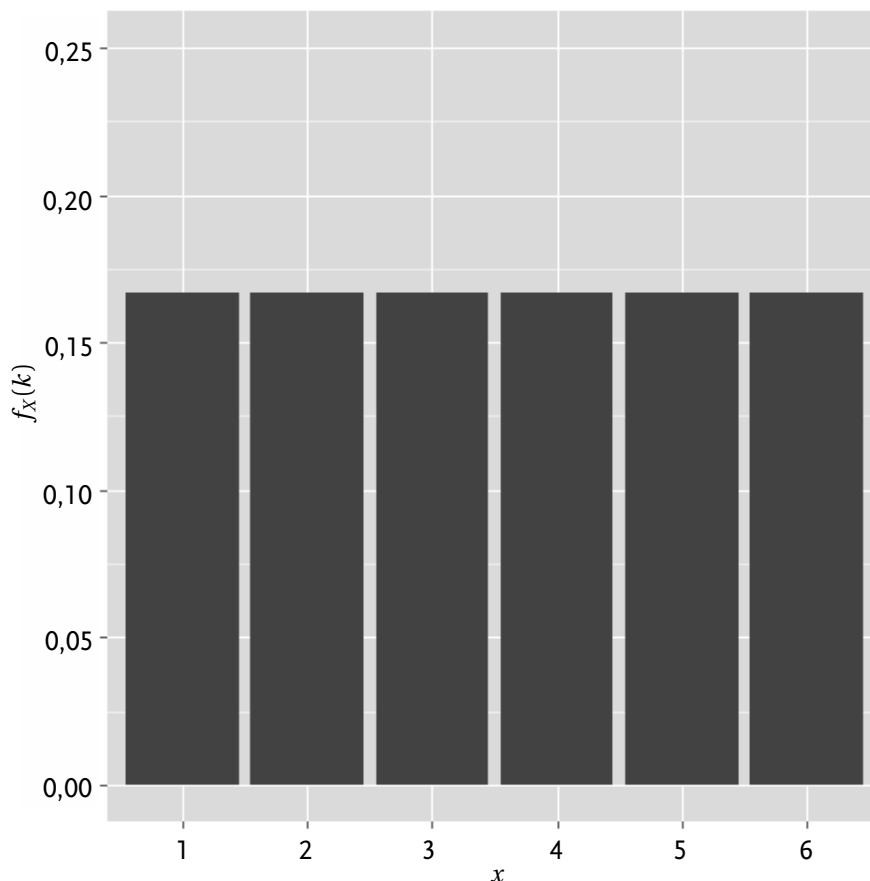


Figura 1. Funció de densitat per a la variable aleatòria «tirada d'un dau».

En general, aquesta distribució s'anomena uniforme discreta i no és reproduïble. Penseu en la tirada de dos daus.

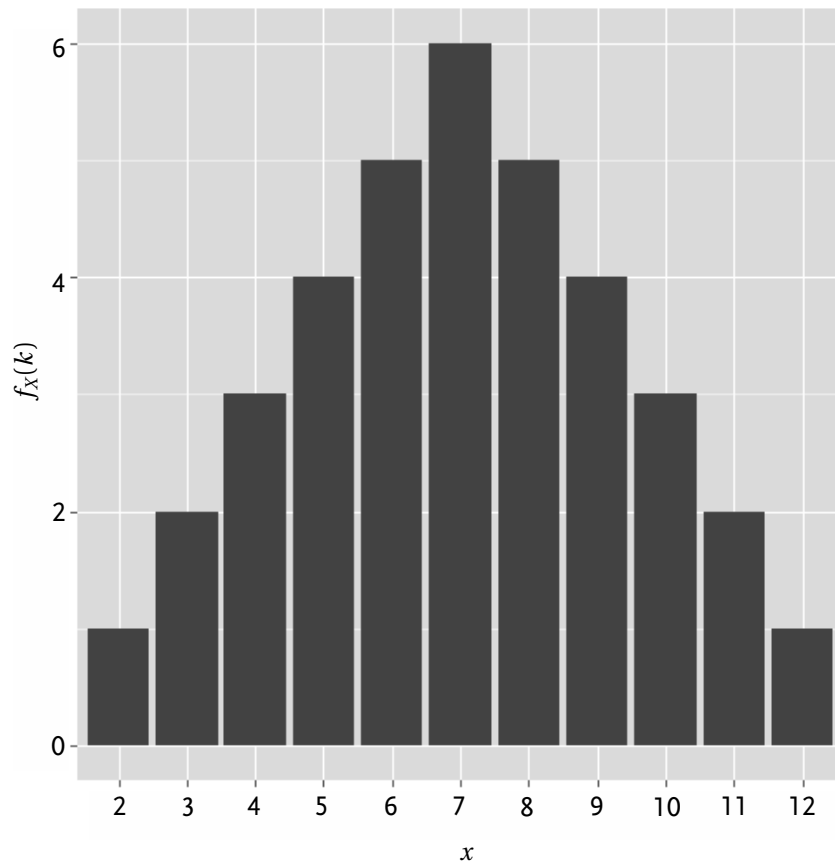


Figura 2. Funció de densitat per a la variable aleatòria tirada de dos daus.

Si sumem el resultat de tirar dos daus ($X_1 + X_2$), en què X_1 és el resultat del primer dau i X_2 és el resultat del segon dau, tenim una funció de densitat de probabilitat que no es pot modelar mitjançant una funció de densitat uniforme i, per tant, no és reproducible:

$$f_X(x) = P(X = x) \neq \frac{1}{12} \quad x = 2, 3, \dots, 12$$

1.1.3. Aproximació de funcions de distribució

La suma de dues variables aleatòries podem ampliar-la a sumes de tres, quatre, ..., n variables aleatòries.

$$\begin{aligned} S_1 &= X_1 \\ S_2 &= X_1 + X_2 \\ &\vdots \\ S_n &= X_1 + \dots + X_n \end{aligned}$$

Cadascuna de les sumes S_n és una variable aleatòria amb una funció de distribució coneguda que convergirà en determinades condicions.

Ja sabem que si la família és reproductiva, la variable suma és de la mateixa família.

D'altra banda, si el nombre de sumands (n) és gran, sota certes condicions podem aproximar la distribució de la variable suma per a una distribució d'una altra família.

Com més gran sigui n , millor serà l'aproximació. Ara veurem alguns exemples de les situacions més típiques.

1.1.4. Aproximació de la distribució binomial a la distribució de Poisson

Com vam veure en el volum anterior, una distribució de Bernoulli es correspon amb una distribució binomial amb un sol intent: $Z \sim \text{Bin}(1, p)$.

Per tant, podem expressar una variable aleatòria binomial $X \sim \text{Bin}(n, p)$ com una suma de n sumands amb distribució de Bernoulli:

$$X = Z_1 + \dots + Z_n$$

Una suma de n variables aleatòries amb distribució Bernoulli amb paràmetre p petit (o una binomial amb paràmetres n i p) s'aproxima a una distribució de Poisson quan $n \geq 30$ i $p < 0,1$. Aleshores:

$$X \sim \text{Bin}(n, p) \rightarrow P(\lambda = np)$$

La distribució de probabilitat de Poisson resultant té paràmetre $\lambda = np$, atès que l'esperança de la distribució binomial és $E(X) = np$ i el paràmetre λ és l'esperança de la funció de distribució de Poisson.

$$X \sim \text{Bin}(n = 32, p = 0,05) \rightarrow P(\lambda = np = 1,6)$$

Aquesta aproximació es pot veure perfectament en la figura següent:

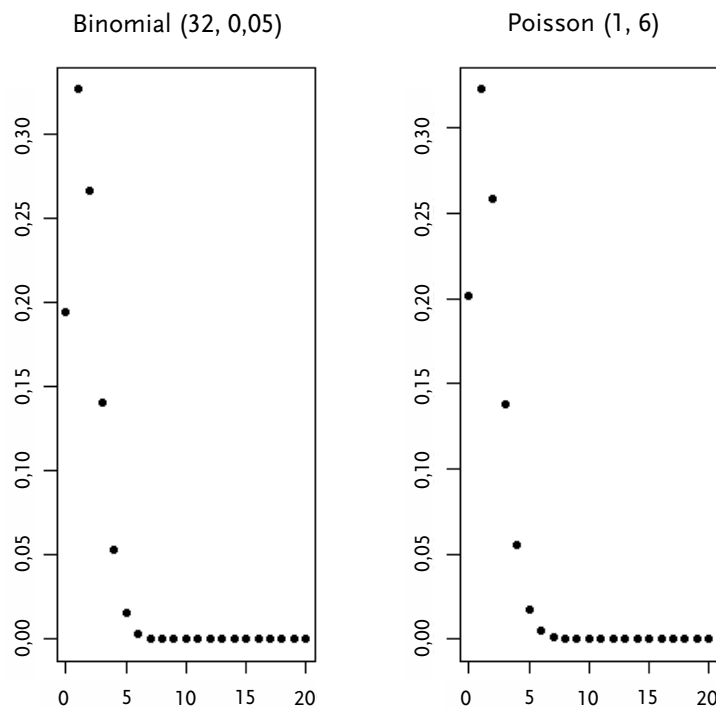


Figura 3. Aproximació de la distribució binomial a la distribució de Poisson; exemple amb el cas: $X \sim \text{Bin}(n = 32, p = 0,05) \rightarrow P(\lambda = np = 1,6)$.

Encara que si utilitzem un ordinador per fer els càlculs no ens calen les aproximacions, aproximar una distribució binomial a una distribució de Poisson ens pot ser útil per facilitar el càlcul de probabilitats quan fem exercicis. Per exemple, seria impossible o molt llarg calcular una probabilitat lligada a la següent funció de distribució binomial: $X \sim \text{Bin}(10.000, 0,003)$ i, en conseqüència, l'aproximem a una distribució $P(\lambda = np = 30)$.

1.1.5. Aproximació de la distribució binomial a la distribució normal

S'utilitza l'aproximació normal per avaluar probabilitats binomials sempre que p no estigui proper a 0 o 1. L'aproximació és excel·lent quan n és gran, i és bastant bona per a valors petits de n si p està raonablement propera a 0,5.

Aquesta aproximació es pot demostrar matemàticament, ja que una suma de n variables de Bernoulli amb paràmetre p mitjà ($0,1 < p < 0,9$) s'aproxima si es compleixen les condicions següents:

$$np \geq 5 \quad \text{o bé} \quad (n > 30) \quad \text{i} \quad (1 < p < 0,9)$$

a una distribució normal de la manera següent:

$$X \sim \text{Bin}(n, p) \rightarrow N(np, \sqrt{npq}), \quad \text{en què} \quad q = 1 - p$$

en què l'esperança i la variància són:

$$\begin{aligned} E(X) &= np & \mu &= E(X) = np \\ \text{Var}(X) &= npq & \sigma &= \sqrt{\text{Var}(X)} = \sqrt{npq} \end{aligned}$$

Com a exercici, veurem què succeeix si s'utilitza una aproximació dolenta de binomial a normal $np < 5$ i posteriorment què passa si s'utilitza una aproximació excel·lent en el cas $np = 15$:

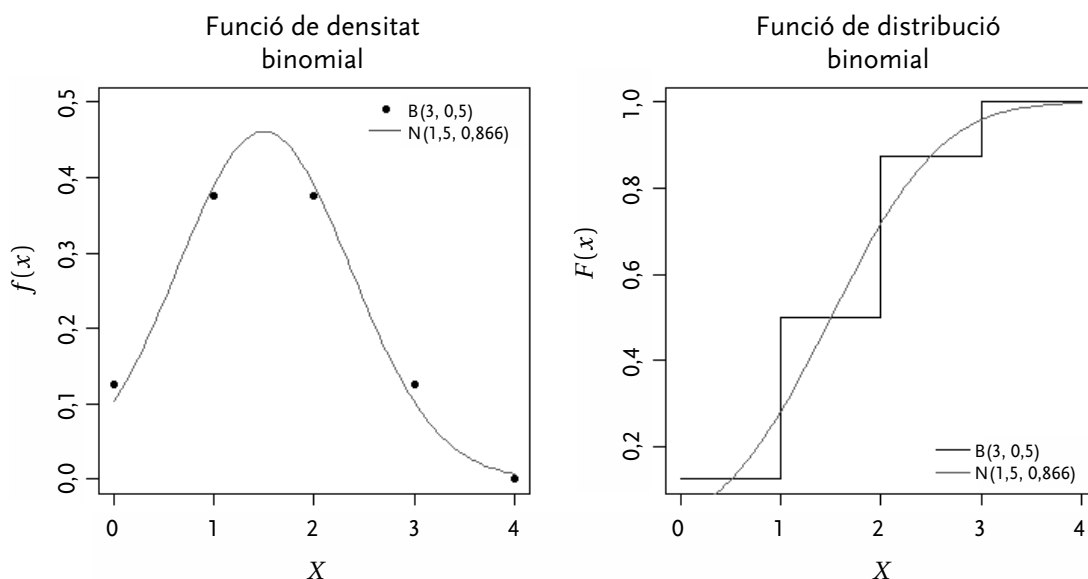


Figura 4. Simulació de l'aproximació incorrecta d'una distribució binomial a una normal

$$X \sim \text{Bin}(n = 3, p = 0,5) \rightarrow N(np = 3, \sqrt{npq} = 0,5).$$

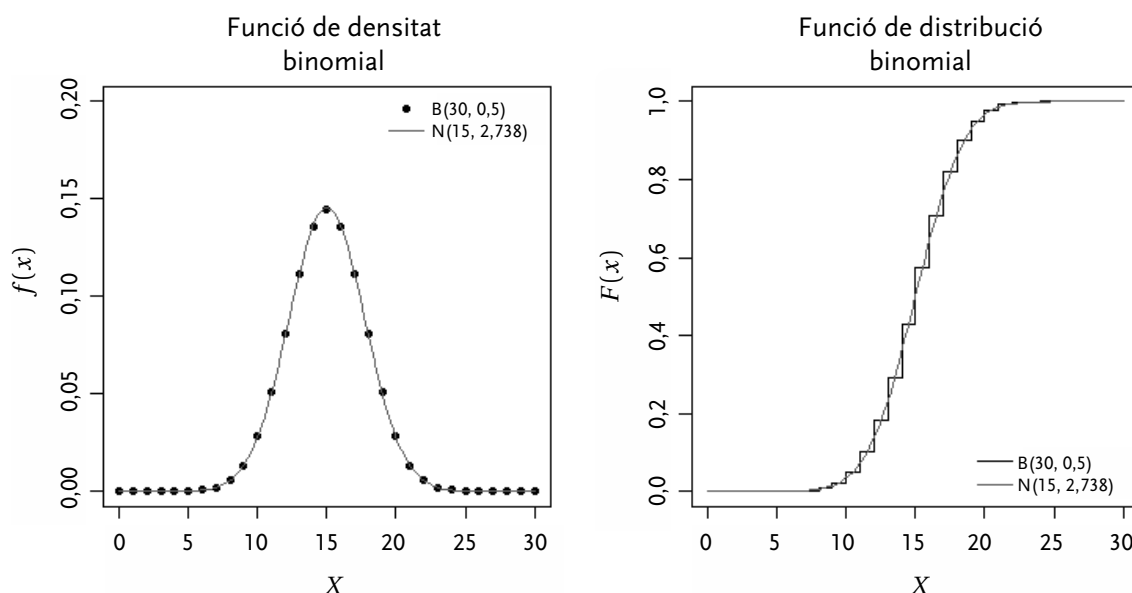


Figura 5. Simulació d'una aproximació excel·lent d'una distribució binomial a una normal
 $X \sim \text{Bin}(n = 30, p = 0,5) \rightarrow N(np = 15, 2, \sqrt{npq} = 2,738)$.

Vegem un exemple de càlcul amb aquesta aproximació de distribucions. La probabilitat que un pacient es recuperi d'una malaltia infecciosa greu és 0,4. Si se sap que 100 persones han contret aquesta malaltia, quina és la probabilitat que almenys 30 persones infectades sobrevisquin?

Si $n = 100$, $p = P(\text{recuperar-se}) = 0,40$

$q = P(\text{no recuperar-se}) = 1 - p = 1 - 0,40 = 0,60$

$\mu = np = 100 \cdot 0,4 = 40$ pacients es recuperen com a mitjana

$\sigma = \sqrt{npq} = \sqrt{100 \cdot 0,4 \cdot 0,6} = 4,899$ pacients com a desviació estàndard

Si indiquem com a X els pacients que es recuperen, $X = 0, 1, 2, \dots, 100$.

Abans de prosseguir la resolució del problema amb l'aproximació normal, és bo aclarir que s'estan avaluant probabilitats associades a una variable discreta X amb una distribució que avalua variables de tipus continu, com és la distribució normal, de manera que hem d'introduir una petita correcció a l'hora de calcular el valor estandarditzat de Z , que correspon a $Z = 30$, per la qual cosa Z pateix un petit canvi, com es mostra a continuació:

$$Z = \frac{(X \pm \frac{1}{2}) - \mu}{\sigma}$$

Aquest és un factor de correcció perquè s'està avaluant una variable discreta amb una distribució contínua, de manera que cal delimitar clarament des de quin punt s'avalua la variable; dit d'una altra manera, en quin límit de la barra (inferior o superior) ens hem de posicionar per determinar la probabilitat requerida. Cada barra de probabilitat que s'ha d'avaluar té com a base la unitat: aquest és el perquè del $\pm 1/2$.

$$Z = \frac{(X - \frac{1}{2}) - \mu}{\sigma} = \frac{(30 - \frac{1}{2}) - 40}{4,899} = \frac{29,5 - 40}{4,899} = -2,1433$$

Aquest valor és negatiu, però habitualment, si ho busquem en la taula de la distribució normal tipificada $N(0, 1)$, sols identifiquem probabilitats per a valors $Z \geq 0$ i $P(Z \leq z)$, per la qual cosa podem fer la transformació següent:

$$P(X \geq 30) = P(Z \geq -2,14) = P(Z \leq 2,14) = 0,9838 = 98,38\%$$

Aquest problema és un dels molts en què pot donar-se aquesta aproximació. Per evitar-lo en l'aproximació $X \sim \text{Bin}(n, p) \rightarrow N(np, \sqrt{npq})$, els valors fixos es corregeixen (correcció de continuïtat de Yates) i se substitueixen per un interval centrat en el punt i de valor 1. En l'esquema següent es mostren totes les situacions possibles i com s'aplica aquesta correcció:

Si $X \sim B(n, p)$ i $X' \sim N(n \cdot p)$

- $P(X = a) = P(a - 0,5 \leq X' \leq a + 0,5)$
- $P(X \leq a) = P(X' \leq a + 0,5)$ (perquè contingui el punt a)
- $P(X < a) = P(X' \leq a - 0,5)$ (perquè no contingui el punt a)
- $P(X > a) = P(X' \geq a + 0,5)$ (perquè no contingui el punt a)
- $P(X \geq a) = P(X' \geq a - 0,5)$ (perquè contingui el punt a)
- $P(a \leq X < b) = P(a - 0,5 \leq X' \leq b + 0,5)$ (perquè contingui el punt a i no el punt b)

1.1.6. Aproximació de la distribució de Poisson a la distribució normal

Les variables aleatòries amb distribució de Poisson independents també són reproductives. Per tant, una variable aleatòria X amb distribució de Poisson pot considerar-se una suma de n variables independents amb distribució de Poisson:

$$X = (X_1 + \dots + X_n) \sim \text{Poisson}(\lambda) \quad \text{i cada} \quad X_i \sim \text{Poisson}(\lambda/n)$$

Encara que per a un espai finit les distribucions de Poisson i la normal no coincideixen, és possible aproximar la primera per mitjà de la segona, d'acord amb la regla següent:

$\lambda \geq 10$	aproximar a la distribució normal de mitjana λ , variància λ
$\lambda < 10$	no aproximar, calcular amb la variable aleatòria original amb distribució de Poisson

Així, una suma de n variables aleatòries amb distribució de Poisson s'aproxima a una distribució normal. Com que qualsevol variable aleatòria amb distribució de Poisson pot ser considerada una suma de variables aleatòries amb distribució de Poisson, les distribucions de Poisson amb paràmetre λ gran (quan $\lambda \geq 10$) poden aproximar-se per a distribucions normals.

$$X \sim \text{Poisson}(\lambda) \rightarrow N(\lambda, \sqrt{\lambda})$$

Amb esperança i variància:

$$\begin{aligned} E(X) &= \lambda \longrightarrow \mu = E(X) = \lambda \\ \text{Var}(X) &= \lambda \longrightarrow \sigma = \sqrt{\text{Var}(X)} = \sqrt{\lambda} \end{aligned}$$

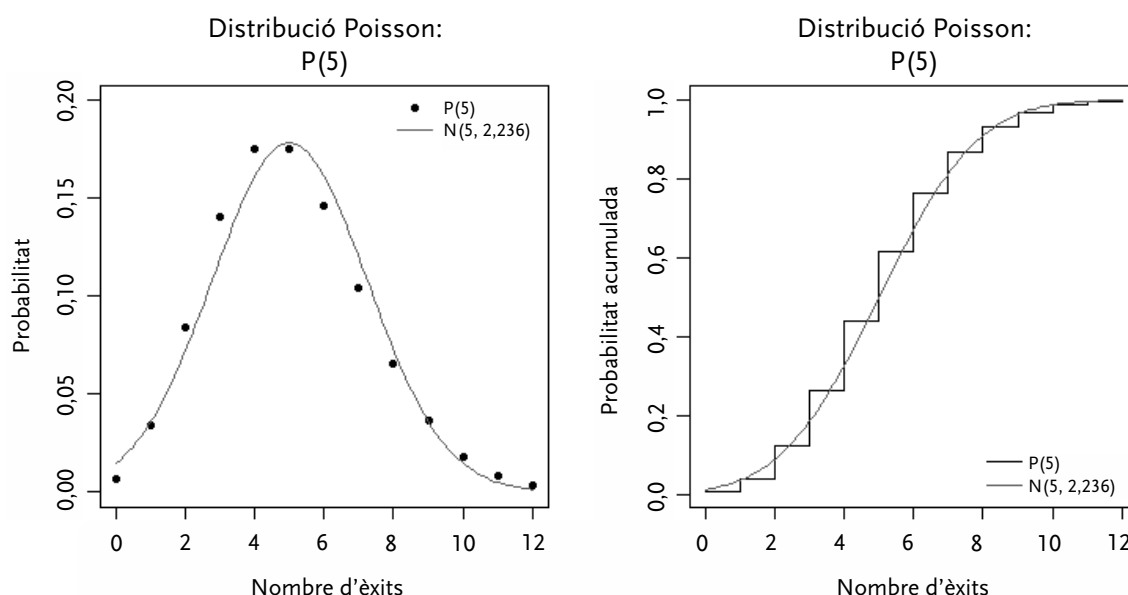


Figura 6. Simulació de l'aproximació incorrecta d'una distribució de Poisson a una distribució normal (a la dreta, la funció de distribució $F(X)$; a l'esquerra, la funció de densitat $f(X)$).

$$X \sim \text{Poisson}(\lambda = 5) \rightarrow N(\lambda = 5, 2.236)$$

Com a exercici (figura 6) veurem què succeeix si s'utilitza una aproximació «dolenta» de distribució Poisson a distribució normal $\lambda = 5$. Aproximació «dolenta» ($\lambda = 5$). Podem veure en la figura corresponent que no s'acaba d'aproximar ni la funció de densitat ni la funció de distribució de la nova variable aleatòria aproximada amb una distribució normal, ja que $\lambda < 10$.

Si ara intentem aproximar una nova variable aleatòria (figura 7) amb una distribució de Poisson amb $\lambda = 15$, veiem que la figura adient sí que presenta una bona aproximació.

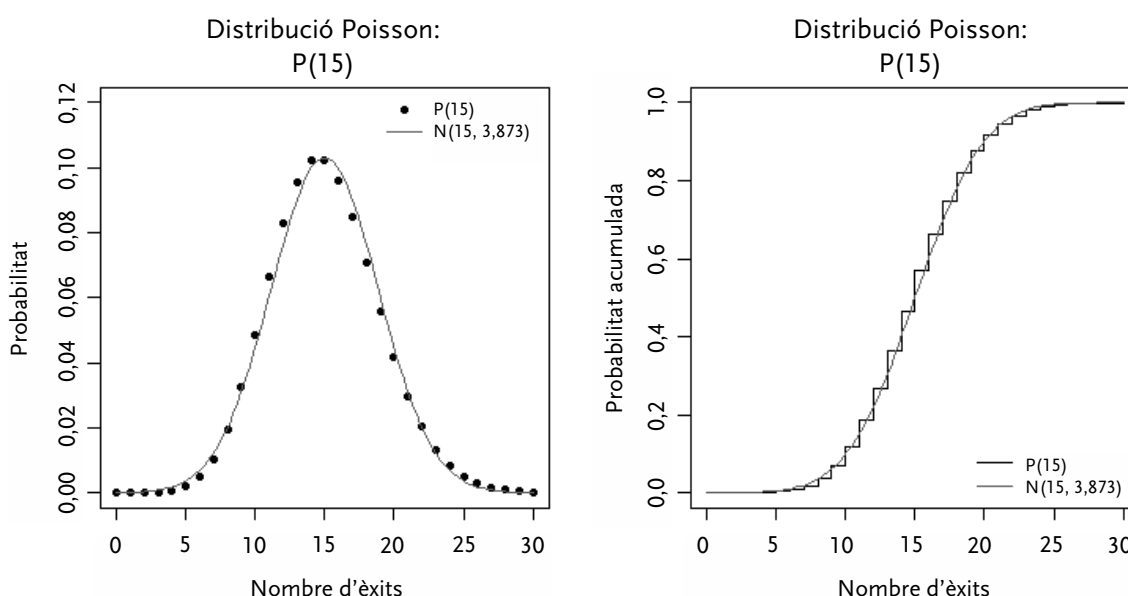


Figura 7. Simulació de l'aproximació d'una distribució de Poisson a una distribució normal correcta (a la dreta, la funció de distribució $F(X)$; a l'esquerra, la funció de densitat $f(X)$).

$$X \sim \text{Poisson}(\lambda = 15) \rightarrow N(\lambda = 15, 3.873)$$