4. Introducció a l'estadística inferencial

Recordem que quan es fa una investigació, les poblacions estadístiques resulten sovint massa àmplies per poder-les estudiar. Una opció és recollir dades només d'una part —una mostra— de la població, però és possible que els resultats obtinguts de l'estudi d'aquesta part no siguin aplicables al conjunt de la població. Aquest problema de no poder abastar tota la població que es vol estudiar és comú en totes les ciències socials i humanes. La solució és utilitzar un conjunt de tècniques que permeten fer inferències sobre una població completa basant-se en l'estudi de dades d'una mostra d'aquesta població. Aquest conjunt de tècniques és el que es coneix com a *estadística inferencial.*

Seguint Welkowitz, Ewen i Cohen (1986), els *objectius de l'estadística inferencial* són:

- Estimar el valor de la mitjana i la variància de la població a partir de les dades d'una mostra de la qual es coneix la mitjana i la variància. Qualsevol estadístic calculat sobre una sola mostra, com la mitjana o la desviació típica ($\bar{X}$ o S), per produir una estimació del paràmetre poblacional corresponent (μ o σ), s'anomena *estimador puntual*.

- Buscar un *interval de confiança*, que és un conjunt de valors amb una probabilitat coneguda de contenir el veritable valor del paràmetre poblacional, ja que a vegades l'estimador puntual és insuficient perquè no ens indica fins a quin punt s'acosta al paràmetre poblacional real.

- Avaluar la probabilitat d'obtenir certs tipus de resultats mostrals quan es compleixen determinades condicions poblacionals. És a dir, conegudes les característiques d'una població (μ o σ), estimar la probabilitat que una mostra extreta d'aquesta població tingui unes determinades característiques ($\bar{X}$ o S).

4.1. Estratègia general de l'estadística inferencial

Per explicar quina és l'estratègia que segueix l'estadística inferencial, posarem un exemple que podeu trobar a Welkowitz, Ewen i Cohen (1986:108-112).

Imaginem-nos que dos amics —en Joan i la Fina—juguen a fer apostes amb una moneda. En Joan aposta 6 euros al resultat de llançar una moneda i la Fina demana l'opció "cara". Llancen la moneda a l'aire i, efectivament, surt cara, de manera que en Joan perd els 6 euros que havia apostat. Decideix que vol recuperar els euros i proposa a la Fina tornar a jugar. La Fina segueix demanant cara i segueix sortint cara en el llançament següent. Ho proven sis vegades seguides i cada vegada surt cara, que és l'opció que demana la Fina.

Després de sis llançaments, en Joan comença a sospitar. La Fina sempre demana cara i sempre li surt el mateix. Si la moneda fos correcta, podria continuar jugant i, a la llarga, aniria recuperant els diners perduts (que ja són 36 euros, encara que, per al que volem explicar, això és el menys important). Però en Joan pensa que si la Fina està fent trampes i la moneda està trucada, hauria de plantar-se i dir a la Fina que és una tramposa.

Per tant, arribats a aquest punt del joc, tenim dues hipòtesis possibles:

La hipòtesi 0: els resultats són deguts a la sort, a l'atzar. La moneda és correcta i l'únic que passa és que la Fina té una bona ratxa.

La hipòtesi 1: la moneda està trucada i té tendència que surti cara. En aquest cas, doncs, els resultats no serien deguts a l'atzar sinó a un factor (la moneda té truc) que fa que els resultats difereixin dels que serien esperables.

A la taula següent podem veure un esquema de la situació en què es troba en Joan, pel que fa a les dues possibles decisions que pot prendre.

		Situació real o estat de la població	
		La hipòtesi 0 és certa: els resultats són deguts a la sort	**La hipòtesi 1 és certa: la moneda està trucada**
Decisió d'en Joan	No rebutja la hipòtesi 0 (segueix jugant)	Decisió correcta	Error: continua donant diners a una amiga que fa trampes
	Rebutja la hipòtesi 0 (abandona el joc i acusa la Fina de fer trampes)	Error: calúmnia injusta a una amiga honesta	Decisió correcta

La situació real o *estat de la població* no es pot conèixer només amb els sis llançaments. Una possibilitat seria observar la població completa. El problema és que la població, en aquest cas, és el conjunt de tots els llançaments de moneda que es poden fer, és a dir, observar infinits llançaments. Si la moneda és correcta, trobaríem que en la meitat dels llançaments surt cara i en l'altra meitat surt creu. Però resulta impossible, en aquest cas, incloure tota la població.

Una altra possibilitat seria observar una mostra de molts llançaments, 100 o 500. Això permetria tenir una mostra prou gran per observar el comportament de la moneda. No obstant això, imaginem-nos que en Joan ha de prendre una decisió a partir d'aquests sis llançaments de la moneda. Per això diem que hi ha dues possibles hipòtesis sobre quin pot ser l'estat de la població. En Joan ha de decidir quina de les dues hipòtesis accepta, és a dir, ha de prendre una decisió sense saber l'estat real de la població.

Per resoldre aquest problema s'han d'explicar —o recordar— algunes definicions i càlculs bàsics sobre la probabilitat d'un succés.

La probabilitat

La probabilitat d'un succés donat es defineix amb la fracció següent:

$$P(\text{succés}) = \frac{\text{Nombre de casos en què el succés pot ocórrer (o nombre de casos favorables)}}{\text{Nombre total de successos (o casos) possibles.}}$$

Per exemple, la probabilitat que surti cara en llançar una moneda a l'aire és la següent:

$$P\ (\text{cara}) = \frac{\text{Nombre de casos en què pot sortir cara en un llançament}}{\text{Nombre de successos possibles en un llançament}} = \frac{1\ (\text{cara})}{2\ (\text{cara o creu})} = 0{,}5$$

Per tant, la probabilitat que surti cara és 0,5, és a dir, del 50 %. Per tant, si la moneda és correcta, sortirà cara la meitat de les vegades i creu l'altra meitat.

Més exemples:

Suposem que tenim un joc de 48 cartes, de les quals 12 són d'or, 12 de bastons, 12 de copes i 12 d'espases.

- Quina és la probabilitat que agafant una carta a l'atzar surti una copa?

$$P\ (\text{copes}) = \frac{12\ \text{copes}}{48\ \text{cartes}} = \frac{1}{4} = 0{,}25$$

Passat a percentatge, podem dir que tenim el 25 % de probabilitat que surti copa.

- Quina és la probabilitat que surti l'as d'or?

$$P\ (\text{as d'or}) = \frac{1}{48} = 0{,}021$$

Tenim un 2,1 % de probabilitat que, agafant una carta del joc a l'atzar, la que ens surti sigui l'as d'oros.

- Quina és la probabilitat que agafant una carta a l'atzar surti un rei?

$$P\ (\text{rei}) = \frac{4}{48} = \frac{1}{12} = 0{,}083$$

Tenim, doncs, una probabilitat del 8,3 % que en extreure una carta del joc a l'atzar ens surti un rei.

La probabilitat no pot ser més petita que 0 ni més gran que 1. Si un succés té probabilitat 0 vol dir que no és possible que ocorri; si té probabilitat 1 vol dir que aquest succés segur que passarà, i no en pot passar cap altre.

Fins ara hem parlat de la probabilitat d'un sol succés cada vegada. Ara parlarem de dos successos que puguin passar alhora.

La probabilitat de A o B

La probabilitat que s'esdevingui un dels dos successos es calcula amb la fórmula següent:

$P(A \text{ o } B) = P(A) + P(B) - P\ (A \text{ i } B)$

Per exemple, si volem saber quina és la probabilitat que en extreure a l'atzar una carta del joc surti un rei o una espasa, la probabilitat serà:

P (rei o espasa) = P(rei) + P(espasa) - P (rei d'espases) = 4/48 + 12/48 – 1/48 = 0,31

Vegem per què és així. Quants casos favorables tenim?

- Per una banda, els 4 reis: rei d'or, rei de copes, rei d'espases i rei de bastons.
- Per l'altra, tenim 12 espases: l'as, el dos, el tres, el quatre, el cinc, el sis, el set, el vuit, el nou, el deu, el cavall i el rei.
- Sembla, doncs, que hi ha 16 casos favorables: 4 reis i 12 espases. No obstant això, fixem-nos que el rei d'espases l'hem comptat dues vegades, una com a rei i l'altra com a espasa. Per tant, per calcular la probabilitat real, a la suma dels dos successos hem de restar-hi la probabilitat dels successos que s'han comptat dues vegades.

Si els successos A i B no poden ocórrer de manera simultània (és a dir, són mútuament excloents), llavors la $P\ (A \text{ i } B)$ és igual a 0. Per exemple, si volem calcular la probabilitat que en treure una carta a l'atzar surti un cavall o un rei, tenim que :

P (cavall o rei) = P (cavall) + P (rei) – P (cavall i rei) = 4/48 +4/48 – 0 = 0,17

La probabilitat que surti alhora un cavall i un rei és 0. Les cartes, o són cavalls o són reis, però no les dues figures al mateix temps.

La probabilitat de A i després B

Suposem que extraiem una carta del joc, la mirem i la tornem al joc. Després barregem les cartes i n'extraiem una altra. Volem saber la probabilitat que en la primera extracció surti un rei i que en la segona extracció surti l'as d'or. Les dues extraccions són independents, és a dir, que tornem la primera carta al joc, de manera que el que hagi passat en la primera extracció no afecta el que pugui passar en la segona. La solució es representa amb la fórmula següent:

P (A i després B) = P (A) · P (B)

En l'exemple, la resposta és:

P (rei i després as d'or) = 4/48 · 1/48 = 4/2304 = 0,0017

Si els dos successos no fossin independents, és a dir, si després de la primera extracció la primera carta no fos retornada al joc, la probabilitat seria:

P (rei i després as d'or) = 4/48 · 1/47 = 4/2256 =0,0018

En aquest cas hem de restar una carta dels casos possibles relatius a l'extracció d'una carta que sigui l'as d'or perquè la primera carta no s'ha retornat, i per tant, al joc només hi queden 47 cartes.

Quina és la probabilitat que en extreure dues cartes a l'atzar la primera sigui un or i la segona un cavall? (Tornem la carta al joc després de la primera extracció.)

P (or i després cavall) = P (or) · P (cavall) = 12/48 · 4/48 = 48/2304 =0,021

Ara tornem als dos amics que estaven jugant a tirar la moneda a l'aire. Havíem dit que en Joan havia de decidir si seguia jugant o no, després de sis llançaments en els quals havia sortit sempre cara. El que pot fer en Joan és calcular la probabilitat que en tirar la moneda surti sis vegades seguides cara.

Es tracta de buscar la probabilitat que ocorrin sis successos independents, és a dir, que surti cara al primer llançament, cara al segon, cara al tercer... i així fins a sis llançaments:

P (cara i després cara i després cara i després cara i després cara i després cara) =

= ½ · ½ · ½ · ½ · ½ · ½ = 1/64 = 0,016

Ara ja sabem que, si la moneda no està trucada, la probabilitat que surti cara sis vegades seguides és de 0,016. Per tant, o bé la moneda és correcta i s'ha esdevingut un succés amb probabilitat 0,016 (hipòtesi 0) o bé la moneda està trucada i la probabilitat que surti cara no és de ½ i, per tant, la probabilitat que surtin 6 cares consecutives no serà del 0,016 (hipòtesi 1). Ara el problema és saber si una probabilitat de 0,016 és prou elevada per acceptar que el fet que hagin sortit 6 cares seguides és fruit de l'atzar, o bé és massa petita i llavors podem sospitar que la moneda està trucada.

Tal com diuen Welkowitz, Ewen i Cohen (1986: 117-118), "el científic segueix l'arbitrària però raonable pràctica de considerar els successos amb probabilitats menors de 0,05 com a 'suficientment improbables per justificar el rebuig de la hipòtesi 0' i els successos amb probabilitats més grans de 0,05 com a 'no suficientment improbables' per justificar aquest rebuig".

Seguint aquesta norma, doncs, haurem de concloure que una probabilitat de 0,016 és massa petita perquè sigui fruit de l'atzar el fet que surtin 6 cares en 6 llançaments consecutius. Per tant, aquest succés és molt improbable i haurem de rebutjar la hipòtesi 0. La millor decisió que pot prendre en Joan davant d'aquest resultat és abandonar el joc i no arriscar-se a seguir perdent. No obstant això, cal fer notar que el que fa en Joan és prendre una decisió a partir de les dades que té. L'estadística li fa veure que el succés és poc probable, però realment ell no sap si la moneda està trucada. En Joan pot estar cometent un error. L'estadística indica el camí cap a la millor decisió sobre la població, a partir dels resultats obtinguts en la mostra.

Fins aquí hem vist l'estratègia que segueix l'estadística inferencial per arribar a decisions sobre la població a partir de les dades d'una mostra. Ho hem vist amb un exemple senzill, que és el llançament d'una moneda. En situacions més complexes, l'estratègia és la mateixa, si bé el càlcul de la probabilitat que ocorri un succés no sol ser tan simple. Ens interessarà buscar, per exemple:

- La probabilitat que sigui degut a l'atzar el fet que la mitjana de les notes d'estadística d'un curs sigui més elevada que la del curs anterior.

- La probabilitat que, aplicant un programa educatiu per millorar les habilitats socials en adolescents, la mitjana d'habilitats d'aquests joves augmenti significativament.

4.2. El model de la corba normal

En el cas de la moneda que hem vist fins ara, és relativament fàcil calcular les probabilitats perquè sabem que a cada llançament la probabilitat que surti cara o creu és de ½. Però què passa quan volem saber quina és la probabilitat que, en extreure una mostra a l'atzar d'una població, ens surti una mitjana determinada?

Una manera de fer-ho —encara que gens pràctica— seria extreure un nombre gran de mostres, totes de la mateixa mida, de la població definida, calcular la mitjana per a cada una i comptar les freqüències amb què surten els valors de cada mitjana. Les mostres han de ser aleatòries i representatives de la població. És a dir, cal que en seleccionar la mostra cada element de la població tingui la mateixa probabilitat de ser escollit i que totes les mostres possibles de la mateixa mida tinguin la mateixa probabilitat de ser escollides. Si calculem la mitjana de cada mostra podríem dibuixar la distribució de les freqüències. Això ens ajudaria a fer-nos una idea de la freqüència amb què podem esperar una mitjana donada.

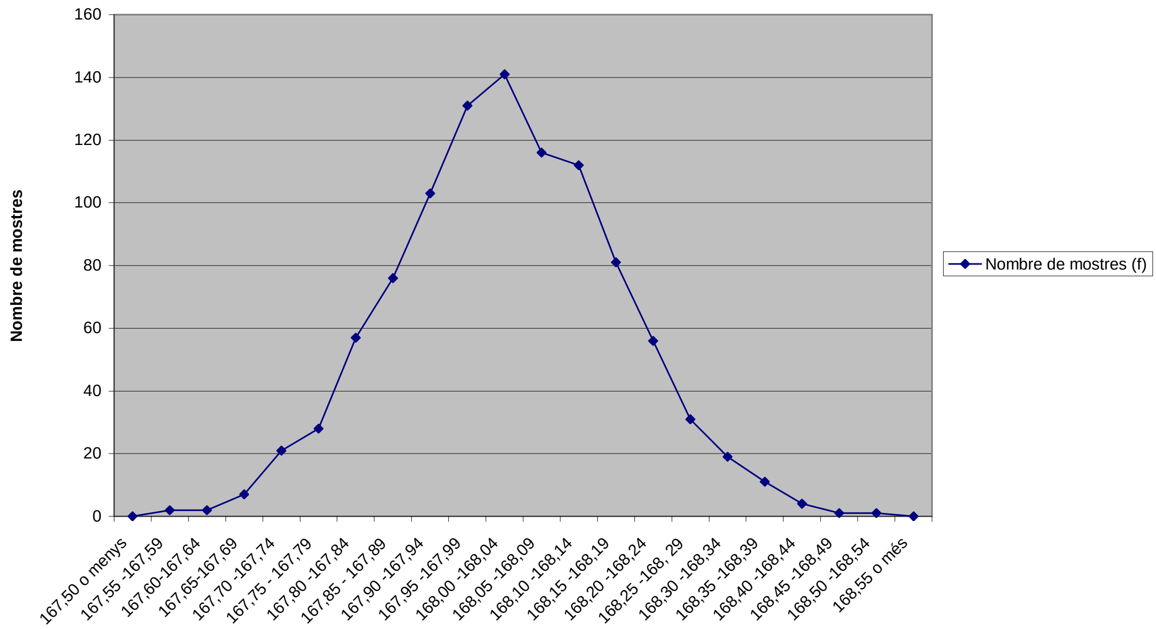
Imaginem-nos que hem buscat 1.000 mostres de la població, cada una de $N = 100$, i hem buscat la mitjana d'alçada de cadascuna. A la taula 4.1 podríem tenir la distribució de freqüències de les alçades mitjanes.

Taula 4.1:Distribució hipotètica de freqüències de les alçades mitjanes en 1.000 mostres de 100 casos cada una[6]

Alçada mitjana (cm)	**Nombre de mostres (f)**	**Proporció de mostres f/N**
167,50 o menys	0	0
167,55 -167,59	2	0,002
167,60-167,64	2	0,002
167,65-167,69	7	0,007
167,70 -167,74	21	0,021
167,75 - 167,79	28	0,028
167,80 -167,84	57	0,057
167,85 - 167,89	76	0,076
167,90 -167,94	103	0,103
167,95 -167,99	131	0,131
168,00 -168,04	141	0,141
168,05 -168,09	116	0,116
168,10 -168,14	112	0,112
168,15 -168,19	81	0,081
168,20 -168,24	56	0,056
168,25 -168, 29	31	0,031
168,30 -168,34	19	0,019
168,35 -168,39	11	0,011
168,40 -168,44	4	0,004
168,45 -168,49	1	0,001
168,50 -168,54	1	0,001
168,55 o més	0	0
Total	1000	1

[6] Aquest exemple s'ha extret de Welkowitz, Ewen i Cohen (1986), pàgina 124.

Gràfic 4.1: Distribució hipotètica de le alçades mitjanes de 1000 mostres de N = 100

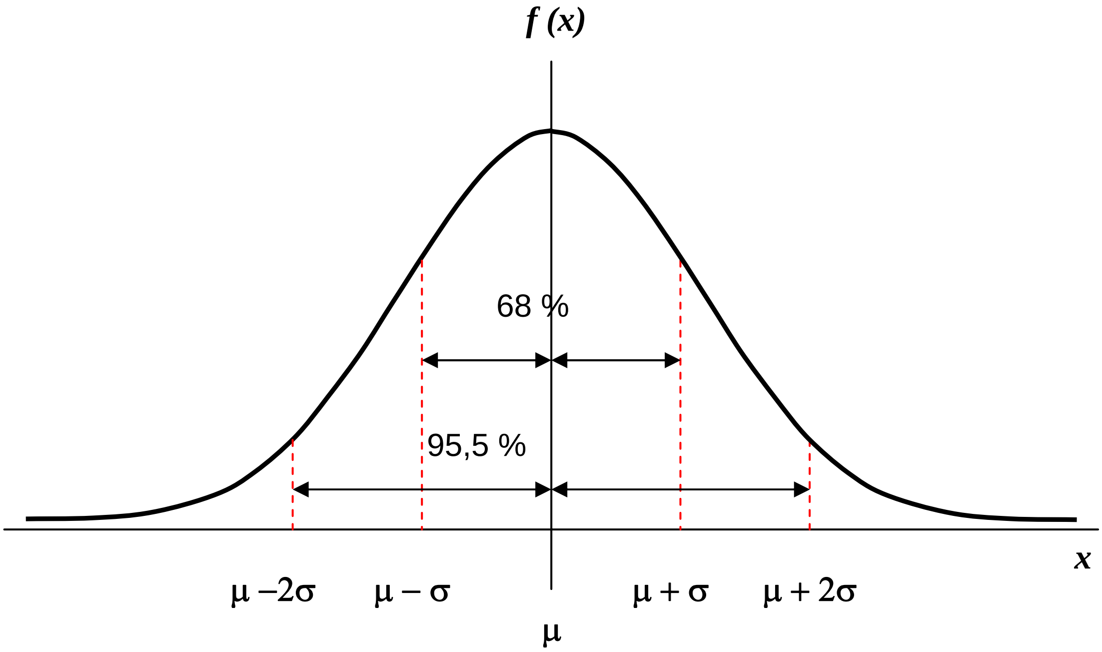


La distribució empírica de mostres serveix com a model estadístic per assignar probabilitats a qualsevol observació d'alçada mitjana. La columna "Proporció de mostres" de la taula 4.1 ens indica la probabilitat que tenim d'obtenir una mostra d'una alçada mitjana determinada. És a dir que, si hem extret de la població 76 mostres amb mitjana d'alçada de 167,85 i 167,89 cm (casos favorables), del total de 1.000 mostres (casos possibles), la probabilitat que qualsevol mostra de mida N = 100 tingui una alçada d'entre 167,85 i 167,89 cm és de 76/1.000 = 0,076. Dit d'una altra manera, si d'aquesta població haguéssim extret moltes mostres de mida N = 100, podríem esperar que aproximadament el 7,6 % d'aquestes mostres tinguessin mitjanes d'entre 167,85 i 167,89 cm. Basant-nos en aquesta distribució, també podríem saber que la probabilitat d'obtenir una mostra amb una mitjana de 168,00 cm o menys és de 0,141 + 0,131 + 0,103 + 0,076 + 0,057 + 0,028 + 0,021+ 0,007 + 0,002 + 0,002 = 0,568, és a dir, del 56,8 %.

De totes maneres, si haguéssim de seguir aquest procediment per cada variable que es vol estudiar seria molt costós. A la pràctica s'utilitzen distribucions mostrals teòriques i no empíriques. És a dir, els matemàtics han pogut mostrar que moltes distribucions mostrals tendeixen a seguir *el model de la corba normal*. És molt freqüent, doncs, que la distribució mostral de les mitjanes segueixi aquest model. Això permet utilitzar el model de la corba normal per calcular probabilitats en molts estudis estadístics. És a dir, no caldrà calcular cap model empíric per calcular probabilitats, sinó que podrem basar-nos en el model de la corba normal que ja ha estat descrit i estudiat pels matemàtics i estadístics. Això ens permetrà comparar els resultats d'una sola mostra amb els del model teòric.

Aquest model de la corba normal ja l'havíem començat a descriure. Com en el cas de qualsevol distribució de freqüències, l'eix horitzontal de la distribució normal representa els valors observats i a l'eix vertical es representen les freqüències observades.

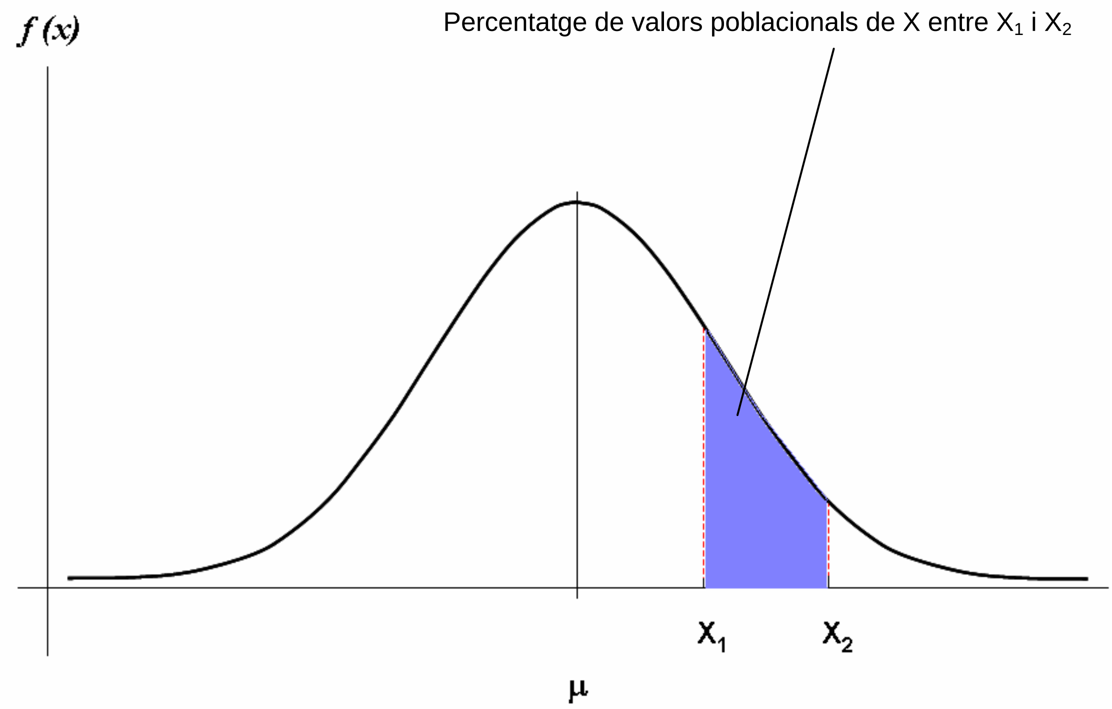
Gràfic 4.2: El model de la corba normal



L'àrea total sota de la corba representa la població total. Aquesta àrea es defineix com el 100 % dels casos de la població.

Suposem que ara ens demanen que trobem la proporció de la població amb puntuacions entre els punts X_1 i X_2 (gràfic 3). La representació gràfica d'aquesta proporció és l'àrea que queda sota la corba i entre aquests dos punts, que equival al percentatge de freqüències d'aquestes puntuacions.

Gràfic 4.3: Àrees sota la corba normal



Els matemàtics calculen les àrees entre dos valors recorrent al càlcul integral, és a dir, fent una integral de l'equació de la corba normal entre aquests dos valors. Per sort, les integrals necessàries ja estan calculades i recollides en una taula, la taula d'àrees sota la corba normal, que trobareu a l'annex 2 d'aquest document. El que cal fer és convertir els valors que volem analitzar en puntuacions típiques *Z*.

A continuació en tenim un fragment que ens servirà per explicar com s'interpreten els valors que surten en aquestes taules. La columna vertical de l'esquerra representa les puntuacions *Z*, expressades amb una xifra decimal. Els valors de dins de la taula representen el percentatge de l'àrea entre la mitjana i el valor *Z*, expressat amb dues xifres decimals.

Per exemple, en la taula 4.2, per trobar el percentatge de casos entre la mitjana i una $Z = 0{,}94$ hem de seleccionar el valor 0,9 de la primera columna i el valor 0,04 de la primera fila. En el punt on es creuin aquests dos valors trobarem el valor de l'àrea, que serà del 32,64 %.

Taula 4.2: Fragment de la taula d'àrees (percentatges) de la corba normal, entre les puntuacions típiques i la mitjana

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	00,00	00,40	00,80	01,20	01,60	01,99	02,39	02,79	03,19	03,59
0,1	03,98	04,38	04,78	05,17	05,57	05,96	06,36	06,75	07,14	07,53
0,2	07,93	08,32	08,71	09,10	09,48	09,87	10,26	10,64	11,03	11,41
0,3	11,79	12,17	12,55	12,93	13,31	13,68	14,06	14,43	14,80	15,17
0,4	15,54	15,91	16,28	16,64	17,00	17,36	17,72	18,08	18,44	18,79
0,5	19,15	19,50	19,85	20,19	20,54	20,88	21,23	21,57	21,90	22,24
0,6	22,57	22,91	23,24	23,57	23,89	24,22	24,54	24,86	25,17	25,49
0,7	25,80	26,11	26,42	26,73	27,04	27,34	27,64	27,94	28,23	28,52
0,8	28,81	29,10	29,39	29,67	29,95	30,23	30,51	30,78	31,06	31,33
0,9	31,59	31,86	32,12	32,38	32,64	32,89	33,15	33,40	33,65	33,89
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

La taula només proporciona els percentatges a un costat de la mitjana, per la qual cosa el percentatge més gran que surt a la taula és el 50 %, és a dir, la meitat de l'àrea sota la corba normal. Però com que sabem que la corba normal és simètrica, és a dir, que l'alçada de la corba per un valor positiu de *Z* és exactament la mateixa pel mateix valor però negatiu, coneixent els valors de la meitat de la corba podem saber totes les àrees que ens interessin.

Cal observar que l'àrea entre *Z* i la mitjana per a un valor de $Z = 1$ és 34,13 %, i que per a $Z = 2$ és 47,72 %. Si recordeu que les *Z* representen desviacions típiques, el percentatge de casos que es troben entre la mitjana i +/- una desviació típica és del 68,26 % i el percentatge de casos que es poden trobar entre la mitjana i +/- dues desviacions típiques és del 95,5 %. Si en lloc de $Z = 2$ agafem $Z = 1{,}96$ veurem que el percentatge de casos entre la mitjana i +/- 1,96 és del 95 %. Podeu veure aquests valors representats al gràfic 4.

Per tant, segons el model de la corba normal, es pot esperar que aproximadament el 95 % de tots els successos possibles se situïn entre la mitjana i +/- 2 desviacions típiques. Per exemple, si suposem que l'alçada de la mitjana dels adults és de 169 cm i la desviació típica és de 6, es pot esperar que el 95 % dels adults tinguin alçades entre

$$169 \pm 2 \cdot 6 = 169 \pm 12$$

és a dir, cal esperar que tinguin alçades entre 157 cm i 181 cm. Només un 2,28 % tindrien alçades superiors a 181 cm i el mateix percentatge —un 2,28 %— tindrien alçades inferiors a 157cm.

Gràfic 4.4: Característiques de la corba normal