

PROBABILITAT I ESTADÍSTICA PER A CIÈNCIES I

Toni Monleón-Getino
Clara Rodríguez Casado

Departament d'Estadística

PROBABILITAT I ESTADÍSTICA PER A CIÈNCIES I

Toni Monleón-Getino
Clara Rodríguez Casado

Departament d'Estadística

Als professors Marta Cubedo, Toni Miñarro i Martín Ríos del Departament d'Estadística de la UB, per revisar el llibre, i a tots els professors del Departament d'Estadística, especialment a Carles Cuadras i Pedro Sánchez, els meus mestres.

Índex

INTRODUCCIÓ	11
1. ESTADÍSTICA DESCRIPTIVA	13
1.1. Tipus de variables	13
1.2. Taules de freqüències	13
1.3. Gràfics de variables qualitatives	14
1.4. Gràfics de variables quantitatives	15
1.5. Nombres d'interval·ls en un histograma. La marca de classe. La regla de Sturges	17
1.6. Tipus d'estadístics	19
1.7. Mesures de posició: percentils i quartils	20
1.8. Mesures de centralitat: mitjana	21
1.9. Mesures de centralitat: mediana	21
1.10. Mesures de centralitat: moda	22
1.11. Variabilitat o dispersió	23
1.12. Mesures de dispersió: rangs	23
1.13. El coeficient de variació (CV): comparar dispersions a escales diferents	25
1.14. Mesures de dispersió: variància i desviació estàndard	25
1.15. Mesures de forma: asimetria	26
1.16. Mesures de forma: curtosi	27
1.17. Problemes proposats	28
Problema 1	28
Problema 2	29
Problema 3	29
Problema 4	29
Problema 5	29
Problema 6	31
Problema 7	31
2. PROBABILITAT I COMBINATÒRIA	33
2.1. Fenòmens aleatoris i esdeveniments	33
Fenòmens aleatoris	33
Esdeveniments i espai mostral	34
Esdeveniments d'un fenomen aleatori	35
Esdeveniment contrari	35
Unió i intersecció d'esdeveniments	36
2.2. Definició i càlcul de la probabilitat	37
Funció de probabilitat	37
Regla de Laplace	38
Definició freqüentista de la probabilitat	39

Definició axiomàtica de Kolmogorov.....	40
Propietats de la probabilitat	40
Exemple: la contaminació atmosfèrica	42
2.3. Probabilitat condicionada i independència estocàstica	43
Probabilitat condicionada	43
Independència estocàstica	44
La fal·làcia de la probabilitat condicionada: fals positiu i fals negatiu	46
Un exemple clínic de probabilitat condicionada en forma de taula	48
Partició de l'espai mostral	48
La regla de la probabilitat total	49
Exemple de la regla de la probabilitat total	50
La regla de Bayes: l'exemple de les llaunes de tonyina defectuoses	52
Un exemple de Bayes d'una prova diagnòstica	52
Problemes proposats	55
Problema 1	55
Problema 2	55
Problema 3	56
Problema 4	56
Problema 5	56
Problema 6	56
Problema 7	57
Problema 8	57
Problema 9	57
Problema 10	57
Problema 11	58
Problema 12	58
Problema 13	58
Problema 14	58
Problema 15	58
Problema 16	59
2.4. Combinatòria	59
2.4.1. Introducció	59
2.4.2. Conceptes fonamentals	60
2.4.3. Permutacions	60
2.4.4. Permutacions amb repetició	60
2.4.5. Permutacions circulars	60
2.4.6. Variacions	61
2.4.7. Variacions amb repetició	61
2.4.8. Combinacions	61
2.4.9. Combinacions amb repetició	62
2.4.10. Aplicacions de la combinatòria	62
Nombre de Bell	62
Teorema del binomi	64
Problemes proposats	64
Problema 1	64
Problema 2	64

Problema 3	64
Problema 4	64
Problema 5	65
Problema 6	65
Problema 7	65
Problema 8	65
Problema 9	65
Problema 10	65
Problema 11	66
Problema 12	66
3. VARIABLES ALEATÒRIES	67
3.1. Definició de <i>variable aleatòria</i>	67
L'espai mostral	67
Variable aleatòria	67
Tipus de variables aleatòries	68
3.2. Funció de densitat i funció de distribució (cas discret)	69
Funció de densitat $f_X(x)$ d'una variable aleatòria discreta	69
Propietats de la funció de densitat d'una variable aleatòria discreta	72
Funció de distribució $F_X(x)$ d'una variable aleatòria discreta	72
Relacions entre la funció de densitat i la funció de distribució $f(x)$ i $F(x)$	72
Propietats de la funció de distribució	73
Probabilitat d'interval	73
Independència de variables aleatòries	74
3.3. Esperança i variància d'una variable aleatòria (cas discret)	74
Caracterització d'una variable aleatòria	74
Esperança d'una variable aleatòria discreta	74
Esperança poblacional <i>versus</i> mitjana mostral	77
Propietats de l'esperança	77
Variància d'una variable aleatòria discreta	77
Variància i desviació típica	80
Propietats de la variància	80
Models de distribució de variables aleatòries	80
Famílies de distribucions	80
3.4. Distribucions discretes: Bernoulli, binomial i Poisson	81
Distribució de Bernoulli	81
Distribució binomial	82
Distribució de Poisson	85
3.5. Variables aleatòries contínues. Distribució normal	87
Variables aleatòries contínues	87
Funció de densitat d'una variable aleatòria contínua	87
Probabilitats d'interval en una variable aleatòria contínua	88
Esperança i variància d'una variable aleatòria contínua	90
La distribució normal	90
Propietats de les variables aleatòries amb distribució normal	91
La funció de distribució del model normal	93
Estandardització o tipificació	93

Problemes proposats	96
Problema 1	96
Problema 2	96
Problema 3	97
Problema 4	97
Problema 5	97
Problema 6	98
Problema 7	98
Problema 8	98
Problema 9	98
Problema 10	98
Problema 11	99
Problema 12	99
Problema 13	99
Problema 14	99
BIBLIOGRAFIA RECOMANADA	101

Introducció

La ciència intenta conèixer els fenòmens naturals i millorar-los. Tal enteniment pot explicar mitjançant un judici d'abstracció i freqüentment s'expressa en termes de lleis, axiomes i teories que permeten predir esdeveniments futurs sense precisar certs límits d'exactitud. Des dels començaments de la civilització hi ha hagut maneres senzilles d'estadística: ja s'utilitzaven representacions gràfiques i altres símbols en pells, roques, pals de fusta i parets de coves per comptar el nombre de persones, animals o certes coses. Cap a l'any 3000 aC els babilònics usaven petites tauletes d'argila per a recopilar dades en taules sobre la producció agrícola i dels gèneres venuts o canviats amb la barata. Els egipcis analitzaven les dades de la població i la renda del país molt abans de construir les piràmides (s. XI aC). Els llibres bíblics de Nombres i Cròniques inclouen, en algunes parts, tractats d'estadística. El primer conté dos censos de la població d'Israel i el segon descriu el benestar material de les diverses tribus jueves. A la Xina hi havia registres numèrics similars d'abans de l'any 2000 aC. Els grecs clàssics feien censos; la informació s'utilitzava cap al 594 aC per a cobrar impostos i així fins avui en dia.

Actualment, la ciència estadística és una eina ineludible per a quantificar la incertesa, i el seu ús és inevitable en l'exploració del coneixement. Els conceptes d'*incertesa* i *aleatorietat* han ofuscat la humanitat des de l'origen dels temps. Així, les societats primitives, encara que de manera rudimentària, ja comptaven i processaven informació numèrica, com el nombre de caps de bestiar que posseïen. Avui dia i molt freqüentment, ens aguaita la incertesa en l'entorn físic i social en què vivim, i encara més a causa del gran volum d'informació i de la incertesa que comporta l'era de la informació i de la informàtica. Suportem les incerteses de la natura i en patim les catàstrofes; només cal recordar la impredictibilitat dels terratrèmols.

Però, què és l'estadística? L'estadística és la ciència que utilitza com a instrument de treball les matemàtiques i el càlcul de probabilitats amb què estudia i prediu el comportament dels fenòmens que depenen de l'atzar, amb la finalitat de detectar comportaments futurs, encara que en molts casos no en queda clar l'origen, la definició ni la fonamentació, que a continuació s'exposen molt breument.

La paraula «estadística» prové del llatí *statisticum Collegium* ('consell d'Estat') i del seu derivat italià *statista* ('home d'Estat' o 'polític'). El terme alemany *Statistik* va ser introduït per Gottfried Achenwall (1719-1772), filòsof i estadístic prussià considerat un dels pares del terme, que apareix en el seu estudi titulat *Staatswissenschaft der vornehmen Europäischen Reiche und Republiken*. «Estadística» designava originàriament l'anàlisi de dades de l'Estat, és a dir, «la ciència de l'Estat» (també anomenada «aritmètica política», de la seva traducció directa de l'anglès). No va ser fins al segle XIX que el terme «estadística» va adquirir el significat de 'recol·lectar i classificar dades', quan va ser introduït a Anglaterra per l'anglès Sir John Sinclair (1754-1835) en el seu treball *Statistical Account of Scotland* (1791-1799) (Viquipèdia, 2009a). Cal no confondre l'«estadista» ('home d'Estat') amb l'«estadístic» ('home de les dades'). A principis del segle XIX, la paraula «estadística» adopta un significat més generalitzat cap a la recollecció i classificació de qualsevol tipus de dades quantitatives.

En l'origen, l'estadística s'associava a les dades que el govern i els cossos de funcionaris administratius, freqüentment centralitzats, havien d'utilitzar. En particular, els censos proveeixen informació regular sobre la població.

Molts conjunts de dades es poden aproximar, amb gran exactitud, utilitzant determinades distribucions probabilístiques, els resultats de les quals es poden utilitzar per a analitzar posteriorment les dades. Així, l'estadística presenta les mateixes bases que el càlcul i l'àlgebra, encara que, depenent de la branca del coneixement en què s'imparteixi la docència, la base matemàtica es deixa de banda o se simplifica, sense tenir en compte l'ús interessat que se'n vulgui fer. L'estadística estudia poblacions i fa afirmacions sobre els seus paràmetres, ja que no es coneix exactament el que ha de succeir; per això l'estadística estudia subconjunts de la població, la **mostra**, i després generalitza les seves conclusions a la població completa, que s'anomena **inferència**. Si s'estudiés tota la població completa no estariem parlant d'estadística, i les conclusions obtingudes serien certes amb una probabilitat del 100% ($p = 1$). La mostra ha de ser representativa de la població, si no es produiria un biaix, per això generalment es trien els individus a l'atzar dins de la població. Aquest procés s'anomena **mostreig**. Així, l'atzar garanteix que la mostra és representativa de la població, però els resultats obtinguts s'enuncien com una probabilitat i per això les afirmacions de l'estadística són probabilístiques.

Com a ciència, l'estadística no explica fenòmens ni estableix relacions causals, sinó associacions estadístiques. Això requereix explicacions addicionals: per exemple, hi ha una associació en certes zones de la Península, entre l'augment de les cigonyes i el nombre de naixements de nens, però això no significa que les cigonyes siguin les causants dels naixements, sinó que són relacions indirectes. No podem prescindir de l'estadística perquè aquesta ens ajuda a estudiar poblacions completes amb una petita mostra, de manera que s'estalvien recursos i temps; així mateix, el nostre desconeixement de realitats molt complexes ens permet almenys establir associacions estadístiques que ens poden proporcionar pistes sobre possibles interpretacions causals que posteriorment seran experimentades.

Finalment, un breu apunt del perquè hem d'estudiar estadística en ciències i fer-ho bé. En ciència es requereix la investigació, tant observacional com experimental, i en cada cas cal sotmetre els resultats a una anàlisi estadística per a confirmar o rebutjar les hipòtesis plantejades inicialment. Les respostes esperades a aquesta pregunta per part de l'investigador són el que es coneix com a **hipòtesi d'estudi**. Un cop definida la hipòtesi, l'investigador n'ha d'avaluar el compliment mitjançant un estudi dissenyat amb aquesta finalitat, emprant la sistemàtica del mètode científic i la metodologia estadística, la qual cosa li permetrà acceptar, modificar o rebutjar la hipòtesi definida. En funció de l'acceptació o no de la hipòtesi, l'investigador ha de formular la llei basant-se en un raonament de tipus inductiu, és a dir que generalitzarà els resultats de l'estudi a la població de referència corresponent.

Les característiques i els continguts dels estudis o assajos que es portin a terme com a part de la investigació d'un fenomen han d'estar definits en un protocol que estableix la raó de ser d'un estudi (objectius, disseny, metodologia, anàlisi prevista dels resultats i condicions de realització).

La mesura de resposta o respostes al tractament triades en l'estudi ha de determinar els mètodes estadístics que s'utilitzaran durant l'anàlisi de les dades de l'estudi, i també la mida de la mostra.

Amb la correcta aplicació de l'anàlisi estadística i una adequada planificació es pot determinar si la relació entre l'exposició a un tractament i un resultat és casual o, per contra, està subjecta a una relació no aleatòria que podria establir una relació de causalitat. L'estadística ajuda a conèixer el paper de l'atzar en les hipòtesis de treball, però no preveu altres errors comuns que es cometien durant la investigació clínica, com són els biaixos de confusió i selecció de la mostra.

1. ESTADÍSTICA DESCRIPTIVA

L'**estadística descriptiva** és una branca de l'estadística que es dedica a recollectar, ordenar, analitzar i representar conjunts de dades, per tal de descriure apropiadament les característiques d'aquestes dades. Aquesta anàlisi és molt bàsica. Encara que l'estadística inferencial generalitza els resultats de les poblacions d'interès i compara les teories sobre la ciència basades en l'evidència empírica expressada en les dades, les primeres conclusions obtingudes després d'una anàlisi descriptiva són bàsiques per a tenir una idea general.

1.1. Tipus de variables

1) **Qualitatius**: els valors no es poden associar «naturalment» amb un número.

- *Nominals*: els valors **no** es poden **ordenar**
 - Colors, sexe, espècie, grup sanguini, fumar (sí/no)
- *Ordinals*: els valors **sí** que es poden **ordenar**
 - Intensitat del dolor, grau tèrmic, nivell de contaminació...

2) **Quantitatius o numèriques**: els valors són numèrics.

- *Discretes*: prenen valors enters (recomptes).
 - Nombre de fills, nombre de minerals, nombre d'organismes...
- *Contínues*: entre dos valors hi ha infinits valors intermedis.
 - Altura, pes, temperatura, concentració d'un producte químic, camp elèctric...

1.2. Taules de freqüències

Exposen la informació recollida en la mostra.

- Freqüències absolutes, (f_i): compten el nombre d'individus de cada valor.
- Freqüències relatives (percentatges o proporcions, f_r): compten el nombre d'individus de cada valor, però dividint pel total.
- Freqüències acumulades: només tenen sentit per a variables amb ordre (ordinals i numèriques). Poden ser absolutes (F_i) o acumulades (F_r).

Exemple: quin percentatge de mostres recollides en el camp tenen tres o mes minerals?

Nombre de minerals	Freqüència (f_i)	% (f_r)	% acumulat (F_r)
0	234	46	46
1	134	26	72
2	97	19	91
3	34	7	98
Més de 3	12	2	100
Total	511	100	

1.3. Gràfics de variables qualitatives

- Diagrames de sectors («formatges»)
 - Només per a variables qualitatives nominals.
 - Les àrees de cada sector són proporcionals a les freqüències.

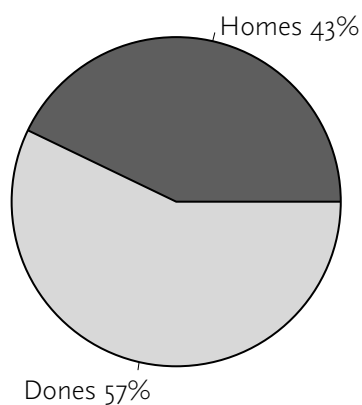


Figura 1. Diagrama de sectors.

- Diagrames de barres
 - Les altures de les barres són proporcionals a les freqüències.
 - Es poden aplicar també a variables discretes.

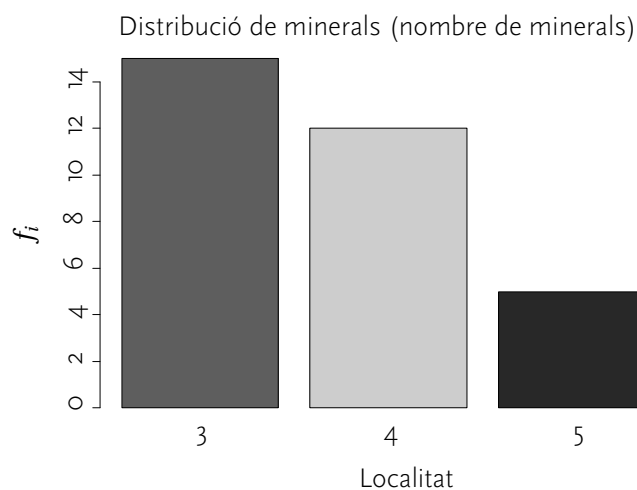


Figura 2. Diagrama de barres.

1.4. Gràfics de variables quantitatives

Són diferents segons si són de variables discretes o de variables contínues:

- Diagrames de barres per a variables discretes
 - Deixem espais per a denotar els valors impossibles.

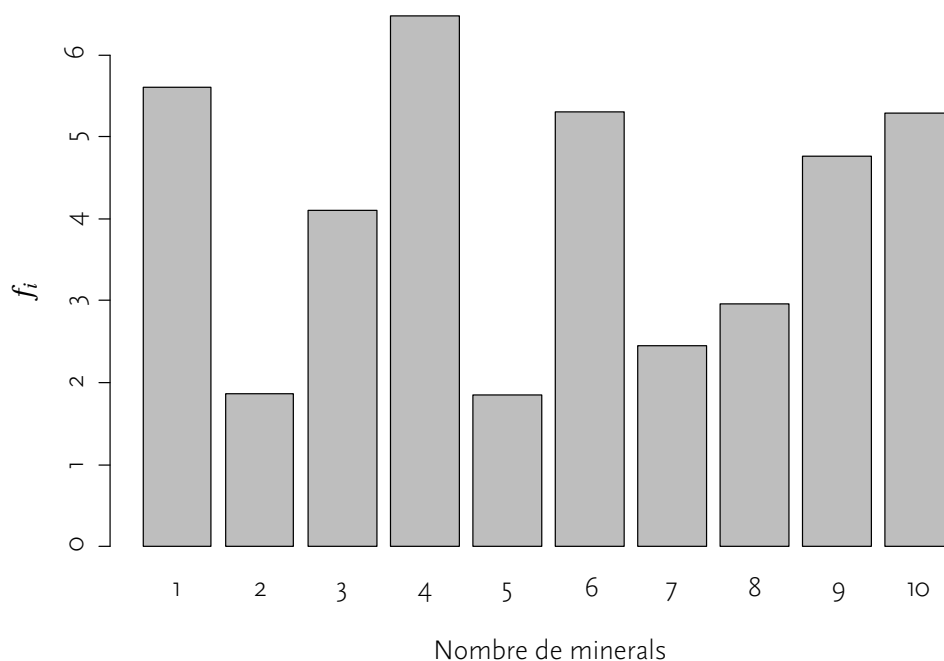


Figura 3. Diagrama de barres d'una variable discreta.

- Histogrames per a variables contínues
 - L'àrea de cada rectangle indica la quantitat (percentatge o freqüència) d'individus de l'interval. Els histogrames es poden indicar en forma de freqüència, ja sigui absoluta o relativa (densitat):

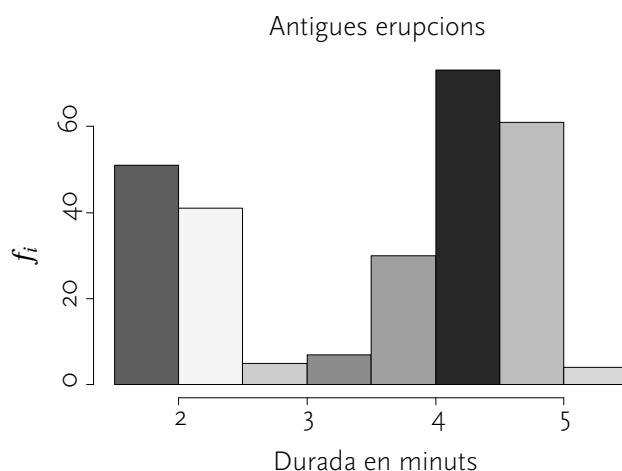


Figura 4. Histograma (en forma de freqüència) d'una variable contínua.

O de freqüència relativa(tant per 1 de la freqüència relativa o anomenat també *densitat*, en anglès):

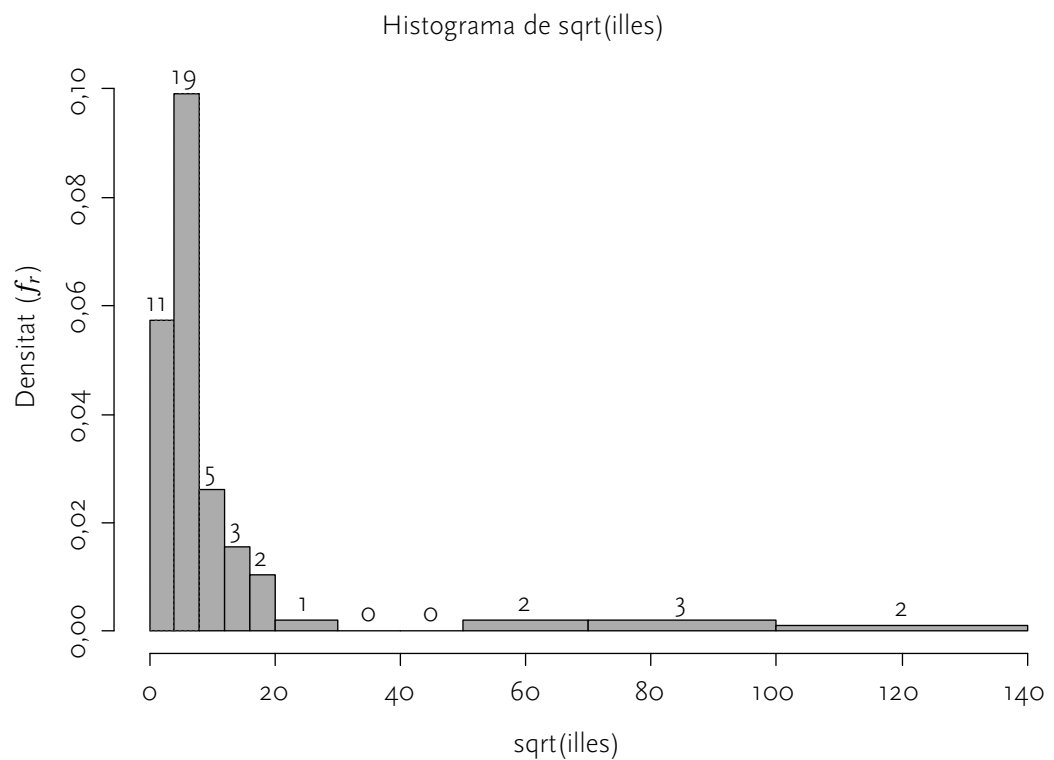


Figura 5. Histograma (freqüència relativa o densitat) d'una variable contínua.

O en diferents intervals:

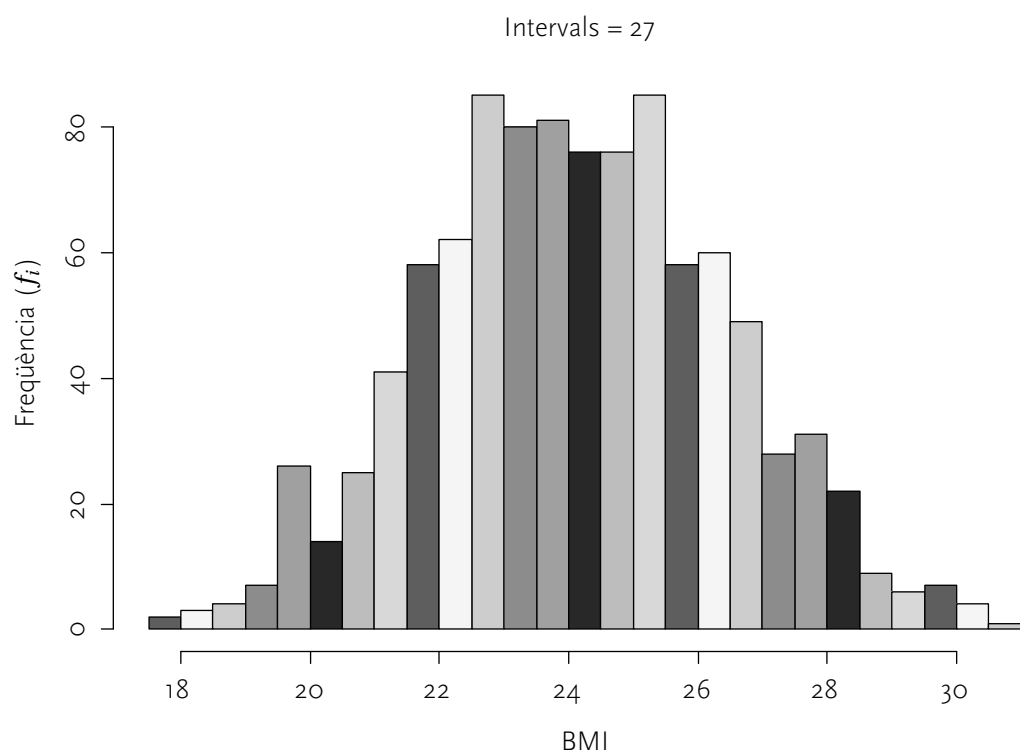


Figura 6. Histograma amb 27 intervals.

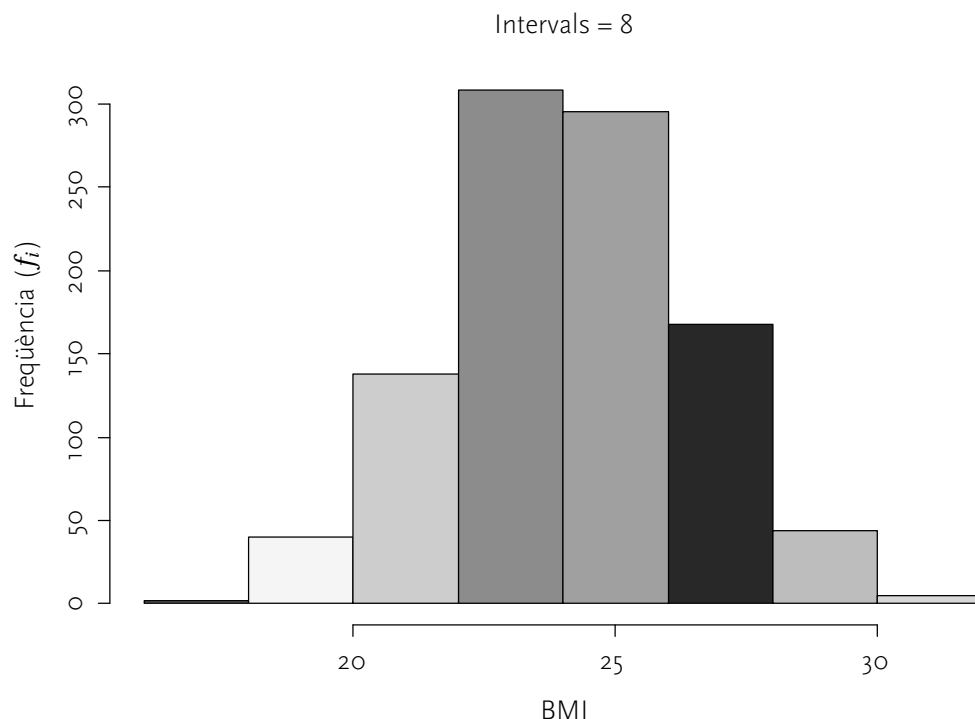


Figura 7. Histograma amb 8 intervals.

1.5. Nombres d'intervals en un histograma. La marca de classe. La regla de Sturges

Quan les dades contenen una gran quantitat d'elements, per a facilitar els càlculs cal agrupar-los. Aquests grups s'anomenen **intervals** o **classes**. Un interval (L_0, L_1) és una sèrie de nombres inclosos entre dos extrems. Per tal de formar una taula agregada de freqüències cal seguir els passos següents:

1. Calculeu el **rang** (R). També se'n diu **recorregut** o **amplitud total**. És la diferència entre el valor més gran i el més petit de les dades:

$$R = \text{màxim} - \text{mínim}$$

2. Seleccioneu el nombre d'**intervals de classe** (n_k). Es recomana que no siguin menys de 5 ni més de 12 per no pertorbar el gràfic que es fa quan s'agrupen dades. Una manera de calcular el nombre òptim de classes és la regla de Sturges:

$$k = 1 + 3,322 \cdot \log(n)$$

On n és la mida mostral.

3. Calculeu l'**amplitud de l'interval** k dividint el rang pel nombre d'intervals:

$$w = \frac{R}{k}$$