INFERÈNCIA ESTADÍSTICA APLICADA

Victòria Alea
Montserrat Guillén
Carme Muñoz
Elizabeth Torrelles
Núria Viladomiu

INTRODUCCIÓ

Aquesta nova edició electrònica d'aquest llibre d'estadística per a les ciències econòmiques, empresarials i socials conté els temes fonamentals que es desenvolupen en un curs de grau a la Universitat.

Tota persona interessada en conèixer les possibilitats que ofereix el tractament de la informació estadística en té prou amb un curs de descriptiva per poder veure com és possible treure conclusions interessants a partir de l'anàlisi de dades.

> Exemples:
>
> Un empresari pot veure l'evolució de les seves vendes al llarg d'un període de temps i treure algunes conclusions només observant l'aparença d'una gràfica. L'investigador en ciències socials pot utilitzar una mitjana aritmètica per a mesurar la valoració pública que mereix un determinat líder polític.

En un primer curs d'estadística es poden veure exemples com els que s'acaben d'esmentar i s'aprèn a utilitzar una metodologia relativament senzilla. L'estudi de l'estadística descriptiva, obre les portes a un seguit de temes d'una qualitat més profunda, la seva fonamentació teòrica i també la inferència estadística.

Sovint, en un segon curs d'estadística es volen resoldre qüestions de caire més avançat.

> Exemples:
>
> L'empresari pot voler decidir si la tendència de les vendes, malgrat les fluctuacions que s'hi aprecien, és creixent. Per altra banda, un sociòleg o un estudiós de la ciència política pot voler contrastar si el baròmetre d'opinió sobre un líder polític ha experimentat un canvi des de l'última enquesta.

Ambdues qüestions, i moltes d'altres de similars, requereixen un coneixement més avançat de les tècniques estadístiques. Cal establir primerament alguns conceptes fonamentals de la teoria de la probabilitat abans de passar als temes pròpiament d'estimació de magnituds i de contrast.

L'objectiu d'aquest llibre és precisament aquest. Pressuposant un coneixement previ d'estadística, especialment en la vessant descriptiva, es tracten temes més

avançats per tal de conèixer els conceptes fonamentals de la inferència estadística.

S'entén com a **inferència estadística** el pas que va més enllà de la descripció de les dades, i permet analitzar estructures de dades sobre una població i extreure conclusions sobre la mateixa.

Al llarg de la formació de grau universitari, les matèries tracten d'introduir a l'estudiant en els continguts bàsics de la seva titulació. Sovint es parla de models de comportament teòrics i serà a partir de l'anàlisi estadística i del coneixement empíric de la realitat que es podran establir resultats més aplicats a cada situació concreta. Per aquesta raó l'estadística és una matèria fonamental i obligatòria en la trajectòria curricular de la majoria d'ensenyaments.

Les autores hem entès aquest llibre com un manual de consulta. Intencionadament hem exclòs referències a exercicis realitzats amb un ordinador. Com que el contingut és predominantment de caire teòric s'han anat introduint exemples al llarg del text i exercicis al final de cada capítol. S'ha optat per un text més clàssic, amb el desig que pugui ser un material d'estudi i no pas de pràctica. Un dels aspectes més útils del present llibre és el tractament sintètic que realitza de cadascun dels conceptes. Habitualment, en una mateixa pàgina es troba la definició del concepte, les seves principals propietats o característiques i seguidament un exemple o gràfic que ajudi a la comprensió del mateix. El text no es fa excessivament retòric i les fórmules més importants s'han destacat amb petits quadres. Igualment, s'inclouen petits quadres sintètics com a resum dels procediments més importants que cal retenir.

El fet que el manual estigui disponible en català permet que es configuri com una de les escasses referències bibliogràfiques sobre la matèria en aquesta llengua i l'única de les seves característiques.

Esperem que la presentació li sigui suficientment amena i mostri l'aplicabilitat de l'estadística en les seves diverses vessants. Esperem que aquesta experiència sigui apassionant.

Les autores volen agrair a la Universitat de Barcelona que va donar el seu suport a la realització d'aquest projecte a través del Gabinet d'Avaluació i Innovació Universitària. Amb l'edició d'aquest manual,

Esperem que aquesta nova edició del llibre sigui d'utilitat per a tota persona interessada en avançar-se una mica més en la matèria.

Barcelona, juliol de 2017.

ÍNDEX

CAPÍTOL I. PROBABILITAT I VARIABLE ALEATÒRIA

1.1 INTRODUCCIÓ

Aquest capítol està dividit en dues parts: a la primera s'efectua una aproximació formal dels principis matemàtics de la ***Teoria de la probabilitat*** i, a la segona, s'introdueix el concepte de ***Variable aleatòria*** amb l'objectiu de poder presentar, al capítol següent, els models matemàtics de probabilitat o ***Distribucions de Probabilitat*** més importants.

1.2 ESTADÍSTICA I PROBABILITAT

L'estadística es pot definir com la ciència empírica que estudia els fenòmens que depenen de l'atzar. Aquests fenòmens aleatoris estan associats a experiments que es poden repetir de forma il·limitada i presenten resultats imprevisibles encara que es realitzin en les mateixes condicions. Els experiments aleatoris, malgrat aquests resultats imprevisibles, es caracteritzen perquè presenten una pauta de comportament o regularitat estadística a llarg termini que, com es veurà, pot, generalment, modelitzar-se mitjançant algun dels models de probabilitat que s'estudien més endavant.

Tots els possibles resultats d'un experiment aleatori formen un conjunt que s'anomena *espai mostral* o espai referencial (E). L'espai mostral pot ser finit, infinit o infinit numerable.

Qualsevol subconjunt de resultats de l'espai mostral rep el nom de *succés* (A, B). Un succés pot ser: *simple* si està format per un únic resultat elemental; *compost* si està format per diferents resultats; *cert* si es verifica per a qualsevol resultat de l'espai mostral (E); i *fals* o *nul* si no pot esdevenir mai, és a dir, si no conté cap dels resultats de l'espai mostral. També són successos aleatoris les operacions realitzades entres successos: la *unió* ($A \cup B$) és el succés format per tots els resultats de A, de B o d'ambdós; la *intersecció* ($A \cap B$) és el succés que recull els resultats comuns de A i de B; aleshores diem que dos successos A i B són *mútuament excloents* o incompatibles quan no tenen cap resultat en comú i, per tant, la seva intersecció és el conjunt nul; el *complementari* d'un succés A dins de l'espai mostral E, que representem per $\overline{A}$, és el succés que conté tots els resultats de E que no estan inclosos en A.

El concepte de ***probabilitat*** permet quantificar la incertesa que acompanya els fenòmens aleatoris ja que assigna a cada succés un indicador de la possibilitat o versemblança de la seva ocurrència. És a dir, la Probabilitat és un nombre real que permet mesurar la possibilitat d'ocurrència d'un succés dins de l'espai

mostral. L'assignació d'aquest valor es pot determinar aplicant algun dels tres criteris següents:

La ***teoria clàssica*** de probabilitat, o regla De Laplace (1789-1827), planteja que, donat un experiment aleatori amb n resultats elementals igualment probables (equiprobables) i mútuament excloents, la probabilitat d'un succés A format per n_A resultats elementals és igual al quocient entre n_A (casos favorables) i n (casos possibles).

$$P(A) = \frac{n_A}{n} = \frac{Nombre\ de\ casos\ favorables}{Total\ de\ casos\ possibles}$$

Per poder aplicar la regla de Laplace és necessari que el nombre de resultats elementals del fenomen aleatori sigui finit i que tots ells tinguin la mateixa probabilitat d'ocórrer (equiprobables). Aquestes limitacions van portar a la determinació de la ***teoria freqüèncialista*** de la probabilitat que va establir formalment Von Mises (1881-1953). Aquesta teoria es fonamenta en la regularitat estadística, és a dir, en la idea que donat un experiment aleatori que es pugui repetir moltes vegades, en condicions aproximadament iguals, la freqüència relativa dels resultats s'aproxima a la corresponent probabilitat a mesura que augmenta el nombre de repeticions.

Així doncs, la probabilitat d'un succés A serà igual al valor al que convergeix la seva freqüència relativa ($\frac{n_A}{N}$) quan s'incrementa infinitament el nombre de repeticions (observacions).

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}$$

Per últim, la necessitat d'assignar probabilitats en situacions en què és impossible la repetició de l'experiment va generar l'aparició de la tercera concepció de probabilitat o ***teoria subjectiva*** (també coneguda com teoria Bayesiana). Aquesta teoria interpreta la probabilitat com el grau de convenciment subjectiu que cada individu pot tenir respecte a l'ocurrència d'un determinat succés. Per tant, aquesta última regla d'assignació de probabilitats, a diferència de les anteriors que són totalment objectives, depèn del criteri personal sobre el fenomen aleatori.

1.3 AXIOMÀTICA DE KOLMOGOROV

La teoria matemàtica de la probabilitat desenvolupada per ***Kolmogorov*** (1903-1987) es basa en els següents axiomes.

Donat un espai mostral (o de referència) E i qualsevol succés A de E diem que P és una funció de probabilitat definida a l'espai mostral E, de forma que a cada succés A se li fa correspondre un nombre real P(A) que mesura la seva possibilitat d'ocurrència, sempre que aquesta funció compleixi els següents axiomes:

Axioma I: $P(A) \geq 0$. Si A és un succés que pertany a un espai mostral E, existeix un nombre real P(A) superior o igual a zero que denominem probabilitat.

Axioma II: $P(E) = 1$. La probabilitat del succés cert o màxima és igual a 1.

Axioma III: Si A_1, A_2, A_3 ... és una successió numerable de successos mútuament excloents ($A_i \cap A_j = \varnothing \ \forall A_i \neq A_j$), aleshores

$P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A_1) + P(A_2) + P(A_3) + ...$

El primer axioma reflecteix la idea intuïtiva que la possibilitat que esdevingui qualsevol resultat ha de ser mesurada per un nombre positiu.

El segon planteja que el màxim possible de probabilitat o la probabilitat associada a l'esdeveniment cert és igual a 1.

I el tercer expressa que la unió d'un conjunt de resultats que no poden esdevenir simultàniament (mútuament excloents) té una probabilitat igual a la suma de les probabilitats individuals de cadascun dels resultats considerats.

Qualsevol regla d'assignació de probabilitat ha de ser, per una banda, consistent amb la idea de possibilitat o versemblança del resultat, però, a més, ha de complir els tres axiomes de Kolmogorov. En aquest sentit, es pot comprovar que qualsevol dels criteris d'assignació anteriors són consistents amb aquestes dues premisses.

Dels tres axiomes es dedueixen els següents teoremes:

Teorema I: $P(\varnothing) = 0$. La probabilitat del succés fals o impossible és zero.

Teorema II: $0 \leq P(A) \leq 1 \ \forall A \in E$. La probabilitat de qualsevol succés és un nombre real positiu inferior o igual a 1.

Teorema III: $P(\overline{A}) = 1 - P(A)$. La probabilitat que no es doni A és igual a 1-P(A).

Teorema IV: Si $A \subset B \Rightarrow P(A) \leq P(B)$. Si un succés A està inclòs dins d'un altre succés B, aleshores la probabilitat de A és com a màxim la de B.

Teorema V: Llei Additiva. P(A∪B) = P(A) + P(B) - P(A∩B). La probabilitat que es doni el succés A o el succés B, o ambdós, és igual a la suma de les seves corresponents probabilitats menys la probabilitat d'ocurrència simultània de A i B. En el cas particular de successos mútuament excloents, la probabilitat de la unió és la suma de probabilitats.

Aquest teorema es pot generalitzar per a més de dos successos. En concret per a n successos qualssevol queda:

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{n} P(A_i \cap A_j) + \sum_{\substack{i,j,k=1 \\ i<j<k}}^{n} P(A_i \cap A_j \cap A_k) - \ldots (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n)$$

Si n=3 $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

1.4 PROBABILITAT CONDICIONADA. TEOREMA DE LA INTERSECCIÓ

Sovint en avaluar la probabilitat d'algun succés es disposa d'alguna informació al respecte. Per exemple, suposem que es vol determinar la probabilitat que un alumne d'ADE, triat a l'atzar, tingui pendent d'aprovar l'assignatura d'Estadística. (Amb el mètode freqüencialista la probabilitat serà, aproximadament, igual al quocient entre el nombre d'alumnes que encara no han aprovat i el nombre total d'alumnes matriculats a ADE.) Si tenim informació addicional respecte a l'alumne triat a l'atzar, per exemple, que aquest és el primer any que es matricula i, per tant, és segur que no té aprovada l'estadística, queda clar que la probabilitat que ens interessa queda modificada i, en aquest cas, serà igual a 1.

La informació addicional sobre el succés modifica (en general redueix) l'espai de referència i, sobre el nou espai, s'obté una probabilitat que rep el nom de condicionada**.**

Definició:

Donats dos successos A i B d'un mateix espai tal que P(B)≠0, la ***probabilitat condicionada de A*** sabent que s'ha esdevingut B, que s'indica P(A/B), és igual a la probabilitat conjunta d'A i B dividida per la probabilitat marginal de B.

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

Exemple 1.1

D'un producte que presenta una probabilitat 0,6 que el comprin, se sap que les probabilitats conjuntes de comprar-lo o no, havent vist o no la seva campanya publicitària, són: probabilitat de comprar i haver vist la publicitat 0,5, i probabilitat de no comprar i no haver vist la publicitat 0,2. Quina és la probabilitat que un client hagi comprat si no ha vist la publicitat?

Solució:

Sigui C = {Comprar} $\overline{C}$ = {No comprar} P = {Haver vist la publicitat} $\overline{P}$ = {No haver-la vist}

Sabem: P(C) = 0,6 P(C∩P) = 0,5 P($\overline{C} \cap \overline{P}$) = 0,2

Podem recollir aquesta informació en un quadre de doble entrada i calcular totes les probabilitats conjuntes i marginals (en negreta dades originals):

	C	$\overline{C}$	
P	**0,5**	0,2	0,7
$\overline{P}$	0,1	**0,2**	0,3
	0,6	0,4	**1**

Per obtenir P(C/$\overline{P}$) apliquem la definició de probabilitat condicionada:

$$P(C/\overline{P}) = \frac{P(C \cap \overline{P})}{P(\overline{P})} = \frac{0{,}1}{0{,}3} = 1/3$$

La probabilitat condicionada compleix els axiomes i teoremes de la probabilitat. Així doncs, si P(B) és diferent de zero:

1. P(A/B) ≥ 0
2. P(E/B) = 1
3. Si A_1, A_2, ... és una successió numerable de successos mútuament excloents ($A_i \cap A_j = \varnothing \ \forall A_i \neq A_j$) aleshores P($A_1 \cup A_2 \cup$ .... /B)= P(A_1/B)+P(A_2/B)+ ....
4. P(∅/B) = 0
5. P($\overline{A}$/B) = 1- P(A/B)
6. Llei Additiva: P[(A∪C)/B] = P(A/B) + P(C/B) - P[(A∩C)/B]

De la definició de probabilitat condicionada es dedueix la llei multiplicativa de probabilitats o ***teorema de la intersecció***. Aquest teorema planteja que la probabilitat conjunta de dos successos A i B, amb P(B)≠0, és igual a la probabilitat marginal de B multiplicada per la probabilitat condicionada de A sota el supòsit que B s'hagi esdevingut.

$$P(A \cap B) = P(B)\, P(A/B)$$

Vegeu que la probabilitat de la intersecció de dos successos de probabilitat no nul·la es pot expressar de dues formes:

$$P(A \cap B) = \begin{cases} P(A)\,P(B/A) \\ \\ P(B)\,P(A/B) \end{cases}$$

El teorema es pot generalitzar per a n successos:

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1)\, P(A_2/A_1)\, P(A_3/A_1 \cap A_2) \ldots P(A_n/\bigcap_{i=1}^{n-1} A_i)$$

Exemple 1.2

D'un arxivador amb 15 factures es sap que només 3 presenten algun error. Si s'han triat a l'atzar dues factures, quin és l'espai mostral dels resultats, i quina és la probabilitat que com a màxim una factura presenti algun error?

Solució:

F = {*Extreure una factura amb algun error*}

$\bar{F}$ = {*Extreure una factura sense errors*}

P(F) = 3/15 = 0,2 P($\bar{F}$) = 1 - P(F) = 0,8

Espai Mostral = { $F\,F,\ \bar{F}\,F,\ F\,\bar{F},\bar{F}\,\bar{F}$ }

La llei multiplicativa permet calcular les probabilitats d'aquests resultats elementals.

P($F\,F$) = P(F) P(F/F) = 3/15·2/14 = 6/210=1/35

P($F\,\bar{F}$) = P(F) P($\bar{F}/F$) = 3/15·12/14 = 36/210 = 6/35

P($\bar{F}\,F$) = P($\bar{F}$) P($F/\bar{F}$) = 12/15·3/14 = 36/210 = 6/35

P($\bar{F}\,\bar{F}$) = P($\bar{F}$) P($\bar{F}/\bar{F}$) = 12/15·11/14 = 132/210 = 22/35

b) A = {*Obtenir com a màxim una factura amb error*}={ $\bar{F}\,F,\ F\,\bar{F},\ \bar{F}\,\bar{F}$ }

$\overline{A}$ = {*Obtenir dues factures amb error*} = { $F\,F$ }

P(A) = 1 - P($\overline{A}$) = 1 - P($F\,F$) = 1 - 1/35 = 34/35

Exemple 1.3

Una prova de selecció de personal consta de dues parts: la primera consisteix en un test psicotècnic i la segona en una entrevista personal. Si de 100 persones

presentades el 60% supera el test i d'aquestes últimes el 30% supera l'entrevista:

a) Quina és la probabilitat que una d'aquestes 100 persones, escollida a l'atzar, hagi superat les dues proves?

b) Quina és la probabilitat que no hagi superat l'entrevista però hagi superat el test?

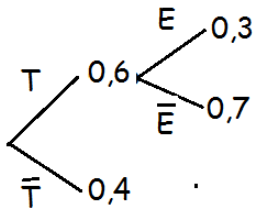
Solució:

T = {*Aprovar el test*} $\overline{T}$ = {*No aprovar el test*}

E = {*Superar l'entrevista*} $\overline{E}$ = {*No superar-la*}

Se sap que P(T) = 0,60 i P(E/T) = 0,30.

Gràfic 1.1 Diagrama d'arbre



a) P(T∩E) = P(T) P(E/T) = 0,6·0,3 = 0,18

b) P(T∩$\overline{E}$) = P(T) P($\overline{E}$/T) = 0,6·0,7 = 0,42

1.4.1 INDEPENDÈNCIA DE SUCCESSOS

En general, quan es parla de probabilitat condicionada implícitament es suposa que els successos són dependents i, per tant, la probabilitat d'un d'ells queda modificada quan succeeix l'altre. Si, pel contrari, entre dos successos A i B, amb probabilitats no nul·les, l'ocurrència o no de qualsevol d'ells no condiciona (no modifica) la probabilitat de l'altre diem que A i B són estocàsticament independents.

$$P(A/B) = P(A)$$
$$P(B/A) = P(B)$$

Per exemple, suposem un joc de cartes que consisteix en esbrinar la carta que s'extraurà aleatòriament. Si apostem per un as la probabilitat de guanyar és P(As)=4/48. Si una vegada extreta la carta ens diuen que ha resultat una copa, voldrem canviar l'aposta? La nova probabilitat condicionada que presenta la nostra aposta és: P(As/Copa) = 1/12, i, en conseqüència, la nostra resposta raonable serà que no volem canviar l'aposta, ja que la probabilitat d'obtenir un as

no queda modificada per la informació addicional. Per tant, el successos anteriors ('*treure as*' i '*treure copa*') són estocàsticament independents.

Definició:

Dos successos A i B són ***estocàsticament independents*** si, i només si, compleixen:

$$P(A \cap B) = P(A)\,P(B)$$

$P(A \cap B) = P(A)\,P(B/A) = P(A)\,P(B)$ ja que $P(B) = P(B/A)$

$P(A \cap B) = P(B)\,P(A/B) = P(B)\,P(A)$ ja que $P(A) = P(A/B)$

Propietats:

1. Si P(A) = 0, A és estocàsticament independent de qualsevol succés B.
2. Si P(A) = 1, A és estocàsticament independent de qualsevol succés B.
3. Si dos successos amb probabilitats no nul·les són independents aleshores mai poden ser mútuament excloents.
4. Si A i B són dos successos estocàsticament independents, aleshores també ho són: $\overline{A}$ i B; A i $\overline{B}$; $\overline{A}$ i $\overline{B}$.

1.4.2 TEOREMA DE LA PROBABILITAT TOTAL

El ***teorema de la probabilitat total*** permet calcular la probabilitat d'un succés B, P(B), a partir de la seva probabilitat condicionada a un conjunt de successos, A_1, A_2,..., A_n, que formen una partició de l'espai mostral.

Es diu que un conjunt de successos A_1, A_2,..., A_n formen una partició de E si, i només si, compleixen les següents condicions:

- Exhaustius: $A_1 \cup A_2 \cup ... \cup A_n = E$, ó $\bigcup_{i=1}^{n} A_i = E$
- Mútuament excloents: $A_i \cap A_j = \varnothing \ \forall i \neq j$.

Si les probabilitats dels successos de la partició, $P(A_1)$, $P(A_2)$,..., $P(A_n)$ i les probabilitats del succés B (on $B \subset E$) condicionades a cada A_i, és a dir, $P(B/A_1)$, $P(B/A_2)$,..., $P(B/A_n)$ són conegudes llavors es pot obtenir la probabilitat de B, P(B), a l'espai mostral E.

La probabilitat del succés B és:

$P(B) = P(B \cap E) = P(B \cap (A_1 \cup A_2 \cup ... \cup A_n)) = P((B \cap A_1) \cup (B \cap A_2) \cup ... \cup (B \cap A_n))$

$P(B) = P(B \cap A_1) + P(B \cap A_2) + ... + P(B \cap A_n) =$

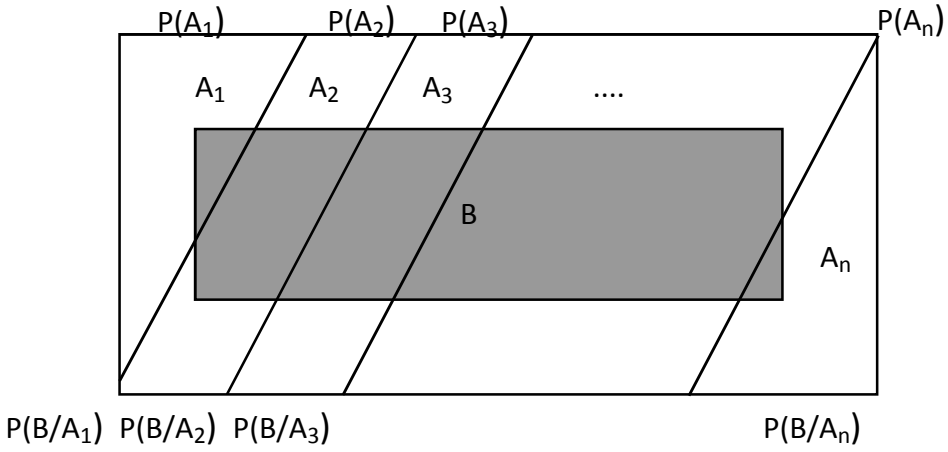
$$= P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + ... + P(A_n)\,P(B/A_n)$$

Per tant queda,

$$P(B) = \sum_{i=1}^{n} P(B/A_i)\, P(A_i)$$

expressió del ***Teorema de la probabilitat total***.

Gràfic 1.2 Partició d'E



Exemple 1.4

La producció total d'una empresa s'obté de 4 tipus de màquines. La probabilitat que una peça fabricada procedeixi del primer tipus de màquina és 0,4 i la probabilitat que provingui de les màquines 2, 3 i 4 és 0,2, 0,35 i 0,05, respectivament. Si la probabilitat de fabricar peces defectuoses per als quatre tipus de màquines anteriors és 0,05, 0,02, 0,08 i 0,01, respectivament, quina és la probabilitat que una peça triada a l'atzar sigui defectuosa?

<u>*Solució:*</u>

A_1 ={*producció de la màquina 1*} $P(A_1) = 0{,}4$

A_2 = {*producció de la màquina 2*} $P(A_2) = 0{,}2$

A_3 = {*producció de la màquina 3*} $P(A_3) = 0{,}35$

A_4 = {*producció de la màquina 4*} $P(A_4) = 0{,}05$

Els successos A_1, A_2, A_3 i A_4 defineixen una partició de E ja que són mútuament excloents i la seva unió forma l'espai mostral. És a dir, compleixen que:

- $P(A_1 \cap A_2) = 0$ $P(A_1 \cap A_3) = 0$ ... Si una peça ha estat fabricada per la màquina 1, per exemple, no pot, simultàniament, haver estat fabricada per cap altra

màquina.

- $P(A_1)+P(A_2)+P(A_3)+P(A_4) = 1$ La producció total procedeix d'alguna d'aquestes quatre màquines.

Es defineix el succés D = {*obtenir peces defectuoses*}

Es sap que als subespais A_1, A_2, A_3 i A_4 poden donar-se peces defectuoses amb les probabilitats:

$$P(D/A_1)=0,05 \quad P(D/A_2)=0,02 \quad P(D/A_3)=0,08 \quad P(D/A_4)=0,01$$

Com que $A_1,...,A_4$ formen una partició de l'espai mostral, la probabilitat que una peça triada a l'atzar sigui defectuosa és:

$$P(D) = P(A_1 \cap D) + P(A_2 \cap D) + P(A_3 \cap D) + P(A_4 \cap D)$$

I les probabilitats de la intersecció les podem expressar en termes ja quantificats:

$P(A_1 \cap D) = P(D/A_1) \cdot P(A_1) = 0,05\ 0,4 = 0,02$

$P(A_2 \cap D) = 0,02\ 0,2 = 0,004$

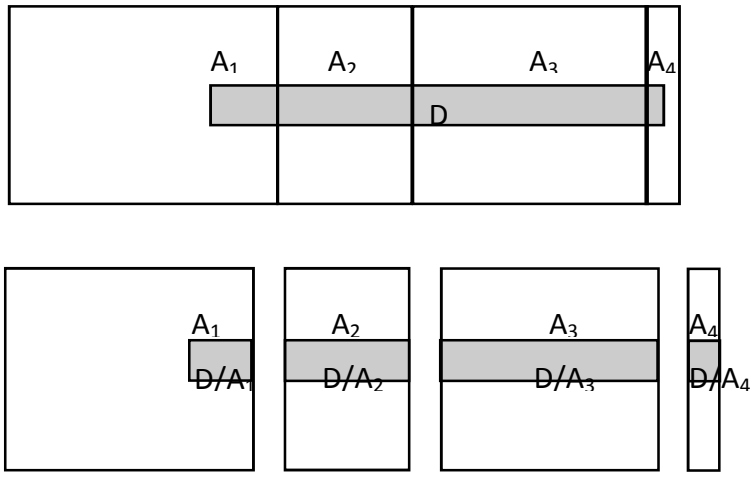
$P(A_3 \cap D) = 0,08\ 0,35 = 0,028$

$P(A_4 \cap D) = 0,01\ 0,05 = 0,0005$

Finalment queda,

$P(D) = 0,02 + 0,004 + 0,028 + 0,0005 = 0,0525$

Gràfic 1.3. Probabilitat total de D.



1.4.3 TEOREMA DE BAYES

Aquest teorema, formulat per Thomas Bayes (1702-1761), ha estat clau en el desenvolupament d'una nova concepció de la Inferència Estadística, l'Estadística Bayesiana. Es basa en el fet que una informació addicional pot modificar la

probabilitat inicial assignada a un determinat succés i, per tant, permet adaptar amb més exactitud la valoració sobre la creença d'ocurrència d'un determinat esdeveniment a mesura que es disposa de més informació.

Assignades unes probabilitats inicials $P(A_1)$, $P(A_2)$,..., $P(A_n)$, o probabilitats *a priori* (que reflecteixen el grau de creença sobre l'ocurrència de A_1, A_2,..., A_n, abans de realitzar un determinat experiment), on els successos A_1, A_2,..., A_n formen una partició de E, es realitza un experiment dins d'aquest espai de referència i s'obté un succés B. Aquest es valora a partir de la seva ocurrència en els successos A_i, és a dir, es calculen les probabilitats condicionades de B a cada succés A_i de la partició, $P(B/A_i)$. Aquesta evidència experimental permet obtenir les noves probabilitats dels A_i condicionades al succés B, $P(A_i/B)$, que es denominen probabilitats *a posteriori* i que reflecteixen el grau de creença corregit sobre l'ocurrència dels successos A_1, A_2,..., A_n donada una evidència experimental.

En efecte:

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B/A_i)}{\sum_{j=1}^{n} P(A_j)P(B/A_j)}$$

Per tant, el ***teorema de Bayes*** formula que:

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{P(B)}$$

és a dir, la probabilitat *a posteriori* és igual a la probabilitat *a priori* multiplicada per un factor modificatiu que depèn del resultat de l'experiment.

Del Teorema de Bayes es pot concloure que:

- si $P(B/A_i) > P(B)$ aleshores $P(A_i/B) > P(A_i)$ Probabilitat *a posteriori* > Probabilitat *a priori*.
- si $P(B/A_i) < P(B)$ aleshores $P(A_i/B) < P(A_i)$ Probabilitat *a posteriori* < Probabilitat *a priori*.

Vegeu que quan els successos A_i i B són independents la probabilitat a priori de A_i no queda modificada per l'ocurrència de B:

$P(A_i/B) = P(A_i) \cdot \frac{P(B/A_i)}{P(B)} = P(A_i)$ ja que $P(B/A_i)=P(B)$ quan són independents.

Exemple 1.5

Un petit comerç que només té a la venda tres tipus de vídeo, V_1, V_2 i V_3, creu que les seves respectives probabilitats de venda són $P(V_1) = 0,5$, $P(V_2) = 0,3$ i

$P(V_3) = 0,2$. Les probabilitats d'avaria dels vídeos durant el període de garantia són 0,25, 0,2 i 0,1 per als tipus V_1, V_2 i V_3, respectivament. Un client torna un vídeo avariat durant el període de garantia i no consta de quin tipus n'és. Calculeu la probabilitat que el vídeo sigui del tipus V_1, V_2 o V_3.

<u>*Solució:*</u>

A = {*avaria*}

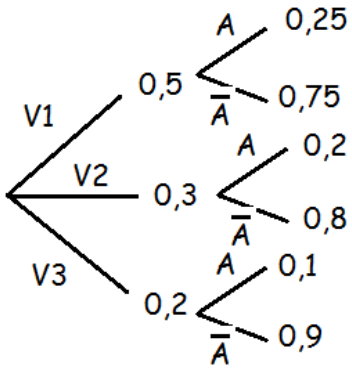
V_1 = {*vídeo tipus 1*} $P(V_1) = 0,5$ i $P(A/V_1) = 0,25$

V_2 = {*vídeo tipus 2*} $P(V_2) = 0,3$ i $P(A/V_2) = 0,2$

V_3 = {*vídeo tipus 3*} $P(V_3) = 0,2$ i $P(A/V_3) = 0,1$

Es volen obtenir les probabilitats condicionades: $P(V_1/A)$ $P(V_2/A)$ $P(V_3/A)$

Gràfic 1.4 Arbre de probabilitat.



Podem expressar $P(V_1/A) = P(V_1 \cap A)/P(A)$ i anàlogament per a V_2 i V_3.

En aplicar la llei multiplicativa es té que les probabilitats intersecció són:

$P(V_1 \cap A) = P(V_1)\, P(A/V_1) = 0,5 \cdot 0,25 = 0,125$

$P(V_2 \cap A) = 0,3 \cdot 0,2 = 0,06$

$P(V_3 \cap A) = 0,2 \cdot 0,1 = 0,02$

Pel Teorema de la Probabilitat Total s'obté la P(A):

$P(A) = P(V_1 \cap A) + P(V_2 \cap A) + P(A \cap V_3) =$

$= P(V_1)\, P(A/V_1) + P(V_2)\, P(A/V_2) + P(V_3)\, P(A/V_3) =$

$= 0,5 \cdot 0,25 + 0,3 \cdot 0,2 + 0,2 \cdot 0,1 = 0,205$

Finalment, les probabilitats *a posteriori* són:

$P(V_1/A) = P(V_1 \cap A)/P(A) = 0,125/0,205 = 0,61$ superior a $P(V_1)$

$P(V_2/A) = 0,06/0,205 = 0,29$ inferior a $P(V_2)$

$P(V_3/A) = 0,02/0,205 = 0,10$ inferior a $P(V_3)$

Observeu que per a aquells successos on $P(A/V_n) > P(A)$, la probabilitat *a posteriori* $P(V_n/A)$ augmenta en relació a la probabilitat *a priori* $P(V_n)$. És a dir, en aquest exemple el video retornat té més probabilitat de ser dels que tenen major proporció d'avaria (major que la proporció global), és a dir la probabilitat *a posteriori* és major que la *a priori*, ja que el vídeo ha resultat defectuós. Pel contrari, aquells vídeos que tenen una tassa de defectuosos inferior a la total tindran menor probabilitat *a posteriori* si se sap que el vídeo ha resultat defectuós.

1.5 VARIABLE ALEATÒRIA

El conjunt de tots els resultats possibles d'un fenomen aleatori constitueix una població estadística que pot estar formada per resultats qualitatius o quantitatius. En qualsevol d'aquest casos resulta convenient associar aquests resultats a valors numèrics per tal de facilitar la seva representació i anàlisi.

El concepte de variable aleatòria permet relacionar cadascun dels resultats d'un experiment aleatori amb un valor numèric o descriure numèricament cadascun dels resultats d'una població estadística. Amb aquesta transformació s'aconsegueix caracteritzar la població estadística mitjançant una funció o model matemàtic que recull les descripcions numèriques dels resultats del fenomen aleatori juntament amb les seves respectives probabilitats.

Definició:

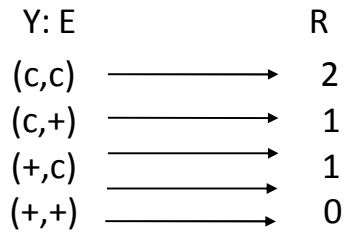
Una ***variable aleatòria*** és una funció de valor real, definida sobre un espai mostral E, que fa correspondre a cadascun dels elements de E un i només un nombre real.

Així, per exemple, si a l'espai mostral format pels resultats obtinguts en el llançament de dues monedes definim la següent aplicació real:

X: E		R
(c,c)	→	0
(c,+)	→	1
(+,c)	→	1
(+,+)	→	2

obtenim una variable aleatòria X que recull el '*nombre de creus*'.

D'altra banda, si en relació al mateix experiment aleatori definim la variable aleatòria Y com el '*nombre de cares*', aquesta s'indueix de l'aplicació:



Com veiem, sobre un mateix espai mostral es poden definir diferents variables aleatòries.

Quan el fenomen que es descriu té un nombre finit o infinit numerable de realitzacions possibles es modelitza mitjançant una ***variable aleatòria discreta***; si el nombre de realitzacions possibles del fenomen és infinit no numerable el model adequat és una ***variable aleatòria contínua***.

El comportament d'una variable aleatòria pot ser explicat per un model estocàstic o probabilístic que rep el nom de *distribució de probabilitat*.

Definició:

La ***distribució de probabilitat o llei de probabilitat*** d'una variable aleatòria és el model matemàtic (teòric) que associa a cadascun dels possibles valors o rangs de valors de la variable la seva corresponent probabilitat. Les probabilitats dels possibles valors de la variable s'estableixen a partir de l'aplicació induïda per X dels elements de E sobre el conjunt dels nombres reals. És a dir, en fer correspondre un valor real a cada succés de l'espai mostral estem, simultàniament, associant la probabilitat del succés al nombre real corresponent.

Exemple 1.6

Considerem la variable aleatòria X definida com el 'nombre de creus obtingudes' en el llançament de dues monedes equilibrades. Com que l'espai mostral associat a l'experiment té un nombre finit de resultats la variable aleatòria X és una variable discreta.

Si tenim en compte els resultats d'aquest experiment i les respectives probabilitats, podem induir les probabilitats associades a cadascun dels valors de X.

Succés	*Probabilitat*	*X*	*P(x)*
{cc}	0,5·0,5	0	0,25
{c+}∪{+c}	2·0,5·0,5	1	0,50
{++}	0,5·0,5	2	0,25

essent, per tant, la distribució de probabilitat de X:

X	P(x)
0	0,25
1	0,50
2	0,25

Aquesta distribució de probabilitat descriu completament el comportament de la variable X. Per exemple, la probabilitat d'obtenir *'almenys una cara'* és:

$P(X \geq 1) = P(X=1) + P(X=2) = P(1) + P(2) = 0,75$

Podem dir que si observem un gran nombre de realitzacions d'aquest fenomen el 75% de les vegades s'observarà la realització *'almenys una cara'*. Així mateix, podem dir que el 25% de les vegades s'observarà la realització *'exactament dues cares'*, etc.

Exemple 1.7

En el context d'un estudi sobre el rendiment laboral, s'observa el temps necessari fins aconseguir l'acabat d'un producte que pot variar entre 35 i 50 minuts.

Ara l'espai mostral conté teòricament un nombre infinit no numerable de resultats:

$E = \{t \mid 35 \leq t \leq 50\}$

I, en conseqüència, la variable X definida com el '*temps necessari per acabar el producte*' és una variable aleatòria contínua, ja que el seu espai mostral està format per tots el valors de l'interval real [35; 50].

1.5.1 VARIABLE ALEATÒRIA DISCRETA. FUNCIÓ DE QUANTIA

Quan resulta possible assignar una probabilitat a cadascun dels valors puntuals que pot prendre una variable aleatòria discreta s'estableix la seva llei de probabilitat que rep el nom de funció de quantia.

Definició:

Donada una variable aleatòria discreta, X, la seva ***distribució de probabilitat*** *o* ***funció de quantia***, que indicarem amb P(x), és el model teòric que assigna una probabilitat, P(x), a cadascun dels seus valors x.

De forma genèrica, la funció de quantia d'una variable discreta es representa com:

X	P(x)
x_1	$P(x_1)$
x_2	$P(x_2)$
...	...
x_i	$P(x_i)$
...	...
x_n	$P(x_n)$

on $P(x_i) = P(X=x_i)$ és la probabilitat que la variable X prengui el valor x_i.

La funció de quantia, per definició, sempre compleix les propietats següents:

- $P(x) \geq 0 \ \forall x$. Sempre pren valors no negatius.
- $\sum_{\forall x} P(x) = 1$. La probabilitat total o probabilitat del succés cert és igual a 1.

Exemple 1.8

Considerem l'espai mostral associat a un experiment que consisteix en observar els clients atesos pel dependent d'un comerç fins que aconsegueix la primera venda.

Si representem amb C el succés '*el client compra*' i amb $\overline{C}$ el succés '*el client no compra*', els resultats possibles que formen l'espai mostral són:

$$E = \{C, \overline{C}\,C, \overline{C}\,\overline{C}\,C, \overline{C}\,\overline{C}\,\overline{C}\,C, ...\}$$

La variable aleatòria definida com el '*nombre de clients atesos fins a la primera venda*' que indueix l'aplicació:

$$\begin{array}{rcl} X: E & \longrightarrow & R \\ (C) & \longrightarrow & 1 \\ (\overline{C}\,C) & \longrightarrow & 2 \\ (\overline{C}\,\overline{C}\,C) & \longrightarrow & 3 \\ \quad ... & & ... \\ (\overline{C}^{\,k-1}\,\overline{C}\,C) & \longrightarrow & k \end{array}$$

és una variable aleatòria discreta, ja que l'espai de referència que la defineix conté un nombre infinit numerable de resultats.

Suposem que es coneix per experiència que la probabilitat que un client qualsevol compri és 0,1 i que les decisions dels diferents clients són

independents, llavors les probabilitats associades a cadascun dels valors de X són les següents:

Succés	*Probabilitat del succés*	*X*	*P(x)*
C	0,1	1	0,1
$\overline{C}$ C	0,9 0,1	2	0,9 0,1
$\overline{C}$ $\overline{C}$ C	$0,9^2$ 0,1	3	$0,9^2$ 0,1
...	...	...	...
$\overline{C}$. $^{k-1}$. C	$0,9^{k-1}$ 0,1	k	$0,9^{k-1}$ 0,1

Aquesta funció és de quantia ja que:

- $P(x_i)\geq 0\ \forall x_i$
- $\sum_{\forall xi} P(x_i) = 0,1+0,1\cdot 0,9+0,1\cdot 0,9^2+...= 0,1\frac{1}{1-0,9}=1$

Observem que aquesta funció de quantia es pot representar mitjançant el següent model:

$$P(x) = \begin{cases} 0,9^{x-1}0,1 & x=1,2,3,... \\ 0 & en\ altres\ casos \end{cases}$$

El comportament de X es descriu totalment amb la funció de quantia, és a dir, es pot trobar la probabilitat associada a qualsevol valor o interval de valors de X a partir de P(x). Així, per exemple:

- La probabilitat que sigui necessari atendre a 20 clients per aconseguir la primera venda és:

  $P(X=20) = P(20) = 0,9^{19}\ 0,1 = 0,0135$
- La probabilitat que calgui atendre com a màxim 3 clients és:

  $P(X\leq 3) = P(1) + P(2) + P(3) = 0,1 + 0,9\ 0,1 + 0,9^2\ 0,1 = 0,271$
- La probabilitat que calgui atendre com a màxim 7 clients si se sap que ja se n'han atès més de 4 és:

  $$P(X\leq 7 / X>4) = \frac{P[(X\leq 7)\cap(X>4)]}{P(X>4)} = \frac{P(5\leq X\leq 7)}{P(X>4)}$$

essent:

$P(5\leq X\leq 7) = P(5) + P(6) + P(7) = 0,9^4\ 0,1 + 0,9^5\ 0,1 + 0,9^6\ 0,1 = 0,1778$

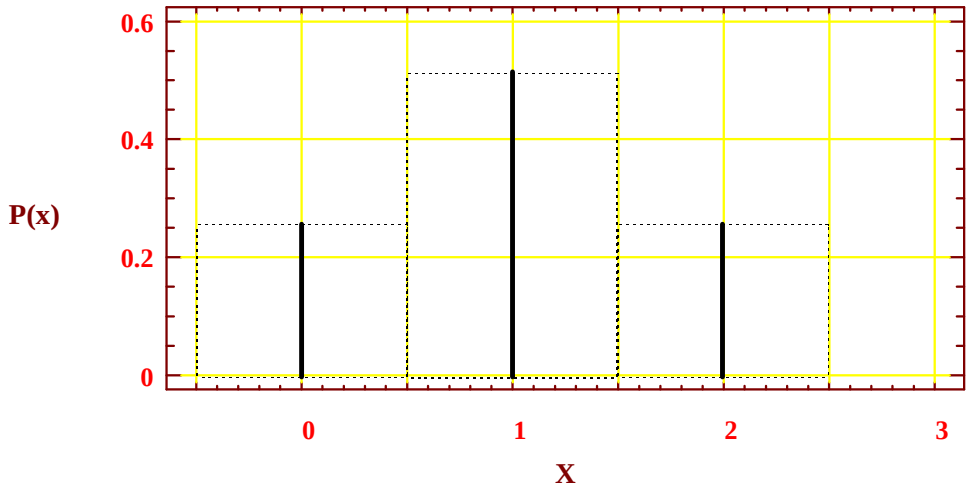
$P(X>4) = 1 - P(X\leq 4) = 1-[P(1)+P(2)+P(3)+P(4)] =$

$= 1- (0,1+0,9\cdot 0,1+0,9^2\cdot 0,1+0,9^3\cdot 0,1) = 0,6561$

$$P(X\leq 7/X>4) = \frac{0{,}1778}{0{,}6561} = 0{,}271$$

La representació gràfica de la distribució de probabilitat d'una variable aleatòria discreta es fa mitjançant un diagrama de barres on a l'eix d'abscisses es representen els valors de la variable i a l'eix d'ordenades les seves probabilitats. En ocasions resulta d'utilitat representar la funció de quantia substituint les barres per rectangles amb base unitària, centrats sobre cadascun dels valors de X, de manera que les probabilitats queden representades per l'àrea del rectangle, tal com veiem en el gràfic 1.5 que recull la funció de quantia de la variable X=*'nombre de creus'* en el llançament de dues monedes:

Gràfic 1.5 Funció de quantia.



1.5.2 VARIABLE ALEATÒRIA CONTÍNUA. FUNCIÓ DE DENSITAT

Una variable aleatòria X és *contínua* quan l'espai mostral associat a l'experiment està format per un nombre infinit de resultats no numerables. Per tant, una variable contínua pot prendre qualsevol valor d'un interval real.

L'espai mostral que defineix la variable aleatòria contínua està format per infinits resultats elementals i, si considerem que tots ells són equiprobables, no és possible assignar una probabilitat a cadascun d'ells, ja que en fer-ho s'incompliria el segon axioma de la probabilitat. És a dir, no es poden assignar probabilitats puntuals als valors d'una variable aleatòria contínua, només es pot avaluar la probabilitat a intervals de valors de la variable, encara que

aquests siguin infinitament petits. És per això que parlem de densitat de probabilitat de la variable.

Definició:

La llei de probabilitat d'una variable aleatòria X contínua ve determinada per una funció f(x) que verifica:

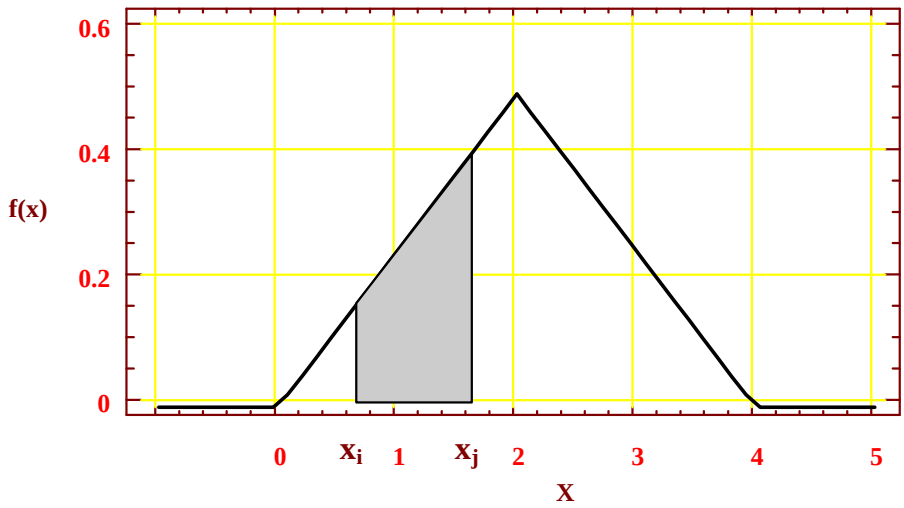
- $f(x) \geq 0 \ \forall x$. Per a tots els valors reals sempre pren valors no negatius.
- $\int_{-\infty}^{+\infty} f(x)dx = 1$. L'àrea definida per aquesta funció és igual a 1 i, per tant, recull la probabilitat total.

i s'anomena ***funció de densitat de probabilitat***.

La probabilitat que X prengui valors en un interval $[x_i, x_j]$ queda determinada per l'àrea definida per f(x) entre $x=x_i$ i $x=x_j$ i es troba integrant la funció de densitat f(x) entre els límits de l'interval:

$$P(x_i \leq X \leq x_j) = \int_{xi}^{xj} f(x)\, dx$$

Gràfic 1.6 Funció de densitat



Quan la variable aleatòria és contínua la probabilitat que prengui un valor dintre d'un interval és la mateixa tant si és obert com si és tancat, ja que la probabilitat en un punt, com ja s'ha dit, és sempre zero.

$P(x_i \leq X \leq x_j) = P(x_i \leq X < x_j) = P(x_i < X \leq x_j) = P(x_i < X < x_j)$.

Exemple 1.9

Donada una variable aleatòria contínua amb la següent funció de densitat:

$$f(x)=\begin{cases}\frac{1}{4}x & 1\le x\le 3\\ 0 \text{ en altres casos}\end{cases}$$

Representeu-la gràficament, comproveu que és funció de densitat i calculeu les probabilitats següents:

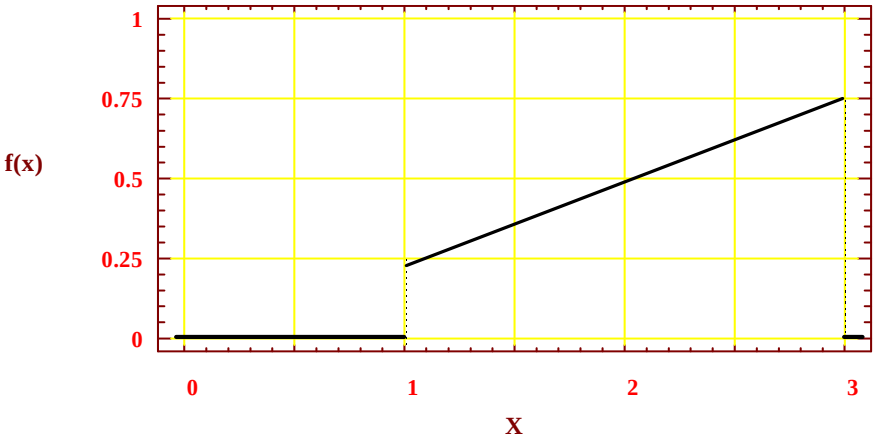
a) Probabilitat que X prengui un valor inferior a 1,5.

b) Probabilitat que X prengui un valor entre 1,5 i 2,7.

c) Probabilitat que X prengui un valor superior a 2,7 si se sap que ha pres un valor superior a 1,5.

Solució:

Gràfic 1.7 Funció de densitat.



- f(x)≥0 ∀x. A partir de la representació anterior és immediat comprovar que per a tots els valors reals la funció queda a sobre o per sobre de l'eix d'abscisses.
- $\int_{-\infty}^{+\infty} f(x)dx=\int_{1}^{3}\frac{1}{4}x\,dx=\frac{1}{4}\frac{x^2}{2}\Big]_1^3=\frac{1}{8}$ (9-1)= 1

a) La probabilitat que X prengui un valor inferior a 1,5 és:

$$P(X\le 1,5)=\int_{1}^{1,5}\left(\frac{1}{4}\right)x\,dx=\left(\frac{1}{4}\right)\frac{x^2}{2}\Big]_1^{1,5}=\left(\frac{1}{4}\right)\left[\frac{2,25}{2}-\frac{1}{2}\right]=0,1562$$

b) La probabilitat que X prengui un valor entre 1,5 i 2,7 és:

$$P(1,5\le X\le 2,7)=\int_{1,5}^{2,7}\left(\frac{1}{4}\right)x\,dx=\left(\frac{1}{4}\right)\frac{x^2}{2}\Big]_{1,5}^{2,7}=\left(\frac{1}{4}\right)\left[\frac{7,29}{2}-\frac{2,25}{2}\right]=0,63$$

c) La probabilitat que X prengui un valor superior a 2,7 si se sap que ha pres un valor superior a 1,5 és:

$$P(X>2,7/X>1,5) = \frac{P[(X>2,7)\cap(X>1,5)]}{P(X>1,5)} = \frac{P(X>2,7)}{P(X>1,5)}$$

Les probabilitats del numerador i denominador de l'expressió anterior són:

$$P(X>2,7)= \int_{2,7}^{3}\left(\frac{1}{4}\right)x\,dx = \left(\frac{1}{4}\right)\frac{x^2}{2}\Big]_{2,7}^{3} = \left(\frac{1}{4}\right)\left[\frac{9}{2}-\frac{7,29}{2}\right]=0,2137$$

$$P(X>1,5)= \int_{1,5}^{3}\left(\frac{1}{4}\right)x\,dx = \left(\frac{1}{4}\right)\frac{x^2}{2}\Big]_{1,5}^{3} = \left(\frac{1}{4}\right)\left[\frac{9}{2}-\frac{2,25}{2}\right]=0,8437$$

i, en substituir, obtenim:

$$P(X>2,7 / X>1,5) = \frac{0,2137}{0,8437} = 0,2533$$

1.5.3 FUNCIÓ DE DISTRIBUCIÓ

Qualsevol variable aleatòria, tant discreta com contínua, té associada una funció que permet el càlcul de probabilitats acumulades fins a un valor x_i. Aquesta funció s'anomena *funció de distribució* i es dedueix a partir de la funció de densitat o a partir de la funció de quantia.

Definició:

Donada una variable aleatòria amb funció de densitat f(x) o amb funció de quantia P(x), la ***funció de distribució*** és una funció que té un valor en x_i igual a la probabilitat acumulada fins x_i:

$$F(x_i) = P(X\leq x_i)$$

Essent,

$F(x_i) = P(X\leq x_i) = \sum_{x\leq x_i} P(x_i)$ si X és una variable aleatòria discreta i

$F(x_i) = P(X\leq x_i) = \int_{-\infty}^{xj} f(x)\,dx$ si X és contínua.

La informació que conté la funció de distribució respecte al comportament de la variable és la mateixa que la continguda a la funció de quantia o de densitat, segons el cas. Així, la probabilitat que una variable prengui valors en un interval

es pot calcular mitjançant la funció de distribució o mitjançant la corresponent funció de probabilitat.

Característiques:

1. La funció és no negativa. $0 \leq F(x) \leq 1$.
2. La funció és no decreixent. Donats dos valors $x_i < x_j$ aleshores $F(x_i) \leq F(x_j)$.
3. La funció convergeix a zero per l'esquerra. $F(-\infty) = 0$.
4. La funció convergeix a u per la dreta. $F(+\infty) = 1$.
5. $P(x_i < X \leq x_j) = F(x_j) - F(x_i)$.
6. $P(X > x_i) = 1 - F(x_i)$.
7. Si la variable és contínua, F(x) és contínua i derivable. La derivada de F(x) és la funció de densitat.
8. Si la variable és discreta, F(x) és una funció escalonada.

Exemple 1.10

Per a la variable X = 'nombre de peces defectuoses en un paquet de 5 peces' s'ha establert la següent distribució de probabilitat:

X	P(x)
0	0,3277
1	0,4096
2	0,2048
3	0,0512
4	0,0064
5	0,0003

La corresponent funció de distribució és

$$F(x) = \begin{cases} 0 & x < 0 \\ 0{,}3277 & 0 \leq x < 1 \\ 0{,}7373 & 1 \leq x < 2 \\ 0{,}9421 & 2 \leq x < 3 \\ 0{,}9933 & 3 \leq x < 4 \\ 0{,}9997 & 4 \leq x < 5 \\ 1 & x > 5 \end{cases}$$

La funció de distribució indica la probabilitat acumulada fins a qualsevol valor de X, així, per exemple F(2) = P(X ≤ 2) = 0,9421, perquè 2 és un valor comprès en l'interval 2 ≤ x < 3, al que correspon el valor F(X)= 0,9421.

Exemple 1.11

Es vol determinar la funció de distribució d'una variable aleatòria que queda caracteritzada per la funció de densitat:

$$f(x) = \begin{cases} \dfrac{1}{2\sqrt{x}} & 0 < x < 1 \\ 0 & \textit{en altres casos} \end{cases}$$

Solució:

En primer lloc comprovem que la funció és de densitat:

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{0}^{1} \frac{1}{2\sqrt{x}}\,dx = \sqrt{x}\,\Big]_{0}^{1} = 1 \quad \text{i} \quad f(x) \geq 0 \quad \forall x$$

Per determinar la funció de distribució hem d'integrar per trams: en primer lloc de -∞ fins a un valor x menor a 0; a continuació de -∞ a un valor x inferior a 1; i, per últim, de -∞ a un valor x superior o igual a 1.

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} 0\,dx = 0 \qquad x<0$$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{x} \frac{1}{2\sqrt{x}}\,dx = \sqrt{x}\,\Big]_{0}^{x} = \sqrt{x} \qquad 0<x<1$$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} \frac{1}{2\sqrt{x}}\,dx + \int_{1}^{x} 0\,dx = \sqrt{x}\,\Big]_{0}^{1} = 1 \quad x>1$$

Per tant, la funció de distribució queda:

$$F(x) = \begin{cases} 0 & x < 0 \\ \sqrt{x} & 0 < x < 1 \\ 1 & x > 1 \end{cases}$$

Quan es treballa amb la funció de distribució d'una variable contínua és indiferent que l'interval per al qual es calcula la probabilitat contingui o no un o els dos extrems, ja que la probabilitat que X prengui exactament un valor és zero. Per tant, donats dos valors de la variable, x_i i x_j, essent $x_i < x_j$ s'ha de tenir en compte que

$P(x_i \leq X \leq x_j) = P(x_i < X < x_j) = P(x_i \leq X < x_j) = P(x_i < X \leq x_j) = F(x_j) - F(x_i)$

En el cas que la variable sigui discreta no és indiferent que l'un o l'altre o ambdós extrems hi estiguin inclosos i s'haurà de tenir en compte que:

- $P(X \leq x_i) = F(x_i)$
- $P(X < x_i) = F(x_{i-1})$
- $P(x_i \leq X \leq x_j) = F(x_j) - F(x_{i-1})$
- $P(x_i < X \leq x_j) = F(x_j) - F(x_i)$
- $P(x_i \leq X < x_j) = F(x_{j-1}) - F(x_{i-1})$
- $P(x_i < X < x_j) = F(x_{j-1}) - F(x_i)$
- $P(X \geq x_i) = 1 - F(x_{i-1})$
- $P(X > x_i) = 1 - F(x_i)$

1.6 CARACTERÍSTIQUES D'UNA VARIABLE ALEATÒRIA

La distribució de probabilitat d'una variable aleatòria presenta certes semblances amb la distribució de freqüències empírica d'un conjunt de dades mostrals. En aquest sentit, per a qualsevol distribució de probabilitat és possible calcular un conjunt de mesures que resultin útils per descriure-la. Així, una distribució de probabilitat es pot sintetitzar utilitzant mesures de posició, de dispersió, de asimetria o d'apuntament. En primer lloc ens centrarem, únicament, en l'obtenció del valor esperat com a mesura de posició i en la variància i desviació estàndard com a mesures de dispersió de la distribució i, a continuació, definirem els moments d'una distribució de probabilitat.

1.6.1 ESPERANÇA MATEMÀTICA

Definició:

L'esperança matemàtica o valor esperat d'una variable aleatòria, que es simbolitza com $E(X) = \mu$, és el valor mitjà teòric de la distribució i s'obté amitjanant tots els valors de la variable ponderats per les seves respectives probabilitats.

Si X és una variable discreta amb funció de quantia P(x), l'esperança matemàtica ve donada per:

$$E(X) = \mu = \sum_{\forall x_i} x_i P(x_i)$$

mentre que si X és una variable contínua amb funció de densitat f(x), l'esperança matemàtica és:

$$E(X)= \mu = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

L'esperança matemàtica d'una distribució s'interpreta com el valor que per terme mitjà prendria la variable aleatòria en el cas que l'experiment es repetís de forma indefinida.

Propietats:

1. E(a) = a. L'esperança d'una constant és la constant.
2. E(X-E(X)) = 0. L'esperança matemàtica és el centre de gravetat de la distribució de probabilitats.
3. $E(g(x)) = \begin{cases} \sum_{\forall x} g(x)P(x) & discreta \\ \int_{-\infty}^{\infty} g(x)f(x)dx & continua \end{cases}$ on g(x) és una transformació lineal de X.

Per tant,

- E(a+X) = a+E(X). L'esperança matemàtica queda afectada pels canvis d'origen.
- E(bX) = b E(X). També queda afectada pels canvis d'escala.
- E(a+b X) = a+b E(X).

4. E(aX+bY) = aE(X) + bE(Y) on X i Y són dues variables aleatòries i, a i b, dues constants.

1.6.2 VARIÀNCIA

A fi de poder mesurar la variabilitat de la distribució és necessari disposar d'una mesura que informi de la seva dispersió. La mesura de dispersió més utilitzada és la variància.

Definició:

La ***variància*** d'una variable aleatòria, que simbolitzarem amb V(X) o σ^2, mesura la dispersió dels valors al voltant de la seva esperança matemàtica i, de forma genèrica, ve donada per l'expressió:

$$V(X) = \sigma^2 = E(X - \mu)^2$$

Si es desenvolupa l'expressió anterior queda:

$V(X) = E(X^2 - 2 \cdot X \cdot \mu + \mu^2) = E(X^2) - 2 \cdot \mu^2 + \mu^2 = E(X^2) - \mu^2$

Per tant, si X és una variable discreta amb funció de quantia P(x), la seva variància és:

$$V(X) = \sigma^2 = E(X - \mu)^2 = \sum_{\forall x_i} (x_i - \mu)^2 \, P(x_i) = \sum_{\forall x_i} x_i^2 P(x_i) - \mu^2$$

mentre que per a una variable contínua:

$$V(X) = \sigma^2 = E(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \, f(x)\, dx = \int_{-\infty}^{\infty} x^2 \, f(x)\, dx - \mu^2$$

Propietats:

1. $V(X) \geq 0$
2. $V(a) = 0$. La variància d'una constant sempre és zero.
3. $V(bX) = b^2 \cdot V(X)$. La variància d'una constant per a la variable és igual al quadrat de la constant per la variància de la variable.
4. $V(a + bX) = b^2 \cdot V(X)$. La variància només queda modificada pels canvis d'escala.

Donat que la variància no presenta les mateixes unitats de mesura que el valor esperat, s'acostuma a calcular la seva arrel quadrada.

Definició:

La ***desviació estàndard*** d'una variable aleatòria, que simbolitzarem com D(X) o σ, mesura la dispersió dels valors al voltant de la seva esperança matemàtica i ve donada per l'arrel quadrada de la variància:

$$D(X) = \sigma = \sqrt{\sigma^2}$$

La desviació estàndard presenta la mateixa interpretació i les mateixes propietats que la variància: és no negativa i només queda modificada pels canvis d'escala.

Exemple 1.12

Un arxivador conté 6 factures de les quals només 4 estan ben classificades. S'extreuen a l'atzar, sense devolució, una a una fins a obtenir-ne una de les correctament classificades. Determineu la funció de quantia, el valor esperat, la

variància i la desviació estàndard de la variable aleatòria X = 'nombre de factures extretes de l'arxivador'.

Solució:

Definim B el succés '*factura ben classificada*' essent P(B) = 4/6,

i $\overline{B}$ el succés '*factura mal classificada*' amb P($\overline{B}$) = 2/6

La funció de quantia de X és:

Successos	X	P(X)
B	1	4/6 = 2/3
$\overline{B}$ B	2	2/6·4/5 = 4/15
$\overline{B}$ $\overline{B}$ B	3	2/6·1/5·1 = 1/15

$E(X) = \mu = 1* 2/3 + 2 *4/15 + 3 *1/15 = 1,4$ factures.

$V(X) = \sigma^2 = E(X^2) - \mu^2 = 1^2 *2/3 + 2^2 *4/15 + 3^2 *1/15 - 1,4^2 = 0,3733$

$D(X) = \sigma^2 = \sqrt{0,3733} = 0,61$

Exemple 1.13

Considerem la variable aleatòria X caracteritzada per la funció de densitat:1

$$f(x) = \begin{cases} \dfrac{1}{2\sqrt{x}} & 0 < x < 1 \\ 0 & \textit{en} \text{ altres casos} \end{cases}$$

Es vol determinar el valor esperat, la variància i la desviació estàndard de X.

Solució:

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx = \int_0^1 \frac{x}{2\sqrt{x}}\, dx = \frac{1}{3}\sqrt{x^3}\Big]_0^1 = 1/3$$

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

$$E(X^2) = \int_{-\infty}^{+\infty} x^2\, f(x)\, dx = \int_0^1 \frac{x^2}{2\sqrt{x}}\, dx = \frac{1}{5}\sqrt{x^5}\Big]_0^1 = 1/5$$

per tant:

$V(X) = E(X^2) - \mu^2 = 1/5 - (1/3)^2 = 0,0889$ i $D(X) = \sqrt{0,0889} = 0,2981$

1.6.3 TEOREMA DE TXEBIXEV

Quan es coneix el model probabilístic, es pot avaluar la probabilitat de qualsevol succés i, en concret, la probabilitat d'obtenir un resultat que pertanyi a un interval centrat en el valor esperat.

Pel contrari, si la informació disponible relativa a la variable aleatòria es redueix únicament als seus principals moments, μ i σ^2, no és possible calcular la probabilitat d'un succés qualsevol. En aquesta situació, l'aplicació de la *desigualtat de Txebixev* permet fixar una cota inferior per a la probabilitat d'obtenir valors que es trobin dins d'un interval centrat en l'esperança matemàtica en funció d'aquests moments.

És a dir, donada una variable aleatòria X, contínua o discreta, amb esperança matemàtica μ i desviació estàndard σ, es pot acotar la probabilitat que presenta qualsevol interval obert centrat en la mitjana $(\mu-k\sigma\ ;\ \mu+k\sigma)$ on k és una constant positiva.

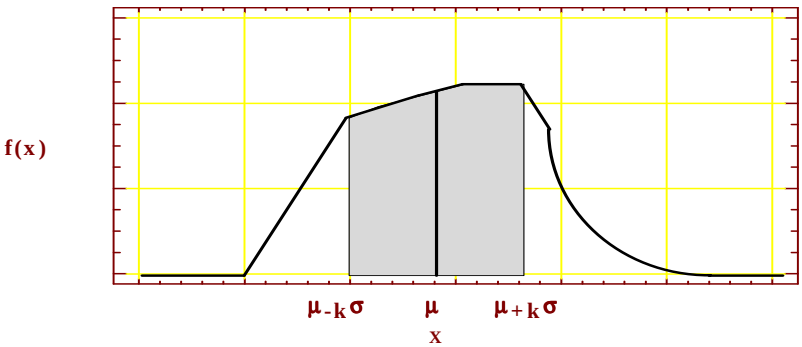
Definició:

Donada una variable aleatòria X, contínua o discreta, amb esperança matemàtica μ i desviació estàndard σ, el ***teorema de Txebixev*** estableix una cota inferior per a la probabilitat en termes dels paràmetres μ i σ de la forma següent:

$$P(\mu-k\sigma < X < \mu+k\sigma) \geq 1-1/k^2 \quad \text{o}$$
$$P(|X-\mu|<k\sigma) \geq 1-1/k^2$$

És a dir, independentment de quina sigui la llei de probabilitat d'una variable aleatòria X, si es coneixen els valors μ i σ, el tant per u d'observacions poblacionals que es troba entre $\mu-k\sigma$ i $\mu+k\sigma$ sempre serà com a mínim $1-1/k^2$.

Gràfic 1.8 Probabilitat en l'interval $(\mu-k\sigma\ \mu+k\sigma)$.



Si considerem ara la probabilitat o percentatge de valors que queda fora de l'interval anterior es pot deduir que:

$$P(|X-\mu|\geq k\sigma) < 1/k^2$$

És a dir, la probabilitat que el valor de la variable aleatòria difereixi de la mitjana com a mínim k vegades la desviació estàndard és menor a $1/k^2$.

Es pot veure com per a qualsevol valor $k\leq 1$ el Teorema es trivial, ja que en aquest cas la cota inferior seria menor o igual a zero i, per tant, únicament té sentit aplicar-lo per a valors de k superiors a 1.

Observeu que les probabilitats mínimes dels intervals centrats en la mitjana $\pm 2\sigma$, $\pm 3\sigma$ o $\pm 4\sigma$ són:

$$k=2 \quad P(\mu-2\sigma < X < \mu+2\sigma) \geq 0,75$$
$$k=3 \quad P(\mu-3\sigma < X < \mu+3\sigma) \geq 0,89$$
$$k=4 \quad P(\mu-4\sigma < X < \mu+4\sigma) \geq 0,94$$

Exemple 1.14

La quantitat a retornar per Hisenda en concepte d'IRPF en una determinada comunitat és una variable aleatòria amb valor esperat 480€. i desviació estàndard 120€. si entre els contribuents que tenen dret a devolució se'n tria un a l'atzar, quina és la probabilitat que la quantitat a retornar a aquest contribuent no difereixi del valor esperat en més de 240€?

Solució:

X = '*Quantitat a retornar*' μ_X =480€ σ_X =120€

$P(\mu-240< X < \mu+240) = P(\mu-k\sigma < X < \mu+k\sigma)$

Com $k\sigma = 240 \Rightarrow k = 240/120= 2$

i l'aplicació de la Desigualtat de Txebixev ens dóna:

$P(\mu-2\sigma \leq X \leq \mu+2\sigma) \geq 1-1/2^2 = 0,75$

Per tant, la probabilitat és igual o superior a 0,75.

1.6.4 ESTANDARDITZACIÓ D'UNA VARIABLE ALEATÒRIA

Una transformació lineal de la variable aleatòria, que té un interès especial, és la tipificació o estandardització de la variable. Aquesta transformació desplaça

el centre de gravetat de la distribució a l'origen de coordenades, modifica la seva dispersió i elimina les unitats de mesura de la variable.

Definició:

Donada una variable aleatòria X (discreta o contínua) amb E(X) = μ i V(X) = σ^2 es defineix la ***variable tipificada o estandarditzada de X***, que es simbolitza amb Z, com la transformació lineal:

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

Per definició aquesta variable sempre compleix:

- E(Z) = 0, ja que $E(\frac{X-\mu}{\sigma}) = \frac{1}{\sigma}E(X-\mu) = 0$.
- V(Z) = 1, ja que $V(\frac{X-\mu}{\sigma}) = \frac{1}{\sigma^2}V(X-\mu) = \frac{1}{\sigma^2}V(X) = 1$.

Aquesta transformació permet comparar la posició d'un mateix individu en dues o més distribucions de probabilitat, ja que la variable tipificada és adimensional i indica la posició relativa de l'individu respecte a μ. La variable tipificada s'interpreta com el nombre de desviacions estàndards que un determinat valor x_i s'allunya de μ.

1.6.5 MOMENTS D'UNA DISTRIBUCIÓ

A més de l'esperança matemàtica i la variància d'una variable aleatòria existeixen altres mesures que serveixen per caracteritzar la distribució. L'obtenció de valors numèrics d'aquestes mesures es pot efectuar mitjançant els anomenats moments de la distribució, que no són més que l'esperança matemàtica de la diferència entre el valor de la variable i un cert valor **a**, elevada a una determinada potència k que determina l'ordre del moment.

$$E(X - a)^k$$

En general, els moments d'una distribució es poden definir al voltant de qualsevol constant **a** però, a la pràctica, es calculen agafant com a valor de referència el zero o bé el valor esperat de la distribució. En el primer cas s'obtenen els moments ordinaris o moments respecte a l'origen que

simbolitzarem amb α_k, mentre que en el segon cas s'aconsegueixen els moments centrals que simbolitzarem mitjançant μ_k.

Definició:

Donada una variable aleatòria X, el ***moment ordinari*** d'ordre k (o k-èsim) és:

$\alpha_k = E(X^k) = \sum_{\forall x_i} x_i^k P(x_i)$ si X és una variable aleatòria discreta

$\alpha_k = E(X^k) = \int_{-\infty}^{\infty} x^k \, f(x)\,dx$ si X és una variable aleatòria contínua

Alguns casos particulars són:

$\alpha_0 = 1$ El moment ordinari d'ordre zero sempre és 1.

$\alpha_1 = E(X) = \mu$

El primer moment ordinari respecte a zero és l'esperança matemàtica de la distribució.

$\alpha_2 = E(X^2)$

Definició:

Donada una variable aleatòria X el ***moment central*** d'ordre k (o k-èsim) ve donat per:

$\mu_k = E(X - \mu)^k = \sum_{\forall x_i} [x_i - E(X)]^k P(x_i) = \sum_{\forall x_i} (x_i - \mu)^k P(x_i)$ discretes.

$\mu_k = E(X - \mu)^k = \int_{-\infty}^{\infty} [x_i - E(X)]^k f(x)\,dx = \int_{-\infty}^{\infty} (x_i - \mu)^k f(x)\,dx$ contínues.

Casos particulars:

$\mu_0 = E(X - \mu)^0 = E(1) = 1$. El moment central d'ordre zero sempre és 1.

$\mu_1 = E(X - \mu)^1 = E(X) - \mu = 0$. El moment central d'ordre u sempre és 0.

$\mu_2 = E(X - \mu)^2 = V(X)$. El moment central d'ordre dos és la variància de la distribució.

Els moments centrals es poden expressar en funció dels moments ordinaris. Per exemple, la variància $\mu_2 = \alpha_2 - \alpha_1^2$.

1.7 EXERCICIS PROPOSATS

Exercici 1. Determineu el valor de la probabilitat en cadascuna de les següents situacions:

a) Probabilitat que un negoci d'alimentació tingui èxit si en termes generals s'estima que per cada negoci d'aquest tipus que fracassa tres tenen èxit.
b) Probabilitat d'obtenir una *copa* en extreure a l'atzar una carta d'una baralla de 48 cartes.
c) Probabilitat d'obtenir una puntuació total superior a 7 en llançar dos daus.
d) Probabilitat d'accident laboral en un sector industrial si d'una mostra aleatòria de 8000 treballadors d'empreses del sector 40 treballadors havien sofert algun tipus d'accident.

Exercici 2. Donats dos successos A i B amb probabilitats 1/2 i 1/3, respectivament, trobeu $P(A\cap B)$ sota els següents supòsits:

a) $P(A\cup B) = 4/5$.
b) A i B són independents.
c) A i B són incompatibles.
d) $P(\overline{A}\cap B) = 1/10$.

Exercici 3. Als habitants de la comarca del Vallès Occidental se'ls va fer una enquesta per determinar el nombre de lectors de dues revistes mensuals (A i B). Els resultats de l'enquesta van ser els següents: el 20% dels habitants llegien la revista A, el 16% llegien la revista B i l'1% llegien ambdues revistes.

a) Són mútuament excloents els successos '*llegir la revista A*' i '*llegir la revista B*'?
b) Són independents els successos '*llegir la revista A*' i '*llegir la revista B*'?
c) Quina és la probabilitat que un habitant triat a l'atzar llegeixi alguna d'aquestes revistes?
d) Si es tria a l'atzar un lector de la revista A, quina és la probabilitat que també llegeixi la revista B?

Exercici 4. Si d'un seminari amb 12 nois i 4 noies es tria una comissió de 3 estudiants a l'atzar, quina és la probabilitat que aquesta estigui composta només per nois?

Exercici 5. D'una població de 2000 contribuents amb sospita d'algun error a les seves declaracions d'IRPF, s'ha observat la presència de desgravacions improcedents en 1500 declaracions, errors de càlcul en 1000 declaracions i la presència simultània dels dos errors en 750 declaracions. Es pot afirmar que els contribuents cometen els anteriors errors de forma independent?

Exercici 6. Donats dos successos independents A i B amb probabilitats 0,2 i 0,6, respectivament, trobeu les següents probabilitats condicionades:

a) $P(A/A \cup B)$
b) $P(A/A \cap B)$
c) $P(A \cap B/A \cup B)$
d) $P(A/\overline{A} \cap B)$
e) $P(A/\overline{B})$
f) $P(\overline{A}/B)$

Exercici 7. Una central té 3 generadors que funcionen independentment. La probabilitat que falli un generador és 0,02. Si per mantenir el subministrament és necessari que funcionin almenys 2 generadors, quina és la probabilitat que no quedi interromput el subministrament?

Exercici 8. Se sap que un 5% dels aparells de TV d'una determinada marca presenta el comandament a distància i el plafó de comandaments defectuosos, i que el 10% només té defectuós el plafó de comandaments. Si un determinat aparell té el plafó de comandaments defectuós, quina és la probabilitat que el seu comandament a distància també sigui defectuós?

Exercici 9. Un 40% dels estudiants d'una classe té aprovada l'assignatura A; un 70% té aprovada l'assignatura B; i un 20% té ambdues assignatures aprovades. Si es procedeix a l'elecció d'un alumne de la classe a l'atzar i resulta que té aprovada l'assignatura B, quina és la probabilitat que no tingui aprovada l'A?

Exercici 10. Una empresa projecta obrir una nova sucursal del seu negoci en aquell municipi de l'àrea metropolitana que li sigui més favorable. Per a prendre aquesta decisió sap que dels municipis considerats un 20% presenten recessió de població, un 30% població en expansió i la resta població estable. A més coneix que dels negocis que van obrir durant l'últim any van prosperar-ne

un 40% als municipis en recessió; un 80% als municipis en expansió i un 65% al tercer tipus de municipis.

a) Quina probabilitat té el nou establiment de no prosperar?
b) Si passat un temps se sap que el negoci ha prosperat, quina és la probabilitat que hagi estat obert a un municipi en recessió?

Exercici 11. En un examen tipus test, on cada pregunta té 4 respostes alternatives de les quals només una és correcta, els alumnes contesten totes les preguntes, o bé a l'atzar o bé perquè coneixen la resposta correcta. Si la probabilitat de conèixer la resposta correcta és 0,3, quina és la probabilitat que un alumne conegués la resposta a una pregunta contestada correctament?

Exercici 12. Les 4 províncies d'una comunitat autònoma presenten les següents xifres de població activa i percentatge d'atur:

Província	A	B	C	D
Població Activa	200.000	600.000	800.000	400.000
% d'aturats	5	8	3	10

Si se sap que un treballador està a l'atur, quina és la probabilitat que sigui de la província A? I de la província B?

Exercici 13. Un balneari amb 20 habitacions accepta un 20% més de reserves durant els caps de setmana ja que ha comprovat que algunes reserves es cancel·len a l'últim moment. Si la distribució de probabilitat de X= '*nombre de reserves que no es cancel·len*' és:

X	16	17	18	19	20	21	22	23	24
P(x)	0,05	0,05	0,15	0,25	0,20	0,15	0,05	0,05	0,05

a) Quina és la probabilitat que tots els clients amb reserva que vagin al balneari un determinat cap de setmana tinguin habitació?
b) Quina és la probabilitat que algun dels clients amb reserva que vagi al balneari un determinat cap de setmana no disposi d'habitació?
c) Quina és la probabilitat que el primer client sense reserva que vagi al balneari un determinat cap de setmana tingui habitació?
d) Quin és el nombre esperat de clients que no cancel·len la reserva?

Exercici 14. Una caixa conté 6 cargols dels quals només 2 són útils per al muntatge d'un determinat producte. Els cargols s'extreuen d'un en un a l'atzar i es proven fins a trobar el primer cargol útil. Si les extraccions són sense devolució:

a) Establiu la distribució de probabilitat de la variable aleatòria X=*'nombre de cargols extrets fins a trobar el primer cargol útil'*.
b) Establiu la distribució de probabilitat de la variable aleatòria Y=*'nombre de cargols que queden a la caixa'*.
c) Trobeu l'esperança matemàtica i la variància de les variables X i Y. Quina d'aquestes variables presenta major dispersió?

Exercici 15. En l'embalatge de paquets de farina de 10 Kg es comet un error aleatori en el pes, X (en kg), amb funció de densitat:

$$f(x)=\begin{cases} kx^2 & -1<x<1 \\ 0 & \textit{en altres casos} \end{cases}$$

a) Determineu el valor de k.
b) Representeu gràficament el comportament de X.
c) Calculeu el valor esperat i la variància de X.
d) Calculeu la probabilitat que un sac de farina superi els 10,5 kg.
e) Quin és el pes mínim que pot presentar un sac si se sap que pertany al 30% dels més pesats?
f) Si l'error d'embalatge és el doble que l'actual, quin és el nou valor esperat i la nova variància de la variable aleatòria?

Exercici 16. Donada una variable aleatòria amb funció de distribució:

$$F(x)=\begin{cases} 0 & x<0 \\ 2x-x^2 & 0\le x\le 1 \\ 1 & x>1 \end{cases}$$

a) Representeu gràficament F(x).
b) Calculeu P(X<1/3), P(X>2/3) i P(X>0,25/X<0,75).
c) Obteniu la funció de densitat.
d) Calculeu l'esperança matemàtica i la variància de X.
e) Calculeu el valor de la mediana i el de la moda de la distribució.

Exercici 17. Una màquina introdueix un producte en envasos que indiquen un pes net de 15,5 gr. La màquina està ajustada de forma que les quantitats introduïdes en els envasos tenen un pes mitjà de 16 gr. amb una desviació estàndard de 0,3 gr. Si l'empresa vol que les quantitats envasades presentin com a mínim el pes indicat a l'envàs i que aquest no excedeixi els 16,5 gr., quin és el percentatge mínim d'envasos que compliran les condiciones fixades per

l'empresa? Si la producció diària és de 100000 envasos, quants s'espera que com a màxim estiguin per sota o per sobre dels límits fixats per l'empresa?

Exercici 18. A causa d'irregularitats en el funcionament d'una màquina el pes (en kg.) de les peces que fabrica és una variable aleatòria amb distribució de probabilitat:

$$f(x) = \begin{cases} -20 + 25x & 0{,}8 < x < 1 \\ 30 - 25x & 1 < x < 1{,}2 \\ 0 & en\ altres\ casos \end{cases}$$

a) Obteniu la funció de distribució.
b) Quina és la probabilitat que una peça pesi menys d'1,05 Kg?
c) Calculeu l'esperança matemàtica i la desviació estàndard de X.
d) Quina és la probabilitat que el pes d'una peça difereixi del valor esperat com a màxim en dues vegades la desviació estàndard?

Exercici 19. A una Estació de Servei se li subministra gasolina una vegada a la setmana. La demanda setmanal X en milions de litres de gasolina segueix la següent funció de densitat:

$$f(x) = \begin{cases} c(3 - x) & 0 < x < 1 \\ 0 & en\ altres\ casos \end{cases}$$

a) Quin és el valor de c perquè f(x) sigui funció de densitat?
b) Quina és la capacitat total dels seus tancs si la probabilitat que s'exhaureixi la gasolina en una determinada setmana és 0,01?

Exercici 20. Els ingressos mensuals de les famílies d'un determinat barri de Barcelona es xifren en 900€ de mitjana amb una desviació estàndard de 60€.
a) Quin percentatge mínim de famílies té uns ingressos mensuals entre 780 i 1020€?
b) Determineu el percentatge màxim de famílies amb uns ingressos mensuals que difereixin de la mitjana en més de 150 €.
c) Trobeu un interval centrat en la mitjana que contingui els ingressos mensuals d'una proporció mínima del 90% de famílies.

CAPÍTOL II. DISTRIBUCIONS UNIDIMENSIONALS DE PROBABILITAT

2.1 INTRODUCCIÓ

Com ja hem vist, el comportament de qualsevol fenomen aleatori es pot descriure en termes de probabilitat mitjançant una variable aleatòria, identificada per la seva distribució de probabilitat.

En aquest capítol s'inclouen diverses distribucions de probabilitat, cadascuna de les quals rep un nom propi donat que l'experiència demostra que nombrosos fenòmens observats en l'àmbit de les ciències socials i experimentals es poden modelitzar adequadament per una d'aquestes distribucions. De fet, cadascun d'aquests models és una família de distribucions de probabilitat caracteritzada per una funció de quantia o de densitat genèrica, essent necessari, en cada cas, no només identificar la família adient, és a dir, el model, sinó també identificar la distribució concreta fixant el valor apropiat dels paràmetres que la caracteritzen.

Alguns d'aquests models són útils per a la descripció de fenòmens de naturalesa discreta i d'altres per a la descripció de fenòmens de naturalesa contínua. En cada cas es descriuen breument les situacions en les quals és adequat utilitzar cada model, així com les característiques més rellevants de cada distribució.

2.2 DISTRIBUCIONS DISCRETES

2.2.1 DISTRIBUCIÓ DE BERNOULLI O DICOTÒMICA

Sovint estem interessats en estudiar el comportament d'alguna variable que només pot prendre dos possibles resultats, per exemple, en observar un producte comprovar si compleix o no unes determinades condicions de qualitat, en observar un pacient veure si presenta o no una determinada malaltia, en llançar una moneda veure si el resultat és cara. Aquest últim és l'exemple més simple, però hi ha infinitat de situacions en les quals els esdeveniments possibles són dues alternatives complementàries. Totes aquestes situacions de dicotomia es poden fer formalment equivalents al llançament d'una moneda, on tenim una probabilitat **p** d'obtenir 'èxit' (o cara) i una probabilitat de 'fracàs' (o creu) **q**=**1-p**. La distribució de probabilitat de variables dicotòmiques s'anomena distribució de Bernoulli, ja que va ser estudiada per J.Bernoulli (1667-1748).

Definició:

Una variable aleatòria discreta X diem que presenta una distribució de ***Bernoulli*** si la seva funció de quantia és:

$$P(X=x)=\begin{cases} p^x q^{1-x} & x=0,1 \\ 0 & \textit{en altres casos} \end{cases}$$

on $0\leq p\leq 1$, $q=1-p$.

Característiques:

1. La variable dicotòmica només pren dos valors: x=1 (èxit) x=0 (fracàs).
2. La distribució dicotòmica depèn únicament del paràmetre **p** o probabilitat d'obtenir èxit.
3. L'esperança matemàtica és $\mu=p$.

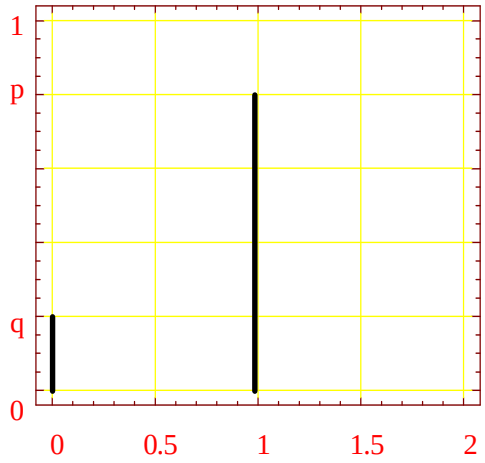
$\mu=E(X)=\sum_{\forall x} xP(x)=1\ p+0\ q=p$

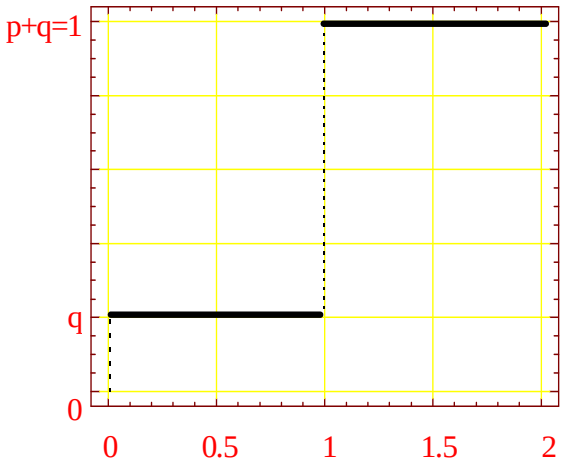
4. La variància és $\sigma^2=p\ q$.

$\sigma^2=V(X)=E(X^2)-\mu^2=\sum_{\forall x} x^2 P(x)-\mu^2=1^2\ p+0^2\ q-p^2=p\ (1-p)=p\ q$

5. La funció de distribució és $F(x)=\begin{cases} 0 & x<0 \\ q & 0\leq x<1 \\ p+q=1 & x\geq 1 \end{cases}$

Gràfic 2.1 Funció de quantia i funció de distribució de X~Bernoulli(p).





2.2.2 DISTRIBUCIÓ BINOMIAL

La distribució binomial s'obté com a generalització del procés de Bernoulli. Per exemple, suposem que es llança una moneda n vegades i que es defineix la variable aleatòria X com el '*nombre de vegades que ha sortit cara en els n llançaments*'. En aquest cas, la variable aleatòria pot prendre els valors sencers de 0 a n. (En els casos extrems, o bé no sortirà cap cara x=0, o bé sempre sortirà cara x=n.)

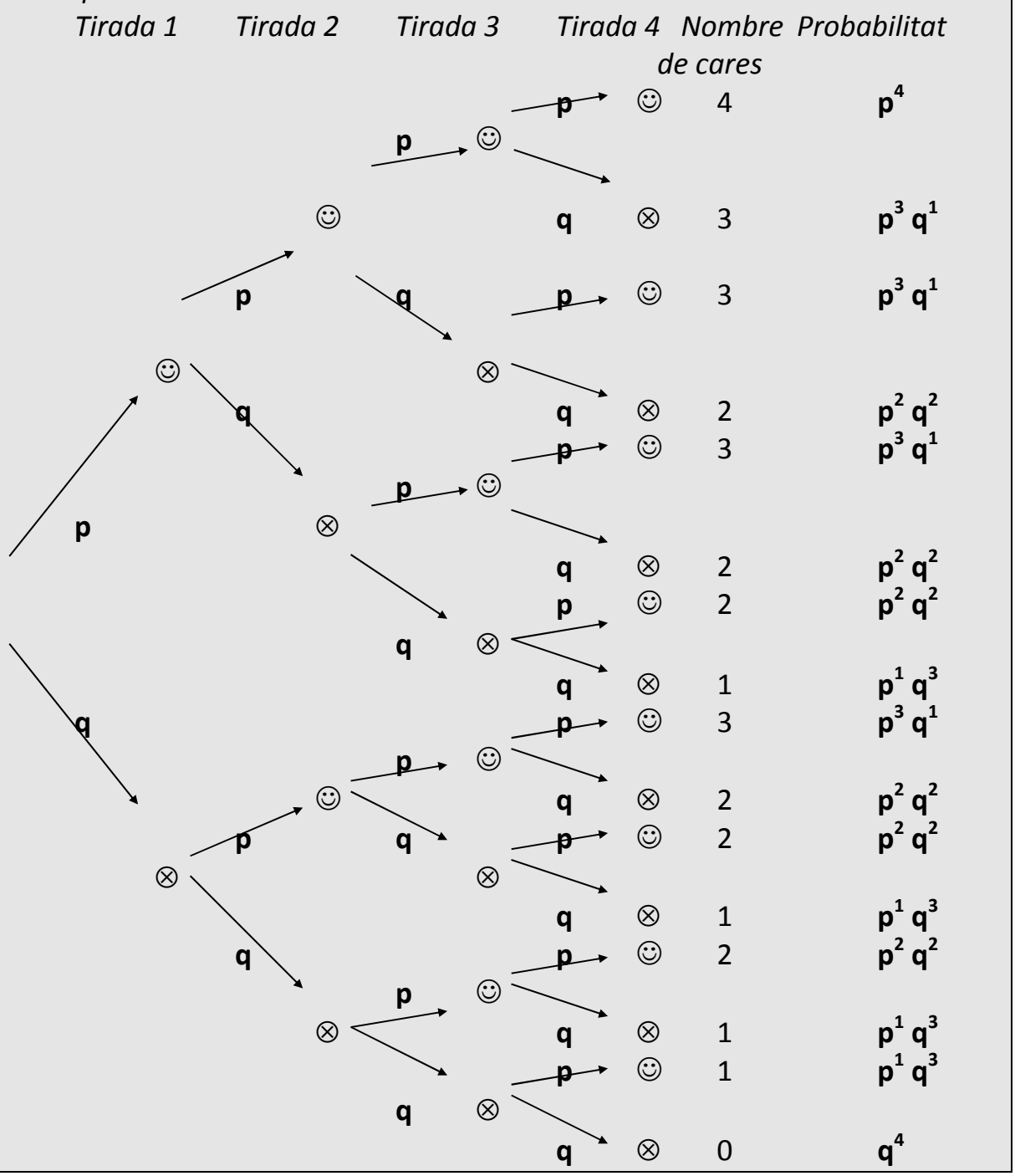
Si suposem que es realitzen n llançaments independents i que la probabilitat d'obtenir cara és un valor p que es manté constant en tots els llançaments, aleshores la variable aleatòria X, que recull el nombre total de cares obtingudes, presenta una distribució binomial de paràmetres n i p i es simbolitza com X~B(n,p).

La probabilitat de cada resultat de l'experiment depèn de **p**, per tant, les probabilitats de cada esdeveniment de la variable '*Nombre de cares en 4 llançaments*' es poden resumir:

Nombre de cares	Probabilitat
0	$\mathbf{q^4 = \binom{4}{0} q^4}$
1	$\mathbf{4\ p^1\ q^3 = \binom{4}{1}\ p^1\ q^3}$
2	$\mathbf{6\ p^2\ q^2 = \binom{4}{2}\ p^2\ q^2}$
3	$\mathbf{4\ p^3\ q^1 = \binom{4}{3}\ p^3\ q^1}$
4	$\mathbf{p^4 = \binom{4}{4} p^4}$

Una manera fàcil de veure quin és el comportament de la variable X és analitzar constructivament com es poden obtenir els diferents resultats.

Taula 2.1 Esdeveniments possibles en 4 llançaments d'una moneda, nombre de cares i probabilitat del succés.



En la taula 2.1 es veuen els esdeveniments possibles que es poden obtenir en llançar quatre vegades una moneda.

La situació anterior es pot generalitzar per obtenir la funció de quantia de la variable aleatòria binomial de paràmetres n i p.

Definició:

La variable aleatòria X que recull el '*nombre d'èxits obtinguts*' en un experiment aleatori que consisteix en realitzar n proves idèntiques i independents, on cadascuna només pot prendre dos possibles resultats que indicarem amb un 1 'èxit' i un 0 'fracàs', és una variable amb distribució ***Binomial***, B(n,p), si la probabilitat d'èxit, p, es manté constant en les n proves.

La seva funció de quantia és:

$$P(X=x)=\binom{n}{x}p^x q^{n-x}$$

per a x= 0,1,2,.....,n amb $0\leq p\leq 1$ i q=1-p.

Es pot comprovar que aquesta funció és sempre no negativa (tant p com q són positius) i compleix que $\sum_{i=0}^{n} P(X=x_i) = \sum_{i=0}^{n}\binom{n}{i}p^i q^{n-i} = (p+q)^n = 1$.

Característiques:

1. La variable binomial pren els valors sencers entre 0 i n.
2. La distribució binomial queda totalment caracteritzada amb dos paràmetres: **n** (nombre de proves independents) i **p** (probabilitat d'èxit).
3. La variable binomial s'obté com a suma de n variables dicotòmiques independents: $X=\sum X_i$ on $X_i \sim B(1,p)\ \forall i=1,...,n$.
4. L'esperança matemàtica és $\mu = np$

$\mu = E(X) = E(X_1+X_2+...+X_n) = E(X_1)+E(X_2)+...+E(X_n) = p+p+...+p = np$.

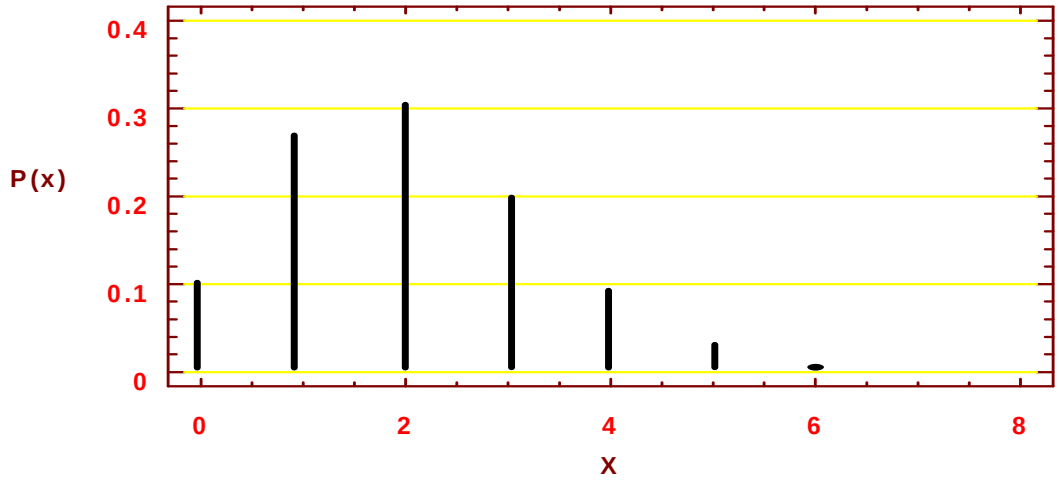
5. La variància és $\sigma^2 = npq$

$\sigma^2=V(X)= V(X_1+X_2+...+X_n) = V(X_1)+V(X_2)+...+V(X_n) = pq+pq+...+pq = npq$.

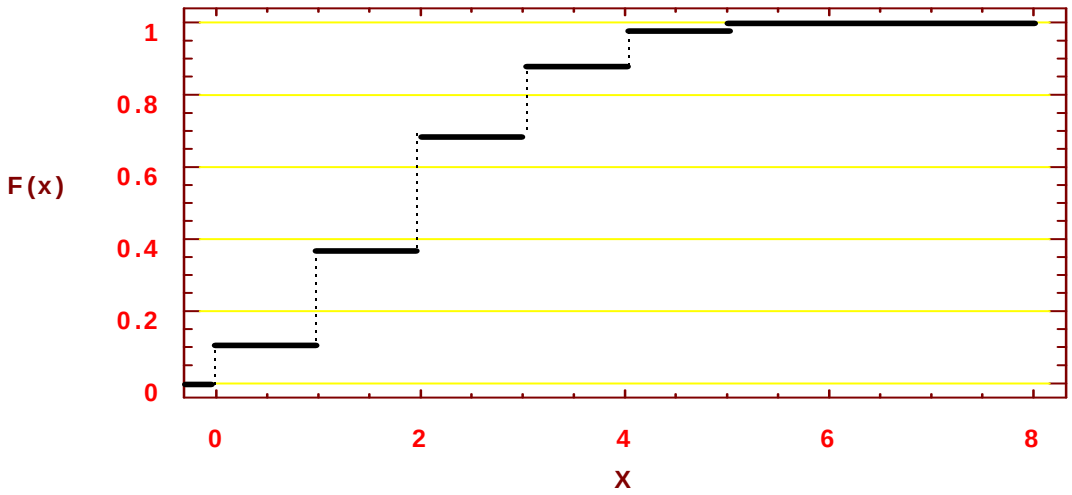
6. La funció de distribució és $F(x)= \begin{cases} 0 & x<0 \\ \sum_{i=0}^{x}\binom{n}{i}p^i q^{n-i} & 0\leq x<n \\ 1 & x\geq n \end{cases}$
7. La distribució binomial és reproductiva en el paràmetre p: en sumar dues o més variables binomials independents amb igual paràmetre p s'obté una nova distribució binomial amb paràmetres $B(\Sigma n_i,p)$.
8. La distribució binomial presenta asimetria positiva quan p<0,5 i asimetria negativa quan p>0,5. La asimetria es va reduint a mesura que p s'aproxima a 0,5, cas en el qual la distribució és simètrica. Així mateix, per a qualsevol valor de p, la asimetria disminueix quan augmenta el valor del paràmetre n.

La funció de quantia d'una distribució binomial es pot representar gràficament amb un diagrama de barres (gràfic 2.2) i la de distribució amb un diagrama escalonat (gràfic 2.3).

Gràfic 2.2 Funció de quantia X~B(10; 0,2).



Gràfic 2.3 Funció de distribució X~B(10; 0,2).



Exemple 2.1

Un fabricant empaqueta els seus articles en caixes de 5 unitats. Si la probabilitat que un article sigui defectuós és del 5%:

a) *Quina és la probabilitat que una caixa contingui més de 2 articles defectuosos?*
b) *Si es considera retornable una caixa quan conté algun article defectuós, quina és la probabilitat que en enviar una caixa triada a l'atzar aquesta es retorni?*

c) Quina és la probabilitat que en enviar 10 caixes a un client més de 2 siguin retornades?

Solució:

X= {*nombre d'articles defectuosos per caixa*} ~ B(5; 0,05)

a) $P(X>2) = P(X=3) + P(X=4) + P(X=5) = \binom{5}{3} 0{,}05^3\, 0{,}95^2 + \binom{5}{4} 0{,}05^4\, 0{,}95^1 +$

$$+\binom{5}{5} 0{,}05^5\, 0{,}95^0 = 10 \cdot 0{,}05^3\, 0{,}95^2 + 5 \cdot 0{,}05^4\, 0{,}95 + 1 \cdot 0{,}05^5 = 0{,}0012$$

Amb la funció de distribució (taula 2):

$P(X>2) = 1-P(X\leq 2) = 1 - F(2) = 1 - 0{,}9988 = 0{,}0012$

b) Probabilitat de retornar una caixa

$$P(X>0) = 1 - P(X\leq 0) = 1-P(X=0) = 1-\binom{5}{0} 0{,}05^0\, 0{,}95^5 = 1 - 0{,}7738 = 0{,}2262$$

c) X' = {*nombre de caixes retornades entre 10*} ~ B(10; 0,2262)

$P(X'>2) = 1-[P(X'=0)+P(X'=1)+P(X'=2)] =$

$$= 1-[\binom{10}{0} 0{,}2262^0\, 0{,}7738^{10} + \binom{10}{1} 0{,}2262^1\, 0{,}7738^9 + \binom{10}{2} 0{,}2262^2\, 0{,}7738^8]$$

$$= 1-0{,}5978=0{,}4022$$

2.2.3 DISTRIBUCIÓ GEOMÈTRICA

La variable aleatòria geomètrica recull '*el nombre de fracassos fins a arribar a aconseguir el primer èxit*' obtinguts en un procés dicotòmic, és a dir, en experiments consistents en realitzar proves idèntiques i independents on només es poden donar dos resultats, èxit (1) o fracàs (0), amb una probabilitat d'èxit **p** que es manté constant a totes les proves.

Per determinar la funció de quantia d'aquesta variable analitzem els resultats que poden aparèixer en realitzar aquest experiment:

- èxit en la primera prova, aleshores X=0 amb una probabilitat P(X=0)= p,
- èxit en la segona prova, X=1 amb una probabilitat P(X=1) = qp,
- èxit en la tercera prova, X=2 amb una probabilitat $P(X=2) = q^2p$,

...

- èxit a la n-èsima prova, X=n-1 amb una probabilitat $P(X=n-1) = q^{n-1}p$.

Definició:

La variable aleatòria discreta X que recull el '*nombre de fracassos abans d'obtenir el primer èxit*' en un procés dicotòmic segueix una distribució ***Geomètrica***, X~G(p), i presenta la següent funció de quantia:

$$P(X=x) = q^x p$$

per a x=0,1,2,...amb $0 \leq p \leq 1$ i q = 1-p.

Es pot comprovar que aquesta funció compleix les condicions necessàries per a ser de quantia: és sempre no negativa (tant p com q són positius) i $\sum_{i=0}^{\infty} q^i p = p\sum_{i=0}^{\infty} q^i = p(1+q+q^2+...) = p\left(\frac{1}{1-q}\right) = 1$.

Característiques:

1. La variable geomètrica és una variable discreta que pren els valors sencers entre 0 i $+\infty$.
2. La distribució de probabilitat queda totalment caracteritzada pel paràmetre **p** (probabilitat d'èxit).
3. L'esperança matemàtica és $\mu = \frac{q}{p}$.
4. La variància és $\sigma^2 = \frac{q}{p^2}$.
5. La moda és x=0.
6. La distribució presenta sempre asimetria a la dreta o positiva.
7. La funció de distribució és $F(x) = \begin{cases} 0 & x < 0 \\ p\sum_{i=0}^{x} q^i = 1 - q^{x+1} & x \geq 0 \end{cases}$
8. La distribució geomètrica no és reproductiva en el paràmetre p. En sumar dues o més v.a. geomètriques independents amb igual paràmetre obtenim una v.a. binomial negativa com veurem a l'apartat següent.
9. La distribució geomètrica no té memòria, és a dir, $P(X \geq a+b / X \geq a) = P(X \geq b)$.

Exemple 2.2

Suposem que el 15% dels aspirants a un determinat lloc de treball tenen el First Certificate (FC). Els aspirants són seleccionats a l'atzar i entrevistats un a un.

a) Calculeu la probabilitat que el primer aspirant amb FC sigui el cinquè entrevistat.

b) Si fins a la dècima entrevista no s'ha trobat cap aspirant amb FC, quina és la probabilitat de trobar-lo a la propera entrevista?

c) Quin és el nombre esperat d'entrevistes que s'han de realitzar per trobar un aspirant amb FC? I quina és la seva variància?

Solució:

X={*nombre d'aspirants entrevistats sense FC fins a trobar el primer amb* FC}, X~G(0,15)

a) $P(X=4)= 0{,}85^4 \cdot 0{,}15=0{,}0783$

b) $P(X=10/X\geq 10)=\frac{P(X=10)}{P(X\geq 10)}=\frac{pq^{10}}{1-P(X\leq 9)}=\frac{pq^{10}}{1-(1-q^{10})}$ =p=P(X=0)=0,15

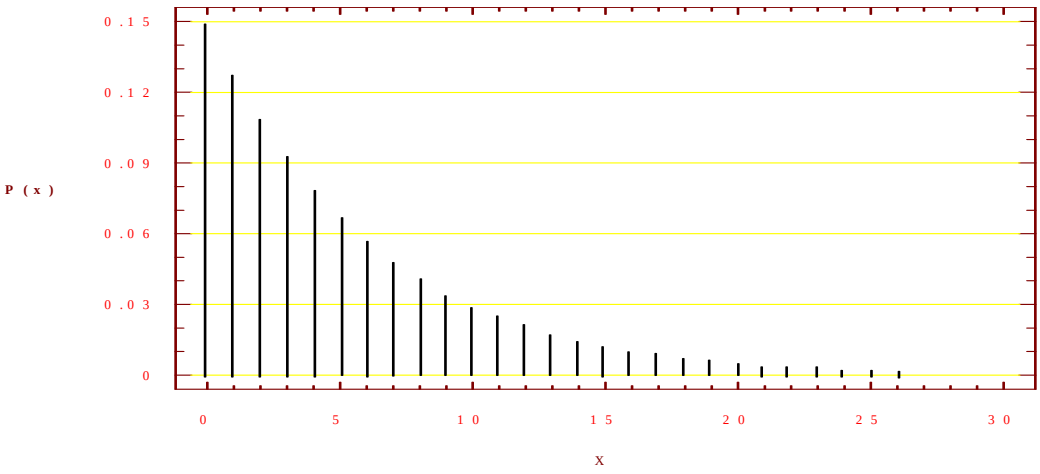
La distribució no té memòria.

c) E(X)=q/p=0,85/0,15=5,67 entrevistes.

$V(X)=q/p^2=0{,}85/0{,}15^2=37{,}78$

σ=6,146 entrevistes.

Gràfic 2.4 Funció de quantia G(p=0,15).



2.2.4 DISTRIBUCIÓ BINOMIAL NEGATIVA

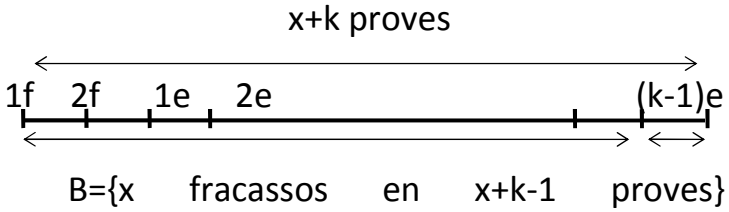
El procés que genera una variable aleatòria binomial negativa és el mateix que el de la variable binomial o el de la geomètrica. És a dir, modelitza experiments consistents en realitzar proves idèntiques i independents, en cadascuna de les quals només es poden obtenir dos resultats, èxit o fracàs, amb una probabilitat

d'èxit **p** que es manté constant a totes les proves. La variable aleatòria binomial negativa recull '*el nombre de fracassos aconseguits fins a arribar al k-èsim èxit*'.

Per deduir la funció de quantia d'aquesta variable o la probabilitat d'obtenir x fracassos abans del k-èsim èxit, P(X=x), definim els successos:

A = {obtenir el k-èsim èxit}

B = {obtenir x fracassos abans del k-èsim èxit en x+k-1 proves}.



B={x fracassos en x+k-1 proves}

P(X=x) = P(A∩B)=P(A)·P(B), ja que A i B són independents.

P(A)=P(X=0)=p.

P(B)=P(obtenir x fracassos en x+k-1 proves)= P(W=x) = $\binom{x+k-1}{x} q^x p^{k-1}$,

on W={Nombre de fracassos en x+k-1 proves}~B(x+k-1; q).

P(X=x)= P(A)·P(B)= p $\binom{x+k-1}{x} q^x p^{k-1} = \binom{x+k-1}{x} q^x p^k$.

Definició:

La variable aleatòria discreta X que recull el '*nombre de fracassos abans d'obtenir el k-èsim èxit*' en un procés dicotòmic segueix una distribució ***binomial negativa***, X~BN(k,p), i presenta la següent funció de quantia:

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

per a x=0,1,2,...amb 0≤p≤1 i q = 1-p.

Es pot comprovar que aquesta funció és sempre no negativa (tant p com q són positius) i la suma de probabilitats és 1: $\sum_{i=0}^{n} P(X=x_i) = \sum_{i=0}^{\infty} \binom{i+k-1}{i} q^i p^k = 1$.

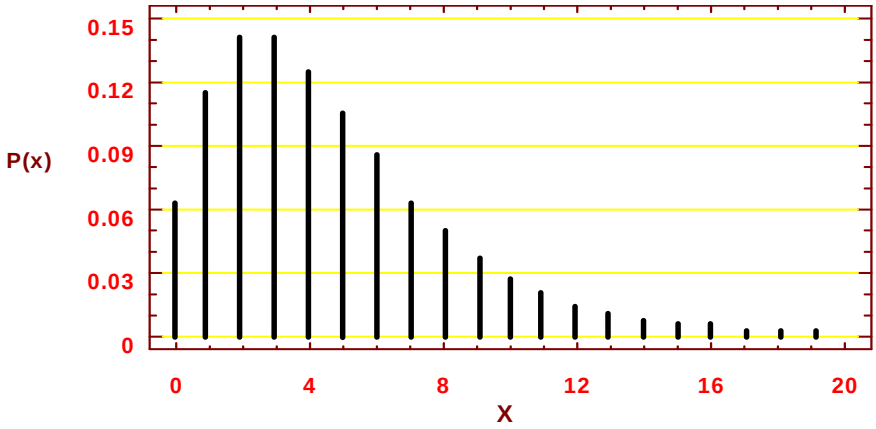
Característiques:

1. La variable binomial negativa és una variable discreta que pren els valors sencers entre 0 i +∞.

2. Queda totalment caracteritzada pels paràmetres **p** (probabilitat d'èxit) i **k** (nombre d'èxits).
3. L'esperança matemàtica és $\mu = k\frac{q}{p}$.
4. La variància és $\sigma^2 = k\frac{q}{p^2}$.
5. La distribució sempre presenta asimetria positiva, si bé la asimetria disminueix en augmentar k.
6. La variable binomial negativa és la suma de n variables geomètriques idèntiques i independents.
7. La variable binomial negativa és reproductiva en el paràmetre p.

El gràfic 2.5 recull el diagrama de barres de la funció de quantia d'una distribució BN(k; p).

Gràfic 2.5 Funció de quantia BN(k=3, p=0,4).



Exemple 2.3

Suposem que el 15% dels aspirants a un determinat lloc de treball tenen el First Certificate (FC). Els aspirants són seleccionats a l'atzar i entrevistats un a un.

a) Calculeu la probabilitat que el tercer aspirant amb FC sigui el vuitè entrevistat.

b) Quin és el nombre esperat d'entrevistes que s'han de realitzar per trobar 3 aspirants amb FC? I quina és la seva variància?

Solució:

X={*nombre d'aspirants entrevistats sense FC fins a trobar el tercer amb FC*}, X~BN(3;0,15).

a) P(X=5) = $\binom{x+k-1}{x} q^x p^k = \binom{7}{5} 0{,}85^5 0{,}15^3 = 0{,}0314$

b) E(X)=k$\frac{q}{p}$ =3·0,85/0,15=17 entrevistes

V(X)=k$\frac{q}{p^2}$ =3·0,85/0,15^2=113,33

σ=10,6 entrevistes

2.2.5 DISTRIBUCIÓ HIPERGEOMÈTRICA

La variable aleatòria hipergeomètrica, a l'igual que la binomial, recull '*el nombre d'èxits obtinguts en realitzar n proves dicotòmiques*' (amb dos únics resultats cadascuna: 0 'fracàs' i 1 'èxit'). A diferència de la binomial, aquí la probabilitat d'èxit, p, no es manté constant al llarg de l'experiment, ja que les proves són dependents, per ser el resultat d'extraccions *sense reposició* realitzades en poblacions finites amb un nombre relativament petit d'elements.

Definició:

Diem que X és una variable aleatòria discreta amb distribució ***hipergeomètrica***, que indicarem per H(N, n, K), si recull el '*nombre d'èxits obtinguts en realitzar n extraccions sense reposició*' d'una població de N elements on K són èxits. La funció de quantia o la probabilitat d'obtenir x èxits en n extraccions és:

$$P(X=x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

per a x=0,1,2,..., n amb $x \leq K$ $n-x \leq N-K$ i N, n i K sencers positius.

Es pot demostrar que aquesta funció compleix les condicions de funció de quantia: $P(X=x) \geq 0 \ \forall x$ i $\sum_{\forall x} P(X=x) = 1$.

Característiques:

1. La variable hipergeomètrica és una variable discreta que pren els valors sencers entre 0 i n.

2. Queda totalment caracteritzada pels paràmetres **N** (grandària de la població), **n** (grandària de la mostra) i **K** (nombre d'èxits a la població).
3. Les extraccions dels n elements es realitzen sense reposició, per tant, són proves dependents.
4. La proporció inicial d'èxits és $p=\frac{K}{N}$ i la proporció inicial de fracassos és $q=1-p=\frac{N-K}{N}$.
5. L'esperança matemàtica és $\mu = np$.
6. La variància és $\sigma^2= npq\frac{N-n}{N-1}$.
7. La funció de distribució és $F(x)=\frac{\sum_{i=0}^{x}\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}$,

per a x=0,1,2,..., n amb $x\leq K$ $n-x\leq N-K$ i N, n i K sencers positius.

8. La distribució hipergeomètrica convergeix a la distribució binomial quan p es manté estable, és a dir, quan N i K són suficientment grans. A la pràctica l'aproximació s'accepta per a $p<0,1$ i $N\geq 50$.

Exemple 2.4

Una fàbrica produeix un determinat article amb dues màquines A i B. La màquina A produeix a un ritme de 20 unitats diàries i la màquina B produeix 15 unitats/dia. Si escollim una mostra de 10 unitats de la producció d'un dia qualsevol, quina és la probabilitat que 6 d'aquestes unitats hagin estat fabricades per la màquina B?

Solució:

X= {*nombre d'unitats diàries produïdes per la màquina B*} X~H(35;10;15)

$P(X=x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \Rightarrow P(X=6) = \frac{\binom{15}{6}\binom{20}{4}}{\binom{35}{10}} = 0,132$

2.2.6 DISTRIBUCIÓ DE POISSON

Sovint la variable d'interès correspon al nombre d'esdeveniments que es produeixen dins d'un suport continu, per exemple, el nombre d'avaries d'una màquina al llarg del temps; el nombre de vehicles que arriben a un peatge en un interval de temps determinat; el nombre d'arbres per m^2; el nombre de pòlisses d'una determinada assegurança gestionades en un període de temps concret; etc... Aquests podrien ser exemples de variables aleatòries amb distribució de Poisson i posen de manifest que són variables discretes on la probabilitat d'obtenir un èxit és molt petita i el nombre de proves realitzades és molt gran ja que s'observen sobre un suport continu. Per tant, es podria dir que són experiments amb distribució binomial amb n gran i p molt petita (és per això, que a la distribució de Poisson també se l'anomena 'd'esdeveniments rars'). És així com S. Poisson (1781-1840) en generalitzar la llei de Bernoulli va formular aquesta distribució.

Definició:

Una variable aleatòria discreta X segueix una distribució de ***Poisson***, que indicarem per X~Pois(λ), quan la podem definir com 'el *nombre d'èxits ocorreguts dins d'un interval continu de temps, d'espai, de superfície,...*', i la seva funció de quantia és:

$$P(X=x)=\begin{cases} e^{-\lambda}\dfrac{\lambda^x}{x!} & x=0,1,2,...\ amb\ \lambda>0 \\ 0 & en\ altres\ casos \end{cases}$$

on λ és la mitjana d'èxits i es manté constant durant tot el procés.

Es pot demostrar que aquesta funció compleix les condicions necessàries per a ser funció de quantia: $P(X=x)\geq 0\ \forall x$ i $\sum_{i=0}^{\infty} P(X=x_i)=1$.

Característiques:

1. La distribució de Poisson depèn del paràmetre λ que es defineix com el nombre mitjà d'esdeveniments ocorreguts per unitat de temps. Aquest paràmetre s'ha de mantenir constant a l'interval.
2. Els esdeveniments han de produir-se de forma aleatòria i independent.
3. La variable aleatòria pren els valors sencers de 0 a +∞. A mesura que s'incrementa el valor de la variable les probabilitats individuals són cada vegada més petites.

4. La funció de distribució és F(x)= $e^{-\lambda}\sum_{i=0}^{x}\frac{\lambda^i}{i!}$ per a x=0,1,2...

5. L'esperança matemàtica i la variància de la distribució tenen el mateix valor i aquest és λ:

$\mu = \sigma^2 = \lambda$.

6. La desviació estàndard és $\sigma=\sqrt{\lambda}$.

7. La distribució de Poisson és el límit de la distribució binomial quan $n\to\infty$ i $p\to 0$. Aquesta relació que existeix entre la distribució de Poisson i la binomial permet que, quan la n és molt gran i la p és molt petita, es puguin aproximar les probabilitats binomials mitjançant els valors de la funció de quantia d'una distribució de Poisson. En concret, podem aproximar el càlcul de probabilitats d'una distribució binomial amb les anteriors característiques fent servir la funció de quantia de la distribució de Poisson, igualant la mitjana de la binomial a la mitjana i a la variància de la Poisson, és a dir, $\lambda = np$.

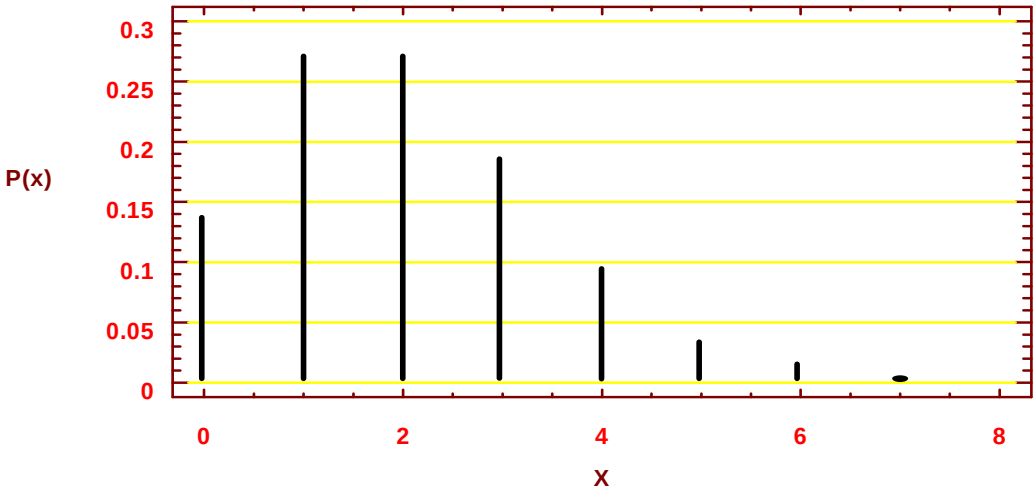
   A la pràctica l'aproximació de la distribució binomial a la de Poisson es recomana per a p<0,1 (o q<0,1) i np<5.

X~B(n,p) p<0,1 i np<5	→ X~Pois(λ=np)

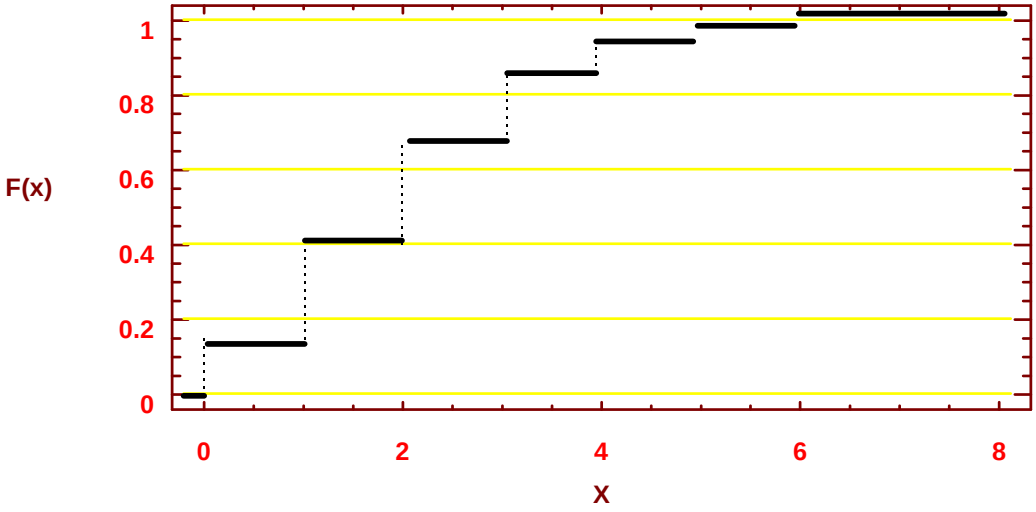
9. La distribució de Poisson és reproductiva. En sumar variables Poisson independents s'obté una nova variable Poisson de paràmetre $\lambda=\Sigma\lambda_i$.

10. La funció de quantia sempre presenta asimetria positiva que es redueix a mesura que augmenta el valor de λ.

Gràfic 2.6 Funció de quantia Pois(λ=2).



Gràfic 2.7 Funció de distribució Pois(λ=2).



Exemple 2.5

Una central telefònica rep per terme mitjà 2 trucades cada 15 minuts entre les 8 i les 22 hores. Si el nombre de trucades és una v.a. Poisson, calculeu:

a) La probabilitat que en un quart d'hora no es rebi cap trucada.

b) La probabilitat que entre les 9:15 i les 10:15 s'hagin rebut més de 6 trucades.

c) Si se sap que durant l'última mitja hora s'han rebut més de 3 trucades, quina és la probabilitat que el total de trucades hagi estat menys de 6?

d) Si aquesta central s'avaria i es passen les trucades a una altra central independent de l'anterior que rep un valor esperat d'1 trucada cada 5 minuts amb distribució Poisson, quina és la probabilitat que en un quart d'hora es rebin com a màxim 8 trucades?

Solució:

a) X={*nombre de trucades rebudes cada 15 minuts*} X~Pois(2)

$$P(X=0) = e^{-2}\frac{2^0}{0!} = 0{,}1353$$

b) Y={*nombre de trucades rebudes cada hora*} Y~Pois(8)

$$P(Y>6) = 1-P(Y\leq 6) = 1 - e^{-8}\left[\frac{8^0}{0!} + \frac{8^1}{1!} + \ldots + \frac{8^6}{6!}\right] = 1 - 0{,}31337 = 0{,}68663$$

c) Z={*nombre de trucades rebudes cada 1/2 hora*} Z~Pois(4)

$$P(Z<6/Z>3) = \frac{P(3<Z<6)}{P(Z>3)} = \frac{P(4)+P(5)}{1-P(Z\leq 3)} = \frac{e^{-4}(\frac{4^4}{4!}+\frac{4^5}{5!})}{1-e^{-4}(\frac{4^0}{0!}+...+\frac{4^3}{3!})} =$$

$$= \frac{0{,}1954+0{,}1563}{1-0{,}43347} = 0{,}6207$$

d) X={*nombre de trucades rebudes cada 15 minuts*} X~Pois(2)

W ={*nombre de trucades rebudes cada 5 minuts*} W~Pois(1)

Per tant, W'~Pois(3 trucades/15')

S=X+W' ~ Pois(5 trucades/15')

P(S≤8) = 0,93191

2.3 DISTRIBUCIONS CONTÍNUES

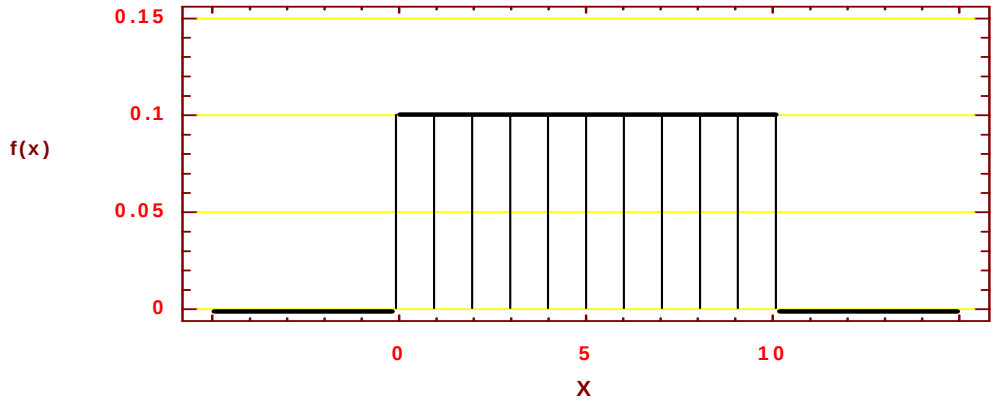
2.3.1 DISTRIBUCIÓ UNIFORME

És el model més senzill del comportament d'una variable aleatòria contínua. Permet modelitzar els resultats d'experiments que poden prendre qualsevol dels infinits resultats reals d'un interval [a; b], sempre que els subintervals d'igual amplitud presentin la mateixa probabilitat.

Suposem que definim la variable X '*temps (en minuts) que un viatger qualsevol haurà d'esperar un tren que surt cada 10 minuts*'. X és una variable contínua que pot prendre qualsevol valor real entre 0 i 10. Si dividim l'interval [0;10] en subintervals d'igual amplitud (per exemple 10 subintervals d'amplitud 1 minut) aleshores podem considerar la probabilitat de cadascun d'ells. Com que la probabilitat d'arribar a l'estació és la mateixa per a qualsevol moment del temps, podem considerar que la probabilitat que el viatger esperi més d'1 minut però menys de 2 és igual a la probabilitat que esperi més de 6 minuts però menys de 7, etc. És a dir, els subintervals anteriors [1;2] [6;7], d'igual amplitud, són equiprobables.

Per tant, un model adequat per a la funció de densitat de X queda recollit per una funció positiva definida en l'interval [0;10] que delimiti la mateixa àrea per a qualsevol subinterval d'igual amplitud.

Gràfic 2.8 Funció de densitat Uniforme U[0,10].



Àrea total = base x altura = 10 x K = 1

Perquè sigui funció de densitat, l'àrea total ha de ser igual a 1. Per tant, K ha de ser 1/10 i la probabilitat d'esperar, per exemple, entre 6 i 7 minuts és igual a 1/10.

Definició:

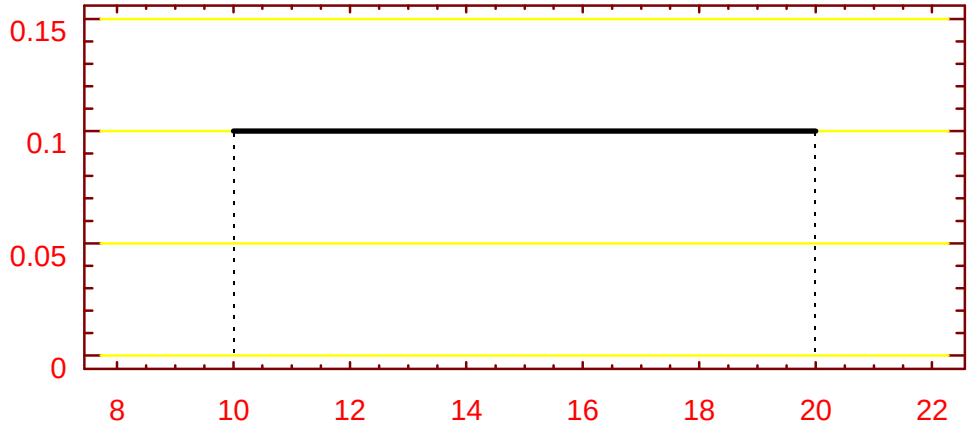
X és una variable aleatòria amb distribució ***Uniforme*** (o Rectangular) a l'interval [a,b], que indicarem per U[a,b], si la seva funció de densitat és constant per a tot el seu recorregut i igual a:

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \textit{en altres casos}\end{cases} \qquad 1$$

Característiques:

1. La distribució uniforme només depèn dels paràmetres **a** i **b,** extrems de l'interval.
2. La funció de distribució és $F(x)=\begin{cases}0 & x<a\\ \dfrac{x-a}{b-a} & a\le x\le b\\ 1 & x>b\end{cases}$
3. L'esperança matemàtica de X és el valor central de l'interval [a,b] $\mu=\dfrac{a+b}{2}$ .2
4. La variància és $\sigma^2=\dfrac{(b-a)^2}{12}$.
5. No és reproductiva.
6. La distribució no té moda.
7. La distribució és simètrica: Me=μ.

Gràfic 2.9 Funció de densitat U[10,20].



Gràfic 2.10 Funció de distribució U[10,20].

Exemple 2.6

Si durant l'última mitja hora un client i només un ha realitzat un ingrés a un determinat caixer automàtic, quina és la probabilitat que aquesta operació hagi estat realitzada durant els últims 5 minuts?

(Suposeu que l'ingrés es pot haver realitzat amb igual probabilitat en qualsevol moment al llarg de l'última mitja hora).

Solució:

X = {*Temps en minuts que ha passat des de la realització de l'ingrés*}

X ~ U[0,30]

$$f(x) = \begin{cases} \frac{1}{30} & 0 \le x \le 30 \\ 0 & \text{en } \textit{altres} \text{ cas os} \end{cases}$$

$P(25<X<30) = P(0<X<5) = \int_0^5 \frac{1}{30}\,dx = \left[\frac{1}{30}x\right]_0^5 = \frac{5}{30} = \frac{1}{6}3$

2.3.2 DISTRIBUCIÓ EXPONENCIAL

Si la distribució de Poisson determina el nombre d'esdeveniments ocorreguts per unitat de temps, la distribució exponencial determina *el temps que hem d'esperar entre un esdeveniment i el següent dins d'un procés de Poisson*. Per exemple, si a un taller de reparacions arriben segons un procés de Poisson 4 peces/hora per terme mitjà, vol dir que hem d'esperar un temps mitjà d'1/4 d'hora o 15 minuts entre l'arribada d'una peça i la següent segons un procés exponencial. Per tant, les dues distribucions estan relacionades entre si i depenen del mateix paràmetre λ.

En general, la distribució exponencial permet modelitzar fenòmens de tipus '*duració de vida d'un component*' (temps transcorregut fins que un element falla) i '*fenòmens d'espera*' (teoria de cues).

Definició:

Una variable aleatòria X presenta una distribució ***Exponencial*** (o exponencial negativa), que indicarem per X~Exp(λ), si es defineix com '*el temps d'espera entre dos esdeveniments consecutius d'un procés de Poisson*' i és una variable aleatòria contínua que té com a funció de densitat:

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0 \lambda>0 \\ 0 & en\ altres\ casos\end{cases}$$

Aquesta funció compleix les condicions necessàries per a ser funció de densitat:

$f(x) \geq 0\ \forall x \in R$ i $\int_0^{\infty} \lambda e^{-\lambda x} = -e^{-\lambda x}\Big]_0^{\infty} = 1$.

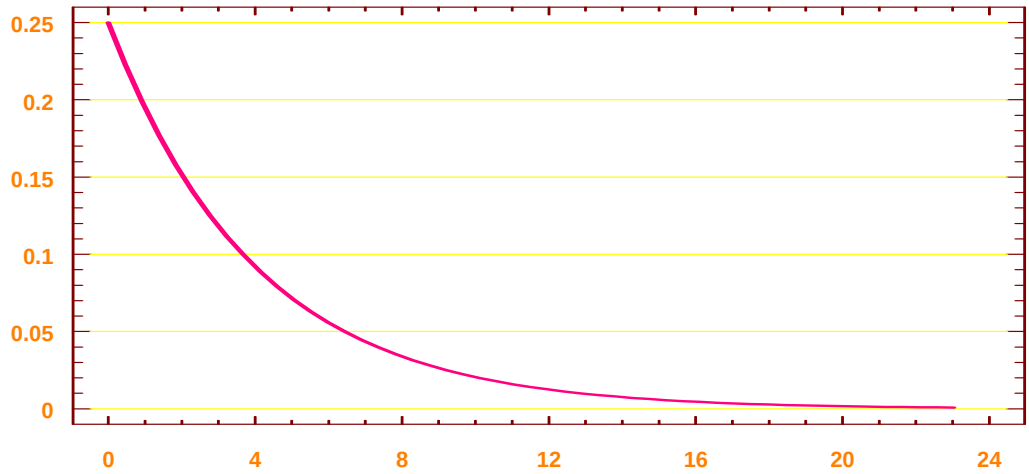
Característiques:

1. La distribució exponencial depèn del paràmetre λ de la distribució de Poisson, per tant, els esdeveniments han de ser independents i λ constant per unitat de temps.
2. La variable pren valors de 0 a $+\infty$.
3. La funció de distribució és $F(x)=P(X \leq x)=1-e^{-\lambda x} \quad x>0$.
4. L'esperança matemàtica és $\mu = \frac{1}{\lambda}$.
5. La variància és $\sigma^2 = \frac{1}{\lambda^2}$.

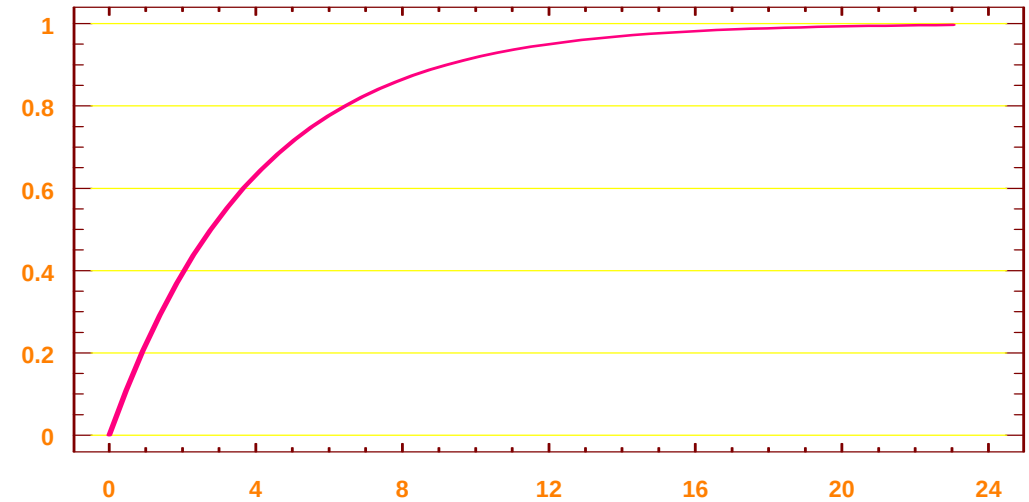
Per tant, l'esperança matemàtica de la variable coincideix amb la desviació estàndard: $\mu=\sigma=\frac{1}{\lambda}$.

6. La distribució exponencial no té memòria, és a dir, la probabilitat d'esperar un temps t_1 addicional és independent del temps d'espera ja transcorregut: $P(X>t_0+t_1/X>t_0) = P(X>t_1)$
7. No és reproductiva.
8. La distribució presenta asimetria positiva.
9. La moda és x=0.
10. La mediana és $x=\frac{-\ln 0{,}5}{\lambda}$.

Gràfic 2.11 Funció de densitat Exp(λ=4).



Gràfic 2.12 Funció de distribució Exp(λ=4).



Exemple 2.7

El departament de reparació d'una empresa rep una mitjana de 4 màquines per dia per a la seva reparació durant la jornada laboral (1 dia laboral = 8 hores). Sota el supòsit que el nombre de màquines que es reben es distribueix amb una llei de Poisson, es demana:

a) La probabilitat que en un dia com a màxim s'hagin de reparar 3 màquines.

b) El temps mitjà que transcorre entre l'arribada de dues màquines consecutives.

c) La probabilitat que una màquina trigui en arribar com a màxim 4 hores després que hagi arribat l'última.

d) Si se sap que durant les últimes 2 hores no s'ha rebut cap màquina, quina és la probabilitat que com a mínim transcorrin 2 hores més abans de rebre la següent?

Solució:

X ={*Nombre de màquines a reparar}* ~ Pois(λ = 4 màquines/dia)

a) P(X≤3)= P(X=0)+P(X=1)+P(X=2)+P(X=3) = $e^{-4}\left[\frac{4^0}{0!}+\frac{4^1}{1!}+\frac{4^2}{2!}+\frac{4^3}{3!}\right]$ = 0,4335

b) Y = {*Temps transcorregut entre l'arribada de dues màquines consecutives*}

Y ~ Exp(λ = 4 màquines/dia) = Exp(λ = 0,5 màquines/hora)

E(Y) = $\frac{1}{\lambda}=\frac{1}{0{,}5}$ = 2 hores

c) P(Y≤4) = 1 - $e^{-0,5\cdot 4}$ = 0,86466

d) P(Y>4/Y>2) = $\frac{P(Y>4)}{P(Y>2)}=\frac{e^{-\lambda 4}}{e^{-\lambda 2}}=e^{-\lambda 4+\lambda 2}=e^{-\lambda 2}$ = P(Y>2) = $e^{-0,5\cdot 2}$ = 0,3679

2.3.3 DISTRIBUCIÓ NORMAL

És la distribució contínua més important perquè recull el comportament poblacional de gran nombre de variables econòmiques, socials, biològiques, etc. i perquè la distribució de probabilitat de la majoria dels estadístics mostrals convergeix en aquesta distribució quan la grandària de la mostra és suficientment elevada.

Va ser A. De Moivre (1667-1754) el primer matemàtic que va deduir empíricament la distribució normal com el límit al qual convergeix la distribució binomial. Posteriorment, Laplace (1749-1827) i F.Gauss (1777-1855) la van formular, per això també s'anomena llei de Gauss-Laplace o campana de Gauss.

Definició:

Diem que la variable aleatòria contínua X presenta una distribució ***Normal*** de paràmetres μ i σ, que indicarem per X~N(μ, σ), si la seva funció de densitat és la següent:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2},\ \forall x\in R$$

on: $\mu\in R$, $\sigma>0$, $\pi=3{,}14$ i $e=2{,}71$.

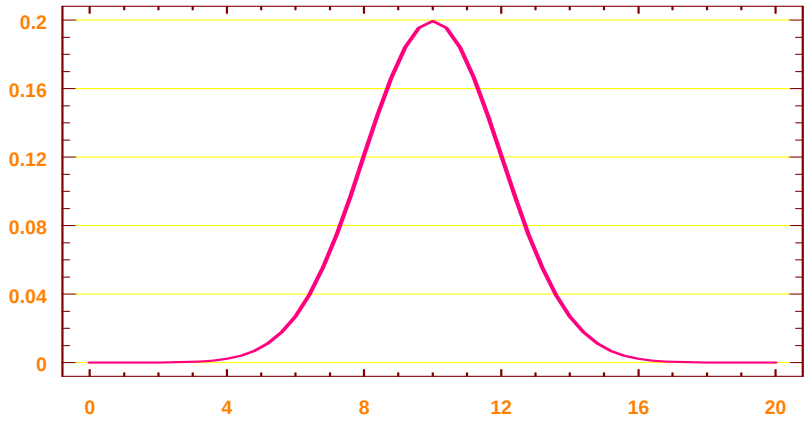
Característiques:

1. La distribució normal depèn de dos paràmetres: μ (esperança matemàtica) i σ (desviació estàndard de la variable).
2. La variable pren valors de -∞ a +∞.
3. La distribució és simètrica:

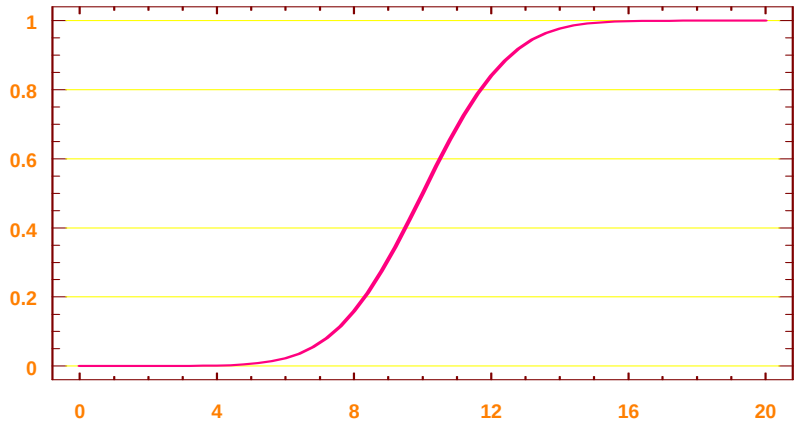
- El coeficient d'asimetria és $g_1=0$.
- Mitjana, mediana i moda coincideixen.
- L'àrea a la dreta i a l'esquerra de la recta x=μ és la mateixa i igual a 0,5.

4. La distribució és mesocúrtica. Presenta un coeficient de curtosi $g_2=0$.
5. La distribució normal presenta dos punts d'inflexió: x=μ-σ i x=μ+σ.
6. És asimptòtica respecte a l'eix d'abscisses.
7. La distribució normal és reproductiva; en sumar o restar dues o més variables normals independents s'obté una nova variable normal amb paràmetres $N(\Sigma\mu, \sqrt{\Sigma\sigma^2})$.

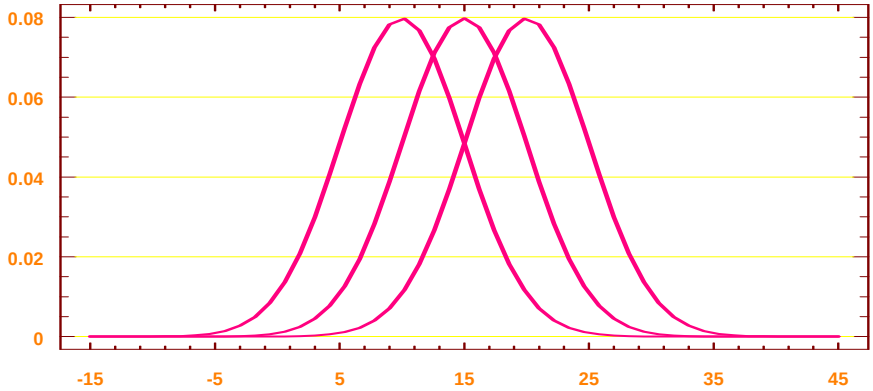
Gràfic 2.13 Funció de densitat d'una normal de paràmetres μ=10 i σ=2.



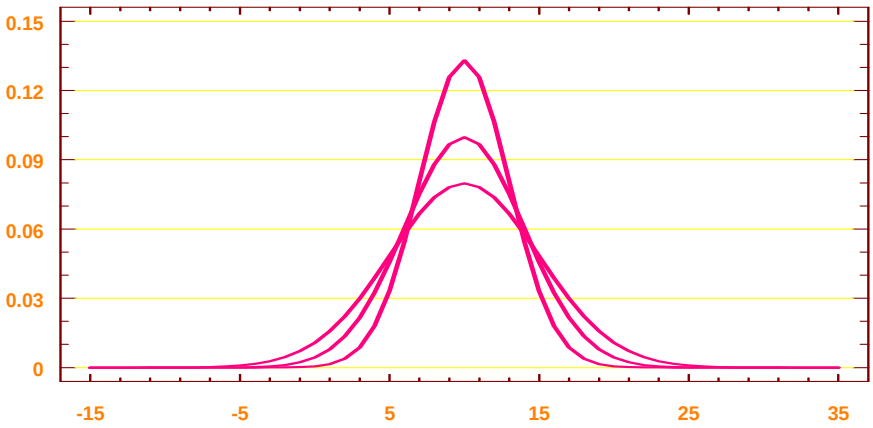
Gràfic 2.14 Funció de distribució d'una normal de paràmetres μ=10 i σ=2.



Gràfic 2.15 Corbes normals amb igual variància i diferents esperances.



Gràfic 2.16 Corbes normals amb igual esperança però diferents variàncies.



Per tal d'obtenir probabilitats referides a qualsevol distribució normal es disposa d'una taula, per a l'ús de la qual és necessari transformar la distribució N(μ, σ) en una distribució normal estandarditzada.

Definició:

Diem que una variable aleatòria normal presenta una distribució ***normal estandarditzada,*** o tipificada, si la seva esperança matemàtica és 0 i la seva variància és 1, i la simbolitzem per Z~N(0, 1).

Característiques:

1. La seva funció de densitat és $f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$ $\forall z\in$ R.
2. Presenta un màxim per a z=0.
3. Té dos punts d'inflexió z=-1 i z=+1.
4. P(Z<0) = P(Z>0) = 0,5.
5. Qualsevol variable normal X de paràmetres μ i σ es pot transformar en una normal estandarditzada simplement estandarditzant o tipificant la variable, és a dir, restant-li la seva esperança i dividint el resultat per la desviació estàndard. Per tant, donada X~N(μ, σ) es pot realitzar la transformació lineal Z=$\frac{X-\mu}{\sigma}$ i obtenir una variable normal estandarditzada, ja que:

E(Z)=E($\frac{X-\mu}{\sigma}$) = 1/σ [E(X) -μ] = 0.

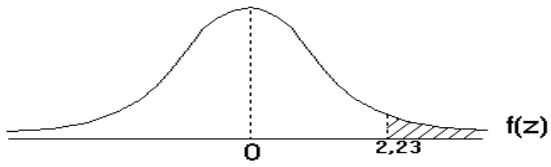
V(Z)=V($\frac{X-\mu}{\sigma}$) = 1/σ^2 V(X) = 1, aleshores Z~N(0,1).

Utilització de les taules.

L'annex Taules Estadístiques conté els valors de la funció de distribució de Z, F(z). És a dir, recull les probabilitats acumulades des de -∞ fins als valors de z especificats a la primera columna/primera fila, de forma que la primera columna hi ha el dígit sencer i el primer decimal de z i la primera fila recull el segon decimal de z. Així, per exemple, la P(Z<1,25) = F(1,25) es troba a la intersecció de la fila corresponent al número z=1,2 i la columna corresponent al 0,05, i s'obté F(1,25) = 0,89435; la P(Z<0,18) es troba a la intersecció de la fila 0,1 i la columna 0,08, F(0,18) = 0,57142.

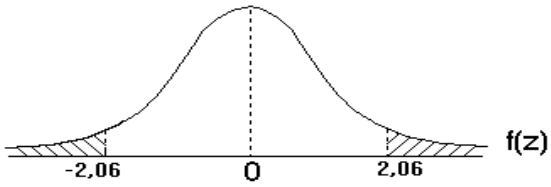
- Si la probabilitat que es vol determinar és a cua superior (probabilitat per excés) es calcularà per diferència. Així, per exemple:

  P(Z>2,32) = 1-P(Z<2,23) = 1-F(2,32) = 1-0,98983 = 0,01017

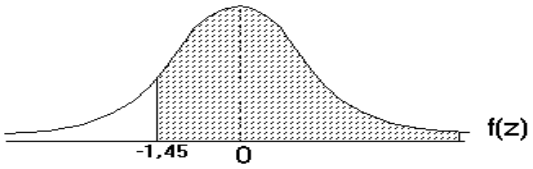


- Per a valors inferiors a 0 s'haurà d'obtenir la probabilitat per simetria ja que aquesta taula només presenta les probabilitats de valors de Z superiors o iguals a zero. Així, per exemple:

  P(Z<-2,06) = P(Z>2,06) = 1 - P(Z<2,06) = 1-0,9803= 0,0197.



- P(Z>-1,45) = P(Z<1,45) = F(1,45) = 0,92647



- Per calcular la probabilitat corresponent a qualsevol altra distribució normal s'haurà de transformar en una distribució estandarditzada. Així, per a X~N(μ, σ), per trobar la probabilitat que X sigui inferior a un determinat valor k, P(X<k):

  1. S'estandarditza: P(X<k)= P(Z<$\frac{k-\mu}{\sigma}$).

  2. S'obté F($\frac{k-\mu}{\sigma}$) en la taula de la normal Estandarditzada.

     Per exemple, donada una X~N(10, 5) es vol calcular la P(X<18):

     P(X<18) = P(Z<$\frac{18-10}{5}$) = P(Z<1,6) = 0,9452

Exemple 2.8

La distribució dels 'salaris anuals dels auxiliars administratius' (X) és una variable aleatòria normal d'esperança matemàtica de 1.500 u.m. i desviació estàndard de 500 u.m. Determineu:

a) La probabilitat que el salari d'un auxiliar triat a l'atzar sigui superior a 2250 u.m.

b) La probabilitat que el salari d'un auxiliar triat a l'atzar estigui entre 1000 i 2000 u.m.

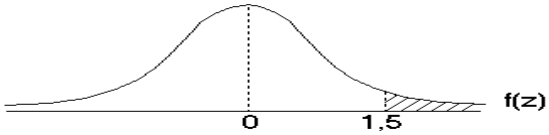
c) Quin salari com a mínim pot cobrar un auxiliar que es troba entre el 70% amb sous més alts?

d) Dels 10 auxiliars que té una empresa, quants es pot esperar que rebin salaris inferiors a 1500 u.m.? I quants reben salaris de com a mínim 1800 u.m.?

Solució:

$X \sim N(1500,500)$

a) $P(X \geq 2250) = P(Z > \frac{2250-1500}{500}4) = P(Z > 1,5) = 1 - F(1,5)$

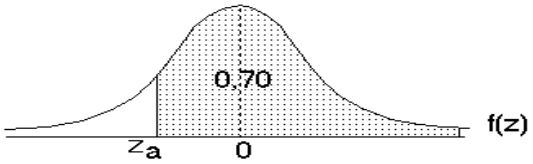


Per tant, $P(X \geq 2250) = 0,06681$

b) $P(1000<X<2000) = P(\frac{1000-1500}{500}5<Z<\frac{2000-1500}{500}6) = P(-1<Z<1) = F(1) - F(-1)$

$= 0,68268$

c) $P(X > a) = 0,70 \Rightarrow P(Z > \frac{a-1500}{500}7) = 0,70$



A les taules de la normal estandarditzada trobem que una probabilitat aproximada a 0,7 s'acumula per al valor z=0,52. Per simetria z_a= -0,52,

$z_a = \frac{a-1500}{500}8 = -0,52 \Rightarrow a = 1500-0,52\cdot500 = 1240$ u.m.

d) $P(X < 1500) = P(Z < 0) = 0,5 \Rightarrow n = 0,5\cdot10 = 5$ treballadors.

$P(X \geq 1800) = P(Z > 0,6) = 1 - F(0,6) = 0,27425 \Rightarrow n = 0,27425\cdot10 = 2,7 \approx 3$ treballadors.

Exemple 2.9

En pintar una porta s'han utilitzat 2 tipus de pintura, una per a la part interior i una altra per a la part exterior. El 'temps d'assecat' (en hores) de la part interior és una variable X~N(10; 2); i el de l'exterior una altra variable Y~N(8; 1,25). Si X i Y són independents,

a) Quina és la probabilitat que al cap de 9 hores almenys una de les dues pintures no s'hagi assecat?

b) Quina és la probabilitat que la pintura de la part interior trigui més del doble que la de l'exterior en assecar-se?

Solució:

X~N(10; 2)

Y~N(8; 1,25)

a) P(almenys una de les dues pintures no s'hagi assecat) = 1- P(les dues s'hagin assecat)

P(assecat l'interior) = P(X<9) = P(Z< $\frac{9-10}{2}$) = P(Z<-0,5) = 0,30854

P(assecat l'exterior) = P(Y<9) = P(Z< $\frac{9-8}{1,25}$) = P(Z<0,8) = 0,78814

P(les dues s'hagin assecat) = P(X<9) P(Y<9) = 0,30854 0,78814 = 0,24317

P(almenys una de les dues pintures no s'hagi assecat) = 1- 0,24317 = 0,75683

b) P(X>2Y) = P(X-2Y>0)

X-2Y ~ N(10-2·8; $\sqrt{2^2 + 2^2 1,25^2}$)

P(X-2Y>0) = P(Z>$\frac{0-(-6)}{\sqrt{10,25}}$) = P(Z>1,87) = 1-0,96926 = 0,03074

2.4 TEOREMA CENTRAL DEL LÍMIT

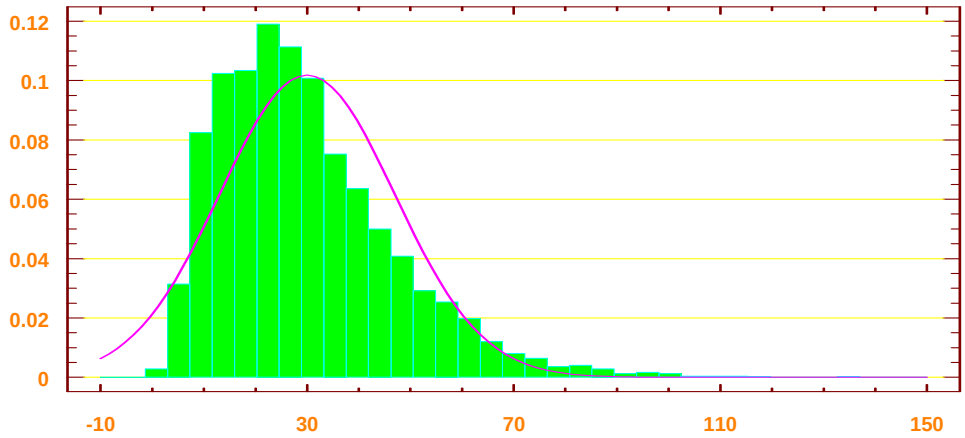
Com ja hem vist, la suma de n variables normals independents té també una distribució normal per a qualsevol grandària de n, però, quina és la distribució de la suma de n variables aleatòries independents quan la seva distribució no és normal?

Els gràfics següents representen la distribució de les sumes de 3 i 20 variables independents amb distribució exponencial de paràmetre λ = 0,1.

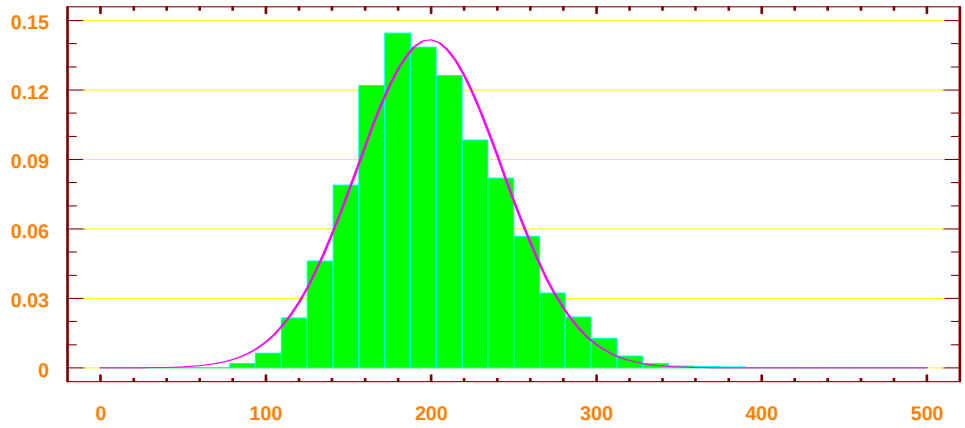
Com es pot veure en els gràfics 2.17 i 2.18, ambdues distribucions presenten un cert grau d'asimetria, que va disminuint a mesura que augmenta el nombre de variables sumades. Així, en sumar un nombre prou gran de variables aleatòries independents observaríem que la distribució és gairebé normal sigui quina sigui la distribució de probabilitat de les variables sumades; de fet un dels teoremes fonamentals de l'estadística, l'anomenat Teorema Central del Límit (TCL),

garanteix la convergència a la distribució normal de la suma d'un nombre prou gran de variables aleatòries independents.

Gràfic 2.17 Suma de tres variables independents amb distribució Exp(0,1).



Gràfic 2.18 Suma de vint variables independents amb distribució Exp(0,1).



TEOREMA de LINDEBERG-LEVY.

Donades n variables aleatòries independents i idènticament distribuïdes, X_1, X_2, ..., X_n, amb iguals valors esperats $\mu_1 = \mu_2 = ... = \mu_n = \mu$ i iguals variàncies $\sigma_1^2 = \sigma_2^2 = \sigma_n^2 = \sigma^2$, la distribució de probabilitat de la variable suma, $S = X_1 + X_2 + ... + X_n$ amb $E(S) = n\mu$ i $V(S) = n\sigma^2$, convergeix a una distribució normal quan n tendeix a infinit.

Aquest teorema té una gran importància a l'estadística donat el gran nombre de problemes on intervé una suma de variables aleatòries (per exemple, el càlcul de la mitjana aritmètica). Si s'aplica el TCL, la distribució de probabilitat de la mitjana aritmètica, o de qualsevol variable suma, es pot aproximar a la distribució normal si n és suficientment gran.

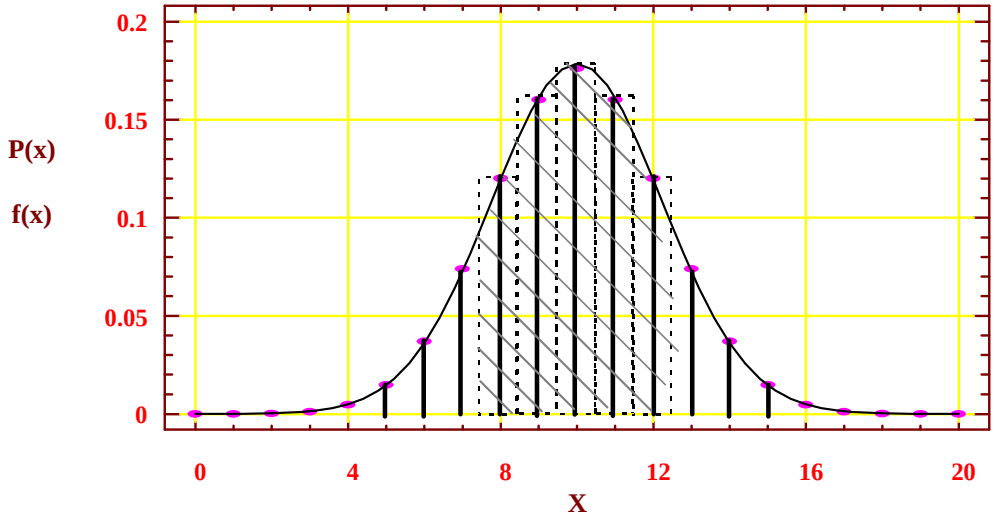
A la pràctica considerarem que per a n>30 la distribució de la variable suma, S, ja es pot aproximar mitjançant la normal:

$$S = X_1 + X_2 + \ldots + X_n \quad (n>30) \Rightarrow S \sim N(n\mu, \sqrt{n\sigma^2})$$

Un cas particular del teorema anterior és el formulat per ***De Moivre*** que va demostrar que si X és una variable binomial de paràmetres n i p, la distribució de $\frac{X-np}{\sqrt{npq}}$ convergeix a la distribució N(0,1). Això vol dir que la distribució binomial B(n,p) es pot aproximar per la distribució normal amb el mateix valor esperat i la mateixa desviació estàndard, és a dir, N(np, $\sqrt{npq}$) per a valors prou grans de n.

Vegeu en el gràfic 2.19 on efectivament l'àrea delimitada per la corba normal, per exemple de paràmetres N(20·0,5; $\sqrt{20 \cdot 0{,}5 \cdot 0{,}5}$), pràcticament coincideix amb l'àrea corresponent als rectangles amb base unitària centrats als valors puntuals de la distribució binomial de paràmetres B(20;0,5).

Gràfic 2.19 Distribucions N(10;2,136) i B(20;0,5).



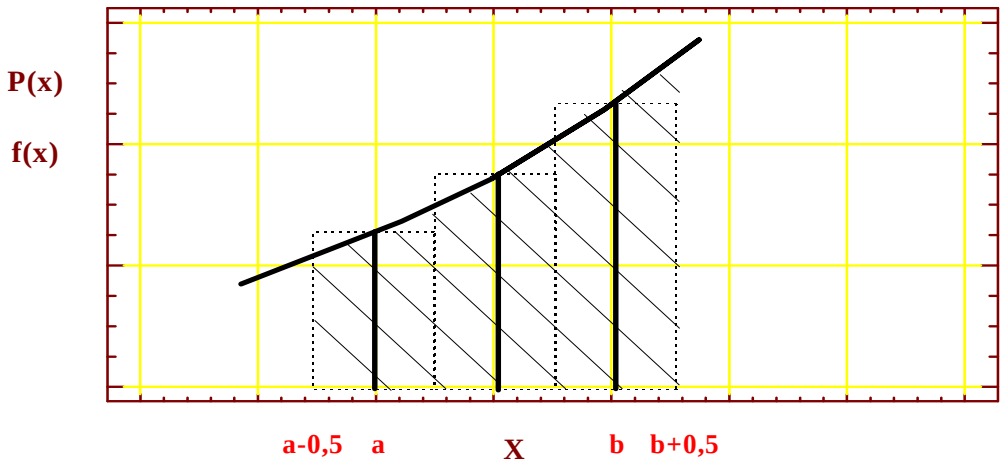
A la pràctica la convergència anterior l'aplicarem si el producte de n, p i q és superior a 5, i els paràmetres de la distribució normal seran:

$$X \sim B(n,p) \quad (npq > 5) \Rightarrow X \sim N(np, \sqrt{npq})$$

Quan s'utilitza aquesta aproximació, per al càlcul de probabilitats d'una binomial s'està cometent un determinat error. Aquest error es redueix sensiblement amb la ***correcció de continuïtat*** que suposa considerar el valor puntual x=a de la variable discreta binomial, per l'interval (a-0,5≤x≤ a+0,5) a la variable contínua normal.

En general, si es vol calcular la probabilitat P(a ≤ X ≤b), on X ~ B(n,p), per aproximació a la normal, N(np, $\sqrt{npq}$), per tal de millorar aquesta aproximació de la probabilitat, com es pot veure al gràfic 2.20, caldrà efectuar la correcció de continuïtat fent correspondre la P(a ≤ X ≤b) de la distribució discreta a la P(a-0,5 ≤ X ≤ b+0,5) de la distribució contínua.

Gràfic 2.20 Correcció de continuïtat.

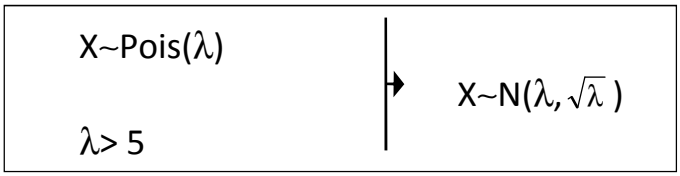


El gràfic 2.20 recull una secció de la superposició del diagrama de barres d'una distribució binomial i la funció de densitat de la normal a la què convergeix la distribució anterior. La probabilitat P(a≤X≤b) de la binomial és igual a la suma de les alçades de les barres incloses en l'interval [a;b] i, per tant, és igual a l'àrea total que presenten els rectangles de base u centrats, representats amb línia discontínua. Si aproximem aquesta àrea amb la continguda per la funció de densitat de la normal, queda clar que la solució òptima és buscar la probabilitat P(a-0,5≤X≤b+0,5).

En els casos d'intervals oberts o semioberts, la correcció de continuïtat és:

- P(a<X<b) ⇒ P(a+0,5≤X≤b-0,5)
- P(a<X≤b) ⇒ P(a+0,5≤X≤b+0,5)
- P(a≤X<b) ⇒ P(a-0,5≤X≤b-0,5)

La distribució normal també permet aproximar la distribució de Poisson quan $\lambda>5$ (amb la correcció de continuïtat):



Exemple 2.10

El departament de màrqueting d'una empresa concerta una mitjana de 5 entrevistes diàries a clients potencials dels seus productes, amb una desviació estàndard de 2 entrevistes/dia. El cap de departament té previst realitzar com a mínim 350 entrevistes per al proper trimestre (65 dies feiners). Si les entrevistes diàries es consideren independents, quina és la probabilitat d'assolir l'objectiu previst?

Solució:

X={*nombre d'entrevistes concertades}* amb $\mu_x=5$ i $\sigma_x=2$

$S=X_1+X_2+...+X_{65}$ $E(S)=65\mu_x=325$ i $V(S)=65\sigma_x^2= 65\cdot4=260$

$n=65 > 30 \Rightarrow S\sim N(325,\sqrt{260})$

$P(S\geq350)=P(Z\geq\frac{350-325}{\sqrt{260}})= P(Z\geq1,55)=1-0,93943=0,06057$

Exemple 2.11

Una empresa coneix per experiència que el 80% de les seves factures presenten pagament ajornat. Si s'extreu una mostra aleatòria de 50 factures, quina és la probabilitat de trobar més de 5 factures amb pagament al comptat?

Solució:

X={*nombre de factures amb pagament al comptat}*~B(50;0,2)

$npq=50\cdot0,2\cdot0,8=8 > 5 \Rightarrow X\sim N(50\cdot0,2;\sqrt{50\cdot0,2\cdot0,8})=N(10;\sqrt{8})$

$P(X>5)=$ {corrección de continuïtat}$=P(X\geq5,5)=P(Z\geq\frac{5,5-10}{\sqrt{8}})= P(Z\geq-1,59)=0,94408$

Exemple 2.12

L'error que es comet en arrodonir un nombre real fins al sencer més proper és una variable aleatòria amb distribució uniforme dins l'interval (-0,5; 0,5). Trobeu l'error màxim que es pot cometre en la suma de 108 nombres reals

triats a l'atzar, cadascun dels quals està arrodonit fins al sencer més proper, amb probabilitat igual a 0,99.

Solució:

Sigui la variable X_i = {*error a la i-èsima mesura}*.

X_i es distribueix segons una uniforme dins l'interval (-0,5; 0,5) per i = 1, 2 ...., 108.

Per tant: $E(X_i) = (0,5 - 0,5)/2 = 0$ i $V(X_i) = (0,5 + 0,5)^2/12 = 0,08333$

S = {*error total comès en sumar 108 nombres}*.

$S = X_1 + X_2 + ... + X_{108}$ és la suma de 108 variables independents i idènticament distribuïdes i, pel TCL, $S \sim N(\mu, \sigma)$ amb:

$\mu = 108 \cdot 0 = 0$

$\sigma^2 = 108 \cdot 0,08333 = 8,999 \approx 9$; $\sigma = 3$

Sigui $|S_{màx}|$ l'error màxim que compleix $P(-S_{màx} \leq S \leq S_{màx}) = 0,99$

En tipificar s'obté que:

$P[(-S_{màx} - 0)/3 \leq Z \leq (S_{màx} - 0)/3] = P (-S_{màx}/3 \leq Z \leq S_{màx}/3) = 0,99 \Rightarrow$

$\Rightarrow P (Z \leq S_{màx}/3) = 0,995$

A les taules de la N(0,1) es troba que $P(Z \leq 2,58) = 0,99506$

Per tant, $S_{màx}/3 = 2,58$ i $S_{màx} = 2,58 \cdot 3 = 7,74$ unitats.

2.5 EXERCICIS PROPOSATS

Exercici 1. Indiqueu la distribució de probabilitat i els paràmetres que permeten modelitzar cadascun dels següents experiments aleatoris:

a) Nombre d'ascensors avariats en un edifici d'oficines amb 4 ascensors idèntics i de funcionament independent, amb una probabilitat d'avaria del 3% per a cadascun.
b) Nombre de persones contagiades d'una determinada malaltia entre 100 triades a l'atzar, si la probabilitat de contagi és del 5% per a tota la població.
c) Nombre de persones contagiades d'una determinada malaltia entre els 10 membres d'una família, si la probabilitat de contagi és constant i del 5% per a tota la població.
d) Nombre de peces examinades, entre les fabricades per una màquina amb un 3% de defectuoses, per arribar a trobar-ne 1 de defectuosa.
e) Nombre de peces examinades, entre les fabricades per una màquina amb un 3% de defectuoses, per arribar a trobar-ne 5 de defectuoses.
f) Nombre de peces defectuoses obtingudes en 10 extraccions d'un lot de 25 peces amb un 20% de defectuoses.
g) Nombre de peces examinades amb reposició d'un lot de 25 amb un 20% de defectuoses, per arribar a trobar-ne 3 de defectuoses.
h) Nombre de clients que han demanat cafè entre 30 triats a l'atzar al bar de la facultat si el 65% demana cafè.
i) Nombre d'avaries que presenta una màquina en un determinat mes, si les avaries es produeixen amb un valor esperat constant de 6 per any.
j) Nombre de comandes examinades entre 15 amb un 30% d'urgents per arribar a trobar-ne 1 d'urgent.
k) Nombre de naixements comptabilitzats fins a obtenir 3 nenes acabades de néixer, si la probabilitat que el nadó sigui nen és del 52%.
l) Nombre de trucades que rep una central telefònica per minut si per terme mitjà es reben 20 trucades cada 15'.
m) Nombre de trucades que rep una central telefònica per minut si ha absorbit les trucades rebudes per tres línies telefòniques que rebien per terme mitjà 15, 5 i 25 trucades cada ¼ d'hora, respectivament, de forma constant al llarg de la jornada laboral.

Exercici 2. Calculeu les següents probabilitats corresponents als fenòmens aleatoris anteriors:

a) Probabilitat que un dia no funcioni cap dels 4 ascensors.
b) Probabilitat que com a màxim siguin 3 els contagiats.
c) Probabilitat que tota la família agafi aquesta malaltia.

d) Probabilitat que abans d'examinar la 20ena peça se n'obtingui una de defectuosa.
e) Nombre mitjà de peces correctes examinades fins a arribar a trobar-ne 5 de defectuoses.
f) Probabilitat que s'obtinguin 2 peces defectuoses.
g) Probabilitat que la 3a peça defectuosa s'obtingui entre les 5 primeres extraccions.
h) Variància i nombre esperat de clients de la mostra que ha demanat cafè.
i) Probabilitat que la màquina no s'hagi avariat durant un mes, si com a màxim pot haver tingut 3 avaries.
j) Nombre esperat de comandes examinades fins a trobar-ne una d'urgent i la seva variància.
k) Probabilitat que el nombre total de naixements comptabilitzats sigui superior a 5 per arribar a trobar 3 nenes acabades de néixer.
l) Probabilitat que durant un minut triat a l'atzar s'hagin rebut més de 3 trucades.
m) Probabilitat que durant un minut triat a l'atzar la central rebi menys de 10 trucades si com a mínim n'ha rebut 5.

Exercici 3. Sigui una v.a. X dicotòmica. Quin és el valor de p que maximitza la seva variància?

Exercici 4. Per experiència sabem que el 20% de les ràdios venudes en un determinat establiment són de la marca ADI. Si un determinat dia es venen 8 ràdios,

a) Quina és la probabilitat que 2 o més siguin de la marca ADI?
b) Quin és el nombre esperat i la desviació estàndard de ràdios ADI venudes?
c) Quin és el nombre més probable de ràdios ADI venudes?
d) Si sabem que s'han venut més de dues ràdios ADI, quina és la probabilitat que hagin estat més de 4?

Exercici 5. El diàmetre en centímetres d'una peça fabricada, X, es distribueix segons la funció de densitat:

$$f(x) = \begin{cases} \frac{3}{26}x^2 & 1 < x \leq 3 \\ 0 & en\ altres\ casos \end{cases}$$

Només es pot utilitzar la peça si el seu diàmetre és superior a 2 cm i inferior a 3 cm.

a) Quina és la probabilitat que una peça sigui utilitzable?

b) Si s'empaqueten les peces de 5 en 5 i només s'accepta un paquet si com a màxim té una peça no utilitzable, quina és la probabilitat de rebutjar un paquet?
c) Si es lliura una comanda de 20 paquets, quina és la probabilitat que ens tornin tots els paquets?

Exercici 6. Una empresa dedicada a la venda a domicili ha comprovat que el 60% de les persones visitades accepten el catàleg i només un 15% d'aquestes últimes realitzen alguna compra. Si diàriament un venedor realitza 10 visites a domicili,

a) Quin és el nombre esperat de catàlegs lliurats en un dia per un venedor?
b) Quina és la probabilitat que més de la meitat dels visitats acceptin el catàleg
c) Si l'empresa té 5 venedors que realitzen les visites de forma independent, quina és la probabilitat que es lliurin com a màxim 40 catàlegs en un dia?
d) Si després d'una setmana de visites a domicili un venedor ha lliurat 20 catàlegs, quina és la probabilitat que realitzi més de 4 vendes?

Exercici 7. D'un arxivador amb 40 lletres de pagament 5 vencen aquest mes. Si es treuen 5 lletres a l'atzar, quina és la probabilitat de trobar-ne almenys 3 de les que han de vèncer aquest mes?

Exercici 8. D'un total de 25 exàmens, 5 presenten la qualificació màxima.

a) Si es trien 3 exàmens a l'atzar, quina és la probabilitat de trobar-ne almenys un amb la qualificació màxima?
b) Quin és el nombre esperat d'exàmens amb qualificació màxima que es trobarà si es seleccionen 5 exàmens a l'atzar?
c) Si la selecció de l'apartat a) es fa amb reposició, quina serà la probabilitat demanada?

Exercici 9. Una companyia d'assegurances ofereix una pòlissa per a un sinistre que presenta una probabilitat del 0,005% d'ocurrència i una indemnització de 250 €. Si la companyia té 80.000 pòlisses:

a) Quina és la probabilitat que es presentin menys de 2 sinistres
b) Quina és la probabilitat que hagi de pagar una indemnització superior a 500€?
c) Quin és el valor esperat i la desviació estàndard de la indemnització?

Exercici 10. El nombre de persones que compren tabac en un estanc és una v.a. Poisson amb paràmetre λ=5 persones per cada 5'. Sota el supòsit que la

persona que atén l'estanc només pot despatxar un màxim de 8 persones cada 5', quina és la probabilitat que es formi cua d'espera a l'estanc (arribin més persones de les que pot despatxar)?

Exercici 11. En un supermercat el nombre de clients que utilitzen la caixa ràpida segueix una distribució Poisson amb mitjana 24 clients/hora.

a) Quina és la probabilitat que més de 6 clients utilitzin la caixa ràpida durant 1/4 d'hora?
b) Quina és la probabilitat que menys de 3 clients utilitzin la caixa ràpida durant els propers 5 minuts?
c) Si en l'últim ¼ d'hora han passat per la caixa ràpida més de 5 clients, quina és la probabilitat que hagin estat exactament 10?
d) Si un dia la caixa ràpida comença a funcionar amb 5 minuts de retard, quina és la probabilitat que hi hagi més de 4 clients que s'esperen
e) Quin és el temps d'espera, per terme mitjà, entre dos clients consecutius de la caixa ràpida?
f) Quina és la probabilitat que, després de l'últim pagament, el proper trigui menys de 5 minuts?

Exercici 12. El departament de control de qualitat d'una empresa proposa els següents mètodes per inspeccionar la producció:

Mètode I: inspeccionar les peces una a una fins a trobar-ne una de defectuosa. Si aquesta apareix abans de la vigèsima inspecció s'haurà d'ajustar la màquina.

Mètode II: triar a l'atzar 150 peces i si se'n troben 3 o més de defectuoses s'haurà d'ajustar la màquina.

Si una determinada màquina produeix un 2% d'unitats defectuoses, quin dels dos mètodes presenta major probabilitat d'haver d'ajustar la màquina?

Exercici 13. Els trens de la costa surten cada 20 minuts de l'estació central. Si suposem que el temps d'espera per a qualsevol viatger és una variable aleatòria uniforme,

a) Quina és la probabilitat que s'hagi d'esperar més de 15 minuts?
b) Quina és l'esperança matemàtica i la variància de la variable anterior?
c) Quina és la probabilitat que més de la meitat d'una mostra de 12 viatgers, triats de forma independent, hagin d'esperar menys de 5 minuts?

Exercici 14. Un inversor vol comprar un local a la zona on li garanteixin, amb major probabilitat, una gran afluència de clients. Se sap que el nombre de clients (per dia) a locals de característiques semblants a la zona A es distribueix

segons una U[100,900] i a la zona B segons una Pois(300). Si vol tenir la garantia amb màxima probabilitat que el visitin un mínim de 250 clients/dia, quina zona triarà?

Exercici 15. Les pàgines dels llibres d'una col·lecció d'Espasa i Bruixeria estan dissenyades de forma que tenen 40 línies de 75 espais per línia. El control de qualitat posa de manifest que per terme mitjà es troba una errada cada 6.000 espais. Si el nombre d'errades es modelitza per una distribució de Poisson,

a) Quin és el percentatge de pàgines sense cap errada?
b) Quina és la probabilitat que en un capítol de 15 pàgines hi hagi exactament 3 errades?

Exercici 16. El consum de carburant que presenta un determinat motor de gas-oil és una v.a. Uniforme de mitjana 6 litres per hora i variància 1/3.

a) Determineu la probabilitat que el consum sigui superior a 5,95 litres/hora.
b) Si al llarg d'una hora determinada se sap que el consum ha estat inferior a 6,25 litres, quina és la probabilitat que hagi estat superior a 5,5 litres?
c) Si el dipòsit del motor anterior té una capacitat de 950 litres, quina és la probabilitat que pugui funcionar 160 hores sense tornar-lo a omplir?

Exercici 17. En un concurs de pilota d'una fira d'atraccions se sap que l'alçada que es pot aconseguir en llançar la pilota és una variable uniforme definida a l'interval [2 metres;10 metres]. Si la pilota supera els 8 metres es guanya un premi de 50 € i, a més, si supera els 9 metres es retornen els diners pagats per concursar. Si concursar costa 10 €, quin és el benefici que espera obtenir per cada concursant el propietari de la parada?

Exercici 18. El temps dels estacionaments a la zona blava de Barcelona durant l'horari de pagament és una variable aleatòria exponencial de mitjana 0,5 hores.

a) Si un cotxe acaba d'aparcar, quina és la probabilitat que el seu estacionament superi les 3 hores?
b) Si un cotxe fa 2 hores que està estacionat, quina és la probabilitat que ho estigui com a mínim 2 hores més?
c) Quina és la probabilitat que l'estacionament d'un determinat cotxe sigui inferior a 30 minuts?
d) Si davant del magatzem on volem descarregar una mercaderia només hi ha 2 aparcaments i són de zona blava, quina és la probabilitat d'haver d'esperar

més de ½ hora per poder estacionar el cotxe en qualsevol dels aparcaments anteriors?

e) Quin és el nombre esperat de cotxes que s'estacionaran en una plaça durant les 9 hores de l'horari de pagament diari i quina és la variància?

f) Quina és la probabilitat que en un dia durant l'horari de pagament (9 hores) es produeixin com a màxim 20 estacionaments?

Exercici 19. La fabricació d'unes determinades peces té una duració que depèn de la marca de la màquina utilitzada. Si es realitza amb la marca A la duració és una v.a. exponencial amb valor esperat 10''; amb la marca B i C també és exponencial però amb valor esperat 12'' i 15'', respectivament. El 40% de la producció total s'obté amb màquines marca A, el 25% amb marca B i la resta amb marca C. Sabent que una peça ha presentat un temps de fabricació inferior a 8'', quina és la marca de la màquina de fabricació més probable (més versemblant)? i quina és aquesta probabilitat?

Exercici 20. El temps transcorregut entre dues avaries consecutives, X, és una variable exponencial amb paràmetre λ=5 màquines/hora.

a) Quin és el valor esperat i la desviació estàndard de X?

b) Quina és la probabilitat que passin més de 10 minuts entre dues avaries consecutives?

c) Sabent que durant els darrers 5 minuts hi ha hagut 1 avaria, quina és la probabilitat que la propera es presenti durant els propers 5 minuts?

d) Quina és la probabilitat que s'avariïn exactament 12 màquines durant les properes 2 hores?

e) Quin és el nombre més probable de màquines avariades en 30 minuts?

Exercici 21. El temps de vida útil (en dies) d'uns retoladors, que es venen en paquets de 5 unitats, es pot modelitzar com una v.a. exponencial de paràmetre 1/2000. Si un retolador es considera defectuós quan té una vida útil inferior a 50 dies, quina és la probabilitat que un paquet no tingui cap retolador defectuós?

Exercici 22. Donades les següents distribucions normals:

a) Determineu les següents probabilitats:

a.1) Per a $X \sim N(10; 2)$ $P(X>11)$

a.2) Per a $X \sim N(1; 2)$ $P(0<X<1,5)$

a.3) Per a $X \sim N(10; 5)$ $P(X<8)$

a.4) Per a $X \sim N(20; 5)$ $P(15<X<18)$

b) Determineu el valor de la constant k que presenta les següents probabilitats:
b.1) Per a X~N(10; 2) P(X<k)=0,87076
b.2) Per a X~N(5; 5) P(X>k)=0,22965
b.3) Per a X~N(1; 2) P(0<X<k)=0,40147
b.4) Per a X~N(20; 5) P(X<k)=0,14007

Exercici 23. Per experiència se sap que els resultats dels tests d'estadística són una v.a. normal amb esperança matemàtica 11,5 i desviació estàndard 5.

a) Si la puntuació mínima per aprovar es fixa en 10, quin percentatge d'alumnes aprovaran?
b) Si es vol aprovar a un 69,5% dels alumnes, quina nota mínima s'ha d'exigir per aprovar?
c) Si les puntuacions mínima i màxima per obtenir un notable són 15,5 i 18, respectivament, quants notables s'espera comptabilitzar en una prova amb 2000 presentats?
d) Si en un grup amb 500 presentats un alumne vol aconseguir estar entre els 5 millors, quina puntuació ha d'obtenir?

Exercici 24. Una empresa produeix unes determinades peces de diàmetre aleatori amb distribució normal d'esperança matemàtica 2,98 i desviació estàndard 0,02 cm. El cost de producció de cada peça és de 1,2 € i es pot vendre si el diàmetre esta comprès entre 2,94 i 3,01 cm. Si el diàmetre es inferior a 2,94 cm la peça no és vendible però té un valor residual de 0,5 €. Si el diàmetre es superior a 3,01 cm es pot rebaixar i ser venuda però això suposa un cost addicional de 0,2 €. Si el preu de venda es fixa en 2,5 €, quin cost i quin benefici espera obtenir l'empresa per peça produïda?

Exercici 25. S'ha comprovat que l'ingrés net setmanal d'un cert establiment s'adapta acceptablement a una llei normal. Les liquidacions de caixa revelen que en un terç de les setmanes l'ingrés supera les 4530 € i que només en una de cada mil setmanes l'ingrés no arriba a les 1000 €. Calculeu la probabilitat que triada una setmana a l'atzar l'ingrés sigui superior a 5000 €

Exercici 26. El procés de producció d'unes gerres està format per modelatge, cocció i decoració. El temps (en minuts) de duració de cadascuna de les tasques anteriors és aleatori i amb distribucions Normals independents de paràmetres:

Temps de modelatge: X ~ N(20 mn, 10 mn)

Temps de cocció: Y ~ N(15 mn, 5 mn)

Temps de decoració: Z ~ N(30 mn., 15 mn)

a) Obteniu la probabilitat que la producció d'una determinada gerra sigui superior a una hora i mitja.
b) Calculeu la probabilitat que la decoració duri més del triple que el modelatge.

Exercici 27. El temps de vida útil d'unes determinades bombetes s'ajusta a una llei Normal amb mitjana 6,68 anys i desviació típica 2. Si totes les bombetes que es fonguin en el període de garantia han de ser substituïdes i l'empresa només vol reposar un 1 per mil de la producció, quin ha de ser el període de garantia màxim que haurà de fixar?

Exercici 28. Tres màquines (A, B, C) fabriquen peces idèntiques de pes aleatori i independent amb distribucions A~ N(20 gr, 4 gr) B~ N(20 gr, 6 gr) i C~ N(19 gr, 5 gr). Les màquines A, B i C fabriquen el 20%, 35% i 45% de la producció total, respectivament. Les mateixes màquines empaqueten les peces de 10 en 10. Calculeu la probabilitat que un paquet (10 peces) triat a l'atzar pesi més de 205 gr?

Exercici 29. Uns grans magatzems tenen a la venda només tres marques (A, B i C) d'un article. Un estudi ha determinat que la probabilitat de venda és del 15%, 30% i 55% per les marques A, B, i C, respectivament. Si un dia, triat a l'atzar, aquest magatzem ha venut 200 d'aquests articles, quina és la probabilitat que exactament la meitat siguin de la marca C?

Exercici 30. Una companyia aèria ha observat que per terme mitjà el 12% de les places reservades no es cobreixen i decideix acceptar un nombre de reserves superior en un 10% a les places disponibles en els avions de 450 places. En quina proporció de vols almenys un passatger amb reserva es quedarà sense plaça?

Exercici 31. A la secció d'electrodomèstics d'uns grans magatzems es venen per terme mitjà 30 rentadores diàries, amb desviació estàndard 5,77. El responsable de la secció té un pla de vendes per al proper semestre (180 dies) segons el qual s'han de vendre més de 5500 rentadores.

Si les vendes diàries són independents:

a) Quina és la probabilitat d'assolir aquest objectiu?
b) Si es volgués garantir amb una probabilitat del 90% la venda de més de 6000 rentadores, quants dies de vendes serien necessaris?

Exercici 32. Un mestressa de casa paga sempre la seva compra setmanal al supermercat amb targeta de crèdit amb càrrec a un compte bancari que a principi d'any té un saldo de 7200 €, i en el qual durant l'any no es fa cap ingrés ni cap altre càrrec que no sigui l'esmentat. L'import de la compra setmanal es pot modelitzar amb una variable uniforme definida en l'interval ($130 \leq X \leq 150$). Si suposem que l'import de la compra és independent d'una setmana a l'altra, quina és la probabilitat que en un any (52 setmanes) el saldo del compte hagi estat suficient?

Exercici 33. Una furgoneta amb capacitat màxima de càrrega de 900 kg transporta paquets de pes independent entre si, essent el pes mitjà de 10 kg, amb desviació estàndard 5 kg. En un viatge la furgoneta transporta 100 paquets. En quin percentatge dels viatges la càrrega total serà superior a 850 kg però no arribarà a superar la càrrega màxima?

Exercici 34. En un paquet de 20 melindros hi consta que el pes aproximat per unitat és de 45-60 gr. Si el pes es distribueix com una Normal i se sap que l'esmentat interval correspon a $\mu \pm 2\sigma$, quina és la probabilitat que un paquet de melindros pesi més d'1kg? (Suposeu que el pes de la bossa és menyspreable.)

Exercici 35. La demanda diària (en milers de litres) d'un determinat tipus de benzina en una estació de servei és una variable aleatòria X amb funció de densitat:

$$f(x)=\begin{cases} k(3-x) & 0 < x \leq 1 \\ 0 & \textit{enaltrecasos} \end{cases}$$

Transcorreguts 100 dies de venda, quina és la probabilitat que la demanda total hagi superat els 50.000 litres?

CAPÍTOL III. INTRODUCCIÓ A LA INFERÈNCIA ESTADÍSTICA

3.1 INTRODUCCIÓ

L'objectiu de la ***inferència estadística*** és induir el comportament d'una població a partir de l'extracció i anàlisi d'una mostra. Per tant, és important que la informació continguda a la mostra sigui representativa de la població que l'ha generada. Aquesta representativitat s'aconsegueix treballant amb les anomenades *mostres aleatòries*. Una mostra és aleatòria quan tots els elements de la població tenen la mateixa probabilitat de ser seleccionats i quan aquests elements s'extreuen de manera independent. En aquesta situació veurem que una mostra aleatòria de grandària n (que procedeix de l'observació de n elements) és una variable aleatòria n-dimensional, amb una funció de quantia o de densitat conjunta que depèn de la distribució de probabilitat de la població. Aquesta característica és fonamental per poder establir les distribucions en el mostratge de tot un seguit de funcions d'observacions mostrals, que s'anomenen *estadístics*. Ens interessaran les distribucions d'alguns estadístics concrets, dels quals depenen els diferents mètodes d'inferència. Com veurem, en el marc de la inferència clàssica, és bàsic fer la hipòtesi que les poblacions mostrejades segueixen distribucions normals (o almenys aproximadament normals), si bé, de vegades, les conseqüències de l'incompliment d'aquest supòsit de normalitat es poden resoldre si les mostres disponibles són grans.

Els mètodes clàssics d'inferència es poden dividir, bàsicament, en dos: *mètodes d'estimació de paràmetres i mètodes de contrast d'hipòtesis*. L'estimació de paràmetres consisteix en fixar valors concrets per als paràmetres que caracteritzen la distribució de probabilitat de la població. El contrast d'hipòtesis permet validar hipòtesis estadístiques que fan referència al valor d'un paràmetre poblacional (el valor esperat, la variància, la proporció d'èxits, etc.) o a la relació que existeix entre paràmetres anàlegs de dues poblacions. En termes generals, els contrastos d'hipòtesis permeten decidir si l'evidència empírica (proporcionada pel comportament de la mostra) és o no compatible amb la hipòtesi referida a la població que s'intenta validar.

3.2 MOSTRA ALEATÒRIA

Donada una població estadística caracteritzada per la variable aleatòria X amb funció de densitat f(x) o de quantia P(x), qualsevol conjunt d'observacions aleatòries i independents, que procedeixen de la població, determina una

mostra aleatòria, que es representa per (X_1, X_2, ..., X_n) on n és la grandària de la mostra o nombre d'observacions que formen la mostra.

Una mostra serà aleatòria si cada element de la població té igual probabilitat (equiprobabilitat) de ser escollit cada cop que s'extregui una observació. El procediment més simple per assegurar l'equiprobabilitat és el mostratge aleatori simple, concretament, el mostratge amb reemplaçament de l'element extret abans de cada nova extracció. En poblacions grans respecte a la grandària de la mostra, pràcticament no hi haurà diferència si es reemplaça o no l'element. Al contrari, en poblacions petites el reemplaçament adquireix una gran importància ja que permet que la població torni al seu estat original abans de cada extracció i, en conseqüència, es garanteix que les extraccions siguin independents (és a dir, la probabilitat d'extreure un element qualsevol no queda condicionada pels resultats de les extraccions anteriors).

En aquesta situació, en una mostra de grandària n cadascuna de les observacions, X_1, X_2,...,X_n, és una *variable aleatòria*, amb distribució de probabilitat idèntica a la de la població, ja que les extraccions són independents (en cada extracció la població manté la composició original).

Així, per exemple, donada una urna que conté 5 boles, dues amb el número 1 i tres amb el número 2, si definim la variable X com el '*número observat en extreure una bola*', la distribució poblacional d'aquesta variable aleatòria és:

X	P(x)
1	2/5
2	3/5

Si s'extreu una mostra amb reposició, el primer element (X_1) pot prendre qualsevol valor de la població (1 o 2), i les probabilitats d'obtenir un 1 o un 2 són 2/5 i 3/5, respectivament, per tant, X_1 és una variable aleatòria que reprodueix la distribució poblacional. Si reemplacem la bola extreta abans d'observar el segon element de la mostra, aquest pot prendre també els valors 1 o 2 i les seves respectives probabilitats són novament 2/5 i 3/5, per tant, X_2 torna a ser una variable aleatòria amb la mateixa distribució de probabilitat que la població. Si reemplacem la bola, de nou, el tercer element de la mostra X_3 serà una variable aleatòria amb la mateixa distribució de probabilitat, i així successivament. (Noteu que els resultats de cadascuna de les extraccions no condicionen el resultat de les altres, per tant, X_1, X_2,.., X_n són variables aleatòries independents.)

***Definició*:**

Una ***mostra aleatòria*** de grandària n és una variable n-dimensional formada per un conjunt de n variables aleatòries (X_1, X_2,..., X_n) independents, amb la mateixa distribució de probabilitat que la població (idènticament distribuïdes) i amb funció de densitat o quantia conjunta igual al producte de les marginals.

Per a la mostra aleatòria (X_1, X_2,..., X_n) es té que:

$f(X_1) = f(X_2) = ...= f(X_n) = f(X)$ (si X és una v.a. contínua) o

$P(X_1)=P(X_2)= .....= P(X_n) = P(X)$ (si X és una v.a. discreta)

La distribució de probabilitat conjunta de la mostra (ja que X_1, X_2,..., X_n són independents) és:

$f(X_1\, X_2 \,...X_n) = f(X_1)\, f(X_2) \,...\, f(X_n)$ (si X és una v.a. contínua) o

$P(X_1\, X_2 \,...X_n) = P(X_1)\, P(X_2) \,...\, P(X_n)$ (si X és una v.a. discreta)

És important diferenciar clarament entre una mostra aleatòria, que és una variable aleatòria n-dimensional, i una realització mostral o mostra concreta. Abans d'observar el valor (valor numèric) del 1r, 2n, ..., n-èsim element de la mostra, X_1, X_2, ...,X_n són variables aleatòries; després de l'extracció mostral, els valors observats a cadascuna de les extraccions constitueixen la realització mostral.

S'ha de puntualitzar que totes les mostres de grandària n tenen la mateixa probabilitat de ser extretes, no obstant això, no totes les realitzacions mostrals (valors numèrics dels elements observats) tenen la mateixa probabilitat.

Exemple 3.1

Un estudi està interessat en les famílies que posseeixen dos cotxes com a màxim. El conjunt d'aquestes famílies és la població estadística i la variable aleatòria que la caracteritza, X, definida com el 'nombre de cotxes per família' té la distribució de probabilitat següent:

X	0	1	2
P(X)	0,2	0,6	0,2

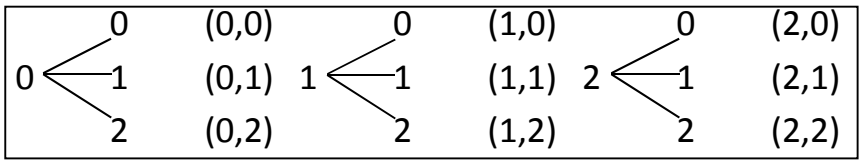
d'on

$$E(X) = \sum_{\forall x_i} x_i P(x_i) = 0\cdot 0{,}2 + 1\cdot 0{,}6 + 2\cdot 0{,}2 = 1$$

$$V(X) = \sum_{\forall x_i} x_i^2 P(x_i) - (E(X))^2 = 0^2 0{,}2 + 1^2 0{,}6 + 2^2 0{,}2 - 1^2 = 0{,}4$$

Prendre una mostra de grandària n=2 vol dir seleccionar a l'atzar 2 famílies del conjunt poblacional i observar el nombre de cotxes.

Així doncs, la mostra de grandària n=2 és la variable aleatòria bidimensional (X_1, X_2) que pot prendre els valors següents:



El conjunt de totes les realitzacions mostrals possibles és:

(0, 0) (0, 1) (0, 2) (1, 0) (1, 1) (1, 2) (2, 0) (2, 1) (2, 2)

amb la distribució de probabilitat següent:

(X_1, X_2)	$P(X_1, X_2)= P(X_1)P(X_2)$
(0,0)	0,2·0,2=0,04
(0,1)	0,2·0,6=0,12
(0,2)	0,2·0,2=0,04
(1,0)	0,12
(1,1)	0,36
(1,2)	0,12
(2,0)	0,04
(2,1)	0,12
(2,2)	0,04
	1

Ja que cada observació mostral, X_1 o X_2, pot prendre qualsevol valor poblacional amb idèntiques distribucions de probabilitat ($P(X_1) = P(X_2) = P(X)$) i la distribució de probabilitat conjunta, com que les variables són independents, és el producte de les probabilitats marginals:

$$P(X_1, X_1) = P(X_1)\, P(X_1)$$
$$P(X_1, X_2) = P(X_1)\, P(X_2)$$

A priori es pot extreure qualsevol de les mostres assenyalades, no obstant això, s'observa que cada realització mostral té diferent probabilitat de ser escollida. Per exemple, la realització (1, 1) és més probable (36%) que la (0, 0), que té una probabilitat del 4%.

3.3 FUNCIÓ DE VERSEMBLANÇA D'UNA MOSTRA

El concepte de versemblança d'una mostra es basa en la idea que poblacions diferents generen mostres diferents. Per tant, una mostra concreta presentarà una determinada probabilitat (versemblança) segons la població d'origen per la qual hagi estat generada. Així, si d'una població Normal amb variància 4 i valor esperat μ desconegut s'extreu una mostra (x_1, x_2,..., x_n) és possible que aquesta hagi estat generada per una N(μ_1,2) concreta, o per una altra N(μ_2,2), o per una N(μ_k,2), per tant, hi ha infinitat de possibilitats.

La funció que dóna la probabilitat que presenta la població d'haver generat la mostra obtinguda s'anomena *funció de versemblança de la mostra*. Aquesta funció queda determinada per la funció de probabilitat conjunta de la mostra particular, on x_1, x_2,..., x_n són valors fixos (observats) i la probabilitat fa referència als diferents valors del paràmetre poblacional.

Definició:

Donada una variable aleatòria X (contínua o discreta) amb distribució de probabilitat f(x) (o P(x)) determinada pel paràmetre ϑ desconegut, la funció (de densitat o de quantia) de probabilitat conjunta de la mostra aleatòria (X_1, X_2,..., X_n) s'anomena ***funció de versemblança de la mostra*** si depèn del paràmetre de la població i es representa:

$$\ell(x_1, x_2,..., x_n; \vartheta) = P(x_1; \vartheta)\, P(x_2; \vartheta) \dots P(x_n; \vartheta) = \prod_{i=1}^{n} P(x_i; \vartheta) \text{ si X és discreta}$$

$$\ell(x_1, x_2,..., x_n; \vartheta) = f(x_1; \vartheta)\, f(x_2; \vartheta) \dots f(x_n; \vartheta) = \prod_{i=1}^{n} f(x_i; \vartheta) \text{ si X és contínua}$$

Aquesta funció està definida per al recorregut del paràmetre poblacional que la determina. Així, per exemple, el domini de $\ell(x_1, x_2,..., x_n; \pi)$, on X~B(1,$\pi$), és [0,1]; el de $\ell(x_1, x_2,..., x_n; \lambda)$, on X~Pois($\lambda$), és [0,+$\infty$]; etc.

(En l'apartat d'estimació es veurà com a partir d'aquesta funció de versemblança s'obté el valor del paràmetre que amb major probabilitat caracteritza la població que ha generat la mostra.)

Exemple 3.2

Es vol determinar la funció de versemblança d'una mostra de grandària n obtinguda d'una població amb distribució Poisson de paràmetre λ desconegut i, en concret, la funció que s'obté de la mostra (5, 4, 3, 6, 4).

Solució:

X~Pois(λ) $P(x) = e^{-\lambda}\frac{\lambda^x}{x!}$

Donada una mostra $(x_1, x_2,..., x_n)$

$P(X=x_i) = e^{-\lambda}\frac{\lambda^{xi}}{x_i!}$ $\forall i=1,...,n$

Per tant, la funció de versemblança és:

$$\ell(x_1, x_2,..., x_n; \lambda) = \prod_{i=1}^{n} P(x_i; \lambda) = e^{-\lambda}\frac{\lambda^{x1}}{x_1!}\, e^{-\lambda}\frac{\lambda^{x2}}{x_2!} \ldots e^{-\lambda}\frac{\lambda^{xn}}{x_n!} = e^{-n\lambda}\frac{\lambda^{\sum xi}}{\prod_{i=1}^{n}(x_i!)}$$

La funció de versemblança que s'obté de la mostra (5, 4, 3, 6, 4) és:

$$\ell(5,4,3,6,4; \lambda) = e^{-n\lambda}\frac{\lambda^{\sum xi}}{\prod_{i=1}^{n}(x_i!)} = e^{-5\lambda}\frac{\lambda^{22}}{5!4!3!6!4!}$$

Exemple 3.3

D'una població amb llei exponencial s'ha obtingut la mostra (2, 4) de grandària 2. El paràmetre λ de la població anterior és desconegut, però se sap que només pot prendre els valors λ=0,2, λ=0,3, λ=0,4 i λ=0,5. D'aquestes quatre possibles poblacions, quina presenta major probabilitat d'haver generat la mostra?

Solució:

X~Exp(λ) $f(x) = \lambda e^{-\lambda x}$ $x \geq 0$

La funció de versemblança és:

$$\ell(x_1, x_2,..., x_n; \lambda) = \prod_{i=1}^{n} f(x_i; \lambda) = \lambda e^{-\lambda x1}\, \lambda e^{-\lambda x2} \ldots \lambda e^{-\lambda xn} = \lambda^n e^{-\lambda \Sigma x}$$

Per a la mostra (2,4) es té:

$\ell(2,4; \lambda) = \lambda^2 e^{-6\lambda}$

I per als valors possibles de λ queda:

λ=0,2 $\ell(2,4; \lambda) = 0,2^2 e^{-1,2} = 0,0120$

λ=0,3 $\ell(2,4; \lambda) = 0,3^2 e^{-1,8} = 0,0149$

λ=0,4 $\ell(2,4; \lambda) = 0,4^2 e^{-2,4} = 0,0145$

λ=0,5 $\ell(2,4; \lambda) = 0,5^2 e^{-3} = 0,0124$

De les quatre poblacions considerades la que presenta major probabilitat d'haver generat la mostra és Exp(0,3) amb una versemblança igual a 0,0149.

Exemple 3.4

Les bosses de 5 caramels variats d'una determinada marca contenen una proporció π de caramels de llimona. A partir de la mostra (2, 0, 1) de grandària 3 (és a dir, 2 caramels de llimona a la primera bossa observada, 0 a la segona i 1 a la tercera) es vol determinar quin dels valors del paràmetre π és el que presenta major probabilitat de ser el poblacional sota el supòsit que π només pot prendre els valors 1/5, 2/5, 3/5 i 4/5.

Solució:

X~B(5,π) i $P(x) = \binom{5}{x}\pi^x(1-\pi)^{5-x} \quad \forall x$

La funció de versemblança d'una mostra de grandària 3 és:

$\ell(x_1, x_2, x_3; \pi) = P(x_1; \pi)\, P(x_2; \pi)\, P(x_3; \pi) =$

$$= \binom{5}{x1}\pi^{x1}(1-\pi)^{5-x1}\binom{5}{x2}\pi^{x2}(1-\pi)^{5-x2}\binom{5}{x3}\pi^{x3}(1-\pi)^{5-x3} =$$

$$= \binom{5}{x1}\binom{5}{x2}\binom{5}{x3}\pi^{x1+x2+x3}(1-\pi)^{15-(x1+x2+x3)}$$

Per a la mostra (2,0,1) queda:

$$\ell(2,0,1;\pi) = \binom{5}{2}\binom{5}{0}\binom{5}{1}\pi^3(1-\pi)^{12}$$

La versemblança de π és:

$$\ell(2,0,1;\pi=1/5) = \binom{5}{2}\binom{5}{0}\binom{5}{1}(1/5)^3(4/5)^{12} = 0{,}02749$$

$$\ell(2,0,1;\pi=2/5) = \binom{5}{2}\binom{5}{0}\binom{5}{1}(2/5)^3(3/5)^{12} = 0{,}00697$$

$$\ell(2,0,1;\pi=3/5) = \binom{5}{2}\binom{5}{0}\binom{5}{1}(3/5)^3(2/5)^{12} = 0{,}00018$$

$$\ell(2,0,1;\pi=4/5) = \binom{5}{2}\binom{5}{0}\binom{5}{1}(4/5)^3(1/5)^{12} \cong 0$$

Es pot concloure que el valor de π més versemblant és 1/5, o, dit d'una altra manera, 1/5 és el valor màxim versemblant de π.

3.4 DISTRIBUCIONS D'ALGUNS ESTADÍSTICS

3.4.1 ESTADÍSTIC MOSTRAL

Com ja s'ha comentat a la introducció la informació mostral té com a objectiu realitzar inferències sobre els paràmetres de la població que es fonamenten en les distribucions de probabilitat dels anomenats *estadístics mostrals*.

***Definició*:**

Un ***estadístic*** és qualsevol funció de les variables aleatòries observades a la mostra sempre i quan no contingui paràmetres desconeguts.

Així, per exemple, d'una població caracteritzada per $f(x;\vartheta)$ on ϑ és un paràmetre desconegut a partir d'una mostra aleatòria $(X_1, X_2,..., X_n)$ es poden definir diferents estadístics. Entre d'altres els següents:

$$\hat{\vartheta} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \hat{\vartheta} = \frac{\max\{X_i\} + \min\{X_i\}}{2}, \quad \hat{\vartheta} = Me, \quad \hat{\vartheta} = \frac{\sum_{i=1}^{n} X_i^2}{n} \ldots$$

En general es representa l'estadístic com: $\hat{\vartheta} = g(X_1, X_2,..., X_n)$, on g és una funció concreta de la mostra aleatòria.

Característiques:

1. Atès que és una funció de variables aleatòries, l'estadístic també és una variable aleatòria i, per tant, queda definit per la seva distribució de probabilitat.
2. La distribució de probabilitat en el mostratge d'un estadístic depèn de la distribució de probabilitat poblacional i de la grandària de la mostra.
3. Per a cada mostra particular s'obté un valor específic de l'estadístic.
4. Si s'empra l'estadístic per estimar un paràmetre poblacional aleshores rep el nom ***d'estimador***, i a cada valor específic que s'obté se l'anomena ***estimació.***

Exemples d'estadístics mostrals són: la mitjana mostral, la proporció mostral, la variància mostral, etc.

3.4.2 DISTRIBUCIÓ DE LA MITJANA MOSTRAL

Sigui una població caracteritzada per la variable aleatòria X amb $E(X)=\mu$ i $V(X)=\sigma^2$. S'extreu una mostra aleatòria de grandària n, $(X_1, X_2,..., X_n)$, i es defineix l'estadístic ***mitjana mostral*** com a:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Aquest estadístic presenta una distribució que queda caracteritzada per:

1. L'esperança matemàtica de $\overline{X}$ és μ

$E(\overline{X})=E(\frac{X_1 + X_2 + ... + X_n}{n})= 1/n\ [E(X_1)+E(X_2)+...+E(X_n)]= 1/n\ [n\mu] = \mu$

Aquest resultat indica que en extreure un nombre elevat de mostres aleatòries de grandària n, l'esperança de les mitjanes mostrals obtingudes tendeix al veritable valor de la mitjana poblacional; malgrat això, el valor concret de la mitjana d'una mostra particular pot ser, i és generalment, diferent a la mitjana poblacional.

2. La variància de $\overline{X}$ és $\frac{\sigma^2}{n}$

$V(\overline{X})=V(\frac{X_1 + X_2 + ... + X_n}{n})= 1/n^2\ [V(X_1)+V(X_2)+...+V(X_n)]= 1/n^2\ [n\sigma^2] = \frac{\sigma^2}{n}$

Aquest resultat implica que la variància de la mitjana mostral és inversament proporcional a la grandària n. Com més observacions tingui la mostra més concentrada estarà la distribució de $\overline{X}$ al voltant de μ.

La desviació estàndard de $\overline{X}$, que rep el nom d'error estàndard, és $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

La distribució de probabilitat de la variable $\overline{X}$ és desconeguda excepte en els casos següents:

- Si la població d'origen és normal, la mitjana mostral presenta també distribució normal donat que és combinació lineal de variables normals.
- Si la població d'origen no és normal però n és suficientment gran (n>30), la mitjana mostral convergeix a la distribució normal pel Teorema Central del Límit.

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Exemple 3.5

Determineu la distribució de probabilitat, l'esperança matemàtica i la variància de la mitjana mostral d'una mostra aleatòria de grandària 2 extreta d'una població caracteritzada per la funció de quantia següent:

X	2	4	6
P(X)	0,2	0,5	0,3

Solució:

Distribució de $\bar{X}$:

Totes les possibles mostres de grandària 2 són:

2	2	(2,2)	4	2	(4,2)	6	2	(6,2)
	4	(2,4)		4	(4,4)		4	(6,4)
	6	(2,6)		6	(4,6)		6	(6,6)

Amb probabilitats:

(X_1, X_2)	$P(X_1, X_2)= P(X_1)P(X_2)$	$\bar{X}$
(2,2)	0,2·0,2=0,04	2
(2,4)	0,2·0,5=0,10	3
(2,6)	0,2·0,6=0,06	4
(4,2)	0,10	3
(4,4)	0,25	4
(4,6)	0,15	5
(6,2)	0,06	4
(6,4)	0,15	5
(6,6)	0,09	6
	1	

Per tant, la distribució de $\bar{X}$ és:

$\bar{X}$	$P(\bar{X})$
2	0,04
3	0,20
4	0,37
5	0,30
6	0,09
	1

Aleshores,

E($\bar{X}$) = 2·0,04+ 3·0,2 + 4·0,37 + 5·0,3 + 6·0,09 = 4,2

V($\bar{X}$) = 2^2 0,04+ $3^3$0,2 + 4^2 0,37 + 5^2 0,3 + 6^2 0,09 - $4,2^2$ = 0,98

$\sigma_{\bar{X}}=\sqrt{0,98}$ =0,9899

També es pot obtenir l'esperança i la variància de $\overline{X}$ a partir dels moments de la població d'origen:

E(X) = 2·0,2+ 4·0,5 + 6·0,3 = 4,2

V(X) = 2^2 0,2+ 4^2 0,5 + 6^2 0,3 - $4,2^2$ = 1,96

σ= 1,4

Com ja s'ha demostrat per a una mostra aleatòria de grandària n = 2:

E($\bar{X}$)= μ = 4,2

V($\bar{X}$)= σ^2/n = 1,96/2 = 0,98

Resultats que coincideixen amb els obtinguts a partir de la distribució de probabilitat de $\bar{X}$.

Exemple 3.6

Als magatzems OSQUI, el nombre de televisors venuts diàriament és una variable aleatòria amb distribució no especificada però amb desviació estàndard coneguda i igual a 30 televisors. Si s'observen les vendes de televisors durant 81 dies triats a l'atzar:

a) Quina és la distribució de probabilitat de la mitjana mostral?

b) Quina és la probabilitat que la mitjana mostral difereixi com a màxim en 10 unitats del veritable valor esperat de les vendes diàries?

<u>Solució</u>:

a) n=81>30 $\Rightarrow$ $\bar{X}$ ~N(μ, $\frac{\sigma}{\sqrt{n}}$) en aquest cas $\bar{X}$ ~N(μ, $\frac{30}{\sqrt{81}}$)

b) P(| $\bar{X}$ -μ |<10)=P(μ-10< $\bar{X}$ <μ+10)=P($\frac{\mu-10-\mu}{\frac{30}{\sqrt{81}}}$ <z< $\frac{\mu+10-\mu}{\frac{30}{\sqrt{81}}}$)=

=P(-3<z<3)=0,9973

3.4.3 DISTRIBUCIÓ DE LA VARIÀNCIA MOSTRAL

Sigui una població caracteritzada per la variable aleatòria X amb $E(X)=\mu$ i $V(X)=\sigma^2$ desconeguts de la qual s'extreu una mostra aleatòria de grandària n, $(X_1, X_2,..., X_n)$. L'estadístic ***variància mostral*** *e*s defineix com a:

$$S^2=\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{n-1}$$

Les característiques de la seva distribució són:

1. L'estadístic S^2 només pot prendre valors positius.

2. L'esperança matemàtica de S^2 és la variància poblacional σ^2:

$$E(S^2)=E(\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{n-1})= 1/(n-1)\ E(\sum_{i=1}^{n}(X_i-\overline{X})^2)=$$

$$=1/(n-1)\ E(\sum_{i=1}^{n}\left[(X_i-\mu)-(\overline{X}-\mu)\right]^2)$$

Com que:

$$\sum_{i=1}^{n}\left[(X_i-\mu)-(\overline{X}-\mu)\right]^2=\sum_{i=1}^{n}\left[(X_i-\mu)^2+(\overline{X}-\mu)^2-2(X_i-\mu)(\overline{X}-\mu)\right]=$$

$$=\sum_{i=1}^{n}(X_i-\mu)^2+n(\overline{X}-\mu)^2-2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)$$

$$\{-2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)=-2(\overline{X}-\mu)(n\overline{X}-n\mu)=-2n(\overline{X}-\mu)^2\}$$

$$=\sum_{i=1}^{n}(X_i-\mu)^2+n(\overline{X}-\mu)^2-2n(\overline{X}-\mu)^2=\sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}-\mu)^2$$

Per tant:

$$E(S^2)= 1/(n-1)\ E[\sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}-\mu)^2]=$$

$$=1/(n-1)\ [\sum_{i=1}^{n}E(X_i-\mu)^2-nE(\overline{X}-\mu)^2]=$$

$$= 1/(n-1)\ [n\sigma^2-n\frac{\sigma^2}{n}]=1/(n-1)\ [(n-1)\ \sigma^2]=\sigma^2$$

Aquest resultat implica que, per a un nombre elevat de mostres aleatòries de grandària n, la mitjana de les variàncies mostrals tendeix a la variància poblacional.

3. La variància de S^2, si la població d'origen és normal, és:

$$V(S^2)=\frac{2\sigma^4}{n-1}$$

La dispersió de l'estadístic disminueix a mesura que s'incrementa la grandària de la mostra.

4. Si la població d'origen és normal la transformació de l'estadístic $\frac{(n-1)S^2}{\sigma^2}$ presenta una distribució de probabilitat anomenada Khi al quadrat amb n-1 graus de llibertat, distribució que s'estudiarà més endavant.

Exemple 3.7

Amb les dades de l'exemple 3.5, determineu la distribució de probabilitat i l'esperança matemàtica de la variància d'una mostra de grandària 2.

Solució:

X	2	4	6
P(X)	0,2	0,5	0,3

El valor esperat i la variància de la població són:

$E(X) = 2\cdot0,2+ 4\cdot0,5 + 6\cdot0,3 = 4,2$

$V(X)=\sigma^2= 2^2\ 0,2+ 4^2\ 0,5 + 6^2\ 0,3 - 4,2^2 = 1,96$

La distribució de S^2 per totes les mostres de grandària 2 és:

(X_1, X_2)	$P(X_1, X_2)= P(X_1)P(X_2)$	S^2
(2,2)	0,2·0,2=0,04	0
(2,4)	0,2·0,5=0,10	2
(2,6)	0,2·0,6=0,06	8
(4,2)	0,10	2
(4,4)	0,25	0
(4,6)	0,15	2
(6,2)	0,06	8
(6,4)	0,15	2
(6,6)	0,09	0
	1	

Per tant, la distribució de S^2 queda:

S^2	$P(S^2)$
0	0,38
2	0,50
8	0,12
	1

Aleshores,

$E(S^2) = 0 \cdot 0{,}38 + 2 \cdot 0{,}5 + 8 \cdot 0{,}12 = 1{,}96$

Resultat que coincideix amb la variància de la població.

3.4.4 DISTRIBUCIÓ DE LA PROPORCIÓ MOSTRAL

Si es realitzen n observacions independents d'un fenomen aleatori dicotòmic on la probabilitat d'obtenir èxit dins de la població és igual a π, llavors la variable aleatòria X definida com '*nombre d'èxits obtinguts en les n proves realitzades*' segueix una llei binomial de paràmetres n i π amb $E(X)=n\pi$ i $V(X)=n\pi(1-\pi)$.

Aquest paràmetre π normalment serà desconegut i per fer inferència sobre la població utilitzarem l'estadístic ***proporció mostral***, és a dir, la proporció d'èxits obtinguts dins de la mostra observada.

Definició:

Si s'extreu una mostra aleatòria de grandària n, (X_1, X_2,..., X_n), d'una població dicotòmica, l'estadístic ***proporció mostral*** es defineix com la proporció d'èxits que presenta la mostra:

$$p=\frac{X}{n}$$

on X és la suma dels èxits obtinguts en n proves independents i, per tant, es distribueix com a $X \sim B(n,\pi)$.

Característiques:

1. La proporció mostral p és una variable aleatòria de paràmetres:

$$E(p)=E\left(\frac{X}{n}\right)=1/n\ E(X)=1/n\ [\ n\ \pi]=\pi$$

$$V(p)=V\left(\frac{X}{n}\right)=1/n^2\ V(X)=1/n^2\ [n\ \pi\ (1-\pi)]=\frac{\pi(1-\pi)}{n}$$

Aquests resultats impliquen que per a un conjunt elevat de mostres la mitjana de la proporció mostral tendeix a la proporció poblacional π i, a més, com més observacions tingui la mostra, més concentrada estarà la distribució de probabilitat de p al voltant del valor π, ja que la dispersió disminueix amb n.

2. La desviació estàndard de la proporció mostral rep el nom d'error estàndard i és igual a:

$\sigma_p=\sqrt{\frac{\pi(1-\pi)}{n}}$

3. Pel Teorema Central del Límit sabem que, si la grandària de la mostra és suficientment elevada, la distribució binomial convergeix a una distribució normal. Això també és cert per a la distribució de probabilitat de p:

X~B(n,π), si $n\,\pi\,(1-\pi)>5 \Rightarrow X\sim N(n\pi,\sqrt{n\pi(1-\pi)}) \Rightarrow p\sim N(\pi,\sqrt{\frac{\pi(1-\pi)}{n}})$

Per tant: $\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$ ~N(0,1) si n és suficientment gran.

Exemple 3.8

Una associació benèfica ha comprovat que el 5% de la població fa anualment alguna donació. Si s'extreu una mostra aleatòria de 300 individus:

a) Quina és la distribució de probabilitat de la proporció mostral d'individus que fan alguna donació anual?

b) Quina és la probabilitat que més del 7,5% de les persones entrevistades facin alguna donació aquest any?

Solució:

a) $\pi=0,05$ i $n=300$.

Comprovem que la mostra és suficientment gran per aproximar a la Normal:

$n\,\pi\,(1-\pi)>5 \Rightarrow 300\cdot0,05\cdot0,95=14,25$

$p\sim N(\pi,\sqrt{\frac{\pi(1-\pi)}{n}}) \Rightarrow p\sim N(0,05;\sqrt{\frac{0,05\cdot0,95}{300}})$

b) $P(p>0,075)=P(z>\frac{0,075-0,05}{\sqrt{\frac{0,05\cdot0,95}{300}}})=P(z>1,987)=1-0,9767=0,0233$

3.5 DISTRIBUCIONS DEDUÏDES DE LA NORMAL

A continuació es presenten les distribucions de probabilitat (característiques, moments i propietats) d'un seguit de variables deduïdes de la normal que tenen gran importància dins del camp de la inferència estadística, ja que modelitzen el comportament probabilístic en el mostratge dels estadístics més utilitzats.

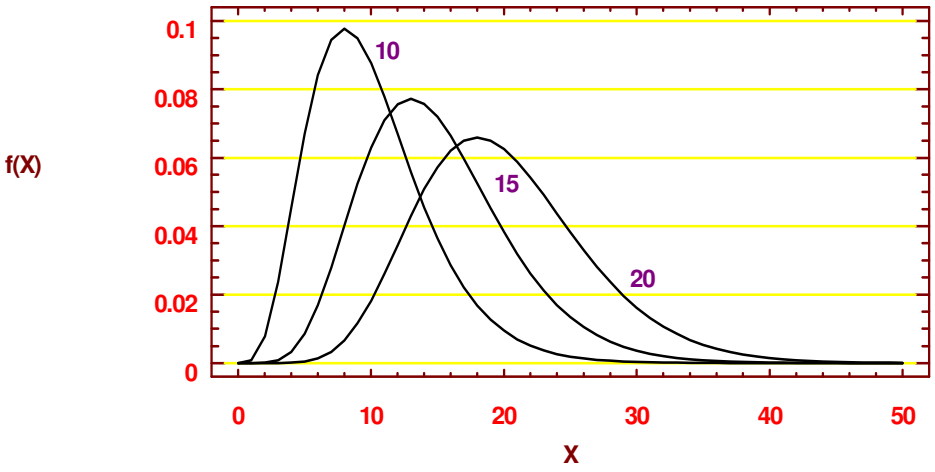
3.5.1 DISTRIBUCIÓ KHI AL QUADRAT

Definició:

Si Z_1, Z_2, ..., Z_n són variables aleatòries independents totes elles distribuïdes segons una normal estàndard, aleshores la variable aleatòria $\sum_{i=1}^{n} Z_i^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ es distribueix segons una **khi al quadrat** amb n graus de llibertat i es simbolitza com χ_n^2:

$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2$$

Gràfic 3.1 Distribucions χ^2 amb 10, 15 i 20 g.l.



Característiques:

1. La variable Khi al quadrat és contínua i el seu recorregut es troba entre 0 i +∞.
2. Existeix un nombre infinit de distribucions Khi al quadrat, una per a cada valor enter positiu de n. (Vegeu gràfic 3.1)

3. La variable Khi al quadrat depèn únicament del paràmetre n o graus de llibertat de la distribució. Aquest paràmetre recull el nombre de variables aleatòries independents que formen la χ^2.
4. L'esperança matemàtica és

   $E(\chi_n^2) = \mu = n$ i la variància $V(\chi_n^2) = \sigma^2 = 2n$.

   Per tant, aquesta distribució queda completament caracteritzada pels seus graus de llibertat.
5. Les distribucions Khi al quadrat tenen asimetria positiva. No obstant això, en incrementar n, la asimetria disminueix.
6. Si Z~N(0,1) aleshores $Z^2 \sim \chi_1^2$.
7. La distribució Khi al quadrat és reproductiva. Si X_1 i X_2 són dues variables Khi al quadrat independents amb m i n graus de llibertat respectivament,

   $X_1 \sim \chi_m^2$, $X_2 \sim \chi_n^2$, aleshores, la suma X_1+X_2 es distribueix també com una Khi al quadrat amb m+n graus de llibertat. Aquesta propietat es verifica per a qualsevol nombre de variables Khi al quadrat independents; és a dir:

   $\sum X_n \sim \chi_{\sum n}^2$.
8. La funció

   $$\frac{(n-1)S^2}{\sigma^2}$$

   de l'estadístic mostral S^2, obtingut a partir d'una mostra de grandària n d'una població N(μ;σ), segueix una distribució Khi al quadrat amb n-1 graus de llibertat.

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left[(X_i - \mu) - (\overline{X} - \mu)\right]^2}{\sigma^2}$$

Com que:

$$\sum_{i=1}^{n}\left[(X_i - \mu) - (\overline{X} - \mu)\right]^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2$$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} - \frac{(\overline{X} - \mu)^2}{\sigma^2/n} =$$

$$= \sum_{i=1}^{n} Z_i^2 - Z^2 \sim \chi_{n-1}^2$$

3.5.2 DISTRIBUCIÓ T DE STUDENT

Definició:

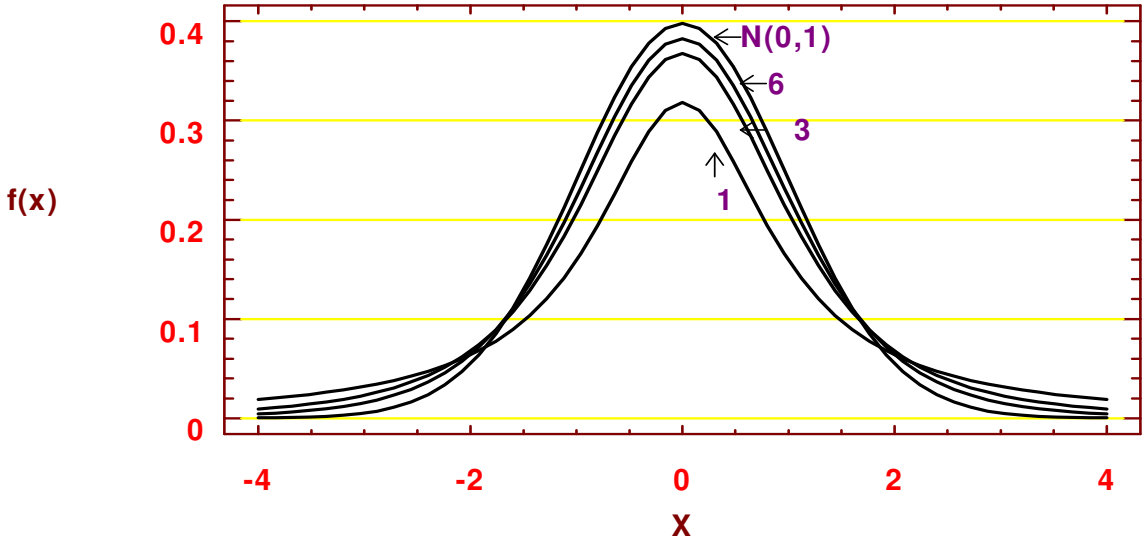
Sigui $\chi^2 \sim \chi^2_n$ i Z~N(0,1), si χ^2 i Z són independents, aleshores la variable aleatòria t igual a:

$$t = \frac{Z}{\sqrt{\chi^2 / n}} \sim t_n$$

segueix una distribució ***t de Student*** amb n graus de llibertat i es simbolitza com t_n.

De la definició anterior es desprèn que la distribució t és el quocient entre una variable normal estàndard i l'arrel quadrada d'una Khi al quadrat dividida pels seus graus de llibertat.

Gràfic 3.2 Distribucions t de Student amb 1, 3 i 6 g.ll.



Característiques.

1. És contínua i el seu recorregut es troba entre $-\infty$ i $+\infty$.
2. Existeix un nombre infinit de distribucions t, una per a cada valor enter positiu de n.
3. Aquesta distribució està completament caracteritzada pels seus graus de llibertat **n**. Així, l'esperança matemàtica i la variància només depenen d'aquest paràmetre i són:

$$E(t) = \mu = 0 \quad n > 1 \qquad V(t) = \sigma^2 = \frac{n}{n-2} \quad n > 2$$

4. Es tracta de distribucions simètriques respecte a l'eix x=0.
5. La distribució t presenta característiques molt semblants a la distribució normal estàndard, si bé, la distribució t presenta sempre major dispersió que la N(0,1), ja que $\sigma = \sqrt{\frac{n}{n-2}} > 1$ per a n>2. A mesura que s'incrementa n la desviació estàndard tendeix a 1.
6. La distribució t convergeix a la distribució N(0,1) quan n tendeix a infinit. A la pràctica considerarem que per a n>30 la N(0,1) dóna una bona aproximació de la t de Student.
7. La funció

   $$\frac{\overline{X} - \mu}{S / \sqrt{n}}$$

   dels estadístics mostrals $\overline{X}$ i S obtinguts a partir d'una mostra de dimensió n d'una població N(μ;σ) presenta una distribució t de Student amb n-1 graus de llibertat.

   Volem demostrar que

   $$\frac{\overline{X} - \mu}{S / \sqrt{n}} = t_{n-1} = \frac{Z}{\sqrt{\chi^2_{n-1} / n-1}}$$

   i sabem que

   $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ i que $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = Z \sim N(0,1)$.

   Per tant,

   $$\frac{\overline{X} - \mu}{S / \sqrt{n}} = \frac{\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}}{\frac{S / \sqrt{n}}{\sigma / \sqrt{n}}} = \frac{Z}{\sqrt{S^2 / \sigma^2}} = \frac{Z}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\frac{1}{(n-1)}}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{(n-1)}}} \sim t_{n-1}$$

3.5.3 DISTRIBUCIÓ F DE SNEDECOR

Definició:

Donades les variables aleatòries independents X_1, X_2, ..., X_m i Y_1, Y_2, ...,Y_n distribuïdes segons una normal estàndard, si definim la variable aleatòria

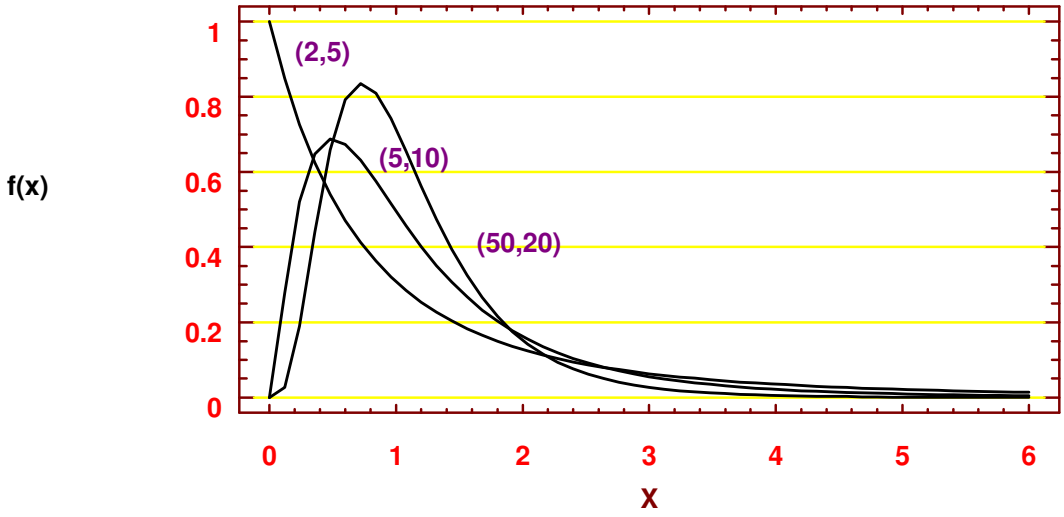
$$F = \frac{(X_1^2 + X_2^2 + ... + X_m^2)/m}{(Y_1^2 + Y_2^2 + ... + Y_n^2)/n}$$

aquesta segueix una distribució ***F de Snedecor*** amb m graus de llibertat al numerador i n al denominador i es simbolitza com $\mathbf{F_{m,n}}$.

De la definició anterior es desprèn que la distribució F és el quocient entre dues variables Khi al quadrat dividida cadascuna d'elles pels seus respectius graus de llibertat.

$$F = \frac{\chi_m^2/m}{\chi_n^2/n} \sim F_{m,n}$$

Gràfic 3.3 Distribucions F de Snedecor amb (2,5), (5,10) i (50,20) g.ll.



Característiques:

1. És una variable contínua i el seu recorregut es troba entre 0 i +∞.
2. Existeix un nombre infinit de distribucions F, una per a cada valor enter positiu de m i de n.
3. Aquesta distribució està completament caracteritzada pels seus graus de llibertat **m** i **n**. Així, l'esperança matemàtica i la variància són:

$$E(F) = \mu = \frac{n}{n-2} \quad n > 2 \qquad V(F) = \sigma^2 = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad n > 4$$

4. Es tracta de distribucions que presenten asimetria positiva. Això no obstant, en incrementar m i n, la seva asimetria disminueix.
5. Si X~ $F_{m,n}$ aleshores $\mathrm{Y} = \frac{1}{\mathrm{X}} \sim F_{n,m}$

   Aquesta característica rep el nom de propietat recíproca.
6. L'estadístic mostral $\frac{S_1^2}{S_2^2}$ obtingut a partir de dues mostres independents d'una població normal, on S_1^2 és la major de les dues variàncies mostrals, presenta una distribució F de Snedecor amb n_1-1 graus de llibertat al numerador i n_2-1 al denominador.

$$\frac{S_1^2}{S_2^2} = \frac{(n_1-1)S_1^2/(n_1-1)\sigma^2}{(n_2-1)S_2^2/(n_2-1)\sigma^2} = \frac{\chi^2_{n1-1}/(n_1-1)}{\chi^2_{n2-1}/(n_2-1)} \sim F_{(n1-1,n2-1)}$$

3.6 EXERCICIS PROPOSATS

Exercici 1. Per a les següents distribucions obteniu el valor x_0 que verifiqui $P(X \geq x_0) = p$:

a) X~ χ^2_3, p=0,005
b) X~ χ^2_{17}, p=0,975
c) X~ χ^2_{60}, p=0,01
d) X~ χ^2_{100}, p=0,025
e) X~ $F_{15,10}$, p=0,05
f) X~ $F_{15,10}$, p=0,01
g) X~ t_{10}, p=0,25
h) X~ t_{10}, p=0,975
i) X~ t_{25}, p=0,95
j) X~ t_5, p=0,20

Exercici 2. Per a les següents distribucions obteniu el valor x_0 que verifiqui $P(X < x_0) = p$:

a) X~ χ^2_3, p=0,005
b) X~ χ^2_{17}, p=0,975
c) X~ χ^2_{60}, p=0,01
d) X~ χ^2_{100}, p=0,025
e) X~ t_{10}, p=0,25
f) X~ t_{15}, p=0,8

Exercici 3. Trobeu les següents probabilitats:

a) P($\chi^2_{13} < 27,69$)
b) P($\chi^2_{13} \geq 27,69$)
c) P($\chi^2_{25} \geq 14,61$)
d) P($\chi^2_{17} < 5,7$)
e) P($t_{10} < -1,812$)
f) P($t_{15} > 2,131$)
g) P($t_{20} > -2,845$)
h) P($F_{15,20} \geq 2,20$)

Exercici 4. Si X_i ~N(0,1), obteniu $P\left(\sum_{i=1}^{14} X_i^2 < 6,57\right)$

Exercici 5. Sigui una població estadística modelitzada per la variable aleatòria discreta següent:

P(x)=0,5 si x=2; P(x)=0,3 si x=4; P(x)=0,2 si x=6 i P(x)=0 en altres casos.

Si s'extreu una mostra aleatòria de grandària 4, quin és el valor esperat i la variància de $\overline{X}$?

Exercici 6. Sigui X una v.a. discreta que només pot prendre els valors 0 i 5 amb probabilitats p i 1-p, respectivament. En formar totes les mostres possibles de grandària 2, l'esperança matemàtica de la v.a mitjana mostral ha estat 3. Quina és la probabilitat que la mitjana mostral prengui el valor 5?

Exercici 7. Sabem que l'edat dels participants d'un campionat de jocs de rol és una variable aleatòria amb esperança matemàtica igual a 16 anys i variància 324. Extreta una mostra aleatòria de 100 participants, quina és la probabilitat que l'edat mitjana mostral estigui compresa entre 14 i 19 anys?

Exercici 8. Una màquina d'omplir paquets de sèmola és ajustada de forma que el pes mitjà sigui de 500 g. i la seva desviació típica sigui de 20 g. L'encarregat ha decidit aturar la producció i reajustar la màquina si en extreure una mostra de 25 paquets el pes mitjà de la mostra és superior a 520 g. o inferior a 480 g. Si es suposa que el pes és una v.a. normal, quina és la probabilitat que s'hagi d'aturar la producció?

Exercici 9. El nombre de visites diàries que rep un assessor jurídic és una variable aleatòria amb una distribució no especificada, però amb una desviació estàndard de 16 visites. Si s'observa el nombre de visites durant 64 dies escollits a l'atzar, quina és la probabilitat que la mitjana mostral es trobi a no més de dues persones del vertader valor mitjà de visites diàries?

Exercici 10. Per al procés de producció d'un cert material, es coneix que la desviació típica del pes del producte és de 25 kg. Si en una mostra aleatòria de 50 productes, la probabilitat que la mitjana mostral prengui un valor superior a 250 kg és de 0,95, llavors quina és l'esperança matemàtica del pes del procés de producció?

Exercici 11. Les comissions que diàriament obté un venedor a domicili és una variable aleatòria normal. Es coneix que la probabilitat d'ingressar més de

1200€ en un mes determinat (25 dies de venda) és 0,15866, mentre que la probabilitat d'ingressar menys de 1080 € és 0,72575. Sota aquestes condicions, quin és el valor esperat i la desviació estàndard de la comissió que ingressa diàriament?

Exercici 12. Siguin X~N(0,σ=2) i Y~N(1,σ=3), variables independents. Si s'extreu una mostra aleatòria de grandària 10 de cadascuna d'aquestes poblacions estadístiques, quina és la probabilitat que la mitjana mostral de X sigui inferior a la de Y?

Exercici 13. En un referèndum, el SÍ va obtenir el 46% dels vots. Quina és la probabilitat que en un mostratge de 200 votants escollits a l'atzar, el SÍ obtingui majoria (més del 50%)?

Exercici 14. Se sap que només el 2% de les declaracions d'IRPF presenta errors de càlcul. Si una agència tributària d'Hisenda ha decidit revisar 400 declaracions triades a l'atzar, quina és la probabilitat que com a mínim un 3% de declaracions presenti aquest error?

Exercici 15. Si llancem un dau equilibrat 300 vegades, quina és la probabilitat que la diferència entre el percentatge de resultats parells dels 150 primers llançaments i el mateix percentatge dels darrers 150 llançaments sigui com a màxim de 0,1?

Exercici 16. El nombre de defectes en l'acabat d'unes peces de confecció segueix una llei de Poisson de paràmetre desconegut λ. S'obté una mostra de 3 peces i s'observa que la primera té 1 defecte, la segona en té 2 i la tercera no presenta cap defecte (1,2,0). Quina és la funció de versemblança de la mostra obtinguda?

Exercici 17.

D'una població exponencial Exp(λ) on λ només pot prendre els valors 1, 2, 3 i 4 s'obté la mostra següent: (0,43; 0,57; 0,36; 0,24; 0,3). Indiqueu la funció de versemblança d'aquesta mostra per als possibles valors de λ.

Exercici 18. Sigui una població caracteritzada per una v.a. X amb funció de densitat:

$$f(X)=\begin{cases}\theta x^{\theta-1} & 0\le x\le 1\ \theta>0\\ 0 & en\ altres\ casos\end{cases}$$

Indiqueu la funció de versemblança de la mostra (x_1, x_2, ...,x_n).

CAPÍTOL IV. ESTIMACIÓ DE PARÀMETRES

4.1 INTRODUCCIÓ

L'objectiu dels mètodes d'estimació és determinar els valors dels paràmetres que caracteritzen la distribució de probabilitat d'una població estadística. Així, per exemple, si sabem que una població es modelitza mitjançant una distribució Normal, per concretar l'esmentada distribució caldrà assignar un valor determinat als paràmetres μ i σ que la caracteritzen; si la població que es modelitza és una distribució de Poisson caldrà, aleshores, fixar un valor per al paràmetre λ, etc. Aquests valors normalment són desconeguts i, per tant, serà necessari fer-ne estimacions.

Per obtenir aquestes estimacions que anomenem estimacions puntuals, caldrà disposar d'estimadors adients, que no són més que estadístics mostrals i, per tant, variables aleatòries caracteritzades per les seves distribucions de probabilitat.

Per tal de seleccionar en cada cas l'estimador més apropiat hi ha establerts uns criteris que permeten valorar quin és, entre tots els possibles estimadors de cada paràmetre, l'òptim. Aquests criteris fan referència al valor esperat de l'estimador (criteri de no esbiaixament), a la seva variància (criteri d'eficiència) i al comportament d'ambdós quan la grandària de la mostra tendeix a infinit (criteri de consistència).

En alguns casos es tracta d'estimar paràmetres per als quals és gairebé intuïtiu quin és l'estimador adequat. En altres casos, on *a priori* no és tan evident l'estimador adient, és necessari disposar de mètodes d'estimació que permetin generar estimadors coherents per als paràmetres.

Quan estimem qualsevol paràmetre poblacional, tot i que disposem d'un estimador amb totes les propietats desitjables, sempre es comet un determinat error d'estimació, ja que normalment l'estimació difereix del valor real del paràmetre. Aquest error es defineix com la diferència entre l'estimació generada i el veritable valor del paràmetre i, per tant, és un valor desconegut ja que el valor del paràmetre ho és. Per això, no sabrem en cada estimació concreta quin ha estat l'error comès; però podrem valorar el grau de precisió assolit quan a partir d'una estimació puntual realitzem una estimació per interval. D'altra banda, també és possible fixar prèviament l'error d'estimació màxim tolerable i determinar quina és la grandària de la mostra adequada en aquest cas.

4.2 ESTIMADOR I ESTIMACIÓ

Suposem, per exemple, que estem interessats en conèixer quin és el valor esperat, μ, d'una població modelitzada per la variable aleatòria X definida com el '*temps transcorregut fins que es produeix la primera avaria d'un aparell electrodomèstic*'. Prescindim, per ara, d'establir quina és la distribució de probabilitat de la població. Per tal d'assignar un valor concret a μ disposem d'una mostra aleatòria de grandària n=3, (X_1, X_2, X_3) on X_i és el '*temps transcorregut fins que s'observa la primera avaria de l'aparell i-èsim*'. Com ja sabem, aquesta mostra és una variable aleatòria tridimensional i X_1, X_2 i X_3 són variables independents i distribuïdes idènticament a la població. Un cop triada la mostra observem, per exemple, la següent realització: x_1=5,0; x_2=6,4 i x_3=5,9. Per estimar μ, és a dir, per assignar un valor concret a la mitjana poblacional, una possibilitat consisteix en fixar com a valor de μ el valor de la mitjana de la mostra que en aquest cas és $\bar{X}$ = 5,77, llavors diem que '*l'estimador de μ és* $\bar{X}$ ', i ho simbolitzem com $\hat{\mu} = \bar{X}$. Com veiem, l'estimació concreta de μ és un valor de la variable aleatòria mitjana mostral i, en aquest cas, '*l'estimació de μ és igual a 5,77*' que s'indica per $\hat{\mu}$=5,77. Com que estem assignant al paràmetre un únic valor numèric estem fent una ***estimació puntual*** del paràmetre.

Definició:

Un ***estimador*** d'un paràmetre poblacional ϑ, que indicarem amb $\hat{\vartheta}$, és una funció de les observacions mostrals que s'utilitza per generar pronòstics (estimacions) del valor real del paràmetre.

$$\hat{\vartheta} = g(X_1, X_2, ..., X_n)$$

Com que és funció de les observacions mostrals, un estimador és una variable aleatòria amb una determinada distribució de probabilitat.

El problema de l'estimació rau en el fet que per a cada paràmetre poblacional pot haver-hi, i de fet hi ha, nombrosos estimadors possibles. Suposem, per exemple, que volem estimar la mitjana de la variable X={*durada (en hores) de la càrrega d'un cert model de bolígraf*} i disposem de la següent mostra de 10 observacions: x_1=23,0; x_2=25,2; x_3=26,3; x_4=27,3; x_5=28,0; x_6=28,4; x_7 =29,2; x_8=30,9; x_9=31,6 i x_{10}=35,1. Considerem els següents estimadors possibles per a μ amb les estimacions associades:

Estimador	*Estimació*
$\hat{\mu} = \bar{X}$	$\bar{x}$ = 28,5
$\hat{\mu}$ = Me (Mediana)	Me = 28,20
$\hat{\mu} = \bar{X}_8$ (Mitjana eliminant els valors màx. i mín)	$\bar{x}_8$ = 28,36

La qüestió que es planteja és: quina entre aquestes estimacions s'apropa més al veritable valor de μ? No podem respondre a aquesta qüestió, ja que desconeixem quin és el veritable valor de μ, però si analitzem les propietats dels estimadors podrem respondre a la següent pregunta: quin d'aquests estimadors, si s'utilitza en mostres successives, tendirà a generar estimacions més properes al veritable valor de μ?

4.3 PROPIETATS DELS ESTIMADORS

Per tal de triar entre estimadors alternatius d'un mateix paràmetre, cal establir un conjunt de propietats que fóra desitjable que complissin els estimadors.

Com que un estimador és una variable aleatòria aquestes propietats fan referència al comportament de la seva distribució de probabilitat.

4.3.1 NO ESBIAIXAMENT

El no esbiaixament és una propietat que fa referència al valor esperat del paràmetre.

Definició:

El ***biaix*** d'un estimador és la diferència entre el seu valor esperat i el veritable valor del paràmetre:

$$\text{Biaix}(\hat{\vartheta}) = E(\hat{\vartheta}) - \vartheta$$

Quan el biaix d'un estimador és positiu l'estimador tendeix sistemàticament a generar estimacions que estan per sobre del vertader valor del paràmetre (sobreestimació); quan el biaix és negatiu les estimacions generades tendeixen

sistemàticament a estar per sota del veritable valor del paràmetre (subestimació).

Suposem, per exemple, que volem estimar el paràmetre **b** d'una població Uniforme definida a l'interval [**0**;**b**]; és a dir, es tracta d'estimar el valor màxim que pot prendre la variable. Si disposem d'una mostra de grandària n=5, amb les següents observacions mostrals x_1=4,2 x_2=0,9 x_3=2,4 x_4=3,9 x_5=1,1, en principi, una possibilitat que sembla raonable és estimar el paràmetre **b** mitjançant l'estimador $\hat{b}$ =màx{x_i}; en aquest cas l'estimació de **b** és 4,2. Aquest estimador té biaix negatiu, és a dir, és esbiaixat i sempre subestimarà el veritable valor de **b**, ja que, com és evident, el valor màxim observat a la mostra mai podrà ser superior al paràmetre **b** que és el valor màxim que pot prendre la variable.

Considerem, d'altra banda, l'estimador $\tilde{b} = 2\overline{X}$, llavors l'estimació serà $\tilde{b}$ =5. Aquest estimador no tendeix a subestimar o sobreestimar sistemàticament el veritable valor del paràmetre. En efecte, com que el valor esperat d'aquesta distribució uniforme és $\mu=\frac{b}{2}$ es pot demostrar que el valor esperat d'aquest estimador coincideix amb el paràmetre poblacional. Efectivament:

E($\tilde{b}$)$= E(2\overline{X})$ = 2 μ = 2 b/2 = b Així doncs, Biaix($\tilde{b}$)=0.

Definició:

Diem que $\hat{\vartheta}$ és un estimador ***no esbiaixat*** del paràmetre ϑ o ***centrat*** en ϑ si el valor esperat de l'estimador és igual a ϑ.

$$E(\hat{\vartheta}) = \vartheta$$

òbviament, $\hat{\vartheta}$ és un estimador no esbiaixat quan Biaix($\hat{\vartheta}$) = 0.

Un estimador no esbiaixat és aquell que per a algunes mostres generarà estimacions per sobre del veritable valor del paràmetre i, per a d'altres, estimacions per sota, però no tendirà sistemàticament a sobreestimar o subestimar el valor del paràmetre.

Com hem vist al capítol anterior E($\overline{X}$)= μ. Per tant, la mitjana mostral és un estimador no esbiaixat del valor esperat poblacional. En canvi, si s'estima la variància poblacional mitjançant l'estimador $\hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n}$ es pot demostrar que E($\hat{\sigma}^2$) $\neq\sigma^2$. En conseqüència la variància de la mostra és un estimador esbiaixat de la variància poblacional. Per això es defineix la variància

mostral com $\hat{\sigma}^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$ que, com ja hem vist al capítol anterior, és un estimador no esbiaixat de σ^2.

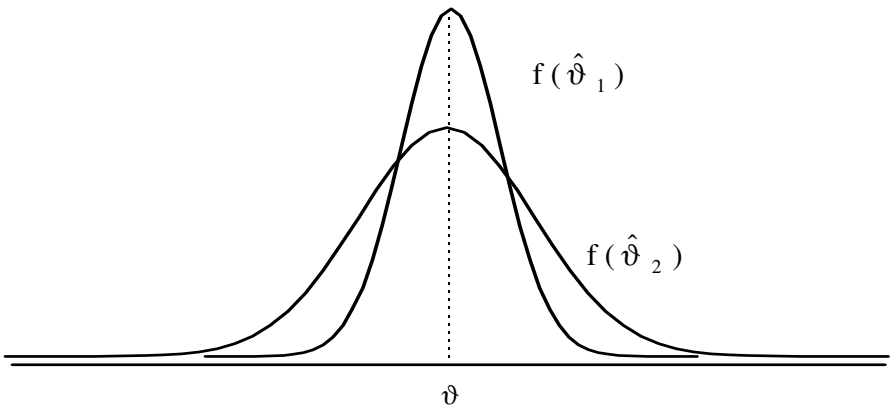
4.3.2 EFICIÈNCIA

El no esbiaixament és una propietat convenient però no suficient per si mateixa per determinar la bondat o qualitat d'un estimador.

Com que tots els estimadors no esbiaixats d'un paràmetre tenen el mateix valor esperat, per seleccionar-ne un sembla natural considerar les seves variàncies i triar el de menor variància o variància mínima.

Així, si per estimar un paràmetre ϑ disposem de dos estimadors no esbiaixats, $\hat{\vartheta}_1$ i $\hat{\vartheta}_2$, amb distribucions de probabilitat f($\hat{\vartheta}_1$) i f($\hat{\vartheta}_2$), respectivament, representades al gràfic 4.1, observem que els dos estimadors presenten distribucions centrades en ϑ (no esbiaixats) però la variància de $\hat{\vartheta}_1$ és més petita que la variància de $\hat{\vartheta}_2$. La qual cosa vol dir que la probabilitat d'obtenir estimacions properes al veritable valor de ϑ és més gran si es generen amb $\hat{\vartheta}_1$ que si es generen amb $\hat{\vartheta}_2$. Per tant, diem que $\hat{\vartheta}_1$ és ***més eficient*** que $\hat{\vartheta}_2$.

Gràfic 4.1 Estimadors no esbiaixats amb variàncies diferents.



Definició:

Diem que $\hat{\vartheta}$ és l'estimador ***eficient*** de ϑ si:

a) $\hat{\vartheta}$ és un estimador no esbiaixat i,

b) V($\hat{\vartheta}$) < V($\hat{\vartheta}_i$), essent $\hat{\vartheta}_i$ qualsevol altre estimador no esbiaixat de ϑ.

De la definició anterior es desprèn que només pot haver-hi un estimador eficient de ϑ, i és aquell que presenta la variància mínima entre els no esbiaixats.

Per al conjunt de tots els estimadors no esbiaixats d'un paràmetre es pot determinar quin és el valor mínim possible de la variància, mitjançant l'anomenada Cota inferior de Cramer-Rao. Així, si la variància d'un estimador no esbiaixat coincideix amb l'esmentada cota sabrem que es tracta de l'estimador de variància mínima o estimador EFICIENT.

Definició:

Donada una mostra aleatòria X_1, X_2, ... , X_n, d'una població amb distribució de probabilitat f(X; ϑ) si $\hat{\vartheta}$ és un estimador no esbiaixat de ϑ, la variància de $\hat{\vartheta}$ compleix la següent desigualtat que s'anomena ***cota de Cramer-Rao***:

$$V(\hat{\vartheta}) \geq \frac{1}{nE\left[\left(\frac{\partial \ln f(X;\vartheta)}{\partial \vartheta}\right)^2\right]}$$

Aquesta expressió indica que qualsevol estimador no esbiaixat de ϑ té variància major o igual a la cota anterior. Per tant, un estimador no esbiaixat és ***eficient*** només si la seva variància coincideix amb la Cota de Cramer-Rao.

Exemple 4.1

Sigui una població amb una proporció d'èxits π. Es vol comprovar que la proporció mostral d'èxits p és l'estimador eficient de π.

Solució:

X~B(1,π) amb P(x) = $\begin{cases} \pi^x (1-\pi)^{1-x} & x = 0,1 \\ 0 & \textit{enaltrescasos} \end{cases}$

E(X)= π i V(X)= $\pi(1-\pi)$.

La proporció mostral p= $\frac{\sum_{i=1}^{n} X_i}{n}$ es distribueix amb E(p)= π i V(p)= $\frac{\pi(1-\pi)}{n}$.

Per tant, és un estimador no esbiaixat.

Cota de Cramer-Rao= $\frac{1}{nE\left[\left(\frac{\partial \ln P(X;\pi)}{\partial \pi}\right)^2\right]}$

Desenvolupant l'expressió anterior per parts es té:

- $\ln P(X;\pi)=\ln[\pi^x (1-\pi)^{1-x}]=X \ln\pi+(1-X)\ln(1-\pi)$
- $\frac{\partial \ln P(X;\pi)}{\partial \pi}=\frac{X}{\pi}-\frac{(1-X)}{(1-\pi)}=\frac{X-\pi}{\pi(1-\pi)}$
- $E\left[\left(\frac{\partial \ln P(X;\pi)}{\partial \pi}\right)^2\right]=E\left[\frac{X-\pi}{\pi(1-\pi)}\right]^2=\frac{1}{[\pi(1-\pi)]^2}E[X-\pi]^2=\frac{V(X)}{[\pi(1-\pi)]^2}=\frac{1}{\pi(1-\pi)}$

Cota de Cramer-Rao$=\frac{1}{nE\left[\left(\frac{\partial \ln P(X;\pi)}{\partial \pi}\right)^2\right]}=\frac{1}{n\frac{1}{\pi(1-\pi)}}=\frac{\pi(1-\pi)}{n}$

La variància de la proporció mostral coincideix amb la Cota de Cramer-Rao, per tant, p és l'estimador eficient de π.

En general, si volem comparar dos estimadors alternatius de ϑ, $\hat{\vartheta}_i$ i $\hat{\vartheta}_j$, considerem l'eficiència relativa que es defineix com el quocient entre les variàncies dels dos estimadors.

Definició:

Donats dos estimadors no esbiaixats del paràmetre ϑ, $\hat{\vartheta}_i$ i $\hat{\vartheta}_j$, diem que $\hat{\vartheta}_i$ és més eficient que $\hat{\vartheta}_j$ si $V(\hat{\vartheta}_i) < V(\hat{\vartheta}_j)$, i *l'**eficiència relativa*** és $\frac{V(\hat{\vartheta}_j)}{V(\hat{\vartheta}_i)}$.

Aquest quocient ens indica l'increment percentual d'eficiència d'un estimador respecte a l'altre. Si ambdós estimadors són igualment eficients, l'eficiència relativa és 1.

Exemple 4.2

Es tracta d'estimar el valor esperat d'una població, μ, amb una mostra de grandària n=3, i disposem de dos estimadors alternatius:

$\hat{\mu}_1=\overline{X}$ *i* $\hat{\mu}_2=\frac{x_1}{4}+\frac{x_2}{4}+\frac{x_3}{2}$. *Quin és més eficient?*

<u>*Solució:*</u>

Els valors esperats i les variàncies dels estimadors són:

$E(\hat{\mu}_1)=\mu$ i $V(\hat{\mu}_1)=\frac{\sigma^2}{3}=0{,}333\ \sigma^2$

$E(\hat{\mu}_2)=\mu$ i $V(\hat{\mu}_2)=\frac{6\sigma^2}{16}=0{,}375\ \sigma^2$

Els dos estimadors són no esbiaixats; $\hat{\mu}_1$ és més eficient que $\hat{\mu}_2$ i l'eficiència relativa és igual al quocient $0{,}375\ \sigma^2/0{,}333\ \sigma^2 = 1{,}126$.

Dintre de la classe dels estimadors no esbiaixats considerem la subclasse dels estimadors LINEALS**,** és a dir, aquells que són funcions lineals de les observacions mostrals, i dintre d'aquests seleccionem l'estimador ÒPTIM.

Definició:

$\hat{\vartheta}$ és l'***estimador lineal, no esbiaixat i òptim*** de ϑ si:

a) $\hat{\vartheta}$ és funció lineal de les observacions mostrals,

b) és no esbiaixat, $E(\hat{\vartheta}) = \vartheta$, i

c) és eficient, $V(\hat{\vartheta}) < V(\hat{\vartheta}_i)$, essent $\hat{\vartheta}_i$ qualsevol altre estimador lineal no esbiaixat de ϑ.

Exemple 4.3

Es tracta de construir l'estimador lineal, no esbiaixat i de variància mínima (òptim) del valor esperat poblacional, μ, a partir d'una mostra de grandària n, $(X_1, X_2, ... X_n)$, on $E(X_i) = \mu$ i $V(X_i)=\sigma^2$ per a $i = 1, 2, ..., n$.

<u>*Solució*</u>:

L'estimador que volem ha d'ésser:

- Lineal, és a dir, de la forma $\hat{\mu} = a_1X_1 + a_2X_2 + ... + a_nX_n$.
- No esbiaixat:

$E(\hat{\mu}) = E(a_1X_1 + a_2X_2 + ... + a_nX_n) = E(a_1X_1)+E(a_2X_2) + ... + E(a_nX_n) =$

$= a_1 E(X_1)+a_2 E(X_2) + ... + a_n E(X_n) = a_1 \mu + a_2 \mu + ... a_n \mu = \mu \sum_{i=1}^{n} a_i = \mu$

Es dedueix, per tant, que per assolir la propietat de no esbiaixament sobre els coeficients $a_1,a_2,...,a_n$ s'ha d'imposar la restricció: $\sum_{i=1}^{n} a_i = 1$.

- Eficient:

La variància de l'estimador és:

$V(\hat{\mu}) = V(a_1X_1 + a_2X_2 + ... + a_nX_n) = a_1^2 V(X_1) + a_2^2 V(X_2) + ... + a_n^2 V(X_n) =$

$= a_1^2 \sigma^2 + a_2^2 \sigma^2 + ...+ a_n^2 \sigma^2$

Perquè la variància sigui mínima, s'ha de minimitzar aquesta expressió subjecta a la restricció $\sum_{i=1}^{n} a_i = 1$.

Es defineix la funció de Lagrange: $\varphi = a_1^2 \sigma^2 + a_2^2 \sigma^2 + ... + a_n^2 \sigma^2 - \lambda (a_1 + a_2 + ... + a_n - 1)$

Perquè φ assoleixi un mínim és condició necessària i suficient (ja que es tracta d'una suma de quadrats) que les primeres derivades de φ respecte a a_1, a_2, ..., a_n siguin nul·les. Així doncs,

$\frac{\delta\varphi}{\delta a_1} = 2a_1 \sigma^2 - \lambda = 0$

$\frac{\delta\varphi}{\delta a_2} = 2a_2 \sigma^2 - \lambda = 0$

.....

$\frac{\delta\varphi}{\delta a_n} = 2a_n \sigma^2 - \lambda = 0$

Si aïllem λ queda: $\lambda = 2a_1 \sigma^2 = 2a_2 \sigma^2 = ... = 2a_n \sigma^2$

Per tant, $a_1 = a_2 = ... = a_n$ i, de la restricció $a_1 + a_2 + ... + a_n = 1$, es dedueix que $a_i = 1/n \ \forall i$.

De forma que l'estimador òptim de μ és $\hat{\mu} = \frac{X_1}{n} + \frac{X_2}{n} + ... + \frac{X_n}{n} = \overline{X}$.

4.4 PROPIETATS ASIMPTÒTIQUES

Les propietats asimptòtiques d'un estimador es refereixen a la seva distribució de probabilitat quan la grandària de la mostra és elevada, o, més exactament, quan la grandària de la mostra tendeix a infinit.

4.4.1 NO ESBIAIXAMENT ASIMPTÒTIC

Definició:

$\hat{\vartheta}$ és un estimador ***asimptòticament no esbiaixat*** de ϑ o ***asimptòticament centrat en*** ϑ si el biaix de $\hat{\vartheta}$ tendeix a zero, o l'esperança de l'estimador tendeix al paràmetre poblacional, quan n tendeix a infinit.

$$\lim_{n\to\infty} Biaix(\hat{\vartheta}) = 0$$
$$\lim_{n\to\infty} E(\hat{\vartheta}) = \vartheta$$

Aquesta propietat garanteix que per a mostres grans el valor esperat de l'estimador tendeix al valor del paràmetre que s'està estimant.

Per exemple, considerem els dos estimadors següents de la mitjana poblacional:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + ... + X_{n-1} + 2X_n}{n} \qquad \hat{\mu}_2 = \frac{X_1 + X_2 + ... + X_n}{n+1}$$

amb valors esperats

$$E(\hat{\mu}_1) = \frac{1}{n} E(X_1 + X_2 + ... + X_{n-1} + 2X_n) =$$

$$= \frac{1}{n}\left[E(X_1) + E(X_2) + ... + E(X_{n-1}) + 2E(X_n)\right] =$$

$$= \frac{1}{n}(\mu + \mu + ... + \mu + 2\mu) = \frac{n+1}{n}\mu \neq \mu$$

$$E(\hat{\mu}_2) = \frac{1}{n+1} E(X_1 + X_2 + ... + X_n) = \frac{1}{n+1}\left[E(X_1) + E(X_2) + ... + E(X_n)\right] =$$

$$= \frac{1}{n+1}(\mu + \mu + ... + \mu) = \frac{n}{n+1}\mu \neq \mu$$

Com veiem, ambdós estimadors són esbiaixats, amb

$$\text{Biaix}(\hat{\mu}_1) = E(\hat{\mu}_1) - \mu = \frac{n+1}{n}\mu - \mu = \frac{\mu}{n}$$

$$\text{Biaix}(\hat{\mu}_2) = E(\hat{\mu}_2) - \mu = \frac{n}{n+1}\mu - \mu = \frac{-1}{n+1}\mu$$

Observem que quan n tendeix a infinit

$$\lim_{n\to\infty} Biaix(\hat{\mu}_1) = \lim_{n\to\infty} \frac{\mu}{n} = 0$$

$$\lim_{n\to\infty} Biaix(\hat{\mu}_2) = \lim_{n\to\infty} \frac{-\mu}{n+1} = 0$$

Com que el biaix de $\hat{\mu}_1$ i el biaix de $\hat{\mu}_2$ tendeixen a zero quan n tendeix a infinit, diem que ambdós són estimadors asimptòticament no esbiaixats de μ.

Anàlogament, podríem analitzar el comportament de E($\hat{\mu}_1$) i de E($\hat{\mu}_2$) quan n tendeix a infinit:

$$\lim_{n\to\infty} E(\hat{\mu}_1) = \lim_{n\to\infty} \frac{n+1}{n}\mu = \mu$$

$$\lim_{n\to\infty} E(\hat{\mu}_2) = \lim_{n\to\infty} \frac{n}{n+1}\mu = \mu$$

Observem que quan n tendeix a infinit, els valors esperats d'ambdós estimadors tendeixen al valor del paràmetre i, per tant, són asimptòticament no esbiaixats.

4.4.2 CONSISTÈNCIA

La consistència fa referència a la dispersió de l'estimador, però no respecte al seu valor esperat (variància de l'estimador), sinó respecte al veritable valor del paràmetre. Més concretament, analitza el comportament de l'esmentada dispersió quan augmenta la grandària de la mostra.

Definició:

*L'**error quadràtic mitjà*** d'un estimador del paràmetre ϑ mesura la dispersió dels valors de l'estimador respecte al valor del paràmetre.

$$\text{EQM} = \text{E}\left(\hat{\vartheta} - \vartheta\right)^2$$

D'aquesta definició es dedueix que és convenient que l'EQM prengui valors petits per garantir la qualitat o bondat de les estimacions, ja que això significa que la dispersió és petita, o que les estimacions estan molt concentrades entorn del veritable valor del paràmetre.

L'error quadràtic mitjà depèn alhora de la variància de l'estimador i del biaix, si en tingués.

$$\text{EQM}(\hat{\vartheta}) = E(\hat{\vartheta} - \vartheta)^2 = E\left[\hat{\vartheta} - E(\hat{\vartheta}) + E(\hat{\vartheta}) - \vartheta\right]^2 = E\left[\hat{\vartheta} - E(\hat{\vartheta})\right]^2 + E\left[E(\hat{\vartheta}) - \vartheta\right]^2 +$$
$$+ 2E\{\left[\hat{\vartheta} - E(\hat{\vartheta})\right]\left[E(\hat{\vartheta}) - \vartheta\right]\}$$

El tercer sumand és igual a zero i, per tant, l'EQM es pot expressar com:

$$\text{EQM}(\hat{\vartheta}) = V(\hat{\vartheta}) + Biaix^2(\hat{\vartheta})$$

En el cas que $\hat{\vartheta}$ sigui un estimador no esbiaixat, l'EQM coincideix amb la variància de $\hat{\vartheta}$.

Definició:

$\hat{\vartheta}$ és un estimador de ϑ ***consistent*** en error quadràtic mitjà si aquest últim convergeix a zero quan la grandària mostral tendeix a infinit.

$$\lim_{n\to\infty} EQM(\hat{\vartheta}) = \lim_{n\to\infty} V(\hat{\vartheta}) + \lim_{n\to\infty} Biaix^2(\hat{\vartheta}) = 0$$

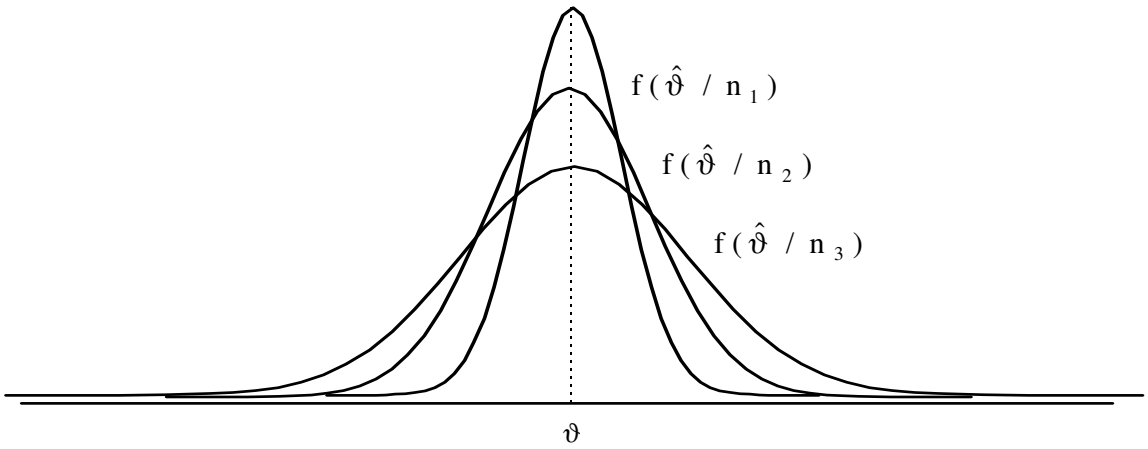
Respecte a la consistència d'un estimador podem considerar les següents situacions:

a) Si $\hat{\vartheta}$ és un estimador no esbiaixat, perquè sigui consistent cal que $\lim_{n\to\infty} V(\hat{\vartheta}) = 0$.

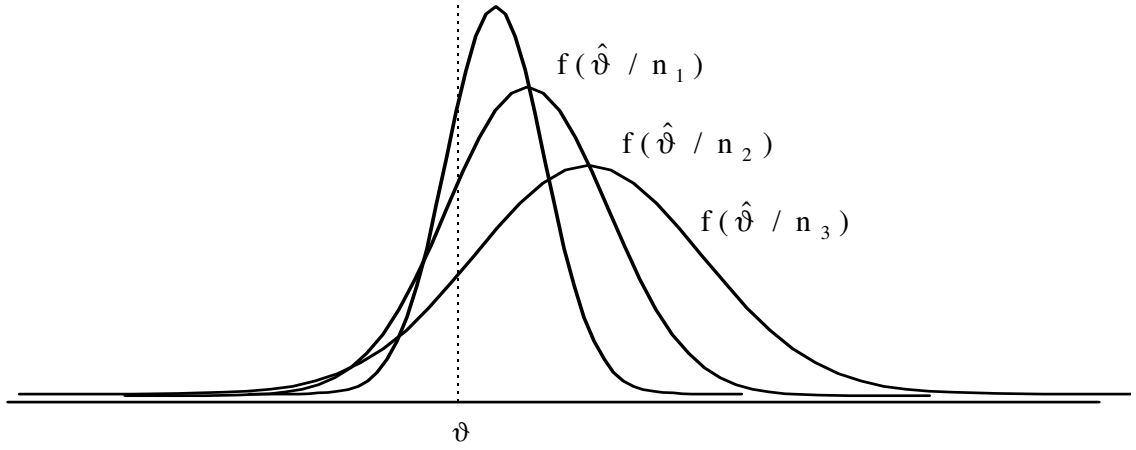
b) Si $\hat{\vartheta}$ és un estimador esbiaixat, perquè sigui consistent cal que $\lim_{n\to\infty} V(\hat{\vartheta}) = 0$ i $\lim_{n\to\infty} Biaix(\hat{\vartheta}) = 0$, simultàniament. Això vol dir que per garantir la consistència és condició necessària però no suficient que l'estimador sigui asimptòticament no esbiaixat.

El gràfic 4.2 il·lustra el comportament asimptòtic d'un estimador no esbiaixat i consistent de ϑ; i el gràfic 4.3 el comportament asimptòtic d'un estimador esbiaixat però consistent de ϑ per a diferents grandàries mostrals ($n_1 > n_2 > n_3$).

Gràfic 4.2 Estimador no esbiaixat i consistent



Gràfic 4.3 Estimador esbiaixat i consistent



Exemple 4.4

Es proposen els següents estimadors de μ i es vol comprovar si són consistents:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + ... + X_{n-1} + 2X_n}{n} \text{ i } \hat{\mu}_2 = \frac{X_1 + X_2 + ... + X_n}{n+1}$$

Solució:

Com hem vist a l'apartat 4.4.1, ambdós són asimptòticament no esbiaixats, per tant, per provar si són consistents només cal observar el comportament de les seves variàncies quan n tendeix a infinit.

Les variàncies són:

$$V(\hat{\mu}_1) = \frac{1}{n^2} V(X_1 + X_2 + ... + X_{n-1} + 2X_n) =$$

$$= \frac{1}{n^2} \left[V(X_1) + V(X_2) + ... + V(X_{n-1}) + 4V(X_n) \right] =$$

$$= \frac{1}{n^2} \left[\sigma^2 + \sigma^2 + ... + \sigma^2 + 4\sigma^2 \right] = \frac{n+3}{n^2} \sigma^2$$

$$V(\hat{\mu}_2) = \frac{1}{(n+1)^2} V(X_1 + X_2 + ... + X_{n-1} + X_n) =$$

$$= \frac{1}{(n+1)^2} \left[V(X_1) + V(X_2) + ... + V(X_{n-1}) + V(X_n) \right] =$$

$$= \frac{1}{(n+1)^2} (\sigma^2 + \sigma^2 + ... + \sigma^2) = \frac{n}{(n+1)^2} \sigma^2$$

Veiem ara quin és el límit on tendeixen $V(\hat{\mu}_1)$ i $V(\hat{\mu}_2)$:

$$\lim_{n\to\infty} V(\hat{\mu}_1) = \lim_{n\to\infty} \frac{(n+3)}{n^2}\sigma^2 = 0$$

$$\lim_{n\to\infty} V(\hat{\mu}_2) = \lim_{n\to\infty} \frac{n}{(n+1)^2}\sigma^2 = 0$$

Així doncs,

$$\lim_{n\to\infty} EQM(\hat{\mu}_1) = \lim_{n\to\infty} V(\hat{\mu}_1) + \lim_{n\to\infty} Biaix^2(\hat{\mu}_1) = 0$$

$$\lim_{n\to\infty} EQM(\hat{\mu}_2) = \lim_{n\to\infty} V(\hat{\mu}_2) + \lim_{n\to\infty} Biaix^2(\hat{\mu}_2) = 0$$

i, per tant, $\hat{\mu}_1$ i $\hat{\mu}_2$ són estimadors consistents de μ.

4.5 MÈTODES D'ESTIMACIÓ

Els mètodes d'estimació estableixen criteris objectius que permeten identificar entre els possibles procediments d'obtenció dels estimadors d'un paràmetre aquells que resultin més plausibles. Els dos mètodes que s'exposen a continuació garanteixen que l'estimador generat presenti almenys la propietat de consistència. Pel que fa a les altres propietats, s'haurà de provar en cada cas si les posseeixen. Ambdós mètodes requereixen el coneixement de la forma de la distribució de probabilitat poblacional.

4.5.1 MÈTODE DELS MOMENTS

El mètode d'estimació dels moments proposat per K. Pearson (1856-1936) és el procediment d'estimació més senzill. Es basa en la idea que els moments poblacionals poden ser estimats a partir dels seus corresponents moments mostrals. Així, l'estimador pel mètode dels moments de la mitjana poblacional, μ, és el primer moment mostral respecte a l'origen o mitjana mostral, $\overline{X}$; el de la variància, σ^2, és el segon moment mostral respecte a la mitjana, S^2, i així successivament.

Definició:

Donada una població estadística, modelitzada per la variable aleatòria X amb distribució de probabilitat f(x) (o P(x)) i caracteritzada pel paràmetre ϑ desconegut, l'expressió $\hat{\vartheta}$ diem que és *l'**estimador de ϑ pel mètode dels moments*** si s'obté de l'equació que resulta d'igualar el primer moment ordinari poblacional amb el seu corresponent moment mostral.

És a dir, si la distribució de probabilitat només depèn d'un únic paràmetre desconegut, ϑ, com que els moments ordinaris poblacionals de X depenen d'aquest paràmetre, per obtenir l'estimador de ϑ pel mètode dels moments caldrà:

1. Especificar la relació que existeix entre el paràmetre ϑ i el primer moment poblacional, $\mu=g(\vartheta)$.
2. Substituir a l'expressió anterior el moment poblacional pel corresponent moment mostral, $\overline{X}=g(\hat{\vartheta})$.
3. Aïllar de l'equació anterior $\hat{\vartheta}$ en funció de $\overline{X}$, $\hat{\vartheta}=f(\overline{X})$.

En extreure una mostra aleatòria (x_1, x_2,..., x_n) particular s'obté una estimació de ϑ quan es substitueix el valor $\overline{X}$ a $\hat{\vartheta}=f(\overline{X})$.

Exemple 4.5

Donada una variable aleatòria X amb funció de densitat:

$$f(x;\vartheta)=\begin{cases}(\vartheta+1)x^{\vartheta} & 0\le x\le 1\\ 0 & \textit{enaltrescasos}\end{cases}$$

a) Determineu l'estimador de ϑ pel mètode dels moments.

b) Calculeu l'estimació generada per la mostra: (0,25; 0,2; 0,1; 0,3; 0,5; 0,25; 0,75; 0,6; 0,4; 0,1).

Solució:

a) Primer moment ordinari poblacional:

$$\alpha_1=\mu=E(X)=\int_0^1 x(\vartheta+1)x^{\vartheta}dx=\int_0^1(\vartheta+1)x^{\vartheta+1}dx=(\vartheta+1)\frac{x^{\vartheta+2}}{\vartheta+2}\Big]_0^1=\frac{\vartheta+1}{\vartheta+2}$$

Primer moment mostral: $a_1=\overline{X}$.

Igualem:

$$\hat{\mu}=\overline{X}\ \Rightarrow\ \overline{X}=\frac{\hat{\vartheta}+1}{\hat{\vartheta}+2}\ \Rightarrow\ \overline{X}(\hat{\vartheta}+2)=\hat{\vartheta}+1\ \Rightarrow\ \hat{\vartheta}\overline{X}-\hat{\vartheta}=1-2\overline{X}$$

De l'equació anterior s'obté l'estimador: $\hat{\vartheta} = \frac{1-2\overline{X}}{\overline{X}-1}$

b) $\overline{X} = \frac{0{,}25 + 0{,}2 + ... + 0{,}1}{10} = 0{,}345 \Rightarrow \hat{\vartheta} = \frac{1-2\cdot 0{,}345}{0{,}345-1} = -0{,}47$

En general, quan la funció de densitat (o de quantia) depèn de k paràmetres desconeguts, s'igualen els k primers moments ordinaris de la mostra amb els corresponents moments de la població fins a obtenir un nombre suficient d'equacions on figuren com a incògnites els paràmetres que es volen estimar. De la solució d'aquest sistema s'obtenen les expressions dels estimadors pel mètode dels moments.

El mètode dels moments només utilitza la informació mostral relacionada amb els seus moments, és a dir, no té en compte la distribució de probabilitat mostral. Això fa que, en general, no es puguin garantir gaires propietats. Tot i així, es demostra que els estimadors pel mètode dels moments són no esbiaixats quan es dedueixen a partir del moment ordinari d'ordre u.

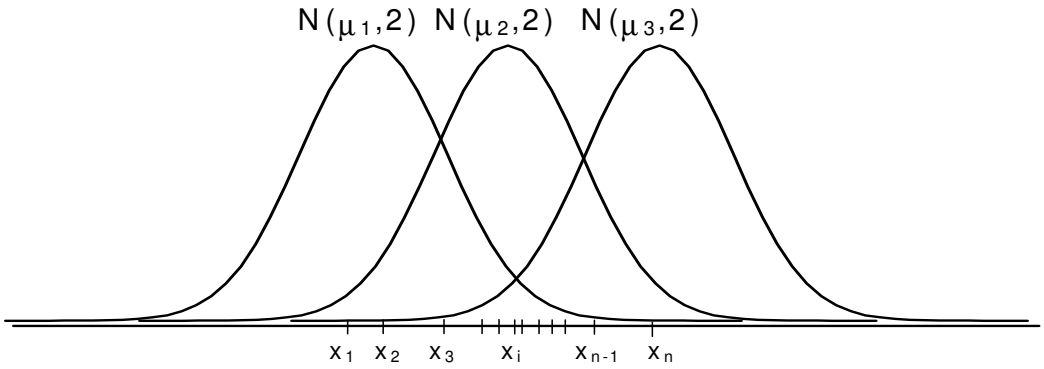
Es demostra que els estimadors pel mètode dels moments, sota condicions molt generals, són CONSISTENTS en EQM. (Evidentment també poden complir altres propietats però, en general, aquesta és l'única que tots ells acostumen a presentar.)

4.5.2 MÈTODE DE LA MÀXIMA VERSEMBLANÇA

El mètode d'estimació de la màxima versemblança, desenvolupat per R.A.Fisher (1890-1964), és un dels mètodes més utilitzats per a l'estimació puntual. Es basa en seleccionar entre tots els valors possibles del paràmetre aquell per al qual sigui màxima la probabilitat d'haver generat la mostra disponible.

Com ja hem vist, poblacions diferents generen mostres diferents i, per tant, una determinada mostra és més probable (versemblant) que provingui d'una població concreta que d'altres. Així, per exemple, si d'una població N(μ,2), amb μ desconegut, s'extreu una mostra (x_1, x_2,..., x_n) és possible que aquesta hagi estat generada per una N(μ_1,2) concreta, o per una altra N(μ_2,2), o per una N(μ_k,2). Hi ha, per tant, infinites possibilitats. Entre totes les alternatives, la funció de versemblança ens permet trobar la població que amb més probabilitat ha generat la mostra donada.

Gràfic 4.4 Tres poblacions Normals que poden haver generat una mostra concreta.



Així, per exemple, si la mostra és la que recull el gràfic 4.4, sembla més versemblant que les observacions mostrals hagin estat generades per la població $N(\mu_2,2)$, tot i que la mostra podria procedir de qualsevol de les altres poblacions. En cert sentit podem dir, doncs, que la mostra 'selecciona' la població que amb més versemblança l'ha generada.

Definició:

D'una població estadística, modelitzada per la variable aleatòria X amb distribució de probabilitat f(x) (o P(x)) i caracteritzada pel paràmetre ϑ desconegut, s'extreu una mostra aleatòria (X_1, X_2,..., X_n). *L'**estimació màxim-versemblant*** és el valor de ϑ que determina la població estadística que amb major probabilitat ha generat la mostra anterior. És a dir, l'estimació màxim-versemblant és el valor de ϑ, que indicarem per $\hat{\vartheta}$, que maximitza la funció de versemblança de la mostra.

Per obtenir l'estimador màxim-versemblant caldrà:

1. Especificar la funció de versemblança: $\ell(X_1, X_2,..., X_n; \vartheta)$
2. Determinar el màxim de la funció anterior igualant a zero la primera derivada respecte a ϑ.

Si la funció de versemblança depèn d'un únic paràmetre, només existirà una primera derivada de la funció ($\frac{\partial \ell}{\partial \vartheta}$). Si la funció depèn de k paràmetres, $\ell(X_1, X_2,..., X_n; \vartheta_1,\vartheta_2,...\vartheta_k)$ hi ha k primeres derivades parcials que igualades a zero permeten determinar els k valors dels estimadors màxim-versemblants, $\hat{\vartheta}_1, \hat{\vartheta}_2,..., \hat{\vartheta}_k$.

A la pràctica la derivada de la funció de versemblança acostuma a ser força complicada i és convenient treballar amb el seu logaritme neperià, ja que ambdues funcions presenten el mateix valor màxim (si existeix) en ser la funció logaritme una transformació monotònica.

Exemple 4.6

Es vol determinar l'estimador màxim-versemblant del paràmetre π d'una població dicotòmica.

Solució:

X~B(1,π) amb P(x) = $\begin{cases} \pi^x (1-\pi)^{1-x} & x = 0,1 \\ 0 & en\ altres\ casos \end{cases}$

- La funció de versemblança d'una mostra de grandària n és:

$\ell(X_1, X_2,..., X_n; \pi)=P(X_1; \pi)P(X_2; \pi)...P(X_n; \pi)=\pi^{x1} (1-\pi)^{1-x1} \pi^{x2} (1-\pi)^{1-x2}... \pi^{xn} (1-\pi)^{1-xn}$

$= \pi^{x1+x2+...+xn} (1-\pi)^{n-(x1+x2+...+xn)} = \pi^{\sum xi} (1-\pi)^{n-\sum xi}$

- El logaritme neperià és:

  $L=\ln \ell(X_1, X_2,..., X_n; \pi) = \ln [\pi^{\sum xi} (1-\pi)^{n-\sum xi}] = (\sum x_i) \ln \pi + (n-\sum x_i) \ln (1-\pi)$

- La primera derivada de L respecte a π:

$$\frac{\partial L}{\partial \pi} = \frac{\partial \left[\sum x_i \ln \pi + (n - \sum x_i)\ln(1-\pi)\right]}{\partial \pi} = \frac{\sum x_i}{\pi} - \frac{n - \sum x_i}{1-\pi}$$

- Per trobar el màxim igualem a zero la primera derivada:

$$\frac{\sum x_i}{\hat{\pi}} - \frac{n - \sum x_i}{1-\hat{\pi}} = 0 \quad \Rightarrow \quad (1-\hat{\pi})\sum x_i = \hat{\pi}(n - \sum x_i) \quad \Rightarrow$$

$$\sum x_i - \hat{\pi}\sum x_i = n\hat{\pi} - \hat{\pi}\sum x_i \quad \Rightarrow \quad \sum x_i = n\hat{\pi}$$

- Per últim, en aïllar $\hat{\pi}$ queda: $\hat{\pi} = \frac{\sum_{i=1}^{n} x_i}{n}$ = p

Per tant, l'estimador màxim-versemblant de la proporció poblacional és la proporció mostral.

Exemple 4.7

Donada una variable aleatòria X amb funció de densitat:

$$f(x;\vartheta) = \begin{cases} (\vartheta+1)x^{\vartheta} & 0 \le x \le 1 \\ 0 & en altres casos \end{cases}$$

a) Determineu l'estimador de ϑ pel mètode de la màxima versemblança.

b) Calculeu l'estimació generada per la mostra següent: (0,25; 0,2; 0,1; 0,3; 0,5; 0,25; 0,75; 0,6; 0,4; 0,1).

Solució:

a) Obtenció de l'estimador màxim-versemblant.

- La funció de versemblança:

$\ell(X_1, X_2,..., X_n; \vartheta) = f(X_1; \vartheta)f(X_2; \vartheta)...f(X_n; \vartheta)= (\vartheta+1)x_1^{\vartheta} (\vartheta+1)x_2^{\vartheta} ... (\vartheta+1)x_n^{\vartheta} = (\vartheta+1)^n \prod x_i^{\vartheta}$

- El logaritme de la funció de versemblança:

$L = \ln [(\vartheta+1)^n \prod x_i^{\vartheta}] = n \ln (\vartheta+1)+ \vartheta\sum \ln x_i$

- La primera derivada de L respecte a ϑ:

$$\frac{\partial L}{\partial \vartheta} = n\frac{1}{\vartheta+1} + \sum \ln x_i$$

- Igualem a zero la primera derivada i aïllem $\hat{\vartheta}$:

$$n\frac{1}{\hat{\vartheta}+1} + \sum \ln x_i = 0 \quad \Rightarrow \quad \frac{1}{\hat{\vartheta}+1} = \frac{-\sum \ln x_i}{n} \quad \Rightarrow \quad \hat{\vartheta}+1 = -\frac{n}{\sum \ln x_i} \quad \Rightarrow$$

$$\hat{\vartheta} = -\left(\frac{n}{\sum \ln x_i} + 1\right)$$

b) Obtenció de l'estimació.

$\sum ln\, x_i = \ln 0{,}25+\ln 0{,}2+\ln 0{,}1+...+\ln 0{,}1= -12{,}599 \Rightarrow$

$$\hat{\vartheta} = -(\frac{10}{-12{,}599} + 1) = -0{,}206$$

Les propietats que, generalment, presenten els estimadors màxim-versemblants són:

- *asimptòticament no esbiaixats*.
- *consistents*.
- *no necessàriament eficients*, però, si existeix un estimador eficient aquest serà l'estimador màxim-versemblant.

Per últim, cal assenyalar que la distribució de probabilitat de l'estimador màxim-versemblant convergeix a la distribució normal quan la grandària de la mostra tendeix a infinit.

4.6 ESTIMACIÓ PER INTERVAL

Els mètodes d'estimació puntual, fins i tot en el millor dels casos, no poden garantir que l'estimació sigui ajustada. És a dir, sabem que la distribució de l'estimador estarà al voltant del paràmetre desconegut ϑ, però cada estimació

pot ésser més gran o més petita que el veritable valor de ϑ i, per tant, s'està cometent un *error d'estimació* que indicarem per:

$$d=\hat{\vartheta}-\vartheta$$

Aquest error és una variable aleatòria que mesura la precisió de l'estimació ja que com més petit sigui l'error més precisa serà l'estimació que realitzem.

Un mètode que ens permet controlar la precisió que s'obté en realitzar una estimació és l'estimació per interval, que consisteix en donar una mesura de l'error d'estimació en termes probabilístics. És a dir, valorar la probabilitat de l'error que es comet quan utilitzem $\hat{\vartheta}$ en lloc de ϑ.

Definició:

Un ***interval de confiança*** és un conjunt de valors que amb una determinada probabilitat (grau de confiança) conté el veritable valor del paràmetre que es vol estimar.

Per determinar l'interval, en primer lloc, s'ha de decidir l'anomenat ***grau o nivell de confiança*****,** que es simbolitza com a $1-\alpha$, i és la probabilitat que té l'interval de contenir el veritable valor del paràmetre. El nivell de confiança normalment ha de ser elevat, però s'ha de tenir en compte que a mesura que s'incrementa aquest, la precisió de l'estimació disminueix; en concret, per a un nivell de confiança igual a 1, l'interval té una amplitud infinita i, per tant, l'estimació no té cap precisió. Generalment es treballa amb una confiança del 90%, 95% o 99%.

Una cop fixat el nivell de confiança, $1-\alpha$, es tracta de determinar els límits de l'interval, que indicarem per a i b, que verifiquin que hi ha una probabilitat $1-\alpha$ que l'interval [**a**; **b**] contingui el veritable paràmetre ϑ:

$$P[a\leq\vartheta\leq b]=1-\alpha$$

Aquests límits dependran de l'estimador del paràmetre, de la seva distribució de probabilitat i de la grandària de la mostra.

Un interval amb una confiança de l'$(1-\alpha)100\%$ de contenir el vertader valor del paràmetre vol dir que, donada una mostra concreta i, per tant, fixat un interval, no podem estar segurs que aquest contingui el paràmetre poblacional però, en obtenir infinites mostres, esperem que aproximadament l'$(1-\alpha)100\%$ dels intervals calculats continguin el vertader paràmetre poblacional i, en conseqüència, que la proporció d'intervals que no continguin el vertader paràmetre sigui de $\alpha\cdot 100\%$. Aquesta probabilitat α s'anomena ***nivell de significació*** i recull el grau d'error que presenta l'interval de confiança.

4.6.1 OBTENCIÓ D'UN INTERVAL DE CONFIANÇA PER A μ

El procediment per obtenir un interval de confiança per al valor esperat μ d'una població normal dependrà del supòsit que la variància poblacional sigui o no coneguda. En primer lloc exposem el cas de variància coneguda i, per tant, només cal estimar la mitjana poblacional a partir de la mostra. En segon lloc, es generalitza el problema anterior i es considera que tant el valor esperat com la variància poblacional són desconeguts i, per tant, serà necessari estimar els dos paràmetres per obtenir l'interval desitjat.

4.6.1.1 VARIÀNCIA POBLACIONAL CONEGUDA

Donada una població estadística modelitzada per $X \sim N(\mu,\sigma)$ on μ és desconeguda i σ és coneguda, es vol determinar l'interval de confiança per al paràmetre μ.

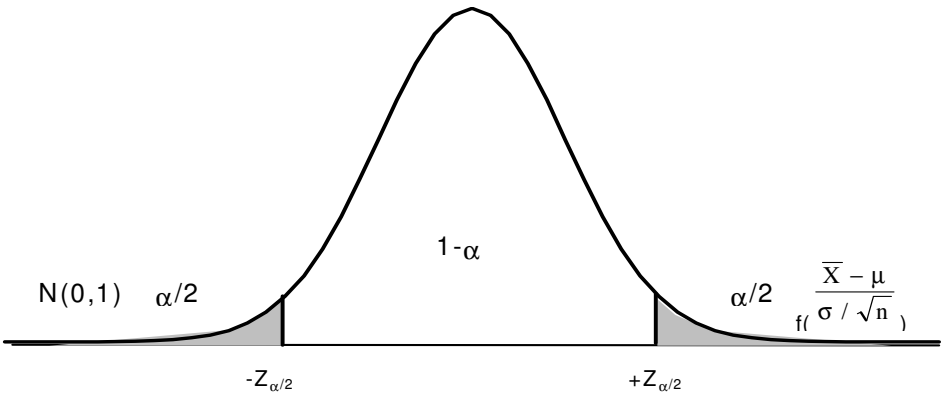
Sabem que el millor estimador del paràmetre μ és $\overline{X}$, estimador que presenta una distribució de probabilitat $N(\mu,\sigma/\sqrt{n})$ i, en estandarditzar, queda

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Fixat el nivell de confiança d'1-α podem determinar els valors crítics $-z_{\alpha/2}$ i $z_{\alpha/2}$ que verifiquen:

$$P(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1-\alpha$$

Aquests valors crítics, com es veu al gràfic següent, determinen el menor interval de la distribució N(0,1) que presenta una probabilitat igual a 1-α. Qualsevol altra combinació de valors z_a i z_b pels quals $P(z_a<Z<z_b)=1-\alpha$ determinen intervals de major amplitud.



Per exemple, si 1-α=0,95 trobem a les taules de la N(0,1) el valor $z_{\alpha/2}$= 1,96 que compleix:

$P(-1{,}96 \leq \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \leq 1{,}96) = 0{,}95$

Per aïllar el paràmetre μ de l'expressió anterior multipliquem els tres membres de la desigualtat per $\sigma/\sqrt{n}$:

$P(-1{,}96 \frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq 1{,}96 \frac{\sigma}{\sqrt{n}}) = 0{,}95$

I, finalment, si restem $\overline{X}$ i canviem el sentit de les desigualtats tenim:

$P(\overline{X} - 1{,}96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1{,}96 \frac{\sigma}{\sqrt{n}}) = 0{,}95$

En general, per a qualsevol nivell de confiança, l'interval és:

$$P(\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

això significa que la probabilitat que l'interval aleatori $[\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$ contingui el veritable paràmetre poblacional μ és $1-\alpha$.

Característiques:

1. L'interval sempre està centrat en la mitjana mostral $\overline{X}$.
2. L'amplitud de l'interval queda determinada per $2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. Per tant, depèn:

- En sentit directe de σ. A major dispersió poblacional major amplitud (menys precisa és l'estimació).
- En sentit invers de **n**. A mesura que augmenta la grandària mostral, l'amplitud de l'interval va disminuint (augmenta la precisió).
- En sentit directe de $\mathbf{z_{\alpha/2}}$ o $\mathbf{1-\alpha}$. Com més gran sigui el nivell de confiança que fixem major serà l'amplitud.

3. L'error màxim d'estimació o precisió de l'estimació és $d = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.
4. Per als nivells de confiança més usuals els valors crítics, $z_{\alpha/2}$, són:

- $1-\alpha = 0{,}90 \quad z_{\alpha/2} = 1{,}645$
- $1-\alpha = 0{,}95 \quad z_{\alpha/2} = 1{,}96$
- $1-\alpha = 0{,}99 \quad z_{\alpha/2} = 2{,}575$

5. El procediment anterior per determinar l'interval de confiança també és vàlid quan la població no és normal però la grandària de la mostra és superior a 30.

Exemple 4.8

Una població estadística es pot modelitzar per la variable aleatòria X que recull 'el pes dels paquets procedents de la màquina A'. Se sap que aquesta població és normal amb valor esperat desconegut i desviació estàndard igual a 200 g. S'extreu una mostra aleatòria de grandària 25 i s'obté $\bar{x}$ *=1050 g.*

a) *Quina serà l'estimació per interval de* μ *si es fixa el nivell de confiança al 90%?*

b) *Si s'augmenta el nivell de confiança al 99% com queda modificada l'estimació anterior?*

Solució:

Dades: X~N(μ,σ=200g) n=25 $\bar{x}$=1050

Estadístic: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$~N(0,1)

a) 1-α= 0,90 $\Rightarrow$ $z_{\alpha/2}$ =1,645

$P(\bar{X}-1{,}645\frac{200}{\sqrt{n}} \leq \mu \leq \bar{X}+1{,}645\frac{200}{\sqrt{n}})= 0{,}90$

Fixada la mostra s'obté l'interval:

[1050-1,645·40 $\leq\mu\leq$ 1050+1,645·40] $\Rightarrow$ $I_{\mu}^{0,90}$ = [984,2; 1115,8]

Resultat que indica que el valor de μ serà superior a 984,2 g. i inferior a 1115,8 g. amb una confiança del 90%.

b) 1-α= 0,99 $\Rightarrow$ $z_{\alpha/2}$ =2,575

$P(\bar{X}-2{,}575\frac{200}{\sqrt{n}} \leq \mu \leq \bar{X}+2{,}575\frac{200}{\sqrt{n}})= 0{,}99$

Fixada la mostra s'obté l'interval:

[1050-2,575·40 $\leq\mu\leq$ 1050+2,575·40] $\Rightarrow$ $I_{\mu}^{0,99}$ = [947; 1153]

L'amplitud de l'interval ha augmentat; s'ha passat de 131,6 (amplitud del $I_{\mu}^{0,90}$) a 206 (amplitud del $I_{\mu}^{0,99}$), per tant, en augmentar la confiança ha disminuït la precisió de l'interval.

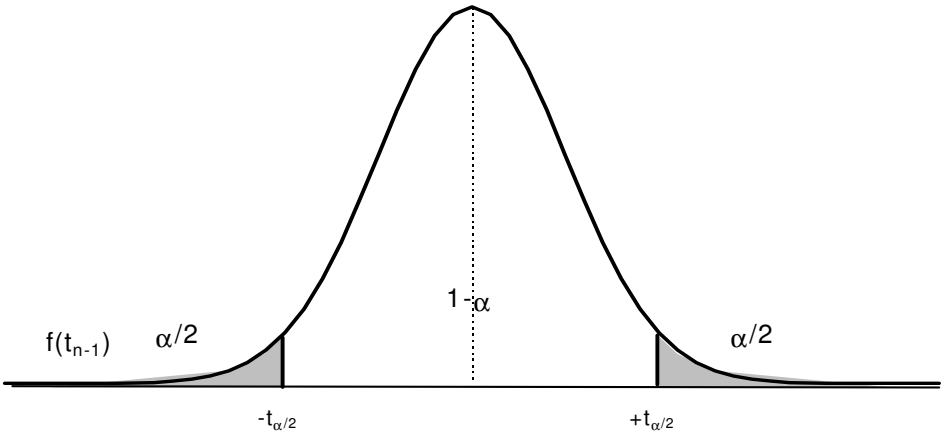
4.6.1.2 VARIÀNCIA POBLACIONAL DESCONEGUDA

A la pràctica si es desconeix l'esperança matemàtica d'una població difícilment es coneixerà la seva variància. Per tant, el cas més freqüent serà que tant μ com σ^2 siguin desconegudes.

Com ja hem vist al capítol anterior, donada una mostra aleatòria $X_1,X_2,...,X_n$ d'una població Normal, l'estadístic $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ es distribueix segons una t de Student amb n-1 graus de llibertat.

Per determinar l'interval de confiança per a μ, fixem el nivell de probabilitat, 1-α, i trobem els valors crítics, $-t_{\alpha/2,n-1}$ i $t_{\alpha/2,n-1}$, a les taules de la distribució t de Student que, com es veu al gràfic següent, verifiquen que:

$$P(-t_{\alpha/2,n-1} \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq +t_{\alpha/2,n-1})=1-\alpha$$



En aïllar μ, l'interval de confiança queda:

$$P(\overline{X}-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X}+t_{\alpha/2,n-1}\frac{S}{\sqrt{n}})= 1-\alpha$$

Fixada una mostra s'obté l'interval $[\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}]$ que amb una confiança d'1-α s'espera que contingui el veritable paràmetre poblacional μ.

Per exemple, el valor crític, per a una mostra de grandària 15 i un nivell de confiança igual a 0,95, que trobem a les taules t de Student és $t_{\alpha/2,n-1}= t_{0,025,14}=2,145$ i l'interval és:

$$P(\overline{X}-2,145\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X}+2,145\frac{S}{\sqrt{n}})= 0,95$$

Característiques:

1. L'interval està centrat en $\overline{X}$.

2. La precisió de l'estimació per interval és $d=t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$.

3. A mesura que augmenta la grandària mostral els valors crítics $t_{\alpha/2,n-1}$ tendeixen als valors $z_{\alpha/2}$ de la distribució N(0,1). Per tant, a la pràctica, per a mostres de grandària superior a 30 podem determinar l'interval de confiança aproximant la distribució t de Student per la normal estandarditzada.

Exemple 4.9

Les vendes diàries d'un comerç es consideren distribuïdes normalment. Amb la finalitat d'estimar la mitjana poblacional, s'extreu una mostra de 10 observacions i s'obté una mitjana mostral de 204.000 u.m. i una desviació estàndard mostral de 90.000 u. m.

Determineu, amb un 1% de significació, un interval de confiança per a les vendes mitjanes diàries del comerç.

Solució:

Dades: n= 10 $\overline{X}$ = 204.000 S = 90.000 α = 0,01

Estadístic: $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ $\qquad \mu \in \left[\overline{X} \pm t_{\alpha/2}\, S/\sqrt{n}\right]$ amb α = 0,01

$\mu \in \left[204.000 \pm 3{,}25 \cdot 90.000/\sqrt{10}\right]$ α = 0,01

Interval: $I_{\mu}^{0,99}$ =[111.503; 296.497]

Podem concloure que amb una confiança del 99% les vendes mitjanes d'aquest comerç es troben entre 111.503 i 296.497 u.m.

4.6.2 OBTENCIÓ D'UN INTERVAL DE CONFIANÇA PER A μ_1- μ_2

En l'àmbit de l'anàlisi estadística ens trobem freqüentment amb situacions que requereixen comparar les mitjanes de dues poblacions. Per exemple, quan un fabricant està interessat en comparar la producció de dues màquines alternatives i estimar la diferència que hi ha entre les seves produccions mitjanes; quan un col·legi professional està interessat en esbrinar si hi ha diferència entre els salaris mitjans dels col·legiats que treballen en dues ciutats diferents, etc.

Donades dues poblacions Normals $X_1 \sim N(\mu_1,\sigma_1)$ i $X_2 \sim N(\mu_2,\sigma_2)$, s'extreuen dues mostres aleatòries independents entre si de grandàries n_1 i n_2 .

D'aquestes mostres s'obté $\overline{X}_1$ i $\overline{X}_2$, respectivament, que com sabem presenten les següents distribucions:

$$\overline{X}_1 \sim N(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}) \text{ i } \overline{X}_2 \sim N(\mu_2, \frac{\sigma_2}{\sqrt{n_2}})$$

Per tant, la variable aleatòria $\overline{X}_1 - \overline{X}_2$ presenta una distribució Normal de paràmetres:

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1-\mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

A partir d'aquesta distribució podem establir l'interval de confiança per a $\mu_1-\mu_2$ que dependrà del fet que siguin o no conegudes les variàncies poblacionals.

4.6.2.1 VARIÀNCIES POBLACIONALS CONEGUDES

Si les variàncies poblacionals σ_1^2 i σ_2^2 són conegudes aleshores:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Fixat el nivell de confiança d'1-α podem determinar els valors crítics $-z_{\alpha/2}$ i $z_{\alpha/2}$:

$$P(-z_{\alpha/2} \leq \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}) = 1-\alpha$$

Per tant, l'interval de confiança és:

$$P[(\overline{X}_1 - \overline{X}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1-\mu_2) \leq (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}] = 1-\alpha$$

Característiques:

1. L'interval sempre està centrat en $\overline{X}_1 - \overline{X}_2$.
2. La precisió de l'estimació per interval és $d=z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
3. Si les dues poblacions tenen la mateixa variància $\sigma_1^2 = \sigma_2^2 = \sigma^2$ l'interval de confiança és:

$P[(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\,\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le (\mu_1-\mu_2) \le (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\,\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,] = 1-\alpha$

4. En el cas de poblacions no Normals es pot utilitzar el procediment anterior quan les grandàries d'ambdues mostres són superiors a 30.

Exemple 4.10

La despesa mensual en béns de consum de les unitats familiars, integrades per dues persones, a la població A presenta una distribució normal amb desviació estàndard igual a 150 €, mentre que a la població B aquesta despesa és també normal però amb desviació estàndard de 80 €. Per tal d'estimar la diferència que existeix entre les despeses mitjanes de les dues poblacions anteriors s'extreuen dues mostres aleatòries i independents de 100 famílies a la població A i de 36 a la població B. De la primera mostra s'obté una despesa mitjana de 3565 € i de 3240 € en la segona.

En base a aquests resultats, obteniu un interval de confiança al 95% per a la diferència de mitjanes.

Solució:

Dades:

$X_1 \sim N(\mu_1, 150)$ $\quad n_1 = 100 \quad \bar{X}_1 = 3565$

$X_2 \sim N(\mu_2, 80)$ $\quad n_2 = 36 \quad \bar{X}_2 = 3240$

Estadístic: $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1-\mu_2, \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}) \Rightarrow \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

$\mu_1-\mu_2 \in [(3565-3240)-1{,}96\sqrt{\frac{150^2}{100}+\frac{80^2}{36}}\,;\,(3565-3240)+1{,}96\sqrt{\frac{150^2}{100}+\frac{80^2}{36}}\,]$

Interval: $I^{0,95}_{\mu_1-\mu_2} = [285{,}66;\ 364{,}33]$

Per tant, es pot concloure al 95% de confiança que existeixen diferències significatives en la despesa mitjana familiar de les dues poblacions o, dit d'una altra manera, la despesa mitjana familiar a la població A supera en 285 € com a mínim i en 364 € com a màxim la de la població B.

4.6.2.2 VARIÀNCIES POBLACIONALS DESCONEGUDES I IGUALS

A la pràctica si els valors esperats de les poblacions que comparem són desconeguts, també ho seran les seves variàncies i, per tant, per establir un interval per a $\mu_1-\mu_2$ caldrà estimar tant els valors esperats com les variàncies.

En general si estem interessats en comparar els valors esperats de dues poblacions per tal de decidir si es poden considerar iguals, prèviament s'haurà de comprovar, o s'haurà de suposar, que presenten la mateixa variància, és a dir, $\sigma_1^2=\sigma_2^2=\sigma^2$. En aquest cas, una estimació no esbiaixada de σ^2, obtinguda a partir de les dues mostres independents, és:

$$S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

i l'estadístic $\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ presenta una distribució t de Student amb n_1+n_2-2 graus de llibertat.

Per determinar l'interval de confiança es fixa el nivell de probabilitat, $1-\alpha$, i es troben els valors crítics $-t_{\alpha/2,n1+n2-2}$ i $t_{\alpha/2,n1+n2-2}$ a la distribució t de Student:

$$P(-t_{\alpha/2,n1+n2-2}\leq\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\leq t_{\alpha/2,n1+n2-2})=1-\alpha$$

L'interval de confiança és:

$$P[(\bar{X}_1-\bar{X}_2)-t_{\alpha/2,n1+n2-2}\,S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\leq(\mu_1-\mu_2)\leq(\bar{X}_1-\bar{X}_2)+t_{\alpha/2,n1+n2-2}\,S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}]=1-\alpha$$

A mesura que augmenta la grandària de les mostres els valors crítics $t_{\alpha/2,n1+n2-2}$ tendeixen als valors $z_{\alpha/2}$ de la distribució N(0,1). A la pràctica per a mostres de grandària superior a 30 es pot aproximar la distribució t de Student per la Normal estandarditzada.

A més, si les grandàries de les mostres són elevades, l'expressió $S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ es pot aproximar per $\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}$, i l'interval de confiança queda:

$$P[(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}] = 1-\alpha$$

Exemple 4.11

D'una enquesta realitzada a 25 famílies de la ciutat A s'ha obtingut una mitjana de 3565 € de despesa mensual en alimentació i una desviació estàndard de 150 €. En una altra ciutat B la mitjana de despesa mensual obtinguda d'una enquesta realitzada a 12 persones ha estat de 3280 € amb desviació estàndard de 170 €

Si es suposa que les despeses mensuals en alimentació a les dues ciutats són normals amb igual variància i les mostres aleatòries i independents, determineu l'interval de confiança al 95% per a la diferència dels valors esperats.

Solució:

Dades:

$n_1 = 25$ $\bar{X}_1 = 3565$ $S_1 = 150$

$n_2 = 12$ $\bar{X}_2 = 3280$ $S_2 = 170$

Estadístic: $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1-\mu_2, \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}) \Rightarrow \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{\alpha/2,n1+n2-2}$

on l'estimador no esbiaixat de σ^2 és:

$$S=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{24\cdot 150^2+11\cdot 170^2}{35}} = 156{,}5613$$

$$\mu_1-\mu_2 \in \left[(3565-3280) \pm 2{,}042\cdot 156{,}5613\sqrt{\frac{1}{25}+\frac{1}{12}}\right]$$

Interval: $I^{0,95}_{\mu_1-\mu_2} = [172{,}73;\ 397{,}27]$

Com que l'interval no conté el valor zero, es pot concloure al 95% de confiança que la diferència en la despesa mensual observada és significativa.

Exemple 4.12

Dues universitats públiques tenen procediments de matriculació diferents i es vol analitzar quin és el més efectiu. Per comparar el temps mitjà que necessiten els alumnes per acabar el tràmit de matriculació, es van seleccionar a l'atzar

100 alumnes a cada universitat i es van obtenir els següents resultats mostrals (en minuts):

$\overline{X}_1 = 135$ $\overline{X}_2 = 122$ $S_1 = 15$ $S_2 = 10$

Si es suposa que les poblacions són normals, d'igual variància i les mostres són independents, obteniu l'interval de confiança al 99% per a la diferència de valors esperats.

Solució:

Dades:

$n_1 = 100$ $\overline{X}_1 = 135$ $S_1 = 15$

$n_2 = 100$ $\overline{X}_2 = 122$ $S_2 = 10$

Estadístic: $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1-\mu_2, \sigma \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}) \Rightarrow \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim N(0,1)$

A partir dels resultats mostrals:

$$\mu_1 - \mu_2 \in \left[(135-122) \pm 2{,}57 \cdot \sqrt{\frac{15^2}{100}+\frac{10^2}{100}} \right]$$

Interval: $I^{0,99}_{\mu_1-\mu_2} = [8{,}37;17{,}63]$

Per tant, es pot concloure que al 99% de confiança existeixen diferències significatives en el temps de matriculació necessari, i aquesta diferència és d'entre 8,4 i 17,6 minuts més a la primera universitat que a la segona.

4.6.3 OBTENCIÓ D'UN INTERVAL DE CONFIANÇA PER A σ^2

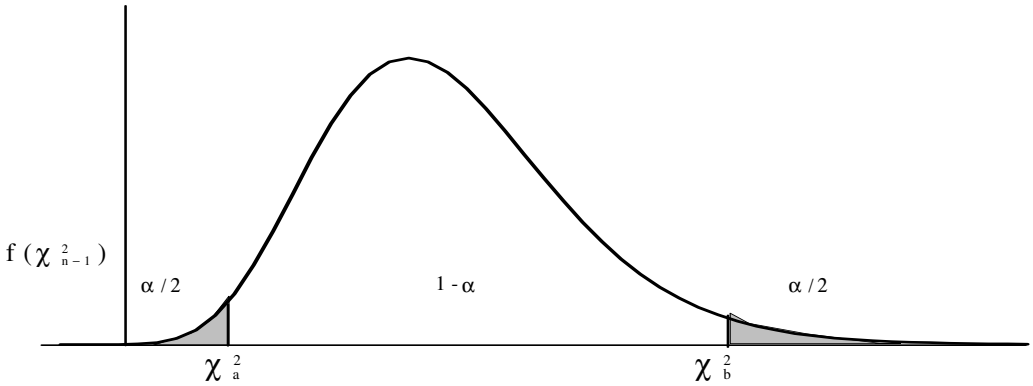
Donada una població estadística caracteritzada per $X \sim N(\mu,\sigma)$, per determinar l'interval de confiança per a la variància poblacional σ^2 partim de l'estimador variància mostral S^2.

Com ja sabem $\frac{(n-1)S^2}{\sigma^2}$ presenta una distribució χ^2 amb n-1 graus de llibertat, i a partir d'aquesta funció podem construir l'interval de confiança per a la variància poblacional.

Per a un nivell de confiança d'1-α trobem a les taules de la funció de distribució de la Khi Quadrat els valors crítics $\chi^2_{a,n-1}$ i $\chi^2_{b,n-1}$ que verifiquen:

$$P(\chi^2_{a,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{b,n-1}) = 1-\alpha$$

Els valors $\chi^2_{a,n-1}$ i $\chi^2_{b,n-1}$ es determinen repartint a parts iguals l'error o nivell de significació α en els extrems de la distribució. És a dir, P($\chi^2 < \chi^2_{a,n-1}$) = $\alpha/2$ i P($\chi^2 > \chi^2_{b,n-1}$) = $\alpha/2$, com es veu al gràfic següent.



Per exemple, per a una mostra de grandària 25 i 1-α=0,95 trobem a les taules de khi quadrat els valors crítics $\chi^2_{a,n-1}$=12,4 i $\chi^2_{b,n-1}$=39,36.

Per tal d'aïllar el paràmetre poblacional σ^2 dividim per $(n-1)S^2$:

$$P\left(\frac{\chi^2_{a,n-1}}{(n-1)S^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{b,n-1}}{(n-1)S^2}\right)=1-\alpha$$

En invertir la desigualtat anterior s'obté l'interval:

$$P\left(\frac{(n-1)S^2}{\chi^2_{b,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{a,n-1}}\right)=1-\alpha$$

que significa que la probabilitat que l'interval aleatori contingui el veritable paràmetre poblacional σ^2 és d'1-α.

Exemple 4.13

D'una mostra triada a l'atzar de 10 ampolles d'oli s'observa que la variància del pes d'aquests envasos és igual a 34 g². Amb un 10% de significació, obteniu un interval de confiança per a la variància poblacional del pes dels envasos d'oli sota el supòsit que aquest segueix una llei normal.

Solució:

Dades: n= 10 S^2 = 34 α = 0,1 graus de llibertat=n-1=9

Estadístic: X~N(μ, σ) $\Rightarrow$ $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$$P\left[\chi_a^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_b^2\right] = 1-\alpha$$

L'interval queda definit:

$$P\left[\frac{(n-1)S^2}{\chi_b^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_a^2}\right] = 1-\alpha$$

Trobem els valors χ_a^2 i χ_b^2 a la taula de la distribució khi quadrat:

$P(\chi^2 \geq \chi_b^2)=0,05 \Rightarrow \chi_b^2=16,92$

$P(\chi^2 \geq \chi_a^2)=0,95 \Rightarrow \chi_a^2=3,33$

Interval: $\left[\frac{34 \cdot 9}{16,92} < \sigma^2 < \frac{34 \cdot 9}{3,33}\right]$

$I_{\sigma^2}^{0,9}$=[18,1; 91,9]

D'on podem concloure que la variància poblacional del pes dels envasos d'oli es troba entre 18 i 92 amb una confiança del 90%.

4.6.4 OBTENCIÓ D'UN INTERVAL DE CONFIANÇA PER A π

Sigui π la proporció poblacional d'elements amb una determinada característica que anomenem 'èxit'. Per exemple: proporció d'individus que estan a favor de la limitació de la publicitat del tabac; proporció de clients que realitzen compres a uns grans magatzems; etc.

Com ja sabem l'estimador màxim-versemblant de π és la proporció mostral p i si n és gran (és a dir, compleix que $n\pi(1-\pi)>5$) la variable aleatòria p presenta una distribució aproximadament Normal de paràmetres:

$$p \sim N\left(\pi; \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Estandarditzant obtenim $\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$.

Aquest resultat ens permet calcular l'interval de confiança per a π.

Fixat el nivell de confiança d'1-α, podem determinar els valors crítics, $-z_{\alpha/2}$ i $z_{\alpha/2}$ (valors que minimitzen l'amplitud de l'interval) a la distribució N(0,1) que compleixen:

$$P(-z_{\alpha/2} \leq \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq z_{\alpha/2}) = 1-\alpha$$

Aïllant π, l'interval queda:

$$P(p-z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq p+ z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}) = 1-\alpha$$

Com veiem aquesta expressió no ens permet determinar els límits de l'interval ja que l'error estàndard és desconegut i, per tant, l'haurem d'estimar aplicant qualsevol de les dues solucions següents:

- Substituir π per l'estimació puntual p.

$$P(p-z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p+ z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}) = 1-\alpha$$

El resultat obtingut és una bona aproximació quan la grandària de la mostra és suficientment elevada.

- Construir l'interval de màxima amplitud, o màxima folgança, que s'obté per a p=0,5, ja que en aquest cas el producte p(1-p) és màxim. Es té doncs que:

$$P(p-z_{\alpha/2}\sqrt{\frac{0{,}5(1-0{,}5)}{n}} \leq \pi \leq p+ z_{\alpha/2}\sqrt{\frac{0{,}5(1-0{,}5)}{n}}) = 1-\alpha$$

En aquest cas l'interval presenta major amplada que el teòric i, per tant, l'estimació és menys precisa.

Característiques:

1. L'interval sempre està centrat en p.
2. L'amplitud de l'interval queda determinada per $2z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ i l'error màxim d'estimació o precisió és d= $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$.

En concret, l'interval de màxima amplitud o variància màxima té associada una precisió igual a:

d=$z_{\alpha/2}\sqrt{\frac{0{,}5(1-0{,}5)}{n}}$.

Exemple 4.14

En una mostra triada de forma aleatòria de 100 sonalls fabricats per una determinada companyia, un 20% no van satisfer les normes de qualitat

establertes. Construïu un interval de confiança del 95% per a la proporció poblacional de sonalls que no satisfan les normes de qualitat.

a) A partir de les dades mostrals.

b) Amb la màxima folgança (amplitud).

Solució:

Dades: p = 0,2 1-α = 0,95 n=100

Estadístic: p ~ $N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

$\pi \in \left[p \pm z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right]$ amb un nivell de significació α.

a) $\pi \in \left[0{,}2 \pm 1{,}96\sqrt{0{,}2 \cdot 0{,}8/100}\right]$

$I_{\pi}^{0,95}$ = [0,1216; 0,2784]

b) $\pi \in \left[0{,}2 \pm 1{,}96\sqrt{0{,}5 \cdot 0{,}5/100}\right]$

$I_{\pi}^{0,95}$ = [0,102; 0,298]

S'estima al 95% de probabilitat que la proporció de sonalls que no satisfan les normes de qualitat es troba entre el 10 i el 30%, aproximadament, amb màxima folgança, o entre el 12 i el 28% en base als resultats mostrals.

4.6.5 OBTENCIÓ D'UN INTERVAL DE CONFIANÇA PER A π_1-π_2

Algunes vegades estem interessats en comparar dues proporcions poblacionals. Per exemple, comparar els percentatges de votants a favor de dos partits polítics dins d'una campanya electoral; comparar la quota de mercat de dos articles substitutius; etc. En aquests casos cal construir l'interval de confiança per a la diferència de proporcions poblacionals a partir de dues mostres aleatòries independents procedents de les seves respectives poblacions.

D'una població, amb una determinada proporció d'èxits π_1 desconeguda, s'extreu una mostra aleatòria de grandària n_1 i s'obté la proporció mostral d'èxits p_1. D'una altra població, amb la proporció poblacional d'èxits π_2, s'extreu una altra mostra aleatòria, independent de l'anterior, de grandària n_2 i s'obté la proporció mostral d'èxits p_2. Per tal d'estimar la diferència entre les

proporcions (π_1-π_2), prendrem com a estimador la diferència de proporcions mostrals (p_1-p_2).

Si n_1 i n_2 són grans, és a dir verifiquen que $n_1\pi_1(1-\pi_1)>5$ i $n_2\pi_2(1-\pi_2)>5$, la variable aleatòria (p_1-p_2) presenta una distribució aproximadament normal de paràmetres:

$$p_1 - p_2 \sim N(\pi_1-\pi_2; \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}})$$

En estandarditzar obtenim l'estadístic $Z=\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$.

A partir del qual podem calcular l'interval de confiança per a π_1-π_2.

Fixat el nivell de confiança d'1-α, podem determinar els valors crítics $-z_{\alpha/2}$ i $z_{\alpha/2}$ a la distribució N(0,1) que verifiquen:

$$P(-z_{\alpha/2} \leq \frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \leq z_{\alpha/2})=1-\alpha$$

L'interval de confiança a l'1-α és:

$$[p_1-p_2-z_{\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}} \leq \pi_1-\pi_2 \leq p_1-p_2+z_{\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}]$$

Com en l'epígraf anterior aquesta expressió no ens permet determinar els límits de l'interval ja que l'error estàndard és desconegut. Si aquest s'estima a partir de les proporcions mostrals la distribució de l'estadístic es pot considerar aproximadament Normal quan les grandàries mostrals són elevades. Llavors l'interval queda definit com:

$$P(p_1 - p_2 - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \leq \pi_1-\pi_2 \leq p_1 - p_2 + z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}) = 1-\alpha$$

Exemple 4.15

D'entre els pacients d'una malaltia se seleccionen de forma independent dos grups de 200 i 100 individus, respectivament. En el primer grup es tracta la malaltia amb medicaments, i s'obté que 25 d'ells experimenten una milloria

immediata. Als pacients del segon grup no se'ls subministra cap medicament i 10 d'ells també milloren. Partint dels resultats anteriors, es pot afirmar amb un 95% de confiança que el tractament és efectiu?

Solució:

Dades: Grup A: $n_1=200$ $p_1=25/200=0,125$

on $n_1 \cdot p_1 \cdot (1-p_1) = 200 \cdot 0,125 \cdot 0,875 = 21,8 > 5$

Grup B: $n_2=100$ $p_2=10/100=0,1$

on $n_2 \cdot p_2 \cdot (1-p_2) = 100 \cdot 0,1 \cdot 0,9 = 9 > 5$

Estadístic:

$$p_1 - p_2 \sim N\left(\pi_1-\pi_2, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

$$\pi_1-\pi_2 \in \left[(0,125-0,1) \pm 1,96 \sqrt{\frac{0,125 \cdot 0,875}{200}+\frac{0,1 \cdot 0,9}{100}}\right]$$

Interval: $\pi_1-\pi_2 \in [0,025 \pm 1,96 \cdot 0,038]$

$$I^{0,95}_{\pi 1-\pi 2} = [-0,05 \; ; \; 0,1]$$

Es pot concloure que no hi ha diferència significativa entre els resultats, ja que el zero es troba dins l'interval calculat i, per tant, no es pot afirmar que el tractament sigui efectiu.

4.7 DETERMINACIÓ DE LA GRANDÀRIA DE LA MOSTRA

Com ja hem dit anteriorment, quan estimem qualsevol paràmetre poblacional, tot i que disposem d'un estimador amb totes les propietats desitjables, generalment l'estimació difereix del veritable valor del paràmetre, és a dir, es produeix un error d'estimació. Aquest error és desconegut, ja que el valor del paràmetre és desconegut. Per aquesta raó construïm intervals de confiança, que ens donen uns límits que es confia continguin el valor del paràmetre amb una probabilitat tan gran com fixem.

Una altra solució al problema de l'error d'estimació rau en determinar prèviament quin és l'error màxim tolerable amb una probabilitat prefixada tan gran com es vulgui i, en funció d'aquests requeriments, determinar quina ha de ser la grandària de la mostra.

4.7.1 ESTIMACIÓ DE μ

Suposem que es tracta d'estimar '*el salari mitjà dels treballadors d'una determinada indústria*' i volem garantir un error màxim d'estimació d=100 € amb probabilitat 0,95. A tal efecte triem una mostra amb la grandària necessària, **n**, i prenem com a estimació la mitjana de la mostra.

La determinació de **n**, la grandària mínima necessària per garantir un error màxim **d** amb una probabilitat d'1-α, depèn del compliment de certes condicions referents a la distribució de la població i del fet que sigui o no coneguda la variància poblacional.

4.7.1.1 VARIÀNCIA POBLACIONAL CONEGUDA

En primer lloc considerem una població modelitzada per una variable Normal X~N(μ, σ) amb σ^2 coneguda, i sigui **n** la grandària mostral que desitgem determinar. Per a aquesta grandària sabem que $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

Garantir un error màxim **d** amb probabilitat 1-α vol dir que la probabilitat que el valor de $\overline{X}$ sigui com a mínim μ-d i com a màxim μ+d és 1-α. És a dir:

$$P(\mu - d \leq \overline{X} \leq \mu + d) = 1 - \alpha$$

En tipificar:

$$P\left(\frac{\mu - d - \mu}{\sigma/\sqrt{n}} \leq z \leq \frac{\mu + d - \mu}{\sigma/\sqrt{n}}\right) = P\left(\frac{-d}{\sigma/\sqrt{n}} \leq z \leq \frac{d}{\sigma/\sqrt{n}}\right) = P\left(\frac{-d\sqrt{n}}{\sigma} \leq z \leq \frac{d\sqrt{n}}{\sigma}\right) = 1 - \alpha$$

Així doncs, el quocient $\frac{d\sqrt{n}}{\sigma}$ és igual al valor $z_{\alpha/2}$ de la distribució N(0,1), que acumula una probabilitat $F(z_{\alpha/2})$ igual a $1-\alpha+\alpha/2$:

$$\frac{d\sqrt{n}}{\sigma} = z_{\alpha/2}$$

i d'aquesta igualtat es dedueix que:

$$n = \left(\frac{z_{\alpha/2}\sigma}{d}\right)^2$$

Exemple 4.16

La Cambra de Comerç de la ciutat A desitja estimar la despesa mitjana (en u.m) per turista. Es tracta de fixar la grandària mínima de la mostra necessària per garantir un error màxim d'estimació de 5 u.m. amb probabilitat 0,99. Es suposa que la despesa per turista es pot modelitzar per una llei Normal amb desviació estàndard de 120 u.m.

Solució:

Dades: $X \sim N(\mu; 120)$ $d = 5$; $1-\alpha = 0{,}99$; $\alpha = 0{,}01$; $\alpha/2 = 0{,}005$.

S'ha de trobar el valor $z_{\alpha/2}$ de la N(0,1) que acumuli una probabilitat $F(z_{\alpha/2}) = 0{,}99 + 0{,}005 = 0{,}995$.

Aquest valor és $z_{\alpha/2} = 2{,}58$ i

$$n = \left(\frac{z_{\alpha/2}\sigma}{d}\right)^2 = (2{,}58^2 \cdot 120^2)/5^2 = 3834{,}08 \approx 3835$$

Així doncs, caldrà disposar d'una mostra de com a mínim 3835 turistes.

En alguns casos l'investigador sap quina és la grandària màxima de la mostra disponible (per exemple, quan a l'obtenció de la informació mostral només s'hi pot destinar una quantitat limitada del pressupost de l'estudi) i, en aquest cas, interessa determinar quin és l'error màxim que es pot garantir amb una probabilitat prefixada $1-\alpha$ o, alternativament, si es desitja assumir un error màxim d, amb quina probabilitat es pot garantir.

En el primer cas l'error màxim quedarà determinat per $d=\frac{z_{\alpha/2}\sigma}{\sqrt{n}}$ i en el segon cas la probabilitat que es pot garantir és la corresponent al valor $z_{\alpha/2}=\frac{d\sqrt{n}}{\sigma}$.

Exemple 4.17

Seguint amb l'exemple 4.16, suposem que la Cambra de Comerç disposa d'un pressupost que li permet treballar amb una mostra de grandària màxima n=1500.

a) Quin és l'error màxim d si es fixa la probabilitat en 0,99?

b) Es desitja fixar l'error màxim d'estimació d=5 u.m., amb quina probabilitat es pot garantir?

Solució:

a) Dades: σ=120; n = 1500; 1-α = 0,99; α = 0,01; $\alpha/2$ = 0,005, $z_{\alpha/2}$ = 2,58.

De l'expressió que ens dóna la grandària mínima de la mostra es dedueix que:

$$d = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{2{,}58 \cdot 120}{\sqrt{1500}} = 7{,}99 \approx 8$$

Per tant, amb aquesta grandària mostral es pot garantir un error màxim d'estimació de la mitjana poblacional de 8 u.m. amb una probabilitat 0,99.

b) Dades: σ=120; n = 1500, d = 5.

$$z_{\alpha/2} = \frac{d\sqrt{n}}{\sigma} = \frac{5 \cdot \sqrt{1500}}{120} = 1{,}61$$

A les taules de la N(0,1) es troba que F(1,61) = 0,9463 = 1-α + $\alpha/2$ i, per tant, $\alpha/2$ = 0,0537;

α = 0,1074 i 1-α = 0,8926

Així doncs, amb n=1500 es pot garantir un error màxim d'estimació de la mitjana poblacional de 5 u.m. amb probabilitat 0,89.

4.7.1.2 VARIÀNCIA POBLACIONAL DESCONEGUDA

Quan en una població normal la variància poblacional és desconeguda es pot determinar, aproximadament, la grandària mínima de la mostra necessària per estimar la mitjana poblacional amb un error màxim d'estimació **d** amb probabilitat 1-α, si es substitueix en la expressió obtinguda a l'apartat anterior la variància desconeguda per una estimació mostral generada per una mostra pilot.

$$n = \left(\frac{z_{\alpha/2} S}{d} \right)^2$$

Exemple 4.18

Un fabricant de pintures té previst afegir al seu producte un additiu per tal de reduir el temps d'assecat (en hores). Per estimar quina serà la reducció mitjana del temps d'assecat desitja obtenir una mostra que li permeti assegurar amb una probabilitat 0,99 un error màxim d'estimació de 0,10 hores. Es pot acceptar

que la variable X = 'reducció del temps d'assecat' és aproximadament normal, amb paràmetres desconeguts. Per disposar d'una estimació de la variància extreu una mostra pilot de 15 observacions i obté S^2= 0,2. Quina haurà de ser la grandària de la mostra?

Solució:

Dades: S^2= 0,2; d = 0,10; 1-α = 0,99; α = 0,01; $\alpha/2$ = 0,005 i 1-α + $\alpha/2$ = 0,995

Per tant $z_{\alpha/2}$= 2,58 i

$$n= \left(\frac{z_{\alpha/2} S}{d} \right)^2 = \frac{2{,}58^2 \cdot 0{,}2}{0{,}1^2} = 133{,}128 \approx 134$$

Per assolir l'objectiu cal una mostra d'aproximadament 134 elements com a mínim.

4.7.2 ESTIMACIÓ DE π

Sigui π la proporció poblacional d'elements amb una determinada característica que anomenem 'èxit'. Es desitja estimar π amb un error màxim d'estimació **d** amb probabilitat 1-α. L'estimador de π és p (proporció mostral d'èxits). Si es compleixen les condicions sota les quals la distribució binomial es pot aproximar per una llei normal (això implica que n ha de ser gran) p~ $N\left(\pi, \sqrt{\pi(1-\pi)/n}\right)$

Es tracta de determinar **n** per tal que es verifiqui:

$$P(\pi-d \leq p \leq \pi+d) = 1-\alpha$$

En tipificar es té:

$$P\left(\frac{\pi - d - \pi}{\sqrt{\pi(1-\pi)/n}} \leq z \leq \frac{\pi + d - \pi}{\sqrt{\pi(1-\pi)/n}} \right) = P\left(\frac{-d}{\sqrt{\pi(1-\pi)/n}} \leq z \leq \frac{d}{\sqrt{\pi(1-\pi)/n}} \right) = 1-\alpha$$

d'on es dedueix que:

$$P\left(z \leq \frac{d}{\sqrt{\pi(1-\pi)/n}} \right) = P\left(z \leq \frac{d\sqrt{n}}{\sqrt{\pi(1-\pi)}} \right) = (1-\alpha) + \frac{\alpha}{2}$$

Trobat el valor $z_{\alpha/2}$ de la N(0,1) per al qual $P(z < z_{\alpha/2}) = (1-\alpha) + \frac{\alpha}{2}$ tenim que

$$\frac{d\sqrt{n}}{\sqrt{\pi(1-\pi)}} = z_{\alpha/2}$$

i, per tant, $\sqrt{n} = \frac{z_{\alpha/2}\sqrt{\pi(1-\pi)}}{d}$ d'on s'obté $n = \frac{z^2_{\alpha/2}\pi(1-\pi)}{d^2}$.

En aquest cas la grandària mínima de la mostra depèn del paràmetre desconegut π. Aquest problema es pot resoldre amb qualsevol dels dos criteris següents:

a) Utilització d'una estimació prèvia de π

Si es substitueix el paràmetre desconegut per una estimació obtinguda d'una mostra pilot, **p**, o per qualsevol altra estimació provisional, la grandària mínima de la mostra necessària per garantir un error màxim **d** amb probabilitat 1-α és, aproximadament:

$$n = \frac{z^2_{\alpha/2}p(1-p)}{d^2}$$

Exemple 4.19

n dels patrocinadors d'un programa de televisió vol estimar quina proporció dels espectadors segueixen el seu programa amb un error màxim del 5% amb probabilitat 0,95. Se sap que, en altres ocasions, programes de contingut similar emesos a la mateixa franja horària han estat seguits, per terme mitjà, per un 25% dels espectadors. Es tracta de determinar quina és la grandària mínima necessària de la mostra.

Solució:

Dades: d = 0,05; 1-α = 0,95; α = 0,05; $\alpha/2$ = 0,025 i (1-α) + $\alpha/2$ = 0,975

Per tant, $z_{\alpha/2}$ = 1,96 i, si es pren p = 0,25 com a estimació provisional de π, resulta:

$$n = \frac{z^2_{\alpha/2}p(1-p)}{d^2} = \frac{1{,}96^2 \cdot 0{,}25 \cdot 0{,}75}{0{,}05^2} = 288{,}12 \sim 289$$

Aproximadament es necessita una mostra de 289 espectadors.

---

b) Grandària de màxima folgança

Per a valors prefixats de l'error màxim tolerable d i de la probabilitat 1-α, la grandària mínima necessària de la mostra depèn del producte $\pi(1-\pi)$. Tot i que π és desconegut l'esmentat producte és màxim quan π és igual a 0,5. Així doncs, si substituïm en l'expressió abans trobada per determinar n el valor de π per 0,5, obtindrem un valor de n que garanteix amb escreix els requeriments exigits.

$$n = \frac{z_{\alpha/2}^2 \cdot 0{,}5 \cdot 0{,}5}{d^2}$$

La grandària **n** així determinada es denomina de màxima folgança.

Exemple 4.20

Es vol examinar un cert nombre d'unitats produïdes en un taller amb la finalitat d'estimar la proporció poblacional d'unitats defectuoses, de manera que l'error de l'estimació sigui com a màxim de 0,01 amb una probabilitat del 80%. Determineu la grandària de la mostra.

Solució:

Dades: Per a un interval de màxima folgança estimem que p = 0,5.

Es fixa un error d'estimació d = 0,01 i un nivell de confiança 1-α = 0,8, per tant, $z_{\alpha/2}$ = 1,28.

$$n = \frac{z_{\alpha/2}^2 \cdot 0{,}5 \cdot 0{,}5}{d^2} = \frac{1{,}28^2 \cdot 0{,}5^2}{0{,}01^2} = 4096$$

4.8 EXERCICIS PROPOSATS

Exercici 1. Es vol estimar la mitjana d'una població, μ, a partir d'una mostra de 4 observacions i es proposen els estimadors següents:

$\hat{\mu}_1 = 1/6 (X_1 + X_2) + 1/3 (X_3 + X_4)$

$\hat{\mu}_2 = 1/5 (X_1 + 2X_2 + 3X_3 + 4X_4)$

$\hat{\mu}_3 = 1/2 (X_1 + X_4)$

a) Trobeu l'esperança matemàtica i la variància dels estimadors proposats. Quins d'aquests són no esbiaixats?
b) Discutiu l'eficiència dels estimadors proposats.
c) Comproveu que el més eficient dels estimadors proposats és menys eficient que l'estimador mitjana mostral.

Exercici 2. Indiqueu quin dels següents estimadors del valor esperat d'una població N(2, 1) minimitza l'EQM i quin és preferible en termes d'eficiència, si es disposa d'una mostra aleatòria de grandària 3.

a) $\hat{\mu}_1 = 0,6X_1 + 0,4X_2$
b) $\hat{\mu}_2 = (4X_1 + 5X_2 + X_3) / 10$
c) $\hat{\mu}_3 = (0,5X_1 + 0,5X_2 + 0,5X_3)/2$

Exercici 3. Donats els següents estimadors de μ d'una població N(0,25; 0,9), comproveu les seves propietats asimptòtiques.

a) $\hat{\mu}_1 = \overline{X} - \frac{0,5}{n}$
b) $\hat{\mu}_2 = 0,3X_1 + X_2 + X_3 + 0,4X_4 - 0,5(X_1+X_4)$
c) $\hat{\mu}_3 = \frac{3}{5}\overline{X} + 0,5$

Exercici 4. La demanda d'un determinat producte és una variable aleatòria amb la funció de densitat següent:

$$f(x) = \begin{cases} \frac{2}{9}x & 0 \leq x \leq 3 \\ 0 & en\ altres\ casos \end{cases}$$

Determineu l'error d'estimació obtingut de la mostra: 1,7; 2,2; 1,8; 2,5; 3; 1,3; 1,6; 2; 2,1; 1,6; si s'estima el valor esperat de X pel mètode dels moments.

Exercici 5. El pes (en gr.) de les peces obtingudes amb un determinat procés de fabricació és una variable aleatòria uniforme definida a l'interval [a;10]. Per tal d'estimar el paràmetre **a** amb una mostra aleatòria de n observacions es proposen dos estimadors: l'obtingut pel mètode dels moments i l'estimador alternatiu $\tilde{a}$ =X_1 + X_n - 10.

Compareu aquests estimadors i indiqueu les propietats que compleixen cadascun d'ells.

Exercici 6. El comportament probabilístic d'una determinada magnitud econòmica s'adapta a la funció següent:

$$f(x) = \begin{cases} \theta^4/4x^5 & x < \theta/2 \\ 0 & en\ altres\ casos \end{cases}$$

Mitjançant el mètode dels moments, obteniu un estimador per al paràmetre desconegut θ i discutiu-ne les propietats asimptòtiques.

Exercici 7. Una variable aleatòria X es distribueix amb la funció de densitat següent:

$$f(x) = \begin{cases} \theta x(1-x)^{\theta-1} & x < 1 \quad \theta > 1 \\ 0 & en\ altres\ casos \end{cases}$$

a) Obteniu l'estimador de θ pel mètode de la màxima versemblança.
b) Trobeu l'estimació de θ a partir de la següent mostra de 5 observacions de la variable X: 0,10; 0,15; 0,22; 0,12; 0,15.

Exercici 8. La variable aleatòria X definida com '*el temps (mesos) que triguen els alumnes de Turisme en trobar el primer lloc de treball*' presenta una distribució exponencial Exp(λ). Obteniu l'estimació màxim versemblant del paràmetre λ a partir de la mostra: 2, 5, 6, 5, 8, 4, 7, 2, 7 i 3 mesos.

Exercici 9. La distribució de probabilitat d'una variable aleatòria X és:

$$f(x) = \begin{cases} k(\theta - x)/2 & 0 < x < \theta \\ 0 & en\ altres\ casos \end{cases}$$

a) Obteniu pel mètode dels moments un estimador del paràmetre θ.
b) Si d'una mostra de grandària 200 s'ha obtingut $\overline{X}$=512, quina és l'estimació puntual per al paràmetre θ?
c) Si posteriorment us diguessin que el vertader valor de θ és 1000, podríeu creure-us-ho? Per què?

Exercici 10. Una companyia asseguradora està convençuda que el nombre de sinistres setmanals d'un determinat tipus s'ajusta a una llei de Poisson. Si en una mostra de 7 setmanes triades a l'atzar es van produir 59 sinistres, quina és l'estimació màxim versemblant del paràmetre λ del model?

Exercici 11. Donada la funció de densitat f(x) = $\frac{5x}{k\beta}$ per a 0<x<β i f(x)=0 en altres casos. Trobeu l'estimador de β pel mètode dels moments.

Exercici 12. Un agent de vendes d'una editorial té un contracte laboral en el qual s'estipula que els seus ingressos són de 15000 u.m. fixes al mes, més una comissió igual a un percentatge de les vendes que realitzi. Aquest agent vol comprovar si la seva feina el compensa econòmicament, per la qual cosa controla els ingressos obtinguts durant 10 mesos triats a l'atzar i n'obté els següents resultats:

26500 ; 22300 ; 30560 ; 25340 ; 24850 ; 22330 ; 26420 ; 28650 ; 29250; 23800

Si la comissió mensual és una variable normal, es demana:

a) Mitjana i desviació típica de les comissions mostrals.
b) Interval de confiança al 95% de la comissió mensual esperada.
c) Interval de confiança al 95% de la variància poblacional de les comissions.

Exercici 13. Una agència de lloguer d'automòbils necessita estimar el nombre de quilòmetres diaris que recorre el seu parc mòbil. Amb aquesta finalitat, durant diversos dies de la setmana, estudia els recorreguts de 71 dels seus vehicles i n'obté una mitjana de 165 km/dia i una desviació estàndard de 5 km/dia. Sota la hipòtesi de normalitat de la característica en estudi *(nombre de quilòmetres per dia):*

a) Construïu un interval per a la mitjana poblacional a un nivell de confiança del 90%.
b) Construïu un interval de confiança del 95% per a la variància poblacional.
c) Si es suposa que la variància poblacional és coneguda i igual a 25, quina és la grandària mostral necessària per reduir a la meitat l'amplitud de l'interval obtingut en l'apartat a)?

Exercici 14. Una empresa que es dedica a la fabricació de tubs d'assaig per a un laboratori posa en mans d'un gabinet tècnic el control de la qualitat del seu producte. Amb aquesta finalitat, s'extreu una mostra aleatòria de grandària 200 i es comprova que un 5% de tubs surten defectuosos.

a) Obteniu l'interval de confiança a un nivell del 90% de la proporció poblacional de tubs defectuosos.
b) Si es treballa amb màxima folgança, quina grandària **n** hauria de tenir la mostra per poder dir que la precisió de l'estimació és del 2% amb un nivell de confiança del 95%?

Exercici 15. Un estadístic vol estimar la mitjana del salari per hora (en €) que reben els treballadors d'una determinada ocupació mitjançant un interval de confiança al 95% d'amplitud 10. Segons altres estudis semblants, està disposat a suposar una distribució normal amb variància 650. Quina grandària de mostra haurà d'utilitzar?

Exercici 16. Es creu que el preu per terme mitjà d'un producte de la marca A és superior al d'una altra marca B. Sota el supòsit que els preus d'ambdues marques són normals amb igual variància, determineu l'interval de confiança al 95% per a la diferència dels preus mitjans a partir dels següents resultats mostrals:

n_A=15 $\overline{X}_A$=254,5 S_A^2= 120 n_B=10 $\overline{X}_B$=216 S_B^2= 135

Es pot acceptar que el preu mitjà de A és superior al preu mitjà de B?

Exercici 17. Es creu que els fumadors presenten més problemes respiratoris que els no fumadors. D'una enquesta realitzada de forma independent a 150 fumadors i 200 no fumadors es va observar que havien patit problemes respiratoris 65 dels fumadors i 60 dels no fumadors. Calculeu l'interval de confiança al 90% per a la diferència de proporcions poblacionals, i indiqueu a quina conclusió arribaríeu.

Exercici 18. D'una mostra aleatòria de 50 proveïdors d'un determinat establiment comercial ha resultat que el 48% ha facturat un import mensual inferior a les 1000 €.

a) Calculeu l'interval al 95% de confiança per a la proporció de proveïdors que facturen per més de 1000 €.
b) Si s'estima, amb un 90% de confiança, la proporció anterior a partir d'una mostra aleatòria de 269 proveïdors, quin serà l'error màxim comès? (Tingueu en compte el resultat de la mostra anterior.)

Exercici 19. Una empresa de serveis a domicili vol comprovar l'efectivitat del correu comercial. Per això comptabilitza el nombre de serveis contractats

durant 65 dies, triats aleatòriament, abans de realitzar la campanya comercial i el nombre de serveis contractats durant 75 dies, triats aleatòriament, després de la campanya i obté els següents resultats:

Nombre de serveis abans de la campanya: $\sum_{i=1}^{65} X_i = 715$ $\sum_{i=1}^{65} X_i^2 = 66365$

Nombre de serveis després de la campanya: $\sum_{i=1}^{75} X_i = 1275$ $\sum_{i=1}^{75} X_i^2 = 80475$

Sota el supòsit que el nombre de serveis contractats abans i després de la campanya presenten distribucions normals amb igual variància, determineu l'interval al 95% de confiança per a la diferència dels valors esperats dels serveis contractats.

Exercici 20. Quina és la grandària de la mostra necessària per estimar la proporció de ciutadans que s'abstindran en les properes eleccions municipals amb una probabilitat del 95% i un error d'estimació màxim de l'1%?

CAPÍTOL V. CONTRAST D'HIPÒTESIS PARAMÈTRIQUES

5.1 INTRODUCCIÓ

Els mètodes de contrast d'hipòtesis tenen com a objectiu comprovar si una determinada afirmació o supòsit respecte a un paràmetre poblacional (o respecte a paràmetres anàlegs de dues o més poblacions) és compatible amb l'evidència empírica que proporciona la informació mostral. Aquests supòsits, que es poden formular respecte a un o més paràmetres, s'anomenen hipòtesis contrastables.

En general, per a qualsevol paràmetre o paràmetres, el contrast es basa en establir un criteri de decisió que depèn en cada cas de la naturalesa de la població, del paràmetre sobre el qual es formula la hipòtesi, de la distribució mostral de l'estimador de l'esmentat paràmetre (o d'una funció de l'estimador) i del control que es desitja fixar *a priori* sobre la probabilitat d'arribar a una decisió errònia.

5.2 ELEMENTS DEL CONTRAST D'HIPÒTESIS

5.2.1 HIPÒTESI NUL·LA I HIPÒTESI ALTERNATIVA

Una hipòtesi estadística és qualsevol supòsit que fa referència a una característica d'una població. Quan aquesta característica és un dels paràmetres poblacionals, la hipòtesi s'anomena *hipòtesi paramètrica*. Així per exemple, si postulem que la proporció d'èxits en una determinada població dicotòmica és del 75%, tenim una hipòtesi paramètrica que expressem com $H:\pi=0,75$ i l'objectiu del contrast és decidir si la hipòtesi que es planteja es pot mantenir com a vàlida. Una hipòtesi com l'esmentada, sobre la que recau la prova del contrast, s'anomena *hipòtesi nul·la*.

***Definició*:**

La ***hipòtesi nul·la*,** que simbolitzem com $H_0:\vartheta=\vartheta_0$ essent ϑ_0 el valor que atribuïm al paràmetre ϑ desconegut, és aquella que recull el supòsit que no hi ha diferència entre el veritable valor del paràmetre i el proposat per la hipòtesi. Aquesta hipòtesi, en principi, es considera certa i només es rebutjarà quan hi hagi suficient evidència empírica en contra. És a dir, la decisió de rebutjar la hipòtesi nul·la estarà en funció que sigui o no compatible amb l'evidència

empírica disponible obtinguda a partir d'una mostra aleatòria procedent de la població.

La proposició contrària a la hipòtesi nul·la rep el nom d'***hipòtesi alternativa*** (H_1) i, generalment, presenta un cert grau d'indefinició. Així, enfront de la hipòtesi nul·la H_0: $\vartheta=\vartheta_0$ es poden proposar diverses hipòtesis alternatives amb un grau d'indefinició que dependrà de la informació *a priori* que tinguem sobre el paràmetre ϑ.

Podem considerar les següents hipòtesis nul·la i alternativa:

- Hipòtesi alternativa d'igualtat:

  H_0: $\vartheta=\vartheta_0$

  H_1: $\vartheta=\vartheta_1$

  En aquest cas, que no és el més freqüent, disposem de molta informació *a priori* sobre ϑ, és a dir, sabem que només pot prendre els valors ϑ_0 o ϑ_1, l'únic dubte que tenim és quin dels dos valors és el veritable.

- Hipòtesi alternativa direccional:

  H_0: $\vartheta=\vartheta_0$

  H_1: $\vartheta>\vartheta_0$ o H_1: $\vartheta<\vartheta_0$

  En aquest cas la informació *a priori* sobre ϑ és menor. L'únic que sabem és que en cas que no sigui cert que el valor de ϑ sigui ϑ_0, el veritable valor de ϑ és superior (o inferior) a ϑ_0. Quan la hipòtesi alternativa és d'aquest tipus el contrast s'anomena "a una cua" o direccional.

- Hipòtesi alternativa no direccional:

  H_0: $\vartheta=\vartheta_0$

  H_1: $\vartheta\neq\vartheta_0$

  En aquest cas la hipòtesi alternativa presenta el màxim grau d'indefinició, la qual cosa implica que no disposem de cap informació *a priori* sobre ϑ, de forma que si no podem acceptar que el veritable valor de ϑ sigui ϑ_0 hem de concloure que ϑ pren qualsevol altre valor. Amb aquesta hipòtesi alternativa el contrast és "a dues cues" o no direccional.

Les hipòtesis que postulen un i només un valor concret per a un paràmetre poblacional són hipòtesis ***simples***. Com veiem, les hipòtesis nul·les les considerem sempre hipòtesis simples. Per contra, les hipòtesis que proposen infinits valors possibles de ϑ, com les hipòtesis alternatives dels dos últims casos, s'anomenen hipòtesis ***compostes***.

5.2.2 ESTADÍSTIC DE PROVA I REGIÓ CRÍTICA

La decisió de rebutjar o no la hipòtesi nul·la depèn de si la informació mostral referent al paràmetre d'interès és o no compatible amb la H_0. L'esmentada informació es resumeix mitjançant l'anomenat ***estadístic de prova*** que, en cada cas, depèn del paràmetre ϑ. Així, per exemple, si desitgem contrastar $H_0:\pi=\pi_0$ contra qualsevol alternativa sintetitzarem la informació mostral relativa a π mitjançant l'estadístic de prova p, proporció d'èxits mostral; si la hipòtesi nul·la fos $H_0:\mu=\mu_0$, l'estadístic de prova seria $\overline{X}$, la mitjana mostral.

***Definició*:**

L'***estadístic de prova*** és un estimador del paràmetre o una funció de l'estimador amb distribució de probabilitat coneguda en el mostratge.

Considerem el cas en què es vol contrastar $H_0:\pi=0,75$ i d'una mostra de grandària n=100 s'obté p=0,73. Aquest valor de l'estadístic de prova sembla compatible, en principi, amb la H_0 formulada, és a dir, si la veritable proporció d'èxits en la població és del 75% no és estrany que una mostra que procedeix d'aquesta població contingui un 73% d'èxits. Per contra, si a la mostra s'obté una proporció d'èxits del 15% considerarem que difereix massa del 75% postulat per la H_0, és a dir, que aquest resultat mostral és incompatible amb l'esmentada hipòtesi nul·la i, en conseqüència, la haurem de rebutjar.

El contrast d'hipòtesis estableix un criteri de decisió referit al valor de l'estadístic de prova. És a dir, determina un rang de valors de l'estadístic que es consideren incompatibles amb la H_0, i, si per a la mostra disponible l'estadístic de prova pren un valor dintre d'aquest rang, la decisió és rebutjar la H_0. Aquest rang o conjunt de valors defineix l'anomenada ***regió crítica*** o regió de rebuig de la hipòtesi nul·la i el valor de l'estadístic de prova que la acota s'anomena ***valor crític***.

Com veurem, per determinar la regió crítica cal establir quina és la distribució de probabilitat de l'estadístic de prova, sota el supòsit que la H_0 és certa, així com el grau de control que volem exercir sobre els possibles valors en la decisió.

5.2.3 ERROR TIPUS I I II. POTÈNCIA DEL CONTRAST

Com hem dit abans, la decisió del contrast pot ser una de les dues següents: o bé no es rebutja la H_0 o bé es rebutja la H_0. Tot i que l'estadístic de prova estigui adequadament ben definit i la regió crítica ben establerta hi ha sempre la possibilitat de prendre una decisió errònia. En aquest sentit es poden cometre

dos tipus d'errors: el primer quan es rebutja una hipòtesi nul·la que és certa i el segon quan no es rebutja una hipòtesi nul·la que és falsa.

El primer error s'anomena ***error de tipus i*** i la seva probabilitat és el ***nivell de significació*** del contrast que es simbolitza com α:

$$\alpha = P(\text{error tipus I}) = P(\text{rebutjar } H_0 / H_0 \text{ certa})$$

El segon error s'anomena ***error de tipus ii*** i la seva probabilitat es simbolitza com β:

$$\beta = P(\text{error tipus II}) = P(\text{no rebutjar } H_0 / H_0 \text{ falsa})$$

La probabilitat de l'error tipus II determina la ***potència*** del contrast, $\eta=1-\beta$, que s'interpreta com la probabilitat que el contrast porti a la decisió correcta:

$$\eta=1-\beta = P(\text{rebutjar } H_0 / H_0 \text{ falsa})$$

Amb hipòtesis alternatives compostes no és possible determinar un únic valor de η, però s'obté una funció de la potència que dependrà dels infinits valors de ϑ continguts a la hipòtesi alternativa.

Exemple 5.1

El cap de vendes d'una empresa sap que l'import de les comissions dels seus venedors es pot modelitzar amb una distribució Normal amb valor esperat de 580 € i desviació estàndard 80.

Per diverses raons, el cap de vendes creu que hi ha hagut un increment de les comissions de forma que el seu import mitjà ha augmentat fins a 600 € tot mantenint-se la desviació estàndard i la normalitat de la població. Per tal de validar la seva creença el cap de vendes té pensat obtenir informació d'una mostra de 100 venedors.

Quin serà el nivell de significació del contrast si el criteri de decisió és rebutjar la H_0:μ=580 davant la H_1:μ=600 quan $\bar{X}$ sigui superior a 590 €? I quina és la potència del contrast?

<u>*Solució:*</u>

En aquest cas es tracta de contrastar una hipòtesi referida a la mitjana poblacional. Se sap que aquest paràmetre només pot prendre dos valors i, per tant, les hipòtesis nul·la i alternativa seran simples:

$$H_0:\mu=580$$

$$H_1:\mu=600$$

L'estadístic de prova és l'estimador de μ, és a dir, la mitjana de la mostra $\overline{X}$. Si la hipòtesi nul·la és certa la distribució de l'estadístic de prova és:

$$\overline{X} \sim N(580; \frac{80}{\sqrt{100}})$$

mentre que si la hipòtesi certa és l'alternativa la distribució de l'estadístic és

$$\overline{X} \sim N(600; \frac{80}{\sqrt{100}})$$

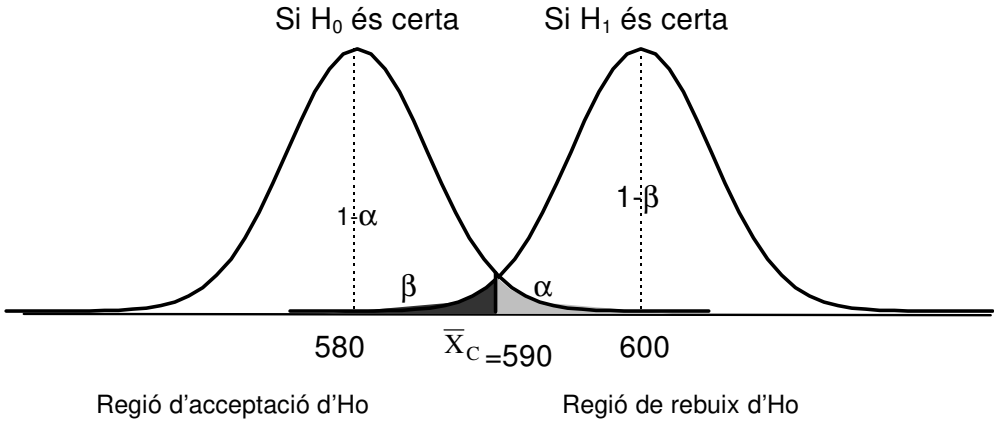
En principi el cap de vendes estableix (arbitràriament) el següent criteri de decisió: si la mitjana de la mostra és superior a 590 € rebutjarà la hipòtesi nul·la en favor de l'alternativa; això vol dir que el valor crític de l'estadístic de prova s'ha fixat en $\overline{X}_C$=590 i la regió crítica o regió de rebuig de la hipòtesi nul·la és la formada per tots els valors possibles de $\overline{X}$ superiors a 590.

En el gràfic següent es veu que, tot i que la H_0 fos certa, la mitjana de la mostra podria ser superior al valor crític i en aquest cas es rebutjaria la H_0 i es cometria un error de tipus I. La probabilitat que això passi correspon a l'àrea situada a la dreta del punt crític sota la distribució de $\overline{X}$ si la H_0 és certa; per tant, el nivell de significació del contrast és:

$$\alpha = P(\text{rebutjar } H_0/H_0 \text{ és certa}) = P[\overline{X} > 590/\overline{X} \sim N(580;8)] = P(z > \frac{590-580}{8}) =$$

$$= P(z > 1{,}25) = 0{,}10565$$

En general el nivell de significació s'indica en percentatge i així direm que en aquest cas el nivell de significació és del 10,56%.



D'altra banda, també és possible que la H_0 no sigui certa i tot i així la mitjana de la mostra sigui inferior a 590; llavors la decisió serà no rebutjar la H_0 i es cometrà, en aquest cas, un error de tipus II amb una probabilitat igual a l'àrea situada a l'esquerra del valor crític sota la distribució de $\overline{X}$ si la H_1 és certa.

Per tant:

$$\beta = P(\text{no rebutjar } H_0/H_0 \text{ és falsa}) = P[\bar{X} < 590/\bar{X} \sim N(600;8)] = P(z < \frac{590-600}{8}) =$$

$$= P(z < -1,25) = 0,10565$$

La potència del contrast, és a dir, la probabilitat d'arribar a la decisió correcta (rebutjar la H_0 essent falsa) és:

$$\eta = 1-\beta = 1-0,10565=0,89435$$

Exemple 5.2

Seguint amb la situació de l'exemple anterior, per tal de reduir el nivell de significació es fixa el criteri de decisió de la següent manera: 'es rebutja la H_0 si la mitjana de la mostra és superior a 595'. Determineu el nou nivell de significació i la potència del contrast.

Solució:

El desplaçament del valor crític cap a la dreta redueix, efectivament, l'àrea sota la distribució de $\bar{X}$ si la H_0 és certa, essent ara:

$$\alpha = P(\text{rebutjar } H_0/H_0 \text{ és certa}) = P[\bar{X} > 595/\bar{X} \sim N(580;8)] = P(z > \frac{595-580}{8}) =$$

$$= P(z > 1,87) = 0,03074$$

Aquesta reducció de la probabilitat d'error tipus I suposa un increment en la probabilitat de l'error tipus II. Efectivament, ara:

$$\beta = P(\text{no rebutjar } H_0/H_0 \text{ és falsa}) = P[\bar{X} < 595/\bar{X} \sim N(600;8)] = P(z < \frac{595-600}{8}) =$$

$$= P(z < -0,625) = 0,26763$$

Aquest nou valor de β, més gran que en el cas anterior, es tradueix en una disminució de la potència del contrast, que en aquest cas queda:

$$\eta = 1- 0,26763 = 0,73237$$

Exemple 5.3

Seguint amb la situació de l'exemple anterior, comproveu com varien el nivell de significació i la potència si fixem la grandària de la mostra en 144 elements.

Solució:

Per tal d'assolir una disminució tant de α com de β cal incrementar la grandària de la mostra.

Amb una grandària de la mostra de 144 elements, la distribució de l'estadístic de prova si la H_0 és certa és

$$\overline{X} \sim N(580; \frac{80}{\sqrt{144}})$$

i la distribució de l'estadístic de prova si la H_1 és certa serà

$$\overline{X} \sim N(600; \frac{80}{\sqrt{144}})$$

Com es pot comprovar, si es fixa el mateix valor crític, $\overline{X}_C$=590, els valors de α i β es redueixen a $\alpha=\beta$=0,06681.

α= P(rebutjar H_0/H_0 és certa) = P[$\overline{X}$ >590/ $\overline{X} \sim N(580;6,67)$] = $P(z>\frac{590-580}{6,67})$ =

= P(z>1,5) = 0,06681

β= P(no rebutjar H_0/H_0 és falsa)=P[$\overline{X}$ <590/ $\overline{X} \sim N(600;6,67)$]=$P(z<\frac{590-600}{6,67})$ =

= P(z<-1,5)=0,06681

5.2.4 VALOR P O NIVELL DE SIGNIFICACIÓ CRÍTIC

En els contrastos d'hipòtesis el valor p (el p-valor, o bé el p-value) estableix la probabilitat d'obtenir una discrepància més gran o igual a la observada a la mostra, si la hipòtesi nul.la fora certa.

***Definició*:**

El ***Valor p*** indica la probabilitat d'obtenir un valor de l'estadístic de prova almenys tan extrem com el que realment s'ha obtingut a la mostra suposant que la hipòtesi nul·la és certa.

El valor p es pot interpretar com la probabilitat d'error si es rebutja la hipòtesi nul·la amb el resultat mostral obtingut. Per tant, quan més petit és aquest valor més evident és que la conclusió adequada serà rebutjar la hipòtesi nul.la perquè la probabilitat d'error serà molt petita.

La determinació del valor p estableix un criteri sobre la decisió de rebutjar o no la hipòtesi nul.la. Es rebutjarà la hipòtesi nul.la si el valor p és igual o inferior al nivell de significació del contrast (normalment α=0,05 o α=0,01).

Per tant, si el valor p és molt petit (pròxim a 0) indica que existeix molt poca probabilitat que el resultat mostral esdevingui d'una població amb H_o certa, és a dir, existeix una evidència molt forta que la H_o no és certa. Per altra banda, un

valor de p gran indica que hi ha evidència suficient per no rebutjar la H_0, és a dir, existeix molt poca probabilitat de que sigui falsa.

Càlcul del valor p

Si es realitza un contrast sobre la mitjana poblacional a una cua:

$H_0: \mu=\mu_0$ o $H_0: \mu=\mu_0$

$H_1: \mu>\mu_0$ $H_1: \mu<\mu_0$

El valor p serà:

Valor-p = P(Z>|z*|) essent z* el valor de l'estadístic de prova provinent d'una determinada mostra.

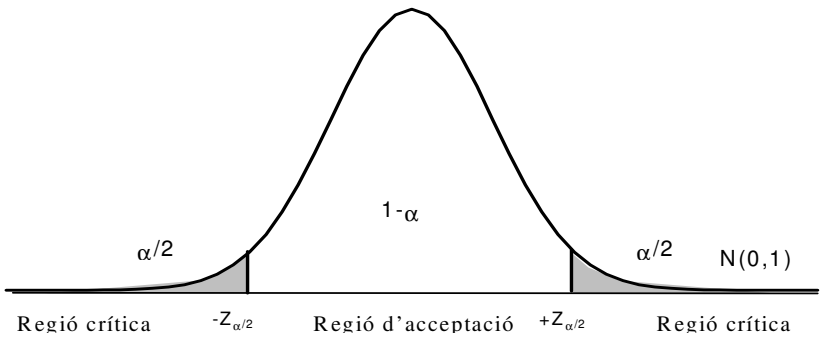
$z^* = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ si σ es coneguda o $z^* = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$ quan n >30

Si el contrast és a dues cues

$H_0: \mu=\mu_0$

$H_1: \mu\neq\mu_0$

El valor p serà: Valor-p = 2 P (Z>| z* |)



La decisió del contrast serà:

- Valor-p < α. Es rebutja la H_0. En aquest cas el Valor-p es suficientment petit i es pot rebutjar la H_0.
- Valor-p > α. No es pot rebutjar la H_0. El Valor-p es massa gran i si es rebutja la H_0 la probabilitat de l'error tipo I seria superior al nivell de significació fixat.

Exemple 5.4

Seguint amb la situació dels exemples anteriors, el cap de vendes de l'empresa sap que l'import de les comissions dels seus venedors es pot modelitzar amb una distribució Normal amb valor esperat de 580 € i desviació estàndard 80.

Per diverses raons, el cap de vendes creu que hi ha hagut un increment de les comissions Per tal de validar la seva creença el cap de vendes a obtingut informació d'una mostra de 100 venedors amb una $\overline{X} = 595$ *€.*

A partir del valor-p, a quina conclusió arribarà el cap de vendes al 5% de significació?

Solució:

En aquest cas es tracta de contrastar una hipòtesi referida a la mitjana poblacional, per tant, les hipòtesis nul·la i alternativa seran simples:

$$H_0: \mu = 580$$

$$H_1: \mu > 580$$

L'estadístic de prova és l'estimador de μ, és a dir, la mitjana de la mostra $\overline{X}$. Si la hipòtesi nul·la és certa la distribució de l'estadístic de prova és:

$$\overline{X} \sim N(580; \frac{80}{\sqrt{100}})$$

Valor-p = $P(z > \frac{595-580}{80/\sqrt{100}}) = P(z > 1{,}87) = 0{,}03074 < 0{,}05$. Per tant, es rebutja H_0 i es pot concloure que ha hagut un increment en les comissions.

5.3 ETAPES DEL CONTRAST

Les etapes del contrast d'hipòtesis tal com es realitzen a la pràctica són les següents:

- Es plantegen les hipòtesis nul·la i alternativa que poden ser:

  $H_0: \vartheta = \vartheta_0$ $H_0: \vartheta = \vartheta_0$ $H_0: \vartheta = \vartheta_0$ $H_0: \vartheta = \vartheta_0$

  $H_1: \vartheta = \vartheta_0$ $H_1: \vartheta > \vartheta_0$ $H_1: \vartheta < \vartheta_0$ $H_1: \vartheta \neq \vartheta_0$

- Es fixa el nivell de significació desitjat. Els valors de α més freqüents són $\alpha=0{,}01$; $\alpha=0{,}05$ i $\alpha=0{,}10$.

- Es tria l'estadístic de prova, que en general és un estimador del paràmetre ϑ o una funció d'un estimador, i s'estableix la seva distribució de probabilitat sota el supòsit que la H_0 és certa.
- Es determina la regió crítica o rang de valors de l'estadístic de prova per als quals es rebutja la H_0.
- S'obté la informació mostral i es troba el valor de l'estadístic de prova.
- Es comprova a quina regió pertany el valor de l'estadístic de prova per tal de prendre una decisió: si el valor de l'estadístic de prova pertany a la regió crítica es rebutja la H_0. En cas contrari no es rebutja, és a dir, es manté provisionalment com a certa mentre no hi hagi informació addicional que la contradigui.

5.4 CONTRAST PER A μ

A partir d'una mostra aleatòria $X_1, X_2, \ldots, X_n$ obtinguda d'una població normal de paràmetres μ i σ desconeguts es vol contrastar la hipòtesi nul·la que la mitjana poblacional és igual a un cert valor μ_0.

La hipòtesi alternativa es pot fixar a doble cua (contrast bilateral), quan es vol comprovar si la mitjana poblacional és diferent al valor postulat a la H_0, o bé a una cua (contrast unilateral), quan es vol comprovar que la mitjana pren un valor superior (contrast a cua superior) o un valor inferior (contrast a cua inferior) a μ_0.

Aquestes hipòtesis les simbolitzarem com:

Doble cua	Cua superior	Cua inferior
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu < \mu_0$

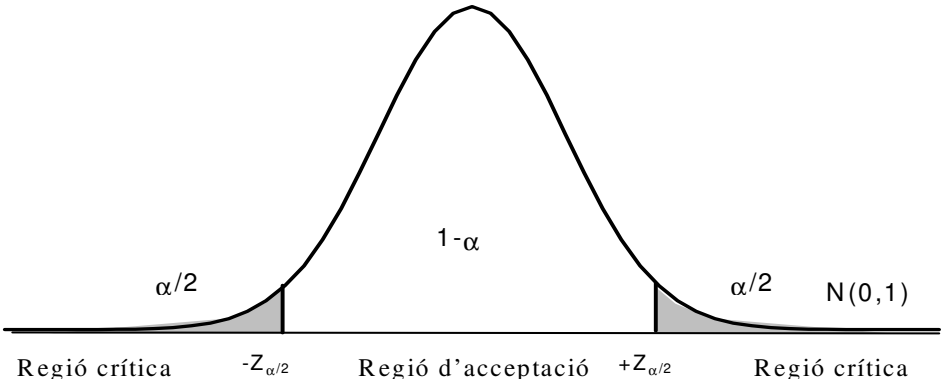
La metodologia a seguir per realitzar aquest contrast depèn de si és o no coneguda la variància poblacional.

5.4.1 VARIÀNCIA POBLACIONAL CONEGUDA

El procediment per establir el criteri de rebuig de la hipòtesi nul·la a partir de la informació mostral es basa en l'estimador òptim de μ, $\bar{X}$. Si la mostra pertany a la població normal postulada a la H_0, aleshores $\bar{X}$ presenta una distribució $N(\mu_0; \sigma/\sqrt{n})$ i, per tant, és molt més probable que $\bar{X}$ i μ_0 presentin valors

propers que valors molt diferents. Pel contrari, si la distribució poblacional no és la formulada a la H_0 segurament s'obtindrà un valor de $\bar{X}$ molt diferent a μ_0. Així, en el contrast a doble cua per establir el criteri de decisió que, amb una determinada probabilitat, ens indiqui quan la diferència, en valor absolut, entre $\bar{X}$ i μ_0 és suficientment gran com per rebutjar la H_0, ens basem en la distribució de $\bar{X}$. Si la hipòtesi nul·la és certa i la variància poblacional és coneguda i igual a σ^2, sabem que la distribució de $\bar{X}$ és $N(\mu_0, \sigma/\sqrt{n})$. Fixada la probabilitat α de rebutjar la hipòtesi nul·la essent certa (nivell de significació), podem trobar els valors crítics $-z_{\alpha/2}$ i $z_{\alpha/2}$ que delimiten la zona d'acceptació.

$$\bar{X} \sim N(\mu_0, \sigma/\sqrt{n}) \Rightarrow Z=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$



- $P(-z_{\alpha/2}<Z< z_{\alpha/2}) = 1-\alpha$ (regió d'acceptació).
- $P(Z<- z_{\alpha/2}) = \alpha/2$ i $P(Z> z_{\alpha/2}) = \alpha/2$ (regió crítica).

Per tant, es rebutjarà la hipòtesi nul·la quan l'estadístic $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ sigui més petit que $-z_{\alpha/2}$ o més gran que $z_{\alpha/2}$, és a dir, si pertany a la regió crítica, i direm que l'evidència empírica no és compatible amb la hipòtesi proposada. Pel contrari, si el valor de l'estadístic de prova pertany a la regió d'acceptació, no es pot rebutjar la hipòtesi nul·la i direm que l'evidència empírica no és suficient per rebutjar que la mitjana poblacional sigui diferent de μ_0.

Hipòtesis:

$H_0: \mu=\mu_0$

$H_1: \mu\neq\mu_0$

Estadístic de prova:

$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Criteri de decisió:

Si $|Z| > z_{\alpha/2}$ es rebutja la H_0

Si la hipòtesi alternativa és a una cua (unilateral), és a dir, si es vol contrastar que la mitjana poblacional és major (cua superior) o menor (cua inferior) al valor postulat a la H_0, la regió crítica per a un nivell de significació α vindrà determinada per un valor z_α que:

- $P(Z>z_\alpha) = \alpha$ (cua superior).
- $P(Z<z_\alpha) = \alpha$ (cua inferior).

Hipòtesis (cua superior):

$H_0: \mu=\mu_0$

$H_1: \mu>\mu_0$

Estadístic de prova:

$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Criteri de decisió:

Si $Z > z_\alpha$ es rebutja la H_0

Hipòtesis (cua inferior):

$$H_0: \mu=\mu_0$$

$$H_1: \mu<\mu_0$$

Estadístic de prova:

$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Criteri de decisió:

Si $Z< -z_\alpha$ es rebutja la H_0

La metodologia del contrast exposat es pot aplicar també a poblacions on els supòsits de normalitat i de variància coneguda no es puguin acceptar, sempre i quan la mostra sigui gran (grandària superior a 30). Així, quan no es compleix el supòsit de normalitat, si n es suficientment gran, sabem pel TCL que $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ presenta una distribució de probabilitat N(0, 1), i en el cas de variància desconeguda sabem que S^2 és un estimador consistent i, per tant, és una bona aproximació de σ^2 quan la grandària de la mostra és elevada.

Exemple 5.5

El fabricant d'un producte sap que fins ara els seus venedors guanyaven per terme mitjà 1000 u.m. mensuals en concepte de comissions, amb una desviació estàndard de 480 u.m. A causa d'un augment de la competència, el fabricant creu que és possible que les comissions s'hagin reduït, però, per altra banda, sospita que les comissions poden haver augmentat a causa de l'increment del preu de venda del seu producte. El fabricant desitja contrastar l'efecte net d'ambdós factors (augment de competència i augment de preu), per això extreu una mostra de 100 observacions que proporcionen unes comissions mitjanes de 925 u.m. Sota el supòsit que la desviació estàndard s'ha mantingut constant:

a) A partir de la informació mostral anterior, a quina conclusió arribarà el fabricant al 5% de significació?

b) A partir de quins valors de la mitjana mostral rebutgem la H_0 si es manté el mateix nivell de significació?

c) Si la sospita del fabricant és que l'efecte net és negatiu (reducció de les comissions) i decideix rebutjar la H_0 per a qualsevol mitjana mostral igual o inferior a 925 u.m., quina és la probabilitat de cometre un error de tipus I?

Solució:

a) Formulació de les hipòtesis:

H_0: $\mu = 1000$

H_1: $\mu \neq 1000$

Selecció de l'estadístic de prova i càlcul del seu valor sota el supòsit que H_0 és certa:

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{925-1000}{480/\sqrt{100}} = -1{,}56$$

Aplicació del criteri de decisió:

Per a α= 0,05, els punts crítics són: $z_{\alpha/2} = \pm 1{,}96$

Com que Z=-1,56 no pertany a la regió crítica es conclou que no hi ha evidència empírica suficient per rebutjar la H_0.

b) Com hem vist el valor crític $z_{\alpha/2}$ és $\pm$ 1,96. Si igualem l'estadístic de prova al valor crític, queda:

$$z_{\alpha/2} = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = \cdots \frac{\overline{X}-1000}{480/\sqrt{100}} = \pm 1{,}96 \Rightarrow \begin{cases} \overline{X} = 1000 - 1{,}96 \cdot 48 = 905{,}92 \\ \overline{X} = 1000 + 1{,}96 \cdot 48 = 1094{,}08 \end{cases}$$

Per tant, els valors de $\overline{X}$ inferiors a 905,92 o superiors a 1094,08 són evidència empírica suficient per rebutjar la H_0 al 5% de nivell de significació.

c) Suposada certa la H_0, $\overline{X} \sim N(1000, 480/\sqrt{100})$, llavors:

α = P(Error tipus I) = P(RH_0 /certa) = P($\overline{X} \leq 925 / \overline{X} \sim N(1000,48)$) =

$$= P\left(Z \leq \frac{925-1000}{48}\right) = P(Z \leq -1{,}56) = 0{,}05938$$

El fabricant ha fixat un nivell de significació del 5,9% o sigui, admet que només en un 5,9% del contrastos realitzats a partir dels resultats obtinguts d'un nombre elevat de mostres (teòricament infinit) rebutjarà la hipòtesi nul·la essent aquesta certa (cometrà l'error tipus I).

5.4.2 VARIÀNCIA POBLACIONAL DESCONEGUDA

En general ens trobarem que quan la mitjana poblacional és desconeguda la variància també ho és. En aquesta situació per poder realitzar el contrast

anterior és necessari, en primer lloc, estimar la variància a partir dels resultats mostrals. Com ja s'ha demostrat l'estimador no esbiaixat de la variància és la variància mostral o $S^2=\frac{\sum (X_i - \overline{X})^2}{n-1}$ i, si la hipòtesi nul·la és certa, l'estadístic de prova $t=\frac{\overline{X} - \mu_0}{S/\sqrt{n}}$ segueix una distribució *t* de Student amb n-1 graus de llibertat.

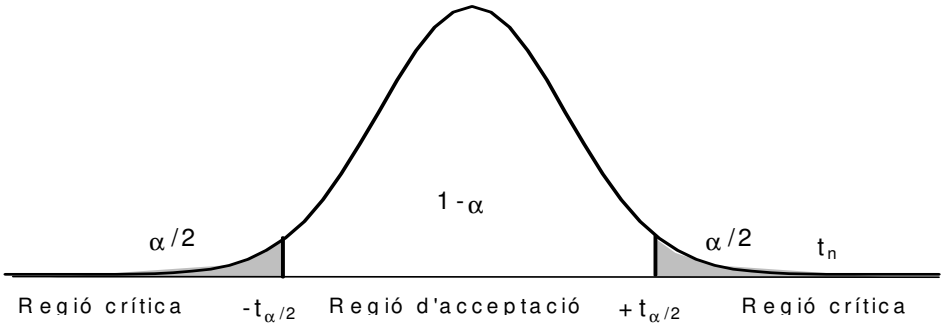
Si el contrast és bilateral:

$$H_0: \mu=\mu_0$$

$$H_1: \mu\neq\mu_0$$

fixada la probabilitat α de rebutjar la hipòtesi nul·la essent certa, podem trobar els valors crítics, que indicarem per $-t_{\alpha/2}$ i $t_{\alpha/2}$, que delimiten la zona d'acceptació a la distribució *t* de Student. Aquests dos valors compleixen que:

- $P(-t_{\alpha/2}<t< t_{\alpha/2}) = 1-\alpha$ (regió d'acceptació).
- $P(t<-t_{\alpha/2}) = \alpha/2$ i $P(t> t_{\alpha/2}) = \alpha/2$ (regió crítica).



Per tant, es rebutjarà la hipòtesi nul·la quan l'estadístic en valor absolut sigui més gran que $t_{\alpha/2}$.

Hipòtesis:

$$H_0: \mu=\mu_0$$

$$H_1: \mu\neq\mu_0$$

Estadístic de prova:

$$t=\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

Criteri de decisió:

Si $|t| > t_{\alpha/2}$ es rebutja la H_0

Si el contrast és unilateral la metodologia a seguir és:

Hipòtesis:

(cua superior)	(cua inferior)
H_0: $\mu=\mu_0$	H_0: $\mu=\mu_0$
H_1: $\mu>\mu_0$	H_1: $\mu<\mu_0$

Estadístic de prova:

$$t=\frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

Criteri de decisió:

Si $t> t_\alpha$ es rebutja la H_0 Si $t < -t_\alpha$ es rebutja la H_0

En el cas que la grandària de la mostra sigui superior a 30, l'estadístic de prova

$$t=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$$

es distribueix aproximadament com una normal estandarditzada i, per tant, els valors crítics són, aproximadament, $-z_{\alpha/2}$ i $z_{\alpha/2}$.

Exemple 5.6

Se sap que el refredat comú es cura sense cap tractament en 8,3 dies de mitjana. Un laboratori farmacèutic ha preparat un nou específic que es creu que serà eficaç per al tractament del refredat. S'agafa una mostra de 7 pacients i se'ls administra el producte. Els símptomes tarden a desaparèixer 8,1 dies de mitjana, amb una desviació estàndard 1,04. Amb un nivell de significació 0,01, i en el cas que la variable temps de curació es distribueixi normalment, quina és la conclusió de l'estudi?

Solució:

Formulació de les hipòtesis:

H_0: $\mu = 8,3$

H_1: $\mu < 8,3$

Selecció de l'estadístic de prova:

$$t = \frac{\overline{X}-\mu_0}{s/\sqrt{n}} \sim t_6 \Rightarrow t=\frac{8,1-8,3}{1,04/\sqrt{7}} = - 0,509$$

Determinació de la regió crítica:

Per a α=0,01, el punt crític és: $t_\alpha = -3,143$

Com que t = -0,509 ∈ [-3,143, +∞] es conclou que no hi ha evidència empírica suficient per rebutjar la H_0.

5.5 CONTRAST PER A σ^2

Donada una població normal de paràmetres μ i σ desconeguts es vol contrastar la hipòtesi nul·la que la variància de la població és igual a un valor específic σ_0^2

H_0: $\sigma^2=\sigma_0^2$

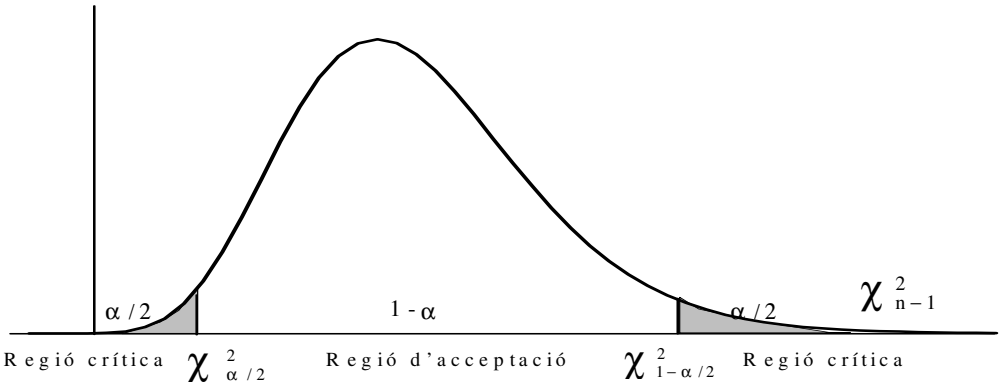
Sota el supòsit que la hipòtesi nul·la és certa la variable aleatòria

$$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}$$

presenta una distribució khi al quadrat amb n-1 graus de llibertat. Si realitzem el contrast a doble cua (H_1: $\sigma^2\neq\sigma_0^2$ bilateral) rebutjarem la H_0 quan χ^2 prengui valors extremadament petits o grans, és a dir, quan sigui significativament diferent a (n-1). Això indica que existeix una diferència significativa entre l'evidència empírica (S^2) i la hipòtesi formulada (σ_0^2).

Fixat un nivell de significació α, el criteri de decisió ve determinat pels valors crítics $\chi^2_{\alpha/2}$ i $\chi^2_{1-\alpha/2}$ que compleixen:

- $P(\chi^2_{\alpha/2}<\chi^2<\chi^2_{1-\alpha/2})=1-\alpha$ (regió d'acceptació).
- $P(\chi^2<\chi^2_{\alpha/2}) = \alpha/2$ i $P(\chi^2>\chi^2_{1-\alpha/2}) = \alpha/2$ (regió crítica).



I, per tant, rebutjarem la hipòtesi nul·la quan $\chi^2<\chi^2_{\alpha/2}$ o $\chi^2>\chi^2_{1-\alpha/2}$,

Hipòtesis:

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

Estadístic de prova:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Criteri de decisió:

Si $\chi^2 < \chi^2_{\alpha/2}$ ó $\chi^2 > \chi^2_{1-\alpha/2}$ es rebutja la H_0

Si el contrast és unilateral la metodologia a seguir és:

Hipòtesis:

(cua superior)	(cua inferior)
$H_0: \sigma^2 = \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 > \sigma_0^2$	$H_1: \sigma^2 < \sigma_0^2$

Estadístic de prova:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Criteri de decisió:

Si $\chi^2 > \chi^2_{1-\alpha}$ es rebutja la H_0 Si $\chi^2 < \chi^2_{\alpha}$ es rebutja la H_0

Exemple 5.7

En un taller funcionen dues màquines, A i B, per a la producció d'unes determinades peces. L'experiència demostra que el pes mitjà de les peces fabricades per A és de 1180 gr., amb una desviació estàndard de 90 gr.. L'encarregat del taller creu que, si bé el pes mitjà de les peces fabricades per B és el mateix que el de les fabricades per A, la màquina B funciona amb una regularitat diferent. Per confirmar-ho tria a l'atzar una mostra de 101 peces fabricades per la màquina B i obté una desviació estàndard de 80 gr.

Proveu la hipòtesi de l'encarregat sota el supòsit de població normal amb α=0,05 i α=0,01.

Solució:

Formulació de les Hipòtesis:

$$H_0: \sigma_B^2 = 90^2$$
$$H_1: \sigma_B^2 \neq 90^2$$

Selecció de l'estadístic de prova:

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma_B^2} = \frac{100 \cdot 80^2}{90^2} = 79{,}01$$

Aplicació del criteri de decisió:

Al 5% de significació i amb 100 g.ll. obtenim a la taula:

$$\chi^2_{\alpha/2} = 74{,}22 \text{ i } \chi^2_{1-\alpha/2} = 129{,}56.$$

Com que $\chi^2 = 79 \in [74{,}22;\ 129{,}56]$ es conclou que no hi ha evidència empírica suficient per rebutjar la H_0.

Amb 100 g.ll. i α=1% obtenim els valor crítics: $\chi^2_{\alpha/2} = 67{,}33$ i $\chi^2_{1-\alpha/2} = 140{,}17$.

Com que $\chi^2 = 79 \in [67{,}33;\ 140{,}17]$ tampoc es pot rebutjar la H_0. Per tant, la dispersió del pes de les peces fabricades per aquestes màquines no són significativament diferent.

5.6 CONTRAST PER A μ_1-μ_2

Suposem que disposem de dues mostres aleatòries i independents de grandàries n_1 i n_2, respectivament, obtingudes de dues poblacions normals de paràmetres μ_1 i σ_1, per a la primera, i μ_2 i σ_2, per a la segona. Es vol contrastar la hipòtesi nul·la que els valors esperats de les dues poblacions són iguals.

- Hipòtesi nul·la:

$$H_0: \mu_1 = \mu_2 \Rightarrow H_0: \mu_1 - \mu_2 = 0$$

La hipòtesi alternativa pot ser no direccional (els valors esperats són diferents) o direccional (el valor esperat d'una població és superior o inferior al de l'altra població).

- Hipòtesi alternativa:

  Contrast a doble cua:

$$H_1: \mu_1 \neq \mu_2 \Rightarrow H_1: \mu_1 - \mu_2 \neq 0$$

  Contrast a cua superior:

$$H_1: \mu_1 > \mu_2 \Rightarrow H_1: \mu_1 - \mu_2 > 0$$

Contrast a cua inferior:

$$H_1: \mu_1<\mu_2 \Rightarrow H_1: \mu_1-\mu_2<0$$

Com ja hem fet a l'apartat 5.4, aquí també considerarem dues situacions:

a) Variàncies poblacionals conegudes.

b) Variàncies poblacionals desconegudes i iguals.

5.6.1 VARIÀNCIES POBLACIONALS CONEGUDES

Per realitzar el contrast de diferència de mitjanes poblacionals ($\mu_1-\mu_2$) ens basem en els resultats mostrals obtinguts a partir de dues mostres aleatòries i independents. Així tindrem $\overline{X}_1$ i $\overline{X}_2$ com les mitjanes de les mostres anteriors de grandària n_1 i n_2, respectivament.

L'estadístic ($\overline{X}_1-\overline{X}_2$), com vam veure en el capítol anterior, es distribueix segons una normal de paràmetres:

$$E(\overline{X}_1-\overline{X}_2)=\mu_1-\mu_2 \text{ i } V(\overline{X}_1-\overline{X}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

I la variable aleatòria estandarditzada

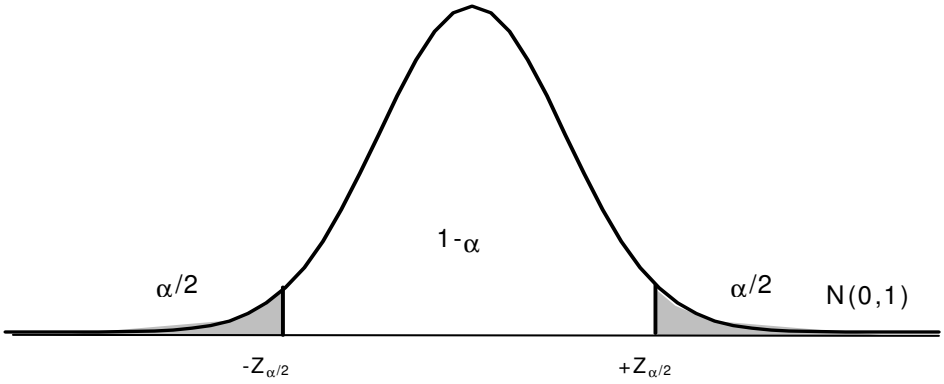
$$Z=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

Si la hipòtesi nul·la s'especifica com H_0: $\mu_1-\mu_2=0$, és a dir, es postula que no hi ha diferències entre les mitjanes poblacionals, l'estadístic de prova sota el supòsit de la H_0 certa és:

$$Z=\frac{(\overline{X}_1-\overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

En un contrast a doble cua haurem de rebutjar la hipòtesi nul·la a favor de l'alternativa ($H_1: \mu_1\neq\mu_2$) quan ($\overline{X}_1-\overline{X}_2$) sigui molt més gran o molt més petit que zero. Pel contrari, si aquest valor és pròxim a zero no la podrem rebutjar.

Fixada la probabilitat α de rebutjar la hipòtesi nul·la essent certa (nivell de significació), podem trobar els valors crítics, $-z_{\alpha/2}$ i $z_{\alpha/2}$, que delimiten la zona d'acceptació i determinen el criteri de decisió.



- $P(-z_{\alpha/2}<Z< z_{\alpha/2}) = 1-\alpha$ (regió d'acceptació).
- $P(Z> z_{\alpha/2}) = \alpha/2$ i $P(Z<-z_{\alpha/2}) = \alpha/2$ (regió crítica).

Per tant, es rebutjarà la hipòtesi nul·la si l'estadístic és més gran que $z_{\alpha/2}$ o més petit que $-z_{\alpha/2}$, és a dir, si pertany a la regió crítica i llavors direm que hi ha diferència significativa entre les mitjanes d'ambdues mostres. Pel contrari, si l'estadístic de prova Z pertany a la regió d'acceptació, no es pot rebutjar la hipòtesi nul·la i, en conseqüència, no rebutgem la igualtat de mitjanes poblacionals.

Hipòtesis:

$H_0: \mu_1-\mu_2=0$

$H_1: \mu_1-\mu_2\neq 0$

Estadístic de prova:

$$Z=\frac{(\bar{X}_1-\bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

Criteri de decisió:

Si $|Z|> z_{\alpha/2}$ es rebutja la H_0

Si la hipòtesi alternativa és a una cua (direccional), és a dir, es vol contrastar que la diferència de mitjanes poblacionals és major que zero (cua superior $H_1: \mu_1>\mu_2$) o menor que zero (cua inferior $H_1: \mu_1<\mu_2$), la regió crítica per a un nivell de significació α ve determinada per un valor z_α que:

- $P(Z>z_\alpha) = \alpha$ (cua superior).
- $P(Z<-z_\alpha) = \alpha$ (cua inferior).

Hipòtesis

(cua superior)	(cua inferior)
H_0: $\mu_1-\mu_2=0$	H_0: $\mu_1-\mu_2=0$
H_1: $\mu_1-\mu_2>0$	H_1: $\mu_1-\mu_2<0$

Estadístic de prova:

$$Z=\frac{(\overline{X}_1-\overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

Criteri de decisió:

Si $Z> z_\alpha$ es rebutja la H_0 Si $Z< -z_\alpha$ es rebutja la H_0

La metodologia del contrast exposat es pot aplicar a poblacions on no es puguin acceptar els supòsits de normalitat i de variància coneguda, sempre i quan la grandària de les mostres sigui superior a 30. Així, en el cas de variàncies desconegudes, si les grandàries mostrals són elevades, S_1^2 i S_2^2 (variàncies mostrals) són estimadors consistents de σ_1^2 i σ_2^2, respectivament, i pel TCL podem aproximar la distribució de ($\overline{X}_1$-$\overline{X}_2$) a la Normal.

Exemple 5.8

Un fabricant de cotxes equipa els seus models amb bateries de la marca A. Per no dependre d'un sol proveïdor es planteja la possibilitat d'equipar alguns dels models amb bateries d'una altra marca B. Atès que el preu de la marca B és superior al de la marca A, el fabricant, abans de prendre una decisió, vol estar segur que la qualitat de B és superior a la qualitat de A. Per experiència creu que la durada mitjana de les diferents marques pot diferir, però la desviació estàndard es pot considerar similar amb un valor aproximat de 4 mesos. Es proven 40 bateries de la marca A i 45 bateries de la marca B i s'obtenen, respectivament, unes durades mitjanes de 30 i 32 mesos. Es pot dir que la diferència entre les mitjanes mostrals és significativa a nivell poblacional? (α=0,05)

Solució:

A: $n_1 = 40$ $\overline{X}_1 \sim N(\mu_1, 4/\sqrt{40})$ $\overline{x}_1 =30$

B: $n_2 = 45$ $\overline{X}_2 \sim N(\mu_2, 4/\sqrt{45})$ $\overline{x}_2 =32$

Formulació de les hipòtesis:

H_0: $\mu_1 = \mu_2$ $\quad$ H_0: $\mu_1 - \mu_2 = 0$

H_1 : $\mu_1 < \mu_2$ $\quad$ H_1: $\mu_1 - \mu_2 < 0$

Selecció de l'estadístic de prova:

Si la H_0 és certa,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{30-32}{\sqrt{\frac{16}{40} + \frac{16}{45}}} = -2{,}3$$

Aplicació dels criteris de decisió:

Per a α=0,05, el punt crític és: $z_\alpha = -1{,}64$

Com que Z = -2,3 $\notin$ [-1,64, $+\infty$] es conclou que hi ha evidència empírica suficient per rebutjar la hipòtesi d'igualtat de mitjanes a favor de l'alternativa que suposa una major qualitat de la marca B.

5.6.2 VARIÀNCIES POBLACIONALS DESCONEGUDES I IGUALS

Si les variàncies poblacionals són desconegudes però iguals a σ^2 ($\sigma_1^2=\sigma_2^2=\sigma^2$**)** s'haurà de determinar l'estimador òptim d'aquest paràmetre. En el capítol anterior vàrem veure que l'estimador no esbiaixat de σ^2 és

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

on S_1^2 i S_2^2 són les variàncies mostrals obtingudes de les mostres aleatòries de grandària n_1 i n_2 extretes de forma independent de les respectives poblacions.

Si la hipòtesi nul·la és certa (H_0: μ_1-μ_2=0), l'estadístic de prova

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

segueix una distribució *t* de Student amb n_1+n_2-2 graus de llibertat. Fixada la probabilitat α de rebutjar la hipòtesi nul·la essent certa, podem trobar els valors crítics que delimiten la zona d'acceptació a la distribució t de Student.

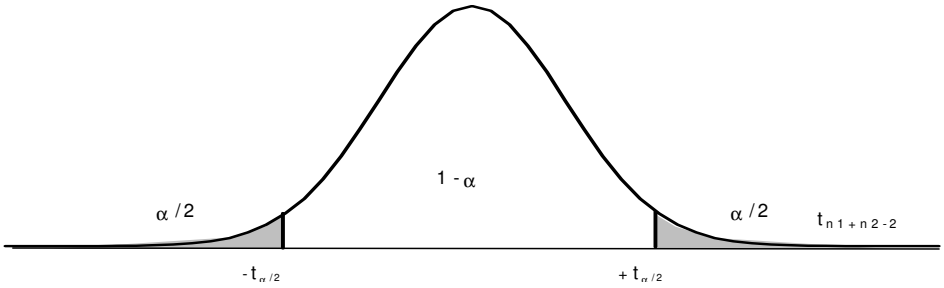
Si el contrast és bilateral:

H_0: μ_1-μ_2=0

H_1: μ_1-$\mu_2 \neq 0$

els dos valors crítics $-t_{\alpha/2}$ i $t_{\alpha/2}$ són:

- $P(-t_{\alpha/2}<t< t_{\alpha/2}) = 1-\alpha$ (regió d'acceptació).
- $P(t> t_{\alpha/2}) = \alpha/2$ i $P(t<-t_{\alpha/2}) = \alpha/2$ (regió crítica).



Per tant, es rebutjarà la hipòtesi nul·la si el valor absolut de l'estadístic t és més gran que $t_{\alpha/2}$.

Hipòtesis:

$H_0: \mu_1-\mu_2=0$

$H_1: \mu_1-\mu_2\neq 0$

Estadístic de prova:

$$t=\frac{(\bar{X}_1-\bar{X}_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n1+n2-2}$$

Criteri de decisió:

Si $|t|> t_{\alpha/2}$ es rebutja la H_0

Si el contrast és unilateral la metodologia a seguir és:

Hipòtesis

(cua superior)	(cua inferior)
$H_0: \mu_1-\mu_2=0$	$H_0: \mu_1-\mu_2=0$
$H_1: \mu_1-\mu_2>0$	$H_1: \mu_1-\mu_2<0$

Estadístic de prova:

$$t=\frac{(\bar{X}_1-\bar{X}_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n1+n2-2}$$

Criteri de decisió:

Si $t> t_{\alpha}$ es rebutja la H_0 Si $t< -t_{\alpha}$ es rebutja la H_0

En el cas que la grandària de la mostra sigui superior a 30, l'estadístic de prova

$$t=\frac{(\bar{X}_1-\bar{X}_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

es distribueix aproximadament com una normal estandarditzada.

Exemple 5.9

Amb la finalitat de determinar si la velocitat que s'assoleix circulant per Barcelona és independent de l'hora en què es realitza el trajecte, s'ha efectuat un seguiment d'una mostra de cotxes en dues franges horàries diferents i s'han obtingut els resultats següents, relatius als quilòmetres que havien recorregut durant 1/2 hora:

Franja 1: n= 15 $\sum X=570$ $\sum X^2=21896$

Franja 2: n= 12 $\sum X=240$ $\sum X^2=5018$

Contrasteu la hipòtesi que la velocitat que es pot assolir és independent de l'hora de circulació, assumint subpoblacions normals amb la mateixa variància i mostres independents.

<u>*Solució:*</u>

Franja 1: n_1= 15 $\bar{X}_1$=38 S_1^2=16,86
Franja 2: n_2= 12 $\bar{X}_2$=20 S_2^2=19,82

$$t=\frac{(\bar{X}_1-\bar{X}_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n1+n2-2}$$

$$S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}=\frac{236{,}04+218{,}02}{15+12-2}=18{,}16$$

Formulació de les hipòtesis:

H_0: $\mu_1-\mu_2=0$
H_1: $\mu_1-\mu_2\neq 0$

Selecció de l'estadístic de prova:

$$t=\frac{(\bar{X}_1-\bar{X}_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{38-20}{\sqrt{18{,}16(\frac{1}{15}+\frac{1}{12})}}=10{,}9$$

Aplicació del criteri de decisió:

Per a $\alpha = 0,05$ i 25 g.ll., el punt crític és $t_{25,\alpha/2} = 2,06$.

Com que $t = 10,9 \notin [-2,06; 2,06]$ es conclou que hi ha evidència empírica suficient per rebutjar la hipòtesi d'igualtat de velocitats i, per tant, la velocitat depèn de la franja horària.

5.7 CONTRAST PER A LA DIFERÈNCIA DE VARIÀNCIES

En comparar dues poblacions de paràmetres desconeguts ens pot interessar, com hem vist a l'apartat anterior, contrastar si hi ha diferència entre les mitjanes poblacionals en el cas que les variàncies poblacionals siguin desconegudes però es puguin considerar iguals. La veracitat d'aquesta última consideració es pot validar a partir d'un contrast d'igualtat de variàncies poblacionals. Òbviament també pot tenir interès fer aquest contrast per si mateix, és a dir, en observar dues poblacions normals podem estar interessats en comparar les dispersions poblacionals.

Donades dues mostres aleatòries independents extretes de dues poblacions normals amb paràmetres desconeguts es vol contrastar la hipòtesi nul·la que les variàncies de les dues poblacions són iguals.

$$H_0: \sigma_1^2 = \sigma_2^2$$

Agafem com a estimadors de σ_1^2 i σ_2^2, les variàncies mostrals S_1^2 i S_2^2 obtingudes de les dues mostres aleatòries i independents de grandàries n_1 i n_2, respectivament. Sabem que l'estadístic $F=\dfrac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$ es distribueix com una *F* de Snedecor amb (n_1-1) graus de llibertat al numerador i (n_2-1) graus de llibertat al denominador. Si la hipòtesi nul·la és certa ($H_0: \sigma_1^2 = \sigma_2^2$), llavors la variable aleatòria F queda

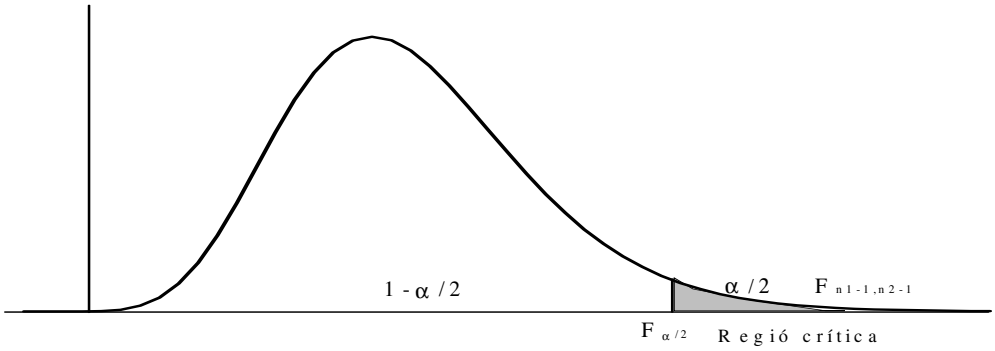
$$F=\frac{S_1^2}{S_2^2}$$

Si realitzem el contrast a doble cua ($H_1: \sigma_1^2 \neq \sigma_2^2$, bilateral) rebutjarem la H_0 quan F prengui valors extremadament diferents a la unitat, això indica que existeix una diferència significativa entre les dues variàncies poblacionals.

El criteri de decisió queda determinat pels valors crítics que es localitzen a les taules de la distribució *F* de Snedecor. Per tal de treballar directament sense necessitat de transformar aquests valors és necessari que F sigui més gran que 1 i, en conseqüència, sempre s'haurà de posar al numerador la major variància mostral observada: $S_1^2 > S_2^2$. Per tant, només es considera la cua superior de la regió critica, que en el cas del contrast a dues cues acumula una probabilitat igual a $\alpha/2$.

Així, fixat un nivell de significació α, el valor crític $F_{\alpha/2}$ compleix:

- $P(F>F_{\alpha/2}) = \alpha/2$ (regió crítica).



I el criteri de decisió és rebutjar la hipòtesi nul·la quan $F > F_{\alpha/2}$.

Hipòtesis:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Estadístic de prova:

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n1-1;n2-1)}$$

Criteri de decisió:

Si $F > F_{\alpha/2}$ es rebutja la H_0.

Si el contrast és direccional, la hipòtesi alternativa es defineix a cua superior: $H_1: \sigma_1^2 > \sigma_2^2$ (ja que com hem dit S_1^2 és la major variància mostral; $S_1^2 > S_2^2$). Fixat un nivell de significació igual a α, el criteri de decisió és rebutjar la hipòtesi nul·la quan $F > F_\alpha$, essent $P(F>F_\alpha) = \alpha$.

Hipòtesis:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

Estadístic de prova:

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n1-1;n2-1)}$$

Criteri de decisió:

Si $F > F_\alpha$ es rebutja la H_0.

Exemple 5.10

Per tal d'establir si les cotitzacions de dos tipus de títols de renda fixa (A i B) presenten la mateixa dispersió, s'obtenen dues mostres aleatòries i independents de 17 dies de cotització cadascuna. Les cotitzacions al tancament de A van presentar una variància de 125,25 i les de B una variància de 638,5. Si les poblacions són normals, podem dir que ambdós títols presenten la mateixa estabilitat en la seva cotització al 10% de significació?

Solució:

B: $n_1 = 17$ $S_1^2 = 638,5$

A: $n_2 = 17$ $S_2^2 = 125,25$

(Recordeu que la major de les dues variàncies va al numerador per tal d'utilitzar directament les taules de la distribució *F* de Snedecor.)

Formulació de les hipòtesis:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Selecció de l'estadístic de prova:

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n1-1;n2-1)} \Rightarrow F = \frac{638,5}{125,25} = 5,098$$

Aplicació del criteri de decisió:

Per a $\alpha = 0,1$ i 16 graus de llibertat al numerador i al denominador el punt crític és $F_{\alpha/2,(16,16)} = 2,33$.

Com que $F = 5,09 > F_{\alpha/2,(16,16)} = 2,33$ es conclou que hi ha evidència empírica suficient per rebutjar la hipòtesi d'igualtat de variàncies i, per tant, les cotitzacions no presenten la mateixa estabilitat.

5.8 CONTRAST PER A π

A vegades es vol contrastar si la proporció poblacional, π, d'elements amb una determinada característica (proporció d'èxits) és igual a un determinat valor π_0.

$$H_0: \pi = \pi_0$$

En aquests casos utilitzem com a estimador de π la proporció mostral d'èxits p. Quan la grandària de la mostra n és suficientment gran ($n\pi_0(1-\pi_0)>5$) i la hipòtesi nul·la és certa, la variable aleatòria p presenta una distribució aproximadament normal de paràmetres:

$$p \sim N\left(\pi_0; \sqrt{\frac{\pi_0(1-\pi_0)}{n}}\right)$$

Estandarditzant obtenim

$$Z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1).$$

Si el contrast és a dues cues, $H_1: \pi \neq \pi_0$, fixat el nivell de significació α, podem determinar els valors crítics $-z_{\alpha/2}$ i $z_{\alpha/2}$ que:

- $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha$ (regió d'acceptació).
- $P(Z > z_{\alpha/2}) = \alpha/2$ i $P(Z < -z_{\alpha/2}) = \alpha/2$ (regió crítica).

i, com a criteri de decisió, rebutjarem la hipòtesi nul·la quan el valor absolut de Z sigui més gran que $z_{\alpha/2}$, és a dir, $|Z| > z_{\alpha/2}$.

Hipòtesis:

$$H_0: \pi = \pi_0$$

$$H_1: \pi \neq \pi_0$$

Estadístic de prova:

$$Z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$$

Criteri de decisió:

Si $|Z| > z_{\alpha/2}$ es rebutja la H_0.

Si la hipòtesi alternativa és a una cua (direccional), és a dir, si es vol contrastar que la proporció poblacional és major o menor que π_0 ($H_1: \pi > \pi_0$, cua superior o

H_1: $\pi<\pi_0$ cua inferior), la regió crítica per a un nivell de significació α vindrà determinada per un valor z_α que:

- $P(Z>z_\alpha) = \alpha$ (cua superior)
- $P(Z<-z_\alpha) = \alpha$ (cua inferior).

Per tant, com a criteri de decisió, rebutjarem la H_0 si $Z>z_\alpha$ a cua superior o bé $Z<-z_\alpha$ a cua inferior.

Hipòtesis:

(cua superior)	(cua inferior)
$H_0: \pi= \pi_0$	$H_0: \pi= \pi_0$
$H_1: \pi>\pi_0$	$H_1: \pi<\pi_0$

Estadístic de prova:

$$Z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$$

Criteri de decisió:

Si $Z> z_\alpha$ es rebutja la H_0. Si $Z<-z_\alpha$ es rebutja la H_0.

Exemple 5.11

Una empresa de productes de neteja es planteja la conveniència del canvi dels actuals envasos de plàstic per uns nous envasos de cartró reciclat. El canvi no es farà tret que més del 60% dels seus clients ho prefereixi. Els resultats d'una enquesta efectuada a 200 clients indiquen que 140 estan a favor del canvi. Verifiqueu amb una significació de l'1% que el canvi s'efectuarà a partir de l'evidència que proporciona la mostra.

<u>*Solució:*</u>

$$n = 200 \quad p \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right) \quad p = 140/200 = 0,7$$

Formulació de les hipòtesis:

$H_0: \pi=0,6$
$H_1: \pi>0,6$

Selecció de l'estadístic de prova:

$$Z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$$

Càlcul de l'estadístic de prova:

$$Z=\frac{0,7-0,6}{\sqrt{\frac{0,6(1-0,6)}{200}}}=2,887$$

Aplicació del criteri de decisió:

Per a $\alpha = 0,01$, el punt crític és $z_\alpha = 2,32$.

Com que $Z = 2,887 \notin [-\infty ; 2,32]$ es conclou que hi ha evidència empírica suficient per rebutjar la hipòtesi nul·la a favor de l'alternativa i, per tant, l'empresa canviarà l'envàs.

5.9 CONTRAST PER A π_1-π_2

En aquest apartat es presenta el contrast que permet comparar dues proporcions poblacionals, π_1 i π_2, a partir de dues mostres aleatòries independents.

Sigui π_1 la proporció d'èxits d'una població de la qual s'extreu una mostra aleatòria de grandària n_1 i s'obté una proporció mostral d'èxits p_1. Sigui π_2 la proporció poblacional d'èxits d'una altra població, d'on s'extreu una mostra aleatòria de grandària n_2 independent de l'anterior, i s'obté una proporció mostral d'èxits p_2.

La hipòtesi nul·la que es planteja és que no hi ha diferències significatives entre les dues proporcions poblacionals:

$$H_0:\pi_1=\pi_2 \quad \Rightarrow \quad H_0:\pi_1-\pi_2=0$$

L'estimador diferència de proporcions mostrals (p_1-p_2), quan n_1 i n_2 són suficientment grans, és a dir, si $n_1\pi_1(1-\pi_1)>5$ i $n_2\pi_2(1-\pi_2)>5$, presenta una distribució aproximadament normal de paràmetres:

$$E(p_1-p_2)=\pi_1-\pi_2$$

$$V(p_1-p_2)=\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}$$

$$p_1-p_2 \sim N\left(\pi_1-\pi_2;\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

Estandarditzant

$$Z=\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\sim N(0,1),$$

obtenim l'estadístic a partir del qual podem realitzar el contrast.

Sota el supòsit que la hipòtesi nul·la és certa (H_0:$\pi_1-\pi_2=0$), és a dir, π_1 i π_2 són iguals a una proporció comú π, aleshores l'estadístic (p_1-p_2) presenta una distribució aproximadament normal amb paràmetres

$$E(p_1-p_2)=0 \text{ i } V(p_1-p_2)=\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

Com que p_1 i p_2 són estimacions alternatives de la proporció poblacional comú π, a partir de la informació de les dues mostres es pot obtenir la següent estimació de π:

$$\hat{\pi}=\frac{n_1p_1+n_2p_2}{n_1+n_2}=\frac{X_1+X_2}{n_1+n_2}$$

on X_1 i X_2 són el nombre d'èxits observats en les respectives mostres.

Per tant, l'estadístic de prova queda:

$$Z=\frac{(p_1-p_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim N(0,1)$$

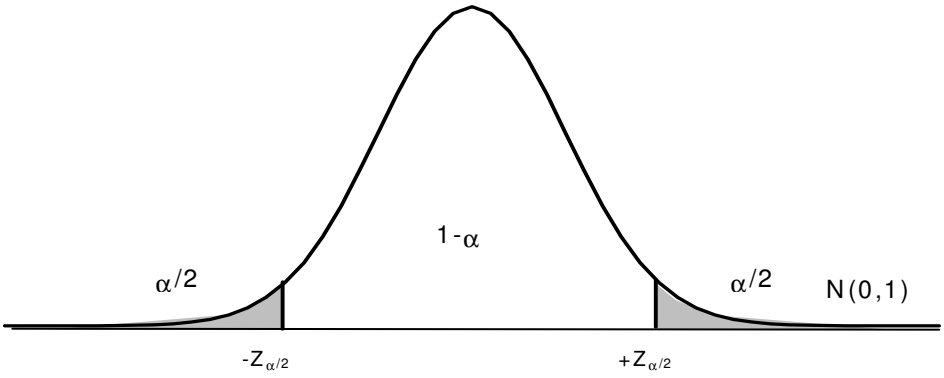
Si el contrast és a dues cues:

$$H_0: \pi_1-\pi_2=0$$

$$H_1: \pi_1-\pi_2\neq 0$$

Fixat el nivell de significació α, podem determinar els valors crítics $-z_{\alpha/2}$ i $z_{\alpha/2}$ que compleixen:

- $P(-z_{\alpha/2}\leq Z\leq z_{\alpha/2})=1-\alpha$ (regió d'acceptació).
- $P(Z>z_{\alpha/2})=\alpha/2$ i $P(Z<-z_{\alpha/2})=\alpha/2$ (regió crítica).



Es rebutjarà la hipòtesi nul·la si el valor absolut de l'estadístic Z és més gran que $z_{\alpha/2}$.

Hipòtesis:

$H_0: \pi_1-\pi_2=0$

$H_1: \pi_1-\pi_2\neq 0$

Estadístic de prova:

$$Z=\frac{(p_1-p_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1}+\frac{1}{n_2})}}\sim N(0,1)$$

Criteri de decisió:

Si $|Z|> z_{\alpha/2}$ es rebutja la H_0.

Si el contrast és unilateral la metodologia a seguir és:

Hipòtesis:

(cua superior)	(cua inferior)
$H_0: \pi_1-\pi_2=0$	$H_0: \pi_1-\pi_2=0$
$H_1: \pi_1-\pi_2>0$	$H_1: \pi_1-\pi_2<0$

Estadístic de prova:

$$Z=\frac{(p_1-p_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1}+\frac{1}{n_2})}}\sim N(0,1)$$

Criteri de decisió:

Si $Z > z_\alpha$ es rebutja la H_0. Si $Z < -z_\alpha$ es rebutja la H_0.

Exemple 5.12

Respecte a una determinada mesura econòmica, es desitja comprovar si hi ha diferència entre la proporció d'assalariats i la d'empresaris que hi són favorables. D'una mostra de 200 empresaris, 118 van manifestar que estaven d'acord amb la mesura i d'una mostra de 250 assalariats, n'hi van estar d'acord 138. La diferència entre les proporcions mostrals és suficient per afirmar que la

proporció d'empresaris favorables a la mesura és més gran que la proporció d'assalariats favorables? ($\alpha = 0,01$)

Solució:

$n_1 = 200 \quad p_1 \sim N(\pi_1, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}) \quad p_1 = 118/200 = 0,59$

$n_2 = 250 \quad p_2 \sim N(\pi_2, \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}) \quad p_2 = 138/250 = 0,552$

Formulació de les hipòtesis:

$H_0: \pi_1 = \pi_2 = \pi \qquad H_0: \pi_1 - \pi_2 = 0$

$H_1: \pi_1 > \pi_2 \qquad H_1: \pi_1 - \pi_2 > 0$

Selecció de l'estadístic de prova:

En el cas que la H_0 sigui certa

$$Z = \frac{(p_1 - p_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}} \sim N(0,1).$$

Càlcul de l'estadístic de prova:

Atès que segons la H_0 les proporcions poblacionals són iguals, el millor estimador de π serà:

$$\hat{\pi} = \frac{118 + 138}{200 + 250} = 0,569$$

$$Z = \frac{0,59 - 0,552}{\sqrt{\frac{0,569 0,432}{200} + \frac{0,569 0,432}{250}}} = 0,809$$

Aplicació del criteri de decisió:

Per a $\alpha = 0,01$, el punt crític és $z_\alpha = 2,32$.

Com que $Z = 0,809 \in [-\infty; 2,32]$ es conclou que no hi ha evidència empírica suficient per rebutjar la hipòtesi d'igualtat de proporcions en favor de l'alternativa i, per tant, l'opinió d'aquests col·lectius no és significativament diferent.

5.10 ANÀLISI DE LA VARIÀNCIA

L'anàlisi de la variància (ANOVA) és, a despit del seu nom, una tècnica dissenyada per contrastar la hipòtesi que les mitjanes de tres o més poblacions

són iguals. L'origen del seu nom rau, com es veurà, en què l'estadístic de prova es basa en el quocient de dos estimadors de la variància poblacional.

Per dur a terme aquesta prova es disposa de k mostres aleatòries i independents obtingudes a partir de k poblacions.

		Població				
		1	2	j		k
Població	Mitjana Variància	μ_1 σ^2	μ_2 σ^2	μ_j σ^2	... ...	μ_k σ^2
Mostra		x_{11} x_{21}	x_{12} x_{22}	x_{1j} x_{2j} ... x_{ij} ...		x_{1k} x_{2k} ... ...
	Grandària	n_1	n_2	n_j		n_k
	Mitjana	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_j$		$\bar{x}_k$
	Variància	s_1^2	s_2^2	s_j^2		s_k^2

on:

x_{ij} és l'observació i-èsima de la mostra j.

n_j és el nombre d'observacions de la mostra j.

Mitjana de la mostra j-èsima: $\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$, j = 1,2, ..., k, on $\sum_{i=1}^{n_j} x_{ij}$ és la suma de totes les observacions mostrals de la població j-èsima.

Mitjana Global: $\bar{x} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n}$ on $\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}$ és la suma de totes les observacions i n és igual a $n_1+n_2+...+n_k$.

Variància de la mostra j-èsima: $s_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{n_j - 1}$

Variància Global: $S^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2}{n-1}$

Amb aquesta informació es vol contrastar la hipòtesi nul·la que les mitjanes d'aquestes k poblacions són iguals:

$$H_0: \mu_1 = \mu_2 = ... = \mu_k = \mu$$

$$H_1: H_0 \text{ no és certa.}$$

Per realitzar aquest contrast cal que es compleixin els següents supòsits:

1. Les k mostres han d'ésser aleatòries i independents entre si.
2. Les poblacions han d'ésser Normals.
3. Les variàncies de les k poblacions han d'ésser idèntiques: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 = \sigma^2$.

L'estadístic de prova que s'empra per contrastar la H_0 es basa en comparar dos orígens de la variabilitat de les dades analitzades: **variació dintre** de les mostres i **variació entre** les mostres.

Per comprendre millor el concepte d'aquestes fonts de variació hem de tenir present que l'ANOVA intenta explicar a què són degudes les diferències entre els valors observats de la variable i la seva mitjana global, és a dir, $x_{ij} - \overline{X}$. Aquestes diferències s'expliquen en part per les divergències existents entre les observacions de cada grup respecte a la seva mitjana, $x_{ij} - \overline{X}_j$, i en part per les diferències entre les mitjanes de cada grup respecte a la mitjana global, $\overline{X}_j - \overline{X}$. Algebraicament, per a una observació qualsevol, podem expressar la relació anterior com:

$$(x_{ij} - \overline{X}) = (x_{ij} - \overline{X}_j) + (\overline{X}_j - \overline{X})$$

Els termes de la dreta de la igualtat anterior són els elements bàsics de la variació dintre i entre mostres, respectivament.

Si elevem al quadrat i sumem els termes de l'expressió anterior obtenim:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{X}_j)^2 + \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X})^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{nj}(x_{ij} - \overline{X}_j)(\overline{X}_j - \overline{X})$$

Si es té en compte que el tercer sumand és zero, queda:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{X}_j)^2 + \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X})^2$$

o el que és el mateix,

Suma de Quadrats Total = Suma de Quadrats Dintre mostres + Suma de Quadrats Entre mostres

que en notació abreujada es pot simbolitzar com:

$$SQT = SQD + SQE$$

La variació total ve donada per SQT. La variació entre mostres, SQE, és l'estadístic que mesura la proximitat entre les mitjanes mostrals, i SQD proporciona una mesura del grau de variabilitat de les dades analitzades.

Dividint les sumes de quadrats anteriors pels seus respectius graus de llibertat obtenim els estadístics següents:

$$MQD = \frac{SQD}{n - k} \quad \text{i} \quad MQE = \frac{SQE}{k - 1}$$

Quan la hipòtesi nul·la és certa es demostra que els estadístics MQD i MQE són dos estimadors no esbiaixats de la variància poblacional i el quocient entre ambdós estadístics,

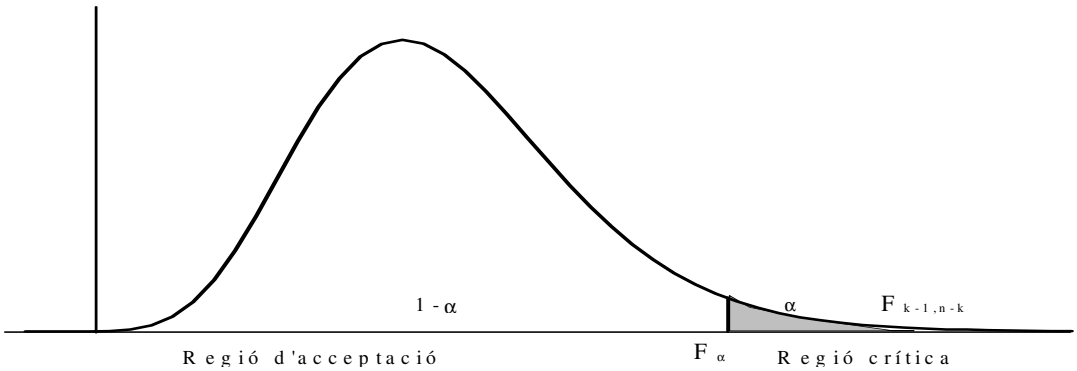
$$F = \frac{MQE}{MQD} = \frac{SQE/(k-1)}{SQD/(n-k)}$$

es distribueix segons una F amb (k-1) graus de llibertat al numerador i (n-k) al denominador.

En el cas que la hipòtesi nul·la sigui certa els valors de MQE i MQD seran molt semblants i l'estadístic de prova serà proper a la unitat. Per contra, si no és certa la H_0 el valor de MQE serà molt més gran que MQD i obtindrem un valor elevat de l'estadístic F que ens durà a rebutjar la H_0.

Fixada la probabilitat α de rebutjar la H_0 essent certa, podem trobar el valor crític F_α que delimita la zona d'acceptació a la distribució F.

- $P(F < F_\alpha) = 1-\alpha$ (regió d'acceptació).
- $P(F > F_\alpha) = \alpha$ (regió crítica).



Per tant, es rebutjarà la hipòtesi nul·la si el valor de l'estadístic és més gran que F_α.

Hipòtesis:

$$H_0: \mu_1 = \mu_2 = ... = \mu_k = \mu$$

$$H_1: H_0 \text{ no és certa}$$

Estadístic de prova:

$$F = \frac{MQE}{MQD} = \frac{SQE/(k-1)}{SQD/(n-k)} \sim F_{k-1,n-k}$$

Criteri de decisió:

Si $F > F_\alpha$ es rebutja la H_0

Per tal de calcular les sumes de quadrats es poden emprar les següents expressions:

$$SQE = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2$$

$$SQT = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{X})^2 = (n-1) S^2$$

$$SQD = SQT - SQE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_j)^2 = \sum_{j=1}^{k} (n_j - 1) S_j^2$$

Els resultats de l'anàlisi de la variància s'acostumen a presentar en forma de taula:

Font de Variació	Graus de Llibertat	Sumes de Quadrats	Sumes de Quadrats Mitjanes	Estadístic F
Entre Mostres	k-1	SQE	MQE=SQE/(k-1)	
Dintre Mostres	n-k	SQD	MQD=SQD/(n-k)	$F = \frac{MQE}{MQD} \sim F_{k-1,n-k}$
Total	n-1	SQT		

Exemple 5.13

Una organització de consumidors manté que la publicitat relativa a la classificació de les marques de tabac en normal, mitjana o baixa, depenent de la seva quantitat de nicotina, és enganyosa, ja que el contingut mitjà de nicotina de

les cigarretes és semblant a totes les marques. L'anàlisi del contingut de nicotina en mg de 24 marques dóna els següents resultats.

	Baixa	Mitjana	Normal
	0,13 0,40 0,42 0,57 0,61 0,67 0,69 0,74 0,76 0,78	0,82 0,86 0,91 0,95 0,97	
Mitjana	0,577	0,902	1,070
Variància	0,04218	0,00387	0,00662
n_j	10	5	9

En base a aquesta informació es vol contrastar el supòsit de l'organització de consumidors.

Solució:

Hipòtesis:

H_0:$\mu_1 = \mu_2 = ... = \mu_k = \mu$

H_1:H_0 no és certa

Estadístic de prova: $F = \dfrac{MQE}{MQD} = \dfrac{SQE/(k-1)}{SQD/(n-k)}$

$$SQD = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_j)^2 = \left\{ S_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_j)^2}{n_j - 1} \right\} = \sum_{j=1}^{k}(n_j - 1)S_j^2 =$$

$$= \sum_{j=1}^{3}(n_j - 1)S_j^2 = (10-1)0{,}04218 + (5-1)0{,}00387 + (9-1)0{,}00662 = 0{,}448$$

$$SQE = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2 = 10(0{,}577 - 0{,}83)^2 + 5(0{,}902 - 0{,}83)^2 + 9(1{,}07 - 0{,}83)^2 =$$

$=1{,}184$

ANÀLISI DE LA VARIÀNCIA				
Origen de la variació	Sumes de Quadrats	Graus de llibertat	Sumes de Quadrats Mitjanes	F
Entre mostres	1,184	2	0,592	27,7
Dintre mostres	0,448	21	0,021	
Total	1,632	23		

Aplicació del criteri de decisió:

Per a α = 0,05 amb 2 graus de llibertat al numerador i 21 al denominador, el valor crític és $F_{\alpha,(2,21)}$=3,47. Com que F = 27,7 > 3,47 es conclou que hi ha evidència empírica suficient per rebutjar la hipòtesi d'igualtat de mitjanes.

5.11 EXERCICIS PROPOSATS

Exercici 1. Un fabricant assegura que les seves caixes de galetes tenen un pes net de 480 gr. Obtinguda una mostra aleatòria de 10 caixes, es va trobar que aquestes presentaven els pesos (en gr) següents: 420, 410, 520, 470, 430, 500, 450, 460, 510, 470.

a) Si el pes de les caixes segueix una distribució Normal amb desviació estàndard de 30 gr, podem considerar certa l'afirmació del fabricant amb un nivell de significació del 5%?
b) Si es desconeix la desviació estàndard de la població, canvia la decisió obtinguda en contrastar la hipòtesi anterior amb el mateix nivell de significació?

Exercici 2. Tres persones diferents afirmen que els ingressos mitjans dels auxiliars administratius són de 720, 600 i 640 euros mensuals, respectivament. A partir d'una mostra de 16 auxiliars administratius s'ha obtingut un sou mitjà de 610 euros, amb una desviació estàndard de 58 euros. Sota el supòsit que la població és Normal:

a) A un nivell de significació del 5%, proveu cadascuna de les afirmacions anteriors.

b) Obteniu l'interval de confiança del 95% per a la mitjana poblacional i raoneu si són acceptables les hipòtesis de l'apartat anterior.

Exercici 3. Mitjançant una mostra aleatòria de grandària 100 extreta d'una població caracteritzada per la variable aleatòria X amb distribució Normal i desviació estàndard 125 (σ=125), es vol contrastar: H_0:μ = 500 i H_1:μ = 525 a un nivell de significació del 5%:

a) Quina és la probabilitat de cometre un error de tipus II?
b) Quina és la potència del contrast?
c) Quina ha d'ésser la grandària mostral per reduir l'error de tipus II a un 0,05 sense que s'incrementi la probabilitat de cometre l'error de tipus I?
d) Com queda modificada la potència del contrast amb el canvi en la grandària mostral de l'apartat anterior?

Exercici 4. La variància d'una determinada mostra aleatòria de 10 observacions ha resultat ser igual a 81,3. Sota el supòsit que la població es distribueix normalment, contrasteu la hipòtesi que el valor de la variància poblacional és 125 al 5% de significació.

Exercici 5. La distància mitjana que ha de recórrer un cotxe d'una determinada marca per aturar-se quan circula a 60 Km/h és de 6,5 m. amb desviació estàndard 0,5. El departament d'enginyeria de la companyia ha dissenyat un nou sistema de frens que es considera més eficaç, tant en precisió com en temps de frenada. Per tal de comprovar-ho s'instal·la el nou sistema en 81 cotxes i les proves demostren que la distància mitjana per aturar un cotxe a una velocitat de 60 Km/h és de 6,34 m amb una desviació estàndard de 0,45 m. Si la població objecte d'estudi té una distribució Normal, es pot acceptar que el nou sistema de frens és més eficaç que l'anterior?

Exercici 6. Una màquina produeix una mitjana de 20 unitats per hora amb variància 81. El seu venedor diu que afegint-li una determinada peça s'incrementarà el seu ritme de producció. Rectificada la màquina s'observa la producció en una mostra de 50 hores per tal de poder verificar l'afirmació del venedor. Si la variància poblacional es manté constant i la població és Normal:

a) A partir de quin valor de la mitjana mostral rebutjarem la H_0 al 5% de significació?
b) Si es fixa el valor crític de la mitjana mostral en 23 unitats per hora, quin és el nivell de significació del contrast?

c) Si la hipòtesi alternativa és $\mu = 25$, quina és la potència del contrast realitzat a l'apartat a? I a l'apartat b?

Exercici 7. Una Companyia de transport per carretera sospita que la duració mitjana de 28000 Km que com a mínim anuncia una marca d'amortidors no correspon a la real en el sentit que és excessiva. Per verificar-ho instal·la aquests amortidors en 40 camions i obté una mitjana de 27563 Km amb una desviació estàndard de 1348 Km. Si la vida útil dels amortidors és una variable aleatòria Normal,

a) Realitzeu el contrast anterior a l'1% de significació.
b) Calculeu el P-value (valor crític P) del contrast anterior.

Exercici 8. Tradicionalment es creia que la variància de la renda familiar en dos municipis era, aproximadament, la mateixa. Una revista afirma que aquesta variància és estrictament superior al municipi A. Per contrastar-ho es trien dues mostres independents de 51 famílies a cada municipi i s'observa una variància mostral 293 i 205 per als municipis A i B, respectivament. Si es suposa que la renda familiar presenta una distribució Normal, es pot acceptar l'afirmació de la revista al 5% de significació?

Exercici 9. Un fabricant vol comparar la resistència mitjana del fil que produeix amb la del que fabrica la competència. Mesura les resistències de dues mostres i obté:

Fabricant: $n_1 = 100$ $\overline{X}_1 = 110{,}8$ $S_1 = 10{,}2$

Competència: $n_2 = 100$ $\overline{X}_2 = 108{,}2$ $S_2 = 12{,}4$

Si les mostres són independents i s'han obtingut de dues poblacions Normals amb iguals variàncies, a quina conclusió arribarà el fabricant al 5% de significació?

Exercici 10. Un economista disposa de les dades següents, relatives al volum de vendes (en milers d'u.m.), extretes de dues mostres aleatòries d'empreses d'indústries associades:

Mostra 1: $n_1 = 26$ $\overline{X}_1 = 350$ $S_1^2 = 400$

Mostra 2: $n_2 = 51$ $\overline{X}_2 = 360$ $S_2^2 = 500$

Si les poblacions són Normals, com es podria justificar el fet d'agrupar els resultats obtinguts d'aquestes dues mostres al 10% de significació?

Exercici 11. Un jugador assegura que el dau amb el que juga està trucat ja que obté menys resultats parells que senars. S'efectuen 100 llançaments del dau i es comptabilitzen 35 resultats parells i 65 de senars. En base a aquest resultat, es pot donar la raó al jugador?

Exercici 12. El fabricant d'un nou tipus de teixit sintètic assegura que aquest presenta la mateixa textura que el cotó. Per tal de contrastar aquesta afirmació es demana a 200 persones que entre dues peces de roba, una del nou teixit i l'altra del de cotó, seleccionin el teixit de cotó només pel tacte.

a) Especifiqueu les hipòtesis nul·la i alternativa adients per contrastar l'afirmació del fabricant.
b) Si 145 persones de les 200 enquestades trien correctament el teixit de cotó, s'acceptarà l'afirmació del fabricant?

Exercici 13. S'ha comprovat que, en la fabricació d'un determinat article, com a mínim un 5% de la producció és defectuosa. Un fabricant assegura que la seva producció presenta menys del 5% d'articles defectuosos. Per contrastar aquesta afirmació s'extreu una mostra aleatòria de 600 articles dels quals 45 presenten algun defecte. Si es treballa al 5% de significació,

a) A quina conclusió s'arribarà?
b) Quina és la proporció mostral que delimita la zona d'acceptació de la hipòtesi del fabricant?

Exercici 14. En un sondeig electoral realitzat en dues ciutats diferents s'han obtingut els resultats mostrals següents:

Ciutat 1: Partit A 300 Partit B 100

Ciutat 2: Partit A 384 Partit B 416

a) Quines hipòtesis nul·la i alternativa plantejaríeu per tal de determinar si hi ha diferències significatives entre la intenció de vot a les dues ciutats?
b) Feu el contrast plantejat a l'apartat anterior i indiqueu a quina conclusió arribeu.

Exercici 15. En una mostra aleatòria de 400 electors d'edats superiors a 21 anys, 200 es van mostrar favorables al candidat X. D'altra banda, en una mostra aleatòria de 100 electors d'edats inferiors a 21 anys, es va observar que 40 estaven també a favor del mateix candidat X. En base a aquests resultats, es pot afirmar que l'edat de l'elector incideix significativament en la proporció d'electors que estan a favor del candidat X? Treballeu al 5% de significació.

Exercici 16. Amb la finalitat de determinar si la velocitat que s'assoleix quan es circula per Barcelona és independent de l'hora en què es realitza el trajecte, s'ha efectuat el seguiment d'una mostra de cotxes en dues franges horàries i s'han obtingut els següents resultats, relatius als quilòmetres que havien fet durant una hora:

Entre 4 i 7 h.: 52, 49, 51, 48
Entre 10 i 13 h.: 30, 30, 31, 34, 35

Contrasteu la hipòtesi que la velocitat que es pot assolir és independent de l'hora de la circulació, si suposem subpoblacions Normals amb igual variància i mostres independents.

Exercici 17. Els alumnes de primer curs d'una determinada escola van estar repartits de manera aleatòria en tres grups. A cada grup se li va ensenyar matemàtiques amb un mètode diferent. Al final del curs, tots els alumnes van ser sotmesos a un mateix examen i, aleatòriament, es van seleccionar algunes de les puntuacions obtingudes pels alumnes de cadascun dels tres grups. Els resultats van ser:

Mètode I: $n=5$ $\Sigma X_i = 116$ $\Sigma X_i^2 = 2728$
Mètode II: $n=5$ $\Sigma X_i = 128$ $\Sigma X_i^2 = 3294$
Mètode III: $n=7$ $\Sigma X_i = 171$ $\Sigma X_i^2 = 4193$

Sota el supòsit de distribucions Normals, variàncies poblacionals iguals i mostres independents, es pot afirmar que el resultat obtingut a l'examen depèn del mètode d'ensenyament?

Exercici 18. Una companyia d'assegurances vol determinar si existeixen diferències significatives en el nombre mitjà de dies d'estada, dels pacients que presenten una mateixa malaltia, a tres hospitals diferents. Amb aquest propòsit, s'han obtingut les tres mostres aleatòries següents:

Hospital 1: 18 16 16 20 20 18
Hospital 2: 21 18 19 21 22 19
Hospital 3: 17 18 16

Si les poblacions són Normals amb variàncies iguals i les mostres són independents, contrasteu la hipòtesi plantejada per la companyia d'assegurances.

Exercici 19. Una màquina està preparada per produir dos tipus de cargols idèntics però de diferent diàmetre segons es programi el seu funcionament. El

diàmetre d'aquests cargols és aleatori amb distribució Normal. S'ha comprovat que el de tipus A presenta un diàmetre mitjà de 12 mm i una desviació estàndard de 4 mm; el de tipus B té la mateixa desviació però el seu diàmetre mitjà és de 15 mm. Durant els darrers dies es creia que la màquina produïa cargols del tipus B, però l'encarregat no n'està segur. Per tal de comprovar-ho tria una mostra de 25 cargols i decideix considerar que la producció de cargols és del tipus A només si el diàmetre mitjà de la mostra és inferior a 13 mm.

a) Quin és el nivell de significació del contrast anterior?
b) Quina és la potència del contrast?
c) Si la mitjana mostral ha resultat 14,2, quin és el P-value (valor crític P)?

Exercici 20. Un anunci afirma que un determinat producte per eliminar taques és efectiu, com a mínim, en el 90% dels casos. Un grup d'usuaris, en canvi, afirma que la informació de l'anunci és falsa. Per determinar si els usuaris tenen raó, una associació de consumidors prova el producte en una mostra aleatòria de grandària 200 i obté que el producte és efectiu en 175 casos.

a) Calculeu el P-value (Valor crític P).
b) Indiqueu a quina conclusió s'arribarà si el contrast es realitza al 5% de significació.
c) Determineu la proporció mostral que delimita la regió crítica al 5% de significació.

CAPÍTOL VI. CONTRASTOS NO PARAMÈTRICS

6.1 INTRODUCCIÓ

Els contrastos que es presenten en aquest capítol s'anomenen ***contrastos no paramètrics*** perquè les hipòtesis contrastades no fan referència a cap paràmetre poblacional.

Els anomenats contrastos de la ***bondat d'ajust*** permeten contrastar la hipòtesi que una mostra procedeix d'una determinada població estadística comparant la informació mostral, recollida en una distribució de freqüències, amb la distribució teòrica postulada per la hipòtesi nul·la, sense cap més condició que aquesta estigui totalment especificada.

Dintre dels contrastos ***d'homogeneïtat*** cal distingir-ne dues classes: d'una banda els equivalents als contrastos d'igualtat de mitjanes poblacionals, que s'utilitzen en els casos en què no és sostenible la hipòtesi de normalitat de les poblacions, o bé quan les mostres no són independents essent les poblacions normals o no. D'altra banda, els contrastos equivalents a l'anàlisi de la variància per aquelles situacions en què s'incompleixen alguns dels supòsits bàsics força restrictius del contrast paramètric.

Tanmateix, s'ha d'advertir que, en general, aquests contrastos d'homogeneïtat són menys potents que els paramètrics, fet que fa aconsellable que davant la possibilitat d'aplicació de l'un o de l'altre sigui preferible la realització del paramètric.

6.2 CONTRAST DE BONDAT D'AJUST

Els ***contrastos de bondat d'ajust*** o d'adherència s'utilitzen per provar la hipòtesi que una mostra procedeix d'una determinada població estadística; és a dir, tenen per objectiu contrastar si una distribució empírica de freqüències s'ajusta a un determinat model teòric de probabilitat que es postula a la hipòtesi nul·la.

Aquestes proves es basen en comparar els resultats de la mostra amb aquells que s'espera observar si la hipòtesi nul·la és correcta. Aquests contrastos es poden aplicar, per exemple, per comprovar si el supòsit de normalitat és acceptable abans de realitzar els contrastos paramètrics.

6.2.1 CONTRAST KHI AL QUADRAT

La ***prova de bondat de l'ajust khi quadrat*** permet contrastar si la informació mostral prové d'una determinada població estadística. Es pot aplicar a qualsevol tipus de dades, quantitatives o qualitatives, sempre que la grandària de la mostra sigui mitjanament elevada.

La hipòtesi nul·la d'aquest contrast recull un model de probabilitat teòric especificat de forma concreta (funció de probabilitat i paràmetres) i postula que la mostra prové d'aquesta població. La hipòtesi alternativa s'expressa simplement com *'la hipòtesi nul·la no és certa'* i, per tant, no necessita cap especificació.

$$H_0: F(X)=F_0(X)$$

$$H_1: F(X)\neq F_0(X)$$

Per calcular l'estadístic de prova del contrast, s'han d'agrupar les observacions mostrals en k classes o categories i l'estadístic recull les discrepàncies existents entre les freqüències empíriques i les teòriques per a les categories anteriors.

L'estadístic de prova es defineix com:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

on:

O_i és la freqüència observada en la mostra a la i-èsima classe,

E_i és la freqüència esperada (teòrica) a la classe i-èsima si la hipòtesi nul·la és certa,

k és el nombre de categories o classes que presenta la variable.

L'estadístic pren un valor petit quan les freqüències observades difereixen poc de les esperades i, en aquest cas, es pot suposar que la distribució postulada a la hipòtesi nul·la correspon a la població que ha generat la mostra. Pel contrari, com més gran sigui el valor de l'estadístic més versemblant serà que la mostra no pertanyi a la població teòrica postulada i, en conseqüència, es rebutjarà la hipòtesi nul·la.

Fixat un nivell de significació α, el criteri de decisió serà rebutjar la hipòtesi nul·la si el valor de l'estadístic χ^2 és superior al valor crític corresponent a la distribució khi quadrat amb k-1 graus de llibertat, χ^2_α, ja que l'estadístic es distribueix aproximadament segons una khi quadrat amb k-1 graus de llibertat per a mostres mitjanament grans i amb més de dues categories. Així doncs, la

regió d'acceptació de la H_0 és la formada pels valors de l'estadístic de prova inferiors a χ^2_α i la regió crítica o de rebuig de la H_0 és la formada pels valors de l'estadístic superiors al valor crític, $\chi^2 > \chi^2_\alpha$.

Hipòtesis:

H_0: $F(X)=F_0(X)$

H_1: $F(X)\neq F_0(X)$

Estadístic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

Criteri de decisió:

Si $\chi^2 > \chi^2_\alpha$ es rebutja la H_0.

Per aplicar la prova Khi quadrat és necessari que:

- Les freqüències teòriques per a cadascuna de les categories o classes siguin iguals o superiors a 5. Quan la freqüència teòrica corresponent a una classe és inferior a 5 serà necessari agrupar classes adjacents perquè les resultants tinguin almenys una freqüència teòrica igual a 5.
- El nombre de classes o categories, k, ha de ser superior a 2.

En el cas que per concretar la distribució postulada a la hipòtesi nul·la sigui necessari estimar els seus paràmetres a partir de la mostra, l'estadístic de prova presentarà k-r-1 graus de llibertat, on r és el nombre de paràmetres a estimar.

Exemple 6.1

Es vol contrastar si el nombre de persones que entren per minut en uns grans magatzems segueix una llei de Poisson. S'ha observat aleatòriament l'entrada i s'han obtingut els següents resultats:

Persones	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Freqüència	*4*	*8*	*24*	*40*	*12*	*0*	*8*	*4*

A quina conclusió s'arribarà si es treballa amb un 5% de significació?

Solució:

Per especificar totalment la distribució Poisson postulada a la hipòtesi nul·la haurem d'estimar el paràmetre λ desconegut. Com que l'estimador màxim versemblant és $\hat{\lambda} = \overline{X}$, en aquest cas l'estimació serà $\hat{\lambda}$=3 i, per tant, les hipòtesis són:

H_0: X~Pois(λ=3)

H_1: X no segueix una Pois(λ=3)

Calculem l'estadístic de prova:

X_i	O_i	E_i *	X_i **	O_i	E_i	$(O_i-E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
0	4	4,98	0	4	4,98	0,9604	0,1928
1	8	14,94	1	8	14,94	48,1636	3,2238
2	24	22,40	2	24	22,40	2,5600	0,1143
3	40	22,40	3	40	22,40	309,7600	13,8286
4	12	16,80	4	12	16,80	23,0400	1,3714
5	0	10,08	5	0	10,08	101,6064	10,0800
6	8	5,04	més de 5	12	8,40	12,9600	1,5428
7	4	2,16					
8	0	0,81					
més de 8	0	0,39					
Total	100	100		100	100		**30,3537**

* E_i = n $P(x_i)$ on $P(x_i) = e^{-\lambda}\ \lambda^{xi}/x_i!$, és a dir, E_0=P(X=0)100 = 0,0498·100; E_1=P(X=1)100=0,1494·100, etc.

** Agrupem les classes amb freqüència teòrica inferior a 5 (considerem que el 4,98 és aproximadament 5 i no l'agrupem).

χ^2=30,3537 i per a α= 0,05 i k-r-1=7-1-1=5 χ^2_α=11,07. Per tant, rebutgem la hipòtesi nul·la, no podem acceptar que la mostra hagi estat generada per una Pois(3).

6.2.2 CONTRAST DE KOLMOGOROV-SMIRNOV

Una altra prova de bondat de l'ajust és la de ***Kolmogorov-Smirnov*** que resulta adequada quan la variable és contínua o quan la grandària de la mostra és petita. Aquesta prova no es pot aplicar quan les observacions són nominals ja que és necessària l'ordenació de les observacions.

La hipòtesi nul·la recull un model de probabilitat teòric especificat de forma concreta (funció de probabilitat i paràmetres). La hipòtesi alternativa és, simplement, *'la hipòtesi nul·la no és certa'*.

H_0: $F(X)=F_0(X)$

H_1: $F(X)\neq F_0(X)$

El procediment del contrast es basa en comparar les freqüències relatives acumulades mostrals amb les corresponents a la funció de distribució de la població plantejada a la hipòtesi nul·la. Si hi ha una diferència suficientment gran entre aquestes freqüències serà difícil sostenir que la mostra prové d'aquesta població i, per tant, es rebutjarà la H_0. Pel contrari, quan les diferències prenguin un valor petit no hi haurà evidència empírica suficient per rebutjar la H_0 i conclourem que la mostra prové de la població especificada.

L'estadístic de prova Kolmogorov-Smirnov es defineix com:

$$D_n = \text{màx}\left|F_n(x_i) - F_0(x_i)\right|$$

on:

$F_n(x_i)$ és la freqüència relativa acumulada de x_i a la mostra,

$F_0(x_i)$ és el valor de la funció de distribució teòrica en x_i.

L'estadístic D_n recull la màxima discrepància en valor absolut entre les freqüències acumulades observades i les teòriques, i presenta una distribució de probabilitat independent del model postulat a la H_0 (la seva funció de distribució només depèn de la grandària mostral).

Fixat un nivell de significació α, el criteri de decisió serà rebutjar la hipòtesi nul·la si el valor de l'estadístic D_n és superior al valor crític D_α que es troba a les taules estadístiques de Kolmogorov-Smirnov. La regió d'acceptació de la H_0 està formada pels valors de l'estadístic de prova inferiors al valor crític, $D_n<D_\alpha$, i la regió crítica o de rebuig està formada pels valors de l'estadístic superiors al valor crític.

Hipòtesis:

H_0: $F(X)=F_0(X)$

H_1: $F(X)\neq F_0(X)$

Estadístic:

$$D_n = \max\left|F_n(x_i) - F_0(x_i)\right|$$

Criteri de decisió:

Si $D_n > D_\alpha$ es rebutja la H_0.

Exemple 6.2

Es vol comprovar si l'import (en €) de la compra en un supermercat segueix una distribució N(14, 3). D'una mostra de 25 clients s'han obtingut els següents resultats:

Import de la compra	*6-8*	*8-10*	*10-12*	*12-14*	*14-16*	*16-18*	*18-20*	*20-22*
Freqüència	*2*	*5*	*7*	*3*	*5*	*2*	*0*	*1*

Si es treballa amb un 5% de significació, a quina conclusió s'arribarà?

Solució:

H_0: X~ N(14, 3)

H_1: X no segueix una N(14, 3).

Determinem la distribució de freqüències acumulades empíriques i teòriques, i calculem l'estadístic de prova:

X_i	n_i	F_n*	F_0**	$D_i = \lvert F_n(x_i) - F_0(x_i) \rvert$
Menys de 6	0	0	0,0038	0,0038
6-8	2	0,08	0,02275	0,05725
8-10	5	0,28	0,09176	0,18824
10-12	7	0,56	0,25143	**0,30857**
12-14	3	0,68	0,5	0,1800
14-16	5	0,88	0,74857	0,13143
16-18	2	0,96	0,90824	0,05176
18-20	0	0,96	0,97725	0,01725
20-22	1	1	0,99621	0,00379
Més de 22	0	1	1	0
Total	25			

* Empíriques: $F_n(8)=2/25=0,08$, $F_n(10)=(2+5)/25=0,28$,
$F_n(12)=(2+5+7)/25=0,56$...

**Teòriques: $F_0(8)=P(X<8)=P(z<\frac{8-14}{3})=P(z<-2)=0,02275$

$F_0(10)=P(X<10)=P(z<\frac{10-14}{3})=P(z<-1,3)=0,09176$

$F_0(12)=P(X<12)= P(z<-0,667)=0,25143$....

D_n=0,30857

Per a α=0,05 i n=25 el valor crític és D_α=0,270. Per tant, rebutgem la H_0, ja que $D_n>D_\alpha$, i no podem acceptar que la distribució poblacional sigui N(14, 3).

6.3 CONTRAST D'HOMOGENEÏTAT PER A DUES MOSTRES

Les proves estadístiques relatives a dues mostres ens permeten establir si existeixen o no diferències a nivell poblacional quan considerem l'existència d'algun factor diferenciador. Per exemple, ens permetria comprovar si les diferències obtingudes entre les respostes de dos grups de pacients tractats amb medicaments diferents són significatives a nivell poblacional.

Aquests contrastos són equivalents al contrast paramètric de diferència de mitjanes poblacionals que, com hem vist, requereix que les poblacions siguin aproximadament Normals, les mostres independents i les observacions mesurades en escala d'interval. Aquests supòsits, en molts casos, no són admissibles per les característiques del problema concret. Així, per exemple, ens podem trobar casos amb poblacions no Normals o mostres dependents o dades ordinals.

6.3.1 LA PROVA DE SUMA DE RANGS DE WILCOXON

La ***prova de suma de rangs*** proposada per ***Wilcoxon*** s'utilitza per analitzar si hi ha diferències entre les distribucions de dues poblacions a partir de dues mostres *dependents*, on cada element d'una de les mostres estigui aparellat amb un element de l'altra i que els components de cada parella s'assemblin tant com es pugui, pel que fa a un seguit de característiques que es consideren rellevants.

Per exemple, si es vol estudiar l'eficàcia d'un nou medicament, l'anàlisi dels resultats serà més objectiva si es trien dos grups de pacients amb individus (composicions de les mostres) similars quant a les característiques que puguin influir en els resultats (edat, gravetat de la malaltia, etc.), de forma que cada element d'una de les mostres sigui similar a un element de l'altra, és a dir, els components de cada parella tinguin les màximes semblances possibles. Fins i tot, si la característica que s'estudia ho permet, es poden aplicar els dos tractaments a una mateixa mostra en diferents períodes de temps. En qualsevol cas, les mostres obtingudes no es poden considerar independents i, per tant, s'incompleix un dels requisits bàsics del contrast paramètric per a la comparació de dues poblacions. Una vegada obtinguts els dos grups s'escollirà a l'atzar el grup al qual s'ha d'administrar el nou tractament i a l'altre se li subministrarà el tradicional.

La hipòtesi nul·la (H_0) postula que no hi ha diferències entre les dues poblacions (les respostes dels dos col·lectius no són significativament diferents). En el nostre exemple, postula que no hi ha diferència entre els dos tractaments. La hipòtesi alternativa (H_1) recull la sospita que hi ha diferències.

Un cop seleccionades les mostres i aplicats els tractaments corresponents, s'analitzen els resultats de cada parella associada d'elements. A aquest efecte, es calculen les diferències d_i entre les respostes dels dos components de cada parella. En el nostre exemple, si l'eficàcia es mesura en nombre de dies que tarda el pacient en recuperar-se, s'analitzaran les diferències de dies per a cada parella de pacients.

S'eliminen les diferències nul·les i la resta s'ordenen de menor a major prescindint dels signes. A continuació se'ls assignen rangs: rang 1 a la primera menor diferència en valor absolut, rang 2 a la segona menor diferència en valor absolut, i així successivament; si es donen dues o més diferències iguals se'ls assigna a totes el rang mitjà dels que els correspondrien si fossin diferents i consecutives.

A continuació s'obté la suma de rangs assignats a les diferències positives i a les negatives per separat[1], $\sum R^+$ o $\sum R^-$. Una diferència considerable entre aquestes dues sumes constituirà una evidència empírica que hi ha diferència entre les dues poblacions, ja que el contrast es basa en el fet que, sota la hipòtesi nul·la és d'esperar que les diferències entre les puntuacions observades a cada parella siguin degudes a factors aleatoris i, per tant, és d'esperar que aproximadament la meitat de les diferències siguin positives i l'altra meitat negatives i, a més, que les diferències positives i negatives amb el mateix valor absolut ocorrin amb la mateixa freqüència.

L'estadístic de prova del contrast de Wilcoxon, T, es defineix com la menor de les sumes dels rangs en valor absolut:

$$T=\text{mín}\{\sum R^+, \sum R^-\}$$

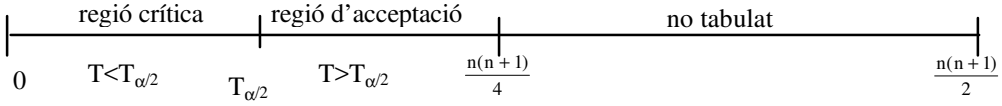
Per a l'aplicació dels criteris de decisió s'ha de tenir en compte si les mostres són petites o grans.

a) Mostres petites

Els valors crítics es troben a la taula de Wilcoxon. Aquesta taula recull els valors de probabilitat corresponents a la cua inferior de l'estadístic de prova. És per això que es pren com a estadístic de prova la menor suma de rangs i la zona d'acceptació de la hipòtesi nul·la, per a un nivell de significació α, està formada

1 Podem comprovar que $\sum R^+ + \sum R^- = \frac{n(n+1)}{2}$ que és la suma dels nombres naturals de l'1 al n.

pels valors superiors al valor crític $T_{\alpha/2}$, que es troba a les taules. Cal tenir en compte que la grandària n indicada a les taules és igual al nombre de diferències no nul·les.



Per tant, els criteris de decisió són no rebutjar la H_0 si $T > T_{\alpha/2}$ (regió d'acceptació) i rebutjar-la si $T < T_{\alpha/2}$ (regió crítica).

També es pot plantejar el contrast de forma direccional o a una cua (en l'exemple anterior es podria formular que el nou tractament és més eficaç que el tradicional). En aquest cas, l'estadístic de prova continua essent la menor suma de rangs, $T=\text{mín}\{\sum R^+, \sum R^-\}$, i la regió d'acceptació de la H_0 està formada pels valors de $T > T_{\alpha}$.

La seva interpretació es fonamenta en el fet que si la ΣR que es pren com a estadístic de prova (+ o -) és suficientment petita com per rebutjar la H_0 és perquè l'altra ΣR (- o +) és suficientment gran per poder concloure que les diferències són significatives.

En el nostre exemple és d'esperar que si sospitem que el nou tractament és més eficaç que el tradicional l'evidència empírica ens proporcioni més diferències negatives que positives, en nombre i en valor absolut; per tant, l'estadístic de prova ΣR^+ resultarà menor que T_{α} i, en conclusió, rebutjarem la H_0 i conclourem que el nou tractament és més eficaç que el tradicional.

Hipòtesis:
H_0: No hi ha diferències entre les dues poblacions.
H_1: Hi ha diferències.
Estadístic:
$T=\text{mín}\{\sum R^+, \sum R^-\}$
Criteri de decisió:
Si $T < T_{\alpha/2}$ es rebutja la H_0.

b) Mostres grans

L'estadístic de prova[2], $T=\text{mín}\{\sum R^+, \sum R^-\}$, té una distribució aproximadament Normal quan el nombre de diferències no nul·les, n, és gran (n > 30) amb els paràmetres següents si la hipòtesi nul·la és certa:

2 T pot ser qualsevol de les sumes de rangs però, per mantenir el mateix criteri que en l'epígraf anterior, prendrem la menor.

$T \sim N(\mu_T, \sigma_T)$

amb $\mu_T = \frac{n(n+1)}{4}$ i $\sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$

Aleshores, l'estadístic de prova serà:

$$Z = \frac{T-\mu_T}{\sigma_T} \sim N(0, 1)$$

I rebutjarem la H_0 quan $Z \leq -z_{\alpha/2}$, en una prova no direccional (a dues cues) i quan $Z \leq -z_\alpha$ si el contrast és a una cua (superior o inferior).

Hipòtesis: H_0: No hi ha diferències entre les dues poblacions. H_1: Hi ha diferències. Estadístic: $Z = \frac{T-\mu_T}{\sigma_T} \sim N(0, 1)$ Criteri de decisió: Si $Z \leq -z_{\alpha/2}$ es rebutja la H_0.

Exemple 6.3

Per comparar el grau de dificultat de dos videojocs amb puntuacions màximes iguals s'han seleccionat 12 nens i s'han observat les puntuacions obtingudes en cada joc. Els resultats han estat:

Nen	Vídeo 1	Vídeo 2
1	94	85
2	78	65
3	89	92
4	62	56
5	49	52
6	78	74
7	75	75
8	80	79
9	82	84
10	62	48
11	53	41
12	79	82

En base a aquesta informació indiqueu si existeixen diferències entre el grau de dificultat dels videojocs, treballant al nivell de significació del 5%.

Solució:

Atès que les mostres són dependents (observeu que les dues mostres estan formades pels mateixos individus), per poder contrastar si hi ha o no hi ha diferències en el grau de dificultat dels dos jocs es farà servir la prova de Wilcoxon.

H_0: el grau de dificultat és similar

H_1: presenten dificultats diferents

Es tracta per tant d'un contrast a dues cues.

En primer lloc, calculem les diferències en valor absolut entre les dues puntuacions obtingudes per cada nen indicant el signe de la diferència.

Tot seguit assignem rang a cadascuna de les diferències, de manera que el rang 1 correspon a la primera menor diferència; el rang 2 a la segona menor; i el rang 11 a la major diferència.

Sumem els rangs positius i els negatius per separat:

$$\Sigma R^+ = 52 \qquad \Sigma R^- = 14$$

Comprovem que $\sum R^+ + \sum R^- = \frac{n(n+1)}{2} \Rightarrow \begin{cases} 52+14=66 \\ \frac{11\cdot 12}{2}=66 \end{cases}$

Nen	Vídeo 1	Vídeo 2	Diferència	Signe	Rang
1	94	85	9	+	8 +
2	78	65	13	+	10 +
3	89	92	3	-	4 -
4	62	56	6	+	7 +
5	49	52	3	-	4 -
6	78	74	4	+	6 +
7	75	75	0		
8	80	79	1	+	1 +
9	82	84	2	-	2 -
10	62	48	14	+	11 +
11	53	41	12	+	9 +
12	79	82	3	-	4 -

L'estadístic de prova és:

$$T = \min \{52, 14\} = 14$$

Per a un 5% de nivell de significació, n=11 i a doble cua obtenim consultant les taules: $T_{0,05/2}$= 11.

Com que, T=14 > $T_{\alpha/2}$=11 $\Rightarrow$ No es pot rebutjar la H_0.

Per tant, no hi ha diferència significativa entre la dificultat que presenten els dos videojocs.

6.3.2 PROVA U DE MANN-WHITNEY

La ***prova U de Mann-Whitney*** permet comprovar si dues mostres *independents* provenen de la mateixa població. Per poder aplicar aquest contrast no cal suposar que les poblacions són Normals, ni que la grandària de les mostres és elevada, només cal que les observacions siguin mesurables, almenys, en una escala ordinal (susceptibles d'ordenació).

Siguin dues poblacions de les quals s'extreuen dues mostres independents de grandàries n_1 i n_2, respectivament. La hipòtesi nul·la H_0 postula que ambdues poblacions presenten la mateixa distribució de probabilitat. La hipòtesi alternativa H_1, que pot ser bilateral o unilateral, estableix que hi ha diferències respecte a la tendència central de les poblacions (es suposa que les distribucions tenen la mateixa forma i dispersió però diferent valor esperat).

Per determinar l'estadístic de prova, el procediment a seguir és el següent:

- Es combinen les observacions de les dues mostres, ja que si és certa la H_0 totes les observacions provenen de la mateixa població, i s'ordenen de menor a major els n_1+n_2 elements.
- S'assignen rangs[3] que aniran des d'1 a n_1+n_2 indicant per a cada rang la mostra a la qual pertany.
- Es sumen els rangs corresponents a les observacions de cadascuna de les mostres per separat que indicarem per R_1 i R_2. Si la H_0 és certa la diferència entre R_1 i R_2 només és conseqüència de l'atzar del mostreig.
- Es calculen els estadístics U_1 i U_2:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

---

3 Si diverses observacions presenten la mateixa puntuació se assignarà a cadascuna el rang mitjà dels seus corresponents rangs.

(Per tal de verificar els càlculs anteriors es pot comprovar que: $U_1+U_2=n_1 n_2$.)

L'estadístic de prova del contrast U de Mann-Whitney es defineix com:

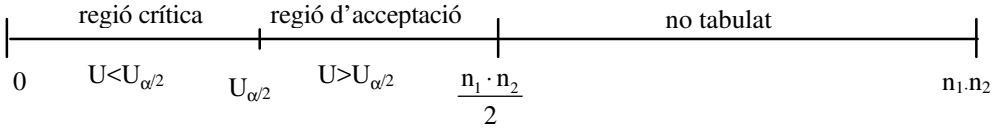
$$U=\text{mín}\{U_1, U_2\}$$

Per a l'aplicació dels criteris de decisió diferenciarem entre mostres petites i mostres grans.

a) Mostres petites

Per definició U_1 i U_2 només poden prendre valors compresos entre 0 i $n_1.n_2$ i sabem que les seves distribucions són simètriques respecte al valor central $\frac{n_1 \cdot n_2}{2}$. Com que l'estadístic és $U=\text{mín}\{U_1,U_2\}$ només està definit per a l'interval $[0, \frac{n_1 \cdot n_2}{2}]$, això permet realitzar la prova tenint en compte només un extrem de la distribució. En concret, l'extrem inferior que és el tabulat a l'annex Taules Estadístiques.

La zona d'acceptació de la hipòtesi nul·la, per a un nivell de significació α, està formada pels valors superiors al punt crític $U_{\alpha/2}$, que es troba a les taules U de Mann-Whitney.



Per tant, els criteris de decisió són no rebutjar la H_0 si $U> U_{\alpha/2}$ (regió d'acceptació) i rebutjar-la en cas contrari ($U< U_{\alpha/2}$ regió crítica).

Podem comprovar que quan $U<U_{\alpha/2}$, la suma de rangs corresponent a la menor U_i és significativament gran respecte a l'altra i, per tant, és difícil sostenir que les dues mostres provenen de la mateixa població.

Si es planteja el contrast de forma direccional (a una cua) l'estadístic de prova és $U=\text{mín}\{U_1, U_2\}$ sigui quina sigui la hipòtesi alternativa. Així doncs, la H_0 es rebutja si $U< U_\alpha$ (regió crítica) i no es rebutja si $U>U_\alpha$ (regió d'acceptació).

Hipòtesis:	
	H_0: No hi ha diferències entre les dues poblacions.
	H_1: Hi ha diferències.
Estadístic:	
	U=mín{U_1, U_2}
Criteri de decisió:	
	Si U< $U_{\alpha/2}$ es rebutja la H_0.

b) Mostres grans

Quan qualsevol de les dues mostres té una grandària superior a 10 i la hipòtesi nul·la és certa, l'estadístic de prova d'aquest contrast, U, es distribueix aproximadament com una normal de paràmetres:

$U \sim N(\mu_U, \sigma_U)$

amb $\mu_U = \frac{n_1 n_2}{2}$ i $\sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$

Aleshores, l'estadístic de prova serà:

$$Z = \frac{U - \mu_U}{\sigma_U} \sim N(0, 1)$$

En aquest cas es podrà escollir U per qualsevol U_i, ja que en ser U_1 i U_2 complementaris s'obtenen resultats equivalents. És a dir, si la distribució de l'estadístic de prova s'aproxima a la normal és indiferent utilitzar U_1 o U_2 per calcular Z, ja que en valor absolut Z serà el mateix en els dos casos. Només el signe i no el valor de Z depèn d'emprar U_1 o U_2.

Si prenem com a estadístic de prova U=mín{U_1, U_2}:

Rebutjarem la H_0 quan $|Z| \geq z_{\alpha/2}$, en una prova bilateral (a dues cues) i quan $Z \leq -z_\alpha$ tant si el contrast és a cua superior com si és a cua inferior.

Hipòtesis:

H_0: No hi ha diferències entre les dues poblacions.

H_1: Hi ha diferències.

Estadístic:

$$Z = \frac{U-\mu_U}{\sigma_U} \sim N(0, 1)$$

Criteri de decisió:

Si $|Z| \geq z_{\alpha/2}$ es rebutja la H_0.

Exemple 6.4

Es vol comprovar si l'opinió sobre els equipaments informàtics d'una escola difereix entre els alumnes dels grups de matí i tarda. La valoració s'efectua amb una escala de 0 (valoració nul·la) a 50 (valoració màxima). Els resultats obtinguts d'una enquesta realitzada a 7 estudiants del matí i 10 estudiants de la tarda són els següents.

Matí	20	25	26	25	29	27	30			
Tarda	18	16	19	22	24	15	27	20	23	30

Realitzeu el contrast adient a un nivell de significació del 5%.

Solució:

Hipòtesis:

H_0: No hi ha diferència en la valoració.

H_1: Hi ha diferència.

En aquest cas el contrast és a dues cues.

Ordenem les dades i assignem rangs.

Dades	Grup	Rang		Dades	Grup	Rang
15	T	1		24	T	9
16	T	2		25	M	10,5
18	T	3		25	M	10,5
19	T	4		26	M	12
20	T	5,5		27	T	13,5
20	M	5,5		27	M	13,5
22	T	7		29	M	15
23	T	8		30	T	16,5

$n_1=7$ $n_2=10$

Sumem els rangs: $R_1=83,5$ (Matí); $R_2=69,5$ (Tarda)

Calculem U_1 i U_2:

$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 7 \cdot 10 + \frac{7 \cdot 8}{2} - 83{,}5 = 14{,}5$

$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 7 \cdot 10 + \frac{10 \cdot 11}{2} - 69{,}5 = 55{,}5$

Comprovem que $U_1+U_2=n_1 n_2 \Rightarrow \begin{cases} 14{,}5 + 55{,}5 = 70 \\ 7 \cdot 10 = 70 \end{cases}$

Prenem com a estadístic de prova U=14,5.

Per a n_1=7, n_2=10 i $\alpha/2$= 0,028, que és el valor més pròxim a 0,025, trobem a la taula U de Mann Whitney el valor crític $U_{\alpha/2}$ =15. Per tant, rebutgem la hipòtesi nul·la ja que $U < U_{\alpha/2}$ (14,5<15).

Si el contrast es realitza al 0,01 de significació el valor crític és $U_{\alpha/2}$ =9 (valor que trobem per a $\alpha/2$=0,005). Per tant, no rebutgem la hipòtesi nul·la.

Exemple 6.5

Un laboratori vol provar que un determinat producte químic afavoreix el creixement de les plantes. Durant un mes es mesura el creixement en mm. que experimenten 12 plantes tractades amb el producte i 9 plantes sense tractar, i s'obtenen les següents observacions:

Sense Tract. (A)	73	87	79	75	82	66	95	75	70			
Amb Tract. (B)	86	81	84	88	90	85	84	92	83	91	53	84

En base a aquestes mesures, a quina conclusió pot arribar el laboratori?

Solució:

Hipòtesis:

H_0: El producte no incideix sobre el creixement.

H_1: El producte afavoreix el creixement.

En aquest cas es tracta d'un contrast a una cua.

n_1=9

n_2=12

Ordenem les dades i assignem rangs.

Dades	53	66	70	73	75	75	79	81	82	83	84
Grup	B	A	A	A	A	A	A	B	A	B	B
Rang	1	2	3	4	5,5	5,5	7	8	9	10	12

Dades	84	84	85	86	87	88	90	91	92	95
Grup	B	B	B	B	A	B	B	B	B	A
Rang	12	12	14	15	16	17	18	19	20	21

Sumen els rangs: R_1=73 (A); R_2=158 (B)

Calculem U_1 i U_2:

$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 9 \cdot 12 + \frac{9 \cdot 10}{2} - 73 = 80$

$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 9 \cdot 12 + \frac{12 \cdot 13}{2} - 158 = 28$

Comprovem que $U_1+U_2=n_1 n_2 \Rightarrow \begin{cases} 80+28=108 \\ 9 \cdot 12 = 108 \end{cases}$

Com que $n_2>10$ podem aproximar per a la distribució normal.

$U \sim N(\mu_U, \sigma_U)$

amb $\mu_U = \frac{n_1 n_2}{2} = 54 \quad \sigma_U^2 = \frac{n_1 n_2 (n_1+n_2+1)}{12} = 198$

Si prenem com a estadístic de prova la menor U_i, U_2 =28, estandarditzant tenim:

$$Z = \frac{U-\mu_U}{\sigma_U} = \frac{28-54}{\sqrt{198}} = -1{,}85$$

Al 5% de significació $-z_\alpha$=-1,64 i $Z<-z_\alpha$. Per tant, rebutgem la H_0. Vegeu que si U_2=28 és suficientment petit com per rebutjar la H_0 és perquè R_2=158 és suficientment gran com per concloure que el producte afavoreix el creixement de les plantes.

6.4 CONTRAST D'HOMOGENEÏTAT PER A MÉS DE DUES MOSTRES

Les proves estadístiques que es presenten a continuació permeten establir si existeixen o no diferències poblacionals en comparar més de dos 'tractaments' o factors diferents. Per tant, són una alternativa a l'anàlisi de la variància simple, però aquí no és necessari que les poblacions siguin aproximadament Normals amb iguals variàncies, ni que les mostres siguin independents, ni que les observacions estiguin mesurades en escala d'interval. És a dir, aquestes proves poden ser utilitzades en condicions molt més generals que l'anàlisi de la variància i no comporten raonaments teòrics tan complexos.

6.4.1 PROVA DE FRIEDMAN

La ***prova de Friedman*** permet comprovar si k mostres *dependents* provenen de la mateixa població, és a dir, si un factor que suposadament subdivideix la població en k grups incideix sobre la mitjana de la població o, pel contrari, les diferències que apareixen entre les mostres no són significatives i, per tant, el factor no genera diferències poblacionals.

H_0: No hi ha diferències entre les k poblacions.

H_1: Almenys una de les poblacions és diferent.

Si disposem les observacions en un quadre de doble entrada que reculli els grups o tractaments per columnes i els elements mostrals o individus per files, el procediment a seguir per determinar l'estadístic de prova és:

- S'assignen rangs (per files) a cadascun dels individus que aniran des d'1 a k.
- Es sumen els rangs (per columnes) corresponents a cada grup per separat i s'obtenen R_1,..., R_k.

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)$$

- Es calcula l'estadístic de prova χ_r^2:

on n és el nombre d'individus i k el nombre de grups.

Per a k> 5 i n>5, si la hipòtesi nul·la és certa, la distribució de l'estadístic χ_r^2s'aproxima a una khi quadrat amb k-1 graus de llibertat i el criteri de decisió és rebutjar la H_0 quan $\chi_r^2 \geq \chi_{1-\alpha}^2$

Hipòtesis:

H_0: No hi ha diferències entre les k poblacions.

H_1: N'hi ha alguna de diferent.

Estadístic:

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)$$

Criteri de decisió:

Si $\chi_r^2 \geq \chi_{1-\alpha}^2$ rebutjarem la H_0.

Exemple 6.6

Es vol comprovar si les vendes a domicili d'un determinat producte depenen de la zona on es realitzen. S'observa aleatòriament el nombre de vendes setmanals obtingudes per tres venedors de l'empresa (1,2,3) en cadascuna de les sis zones considerades (A,B,C,D,E,F) i s'obtenen els següents resultats:

	A	B	C	D	E	F
1	5	7	9	3	2	4
2	2	7	6	6	3	1
3	4	8	3	9	1	3

A quina conclusió arribarem al 5% de significació?

Solució:

Hipòtesis:

H_0: La zona no incideix sobre les vendes.

H_1: La zona incideix sobre les vendes.

Assignem rangs per files i sumem els rangs per grups (columnes):

	A	B	C	D	E	F
1	4	5	6	2	1	3
2	2	6	4,5	4,5	3	1
3	4	5	2,5	6	1	2,5
R_i	10	16	13	12,5	5	6,5

Calculem l'estadístic de prova:

$$\chi_r^2 = \frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2 - 3n(k+1) =$$

$$= \frac{12}{3\cdot 6\cdot 7}\left[10^2 + 16^2 + 13^2 + 12{,}5^2 + 5^2 + 6{,}5^2\right] - 3\cdot 3\cdot 7 = 8{,}2857$$

Criteri de decisió: $\chi^2_{1-\alpha}$= 5,99, $\chi_r^2 \geq \chi^2_{1-\alpha}$. Per tant, acceptem que la zona incideix sobre les vendes.

6.4.2 PROVA DE KRUSKAL-WALLIS

La ***prova H o de Kruskal-Wallis*** permet comprovar si k mostres *independents* provenen de la mateixa població o, pel contrari, si hi ha alguna població que en mitjana es pot considerar diferent.

H_0: No hi ha diferències entre les k poblacions.

H_1: N'hi ha alguna de diferent.

Per determinar l'estadístic de prova, el procediment a seguir és semblant als anteriors:

- Es combinen les observacions dels k grups, ja que si és certa la H_0 totes les observacions provenen de la mateixa població, i s'ordenen de menor a major els $n=n_1+n_2+...+n_k$ elements.
- S'assignen rangs que aniran des d'1 a n i s'identifica el grup al que pertanyen.
- Es sumen els rangs corresponents a les observacions de cada grup per separat i s'obtenen $R_1,..., R_k$. (Es pot comprovar que els resultats són correctes verificant que $\sum_{i=1}^{k} R_i = \frac{n(n+1)}{2}$.)
- Es calcula l'estadístic H:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

- Per a mostres de grandària superior a 5 si la hipòtesi nul·la és certa, la distribució de l'estadístic H s'aproxima a una khi quadrat amb k-1 graus de llibertat. Per tant, el criteri de decisió és:

  Rebutjar la H_0 quan $H \geq \chi^2_{1-\alpha}$

Hipòtesis:

H_0: No hi ha diferències entre les k poblacions.

H_1: N'hi ha alguna de diferent.

Estadístic:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

Criteri de decisió:

Si $H \geq \chi^2_{1-\alpha}$ es rebutja la H_0.

Exemple 6.7

Es vol comprovar que tres marques d'un mateix tipus de màquina presenten produccions similars. S'observa aleatòriament la producció (unitats/hora) obtinguda per les màquines anteriors i s'obté:

Producció (Unitats/hora)						
A	28	30	31	28	27	
B	26	27	34	32	22	36
C	27	28	26	30	29	

A quina conclusió arribarem al 5% de significació?

Solució:

Hipòtesis:

H_0: No hi ha diferències de producció entre les màquines.
H_1: Existeixen diferències entre les màquines.

Assignem rangs combinant totes les observacions:

Rangs						
A	8	11,5	13	8	5	
B	2,5	5	15	14	1	16
C	5	8	2,5	11,5	10	

Sumem els rangs per grups (files): R_A= 45,5 R_B= 53,5 R_C= 37.

Comprovació:

$$\sum_{i=1}^{k} R_i = \frac{n(n+1)}{2} \Rightarrow \begin{cases} 45{,}5 + 53{,}5 + 37 = 136 \\ \dfrac{16 \cdot 17}{2} = 136 \end{cases}$$

Calculem l'estadístic de prova:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{16 \cdot 17} \left[\frac{45{,}5^2}{5} + \frac{53{,}5^2}{6} + \frac{37^2}{5} \right] - 3 \cdot 17 = 0{,}3923$$

Criteri de decisió: $\chi^2_{1-\alpha}$= 5,99, H < $\chi^2_{1-\alpha}$Per tant, no podem rebutjar la hipòtesi nul·la.

6.5. EXERCICIS PROPOSATS

Exercici 1. L'encarregat d'un taller vol determinar si la productivitat es troba equidistribuïda durant els 5 dies de treball de la setmana. Obtinguda una mostra aleatòria de 4 setmanes completes de treball, s'enregistre el següent nombre de peces produïdes:

Dilluns	Dimarts	Dimecres	Dijous	Divendres
49	35	32	39	45

Amb una α = 0,05, hi ha alguna raó per creure que la productivitat no es distribueix uniformement?

Exercici 2. Una societat d'inversió desitja contrastar si la variable "rendibilitat interna d'un títol de renda fixa" es pot considerar que es distribueix normalment amb valor esperat 0,1 i variància 0,0004. Per contrastar-ho es disposa d'una mostra de 100 títols que presenta la distribució de freqüències següent:

Rendibilitat	N. Títols
fins a 0,06	13
0,06-0,09	23
0,09-0,11	27
0,11-0,12	20
0,12-0,14	15
+ de 0,14	2

Realitzeu el contrast adient al 5% de significació.

Exercici 3. A partir de la següent mostra de qualificacions obtingudes en un examen per 100 alumnes triats aleatòriament:

Qualificació	n_i
de 0 a 10	10
de 10 a 20	19
de 20 a 30	34
de 30 a 40	25
de 40 a 50	12

Proveu que aquesta mostra procedeix d'una població Normal.

Exercici 4. Una companyia d'assegurances té contractat un gran nombre de pòlisses d'assegurança obligatòria de vehicles amb persones de menys de 25

anys. La companyia desitja saber si hi ha relació entre l'edat dels conductors i la sinistralitat. Se sap que els conductors menors de 25 anys es classifiquen de la forma següent:

Edat	15-16	17-18	19-20	21-22	23-24
%	10	20	20	25	25

S'obté una mostra aleatòria de 40 comunicats de sinistres a partir de la qual la distribució de freqüències és:

Edat	15-16	17-18	19-20	21-22	23-24
n_i	5	12	10	8	5

Proveu al 5% de significació si hi ha relació entre l'edat dels conductors i la sinistralitat.

Exercici 5. Es vol contrastar que un determinat examen amb 5 preguntes tipus test (5 respostes alternatives amb una única correcta) no ha estat realitzat a l'atzar. Es selecciona una mostra aleatòria de 100 alumnes i s'obté la següent distribució de preguntes correctes:

Preguntes Correctes	N. alumnes
0	19
1	30
2	20
3	20
4	10
5	1

Proveu la hipòtesi al 5% de significació.

Exercici 6. S'ha obtingut una mostra de 100 mestresses de casa a les quals es va proposar que triessin una entre 5 marques de sabó en pols per a rentadores (d'igual qualitat i preu).

Marques	Freqüències
A	19
B	9
C	25
D	30
E	17

Comproveu si les mestresses de casa tenen, en realitat, marques preferents al 5% de nivell de significació.

Exercici 7. Un estudi encaminat a contrastar la hipòtesi que el nombre de naixements de nens i el de nenes no és significativament diferent va obtenir, d'una mostra de 320 famílies amb 3 fills, els resultats següents:

N. nens i nenes	3 nens 0 nenes	2 nens 1 nena	1 nen 2 nenes	0 nens 3 nenes
N. famílies	33	112	125	50

Fent servir el contrast Khi quadrat a quina conclusió s'arriba?

Exercici 8. Es vol saber si el rendiment d'un col·lectiu de treballadors és el mateix quan es treballa amb aire condicionat que quan es treballa sense aire condicionat. Per fer-ho, s'ha triat una mostra aleatòria de 16 treballadors, als quals s'ha fet treballar una setmana sense aire condicionat (A) i una altra setmana amb aire condicionat (B). El nombre mitjà d'unitats produïdes per cada treballador en les dues condicions és la següent:

A	89	103	93	99	81	103	81	103	116	92	100	98	112	98	76	86
B	87	103	93	98	79	101	78	106	115	90	97	98	113	97	78	85

Realitzeu el contrast adient al 5% de significació.

Exercici 9. Es vol contrastar si el consum d'un determinat producte a l'estiu i a l'hivern varia significativament. Es trien 45 famílies a l'atzar i s'observa el nombre d'unitats consumides a l'estiu i a l'hivern per cadascuna. Dels resultats obtinguts s'han calculat:

$\Sigma R+ = 360 \quad \Sigma R- = 630 \quad$ Empats = 1

Realitzeu el contrast de Rangs de Wilcoxon amb $\alpha = 0{,}05$.

Exercici 10. S'han tractat 10 tipus de plantes amb dos productes diferents i s'ha observat els següents creixements mensuals (en mil·límetres):

Tipus	1	2	3	4	5	6	7	8	9	10
Producte 1	21	20	27	30	28	27	31	33	30	15
Producte 2	29	24	25	33	29	25	32	34	33	16

Pot afirmar-se, amb una confiança del 95%, que ambdós productes presenten la mateixa eficàcia?

Exercici 11. Triats a l'atzar 7 pobles de la comarca A i 5 de la comarca B, es van calcular les taxes d'atur següents (en %):

Comarca A	11,3	12,9	10,5	11,4	12,7	12,0	13,5
Comarca B	10,9	11,8	11,2	11,1	11,55		

Comproveu, amb un 99% de confiança, si existeixen diferències significatives entre les taxes d'atur de les comarques, utilitzant el contrast no paramètric adient.

Exercici 12. Es tracta de provar la hipòtesi que dos equips de manteniment triguen el mateix temps en reparar una avaria. Es trien a l'atzar dues mostres independents: la primera de 14 observacions corresponents a l'equip A; la segona de 16 observacions de l'equip B. Ordenades les mostres s'han calculat les sumes dels rangs següents: $\Sigma R_A=173,5$ $\Sigma R_B=291,5$.

Si es fa servir el contrast U de Mann Whitney amb $\alpha = 0,05$, pot afirmar-se que ambdós equips són igual d'eficaços?

Exercici 13. S'han provat dos mètodes d'estudi diferents per ajudar als estudiants a aprovar l'estadística. Per comparar l'efectivitat dels mètodes, s'han seleccionat 8 escoles d'empresarials i 50 alumnes a cada escola. Als alumnes de 4 escoles se'ls va ensenyar amb el mètode A i als de les altres 4 amb el mètode B. Després dels exàmens finals de juny el nombre d'alumnes aprovats en les diferents escoles va ser:

Mètode A	Mètode B
28	33
31	29
27	35
25	30

Proveu si hi ha o no hi ha diferències entre els dos mètode d'estudi.

Exercici 14. Per analitzar si existeixen diferències entre 2 procediments d'obtenció de cert producte, s'observen les produccions diàries obtingudes amb el procediment A i amb el B:

A	26	25	20	25	30	27	29			
B	27	20	18	16	19	23	30	24	22	15

La diferència observada entre aquests procediments és significativa? Utilitzeu el contrast no paramètric adient amb $\alpha = 0,01$.

Exercici 15. Per determinar si existeixen diferències en el preu de venda dels habitatges entre els barris d'un determinat districte de Barcelona s'han observat, de forma aleatòria i independent, la següent mostra:

Barri	Preu de Venda (en desenes de milers de €)				
A	18,5	12,6	15,4	16,4	
B	20,1	13,5	15,8		
C	18,4	20,9	12,2	14,8	17,3

Determineu si existeixen diferències en el preu de venda dels habitatges en els barris anteriors amb un nivell de significació de l'1%.

Exercici 16. Tres crítics literaris (A, B i C) van valorar les novel·les de 8 finalistes a un premi literari. Els resultats de la valoració van ser els següents:

	Crític		
Novel·la	A	B	C
1	8	9	6
2	6	5	4
3	4	7	8
4	5	5	5
5	9	8	7
6	7	4	9
7	3	2	3
8	6	10	8

Per a α=0,05, determineu si les diferències entre les valoracions efectuades pels tres crítics són significatives.

Exercici 17. Es vol determinar si existeixen diferències en el grau d'interès que presenta la programació cinematogràfica de tres cadenes de televisió. Amb aquesta finalitat s'han triat aleatòriament 3 mostres independents d'espectadors que han aplicat les següents puntuacions (de l'1 al 10):

Cadena	Puntuació						
A	5	8	7	3	9	1	7
B	4	5	8	7	6	2	
C	6	7	8	4	9		

Determineu si existeixen diferències en el grau d'interès que presenta la programació cinematogràfica de les cadenes considerades al 5% de significació.

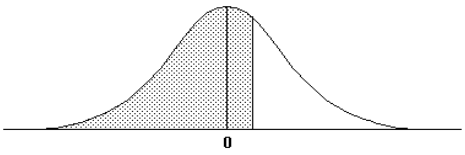
Exercici 18. Un estudi de mercat està encaminat a determinar l'opinió dels consumidors respecte a 4 marques d'un mateix tipus de detergent. Es

seleccionen a l'atzar 10 persones i se'ls demana que provin els 4 detergents i els valorin d'1 a 4 (de menys a més qualitat). Realitzada la prova s'han comptabilitzat les següents puntuacions:

	Puntuacions			
Persona	A	B	C	D
1	1	3	2	4
2	1	4	3	2
3	1	3	4	2
4	2	4	3	1
5	3	1	2	4
6	1	2	3	4
7	2	1	4	3
8	1	2	4	3
9	2	3	4	1
10	1	2	3	4

Determineu si existeix alguna relació entre l'opinió d'aquestes persones i, per tant, si les diferències observades no són estadísticament significatives al 5%.

TAULES ESTADÍSTIQUES: DISTRIBUCIÓ NORMAL TIPIFICADA



$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t/2} \, dt$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

SOLUCIONS

CAPÍTOL I:
PROBABILITAT

1.
a) 3/4
b) 12/48
c)15/36
d) 40/8000

2.
a) 1/30
b) 1/6
c) 0
d) 7/30

3.
a) No
b) No
c) 0,35
d) 0,05

4. 0,393

5. Si

6.
a) 10/34
b) 1
c) 3/17
d) 0
e) 0,2
f) 0,8

7. 0,9988

8. 1/3

9. 5/7

10.
a) 0,355
b) 0,124

11. 12/19

12. 0,082 ; 0,393

13.
a) 0,7
b) 0,3
c) 0,5
d) 19,7

14.
a)

x	1	2	3	4	5
P(x)	5/15	4/15	3/15	2/15	1/15

b)

y	1	2	3	4	5
P(y)	1/15	2/15	3/15	4/15	5/15

c) E(X)=7/3 V(X)=14/9 CV(X)=53,4%
E(Y)=11/3 V(Y)=14/9 CV(Y)=34,0%

15.
a) k=3/2
c) E(X)=0 V(X)=0,6
d) P(X>0,5)=0,4375
e) 10,7368 kg.
f) Y=2X E(Y)=0 V(Y)=2,4

16.
b) 0,55; 0,11; 0,533
c) f(x)=2(1-x) $0 \leq x \leq 1$ i f(x)=0 en altres casos
d) E(X)=1/3 V(X)=1/18
e) Me=0,29289 Mo=0

17. Com a mínim el 64%; 36.000 envasos com a màxim.

18.
a) F(x)=0 $x<0,8$;
F(x)=$-20x+12,5x^2+8$ $0,8 \leq x \leq 1$;
F(x)=$30x-12,5x^2-17$ $1 \leq x \leq 1,2$ i
F(x)=1 $x \geq 1,2$
b) 0,71875
c) E(X)=1 D(X)=0,082
d) 0,966

19.
a) 2/5
b) 0,9875 milions de litres

20.
a) Com a mínim 0,75
b) Com a màxim 0,16
c) 710,26; 1089,73

CAPÍTOL II:
DISTRIBUCIONS DE PROBABILITAT

1.
a) B(4;0,03)
b) B(100;0,05) ≈ Pois(5)
d) G(0,03)
e) BN(5;0,03)
f) H(25;5;10)
g) BN(3;0,2)
h) B(30;0,65)
i) Pois(0,5)
k) BN(3;0,48)
l) Pois(20/15)
m) Pois(45/15)

2.
a) $0,03^4$
b) 0,265
d) 0,4394
e) 161,67
f) 0,385
g) 0,05792
h) 19,5 i 6,825
i) 0,6076
k) 0,53746
l) 0,0465
m) 0,994

3. p = 0,5

4.
a) 0,4967
b) 1,6 i 1,13
c) 1 ràdio

d) 0,0512

5.

a) 0,73
b) 0,4093
c) $0{,}4093^{20}$

6.

a) 6
b) 0,6331
c) 0,99878
d) 0,17

7. 0,0093

8.

a) 0,5043
b) 3/5
c) 0,488

9.

a) 0,09158
b) 0,7619
c) 1.000 i 500

10. 0,06809

11.

a) 0,3937
b) 0,67668
c) 0,0745
d) 0,05265
e) 2,5mn
f) 0,8647

12. Mètode II

13.

a) 0,25
b) 10 i 33,33
c) 0,0143

14. B

15.

a) 60,65%
b) 0,0389

16.

a) 0,525
b) 0,6
c) 0,0853

17. -3,75

18.

a) 0,00248
b) 0,0183
c) 0,6321
d) 0,13535
e) E(X) = 18 V(X) = 18
f) 0,7224

19. A i 0,4527

20.

a) 12mn
b) 0,4346
c) 0,34076
d) 0,09478
e) 2 màquines

21. 0,8825

22.

a.1) 0,3085
a.2) 0,2902
a.3) 0,3446
a.4) 0,1859
b.1) 12,26
b.2) 8,7
b.3) 2,1
b.4) 14,6

23.

a) 0,6179
b) 8,95
c) 230
d) Més de 23,15 punts

24. 1,213 i 1,241 €

25. 0,18141

26.

a) 0,0901
b) 0,1867

27. 6 mesos

28. 0,285

29. 0,0204

30. 0,0197

31.

a) 0,09853
b) 204

32. 0,02743

33. 2,14%

34. 0,9985

35. 0,12302

CAPÍTOL III: INFERÈNCIA ESTADÍSTICA

1.

a) 12,84
b) 7,56
c) 88,38
d) 129,56
e) 2,85
f) 4,56
g) 0,7
h) -2,228
i) -1,708
j) 0,92

2.

a) 0,07
b) 30,19
c) 37,48
d) 74,22
e) -0,7
f) 0,866

3.

a) 0,99

b) 0,01
c) 0,95
d) 0,005
e) 0,05
f) 0,025
g) 0,995
h) 0,05

4. 0,05

5. 3,4 i 0,61

6. 9/25

7. 0,81904

8. 0

9. 0,68268

10. Aproximadament 256

11. μ=1075, σ=125

12. 0,81057

13. Aproximadament 0,12

14. Aproximadament 0,08

15. 0,95818

16. $\ell(1,2,0; \lambda)= e^{-3\lambda} \frac{\lambda^3}{2}$

17. (1;0,1495) (2;0,7158) (3;0,8131) (4;0,51245)

18. $\ell(x_1, x_2, ...,x_n; \theta)= \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$

CAPÍTOL IV:
ESTIMACIÓ DE PARÀMETRES

1.

a) $E(\hat{\mu}_1)=\mu$ $E(\hat{\mu}_2)=2\mu$ $E(\hat{\mu}_3)=\mu$
$V(\hat{\mu}_1)=5/18\sigma^2$ $V(\hat{\mu}_2)=6/5\sigma^2$ $V(\hat{\mu}_3)=1/2\sigma^2$

b) $\hat{\mu}_1$

2. $\hat{\mu}_2$
$EQM(\hat{\mu}_1)=0,52$
$EQM(\hat{\mu}_2)=0,42$
$EQM(\hat{\mu}_3)= 0,4375$

3. $\hat{\mu}_1$: no esbiaixat i consistent;
$\hat{\mu}_2$: esbiaixat i no consistent;
$\hat{\mu}_3$: esbiaixat i no consistent.

4. 0,02

5. Alternatiu: no esbiaixat i no consistent;
Moments: no esbiaixat, consistent i més eficient per a n>2.

6. $\hat{\theta} = \frac{3}{2}\overline{X}$ No esbiaixat i consistent

7.

a) $\hat{\theta} = \frac{-n}{\sum \ln(1-x_i)}$

b) 6,198

8. $\hat{\lambda}$=0,2

9.

a) $\hat{\theta} = 3\overline{X}$

b) 1536

c) Si

10. 8,43

11. $\hat{\beta} = \frac{3}{2}\overline{X}$

12.

a) 11.000, s=2839,44

b) $I_{\mu}^{0,95}$=[8968,9; 13031,1]

c) $I_{\sigma^2}^{0,95}$=[3.815.036,8; 26.874.814,8]

13.

a) $I_{\mu}^{0,9}$=[164; 166]

b) $I_{\sigma^2}^{0,95}$=[18,42; 35,89]

c) Aproximadament 270

14.

a) $I_{\pi}^{0,9}$=[0,025; 0,075]

b) 2401

15. 100

16. $I_{\mu_A-\mu_B}^{0,95}$=[29,5; 47,7]

17. $I_{\pi_1-\pi_2}^{0,9}$=[0,048; 0,218]

18.

a) $I_{\pi}^{0,95}$=[0,38; 0,66]

b) Aproximadament 5%

19. $I_{\mu_2-\mu_1}^{0,95}$=[-3,68; 15,68]

20. 9604

CAPÍTOL V:
CONTRAST PARAMÈTRIC

1.

a) Z=-1,686 NRH_0

b) t=-1,34 NRH_0

2.

a) t=-7,58; t=0,69; t=-2,06

b) $I_{\mu}^{0,95}$=[579,1; 640,9]

3.

a) 0,35942

b) 0,64058

c) 269

d) Incrementa aproximadament un 31%

4. χ^2 =5,85 NRH_0

5. χ^2 =64,8 NRH_0 Z=-3,2 RH_0

6.

a) 22,087
b) 0,00914
c) 0,98899 i 0,94179

7.

a) Z=-2,05 NRH_0
b) 0,02018

8. F=1,429 NRH_0

9. Z=1,62 NRH_0

10. F=1,25 NRH_0 t=-1,92 RH_0

11. Z=-3 RH_0

12.

a) $H_0:\pi=0,5$; $H_1:\pi\neq0,5$
b) Z=6,36 RH_0

13.

a) Z=2,81 RH_0
b) p=0,0646

14.

a) $H_0:\pi_1-\pi_2=0$; $H_1:\pi_1-\pi_2\neq0$
b) Z=8,9 RH_0

15. Z=1,79 NRH_0

16. t=12,55 RH_0

17. F=1,44 NRH_0

18. F=4,32 RH_0

19.

a) 0,00621
b) 0,89435
c) 0,15866

20.

a) 0,119
b) NRH_0
c) p=0,86521

CAPÍTOL VI:
CONTRAST NO PARAMÈTRIC

1. χ^2_{k-1}=4,9. Com que 4,9<9,49 NRH_0

2. D = 0,10725. Com que D < D_α =0,136 NRH_0

3. D≈0,018. Com que D < D_α =0,136 NRH_0

4. D=0,175. Com que D < D_α =0,215 NRH_0. χ^2_{k-1}=5,48< $\chi^2_{3,\alpha}$ =7,81 NRH_0

5. χ^2_{k-1}=118,5> $\chi^2_{3,\alpha}$ =7,81 RH_0

6. χ^2_{k-1}=12,8> $\chi^2_{4,\alpha}$ =9,49 RH_0

7. χ^2_{k-1}=4,46< $\chi^2_{3,\alpha}$ =7,81 NRH_0

8. T=23>T_α=17 NRH_0

9. Z=-1,57 NRH_0

10. T=11>T_α=8 NRH_0

11. U=9>$U_{\alpha/2}$=2 NRH_0

12. Z=--1,808 RH_0

13. U=2>$U_{\alpha/2}$=0 NRH_0

14. U=14,5>$U_{\alpha/2}$=9 NRH_0

15. H=0,118< $\chi^2_{2,\alpha}$ =9,21 NRH_0

16. Friedman=0,0625< $\chi^2_{2,\alpha}$ =5,99 NRH_0

17. H=1,108< $\chi^2_{2,\alpha}$ =5,99 NRH_0

18. Friedman=9,48> $\chi^2_{3,\alpha}$ =7,81 RH_0

BIBLIOGRAFIA

ALEA, V. et al. (2011) *Estadística I: Cuestiones tipo test con R Commander.* Barcelona. Edicions UB. Textos docents 368

ALEA, V. et al. (2014) *Guía para el análisis estadístico con R Commander.* Barcelona. Edicions UB. Textos docents 391

FREEDMAN, D., et al. (1993) *Estadística*. Barcelona. A.Bosch Ed.

FREIXA, M., et al. (1992) *Análisis exploratorio de datos: Nuevas técnicas estadísticas*. Barcelona. PPU.

HILDEBRAND, D.K. i OTT, R.L: (1997) *Estadística Aplicada a la Administración y a la Economía.* Madrid, Addison-Wesley Iberoamericana.

LARSON, H.J. (1988) *Introducción a la Teoría de Probabilidades e Inferencia Estadística*. México. Limusa.

LIND, D.A. et al. (2015) *Estadística aplicada a los negocios y a la economía.* México. McGraw-Hill.

MARTÍN-GUZMÁN, P. et al. (2006) *Manual de Estadística descriptiva*. Navarra. Thomson.

NEWBOLD, P. (2008) *Estadística para administración y la economía.* Madrid. Prentice Hall.

PEÑA, D. i ROMO, J. (1997) *Introducción a la Estadística para las ciencias sociales.* Madrid, McGraw-Hill/Interamericana de España.