Càlcul de Probabilitats

Bloc 1 – Probabilitat i Estadística

Setembre 2014

Índex

Objectiu

Diferenciar el tipus de probabilitat segons el denominador

Programa	Compila	No compila	Total
C++	72	48	120
Java	64	16	80
Total	136	64	200

Programa	Compila	No compila	Total
C++	0.36	0.24	0.60
Java	0.32	0.08	0.40
Total	0.68	0.32	1.00

Quina és la probabilitat de que s'executi en C++ i compili? **0.36 (72/200)**

Programa	Compila	No compila	Total
C++	0.60	0.40	1.00
Java	0.80	0.20	1.00
Total	-	-	-

Quina és la probabilitat de que compili un programa en C++? **0.60 (72/120)**

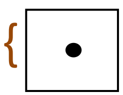
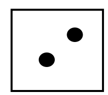
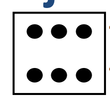
Programa	Compila	No compila	Total
C++	0.53	0.75	-
Java	0.47	0.25	-
Total	1.00	1.00	-

Quina és la probabilitat que provingui de C++ si ha compilat? **0.53 (72/136)**

Experiència aleatòria. Definicions

- Els **fenòmens deterministes** porten a uns mateixos resultats a partir d'unes mateixes condicions inicials. [Ex: si poso la mà al foc, em cremaré]

- Els **fenòmens aleatoris** tenen una certa incertesa en el resultat d'una propera realització de l'experiència aleatòria. [Ex: Si llenço un dau, no sé quin número sortirà]

- Tota experiència aleatòria té associat un **conjunt de resultats possibles** (Ω ={w1,w2,...}) [Ex: en un dau, Ω ={ ... }]

- Qualsevol subconjunt de Ω és un **esdeveniment o succés** (A,B,...). [Ex: Ω (segur) o $\varnothing$ (impossible)]

- Una **partició** és un conjunt d'esdeveniments $A_i \neq \varnothing$, disjunts i que la seva unió és Ω. [Ex: en un dau,
  A_1 = # parell ; A_2 = # senar → $A_1 \cup A_2 = \Omega$ i $A_1 \cap A_2 = \varnothing$]

Experiència aleatòria. Exemples

- Simples
  - llençar una moneda → Ω ={cara, creu}
  - llençar un dau → Ω ={1,2,3,4,5,6}
  - extreure una bola d'una urna (blanca, negra) → Ω ={b,n}

- Complexes
  - extreure amb reposició dues boles d'una urna → Ω ={bb,bn,nb,nn}
  - segons el resultat del llançament d'una moneda (c,+), escollir una urna d'entre dues amb composició diferent, i extreure'n una bola (b,n):

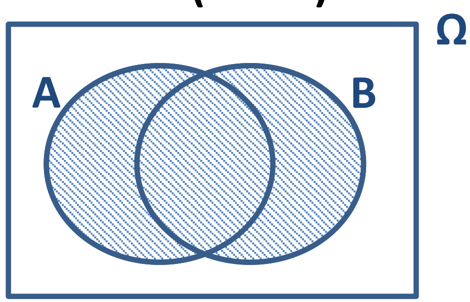
    Ω ={cb, cn, +b, +n}
  - cas servidor i xarxa: possibilitats segons si servidor i/o xarxa funcionen (y) o no (n)

    Ω ={yy, yn, ny, nn}

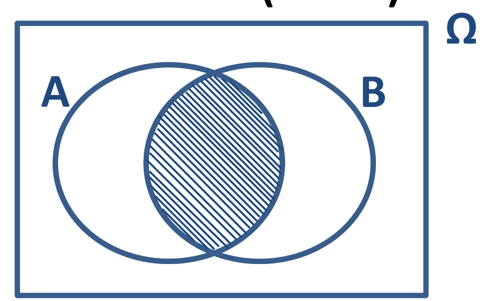
Experiència aleatòria. Operacions amb conjunts

Com que els esdeveniments són conjunts, totes les operacions dels conjunts es poden aplicar, i el resultat és un altre esdeveniment.

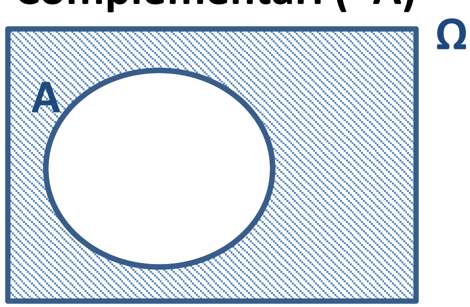
Unió (A U B)



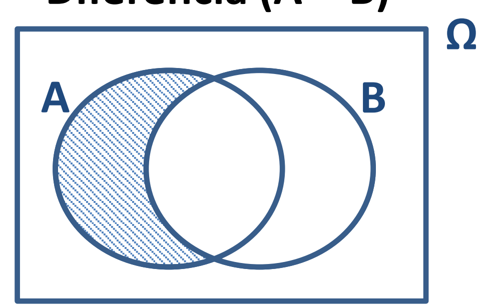
Intersecció (A ∩ B)



Complementari (¬A)



Diferència (A – B)



Definicions: Dos conjunts, A i B són **complementaris** (o formen una **partició**) si $A \cap B = \varnothing$ i $A \cup B = \Omega$
Dos conjunts, A i B són **disjunts** si $A \cap B = \varnothing$

Experiència aleatòria. Exercicis

- Trobar Ω en els següents casos
  - nombre de defectes (“tares”) en una peça industrial

    Solució: $\Omega = \{ \qquad \}$
  - treure dues boles d’una urna amb 4 boles negres i una blanca

    Solució: $\Omega = \{ \qquad \}$
  - diferència en valor absolut entre el nombre de cares i creus en 10 tirades

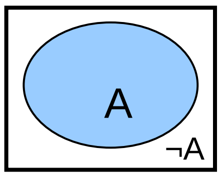
    Solució: $\Omega = \{ \qquad \}$

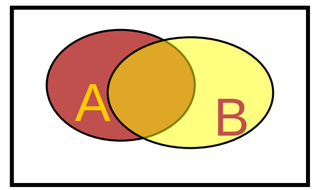
- Operacions amb conjunts. Verifiqueu:
  - $A \cap B = \neg(\neg A \cup \neg B)$
  - $A \cup B = \neg(\neg A \cap \neg B)$

Experiència aleatòria. Representacions de succesos

Usem representacions gràfiques per visualitzar millor el procés d'una experiència aleatoria

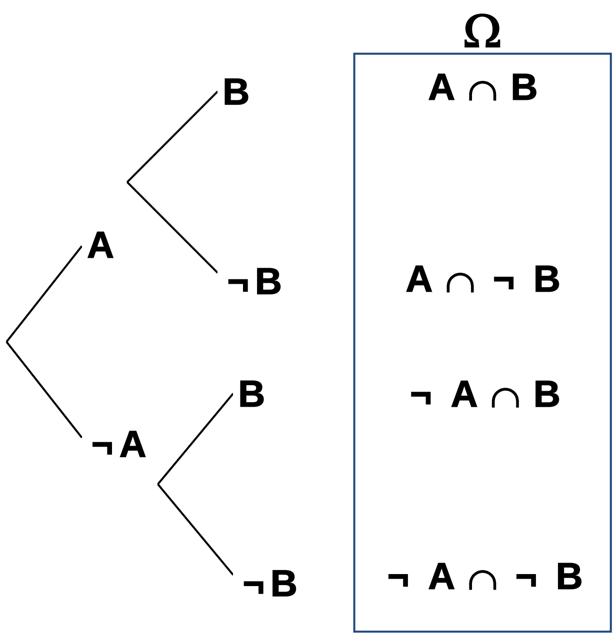
- Conjunts/diagrames de Venn



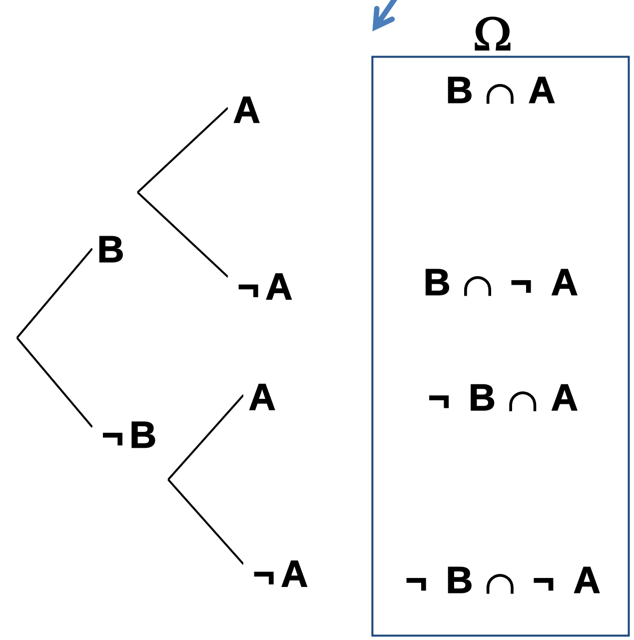


Atenció: No tenen perquè estar endreçats cronològicament!

- Arbres



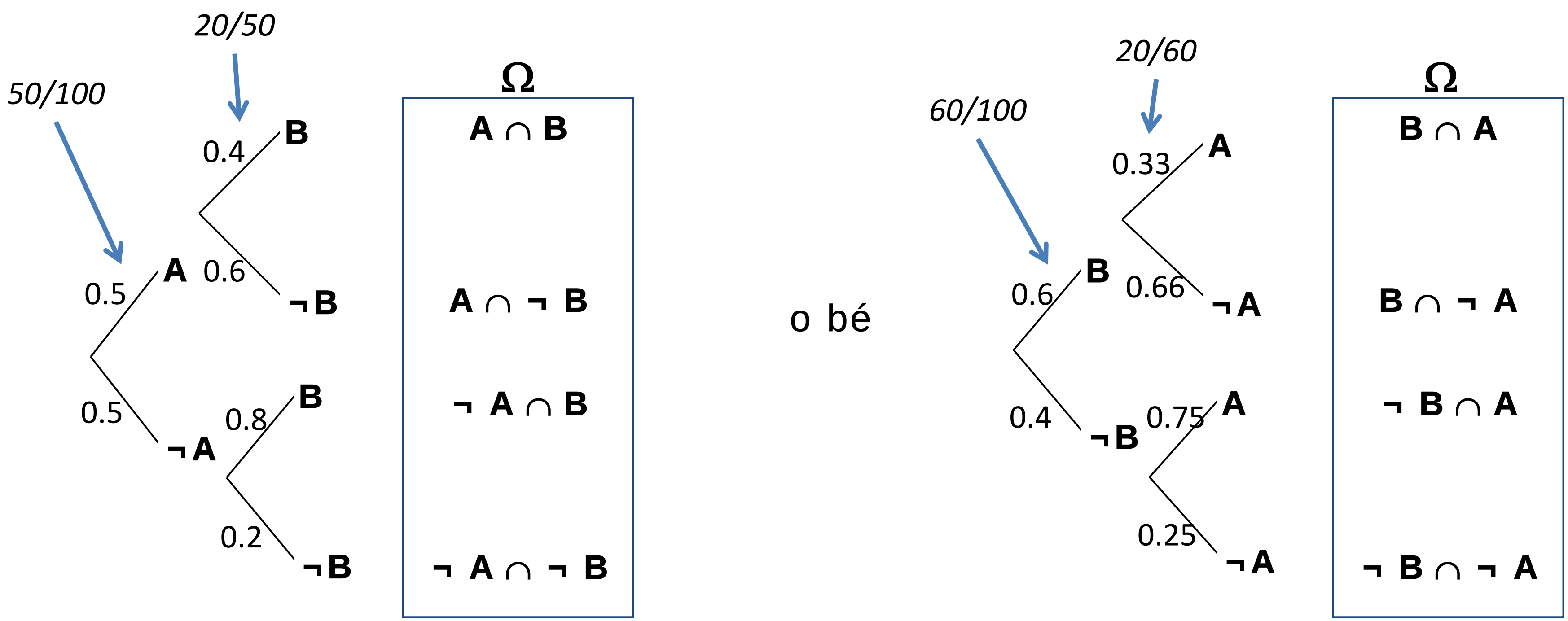
o bé



Experiència aleatòria. Arbre a partir de taula 2x2

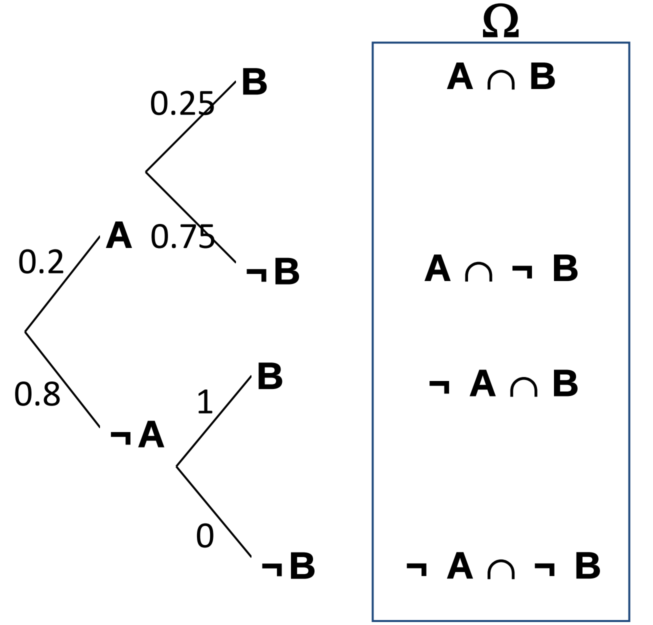
Amb aquesta taula, es poden construir 2 arbres

	B	¬B	Total
A	20	30	50
¬A	40	10	50
Total	60	40	100



Exp. aleatòria. Exercici: Taula 2x2 a partir d'arbre

Amb aquest arbre i el número total (n=100) es pot construir la taula. **Prova-ho!**



	B	¬B	Total
A			
¬A			
Total			100

Probabilitat. Definició i propietats

- Per quantificar la incertesa podem definir **una aplicació** que **a cada succés li fa correspondre un valor entre 0 i 1** que anomenem probabilitat
- Propietats per definició:
  - $0 \leq P(A) \leq 1$
  - $P(A_0 \cup A_1 \cup \ldots \cup A_n) = P(A_0) + P(A_1) + \ldots + P(A_n)$ si $A_i \cap A_j = \varnothing$ per $i \neq j$
  - $P(\Omega) = 1$
- Propietats deduïdes:
  - $P(\neg A)=1 - P(A)$
  - $P(\varnothing) = 0$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

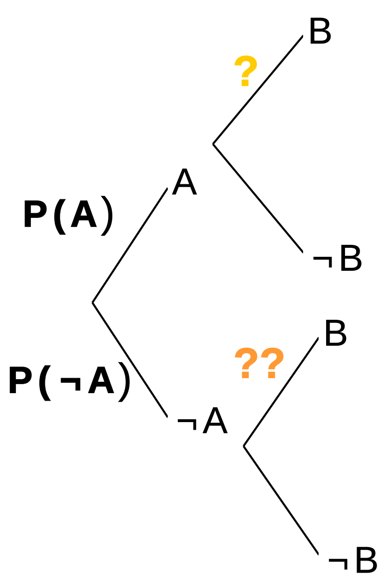
NOTA: El cas particular en qué la probabilitat s'obté de casos favorables/casos totals es presenta en condicions d'*equiprobabilitat* (tots els succesos elementals tenen la mateixa probabilitat). No es pot generalitzar a qualsevol experiència! [Ex: es podria aplicar a un llançament d'una moneda però no als possibles resultats d'una travessa]

Probabilitat. Independencia

Independència aplicat a 2 (o més) esdeveniments és defineix com:

A i B són independents sii $P(A \cap B) = P(A)\cdot P(B)$

[***3 o més succesos són independents si són independents 2 a 2***]



$\mathbf{P(A \cap B)} = P(A) \cdot P(B)$

$\mathbf{P(A \cap \neg B)} = P(A) \cdot P(\neg B)$

$\mathbf{P(\neg A \cap B)} = P(\neg A) \cdot P(B)$

$\mathbf{P(\neg A \cap \neg B)} = P(\neg A) \cdot P(\neg B)$

Podem associar probabilitats a cada branca.

Si A i B són independents, obtenim que a l'arbre:

a) "?" és P(B)

b) "??" és P(B)

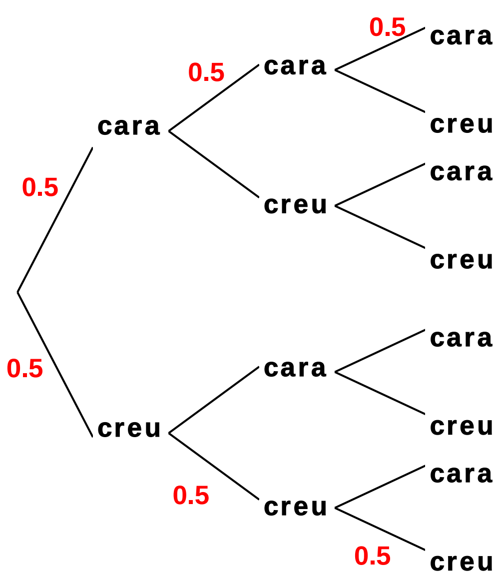
c) "?" = "??"

Probabilitat. Independencia. Exemple

- Estudiarem l'experiència aleatòria de **llençar una moneda equilibrada <u>tres</u> vegades.**
- Abans de fer cap realització i coneixent les característiques de l'experiència, calcularem:
  - P(A) sent A = "obtenir **2 cares"**
  - P(B) sent B = "obtenir **almenys 2 cares"**

(*al bloc 2 es calcularan les probabilitats mitjançant una variable aleatòria*)

Independència



(Cara,Cara,Cara) = w1	P(w1) = P(C)·P(C)·P(C) = 1/8
(Cara,Cara,Creu) = w2	P(w2) = 1/8
(Cara,Creu,Cara) = w3	P(w3) = 1/8
(Cara,Creu,Creu) = w4	P(w4) = 1/8
(Creu,Cara,Cara) = w5	P(w5) = 1/8
(Creu,Cara,Creu) = w6	P(w6) = 1/8
(Creu,Creu,Cara) = w7	P(w7) = 1/8
(Creu,Creu,Creu) = w8	P(w8) = 1/8

Probabilitat. Independencia. Exemple

Calcular:

- A = “Obtenir dues cares”
- B = “Obtenir almenys dues cares”

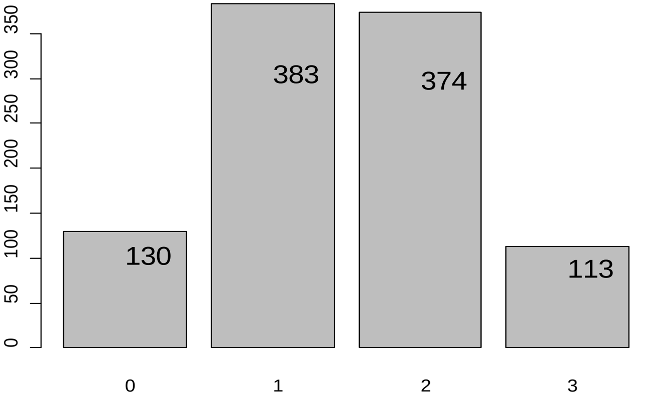
Solució: si es coneixen tots els resultats d’una experiència aleatòria, i la probabilitat de cada esdeveniment elemental, acumulem probabilitats:

$$P(A) = P(w2) + P(w3) + P(w5) = 1/8+1/8+1/8 = 3/8$$

$$P(B) = P(w2) + P(w3) + P(w5) + P(w1) = 1/8+1/8+1/8+1/8 = 1/2$$

Nota: Per altra part, si l’experiència es realitza repetidament, *observarem* unes freqüències semblants a les probabilitats de l’esdeveniment.

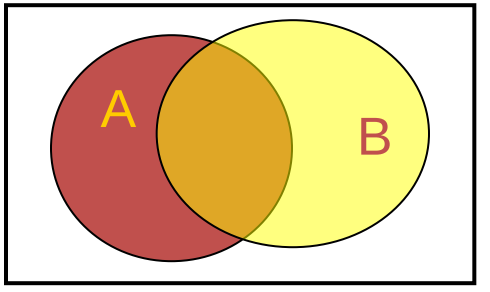
(*però això és un tema d’estadística descriptiva*)



Probabilitat condicionada

- Quan un esdeveniment afecta a l'expectativa d'un altre parlem de **probabilitat condicionada P(A|B)** que es llegeix com a "probabilitat d'observar A tenint en compte que s'ha realizat B" (o senzillament, "probabilitat de A condicionada per B") [Ex: A="Ploure"/B="Estar ennuvolat"]
- Es defineix P(A|B), si P(B)>0:

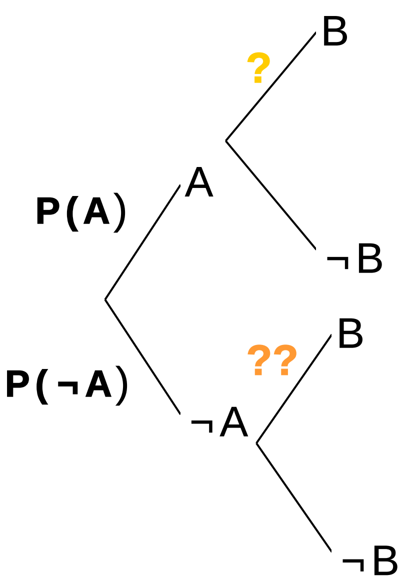
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



- A la pràctica, **condicionar per B significa que reduïm a B el conjunt de resultats observables**, i les probabilitats han de recalcular-se respecte P(B)
- En considerar P(A|B), **cada esdeveniment juga un paper diferent**: A és incert però B és conegut
- En general, **P(A|B) ≠ P(B|A) ≠ P(A ∩ B)** [Ex: A="Fumar"/B="Tenir cancer de pulmó". La probabilitat de ser fumador si tens cancer de pulmó és més alta que no la inversa]

Probabilitat condicionada. Representació

Els arbres d’esdeveniments incorporen les probabilitats condicionades.



$\mathbf{P(A \cap B)} = P(A) \cdot P(B|A)$

$\mathbf{P(A \cap \neg B)} = P(A) \cdot P(\neg B|A)$

$\mathbf{P(\neg A \cap B)} = P(\neg A) \cdot P(B|\neg A)$

$\mathbf{P(\neg A \cap \neg B)} = P(\neg A) \cdot P(\neg B|\neg A)$

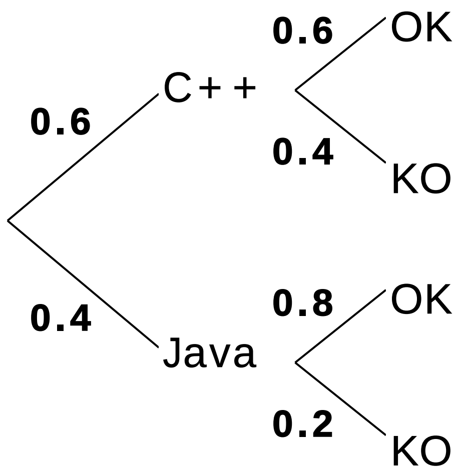
La branca que va de ‘A’ a ‘B’ porta una probabilitat condicionada P(B|A) (?): estem suposant que ‘A’ ha passat.

P(A∩B) és el producte de les probabilitats en el camí des del node arrel fins al node A∩B:

P(A∩B) = P(A) P(B|A)

Prob. condicionada. Representació. Exemples

Les probailitats condicionades es coloquen a les branques dels arbres:



El 80% dels programes Java compila a la primera… però **quin % dels programes que compilen a la primera estan escrits en Java?**

Sol: 32/68

	OK	¬OK	Total
C++	0.36	0.24	0.6
Java	0.32	0.08	0.4
Total	0.68	0.32	1

Independència i Probabilitat Condicionada

- Si A i B són independents

$$P(A|B) = P(A\cap B)/P(B) \overset{\text{Independents}}{=} P(A)\cdot P(B)/P(B) = P(A)$$

- Que passi B no canvia l'expectativa de A i viceversa, que passi A no canvia l'expectativa de B. Si A i B són independents, llavors:

$$P(B|A) = P(B|\neg A) = P(B)$$

- La idea d'esdeveniments independents està lligada a la de la *informació* que un aporta sobre l'altre: A i B són independents quan la probabilitat d'A és la mateixa, indiferentment del que passi amb B.

Independència i Prob. Condicionada. Exemple

Per estudiar l'eficiència en un aeroport, una primera aproximació ens porta a estudiar les probabilitats d'haver-se d'esperar a l'hora de facturar i a l'hora d'embarcar. Considerem el cas d'un aeroport on s'ha comprovat que per a un viatger que arriba, la probabilitat de trobar cua a facturació és 0.4; i de trobar cua a l'embarcament és 0.6 si va trobar cua a facturació, i 0.2 si no en va trobar.

Calculeu les següents probabilitats:

a) de trobar cua a la facturació i a l'embarcament

b) de trobar cua a l'embarcament

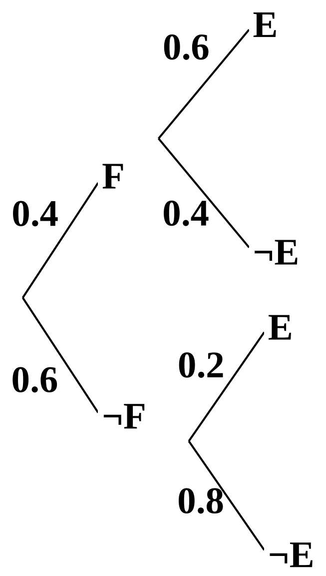
c) de trobar cua a l'embarcament si s'ha trobat cua a la facturació

d) d'haver trobat cua a la facturació si no ha trobat cua a l'embarcament

Independència i Prob. Condicionada. Exemple

F és "trobar cua a facturació"
¬F és "no trobar cua a facturació"

E és "trobar cua a embarcament"
¬E és "no trobar cua a embarcament"



a) $P(F \cap E) = 0.4 \cdot 0.6 = 0.24$

b) $P(E) = P(E \cap F) + P(E \cap \neg F) = 0.4 \cdot 0.6 + 0.6 \cdot 0.2 = 0.36$

c) $P(E|F) = 0.6$

d) $P(F|\neg E) = P(F \cap \neg E)/P(\neg E) = 0.16/0.64 = 0.25$

Independència i Prob. Condicionada. Exercici

Un client es vol connectar amb un servidor remot mitjançant una xarxa de comunicacions. El procés consisteix en realitzar ***n intents de connexió*** *a la xarxa en un període determinat. Tenim èxit si, en algun intent, hem trobat un camí per la xarxa fins al servidor i si el servidor està en marxa.*

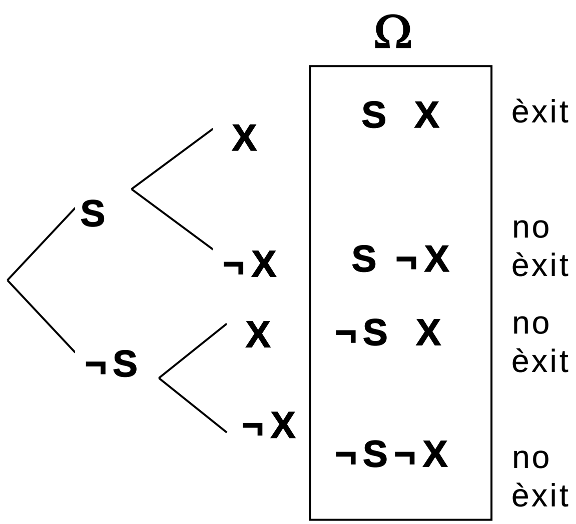
En primer lloc representarem l'experiència pels casos de 1 i 2 intents:

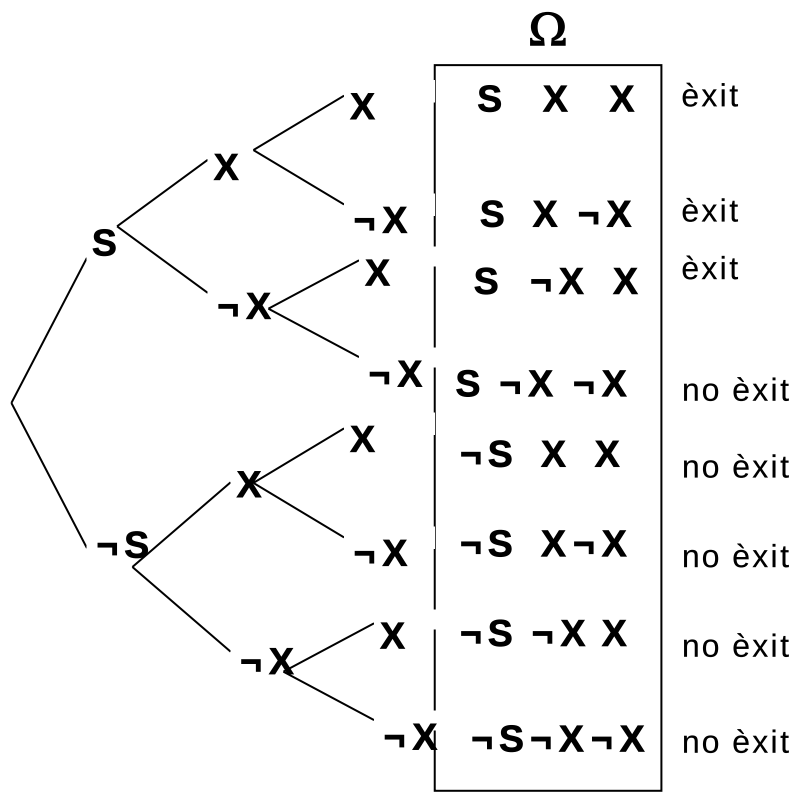
S = "el servidor està en marxa" (respon)
¬S = "no en marxa" (no respon)
X = "la petició del client ha trobat camí per la xarxa"
¬X = "no camí a la xarxa"

1) Representeu l'arbre pels casos 1 i 2 intents

Independència i Prob. Condicionada. Exercici

Arbres de probabilitat





Source: (repàs de combinatòria: Lecture4 a Instructor Resources a http://www.janehorgan.com/)

Independència i Prob. Condicionada. Exercici

Si en aquest exemple podem suposar:

- en n intents, el servidor no canvia d'estat
- els estats del servidor i de la xarxa són independents
- els n intents de connexió són independents uns d'altres
- el servidor falla, a l'atzar, 1 de cada 10 vegades: $P(\neg S)=p1=1/10$
- i la xarxa una de cada 5 vegades $P(\neg X)=p2=1/5$

Podem calcular la probabilitat d'èxit, és a dir, de contactar i poder treballar amb el servidor si només es realitza un intent, fent:

$$P(\text{el servidor funciona} \cap \text{la xarxa funciona}) = P(S \cap X) =$$

$$=$$

Independents

Independència i Prob. Condicionada. Exercici

Si definim T_i com l'esdeveniment connectar en algun dels i intents, llavors podem calcular la probabilitat d'èxit si:

– només es realitza un intent

$$P(T_1) = P(S \cap X) = P(S) \cdot P(X) = (1 - p1) \cdot (1 - p2)$$

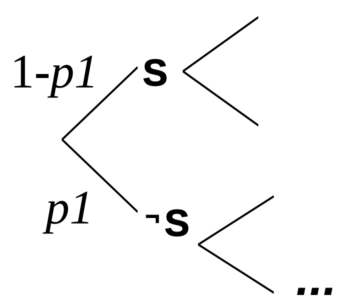
– es realitzen dos intents (1–Prob("no èxit en 2 intents"))

$$P(T_2) = 1 - P(\neg T_2) = 1 - (p1 + (1 - p1) \cdot (p2)^2)$$

$$on \quad P(\neg T_2) = P(\overline{S}) + P(S \cap \overline{X} \cap \overline{X}) = p1 + (1 - p1) \cdot (p2)^2$$

– es realitzen n intents (1–Prob("no èxit en n intents"))

$$P(T_n) = 1 - P(\neg T_n) = 1 - (p1 + (1 - p1) \cdot (p2)^n) \to_{n \to \infty} 1 - p1$$

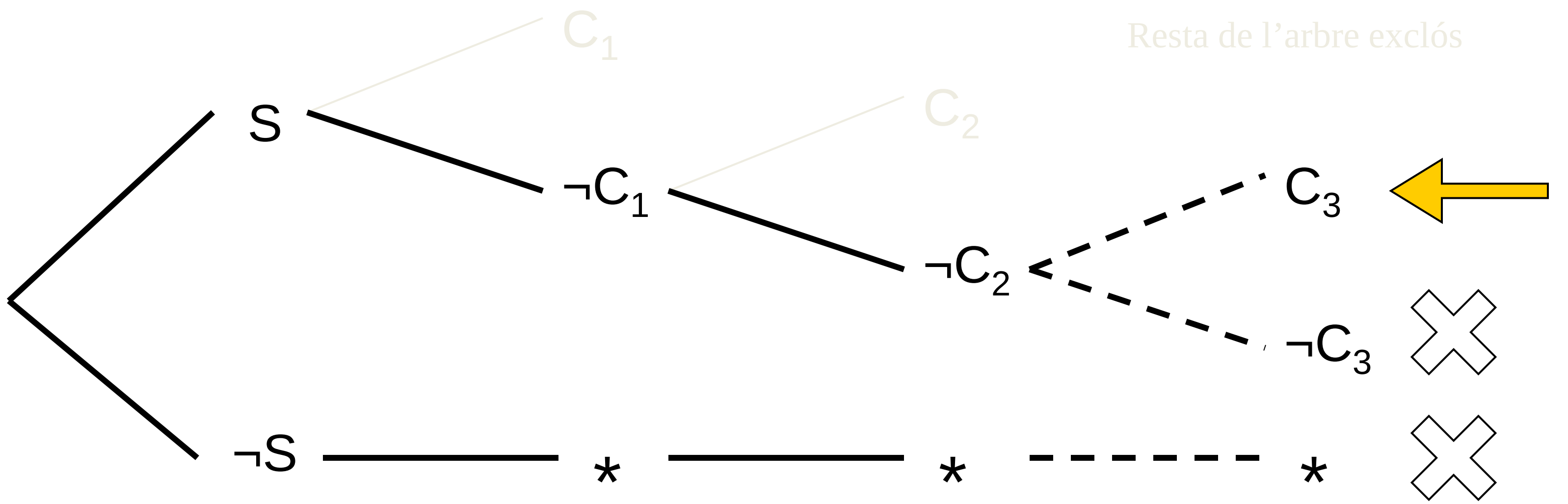


Si *n* augmenta, la probabilitat d'èxit tendeix a ***1-p1***

(tendeix a P(S), la probabilitat que el servidor funcioni, ja que si *n* és gran la probabilitat que en algun intent la xarxa no falli tendeix a 1)

Independència i Prob. Condicionada. Exercici

Ara suposem que s'han realitzat dos intents de connexió sense èxit (no sabem si per causa de la xarxa o del servidor). ¿Quina és la probabilitat de connectar en un tercer intent?



Sabent que teníem definit T_i com l'esdeveniment "connectar en algun dels i intents", llavors $\neg T_2$ serà "no connectar en 2 intents". Així, cal calcular:

$$P(S \cap X_3 \mid \neg T_2) = \frac{P(S \cap X_3 \cap \neg T_2)}{P(\neg T_2)}$$

Independència i Prob. Condicionada. Exercici

En primer lloc calculem la probabilitat del numerador. Com que:

$$S\cap X_3\cap\neg T_2 = S\cap X_3\cap [\neg S\cup(S\cap\neg X_1\cap\neg X_2)] = S\cap\neg X_1\cap\neg X_2 \cap X_3$$

llavors la seva probabilitat és:

$$P(S\cap X_3\cap\neg T_2) = 0.9\cdot 0.2\cdot 0.2\cdot 0.8 = 0.0288$$

Fallar en dos intents i tenir èxit en el tercer (amb el servidor en marxa) és un succés amb poques "possibilitats" (menys del 3%). I així, la probabilitat que busquem es compensa quan es compara amb el succés que condiciona (¬T2 que tampoc és molt freqüent):

$$P(S\cap X_3\,|\,\neg T_2) = \frac{P(S\cap X_3\cap\neg T_2)}{P(\neg T_2)} = \frac{0.0288}{0.136} = 0.2118$$

Aquesta probabilitat també la podem comparar amb la de connectar en un tercer intent sense tenir cap informació prèvia (probabilitat "bruta" de connectar en el tercer intent és $P(S \cap X_3)=0.72$). Per tant, no connectar en els dos intents previs és una mala senyal: ¡les possibilitats baixen del 72% a un 21%!

Probabilitat a posteriori. Fórmula de Bayes

A partir de la definició de probabilitat condicionada:

$$\mathrm{P(A\,|\,B)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(B)}}$$

i de la probabilitat de la intersecció aïllada de la condicionada complementària:

$$\mathrm{P(B\,|\,A)} = \frac{\mathrm{P(B \cap A)}}{\mathrm{P(A)}} \rightarrow P(B \cap A) = P(A \cap B) = P(B\,|\,A) \cdot P(A)$$

es dedueix la fórmula de Bayes:

$$\mathrm{P(A\,|\,B)} = \frac{\mathrm{P(B\,|\,A) \cdot P(A)}}{\mathrm{P(B)}}$$

que, coneixent P(A) i P(B), permet passar de P(A|B) a P(B|A) i viceversa (usualment en l'enunciat del cas, les probabilitats condicionades en un sentit són conegudes i interessa calcular les condicionades complementàries). [Exemple: si conec la probabilitat de pluja si està ennuvolat i vull conèixer la probabilitat d'ennuvolat si ha plogut]

Prob. a posteriori. Probabilitats totals i Bayes

Podem calcular la probabilitat d'un succés B_k a partir de les probabilitats de les interseccions d'aquest amb una partició $A_1, A_2, \ldots A_j$ de W:

$$P(B_k) = P(B_k \cap A_1) + P(B_k \cap A_2) + \ldots + P(B_k \cap A_j) =$$

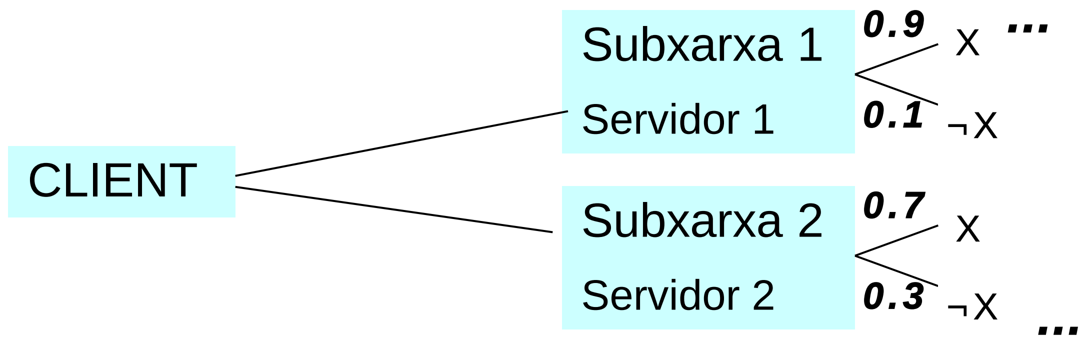
$$= P(B_k|A_1) \cdot P(A_1) + P(B_k|A_2) \cdot P(A_2) + \ldots + P(B_k|A_j) \cdot P(A_j)$$

Llei de probabilitats totals (LPT). S'aplica quan disposem d'una partició, i la probabilidad del succés d'interès és senzilla d'obtenir si està condicionat per un element qualsevol de la partició.

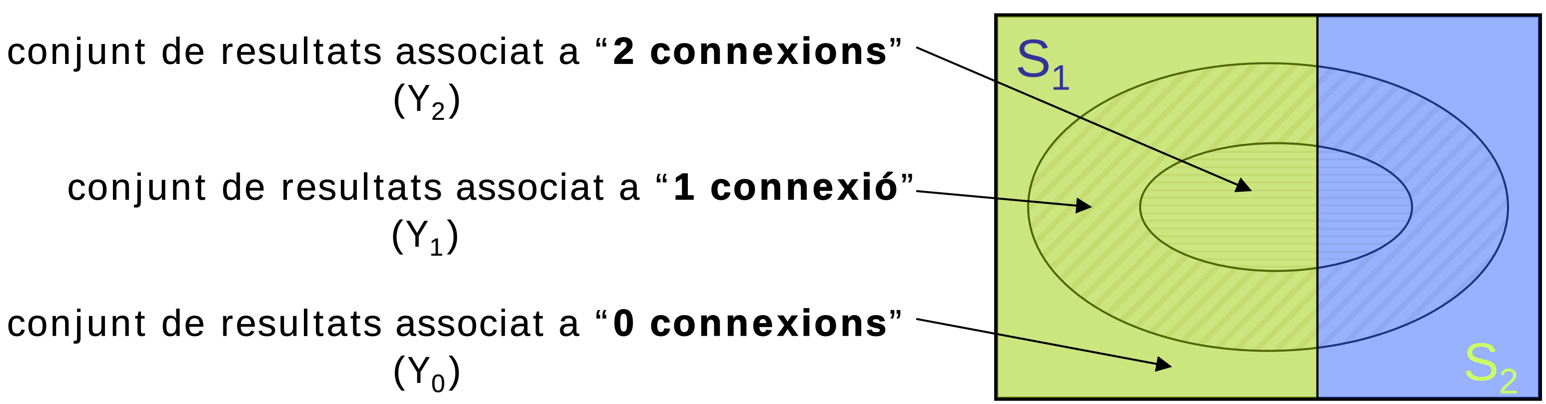
Combinant la fórmula de Bayes amb la llei de probabilitats totals (i una partició {Ai} adequada) s'obté el teorema de Bayes:

$$P(A_i | B_k) = \frac{P(B_k | A_i) \cdot P(A_i)}{\sum_j P(B_k | A_j) \cdot P(A_j)}$$

Probabilitat a posteriori (LPT). Exemple



Ara considerem que tenim dues subxarxes amb dos servidors i en triem una o altra a l'atzar (50%). I llavors fem els intents sempre sobre la mateixa xarxa (en la primera falla la connexió 1 de cada 10 cops, i en la segona 3 de cada 10). Per a n=2 (2 intents de connexió). ¿Com calcular la probabilitat d'obtenir Y = 0, 1, 2 connexions?



Probabilitat a posteriori (LPT). Exemple

En aquest cas podem calcular fàcilment les probabilitats de Y_0, Y_1 o Y_2 condicionades pel servidor:

$P(Y_0 \mid S_1) = 0.1 \cdot 0.1 = 0.01$
$P(Y_0 \mid S_2) = 0.3 \cdot 0.3 = 0.09$
$P(Y_1 \mid S_1) = 0.1 \cdot 0.9 + 0.9 \cdot 0.1 = 0.18$
$P(Y_1 \mid S_2) = 0.3 \cdot 0.7 + 0.7 \cdot 0.3 = 0.42$
$P(Y_2 \mid S_1) = 0.9 \cdot 0.9 = 0.81$
$P(Y_2 \mid S_2) = 0.7 \cdot 0.7 = 0.49$

Quines probabilitats han de sumar 1?

I com que Y_0, Y_1 o Y_2 poden ser expressats com a unió de conjunts disjunts (ja que $\{S_1, S_2\}$ és una *partició*):

LPT

$$Y_i = (Y_i \cap S_1) \cup (Y_i \cap S_2); \quad \Rightarrow P(Y_i) = P(Y_i \cap S_1) + P(Y_i \cap S_2) = P(Y_i \mid S_1) \cdot P(S_1) + P(Y_i \mid S_2) \cdot P(S_2)$$

Calcula:

$P(Y_0) \rightarrow$ zero connexions, 1 de cada 20 vegades

$P(Y_1) \rightarrow$ una connexió, 3 de cada 10 vegades

$P(Y_2) \rightarrow$ dos connexions, 13 de cada 20 vegades

Probabilitat a posteriori. Exercici

I ara suposant que s'han aconseguit k connexions, **amb quina probabilitat hem estat atesos pel servidor *i*?**

Agafant el nombre de connexions com una partició i els dos servidors com una altra partició podem aplicar:

$$P(A_i | B_k) = \frac{P(B_k | A_i) \cdot P(A_i)}{\sum_j P(B_k | A_j) \cdot P(A_j)}$$

Calculeu les probabilitats a posteriori

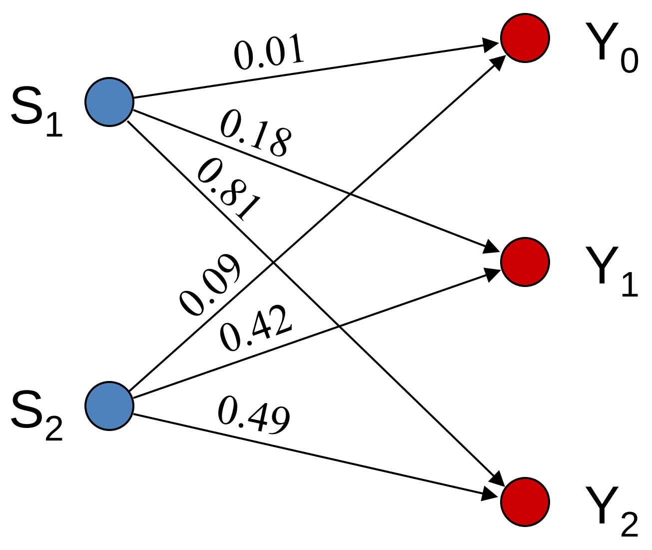
A priori	*i=1*	*i=2*	**P(Yk)**		A posteriori	*i=1*	*i=2*
$P(Y_0 \mid Si)$	0.01	0.09	0.05		**$P(Si \mid Y_0)$**		
$P(Y_1 \mid Si)$	0.18	0.42	0.30		**$P(Si \mid Y_1)$**		
$P(Y_2 \mid Si)$	0.81	0.49	0.65		**$P(Si \mid Y_2)$**		
P(Si)	0.5	0.5					

*Si s'han aconseguit dues connexions, creiem que hem utilitzat el primer servidor amb probabilitat **.....***

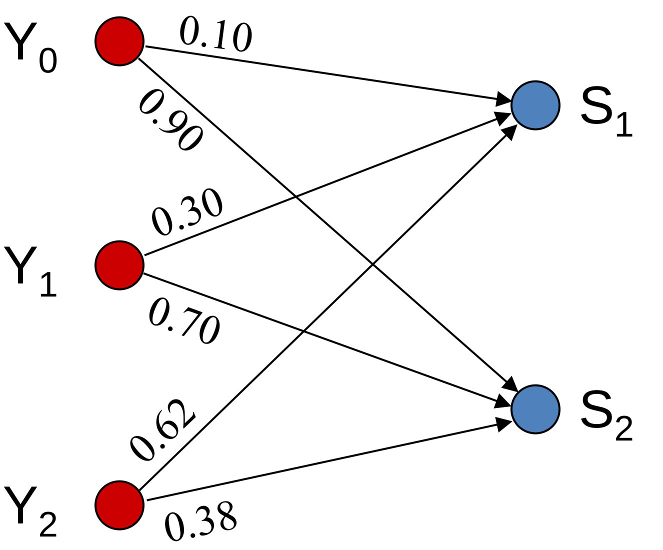
Probabilitat a posteriori. Teorema de Bayes

El teorema de Bayes transforma unes probabilitats condicionades en unes altres

$P(Y_k \mid S_i)$



$P(S_i \mid Y_k)$

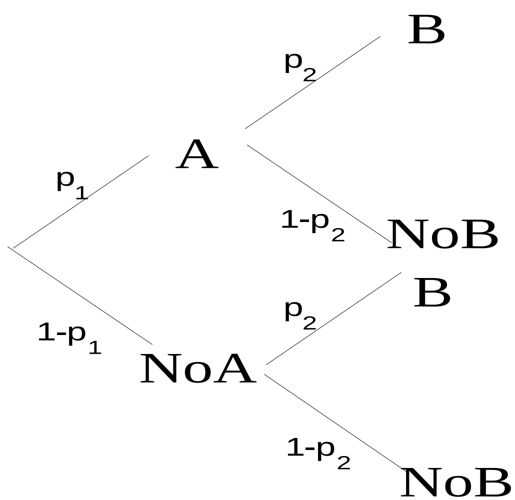


Coneixent el servidor utilitzat, calculem la probabilitat d'obtenir un número determinat de connexions.

Sabent el número de connexions aconseguides, calculem la probabilitat d'haver utilitzat determinat servidor.

PROBABILITAT condicionada, conjunta i marginal

Amb independència



Condicionada

Marginal

Conjunta

	B	¬B	
A	A∩B	A∩¬B	p_1
¬A	¬A∩B	¬A∩¬B	$1-p_1$
	p_2	$1-p_2$	

Les condicionades són iguals a les marginals:

$P(B \mid A) = P(B \mid \neg A) = P(B) = p_2$

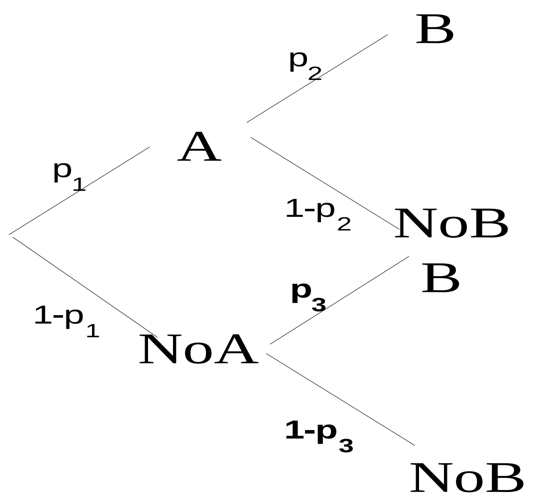
$P(\neg B \mid A) = P(\neg B \mid \neg A) = P(\neg B) = 1-p_2$

La conjunta (intersecció) és igual al producte de les marginals:

$P(A \cap B) = P(A) \cdot P(B) = p_1 \cdot p_2$

PROBABILITAT condicionada, conjunta i marginal

Sense independència



	B	¬B	
A	A∩B	A∩¬B	p_1
¬A	¬A∩B	¬A∩¬B	$1\text{-}p_1$
	p2	1-p2	

Les condicionades NO són iguals a les marginals:

$P(B \mid A) = p_2 \neq P(B \mid \neg A) = p_3 \neq P(B)$

$P(\neg B \mid A) = 1 - p_2 \neq P(\neg B \mid \neg A) = 1 - p_3 \neq P(\neg B)$

La conjunta (intersecció) NO és igual al producte de les marginals:

$P(A \cap B) \neq P(A) \cdot P(B)$

PROB cond., conj. i marg. Exemple amb probs.

Si les probabilitats conjuntes són:

	A	No A	
B	0.06	0.54	0.60
No B	0.04	0.36	0.40
	0.10	0.90	

les condicionades per columnes són:

	A	No A
B	0,6	0,6
No B	0,4	0,4
	1	1

SI independència

Si les probabilitats conjuntes són:

	A	No A	
B	0.05	0.55	0.60
No B	0.05	0.35	0.40
	0.10	0.90	

les condicionades per columnes són:

	A	No A
B	0,5	0,61
No B	0,5	0,39
	1	1

NO independència

PROB cond., conj., marg. Exemple de freqüencies

2 mostres (100 obs.), gènere i edat en 3 categories

	1	2	3	
masc.	6	24	30	60
fem.	4	16	20	40
	10	40	50	

	1	2	3
masc.	0,6	0,6	0,6
fem.	0,4	0,4	0,4
	1	1	1

SI independència

	1	2	3	
masc.	5	30	25	60
fem.	5	10	25	40
	10	40	50	

	1	2	3
masc.	0,5	0,75	0,5
fem.	0,5	0,25	0,5
	1	1	1

NO independència

Aplicacions de Bayes

- Hardware Fault Diagnosis
  - Extreient d'una Base de Dades les probabilitats de certs *problemes* ($P(A_i)$), i coneixent també la probabilitat de *fallada* segons el problemes ($P(F|A_i)$), es pot calcular quin problema és més probable quan es dóna una fallada ($max(P(A_i|F)$)
- Machine Learning (cas d'algoritmes de classificació d'aprenentatge supervisat)
  - Coneixent a priori les probabilitats de pertanyer a unes certes *classes* ($P(A_i)$), i coneixent també la probabilitat de certa *característica* segons la classe ($P(F|A_i)$), es pot calcular quina classe és més probable quan es dóna certa característica ($max(P(A_i|F)$)
- Reliability. System reliability
  - Series System (la probabilitat de funcionar el sistema és el producte de les probabilitats de funcionar dels components)
  - Parallel Systems (la probabilitat de funcionar el sistema és 1 menys la probabilitat de que fallin tots - que és el producte de les probabilitats de fallar dels components)

Font: Lectures 7 y 8 a Instructor Resources a http://www.janehorgan.com/

Aplicació de Bayes. Exemple

Alguns processadors utilitzen un tipus especial de caches, on la memòria està distribuïda en quatre bancs. Aquests processadors són capaços de fer dos accessos simultàniament, mentre no tinguin que accedir al mateix banc. Quina és la probabilitat de conflicte?

Una situació simple assumeix que no hi ha cap relació entre el banc accedit per un accés i l'altre.

	Accés #2			
Accés #1	1/16	1/16	1/16	1/16
	1/16	1/16	1/16	1/16
	1/16	1/16	1/16	1/16
	1/16	1/16	1/16	1/16

Probabilitat conflicte = 1/4

Però potser el comportament del processador és més complex. Per exemple: certa propensió a utilitzar pel segon accés el següent banc:

	Accés #2			
Accés #1	0.05	**0.10**	0.05	0.05
	0.05	0.05	**0.10**	0.05
	0.05	0.05	0.05	**0.10**
	0.10	0.05	0.05	0.05

Probabilitat conflicte = 0.20

Aplicació de Bayes. Exercici

Per estudiar l'eficiència en un aeroport, una primera aproximació ens porta a estudiar les probabilitats d'haver-se d'esperar a l'hora de facturar i al control de passaports (CP). Considerem el cas d'un aeroport on s'ha comprovat que per a un viatger que arriba, la probabilitat de trobar cua a facturació és 0.64. Posteriorment s'ha comprovat que les probabilitats d'esperar o no per el CP depenen del fet d'haver esperat al moment de facturar, com mostra la següent taula:

	si per facturar s'ha hagut d'esperar	si per facturar no s'ha hagut d'esperar
no espera	0.187	0.238
espera el primer a la cua	0.687	0.461
espera perque ja hi ha una cua	0.126	0.301

1. la probabilitat de no esperar a facturació ni a CP.
2. la probabilitat de ser atès immediatament al CP (és a dir, no hi ha ningú passant el control).
3. la probabilitat d'haver d'esperar a facturació o a CP.
4. la probabilitat d'haver d'esperar a un dels llocs (però només a un).
5. Si un viatger arriba al CP i ha d'esperar perquè hi ha una persona passant el control, i cap més: quina és la probabilitat de haver esperat a facturació?
6. Els passatgers A i B arriben a l'aeroport en dos moments independents per agafar els seus vols. Trobeu la probabilitat que els dos hagin de posar-s'hi a la cua al CP.

Variable aleatòria

Bloc 2 – Probabilitat i Estadística

Març 2015

Índex

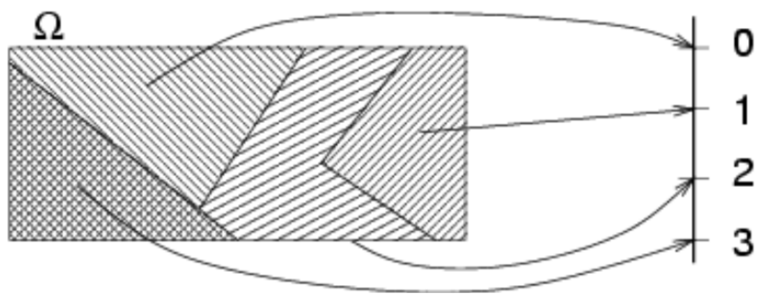
Variable aleatòria. Definició

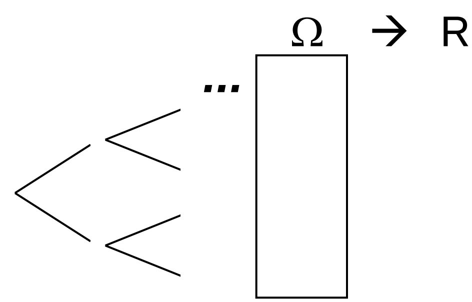
- La majoria d' experiències aleatòries porten a resultats interpretables com un número. Ens interessa l' estudi de l' experiment des del **punt de vista numèric**.
- Una **variable aleatòria** X és una aplicació entre el conjunt Ω i la recta real:

$$X: \Omega \rightarrow \mathbf{R}$$

Notació: les anomenarem amb un símbol tal com *X, Y, Z,* ...(lletra llatina majúscula)

- Una variable indueix una partició de Ω amb els valors que adopta:





- Les probabilitats definides en Ω es transfereixen als valors de la VA X, definint unes **funcions de probabilitat** (o de densitat) i de **distribució de probabilitat** :

$$\Omega \xrightarrow{X} \mathbf{R} \xrightarrow{\text{prob.}} [0,1]$$

Variable aleatòria. Tipus

Distingim dos tipus de VA:

- Si el conjunt de valors que poden agafar és enumerable (és a dir, un interval de valors enters (0..n), o el conjunt dels naturals N), la VA és **discreta (VAD)**

  Per exemple, en l'experiència de llençar una moneda 3 vegades:

  – VAD X = "cara o creu en l'últim llançament" (possibles valors: 0,1; probabilitats: ½, ½)

  – VAD Y = "número de cares" (possibles valors: 0,1,2,3; probabilitats: ?)

  – VAD Z = "número de caigudes del sistema" (possibles valors: 0,1,2,3,4,5,6,…; probabilitats:?)

- Si agafa valors d'un conjunt no discret (és a dir, normalment la recta real o un segment d'ella) la VA és **contínua (VAC)**

  – VAC X = "temps entre caigudes del sistema" (possibles valors: R+)

  – En general, les VAC són mesures físiques de temps, longituds,....

Variable aleatòria. Funció de prob. i distr. en VAD

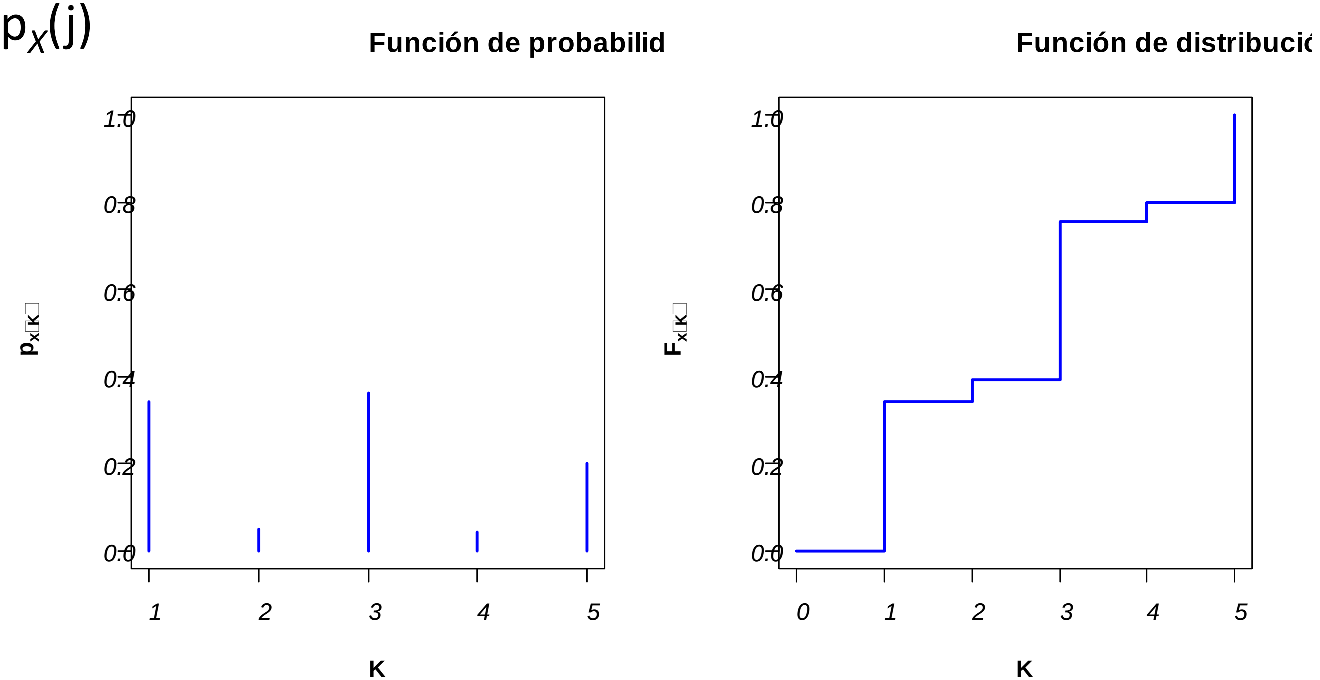
Les funcions de **probabilitat i distribució** es defineixen segons si són VAC o VAD

- La **funció de probabilitat** (p_X) en una VA DISCRETA (VAD) defineix la probabilitat puntual de cada un dels possibles valors k

$p_X(k) = P(X = k)$ (complint $\Sigma_{\forall k}\, p_X(k) = 1$)

- La **funció de distribució** (F_X) de probabilitat en una VAD defineix la probabilitat acumulada, és a dir:

$F_X(k) = P(X \leq k) = \Sigma_{j \leq k}\, p_X(j)$



Variable aleatòria. Funció de prob. i distr. en VAC

- La **funció de densitat** de probabilitat (**f_X**) d'una VA CONTINUA és la funció que recobreix l'àrea on està definida la variable complint:

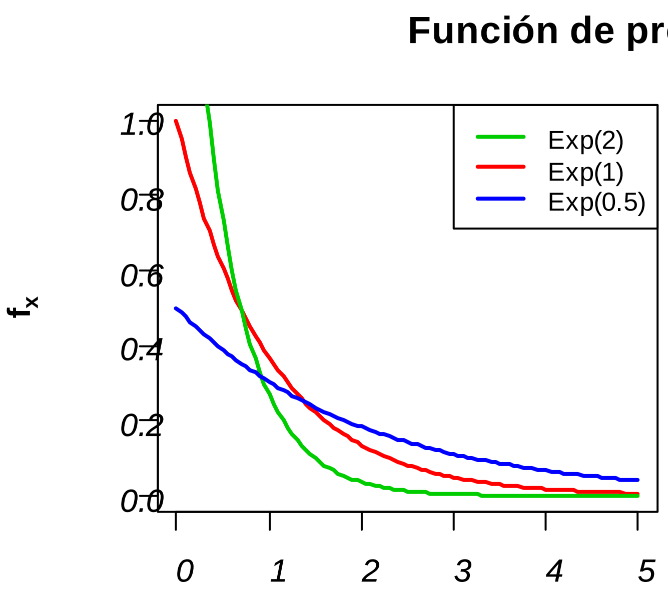
$$\int_{-\infty}^{+\infty} f_X(x)dx = 1$$

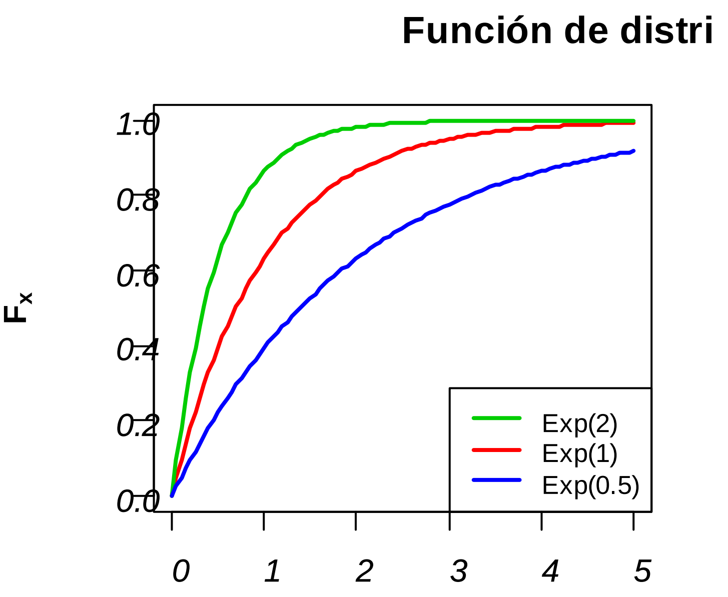
Observem que $f_X(k)$ és el valor de la funció en k, però no és una probabilitat puntual, $f_X(k) \neq P(X=k)$, i $P(X=k)$ val 0

- La **funció de distribució** de probabilitat (**F_X**) d'una VA CONTINUA defineix la probabilitat acumulada, és a dir:

$$F_X(k) = P(X \leq k) = \int_{-\infty}^{k} f_X(x)dx$$

Observem que $f_X(x) = \dfrac{dF_X(x)}{dx}$





Variable aleatòria. Funció de prob. i distr. en VAC

És a dir, en el cas VA CONTINUA, qualsevol **funció positiva**, $f_X(x) \geq 0$, que compleix:

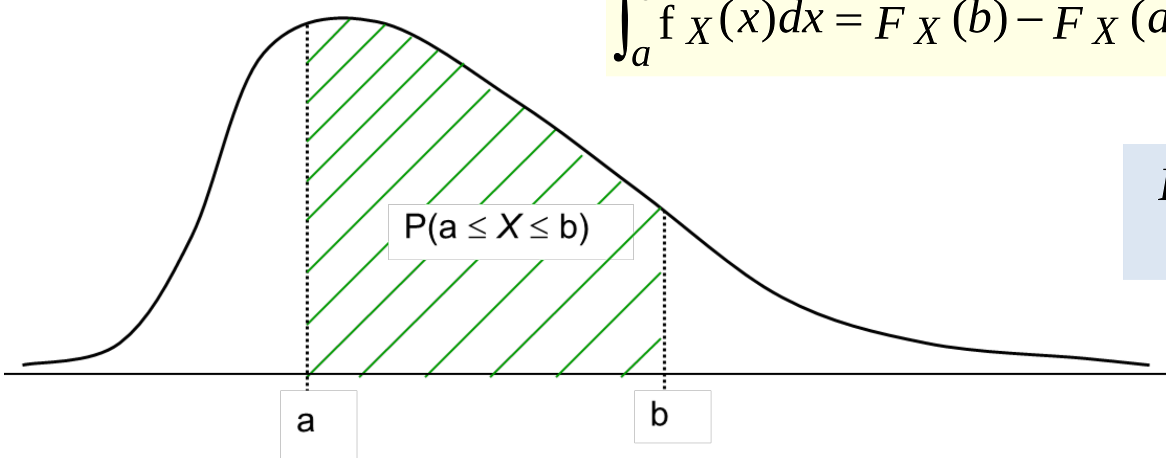
$$\int_{-\infty}^{+\infty} f_X(x)dx = 1$$

és una funció de densitat vàlida, és a dir, caracteritza la variable. La funció de distribució s'obté amb:

$$F_X(u) = \int_{-\infty}^{u} f_X(x)dx$$

Tota l'àrea sota la funció de densitat és sempre igual a 1. En particular, l'àrea que hi ha sota la funció de densitat entre els límits *a* per l'esquerra i *b* per la dreta és:

$$\int_a^b f_X(x)dx = F_X(b) - F_X(a)$$



L'àrea sota f_X equival a una probabilitat!

Variable aleatòria. Càlcul de probabilitats

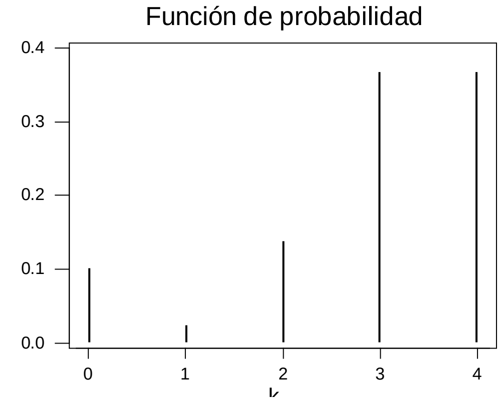
- **VAD**
  - $P(X = k) = p_X(k)$
  - $P(X \leq k) = F_X(k) = S_{j}\leq k\ p_X(j)$
  - $P(X<k) = P(X \leq k\text{-}1) = F_X(k\text{-}1)$
  - $P(a < X \leq b) = P(X \leq b) – P(X \leq a) = F_X(b) − F_X(a)$
  - $P(a \leq X \leq b) = P(X \leq b) – P(X < a) = F_X(b) − F_X(a\text{-}1)$
  - $P(a < X < b) = P(X < b) – P(X \leq a) = F_X(b\text{-}1) – F_X(a)$
- **VAC**
  - $P(X = k) = 0 \quad (\neq f_X(k))$
  - $P(X \leq k) = F_X(k)$
  - $P(X<k) = P(X \leq k) = F_X(k)$
  - $P(a \leq X \leq b) = P(a < X \leq b) = P(a < X < b) = F_X(b) − F_X(a)$

Variable aleatòria. Exemple de VAD

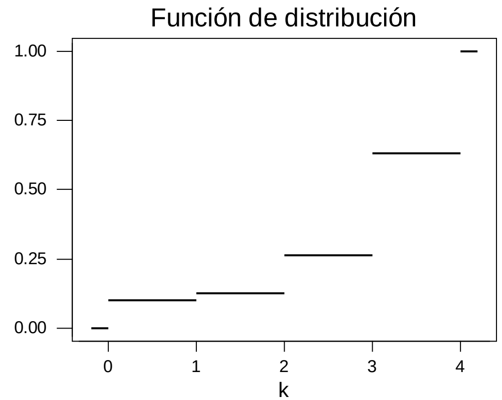
Reprenem l'exemple del servidor i la xarxa. Ara amb 4 intents de connexió, probabilitat de funcionar el servidor $p_1=0.9$ i probabilitat de funcionar la xarxa $p_2= 0.8$. Definim la VAD X = "nº de connexions aconseguides"

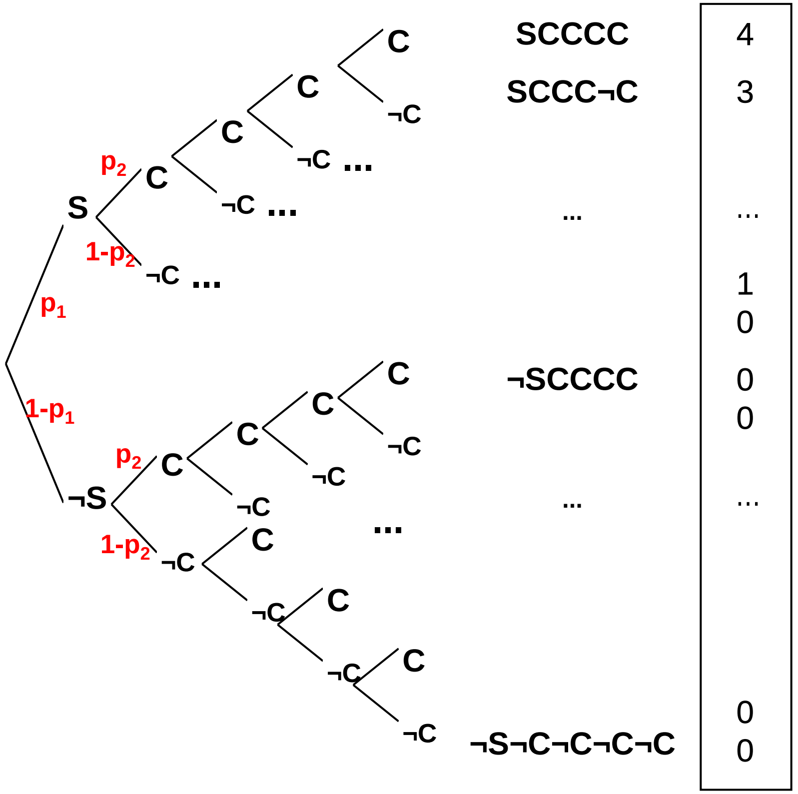
$$P(X=0)= 1- p_1+ p_1\cdot(1- p_2)^4 \rightarrow \qquad P(X=k)= p_1\cdot p_2^k\cdot(1- p_2)^{4-k}\binom{4}{k} \qquad k>0$$

k	$P_X(k)$
0	0.10144
1	0.02304
2	0.13824
3	0.36864
4	0.36864



k	$F_X(k)$
$(-\infty,0)$	0.00000
[0,1)	0.10144
[1,2)	0.12448
[2,3)	0.26272
[3,4)	0.63136
$[4,+\infty)$	1.00000



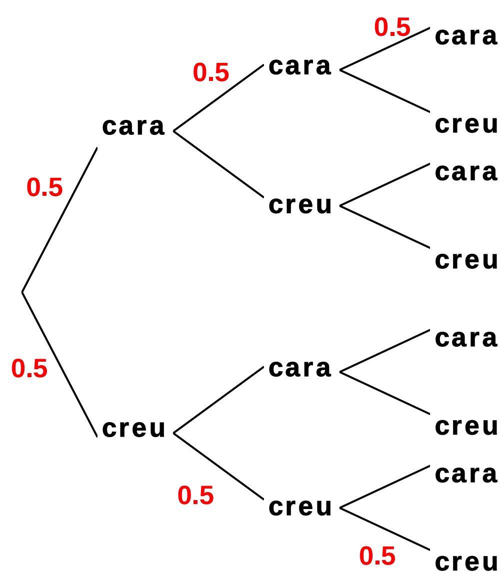


Variable aleatòria. Exemple de VAD

Estudiarem l'experiència aleatòria de **llençar una moneda equilibrada <u>tres</u> vegades.** Calcularem:

- P(A) sent A = "obtenir **2 cares**"
- P(B) sent B = "obtenir **almenys 2 cares**"

[al bloc 1 s'han calculat les probabilitats a partir de l'arbre i dels successos; ara ho calcularem definint la variable aleatòria X = "número de cares"]

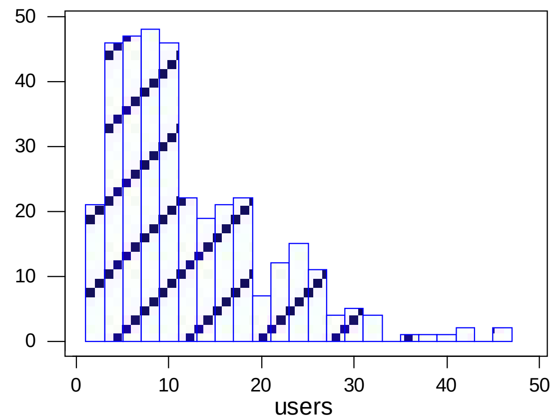


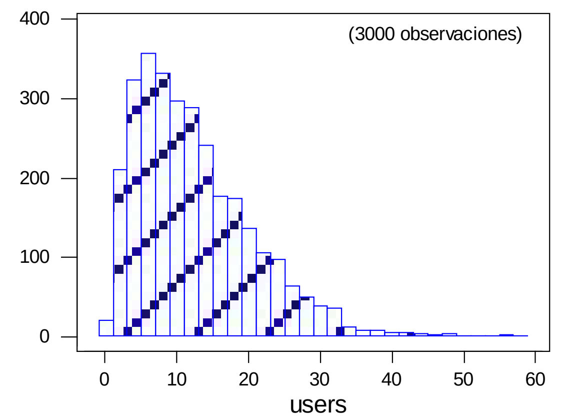
(Cara,Cara,Cara) = w1	P(w1) = 1/8	X = 3
(Cara,Cara,Creu) = w2	P(w2) = 1/8	X = 2
(Cara,Creu,Cara) = w3	P(w3) = 1/8	X = 2
(Cara,Creu,Creu) = w4	P(w4) = 1/8	X = 1
(Creu,Cara,Cara) = w5	P(w5) = 1/8	X = 2
(Creu,Cara,Creu) = w6	P(w6) = 1/8	X = 1
(Creu,Creu,Cara) = w7	P(w7) = 1/8	X = 1
(Creu,Creu,Creu) = w8	P(w8) = 1/8	X = 0

Variable aleatòria. Exemple de VAD

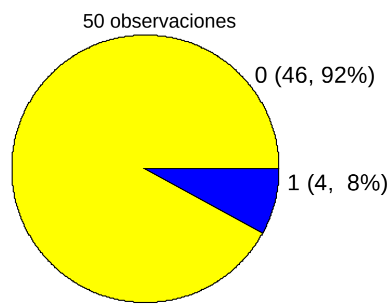
La repetició d'una experiència (en condicions *estables*) posa de manifest la **variabilitat dels resultats**. No es poden esperar sempre els mateixos resultats, l'aparença de la distribució dels valors serà diferent, però amb un número suficientment gran d'observacions serà *semblant* a la situació "teòrica".

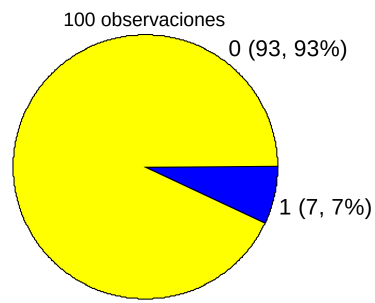
Per exemple, si un tècnic de sistemes recull informació del número d'usuaris, tindrà una aproximació a la VA teòrica del número d'usuaris

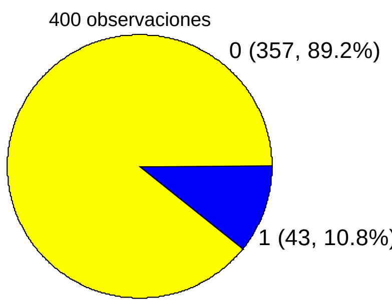




I si recull informació sobre si la paginació és normal (0) o elevada (1)

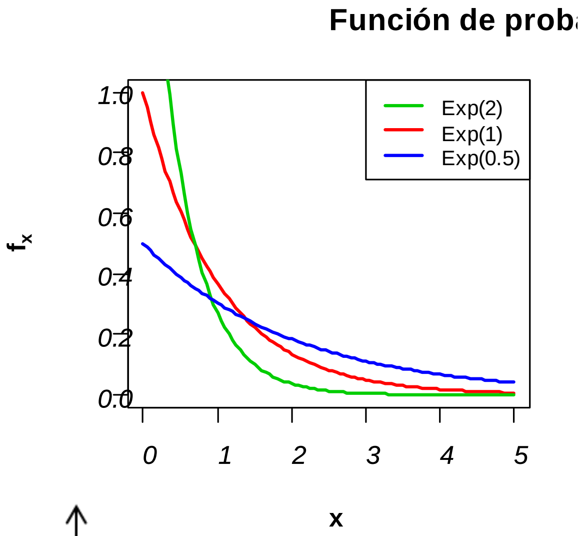




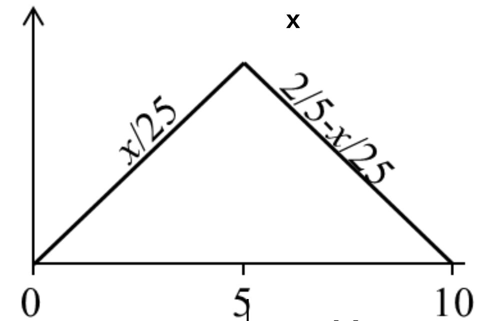


Variable aleatòria. Exemple de VAC

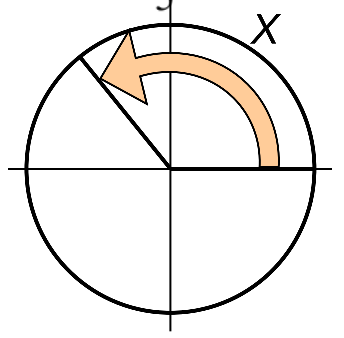
Exemple 1. La “vida útil d’un transistor en anys” segueix una distribució que decreix exponencialment.



Exemple 2. “L’esforç requerit per desenvolupar un projecte” es pot mesurar en homes/mes, per exemple segons la figura.



Exemple 3. “L'angle X senyalat per la agulla d’una ruleta al parar-se”, $0 \leq X \leq 2\pi$.



$F_X(k) = P(X \leq k) = k/(2\pi)$.

Probabilitats acumulades. Quantils

- Sigui X una variable aleatòria, i α un valor real $(0 \leq \alpha \leq 1)$ diem que x_α és el **quantil α** de X si es compleix: $F_X(x_\alpha) = \alpha$
- Calcular un quantil és el problema invers al càlcul de **probabilitats acumulades**. La funció inversa de la funció de distribució ens retorna x_α
- L'equivalent mostral dóna lloc als **quantils** utilitzats en estadística descriptiva. Expressats en percentatge dóna lloc als **percentils**.
- Exemples: el primer quartil (Q1) és el percentil 25, el tercer quartil (Q3) és el percentil 75, la mediana (Me) és el percentil 50.

Probabilitats acumulades. Exemple amb VAC

Exemple 2 de diapositiva 12:

Quants homes/mes són necessaris per al 90% dels projectes? Quin és el quantil 0.9 per a la variable "esforç per desenvolupar projecte"? Amb quin valor x fem que $F_X(x) = 0.9$?

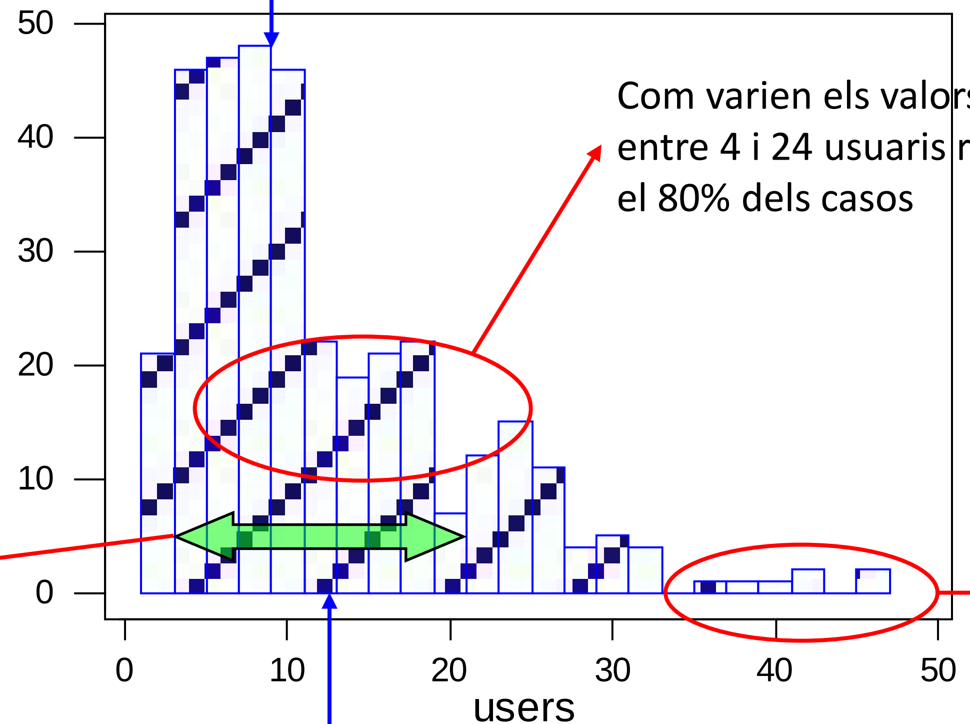
Pista: podem trobar analíticament F_X, i resoldre l'equació, o en aquest cas podem aprofitar que l'àrea sota la funció de densitat és un triangle...

Solució: 7.76 homes/mes.

Probabilitats acumulades. Exemple amb VAD

El 50% dels casos hi ha
com a molt 9 usuaris en
el sistema

Com varien els valors usuals:
entre 4 i 24 usuaris representen
el 80% dels casos



La desviació
típica és
d'uns 8.5
usuaris

Un 2% dels casos
correspon a més de
35 usuaris

La carga mitjana és
de 11.5 usuaris

Indicadors de V.A. Quins són?

- Indicadors en una mostra: Una mostra de valors expressa amb *n* observacions la variabilitat d'una experiència; si volem resumir aquestes dades utilitzarem la **mitjana mostral** i la **desviació estàndard mostral** S_x [tal com es veu a l'Estadística descriptiva]

- Indicadors en una variable aleatòria: Una v.a. *X* no representa una determinada mostra, i per això ha d'utilitzar la distribució de probabilitat dels valors que pot agafar:
  - com indicador central agafem el **valor esperat o esperança E(X)** **(μ_X)**
  - com mesura de dispersió la **variància V(X) (σ^2_X)** o la **desviació estàndard (σ_X)**

Indicadors de V.A. Com es calculen?

- **Esperança de X**

$$\boldsymbol{VAD} \rightarrow E(X) = \mu_X = \sum_{\forall k} (k \cdot p_X(k))$$

$$\boldsymbol{VAC} \rightarrow E(X) = \mu_X = \int_{-\infty}^{+\infty} x \cdot f_X(x) dx$$

- **Variança de X**

$$\boldsymbol{VAD} \rightarrow V(X) = \sigma_X^2 = \sum_{\forall k} \left[\left(k - E(X)\right)^2 \cdot p_X(k) \right] \rightarrow \sigma_X = \sqrt{\sigma_X^2} = \sqrt{V(X)}$$

$$\boldsymbol{VAC} \rightarrow V(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} \left(x - E(X)\right)^2 \cdot f_X(x) dx \rightarrow \sigma_X = \sqrt{\sigma_X^2} = \sqrt{V(X)}$$

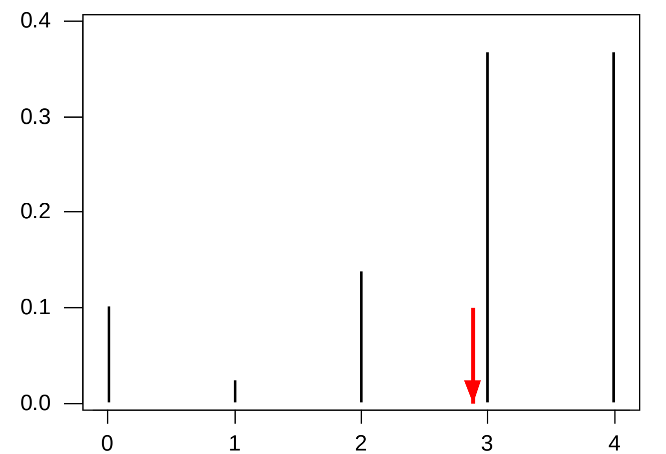
- **Relació entre Esperança i Variança (en VAD i VAC):**

$$V(X) = E[X - E(X)] = E(X^2) - E(X)^2$$

Indicadors de V.A. Exemple

En l'exemple del servidor i de la xarxa, analitzem l'esperança i variància de la variable *X* "número de connexions al servidor amb *n*=4"

k	$P_X(k)$	$k \cdot P_X(k)$	$(k-\mu)^2$	$(k-\mu)^2 \cdot P_X(k)$
0	0.10144	0	8.2944	0.84138
1	0.02304	0.02304	3.5344	0.08143
2	0.13824	0.27648	0.7744	0.10705
3	0.36864	1.10592	0.0144	0.00531
4	0.36864	1.47456	1.2544	0.46242
		$\mu=\sum k \cdot P_X(k)=2.88$		$\sigma^2=\sum(k-\mu)^2 \cdot P_X(k)=1.50$
				$\sigma=\sqrt{\sigma^2}=1.22$



L'esperança 2.88 indica que s'espera un promig de 2.88 connexions si l'experiència es repetís un gran número de vegades. No s'ha d'arrodonir a un enter. L'esperança s'associa al centre de gravetat.

La desviació 1.22 informa sobre la magnitud de la dispersió de la variable. Major desviació suposa major probabilitat de buscar un valor allunyat del valor esperat. Menor desviació, major concentració.

Indicadors de V.A. Propietats

Siguin *X* i *Y* dues variables aleatòries, i *a* i *b* dos escalars

Propietats de l'Esperança	**Propietats de la Variància**
$E(a+X) = a + E(X)$	$V(a+X) = V(X)$
$E(bX) = b \cdot E(X)$	$V(bX) = b^2 \cdot V(X)$
$E(a+bX) = a + b\,E(X)$	$V(a+bX) = b^2 \cdot V(X)$
$E(X+Y) = E(X) + E(Y)$	$V(X+Y) = V(X) + V(Y)$ si són ind.
$E(X \cdot Y) = E(X) \cdot E(Y)$ si són ind.	$V(X \cdot Y) = ?$

Indicadors de V.A. Exemple

En l'experiència aleatòria de **llençar una moneda equilibrada <u>tres</u> cops**, hem calculat probabilitats definint la variable aleatòria "número de cares".

X	$P_X(X)$
0	1/8
1	3/8
2	3/8
3	1/8

Calcula la esperança i la variància

$E(X) = 0·1/8+1·3/8+2·3/8+3·1/8 = 1.5$

$V(X) = (0 – 1.5)^2·1/8 + (1 – 1.5)^2·3/8 + (2 – 1.5)^2·3/8 + (3 – 1.5)^2·1/8 = 0.75$

Nota: NO confondre amb l'*<u>indicador de tendència central d'una mostra</u>:* **mitjana** d'unes dades recollides al realitzar repetidament l'experiència aleatòria. [Per exemple, si recollim 8 realitzacions en 50 voluntaris — cadascú repeteix 8 vegades, els tres llençaments i apunta el promig del nombre de cares en les 8 experiències. Una possiblitat és: dos voluntaris han tret de mitjana 0.75 cares; un ha tret 2.125 cares, i la resta han quedat entremig d'aquests valors, sent el cas més repetit, el treure 1.375 cares de mitjana]

Conclusió : La mitjana mostral pot variar; l'esperança no.

Indicadors de V.A. Exercici

Considerem el conjunt de tots els paquets de 3 bits que es poden enviar per una línia de comunicació:

Ω = {000, 001, 010, 011, 100, 101, 110, 111}

Suposem que totes les seqüències són equiprobables.

Es defineixen dues variables aleatòries:

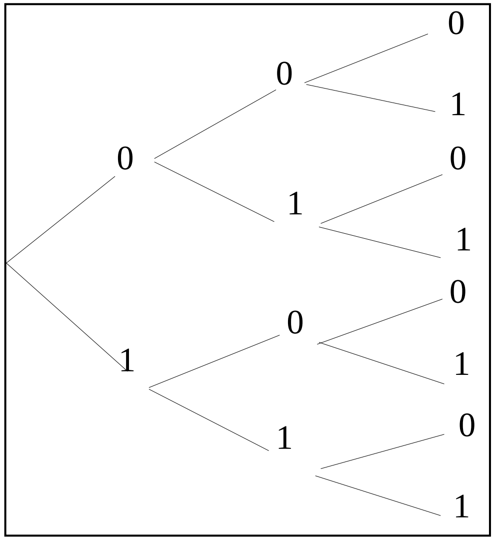
X: "suma dels 3 bits" $\rightarrow$ $\Omega_X = \{0,1,2,3\}$

Y: "alternances en la seqüència de bits". $\rightarrow$ $\Omega_Y = \{0,1,2\}$

1) Construiu l'arbre de probabilitats
2) Definiu la funció de probabilitat de les variables X i Y
3) Calculeu els valors esperats
4) Calculeu les variàncies

Indicadors de V.A. Exercici - Solució

1) Arbre de probabilitats



Possibilitats	X (suma)	Y (#alternances)
000		
001		
010		
011		
100		
101		
110		
111		

3) Esperances

E(X) =

E(Y) =

2) Funcions de probabilitat

x	$P_X(x)$
0	
1	
2	
3	

y	$P_Y(y)$
0	
1	
2	

4) Variàncies

V(X) =

V(Y) =

Indicadors de V.A. Exercici

Continuant amb l'exemple de l'aeroport, ens plantegem estudiar determinades variables aleatòries per a modelar i obtenir informació en diversos aspectes del funcionament d'un aeroport.

En primer lloc, s'ha establert que la distribució de X: "nombre de viatgers que arriben a un punt de facturació per minut" és com es veu a continuació:

k	$P_X(X = k) = Px(k)$	$P_X(X \leq k) = Fx$
5	0.12	0.12
6	0.32	0.44
7	0.48	0.92
8	0.08	1

Calculeu la probabilitat que arribin 7 viatgers;

menys de 7 viatgers;

més de 7 viatgers;

entre 7 i 8 viatgers.

Trobem també l'esperança, la variància i la desviació tipus:

E(X) =

V(X) = $\rightarrow \sigma_X$ =

Indicadors de V.A. Exercici

D'una altra banda, quan s'ha estudiat el temps que un viatger roman en el taulell de facturació, s'ha trobat que la següent funció:

$$f_X(k) = 0.2 \cdot e^{-0.2k} \quad \text{per x > 0}$$

és un model adequat per a representar la variable aleatòria T de "temps (en minuts)".

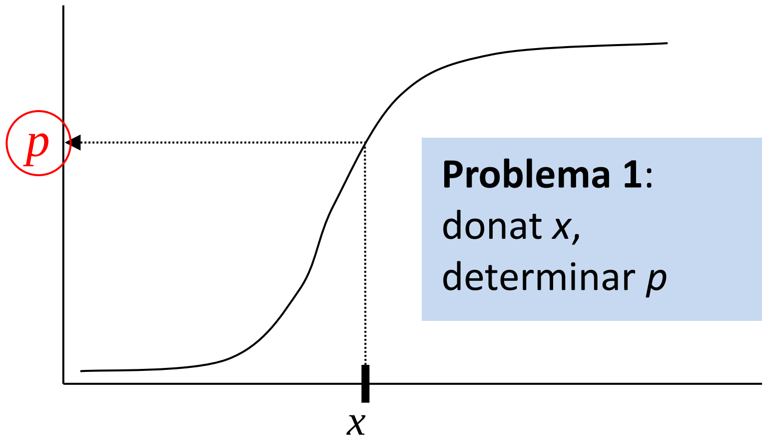
Calculeu la probabilitat que
- el temps sigui 7 minuts;
- menys de 7 minuts;
- més de 7 minuts;
- entre 7 i 8 minuts

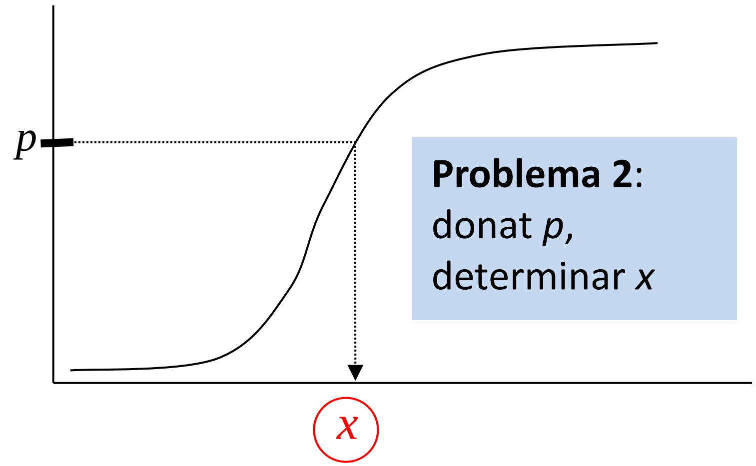
Com podem saber quin és el temps que, en mitjana*, un viatger roman en facturació?

* Nota: la paraula "mitjana" tant pot referir-se a l'esperança com a la mitjana mostral comú. En quin contexte s'està aplicant aquí?

Recapitulació. Tipus de problemes en VAC

El cas de VAC, és habitual plantejar problemes en els dos sentits:





Donat ***x*** calcular la probabilitat ***p*** tq:

$$p = F_X(x) = P(X \le x)$$

Donada una probabilitat ***p*** calcular ***x*** tq:

$$x = F_X^{-1}(p) \qquad (P(X \le x) = p)$$

Exemple: si els llits dels hotels mesuren 200 cm, quina proporció de congressistes poden dormir ben estirats?

Exemple: si desitgem que pugin dormir ben estirats el 98% dels congressistes, quina longitud han de tenir els llits?

Recapitulació. Tipus de problemes en VAD

Problema 1: donat *x*, determinar *P(X=x)*

Problema 2: donat *x*, determinar *P(X≤x)*

Problema 3: donat *P(X≤x)*, determinar *x*

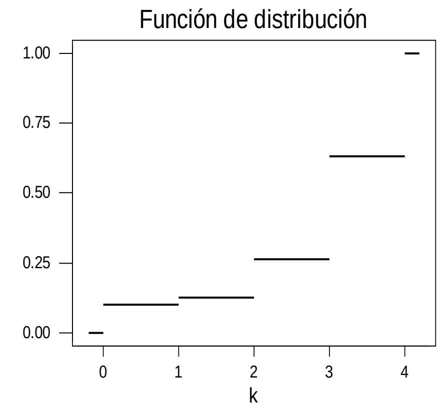
El cas del càlcul de quantils en VAD la solució pot ser aproximada:

Exemple:

Quin és el quantil de 0.25?
[o quin *és el percentil* 25? o del 25%]

2 és el valor on es troba una probabilitat de 0.25

k	$F_X(k)$
$(-\infty,0)$	0.00000
[0,1)	0.10144
[1,2)	0.12448
[2,3)	0.26272
[3,4)	0.63136
$[4,+\infty)$	1.00000

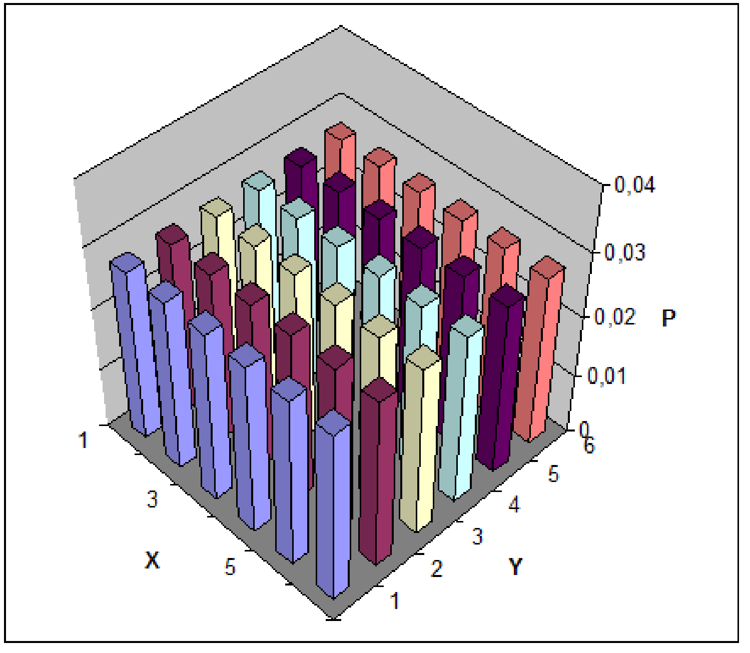


Opció a seguir: Agafarem el primer valor tal que la funció de distribució superi la probabilitat desitjada

Parells de variables. VAD

- Quan d'una experiència s'obtenen dues variables *X* i *Y*, quines relacions aleatòries es duen a terme entre elles?
- Per exemple, llancem dues vegades un dau equilibrat, i anomenem *X* al "primer resultat" i *Y* al "segon resultat".
- Raonablement, els dos llançaments són independents, llavors:

$$P(X=x \cap Y=y) = P(X=x) \cdot P(Y=y) = 1/6 \cdot 1/6 = 1/36, \quad \text{per} \quad x, y = 1, 2, \ldots, 6.$$



Funcions de probabilitat en parell de VAD

- Definim **funció de probabilitat conjunta** de *X* i *Y:*

$$P_{X,Y}(x,y) = P(X{=}x \cap Y{=}y)$$

- Definim la **funció de probabilitat de X condicionada per Y**:

$$P_{X/Y=y}(x) = P_{X,Y}(x,y) \;/\; P_Y(y)$$

- *X* i *Y* són **V.A. independents si:**

$$P_{X,Y}(x,y) = P_X(x) \cdot P_Y(y) \quad \Leftrightarrow \quad P_{X/Y=y}(x) = P_X(x) \quad \Leftrightarrow \quad P_{Y/X=x}(y) = P_Y(y)$$

En quant als indicadors esperança i variància, si *X* i *Y* són **independents** , llavors:

- $E(X \cdot Y) = E(X) \cdot E(Y)$
- $V(X \pm Y) = V(X) + V(Y)$ → ATENCIÓ: A l’expressió de la dreta sempre és un **“+”**

Funcions de prob. en parell de VAD. Exemple

En l'exemple de llançar dues vegades un dau equilibrat, a partir de les dues variables X "primer resultat" i Y "segon resultat" definim unes noves variables:

S = "suma dels dos resultats"

D = "diferència en valor absolut dels dos resultats"

Nota: Això **NO** són les taules de probabilitat conjunta, sinò unes taules auxliars amb els resultats possibles en les dues tirades

S=X+Y	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

K	**$P_S(S=k)$**
2	0.028
3	0.056
4	0.083
5	0.111
6	0.139
7	0.167
8	0.139
9	0.111
10	0.083
11	0.056
12	0.028

Nota: Per obtenir aquestes probabilitats es pot dividir els casos favorables entre els casos possibles només quan els successos siguin equiprobables.

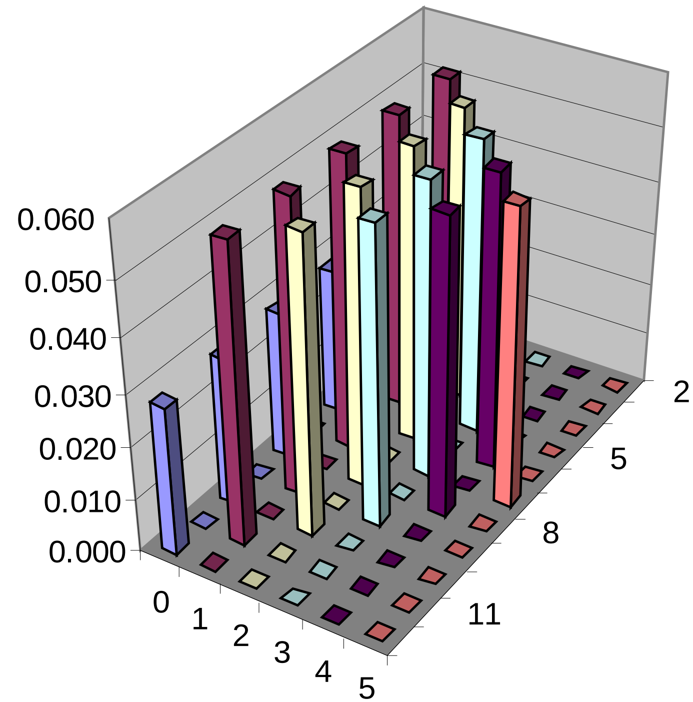
D=|X-Y|	**1**	**2**	**3**	**4**	**5**	**6**
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

K	**$P_D(D=k)$**
0	0.167
1	0.278
2	0.222
3	0.167
4	0.111
5	0.056

Funcions de prob. en parell de VAD. Exemple

A continuació, s'ha de trobar la funció P(*S*=s ∩ *D*=d), buscant els resultats coincidents.

S/D	0	1	2	3	4	5	
2	0.028	0.000	0.000	0.000	0.000	0.000	0.028
3	0.000	0.056	0.000	0.000	0.000	0.000	0.056
4	0.028	0.000	0.056	0.000	0.000	0.000	0.083
5	0.000	0.056	0.000	0.056	0.000	0.000	0.111
6	0.028	0.000	0.056	0.000	0.056	0.000	0.139
7	0.000	0.056	0.000	0.056	0.000	0.056	0.167
8	0.028	0.000	0.056	0.000	0.056	0.000	0.139
9	0.000	0.056	0.000	0.056	0.000	0.000	0.111
10	0.028	0.000	0.056	0.000	0.000	0.000	0.083
11	0.000	0.056	0.000	0.000	0.000	0.000	0.056
12	0.028	0.000	0.000	0.000	0.000	0.000	0.028
	0.167	0.278	0.222	0.167	0.111	0.056	1.000



Quan *S*=12 i *D*=0? Sols si *X*=6 i *Y*=6 → P(S=12,D=0) = 1/36

Quan *S*=9 i *D*=3? Quan *X*=6 i *Y*=3, o si *X*=3 i *Y*=6 → P(S=9,D=3) = 1/36 + 1/36=1/18

Quan *S*=7 i *D*=4? No existeix un resultat on es pugui produir aquesta combinació → P(S=7,D=4) = 0

Funcions de prob. en parell de VAD. Exemple

Són S i D independents? **No**. Es pot mirar de diverses maneres:

1) $P(S|D=i) \neq P(S)$ per alguna i
2) $P(D|S=i) \neq P(D)$ per alguna i
3) $P(D=i, S=j) \neq P(D=i) \cdot P(S=j)$ per alguna i i alguna j

Nota: És molt més fàcil demostrar la NO independència perquè només s'ha de trobar un cas que corrobori que no compleix la condició [P.ex, per demostrar que són independents, amb la primera propietat s'hauria de demostrar que $P(S|D=i) = P(S)$ per qualsevol ***i***]

	$P_{S\|D}(S\|D=0)$	$P_{S\|D}(S\|D=1)$	$P_{S\|D}(S\|D=2)$	$P_{S\|D}(S\|D=3)$	$P_{S\|D}(S\|D=4)$	$P_{S\|D}(S\|D=5)$
2	0.17	0.00	0.00	0.00	0.00	0.00
3	0.00	0.20	0.00	0.00	0.00	0.00
4	0.17	0.00	0.25	0.00	0.00	0.00
5	0.00	0.20	0.00	0.33	0.00	0.00
6	0.17	0.00	0.25	0.00	0.50	0.00
7	0.00	0.20	0.00	0.33	0.00	1.00
8	0.17	0.00	0.25	0.00	0.50	0.00
9	0.00	0.20	0.00	0.33	0.00	0.00
10	0.17	0.00	0.25	0.00	0.00	0.00
11	0.00	0.20	0.00	0.00	0.00	0.00
12	0.17	0.00	0.00	0.00	0.00	0.00

Funcions de prob. en parell de VAD. Exemple

- Què passa amb **μ** i **σ** per la variable *S* quan està condicionada per *D* [llavors parlem d'*esperances* i *variàncies condicionades* perquè utilitzem $P_{X|Y=y}(x)$ en lloc de $P_X(x)$)]
- Per *D*=0, S té un valor esperat de 7, i una desviació estàndard de 3.42 (gran dispersió).
- A mesura que *D* creix, la esperança es manté, però la desviació disminueix perquè la probabilitat es concentra al voltant del 7

	$P_{S\|D}(S\|D=0)$	$P_{S\|D}(S\|D=1)$	$P_{S\|D}(S\|D=2)$	$P_{S\|D}(S\|D=3)$	$P_{S\|D}(S\|D=4)$	$P_{S\|D}(S\|D=5)$
2	0.17	0.00	0.00	0.00	0.00	0.00
3	0.00	0.20	0.00	0.00	0.00	0.00
4	0.17	0.00	0.25	0.00	0.00	0.00
5	0.00	0.20	0.00	0.33	0.00	0.00
6	0.17	0.00	0.25	0.00	0.50	0.00
7	0.00	0.20	0.00	0.33	0.00	1.00
8	0.17	0.00	0.25	0.00	0.50	0.00
9	0.00	0.20	0.00	0.33	0.00	0.00
10	0.17	0.00	0.25	0.00	0.00	0.00
11	0.00	0.20	0.00	0.00	0.00	0.00
12	0.17	0.00	0.00	0.00	0.00	0.00
μ	**7.00**	**7.00**	**7.00**	**7.00**	**7.00**	**7.00**
σ	**3.42**	**2.83**	**2.24**	**1.63**	**1.00**	**0.00**

Funcions de prob. en parell de VAD. Exercici

En l'exemple anterior dels paquets de 3 bits que es poden enviar a través d'una línea de comunicació [Ω = {000, 001, 010, 011, 100, 101, 110, 111}] teniem les v.a.:

X: "suma dels 3 bits"

Y: "número d'alternances en la seqüència de bits"

Ompliu la taula amb la funció de probabilitat conjunta de les variables X i Y, a partir de la taula següent:

Possibilitats	X (suma)	Y (#alternances)
000	**0**	**0**
001	**1**	**1**
010	**1**	**2**
011	**2**	**1**
100	**1**	**1**
101	**2**	**2**
110	**2**	**1**
111	**3**	**0**

P_{YX}	*X=0*	*X=1*	*X=2*	*X=3*	
Y=0					1/4
Y=1					1/2
Y=2					1/4
	1/8	3/8	3/8	1/8	

Quant val E(Y|X=0)?

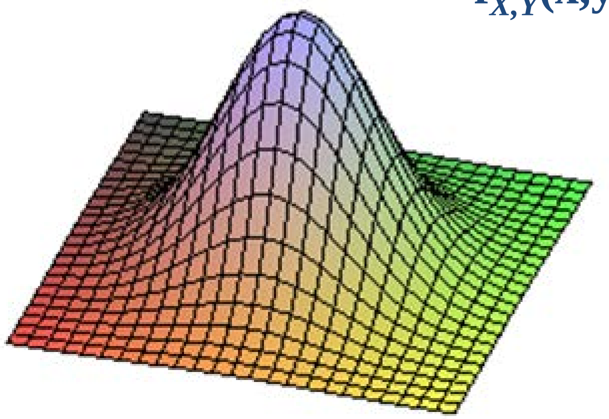
I V(Y|X=0)?

Parell de Variables. VAC

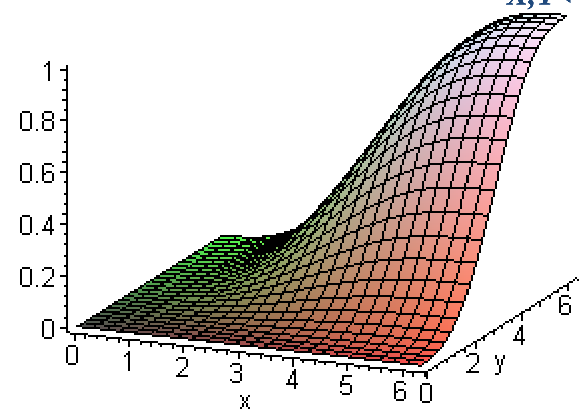
En cas de que existeixin dos variables contínues *X* i *Y* en la mateixa experiència, la relació comuna es reflecteix a través de la **funció de densitat conjunta, $f_{X,Y}(x,y)$.**

- Sigui $F_{X,Y}(x,y) = P(X \leq x \cap Y \leq y)$, *funció de distribució conjunta* de les variables aleatòries
- Si derivem $F_{X,Y}(x,y)$ respecte a les variables (x i y), obtenim $f_{X,Y}(x,y)$
- La definició de funcions condicionades és idéntica que per a V.A.D: $f_{X|Y=y}(x) = f_{X,Y}(x,y) / f_Y(y)$
- El volum total tancat sota $f_{X,Y}(x,y)$ és igual a 1, i una porció d'ell equival a una probabilitat

$f_{X,Y}(x,y)$



$F_{X,Y}(x,y)$



Indicadors de parell de V.A.

- A partir d'un parell de variables *X* i *Y* definim indicadors de la seva relació bivariant (equivalents als mostrals que es veuen a estadística descriptiva)

- La covariància indica si existeix relació lineal o no, a partir del producte, per cada parell de valors, de la diferència respecte al seu valor esperat

$$\boldsymbol{VAD} \rightarrow Cov(X,Y) = \sum_{\forall x} \sum_{\forall y} (x - E(x))(y - E(y)) \cdot p_{XY}(x,y)$$

$$\boldsymbol{VAC} \rightarrow Cov(X,Y) = \iint_{\forall x,y} (x - E(x))(y - E(y)) \cdot f_{XY}(x,y)\,\mathrm{dxdy}$$

- La correlació indica si existeix relació lineal o no relativitzant-ho a valors entre -1 i 1 (a partir de la covariància i dividint per les desviacions corresponents)

$$corr(X,Y) = \rho = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Nota: per qualsevol parell de variables *X* i *Y* → $-1 \leq \rho_{X,Y} \leq 1$

Nota: La correlació és més interpretable per estar estandaritzada

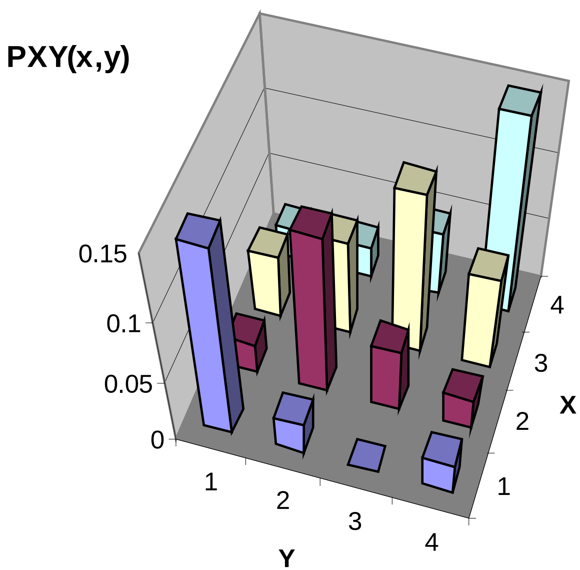
Indicadors de parell de V.A. Exemple

X/Y	1	2	3	4	
1	0.150	0.025	0.000	0.025	0.200
2	0.025	0.125	0.050	0.025	0.225
3	0.050	0.075	0.125	0.075	0.325
4	0.025	0.025	0.050	0.150	0.250
	0.250	0.250	0.225	0.275	1.000

Suposem dos variables *X* i *Y* que presenten la funció $P_{X,Y}(x,y)$ de la taula.
Veiem que:

$\mu_X = 2.625$ $\sigma_X = 1.065$

$\mu_Y = 2.525$ $\sigma_Y = 1.140$



Observeu la relació entre les variables: en general, el valor de Y és de la magnitud de *X*, hi ha una *relació directa*, encara que no sigui determinista [*Y* = *X*]

covariància → **Cov(*X*,*Y*) = 0.647**

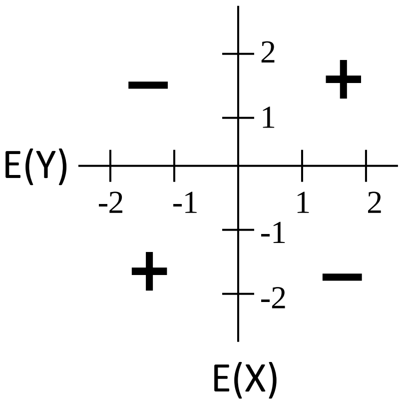
[idea de relació positiva]

correlació → $\rho_{X,Y}$ **= 0.533**

[indica *magnitud* de la relació, ja que sabem que l'indicador està entre -1 i 1]

Indicadors de parell de V.A. Exemple

Si una variable té $\mu = 0$ i $\sigma = 1$ es diu que és una variable *centrada* i *reduïda* (estandaritzada). Igualment, diem que la correlació ve estandaritzada perquè pren valors entre -1 i 1.



Aquests quadrants indiquen el signe del producte resultant. Si la probabilitat es reparteix preferentment entre els quadrants positius, la relació entre *X* i *Y* és *directa*. En cas contrari la relació és *inversa* (si *X* augmenta, *Y* tendeix a disminuir).

- Si $|\rho_{X,Y}|=1$, la relació és total i lineal: $\boldsymbol{Y = a+b\cdot X}$ (signe $\rho_{X,Y}$ = signe *b*)
- $|\rho_{X,Y}|$ a prop de 1 $\Rightarrow$ *X* i *Y* estan molt relacionades linealment
- *X* i *Y* independents $\Rightarrow \rho_{X,Y}=0$, però a la inversa no és cert
- La magnitud de la covariància depen de l'escala agafada per les variables [Per exemple, si canvio les unitats de metres a quilòmetres, la covariància canviarà però la correlació, no]

Indicadors de parell de V.A. Exercici

Tenim la distribució de probabilitat conjunta entre "Memòria d'un ordinador entre 1 i 6 GB" (M) i "nombre (entre 0 i 5) de bloquejos o incidències mensuals" (B)

Distribució conjunta M/B	0	1	2	3	4	5
1	0.00	0.020	0.030	0.050	0.060	0.08
2	0.00	0.015	0.015	0.035	0.045	0.03
3	0.00	0.030	0.060	0.050	0.030	0.03
4	0.06	0.050	0.045	0.030	0.030	0.00
6	0.10	0.075	0.030	0.000	0.000	0.00

k	$P_M(k)$
1	0.24
2	0.14
3	0.20
4	0.215
6	0.205

k	**0**	**1**	**2**	**3**	**4**	**5**
$P_B(k)$	0.16	0.19	0.18	0.165	0.165	0.14

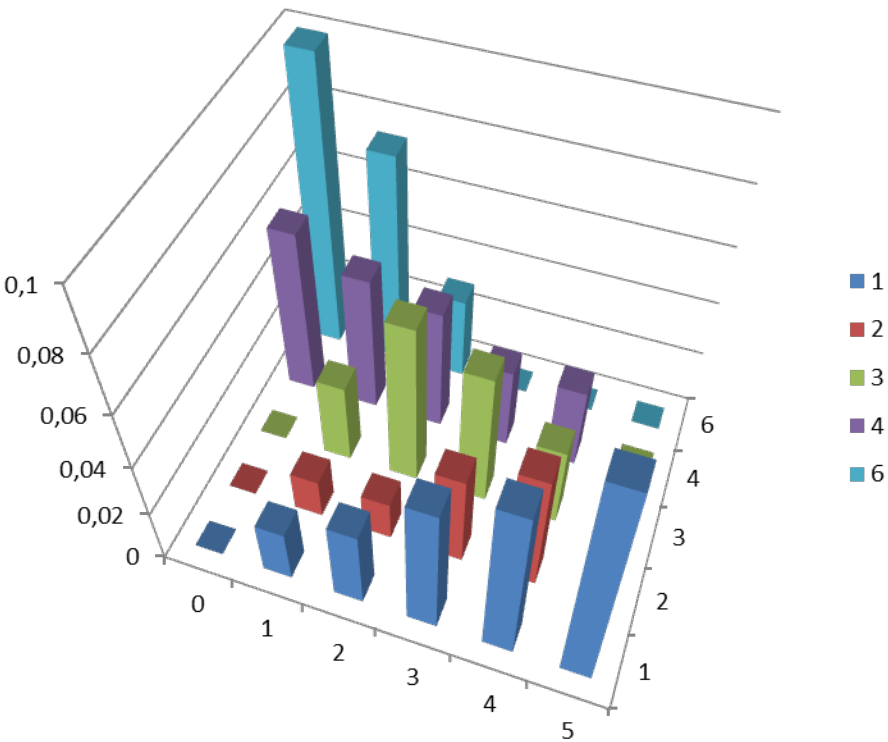
Distribucions marginals

1. Representeu $P_{M\,|\,B=3}()$ i $P_{M\,|\,B=4}()$. Que es dedueix d'això?
2. Representeu $P_{B\,|\,M=2}()$ i $P_{B\,|\,M=4}()$. Que es dedueix d'això?
3. Calculeu prob. que un ordinador tingui menys de 3 GB i més de 2 incidències
4. Calculeu prob. que tingui més de 2 incidències si té menys de 3 GB
5. Calculeu prob. que tingui més de 3 GB si ha tingut menys de 2 incidències

Indicadors de parell de V.A. Exercici

En mitjana, els ordinadors tractats tenen una memòria de 3.21 GB, i han patit una mitjana de 2.405 incidències mensuals. La distribució és bastant uniforme, amb desviacions estàndards de 1.765 GB i 1.66 inc./m.

Les dues variables no són independents, la funció $P_{M,B}()$ mostra una clara relació inversa (com més memòria, menys incidències):



Quant val la Covariància?

Quant val la Correlació?

Si $|\rho_{X,Y}|$ és alta, la variabilitat de *B* condicionada amb *M* és sensiblement menor que la global de *B*.

Indicadors de parell de V.A. Propietats

Siguin X i Y dues variables aleatòries, i a i b dos escalars:

- **Esperança**
  - $E(X \cdot Y) = E(X) \cdot E(Y) + Cov(X,Y)$
- **Variància**
  - $V(X+Y) = V(X) + V(Y) + 2\,Cov(X,Y)$
  - $V(X-Y) = V(X) + V(Y) - 2\,Cov(X,Y)$
- **Covariància**
  - $Cov(aX, bY) = a \cdot b \cdot Cov(X,Y)$
  - $Cov(X,X) = V(X)$

Models de variable aleatòria (VAC i VAD)

Bloc 3 – Probabilitat i Estadística

Març 2015

Índex

Models de VAD i VAC

- **Definició Wikipedia**: *List of probability distributions: "Many probability distributions are so important in theory or applications that they have been given specific names".*
- **Estudiarem les VAD**: Binomial, Poisson, Bernoulli, Geomètrica, Binomial Negativa
- **Estudiarem les VAC**: Exponencial, Normal, Uniforme
- A partir dels paràmetres de cada model es calculen **indicadors**
  - Esperança → E(X) = μ_X
  - Variància → V(X) = σ_X^2
- A partir de les **funcions de probabilitat i distribució** de probabilitat es calculen probabilitats:

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

- 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: $F_X(x_\alpha) = \alpha$ $(0 \leq \alpha \leq 1)$ (en VAD el càlcul és aproximat)

Model de Bernoulli

- **Definició**: Número d'èxits en la realització d'1 únic experiment amb 2 possibles resultats: 0 ("no èxit") i 1 ("èxit")
- **Notació**: X~Bern(p)
- **Paràmetres**: p (probabilitat d'observar 1 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*

Model associats a la Bernoulli

- En una experiència aleatòria que implica repetició de proves Bernoulli independents, es plantegen com a distribucions interessants les següents **VAD**:
  - sobre n repeticions, número "d'èxits" totals (dist. **binomial**)
  - número de repeticions fins observar el primer "èxit" (dist. **geomètrica**)
  - número de repeticions fins observar el r-èssim "èxit" (dist. **binomial negativa**)

- 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:
  - número "d'èxits" en l'interval (dist. de **Poisson**) → **VAD**
  - temps entre "èxits" (dist. **Exponencial**) → **VAC**

- En aquests darrers casos parlem de 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 la de temps entre "èxits" (http://en.wikipedia.org/wiki/Poisson_process)

Model Binomial

- **Definició**: Número de èxits en la repetició de *n* proves de Bernoulli independents amb probabilitat *p*
- **Notació**: X~B(n,p)
- **Paràmetres**: n (nombre de repeticions), p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat**:

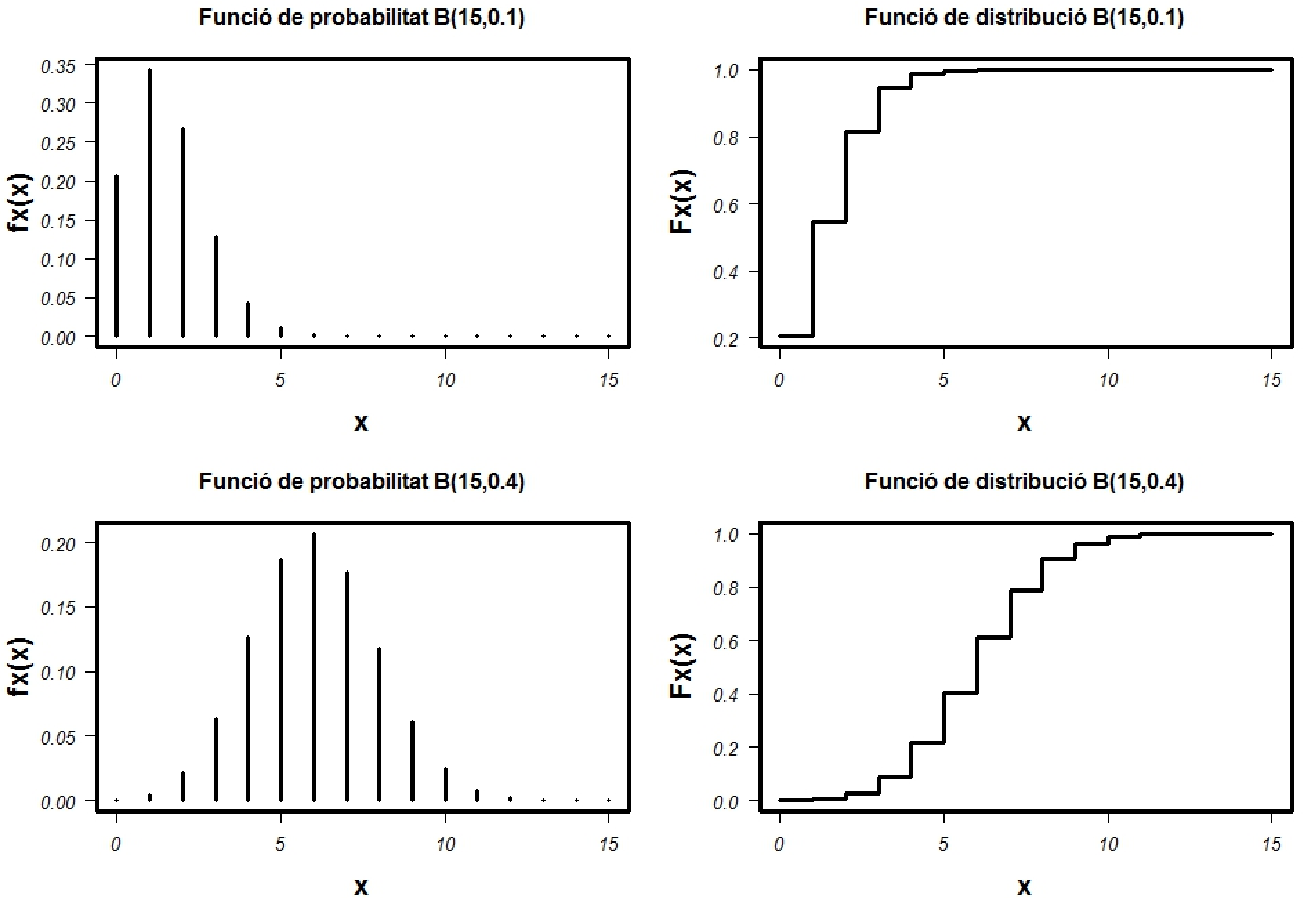
$$P(X = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k} \quad amb \quad k = 0, 1, \dots, n$$

- No té **funció de distribució** analítica → S'utilitzen taules o R
- **Indicadors:**
  - E(X) = n·p
  - V(X)=n·p·q

R: dbinom, pbinom, qbinom

Model Binomial. Representació gràfica

Ex: Com es distribueix el <u>número de correus *spam* entre els 15 primers rebuts al dia</u> segons si la probabilitat de que un correu sigui *spam* és...?



Model Binomial. Exemple

Suposem la VAD X: “número de uns en 4 tirades d’un dau”

X~B(n=4, p=1/6)

resultat				X	probabilitat			
0	0	0	0	0	q	q	q	q
0	0	0	1	1	q	q	q	p
0	0	1	0	1	q	q	p	q
0	0	1	1	2	q	q	p	p
0	1	0	0	1	q	p	q	q
0	1	0	1	2	q	p	q	p
0	1	1	0	2	q	p	p	q
0	1	1	1	3	q	p	p	p
1	0	0	0	1	p	q	q	q
1	0	0	1	2	p	q	q	p
1	0	1	0	2	p	q	p	q
1	0	1	1	3	p	q	p	p
1	1	0	0	2	p	p	q	q
1	1	0	1	3	p	p	q	p
1	1	1	0	3	p	p	p	q
1	1	1	1	4	p	p	p	p

- Contarem 0 i 4 → 1 vegada,
- Contarem 1 i 3 → 4 vegades
- Contarem 2 → 6 vegades.
- Els resultats no són equiprobables excepte si p=0.5:

- obtenir 4 cares diferents de “1” té prob. de $(5/6)^4$=0.48
- obtenir 4 “1” té prob. de $(1/6)^4$=0.00077

Model Binomial. Ex. de càlcul de probabilitats

…	
12	0.8684
13	0.9423
14	**0.9793**
15	0.9941
16	0.9987
17	0.9998
…	

Sigui $X \sim$ B(n=20,p= 0.5) i alguna de les probabilitats tabulades:

- **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 taules → $P(X=14)=P(X\leq 14)-P(X\leq 13)=\mathbf{0.037}$
  - 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 taules → $P(X\leq 14)=\mathbf{0.979}$
  - 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 taules →$P(X\leq x_{0.95})=0.95\rightarrow x_{0.95}=\mathbf{14}\ ja\ que\ P(X\leq 14)=0.9793>0.95$
  - Amb R →$P(X\leq x_{0.95})=0.95\rightarrow qbinom(p=0.95,size=20,prob=0.5)=\mathbf{14}$

Nota: A les taules estadístiques trobarem les probabilitats acumulades per a alguns valors dels paràmetres *n* i *p* (per exemple $2 \leq n \leq 20$ i *p* valors entre 0.05 i 0.5).

Nota: Per valors de $p > 0.5$ es poden obtenir probabilitats acumulades amb una Binomial complementària:

$P(X \leq k) = 1 - P(Y \leq n-k-1)$ on $X \sim B(n, p)$ i $Y \sim B(n, 1-p)$

Model Binomial. Exercici

La cabina de discos d'un servidor conté 18 discs idèntics. Un disc de cada 5 necessita ser substituït al cap d'un any per ser reparat

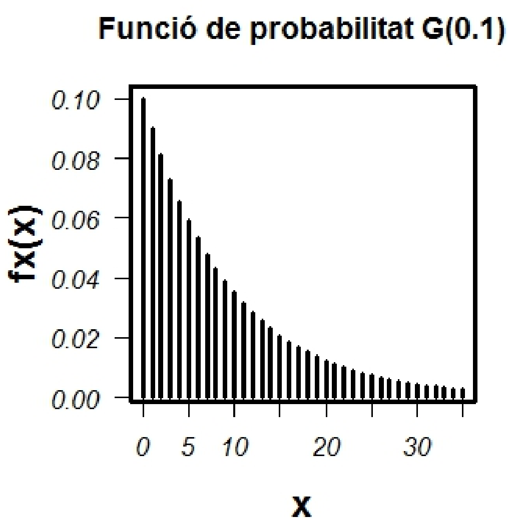
VAD que compta substitucions:

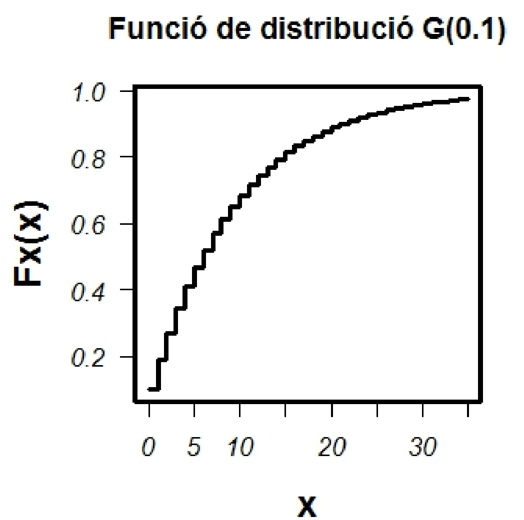
R: "número de discs que han de ser substituïts al cap de l'any"

Model per a la VAD:	R ~ B(n=18, p=0.2)
Número esperat de substitucions anuals:	E(R)=3.6
Variància i desviació típica:	V(R)=2.88 → $\sigma_x = 1.697$
Prob. d'observar 4 casos:	P(R=4)=0.2153
Prob. sols una substitució:	P(R=1) =0.0811
Prob. de tenir-ne menys de 3:	P(R<3)=0.2713
Prob. que almenys 6 discs siguin substituïts:	P(R≥6)=0.1329

Model Geomètrica

- **Definició**: Número d'intents (k) d'un experiment de Bernoulli fins observar el primer èxit
- **Notació**: X~Geom(p)
- **Paràmetres**: p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat** : $P(X = k) = q^{k-1} \cdot p$





R: dgeom, pgeom, qgeom

Nota: no està acotada, qualsevol valor enter > 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 traslladada a valors més baixos, i $F_X(k)$ creix més ràpid.

Model Binomial negativa

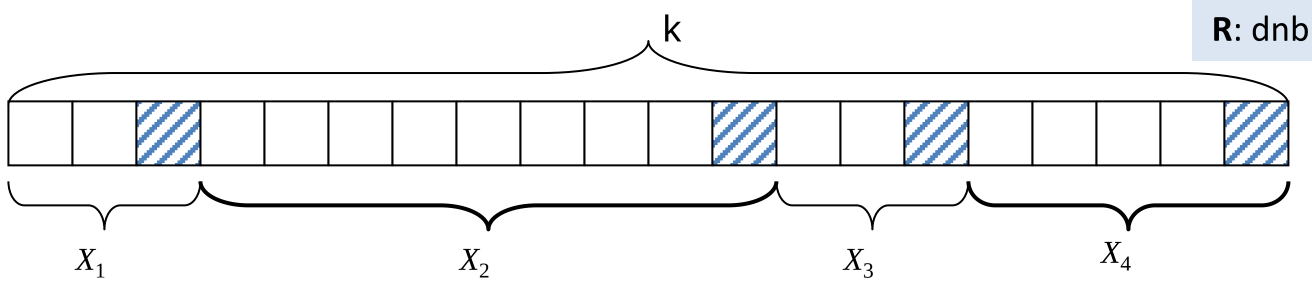
- **Definició**: Número de repeticions (k) d'un experiment de Bernoulli fins observar *r* èxits
- **Notació**: X~BN(r , p)
- **Paràmetres**: r (nombre d'èxits), p (probabilitat d'observar 1 èxit)
- **Funció de probabilitat**:

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

Sense comptar l'últim intent (que ha de ser un èxit), són *r* -1 èxits barrejats en qualsevol combinació amb k-*r* fracassos. $P_X(k)$, amb $k \geq r$ és

- En general, podem pensar que una BN és una suma de r geomètriques independents.

R: dnbinom, pnbinom, qnbinom



- Ex: Quantes repeticions fan falta per aconseguir r èxits (r>1)? Si r=1, tenim la distribució geomètrica.

Model Poisson

- **Definició**: Número d'ocurrències favorables en un determinat interval de temps o espai
- **Notació**: X~P(λ)
- **Paràmetres**: λ (taxa d'aparició de l'event)
- **Funció de probabilitat**:

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

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: una trucada a una centraleta)

- **Indicadors:**
  - E(X) = λ
  - V(X)= λ

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.

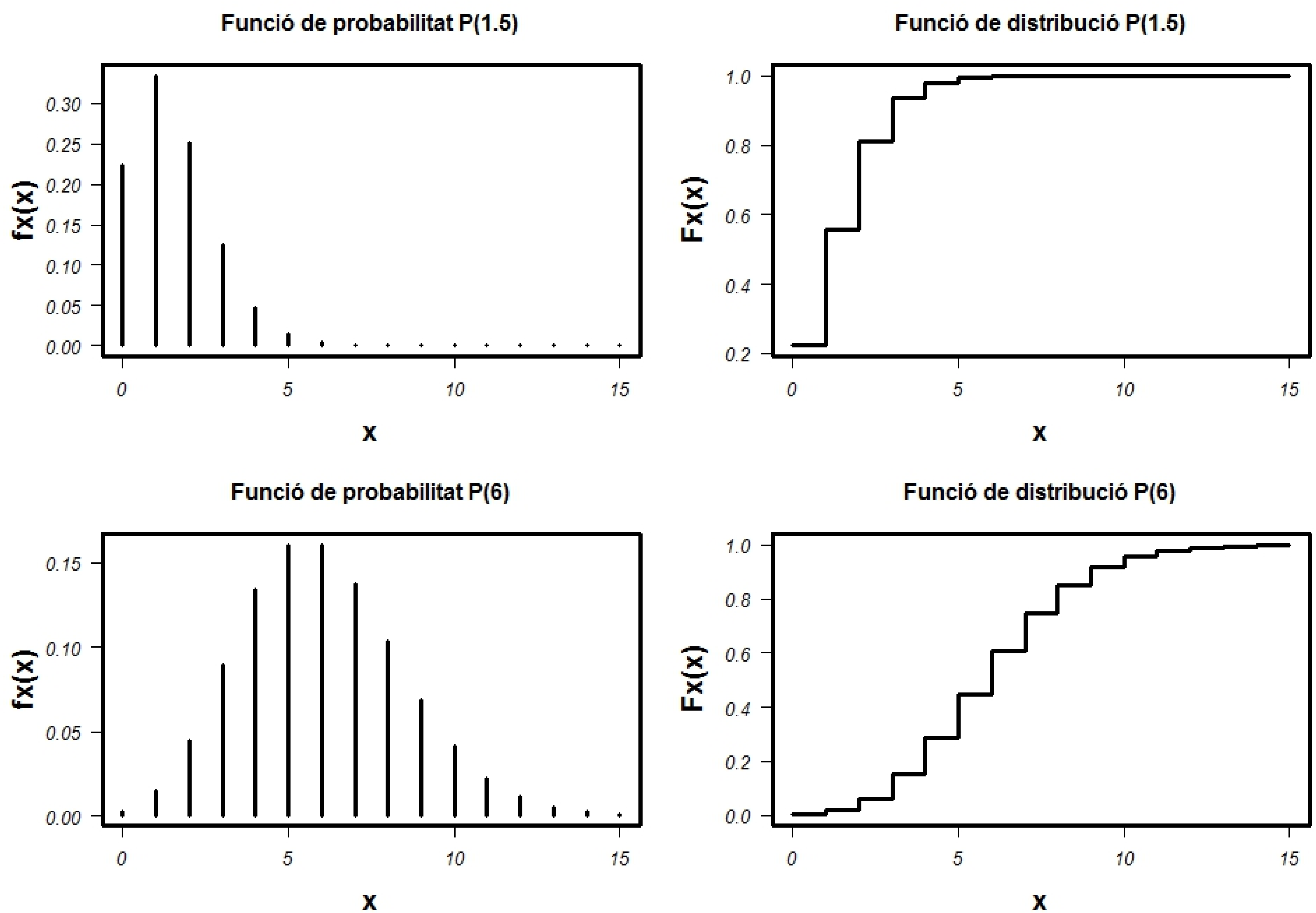
- **λ** és un número real positiu que representa la taxa mitjana d'ocurrències per unitat considerada

  [Ex: 10 trucades/hora]

R: dpois, ppois, qpois

Model Poisson. Representació gràfica

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



Model Poisson. Ex. de càlcul de probabilitats

λ	0	1	2	3	4	5	6	7	8
2	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000

Sigui $X \sim P(\lambda = 2)$ i alguna de les prob. tabulades:

- **Probabilitat puntual**. Quina és la probabilitat de 3?
  - Amb fórmules → $P(X=3) = \frac{e^{-2} \cdot 2^3}{3!} = \mathbf{0.1804}$
  - Amb taules → $P(X=3) = P(X \leq 3) - P(X \leq 2) = \mathbf{0.18}$
  - 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 taules → $P(X \leq 3) = \mathbf{0.857}$
  - 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 taules → $P(X \leq x_{0.95}) = 0.95 \rightarrow x_{0.95} = \mathbf{5}\ ja\ que\ P(X \leq 5) = 0.983 > 0.95$
  - Amb R → $P(X \leq x_{0.95}) = 0.95 \rightarrow qpois(p = 0.95, lambda = 2) = \mathbf{5}$

Nota: A les taules estadístiques només trobarem les probabilitats acumulades per a alguns valors del paràmetre λ
Nota: La suma de VAD Poisson és una VAD Poisson amb paràmetre λ igual a la suma dels paràmetres de cadascuna. Per això 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)

Model Poisson. Exercici

El centre de càlcul d'una empresa atén les incidències que sorgeixen als treballadors. S'ha observat que aquestes apareixen esporàdicament, encara que l'elevat número d'usuaris implica que el volum de problemes a tractar diàriament sigui considerable (s'ha suposat una mitjana de 2.35 incidències/dia).

Quina és la distribució de la v.a "número d'incidències diàries"?

X: "número d'incidències al dia" → $X \sim P(\lambda=2.35)$

Quina és la seva Esperança? I la seva desviació típus?

$E(X) = 2.35$ incidències i $\sigma_X=1.53$ incidències

Quina és la distribució de la v.a "número d'incidències en 5 dies"? I en 7?

X_5: "número d'incidències 5 dies" → $X_5 \sim P(\lambda=11.75)$; $X_7 \sim P(\lambda=16.45)$

Probabilitat que en un dia es produeixi 3 incidències

$P(X=3) = 0.2063$

Probabilitat d'observar menys de 3 incidències en un dia:

$P(X<3) = P(X\leq 2) = 0.5828$

Entre el dilluns i el dimarts s'han rebut sis incidències. Quina és la probabilitat que de dilluns a divendres es tractin no més de 15?

$P(X_3\leq 15-6) = P(X_3\leq 9) = 0.8254$

Quina és la probabilitat que cap dia de la setmana laboral presenti incidències?

$P(X_7=0) = 0.0000079$

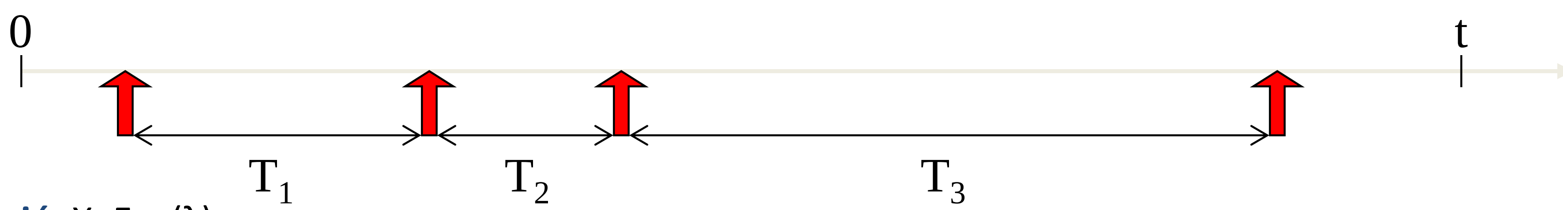
TAULA resum de models de VAD

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

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

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

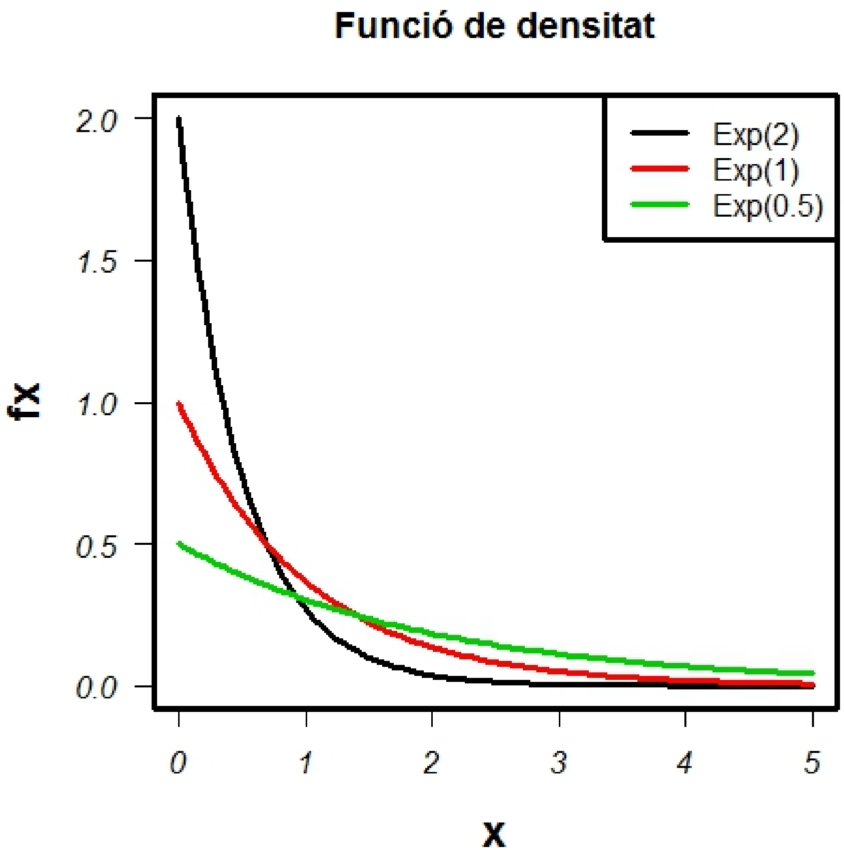
$$\mathrm{f}_{\mathrm{X}}(x) = \lambda \cdot e^{-\lambda x} \qquad amb \quad x > 0$$

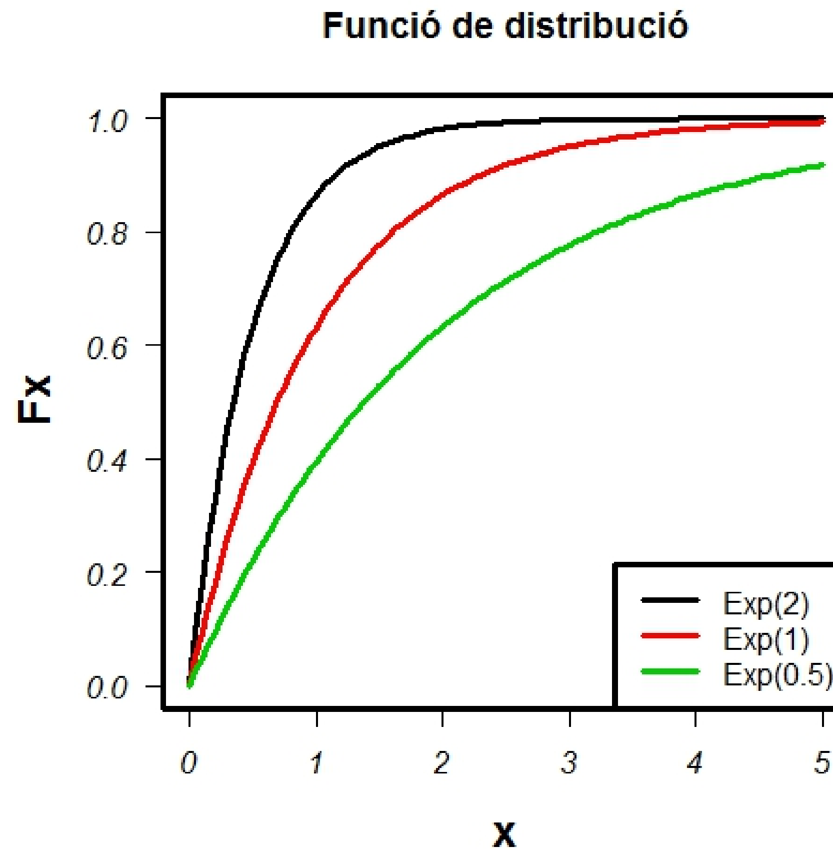
$$F_{\mathrm{X}}(x) = 1 - e^{-\lambda x} \qquad amb \quad x > 0$$

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

Model Exponencial. Representació gràfica

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

La distribució Exponencial no té taules, perquè la funció de distribució té expressió analítica

Model Exponencial. Exemples

Siguin les següents variables aleàtories:

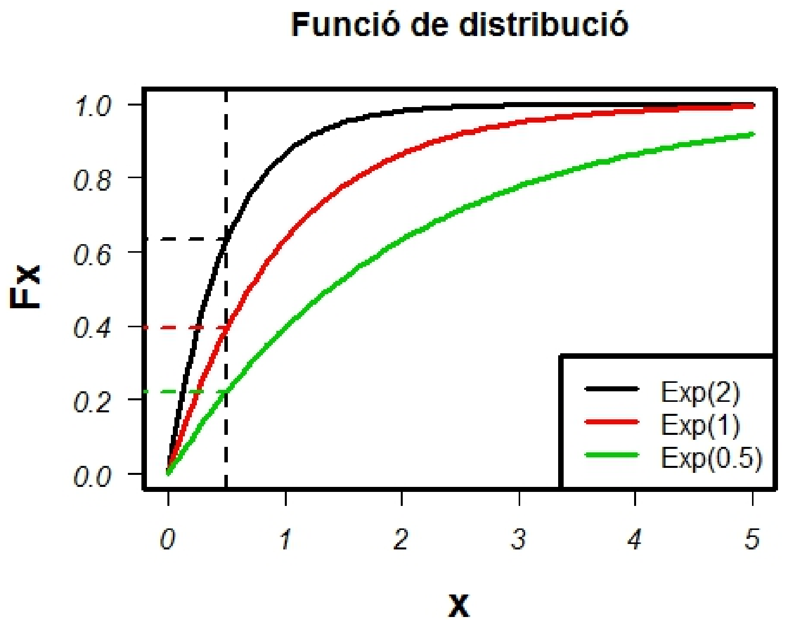
N: "número de peticions/seg a un servidor de BBDD" → N ~ P(λ)

T: temps (seg.) transcorregut entre dos peticions consecutives

Si λ = 2 arr./s. → P(T<0.5) = 0.632

Si λ = 1 arr./s. → P(T<0.5) = 0.393

Si λ = 0.5 arr./s. → P(T<0.5) = 0.221



Nota: Si les arribades són més freqüents (λ alt), el temps entre arribades són més curts. Per tant, és més probable trobar temps per sota de mig segon quan λ és major (*F_X creix més ràpid*).

Model Exponencial. Observacions

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

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

- No tindrem taules estadístiques per probabilitats acumulades en el model Exp. 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 anteriorietat al moment present. Matemàticament:

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

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

Solució: Igual

Model Exponencial (i Poisson). Exercici

El centre de càlcul d'una important empresa atén les incidències que els sorgeixen als treballadors. Se suposa una mitjana de 4 incidències/dia, i 8 hores laborables al dia.

Quina és la distribució de la v.a "número d'incidències diàries"?

X: "número d'incidències al dia" → $X \sim P(\lambda=4)$

Quina és la distribució de la v.a "número d'incidències per hora"?

Xh: "número d'incidències per hora" → $Xh \sim P(\lambda=4/8) = P(\lambda=1/2)$

Probabilitat de rebre 0 incidències en un dia:

$P(X = 0) = 0.0183$

Probabilitat de rebre 0 incidències en una hora:

$P(Xh = 0) = 0.607$

Quina és l'esperança de la variable temps (en hores) entre incidències?

T: "Temps en hores entre incidències" $\sim Exp(\lambda = 1/2) \rightarrow E(T) = 2$ hores

I la desviació tipus?

$V(X) = 1/\lambda^2 \rightarrow V(X) = 4 \rightarrow \sigma_x = 2$ hores

Probabilitat d'estar 8 o més hores sense rebre incidències:

$P(T>8) = 0.0183$

Model Exponencial. Aplicacions

- **"Failure Rate and Reliability"** (Jane Horgan lecture 17)

  $X \sim Exp(\lambda)$ $\quad F_X(x)=P(X \le x)=1 - e^{-\lambda \cdot x}$ $\quad R_X(x)= P(X > x) = e^{-\lambda \cdot x}$

  – R és la funció de "reliability" o de fiabilitat (probabilitat de durar més de..)

  – λ és "failure rate" o taxa d'error

  – E(X) = 1/λ és "MTTF" o "Mean Time To Failure"

- **"Modelling Response Times: M/M/1"** (Jane Horgan lecture 17)

  – La teoria de cues permet estudiar sistemes en que interaccionen més d'una variable (per exemple temps de resposta en un sistema d'espera amb cues).

El model M/M/1 indica: 1 per contemplar una sola cua no finita, i dos M's pels dos temps d'arribada i servei exponencials (*Exp*(λ) i *Exp*(μ) o número d'arribades **P**(λ) i serveis **P**(μ)), que compleixen la propietat de Markov de no tenir memòria.

Amb l i m es defineix un indicador del sistema com factor de càrrega o "traffic intensity" (λ / μ):
- si λ /μ > 1 (λ > μ) → la taxa d'arribades és superior a la de sortides (sobreutilització del sistema)
- si λ /μ = 1 (λ = μ) →la taxa d'arribades s'iguala amb la taxa de sortides
- si λ /μ < 1 (λ < μ) → la taxa d'arribades és inferior a la de sortides (infrautilització del sistema)

(el model M/M/1 permet un tractament per teoria de cues o per simulació. Models més complexos poden permetre sols el tractament per simulació)

Model Exponencial (i Poisson). Exercici fiabilitat

El centre de càlcul d'una important empresa garanteix treballar amb una mitjana de 2 hores entre incidències.

Quina és la distribució de la v.a "temps en hores entre incidències"?

T: "temps en hores entre incidències" → $E(T) = 1/\lambda = 2$ → $T \sim Exp(\lambda=0.5)$

Quina és la funció de distribució?

$F_T(t) = P(T \leq t) = 1-e^{-t/2}$

Quina és la funció de fiabilitat?

$R_T(t) = P(T > t) = e^{-t/2}$

Quina és la taxa d'errors?

$\lambda = 0.5$ errors/hora

Quin és el MTTF?

MTTF = E(T) = 2 hores

Quin és el valor d'hores entre incidències que podem garantir que es superarà amb una fiabilitat del 78%?

$P(T > t_{0.78})=0.78$ → $t_{0.78} = 0.5$

Model Exponencial (i Poisson). Exercici

Continuem analitzant el cas de l'aeroport, donant importància al procés d'arribades de passatgers als punts de facturació. Respon a les següents preguntes.

1. Un determinat punt de facturació es caracteritza perquè el número de viatgers que arriben per minut es distribueix segons una Poisson amb mitjana de 9.5. Calcula la probabilitat que aquest número sigui menor que 7 [0.165]
2. En aquest punt de facturació, quina és la probabilitat d'observar exactament 10 arribades en un minut? [0.123]
3. En el mateix punt de facturació, quina és l'esperança de la variable temps (*en segons*) entre dues arribades? [6.316 sg]
4. Al punt de facturació, quina és la probabilitat d'estar menys de 4 segons sense arribades? [0.469]
5. Considerant 18 punts de facturació caracteritzats per una probabilitat 0.8 d'observar exactament 0 arribades en un minut, quina és la probabilitat de tenir més de 14 punts amb 0 arribades? [0.501]

Model Uniforme

- **Definició**: VAC amb funció de densitat constant en un determinat rang [la probabilitat de pertanyer a un interval concret dins d'aquest rang només depén de la longitud del interval]
- **Notació**: X~U(a,b)
- **Paràmetres**: a (valor mínim del rang de X), b (valor màximdel rang de X)
- **Funció de denistat 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 $x > b$ → $$F_X(x) = \frac{x-a}{b-a} \qquad amb \quad a < x < b$$

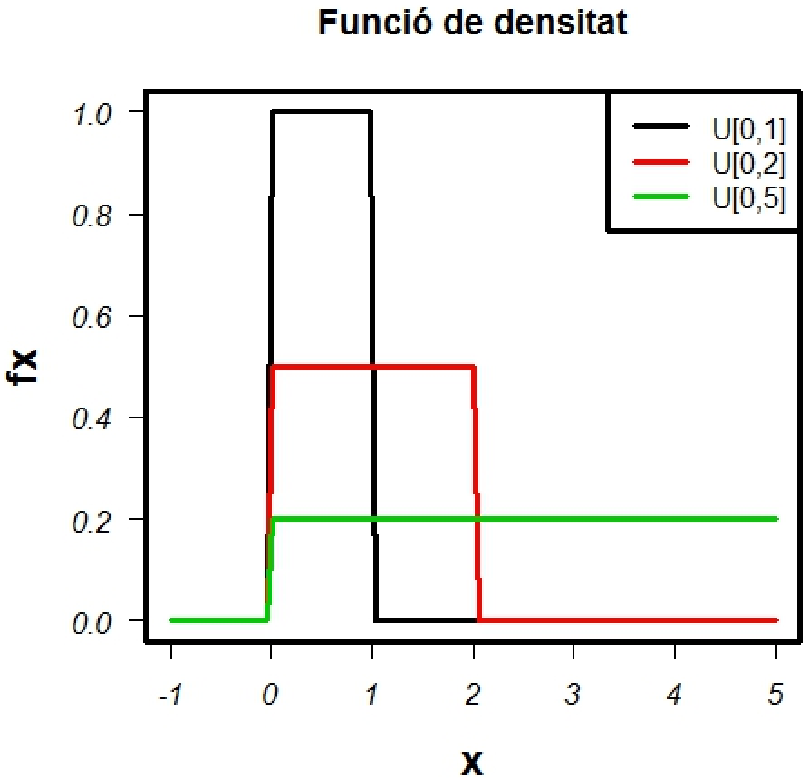
- **Indicadors:**
  - E(X) = (b + a)/2
  - V(X) = $(b - a)^2/12$

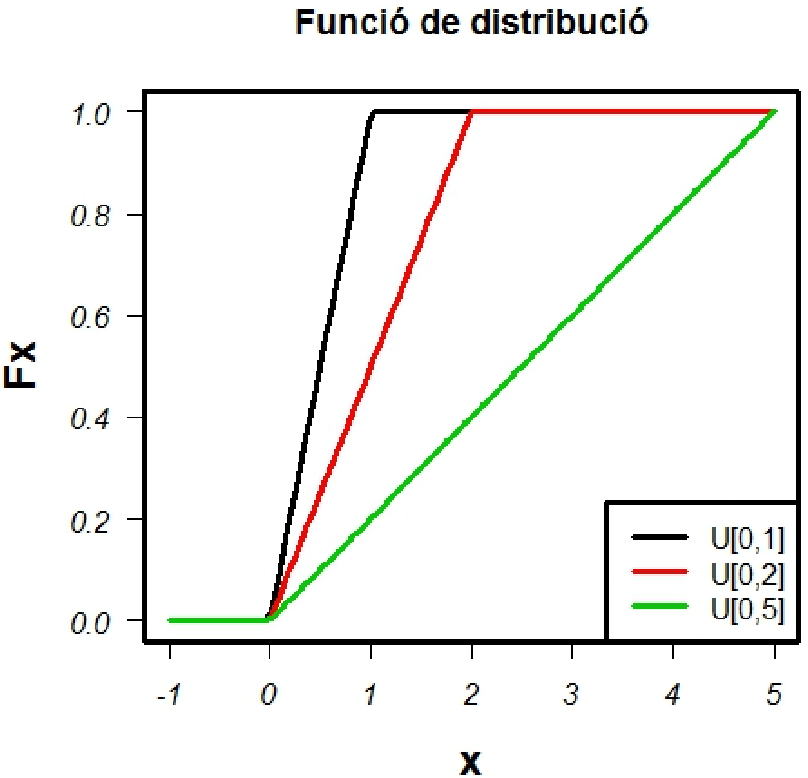
$$E(X) = \mu_X = \int_{-\infty}^{+\infty} x \cdot f_X(x)dx = \int_a^b x \cdot \frac{1}{b-a} dx =$$
$$= \frac{1}{b-a} \cdot \frac{x^2}{2}\bigg]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

Intuïtivament ja es veu que la mitjana ha de ser el centre del interval

Model Uniforme. Representació gràfica

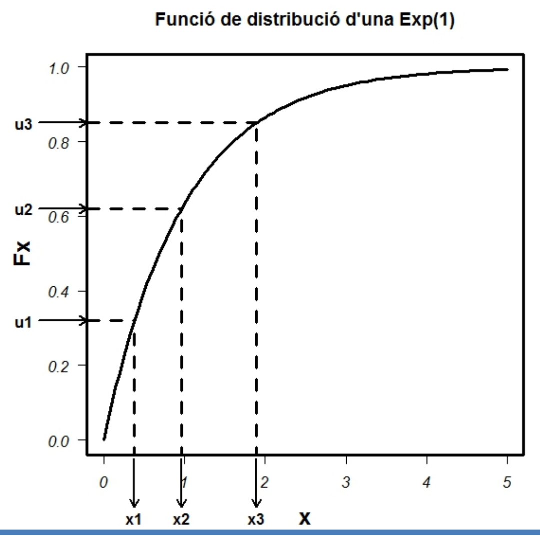
Ex: Com es distribueixen el nombres aleatoris entre 0 i 1? I entre 0 i 2? I entre 0 i 5?





Model Uniforme i Exponencial. Aplicació

- Generar números pseudo-aleatoris amb distribució U[0,1] és relativament complicat
- No obstant, generar números aleatoris de qualsevol distribució, un cop tenim els nombres pseudo-aleatoris amb distribució U[0,1] és senzill:
  - Per generar Y ~ U[a,b] → Y = a + (b-a)·u~U[a,b] on u són valors d'una U[0,1]
  - Per generar T ~ exp(λ) → T = $F^{-1}(u) = -\ln(1-u)/\lambda \sim \exp(\lambda)$ on u són valors d'una U[0,1]
- En general, el mètode de la transformació inversa (que empra F^{-1}) permet generar valors de qualsevol distribució (no és necessari que F tingui expressió analítica)



Generar valos d'una exponencial:
- Es parteix de $F_X(x) = 1 - e^{-\lambda \cdot x}$
- S'agafen valors $u_i \sim U[0, 1]$
- Es troben les x_i tals que $F_X(x_i) = u_i = 1 - e^{-\lambda \cdot x_i}$
- Els x_i segueixen una distribució Exponencial(λ)

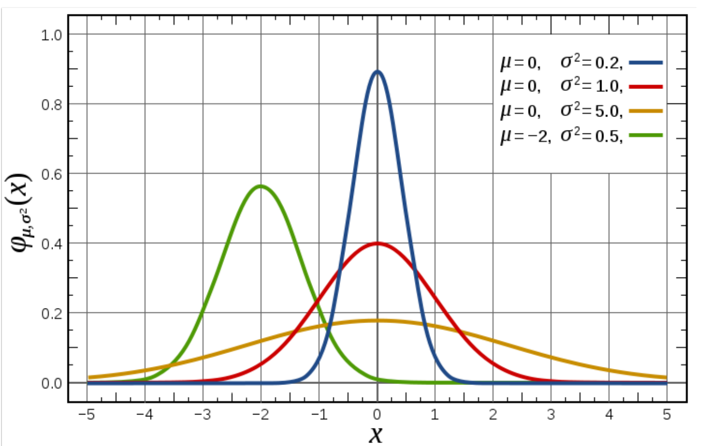
R: Per generar valors psudoaleatoris tenim funcions per a cada distribució: rbinom, rpois, runif, rexp ....

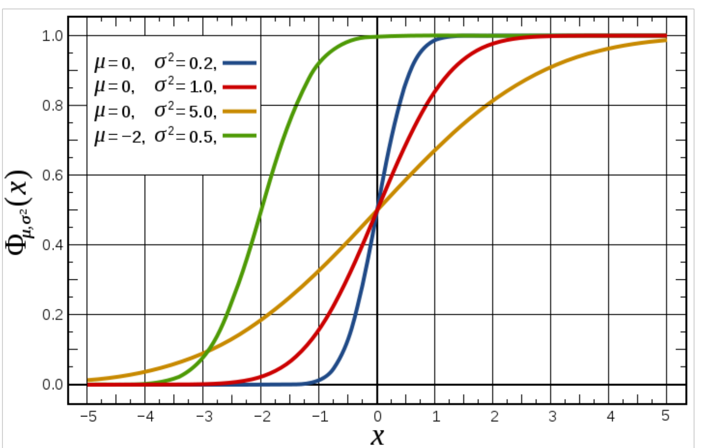
Model Normal (o de Gauss). Introducció

(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 s'ajusta a 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) [a vegades s'usa σ^2 en comptes de σ com a paràmetre]
- **Paràmetres**: μ (esperança), σ (desviació estàndard)
- **Funció de densitat**:

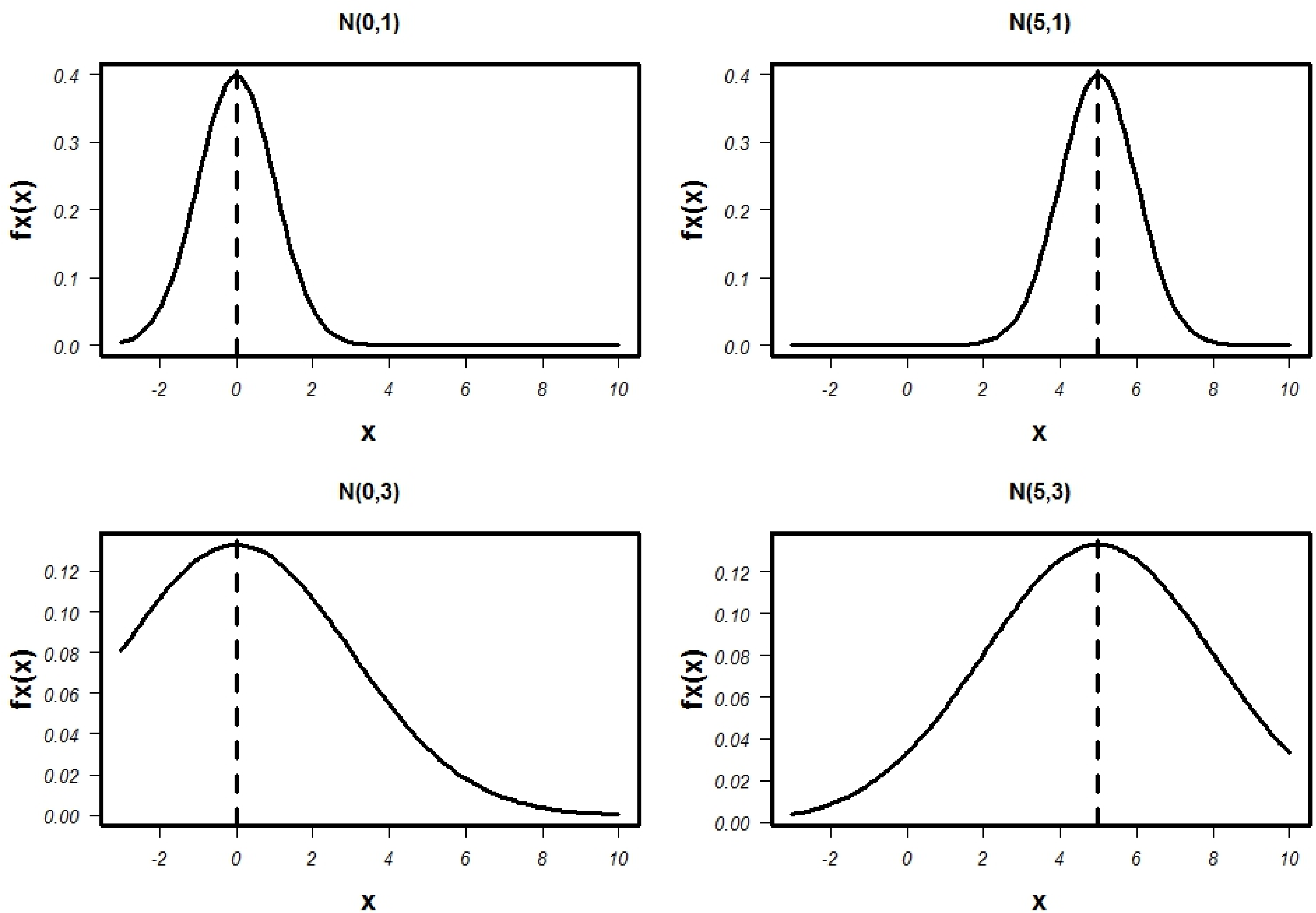
$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad amb \quad x \in R$$

- **La funció de distribució** no té expressió analítica → Taules
- **Indicadors:**
  - E(X) = μ
  - V(X) = σ^2

La Normal amb paràmetres μ = 0 i σ=1 s'anomena **Normal estàndard** i és la que apareix a les taules

Model Normal. Representació gràfica

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



Model Normal. Ex. 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

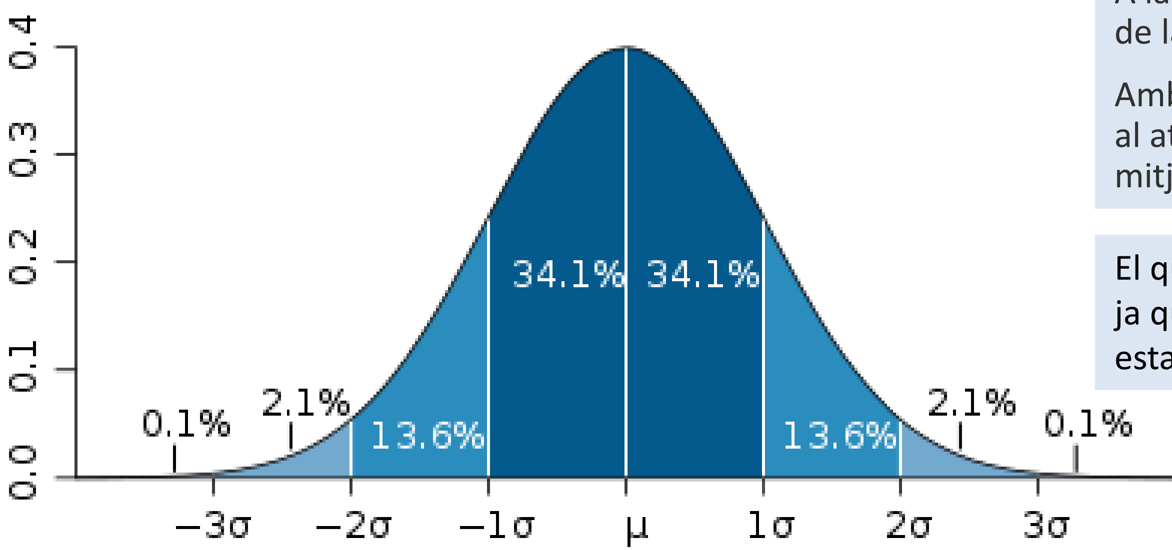
- **Probabilitat acumulada**. Quina és la probabilitat de 2 o menys?
  - Amb fórmules →**No es pot fer!!!**
  - Amb taules →$P(X \leq 2) = \mathbf{0.977}$
  - Amb R → $P(X \leq 2) = pnorm(q = 2, \text{mean} = 0, sd = 1) = \mathbf{0.9772499}$

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

Nota: Només hi ha taules de la Normal estàndar, **N(0,1)**. En la resta de casos, s'ha **d'estandaritzar** per fer els càlculs amb taules

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.
- Els punts d'inflexió es troben a una desviació estàndar de la μ ($x=\mu-\sigma$ i $x=\mu+\sigma$)
- Els quantils de la Normal estàndar Z~N(0,1), normalment, es denoten amb z_p. El quantil z_p representa aquell valor tal que en una Normal estàndar 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 $Z_{0.975} = 1.96$ ja que s'utilitza molt en la inferència estadística

Model Normal. Estandarització

- La **combinació lineal de variables Normals** és Normal:
  - Sigui a i b, dos escalars i $X \sim N(\mu_X, \sigma_X) \rightarrow Y = a \cdot X + b \sim N(\mu_Y = a \cdot \mu_X + b, \sigma_Y = a \cdot \sigma_1)$
  - Sigui a i b, dos escalars, $X \sim N(\mu_X, \sigma_X)$ i $Y \sim N(\mu_Y, \sigma_Y) \rightarrow$

    $\rightarrow W = a \cdot X + b \cdot Y \sim N(\mu_W = a \cdot \mu_X + b \cdot \mu_Y, \sigma_W = \sqrt{a^2\sigma_1{}^2 + b^2\sigma_2{}^2 + 2 \cdot ab \cdot \rho_{XY} \cdot \sigma_1 \cdot \sigma_2}$

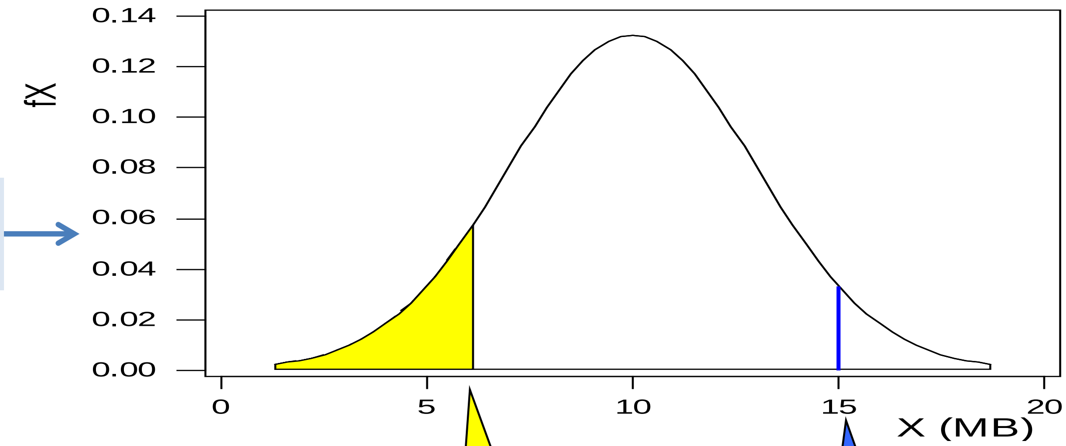
- Aquesta propietat permet relacionar distribucions Normals a base de translacions i escalars. En particular, transformar a la Normal estàndard $Z \sim N(0,1)$, **estandarditzar**, permet buscar en les taules de Z, probabilitats de qualsevol Normal

- 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:

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

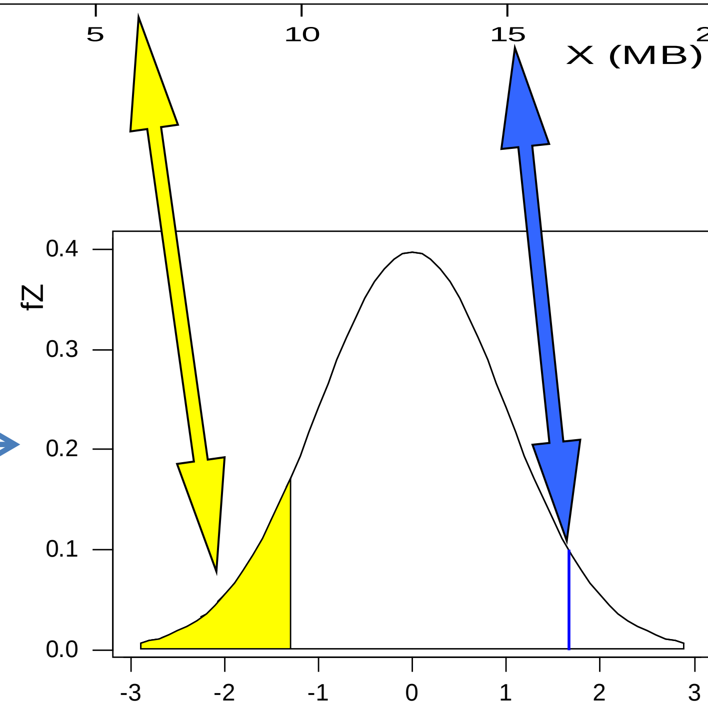
Model Normal. Estandarització



Variable X: situació real (unitats reals, per exemple MB)

$$Z = \frac{X - \mu}{\sigma}$$

Variable Z: situació estandarditzada, sense unitats, centrada en 0, dispersió tipificada (igual a la unitat)



Model Normal. Exemple d'estandarització

Exemple 1: [**probabilitats inferiors a 0.5**] . Sigui Z ~ N(0, 1). Quin és el valor *z* que deixa una probabilitat per sota de p = 0.25?

Com 0.25 < 0.5, no ve en taules, utilitzar simetries:

$$P(Z < -|z|) = P(Z > |z|) \to F_Z(0.6745) = 0.75 \to z = -0.6745$$

Exemple 2: [**Estandarització**] Sigui X: "Increment diari espai disc" ~ N(10 MB, 3 MB)

1) Quant val P(*X*>15)?

$$\mathrm{Z} = \frac{\mathrm{X} - 10}{3} \sim \mathrm{N}(0,1) \to \mathrm{P}(\mathrm{X} > 15) = \mathrm{P}\left(\mathrm{Z} > \frac{15 - 10}{3}\right) = \mathrm{P}\left(\mathrm{Z} > \frac{5}{3}\right) = 1 - \mathrm{P}\left(\mathrm{Z} \leq \frac{5}{3}\right)$$

$$= 1 - \mathrm{F_Z}\left(\frac{5}{3}\right) = 1 - 0.9522 = 0.0478$$

2) 1 de cada 10 dies, l'augment és inferior a quant? (Quant val *t* tal que P(*X* < *t*) = 0.1?)

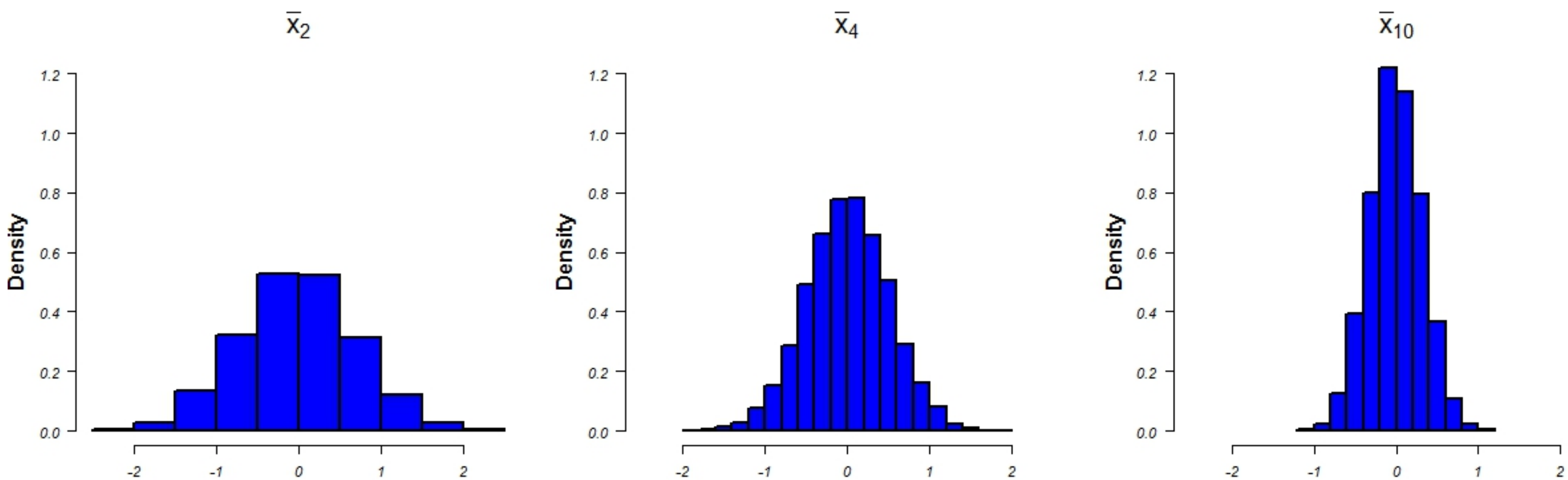
$$P(Z < t') = 0.1 \to t' = F_Z^{-1}(0.1) = -1.2816 \to t = 10 + 3 \cdot t' = 6.15\ MB$$

Distribució de la mitjana de v.a. Model Normal

- 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)

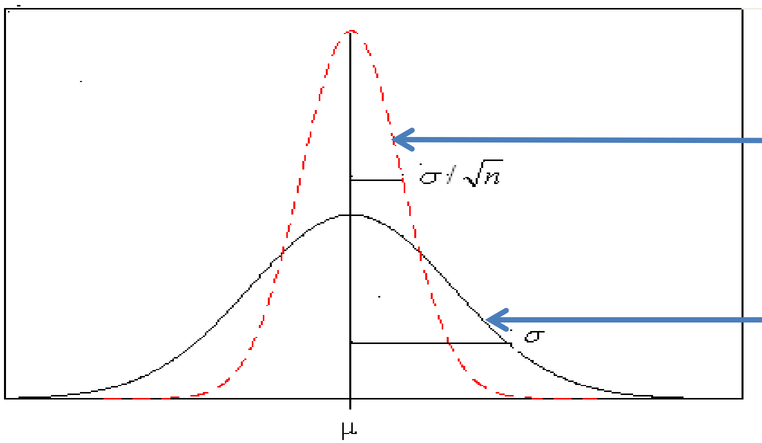
- Siguin X_1, X_2, …, X_n independents i idènticament distribuides amb esperança μ i desviació típica σ. Llavors:

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

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

- 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**.

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

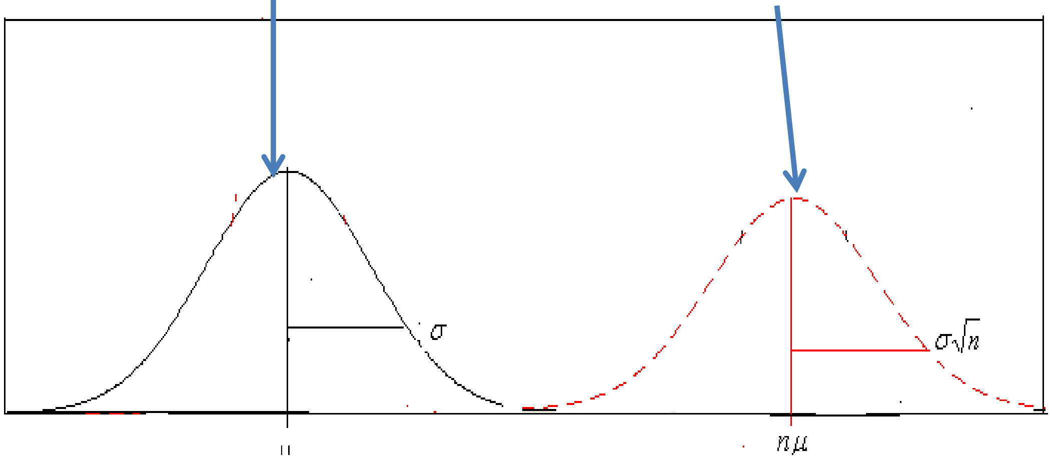
$$S_n = \sum X_i \xrightarrow[n\ gran]{} N(n\mu, \sigma\sqrt{n})$$

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

Només han de complir:

E(X) = μ

V(X) = σ^2



Teorema Central del Límit (TCL). Exemple

El treball de CPU per fer un backup presenta unes característiques diàries de mitjana 30'/dia i una desviació de 15'/dia. Si volem calcular probabilitats sobre el consum de CPU total mensual (suposant independència entre els 30 dies del mes), haurem de plantejar-nos la variable:

$S_n = X_1 + X_2 + \ldots + X_{30}$ (no ens donen la distrib. de X_i, sols μ i σ)

1) Quina és la distribució de S_n?

$$S_n = \sum X_i \xrightarrow[n\ gran]{} N(n\mu, \sigma\sqrt{n}) = N(900\,, 82.15)\ min = N(15{,}1.37)\ hores$$

2) Quina és la probabilitat d'un consum total mensual de més de 18 hores de CPU?

$$P(S_n > 18) = P\left(\frac{S_n - 15}{1.37} > \frac{18-15}{1.37}\right) = P(Z > 2.19) = 1 - p(Z < 2.19) = 1 - 0.9857 = 0.0143$$

[un de cada 70 mesos]

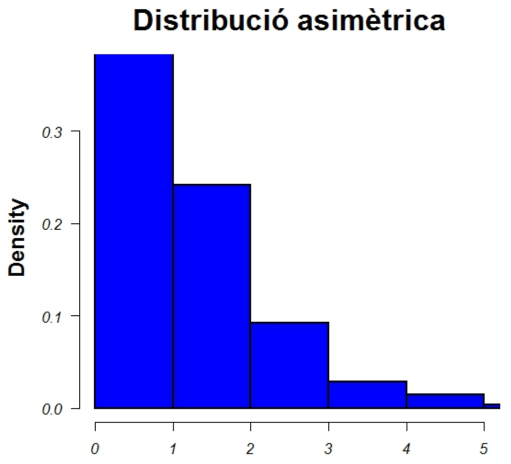
3) I la probabilitat de, en un mes, un consum mitjà diari inferior a 36'?

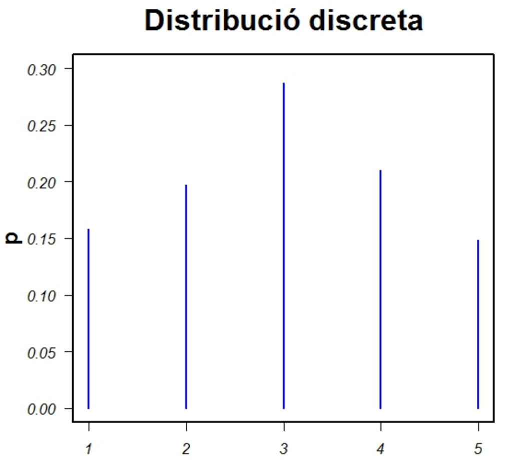
$$\bar{X}_{30} \sim N\left(\mu, \frac{\sigma}{\sqrt{30}}\right) = N(30\,, 2.74)\ min \rightarrow P(\bar{X}_{30} < 36) = P\left(Z < \frac{36-30}{2.74}\right) = P(Z < 2.19) = 0.9857$$

Teorema Central del Límit (TCL). ”n”

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

- Depèn de com sigui la distribució original i de que 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 agafen pocs valors):





- Aplicacions del TCL: la normal aproxima bé certes distribucions. [Exemple: variable de Poisson, si λ és gran. La t de Student, i la χ^2, que es veuran més endavant, també]

Teorema Central del Límit (TCL). Exercici

L'error de mesura del temps d'un procediment és normal amb σ = 1/4 seg. i mitjana 0. Es considera repetir les mesures de forma independent.

1) VAC d'una mesura de l'error: E_n: "error en la n-èssima mesura"
2) Model de la VAC: $E_n \sim N(\mu = 0, \sigma = 0.25)$
3) Prob. error menor de 0.1 s.:

$$P(|E_1| \leq 0.1) = P(-0.1 \leq E_1 \leq 0.1) = F_{E_1}(0.1) - F_{E_1}(-0.1) = 0.655 - 0.435 = 0.311$$

4) Error màxim (amb prob 0.95) en una mesura:

$$P(-e \leq E_1 \leq e) = 0.95 \ \rightarrow P(E_1 \leq e) = 0.975 \rightarrow e = 0.49\, s.$$

5) Idem per VAC de mitjana de 10 mesures:

$$X_{10} \sim N\left(\mu = 0, \sigma = \frac{0.25}{\sqrt{10}}\right) \rightarrow P(-f \leq X_{10} \leq f) = 0.95 \rightarrow f = 0.156 \text{ s.}$$

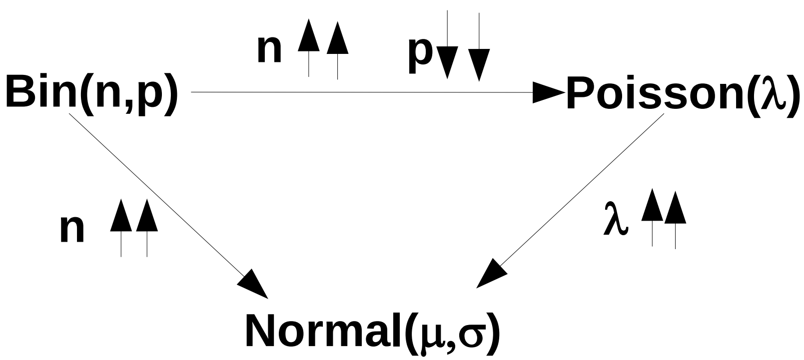
6) Número n mínim de mesures per tal que l'error màxim de la mitjana de les *n* mesures (amb prob 0.95) sigui inferior a 0.1 s. És a dir n tal que

$$P(-0.1 \leq X_n \leq 0.1) \geq 0.95 \rightarrow P\left(\frac{-0.1-0}{\frac{0.25}{\sqrt{n}}} \leq Z \leq \frac{0.1-0}{\frac{0.25}{\sqrt{n}}}\right) \rightarrow \frac{0.1}{\frac{0.25}{\sqrt{n}}} = 1.956 \rightarrow n = 24.01 \rightarrow n = 25$$

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** 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
  - $\mu = n \cdot p$
  - $\sigma^2 = n \cdot p \cdot (1-p)$
- La **distribció de Poisson** amb paràmetre *λ* es pot aproximar per una Normal quan la *λ* és prou gran. Llavors els paràmetres de la Normal són:
  - $\mu = \lambda$
  - $\sigma^2 = \lambda$



Nota: la Binomial es pot aproximar a una Poisson quan la *n* és prou gran i la *p* prou petita

TCL. Relacions entre distribucions. Exemple

S'ha comprovat que la probabilitat que el temps de resposta d'una determinada pàgina web al llarg d'un dia sigui inadequat és d'un 5%. Per calcular probabilitats del número de dies en que el servei és inadequat durant una setmana (7 dies), un mes (30 dies) o cinc anys (1825 dies), necessitem les següents variables aleatòries per les quals podem identificar el model més adequat:

X_{set}="nombre de dies en 1 setmana amb servei inadequat"

X_{set}~Bin(n=7,p=0.05)

X_{mes}="nombre de dies en 1 mes amb servei inadequat"

X_{mes} ~ Bin(n=30,p=0.05)

X_{mes} ~ P(λ=1.5)

X_{5anys}="nombre de dies en 5 anys amb servei inadequat"

X_{5anys} ~ Bin(n=1825, p=0.05)

X_{5anys} ~ P(λ=91.25)

X_{5anys} ~ N(μ=91.25, σ=9.55) [Encara que p és petita, n és molt gran]

TCL. Exercici

Suposem que s’ha establert que el pes del equipatge d’un viatger segueix una distribució Normal amb mitjana 18.9 Kg. i desviació 4 Kg. Habitualment, si el equipatge d’un viatger sobrepassa els 20 Kg., llavors té un sobrepreu que depèn de l’excés de pes. Contesta les següents qüestions:

1. Troba la probabilitat que un viatger hagi de pagar sobrepreu per excedir el seu equipatge els 20 Kg. de pes. 0.392 [1-pnorm(20,18.9,4)]

2. Quin és el pes que podem assegurar , amb un 95% de probabilitat , que un equipatge no superarà? 25.479 [qnorm(0.95,18.9,4)]

3. Indica els paràmetres (μ , σ) de la variable pes <u>total</u> de 10 equipatges

$N(\mu = 189, \sigma = 4\sqrt{10} = 12.649)$

4. Indica els paràmetres (μ , σ) de la variable pes <u>promig</u> de 10 equipatges

$N(\mu = 18.9, \sigma = 4/\sqrt{10} = 1.265)$

TAULA resum de tots els models rellevants

Distribució	Declaració	Funció de probabilitat o de densitat	Funció distribució	Esperança E(X)	Variància V(X)
Bernoulli	X~Bern(p)	$P_X(k)=\begin{cases} q & k=0 \\ p & k=1 \end{cases}$	$F_X(k)=\sum_{i<=k} P_X(i)$	p	$p\cdot q$
Binomial	X~B(n,p)	$P_X(k)=\binom{n}{k}\cdot p^k\cdot q^{n-k}$ $k=0,1,\ldots,n$ **(R:dbinom(k,n,p))**	$F_X(k)=\sum_{i<=k} P_X(i)$ (taules estadístiques) **(R:pbinom(k,n,p))**	$p\cdot n$	$p\cdot q\cdot n$
Poisson	X~P(λ) *	$P_X(k)=\frac{\lambda^k\cdot e^{-\lambda}}{k!}\quad k=0,1,2,\ldots$ **(R:dpois(k,λ))**	$F_X(k)=\sum_{i<=k} P_X(i)$ (taules estadístiques) **(R:ppois(k,λ))**	λ	λ
Exponencial	X~Exp(λ) *	$f_X(x)=\lambda\cdot e^{-\lambda\cdot x}\quad x>0$ **(R:dexp(x,λ))**	$F_X(x)=1-e^{-\lambda\cdot x}$ **(R:pexp(x,λ))**	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Uniforme	X~U[a,b]	$f_X(x)=\frac{1}{b-a}\quad a<x<b$ **(R:dunif(k,a,b))**	$F_X(x)=\frac{x-a}{b-a}$ **(R:punif(k,a,b))**	$\frac{(a+b)}{2}$	$(b-a)^2/12$
Normal	X~N(μ,σ)	$f_X(x)=\frac{1}{\sigma\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$ **(R:dnorm(k,μ,σ))**	$F_X(x)=?$ (taules estadístiques N(0,1)) **(R:pnorm(k,μ,σ))**	μ	σ^2

$0<p<1$ $\quad q=1-p$ $\quad n\in\mathbb{N}$ $\quad r\in\mathbb{N}$ $\quad \lambda\in\Re^+$

*λ paràmetre del procés Poisson: variables Poisson i Exponencial