POLITEXT

José A. González - Erik Cobo
Pilar Muñoz - Manuel Martí

Estadística per a enginyers informàtics

EDICIONS UPC

POLITEXT 183

Estadística per a enginyers informàtics

POLITEXT

José A. González - Erik Cobo
Pilar Muñoz - Manuel Martí

Estadística per a enginyers informàtics

EDICIONS UPC

Aquesta obra fou guardonada en el 10è. concurs
"Ajut a l'elaboració de material docent" convocat per la UPC.

Primera edició: juliol de 2008

Aquesta obra compta amb el suport
de la Generalitat de Catalunya

Disseny de la coberta: Manuel Andreu

Edicions de la Universitat Politècnica de Catalunya, SL
Jordi Girona Salgado 1-3, 08034 Barcelona
Tel.: 934 137 540 Fax: 934 137 541
Edicions Virtuals: www.edicionsupc.es
E-mail: edicions-upc@upc.edu

Producció: TECFOTO, S.L.
Ciutat de Granada 55, 08005 Barcelona

Dipòsit legal: B-25481-2008
ISBN: 978-84-8301-953-5

Als nostres éssers més estimats,
sense el seu suport i la seva comprensió no ho hauríem aconseguit.

Pròleg

És obvi que vivim en un món dinàmic, i no és exagerat afirmar que cada vegada els canvis que experimenta el nostre entorn són més ràpids. Per als centres universitaris, aquest fet té una gran transcendència, ja que ha donat lloc a una tremenda convulsió en el pla docent, i a preguntar-nos com hem ensenyat, i com hem de fer-ho a partir d'ara; què necessiten els nostres estudiants per la seva vida professional; quina és l'empremta que la universitat ha de deixar en ells.

Els alumnes passen tan sols una petita part de la seva vida adulta a la universitat, però aquest període té un impacte fonamental per al seu futur. Fins no fa massa temps, a la universitat s'aprenia el que es precisava per desenvolupar una professió, i els coneixements adquirits caducaven lentament, podríem dir. El paper de la universitat era més o menys com un transmissor del saber: els alumnes venien a escoltar als mestres i anaven a posar en pràctica aquests coneixements.

En la actualitat, el nou paradigma seria "aprendre al llarg de la vida". El cas de les noves tecnologies és molt clar: la tecnologia digital avança tan ràpidament que cada pocs anys deixa totalment obsoleta la generació anterior. Però els canvis tecnològics arrosseguen uns altres canvis: per exemple, l'aparició d'ordinadors més potents, amb immenses memòries, fan possible l'aplicació d'algorismes que no eren aplicables o que es limitaven a resoldre problemes de joguina, i aquesta mateixa aplicació indueix la millora i el desenvolupament de mètodes més sofisticats. El fenomen d'internet és encara més interessant, perquè no tan sols s'associa a canvis tecnològics i metodològics, possibilitant la disponibilitat immediata de la informació acumulada, sinó que ha suposat una autentica revolució en el pla de les idees. No crec que cap altre element en el món de la informàtica hagi destacat tant el concepte d'*informació*. Ja fa molt de temps que es deia: qui té la informació té el poder.

Bé, aquestes idees em vénen al cap quan llegeixo temes al voltant de variabilitat i decisió, informació i dades, estimació i previsió. El llibre que teniu a les mans parla d'això, de conceptes essencials en el nostre temps, i és segur que tindran més importància cara al futur. És el llenguatge en el que parla l'Estadística, però no és exclusiu d'aquesta branca, són termes que comparteixen la generalitat de les ciències i les tècniques. Tot en el nostre entorn és variable, i tot presenta un grau d'incertesa.

Tenir coneixements suficients d'Estadística no vol dir solament saber utilitzar uns programes que fan gràfics espectaculars i càlculs obscurs. Significa tenir una visió especial sobre els problemes que ens envolten: apreciar en la mesura correcta, sense biaixos, l'existència de components aleatòries i d'altres més regulars; ser capaços de fer estimacions de l'error que suposa la incertesa; construir models matemàtics raonables, i no creure's més el model que el món real; dubtar metòdicament (que no és igual que desconfiar) de suposicions sense un fonament robust; planificar, mesurar, contrastar, ... I penso que totes aquestes qualitats són imprescindibles en els titulats en enginyeria informàtica, perquè són capacitats sense data de caducitat, perquè els ajudaran a continuar la seva formació en el futur, perquè són transversals en moltes disciplines (i perquè la vida dóna moltes voltes).

Maria Ribera Sancho
Degana de la Facultat d'Informàtica de Barcelona
Març de 2008

Presentació

En Enginyeria Informàtica existeixen molts fenòmens que no son deterministes, és a dir, que estan subjectes a error aleatori, com poden ser el temps de resposta d'un ordinador en sistemes complexos. Per poder modelar aquests fenòmens és necessari usar procediments que permetin quantificar i, si és possible, reduir el terme d'error. L'objectiu del llibre es presentar al lector la metodologia estadística bàsica que permeti deduir —inferir— el comportament de la població que s'està estudiant, partint d'un nivell elemental de coneixements de probabilitat. Els autors fan ús extensiu d'exemples provinents del camp informàtic, anotacions i comentaris per recolzar els continguts i facilitar la lectura i la comprensió dels conceptes teòrics. Completen l'obra molts treballs i exercicis, la majoria solucionats i amb discussions útils. Després d'un capítol dedicat a introduir el concepte d'inferència estadística, s'exposen al capítol 2 els principis fonamentals, mostra i estimador. A continuació es presenten les tècniques del interval de confiança (cap. 3) i la prova de significació i contrast d'hipòtesis (cap. 4). El capítol 5 es dedica a explicar el tema de la comparació de dues poblacions, i el capítol 6 desenvolupa les relacions entre variables sota l'òptica del model lineal. El capítol 7 presenta els aspectes relatius a validació i previsió i, finalment, al capítol 8 s'introdueixen les proves de Pearson per l'anàlisi de variables categòriques.

ÍNDEX

1 Introducció a la inferència estadística

La paraula *estadística* s'associa generalment a mitjanes, taules i gràfics acolorits, quan aquests recursos són només algunes de les eines que els mètodes estadístics —entre ells, la inferència estadística— tenen per resumir un conjunt de dades i extreure'n informació. Els mitjans de comunicació inclouen sovint informació d'aquest estil, especialment si es tracta d'algun tema polèmic ("Es doblen les morts de trànsit a Barcelona"), per subtilment donar a entendre un efecte sorprenent. El grau de credibilitat de lesn notícies il·lustrades amb dades pseudocientífiques és molt alt, i poden donar suport a posicions que, en rigor, són injustificables. És molt sa adoptar la postura del *dubte metòdic*, plantejant-se si existeixen explicacions alternatives a la interpretació, més o menys explícita, que figura a la informació. La casualitat —l'atzar— sovint pot ser una explicació, i la inferència estadística ens aporta eines per tractar l'indeterminisme present a gairebé qualsevol situació i contrastar diferents opcions. L'objectiu del capítol, aplicable a tot el llibre, és presentar i justificar el mètode d'inferència estadística, situant-lo en un context científic. En acabar el capítol, el lector haurà après els conceptes fonamentals (població, mostra, paràmetre, etc.) que sustenten el mètode, i veurà l'estadística com l'opció més correcta per tractar, de manera sistemàtica, problemes que es poden presentar en l'àmbit professional d'un enginyer.

1.1 Què és la inferència estadística?

La inferència estadística és un procediment per incorporar l'evidència empírica en un cos de coneixement més ampli, quan aquesta evidència prové d'unes dades o d'unes proves experimentals. Si, per exemple, volem esbrinar quina és la relació funcional entre el temps invertit per un algorisme d'ordenació i el nombre d'elements que s'han d'ordenar, podem fer servir dos procediments. El primer, teòric, consisteix a *deduir* aquesta expressió funcional a partir de l'anàlisi de l'algorisme. El segon procediment, empíric, consisteix a "provar-ho": realitzar una sèrie d'execucions de l'algorisme i estudiar-ne les observacions.[1]

[1] Per a molts algorismes d'ordenació, aquesta relació és quadràtica en la mida del vector a ordenar (n). La notació per designar un cost de tipus quadràtic és $O(n^2)$. Alguns algorismes, com el *quicksort*, són sensiblement més eficients, ja que el seu cost mitjà és $O(n \cdot \log(n))$.

Ara bé, fins a quin punt unes quantes proves ens aporten informació sobre el funcionament general de l'algorisme? El procediment inductiu pretén inferir a un tot les propietats establertes en unes parts d'aquest tot. Fins fa ben poc, els filòsofs s'han queixat de la manca d'eines tècniques que permetessin aquest salt de les parts al tot. Per a Hume, la inferència era simplement impossible i, per a Russell, la inducció continuava essent un problema de lògica no resolt. A mitjan segle passat, Popper va aportar un punt de vista una mica més positiu: "només la refutació d'una teoria es pot inferir de dades empíriques i aquesta inferència és purament deductiva". Avui en dia, ja està plenament acceptat que la metodologia estadística fa possible la inferència sempre que s'acceptin alguns riscos. Al llarg d'aquest llibre, veurem com la inferència estadística proposa formalitzar aquest procés, cosa que requereix, necessàriament, definir, quantificar i fitar els riscos que comporta.

Observem a la taula 1.1 algunes preguntes que poden sorgir en experiments informàtics i que es poden contestar amb l'ajut de la metodologia estadística. L'exemple més senzill tracta de la distribució d'una sola variable: quant triga el meu algorisme a ordenar vectors d'una mida concreta? O bé, quant espai de memòria requereix la meva implementació de l'algorisme de Kruskal? O també, quantes interrupcions haig d'esperar en el meu treball, degudes a trucades de clients? Si no hi hagués variabilitat —si el meu algorisme trigués sempre el mateix, o el de Kruskal ocupés exactament el mateix espai, o em truquessin sempre el mateix número de clients—, la inferència seria immediata: amb una observació en tindríem suficient per conèixer el comportament de cada una d'aquestes variables. Però si el fenomen estudiat presenta variabilitat, hi apareix incertesa, ja que el seu comportament no és previsible sense error. La inferència estadística tracta, doncs, una situació més general, que inclou una estimació del grau de variabilitat entre les unitats mesurades (que, genèricament, es coneixen com a *individus*).

És, doncs, la variabilitat la que justifica l'ús de l'estadística per aprofundir determinats problemes que es troben en qualsevol àmbit, des de l'Administració fins als camps tècnic i científic.

Taula 1.1 Preguntes típiques de la inferència estadística

Pregunta típica	Amb relació a...	Lloc dins del llibre
Quant triga el meu programa a...? Quant ocupa...?	Com s'estima un paràmetre...? Com es distribueix la variable X?	Intervals de confiança
Relació entre temps de CPU i mida? Temps de CPU en canviar el sistema operatiu?	Quina és la relació entre X i Y? Què passa amb Y quan canvio X?	Comparació de dues mostres Anàlisi de la variància Regressió

El segon tipus d'exemple aborda l'estudi simultani de dues variables. Quina relació funcional hi ha entre el nombre d'elements d'un vector i el temps utilitzat a ordenar-lo? O bé, quina relació hi ha entre la nota final de la carrera i el nivell d'ingressos, dos anys després d'obtenir el títol? L'establiment d'una relació funcional té dues utilitats potencials: la previsió i la intervenció. Per preveure el valor d'una variable a partir del valor d'una altra, és suficient establir una relació numèrica entre ambdues

variables. Per exemple, permet preveure una aproximació del sou d'un enginyer que just ha acabat els estudis. Però, buscant una millora salarial, no aconseguim res manipulant el seu expedient acadèmic. Quan existeix una relació anomenada de causa-efecte, intervenir sobre el factor *X* origina canvis en la distribució dels valors de *Y*: per exemple, les hores dedicades a l'estudi de la matèria i la qualificació obtinguda a l'examen. Si un alumne incrementa el seu esforç estudiant més temps, probablement millorarà la seva nota. L'efecte causal s'estableix estudiant els canvis que es manifesten a *Y* (variable resposta) quan realitzem canvis a *X* (variable intervinguda) i mantenim a un nivell fix tota tercera variable *Z* (condicionant o covariant) que pugui influir en el procés.

En el cas bivariant (o multivariant), és habitual que la incertesa afecti únicament la variable resposta; l'atzar pot també afectar les altres variables, sota certes condicions i normalment si no s'estudia un efecte causal, és a dir, si només es vol establir un model per a la relació existent entre *X* i *Y*.

L'*exemple* més senzill de relació causa-efecte és el de l'interruptor (*X*), que ens permet encendre l'ordinador (*Y*) sempre que la resta del sistema (*Z*) estigui en ordre (en aquest cas, la relació no és aleatòria sinó determinista).

Com a *exemple* de relació no causal, es pot considerar el volum (*X*) de la caixa d'un ordinador i la capacitat de càlcul (*Y*) d'aquest. Possiblement, a major volum, major capacitat, fet que podria, fins i tot, permetre certa predicció.[2] Però segur que a ningú no se li ocorre fer un ordinador més voluminós (*X*) per incrementar les seves prestacions (*Y*), sense modificar els seus components (*Z*).

Un *tercer exemple* poden ser els resultats acadèmics dels alumnes d'una assignatura, en funció de si són repetidors o nous, i de si han realitzat o no com a mínim 4 dels problemes proposats durant el curs. El quadre següent dóna la mitjana de la nota a l'examen segons aquests factors:

Nota mitjana a l'examen	≥ 4 probs.	< 4 probs.
repetidor	5.366	4.387
nou	5.891	4.318

Sembla que els millors resultats acadèmics dels alumnes d'aquesta assignatura corresponen als alumnes nous que fan els problemes. Per tant, podem fer certa *previsió*: si algun alumne fa els problemes i és nou, es preveu una millor nota. Ara bé, hi podem *intervenir*? Un alumne pot millorar les seves expectatives realitzant els problemes, però un repetidor no pot deixar de ser-ho. Per tant, la variable 'repetidor' és una condició, és una tercera variable *Z* que pot servir per predir, però, mentre no es pugui decidir el seu valor i canviar-la, no serveix per intervenir.

Vegem ara la variable 'realitzar almenys quatre problemes'. En podem decidir el valor i, per tant, potencialment podria servir per intervenir. Ara bé: ¿existeix alguna altra diferència

[2] El mot *predicció* no és molt apropiat, perquè té connotacions de ciència esotèrica. Malgrat tot, s'empra freqüentment com a sinònim de previsió i, quan no hi hagi confusió, l'utilitzarem eventualment amb el mateix sentit.

entre els que fan els problemes i els que no els fan que pugui també explicar aquest millor rendiment? Potser els que fan els problemes també van a classe, estudien, fan les pràctiques i revisen els exàmens passats. Potser estan motivats i tenen ganes d'aprendre i d'aprovar en un curs. Totes aquestes variables en poden ser l'autèntica causa, en lloc dels problemes fets. Per això, per estar segurs que els problemes són l'autèntica causa i la variable a intervenir, el millor mecanisme seria realitzar un experiment: es trien a l'atzar dos grups d'alumnes nous —un grup fa els problemes i l'altre no. Com que s'han seleccionat a l'atzar els grups, la resta de variables estan equilibrades i no poden servir d'explicació alternativa als problemes del diferent rendiment en les notes. D'aquí la importància d'un bon disseny del nostre experiment, on s'inclogui, si és possible, l'assignació aleatòria al grup (el que es coneix com a *aleatorització*).

1.2 Població, mostra i individu

Per comprendre millor el raonament que constitueix la inferència estadística, introduïm alguns conceptes fonamentals:

- Entenem per *població* el conjunt de tots els elements que compleixen certes propietats i dels quals es desitja estudiar un determinat aspecte.

- Un *individu, element o cas* és cada una de les unitats que componen la població. S'ha de tenir en compte que aquests "individus" no han de ser, necessàriament, persones. Poden ser ordinadors, programes, execucions...

- Finalment, *mostra* és el conjunt d'individus seleccionats de la població que és estudiat i a partir del qual es treuen conclusions per inferir característiques de la població.

És molt important definir amb molta cura aquestes unitats ja que, segons com les definim, podem arribar a conclusions diferents.

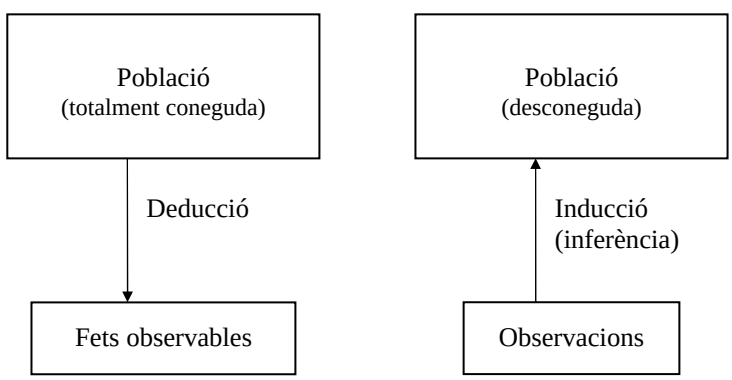


Figura 1.1 Procés deductiu i procés inductiu

Ara que s'han descrit aquests nous termes, podem dir que la inferència estadística pretén formalitzar el salt des de la mostra a la població. És necessari recordar que també es produeix un salt, però en sentit invers, quan arribem a conclusions particulars basant-nos en lleis generals, per exemple, quan calculem quina és la probabilitat de llançar un parell de daus i obtenir-ne dos números iguals.

Comentari: El procés cíclic que il·lustra la figura 1.1 (inducció de coneixement a partir de dades, deducció a partir de lleis de noves propietats a comprovar...) ha permès un progrés espectacular en gairebé totes les ciències. Avui en dia, encara que pot trobar-se amb variants menors, s'accepta com a model de raonament científic el contingut en el esquema de la taula 1.2.

Taula 1.2 Passos del mètode científic

1	Descriure el problema a investigar.
2	Documentar i definir el problema o hipòtesi.
3	Deduir conseqüències contrastables de les hipòtesis.
4	Dissenyar l'observació o experimentació.
5	Recollir les dades procedents d'una mostra.
6	Inferir estadísticament propietats de la població.
7	Establir-ne les conclusions.
8	Integrar les conclusions en el cos de coneixement.

Observeu que aquest procediment combina els raonaments deductius necessaris, com per exemple, els que calen per dissenyar la recollida de dades, amb els inductius que es requereixen per generalitzar les observacions obtingudes a partir dels elements de la mostra.

Podeu ampliar aquesta informació a: <http://en.wikipedia.org/wiki/Scientific_method>

1.3 Paràmetre, estadístic i estimador

De fet, observar un individu de la població significa anotar el valor d'una o de diverses característiques que li són pròpies: per exemple, un atribut com el sexe (masculí, femení), o una mesura numèrica com l'estatura (175 cm). Una observació, que es denota com x_i, és un valor determinat procedent d'un individu concret. D'alguna manera, no ens interessa l'individu en si (és solament l'individu *i*-èsim), sinó només els valors que ens proporciona.

Aquestes característiques tenen variabilitat a la població, la qual cosa es posa de manifest repetint les observacions. Per exemple, si a la població d'estudiants de la Universitat X (UX) hi ha un 40% de dones, és normal que una mostra ben escollida (= a l'atzar) d'estudiants contingui homes i dones, i normalment

més homes que dones. Naturalment, si la mostra és petita, és perfectament possible que només hi estigui representat un dels dos sexes. Això es pot il·lustrar amb aquest cas: si es pren una mostra aleatòria de tan sols tres estudiants, hi ha una probabilitat del:

$$P(\text{home})\cdot P(\text{home})\cdot P(\text{home}) = 0.60\cdot 0.60\cdot 0.60 = 0.216$$

és a dir, del 21.6% que no hi hagi cap dona (en aquest moment fem una *deducció*, un salt de la població a la mostra). Aquest valor ens diu que no és gens difícil trobar-nos amb aquest cas "extrem". És clar que mai no hem de prendre decisions si la informació empírica és tan pobra. Però, estrictament parlant, podríem dir que l'estimació que s'ha trobat del percentatge d'estudiants de sexe femení a la UX és ¡0%! (i aquest comentari és un exemple d'*inducció*, encara que poc afortunada).

De pas, afegim que els professionals de l'estadística solen utilitzar models matemàtics per manegar més còmodament situacions reals complexes (com quan hi ha incertesa) i facilitar el tractament del coneixement. Així, podríem assignar una distribució de probabilitat (coneguda com Bernoulli) a una variable que associés el sexe d'un estudiant de la UX a un número (0 a homes i 1 a dones). El valor del 40% (o 0.40) seria un paràmetre rellevant —l'únic necessari—per descriure com es distribueix aquesta variable.

Recapitulem: una característica mesurable dels individus (que anomenem *variable*) es manifesta en diferents valors. Per qüestions pràctiques, ja que sovint no és viable tractar exhaustivament cada possible valor de la variable, utilitzem quantitats (que anomenem *paràmetres*) per resumir un tret determinat de la variable. Per exemple, si la variable és numèrica, el valor de la mitjana o la dispersió al voltant de la mitjana. Els paràmetres són propis de la població, cosa que equival a dir que si no es coneix perfectament la població tampoc no es coneix el valor exacte d'un paràmetre concret. Els *estimadors* són una funció de la mostra, és a dir: valors que obtenim operant els valors numèrics de les observacions d'una mostra, i que poden aproximar-se, d'alguna manera, al paràmetre de referència. Cal deixar clar des del principi que un estimador pot retornar valors diferents, ja que una mostra escollida aleatòriament tindrà observacions diferents d'una altra extreta a continuació. Finalment, s'anomena *estadístic* qualsevol funció matemàtica que construïm amb una mostra. Està clar que un estimador és un estadístic, però qualsevol estadístic no és l'estimador d'un paràmetre de la població. Potser no és tan intuïtiu afirmar que poden existir molts estimadors per al mateix paràmetre. En aquest cas, és important escollir el millor (ja veurem què significa aquest terme "millor estimador").

Taula 1.3 Comparació entre paràmetre i estimador

		Paràmetre	**Estadístic, estimador**
Símbol genèric		θ	$\hat{\theta}$
Parlem de...		La població	Una mostra
Nom	Mitjana	μ	$\bar{x}$
	Desviació tipus	σ	s
	Proporció	π	p

Com veiem a la taula 1.3, els paràmetres es denoten generalment amb lletres gregues: mu (μ) per a un indicador concret de la tendència central dels valors de la variable, conegut com *esperança* o *valor esperat*; sigma (σ) designa la *desviació tipus*, un indicador de dispersió, i pi (π) s'utilitza per referir-se a una *probabilitat*, és a dir, a la part de la població que satisfà una determinada propietat (com ser dona). Si volem referir-nos a un paràmetre no especificat, utilitzem la lletra theta (θ). Un estimador d'aquests paràmetres a vegades es mostra amb el mateix símbol, coronat amb un accent circumflex (^), però els estimadors més habituals utilitzen lletres llatines: $\bar{x}$, s, p. Sembla oportú recordar aquí com s'obtenen aquests valors a partir d'una mostra de n observacions (x_1, x_2, ..., x_n):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

REPÀS: $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$p = \frac{1}{n}\sum_{i=1}^{n} x_i \text{, on } x_i = 1 \text{ si l'individu } i \text{ compleix la propietat, } x_i = 0 \text{ en altre cas.}$$

Exemple. Què podem obtenir si volem estimar la proporció de noies estudiants a la UX amb una mostra de tan sols tres individus? Anem a veure que, tot i que la mostra és massa petita, el procediment és a la llarga adequat. Totes les possibilitats s'enumeren a la taula 1.4.

Taula 1.4 Mostres possibles de n=3, si la probabilitat d'observar un noi (M) és 0.60

Sexe (1er-2on-3er)	Nombre de dones (k)	Proporció mostral ($p = k/n$)	Probabilitat de la mostra	En percentatge
M-M-M	0	0.00	0.60·0.60·0.60	21.6%
M-M-F	1	0.33	0.60·0.60·0.40	14.4%
M-F-M	1	0.33	0.60·0.40·0.60	14.4%
M-F-F	2	0.67	0.60·0.40·0.40	9.6%
F-M-M	1	0.33	0.40·0.60·0.60	14.4%
F-M-F	2	0.67	0.40·0.60·0.40	9.6%
F-F-M	2	0.67	0.40·0.40·0.60	9.6%
F-F-F	3	1.00	0.40·0.40·0.40	6.4%

Un càlcul rigorós aconsella, en primera instància, considerar l'ordre dels individus a la mostra, però, com es pot veure, les vuit mostres possibles es poden reduir a només quatre casos d'interès pràctic: que no hi hagi cap noia, que hi hagi una, dues o tres. Cadascun d'aquests casos té una probabilitat que es pot obtenir fàcilment, sumant les probabilitats de les mostres adients. Per exemple: la probabilitat que hi hagi una dona és 0.144+0.144+0.144=0.432.

La proporció mostral condueix algunes vegades a una subestimació, i altres a una sobreestimació, de la proporció real de dones, indicada pel valor π=0.40. Mostrem que, si es repetís moltes vegades l'experiència, el resultat global no apuntaria cap desviació del paràmetre:

Taula 1.5 Distribució de la proporció mostral respecte al paràmetre

Nombre dones (k)	Proporció mostral p	Desviació ($p - 0.40$)	Probabilitat de p	En percentatge	Producte prob.×($p - 0.40$)
0	0.00	–0.400	0.216	21.6%	–0.086
1	0.33	–0.067	3·0.144	43.2%	–0.029
2	0.67	0.267	3·0.096	28.8%	0.077
3	1.00	0.600	0.064	6.4%	0.038
			Suma:	100%	0.000

Les distintes desviacions sumades (ponderant per la probabilitat: per això fem el producte a la darrera columna) resulten en suma zero, la qual cosa vol dir que les proporcions trobades en diferents mostres s'equilibren entre sí, encara que faria falta un nombre molt gran de mostres per verificar-ho a la pràctica. Evidentment aquesta propietat no és un gran consol, i no és convenient confiar en una mostra tan petita. Però la qüestió de quina és la grandària apropiada requereix més coneixements.

Addicionalment, assenyalem que el desenvolupament del cas ha fet aparèixer la distribució de probabilitat *binomial*, aplicada a la variable aleatòria "Nombre de dones en una mostra de grandària 3", que assigna la probabilitat per al valor k com:

$$\binom{3}{k} 0.40^k \, 0.60^{3-k} \, .$$

Finalment, hem d'apuntar una situació que hi ha a la base de quasi tota experiència indeterminista. Suposem que el mostreig dels estudiants es fa a la sortida del gimnàs (assumim que tots els socis són estudiants de la UX i que si un no és estudiant es pot detectar) o a la biblioteca. Potser aquests mètodes són deficients, si els nois van més al gimnàs que les noies, o si les noies van més a la biblioteca que els nois. Un cas condueix a una subestimació en la proporció de noies, i l'altre a una sobreestimació. I una repetició sostinguda del procés de mostreig no corregiria l'estimació, al contrari que en l'exemple previ.

Veiem que les desviacions, que també s'anomenen *errors* en la terminologia estadística (sense el matís negatiu que té en altres àmbits), poden ser de dos tipus: els errors *aleatoris*, deguts exclusivament a les fluctuacions de l'atzar, i tota la resta, coneguts com a errors *sistemàtics* o *biaixos*. L'estadística que veurem al llarg d'aquest llibre ajuda a quantificar la magnitud dels primers. Quant als segons, en tenen la responsabilitat compartida tant l'estadístic com l'expert de l'àrea que s'investiga, assenyalant possibles fonts de biaixos específics del problema.

1.4 Inferència estadística a la informàtica

La consolidació de l'estadística no és tan recent com la informàtica, però tampoc no és molt antiga. Prové dels inicis del segle XX i va contribuir a donar un nou impuls a disciplines com l'agronomia, l'enginyeria, l'econometria, les ciències socials o la psicometria, per no citar-ne d'altres. Com és natural, a mesura que la informàtica es va anant desenvolupant com a tècnica (i també com a ciència), s'acull a la metodologia científica que hem presentat i adopta com a propis aquests procediments. Però totes les ciències progressen i l'estadística ha fet grans avenços en les darreres dècades, i no és cap sorpresa dir que la informàtica hi té bona part de culpa. En efecte, el constant increment de la potència de càlcul ha obert moltes portes que ni es somniava que fossin accessibles algun dia. Això ha permès provar mètodes i models teòrics inaplicables per la gran quantitat de càlculs que requereixen. Un exemple concret: els pronòstics meteorològics s'elaboren introduint a ordinadors potentíssims grans quantitats de dades per tal d'ajustar models matemàtics molt complexos, i simular l'evolució temporal d'aquest clima virtual. Actualment, s'elaboren pronòstics fins a deu dies, però només són altament fiables (amb un índex superior al 80% a Europa) fins al tercer dia. I això és molt bo, en comparació dels resultats de fa pocs anys.

Si bé en aquest cas situem l'estadística com una eina al servei de la meteorologia (què es pot imaginar més incert que el clima?), també es poden citar exemples en els quals l'estadística és útil a la ciència informàtica, o en què col·laboren juntament. La llista següent no pretén ser exhaustiva:

- Intel·ligència artificial: podeu consultar una amplia col·lecció de temes en comú a: <http://www.dcs.bbk.ac.uk/~mark/web.html>

- Robòtica: la incorporació de sistemes de percepció i algorismes de control en entorns reals requereix un tractament adequat de les dades recollides de l'exterior, per tal d'assegurar la repetitivitat i la fiabilitat dels moviments.

- Visió per ordinador: el soroll introduït a les imatges captades en temps real dificulta la identificació correcta dels objectes. S'utilitza l'estadística per caracteritzar i calibrar el color, i per calcular la localització i la inclinació dels objectes.

- Tecnologia de la producció: tracta sobre la simulació dels sistemes productius i l'optimització dels processos productius.

- Control: l'aportació principal de l'estadística és el modelatge. S'aplica sobretot per verificar la validesa de les dades procedents d'una xarxa de sensors, i també per detectar i corregir errors de lectura.

- Qualitat del software: actualment, l'enginyeria del software adopta un enfocament més empíric per tractar el problema de la fiabilitat dels programes informàtics.

- Mineria de dades: extracció i construcció d'informació no representada explícitament en ingents volums de dades. Implica diverses àrees d'estudi, com la visualització gràfica i les bases de dades. Cada vegada adquireixen més força subbranques com el *web-mining* o l'estudi de les citacions bibliogràfiques.

- Bioinformàtica: la biologia computacional consisteix bàsicament a posar en marxa models i mètodes numèrics en l'àmbit de l'enginyeria genètica, on l'atzar juga un paper molt rellevant.

- Arquitectura de computadors: la simulació és un recurs molt utilitzat per avaluar i experimentar amb noves configuracions de processadors, xarxes de comunicacions o supercomputació.

- Gràfics, visualització i multimèdia.

- Complexitat d'algorismes.

El ventall d'opcions és ampli, i no sempre trobem els exemples en camps d'investigació puntera. Qualsevol informàtic ha d'afrontar alguna vegada un problema que l'obliga a plantejar-se unes hipòtesis, una recollida de dades i una anàlisi metòdica per trobar arguments robustos que recolzin les seves conclusions. De fet, és probable que al llarg de la seva trajectòria professional assumeixi cada cop més responsabilitats, i hagi de prendre decisions importants amb freqüència. Emprarà menys temps en qüestions tècniques (programar, per exemple) i més a establir objectius, planificar i organitzar. El raonament estadístic és una eina útil per desenvolupar habilitats que el professional necessitarà en el món laboral, com el sentit crític, el domini del llenguatge quantitatiu, la capacitat d'anàlisi i síntesi, o l'aptitud per comunicar-se efectivament.

1.5 Problemes

1. Quins paràmetres són rellevants per a les següents distribucions de probabilitat: *a*) de Poisson; *b*) exponencial; *c*) normal; *d*) uniforme; *e*) triangular?

2. Se suposa que el temps d'accés a dades emmagatzemades al disc dur és uniforme entre 0 i un temps de T mil·lisegons, i es vol estimar aquest temps màxim. Proposeu almenys dos mètodes diferents d'estimació de T, i descriviu amb precisió què són la població, la mostra i els altres conceptes que s'han presentat.

1.6 Solució dels problemes

1.

a) Per a una distribució de Poisson, el paràmetre és el valor λ (taxa d'esdeveniments per unitat emprada: típicament, de temps). Coincideix amb la mitjana.

b) Per a una distribució Exponencial, el paràmetre és el valor λ o, més habitualment, el seu invers $1/\lambda$, que coincideix amb la mitjana. Si parlem d'una variable que mesura el temps, $1/\lambda$ seria el temps mitjà entre dos esdeveniments consecutius.

c) Per a una distribució normal, els dos paràmetres són la mitjana μ i la desviació tipus σ.

d) Per a una distribució uniforme, necessitem el mínim i el màxim valors possibles.

e) Per a una distribució triangular, si és simètrica, n'hi ha prou amb el mínim i el màxim. Si és asimètrica, necessitem també el valor de la moda (és el punt on se situa l'extrem del triangle, és a dir, el màxim de la funció de densitat).

2. Estem parlant d'una distribució uniforme entre 0 i *T*. Com que el valor mínim està fixat, l'únic paràmetre d'interès és el valor màxim (és bastant lògic que sigui així: el temps mínim de posicionament del lector del disc dur pot ser negligible i zero seria una opció adequada, encara que potser no és exacta, però no seria tan senzill trobar una proposta per al temps màxim que el lector pot trigar per situar-se, almenys sense conèixer millor altres paràmetres del funcionament del disc).

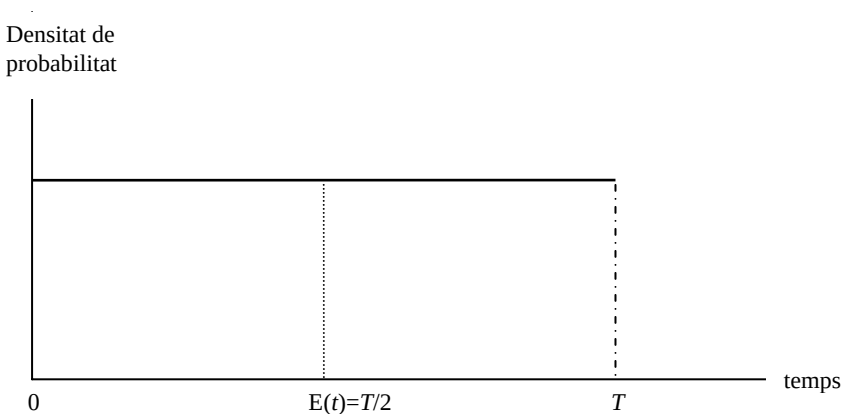


Figura 1.2 Distribució del temps d'accés, uniforme entre 0 i T

La població de referència seria el conjunt de tots els accessos que el lector del disc dur pot realitzar per accedir a dades escrites en ell. Un disc dur està estructurat com molts cilindres concèntrics, cada un dels quals pot encabir un gran nombre de sectors, que és l'espai on pot dirigir-se el lector. Per la informació de l'enunciat, hem de suposar que les dades es troben repartides pel disc de manera que els temps d'accés observats es distribuirien com a la figura 1.2. Podria donar-se una altra la situació: potser si el disc estigués menys ple seria més fàcil obtenir temps menors, i llavors la distribució no seria simètrica i farien falta uns altres paràmetres per precisar exactament com serien aquests temps.

Una mostra de temps consistiria en una selecció a l'atzar de n sectors que contenen dades, i un individu d'aquesta mostra seria un accés particular, del qual mesuraríem el temps amb l'instrument adequat. Encara que hem assumit que la variable temps és contínua, és segur que l'instrument de mesura no pot distingir entre temps molt semblants i donarà resultats amb precisió finita, per exemple, d'un mil·lisegon. Per exemple, una mostra de deu observacions podria donar-nos aquests valors (ms):

$$4,\quad 16,\quad 13,\quad 12,\quad 14,\quad 23,\quad 26,\quad 0,\quad 1,\quad 21.$$

El primer estimador de T que proposarem és bastant lògic: consisteix a agafar el valor més gran de la mostra:

$$t_1 = \text{màx}(x_1, x_2, ..., x_n)$$

La segona proposta que podem fer tampoc no és difícil. Observeu que la mitjana de la distribució està justament al mig: llavors, té sentit pensar que la mitjana mostral multiplicada per dos serà un valor semblant a T:

$$t_2 = 2\,\bar{x}$$

Seguint per la mateixa idea, com que la mediana també és la meitat de T, tenim un tercer estimador:

$$t_3 = 2\cdot\text{Med}(x_1, x_2, ..., x_n)$$

Però l'estimador més fiable de tots és el quart, perquè considera el primer i li afegeix una petita correcció, atès que t_1 sempre és més petit o igual que T:

$$t_4 = t_1\cdot(n+1)/n$$

(al capítol següent justificarem aquesta correcció). Vegem ara quines són les diverses estimacions que proporcionen els quatre estimadors proposats:

t_1	t_2	t_3	t_4
26	26	27	28.6

2 Principis d'inferència estadística

Al capítol anterior s'ha dit que la inferència estadística permet establir conclusions aplicables a la població des de la informació obtinguda d'una mostra. Els instruments principals són els estimadors, que calculem a partir de la mostra. Com que és un fet que els estimadors varien d'una mostra a una altra, en aquest tema es defineix formalment el concepte de *mostra aleatòria simple* (MAS), i s'estudien les propietats desitjables d'un estimador per saber quant oscil·la un estadístic a causa del procés de mostreig. D'aquesta manera, es podrà quantificar millor la relació informació-soroll que aporta un estadístic observat en una mostra concreta. Al llarg del tema, s'utilitza el cas de la mitjana per il·lustrar amb més claredat l'exposició de les idees.

2.1 Mostra aleatòria simple

En qualsevol text introductori d'estadística[3] es pot trobar una part dedicada al càlcul de probabilitats i a les variables aleatòries. Recordem que es defineix *variable aleatòria* (v.a.) com una aplicació X, des d'un conjunt d'esdeveniments Ω (que representa la població) a la recta dels reals $\mathbf{R}$, tal que a cada element ω_i li assignem un valor x_i de $\mathbf{R}$.

$$X\colon\ \Omega \rightarrow \mathbf{R}$$
$$\omega_i \rightarrow\ X(\omega_i) = x_i$$

Exemple. Com seria de fluida l'entrada a una gran ciutat si els vehicles portessin més ocupants? Tots hem patit una retenció de trànsit alguna vegada, si és que no la patim cada dia, i hem observat amb malestar els conductors que viatgen sols al seu cotxe (aquest malestar s'experimenta encara que nosaltres també siguem l'únic ocupant, i és especialment intens si viatgem en autobús: quan un fa l'esforç d'agafar el transport públic no li agrada quedar atrapat en una marea de vehicles infrautilitzats). Compteu molts cotxes amb un ocupant, uns quants amb dos, alguna vegada en veieu un amb tres persones... Bé, doncs teniu la sort d'estar observant com es distribueix una variable aleatòria! Es tractaria de la v.a. "ocupació d'un vehicle a l'entrada de la ciutat". El nom que donem a una v.a. hauria

[3] Per exemple: Daniel Peña, *Fundamentos de estadística*, Alianza Editorial, Madrid, 2001.

d'identificar-la plenament, és a dir, hauria de referenciar sense ambigüitat la població i la forma en què cada individu —un vehicle— d'aquesta s'associa a un nombre real. A la pràctica, precisar la població de partida és molt complicat. A quin conjunt de vehicles ens referim?

- A tots els vehicles o només als privats?
- Als que entren per una via concreta o per totes les entrades?
- Als que entren en hora punta o a qualsevol hora?
- I quin seria el període que anomenem "hora punta"?

I encara hi ha més matisos per considerar. Les característiques que vulguem incloure tenen una repercussió immediata en el procés d'observació d'una mostra, ja que condicionen la selecció d'un vehicle (per exemple, una furgoneta es podria elegir o no, depenent de si s'hi inclouen només els turismes).

Per simplicitat, prenguem vehicles fins a cinc places i suposem que ja tenim clars altres condicionaments que cal tenir en compte. Aquesta variable s'assigna únicament als valors 1, 2, 3, 4 i 5. Com trobareu al text que hagueu consultat, el nombre d'ocupants (X) és una variable aleatòria discreta i té una funció de probabilitat, $p_X(\cdot)$, que assigna a cada valor possible una probabilitat. Per exemple, $p_X(1)$ seria la probabilitat que un vehicle a l'atzar portés solament el conductor. El coneixement exacte de la funció de probabilitat és un ideal, ja que portaria a conèixer perfectament els paràmetres de la població que ens interessen, i el procés d'inferència seria innecessari. Així, el nombre esperat d'ocupants en cada vehicle seria calculable com:

$$\mu = \sum_{x=1}^{5} x\, p_X(x)$$

Recordeu que μ també es pot escriure acompanyat del símbol de la v.a. a què fa referència: μ_X. Exactament igual succeeix amb la variància: $V(X) = \sigma^2 = \sigma_X^2$. Quan no hi pugui haver confusió, normalment ometrem el símbol de la v.a. Després de repassar el concepte de variable aleatòria, estem en condicions de definir el de mostra aleatòria. Per a gran part de les aplicacions de l'estadística, és suficient definir la *mostra aleatòria simple*, o de forma abreujada, MAS.

Sigui la v.a. X, definida a Ω. Anomenem MAS de grandària n de la v.a. X la funció vectorial $M = (X_1, X_2, \ldots, X_n)$:

$$\begin{array}{lccc} M: & \Omega^n & \rightarrow & R^n \\ & \omega = (\omega_1, \omega_2, \ldots, \omega_n) & \rightarrow & M(\omega) = (X_1, X_2, \ldots, X_n) \end{array}$$

$X_1, \ldots, X_n$ tenen la mateixa distribució de probabilitat que X i són estadísticament independents entre si. Aquestes propietats es compleixen si cada vegada que observem la població per extreure'n un individu no es modifica la probabilitat que cadascun dels individus té de ser elegit. Típicament, això s'aconsegueix si hi ha reposició després de cada extracció (l'individu torna al conjunt d'elegibles, igual que abans de l'observació) o si la població és molt gran (la supressió d'un individu no afecta substancialment les probabilitats dels altres membres de la població).

Nota. És habitual que a cada individu de la població li assignem la mateixa probabilitat de ser escollit. Això no ha de significar necessàriament que la variable aleatòria X hagi de ser equiprobable en els valors que adopta. Per exemple, adoptem un sistema que permet que qualsevol vehicle de la població sigui elegit, amb la mateixa probabilitat, per comptar el nombre d'ocupants; però un sistema com aquest no obliga que hi hagi la mateixa proporció dels valors 'un ocupant', 'dos ocupants', ..., 'cinc ocupants'.

La mostra reprodueix n vegades el comportament de la variable i, si cada vehicle que s'observa s'escull de forma independent de la resta, el valor resultant tampoc no té relació amb els altres. En el nostre cas, sembla bastant raonable que els vehicles que ocupen la via pública no tenen res a veure entre ells (excepte si formen una caravana per algun esdeveniment especial, com ara un casament); llavors, és fàcil suposar que les ocupacions que veiem siguin independents. Però si el mecanisme de selecció és subjectiu i dóna alguna llibertat a l'observador per elegir el vehicle, aquest podria posar en marxa un procés "compensador" i *buscar* cotxes una mica plens si ha observat massa valors baixos, o viceversa. D'aquesta manera, es trencaria la suposició que la mostra es compon de valors independents.

Tinguem en compte que seleccionar individus per formar una MAS pot resultar complex, especialment si no disposem d'una definició operativa de la població. La forma més simple de selecció es dóna quan tenim una llista exhaustiva de tots els individus. Llavors, podem ordenar-los, donar-los un número d'1 a N i extraure'n aleatòriament números repartits uniformement entre 1 i N. Però sovint aquesta llista no es pot formar, o és incorrecta: per exemple, cap ciutat important no pot mantenir al dia un cens dels seus habitants, ja que la dinàmica d'aquests en fa variar contínuament la població, sense mencionar tots els individus en situació irregular que no es poden tenir en compte, i els errors que s'hi introdueixen inevitablement.

També podem trobar-nos davant de poblacions infinites, per les quals és impossible construir una llista. Llavors, hem d'utilitzar algun mètode de selecció basat en una correspondència existent entre els individus i elements que puguem generar fàcilment a l'atzar: per exemple, números en l'interval [0, 1], o una combinació de diversos elements, com nombres enters i reals.

Exemple. Com que no és creïble disposar d'una llista de tots els vehicles que entren a la ciutat, hem de dissenyar un sistema que ens permeti *innocentment* escollir vehicles. Una possibilitat és: donada una seqüència aleatòria de números 0 o 1 (que es coneix com a *procés de Bernoulli*), assignar cada número a un vehicle i obtenir el nombre d'ocupants pels vehicles que s'han marcat amb un 1. Un altre procediment seria: cada 30 segons, elegir el primer vehicle que passi davant del lloc d'observació amb una matrícula que acabi amb el número zero (un factor independent del nombre d'ocupants). Aquest procediment no es basa en l'atzar, perquè és sistemàtic, però és vàlid si es pot considerar que els vehicles arriben en ordre aleatori.

Contraexemple. Si volem estimar el nombre d'usuaris en un equip multiusuari, no hem de mesurar cada 30 segons, sinó amb una cadència molt superior, ja que un usuari normalment estableix una connexió per un període prolongat. D'aquesta manera, ens assegurem que observem el sistema en condicions suficientment diferents.

En endavant utilitzarem aquesta definició de MAS per aprofundir el concepte d'estimador que s'ha introduït al capítol 1, i la relació indissoluble entre estimador i paràmetre. Concretament, ens referirem a la qüestió essencial del problema de la inferència: com estimar característiques de la població basant-se en les dades recollides en una mostra.

2.2 Distribució de la mitjana mostral

Des del moment que es pot entendre que una mostra (X_1, ..., X_n) té caràcter aleatori, és vàlid atribuir a qualsevol estadístic que se'n pugui derivar les propietats d'una variable aleatòria, com posseir una distribució de probabilitat, esperança o variància.[4] Concretament, tractarem l'estadístic $\overline{X}$, per la seva importància com a estimador del paràmetre de la tendència central de la variable X, μ. És fàcil imaginar com la mitjana mostral varia d'una mostra a una altra prenent diferents valors.

Exemple. La qüestió de quantes persones entren a la ciutat en un cotxe, per una via determinada, entre les 8 i les 9 del matí, és una pregunta difícil. Generalment, no se'n busca la resposta exacta (12708!), però l'aproximació que necessàriament podrem donar tampoc no ha de ser molt grollera. Com que hi ha dispositius que poden comptar automàticament el nombre de vehicles que passen per un detector, en suposem un número arbitrari, que pot ser determinat cada vegada que es fan les observacions: per exemple, 6000 cotxes per hora. La resposta més *naïf* seria: com que cada cotxe pot portar entre 1 i 5 passatgers, en una hora entraran entre 6000 i 30000 persones. Òbviament, és una informació molt poc precisa, i no hem tingut en compte cap observació real.

Si disposem d'una mostra de 12 vehicles escollits a l'atzar, podem observar valors com:

$$1, 2, 1, 1, 1, 3, 1, 1, 2, 1, 1, 1,$$

i es pot fer la deducció següent: suposant que la "realitat" es comportés com la mostra, hi hauria 4500 vehicles (9 de 12, multiplicat per 6000) amb un sol passatger, 1000 (2 de 12, per 6000) amb dos ocupants i 500 amb tres ocupants. Total: 8000 persones que entren en aquest temps. També es pot obtenir el resultat trobant l'ocupació mitjana (mostral) per cotxe i fent el producte amb els 6000 cotxes:[5]

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{16}{12} = 1.333...$$
$$6000 \cdot 1.333... = 8000$$

Sembla un valor massa precís: estem dient que no són 7999 ni 8001 les persones que entraran de 8 a 9! En realitat, veiem ràpidament que donem una falsa sensació d'exactitud, ja que és fàcil que una altra mostra amb la mateixa grandària ens doni un resultat diferent.

---

4 Si és que matemàticament existeixen; recordem que, per a v.a. contínues, aquests valors s'obtenen resolent una integral definida, que pot no tenir solució.

5 Atenció: utilitzem el símbol $\bar{x}$ en minúscula perquè aquí designa el nombre real observat, no un valor aleatori, que designarem amb lletra majúscula, $\overline{X}$, com qualsevol variable aleatòria. Aquesta distinció, a vegades, és molt subtil.

Possiblement, és una referència vàlida per tenir idea de la magnitud de la resposta (més aviat un valor petit que no proper a 30000), però ens falta informació que ens pugui orientar sobre la importància de la desviació comesa respecte del valor correcte. És un error de desenes de persones? De centenars? De milers?

Així, les mesures aniran variant d'una mostra a una altra. Si volem utilitzar l'estadístic *mitjana* com a estimador del paràmetre poblacional *esperança*, aquesta variabilitat ens induirà a errors que, per descomptat, mai no són desitjables. Ara bé, podem quantificar aquests errors? Almenys, podem delimitar-ne la magnitud? Per tal de respondre aquestes dues preguntes crucials, abans n'hem de respondre altres de més senzilles:

a) Al voltant de quin valor varien les mitjanes mostrals?
b) Varien molt o poc al voltant d'aquest valor?
c) Quina forma té la distribució de $\overline{X}$?

S'ha de tenir en compte que, una vegada acceptat que l'estadístic *mitjana* o *mitjana mostral* $\overline{X}$ té una distribució determinda, les dues primeres preguntes es redueixen a calcular l'esperança i la variància de l'estadístic $\overline{X}$. Calculem-les.

Repàs**:** important per seguir sense problemes les demostracions que vénen a continuació.

$E(aX) = a \cdot E(X)$, on a representa una constant.
$E(X + Y) = E(X) + E(Y)$
$V(aX) = a^2 \cdot V(X)$
$V(X + Y) = V(X) + V(Y)$, si X i Y són independents.

2.2.1 Esperança de la mitjana mostral

Donada una MAS, $(X_1, ..., X_n)$, calculem el "centre" de la distribució de l'estadístic $\overline{X}$:

$$\begin{aligned} E(\overline{X}) &= E(\Sigma X_i / n) = \\ &= [E(\Sigma X_i)] / n = \\ &= [\Sigma E(X_i)] / n = \\ &= [\Sigma E(X)] / n = \text{ [ja que les } X_i \text{ són v.a. idènticament distribuïdes]} \\ &= [n E(X)] / n = E(X) = \mu \end{aligned}$$

Per tant, el centre de l'estadístic $\overline{X}$, que representem per $E(\overline{X})$, coincideix amb el centre de la v.a. X, que representem per $E(X) = \mu$. És a dir: per a qualsevol v.a., la distribució dels valors de la mitjana que es poden obtenir amb totes les mostres possibles es troba centrada en el valor del paràmetre que es pretén estimar. També es pot dir que té per esperança el valor del paràmetre desconegut.

Té alguna utilitat aquesta coincidència? Suposem que fem servir una mitjana mostral ($\bar{x}$) com a estimador de l'esperança matemàtica (μ). Els nostres errors poden ser tant per excés com per defecte. Però, atès que el centre de la variable $\overline{X}$ coincideix amb l'objectiu de la nostra estimació, el conjunt dels nostres possibles errors positius i negatius està equilibrat.

Exemple. Per comprendre millor el procés inductiu que estem seguint, suposem el punt de vista oposat: coneixem exactament la distribució de l'ocupació dels vehicles en les condicions esmentades, i estudiem què es pot "deduir" d'aquesta informació. La v.a. X, que designa quants passatgers viatgen en un vehicle escollit a l'atzar, tindrà la funció de probabilitat següent (a la figura 2.1 es veu gràficament com la probabilitat es distribueix segons els valors considerats):

Taula 2.1 Distribució de probabilitat de la v.a. X

x_i	$p_X(x_i)$
1	0.60
2	0.25
3	0.09
4	0.04
5	0.02

Recordem que aquesta informació no ens serveix per "endevinar" el que ha de succeir, com el nombre d'ocupants del cotxe que passarà dintre de 30 segons, encara que coneguéssim la composició de tota la població, perquè quan l'atzar hi intervé fa imprevisibles els esdeveniments futurs. Aquesta informació privilegiada és útil per calcular probabilitats exactes de qualsevol esdeveniment, o per trobar el valor d'un paràmetre de referència, com el valor esperat. Tal com hem expressat prèviament, aquest és:

Taula 2.2 Càlcul de E(X)

x_i	$p_X(x_i)$	$x_i \cdot p_X(x_i)$
1	0.60	0.60
2	0.25	0.50
3	0.09	0.27
4	0.04	0.16
5	0.02	0.10
		$E(X) = 1.63$

Una mostra concreta representa els valors observats en les proporcions, més o menys, que dóna la funció de probabilitat, i difícilment coincideixen plenament. Passa el mateix quan llancem una moneda no carregada: hem d'esperar igual nombre de cares que de creus, però rarament és així, encara que la diferència sol ser petita. També hem d'esperar que les proporcions mostrals dels valors que la v.a. adopta siguin semblants als que surten a la funció de probabilitat (que podem anomenar *proporcions poblacionals* o *probabilitats*).

Respecte a la mitjana dels ocupants per vehicle, hem trobat que l'esperança és 1.63 (no farem acudits sobre quants dits, cames o braços ha de tenir el segon passatger): si fossin 100 cotxes, l'ocupació esperada seria de 163 persones, distribuïts com sigui. Per a 6000

cotxes, parlem d'unes 10000 persones (9780, exactament). La mitjana de la mostra anterior era de 1.333, i hem de tenir clar que una altra mostra de 12 vehicles podria donar un valor diferent. (Punt de reflexió: podríem observar en una mostra d'aquesta grandària 20 persones? I 30? I 40?).[6]

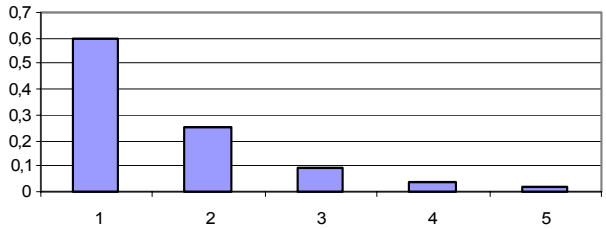


Figura 2.1 Distribució de l'ocupació d'un vehicle (nombre de persones)

Si escollim dos cotxes i sumem el nombre d'ocupants que observem en ambdós vehicles tindrem un valor que estarà entre 2 i 10. Podem veure aquest valor com una v.a. que correspon a aquest procediment, o com un estadístic "suma" aplicat a una mostra de grandària n=2. En qualsevol cas, és important precisar que els dos vehicles són independents entre si o que la mostra és MAS. D'aquesta manera, el càlcul de les probabilitats de cada cas és senzill. Per exemple, sigui Y_2 la variable "suma d'ocupants de dos vehicles"; concretament, si X_1 correspon al primer vehicle i X_2 al segon, llavors $Y_2 = X_1 + X_2$. La probabilitat d'observar el valor 2 és:

$$\begin{aligned} P(Y_2 = 2) &= P(X_1 = 1 \cap X_2 = 1) \qquad \text{[com que els dos cotxes són independents...]} \\ &= P(X_1 = 1) \cdot P(X_2 = 1) \\ &= 0.60 \cdot 0.60 = 0.36 \end{aligned}$$

Altres valors, tots excepte el 10, han de considerar diverses possibilitats mútuament excloents (com que no tenen intersecció, la probabilitat del conjunt es descompon en suma de probabilitats). Trobem la probabilitat d'observar el valor 3:

$$\begin{aligned} P(Y_2 = 3) &= P((X_1 = 1 \cap X_2 = 2) \cup (X_1 = 2 \cap X_2 = 1)) \\ &= P(X_1 = 1 \cap X_2 = 2) + P(X_1 = 2 \cap X_2 = 1) \\ &= 0.60 \cdot 0.25 + 0.25 \cdot 0.60 = 0.30 \end{aligned}$$

Taula 2.3 Distribució de probabilitat de la v.a. Y2

y_i	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$p_{Y_2}(y_i)$	0.36	0.30	0.1705	0.093	0.0521	0.0172	0.0052	0.0016	0.0004

La taula 2.3 ens mostra totes les probabilitats com a funció de probabilitat de la v.a. Y_2, i la figura 2.2 mostra la disposició de la probabilitat dels valors considerats. S'aprecia clarament la poca rellevància d'alguns d'ells, situats a la dreta de la distribució.

[6] Cal tenir en compte que 20/12 suposa una mitjana de 1.667, que és molt semblant al valor de l'esperança, però 40/12 val 3.333, que és massa allunyat. És molt difícil que trobem 12 cotxes tan plens.

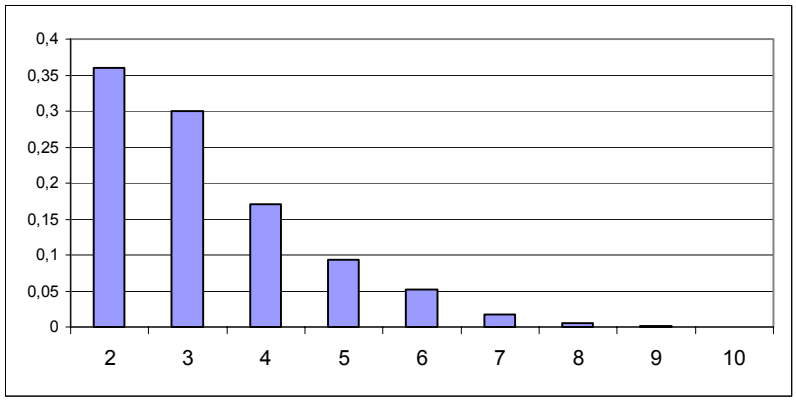


Figura 2.2 Distribució de l'ocupació de dos vehicles (nombre de persones)

Amb una mica de paciència i una eina per fer els càlculs, és possible fer el mateix i trobar les distribucions de la suma de 3, 4, 5 o 6 cotxes (respectivament, les variables Y_3, Y_4, Y_5 i Y_6). A la taula 2.4, veiem la funció $\mathrm{p}_{Y_6}(y_i)$:

Taula 2.4 Distribució de probabilitat de la v.a. Y_6

y_i	$\mathrm{p}_{Y_6}(y_i)$	y_i	$\mathrm{p}_{Y_6}(y_i)$	y_i	$\mathrm{p}_{Y_6}(y_i)$	y_i	$\mathrm{p}_{Y_6}(y_i)$	y_i	$\mathrm{p}_{Y_6}(y_i)$
6	0.0467	11	0.126	16	0.0093	21	0.0001	26	$2\cdot10^{-7}$
7	0.1166	12	0.0893	17	0.0044	22	$3\cdot10^{-5}$	27	$3\cdot10^{-8}$
8	0.1635	13	0.0573	18	0.0019	23	$1\cdot10^{-5}$	28	$6\cdot10^{-9}$
9	0.1736	14	0.0339	19	0.0008	24	$3\cdot10^{-6}$	29	$8\cdot10^{-10}$
10	0.158	15	0.0184	20	0.0003	25	$7\cdot10^{-7}$	30	$6\cdot10^{-11}$

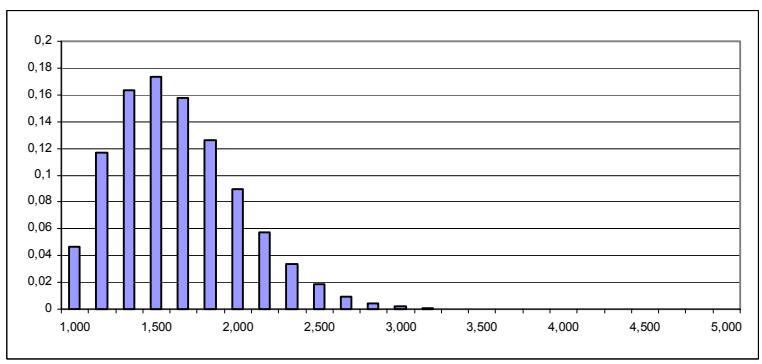


Figura 2.3 Distribució de l'ocupació mitjana de sis vehicles (nombre de persones/cotxe)

Observeu que, en una mostra de sis cotxes, és difícil comptar més de quinze persones. Si parlem de mitjanes, això vol dir que rarament veurem un promig superior a 2.5 (15/6) persones per cotxe. Penseu que és molt senzill representar gràficament la distribució de les mitjanes (de totes les mitjanes de mostres de sis cotxes): simplement, hem de dividir y_i per 6. Examineu el diagrama de la figura 2.3.

Tornem a la idea de variabilitat. La distribució de la variable mitjana mostral $\overline{X}$ fa patent que unes vegades trobarem valors alts i, d'altres, valors més baixos, igual que de vegades veiem cotxes més plens o més buits. Si estudiem la figura 2.3 amb deteniment, apreciarem que el "centre" de la distribució ha d'estar (ho està!) just al valor de μ, 1.63. És innegable que cada mostra real de sis cotxes ens retorna una estimació més o menys propera, però errònia.

Veiem que les nostres estimacions, encara que es distribueixin al voltant del paràmetre d'interès, no acaben d'encertar. De fet, si treballem amb una variable contínua, de la recta dels reals (imaginem que té tants decimals com volem), el valor estimat mai no coincidirà amb el paràmetre d'interès: sempre ens equivocarem. Quin horror! Ara bé, quant ens equivoquem? És tolerable la magnitud de l'error d'acord amb els nostres objectius? Això ens porta a la pregunta següent: com podem mesurar o quantificar aquest error? A continuació, veurem si la variància d' $\overline{X}$ ens aporta informació sobre aquest error.

2.2.2 Variància de la mitjana mostral

Què representa la variància d'una variable? La seva dispersió. A l'exemple anterior hem vist que la mitjana mostral s'allunya menys de μ que un valor individual (la dispersió serà menor perquè les desviacions de la mitjana es compensen). Per tant, hauria de tenir una variància més petita que la de X. Vegem ara com s'expressa la variància de $\overline{X}$. Per definició,

$$\mathrm{V}\left(\overline{X}\right) = \mathrm{E}\left[\overline{X} - \mathrm{E}\left(\overline{X}\right)\right]^2 = \mathrm{E}\left[\overline{X} - \mu\right]^2$$

Per tant, la variància de $\overline{X}$ valora l'error que es cometria en cada mostra en estimar el paràmetre poblacional $\mathrm{E}(X) = \mu$. La variància quantifica la magnitud de la mitjana de l'error elevat al quadrat, i d'aquesta manera es té una mesura de quin és en promig la desviació d'una estimació puntual.
Calculem la "dispersió" teòrica de $\overline{X}$:

$$\begin{aligned} \mathrm{V}(\overline{X}) &= \mathrm{V}(\Sigma X_i / n) = \\ &= [\,\mathrm{V}(\Sigma X_i)\,] / n^2 = \\ &= [\,\Sigma\,\mathrm{V}(X_i)\,] / n^2 = \end{aligned}$$

Cal recordar que el pas que acabem de fer (la variància d'una suma com a suma de variàncies) es justifica per la independència de les variables. Si no fos així, caldria afegir-hi els termes corresponents a les covariàncies entre les X_i.

$$\begin{aligned} &= [\,\Sigma\,\mathrm{V}(X)\,] / n^2 = \\ &\quad [\text{ja que les } X_i \text{ v.a. són idènticament distribuïdes}] \\ &= [\,n\,\mathrm{V}(X)\,] / n^2 = \\ &= \mathrm{V}(X) / n = \sigma^2 / n \end{aligned}$$

Aquest resultat és molt rellevant. En primer lloc, s'interpreta fàcilment que la dispersió de la variable original X està relacionada directament amb la dispersió de la mitjana mostral. La segona part d'aquesta fórmula ens diu que la variància de $\overline{X}$ és inversament proporcional a la grandària de la mostra.

No és una gran sorpresa: ja sabíem que, quant més gran és la grandària de la mostra, més informació aporta sobre la població i, per tant, hi ha menys probabilitat que $\overline{X}$ s'allunyi del paràmetre $E(X) = \mu$.

Exemple. El temps que els PC de l'aula S03 triguen a carregar el sistema operatiu té una variància de 250 segons^2. Si, per conèixer el temps mesurat que trigaran, obtenim una mostra en 10 ordinadors, la variància de la variable mitjana del temps que triguen 10 PC val:

$$V(\overline{X}) = V(X)/n = 250\ s^2 / 10 = 25\ s^2$$

En conseqüència, l'error (quadrat) que s'ha d'esperar de la nostra observació és de 25 segons (quadrats).

De la mateixa forma que definim la *desviació tipus*[7] com l'arrel quadrada de la variància d'una v.a. qualsevol, ara definirem l'*error tipus*[8] com l'arrel de la variància de la variable mitjana mostral; d'aquesta manera, treballem amb un valor que té les mateixes unitats que la v.a. original:

$$\sigma_{\overline{X}} = \sqrt{V(\overline{X})} = \sigma_X / \sqrt{n}$$

Exemple (continuació). L'error tipus de la nostra estimació de la mesura del temps que els PC de l'aula S03 trigaran a carregar el sistema operatiu serà:

$$\sigma_{\overline{X}} = \sqrt{V(\overline{X})} = \sqrt{25\ s^2} = 5\,s$$

És a dir, el valor de la mitjana que obtinguem amb una mostra de deu PC estarà afectat per un error tipus d'estimació de 5 segons.

S'ha de tenir en compte que, mentre el terme *desviació tipus* no té cap connotació positiva ni negativa, l'*error tipus* ja deixa clar, des del primer moment, que es tracta d'alguna cosa no desitjable: l'error que es pot cometre en estimar la mitjana poblacional a partir de la mitjana mostral.

Nota. Amb dades mostrals, podem estimar l'error tipus mitjançant l'estimador *s*, dividit per l'arrel de *n*. Mireu els exercicis 1 i 2 al final d'aquest capítol.

---

7 La denominació en castellà és *desviación típica* i en anglès, *standard deviation.*

8 La denominació en castellà és *error típico* i en anglès *standard error*.

Exemple. Calculem la variància i la desviació tipus de l'ocupació d'un vehicle.

L'arrel quadrada de V(X) (taula 2.6) és 0.9450. Destaquem que, com que la variància de la mitjana és inversament proporcional a la grandària de la mostra n, l'error tipus disminueix amb l'arrel quadrada de n. Això també significa que la desviació típica de la suma de n variables independents i idènticament distribuïdes augmenta amb l'arrel quadrada de n.

Taula 2.6 Càlcul de V(X); recordem que $\mu = 1.63$

x_i	$p_X(x_i)$	$(x_i - \mu)^2$	$(x_i - \mu)^2 \cdot p_X(x_i)$
1	0.60	0.3969	0.23814
2	0.25	0.1369	0.034225
3	0.09	1.8769	0.168921
4	0.04	5.6169	0.224676
5	0.02	11.3569	0.227138
			V(X) = 0.8931

Taula 2.7 Paràmetres de suma (ocupació total) i mitjana depenent de n

Nombre de cotxes	Ocupació total esperada	Desviació tipus	Ocupació esperada de $\overline{X}$	Error tipus
1	1.63	0.9450	1.63	0.9450
10	16.3	2.988	1.63	0.2988
100	163	9.450	1.63	0.0945
1000	1630	29.88	1.63	0.02988
10000	16300	94.50	1.63	0.009450
n	1.63 n	0.9450 $\sqrt{n}$	1.63	0.9450/$\sqrt{n}$

La conclusió a la qual arribaríem de la lectura de la taula 2.7 seria: amb mostres grans, l'ocupació total presenta una desviació tipus que creix més lentament que la tendència central (quan una creix cent vegades, l'altra solament ho fa deu vegades), cosa que permet relativitzar la importància de la pertorbació aleatòria que afecti l'estimació realitzada, perquè un error d'unes 100 persones respecte de 16300 és millor que un error de 10 respecte de 163.

Per la mateixa raó, una mitjana obtinguda amb una mostra gran està menys afectada per pertorbacions aleatòries i ofereix una estimació de μ més precisa. En concret, amb 10000 observacions podríem comprovar que la diferència entre la mitjana mostral i el valor poblacional és típicament de l'ordre d'una centèsima (0.009450 és quasi 0.01).

Ara, podem tornar a la realitat de l'exemple plantejat. No sabem ni sabrem mai la distribució autèntica de la variable que ens ocupa, però en tenim una pista gràcies al concepte d'error tipus. És veritat que tampoc no coneixem el paràmetre σ, però podem fer una estimació de la desviació tipus a partir de la mostra, i llavors tindrem una idea aproximada de l'error probable.

2.2.3 Forma de la distribució de la mitjana mostral

Ja sabem quins són el centre i la dispersió de la variable $\overline{X}$. No obstant això, aquesta informació és insuficient per fer estimacions que considerin la incertesa de la mostra: per exemple, una estimació amb un petit marge d'error, determinat per l'experimentador (treballar amb confiança o fiabilitat 100% és impossible, excepte si donem estimacions tan òbvies que no tinguin cap interès). Veurem que els càlculs necessaris per incorporar aquest aspecte requereixen calcular probabilitats, és a dir, conèixer la forma de la distribució de la mitjana mostral.

Exemple. Quins valors pren la mitjana de sis cotxes, en el 99% dels casos? Si examinem la taula 2.4, hauríem de refer la pregunta i dir: quins valors de la cua superior acumulen una probabilitat de 0.01? La millor aproximació correspon als valors del 16 fins al 30, que suposen una proporció de 0.75%. En termes de mitjanes, direm que 99 de cada 100 vegades la mitjana mostral val entre 1 i 2.5, perquè en un 99% dels casos l'ocupació de sis cotxes oscil·la entre 6 i 15. Però, és clar, en aquest cas coneixem la distribució de la mitjana. Si la mostra tingués grandària 12, la resposta no seria trivial.

Per sort, en determinades condicions, la distribució de probabilitat de la mitjana mostral és coneguda o, almenys, es pot aproximar per una llei de probabilitat coneguda. L'enunciat d'aquest fet capital s'anomena *teorema del límit central* o TLC, del qual donem una versió adaptada per al cas de MAS:

Teorema. Siguin X_1, ..., X_n n v.a. independents i amb la mateixa distribució, v.a.i.i.d, essent $E(X_i)=\mu$ i $V(X_i)=\sigma^2$. Llavors, la variable centrada i reduïda:

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

tendeix a seguir la distribució de probabilitat N(0,1) quan la n tendeix a infinit. És a dir:

$$\lim_{n\to\infty} P(Z_n \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Recordem que s'entén per variable centrada i reduïda la que té mitjana 0 i desviació tipus 1, i que la distribució N(0,1) fa referència a la llei de Laplace-Gauss, més coneguda per normal, i que segueix la coneguda forma de la campana de Gauss.

S'ha de tenir en compte que hem parlat de variables aleatòries independents i idènticament distribuïdes (v.a.i.i.d.), però no hem dit res sobre quina distribució segueixen. És a dir, el resultat que la suma de v.a.i.i.d. convergeix fins a una normal succeeix, sigui quina sigui la distribució que cada una de les v.a. segueixi. També és cert que la semblança amb la normal succeeix amb menor grandària mostral com més s'acosta la distribució de la v.a. original a la normal.

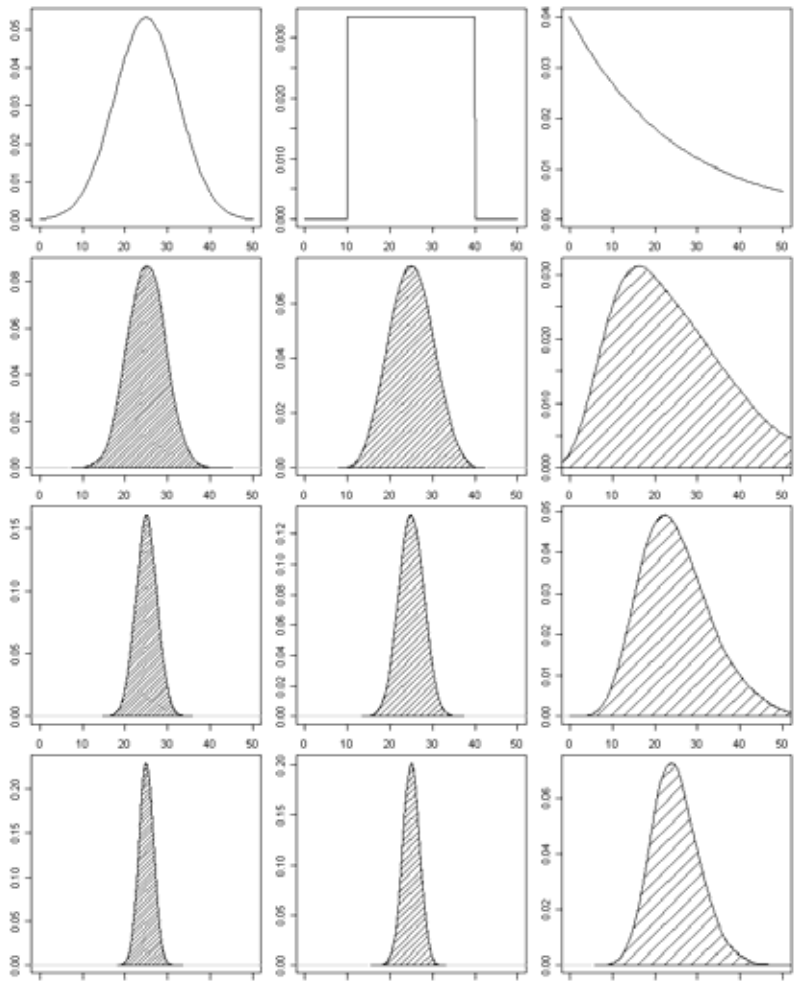


Figura 2.4 Il·lustració del TLC per diverses distribucions. En vertical: la distribució a la població, la mitjana de 3 observacions, la de 9 i la de 20. Observeu que (1) la mitjana mostral acaba prenent la forma de campana en tots els casos i (2) en distribucions asimètriques la convergència a la normal és més lenta.

La figura 2.4 il·lustra perfectament aquesta situació. Les files marquen la grandària de la mostra, prenent n = 1, 3, 9 i 20. La primera fila, per a una mostra de grandària unitària, coincideix amb la forma de la variable aleatòria a la població. Les columnes marquen tres lleis diferents. La primera columna representa una variable aleatòria amb distribució normal; la segona, una uniforme, i la tercera té una distribució asimètrica (exponencial).

En el cas de la v.a. amb distribució normal (primera columna), es pot veure que, per a qualsevol grandària de la mostra, la distribució de la variable mitjana mostral $\overline{X}$ és sempre normal. En canvi,

per a les altres dues variables, es necessita una grandària mostral mínima, superior en el cas de l'asimètrica, perquè la semblança amb la normal es pugui considerar "acceptable".

Com es pot observar, aquest gràfic també mostra les propietats que hem vist anteriorment: a mesura que la grandària de la mostra augmenta, la variància es va fent més petita i es concentra al voltant d'una esperança que sempre coincideix amb l'esperança de la variable original.

Exercici. Connecteu-vos a la pàgina web:

<http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/>

i proveu l'applet que hi trobareu. Simulant l'extracció de moltes observacions d'una població hipotètica (representada per una distribució de probabilitat, que podeu canviar), veureu com es distribueix la mitjana mostral amb diferents mides de mostra (preneu els valors n = 2, 5, 10 i 25).

En resum, el teorema del límit central estableix que, si es prenen mostres de grandària n d'una població de mitjana μ i desviació típica σ, a mesura que n creix la distribució de $\overline{X}$ s'aproxima a la d'una normal de mitjana μ i desviació típica $\sigma/\sqrt{n}$.

$$\overline{X} \sim \mathrm{N}\,(\mu,\ \sigma/\sqrt{n})$$

Llavors:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

Taula 2.8 Valors crítics d'una v.a. N(0,1)

α	$1-\alpha$	$z_{\alpha/2}$	$z_{1-\alpha/2}$
0.001	0.999	–3.291	3.291
0.01	0.99	–2.576	2.576
0.05	0.95	–1.960	1.960
0.10	0.90	–1.645	1.645

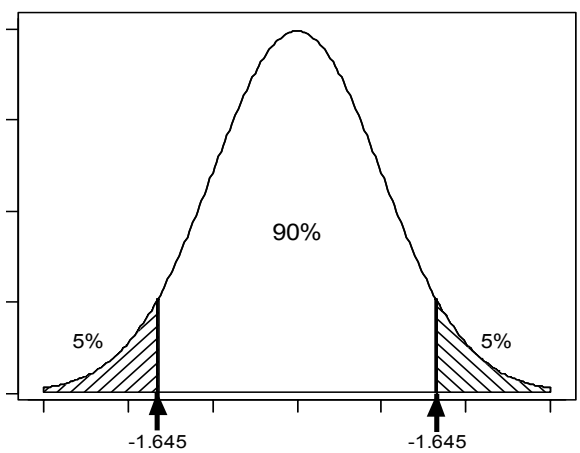


Figura 2.5 La probabilitat α (0.10) se situa als extrems de la distribució i deixa entre els punts $z_{\alpha/2}$ i $z_{1-\alpha/2}$ una probabilitat $1-\alpha$ (0.90).

Així doncs, sabem quant i com varia $\overline{X}$. Atès que en coneixem la distribució, podrem definir intervals que tinguin una certa probabilitat $1-\alpha$ de contenir la mitjana mostral $\overline{X}$:

$$\mathrm{P}\left(\mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

on $z_{\alpha/2}$ és el valor que té una probabilitat de $\alpha/2$ de quedar per damunt d'un valor observat d'una distribució N(0,1). El valor $z_{1-\alpha/2}$ s'interpreta igual, però s'ha de notar que, com que es tracta de la probabilitat complementària, és el valor $z_{\alpha/2}$ amb signe positiu. Vegem-ne uns exemples i a la figura 2.5 una il·lustració d'aquest fet.

Exemple. Suposem que recollim una mostra de 100 vehicles, i que la mitjana mostral amb aquesta grandària es pot suposar que és normal, a tots els efectes. Volem construir, per als valors de α de la taula 2.8, els intervals que continguin les proporcions respectives de les possibles mitjanes.

- El 99.9% de les mitjanes cau a $1.63 \pm 3.291 \cdot 0.945/10 = [1.32, 1.94]$
- El 99% de les mitjanes cau a $1.63 \pm 2.576 \cdot 0.945/10 = [1.39, 1.87]$
- El 95% de les mitjanes cau a $1.63 \pm 1.960 \cdot 0.945/10 = [1.44, 1.82]$
- El 90% de les mitjanes cau a $1.63 \pm 1.645 \cdot 0.945/10 = [1.47, 1.79]$

2.2.4 Quina grandària mostral és necessària per al teorema del límit central?

Depèn de la distribució de la variable original, que afecta a la *velocitat* la qual la distribució de la mitjana s'apropa a la de la normal. Això fa que diferents llibres de text recomanin diferents mides mostrals. No hi ha valors crítics que actuïn com a frontera, però podem establir uns criteris raonables per considerar com a vàlida una aproximació a la distribució normal:

1) Si X és normal $\Rightarrow$ $\overline{X}$ ~ N per a qualsevol n,

   ja que una combinació lineal de v.a. normals independents sempre és normal.

2) Si X és contínua $\Rightarrow$ $\overline{X}$ ~ N si $n > 30$

   Tenint en compte que com s'ha vist als gràfics anteriors, com més s'assembla X a una llei normal, abans s'arriba a una aproximació acceptable. Així, si X fos uniforme, una mostra de grandària 20 seria suficient. Però si X fos extremament asimètrica, seria prudent que el valor de n estigués una mica per damunt de 30.

3) Si X és discreta $\Rightarrow$ $\overline{X}$ ~ N si $n >> 30$

   A part del factor *simetria* de la distribució, que afecta igual que les v.a. contínues, l'altre factor rellevant per a la grandària mínima de la mostra és la quantitat de valors diferents que pot prendre X amb probabilitat significativa (perquè algunes lleis prenen teòricament

tots els valors naturals, com la de Poisson o la geomètrica, però no tots tenen possibilitats reals de ser observats alguna vegada). Quant menys valors pugui prendre una v.a., més lenta és la convergència de la mitjana a la distribució normal.

Exemple. S'assumeix que el temps transcorregut entre dues fallades del sistema segueix una distribució asimètrica per la dreta, molt semblant a l'exponencial negativa. Si obtenim la mitjana de 50 d'aquests temps, la seva distribució serà normal.

Exemple. Definim l'indicador de Bernoulli X per marcar si un espai està lliure (0) o ocupat (1). N'obtenim una mostra de 1000 espais. La variable mitjana de l'indicador tindrà una distribució normal. (Punt de reflexió: quin significat té aquesta mitjana?)

Observació. El paradigma de distribució difícil —convergència lenta fins a assemblar-se suficientment a una normal— és una llei de Bernoulli (discreta i només de dos valors: 0 i 1) amb paràmetre π molt proper a 0 o a 1 (forta asimetria).

2.3 Propietats d'un estimador

Ja hem introduït l'estimador com un estadístic que utilitzem amb finalitats inferencials per aproximar-nos al valor d'un paràmetre. Més formalment direm que un estimador $\hat{\theta}$ del paràmetre desconegut θ, a partir de la mostra $M(\omega_i) = (X_1 , X_2 , \dots , X_n)$, és una funció de les v.a. que defineixen la mostra:

$$\hat{\theta} = f (X_1 , X_2 , \dots , X_n)$$

El valor que l'estimador $\hat{\theta}$ pren per a una mostra concreta rep el nom d'*estimació puntual:*

$$\hat{\theta}_i = f (x_1 , x_2 , \dots , x_n)$$

Exemple. La variància mostral s^2 és un estimador de la variància poblacional σ^2. A partir de la mostra de 12 vehicles dels quals hem observat l'ocupació:

1, 2, 1, 1, 1, 3, 1, 1, 2, 1, 1, 1,

obtenim una estimació puntual de la variància σ^2 a partir de s^2=0.4242...

Exemple. En el cas d'una v.a. amb distribució normal, la mitjana i la mediana poblacionals coincideixen. Està justificat, per tant, utilitzar la mediana mostral (el valor x_i tal que deixa el 50% de la mostra per sota) per tal d'estimar el valor de la mitjana poblacional. Queda clar que la mediana mostral i la mitjana mostral no tenen per què ser iguals.

Estudiem, a continuació, quines propietats ha de complir un estadístic per ser un "bon" estimador o, per dir-ho d'una altra manera, per ser preferible a un altre.

2.3.1 Biaix d'un estimador

Definim el *biaix*[9] com la diferència entre l'esperança matemàtica d'un estimador i el valor del paràmetre que es vol estimar:

$$\text{Biaix} = \text{E}(\hat{\theta}) - \theta$$

Diem que un estimador $\hat{\theta}$ del paràmetre θ és *no esbiaixat*[10] si el biaix és zero. És a dir, quan el valor esperat de $\hat{\theta}$ coincideix amb el valor del paràmetre que es vol estimar, θ.

Exemple. S'ha vist a l'apartat 2.2.1 que, en una mostra aleatòria simple, l'esperança de l'estadístic *mitjana mostral* coincideix amb l'esperança de la v.a. que s'està estudiant:

$$\text{E}(\overline{X}) = \text{E}(X) = \mu$$

Per tant, en una MAS la mitjana mostral $(\overline{X})$ és un estimador no esbiaixat de la mitjana poblacional (μ).

A continuació, mostrem un exemple d'estimador amb biaix i, de pas, veurem per què utilitzem la variància mostral definida prèviament. Suposem l'estimador següent per a la variància σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum x_i^2}{n} - 2\frac{\sum x_i}{n}\bar{x} + \frac{\sum \bar{x}^2}{n} =$$
$$= \frac{\sum x_i^2}{n} - 2\bar{x}^2 + \bar{x}^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$$

És una opció lògica dividir per *n*, tal com es fa amb la mitjana, en lloc de fer-ho per *n*–1, perquè al cap i a la fi s'estan sumant *n* termes. Però, quina esperança té aquest nou estimador? Si fos σ^2, pensaríem que hem trobat l'estimador ideal, sense biaix. Provem d'obtenir-ne el seu valor esperat.

Abans de res, recordem una propietat que utilitzarem més endavant:

$$\text{V}(X) = \text{E}(X^2) - (\text{E}(X))^2 \;\Rightarrow\; \text{E}(X^2) = \text{V}(X) + (\text{E}(X))^2 = \sigma^2 + \mu^2$$

El valor esperat del quadrat d'una v.a. té sentit quan busquem l'esperança d'un estimador de la variància, com veiem tot seguit:

$$\text{E}(\hat{\sigma}^2) = \text{E}\left(\frac{\sum X_i^2}{n}\right) - \text{E}(\overline{X}^2) = \frac{1}{n}\sum \text{E}(X_i^2) - \text{E}(\overline{X}^2)$$
$$= \frac{1}{n}\left(n(\sigma^2 + \mu^2)\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\left(1 - \frac{1}{n}\right)$$

9 En castellà *sesgo*; en anglès *bias.*

10 En castellà *insesgado*; en anglès *unbiased*.

El resultat anterior ens diu que l'estimador presentat té tendència a subestimar el valor de la variància real, per un factor 1–1/*n*, que també es pot escriure com $(n-1)/n$. Per tant, traiem el biaix de l'estimador $\hat{\sigma}^2$ si el denominador del quocient és $n-1$, en lloc de *n*. Això ens porta a l'estimador habitual de la variància, que ja hem introduït al capítol anterior.

Comentari. És comprensible la pèrdua d'un element en el denominador, ja que podem observar que els *n* sumands inclouen el terme de la mitjana mostral (que també s'obté dels valors de la mostra). La redundància que es produeix en aquesta situació es tradueix en una estimació deficitària de la dispersió, que no es produiria si coneguéssim la mitjana poblacional i aparegués a l'expressió de $\hat{\sigma}^2$.

2.3.2 Eficiència d'un estimador

Vegem ara perquè interessa estudiar la variància d'un estimador. Assumim que disposem de dos estimadors no esbiaixats per a un paràmetre determinat. Els possibles valors del que tingui menor variància es trobaran més concentrats al voltant del paràmetre d'interès. En absència de biaix, és clar que serà preferible utilitzar el que ens pugui proporcionar estimacions més properes al valor desconegut.

Entre dos estimadors no esbiaixats, es diu que és més *eficient* el que té una variància menor. Observeu que s'està definint l'eficiència de forma relativa, ja que es disposa de dos estimadors, però que no es pretén quantificar l'eficiència d'un estimador concret.

Exemple. Donada una mostra de grandària *n* parell, i essent $m = n/2$:

$$\overline{X}_{\text{senars}} = (X_1 + X_3 + \ldots X_{n-1}) / m$$

on:

$$\mathrm{E}\left(\overline{X}_{\text{senars}}\right) = \mathrm{E}(X)$$

i:

$$\mathrm{V}\left(\overline{X}_{\text{senars}}\right) = \mathrm{V}(X) / m = \mathrm{V}(X) / (n/2) = 2\,\mathrm{V}(X) / n$$

Per tant, $\overline{X}$, amb variància $\mathrm{V}(X)/n$, és més eficient que $\overline{X}_{\text{senars}}$.

L'expressió "més eficient" té connotacions econòmiques i vol dir que la relació qualitat/cost serà més favorable. L'estimador més eficient proporciona més informació (té menys soroll o error aleatori) per a una mateixa grandària mostral (que equival a cost). O també pot obtenir la mateixa quantitat d'informació amb una mostra més petita (amb cost menor).

Exemple. Un escàner recull informació mitjançant uns miralls i unes lents que projecten la llum reflectida per l'objecte en un sensor CCD. Això, i una il·luminació no uniforme, pot provocar que la lectura del color de l'objecte estigui afectada per alguna aberració, que es tradueix en un canvi en els valors numèrics que defineixen el color del punt examinat.

Per estudiar la magnitud de les aberracions cromàtiques de dos models d'escàner, s'utilitza un full de tonalitat uniforme i es comparen punts a l'atzar de la superfície escanejada per ambdós aparells. Com que cada punt és una combinació de tres colors, simplifiquem considerant només la tonalitat o matís (*hue*, en anglès).

Per tant, cada escàner representa la forma que capta la tonalitat del full (un número determinat, com 128: el paràmetre) com una distribució de valors més o menys allunyats del valor correcte. Si aquesta distribució té per esperança el valor exacte del full, aquest escàner produeix estimacions no esbiaixades del matís del punt escanejat; si l'esperança fos més alta, la tonalitat captada tendiria a ser més blavosa i, si fos més baixa, seria més verdosa.

Suposant que els dos escàners no tinguessin biaix de tonalitat, veuríem quin és més eficient per representar amb fidelitat el color original comparant la dispersió de les estimacions respectives, és a dir, veient quina de les dues distribucions presenta la menor desviació tipus.

Exemple. Dos informàtics han dissenyat dos experiments per comparar els rendiments, en temps, de dos algoritmes d'inversió de matrius. El primer ha generat dues mostres de matrius, una per a cada algoritme, i compara les mesures de les dues mostres $(\bar{x}_1 - \bar{x}_2)$. El segon ha utilitzat la mateixa matriu per a les proves amb ambdós algoritmes i calcula la mitjana de les diferències ($\bar{x}_D$), cosa que redueix la variabilitat deguda a la matriu (els detalls formals es veuran al tema 5). La figura 2.6 mostra que, si bé ambdós experiments són no esbiaixats, el segon és més eficient.

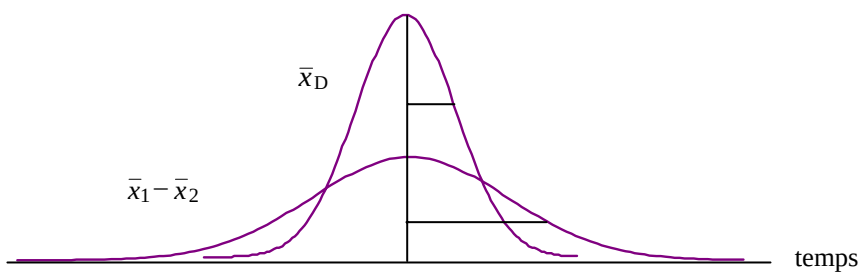


Figura 2.6 Distribució de la diferència de mitjanes dels temps d'inversió de matrius (corba més aplanada) i distribució de la mitjana de la diferència dels temps d'inversió de matrius (corba més apuntada)

Finalment, esmentem una última propietat que no es desenvoluparà en aquest text: la *consistència* d'un estimador. Es diu que un estimador és consistent si, a mesura que la grandària de la mostra creix, la seva distribució es concentra progressivament en una banda més estreta al voltant del paràmetre objectiu. Per tant, és una propietat que es defineix en el límit, si *n* tendeix a infinit. Un estimador pot tenir biaix o no ser el més eficient, però si es comprova que, quan n creix, tant el biaix com la seva variància tendeixen a zero, llavors l'estimador és consistent (aquests dos indicadors compon

l'anomenat *error quadràtic mitjà*, que es defineix com EQM = biaix al quadrat de l'estimador + variància de l'estimador).

Exercici. Demostreu que els estimadors que hem introduït com a exemples són consistents.

2.4 El concepte d'estimació per interval de confiança

Hem vist en un exemple anterior que podem fer un interval que contingui una proporció fixada de totes les possibles mitjanes mostrals, al voltant de μ. La figura 2.7 n'il·lustra el resultat. Concretament, l'interval que conté el 95% de les $\overline{X}$ de mostres de grandària n és:

$$\mu \pm z_{0.975}\,\sigma/\sqrt{n} = \mu \pm 1.96\,\sigma/\sqrt{n}$$

A la pràctica, el problema és el contrari: coneixem $\overline{X}$ i volem fer l'estimació d' $\mathrm{E}(X)$. En general, disposem d'una mostra o d'estadístics suficients per descriure el que ens interessa d'una mostra, però volem estimar un determinat paràmetre θ. Una bona idea seria fer un interval que ens informi, amb certesa controlable per part nostra, on es troba el valor del paràmetre d'interès. Si exigim un nivell de seguretat molt alt, n'obtindrem un interval ample, i potser poc útil, però per a nivells raonables l'interval ens proporcionarà una idea acceptable de la magnitud del paràmetre.

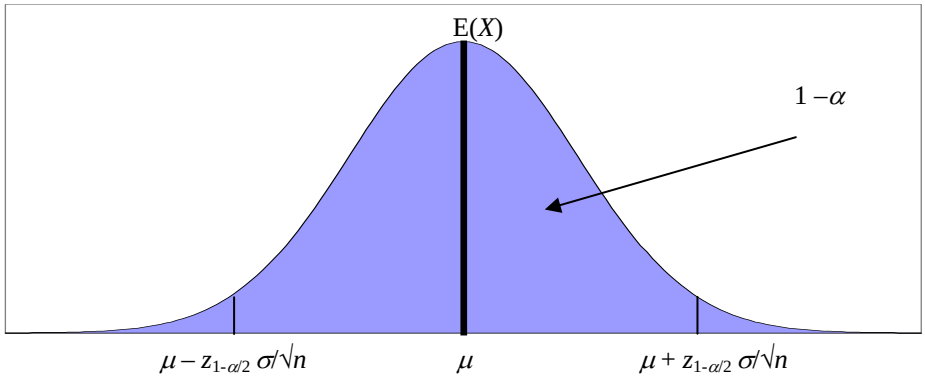


Figura 2.7 Interval que conté una proporció 1–α de mitjanes mostrals

De tota manera, subratllem ara alguns obstacles que hem de superar encara:

- Com donem la volta a l'interval per a la mitjana mostral i el convertim en interval per a la mitjana poblacional?
- Com ho farem per a d'altres paràmetres, tal com σ^2 o π?
- Dóna la impressió que fa falta conèixer la variància poblacional per estimar el paràmetre E(X). En realitat, solament podem comptar amb els estimadors mostrals $\overline{X}$ i s^2. Això porta a la qüestió següent (molt subtil): quina distribució de probabilitat tindrà l'estadístic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim ?$$

El problema és que el nou estadístic T no és una variable aleatòria de distribució normal centrada i reduïda, ja que el denominador no és el seu error tipus poblacional (un paràmetre, constant), sinó la seva estimació mostral (un estadístic, amb dispersió). Com a conseqüència, tenim una variable, que és el resultat de dividir dues variables aleatòries, la distribució de la qual és, per ara, desconeguda.

En el proper tema estudiarem unes distribucions derivades de la normal que permeten solucionar aquestes dificultats.

2.5 Problemes

1. Recordeu el càlcul de la mitjana i de la variància d'una mostra?

> Repàs: $s^2 = \Sigma_{i=1,n} (x_i - \bar{x})^2 \ / \ (n-1)$
> $s^2 = [\Sigma x_i^2 - (\Sigma x_i)^2/n \ ] / (n-1)$ (més senzill per a càlculs manuals)

   Practiqueu ambdues fórmules, a mà i amb la calculadora, per a un cas senzill de $n = 5$ alumnes que contesten, respectivament, que a la seva família són 1, 2, 3, 4 i 5 germans.

2. Suposeu ara que esteu interessat a conèixer la mitjana de germans de les famílies d'aquests alumnes.

   Si es considera que aquesta mostra de n=5 és una mostra aleatòria representativa de totes les famílies, quin error s'ha d'esperar per a la mitjana observada a la mostra en estimar la mitjana poblacional?

3. Els psicòlegs mesuren la intel·ligència amb el quocient intel·lectual (QI), que segueix, per a la població adulta general, una N(100,15). És a dir, $\mu = 100$ i $\sigma = 15$. Si s'han recollit moltes mostres de grandària $n = 9$, i de cada mostra i se'n calcula la mitjana $\overline{x}_i$:

   a) Com varien les mitjanes $\overline{x}_i$ de les mostres?

   b) A la Facultat d'Informàtica, hem recollit una mostra de grandària $n = 9$ i hem observat $\overline{x}_i = 104$. Es tracta d'un exemple: (1) acceptablement típic; (2) especialment afortunat, molt a prop de μ; (3) tan estrany i allunyat que sospitem que els informàtics no són d'aquella població?

   c) Repetiu els dos apartats anteriors, però amb $n = 25$ i $n = 225$.

4. Els ingressos familiars[11] van presentar la distribució següent, agrupada *grosso modo* (unitats de X, 10^4 dòlars americans):

	Ingressos: x	Proporció: $p(x)$
A	2	0.50
B	4	0.30
C	8	0.20

a) Calculeu la mitjana μ i la desviació tipus σ i mostreu-les en una gràfica de la distribució de la població.

b) Suposeu que s'extrau a l'atzar una mostra de $n = 2$ ingressos: X_1 i X_2. Tabuleu la distribució conjunta de X_1 i X_2. Calculeu l'esperança i la variància de $\overline{X}$. Mostreu-les en una gràfica de la distribució mostral de $\overline{X}$.

c) Calculeu, a partir de les fórmules, E($\overline{X}$) i V($\overline{X}$) per a: (1) $n = 2$. Concorda aquest resultat amb el que s'ha obtingut amb la distribució conjunta? Repetiu-ho, només amb les fórmules, per a: (2) $n = 5$; i (3) $n = 20$.

5. Estudieu les propietats dels estimadors proposats a l'exercici 2 del capítol anterior.

2.6 Solució dels problemes

1.

x_i		$(x_i - \overline{x})^2$		x_i^2	
1		4		1	
2		1		4	
3		0		9	
4		1		16	
5		4		25	
15	$\overline{x} = 15/5 = 3$	10	$s^2 = 10/4 = 2.5$	55	$s^2 = [\ 55 - 15^2/5\ ] / 4 = 2.5$

És a dir, la mitjana mostral és de 3 germans, la variància mostral és de 2.5 germans² i la desviació tipus mostral és, aproximadament, d'1.6 germans (arrel quadrada de 2.5). Podem imaginar que la distància (o desviament) respecte de la mitjana d'una família "tipus" és d'1.6 germans.

2.

$$s^2_{\overline{x}} = s^2/n = 2.5 / 5 = 0.5$$
$$s_{\overline{x}} = \sqrt{s^2_{\overline{x}}} = \sqrt{0.5} \approx 0.707$$

[11] Adaptat de Wonnacott & Wonnacott, *Introducción a la estadística*, 5a ed., Limusa, pàg. 234.

Si estimem la mitjana de la població com 3 germans (és a dir, si decidim aproximar-nos a la mitjana poblacional a partir de la mitjana mostral), l'error esperat en fer aquesta afirmació és de 0.7 germans.

3. a) $V(\overline{X}) = V(X) / n = 15^2/9 = 25 \text{ un}^2 \quad \rightarrow \quad \sigma_{\overline{x}} = 5 \text{ un}$

La variabilitat de les mitjanes mostrals és la tercera part de la variabilitat de la variable.

b) $[\overline{x}_i - E(X)]^2 = [104 - 100]^2 = 4^2 \text{ un}^2$

104 és una xifra "raonable" ja que podem esperar una dispersió similar (5^2 un^2).

c) Si $n = 25 \qquad V(\overline{X}) = V(X) / n = 15^2/25 = 9 \text{ un}^2 \rightarrow \quad \sigma_{\overline{x}} = 3 \text{ un}$

Continua essent una xifra "raonable", ja que el seu valor esperat era de 3^2 un^2.

Si $n = 225 \quad V(\overline{X}) = V(X) / n = 15^2 / 225 = 1^2 \text{ un}^2 \rightarrow \quad \sigma_{\overline{x}} = 1 \text{ un}$

Ara ja no és una xifra raonable, ja que el seu valor esperat era d'1 un.

4. a)

	Ingressos: x	$p(x)$	Mitjana μ $x \cdot p(x)$	Variància σ^2 $(x-\mu)^2 \cdot p(x)$
A	2	0.50	1	1.620
B	4	0.30	1.2	0.012
C	8	0.20	1.6	3.528
			μ=3.8	$\sigma^2 = 5.16$; $\sigma \approx 2.27$

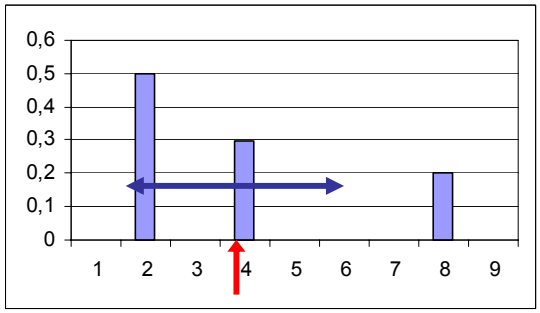


Figura 2.8 Distribució de probabilitat dels ingressos. La fletxa vertical marca el punt de la mitjana i la fletxa doble horitzontal l'extensió de la desviació tipus (una σ cap amunt, una σ cap avall).

b)

	A_2 X_2=2 $P(X_2=2)$=0.5	B_2 X_2=4 $P(X_2=4)$=0.3	C_2 X_2=8 $P(X_2=8)$=0.2
A_1 X_1=2 $P(X_1=2)$=0.5	$\overline{x}$ =2 $P(X_1=2 \wedge X_2=2)$ =0.25	$\overline{x}$ =3 $P(X_1=2 \wedge X_2=4)$ =0.15	$\overline{x}$ =5 $P(X_1=2 \wedge X_2=8)$ =0.1
B_1 X_1=4 $P(X_1=4)$=0.3	$\overline{x}$ =3 $P(X_1=4 \wedge X_2=2)$ =0.15	$\overline{x}$ =4 $P(X_1=4 \wedge X_2=4)$ =0.09	$\overline{x}$ =6 $P(X_1=4 \wedge X_2=8)$ =0.06
C_1 X_1=8 $P(X_1=8)$=0.2	$\overline{x}$ =5 $P(X_1=8 \wedge X_2=2)$ =0.1	$\overline{x}$ =6 $P(X_1=8 \wedge X_2=4)$ =0.06	$\overline{x}$ =8 $P(X_1=8 \wedge X_2=8)$ =0.04

Ingressos $\bar{x}$	Observacions possibles	Probabilitat $p(\bar{x})$		Mitjana $\mu_{\bar{X}}$ $\bar{x}\cdot p(\bar{x})$	Variància $\sigma^2_{\bar{X}}$ $(\bar{x}-\mu_{\bar{X}})^2 p(\bar{x})$
2	$A_1 \cap A_2$	0.5·0.5	=0.25	0.50	0.8100
3	$A_1 \cap B_2 \cup B_1 \cap A_2$	0.5·0.3 + 0.3·0.5	=0.30	0.90	0.1920
4	$B_1 \cap B_2$	0.3·0.3	=0.09	0.36	0.0036
5	$A_1 \cap C_2 \cup C_1 \cap A_2$	0.5·0.2 +0.2·0.5	=0.20	1.00	0.2880
6	$B_1 \cap C_2 \cup C_1 \cap B_2$	0.3·0.2 +0.2·0.3	=0.12	0.72	0.5808
8	$C_1 \cap C_2$	0.2·0.2	=0.04	0.32	0.7056
				$\mu_{\bar{X}}$ =3.8	$\sigma^2_{\bar{X}}$ = 2.58 $\sigma_{\bar{X}}$ = 1.61

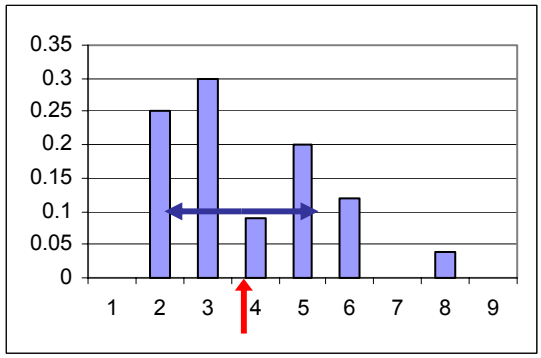


Figura 2.9 Distribució de probabilitat de la mitjana dels ingressos de dos individus. Es pot apreciar que el valor esperat no ha canviat i que la desviació tipus ha disminuït.

c)

n=2	*n*=5	*n*=20
$\mu_{\bar{X}} = \mathrm{E}(\bar{X}) = \mathrm{E}(X) = 3.8$	$\mu_{\bar{X}} = \mathrm{E}(\bar{X}) = \mathrm{E}(X) = 3.8$	$\mu_{\bar{X}} = \mathrm{E}(\bar{X}) = \mathrm{E}(X) = 3.8$
$\sigma_{\bar{X}} = \mathrm{V}(\bar{X}) = \sigma/\sqrt{n} =$ $2.27/\sqrt{2} = 1.61$	$\sigma_{\bar{X}} = \mathrm{V}(\bar{X}) = \sigma/\sqrt{n} =$ $2.27/\sqrt{5} = 1.02$	$\sigma_{\bar{X}} = \mathrm{V}(\bar{X}) = \sigma/\sqrt{n} =$ $2.27/\sqrt{20} = 0.51$
Coincideix amb el resultat anterior		

S'ha de tenir en compte que $\bar{X}$ és sempre un estimador no esbiaixat de μ, per a tot n:

$$\mu_{\bar{X}} = \mathrm{E}(\bar{X}) = \mathrm{E}(X) = \mu$$

A més a més, $\sigma_{\bar{X}}$ (que es pot calcular directament, sense recórrer a la distribució conjunta) disminueix a mesura que n augmenta ($\sigma_{\bar{X}}$ és inversament proporcional a $\sqrt{n}$).

5. Recordem que s'havien proposat quatre estimadors:

$t_1 = \text{màx}(x_1, x_2, ..., x_n)$
$t_2 = 2\,\bar{x}$
$t_3 = 2 \cdot \text{Med}(x_1, x_2, ..., x_n)$
$t_4 = t_1 \cdot (n+1)/n$

Vegem si el primer té biaix: n'obtenim la funció de distribució i, a continuació, la funció de densitat amb la qual calcularem el valor esperat.

$$\begin{aligned} F_{t1}(x) &= P(t_1 \le x) = P(x_1 \le x \cap ... \cap x_n \le x) \\ &= P(x_1 \le x) \times ... \times P(x_n \le x) && \text{[MAS són variables i.i.d.]} \\ &= P(X \le x)^n = (x/T)^n && 0 \le x \le T \\ f_{t1}(x) &= n/T \cdot (x/T)^{n-1} && 0 < x < T \\ E(t_1) &= n \cdot T/(n+1) && \text{[Deixem la demostració com a exercici per al lector.]} \end{aligned}$$

És a dir, el primer estimador presenta biaix: hi ha una diferència amb el paràmetre igual a $T/(n+1)$. Podem fer que coincideixin si multipliquem t_1 per $(n+1)/n$: d'aquesta manera obtenim el quart estimador. Si bé aquesta operació fa que incrementi la seva variància, hem eliminat el biaix totalment.

Trobem les variàncies de t_1 i t_4:

$$V(t_1) = n \cdot T^2 / ((n+1)^2 (n+2)) \qquad \text{[Deixem la demostració com a exercici per al lector.]}$$
$$V(t_4) = ((n+1)/n)^2 \, V(t_1) = T^2 / (n\,(n+2))$$

El segon estimador no te biaix perquè s'ha construït precisament per complir aquesta propietat (el valor esperat del temps és $T/2$). Com és la mitjana mostral multiplicada per 2, trobem fàcilment la seva variància:

$$V(t_2) = 4 \cdot V(\bar{X}) = 4 \cdot T^2/(12\,n) = T^2/(3\,n) \qquad \text{[Ja que la variància d'un temps uniforme [0, } T\text{] val } T^2/12]$$

Si el comparem amb el quart estimador, veiem que t_4 és més eficient per a qualsevol grandària mostral superior a 1. Respecte al tercer estimador, basat en la mediana, direm només que també és no esbiaixat, però que en ser un estadístic basat en rangs presenta una complexitat molt superior a l'hora de fer els càlculs que hem presentat fins ara. Solament afegirem que la seva variància és més gran encara que t_2, i, per tant, és el menys eficient dels tres no esbiaixats.

3 Intervals de confiança

Al capítol anterior, hem deixat obertes unes qüestions al voltant de la possibilitat de fer estimacions mitjançant intervals amb un grau de certesa (que anomenarem nivell de *confiança*) donat, que es resoldran en aquest capítol. El procediment de partida per a la mitjana de mostres grans es basa en el teorema del límit central, si X no és normal, que estableix les condicions de convergència fins a la distribució normal centrada i reduïda N(0,1) de l'estadístic:

$$\hat{Z} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Ara bé, s'ha de tenir present que l'ús d'aquest estadístic implica que, per poder estimar la mitjana poblacional, necessitem conèixer prèviament la variància de la variable. Aquesta situació es pot donar en alguna ocasió, però no és, ni molt menys, la situació habitual.

Substituir el paràmetre σ per l'estadístic s implica substituir una constant, que té un únic valor, per una variable aleatòria, amb tota una distribució de valors. Però la distribució de l'estadístic:

$$T = \frac{\overline{X} - \mu}{\sqrt{s^2/n}} \sim ?$$

és més complicada, ja que és el quocient de dues variables aleatòries. En aquest capítol introduirem, de forma pràctica, la distribució de χ^2 (khi quadrat) i la de la t de Student, ambdues derivades de la normal. Aquestes dues distribucions, juntament amb la F de Fisher-Snedecor, que estudiarem més endavant, permeten realitzar la inferència més habitual.

3.1 Interval de confiança per μ amb σ coneguda

Quan volem trobar un interval que molt probablement contingui l'esperança d'una variable aleatòria, hem de partir primer del resultat exposat al final del tema anterior:

$$\mu \pm z_{1-\alpha/2}\,\sigma/\sqrt{n} \qquad \text{conté una proporció } 1-\alpha \text{ de les mitjanes mostrals}$$

Hem de veure, llavors, què pot passar si s'observa un nombre gran de mostres. Teòricament, alguna de les mitjanes mostrals (un percentatge 100α) sortirà fora dels límits establerts per l'expressió anterior.

El gràfic de la figura 3.1 mostra el resultat d'afegir els valors $\pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$ al voltant de cada una de les mitjanes mostrals $\overline{x}_i$ que hem posat d'exemple, on α val 0.05. Observeu que, per les mesures de les mostres 2, 4, 5, 6 i 7 (de dalt a baix), l'interval dibuixat passa per sobre del punt que representa el valor del paràmetre $E(X)=\mu$. Aquests intervals estarien encertant en la seva inferència: realment contenen el paràmetre d'interès μ. El mateix succeiria amb totes les mesures mostrals contingudes entre els límits establerts més amunt, és a dir, amb el 95% de les possibles mostres. En canvi, els intervals de les mitjanes de les mostres 1 i 3 no contenen el paràmetre. Representen el 5% de les possibles mostres que fallarien en la seva apreciació.

Un interval construït així té, per tant, $1-\alpha$ de possibilitats de contenir el paràmetre poblacional, i per això rep el nom d'*interval de confiança (IC) $1-\alpha$*. Dit d'una altra manera: s'espera que el $100(1-\alpha)\%$ de les mostres donin lloc a intervals que contenen correctament μ.

Així doncs, l'interval de confiança $1-\alpha$ de μ és:

$$\text{IC}(\mu,\, 1-\alpha) \;=\; \overline{X} \pm z_{1-\alpha/2}\,\sigma_{\overline{x}} \;=\; \overline{X} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$$

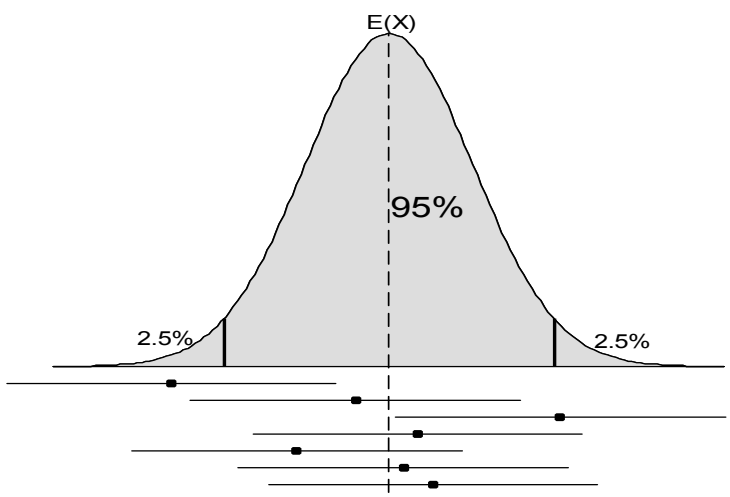


Figura 3.1 Exemples d'intervals de confiança, al voltant de diferents mitjanes mostrals: noteu que les que es troben fora dels límits del 95% no formen intervals que contenen el paràmetre.

Cal tenir en compte que, per poder afirmar que la variable mitjana mostral, una vegada centrada i reduïda, segueix una normal de mitjana zero i variància u:

$$Z = (\overline{X} - \mu) / (\sigma/\sqrt{n}) \sim N(0,1),$$

s'han de complir les condicions establertes a l'apartat 2.2.4.

Exemple. Reprenem el cas de l'ocupació dels vehicles que entren en hora punta. Hem determinat que la mostra tindrà grandària 50 per garantir la hipòtesi de normalitat de la mitjana amb suficient aproximació. (Compte: un error molt habitual és dir que, amb la grandària mostral escollida, *la variable és normal*. No! Si la distribució de la variable no és normal, això no ho canviarà el nombre d'observacions que en fem. S'ha de dir que és la mitjana qui es distribueix —aproximadament— com una normal.)

Les dades recollides són aquestes:

Ocupants	1	2	3	4	5
Freqüència	32	13	3	2	0

El nombre d'ocupants comptabilitzats és 75, és a dir, la mitjana mostral val 1.5. Coneixem que la desviació tipus σ val 0.9450: llavors, el IC(μ, 95%) resultant s'obté com:

$$1.5 \pm 1.96 \cdot 0.9450/\sqrt{50} = 1.5 \pm 0.2619 = [1.238, 1.762]$$

Per tant, creiem, amb una confiança del 95%, que l'autèntica mitjana poblacional està entre 1.238 i 1.762 ocupants per vehicle. En aquesta ocasió, l'interval ha encertat (perquè en coneixem el valor vertader, 1.63). Si es repetís l'experiment moltes vegades, podríem comprovar que aproximadament una de cada 20 vegades l'interval no conté el valor real. És el preu que hem acceptat de pagar.

Una reflexió més: no hem de fer la interpretació "tenim una probabilitat del 95% que el paràmetre estigui dintre de l'IC", entenent com a IC el resultat trobat. El paràmetre és fix i no varia aleatòriament: és l'interval [$\overline{X} - 1.96\sigma_{\overline{x}}$, $\overline{X} + 1.96\sigma_{\overline{x}}$] el que és aleatori, mentre no s'hagi observat la mostra, és clar. L'interval [1.238, 1.762] ja no és aleatori i, per tant, o conté el valor autèntic de la població o no el conté, però aquesta incertesa no és cap signe d'aleatorietat de μ, sinó de desconeixement del seu valor (quan es coneix, com en el cas de l'exemple, podem comprovar si un interval concret ha encertat o no).

Exercici. L'interval de confiança és un procediment robust que garanteix el nivell d'encert que establim en l'estimació del paràmetre. Per comprovar això, us suggerim que entreu a la pàgina web <http://www.kuleuven.ac.be/ucs/java/gent/Ap5a.html>, on trobareu una demostració empírica que il·lustra el funcionament del procediment mitjançant la simulació de mostres.

Quan entreu a la pàgina, pitgeu el botó Run, i generareu cent mostres que donaran lloc a cent IC representats respecte del paràmetre —que és conegut i val zero. Uns quants, dibuixats en color vermell, no tallen la línia vertical que representa el paràmetre. És d'esperar que al voltant del 95% dels IC continguin efectivament el valor zero.

3.2 Distribucions originades pel mostreig

Estudiarem tres distribucions derivades de la normal que utilitzarem en aquest curs. Són especials, ja que generalment no s'observen directament de fenòmens reals, sinó indirectament a través d'una mostra. En aquest capítol, definirem les distribucions χ^2 i t, que ens serviran per trobar intervals de confiança per a la variància i per a l'esperança, en cas que la variància no sigui coneguda. La tercera, la F de Fisher-Snedecor, la introduirem més endavant.

3.2.1 Distribució χ2 (khi quadrat)

Sigui X una variable aleatòria amb distribució normal centrada i reduïda, $X \sim N(0,1)$. Llavors, el seu quadrat, X^2, segueix una distribució de khi quadrat amb 1 grau de llibertat:

$$X^2 \sim \chi_1^2$$

Convé assenyalar que, en ser el quadrat d'un número real, tots els seus valors seran positius. És una distribució molt asimètrica, amb el màxim de la funció de densitat al punt x=0 (v. figura 3.2).

Exemple. Sigui X una v.a. N(0,1),

sabem que	$P(X > 1.96) = P(X < -1.96) = 0.025$
o també que	$P(\|X\| > 1.96) = 0.05$
Per tant	$P(\|X\|^2 > 1.96^2) = P(X^2 > 3.84) = 0.05$

Exemple. Sigui W una v.a. amb distribució χ_1^2. A la taula de la distribució de χ^2, primera línia, podem veure que $P(W > 5.024) = 0.025$.

Exercici. Comproveu que s'arriba als mateixos resultats amb unes taules estadístiques que amb Minitab o un full de càlcul.

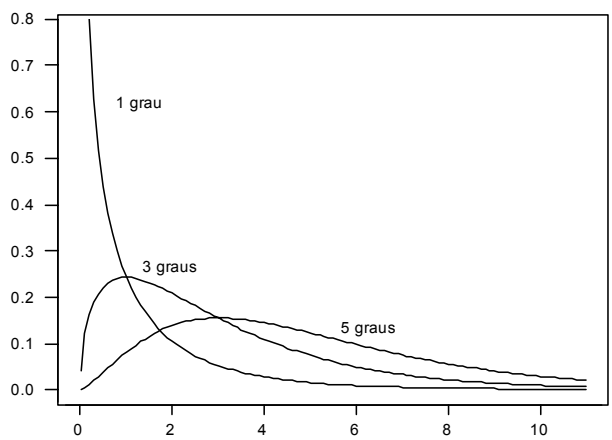


Figura 3.2 Perfil típic de la distribució de χ^2 amb un, tres i cinc graus de llibertat (la asimetria es redueix quan el nombre de graus de llibertat augmenta)

Siguin ara *n* variables aleatòries independents amb distribució normal, centrada i reduïda:

$$X_1 , X_2 , ... , X_n \sim \text{N}(0,1) \quad \text{v.a.i.i.d}$$

llavors, la suma dels seus quadrats segueix una distribució de khi quadrat amb *n* graus de llibertat:

$$\Sigma_i X_i^2 = X_1^2 + X_2^2 + ... + X_n^2 \sim \chi_n^2$$

Aquesta distribució té una forma asimètrica, tal com mostra la figura 3.2, malgrat que l'asimetria s'atenua si *n* augmenta.

Nota. Tres propietats que cal recordar. Si $Y \sim \chi_n^2$, llavors:

- $\text{E}(Y) = n$
- $\text{V}(Y) = 2n$
- quan *n* tendeix a infinit, la distribució de *Y* tendeix a una $\text{N}(\mu = n, \sigma^2 = 2n)$.

Normalment, podem utilitzar les taules per consultar valors de la distribució de khi quadrat a partir de probabilitats fixes. La darrera propietat es tindrà en compte quan el paràmetre *n* sigui molt gran (n>100) i l'aproximació del TLC sigui viable.

Exemple. Sigui l'estadístic següent, molt semblant a l'estadístic s^2, estimador de la variància, però utilitzant el paràmetre μ en lloc de l'estadístic $\overline{X}$ i dividint per *n* en lloc de per *n*–1:

$$\hat{\sigma}^2 = \frac{\sum_{i=1,n}(x_i - \mu)^2}{n}$$

Llavors, si X_i té distribució normal,

$$n\frac{\hat{\sigma}^2}{\sigma^2} = n\frac{\sum_{i=1,..,n}(x_i - \mu)^2 / n}{\sigma^2} = \frac{\sum_{i=1,..,n}(x_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

ja que cada sumand seguirà, independent de la resta, una N(0,1) al quadrat.

a) Distribució de l'estadístic s^2 per a una distribució normal

En aquest apartat, i a partir del resultat anterior, arribarem a deduir quina és la distribució de probabilitat de l'estadístic variància mostral d'una MAS obtinguda d'observacions procedents d'una variable normal. Es pot demostrar que

$$Y_{n-1} = (n-1)\frac{s^2}{\sigma^2} = (n-1)\frac{\sum_{i=1,n}(x_i - \overline{x})^2 / n-1}{\sigma^2} = \frac{\sum_{i=1,n}(x_i - \overline{x})^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(x_i - \overline{x})^2}{\sigma^2} \sim \chi_{n-1}^2$$

Aquest resultat és rellevant, ja que implica que coneixem la distribució d'una transformació lineal de l'estadístic s^2, estimador de la variància. Això ens permetrà calcular intervals de confiança. És important ressaltar que, perquè sigui cert que aquest estadístic segueix la distribució de khi quadrat,

cada sumand ha de seguir una distribució N(0,1) al quadrat, és a dir, $X_i \sim \mathrm{N}(\mu, \sigma)$. Per tant, la condició necessària per utilitzar aquest estadístic és que la distribució de la variable mesurada X sigui normal.

Fixeu-vos que es perd un grau de llibertat perquè l'ús de la mitjana mostral en lloc de μ introdueix una restricció a la mostra (és a dir, si coneixem n–1 valors de la mostra i la mitjana mostral, l'n-èsim valor de la mostra *no és lliure*, no pot ser qualsevol, està fixat per les condicions anteriors).

b) Interval de confiança de $\sigma^2 = \mathrm{V}(X)$ per a una v.a. X amb distribució normal

Hem vist que, si la variable X en estudi segueix una distribució normal, llavors:

$$(n-1)\, s^2 / \sigma^2 \sim \chi^2_{n-1}$$

Per tant (atenció al canvi de límits de la distribució en invertir):

$$1-\alpha \;=\; \mathrm{P}\left(\chi^2_{n-1,\ \alpha/2} \;\le\; \frac{s^2(n-1)}{\sigma^2} \;\le\; \chi^2_{n-1,\ 1-\alpha/2} \right)$$

$$1-\alpha \;=\; \mathrm{P}\left(\frac{1}{\chi^2_{n-1,\ 1-\alpha/2}} \;\le\; \frac{\sigma^2}{s^2(n-1)} \;\le\; \frac{1}{\chi^2_{n-1,\ \alpha/2}} \right)$$

$$1-\alpha \;=\; \mathrm{P}\left(\frac{s^2(n-1)}{\chi^2_{n-1,\ 1-\alpha/2}} \;\le\; \sigma^2 \;\le\; \frac{s^2(n-1)}{\chi^2_{n-1,\ \alpha/2}} \right)$$

obtenim un interval que contindrà, en un 100 (1– α) % dels casos, el valor del paràmetre σ^2.

Exemple. En 25 mesures preses amb un escàner d'una superfície perfectament monocroma, s'ha observat una variància $s^2=6^2$ unitats de matís (el matís, o *hue*, es representa en una escala de 0 a 255). Què sabem sobre l'autèntic valor de la variància poblacional? Una forma de contestar aquesta pregunta és construir un interval de confiança que, si no se'ns diu el contrari, farem del 95%:

$$\mathrm{IC}(\sigma^2, 0.95) = \left(\frac{s^2(n-1)}{\chi^2_{n-1,\ 1-\alpha/2}}, \frac{s^2(n-1)}{\chi^2_{n-1,\ \alpha/2}} \right) = \left(\frac{36(25-1)}{\chi^2_{n-1,\ 0.975}}, \frac{36(25-1)}{\chi^2_{n-1,\ 0.025}} \right) =$$

$$= \left(\frac{864}{39.36}, \frac{864}{12.4} \right) = \left(21.95, 69.68 \right)$$

Per tant, havent observat una variància mostral $s^2=36$, el que podem afirmar sobre la variància poblacional σ^2 és que, amb una confiança del 95%, es tracta d'algun dels valors

compresos entre 21.95 i 69.68. Hi ha dos aspectes que es poden ressaltar: l'asimetria de l'interval al voltant de l'estimació puntual (36) i la seva gran magnitud: encara que la mostra no és molt petita (n=25), el grau d'incertesa sembla notable. Com que seria complicat interpretar els valors amb les unitats al quadrat, calcularem l'interval de confiança de la desviació tipus:

$$\text{IC}(\sigma, 95\%) = [4.69, 8.35]$$

L'interval continua essent asimètric al voltant de l'estimació puntual, que era 6. Ara, sense quadrats, és més fàcil d'interpretar: amb una confiança del 95%, la desviació tipus poblacional de l'error de mesura del matís per part de l'escàner és algun valor comprès entre 4.7 i 8.3 punts.

3.2.2 Distribució t de Student

Comentari. William Gosset era un químic que treballava per a Guinness investigant quina era la millor varietat d'ordi per fer cervesa. Gosset calculava l'estadístic t anterior amb mostres de grandària 4, perquè treballava amb dades de quatre granges. Per poder detectar i rebutjar els lots de cervesa que no complien les especificacions desitjades, havia acceptat el cost de rebutjar un 5% dels lots que sí les complien. Així, rebutjava aquelles mostres en què el valor resultant es trobava fora dels límits ±1.96. Aviat va començar a sospitar que estava rebutjant massa lots de cervesa, i va veure que fora dels límits ±1.96 hi havia més del 5% de lots correctes. Es va adonar que s era un estadístic i no un paràmetre, i va proposar una distribució una mica més aplanada que la normal, seguint la qual rebutjava el α % desitjat de lots correctes. L'empresa Guinness estava satisfeta amb els resultats, però no li va permetre que publiqués amb el seu nom l'article en el qual aquests es descrivien amb el seu nom. El va firmar amb el pseudònim d'"estudiant", raó per la qual la distribució que va proposar es coneix com la t de Student.

Sigui Z una variable aleatòria amb distribució normal centrada i reduïda, i Y una variable amb una distribució de khi quadrat amb n graus de llibertat, independents entre elles. Llavors:

$$\frac{Z}{\sqrt{Y/n}} \sim \text{t}_n$$

segueix una distribució t de Student amb n graus de llibertat.

La distribució t de Student és simètrica al voltant del zero, molt semblant a la normal, especialment per a valors grans de n (v. figura 3.3). Com més petit és el nombre de graus de llibertat, més es diferencia de la normal, fent-se més aplanada, i amb cues lleugerament més llargues. Quan n creix, la t de Student tendeix a coincidir amb una N(0,1).

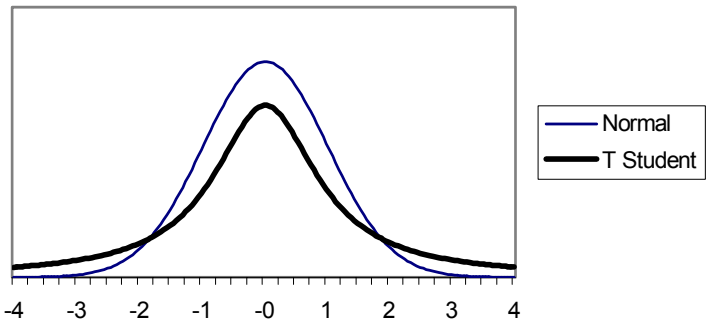


Figura 3.3 Perfil típic de la distribució t *(línia gruixuda), comparada amb la campana de Gauss (línia fina)*

Exercici. Comproveu, amb les taules impreses, l'equivalència de l'última fila de la taula de la t amb la taula de la N(0,1). Repetiu l'exercici amb un paquet informàtic (un full de càlcul o Minitab, per exemple).

Exercici. Sigui T una variable aleatòria amb distribució t de Student amb 12 graus de llibertat. Trobeu a les taules que P ($T > 1.796$) = 0.05.

Ara ja estem en condicions de resoldre el problema que hem deixat obert al capítol anterior. Per fer un interval de confiança de la mitjana poblacional μ, utilitzem l'estadístic Z, que segueix una distribució normal, centrada i reduïda:

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Però si no coneixem σ, haurem de substituir-la per s. Quina distribució segueix el nou estadístic?

$$T = \frac{\overline{X} - \mu}{\sqrt{s^2/n}} \sim ?$$

Dividint el numerador i el denominador per l'error tipus ($\sqrt{\sigma^2/n}$):

$$T = \frac{\overline{X} - \mu}{\sqrt{s^2/n}} = \frac{\dfrac{\overline{X} - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{\dfrac{s^2/n}{\sigma^2/n}}} = \frac{\dfrac{\overline{X} - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{\dfrac{s^2}{\sigma^2}}}$$

Observeu que l'expressió del numerador correspon a una variable N(0,1). Desenvolupem ara l'expressió del denominador:

$$\sqrt{\frac{s^2}{\sigma^2}} = \sqrt{\frac{\sum_{i=1,n}(X_i - \overline{X})^2}{n-1} \Big/ \sigma^2} = \sqrt{\frac{\sum_{i=1,n}(X_i - \overline{X})^2}{\sigma^2} \Big/ n-1}$$

Comprovem que, reordenat d'aquesta manera, tenim al denominador l'arrel quadrada d'una variable que segueix una distribució de khi quadrat dividida, pels seus graus de llibertat. El quocient és la definició prèvia de la t de Student, concretament una t amb n–1 graus de llibertat.

És molt important recordar que, per a què l'estadístic Y del denominador segueixi una distribució de khi quadrat, cada un dels sumands que el compon ha de seguir una distribució normal i, per tant, la variable X que estem estudiant *ha de seguir una distribució normal*. Volem ressaltar que abans, per a l'estadístic Z, només calia la normalitat de $\overline{X}$ que aconseguíem, ja fos per disposar d'una mostra gran, n>30, o bé perquè la variable X ja era normal. Ara la primera premissa, una mostra gran, ja no serveix.

Nota. Per a què l'estadístic T segueixi una t de Student, és necessari també que les variables del numerador i del denominador siguin independents. Atès que les úniques variables aleatòries rellevants són $\overline{X}$ i s^2, on X és normal i la mostra és MAS, això es redueix a demostrar la seva independència, aspecte que no demostrarem aquí. El lector interessat a trobar més informació pot consultar el teorema de Cochran a la Wikipedia:

<http://en.wikipedia.org/wiki/Cochran%27s_theorem>

Comentari. Els estadístics Z i T tenen una estructura molt similar. El numerador representa la distància entre el valor de la mostra $\overline{X}$ i el paràmetre μ de la població: l'error absolut d'estimació. El denominador d'ambdós estadístics informa de l'error tipus (*típic*) de $\overline{X}$. Ambdós estadístics Z i T representen una mesura d'error relatiu, adimensional, que permet valorar la desviació de la mitjana mostral independentment de les unitats de la variable observada.

Més endavant, al capítol dedicat als contrastos d'hipòtesis, veurem que el quocient *error d'estimació/error tipus* s'utilitza per posar a prova formalment si és admissible que el paràmetre μ prengui un determinat valor. Llavors, tindrem en compte si les pertorbacions que el mostreig aleatori afegeix poden ser l'única font que expliqui la desviació de la mitjana respecte de l'hipotètic valor de μ.

a) Interval de confiança de μ d'una variable normal, sense conèixer σ

Acabem de veure que, en substituir σ per s, hem de recórrer a la t de Student en lloc de la distribució normal.

$$T = \frac{\overline{X} - \mu}{\sqrt{s^2/n}} \sim \mathrm{t}_{n-1}$$

Llavors, l'estadístic T queda confinat a una regió amb probabilitat $1-\alpha$. Per similitud amb la situació on la σ és coneguda, i tenint en compte que la distribució t de Student és simètrica, com la normal, l'interval de confiança $1-\alpha$ de μ serà:

$$\mathrm{IC}(\mu, 1-\alpha) = [\overline{X} - t_{n-1,\,1-\alpha/2}\, s/\sqrt{n},\ \overline{X} + t_{n-1,\,1-\alpha/2}\, s/\sqrt{n}]$$

Exemple. Quan reproduïm un objecte d'àudio (una cançó, per exemple) des d'un lector de CD, es produeix un flux d'informació que fluctua amb distribució normal i que es mesura en kBps. S'han pres una sèrie d'observacions independents a partir d'objectes d'àudio seleccionats i s'ha observat: $n = 12$, $\bar{x}$ =108.4 kBps, s = 27.2 kBps. Com que el llindar del 97.5% per a una t amb 11 graus de llibertat val 2.200, trobem un IC(95%) per a la mitjana del flux:

$$\begin{aligned}\mathrm{IC}(\mu, 95\%) &= [108.4 - 2.200 \cdot 27.2/\sqrt{12},\ 108.4 + 2.200 \cdot 27.2/\sqrt{12}] \\ &= [91.13, 125.67]\end{aligned}$$

Per tant, creiem, amb aquesta confiança, que el flux mitjà pot estar entre 91.13 i 125.67 Kbps.

Exemple. Repetim ara el càlcul de l'interval de confiança per a l'ocupació mitjana d'un vehicle, sense utilitzar cap informació suposadament poblacional. Recordem que la mostra disponible es resumeix en aquest quadre:

Ocupants	1	2	3	4	5
Freqüència	32	13	3	2	0

Ja sabíem que l'ocupació mitjana observada era 1.5, però ara necessitem l'estimació de la variància. Es pot calcular amb relativa facilitat:

Taula 3.1 Càlcul de la variància mostral

Ocupants (x_i)	Freqüència (n_i)	$(x_i - \bar{x})^2$	Acumulat parcial $n_i\,(x_i - \bar{x})^2$
1	32	0.25	8
2	13	0.25	3.25
3	3	2.25	6.75
4	2	6.25	12.5
	n=50		30.5 s^2=30.5/49=0.622

El valor de la desviació tipus mostral és 0.789. El llindar del 95% per a una distribució t de Student amb 49 graus de llibertat, trobat amb un full de càlcul, val 2.0096. Per tant, l'IC(μ, 95%) resultant s'obté com:

$$1.5 \pm 2.0096 \cdot 0.789/\sqrt{50} = 1.5 \pm 0.2242 = [1.276, 1.724]$$

Creiem, amb una confiança del 95%, que la mitjana poblacional del nombre d'ocupants d'un vehicle és algun dels valors compresos entre 1.276 i 1.724.

Nota. El lector atent s'haurà adonat que la variable de l'exemple anterior no s'assembla gens a la distribució normal. Per tant, el resultat està afectat per la sospita que prové del fet que una de les premisses del procediment no es compleix. A continuació, es comenten les implicacions d'aquest aspecte.

b) Premisses per fer una estimació de μ sense conèixer σ.

Nota. Una premissa és una proposició que ha estat provada anteriorment o s'ha donat com a certa i que serveix de base a un argument o discussió. La paraula que s'utilitza en anglès és *assumption*, i altres denominacions són: assumpcions, hipòtesis prèvies necessàries, requisits, condicions d'aplicació...

Per poder afirmar que l'estadístic *T* segueix una distribució t de Student amb *n*-1 graus de llibertat, ens basem en la premissa que la variable en estudi segueix una distribució normal.

Ara bé, si la grandària mostral creix, l'estimació s^2 de σ^2 és millor —s'apropa al valor real—, amb la qual cosa la substitució de σ^2 per s^2 té implicacions menors. Per aquesta raó, encara que la variable mostrada no sigui normal, es pot assumir que l'estadístic *T* s'apropa a la normal per a grandàries mostrals molt grans (com a mínim, de l'ordre d'una centena).

Exemple. Quan hem trobat un interval de confiança per a l'ocupació mitjana d'un vehicle, sense utilitzar la desviació poblacional, hem ignorat una premissa, ja que la variable és clarament no normal (és discreta i ni tan sols simètrica). No sabem si la grandària, *n*=50, es pot considerar segura, ni quines conseqüències té la transgressió comesa. Per tant, encara que és un resultat millor que no saber res, hem d'interpretar el resultat amb prudència. Per procediments simples, basats en la simulació de mostres, podríem veure que la confiança real se situa 1 o 2 punts per sota del nivell del 95%. És a dir, el procediment de l'IC(95%) aplicat a mostres de grandària 50 d'aquesta població proporciona intervals que encerten 93-94% de les vegades (més endavant veurem amb més detall com s'arriba a aquest resultat).

Així doncs, en el cas de la mitjana, sabem inferir els resultats d'una mostra a la població si disposem d'una variable amb distribució normal, o bé si la mostra és suficientment gran. Per tant, aquestes fórmules ja ens serveixen per solucionar la majoria de les situacions, si el cost d'obtenció de la mostra no és excessiu.

En cas que no es compleixi cap de les dues condicions anteriors (no disposem d'una mostra gran ni d'una variable amb distribució normal), podem recórrer a dos grans grups de solucions: 1) transformar la variable per aconseguir la seva normalitat, o 2) recórrer a procediments estadístics que no requereixin aquesta distribució.

Hi ha diverses transformacions útils per canviar la forma de la distribució de variables no normals. Per a variables definides positives (com "el temps fins a..." o "l'espai des de..."), la transformació logarítmica acostuma a corregir la seva asimetria habitual i aconsegueix distribucions prou aproximades a la normal. A més, com es veurà en casos concrets més endavant, pot facilitar la interpretació dels resultats. Per exemple, si s'assumeix que la desviació tipus del temps d'execució d'un algoritme és proporcional al mateix temps, la transformació logarítmica permet convertir en additiu aquest efecte multiplicador.

Si disposem d'un recompte de fenòmens estranys (de baixa probabilitat), que solen seguir una distribució de Poisson, la transformació "arrel quadrada" acostuma a funcionar bé. Quan es tracta d'estabilitzar la variància d'una binomial (en el supòsit que les dades siguin poc homogènies), Fisher proposa no utilitzar la proporció mostral P directament sinó una transformació x, basada en l'arrel quadrada i l'arcsinus:

$$x = \arcsin\left(\sqrt{P}\right)$$

Nota. El llibre de Peña, esmentat a la bibliografia, exposa la família de transformacions de potències de Box-Cox que permet escollir aquella transformació que millora les propietats estadístiques de la variable estudiada. El logaritme, l'arrel o la inversa formen part d'aquesta família.

Quan tenim una distribució no normal, la segona possibilitat consisteix a recórrer a procediments que no imposin el requisit de la normalitat. En aquest cas, s'ha de consultar la bibliografia, ja que no formen part del plantejament d'aquesta obra.

Nota. Hi ha procediments que, mitjançant el càlcul combinatori, determinen la distribució exacta d'un estadístic sense recórrer a les distribucions de referència que hem vist (normal, khi, t). Com que requereixen càlcul intensiu, fins que no s'ha generalitzat l'ús de l'ordinador han estat molt poc utilitzades. Un altre tipus de proves recorren a estadístics que no impliquen una mètrica per a la variable X en estudi, sinó que es basen en l'ordre de les observacions, i se solen denominar mètodes basats en rangs o també no paramètrics.

3.3 Interval de confiança de π (probabilitat en una distribució binomial)

Les variables categòriques, qualitatives o particions, com ara el tipus d'ordinador, el tipus de sistema operatiu, el tipus d'avaria o el tipus d'algoritme, no prenen valors numèrics, sinó classes, o categories, o atributs nominals. Una variable categòrica pot ser reduïda a variable *dicotòmica* (amb només dues classes) si existeix una categoria de la partició que tingui una rellevància especial. Per exemple, el sistema operatiu d'un ordinador personal podria reduir-se a "Windows o a un altre".

Suposem que un cert ordinador té una probabilitat π d'espatllar-se en el període d'un mes. Al mes de febrer, el resultat pot ser que l'ordinador s'espatlli o no. La variable indicadora de si l'ordinador s'ha espatllat segueix una distribució de Bernoulli (recordem que aquestes variables sí que són quantitatives, i prenen els valors 0 i 1). Si disposem de n ordinadors, la variable X: "recompte del nombre d'ordinadors que s'han espatllat durant un mes" seguirà la distribució *binomial*:

$$X \sim \mathrm{B}(n, \pi),$$

sempre que puguem assumir que: (1) tots els ordinadors vénen de la mateixa població (posem per cas, que tenen la mateixa antiguitat, i potser són del mateix proveïdor); per tant, tots tenen la mateixa probabilitat π d'espatllar-se, i que (2) el fet que un ordinador s'espatlli no modifica la probabilitat que un altre s'espatlli. En resum, podem assumir que el conjunt de les n variables indicadores d'avaria són variables aleatòries independents i idènticament distribuïdes.

Recordem que la distribució binomial té com a esperança i variància:

$$\mu = \mathrm{E}(X) = \pi \cdot n$$
$$\sigma^2 = \mathrm{V}(X) = \pi \cdot (1-\pi) \cdot n$$

Exemple. Si tenim un parc de 100 ordinadors i la probabilitat π d'haver de reinstal·lar el sistema operatiu és de 0.15; el número esperat d'ordinadors que necessiten reinstal·lació en un mes determinat és de $0.15 \cdot 100 = 15$, i la desviació tipus al voltant d'aquest valor (una desviació "mitjana" que esperem observar en el mes) és de $\sqrt{(0.15 \cdot 0.85 \cdot 100)} = 3.57$. Això vol dir que seria normal que en un mes s'haguessin de reinstal·lar 20 ordinadors, i no ho seria si fossin solament 5.

A vegades, pot ser més interessant parlar no del nombre absolut, sinó de la proporció del nombre d'ordinadors afectats. En aquesta situació, podem definir la variable proporció mostral:

$$P = X / n$$

S'ha de tenir present que, igual que $\overline{X}$, P és un estadístic. Es tracta d'un resultat observat en una mostra de n observacions i, com a tal, té una certa distribució de probabilitat. Com que n representa la grandària de la mostra, es tracta d'una constant amb un valor que decidim nosaltres. Per tant, n no és una variable aleatòria, i la forma de la distribució de P coincideix amb la de X. L'esperança de P és:

$$\mathrm{E}(P) = \mathrm{E}(X / n) = \mathrm{E}(X)/n = \pi \cdot n / n = \pi$$

Per tant, P és un estimador sense biaix de π, amb variància:

$$\mathrm{V}(P) = \mathrm{V}(X / n) = \mathrm{V}(X)/n^2 = \pi \cdot (1-\pi) \cdot n/n^2 = \pi \cdot (1-\pi)/n$$

De la mateixa manera que per a l'estimador de la mitjana, la dispersió de l'estimador de la probabilitat va disminuint a mesura que la grandària de la mostra augmenta, fet que demostra formalment el raonament intuïtiu que ja teníem.

Nota. Es podria calcular l' IC(π, 95%) mitjançant un càlcul exacte basat en la binomial.

Suposem que estem auditant la qualitat de la documentació del codi d'un departament de programació que implementa centenars d'algoritmes a l'any. Un cop estudiats 10 programes, en trobem 8 que sí que compleixen les normes de qualitat. Què sabem sobre l'autèntica probabilitat π que un programa d'aquesta població estigui ben documentat? No fa falta fer molts càlculs per saber que π no pot ser 0, ni tampoc 1. Vegem quins altres valors poden ser raonables i quins no. Si assumim que π=0.8, la probabilitat d'observar en una mostra de n=10, X=8, val:

$$\mathrm{P}[X = 8 \mid X \sim \mathrm{B}(10, 0.8)] = C_8^{10}\pi^8(1-\pi)^2 = 0.302$$

Ara bé, si π fos 0.3:

$$\mathrm{P}[X = 8 \mid X \sim \mathrm{B}(10, 0.3)] = C_8^{10}\pi^8(1-\pi)^2 = 0.001$$

I la d'observar-ne 8 o més seria:

$$\mathrm{P}[X \geq 8 \mid X \sim \mathrm{B}(10, 0.3)] = C_8^{10}\pi^8(1-\pi)^2 + C_9^{10}\pi^9(1-\pi)^1 + C_{10}^{10}\pi^{10}(1-\pi)^0 = 0.002$$

Per tant, π=0.3 no és un valor raonable.

Podem proposar com a estimadors $\hat{\pi}$ aquells valors de π pels quals la probabilitat d'observar 8 observacions o més extremes es troba al nivell desitjat α.

Per exemple:

$$\text{Límit inferior } \mathrm{IC}(\pi, 95\%) = \hat{\pi} \mid \{\mathrm{P}[X \geq 8 \mid X \sim \mathrm{B}(10, \hat{\pi})] = 0.025\} = 0.444$$
$$\text{Límit superior } \mathrm{IC}(\pi, 95\%) = \hat{\pi} \mid \{\mathrm{P}[X \leq 8 \mid X \sim \mathrm{B}(10, \hat{\pi})] = 0.025\} = 0.975$$

És a dir, valors del paràmetre π per sota de 0.444 o per sobre de 0.975 fan poc probables (<0.05) mostres amb 8 observacions extremes, o més. Per tant, l'interval del 95% de confiança del paràmetre π va de 0.444 a 0.975:

$$\mathrm{IC}(\pi, 95\%) = [0.444, 0.975]$$

En altres paraules: havent observat 8 de 10 programes amb una documentació correcta, l'únic que podem garantir (amb un risc α = 0.05) és que l'autèntica probabilitat que un programa estigui ben documentat en aquest departament és algun valor entre 0.444 i 0.975.

Fixem-nos en la gran amplitud d'aquest interval com a resultat d'una grandària mostral petita per a una variable dicotòmica (ja que aquestes porten la menor informació possible).

Un càlcul exacte és complicat, perquè comporta resoldre equacions no lineals molt complexes. Si disposem de mostres grans, és més còmode recórrer a la convergència de la distribució binomial a la distribució normal.

$$P \sim \mathrm{N}\left(\pi, \pi\cdot(1-\pi)/n\right)$$

Es pot veure que l'aproximació de la binomial a la normal és millor com més gran és el nombre d'observacions, i més allunyat de 0 i de 1 està el valor de π. Normalment, s'accepten com a *condicions d'aplicació* de l'aproximació normal:

$$\pi \cdot n \geq 5$$
$$(1-\pi) \cdot n \geq 5$$

Fent servir la convergència a la normal, el càlcul de l'interval de confiança és molt més simple i quasi idèntic al de μ:

$$\mathrm{IC}(\pi, 1-\alpha) = P \pm z_{1-\alpha/2}\,\sigma_P = P \pm z_{1-\alpha/2}\sqrt{[\pi\cdot(1-\pi)/n]}$$

Recorda aquella situació paradoxal en què per fer l'estimació de μ era necessari conèixer σ? Doncs ara ens hem superat: necessitem conèixer π per poder estimar la variabilitat de π, la qual, al mateix temps, necessitem per poder estimar π. Vegem-ne dues solucions possibles. La primera té en compte que la funció $f(\pi)=\pi\cdot(1-\pi)$ té un màxim en 0.5: es demostra derivant aquesta funció i igualant a zero, i es pot comprovar a la taula 3.2.

$$f\,'(\pi) = 1-2\pi = 0 \quad \Rightarrow \quad \pi = ½$$

Taula 3.2 Variació del producte $\pi(1-\pi)$ segons el valor de π

π	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$1-\pi$	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$\pi(1-\pi)$	0.09	0.16	0.21	0.24	0.25	0.24	0.21	0.16	0.09

Per tant, es pot adoptar una actitud conservadora i dir que, en una mostra de grandària n, la dispersió de l'estadístic P val, com a molt:

$$\sigma_P = \sqrt{[\pi(1-\pi)/n]} = \sqrt{[0.5(1-0.5)/n]} = 0.5/\sqrt{n}$$

per la qual cosa el càlcul de l'interval de confiança $1-\alpha$ de π és:

$$\mathrm{IC}(\pi, 1-\alpha) = P \pm z_{1-\alpha/2}\,\sigma_P = P \pm z_{1-\alpha/2}\sqrt{[0.5\cdot(1-0.5)/n]} = P \pm z_{1-\alpha/2}\,0.5/\sqrt{n}$$

Exemple. Si tirem 100 vegades una moneda a l'aire i observem 56 cares:

$$\begin{aligned}\mathrm{IC}(\pi, 95\%) &= P \pm z_{1-\alpha/2}\sqrt{[0.5\cdot 0.5\,/n]}\,) = \\ &= 0.56 \pm 1.96\cdot 0.5/\sqrt{100} \approx \\ &\approx 0.56 \pm 0.10 = \\ &= [0.46, 0.66] \\ \text{o millor: } &[0.4620, 0.6580]\end{aligned}$$

Per això afirmarem, amb una confiança del 95%, que la proporció de vegades que surtirà cara amb aquesta moneda està dins del interval [0.46, 0.56].

Es pot dir que $0.5/\sqrt{n}$ és el valor de l'error tipus de P en la situació de màxima indeterminació. Té l'avantatge que, donada una grandària mostral determinada, tenim el mateix valor per a qualsevol variable binomial que estimem. L'inconvenient s'aprecia ràpidament a la taula 3.2: si la π fos molt petita o molt gran, la variància estaria sobreestimada i l'IC seria considerablement més gran del que caldria, amb la pèrdua de precisió que això comporta.

La segona possibilitat consisteix a substituir π per P, tal com vam fer amb σ^2 per s^2. Vista l'expressió de l'error tipus de l'estimador P ($\sigma_P = \sqrt{[\pi(1-\pi)/n]}$), podem afirmar que l'error introduït en substituir π per P tendeix a zero a mesura que n augmenta; per tant, aquesta substitució serà vàlida per a mostres grans. Ara, el càlcul de l'interval de confiança $1-\alpha$ de π és:

$$\text{IC}(\pi, 1-\alpha) = P \pm z_{1-\alpha/2}\, \sigma_P = P \pm z_{1-\alpha/2} \sqrt{[P(1-P)/n]}$$

A *l'exemple* anterior del llançament de 100 vegades d'una moneda, en què observem 56 cares:

$$\text{IC}(\pi, 95\%) = P \pm z_{1-\alpha/2} \sqrt{[P(1-P)/n]}) = P \pm z_{1-\alpha/2} \sqrt{[0.56 \cdot 0.44 /n]}) \approx$$
$$\approx 0.56 \pm 0.10 = [0.46, 0.66] \quad \text{o millor: } [0.4627, 0.6573]$$

S'ha de tenir present que, al nivell de precisió que estem treballant, ambdós procediments condueixen quasi al mateix interval. Això és perquè, en aquesta situació, ens trobem a prop del màxim. Si estiguéssim estimant un fenomen més estrany, amb una π allunyada de 0.5, la concordança entre ambdós procediments seria menor.

Exemple. Com podem saber si l'interval de confiança per a μ respon al nivell establert, per exemple, al 95%? Això es garanteix si les premisses es compleixen en la població estudiada però, si no és així, quines conseqüències porta l'incompliment de la premissa? Examinarem el cas de l'ocupació dels vehicles, que s'ha resolt —inadequadament— en un exemple anterior.

Amb Minitab hem simulat l'observació de 5000 mostres, cadascuna de 50 vehicles, i n'hem calculat els 5000 intervals corresponents, sense fer ús de la desviació poblacional (vegeu-ne l'exemple de l'apartat 3.2.2.a). El resultat obtingut de la simulació és que 4686 intervals (el 93.72%) contenien el veritable valor de la mitjana, 1.63. Si amb les dades de la simulació volem construir un interval de confiança per a la veritable confiança que s'ha d'esperar amb les condicions de l'experiment dels vehicles, resulta que:

$$\begin{aligned}\text{IC}(\pi, 95\%) &= P \pm z_{1-\alpha/2} \sqrt{[\, P\,(1-P)/n]}) = \\ &= 0.9372 \pm z_{1-\alpha/2} \sqrt{[0.9372 \cdot 0.0628 /5000]}) = \\ &= 0.9372 \pm 0.0067246 = \\ &= [0.93048, 0.94392]\end{aligned}$$

Per aquesta raó, hem dit que el nivell de confiança s'ha reduït en 1 o 2 punts, que tampoc no sembla una reducció rellevant. Tindrem en compte que els intervals per a l'ocupació esperada d'un vehicle haurien de ser lleugerament més amples.

Per cert, aquesta vegada no seria correcte un càlcul conservador per a l'interval de confiança, ja que no es pot assumir que el nivell de confiança sigui a prop de 0.5. Si es fes, trobaríem un interval més ample que sí queconté el valor 0.95.

3.4 Grandària mostral

Quan un experiment inclou un cost relacionat amb el nombre d'observacions que es realitzaran, és important trobar abans el valor de la grandària de la mostra per estimar un determinat paràmetre amb *precisió* suficient. Un valor més alt comporta un encariment de l'experiment innecessari, i aquest pot ser inútil si no es treballa amb un volum de dades adequat.

Recordem que l'amplitud de l'interval de confiança depèn de l'error tipus de l'estimador i del nivell de confiança. Si estem interessats a fitar la nostra incertesa, és a dir, a delimitar l'amplitud d'aquest interval sense alterar-ne el nivell de confiança $1-\alpha$, la solució passa per ajustar l'error tipus d'estimació del paràmetre, i trobar-ne la n òptima. A continuació, estudiem aquests paràmetres.

En el cas de l'estimació de la mitjana poblacional o esperança matemàtica, l'interval de confiança, assumint que coneixem σ, és:

$$\mathrm{IC}(\mu, 1-\alpha) = \bar{x} \pm z_{1-\alpha/2}\, \sigma/\sqrt{n}$$

Si, per exemple, volguéssim que la semiamplitud (una mesura de precisió) de l'interval de confiança valgui A:

$$\begin{aligned} A &= \tfrac{1}{2}\,(\text{límit superior IC} - \text{límit inferior IC}) = \\ &= \tfrac{1}{2}\,(\bar{x} + z_{1-\alpha/2}\,\sigma/\sqrt{n} - (\bar{x} - z_{1-\alpha/2}\,\sigma/\sqrt{n})) = \\ &= z_{1-\alpha/2}\,\sigma/\sqrt{n} \end{aligned}$$

d'on podem aïllar n per conèixer quina grandària mostral ens permet obtenir la precisió desitjada:

$$n = (z_{1-\alpha/2}\,\sigma/A)^2$$

Exemple. Quants vehicles hem d'observar si volem aplicar el resultat per obtenir una estimació del nombre de viatgers que entra en una hora (comptem 6000 cotxes), amb una tolerància de ±500 persones? El nivell de confiança ha de ser del 95%.

Admetem que la desviació poblacional real és 0.9450, tal com hem fet abans. La semi-amplitud de l'interval per a l'*ocupació mitjana per vehicle* (dividint pel nombre de cotxes observats) és:

$$A = \frac{500}{6000} = 1/12$$

Llavors, la grandària de la mostra que hem d'observar serà:

$$n = (1.96 \cdot 0.9450 \,/\, (1/12))^2 = 494$$

Vegem ara el cas de l'estimació d'una probabilitat. L'interval de confiança, en la situació de màxima incertesa (no tenim la *P* que ens permeti una estimació de l'error tipus: recordeu que encara no s'ha fet el mostreig), és:

$$\mathrm{IC}(\pi, 1-\alpha) = P \pm z_{1-\alpha/2} \sqrt{[0.5 \cdot 0.5 / n]}$$

D'on:

$$A = z_{1-\alpha/2} \sqrt{[0.5 \cdot 0.5 / n]}$$
$$n = (0.5\, z_{1-\alpha/2} / A)^2$$

Exemple. Per conèixer el percentatge de vots d'un partit polític, amb una precisió (semiamplitud) per a l'interval de confiança (95%) igual a 1%, quants casos es necessiten? I si en tenim prou amb 5%?

$$n = (0.5\, z_{1-\alpha/2} / A)^2 =$$
$$= (0.5 \cdot 1.96 / 0.01)^2 \approx 9604 \text{ casos}$$

$$n' = (0.5\, z_{1-\alpha/2} / A')^2 =$$
$$= (0.5 \cdot 1.96 / 0.05)^2 \approx 384.16 \rightarrow 385 \text{ casos}$$

3.5 Resum

El quadre següent reflecteix la mecànica de la construcció d'intervals de confiança. Cada un d'aquests passos s'ha de reflectir quan se soluciona un problema.

Pas	ESQUEMA DE LA SOLUCIÓ
1	Definir l'estadístic que cal utilitzar
2	Especificar-ne la distribució
3	Dir les condicions o premisses necessàries
4	Delimitar el nivell de confiança (usualment, $1-\alpha$ =95%)
5	Calcular-ne l'interval
6	Interpretar-ne els resultats

La taula 3.3 reprodueix els punts 1, 2, 3 i 5 per als intervals estudiats en aquest capítol.

Taula 3.3 Intervals de confiança

Dist. X	Paràmetre	Estadístic	Premisses	Distribució	Interval de confiança ($1-\alpha=0.95$)
Normal ($X \to$ N)	μ	$Z = \dfrac{(\bar{x}-\mu)}{\sqrt{\sigma^2/n}}$	σ coneguda	$Z \sim$ N(0,1)	$\mu \in (\bar{x} \pm z_{0.975}\sqrt{\sigma^2/n})$
	μ	$T = \dfrac{(\bar{x}-\mu)}{\sqrt{s^2/n}}$		$T \sim t_{n-1}$	$\mu \in (\bar{x} \pm t_{n-1,\,0.975}\sqrt{s^2/n})$
	σ^2	$X^2 = \dfrac{s^2(n-1)}{\sigma^2}$		$X^2 \sim \chi^2_{n-1}$	$\sigma^2 \in \left(\dfrac{s^2(n-1)}{\chi^2_{n-1,\ 0.975}}, \dfrac{s^2(n-1)}{\chi^2_{n-1,\,0.025}}\right)$
Altres	μ	$Z = \dfrac{(\bar{x}-\mu)}{\sqrt{s^2/n}}$	n "gran" (≥ 100)	$Z \sim$ N(0,1)	$\mu \in (\bar{x} \pm z_{0.975}\sqrt{s^2/n})$
	π (Binomial)	$Z = \dfrac{(p-\pi)}{\sqrt{\pi(1-\pi)/n}}$	$(1-\pi)\ n \geq 5$ $\pi \cdot n \geq 5$	$Z \sim$ N(0,1)	$\pi \in (P \pm z_{0.975}\sqrt{\hat{\pi}(1-\hat{\pi})/n})$ $\hat{\pi} = P \quad o \quad \hat{\pi} = 0.5$
	λ (Poisson)	$Z = \dfrac{(L-\lambda)}{\sqrt{\lambda}}$	$\lambda \geq 5$	$Z \sim$ N(0,1)	$\lambda \in (L \pm z_{0.975}\sqrt{L})$

3.6 Preguntes tancades de resposta única

A continuació, s'inclouen algunes preguntes de tipus test, que poden servir per repassar els temes que s'han tractat als primers capítols.

1. Pel carrer del Tinent Coronel Valenzuela, a les 7.55h del matí, els alumnes de la UPC caminen a una velocitat de mitjana de 3 km/hora. Quina desviació tipus, aproximada, us sembla més raonable?:

a) 0 km/h
b) 1 km/h
c) 3 km/h
d) Totes les respostes anteriors donen valors impossibles.

2. El temps de posada en marxa de 4 portàtils ha estat 10", 20", 20" i 30". L'error tipus de la mitjana val, aproximadament:

a) 4.08 "
b) 7.07 "
c) 8.16 "
d) 14.14 "

3. Tenim una mostra de set observacions:

3.9 7.4 5.5 10 9.5 6.7 8

La suma dels valors és 51 i la suma dels quadrats és 399.36.
Quin és l'error tipus en aquesta mostra?

a) 0.813408
b) 1.373454
c) 2.152075
d) 4.631429

4. Amb un IC(α=5%) per a un paràmetre, podem afirmar que:

a) El 95% de les observacions de la mostra són dins de l'interval.
b) El 95% de les possibles realitzacions de la variable caurien dins de l'interval.
c) El 5% de les observacions d'una nova mostra caurien fora de l'interval.
d) Hi ha una probabilitat d'un 5% que el paràmetre no estigui dins de l'interval.

5. Per fer un interval de confiança de la variància d'una variable:

a) És suficient amb una mostra de 30 observacions.
b) Cal que la variable sigui normal.
c) Cal que la variable sigui una distribució simètrica.
d) Totes les respostes anteriors són certes.

6. Amb 100 bateries de portàtil de la marca ACME, hem calculat un interval de confiança al 95% per a la μ de la variable "durada de les bateries". L'interval obtingut ha estat (10.5 h, 13.5 h). Quina és la mitjana de les 100 bateries?

a) 10.5 h
b) 12.0 h
c) 13.5 h
d) No es pot saber a partir de l'interval de confiança.

7. Amb 100 bateries de portàtil de la marca ACME, hem calculat un interval de confiança al 95% per a la μ de la variable "durada de les bateries". L'interval obtingut ha estat (10.5 h, 13.5 h). Podem calcular un interval de confiança per a la desviació estàndard de la variable en qüestió amb la informació que tenim?

a) Sí, òbviament.
b) És impossible, ni suposant normalitat.
c) Només si suposem que la variable durada és normal.
d) No, perquè les bateries són de la marca ACME.

8. La mitjana mostral seguirà una distribució normal...

a) ... sempre.
b) ... quan coneixem la variància.
c) ... quan la variable original sigui normal.
d) Totes les respostes anteriors són correctes.

9. Les condicions que afavoreixen la normalitat de la mitjana mostral són:

a) La normalitat de la variable en estudi.
b) Una major grandària mostral.
c) L'aleatoritat de la mostra i la independència de les observacions.
d) Totes les respostes anteriors són correctes.

10. Quina de les afirmacions següents és certa?

a) Entre les propietats desitjades dels estimadors hi ha el biaix i la variància.
b) La variància de la mitjana mostral s'anomena error tipus de la mitjana.
c) Entre dos estimadors no esbiaixats escollim el de variància mínima.
d) Totes les respostes anteriors són correctes.

11. En estimar la mitjana, l'interval de confiança del 99%:

a) Inclou el 99% de les mitjanes poblacionals.
b) Inclou el 99% de les mitjanes mostrals.
c) Inclou la mitjana poblacional el 99% de les vegades.
d) Inclou la mitjana mostral el 99% de les vegades.

12. Donada una variable aleatòria normal de desviació tipus 1.2, quina grandària mostral necessitem per determinar la mitjana poblacional amb una precisió (la meitat de l'amplada de l'IC) de 0.3 si treballem a un nivell de confiança del 95%?

a) 8
b) 60
c) 62
d) 100, perquè altrament no podem aproximar la mitjana per una normal

13. Donada una mostra d'una població que segueix una distribució normal, volem calcular un interval de confiança per a la mitjana poblacional. Per fer-ho, emprarem un estadístic que segueixi una distribució:

a) Si coneixem la desviació tipus poblacional, és millor utilitzar una N(0,1).
b) Si coneixem la desviació tipus poblacional, és millor utilitzar la t de Student.
c) Si coneixem la desviació tipus poblacional, no és pot utilitzar una N(0,1).
d) Sempre una N(0,1).

14. Donada una mostra aleatòria simple de grandària 200 d'una variable X de variància 25, la mitjana mostral és 54. Llavors, podem afirmar que...

a) ... l'esperança de X es troba dins de l'interval [53.31, 54.69] (α =5%).
b) ... l'esperança de X es troba dins de l'interval [53.23, 54.77] (α =5%).
c) ... l'esperança de X és 55.
d) ... la variable és normal perquè n és gran.

15. Volem estimar la mitjana del temps d'execució d'un algorisme amb un interval de confiança al 95% i, disposant d'una n gran, en què cada banda sigui la desena part de la desviació de tipus mostral (és a dir, la semiamplada val 0.1·s), quina grandària mostral necessitarem?

a) 20
b) 39
c) 196
d) 384

16. Volem estimar per interval de confiança l'esperança d'una variable de la qual disposem de la mostra següent: 12.3, 9.9, 3.0, 2.2, 2.0, 1.8, 0.5, i sabem que σ és 3.

a) (2.31, 6.75) al 95%.
b) (0.27, 8.79) al 95%.
c) No seria fiable; la mostra és massa petita.
d) No podem treballar; no sembla venir d'una distribució normal.

17. Tenim aquesta sortida d'un interval de confiança donat per Minitab:

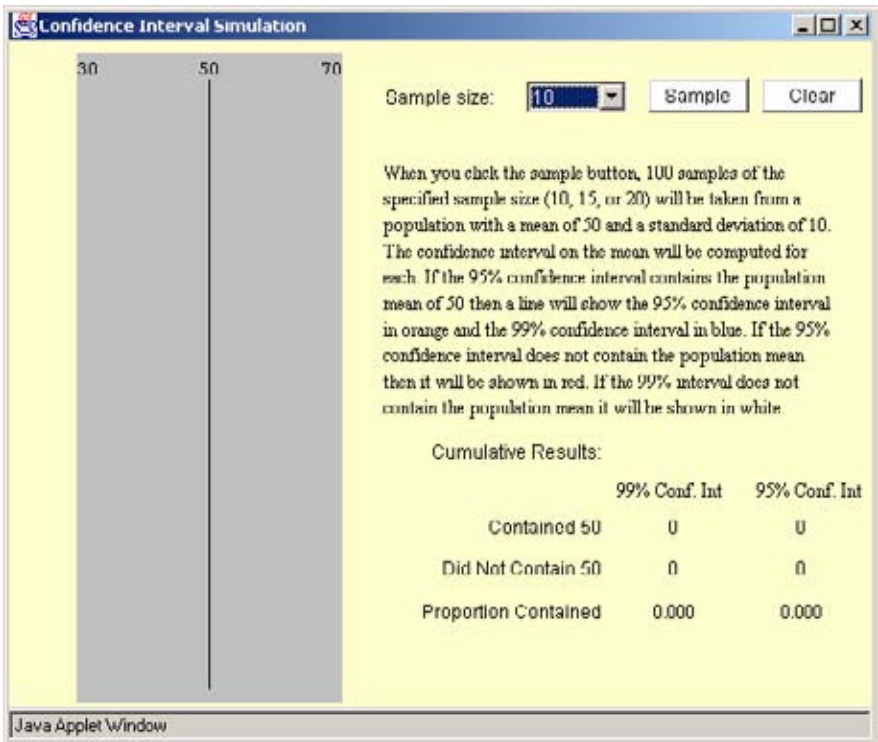
```
Variable     N      Mean    StDev  SE Mean      95.0 % CI
C3           7       ***     ***     1.01     (8.31, 13.26)
```

a) La mitjana mostral és 10.785 i la desviació típica mostral és 2.676.
b) La mitjana mostral és 10.785 i la desviació típica mostral és 3.341.
c) La mitjana poblacional és 10.785 i la desviació típica poblacional és 2.676.
d) La mitjana poblacional és 10.785 i la desviació típica poblacional és 3.341.

Respostes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	b	a	a	d	b	b	c	c	d	c	c	c	a	a	d	a	a

3.7 Guia de treball

Aneu a la pàgina web <http://www.ruf.rice.edu/~lane/stat_sim/conf_interval/index.html> i feu clic a Begin. Veureu una finestra com la següent:



Com podeu llegir, es tracta de simular, repetides vegades, mostres de grandàries 10, 15 o 20, d'una distribució N(50,10). Amb cada mostra es calcula l'IC amb σ desconeguda, que es representa amb una línia (de fet, es representen dos IC: el del 95% de confiança i el del 99% de confiança). Cada vegada que es prem el botó Sample, es generen 100 mostres, i els resultats s'acumulen a la taula de la part inferior, on podeu comprovar la coincidència dels resultats empírics amb els esperats.

Recordeu que l'IC es troba amb l'expressió:

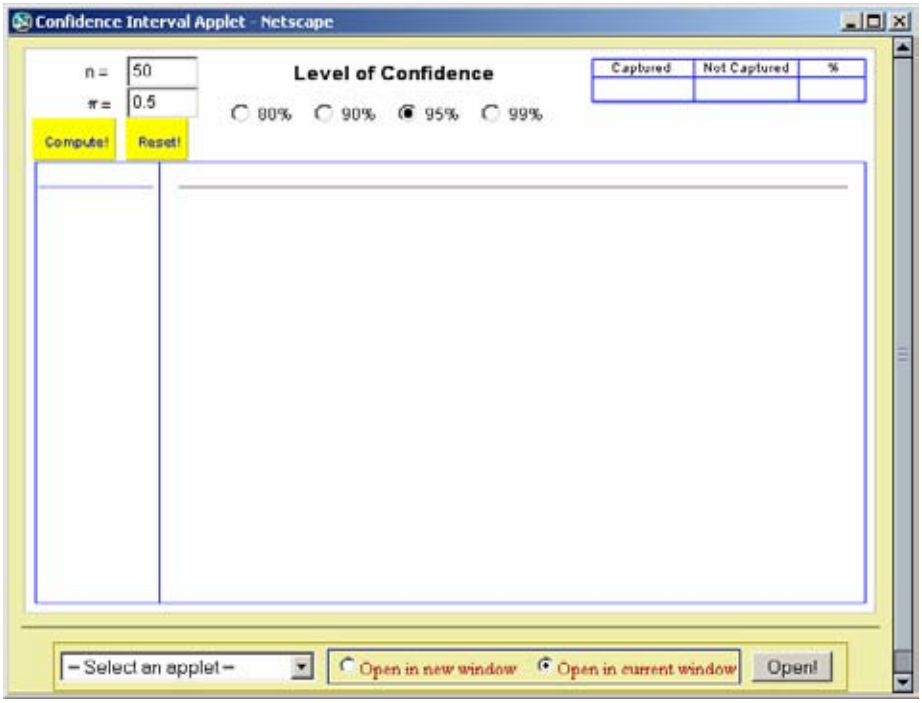
$$\mathrm{IC}(\mu,1-\alpha)=\bar{x}\pm t_{n-1,\alpha/2}s_{\bar{X}}=\bar{x}\pm t_{n-1,\alpha/2}s/\sqrt{n}$$

Qüestions

1. Per què els intervals estan desplaçats a la dreta i l'esquerra?
2. Per a una mida mostral n = 10, trobeu la proporció exacta dels intervals que tindran el centre per damunt de 55.
3. Una mostra determinada ens ha donat una mitjana μ = 55 i una desviació tipus s = 8. Trobeu els intervals de confiança del 95 i el 99%.

4. Preneu la mida de la mostra igual a 10, i genereu-ne 100 mostres. Quin percentatge dels IC(μ, 95%) s'espera que incloguin el valor 50? Quants l'inclouen realment?
5. Esborreu (Clear) i genereu 1000 mostres (premeu deu vegades el botó Sample). Quants IC(μ, 95%) que no continguessin el valor 50 esperàveu, i quants n'heu observat?
6. Quin és més ample, l'interval (95%) o l'interval (99%)? Per què?
7. Si dupliquem la mida de la mostra, com afecta l'ample de l'IC? Comproveu-ho.
8. Augmentarà o disminuirà el nombre esperat d'IC(95%) que inclouen el paràmetre?
9. Per a una mida donada, l'ample dels intervals (per a una confiança determinada) varia una mica. Per què?
10. Suposeu que aquest applet us deixa modificar el valor de σ. Quins canvis s'apreciarien en una simulació?

Ara aneu a: <http://kitchen.stat.vt.edu/~sundar/java/applets/ConfIntApplet.html>. Veureu una finestra com la següent:



El problema ara és estimar una proporció, la probabilitat que es presenti un esdeveniment determinat, anomenat genèricament *èxit*. Com que suposadament podem repetir l'experiència en condicions iguals i independents, el nombre d'èxits X a n proves es distribueix segons la llei binomial. Podem estimar per IC la proporció (o paràmetre π):

$$\text{IC}(\pi, 1-\alpha) = P \pm z_{1-\alpha/2}\,\sigma_P = P \pm z_{1-\alpha/2}\sqrt{[0.5\cdot(1-0.5)/n]} = P \pm z_{1-\alpha/2}\,0.5/\sqrt{n}, \qquad \text{o bé:}$$
$$\text{IC}(\pi, 1-\alpha) = P \pm z_{1-\alpha/2}\,\sigma_P = P \pm z_{1-\alpha/2}\sqrt{[P(1-P)/n]}$$

L'applet anterior utilitza el segon procediment. Recordeu que l'ús de la variable normal en les expressions anteriors es justifica perquè la variable binomial tendeix a assemblar-se a la distribució normal per a n grans. Les condicions:

$$\pi \cdot n \geq 5 \text{ i}$$
$$(1-\pi)\cdot n \geq 5$$

són criteris per excloure casos molt inadequats, però no en garanteixen una aproximació perfecta.

Qüestions

1. Com podem veure que l'applet utilitza el segon procediment?
2. Poseu n=100, π=0.25, confiança=95%, i premeu Compute! fins a acumular 500 mostres (vint vegades). Quin percentatge d'intents no han inclòs el paràmetre?
3. Si us hi fixeu bé, quan la x (el recompte, que trobareu a l'esquerra) és molt petita, l'IC no inclou el valor 0.25, i tampoc si és molt gran. Digueu quins són els valors (x_1 i x_2) que pot prendre x per tal que l'IC contengui el valor de π.
4. Preneu una eina de càlcul per calcular la probabilitat exacta que una variable B(100, 0.25) caigui entre x_1 i x_2. És igual o semblant a 0.95?
5. Empreu ara n=25, i repetiu el mateix procés. Com interpreteu l'eficàcia de l'IC per estimar el paràmetre en ambdós casos?
6. Quina és la variable normal (gaussiana) que aproxima la binomial de l'apartat amb n=100? Quina és l'amplada de l'IC que correspon segons aquesta aproximació? I amb n=25?

Respostes de la guia de treball

IC (mitjana de variable normal)

1. Perquè estan centrats a la mitjana mostral, que varia cada vegada.

2. $\overline{X} \sim N(50,\sqrt{10})$: ¿$P(\overline{X} > 55)$?
   $P(\overline{X} > 55) = 1-P(\overline{X} < 55) = 1-P(Z <(55-50) / \sqrt{10}) = 1-P(Z < 1.58) = 0.057$

3. $\bar{x} = 55$, $s = 8$, n=10
   IC(μ, 95%) = $55 \pm t_{9,\,0.025}\ 8/\sqrt{10} = 55 \pm 2.2622 \cdot 2.53 = [49.28, 60.72]$
   IC(μ, 99%) = $55 \pm t_{9,\,0.005}\ 8/\sqrt{10} = 55 \pm 3.2498 \cdot 2.53 = [46.78, 63.22]$

4. Se n'esperen el 95%. Realment, depèn de la simulació, però hauria de ser un nombre al voltant de 95 (més/menys tres o quatre, com a molt).

5. Un 5% no haurien d'incloure el valor 50, és a dir, 50 intervals. Se n'observaran uns 50, amb fluctuació més gran (més/menys 12–15).

6. És més ample el de 99%, perquè el terme de la t de Student és més gran en valor absolut:
$$t_{9,\,0.975} = 2.262;\ t_{9,\,0.995} = 3.250.$$

7. Si n=20, l'ample de l'IC és menor, ja que l'error tipus ($s/\sqrt{n}$) és més petit. Els resultats de la simulació ho mostren clarament.

8. No canvia! L'IC es construeix per una confiança predeterminada, sigui quina sigui la mida mostral.

9. Si coneguéssim σ, l'ample de l'IC seria constant, però s'utilitza una estimació de la desviació tipus que és diferent a cada mostra.

10. Si augmentéssim σ, les mostres tindrien una dispersió més gran i els IC serien més amples per poder satisfer una confiança donada. I a la inversa si disminuïm σ.

IC (proporció d'un esdeveniment)

1. Perquè els IC són clarament desiguals en amplada, una prova que P intervé en l'estimació de l'error tipus.

2. Al voltant d'uns 25, que són el 5% de 500 (10, més o menys seria l'esperable).

3. Mirant els resultats, veiem que si $x < 18$, l'IC queda a l'esquerra de 0.25, sense tocar aquest valor. I si $x > 34$, l'IC queda estrictament a la dreta de 0.25. Així, x_1=18 i x_2=34.

4. $X \sim B(100, 0.25)$. $P(18 \le X \le 34) = P(X \le 34) - P(X < 18) = F_X(34) - F_X(17) =$
$= 0.98357 - 0.03763 = 0.9459$
No és exactament el 95%, però s'hi assembla molt.

5. Amb n=25 potser observareu que costa apropar-se a la ratio teòrica del 5% de fallades. Aquestes es produeixen quan la x és inferior a 4 o superior a 11. Adoneu-vos que sempre que la x val 3, l'interval és el mateix: (–0.007, 0.247), i si la x és 12, l'interval val (0.284, 0.676). No hi ha valors enters que donin IC amb l'extrem just al valor 0.25. Per tant, la probabilitat de fallada és mes gran que el 5% previst (el doble). $P(4 \le X \le 11) = F_X(11) - F_X(3) = 0.8931$, si $X \sim B(25, 0.25)$, i $1–0.8931 = 0.1069$.

   Observeu que, en aquest cas, $n \cdot \pi > 5$. Clarament, si n és molt gran, l'aproximació de la binomial a una normal és millor i, per tant, les probabilitats es corresponen amb més exactitud.

6. Si n=100, s'aproxima per una llei $N(100 \cdot 0.25, \sqrt{100 \cdot 0.25 \cdot 0.75}) = N(25, 4.33)$. L'amplada de l'interval (95%) per aquesta llei és: $2 \cdot 1.96 \cdot 4.33 = 16.97$, que s'assembla molt a 34–17.

   Si n=25, s'aproxima per una llei $N(25 \cdot 0.25, \sqrt{25 \cdot 0.25 \cdot 0.75}) = N(6.25, 2.165)$. L'amplada de l'interval (95%) per aquesta llei és: $2 \cdot 1.96 \cdot 2.165 = 8.487$, que no s'assembla tant a 11–3.

3.8 Problemes

1. Un programa comercial de traducció automàtica que tradueix de l'anglès al català es basa en un diccionari bilingüe electrònic (que anomenarem A). Un equip d'investigadors en processament del llenguatge natural es proposa testejar l'efectivitat d'aquest diccionari. Per fer-ho, s'agafa un text de 1545 paraules i es passa pel bilingüe electrònic. La traducció proposada amb el diccionari és validada per un filòleg, que considera correcta la traducció de 1154 termes.

   Quina és la proporció amb què el diccionari proposa una traducció incorrecta? Per respondre aquesta pregunta, seguiu les indicacions següents:

   a) Indiqueu quin estimador puntual utilitzeu (no es pot considerar la traducció incorrecta del bilingüe com un fenomen rar).

   b) Feu una estimació puntual d'aquesta proporció.

   c) Indiqueu l'estadístic associat (i llei que segueix) per calcular l'interval de confiança.

   d) Quines condicions són necessàries?

   e) Construïu l'interval de confiança al 90% per a la proporció demanada.

   f) Interpreteu-lo.

   g) Pot ser que la proporció real de traduccions incorrectes no pertanyi a aquest interval?(S/N) Per què?

2. La fiabilitat del software és l'atribut de qualitat del software estudiat més extensament. Es pot definir com la probabilitat que un sistema executi satisfactòriament les funcions requerides en un període de temps especificat (IEEE, 1990). Entre les mètriques emprades per descriure aquesta fiabilitat destaca el temps de reparació del software (ITRS). Treballem en una empresa de software i, per a un cert producte nostre, ens comprometem a fer-ne el telemanteniment. Un client vol la garantia que, en general, aquests temps no sobrepassen els 1000 minuts (16h 40m). Atès que aquests temps solen tenir una distribució asimètrica, amb una cua més llarga per la dreta, aquest client accepta que es treballi amb la transformació logarítmica del temps en minuts i demana com a garantia que: $E(Log_{10}(ITRS)) < 3$. Les dues parts acorden una confiança del 90%. Hem obtingut quatre observacions d'aquesta variable:

ITRS	990	320	500	400
Log10(ITRS)	2,996	2,505	2,699	2,602

   Doneu un interval de confiança que permeti comprovar si es compleixen els requisits de la garantia.

3. En Pere és administrador d'una xarxa docent distribuïda en tres aules. Ha comprat una aplicació que li calcula el nombre total de sessions en xarxa per terminal. Per a la primera setmana d'ús d'aquesta aplicació ha obtingut els resultats següents. Calculeu l'interval de confiança de la

mitjana poblacional del nombre total setmanal de sessions en xarxa en l'aula 2. Suposem que la variable segueix una distribució normal.

```
                                  Individual 95% CIs For Mean
Level       N      Mean     StDev ----+---------+---------+---------+--
Aula 1     10     6,800     2,201  (------*------)
Aula 2     10    11,000     2,261
Aula 3     10     9,900     2,025                (-------*------)
                                  ----+---------+---------+---------+--
                                     6,0       8,0      10,0      12,0
```

4. Al Marc Finestra li agradaria tenir una connexió UMTS per al seu ordinador, i no dependre d'una línia telefònica convencional per accedir al seu correu personal, ja que li sembla que usar el mòbil és prohibitiu. Ha vist una oferta interessant, 50 MB/mes per 10 €, i 0.60 €/mes per cada MB addicional (per exemple, un consum de 55 MB costaria 13 €). Naturalment, pensa que ha d'estudiar amb deteniment les seves necessitats, perquè veu clar que excedir-se del límit de 50 MB és bastant car.

   Finalment, el Marc decideix contractar l'oferta per provar-la sis mesos sense compromís. Els costos d'aquests primers mesos resulten ser els següents:

	1	2	3	4	5	6
Consum	51.4705	58.0895	55.1058	53.7371	41.1703	52.8702
Excedent	1.4705	8.0895	5.1058	3.7371	0	2.8702
Facturat	10.88	14.85	13.06	12.24	10.00	11.72

   Amb els valors dels consum observats ($\Sigma\, x_i$ = 312.44, $\Sigma\, x_i^2$= 16438.187), trobeu l'interval de confiança al 95% per el valor del consum mitjà mensual.

3.9 Solució dels problemes

1.

a) $p = \dfrac{\#\,\text{traduccions incorrectes}}{\#\,\text{total de traduccions}}$

b) $p = \dfrac{1545-1154}{1545} = 1 - \dfrac{1154}{1545} = 0.253$

c) Com que no es tracta d'un fenomen rar, aproximem a la normal. $\dfrac{p-\pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}} \sim Z$

d) n gran, p no massa extrema, [$(1-\pi)n \geq 5$ i $(\pi n) \geq 5$] i mostra aleatòria simple.

e) $$\text{IC}(\pi, 1-\alpha) = \left[p \pm z_{1-\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}} \right]$$

Aproximant π per p:

$$\mathrm{IC}(\pi,1-\alpha)=\left[p\pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right]=\left[0.253\pm 1.645\sqrt{\frac{0.253\cdot 0.746}{1545}}\right]=$$

$$=\left[0.253\pm 1.645\cdot 0.011\right]=\left[0.253\pm 0.018\right]=\left[0.235,\,0.271\right]$$

f) El percentatge de traduccions incorrectes del bilingüe es troba entre el 23.5 i el 27.1%, amb una confiança del 90%.

g) Sí. Perquè la finalitat de l'estimació depèn de la representativitat de la mostra observada. De fet, hi ha un 10% de realitzacions mostrals que donarien lloc a intervals de confiança que no contindrien l'autèntic valor de π. No sabem si l'interval de confiança calculat és un d'aquests, o no.

2. IC(μ, 90%) : $\bar{x} \pm t_{3,\,0.05}\, s/\sqrt{n} = 2.700 \pm 2.353 \cdot 0.212/\sqrt{4} = [2.451,\, 2.950]$

Es pot acceptar que les condicions de la garantia es respecten perquè confiem que el valor de la mitjana poblacional estigui per sota de 3. Premisses: mostra aleatòria simple i Log_{10}(ITRS) amb distribució normal.

3.

$$\begin{aligned}\mathrm{IC}(\mu,\,95\%) &= \bar{x} \pm t_{9,\,0.025}\, s/\sqrt{n} = \\ &= 11.0 \pm 2.262 \cdot 2.261/\sqrt{10} \approx \\ &\approx 11 \pm 1.617 \approx \\ &\approx [9.383,\, 12.617]\end{aligned}$$

4. Mitjana de la mostra: $\bar{x}$ =312.44/6 = 52.07
Variància mostral: s^2 = (16438.187 – (312.44)^2/6)/5 = 33.68; la desviació estàndard val 5.80 €.
Error tipus: $s/\sqrt{n}$ = 5.80/$\sqrt{6}$ = 2.37
Valor crític de la distribució t amb 5 graus de llibertat: $t_{5,0.975}$ = 2.571
Interval de confiança: 52.07 ± 2.571· 2.37 = [45.99, 58.16]

Veiem que el consum mitjà del Marc està entre 45.99 € i 58.16 € amb confiança del 95%. Si no augmenta el seu consum amb el temps, pot confiar en no trobar-se amb una despesa desmesurada. En tot cas, la petita dimensió de la mostra es tradueix en un interval massa ample, que és similar, fins i tot, al rang de la pròpia mostra. En Marc no pot descartar que, en mitjana, es trobi bastant per amunt del límit dels 50MB i, per tant, les seves factures incorporarien, gairebé sempre, un sensible increment sobre el preu base dels 10 €.

Estadísticament parlant, hauríem de preguntar-nos si es plausible la premissa de normalitat sobre el consum de cada mes. Això no pot deduir-se de la petita mostra que hem recollit; necessitaríem més informació relativa als hàbits del Marc. Per exemple, si el seu consum diari és molt irregular és difícil que a la fi del mes es distribueixi en forma de campana.

4 Prova de significació i contrast d'hipòtesis

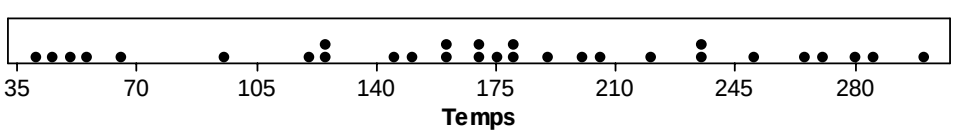
Al capítol anterior hem vist com estimar el valor del paràmetre d'una població a partir de l'estadístic d'una mostra i com obtenir un interval de confiança per a un risc d'error determinat. Volíem saber quins valors del paràmetre són versemblants d'acord amb l'evidència empírica que les dades aporten.

4.1 Prova de significació

En aquest capítol, veurem com contestar preguntes i prendre decisions sobre un valor concret del paràmetre. Partim, doncs, d'un valor que defineix la nostra *hipòtesi* i la posem a *prova*, confrontant-la amb la informació que les dades ens proporcionen.

Exemple. El responsable d'un espai a la web rep moltes pressions perquè els usuaris disposin de la informació sol·licitada sense haver de esperar massa. El problema és que la pàgina més utilitzada fa peticions complexes a altres servidors, i aquestes peticions poden fer que la pàgina trigui a carregar-se un temps gairebé inacceptable (en realitat, estem parlant de segons, però l'usuari té poca paciència quan és davant de la pantalla). El responsable ha acordat amb els seus superiors que un temps de 150 centèsimes de segon (cs) és una mitjana tolerable, però tem que diversos factors hagin incrementat aquest temps.

Finalment, decideix mesurar uns quants temps de resposta de la referida pàgina per mirar de trobar proves que potser el sistema ha canviat i ja no és com ell pensava. Defineix acuradament un protocol per escollir les dades de la mostra que preservi la independència dels valors, i n'obté els següents:



Dels 30 temps observats, 19 superen les 150 cs i el més llarg arriba a 3s! El responsable veu difícil convèncer els seus superiors que el servei no ha empitjorat.

No obstant això, continua buscant algun element dissonant en les dades. Recorda que l'acord que ha establert es refereix al temps mitjà de resposta, i que valors individuals alts es compensen amb altres valors petits (de fet, hi ha temps bastant petits). Calcula la mitjana mostral i la desviació tipus de la mostra:

```
mitjana = 172.9 cs
desv. tipus = 88.3 cs
```

Ara el responsable està molt desanimat. La mitjana és troba per sobre de 150 i s'haurà de passar algunes nits buscant una solució urgent. I si provés amb una altra mostra? Potser aquesta no és vàlida. Però això no estaria bé: està segur que la mostra s'ha observat correctament, solament que l'atzar ha fet que la mitjana sigui gran en lloc de.... Un moment! Igual la desviació de la mitjana respecte del temps desitjat de 150 és casual. La desviació tipus és quasi un segon. Ell entén perfectament què vol dir, perquè està acostumat a veure grans diferències entre uns temps i altres. Dos i tres segons de diferència no són rars. També pot ser que, en una mostra de 30 observacions, una desviació per la mitjana de 23 cs no sigui gens estranya. El que ha de fer el responsable ara és calibrar la importància d'aquesta desviació en comparació de la magnitud dels errors de mostreig.

El responsable sap que la manera d'establir la comparació és construir una hipòtesi de treball:

$$H_0: \mu = 150$$

segons la qual els paràmetres de la població no han canviat, i comprovar amb les dades si la hipòtesi es pot mantenir. Com que sembla plausible acceptar que el temps és normal, aplica el quocient (que hem vist quan fèiem intervals de confiança):

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}},$$

de distribució coneguda —t de Student amb 29 graus de llibertat— si la hipòtesi és correcta. Ara ja sap què ha de dir a l'informe que presentarà als seus superiors. Quan el servei funciona normalment, és a dir, si el temps mitjà de servei és 150 cs, podem observar valors de la mitjana que no s'allunyen massa del valor central μ, i l'estadístic de referència T pren valors d'acord amb la distribució t de Student. En el 95% dels casos amb mostres de grandària 30, el valor de T és inferior a 1.70. En el cas presentat:

$$t = \frac{172.9 - 150}{88.3/\sqrt{30}} = 1.42$$

El valor *t* observat no representa un argument fort per dubtar de la suposició que defensa el responsable, és a dir, que el temps mitjà no ha variat. La figura 4.1 mostra la posició d'aquest resultat respecte de la distribució de tots els valors que pren una t amb 29 graus de llibertat.

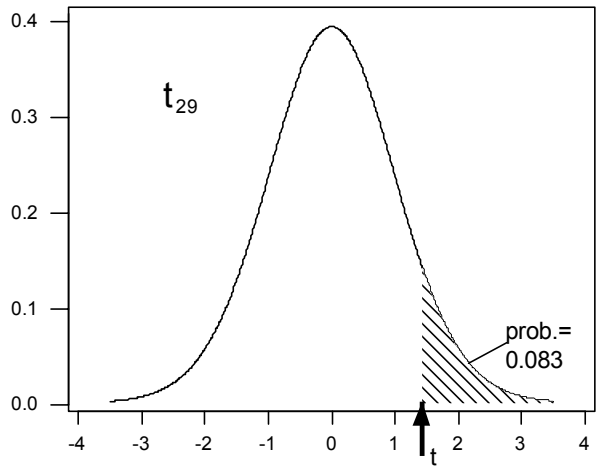


Figura 4.1 Distribució de probabilitat d'una t *de Student amb 29 graus de llibertat. La fletxa assenyala el punt t=1.42.*

Calculant l'àrea per damunt d'1.42 trobem que equival a una probabilitat de 0.083. És a dir, una de cada 12 vegades que es fes la mateixa experiència, assumint que *el temps mitjà de servei* val 150 cs, trobaríem una mitjana mostral més gran que 172.9, i no sembla convincent dir que aquesta freqüència és massa baixa. El responsable no pot assegurar que el sistema funciona tal com s'havia establert, però amb les dades disponibles no hi ha prou força per denunciar una situació fora de control.

A l'exemple anterior veiem com s'utilitza la informació empírica per trobar evidència estadística —en forma de probabilitat— a una situació amb forta presència de l'atzar (perquè és clar que les observacions preses tenen una variabilitat important, que dificulta notablement la presa de decisions). No s'ha considerat convenient utilitzar la via dels intervals de confiança, encara que el resultat seria útil per recolzar la postura del responsable (l'IC al 95% és [139.9, 205.9], que inclou el valor 150: per tant, es podria defensar que el valor acordat és un dels possibles valors de la vertadera mitjana).

La diferència més important entre un IC i una prova de significació és que els intervals proporcionen estimacions del paràmetre i les proves donen una informació quantitativa sobre una determinada hipòtesi H_0 (anomenada *hipòtesi nul·la*). La lògica que s'amaga darrera d'aquesta operació es pot descriure així:

- Si és bastant probable trobar una altra mostra més "extremada" que la disponible, també vol dir que aquesta no deu ser gaire estranya: és compatible amb la hipòtesi nul·la (l'atzar pot explicar sense esforç les diferències trobades).

- Si és poc probable trobar una altra mostra més "extremada" que la disponible, llavors serà més difícil justificar que les diferències entre la mostra observada i la hipòtesi nul·la es deuen només a l'atzar. En el cas que la probabilitat sigui realment petita, la hipòtesi nul·la i la realitat observada es mostren poc compatibles —en major o menor grau—, i es pot dubtar de la versemblança de H_0.

La probabilitat a què fem referència s'anomena *p-valor* (*p-value*, en anglès).

> *Definició*. P-valor és la probabilitat d'obtenir per atzar un resultat més "extrem" que el de la mostra observada, en les condicions que figuren a la hipòtesi nul·la. Si és molt improbable que una cosa succeeixi per casualitat, diem que és estadísticament significativa, amb $p<0.001$, per exemple (és un resultat que s'observa menys d'una vegada cada 1000 intents).

El sentit del que s'ha d'entendre per "extrem" depèn molt del tipus de prova que s'estigui fent. Si es tracta d'una prova sobre el paràmetre μ, llavors pot ser que *extrem* vulgui dir "prendre un valor alt" (com a l'exemple anterior: un temps extrem és un temps alt); pot ser que vulgui dir tot el contrari, és a dir, un valor baix (per exemple, un preu de venda extrem és un preu baix, perquè és el que el venedor no vol), i pot ser que vulgui dir senzillament "allunyat", per sobre i també per sota (per exemple, per a un fabricant de peces, una mida extrema és un valor massa gran o massa petit del que hauria de ser). Es pot pensar que, darrere de cada hipòtesi nul·la, hi ha una altra hipòtesi H_1 (que anomenem *alternativa*, perquè suposa una *altra* opció per al paràmetre estudiat), la forma de la qual es veu a la taula 4.1.

Taula 4.1 Proves de significació segons la hipòtesi alternativa

H. alternativa	p-valor	denominació
H_1: $\mu > \mu_0$	$P(T > t)$	Unilateral per la dreta
H_1: $\mu < \mu_0$	$P(T < t)$	Unilateral per l'esquerra
H_1: $\mu \neq \mu_0$	$P(\|T\| > \|t\|)$	Bilateral

Nota. Amb la disponibilitat actual d'eines de càlcul potents, com ordinadors o fins i tot calculadores de butxaca, ja no resulta difícil trobar valors exactes per a distribucions complexes com ara la t de Student o la khi quadrat. Tanmateix, l'ús tradicional de taules estadístiques (pensades sobretot per fer intervals de confiança) és molt estès encara, i el càlcul de p-valors (i l'ús consegüent de proves de significació) es veu obstaculitzat d'alguna manera.

Exemple. Les bateries noves d'un model d'ordinador portàtil proporcionen una autonomia mitjana de 3 hores. Se sap que, amb el temps, les bateries perden capacitat per emmagatzemar energia, i el seu rendiment cau ràpidament. Una companyia que ha adquirit aquest model per als seus treballadors vol determinar si, al cap d'un any, les bateries

conserven la seva capacitat. Sis empleats experimentats, i amb un perfil d'usuari similar, són escollits perquè participin a la prova, que consisteix a mesurar el temps de treball que permeten les bateries del portàtil fins a l'esgotament total. El resultat és:

Treballador	1	2	3	4	5	6
Temps (min)	110	115	205	145	130	135

La mitjana del temps a la mostra val 140 minuts, i la desviació tipus, 34.35 minuts. Amb aquesta mostra, l'estadístic T pren per valor:

$$t = \frac{140 - 180}{34.35/\sqrt{6}} = -2.85$$

El p-valor d'obtenir un temps mitjà inferior o igual a 140, és a dir, per sospitar que els temps útils han disminuït, val:

$$P(T < -2.85) = 0.018$$

Ens trobem amb un dilema. Si les bateries no haguessin minvat la seva capacitat en un any, seríem davant d'una situació possible però poc habitual (el p-valor equival a menys de 2 casos entre 100). L'altra opció és suposar que la capacitat mitjana ha variat (al cap d'un any és menor). La companyia hauria de valorar si la possibilitat d'una intervenció exclusiva de l'atzar no s'ha de descartar, o si s'ha d'admetre que al cap d'un any les bateries ja mostren pèrdues importants de capacitat (admetent que la mostra era realment representativa, potser els treballadors, conscients de l'experiment, consumien massa recursos?).

Exemple. Analitzem el voltatge de sortida (en volts) d'uns transformadors de corrent que s'adjunten amb un model d'ordinador portàtil. L'especificació diu que s'han d'obtenir 20 V i, com que l'ordinador és molt sensible a petites pertorbacions en la diferència de potencial, les partides que vénen de la fàbrica es controlen acuradament. Una mostra de 16 transformadors ha donat com a resultat una mitjana de 20.0243 V i una desviació tipus de 0.1194 V.

L'estadístic t és igual a $(20.0243 - 20)/(0.1194/\sqrt{16}) = 0.81$. Com que es voldrien desestimar tant els transformadors que donen una tensió massa baixa com els que la donen massa alta, hem de valorar la dificultat d'obtenir una mostra de la mateixa dimensió que aquesta però més extrema (també podríem dir "més sospitosa"). És a dir:

$$P(|T| > |t|) = P(T < -|t|) + P(T > |t|) = 2 \cdot P(T < -0.81) = 0.429$$

El valor obtingut ens diu que és força probable obtenir mostres procedents de transformadors ideals amb desviacions més grans que l'observada. Per tant, la partida analitzada no ens aporta cap evidència que la mitjana poblacional pugui ser diferent de 20.

4.1.1 Proves de significació amb un paràmetre

Veiem a la taula 4.2 un resum de les principals proves que es poden realitzar sobre un paràmetre segons les premisses. Les que s'han vist als exemples anteriors corresponen a la segona fila de la taula. Si es coneix la desviació poblacional (i la variable és normal, o disposem d'un nombre gran d'observacions), és el cas de la primera fila. El cas de la variància de la variable *Y* és a la fila tercera. Finalment, s'inclouen els estadístics i les premisses de les proves d'altres paràmetres: proporció, i taxa d'una variable de Poisson.

Taula 4.2 Proves més usuals sobre un paràmetre de la població

Dist. *Y*	Paràmetre	Hipòtesi	Estadístic	Premisses	Distribució sota H_0
Normal ($Y \to$ N)	μ	$H_0 : \mu = \mu_0$	$Z = \dfrac{\bar{y} - \mu_0}{\sqrt{\sigma^2/n}}$	σ coneguda	$Z \sim N(0,1)$
	μ	$H_0 : \mu = \mu_0$	$T = \dfrac{\bar{y} - \mu_0}{\sqrt{s^2/n}}$		$T \sim t_{n-1}$
	σ^2	$H_0 : \sigma^2 = \sigma_0^2$	$X^2 = \dfrac{s^2(n-1)}{\sigma_0^2}$		$X^2 \sim \chi^2_{n-1}$
Altres	μ	$H_0 : \mu = \mu_0$	$Z = \dfrac{\bar{y} - \mu_0}{\sqrt{s^2/n}}$	n "gran" (≥ 100)	$Z \sim N(0,1)$
	π (binomial)	$H_0 : \pi = \pi_0$	$Z = \dfrac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$	$(1-\pi_0)\, n \geq 5$ $\pi_0\, n \geq 5$	$Z \sim N(0,1)$
	λ (de Poisson)	$H_0 : \lambda = \lambda_0$	$Z = \dfrac{f - \lambda_0}{\sqrt{\lambda_0}}$	$\lambda_0 \geq 5$	$Z \sim N(0,1)$

Exemple. Suposem que el control de qualitat sobre els transformadors consta de dues proves, una sobre la tensió mitjana, que s'espera que sigui de 20 volts, i una altra sobre la desviació de la fluctuació dels transformadors, és a dir, σ, que no hauria de ser superior a 0.05 volts. Aquesta prova es farà sobre una altra mostra. La desviació mostral que hem observat a la segona mostra, de 16 elements també, és 0.0963 V.

La hipòtesi correspon a: H_0: $\sigma^2 = (0.05)^2$ i l'estadístic que es defineix a la taula val:

$$x^2 = \frac{s^2(n-1)}{\sigma_0^2} = \frac{0.0963^2 \cdot 15}{0.05^2} = 55.64$$

Hem de calcular la significació d'aquesta prova com la probabilitat que una distribució de khi quadrat amb 15 graus de llibertat pugui superar el valor 55.64. Aquesta probabilitat és

molt petita: 0.000001394. Segurament, la tensió de sortida d'aquests transformadors varia massa (encara que ho faci al voltant d'un valor acceptable, de mitjana) i hem d'interpretar-ho com que no s'està complint l'especificació.

4.2 Contrast de dues hipòtesis

La prova de significació proporciona una informació molt útil per comprendre millor la qüestió de si una mostra pot provenir d'una població especificada per paràmetres que prenen determinats valors. La prova de significació no dóna una resposta categòrica, ja que el p-valor no és ni un 'sí' ni un 'no', encara que es pot entendre sense dificultat que com més petit sigui el p-valor, més difícil és acceptar la població hipotètica com a població d'origen de la mostra.

Sovint, darrere d'una prova d'aquest tipus hi ha un problema de decisió. El plantejament dels contrastos d'hipòtesis (en anglès, *tests*) és utilitzar les dades, no per obtenir informació sobre el paràmetre o la població, sinó per poder decidir amb arguments sòlids. Per exemple, es podria utilitzar per posar en marxa un procediment que permetés detectar els problemes d'un servei web, i restablir-lo a unes condicions que asseguressin la qualitat del servei. O per considerar la conveniència de buscar un altre proveïdor de bateries, o de transformadors, per al model de portàtil que es comercialitza. Aquestes són opcions que s'obren en el cas que la hipòtesi nul·la sembli insostenible, inversemblant davant l'evidència empírica.

És obligat tenir present que tots els processos de decisió comporten uns determinats riscos. Si un ha de decidir, també es pot equivocar. I és clar que les equivocacions es paguen (si els errors no tinguessin un *cost* associat, errar no tindria cap importància). Imagineu que es conclou que el temps de resposta de la pàgina web s'ha incrementat notablement, i que es creu oportú que el responsable ha de modificar el servidor perquè funcioni com es va establir al seu moment. Potser, després de moltes hores buscant infructuosament un motiu, resulta que el servei continua funcionant amb normalitat. S'hauria perdut el temps d'un tècnic qualificat (potser a costa d'endarrerir altres tasques) per a no res. En aquest cas, hauríem rebutjat la hipòtesi nul·la quan aquesta era correcta.

Però també és antieconòmic equivocar-se en sentit contrari. Suposeu que els transformadors d'una partida tenen un defecte que fa que la tensió de sortida mitjana estigui una mica per sobre dels 20 V ideals, de manera que la gran majoria dels aparells continuïn essent vàlids, però s'incrementi sensiblement la proporció dels que són potencialment perillosos per als circuits de l'ordinador (si aquesta proporció fos, per exemple, d'un 1‰, passar a l'1% seria inacceptable, per la imatge de la marca i pels costos de compensar el defecte als usuaris). Rebutjar la partida seria la decisió correcta; no detectar el problema comporta complicacions i efectes no desitjats.

Els textos d'estadística es refereixen als errors descrits abans en funció de la hipòtesi *alternativa* (totalment excloent de la nul·la), que s'expressa com H_1. A la taula 4.3, veiem l'esquema que s'aplica en un procés de decisió.
També és habitual definir el riscos de primera espècie i de segona (respectivament, α i β) com:

- α: la probabilitat que la hipòtesi nul·la es rebutgi quan en realitat aquesta és certa,
- β: la probabilitat que la hipòtesi alternativa es rebutgi quan en realitat aquesta és certa.

Taula 4.3 Tipus d'errors i riscos

		Decisió	
		H_0 no es pot rebutjar	H_0 es pot rebutjar
Hipòtesi correcta	H_0	OK	Tipus I (risc α)
	H_1	Tipus II (risc β)	OK

Aquestes probabilitats estan relacionades: si volem disminuir el risc α, llavors el risc β augmenta, i viceversa. Quin hauria de ser el criteri per prendre la decisió correcta la major part de les vegades? Òbviament, hem d'esperar que els dos riscos siguin tan petits com sigui possible, però també volem minimitzar el cost que implica qualsevol dels dos errors.

Malauradament, resoldre adequadament aquesta qüestió no queda a l'abast d'aquesta obra. La raó principal és que sovint la hipòtesi alternativa no és una hipòtesi simple, concreta, com la nul·la. Així, la forma habitual d'un contrast d'hipòtesis, amb alternativa *composta*, és:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Aquesta forma vindria a dir: "A la hipòtesi nul·la, el paràmetre estudiat pren el valor θ_0 i, si no, llavors no sabem quin valor prendrà (un altre, és clar)". Aquest argument és inqüestionable, i de vegades és l'única resposta quan no hi ha gaire informació del problema que tractem de resoldre. Com a conseqüència, en no poder precisar quin és l'escenari en cas que H_0 sigui falsa, no es pot quantificar el risc β, és a dir, la possibilitat de quedar-nos erròniament amb la hipòtesi nul·la.

4.2.1 Comparació de dues hipòtesis simples

A continuació, veurem una situació que plantejarem de forma simplificada: es tracta d'un cas on la hipòtesi alternativa també és una hipòtesi simple, i també indicarem uns costos determinats aplicables a les decisions equivocades.

Exemple. Un robot classifica unes peces de dimensions variades com a circulars o quadrades. El mètode es basa en l'anàlisi de la imatge captada per una càmera adossada: es compara el perímetre del perfil de la peça amb el diàmetre màxim. El valor esperat d'aquesta ràtio Y per a un cercle és la constant π (3.14...) i per a un quadrat, la constant $\sqrt{8}$ (2.828…). Per les inexactituds derivades del tractament de la imatge, la mesura conté variacions que equivalen a una desviació tipus igual a 0.1, i es dóna per bo que les ràtios Y segueixen segueixen una distribució gaussiana. La figura 4.2 mostra juntes les distribucions

de les mesures obtingudes per cercles i quadrats (centrades, respectivament, a $\sqrt{8}$ i π), i s'han il·lustrat els riscos α i β suposant arbitràriament que el punt crític per decidir entre les dues formes és igual a 3.

Les peces arriben al sistema amb la mateixa proporció (tants quadrats com cercles), però a l'atzar. Alguna vegada el robot fa una mala classificació, i s'ha determinat que la correcció posterior ocasiona unes pèrdues valorades en 12 unitats si l'objecte era un cercle i 8 unitats si era un quadrat. Amb aquesta informació, hem de determinar quin és el llindar Δ òptim per decidir a cada passa.

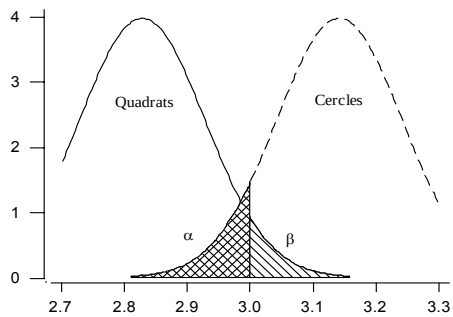


Figura 4.2 Distribució de la ràtio perímetre/diagonal màxima per a quadrats i cercles (les corbes continuen per sota de 2.7 i per sobre de 3.3)

Decidim arbitràriament[12] quina és la hipòtesi nul·la. Per exemple, d'acord amb la il·lustració de la figura 4.2, H_0 és "tenim un cercle"; llavors, H_1 és "tenim un quadrat". En termes del paràmetre μ, valor esperat de la ràtio:

$$\begin{cases} H_0 : \mu = 3.1416... \\ H_1 : \mu = \sqrt{8} \end{cases}$$

La informació empírica disponible es redueix a una única observació, *Y*. Afortunadament, es coneix la desviació poblacional, de manera que podem construir un estadístic de referència. Sota la hipòtesi nul·la,

$$Z = \frac{Y - \pi}{\sigma} = \frac{Y - 3.1416...}{0.1} \sim N(0,1)$$

El p-valor es calcularia com $P(Z < z)$, ja que l'opció plantejada per la hipòtesi alternativa orienta els valors més oposats a H_0 cap avall.

[12] En aquest cas treballem amb dues hipótesis. Si una de les hipótesis és composta, la simple es posa sempre com a nul·la.

L'error tipus I consisteix a confondre un cercle amb un quadrat. Succeirà quan Y sigui petita, o quan Z se situï a les posicions inferiors de la corba normal. Per tant, si decidim que una figura ha de ser un quadrat quan Y sigui inferior a Δ, el risc α val:

$$P\left(Z \leq \Delta - 3.1416.. \Big/ 0.1\right)$$

L'error tipus II consisteix a confondre un quadrat amb un cercle. En aquest cas, no som sota la hipòtesi nul·la, perquè els valors de la mesura Y es distribueixen com una N($\sqrt{8}$, 0.1); per tant, el risc β val:

$$P\left(Z \geq \Delta - \sqrt{8} \Big/ 0.1\right)$$

El cost esperat de qualsevol decisió depèn únicament dels errors que puguem cometre, és a dir:

Cost esperat =12·P(cercle classificat com a quadrat)+8·P(quadrat classificat com a cercle) =
$= 12 \cdot P(C \cap Q_{clas}) + 8 \cdot P(Q \cap C_{clas})$

Amb Q_{clas} expressem que la decisió ha estat classificar la peça com un quadrat, i amb C_{clas} que l'hem classificat com un cercle.

$$= 12 \cdot P(Q_{clas} \mid C) \cdot P(C) + 8 \cdot P(C_{clas} \mid Q) \cdot P(Q)$$
$$= 12 \cdot \alpha \cdot P(C) + 8 \cdot \beta \cdot P(Q) = 12 \cdot \alpha \cdot 0.5 + 8 \cdot \beta \cdot 0.5,$$

ja que sabem que la probabilitat que arribi al robot un cercle, P(C), és la mateixa que hi arribi un quadrat, P(Q). Finalment, tenim una expressió per la qual hauríem de trobar el valor mínim, i per quin valor de Δ correspon. Tenint en compte que hi intervé la funció de distribució per a una normal, una funció sense expressió algebraica senzilla, hem de trobar aquest mínim provant per diversos valors. A la taula 4.4 se'n mostren uns quants.

Taula 4.4 Riscos α i β, i cost esperat en funció del punt crític Δ

Δ	z sota H_0	z sota H_1	α	β	Cost esperat
2.968	−1.736	1.396	0.041	0.081	0.5733
2.969	−1.726	1.406	0.042	0.080	0.5726
2.970	−1.716	1.416	0.043	0.078	0.5722
2.971	−1.706	1.426	0.044	0.077	0.5719
2.972	−1.696	1.436	0.045	0.076	0.5718
2.973	−1.686	1.446	0.046	0.074	0.5719
2.974	−1.676	1.456	0.047	0.073	0.5722
2.975	−1.666	1.466	0.048	0.071	0.5726
2.976	−1.656	1.476	0.049	0.070	0.5732

Tal com es pot veure, el mínim cost esperat es troba prop del punt Δ=2.972. En conclusió, la política més avantatjosa és decidir que si la ràtio Y és superior a 2.972 (o l'estadístic $Z \geq -1.696$) suposem que és un cercle i si és inferior suposem que és un quadrat. El procediment té associat uns riscos: a la llarga, el 4.5% dels cercles seran identificats com quadrats i el 7.6% dels quadrats seran confosos amb una figura circular.

A la pàgina <http://www.kuleuven.ac.be/ucs/java/gent/Ap6a.html> trobareu un applet que simula moltes extraccions d'una variable normal, amb la grandària i els paràmetres que escolliu (vegeu *Settings*). Si no toqueu res i premeu el botó *Run*, veureu com van apareixent a dalt les mostres, d'una N(0,1), i a sota els estadístics (una t amb 29 graus de llibertat). El contrast proposat té com a H_0 una mitjana igual a 1 (ho veieu com una línia blava al gràfic superior). L'objectiu és adonar-se que, com que la població real no coincideix amb la hipotètica, els estadístics no es distribueixen d'acord amb la línia teòrica (corba blava, a sota). Tracteu de reproduir l'experiment anterior, però suposant que es prenen dues mesures independents de la peça (n=2).

4.2.2 Hipòtesi simple contra hipòtesi composta

Malgrat les dificultats que presenta una situació real, on és habitual no disposar d'una alternativa concreta, cal establir mecanismes que permetin resoldre el problema d'una manera satisfactòria. Imaginem que tenim el contrast d'hipòtesis següent:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

on μ_0 és algun valor conegut. A partir d'aquest esquema, es poden trobar uns quants punts conflictius:

- Com es pot prendre una decisió òptima o, almenys, encertada? Encara que coneguem el cost que implica un error de tipus I, sense saber exactament a què es refereix la hipòtesi alternativa és difícil concretar el cost d'un error de tipus II.
- Què vol dir $\mu > \mu_0$? Si el temps mitjà de servei per atendre un usuari que reclama una pàgina web fos 155 cs, el responsable hauria de posar en marxa mesures especials per rebaixar cinc centèsimes de segon aquest temps? Val la pena?
- En realitat, es pot pensar que alguna vegada acceptarem correctament la hipòtesi nul·la, especialment si es tracta d'un contrast bilateral? No existirà sempre alguna diferència numèrica, en contra de H_0? És una hipòtesi impossible de demostrar?

Encara que l'enfocament original (de Ronald Fisher) només considerava les proves de significació i, per tant, no es tractava de decidir entre un parell d'opcions, sembla que el criteri per decidir hauria de tenir en compte el principi que un p-valor petit va en contra de la hipòtesi nul·la i, en conseqüència, a favor de l'alternativa. Això no vol dir de cap manera que un p-valor gran vagi en contra de H_1 perquè, en primer lloc, no sabem quina és aquesta hipòtesi i, sobretot, perquè aquesta és una probabilitat calculada des de la suposició que la hipòtesi nul·la és certa. Si, en lloc de calcular justificadament quin és el valor més adient per a α, especifiquem el seu valor en el protocol (és a dir, abans d'observar la

mostra, per tal de no afectar la nostra objectivitat), podem utilitzar-lo com a referència per decidir quan el p-valor és prou petit. Llavors, si una vegada s'ha observat la mostra i s'ha calculat el p-valor, aquest és més petit que α, pensaríem que H_0 no explica satisfactòriament el comportament de les dades perquè sembla difícil que l'estadístic s'hagi desviat de forma considerable solament per causa de l'atzar i hauríem de rebutjar la hipòtesi nul·la. Aquest sistema pot ser que no sigui el procés més just, a la llarga, però te l'avantatge de la simplicitat.

Exemple. Per il·lustrar les explicacions anteriors, replantegem l'exemple de les bateries de portàtil com un contrast.

H_0: Hipòtesi nul·la: $\mu = 180$
H_1: Hipòtesi alternativa: $\mu < 180$

Suposem que les premisses (mostra aleatòria simple i normalitat de la variable observada) són acceptables i, per tant, sota H_0, l'estadístic establert segueix una t de Student amb n–1 graus de llibertat (fins ara, no ha canviat res). Es decideix adoptar α =0.05, de manera que si el p-valor és inferior rebutjarem H_0. També podem decidir de manera equivalent, comparant l'estadístic t i un cert valor límit corresponent al percentil α d'una t de Student. Com que la nostra mostra consta de 6 observacions, hem de prendre 5 graus de llibertat, i el valor límit és –2.015 (entenem que un valor més petit que aquest és una evidència negativa per a la hipòtesi nul·la, v. figura 4.3).

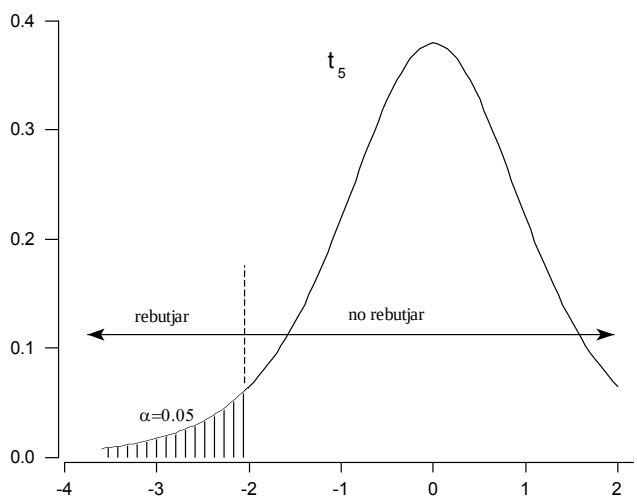


Figura 4.3 Distribució de l'estadístic t (5 graus de llibertat), amb zona de rebuig per l'esquerra definida per a risc α=0.05 (àrea ratllada sota la corba)

La mostra que havíem recollit presentava un estadístic igual a –2.85; per tant, inferior al límit preestablert. Així doncs, amb un nivell de risc de primera espècie del 5% (que vol dir que, si les bateries d'un any conservessin la capacitat de les noves, es rebutjaria H_0 una de

cada vint vegades), s'observen evidències en contra de la hipòtesi nul·la. També podem concloure dient que el p-valor (0.018) és menor que α, per la qual cosa sembla aconsellable rebutjar que la capacitat de les bateries es mantingui al cap d'un any i creure que ha minvat.

L'adopció d'un valor predeterminat per a α té una sèrie d'implicacions que val la pena destacar:

- Primer de tot, el mètode pot ser sensible al valor escollit per α, i portar a decisions arbitràries. Un experiment podria fer rebutjar H_0 al nivell del 10% i no fer-ho per al 5%.

- Es pot fixar el valor de α, però no el de β. Així, controlem la proporció dels contrastos en què H_0 es rebutjarà equivocadament, però no sabem res de les vegades que fallarem conservant una hipòtesi nul·la que no és certa.

- De fet, quan el p-valor és més gran que α, que és el mateix que quan l'estadístic cau dins els límits no crítics, i diem que no es pot rebutjar la hipòtesi nul·la, moltes vegades podem cometre un error subtil: per estalviar paraules, o per creure que el missatge serà més comprensible, diem "acceptem H_0", com si la hipòtesi fos correcta o, pitjor encara, per donar *valor a una confirmació* de la hipòtesi. Aquesta expressió i, més encara, aquesta intenció s'han d'evitar.

- La indefinició anterior no és tal si arribem al cas contrari: quan el p-valor és petit, la hipòtesi nul·la esdevé poc creïble i es rebutja. Com a efecte, donem suport a la hipòtesi alternativa, perquè es tracta d'una situació oposada a la nul·la (sovint, es diu que s'ha trobat una diferència *estadísticament significativa*, en el sentit que és difícil que tingui origen en l'atzar). Amb arguments probabilístics i, en particular, en els contrastos d'hipòtesis, mai no es troben evidències a favor d'una hipòtesi: podem trobar quelcom en contra de H_0, o no poder dir res.

És recomanable, en qualsevol cas, incorporar l'interval de confiança al resultat del contrast d'hipòtesis. D'aquesta manera, l'experimentador fa públic el càlcul de l'estimació del paràmetre, amb la qual cosa es pot apreciar millor: (1) la magnitud de la incertesa present a les dades i (2) la magnitud de la possible desviació del paràmetre respecte del valor estipulat.

Exemple. L'interval de confiança (95%) per la durada mitjana de les bateries d'un any, a partir de les dades recollides, és:

$$\begin{aligned} \mathrm{IC}(\mu,1-\alpha) &= [\,\bar{x} - t_{n-1,\,1-\alpha/2}\; s/\sqrt{n},\; \bar{x} + t_{n-1,\,1-\alpha/2}\; s/\sqrt{n}\,] \\ &= [140 - 2.5706\cdot 34.35/\sqrt{6},\; 140 + 2.5706\cdot 34.35/\sqrt{6}] = [104, 176] \end{aligned}$$

L'interval de confiança troba que el temps mitjà pot anar d'1 h 44 min a 2 h 56 min que, per una banda, és un rang bastant ample però, per l'altra banda, no conté el valor de les 3 hores (encara que no hi és molt lluny). Si, en lloc d'aquest resultat tinguéssim un IC com [130, 150], la nostra lectura de les dades seria molt diferent: disposaríem d'una estimació satisfactòria i estaríem més convençuts que el temps afecta la capacitat de les bateries (la pèrdua estimada per al primer any seria d'entre 30 i 50 minuts).

4.3 Potència d'un contrast

Si la situació en estudi no s'ajusta a la hipòtesi nul·la, quines possibilitats té el contrast de detectar la diferència? Veiem ràpidament que un factor elemental és la semblança entre la hipòtesi nul·la i la realitat: quan la diferència és irrellevant, és molt probable que el contrast acabi determinant que no hi ha evidències en contra de H_0, però, si aquesta hipòtesi planteja una suposició discutible, probablement serà rebutjada amb contundència. Per exemple, si es vol comprovar si bateries de cinc anys duren encara tres hores de mitjana, és fàcil suposar que les dades aportaran molta evidència en contra, per la poca durada real associada a bateries tan velles.

Quan parlem de *semblança*, no hem d'oblidar que es tracta d'un terme relatiu a la variabilitat del fenomen, és a dir, a la dispersió de la variable observada. Si una pàgina web complexa triga una mitjana de 190 cs a mostrar-se (en lloc de 150), aquesta discrepància es pot detectar fàcilment si gairebé totes les pàgines triguen un valor semblant (desviació tipus petita), o difícilment si hi ha peticions que són molt ràpides i altres que triguen quatre o cinc segons.

La grandària de la mostra és un altre factor important. Quan no es pot detectar una diferència existent per l'atzar present a la mostra (gran variabilitat dels individus), la solució passa per observar més individus. D'aquesta manera, l'error tipus de la mitjana disminueix. Finalment, és evident que el valor de α també hi intervé: si prenem un valor molt petit, estem reduint el risc de rebutjar equivocadament una hipòtesi vertadera, però si H_0 no fos certa igualment estaríem fent més difícil trobar rebutjable la hipòtesi nul·la (i incrementant el risc β).

S'anomena *potència* d'un contrast el valor $1-\beta$. Una potència gran és desitjable, perquè això significa que el contrast tindrà moltes possibilitats de rebutjar una hipòtesi falsa i, per tant, de detectar una diferència que pot implicar, per exemple, aplicar mesures correctores o confirmar un progrés respecte de la situació anterior. Com hem dit, el valor de β depèn que la hipòtesi alternativa faci referència a un valor concret per al paràmetre. Així, en un contrast amb alternativa composta, la potència es pot pensar com una funció del possible valor real del paràmetre, fent-lo variar. Il·lustrem la idea amb un cas suposat.

Exemple. Quin poder[13] tenen les 30 mesures del responsable de la pàgina web? Per estudiar el risc β d'un contrast sobre la mitjana del temps, assumeix que la desviació tipus poblacional és 100, un valor que la seva experiència troba raonable. Pren com a risc α=0.05. Es planteja un contrast unilateral per la dreta (no hi ha sospites ni preocupació de que els temps s'hagin escurçat):

H_0: Hipòtesi nul·la: $\mu = \mu_0 = 150$
H_1: Hipòtesi alternativa: $\mu = \mu_1 > \mu_0$

i, atès que es coneix σ, es compleix que l'estadístic següent segueix una llei normal:

[13] *Poder* i *potència* tenen la mateixa etimologia. En anglès, és difícil fer un joc de paraules, ja que ambdues són *power*.

$$Z = \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

Per tant, H_0 es rebutja si l'estadístic $z = (\overline{X} - 150)/(100/\sqrt{n})$ està per sobre del valor crític $z_{0.95} = 1.645$ o, per dir-ho d'una altra forma, si la mitjana observada és superior a $150 + 1.645 \cdot 100/\sqrt{30} = 180.0$.

Llavors, el risc β és la probabilitat $P(\overline{X} < 180.0 \mid \overline{X} \sim N(\mu_1, 100/\sqrt{30}))$. Si estandarditzem la variable:

$$\beta = \mathrm{P}\left(Z \le \frac{180 - \mu_1}{100/\sqrt{30}} \right) = \mathrm{P}(Z \le z_1)$$

Posem uns quants valors per observar com canvia el risc i, en conseqüència, la potència:

Taula 4.5 Risc β i potència segons el valor hipotètic del paràmetre

μ_1	z_1	β	Potència
160	1.0954	0.8633	0.1367
170	0.5477	0.7081	0.2919
180	0.0000	0.5000	0.5000
190	-0.5477	0.2919	0.7081
200	-1.0954	0.1367	0.8633
210	-1.6432	0.0502	0.9498

Ara podem analitzar diversos supòsits. Si s'hagués produït un retard en la mitjana de 20 centèsimes sobre l'estàndard 150, un contrast com el que es prepara solament detectaria el retard el 29% de totes les vegades que es produís (suposant que es pogués repetir indefinidament). Però, si en realitat el retard fos de mig segon (50 cs), la potència arribaria a ser del 86%, que és un valor molt satisfactori.

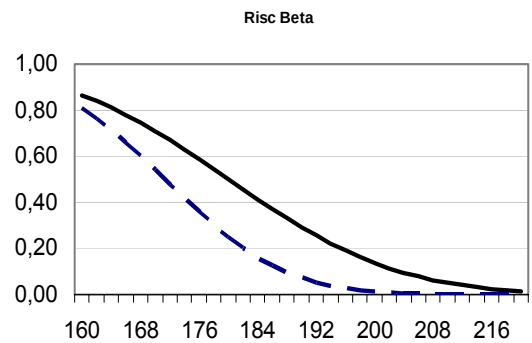


Figura 4.4 Risc de segona espècie de l'exemple, amb n=30 (línia contínua) i n=60 (línia discontínua). L'eix d'abscisses mostra diferents valors de μ_1, el paràmetre en la hipòtesi alternativa.

La figura 4.4 mostra dues corbes del risc β per a aquest cas: la línia contínua és el risc amb grandària mostral 30, mentre que la línia discontínua empra una grandària igual al doble. Es posa de manifest que, com més gran és la mostra, més petit és el risc β i més fiable és el contrast per detectar diferències respecte del valor hipotètic a la hipòtesi nul·la.

Exercici. Trobeu sobre el gràfic de la figura 4.4 quin retard aproximat es podria detectar amb α=0.05 i risc β 20%, prenent mostres de mida 30 i 60, respectivament. Repetiu l'operació, però fent-ne el càlcul (solució exacta: 45.4 cs i 32.1 cs).

4.4 Preguntes tancades de resposta única

A continuació, s'inclouen, a tall d'entrenament, unes preguntes amb quatre opcions.

1. Escriviu la hipòtesi nul·la per poder comprovar la situació següent: "Es vol demostrar que una nova versió del sistema operatiu millora el rendiment d'aquest."

a) H_0: la nova versió no en millora el rendiment.
b) H_0: la nova versió en millora el rendiment.
c) H_0: el rendiment amb la nova versió és superior a l'anterior.
d) H_0: totes les respostes anteriors són falses.

2. En un test d'hipòtesi, si la hipòtesi nul·la H_0 és certa...

a) ...és possible cometre dos errors, el de tipus I i el de tipus II.
b) ...només es pot produir l'error de tipus I.
c) ...només es pot produir l'error de tipus II.
d) ...no és possible cap error ja que H_0 és certa.

3. Un fabricant ens ha ofert una tarja amb una velocitat mitjana d'accés de 100 ms i una desviació tipus de 10 ms. Ara ens n'ofereix una de nova amb paràmetres μ =50 ms i σ =5 ms. Com que sospitem que, de tant en tant, ens envia un lot de les targes velles, decidim fer-ne un control de qualitat per tal de garantir que només un 5% dels lots de targes velles passi el control. Si decidim fer-ho amb una prova d'hipòtesi, en H_0 posarem:

a) μ =100 ms i σ =10 ms
b) μ =100 ms i σ =5 ms
c) μ =50 ms i σ =10 ms
d) μ =50 ms i σ =5 ms

4. Un fabricant ens ha ofert una tarja amb una velocitat mitjana d'accés de 100 ms i una desviació tipus de 10 ms. Ara, ens n'ofereix una de nova amb paràmetres μ =50 ms i σ =5 ms. Siguin H_0 : μ =100 ms i σ =10 ms (tarja vella) i H_1: μ=50 ms i σ =5 ms (tarja nova). El risc α de cometre un error de primera espècie és:

a) La probabilitat que una tarja sigui nova.
b) La probabilitat de concloure que una tarja vella és de les noves.
c) La probabilitat de concloure que una tarja nova és de les velles.
d) Totes les respostes anteriors són falses.

5. Un plantejament unilateral (vs. bilateral) de la hipòtesi alternativa... :

a) ... canvia la regió crítica.
b) ... modifica el p-valor.
c) ... pot canviar les conclusions.
d) Totes les respostes anteriors són correctes.

6. Quina és falsa?

a) Per calcular el risc β necessitem una hipòtesi alternativa simple.
b) El risc α i el p-valor fan referència a l'error tipus I.
c) Per disminuir el risc β podem augmentar el risc α o la grandària de la mostra.
d) $1-\alpha$ és la probabilitat que la hipòtesi nul·la sigui certa.

7. "El plantejament d'un test d'hipòtesi clàssic és conservador en el sentit que s'assumeix certa la hipòtesi nul·la fins que no hi hagi una evidència clara del contrari."

a) La frase és certa.
b) La frase és falsa.
c) La frase no es pot respondre, ja que no és conservador.
d) La frase no es pot respondre, ja que el que es vol demostrar és la veritat de la hipòtesi H_1.

8. El p-valor d'un test d'hipòtesi...

a) Quantifica la versemblança de H_0.
b) Dóna la probabilitat d'acceptar la hipòtesi nul·la, quan aquesta és falsa.
c) Les respostes a) i b) són certes.
d) Totes les respostes anteriors són falses.

9. Quines són les assumpcions que s'han de verificar per poder utilitzar la prova d'hipòtesi Z per a una mostra?

a) El nombre d'observacions ha de ser gran i assumir que la desviació tipus mostral aproxima σ.
b) Qualsevol nombre d'observacions, si la distribució de la variable aleatòria és normal i es coneix la desviació tipus poblacional.
c) El nombre d'observacions ha de ser gran i conèixer la desviació tipus poblacional.
d) Totes les respostes anteriors són vàlides.

10. Amb 100 bateries de portàtil de la marca ACME, hem calculat un interval de confiança del 95% per a la μ de la variable "durada de les bateries". L'interval obtingut ha estat (10.5 h, 13.5 h). Plantegem ara els dos tests d'hipòtesi següents, amb $\alpha = 0.05$:

(i) H_0: $\mu = 11.0$ (ii) H_0: $\mu = 10.0$
H_1: $\mu \neq 11.0$ H_1: $\mu \neq 10.0$

Llavors, tenim que els resultats són:

a) (i) acceptem H_0; (ii) acceptem H_0
b) (i) acceptem H_0; (ii) rebutgem H_0
c) (i) rebutgem H_0; (ii) acceptem H_0
d) (i) rebutgem H_0: (ii) rebutgem H_0

11. Si volem contrastar amb un estadístic que segueix una distribució N(0,1):
"$H_0 : \mu \leq 0$ *versus* $H_1 : \mu > 0$", i en avaluar-lo sota la hipòtesi nul·la ens dóna –2, conclourem que...

a) ...la mitjana poblacional és 0.
b) ...la mitjana poblacional és més gran que 0 (amb un marge d'error del 5%).
c) ...res no s'oposa a acceptar H_0.
d) ...hi ha un 95% de possibilitats que la mitjana poblacional sigui 0.

12. Tenim els temps de CPU que el compilador GNU-C requereix per compilar una mostra de 200 programes de mides similars. Volem fer un test d'hipòtesi sobre la μ de la variable "temps de CPU". Els temps de CPU obtinguts no són normals. Quina seria la distribució de referència de l'estadístic que calcularíem?

a) Una *Z*, perquè $n>100$ (llavors, suposem $\sigma = s$ i, pel TCL, també normalitat).
b) Una *Z*, perquè $n>30$ i podem aplicar el TCL.
c) Una t, perquè no coneixem la variància de la variable estudiada.
d) No es pot fer aquest test d'hipòtesi amb la informació que se'ns dóna.

13. Suposem que avaluem una prova d'hipòtesi sobre l'esperança d'una variable *X*:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

i que coneixem la desviació tipus de *X*. L'estadístic de referència de la prova ens retorna el valor –1.4. Si hem pres un risc $\alpha = 5\%$, la conclusió és:

a) Rebutgem la hipòtesi nul·la (p-valor=0.0404).
b) Rebutgem la hipòtesi nul·la (p-valor=0.000).
c) No podem rebutjar la hipòtesi nul·la (p-valor=0.1616).
d) No podem rebutjar la hipòtesi nul·la (p-valor=0.0808).

14. Volem contrastar, amb un nivell de confiança del 95% i segons un estadístic que segueix una t de Student amb 21 graus de llibertat, "H_0: μ=0 *versus* H_1: $\mu \neq 0$". Si anomenem *X* l'estadístic en el què hem substituït μ per 0, decidirem rebutjar H_0 si:

a) $P(X < 2.08) = 0.975$
b) $P(X > 2.08) = 0.025$
c) $X < -2.08$ o $X > 2.08$
d) Totes les respostes anteriors són correctes.

15. Entre les assumpcions d'una prova sobre una proporció, "$n\cdot\pi > 5$ i $n\cdot(1-\pi) > 5$", és una manera de dir…

a) …que la mostra no ha de ser petita i π no ha de ser un valor extrem.
b) …que la mostra no ha de ser ni molt gran ni molt petita.
c) …que si n és gran, π ha de ser proper a 0 o a 1.
d) No té sentit, perquè volem estimar π.

Respostes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	a	b	a	b	d	d	a	a	d	b	c	a	c	c	a

4.5 Problemes

1. Els usuaris d'una biblioteca porten anys protestant per les prestacions del sistema de recerca disponibles per fer les seves consultes. Els responsables de la biblioteca decideixen valorar la possibilitat de canviar el sistema. Durant el període de prova, han realitzat un experiment comparant ambdós sistemes mitjançant una escala que mesuri la satisfacció dels usuaris. Fan la prova anterior de comparació de mitjanes i en resumeixen els resultats amb la frase següent: *el nou sistema genera més satisfacció en els usuaris (P< 0.01).* Quines de les afirmacions següents són certes?:

   a) Es rebutja la hipòtesi que les mitjanes de l'escala de satisfacció són iguals en ambdós grups.

   b) Suposant que ambdós sistemes generin la mateixa satisfacció, la probabilitat d'haver obtingut un resultat tan extrem o més que el que s'ha observat és inferior a l' 1%.

   c) Creiem que el resultat observat reflecteix una diferència poblacional, en el conjunt de tots els casos possibles, d'ambdues mitjanes de satisfacció.

   d) La proporció de casos més satisfets amb el sistema antic que amb el nou és inferior a l'1%.

   e) Quan diem que el nou sistema és millor tenim una probabilitat d'equivocar-nos inferior al 0.01.

2. (continuació del problema 2 del tema anterior) Poseu a prova el requisit de la garantia: $E(Log_{10}(ITRS)) < 3$. (Pista: per demostrar-ho, s'ha de rebutjar el contrari.)

3. Una empresa financera ha d'introduir mensualment un gran volum de dades referents a factures i clients. Òbviament, aquests conjunts de dades han de reflectir amb fiabilitat la informació original. Les bases de dades les van introduir dues empreses, per poder confrontar les concordances. S'hi van detectar un total de 128 discordances (diferents dades introduïdes per cada empresa), de les quals es va comprovar que 70 eren errors de l'empresa A i 58, de l'empresa B. Partint de la hipòtesi que totes dues empreses tenen la mateixa qualitat, la probabilitat que un error vingui de l'empresa A seria de 0.5.

a) Poseu a prova la hipòtesi que aquesta probabilitat és 0.5.

b) Si es volguessin trobar diferències en el cas que una empresa fes el doble d'errors que l'altra, la prova a contrastar seria [H_0: π=0.5 versus H_1: π=0.667], quant val el risc de segona espècie?

c) Si les discordances fossin 200, canviaria algun dels riscos de la prova? Com es modifiquen i per què?

4. Per planificar els experiments destinats a estudiar el voltatge dels transformadors per a ordinadors portàtils que produeix una empresa, es determina, en primer lloc, el nombre d'observacions que cal prendre. El criteri és detectar una partida de transformadors amb una tensió mitjana perillosament diferent de l'estàndard de 20 V.

a) Quantes observacions es necessiten per distingir una desviació de la mitjana igual o superior a una desviació tipus ($1\cdot\sigma$), considerant uns riscos α=0.05 i β=0.20?

b) Com que es pensa que l'error d'admetre un transformador defectuós com a bo és prou greu, es disminueix el risc β fins a 0.05. Quina grandària mínima és recomanable ara?

c) El departament tècnic demana més sensibilitat, per detectar efectes més petits que també podrien malmetre els equips. Suggereix una desviació $0.5\cdot\sigma$. Quina grandària necessita en aquestes condicions?

Podeu reproduir el problema amb l'eina que trobareu a la web:
<http://wise.cgu.edu/power/power_applet.html>

4.6 Solució dels problemes

1. Les tres primeres són correctes, la quarta és una autèntica ximpleria, que no té res a veure, i la cinquena és un error molt habitual d'interpretació del risc α i del *p-valor*. Ambdós mesuren la probabilitat d'uns resultats condicionada a una certa hipòtesi, no la probabilitat que sigui certa una hipòtesi condicionada a uns resultats.

2. Sigui $\text{Log}_{10}(\text{ITRS}) = Y$ (treballem amb logaritmes, ja que sense transformar no es pot assumir normalitat)

$H_0 : \mu \geq 3$, o bé $H_0 : \mu = 3$
$H_1 : \mu < 3$

Estadístic $T = (\bar{y} - \mu_0) / (s/\sqrt{n}) \rightarrow t_{n-1}$
Distribució sota H_0 : $T \rightarrow t_{n-1}$
Premissa: Y normal
Regla de decisió: rebutjar H_0 (α=0.05) si $t < t_{3,\,0.05} = -2.353$
Càlculs: $\bar{y} = \Sigma_{i=1,\ldots 4} y_i / n = (2.996 + 2.505 + 2.699 + 2.602)/4 = 2.700$
$s^2 = [(2.996 - 2.700)^2 + \ldots]/(4-1) \cong 0.045 \cong 0.212^2$
$t = (2.700 - 3) / (0.212/\sqrt{4}) \cong -2.825$

Decisió: Atès que $-2.825 < -2.353$, es rebutja que $\mu = 3$ amb risc $\alpha = 0.05$.
Conclusió pràctica: No hi ha res que s'oposi al requisit de $\mu \leq 3$.

Ja havíem trobat l'IC del 90% de confiança al tema anterior: [2.451, 2.950]. Com que el contrast és unilateral i el risc α del 5%, l'interval adequat és aquest, que deixa un 5% a cada banda, o bé un IC del 95% però sense fita inferior: $\mu < 2.950$.

3.

a) H_0: $\pi=0.5$
H_1: $\pi\neq0.5$
$Z = (P-\pi)/\sqrt{(\pi(1-\pi)/n)} \sim N(0,1)$
Premisses: MAS i mostra gran: $\pi n > 5$ i $(1-\pi)n > 5$
$Z = (58/128-0.5)/\sqrt{(0.5\cdot0.5/128)} \approx -1.0607 > -1.96 = z_{0.025}$
Res no s'oposa a acceptar que ambdues cometin la mateixa proporció d'errors.
[Nota. També per IC(95%):
$0.546 \pm 1.96\sqrt{(0.546\cdot0.454/128)} \approx 0.546 \pm 0.086 \approx [0.450, 0.632]$
Com que inclou el valor de H_0, res no s'oposa a acceptar $\pi=0.5$]

b) H_0: $\pi_0=0.5$
H_1: $\pi_1=0.667$
β = Prob (regió acceptació | certa H_1)
Límit de la regió crítica $= \pi_0 + z_{0.95}\sqrt{(\pi_0\cdot(1-\pi_0)/n)} =$
$= 0.5 + 1.645\sqrt{(0.5\cdot0.5)/128} \approx 0.5 + 0.0727 = 0.5727$
$\beta = P(P < 0.5727 \mid H_1) = P[Z < (0.5727 - \pi_1) / \sqrt{(\pi_1 \cdot (1-\pi_1)/n)}] =$
$= P[Z < (0.5727 - 0.667)/\sqrt{(0.667\cdot0.333/128)}] =$
$\approx P[Z < -2.2638] = 0.0118$

c) H_0: $\pi_0=0.5$
H_1: $\pi_1=0.667$
Límit de la regió crítica $= \pi_0 + z_{0.95}\sqrt{(\pi_0\cdot(1-\pi_0)/200)} =$
$= 0.5 + 1.645\sqrt{(0.5\cdot0.5)/200} \approx 0.5 + 0.0685 = 0.5685$
$\beta = P(P < 0.5727 \mid H_1) = P[Z < (0.5685 - \pi_1) / \sqrt{(\pi_1 \cdot (1-\pi_1)/n)}] =$
$= P[Z < (0.5685 - 0.667)/\sqrt{(0.667\cdot0.333/200)}] =$
$\approx P[Z < -2.9557] = 0.0016$

El risc α resta invariable, ja que depèn del plantejament inicial, fix, del problema. En canvi, el risc β ha disminuït per la menor variabilitat de l'estimació, originada per la major grandària mostral.

4.

a) Anomenem X la variable aleatòria "tensió de sortida dels transformador mesurat", expressada en volts. El contrast que es planteja és:

$H_0 : \mu = \mu_0 = 20$
$H_1 : |\mu - \mu_0| > \sigma$, que vol dir: $\mu > 20 + \sigma$, o bé $\mu < 20 - \sigma$

perquè es vol detectar tant un excés de tensió com un defecte. Sota la hipòtesi nul·la, i suposant que σ és coneguda, la mostra proporciona un estadístic:

$$Z = \frac{(\bar{x} - \mu_0)}{\sqrt{\sigma^2/n}} \sim \mathrm{N}(0,1)$$

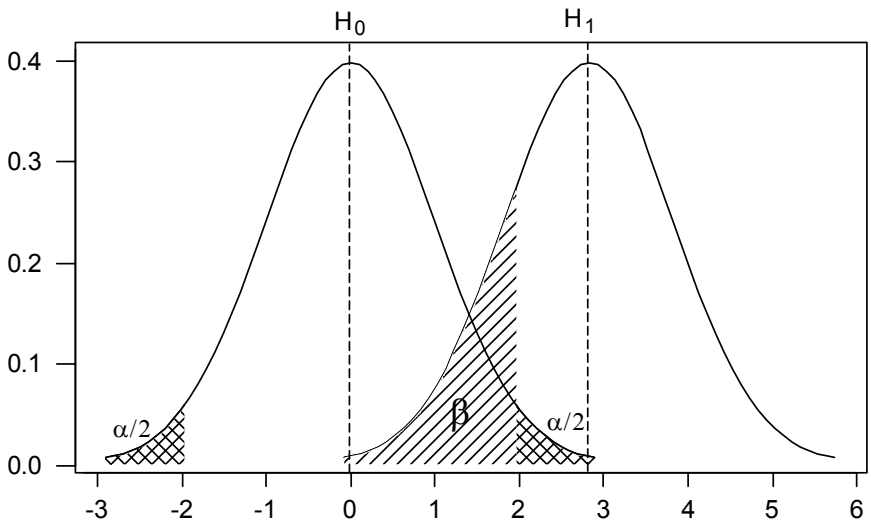
Per tant, si Z fos superior a $z_{1-\alpha/2}$ o inferior a $z_{\alpha/2}$, estaríem rebutjant la hipòtesi nul·la. En canvi, si fos a l'interior de l'interval central (entre –1.96 i 1.96, en el cas que ens ocupa), no rebutjaríem que la tensió mitjana fos de 20 V, i això passa algunes vegades amb partides de transformadors defectuosos. El que volem és trobar una grandària suficient per evitar que això succeeixi excepte en una proporció no molt gran i assumible (β).

Observeu que si X tingués una mitjana almenys igual a $\mu_0 + \sigma$ (sota H_1), l'estandardització que dóna lloc a una N(0,1), si fos certa la hipòtesi nul·la, dóna en aquest cas, una variable amb mitjana $\sqrt{n}$:

$$Z_1 = \frac{(\bar{x}_1 - \mu_0)}{\sqrt{\sigma^2/n}} = \frac{(\bar{x}_0 + \sigma - \mu_0)}{\sqrt{\sigma^2/n}},$$

$$\mathrm{E}(Z_1) = \frac{\sigma}{\sqrt{\sigma^2/n}} = \sqrt{n}$$

Gràficament, es pot veure la disposició dels riscos i de les dues distribucions analitzades, que tenen desviació tipus igual a 1 per a qualsevol n (si la mitjana real estigués per sota de μ_0, el plantejament seria molt semblant):



La de l'esquerra és N(0,1), la de la dreta és N($\sqrt{n}$, 1). Per tant, atès que z_β ($z_{0.20}$) val –0.8416, el punt crític de la dreta es pot escriure com:

$$z_{0.975} = 1.96 = \sqrt{n} + z_\beta\, \sigma_z = \sqrt{n} - 0.8416 \cdot 1 \rightarrow \sqrt{n} = 1.96 + 0.8416 = 2.8016,\ n \approx 7.85$$

Com que hem de prendre un valor enter, la grandària buscada és 8.

b) Si el risc de segona espècie baixa, és perquè volem més potència per tal de detectar millor els transformadors problemàtics. Visualment, el que es vol és allunyar la distribució sota H_1 perquè l'àrea β es faci més petita. Això comporta necessàriament una grandària més gran (ja que la mitjana $\sqrt{n}$ augmenta).

Analíticament, n'hi ha prou a refer el càlcul anterior, però ara z_β ($z_{0.05}$) val –1.645:

$$z_{0.975} = 1.96 = \sqrt{n} + z_\beta\,\sigma_z = \sqrt{n} - 1.645\cdot 1 \rightarrow \sqrt{n} = 1.96 + 1.645 = 3.605, \quad n \approx 12.995$$

El resultat és 13 observacions.

c) En general, si es vol detectar un efecte a la mitjana almenys igual a Δ, hem de veure que l'estadístic es desplaçarà una certa quantitat del centre 0, que correspon a una hipòtesi nul·la estrictament certa. Si l'efecte fos exactament Δ volts en excés, la mitjana de l'estadístic seria:

$$\mathrm{E}(\mathrm{Z}_1) = \frac{\Delta}{\sqrt{\sigma^2/n}} = \sqrt{n}\,\Delta/\sigma$$

I, per tant, la determinació de la grandària òptima s'obté de:

$$z_{1-\alpha/2} = \sqrt{n}\,\Delta/\sigma + z_\beta \;\rightarrow\; \sqrt{n}\,\Delta/\sigma = z_{1-\alpha/2} - z_\beta \;\rightarrow\; \sqrt{n} = (z_{1-\alpha/2} - z_\beta)\,\sigma/\Delta \rightarrow$$

$$n = \left\lceil \frac{\sigma^2}{\Delta^2}\left(z_{1-\alpha/2} - z_\beta\right)^2 \right\rceil$$

En el nostre cas:

$$n = \left\lceil \frac{1}{0.5^2}\left(1.96 + 1.645\right)^2 \right\rceil = 52$$

5 Comparació de dues poblacions normals

Als capítols anteriors, hem vist com inferir els resultats des d'una mostra a una població. Hem après a realitzar intervals de confiança i a contrastar dues hipòtesis sobre un paràmetre, com la mitjana μ. Però només hem exposat l'estudi d'una mostra, que limita les possibilitats de comparar, per exemple, resultats entre dos procediments amb els respectius paràmetres desconeguts.

A continuació, es presenten els fonaments que permeten la inferència estadística orientada a comparar els paràmetres de dues distribucions normals, és a dir, mitjanes i variàncies. Veurem que l'esquema bàsic per a la construcció d'intervals de confiança o per a la realització de proves d'hipòtesis continua essent vàlid, amb les particularitats de l'estadístic apropiat per a cada cas.

5.1 Prova de $\mu_1=\mu_2$. Mostres independents

Assumim, per començar, que es coneixen les variàncies poblacionals i que les distribucions de les dues variables que estem comparant són normals. Suposem, també, que els processos de mostreig d'ambdues poblacions són aleatoris i independents:

$$Y_1 \sim \mathrm{N}(\mu_1, {\sigma_1}^2)$$
$$Y_2 \sim \mathrm{N}(\mu_2, {\sigma_2}^2)$$

i n'obtenim dos MAS independents, de grandàries respectives n_1 i n_2:

$$\left(y_{11}, y_{12}, ..., y_{1n_1}\right) \text{ i } \left(y_{21}, y_{22}, ..., y_{2n_2}\right)$$

Per estudiar la diferència entre μ_1 i μ_2 es recorre a $\overline{Y}_2 - \overline{Y}_1$, la diferència de les mitjanes mostrals. Aquest estadístic segueix una distribució *normal*, ja que és una combinació de dues v.a. normals independents, amb *esperança* la diferència de les esperances o mitjanes poblacionals de Y_1 i Y_2. Per les propietats de l'operador *esperança*:

$$\mathrm{E}\left(\overline{Y}_2 - \overline{Y}_1\right) = \mathrm{E}\left(\overline{Y}_2\right) - \mathrm{E}\left(\overline{Y}_1\right) = \mu_2 - \mu_1$$

La *variància* de la diferència de les mitjanes mostrals es converteix, en aquesta situació de mostres independents, en la suma de les variàncies de les mitjanes:

$$\begin{aligned} \mathrm{V}\left(\overline{Y}_2 - \overline{Y}_1\right) &= \mathrm{V}\left(\overline{Y}_1\right) + \mathrm{V}\left(\overline{Y}_2\right) - 2\mathrm{Cov}\left(\overline{Y}_2, \overline{Y}_1\right) = \\ &= \mathrm{V}\left(\overline{Y}_1\right) + \mathrm{V}\left(\overline{Y}_2\right) = \\ &= \sigma_1^2 / n_1 + \sigma_2^2 / n_2 \end{aligned}$$

perquè la independència de les mostres implica la independència de les mitjanes mostrals i, per tant, aquestes tenen covariància zero. En conseqüència, coneixem la distribució de la diferència de les mitjanes mostrals:

$$\overline{Y}_2 - \overline{Y} \sim \mathrm{N}\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Exemple. Al Departament de Sistemes hi ha controvèrsia amb relació a quina versió de Linux (*A* o *B*) és més eficient per a un servidor web. El responsable del departament encarrega que es recullin dades en dos servidors bessons, però sense relació entre si, del temps que triga a carregar-se una pàgina bastant complexa (v. cas del tema anterior). Tant a la versió *A* com a la versió *B* s'obtenen 30 valors independents entre si. Suposem que els temps són normals en els dos casos i que la desviació tipus del temps val 80 cs per a les dues versions. A la figura 5.1 se n'observa el resultat.

La mitjana per a la versió *A* ha estat de 159.4 cs i per a la versió *B*, de 189.8 cs. Quina informació aporten aquestes dades? Es pot concloure que *A* és superior a *B*? Fins a quin punt podríem observar un resultat absolutament oposat?

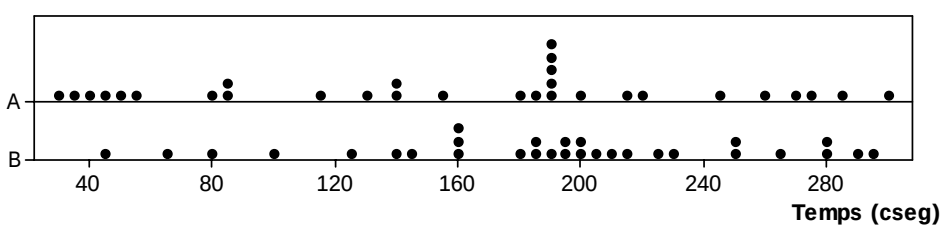


Figura 5.1 Dades de les dues mostres independents per a la versió A i per a la versió B

La diferència observada a les dades, $\overline{Y}_B - \overline{Y}_A$ =189.8–159.4 = 30.4 cs s'ha de valorar amb relació a l'error tipus que presenta la diferència de mitjanes de mostres de grandària 30, que obtenim a partir de:

$$\mathrm{V}\left(\overline{Y}_B - \overline{Y}_A\right) = \sigma_A^2 / n_A + \sigma_B^2 / n_B = 80^2 / 30 + 80^2 / 30 = 426.67$$

és a dir, l'error tipus és l'arrel quadrada de 426.67, o sigui, 20.66 cs. Quan μ_1 i μ_2 siguin iguals, una diferència de mitjanes al voltant de 20 cs serà típica, i sembla que 30 cs no és una diferència que destaqui.

5.1.1 Variàncies conegudes

A partir de la diferència de mitjanes mostrals, es pot arribar a construir un estadístic de referència per a la prova amb distribució coneguda:

$$Z = \frac{(\overline{Y}_2 - \overline{Y}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

Coneguda la distribució de $\overline{Y}_2 - \overline{Y}_1$, ja es pot estimar la diferència existent entre les dues mitjanes fent un interval de confiança:

$$\mathrm{IC}(\mu_2-\mu_1,\, 1-\alpha): \quad (\mu_2-\mu_1) \in \bar{y}_2 - \bar{y}_1 \pm z_{\alpha/2} \cdot \sqrt{(\sigma_1^2/n_1 + \sigma_1^2/n_2)}$$

O bé, mitjançant un contrast d'hipòtesis:

$$\begin{cases} \mathrm{H}_0 : \mu_1 = \mu_2 \\ \mathrm{H}_1 : \mu_1 \neq \mu_2 \end{cases}$$

Sota H_0:

$$Z = \frac{(\bar{y}_2 - \bar{y}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{y}_2 - \bar{y}_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

Exemple (continuació). En el nostre cas, l'estadístic de referència val:

$$z = 30.4/20.66 = 1.47$$
$$\text{p-valor} = \mathrm{P}(|Z| > |z|) = 2 \cdot \mathrm{P}(Z < -1.47) = 0.141$$

que no es pot considerar atípic en una distribució N(0,1). El p-valor trobat és massa gran per poder argumentar que la diferència observada sigui significativa. Podria ser una conseqüència de l'atzar. L'interval de confiança s'interpretaria en el mateix sentit: com que el zero pertany a l'interval, la diferència real entre les mitjanes tant podria ser a favor de *A* com de *B*:

$$\mathrm{IC}(\mu_B-\mu_A,\, 95\%) = \bar{y}_B - \bar{y}_A \pm z_{0.025} \cdot \sqrt{(\sigma_A^2/n_A + \sigma_B^2/n_B)} = 30.4 \pm 1.96 \cdot 20.66 = (-10.1,\, 70.9)$$

Nota. Observeu que, si les variables mesurades no són normals però la grandària de les mostres és prou gran perquè es pugui aplicar el TCL a cada mostra, es pot seguir utilitzant l'estadístic anterior. Els requisits demanats són: la normalitat de les mitjanes mostrals, conèixer les variàncies de cada població i la independència estadística, que és una qüestió de disseny de recollida de les dades, i no de com es distribueixen les variables implicades.

5.1.2 Variàncies desconegudes però idèntiques

Ara bé, usualment σ_1 i σ_2 són desconegudes i, per tant, s'han d'estimar. Així, si podem assumir igualtat de variàncies (aquest concepte s'anomena *homoscedasticitat*: $\sigma_1^2 = \sigma_2^2 = \sigma$), llavors $s_1^2 = s_2^2$ són dos estimadors del mateix paràmetre σ^2, i l'estimació més eficient de σ^2 s'obté donant més importància a la mostra més gran:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

Observeu que la ponderació que utilitzem és la grandària de la mostra restant 1, per la qüestió del grau de llibertat que es perd en l'estimació de la variància. Observeu igualment que en l'estimació de la variància comuna es calculen quadrats de diferències respecte de la mitjana, però que cada grup utilitza la seva mitjana.

Exercici. Simplifiqueu aquest estadístic per al cas particular en què ambdues mostres siguin de la mateixa mida. (Solució: l'estimació de la variància comuna coincideix amb la mitjana aritmètica de s_1^2 i s_2^2.)

Ara, en substituir σ per la seva estimació s, obtenim finalment:

$$T = \frac{(\bar{y}_2 - \bar{y}_1)}{s\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

que també ens permet construir intervals de confiança per la diferència de mitjanes, o resoldre proves o contrastos.

La premissa que dues variàncies són iguals malgrat que no sabem quin valor tenen pot semblar estranya, però no hi ha motiu per desconfiar d'una suposició així. Hi poden haver raons per justificar que els dos procediments per comparar potser són diferents en la mitjana, però no en la dispersió. O, com veurem més endavant, ens podem basar en una prova formal que ens digui que no hi ha cap evidència en contra de la igualtat de variàncies.

Exercici. Comproveu que, si disposem de n unitats, el disseny de mostres independents més eficient consisteix a dividir-les en parts iguals (és a dir, heu de veure com repartiu els n individus per tal que l'error tipus de la diferència de mitjanes mostrals sigui el mínim).

Nota. Hi ha situacions en què, comparant dos procediments, els individus experimenten un efecte variable que depèn de la seva posició en la distribució del procediment de referència. Per exemple, una eina d'autoaprenentatge que oferim als estudiants no afectarà igual l'estudiant molt bo (que difícilment pot pujar sensiblement la nota), l'estudiant gens motivat (que no canviarà l'actitud) o l'estudiant mitjà, possiblement el màxim beneficiari i l'objecte real de l'experiment. En canvi, en moltes altres situacions sí que es produeix el que anomenem *efecte fix*.

L'anomenem "fix" perquè no és una fluctuació que variï d'un individu a un altre (també es coneix com a *efecte additiu*). La idea d'estimar la diferència de mitjanes ve de la necessitat de saber quin canvi podria experimentar qualsevol individu si estigués en la situació oposada. Les proves de comparació de dues mitjanes, en el cas particular d'homoscedasticitat, són l'eina habitual per estimar un efecte fix.

S'entén fàcilment si pensem que, en les condicions descrites, la distribució de la v.a. Y_2 és la mateixa que la de la v.a. Y_1, però desplaçada una quantitat $\Delta = \mu_2 - \mu_1$. Això és aplicable també a qualsevol individu que hagi estat observat: si en lloc de pertànyer a la població 1 fos de la població 2, el seu valor observat s'hauria incrementat en la mateixa quantitat.

El plantejament de fons es pot representar com un model matemàtic:

$$Y_{mi} = \mu_m + \varepsilon_{mi}$$

on Y_{mi} és una observació per a l'individu i-èsim de la població m (m =1, 2); μ_m és la referència central per a la població m, i ε_{mi} és una pertorbació aleatòria normal, de mitjana 0 i desviació tipus σ. Aquest model no és l'únic possible; es podria formular d'altres maneres: per exemple, si Δ —l'efecte— val $\mu_2 - \mu_1$, un model alternatiu seria:

$$Y_{mi} = \mu + s_m \cdot \Delta + \varepsilon_{mi}$$

on μ és una referència central fixada en el valor $(\mu_1 + \mu_2)/2$, i:

$$s_m = \begin{cases} -1/2 & \text{si } m = 1 \\ 1/2 & \text{si } m = 2 \end{cases}$$

Exemple. A l'empresa J&J, els ordinadors personals dels usuaris tenen instal·lades dues marques diferents (diguem-ne *A* i *B*) i incompatibles de programes antivirus. Quan la direcció decideix uniformar el software, el cap d'informàtica ha d'escollir quina de les dues marques és millor. La variable primària *Y* que cal tenir en compte és el nombre de fitxers examinats per minut en un examen exhaustiu del disc dur. És vàlid admetre que *Y* segueix una distribució normal, amb la mateixa variància per a qualsevol dels dos antivirus.

Com que en el mateix ordinador no es poden instal·lar els dos antivirus, el cap vol comparar un grup d'ordinadors seleccionats a l'atzar dels que fan servir *A* i un altre grup per a *B* (també escollit a l'atzar i independent de l'anterior). Prèviament, s'ha comprovat que no hi ha diferències importants entre les màquines que utilitzen un programa o un altre en antiguitat, ús i potència de la màquina, factors que també afecten la distribució de *Y* i que, en cas de ser presents de forma desigual, introduirien una relació que es podria

atribuir erròniament al tipus d'antivirus instal·lat (aquestes variables se les coneix com a *variables confusores*, perquè confonen a qui dedueix que la resposta Y varia segons si es tracta de la població A o de B, i no aprecia que és un altre factor el causant dels canvis).

Les dades obtingudes amb deu ordinadors per a cada grup són:

	1	2	3	4	5	6	7	8	9	10
A	221	234	267	216	482	308	395	264	386	314
B	335	353	431	413	262	303	306	402	338	295

Recalquem que, com que són dues mostres independents entre si, la dada número 1 del grup A no té res a veure amb la dada número 1 de B, etc. Si els dos grups es desordenessin per separat, no es perdria cap informació.

Mostrem ara la resolució d'un contrast d'hipòtesis bilateral, perquè a priori no tenim cap informació que un antivirus pugui ser millor que l'altre.

Solució:

1. $\mathrm{H}_0 : \mu_A = \mu_B$
   $\mathrm{H}_1 : \mu_A \neq \mu_B$

2. Estadístic
$$t = \frac{(\bar{y}_A - \bar{y}_B)}{s\sqrt{1/n_A + 1/n_B}}$$

3. Distribució sota H_0: $t \sim \mathrm{t}_{n_A+n_B-2}$

   Premisses: Y_A i Y_B, normals; $\sigma_A^2 = \sigma_B^2$ MAS i independents

4. Regla de la decisió:

   rebutjar H_0 ($\alpha = 0.05$) si $t < t_{n_A+n_B-2,0.025} = -2.1009$
   o bé si $t > t_{n_A+n_B-2,0.975} = 2.1009$

5. Càlculs. Els estadístics per a cada grup són:

	Mitjana	Desviació tipus
A	308.7	87.6
B	343.8	55.9

   L'estimació de la variància comuna val:

$$s^2 = \frac{9 \cdot 87.6^2 + 9 \cdot 55.9^2}{18} = 73.5^2 \qquad t = \frac{308.7 - 343.8}{73.5\sqrt{1/10 + 1/10}} = -1.07$$

6. Decisió: atès que $|t| < 2.1009$
no hem trobat cap evidència contra $\mu_1=\mu_2$, amb risc $\alpha=0.05$

El p-valor del resultat $t = -1.07$ val gairebé 0.30. És evident que es tracta de diferències que podrien provenir de l'atzar del mostreig.

7.
$$\text{IC}(\mu_A - \mu_B, 95\%) = \bar{y}_A - \bar{y}_B \pm t_{n_1+n_2-2,0.975}\sqrt{s^2/n_1 + s^2/n_2}$$
$$= -35.1 \pm 2.1009 \cdot 32.9 = -35.1 \pm 69.06 = [\,-104.2\,,\,34.0\,]$$

8. Conclusió pràctica: no tenim prou informació per decidir si l'antivirus *A* processa, de mitjana, més fitxers per minut que l'antivirus *B*. Podríem esperar tant un efecte favorable a *A* (fins a 104 fitxers/min més) com a *B* (fins a 34 fitxers/min més).

Què ha de fer el cap? Decidir amb una moneda? Això seria una bona solució si haguéssim demostrat que ambdós programes són equivalents, però *no ho hem fet* (no hem pogut rebutjar que són equivalents). És bastant raonable creure que alguna diferència hi ha entre els dos. Podem pensar que potser la diferència real és irrellevant, però l'amplada de l'IC no permet donar aquesta suposició com un escenari possible. Més aviat hem d'interpretar el resultat com un dèficit d'informació: la població és molt heterogènia, la variància de *Y* deu ser bastant gran i la mostra ha resultat petita per detectar alguna diferència de pes. Sembla que l'única alternativa que s'ofereix al cap és repetir l'experiència, però amb molts més ordinadors (amb les molèsties que això comporta).

5.2 Prova de $\mu_1=\mu_2$. Mostres aparellades

Per comparar dos mètodes diferents, un disseny molt simple i molt intuïtiu és el disseny de dades aparellades. Consisteix a agafar dues dades de cada un dels *n* individus que componen la mostra:

$$\begin{matrix}(y_{11}, y_{21})\\ \vdots \\ (y_{1i}, y_{2i})\\ \vdots \\ (y_{1n}, y_{2n})\end{matrix}$$

i calcular l'estadístic *diferència* $d_i = y_{2i} - y_{1i}$, per $i=1$ a n. D'aquesta manera, obtenim una única mostra construïda a partir de les mesures originals. La clau del procediment de mostres aparellades és disposar de parells de mesures preses en condicions molt semblants. Això es pot aconseguir si les dues mesures es poden prendre al mateix temps, sense que una interfereixi en l'altra, o (si es prenen successivament) que la primera mesura no afecti l'estat de la unitat experimental. Per exemple, per comparar dos mètodes educatius no es pot emprar aquest disseny, perquè no es pot separar l'impacte de cada mètode per mesurar en el mateix alumne l'eficiència de cada un.

Exemples. 1: El degà d'una facultat vol saber si els alumnes d'una assignatura determinada rebrien els mateixos coneixements si l'ensenyament fos no presencial. Aquí tindríem un problema greu si volguéssim aplicar el disseny aparellat perquè, encara que els alumnes estiguessin disposats a repetir l'assignatura en format no presencial —cosa molt dubtosa—, seria complicat distingir entre els coneixements adquirits la primera vegada i els de la segona.

2: Es vol comparar el temps mitjà de compilació de programes escrits en Java per dos plataformes de desenvolupament. Es disposa de suficients programes per provar, de longitud i temàtica molt diversa. Un disseny de mostres independents assignaria a l'atzar uns programes a la primera plataforma i uns altres a la segona (si no són massa pocs, hem de confiar que l'atzar reparteix equitativament factors tan importants per al temps de compilació com la longitud del programa, però si no hem d'estar segurs, és millor distribuir-los utilitzant un criteri just i objectiu). Un disseny de mostres aparellades seleccionaria els programes apropiats, es compilarien a les dues plataformes i es prendria la diferència entre els dos temps obtinguts per cada programa.

Si el disseny aparellat és aplicable, tindrem la fortuna de poder aplicar un mètode molt senzill per a l'anàlisi estadística. En efecte, amb les dades obtingudes, podem trobar l'increment observat $d_i = y_{2i} - y_{1i}$, de manera que el problema original ha quedat reduït a un problema sobre una única mostra. Si es defineix la variable diferència $D = Y_2 - Y_1$, és obvi que

$$\mathrm{E}(D) = \mu_D = \mathrm{E}(Y_2 - Y_1) = \mu_2 - \mu_1$$

La prova quedaria transformada com es veu a continuació:

$$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \rightarrow \quad \mathrm{H}_0 : \mu_2 - \mu_1 = 0 \quad \rightarrow \quad \mathrm{H}_0 : \mu_D = 0$$

Per tant, un disseny de mostres aparellades es resol com un problema d'inferència amb un sol paràmetre. Hi podem aplicar les tècniques exposades al capítol anterior. Hem de recordar que la hipòtesi de normalitat s'aplica únicament a D, i no és necessari que Y_1 i Y_2 es distribueixin com una normal.

Exemple. El cap d'informàtica de J&J veu com a solució per decidir entre els dos antivirus implementar un disseny de dades aparellades sobre 10 ordinadors escollits a l'atzar. Amb molta cura, es desinstal·len els antivirus originals i s'instal·len de nou amb procediments estàndards (5 per a *A* i 5 per a *B*, independentment de quin programa hi havia inicialment). Es mesura el nombre de fitxers examinats per minut amb el primer antivirus, i es repeteix de nou després de treure'l i instal·lar el programa oposat, amb el qual s'obté la segona mesura. D'aquesta manera, s'assegura que la configuració de l'ordinador i el contingut del disc dur resten pràcticament iguals, a excepció dels fitxers propis dels antivirus, com és natural. Les noves dades són ara:

	1	2	3	4	5	6	7	8	9	10
	AB	*BA*	*AB*	*BA*	*AB*	*BA*	*AB*	*BA*	*AB*	*BA*
A	230	341	261	309	275	367	317	272	298	317
B	252	347	303	321	339	398	325	267	360	412
B–A	22	6	42	12	64	31	8	–5	62	95

La primera fila indica l'ordinador emprat; la segona és l'ordre en què s'han provat els programes (per comprovar després si el possible rastre que deixa l'antivirus, malgrat totes les precaucions de l'informàtic, pot influir en la resposta. No és aquest el cas, com es comprovarà més endavant). La tercera fila i la quarta són les dades obtingudes, i la darrera, la diferència entre *B* i *A*.

El nombre mitjà de fitxers examinats per minut que avantatja *B* a *A* és 33.7, i la desviació tipus de la diferència val 31.8, de manera que l'estadístic *t* és:

$$t = \frac{33.7 - 0}{31.8/\sqrt{10}} = 3.35$$

Un valor com aquest, en relació amb una distribució t de Student amb 9 graus de llibertat, té associat un p-valor (bilateral) igual a 0.008, i l'IC al 95% per la diferència mitjana val (11.0, 56.4) fitxers/minut: sembla que l'antivirus *B* té algun avantatge sobre *A*, i l'evidència trobada es pot qualificar com a forta.

Nota. El model matemàtic de les dades aparellades es podria escriure com:

$$Y_{mi} = \mu + \Delta_m + \zeta_i + \varepsilon_{mi} \qquad (m = 1, 2;\ i = 1, ..., n)$$

on ε i ζ són variables de mitjana zero, normals i independents, que representen pertorbacions a l'atzar; ε denota la variabilitat particular d'un individu (observat diverses vegades, dóna valors diferents), i ζ, en canvi, reflecteix la diversitat a la població (el fet de la resposta varia perquè observa individus diferents). La variància de ε $\left(\sigma_\varepsilon^2\right)$ s'anomena *intraindividual*, i la de ζ $\left(\sigma_\zeta^2\right)$ és la variància *interindividual*. La variància total d'una observació equival a $\sigma_Y^2 = \sigma_\varepsilon^2 + \sigma_\zeta^2$.

Podem comprovar que:

$$D_i = Y_{2i} - Y_{1i} = \Delta_2 - \Delta_1 + \zeta_i - \zeta_i + \varepsilon_{2i} - \varepsilon_{1i} = \mu_2 - \mu_1 + \varepsilon_{2i} - \varepsilon_{1i}$$

i que l'expressió anterior té variància:

$$\mathrm{V}(D_i) = \mathrm{V}(\varepsilon_{2i} - \varepsilon_{1i}) = 2\sigma_\varepsilon^2$$

Fixeu-vos que, en prendre la diferència de dues mesures en el mateix individu, es cancel·la la component ζ_i. Per tant, deixem de considerar molts factors que intervenen directament en la variabilitat de la resposta (ordinadors vells i nous, amb el disc molt ple o molt buit, i altres factors que representen diferències *entre* ells), però que no són rellevants per al nostre objectiu. D'altra banda, com que cada unitat s'observa dues vegades, la part de variabilitat intraindividual es duplica.

Perquè un disseny amb dades aparellades sigui interessant, comparat amb un disseny basat en mostres independents, s'ha de complir que V(D_i) sigui més petita que σ^2. Això implica que σ_ε^2 ha de ser menor que σ_ζ^2, i usualment ho és. No és una condició molt exigent: equival a dir que les variacions en un mateix individu han de ser més petites que les variacions entre individus diferents.

En el cas anterior, l'error tipus de la mitjana de la diferència val 31.8/√10 = 10.05. Si amb les mateixes dades haguéssim efectuat —indegudament— una anàlisi de mostres independents, l'error tipus seria $46.0\sqrt{1/10 + 1/10}$ =20.57, més del doble. Per tant, com que el numerador és el mateix, hauríem aconseguit un estadístic *t* menys allunyat de zero i, per tant, la prova aparellada és més sensible per trobar una diferència —és a dir, més potent. Si es comprova, l'anàlisi de mostres independents no podria rebutjar la igualtat de mitjanes.

El disseny de mostres aparellades presenta un lleuger desavantatge. Com que el nombre de graus de llibertat és menor, el punt crític per rebutjar la hipòtesi nul·la és més gran (en aquest cas, $t_{9,\,0.975}$ = 2.26 enfront de $t_{18,\,0.975}$ = 2.10). Si les grandàries de les mostres són suficientment grans, aquest efecte és negligible.

5.3 Prova de $\mu_1=\mu_2$ amb efecte multiplicatiu i disseny aparellat

Hi ha una situació molt habitual en informàtica —i en altres àmbits— que no s'adapta a la premissa de l'efecte fix, tal com l'hem vist. Suposem que es vol determinar si dos procediments són equivalents en temps d'execució. Les tècniques d'anàlisi de la complexitat d'un algorisme poden establir una expressió per al cost mitjà, en temps o en espai consumit, en funció de la grandària de l'entrada que es processa, però normalment arriben a determinar l'ordre superior del cost (logarítmic, lineal, quadràtic, etc.). D'aquesta manera, es conclouria que dos algorismes són O(n^2), sense precisar coeficients ni termes d'ordre inferior, que poden, malgrat tot, marcar una diferència substancial en el rendiment de la tasca.

El problema és que la variabilitat de la resposta (usualment, el temps) depèn profundament de la mida de l'*input*: entrades grans comporten temps grans, i també variacions més grans que les que vénen associades a les entrades de grandària menor. El canvi esperat en el temps per utilitzar el programa 1 en lloc del 2 per processar un vector petit pot ser unes centèsimes de segon. Però, si el vector és cent vegades més gran, el canvi observat hauria de ser de l'ordre de segons, almenys (això, per a algoritmes de cost lineal; si fos un cost superior podríem parlar de minuts o hores).

Per tant, quan veiem que el canvi de població no es pot descriure adequadament amb un efecte constant que se suma a la tendència comuna (i a les pertorbacions de l'atzar), sinó que ve donat per un efecte *proporcional* a la magnitud de la resposta, parlarem d'un efecte multiplicatiu. Vegem amb un exemple com podem estudiar aquest tipus d'efecte.

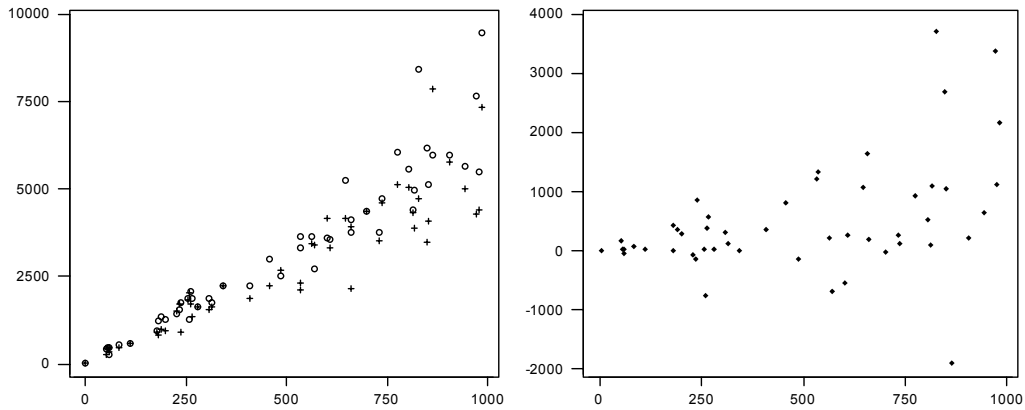


Figura 5.2 A l'esquerra: temps d'execució per a dos programes (cercles i creus) que processen aparelladament entrades de la mida representada a l'eix d'abscisses. Observeu que les diferències entre els programes (gràfic de la dreta) augmenten amb la mida de l'entrada.

Exemple. Existeixen molts algoritmes per ordenar vectors. Volem comparar dues implementacions de dos d'ells, Quicksort (*Q*) i Mergesort (*M*). Com que no té molt sentit establir la diferència per a una grandària de vectors donada, es dissenya un experiment en el qual els dos programes ordenaran 50 vectors donats, de diverses mides, fins a un milió d'elements. Es mesura el temps d'execució en mil·lisegons (v. figura 5.2). Els estadístics de cada programa són:

(ms)	Mitjana	Desviació tipus
Q	3230	2267
M	2745	1872
Q–M	485	961

Si bé s'observa un temps millor per al Mergesort, no es veu clarament quin pot ser el millor, però sí que resulta obvi que establir un efecte additiu seria un error, ja que l'efecte real ha de ser petit per a vectors petits i gran per a vectors grans (amb tots els graus intermedis). Per simplicitat, es vol enfocar l'efecte en un sol paràmetre senzill d'interpretar, i no en una funció d'un factor aliè a la comparació dels algoritmes, com és la mida del vector.

Ja sabem, per la forma com s'han recollit les dades, que no podem treballar com si fossin mostres independents. Si ho fossin, aquestes dades no aportarien cap evidència contra la hipòtesi d'igualtat dels dos programes, amb p-valor=0.246. Un altre error seria obtenir les diferències entre *Q* i *M*, i estimar-ne la diferència mitjana. Ara sí que hi trobem diferència significativa (p=0.001), i un IC que no conté el zero: (212, 758), però l'IC no té cap significat, ja que no seria aplicable ni per a vectors petits ni per als grans.

Hi ha dues aproximacions. La primera consisteix a estudiar la ràtio entre les dues variables:

$$W = Y_Q/Y_M$$

Aquesta alternativa té l'inconvenient que la distribució de W pot ser molt estranya i afectar les premisses per inferir correctament. La segona opció consisteix a treballar amb una transformació no lineal de les dades, que respecta molt millor la premissa de normalitat de la variable resposta:

$$V = \log(Y_Q/Y_M) = \log(Y_Q) - \log(Y_M)$$

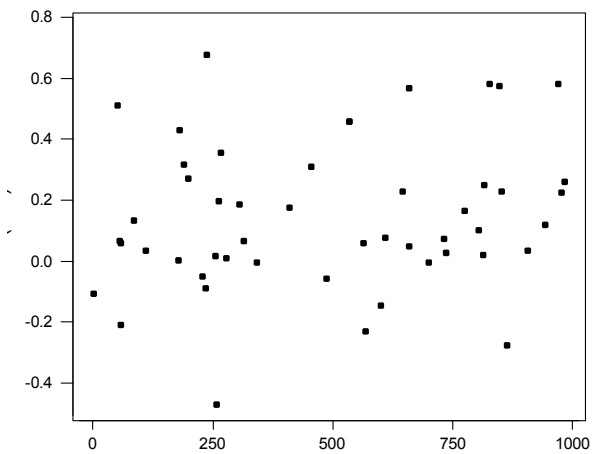


Figura 5.3 Valors de $\log(Y_Q) - \log(Y_M)$ *en funció de la mida del vector*

Resulta que aquesta via ens condueix a estudiar una diferència, com el mètode de l'apartat anterior. Per tant, estimarem el valor mitjà de la diferència dels logaritmes dels temps obtinguts, i ara sí que podem veure que es tracta d'un efecte constant, independent de la mida del vector (a la figura 5.3, els valors observats de V es distribueixen amb la mateixa dispersió, sigui quina sigui la mida del vector).

Després de fer les transformacions, trobem que la mitjana mostral de V val 0.14566 i que la desviació tipus mostral és 0.24401. Amb això podem trobar que l'estadístic t de la prova de significació val 4.22. Aquest valor ens permet rebutjar amb fermesa que:

$$\mathrm{E}(V) = 0 \Leftrightarrow \mathrm{E}(\log(Y_Q)) = \mathrm{E}(\log(Y_M))$$

Teòricament, d'aquí no podem afirmar que l'esperança de Y_Q és diferent de l'esperança de Y_M (per la transformació logarítmica que hi hem introduït) però, si creiem que les dues variables no segueixen cap distribució estranya, no hi ha cap problema a estendre la conclusió a rebutjar la igualtat de mitjanes de Y_Q i Y_M. Ja hauríem resolt que els dos programes no són igual d'eficients a l'hora d'ordenar un vector.

D'altra banda, hi ha el problema d'estimar el factor diferencial entre ells (ja que no és apropiat parlar de *diferència*). Utilitzarem l'interval de confiança per a la mitjana de V (el valor 2.0096 prové de la distribució t de Student amb 49 graus de llibertat, per al nivell del 97.5%):

$$IC(\mu_V, 95\%) = 0.14566 \pm 2.0096 \cdot 0.24401/\sqrt{50} = (0.07632, 0.21501)$$

Si prenem l'exponencial de l'interval, podem obtenir una bona aproximació a l'IC de $E[Y_Q/Y_M]$: (1.0793, 1.2399). Per tant, la conclusió és que el procediment Q incrementa, de mitjana, entre un 8% y un 24% el temps que s'aconseguiria amb el procediment M. Parlem d'un efecte multiplicatiu, ja que no es tracta d'un increment absolut sinó d'un increment relatiu.

5.4 Prova de $\sigma_1^2 = \sigma_2^2$ en mostres independents

S'ha vist la comparació de les mitjanes de dues variables amb distribució normal. Ara es veurà la comparació de les seves variàncies, per remarcar que el fet que les variàncies siguin diferents és també una informació rellevant sobre l'efecte en estudi (que ja no pot ser fix). Per fer aquesta prova, primer és necessari introduir una nova distribució de probabilitat.

5.4.1 La distribució F de Fisher-Snedecor

Siguin X_1 i X_2 dues variables aleatòries independents amb distribució de khi quadrat, amb n i m graus de llibertat, respectivament:

$$X_1 \sim \chi^2_n$$
$$X_2 \sim \chi^2_m$$

Llavors, el quocient d'ambdues, dividides prèviament pels seus graus de llibertat, segueix una distribució F, anomenada de Fisher —o bé de Fisher-Snedecor—, amb n i m graus de llibertat.

$$Y = (X_1 / n) \ / (X_2 / m) \sim F_{n,m}$$

El valor esperat de Y sempre és 1, siguin quins siguin n i m. S'ha de tenir present que la inversa d'una variable amb distribució F, amb n i m graus de llibertat, és també una F amb m i n graus de llibertat:

$$W = 1 / Y \sim F_{m,n}$$

Cal tenir present, també, que el quadrat d'una variable amb distribució t de Student de n graus de llibertat és una variable que segueix una distribució F, amb 1 i n graus de llibertat:

$$\text{Si} \qquad t \sim t_n \qquad \Rightarrow \qquad t^2 \sim F_{1,n}$$

El fet que aquesta distribució tingui dos paràmetres fa que les taules impreses hagin de ser més extenses. Llavors, les taules es concentren en els riscos α més usuals (0.01, 0.025, 0.05...).

Sigui $\tilde{\tilde{F}} \sim F_{1,\infty}$ $P(\tilde{\tilde{F}} > 3.84) = 0.05$
Sigui $\hat{F} \sim F_{10,5}$ $P(\hat{F} > 10.10) = 0.01$
Sigui $F \sim F_{10,5}$ $P(F > 3.297) = 0.10$
Sigui $G \sim F_{5,10}$ $P(G > 0.3033) = P(G > 1/3.297) = P(1/F > 1/3.297)$
$= P(F < 3.297) = 0.90$

Exercici. Repetiu aquests càlculs amb Minitab o R.

5.4.2 Comparació de variàncies de dues poblacions normals

Ara ja es poden comparar dues variàncies. Siguin dues mostres aleatòries simples i independents, de mides n_1 i n_2, de dues variables amb distribució normal:

$$Y_1 \sim N(\mu_1, \sigma_1^2)$$
$$Y_2 \sim N(\mu_2, \sigma_2^2)$$

Sigui també la prova d'hipòtesi de la igualtat de les seves variàncies.

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Llavors, sota H_0, el quocient entre les estimacions mostrals de les seves variàncies segueix una distribució F, amb n_1-1 i n_2-1 graus de llibertat:

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$

Observeu que aquest estadístic significa: "Si es prenen dues mostres de la mateixa població, les seves variàncies mostrals respectives poden ser diferents —i sempre ho són—, però la seva ràtio fluctua dins el rang d'una distribució de probabilitat coneguda que (*a*) té per valor esperat 1 i en què (*b*) com més grans són els graus de llibertat —és a dir, la grandària de les mostres— més a prop del valor 1 es troba la major part de la distribució." Encara més senzillament, podríem dir: "Si la població és única, les variàncies mostrals s'han d'assemblar, i si les mostres són grans s'assemblaran més encara."

A partir d'ara, podem delimitar la regió crítica, rebutjant H_0 a partir dels límits $F_{n_1-1, n_2-1, \alpha/2}$ i $F_{n_1-1, n_2-1, 1-\alpha/2}$. És a dir, si fixem α=0.05, es rebutjarà H_0 sempre que $F < F_{n_1-1, n_2-1, 0.025}$ o bé que $F < F_{n_1-1, n_2-1, 0.975}$.

Ja s'ha vist que l'invers d'una variable amb distribució F segueix també una F (si $F \sim F_{n_1-1, n_2-1}$, llavors $1/F \sim F_{n_2-1, n_1-1}$). Per això, es pot obviar a les taules el límit inferior posant sempre al numerador de l'estadístic F la variància mostral de mida més gran:

> Siguin s_A^2 la variància major i s_B^2 la menor
> Farem servir $F = s_A^2 / s_B^2 \sim F_{n_A-1, n_B-1}$ i es
> rebutjarà H_0 sempre que $F > F_{n_A-1, n_B-1, 1-\alpha/2}$

Cal tenir en compte que el caràcter bilateral del plantejament de la hipòtesi no ha canviat. Senzillament, hem simplificat el procediment a l'hora de buscar els límits a les taules concentrant-nos en la meitat superior. Per això, encara que només mirem el límit superior, el valor de les taules és el corresponent a $\alpha/2$ (per exemple, 0.025) per a una prova amb risc α (0.05, per al mateix exemple).

Exemple. Es té interès a comparar la durada dels recanvis dels cartutxos de tinta de dues marques: ORIGINAL, S.A. i MOI_AUSSI, S.L.. S'ha comprovat que la marca MOI_AUSSI té una durada mitjana més gran que la marca ORIGINAL. Però se sospita que la seva variabilitat pot ser diferent. En dues mostres, s'han obtingut els resultats següents:

	n	$\bar{y}_i$	s_j^2
ORIGINAL	8	263	64
MOI_AUSSI	6	407	144

Solució: $H_0 : \sigma^2_{OR} = \sigma^2_{M\_A}$
$H_1 : \sigma^2_{OR} \neq \sigma^2_{M\_A}$

Punt crític superior (α=0.05) per a una distribució amb 5 i 7 g.l.= 5.29
Premisses: normalitat i MAS independents
$F = s_{\text{major}}^2 / s_{\text{menor}}^2 = 144/64 = 2.25 < 5.29 = F_{5,7,0.975}$

La hipòtesi nul·la no pot ser rebutjada. No s'ha pogut determinar si la marca MOI_AUSSI fabrica cartutxos amb durada més variable que l'altra marca.

Si el plantejament de la hipòtesi fos unilateral, abans de recollir les dades ja es decidiria quina variància aniria al numerador (independentment que després resulti ser la menor) i després el risc α es deixaria només al límit superior. Continuant amb l'exemple, podríem afegir la informació següent. Se sap que els processos de fabricació dels productes anteriors són idèntics, amb l'excepció d'un pas de producció, que només té MOI_AUSSI i que, d'influir en la fabricació, només ho pot fer afegint-hi variabilitat, però sense eliminar-la. Aquesta situació canvia el punt crític, ja que ara tot el risc α s'acumula a la part de la dreta de la distribució:

$$H_0 : \sigma^2_{OR} = \sigma^2_{M\_A}$$
$$H_1 : \sigma^2_{OR} < \sigma^2_{M\_A}$$

$$F = s_{M\_A}^2 / s_{OR}^2 = 144/64 = 2.25 < 3.97 = F_{5,7,0.95}$$

Amb aquesta informació addicional, la hipòtesi nul·la tampoc no pot ser rebutjada.

Exercici. Si amb altres mostres trobéssim la mateixa ràtio de variàncies (2.25), quina seria la grandària mínima de les mostres que permetés rebutjar la hipòtesi nul·la en els dos casos anteriors? Fixeu-vos que el que es demana no es pot fer amb un càlcul directe, i que s'ha de buscar a les taules —o, millor, en un full de càlcul— una intersecció propera al valor de F.
(Solució. Bilateral: 20 per a la marca del numerador i 30 per la del denominador. Unilateral: 14 per a la mostra de la marca MOI_AUSSI i 21 per a la mostra d'ORIGINAL.)

Nota. L'ús d'intervals de confiança per a la comparació de variàncies s'aplica a la ràtio d'aquestes. Podeu comprovar que, per avaluar el rang de variació del quocient σ_1^2/σ_2^2 obtenim que:

$$\mathrm{IC}\left(\sigma_1^2/\sigma_2^2, 1-\alpha\right) = \left(\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}},\ \frac{s_1^2}{s_2^2} \cdot F_{n_1-1,n_2-1,1-\alpha/2} \right)$$

Fixeu-vos en el intercanvi de graus de llibertat per els factors que multipliquen el rati de variàncies mostrals: està fet per tal que sigui fàcil trobar aquests valors a les taules.

En el cas anterior, per un IC al 90%, tenim que:

$$F_{5,7,0.95} = 3.97$$
$$F_{7,5,0.95} = 4.88$$

$$\mathrm{IC}\left(\sigma_{M\_A}^2/\sigma_{OR}^2, 90\%\right) = \left(2.25\frac{1}{3.97},\ 2.25 \cdot 4.88 \right) = (0.57, 11.0)$$

Observem: 1) que l'interval conté el valor 1, que indicaria que no es pot rebutjar l'equivalència de variàncies, i 2) que la ràtio de desviacions tipus oscil·la entre 0.75 i 3.31.

5.5 Càlcul de la grandària mostral per comparar dues mitjanes

Suposem que, en la comparació de dues mitjanes, estem interessats a prendre una decisió entre dos valors concrets, per exemple 0 i Δ, que situem a les hipòtesis nul·la i alternativa:

$$\begin{cases} \mathrm{H}_0 : \mu_A - \mu_B = 0 \\ \mathrm{H}_1 : \mu_A - \mu_B = \Delta \end{cases}$$

Per exemple, podem imaginar que, interessats a avaluar el rendiment d'un nou programa *B* respecte a la versió clàssica *A*, Δ representa aquella millora en *B* que en fa rendible el desenvolupament i la substitució de *A*, mentre que el valor de la hipòtesi nul·la indica l'absoluta igualtat entre ambdós.

Suposeu també que es coneix, en les unitats experimentals on les volem comparar, el grau de dispersió (σ) existent entre els resultats de diferents execucions d'un mateix programa. Per simplicitat, considerem la situació (de màxima eficiència) en què es realitzen, exactament, el mateix nombre d'execucions amb el programa *A* que amb el programa *B*: $n_A = n_B = n$. En aquesta situació, la variància de la diferència de les mitjanes per a mostres independents és (assumint iguals les n i les σ de cada programa):

$$\mathrm{V}\left(\bar{Y}_A - \bar{Y}_B\right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}$$

La figura 5.4 representa la distribució dels valors observables de la diferència de mitjanes, tant sota la hipòtesi nul·la, corba centrada en 0, com sota l'alternativa, corba centrada en Δ. Si es defineix un risc de primera espècie α bilateral, el punt crític es troba a una distància $z_{1-\alpha/2}\cdot\sigma\sqrt{(2/n)}$ del punt 0 (el centre de la distribució de la esquerra) i a una distància $|z_\beta|\cdot\sigma\sqrt{(2/n)}$ del punt Δ. Com que usualment β és un valor <0.5, z_β és un valor negatiu, de manera que $|z_\beta| = -z_\beta$. La distància total que separa els centres de les dues distribucions es pot escriure com:

$$\Delta = z_{1-\alpha/2}\,\sigma\sqrt{\frac{2}{n}} + \left|z_\beta\right|\sigma\sqrt{\frac{2}{n}} = \left(z_{1-\alpha/2} - z_\beta\right)\sigma\sqrt{\frac{2}{n}}$$

Només resta calcular la grandària mostral *n* que necessitem per a cada mostra, aïllant la variable de l'expressió anterior:

$$n = \frac{2\,\sigma^2\left(z_{1-\alpha/2} - z_\beta\right)^2}{\Delta^2}$$

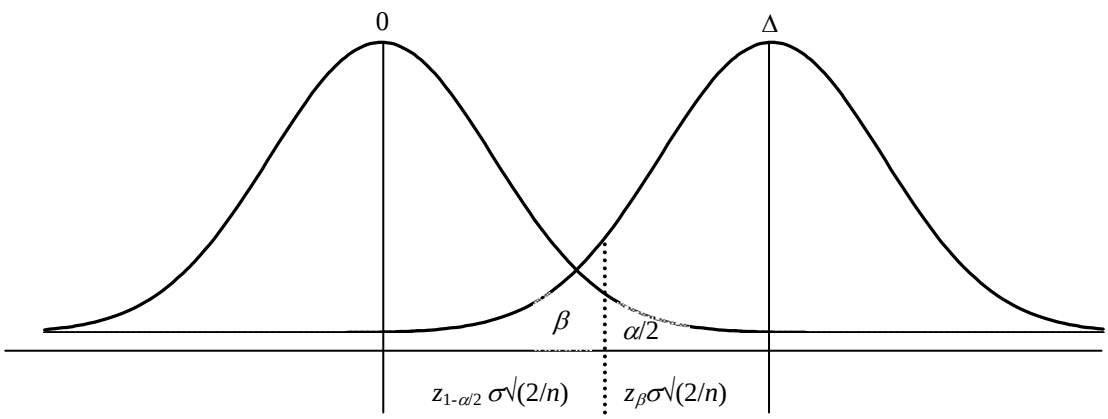


Figura 5.4. Distribució de $\overline{Y}_A - \overline{Y}_B$: a l'esquerra, sota la hipòtesi nul·la; a la dreta, sota la hipòtesi alternativa. La línia de punts marca el límit de la regió crítica per a un risc determinat.

D'aquí també es dedueix que, si ha de ser una prova unilateral, llavors la grandària és:

$$n = \frac{2\,\sigma^2\left(z_{1-\alpha} - z_\beta\right)^2}{\Delta^2}$$

que serà menor que el valor trobat amb la primera expressió, ja que $z_{1-\alpha}$ és més petit que $z_{1-\alpha/2}$.

Recordeu que la potència d'un estudi per establir una alternativa d'interès pren el valor complementari del risc β. I que moltes vegades s'estableixen les condicions de l'estudi dient, per exemple, que "es vol garantir una potència mínima del 80%...", que significa que β ha de valer 0.20.

Exemple. Quina grandària mostral seria necessària per detectar un increment en la nota final mitjana d'estudiants que fan ús d'una eina informàtica d'autoaprenentatge igual a 1 punt? Assumim que la desviació típica de la nota final en els estudiants és 1.75 punts, i prenem els riscos $\alpha = 0.05$, bilateral, i $\beta = 0.20$. També se suposa que les mostres són independents i que la nota final segueix una distribució normal.

$$n = [\, 2 \cdot 1.75^2 (1.96 - (-0.8416))^2] \,/\, 1^2 = 48.07$$

Es necessiten 49 casos per grup. Llavors, si realment l'eina mostra l'eficàcia considerada com a mínimament desitjable, amb un centenar d'estudiants podem posar en marxa un estudi que confirmaria empíricament la seva superioritat com a mètode docent en un 80% dels experiments.

5.6 Resum

La taula 5.1 mostra esquemàticament el procés de resolució dels diversos contrastos d'hipòtesis associats als paràmetres d'una distribució normal (μ i σ). Pel que fa a la variància, la hipòtesi de normalitat és fonamental, ja que si ens referíssim a una població no normal els riscos adoptats no estarien reflectint correctament els riscos reals i, per tant, augmentaria la probabilitat de prendre decisions errònies. Quant a la mitjana, i en el cas de tractar distribucions no normals, atès que són proves que es basen en el TCL, són molt més fiables mentre la grandària de les mostres sigui suficientment alta.

Taula 5.1 Proves de comparació de dos paràmetres més usuals

Paràmetre	Hipòtesi	Estadístic	Premisses	Distrib.(H_0)	Decisió α=0.05
μ	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$Z = \dfrac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$	$Y_1, Y_2 \sim$ N, MAS ind. i σ_1, σ_2 coneg.	$Z \sim N(0,1)$	Rebutjar si $\lvert Z \rvert > 1.96$
μ	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$T = \dfrac{(\bar{y}_1 - \bar{y}_2)}{s\sqrt{1/n_1 + 1/n_2}}$ $s^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$	$Y_1, Y_2 \sim$ N $\sigma_1 = \sigma_2$ desconeg. MAS indep.	$T \sim t_{n1+n2-2}$	Rebutjar si $\lvert T \rvert > t_{n-1,\,0.975}$
μ	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$T = \dfrac{\overline{D}}{\sqrt{s_D^2/n}}$	$D \sim$ N MA aparellada	$T \sim t_{n-1}$	Rebutjar si $\lvert T \rvert > t_{n-1,\,0.975}$
σ	$H_0 : \sigma^2_1 = \sigma^2_2$ $H_1 : \sigma^2_1 \neq \sigma^2_2$	$F = \dfrac{s_A^2}{s_B^2}$ amb $s_A{}^2 > s_B{}^2$	$Y_1, Y_2 \sim$ N MAS indep.	$F \sim F_{n_A-1, n_B-1}$	Rebutjar si $F > F_{n_A-1, n_B-1, 0.975}$
Les proves unilaterals corresponents es fan acumulant el risc α a un costat.					

5.7 Preguntes tancades de resposta única

1. En comparar el rendiment de dos sistemes, *A* i *B*, s'obté un interval del 95% de confiança de la diferència de les seves mitjanes, que va de L_1 a L_2 $\{IC_{95\%}(\mu_A - \mu_B)=(L_1, L_2)\}$. Creiem, amb una confiança del 95%, que:

a) L'autèntica diferència de mitjanes poblacionals es troba entre L_1 i L_2.
b) L'autèntica diferència de mitjanes poblacionals es troba entre L_1 i L_2, en el 95% de les mitjanes.
c) En el 95% de les execucions, la diferència de rendiment està entre L_1 i L_2.
d) Totes les respostes anteriors són correctes.

2. L'estadístic $$\hat{T} = \frac{(\bar{y}_1 - \bar{y}_2)}{s\sqrt{1/n_1 + 1/n_2}}$$

segueix una t de Student amb:

a) n_1+n_2-2 graus de llibertat
b) $n-1$ graus de llibertat
c) 30 graus de llibertat
d) Aquest estadístic no segueix una t de Student.

3. Perquè l'estadístic $$\hat{T} = \frac{(\bar{y}_1 - \bar{y}_2)}{s\sqrt{1/n_1 + 1/n_2}}$$

segueixi una t de Student amb n_1+n_2-2 graus de llibertat, les premisses necessàries són:

a) MAS i independents
b) Normalitat de la variable original
c) Homoscedasticitat o igualtat de variàncies
d) Totes les respostes anteriors són necessàries.

4. L'estimació mitjançant dades aparellades és més eficient que la de dades independents perquè:

a) Augmenta la diferència entre les mitjanes.
b) Disminueix la diferència entre les mitjanes.
c) Augmenta l'error tipus d'estimació.
d) Disminueix l'error tipus d'estimació.

5. La prova definitiva per saber si unes dades són aparellades és:

a) Que les mides mostrals siguin iguals.
b) Que les dades vinguin per parelles.
c) Que les parelles s'hagin recollit molt juntes.
d) Que hi hagi alguna similitud o correlació entre els valors de les parelles.

6. La prova de comparació de dues variàncies amb dades independents és:

   a) Unilateral i es mira unilateral a les taules.
   b) Bilateral i es mira bilateral a les taules.
   c) Si fos bilateral, es miraria unilateral a les taules invertint la F.
   d) Si fos unilateral, es miraria bilateral a les taules invertint la F.

Respostes	1	2	3	4	5	6
	a	a	d	d	d	c

5.8 Problemes

Quatre problemes trets d'exàmens passats, i un en relació amb el text.

1. Una multinacional vol renovar la seva xarxa d'ordinadors. Dues empreses del sector informàtic ofereixen un nou tipus de CPU i, abans de fer la comanda, la multinacional executa un programa en 40 unitats de cada proveïdor i n'obté els resultats següents: el temps mitjà d'execució del primer proveïdor és de 23.36 segons; el temps mitjà d'execució del segon proveïdor és de 25.2 segons, $s_x^2 = 3.42$ segons^2 i $s_y^2 = 2.57$ segons^2 . Suposem que el temps d'execució dels dos models segueix una distribució normal $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$:

   a) Podem afirmar que la variabilitat del temps d'execució dels ordinadors de la primera empresa és igual a la de la segona, amb un risc del 5 per cent? (Plantegeu i resoleu el test d'hipòtesi adient. Digueu quin estadístic heu utilitzat. Quina distribució de probabilitat segueix aquest estadístic? Per què?)

   b) Es pot afirmar que el temps d'execució dels ordinadors de la primera empresa és inferior al de la segona, amb un risc del 5%? Amb quin model es quedarà la multinacional?

   c) Trobeu un interval de confiança per a μ_1 i per a μ_2. Feu els càlculs amb un risc del 5%.

   d) Calculeu un IC amb un risc del 5% per a la desviació típica del temps d'execució dels ordinadors triats per la multinacional.

   e) Estudis més precisos han demostrat que la distribució poblacional del temps d'execució dels ordinadors de la primera empresa és N(23, 3.25) i el de la segona N(25, 2.75). Executem un programa en cinc ordinadors d'una mateixa empresa.

      e1) Quin és el valor màxim del temps d'execució mitjà per acceptar que són de la primera empresa, amb un risc del 5%?

      e2) Quina és la probabilitat d'acceptar que els ordinadors són de la primera empresa quan, en realitat, són de la segona?

2. Volem estudiar el rendiment, en temps, de dos sistemes multiusuari (*A*, *B*), en l'execució d'un algorisme de compressió d'imatges. Concretament, volem estudiar el temps d'entrada i sortida (*Y*), definit com la diferència entre el temps real i el temps de CPU.

   El tècnic de sistemes ha decidit recollir, amb cada sistema, sis compressions de la mateixa imatge, una cada 4 hores, al llarg d'un dia. Aquests en són els resultats:

Sistema *A*:	16.2	13.0	14.0	35.8	24.3	32.8
Sistema *B*:	11.5	5.8	12.5	27.7	23.0	25.4

$\Sigma Y_A = 136.1$ $\Sigma Y_B = 105.9$ $\Sigma Y_A^2 = 3575.41$ $\Sigma Y_B^2 = 2263.59$

   a) Compareu-ne els rendiments, en temps, tot assumint que les dades són independents.

   b) Feu el mateix, però assumint que les dades són aparellades.

   c) Comenteu les eines de què disposa el tècnic de sistemes per decidir si es tracta de dades aparellades o independents. Heu d'especificar tant els raonaments de tipus teòric, segons l'obtenció de la mostra, com els procediments empírics o estadístics.

   d) Suposeu que els resultats obtinguts en les 6 i 6 observacions van ser:

	mitjana	variància
A	341	200
B	418	200
Dif	77	40

   Què opineu sobre l'estructura d'aquestes dades? Es tracta de dades aparellades o independents? (No es demana fer una prova d'hipòtesi; tan sols es demana cap a on apunten aquests resultats.)

3. (Continuació del problema 2 del capítol 3 i del problema 2 del capítol 4) Per reduir el cost, la gerència ens demana comparar el telemanteniment amb el sistema clàssic (ITRS) i un sistema alternatiu (WTRS) més barat. Les dades per fer la comparació provenen de les mateixes avaries.

ITRS	990	320	500	400
Log_{10}(ITRS)	2.996	2.505	2.699	2.602
WTRS	970	360	600	470
Log_{10}(WTRS)	2.987	2.556	2.778	2.672

   a) Raoneu i justifiqueu (amb dades concretes però sense fer inferència) els avantatges d'utilitzar les mateixes avaries.

b) Estimeu les possibles diferències entre els dos mètodes de telemanteniment.

c) De fet, els criteris de cost suggereixen que el segon sistema seria acceptable sempre que no incrementés més d'un 25% el temps de reparació de cada avaria. És a dir, sempre que: WTRS ≤ 1.25·ITRS. Creieu que es dóna aquesta situació? (Nota: per la seva construcció, sabem que es impossible que el segon trigui menys.)

4. Els professors d'Estadística, preocupats per la qualitat en la docència de l'assignatura que imparteixen, aquest quadrimestre han passat l'enquesta proposada per l'ICE, amb l'objectiu de detectar-ne els punts forts i febles, i poder dur a terme propostes de millora. Aquesta enquesta s'ha passat als tres grups i se n'han obtingut en total 83 respostes vàlides.

   En primer lloc, ens centrarem en el factor *aprenentatge*. Volem comparar l'ítem P02: *He après coses que considero valuoses (1=Molt en desacord / 5=Molt d'acord)* i l'ítem P36: *El teu nivell d'interès per la matèria abans d'aquest curs era (1=Molt petit /5=Molt gran).* Considerem que l'ítem P02 és una mesura de l'interès per la matèria després del curs, tal com el P36 ho és abans de l'inici. Volem veure si ha augmentat l'interès per la matèria un cop finalitzat el curs. Se n'han obtingut els resultats següents:

	n	Mean	StDev	SE Mean
P02	83	3.108	1.048	0.115
P36	83	2.060	0.846	0.093
P02-P36	83	1.048	1.209	0.133

   a) Per dur a terme aquesta comparació, quina anàlisi s'hauria de dissenyar (dades aparellades o independents? Per què?

   b) Creieu que realment s'haurà incrementat l'interès? Plantegeu i resoleu el test d'hipòtesi adient, amb un nivell de confiança del 95%

   c) Continuant amb aquest mateix factor d'aprenentatge, ítem P02, volem veure si els resultats són similars entre un grup del matí i un grup de la tarda. Quin disseny s'hauria de seleccionar, dades aparellades o independents? Per què?

   d) Calculeu l'interval de confiança, al 95%, per a la diferència de mitjanes poblacionals de respostes d'aquest ítem. Especifiqueu les premisses necessàries per poder efectuar aquest càlcul i verifiqueu les que es puguin amb la informació proporcionada.
   (Variables: P02_A, P02_B). L'estadística descriptiva que s'obté és:

	n	Mean	StDev	SE Mean
P02_A	34	3.029	1.029	0.177
P02_B	34	3.118	1.200	0.206
P02_A-P02_B	34	-0.088	1.694	0.291

   e) A partir d'aquest interval, podeu considerar que hi ha diferències significatives pel que fa a l'aprenentatge d'aquests grups? Justifiqueu la vostra decisió.

5. A l'exemple del apartat 5.3 (mostres aparellades), s'ha considerat oportú incloure el factor *ordre de la seqüencia* a l'hora de provar els antivirus: primer l'*A* o primer el *B*. Això vol dir que un antivirus pot malmetre el sistema, de manera que el seu efecte repercuteixi en el programa que s'executa després (un efecte *retardat*, que s'atribuiria erròniament al segon antivirus). Es pot controlar d'alguna manera comparant les execucions efectuades en primer lloc o en segon lloc, és a dir, prenent els dos grups:

AB	*BA*
22	6
42	12
64	31
8	-5
62	95

Comproveu que l'ordre no representa un factor rellevant.

5.9 Solució dels problemes

1.

a) Test bilateral de comparacions de variàncies:

$$H_0: \sigma_x^2 = \sigma_y^2 \qquad \frac{s_x^2}{s_y^2} \sim F_{n_1-1,n_2-1} = F_{39,39}$$
$$H_1: \sigma_x^2 \neq \sigma_y^2$$

Valor de l'estadístic: $\frac{s_x^2}{s_y^2} = \frac{3.42}{2.57} = 1.33$

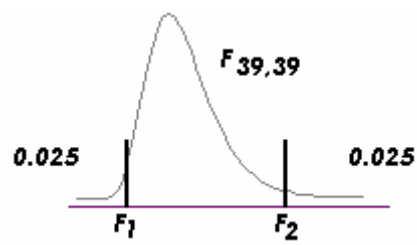
Punts crítics: la distribució de probabilitat és una F de Fisher, quocient de dues khi quadrat.

$0.025 = P(F_{39,39} \geq F_2) \Rightarrow F_2 = 1.88$

A la taula, hem mirat el valor de $F_{40,40}$:

$$0.025 = P(F_{39,39} \leq F_1) = P\left(\frac{1}{F_1} \leq \frac{1}{F_{39,39}}\right) = P\left(\frac{1}{F_1} \leq F_{39,39}\right)$$

$$\Rightarrow F_1 = \frac{1}{F_2} = \frac{1}{1.88} = 0.53$$



Zona d'acceptació: (0.53, 1.88)

Com que el valor de l'estadístic $1.33 \in (0.53, 1.88) \Rightarrow$ acceptem H_0, és a dir, res no s'oposa a acceptar que les variàncies són iguals (que no vol dir que haguem demostrat que ho siguin).

b) Test de comparació de mitjanes amb variàncies iguals i desconegudes:

H_0: $\mu_x = \mu_y$

H_1 $\mu_x < \mu_y$ (per l'apartat 1)

$$\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2} \qquad \text{on} \quad s=\sqrt{\frac{(n_1-1)s_x^2+(n_2-1)s_y^2}{n_1+n_2-2}}$$

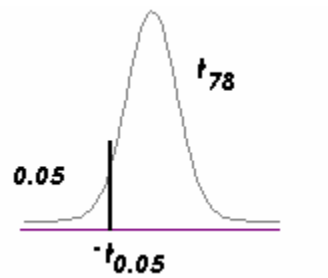
Valor de l'estadístic sota H_0: $\mu_x = \mu_y$, H_0: $\mu_x - \mu_y = 0$

$$s^2 = \frac{(n_1-1)s_x^2+(n_2-1)s_y^2}{n_1+n_2-2} \underset{\uparrow\, n_1=n_2}{=} \frac{s_x^2+s_y^2}{2} = \frac{3.42+2.57}{2} = 2.995 \Rightarrow s = \sqrt{2.995} = 1.73$$

$$t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{23.36-25.2}{1.73\sqrt{\frac{2}{40}}} = -4.756$$

Punt crític:

$0.05 = P(t_{78} \le t_\alpha) \Rightarrow t_\alpha \cong -1.66$



Observació: $t_{60,0.05} = -1.671$, $t_{120,0.05} = -1.658$

Com que $t = -4.756 \in (-\infty, -1.66) \Rightarrow$ rebutgem H_0

Podem afirmar que el temps d'execució del primer CPU és inferior al del segon, amb un risc de 5%; per tant, la multinacional es quedarà amb el primer model.

c)

$$\frac{\bar{x}-\mu_1}{\frac{s_x}{\sqrt{n_1}}} \sim t_{n_1-1} \Rightarrow IC(\mu_1;0.05) = \left(\bar{x} - t_{1-\alpha/2}\frac{s_x}{\sqrt{n_1}}, \bar{x} + t_{1-\alpha/2}\frac{s_x}{\sqrt{n_1}}\right)$$

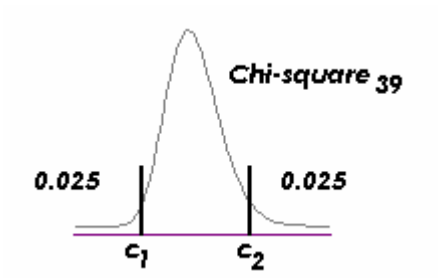
on $0.025 = P(t_{39} \le t_{1-\alpha/2}) \Rightarrow t_{1-\alpha/2} = 2.021$ (Valor de t_{40})

$$IC(\mu_1;\alpha=0.05) = \left(23.36 - 2.021\frac{\sqrt{3.42}}{\sqrt{40}},\ 23.36 + 2.021\frac{\sqrt{3.42}}{\sqrt{40}}\right) = (22.76,\ 23.95)$$

$$IC(\mu_2;\alpha=0.05) = \left(25.2 - 2.021\frac{\sqrt{2.57}}{\sqrt{40}},\ 25.2 + 2.021\frac{\sqrt{2.57}}{\sqrt{40}}\right) = (24.68,\ 25.71)$$

d)

$$\frac{(n_1-1)s_x^2}{\sigma_x^2} \sim \chi^2_{n_1-1}$$



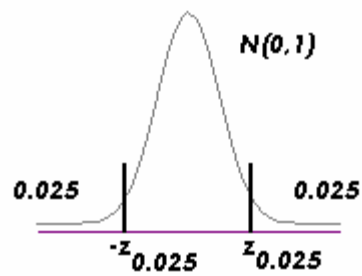
$$0.025 = \mathrm{P}(\chi^2_{39} \le \chi_1) \Rightarrow \chi_1 \cong 16.791$$
$$0.975 = \mathrm{P}(\chi^2_{39} \le \chi_2) \Rightarrow \chi_2 \cong 46.979$$

Observació. A les taules, hem agafat els valors corresponents a χ^2_{30}.

$$\chi_1 \le \frac{(n_1-1)s_x^2}{\sigma_x^2} \le \chi_2 \Rightarrow \frac{(n_1-1)s_x^2}{\chi_2} \le \sigma_x^2 \le \frac{(n_1-1)s_x^2}{\chi_1} \Rightarrow$$
$$\Rightarrow \sqrt{\frac{(n_1-1)s_x^2}{\chi_2}} \le \sigma_x \le \sqrt{\frac{(n_1-1)s_x^2}{\chi_1}} \Rightarrow \sqrt{\frac{39.342}{46.979}} \le \sigma_x \le \sqrt{\frac{39.342}{16.791}} \Rightarrow$$
$$\Rightarrow 1.68 \le \sigma_x \le 2.81$$

e1)

H_0: $\mu = 23$
H_1: $\mu = 25$



$$x_i \sim N(23, 3.25) \Rightarrow \bar{x} \sim N(23, \frac{3.25}{n}) \Rightarrow Z = \frac{\bar{x}-23}{\frac{\sqrt{3.25}}{\sqrt{n}}} \sim N(0,1)$$

$$0.025 = \mathrm{P}(Z \le z_{\alpha/2}) \Rightarrow z_{\alpha/2} = -1.96$$
$$-1.96 \le \frac{\bar{x}-23}{\sqrt{3.25/5}} \le 1.96 \Rightarrow \bar{x} \le 23 \pm 1.96\sqrt{\frac{3.25}{5}} = 24.58$$

e2)

Regió d'acceptació: $\bar{x} \in (23-1.96\sqrt{\frac{3.25}{5}}, 23+1.96\sqrt{\frac{3.25}{5}}) \Rightarrow \bar{x} \in (21.41, 24.58)$

$$\beta = \mathrm{P}(\text{acceptar } \mathrm{H}_0 \mid \mathrm{H}_1 \text{ cert}) = \mathrm{P}(21.41 \le \bar{x} \le 24.58 \mid \mu = 25) =$$
$$= \mathrm{P}\left(\frac{21.41-25}{\sqrt{2.75/5}} \le \frac{\bar{x}-25}{\sqrt{2.75/5}} \le \frac{24.58-25}{\sqrt{2.75/5}}\right) = \mathrm{P}(-4.84 \le Z \le -0.56) =$$
$$= F(-0.56) - F(-4.84) = 1 - F(0.56) = 1 - 0.7123 = 0.2877$$

2.

a)

$$\overline{Y}_A = 136.1 / 6 = 22.68$$
$$\overline{Y}_B = 105.9/6 = 17.65$$

$$s_A^2 = (3575.41 - 136.1^2 / 6) / 5 = 97.64$$
$$s_B^2 = (2263.59 - 105.9^2 / 6) / 5 = 78.89$$
$$s^2_{CONJUNTA} = (97.64 \cdot 5 + 78.89 \cdot 5) / 10 = 88.27$$

Hem de contrastar:

$$\begin{cases} \mathrm{H}_0 : \mu_A = \mu_B \\ \mathrm{H}_1 : \mu_A \neq \mu_B \end{cases}$$

$$t = (22.68 - 17.65) / \sqrt{97.64/6 + 78.89/6} = 5.03 / 5.42 = 0.93$$

Premisses:

- Normalitat (podria estudiar-se amb un *normal probability plot*, NPPlot). S'assumeix.
- Homoscedasticitat (sembla raonable: $97.64 / 78.89 \approx 1$)
- Independència de les observacions (per l'enunciat)

Com que $t = 0.93 < 2.23 = t_{10,0.975}$, res no s'oposa a acceptar H_0.
No hem aconseguit demostrar que hi hagi diferències entre ambdós sistemes.

b)

Diferència:	4.7	7.2	1.5	8.1	1.3	7.4

$\Sigma D = 30.2$ $\quad\quad$ $\overline{D} = 30.2 / 6 = 5.03$

$\Sigma Y_A^2 = 198.24$ $\quad\quad$ $s_A^2 = (198.24 - 30.22^2 / 6) / 5 = 9.25$

Hem de contrastar:

$$\begin{cases} \mathrm{H}_0 : \mu_D = 0 \\ \mathrm{H}_1 : \mu_D \neq 0 \end{cases}$$

$$t = 5.03 / \sqrt{9.25/6} = 5.03 / 1.24 = 4.06$$

Premisses:

- Normalitat (podria estudiar-se amb un *normal probability plot*). S'assumeix.
- Dades aparellades (per l'enunciat)
- Independència en les diferències. S'assumeix.

Com que $t = 4.06 > 2.57 = t_{5,\,0.975}$, es refusa H_0. Hem demostrat que un sistema és més ràpid que l'altre.

c)

Teòrics: raonar si la forma d'obtenció de les dades pot originar algun tipus de similitud o relació entre les parelles d'observacions. Es tracta de buscar fonts de variació que siguin específicament comunes per a les parelles. És a dir, que les comparteixin, exclusivament, els elements de les parelles (coses que tinguin en comú els dos elements de la mateixa parella, però que siguin diferents de la resta d'individus de les altres parelles).

Empírics: estudiar el grau de relació entre ambdues mostres, ja sigui mitjançant l'estadístic de correlació de Pearson o mitjançant la inspecció del gràfic *A versus B*.

d) El fet que la variància de la diferència sigui més petita apunta que les dades són aparellades (s'ha eliminat la font de variació que aparella les dades). Més formalment, ho podem justificar amb els càlculs següents:

$$V(D) = V(A) + V(B) - 2\,Cov(A,B)$$

$$Cov(A,B) = [\,V(A) + V(B) - V(D)\,] / 2 = (200+200-40)/2 = 180$$

$$r(A,B) = Cov(A,B) / (s_A \cdot s_B) = 180 / (\sqrt{200} \cdot \sqrt{200}\,) = 0.9$$

Aquests resultats apunten una relació positiva entre les variables *A* i *B*. Es tractaria de dades aparellades.

3.

a)

Log_{10}(ITRS)	2.996	2.505	2.699	2.602
Log_{10}(WTRS)	2.987	2.556	2.778	2.672
Dif	-0.0089	0.0512	0.0792	0.0700

$$\overline{D} = \Sigma_{i=1,\ldots 4}\, D_i / n = (-0.0089 + 0.0512 + 0.0792 + 0.0700)/4 = 0.0479$$

$$s_D^2 = [(-0.0089-0.0479)^2 + \ldots]/(4\text{-}1) \cong 0.00157 \cong 0.0396^2$$

Permeten dades aparellades i treballar amb diferències.
Elimina la variabilitat deguda al diferent tipus d'avaria.
En disminuir la variància, augmenta la precisió i la potència estadística.

$$s_D^2 \cong 0.00157 \;<<<\; s^2 \cong 0.045$$

b) Estimeu les possibles diferències entre els dos mètodes de telemanteniment.

Estadístic $T = (\overline{D} - \mu_D) / (s_D/\sqrt{n}) \rightarrow t_{n-1}$
Premissa: diferència normal

$$IC_{95\%}(\mu_D) = \overline{D} \pm t_{3,\,0.025}\, s_D/\sqrt{n} = 0.0479 \pm 3.182 \cdot 0.0396/\sqrt{4} = 0.0479 \pm 0.0630 =$$
$$= [-0.0151,\ 0.1109]$$

Conclusió. Amb una confiança del 95%, la diferència mitjana entre ambdós mètodes es troba entre aquests valors (o també: la mitjana de la ràtio W/I està entre 0.966 i 1.291, al 95% de confiança; recordeu que la base de l'operació exponencial aquí és 10).

c)

$$WTRS \le 1.25 \cdot ITRS\ ;$$
$$WTRS\ /\ ITRS\ \le 1.25;$$

$$\log_{10}(WTRS\ /\ ITRS) \le \log_{10} 1.25;$$
$$\log_{10} WTRS - \log_{10} YTRS\ ) \le 0.097\ ;$$

L'interval anterior (–0.0151, 0.1109) inclou aquest punt (0.097); per tant, és plausible que, de mitjana, el nou sistema incrementi en un 25% o més el temps de l'anterior.

$$H_0 : \mu_D \le 0.097$$
$$H_1 : \mu_D > 0.097$$

Regla de decisió: refusar H_0 (α=0.05) si $t > t_{3,\,0.95} = 2.353$

Càlcul: $t = (0.0479 - 0.097) / (0.0396/\sqrt{4}) \cong -2.477$

Decisió: atès que t no és més gran que 2.353, no es pot refusar que μ_D sigui més petit que 0.097, amb un risc α=0.05. Per poder rebutjar, la mitjana de D hauria de ser almenys de $0.097 + 0.0198 \cdot 2.353 = 0.1436$.

Conclusió pràctica: no es pot dir que el nou sistema incrementi els temps en un 25%.

4.

a) Dades aparellades: és el mateix estudiant que respon els dos ítems.

b)

$$H_0: \mu_{P02} = \mu_{P36}$$
$$H_1: \mu_{P02} > \mu_{P36}$$

Estadístic: $$T = \frac{\overline{D} - \mu_0}{s_D / \sqrt{n}}$$

Distribució sota H_0 : $t \to t_{82}$

Premisses: diferències aparellades, normals i MAS

Regla de decisió: rebutjar H_0 ($\alpha = 0.05$) si O bé $p < \alpha = 0.05$
O bé $t > t_{82,\,0.95} = 1.65$

Càlculs: $$t = \frac{\overline{D} - \mu_0}{s_D / \sqrt{n}} = \frac{1.048 - 0}{1.209 / \sqrt{83}} = 7.9$$

Decisió: amb taules, com que $t = 7.9 > 1.65 = t_{82,\,0.95}$, es rebutja $\mu_{P02} = \mu_{P36,}$ amb risc $\alpha = 0.05$.

Conclusió pràctica: els estudiants han incrementat l'interès per la matèria.

c) En aquestes dades, independents, ja que són estudiants diferents que responen en grups diferents.

d) Per poder dur a terme el càlcul, cal que es compleixin les premisses següents:

- Les respostes obtingudes a l'ítem P02, tant en el grup *A* com en el grup *B*, han de seguir una distribució normal, i aquesta premissa s'assumeix, ja que no es pot comprovar amb les dades del problema.
- Les dues mostres són independents, pel fet que els estudiants del grup *A* no han respost les enquestes del grup *B*, i viceversa.
- Cal comprovar si es pot assumir igualtat de variàncies poblacionals, és a dir:

$$H_0: \sigma^2_{P02\_A} = \sigma^2_{P02\_B}$$
$$H_1: \sigma^2_{P02\_A} \neq \sigma^2_{P02\_B}$$

Estadístic**:** $$F = \frac{s^2_{\text{major}}}{s^2_{\text{menor}}}$$

Distribució sota H_0 : $F \rightarrow F_{33,33}$

Premisses: les respostes obtingudes a l'ítem P02, tant en el grup *A* com en el grup *B*, han de seguir una distribució normal.

Regla de decisió: rebutjar H_0 (α=0.05) si $F > F_{33,33\ 0'975} \approx 2$

Càlculs: $$F = \frac{s^2_{\text{major}}}{s^2_{\text{menor}}} = \frac{1.200^2}{1.029^2} = 1.36$$

Decisió: amb taules, com que $F = 1.36 < 2 = F_{33,33\ 0.975}$, s'accepta $\sigma^2_{P02\_A} = \sigma^2_{P02\_B}$, amb risc $\alpha = 0.05$.

Conclusió pràctica: res no s'oposa a acceptar que la variabilitat poblacional de les respostes obtingudes a l'ítem P02 sigui la mateixa per a ambdós grups.

Calculem ara l'interval de confiança al 95% per a la diferència de mitjanes poblacionals.

Arran del resultat obtingut en el test d'hipòtesi precedent, cal calcular:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{33 \cdot (1.2^2 + 1.029^2)}{66} = 1.25$$

Per tant, $s=\sqrt{1{,}25}=1{,}12$ $\quad \bar{y}_{\mathrm{P02\_}A}\ \bar{y}_{\mathrm{P02\_}B}$

$$\begin{aligned}\mathrm{IC}(\mu_{\mathrm{P02}A}-\mu_{\mathrm{P02}B},\,0{,}95) &= (\bar{y}_{\mathrm{P02\_}A}-\bar{y}_{\mathrm{P02\_}B})\pm t_{66,\,0.975}\cdot s\sqrt{(1/n_{\mathrm{P02}A}+1/n_{\mathrm{P02}B})}\\ &= -0.088\pm 1.96\cdot 1.12\,\sqrt{2/34}\ \approx\\ &\approx -0.088\ \pm 0.532\approx\\ &\approx [-0.620,\,0.444]\end{aligned}$$

e) Amb una confiança del 95%, es pot considerar que no hi ha diferències significatives, ja que el 0 pertany a l'interval de confiança [–0.620, 0.444]

5. Es tracta d'una comparació de dues mostres independents. Les diferències $B-A$ tenen una mitjana observada, a la seqüència *AB*, igual a 39.6 (s^2=602.8), i a la seqüència *BA* és 27.8 (s^2=1582). L'estadístic *t* val 0.56 i el p-valor equivalent és 0.588. Per tant, no s'observa cap indici que l'ordre pugui introduir un efecte significatiu a les diferències entre *A* i *B*.

6 Relacions entre variables. Model lineal

6.1 Relació entre dues variables numèriques

En aquest capítol, estudiarem la relació existent entre una variable *Y*, anomenada *resposta*, i una variable *X*, que rep el nom de variable *independent, predictora* o *explicativa*, amb la qual volem explicar les variacions de la primera. El model més senzill s'expressa com una funció lineal del valor esperat de *Y* respecte de *X*; és, doncs, una relació no determinista, on la resposta *Y* té variabilitat, encara que es fixi el valor de *X*.

Ens concentrarem en el cas en què només tenim una variable independent *X* i una variable resposta *Y*. Els models que utilitzen més d'una variable explicativa es descriuen en textos més avançats[14] d'estadística, i aquí no es mencionaran.

A l'apartat 6.6, veurem amb detall un cas particular del model lineal, on la variable *X* no es considera quantitativa sinó que pren valors entre un conjunt de categories. En tot cas, la resta de premisses són les mateixes que en el cas anterior, excepte la relació funcional entre *Y* i *X*, per raons evidents.

El tipus de dades que veurem ara pot ser com el que es mostra a la figura 6.1. La resposta és el salari en dòlars per hora i *X* és el nombre d'anys d'educació de la persona observada (dades[15] del cens de 1985 dels Estats Units).

A la vista de la figura 6.1, contestar la pregunta de si *el salari augmenta amb els anys d'educació de la persona* no és trivial. En podem assenyalar dos trets principals: primer, la tendència és positiva; en general, el sou és més alt com més educació ha rebut l'individu. Segon, la norma es trenca sovint; més anys no implica sempre més salari; hi observem una dispersió notable que implica l'abundància d'excepcions a la regla anterior.

14 Per exemple: Fox, John. *Applied Regression Analysis, Linear Models, and Related Methods*. 1997. Sage Publications.

15 Referència: Berndt, E.R. *The Practice of Econometrics*. 1991. Nova York: Addison-Wesley.

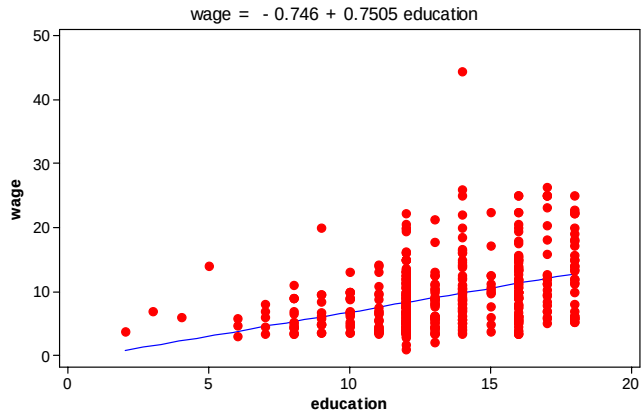


*Figura 6.1 Salari per hora (*wage*) respecte dels anys d'educació rebuda per l'individu (*education*)*

Atès que no és possible respondre categòricament a la qüestió, considerarem una proposta més acurada. L'expressió que figura al gràfic, wage = –0.746 + 0.7505·education, és l'equació de la recta que travessa el núvol de punts, que ve a indicar *grosso modo* (ja veurem com) la relació existent entre anys i salari. El terme independent és negatiu, però sembla suficientment a prop de zero com per deduir que no hi ha salari base separat dels anys de formació. Quant al terme lineal, vol dir que el salari per hora augmenta (com a tendència general) en 75 cèntims de dòlar per cada any extra que el treballador aporta. Observeu que aquesta proposta no encerta molt bé amb els sous de les persones amb cinc anys o menys de formació.

De qualsevol forma, és necessari afegir que aquesta recta no incorpora tota la realitat. Al voltant d'aquest patró es produeix una fluctuació considerable, potser més petita per sota dels 12 anys. De manera aproximada, es podria dir que una variació de cinc dòlars al voltant del que prediu la recta no seria gens estrany.

Haver de parlar d'un terme associat a la variabilitat pot semblar incòmode, però és del tot necessari quan les dades reflecteixen una realitat complexa. Totes les persones amb 14 anys de formació han de cobrar el mateix? De fet, no és el temps invertit en l'educació l'únic factor que importa; se'n podrien enumerar molts d'altres, encara que el cens no recull més que la mínima informació (sector professional, categoria ocupacional, sexe, raça, edat,...). No tractar tots aquests aspectes explícitament es tradueix en incertesa o fluctuació aleatòria. El model lineal del qual parlarem resumeix tota aquesta realitat en tres elements: els dos coeficients de la recta i la variància residual, una estimació de la variabilitat de Y per a cada valor de X. Amb això en tenim prou per afrontar preguntes com:

- Quina part de la variabilitat del salari dels americans s'explicaria pel temps d'educació?
- Quin és l'increment esperat de sou per any d'educació addicional?
- Quant es guanya, de mitjana, amb 15 anys d'educació?

Abans hauríem d'explicar en què es fonamenta el model lineal, quina participació hi té l'atzar, com es troben els coeficients esmentats, com s'aborda la qüestió de la inferència estadística, què significa fer previsions, com sabem que el model és bo...

6.2 Regressió lineal simple

Malgrat que la paraula *regressió* ve de finals del segle XIX, de Francis Galton, es tracta d'un mètode que Gauss, Legendre i Adrain van desenvolupar per separat a la primera dècada del segle, aplicat a la determinació de la posició en el cel de cossos espacials. El mot prové de la paraula llatina *regressio* que significa "retorn", en una interpretació molt particular que en fa Galton, i ha perdurat fins als nostres dies, encara que el sentit original ja no es conserva.

Vegeu-ne més informació a aquesta adreça:
<http://campusvirtual.uma.es/est_fisio/apuntes/ficheros/estad_uma_03.pdf>

La *regressió lineal simple* (un sol regressor X) es pot formular com:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

Y_i	valor observat de la resposta en el cas i
X_i	valor de la variable X en el cas i
β_0	constant o terme independent
β_1	pendent de la recta
ε_i	terme d'error aleatori, interpretat com la desviació del valor de Y_i respecte de la recta

Aquest model té dues components, la determinista, formada per l'equació de la recta, i l'aleatòria o estocàstica, representada pel terme d'error:

$$Y_i = \mu_i + \varepsilon_i \quad \text{on} \quad \mu_i = \beta_0 + \beta_1 \cdot X_i \quad \text{part determinista (lineal) de } Y$$
$$\varepsilon_i \quad \text{part aleatòria de } Y$$

Com es veu, els valors esperats de la resposta per a diferents X_i (μ_i) se situen sobre la recta definida pels dos paràmetres, β_0 i β_1. Al voltant del valor μ_i hi ha una certa oscil·lació ε_i, amb esperança zero i de variància constant (sigui quin sigui el valor de X) $V(\varepsilon_i)=\sigma^2$, també coneguda com a variància residual.

Un *exemple* senzill de tot el que hem dit fins ara és: el pes en quilograms (Y) d'homes adults i sans de Barcelona, en funció de la seva alçada en centímetres (X). Suposem que el pendent és 1 kg/cm i que la constant és –100 kg:

$$\beta_0 = -100 \text{ kg} \qquad \beta_1 = +1 \text{ kg/cm}$$

El pendent indica que cal esperar 1 kg més de pes per a cada 1 cm més d'alçada per a cada individu. I la constant s'interpreta recorrent a la seva definició geomètrica: –100 kg és el valor de Y en què la recta creua l'eix d'ordenades (no és el pes esperat d'un senyor de 0 cm d'alçada). En resum, l'equació d'aquesta recta és:

$$\mu_i = -100 \text{ kg} + 1 \text{ kg/cm} \cdot X_i$$

Així, per exemple, el pes predit amb aquest model per a un individu de 180 cm és 80 kg. Fins ara no hi ha res nou: s'ha aplicat l'equació de la recta. Però l'estadística no espera que

tots els individus de 180 cm pesin 80 kg. Hem de incloure un nou paràmetre en el model: la dispersió al voltant d'aquest valor predit. Suposem que és $\sigma^2 = 36$ kg^2 o, simplement, $\sigma = 6$ kg.

Així, la desviació tipus d'un cas de 180 cm al voltant del valor predit de 80 kg és de 6 kg. Si s'espera una desviació de 6 kg, no seran estranyes desviacions de 2 kg, ni de 10 kg, però sí que sorprendria una desviació de 25 kg. Fins i tot… si es pogués aplicar una distribució normal, es podria dir que el 95% dels individus de 180 cm pesen entre 68 i 92 kg:

$$\text{IC(pes, 95\%)} = \mu \pm z_{\alpha/2} \cdot \sigma = 80 \text{ kg} \pm 1.96 \cdot 6 \approx [68, 92]$$

Les qüestions que resten pendents ara són les següents: (1) Com s'estimen els valors de β_0 i β_1? (2) Fins a quin punt es pot creure que el model és real i correcte? A l'exemple, el pendent d'1 kg/cm es manté tant al voltant de 170 cm com de 140 cm? (3) Es pot quantificar la qualitat, és a dir, la capacitat predictiva d'aquest model?

Nota. Els paràmetres de la recta s'han d'interpretar d'acord amb les seves unitats i segons es pugui parlar o no de relació causal. Així, si les condicions de la recollida de les dades ho permeten, el pendent es pot interpretar causalment: la desposta Y tindrà un canvi esperat de β_1 (unitats de Y) per a cada increment d'una unitat que es provoqui en la causa X. Per exemple, cada hora d'estudi pot incrementar la nota final en un examen en 0.05 punts.

En canvi, si X és un atribut dels casos i no una condició que depengui de l'investigador, o si se sospita que terceres variables Z poden ser l'explicació de l'associació observada, es diu: un canvi d'una unitat en la variable X s'associa (o acompanya) amb un canvi de β_1 unitats en la variable Y. Per exemple, cada punt de més en l'assignatura de llengua s'associa a un increment de 0.75 punts en la nota de matemàtiques.

6.2.1 Premisses

Naturalment, aquest model no és universal. A continuació, es veuen les condicions que s'han de donar perquè pugui ser raonable adoptar i utilitzar aquest model.

La premissa que fa referència a la part determinista del model és aquella que determina que la relació funcional entre les variables és la línia recta (la ja comentada *linealitat*). La informació més rellevant està inclosa aquí (per exemple, un increment en el nombre de processos en el sistema causa un increment en el temps que triga a finalitzar un programa determinat). La part aleatòria del model es basa en *pertorbacions* ε_i que afecten la resposta com a soroll blanc (sense informació). Les premisses necessàries per desenvolupar el model que presentem són tres, que s'han d'afegir a l'anterior de linealitat:

- *Normalitat* de la distribució dels errors.
- *Homoscedasticitat*, o variància constant de l'error, sigui quin sigui el valor de X.
- *Independència* dels errors. Les observacions han estat recollides de forma independent.

D'altra banda, hi ha una altra premissa que fa referència a la forma com s'han recollit les dades: els valors de les X no són aleatoris. És a dir, no són el resultat d'una observació, sinó que són escollits per l'investigador, que ha assignat el seu valor als casos.

Al capítol següent, es tractarà com analitzar la validesa de les premisses i es comentarà les conseqüències d'una manca de compliment.

Nota. Aquestes premisses estableixen que la variable aleatòria ε té distribució N(0, σ^2), per la qual cosa la variància de la desposta Y, condicionada per qualsevol valor concret de X, coincideix amb la variància de les pertorbacions ε:

$$\begin{aligned} V(Y_i) \equiv V(Y \mid X_i) &= V(\beta_0 + \beta_1 \cdot X_i + \varepsilon_i) = \\ &= V(\beta_0) + V(\beta_1 \cdot X_i) + V(\varepsilon_i) = \\ &= 0 + 0 + \sigma^2 = \sigma^2 \end{aligned}$$

De tota manera, cal recordar que les n observacions —amb diferents X— proporcionen una variància de la resposta usualment més gran (v. il·lustració de la figura 6.8) perquè, a part de la variabilitat de les pertorbacions, inclou la variabilitat del factor X, si és que aquest està relacionat amb Y. Tornarem sobre el tema a l'apartat 6.4.

6.3 Estimació dels paràmetres

Exemple. La pantalla de l'ordinador portàtil és l'element que consumeix més energia del sistema. Per estudiar l'impacte que el nivell de brillantor de la pantalla (que l'usuari pot graduar) té en la durada de la bateria, treballant amb tasques quotidianes, es mesura el temps que l'ordinador triga des que arrenca amb la bateria totalment carregada fins que avisa per manca d'energia suficient per continuar. Els resultats obtinguts figuren a continuació:

Brillantor	1	2	3	4	5	6	7	8	9	10
Durada (min)	241	193	205	169	174	134	163	124	111	92

És obvi que l'efecte existeix, ja que s'aprecia clarament que la bateria disminueix la durada amb nivells de brillantor creixents, malgrat que hi ha una fluctuació aleatòria al voltant d'una trajectòria descendent, possiblement rectilínia, com es pot veure a la figura 6.2.

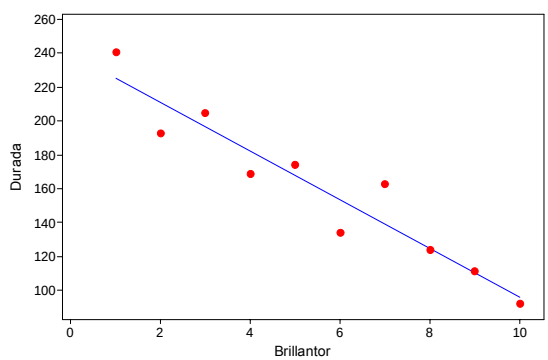


Figura 6.2 Diagrama amb la posició de les deu observacions de brillantor i durada de la bateria. La recta que s'ha traçat correspon a la recta de regressió.

> Per tant, la qüestió és determinar com varia la durada de la bateria quan aquest nivell es modifica. Això suposa obtenir estimadors dels tres paràmetres desconeguts en joc: β_0, β_1 i σ^2. És a dir, a partir d'una mostra representativa de la població de temps de la bateria, volem inferir un valor per a cada paràmetre, tenint en compte l'error comès pel mostreig, i utilitzar el resultat per obtenir previsions.

No és molt difícil traçar una línia recta que segueixi, aproximadament, la tendència d'un núvol de punts. Intuïtivament, tractem de dibuixar una línia que no passi lluny de cap punt. Aquest mateix criteri es pot formular matemàticament, per tal de trobar la recta que s'aproxima més al núvol. En primer lloc, definim el conjunt de totes les rectes candidates:

$$\{(b_0, b_1) \mid b_0, b_1 \in \mathbf{R}\}$$

on b_0 designa el terme independent i b_1 el pendent.

La recta estimada seria: $\hat{y}_i = b_0 + b_1 \cdot X_i$ i amb aquesta recta es pot trobar una predicció per a cada punt. Anomenem error de la predicció de la i-èsima observació (e_i) la diferència entre l'observació i la predicció $(y_i - \hat{y}_i)$.

Una bona aproximació seria una recta que fa petites les diferències $|y_i - \hat{y}_i|$. Encara que es poden aplicar diversos criteris, el més comú i també el que té millors propietats és el que es coneix com a *estimació mínim quadràtica*, perquè escull el parell (b_0, b_1) que minimitza la suma dels errors de previsió al quadrat:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot X_i)^2$$

Altres opcions, com fer mínima la suma dels valors absoluts o la major diferència, són notablement més complexes, ja que no es tracta de funcions derivables.

La recta trobada és una recta d'equació $Y = b_0 + b_1 \cdot X$, que esperem que sigui molt semblant a la recta que hipotèticament es dóna a la població: $Y = \beta_0 + \beta_1 \cdot X$ (però el grau de semblança depèn de la grandària de la mostra i també —per què no?— de la sort). Novament, ens trobem amb una situació en la qual els estimadors tenen incertesa i varien segons les dades recollides.

Trobem les expressions per calcular el valor dels estimadors per a una mostra concreta. El desenvolupament complet es troba a l'annex 1 del capítol. De les equacions resultants obtindrem la solució:

$$\boxed{b_0 = \overline{Y} - b_1 \cdot \overline{X}}$$

$$\boxed{b_1 = s_{XY} \Big/ s_X^2 = r_{XY} \, s_Y \Big/ s_X}$$

(Recordatori: s_{XY} designa la covariància mostral; r_{XY}: correlació mostral o estimada; s_Y: desviació tipus mostral de Y; s_X^2: variància mostral de X)

Una propietat que cal recordar és que utilitzar aquests estimadors implica que la recta passa pel centre del núvol (definit per les seves mitjanes $\overline{X}$, $\overline{Y}$).

Nota. Recordeu que es defineix el coeficient de correlació lineal ρ_{YX} com la covariància, dividida per les desviacions estàndards d'ambdues variables: $\rho_{XY} = \sigma_{XY} / \sigma_X \cdot \sigma_Y$, tipificant el resultat a un valor entre –1 i 1; s'estima per: $r_{XY} = s_{XY} / s_X s_Y$.

El coeficient de correlació atorga un paper simètric a les dues variables en estudi: ambdues són aleatòries. En endavant, no es treballarà amb el coeficient de correlació, però trobarem un resultat que es relaciona directament amb el coeficient de correlació lineal (el coeficient de determinació R^2).

Exemple. Calculem l'estimació del terme independent i del pendent de la recta per a les dades de la durada de la bateria, segons el nivell de brillantor seleccionat.

Mitjana *X*	5.5	Desv. tipus *X*	3.028	Correlació *X*, *Y*	–0.95082
Mitjana *Y*	160.6	Desv. tipus *Y*	45.9		

D'acord amb aquestes dades, el pendent estimat val:

$$b_1 = \frac{-0.95082 \cdot 45.9}{3.028} = -14.412$$

Això suposa que cada grau que apugem la brillantor de la pantalla significa uns catorze minuts menys de temps que la bateria podrà alimentar l'ordinador. El terme independent serà:

$$b_0 = 160.6 - (-14.412) \cdot 5.5 = 239.9$$

Si hi hagués un grau 0 de brillantor, l'ordinador podria treballar durant unes quatre hores (perquè l'estimació amb la recta seria $239.9 - 14.412 \cdot 0 = 239.9$ minuts, gairebé 4 hores).

6.3.1 Distribució dels estimadors mínim quadràtics

A continuació, s'estudia quina és la distribució de probabilitat dels estimadors b_0 i b_1, indispensable per aplicar inferència. Comencem per b_1, estimador del pendent.

a) Distribució de l'estimador b_1 del pendent

Els detalls de la demostració es troben a l'annex 2 del capítol. En primer lloc, cal recalcar emfàticament que la variable b_1 seguirà, al llarg de les possibles mostres, una distribució normal.

Vegem ara l'esperança i la variància d'aquesta variable b_1:

$$\mathrm{E}(b_1) = \beta_1$$

$$\mathrm{V}(b_1) = \frac{\sigma^2}{(n-1)s_X^2}$$

En resum, l'estimador del pendent segueix una distribució normal de paràmetres:

$$b_1 \sim \mathrm{N}(\beta_1, \sigma^2 / (n-1)s_X^2)$$

Destaquem: (1) es tracta d'un estimador sense biaix; (2) la dispersió de l'estimador és proporcional a la dispersió de les pertorbacions, però en relació inversa a la dispersió de les X i al nombre d'observacions. Per obtenir una bona estimació, cal evitar una concentració excessiva d'observacions en un rang molt estret per a la variable X.

Una vegada es coneix la distribució de l'estimador b_1, s'està en condicions de realitzar inferència estadística mitjançant els procediments habituals (intervals de confiança, contrastos d'hipòtesis), atès que l'estadístic següent seguirà una distribució normal centrada i reduïda:

$$(b_1 - \beta_1) / \sqrt{(\sigma^2 / (n-1)\, s_X^2)} \sim \mathrm{N}(0,1)$$

Si bé, com que també s'estimarà σ^2 per s^2 (ja veurem com és s^2), la distribució de referència serà una t de Student:

$$t_{b_1} = (b_1 - \beta_1) / \sqrt{(s^2 / (n-1)\, s_X^2)} \sim \mathrm{t}_{n-2}$$

Un aspecte destacat de la inferència que es farà és determinar si es pot rebutjar que β_1 és diferent de 0 (perquè si fos 0, no hi podria haver una relació lineal entre X i Y).

Nota. Observeu que els graus de llibertat són "n–2", ja que es disposa de n dades i s'han imposat dues restriccions. (Vegeu les equacions que hem anomenat normals a l'annex 1.)

Exemple. En una mostra de 50 casos sobre les hores d'estudi i la nota obtinguda en una assignatura (en una escala de 0 a 10), s'han obtingut els resultats següents:

$$b_1 = 0.12 \text{ punts/hora}$$
$$\sum \left(X_i - \bar{X}\right)^2 = 1045 \text{ hores}^2$$
$$s^2 = 2.25 \text{ punts}^2$$

Calculem un IC al 95% de confiança per al pendent:

$$\begin{aligned} \mathrm{IC}(\beta_1,\ 95\%) &= b_1 \pm t_{n-2,\alpha/2} \cdot s_{b_1} = \\ &= b_1 \pm t_{n-2,\alpha/2} \cdot [s^2 / (n-1)\, s_X^2]^{1/2} = \\ &= b_1 \pm t_{n-2,\alpha/2} \cdot [s^2 / \Sigma\, (X_i - \bar{X})^2]^{1/2} \approx \\ &\approx 0.12 \pm 1.97\, [2.25/1045]^{1/2} \approx \\ &\approx 0.12 \pm 0.0914 \approx \\ &\approx [0.0286,\ 0.2114] \end{aligned}$$

Conclusió. Amb confiança del 95%, β_1 és un possible punt de l'interval [0.0286, 0.2114], i la interpretació és que l'autèntic increment en el nombre de punts per hora d'estudi és algun dels valors compresos entre 0.0286 i 0.2114 punts per hora d'estudi. Penseu que aquest interval és bastant ample (un optimista pot pensar que en traurà un bon profit estudiant 10 hores; un pesimista, que tres dècimes no valen la pena).

D'altra banda, i com es veurà més endavant, convindria estudiar si es compleixen les premisses. La de linealitat estaria especialment sota sospita (representa el mateix benefici en nota passar d'estudiar de 0 a 1 hora, que passar d'estudiar de 10 a 11 hores?).

Exercici. Trobeu un IC al 95% per al pendent de la recta que modela la durada de la bateria en funció de la brillantor de la pantalla, sabent que s^2 val 227.30.

Solució. Cada punt de brillantor més pot suposar entre 10 min 35 s i 18 min 15 s menys de temps de treball.

b) Distribució de l'estimador b_0 de la constant

Els detalls de la demostració es troben a l'annex 3 del capítol. Cal subratllar també que la variable b_0 seguirà, al llarg de les possibles mostres, una distribució normal.

Es conclou que l'esperança i variància de l'estimador valen:

$$\mathrm{E}(b_0) = \beta_0$$
$$\mathrm{V}(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)\,s_X^2}\right]$$

Per tant,

$$b_0 \sim \mathrm{N}(\beta_0,\ \sigma^2(1/n + \overline{X}^2 / (n-1)\,s_X^2))$$

i

$$t_{b_0} = (b_0 - \beta_0) / \sqrt{[\ s^2(1/n + \overline{X}^2 / (n-1)\,s_X^2)\ ]} \sim \mathrm{t}_{n-2}$$

L'expressió de variància de l'estimació de la constant subratlla la font de fluctuació que pot experimentar l'estimador: d'una banda, la variabilitat de la mitjana de la resposta (s^2/n) i, d'altra banda, la variabilitat de l'estimació del pendent, ($s^2 / (n-1)s_X^2$), unit al fet que com més gran sigui la mitjana de X més s'amplifica l'error del pendent, cosa que dóna com a resultat una estimació b_0 més dispersa.

Exemple. Amb les dades de les pantalles d'ordinador, provem d'estimar la variància de b_0 (com que no coneixem σ^2, hem de considerar que tan sols disposarem d'una aproximació de $\mathrm{V}(b_0)$: dit amb altres paraules, estimem la variància de l'estimador del terme independent):

$$s_{b_0}^2 = \hat{\mathrm{V}}(b_0) = s^2\left(\frac{1}{n} + \frac{\overline{X}^2}{(n-1)\,s_X^2}\right) = 227.30\left(\frac{1}{10} + \frac{5.5^2}{(10-1)\,3.028^2}\right) = 106.05$$

L'arrel quadrada val 10.30; per tant, l'interval de confiança serà:

$$\mathrm{IC}(\beta_0, 95\%) = b_0 \pm t_{n-2,\alpha/2} \cdot s_{b_0} = 239.9 \pm 2.306 \cdot 10.30 = [216.1, 263.6] \text{ min.}$$

La intersecció a l'origen és algun punt en l'interval (3 h 36 min, 4 h 23 min), amb confiança del 95%.

c) Variància residual

s^2 es pot representar també com a s_R^2 (de *residual*) o s_e^2 (d'*error*). L'estimador de la variància residual és la mitjana dels errors al quadrat, tenint en compte que, com que prèviament s'han estimat els dos paràmetres de la recta, els graus de llibertat que han de figurar al denominador són n–2:

$$s^2 = \Sigma e_i^2 /(n–2) = \Sigma (y_i – \hat{y}_i)^2 / (n–2)$$

Encara que no es demostri formalment, s'ha de tenir present que si:

$$[y_i – \mathrm{E}(y_i)] / \sigma \sim \mathrm{N}(0,1)$$

llavors s'ha d'esperar que:

$$\Sigma (y_i – \hat{y}_i)^2 / \sigma^2 = \Sigma e_i^2 / \sigma^2 \sim \chi^2_{n–2}$$

Tenint en compte que l'esperança d'una distribució khi quadrat correspon als seus graus de llibertat:

$$\mathrm{E}(\Sigma e_i^2 / \sigma^2) = n–2$$
$$\mathrm{E}(\Sigma e_i^2 / (n–2)) = \sigma^2$$

es té que s^2 es distribueix també sense biaix.

Exemple. Trobem com s'ha estimat el valor de la variància residual utilitzat a l'exemple de la pantalla de l'ordinador. Necessitem calcular, per a cada observació, el valor predit amb la recta, i d'aquí derivar-ne els errors e_i:

X_i	1	2	3	4	5	6	7	8	9	10
y_i	241	193	205	169	174	134	163	124	111	92
Y_i	225.45	211.04	196.63	182.22	167.81	153.39	138.98	124.57	110.16	95.75
e_i	15.55	–18.04	8.37	–13.22	6.19	–19.39	24.02	–0.57	0.84	–3.75
e_i^2	241.7	325.5	70.1	174.7	38.4	376.1	576.9	0.3	0.7	14

Si sumem tots els valors de la darrera fila, obtindrem 1818.4, que dividit per 8 dóna s^2=227.30, el valor que havíem utilitzat. Recordem que l'arrel quadrada de s^2 (15.1) és una estimació de la desviació tipus de la fluctuació al voltant de la recta: és a dir, podem esperar com a típiques variacions de quinze minuts respecte de les previsions que ens doni el model lineal. I veurem que l'anàlisi dels errors ens pot ajudar a validar les condicions que fonamenten la seva aplicació.

6.4 Descomposició de la variabilitat

El model lineal combina una part determinista amb una part aleatòria. Suposem que s'ha d'endevinar el valor que prendrà la variable Y en l'observació següent, i oblidem-nos de moment de X. Com que l'atzar intervé en la resposta, no és possible encertar, però té sentit utilitzar μ (o, en el seu cas, $\overline{Y}$)per

predir el valor de l'observació següent de Y perquè és una opció que, a la llarga, minimitza el valor esperat de l'error quadràtic —entenent com a error la diferència entre la predicció i l'observació.

Ara bé, una vegada establerta una relació lineal entre X i Y amb els coeficients de la recta b_0 i b_1, aquesta predicció es pot millorar utilitzant, en lloc de la mitjana, el valor predit per la recta $\hat{y}_i$. Quant disminueix l'error de predicció? Es pot demostrar que aquest error disminueix i no augmenta? A continuació, es descompon la variabilitat observada de la desposta Y en dues variabilitats, una explicable pel valor de la variable independent X i l'altra, la variabilitat residual, sense explicació (el seu origen és incert).

Vegeu aquesta idea a la figura 6.3. S'ha dit que la millor predicció, sense tenir en compte el valor de X, és la mitjana. Es pot representar per una línia horitzontal, que no requereix pendent. Si afegim a aquesta recta horitzontal el paràmetre del pendent, ja es té una recta de regressió, que se simbolitza per la predicció $\hat{y}_i$ —que sí que té en compte X.

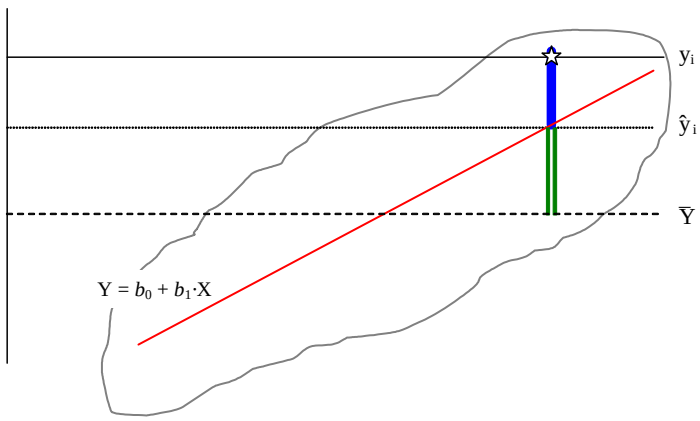


Figura 6.3 Esquema de la descomposició de la variabilitat de Y: la distància entre l'observació y_i i la predicció més simple $\overline{Y}$ té una part on X intervé (segment doble: $\hat{y}_i - \overline{Y}$) i una part impredecible (segment gros: $y_i - \hat{y}_i$).

Centreu l'atenció en l'observació y_i. L'error de predicció si s'utilitza la mitjana $\overline{Y}$ serà:

$$y_i - \overline{Y}$$

Mentre que, si s'utilitza X_i, obtenim el valor $\hat{y}_i$ predit per la recta, i el nou error serà:

$$y_i - \hat{y}_i$$

El que haurà millorat la predicció en aquesta observació serà:

$$\hat{y}_i - \overline{Y}$$

En resum:

$$\text{error original} = \text{error amb la recta} + \text{error explicat}$$
$$y_i - \overline{Y} \quad = \quad y_i - \hat{y}_i \quad + \quad \hat{y}_i - \overline{Y}$$

Les sumes de quadrats (*SQ*) s'obtenen elevant al quadrat i sumant per a totes les observacions:

$$\Sigma (y_i - \overline{Y})^2 = \Sigma (y_i - \hat{y}_i)^2 + \Sigma (\hat{y}_i - \overline{Y})^2$$
$$SQ_{Total} = SQ_{Residual} + SQ_{Explicada}$$
$$SQ_T = SQ_R + SQ_E$$

Prèviament, es demostra (v. annex 4) que el sumatori dels productes creuats, $\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \overline{Y})$, s'anul·la.

Per tant, es pot dir que la variabilitat de la desposta *Y* en una mostra de *n* observacions es descompon com a suma d'una variabilitat explicada —per *X*— i una variabilitat residual, no explicada, o d'origen incert.

Ara ja estem en condicions d'estudiar, per a una certa resposta *Y*, quina part de la variabilitat s'atribueix a una funció determinista i quina part no es pot predir. A continuació, es proposa una mesura sobre el grau de determinisme del model. Abans, ordenarem tots els càlculs necessaris per fer aquesta descomposició de la variabilitat en forma de taula.

No és difícil demostrar que:

$$\hat{y}_i - \overline{Y} = b_1 (X_i - \overline{X})$$

Llavors:

$$SQ_E = \Sigma (\hat{y}_i - \overline{Y})^2 = b_1^2 \Sigma (X_i - \overline{X})^2 = b_1^2 (n-1) s_X^2 = b_1 (s_{XY}/s_X^2)(n-1)s_X^2 = b_1 (n-1) s_{XY}$$

Això significa que, a partir de les covariàncies (o de la variància de les variables explicatives) i del pendent, es pot calcular $SQ_{Explicada}$. Com que SQ_{Total} també és fàcil d'aconseguir a partir de la variància de *Y*, $SQ_{Residual}$ es pot calcular per diferència, sense necessitat de fer les prediccions i calcular els residus. Els graus de llibertat totals són *n*–1, i els corresponents a la variabilitat explicada 1, perquè solament és rellevant el pendent: per tant, corresponen a *n*–2 els graus de llibertat per la part residual. Els càlculs a la taula de l'anàlisi de la variància (ANOVA, de l'anglès *analysis of variance*) són a la taula 6.1.

Taula 6.1 Taula ANOVA del model lineal simple. Les abreviatures entre parèntesis són termes equivalents en anglès (sum of squares, degrees of freedom, mean squares)

Font	*SQ* (*SS*)	*GdL* (*DF*)	*QM* (*MS*)	Raó
Explicada	$\Sigma (\hat{y}_i - \overline{Y})^2 = b_1^2 (n-1) s_X^2 = b_1 (n-1) s_{XY}$	1	SQ_E / GdL_E	$F = QM_E / QM_R$
Residual	$\Sigma (y_i - \hat{y}_i)^2 = SQ_T - SQ_E$	$n-2$	SQ_R / GdL_R	
Total	$\Sigma (y_i - \overline{Y})^2$	$n-1$		

*Nota***.** El quocient entre *SQ* i els graus de llibertat (*GdL*) no rep el nom de *variància*, sinó de *quadrats mitjos* (*QM*). Sí que es tracta d'una variància en el cas dels residus, ja que QM_R estima directament la variància residual, perquè equival a s_e^2. Però per a QM_E no hi ha cap variable aleatòria associada (precisament, perquè aquí no hi intervé l'atzar).

S'ha de tenir present que, com més gran sigui QM_E, més indici tenim que la recta de regressió aporta informació, i com més gran sigui el denominador $QM_R=s^2$, més gran és el soroll que té l'estimació.

Sota la hipòtesi H_0 en què el pendent és nul ($\beta_1=0$), és a dir, si la recta no aporta res, el quocient *F* segueix una distribució F de Fisher, amb 1 i *n*–2 graus de llibertat (sempre podem esperar una petita contribució deguda a un pendent estimat de magnitud insignificant; per això, sota H_0, la raó pot ser més gran que zero).

Així doncs, s'han vist dos estadístics per posar a prova la hipòtesi que el pendent β_1 de la recta és nul: el quocient *F* i l'estadístic t_{b_1} de l'apartat 6.3.1.*a*. De fet, l'alumne interessat pot comprovar que, algebraicament, $F = (t_{b_1})^2$, que determina la identitat total de les dues conclusions (per exemple, el mateix p-valor).[16]

6.5 Coeficient de determinació

Com que les *SQ* totals es descomponen en les que s'expliquen per la part explicada del model i la part residual, ara es pot quantificar la capacitat de previsió del model. La pregunta és: de tota la variabilitat de les *Y*, quina part ve associada (és a dir, explicada, atribuïda, deguda) a la variable *X*? La forma més simple de quantificar-ho és definint el *coeficient de determinació*, que equival a la proporció existent entre la suma de quadrats explicada per *X* respecte a la suma de quadrats total:

$$R^2 = SQ_E \,/\, SQ_T = \\ = \Sigma (\hat{y}_i - \overline{Y})^2 \;/\; \Sigma (y_i - \overline{Y})^2$$

Alternativament:

$$R^2 = SQ_E \,/\, SQ_T = \\ = (SQ_T - SQ_R) \,/\, SQ_T = \\ = 1 - SQ_R / SQ_T$$

Així, com més gran és el valor de R^2, millor representa el model lineal la relació entre *X* i *Y*. En el cas que el model sigui perfectament determinista, la *SQ* residual es fa zero, i R^2 aconsegueix el valor màxim d'1. En la situació contrària, en què *X* és indiferent a la variació de *Y*, la *SQ* explicada serà zero i el R^2 prendrà el seu valor mínim de 0.

16 Recordeu que el quadrat d'una variable amb distribució t de Student, amb *n*–2 graus de llibertat, és una variable amb distribució F de Fisher, amb 1 i *n*–2 graus de llibertat (v. apartat 5.5.1 del capítol anterior).

Nota. El coeficient de determinació per a una regressió lineal simple equival al quadrat del coeficient de correlació lineal:

$$R^2 = r_{XY}^2$$

Tingueu present que la correlació és una mesura descriptiva del grau d'associació lineal entre dues variables aleatòries, que és un altre punt de vista diferent del de R^2.

És important remarcar que R^2 sempre s'ha d'interpretar tenint en compte exclusivament una hipotètica relació lineal entre Y i X. Dit d'una altra manera: les dues variables poden estar estretament relacionades entre si —i, per tant, es pot fer una bona previsió de Y amb X— mitjançant un model *no lineal* i, no obstant això, tenir un coeficient R^2 quasi nul. En aquest cas, no s'ha de dir que les dues variables no estan relacionades, sinó que un model lineal no és apte per predir la resposta (v. figura 6.4).

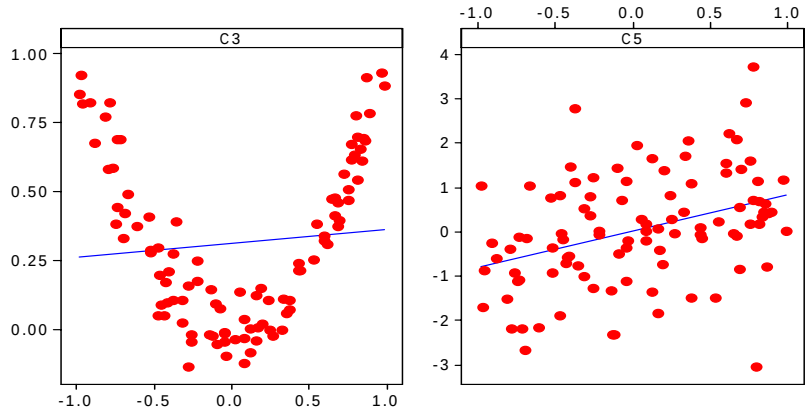


Figura 6.4. A l'esquerra, dues variables molt relacionades, però amb R^2 *petit; a la dreta, dues variables que també tindran un coeficient de determinació pobre però que sí que mostren un comportament lineal.*

Vegeu a la figura 6.5 uns exemples on es mostra quina forma (sobretot, quina dispersió al voltant d'una línia recta) pot prendre un núvol de punts per a diferents graus del coeficient R^2.

Nota. Utilitzant els quadrats mitjos *QM* en lloc de les sumes de quadrats *SQ*, es defineix el coeficient de determinació corregit o ajustat pels graus de llibertat com

$$R^2_{\text{corregit}} = 1 - QM_R / QM_T = 1 - [\Sigma (y_i - \hat{y}_i)^2/(n-2)] / [\Sigma (y_i - \overline{Y})^2 /(n-1)]$$

Aquest coeficient de determinació és útil en el cas que hi hagi un nombre reduït d'observacions, una situació a la qual R^2 és sensible, i que pot conduir a una mala interpretació.

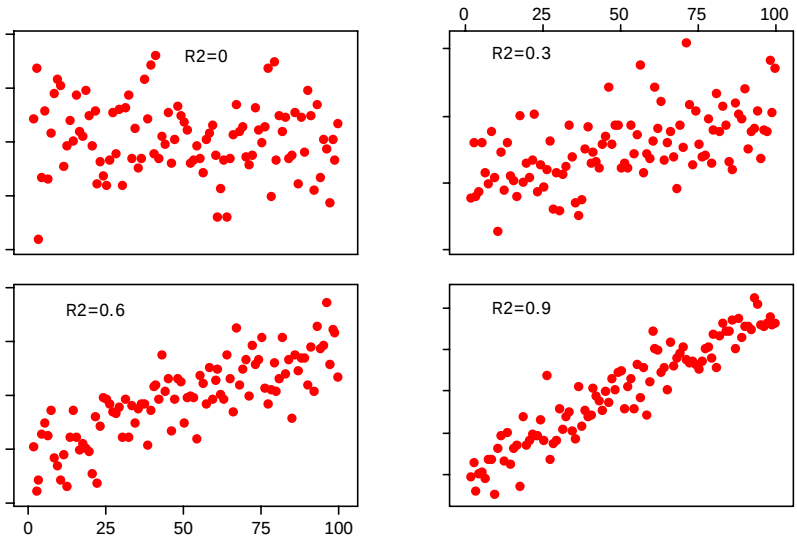


Figura 6.5 Quatre exemples de 100 observacions cadascun, amb coeficients de determinació de 0, 0.3, 0.6 i 0.9, respectivament (els valors de les X són els mateixos sempre)

Exemple. Fem la descomposició de la variabilitat de la durada de la bateria de l'ordinador i obtenim el coeficient de determinació per respondre a la pregunta: quina part de la variació de la durada de la bateria s'explica per l'ajustament de la brillantor de la pantalla? No fem ús del càlcul anterior de la variància residual, perquè volem mostrar que hi ha una opció més senzilla.

$$SQ_T = (n-1)s_Y^2 = 9 \cdot 45.9^2 = 18954.4$$
$$SQ_E = b_1^2(n-1)s_X^2 = (-14.412)^2 \cdot 9 \cdot 3.028^2 = 17135.7$$
$$SQ_R = SQ_T - SQ_E = 1818.7$$

I, per tant, la variància residual és 1818.7/8 = 227.34, com havíem obtingut anteriorment.

Font de variació	*SQ* (*SS*)	*GdL* (*DF*)	*QM* (*MS*)	Raó
Explicada	17135.7	1	17135.7	75.37
Residual	1818.7	8	227.3	
Total (o total corregida)	18954.4	9		

La raó *F* obtinguda (75.37) és un valor molt gran per a una F de Fisher amb 1 i 8 graus de llibertat, cosa que significa que la brillantor influeix, sens duote, en la durada de la bateria. L'arrel quadrada d'aquest valor (8.68) és el resultat de l'estadístic:

$$|t_{b_1}| = |b_1| / \sqrt{(s^2 / (n-1)s_X^2)},$$

on b_1 té signe negatiu perquè la relació entre *X* i *Y* és clarament inversa.

Finalment, es calcula, almenys, un dels dos coeficients de determinació:

$$R^2 = SQ_E / SQ_T = 17135.7 / 18954.4 = 0.904 = 90.4\%$$
$$R^2_{\text{corregit}} = 1 - 227.3 / 2106.04 = 0.892 = 89\%$$

Podem finalitzar dient que resta encara un 10%, aproximadament, de variabilitat de la durada que no depèn de la brillantor de la pantalla sinó d'altres factors (ús desigual d'altres recursos de l'ordinador que consumeixen energia...).

6.6 Descomposició de la variabilitat amb un factor qualitatiu (ANOVA)

Hi ha ocasions en què es vol comparar entre si un nombre determinat de grups, usualment no gaire gran, per a cada un dels quals tenim generalment bastants observacions. És una situació que es diferencia notablement del cas apte per regressió, on el supòsit típic és una sola observació per a cada valor X_i. No obstant això, la distinció principal és que, en aquelles ocasions, el factor X no pren valors numèrics sinó el que en diríem *classes* o *categories*.

Llavors, no té sentit plantejar una regressió lineal, perquè no podrem concloure amb una interpretació racional de β_1: no es pot dir que la resposta tindrà un increment esperat de β_1 quan X s'incrementi en una unitat (què vol dir?). Tampoc no és una bona solució assignar arbitràriament valors numèrics a cada grup. Encara que es podria estimar un model amb els valors nous, el resultat seria igualment arbitrari. No obstant això, independentment de com prenguem X, sí que ens interessa comprovar si la relació entre X i Y és nul·la: això suposaria que tots els grups serien iguals, de mitjana.

Exemple. Es vol comprovar si els tres grups d'una assignatura obtenen qualificacions similars. Formalment, té interès examinar la hipòtesi següent:

$$H_0: \mu_{10} = \mu_{20} = \mu_{30},$$

i es disposa de la informació següent:

Grup	Nombre d'alumnes	Mitjana	Desviació tipus
10	32	6.15	1.8
20	28	5.73	1.5
30	25	5.48	2.0

No cal dir que el nom dels grups podria ser A, B i C, i que la seva representació numèrica no li confereix cap valor quantitatiu. De fet, si calculéssim la recta de regressió (que sí que és possible), no podríem interpretar els estimadors trobats. Posem, a continuació, les mesures necessàries per estimar la recta (les "mitjanes" es proposen com a exercici per al lector):

"Mitjana" de grup	19.176	"Desv. tipus" de grup	8.196	Covariància	–2.2736
Nota mitjana	5.8146				

Amb això deduïm que la recta estimada és nota = 6.464 – 0.0338· grup, que no vol dir que si es crea el grup 21 se li prevegi una nota mitjana tres centèsimes menor que al grup 20!

En qualsevol cas, es vol determinar si H_0 és creïble sense fer ús d'una variable *X* quantitativa, ja que el model lineal utilitza l'ordre de les classes segons un ordre numèric, quan realment aquesta relació d'ordre entre els grups pot no existir. Imagineu que s'intercanvien el 20 i el 30: tindríem una altra recta! I el resultat no pot dependre d'això.

En lloc d'aprofitar la via de la recta, provarem de mirar el problema des d'un altre punt de vista, que ja hem vist amb el model lineal. Es tracta de calcular la descomposició de la variabilitat de la resposta *Y* en una part explicable pel factor i una part deguda a fluctuacions aleatòries i, per tant, no associada a *X*. Si s'arriba a produir un canvi notable entre la variabilitat inicial i la de la part d'origen desconegut, el factor contribueix a explicar aquest descens i, per tant, les poblacions de cada grup són diferents. En aquest plantejament, no hi ha cap necessitat d'introduir un model lineal si es fan unes poques modificacions per tal d'adequar-les al cas present.

En primer lloc, suposarem que tenim *K* grups, o poblacions, de les quals s'han observat *n* individus per a cada grup, de forma independent a cada mostra i també entre mostres. Denominem:

y_{ij}	$i = 1...\ n,\ j = 1...\ K$:	observació de l'individu *i* al grup *j*
$\overline{Y}_j$	$j = 1...\ K$:	mitjana al grup *j*
$\overline{\overline{Y}}$		gran mitjana, o mitjana de totes les observacions

Llavors, les sumes de quadrats que apareixien a la descomposició de la variabilitat en el model lineal es transformen com es veu a la taula 6.2. Hem de tenir en compte que la previsió per a cada observació ve donada per la mitjana del grup al qual pertany l'observació.

Taula 6.2 Comparació de les sumes de quadrats per model lineal i model ANOVA

Variabilitat	Model lineal	Model ANOVA
Explicada SQ_E	$\Sigma\ (\hat{y}_i - \overline{Y}\ )^2$	$\sum_{j=1}^{K}\sum_{i=1}^{n}\left(\overline{Y}_j - \overline{\overline{Y}}\right)^2 = n\sum_{j=1}^{K}\left(\overline{Y}_j - \overline{\overline{Y}}\right)^2$
Residual SQ_R	$\Sigma\ (y_i - \hat{y}_i)^2$	$\sum_{j=1}^{K}\sum_{i=1}^{n}\left(y_{ij} - \overline{Y}_j\right)^2 = (n-1)\sum_{j=1}^{K} s_j^2$
Total SQ_T	$\Sigma\ (y_i - \overline{Y}\ )^2$	$\sum_{j=1}^{K}\sum_{i=1}^{n}\left(y_{ij} - \overline{\overline{Y}}\right)^2$

Tinguem present que, en aquest cas i només si la grandària és la mateixa a les *K* mostres, la gran mitjana també és la mitjana de les *K* mitjanes. Així, SQ_E també es pot escriure com $n(K-1)s^2_{\overline{Y}}$, ja que el sumatori equival al numerador de la variància mostral de les mitjanes. Ara suposem que totes

les poblacions tenen una mitjana comuna μ: llavors, tindríem K estimacions independents d'aquesta mitjana, i la fluctuació natural d'aquestes mitjanes la coneixem: és σ^2 / n, on σ^2 és la variància de la resposta a qualsevol grup (d'acord amb la premissa que ja aplicàvem amb el model lineal i que seguim utilitzant aquí). Per tant, sota la hipòtesi nul·la d'igualtat de mitjanes, SQ_E estima la variància de Y, multiplicat per (K–1). Assumim que aquest és el nombre de graus de llibertat d'aquesta suma de quadrats.

D'altra banda, SQ_R és una combinació de les variàncies mostrals de cada grup, que també són diferents estimacions de σ^2. En aquest cas, el nombre de graus de llibertat és (n–1)·K, o n·K–K.

També en aquest cas, es pot demostrar que la suma de quadrats total és la suma dels dos termes descrits abans. Veiem que els graus de llibertat que hem assignat a cada part són coherents amb el que correspon a SQ_T:

$$GdL_T = GdL_E + GdL_R = K–1 + n·K–K = n·K–1 = N–1,$$

si denotem per N el nombre total d'individus que s'han observat en l'experiència (recordeu que SQ_T equival a $(N–1)s_Y^2$).

Es pot demostrar que, sota H_0, la ràtio F dels quadrats mitjos segueix una F de Fisher, com al model lineal, però amb K–1 i N–K graus de llibertat, respectivament, a numerador i denominador. Aquest resultat és vàlid fins i tot si les mostres no són de la mateixa mida per a cada grup. En aquest cas, més general, la taula de descomposició de la variabilitat apareix a la taula 6.3.

Taula 6.3 Taula ANOVA per a l'anàlisi d'igualtat de mitjanes en K poblacions

Font	SQ	GdL	QM	Raó
Explicada	$\sum_{j=1}^{K} n_j \left(\overline{Y}_j - \overline{\overline{Y}}\right)^2$	K–1	SQ_E / GdL_E	$F = QM_E \big/ QM_R$
Residual	$\sum_{j=1}^{K} (n_j - 1)s_j^2$	$N - K$	SQ_R / GdL_R	
Total	$\sum_{j=1}^{K} \sum_{i=1}^{n_j} \left(y_{ij} - \overline{\overline{Y}}\right)^2$	N –1	$\overline{\overline{Y}} = \sum_{j=1}^{K} n_j \overline{Y}_j \Big/ \sum_{j=1}^{K} n_j$	

Recordem que aquest model es fonamenta en unes condicions particulars:

- les mostres són MAS, i independents entre si
- la variable resposta és normal en cada grup
- la variable resposta té la mateixa variància σ^2 a qualsevol grup

Diversos autors (vegeu el llibre de Wonnacott i Wonnacott, *Introducción a la estadística*, Limusa) constaten que el model continua essent acceptable amb poblacions no normals i, fins i tot, si les variàncies presenten diferències, sempre que la mida de les mostres sigui aproximadament igual, cosa que demostra la fortalesa del mètode. De tota manera, sempre cal comprovar que les possibles desviacions de les premisses del model no siguin tan pronunciades com per posar en risc les conclusions obtingudes.

Exemple. Fem la descomposició de la variabilitat de les dades de l'exemple anterior. Suposem que, efectivament, són dades independents i que es pot suposar la distribució normal. Per les variàncies mostrals, ja podem comprovar que una mateixa variància poblacional per a tots els grups és possible.

Grup	$\overline{Y}_j - \overline{\overline{Y}}$	$n_j\left(\overline{Y}_j - \overline{\overline{Y}}\right)^2$	s_j	$(n_j - 1)s_j^2$
10	6.15–5.8146 = 0.3354	3.5998	1.8	100.44
20	5.73–5.8146 = –0.0846	0.2004	1.5	60.75
30	5.48–5.8146 = –0.3346	2.7989	2.0	96.00
		SQ_E = 6.599		SQ_R = 257.19

Font	*SQ*	*GdL*	*QM*	Raó
Explicada	6.599	2	3.300	1.052
Residual	257.19	82	3.136	
Total	263.789	84		

$$H_0: \mu_{10} = \mu_{20} = \mu_{30}$$

Suposant que aquesta hipòtesi és correcta, l'estadístic F que obtenim segueix una distribució F de Fisher, amb 2 i 82 graus de llibertat. El p-valor, calculat com P(F > 1.052) perquè l'ANOVA és una prova inherentment unilateral, és 0.354, que indica que no hi ha res que ens faci pensar que les notes mitjanes de cada grup hagin de ser diferents. Per tant, no es pot rebutjar la hipòtesi nul·la. De les dades anteriors, també deduïm que l'estimació de la variància residual és 3.136 (la desviació tipus estimada seria 1.77). Admetent que la nota mitjana per als tres grups és única, es pot estimar aquesta nota amb l'interval de confiança següent:

$$\text{IC}(\mu, 95\%) = \overline{\overline{Y}} \pm t_{N-K,1-\alpha/2} \cdot \sqrt{QM_R / N} = 5.8146 \pm 1.9893 \cdot 0.192 = (5.432, 6.197)$$

Si haguéssim trobat que H_0 era rebutjable, es podria calcular un IC per la mitjana d'un grup determinat amb l'expressió següent, que es beneficia d'una estimació més robusta de la variància i d'un factor t amb més graus de llibertat que el que permetria el càlcul limitat a les dades del propi grup:

$$\text{IC}(\mu_j, 95\%) = \bar{Y}_j \pm t_{N-K,1-\alpha/2} \cdot \sqrt{QM_R / n_j}$$

Així, el grup 10 obtindria l'IC (5.527, 6.773), el grup 20 l'IC (5.064, 6.396) i el grup 30 l'IC (4.775, 6.185). A la figura 6.6 s'aprecia que l'estimació conjunta és força més precisa i que coincideix aproximadament amb la intersecció dels intervals parcials, tot i que aquest efecte és casual. El que sí que és habitual és que no hi hagi intersecció entre els IC quan el p-valor és petit.

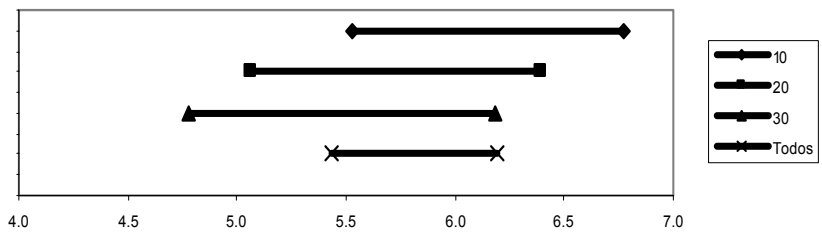


Figura 6.6 De dalt a baix: IC(μ_{10}, 95%), IC(μ_{20}, 95%), IC(μ_{30}, 95%) i IC(μ, 95%)

6.7 Recapitulació

Recordeu la màxima que diu: "Els models són per a fer-los servir, no per a creure-hi" Els models no són la veritat, sinó instruments útils per fer avançar el coneixement. Dir que la terra és una bola rodona de 6371 km de radi no és exactament cert, però aquesta aproximació és molt millor que no saber res. Un model sempre pot ser refinat a costa de perdre simplicitat: podríem adoptar per al model de la terra una geometria el·lipsoïdal, però llavors necessitarem dos radis en lloc d'un sol. De la mateixa manera, fem hipòtesis com que una variable aleatòria es distribueix com una normal (un altre model), o que la seva variància en dues poblacions és la mateixa. O que la mitjana d'una variable resposta varia linealment respecte dels valors que pren la variable *X*.

La ciència és útil perquè serveix per fer previsions. Però aquestes no poden venir solament d'un conjunt d'observacions. És necessari desenvolupar teories que ajudin a explicar el coneixement adquirit, i que puguin també proposar fets, possiblement no observats, per a ser verificats (a la introducció del llibre, ja mencionàvem aquest procediment, el *mètode científic*). Els models formals són imprescindibles per descriure adequadament teories consistents; d'altra banda, un model aplicat a una situació extrema però raonablement factible pot donar lloc a un fet per constatar. Però si la realitat contradiu una conseqüència derivada de la teoria, aquesta s'ha de qüestionar seriosament.

Exemple. Quan el sistema solar conegut acabava a Saturn, el model de Newton per explicar el moviment dels planetes funcionava molt bé per als primers planetes, però no encaixava del tot per a l'òrbita del darrer, a causa d'algunes pertorbacions observades amb el telescopi que no s'ajustaven a les equacions del model. Estenent la teoria a un hipotètic planeta situat més enllà, es va detectar Urà, que efectivament exercia una força gravitacional sobre Saturn que podia explicar les anomalies (posteriorment, es van trobar pel mateix mètode Neptú i Plutó).

Nota. Segons Popper, "el criteri per establir l'estatus científic d'una teoria és la seva refutabilitat o testabilitat", és a dir, una teoria no és millor com més confirmacions rebi, sinó si suporta els intents de refutar-la (de demostrar-ne la falsedat). Des del seu punt de vista, la ciència avança perquè parteix d'un problema, n'assaja possibles solucions i en rebutja les errònies (una teoria no ho pot explicar tot i, si ho fa, no és científica).

Així, hem presentat els contrastos d'hipòtesis no com a proves per confirmar una teoria (H_0) sinó com a intents per rebutjar-la trobant-hi evidències en contra. Els paràmetres que intervenen en un contrast, ja sigui per a una mostra, per a dues, per a K grups o per a un model lineal, són els mateixos que s'utilitzen al model matemàtic desenvolupat per a la situació exposada.

Ja hem vist un exemple de model per a dues poblacions:

$$Y_{mi} = \mu + \Delta_m + \varepsilon_{mi}$$

Plantegem que la i-èsima observació a la població m (Y_{mi}) és una combinació d'una hipotètica mitjana global (μ), més una correcció particular de la població (Δ_m), més una pertorbació aleatòria (ε_{mi}), per variabilitat dels individus o per errors de mesura. És usual imposar que la suma de les correccions sigui 0, ja que el nombre de paràmetres que realment hi intervenen és dos, no tres. La igualtat de mitjanes es podria plantejar, per tant, com la hipòtesi:

$$H_0: \Delta_1 = \Delta_2 = 0$$
$$H_1: \Delta_1 = -\Delta_2 \neq 0$$

Sota H_0, el model se simplificaria fins a quedar com: $Y_{mi} = \mu + \varepsilon_{mi}$: una distribució aleatòria i idèntica al voltant d'un mateix punt, per a qualsevol població. És a dir, les dues poblacions són una.

Des d'un altre punt de vista, H_0 proposa que tota la variabilitat té origen aleatori, inherent als individus observats. Per contra, la hipòtesi alternativa defensa un doble origen: el primer és indeterminista (l'atzar, com abans) i el segon no (la pertinença a una població o a l'altra).

Ara podríem afegir el model per a ANOVA, que és molt semblant a l'anterior:

$$Y_{ij} = \mu + \Delta_j + \varepsilon_{ij}$$

Per aquesta expressió comprenem per què es diu que ANOVA és un model lineal (per la forma additiva dels seus components). La figura 6.7 mostra dues situacions d'interès: quan no hi ha diferència real entre els grups respecte de la distribució de la resposta Y, o quan aquesta es distribueix de forma distinta —concretament, amb diferents mitjanes— segons el grup. El model ens condueix directament a expressar un contrast d'aquestes dues hipótesis, expressat de la manera següent:

$$H_0: \Delta_1 = \Delta_2 = \ldots = 0$$
$$H_1: \text{existeix alguna } j \text{ tal que } \Delta_j \neq 0,$$

que és en tot equivalent a la formulació, potser més habitual, expressada en termes de les mitjanes de cada grup (vegeu l'exemple de l'apartat 6.6).

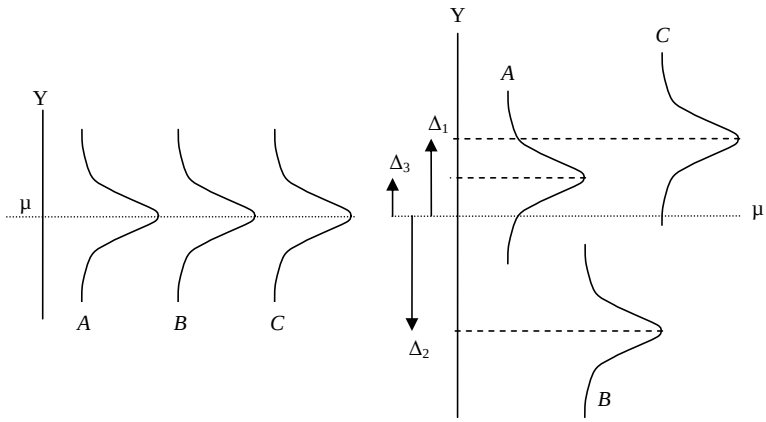


Figura 6.7 A l'esquerra, hipòtesi d'igualtat de mitjanes a tres poblacions; a la dreta, suposant diferents mitjanes. En el darrer cas, es pot veure que un individu de la població 2 té més probabilitats d'estar més distanciat de la mitjana global μ que si fos de les altres poblacions.

Què aporta un model lineal? Suposem que no tenim dues poblacions sinó un nombre qualsevol *n*. Suposem, a més, que cada població, o classe, està associada a un nombre escalar: X_1, X_2, ..., X_n (els valors es poden repetir). Llavors, es planteja si la distribució de la resposta *Y* es pot relacionar per mitjà d'una expressió algebraica (com l'equació d'una recta) amb els valors de *X*. Si parlem de model lineal, aquesta expressió és com:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

L'equació d'una recta s'expressa només amb dos paràmetres, el pendent β_1 i la intersecció a l'origen β_0, per a qualsevol nombre *n* d'observacions. ε_i continua essent una pertorbació que afecta la resposta. Notem que el model es pot transformar per tal que el terme independent representi una mitjana global per a *Y*, desplaçant les *X* una distància determinada:

$$Y_i = \mu + \beta_1 \cdot (X_i - \overline{X}) + \varepsilon_i$$

Per tant, continuant el paral·lelisme amb l'exemple anterior, es podria estudiar si el factor *X* està relacionat amb la resposta *Y* posant a prova si el pendent és nul: en haches cas, tota la variabilitat de la resposta seria d'origen individual, i no tindria res a veure amb l'existència de diferents poblacions relacionades amb *X*. Però si el pendent no és nul, ha d'existir alguna relació entre *X* i *Y*, i una part de la variabilitat de la resposta seria explicable pels valors que pren *X* (v. figura 6.8).

Hem de ser prudents quan estudiem un model estadístic que possiblement té associat un contrast d'hipòtesis típic. Per exemple, amb el model lineal, és habitual examinar la hipòtesi $\beta_1=0$, perquè aquesta és la forma de veure si el factor *X* està relacionat amb la desposta *Y*. D'aquí no s'ha de pensar que tot el model lineal està orientat al voltant d'aquesta hipòtesi, o que l'única cosa d'importància amb un model lineal és saber si el pendent és nul o no. Com en altres models, però potser en especial amb el lineal, és molt important estimar els paràmetres rellevants, ja que és el mitjà que permet utilitzar el model per fer previsions.

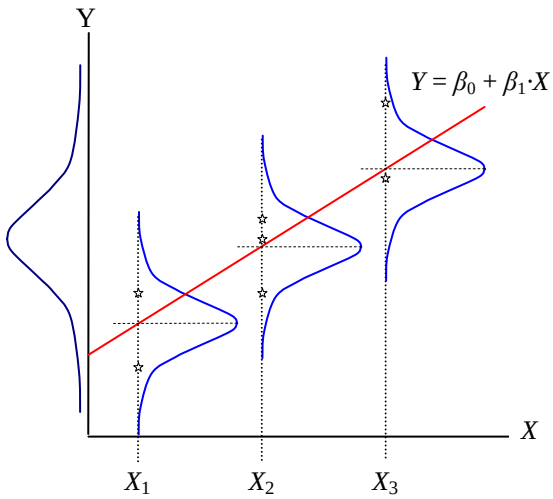


Figura 6.8 Esquema del model lineal amb un regressor X. La resposta Y presenta una variabilitat total (la distribució de l'esquerra, al marge de l'eix) que és realment una combinació de diferents Y observades en condicions X distintes, on s'aprecia una variabilitat menor (a la dreta, per X_1, X_2 i X_3).

Posem per cas, quina capacitat del disc dur tindran els ordinadors personals dintre de dos anys? Aquí, la qüestió de si "el pendent és nul" (és a dir, si la capacitat del disc canvia amb el temps o, més ben dit, amb els nous models) no necessita resposta: és clar que sí; cada vegada els discs són més grans. Però no podem dir simplement: "dintre de dos anys els discs seran més grans". Quant, quina capacitat serà la típica? Per respondre quantitativament es necessita: (1) trobar un model apropiat per descriure com augmenta la capacitat del disc amb el temps; (2) estimar els paràmetres del model, i (3) fer la previsió pròpiament dita.

Ens hem deixat un pas molt important. És bo el nostre model? Cal recordar que tot model és una simplificació i, per tant, suposa respectar una sèrie de condicions, sense les quals la credibilitat d'aquest se'n ressentiria. Per al model lineal, hem enunciat les següents:

- que en el rang estudiat la resposta segueixi un comportament lineal, de mitjana (i, per a previsions fora del rang, que sigui versemblant que la linealitat es pugui extrapolar;
- que les observacions obtingudes siguin independents, i
- que la resposta condicionada es distribueixi com una normal, amb la mateixa variància.

Per al model ANOVA, són les mateixes, a excepció de la primera, com és natural.

Aquestes parts fonamentals, la de la validació de les premisses i la de la previsió amb el model lineal, les veurem al capítol següent.

6.8 Formulari de model lineal

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \mathrm{V}(\varepsilon_i) = \sigma^2 \qquad s_X^2 = \frac{\sum \left(X_i - \overline{X}\right)^2}{n-1} = \frac{\sum X_i^2 - n\,\overline{X}^2}{n-1}$$

$$\hat{Y}_i = b_0 + b_1 X_i$$

$$b_0 = \overline{Y} - b_1 \overline{X} \qquad b_1 = \frac{s_{XY}}{s_X^2} \qquad s_Y^2 = \frac{\sum \left(Y_i - \overline{Y}\right)^2}{n-1} = \frac{\sum Y_i^2 - n\,\overline{Y}^2}{n-1}$$

$$s_R^2 = \frac{\sum e_i^2}{n-2}, \qquad e_i = Y_i - \hat{Y}_i \qquad s_{XY} = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)^2}{n-1}$$

Estim.	Esperança	Variància	Estimació de la variància	IC(95%)
B_0	$\mathrm{E}(b_0)=\beta_0$	$\mathrm{V}(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)s_X^2}\right]$	$s_{b_0}^2 = s^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)s_X^2}\right]$	$\left\lfloor b_0 \pm t_{n-2,\,0.975} \cdot s_{b_0} \right\rfloor$
B_1	$\mathrm{E}(b_1)=\beta_1$	$\mathrm{V}(b_1) = \frac{\sigma^2}{\sum \left(X_i - \overline{X}\right)^2}$	$s_{b_1}^2 = \frac{s^2}{\sum \left(X_i - \overline{X}\right)^2}$	$\left\lfloor b_1 \pm t_{n-2,\,0.975} \cdot s_{b_1} \right\rfloor$

$$b_0 \sim \mathrm{N}\left(\beta_0, \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum \left(X_i - \overline{X}\right)^2}\right)}\right) \Leftrightarrow \frac{b_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum \left(X_i - \overline{X}\right)^2}}} \sim \mathrm{N}(0,1) \quad \text{o bé} \quad \frac{b_0 - \beta_0}{s_R\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum \left(X_i - \overline{X}\right)^2}}} \sim \mathrm{t}_{n-2}$$

$$b_1 \sim \mathrm{N}\left(\beta_1, \sqrt{\frac{\sigma^2}{\sum \left(X_i - \overline{X}\right)^2}}\right) \Leftrightarrow \frac{b_1 - \beta_1}{\sigma\sqrt{\frac{1}{\sum \left(X_i - \overline{X}\right)^2}}} \sim \mathrm{N}(0,1) \quad \text{o bé} \quad \frac{b_1 - \beta_1}{s_R\sqrt{\frac{1}{\sum \left(X_i - \overline{X}\right)^2}}} \sim \mathrm{t}_{n-2}$$

Taula ANOVA per a *regressió lineal simple*

Font	*SQ*	*GdL*	*QM*	Raó
Explicada	$\Sigma\,(\hat{y}_i - \overline{Y})^2 = b_1^2(n-1)s_X^2 = b_1\,(n-1)s_{XY}$	1	$SQ_E\,/\,GdL_E$	$F = QM_E \big/ QM_R$
Residual	$\Sigma\,(y_i - \hat{y}_i)^2 \;=\; SQ_T \;-\; SQ_E$	$n-2$	$s^2 = SQ_R\,/\,GdL_R$	
Total	$\Sigma\,(y_i - \overline{Y})^2$	$n-1$		

$$\mathrm{R}^2 = SQ_E\,/\,SQ_T = r_{XY}^2$$

Taula ANOVA per a *un factor X qualitatiu* (comparació de *K* mostres independents)

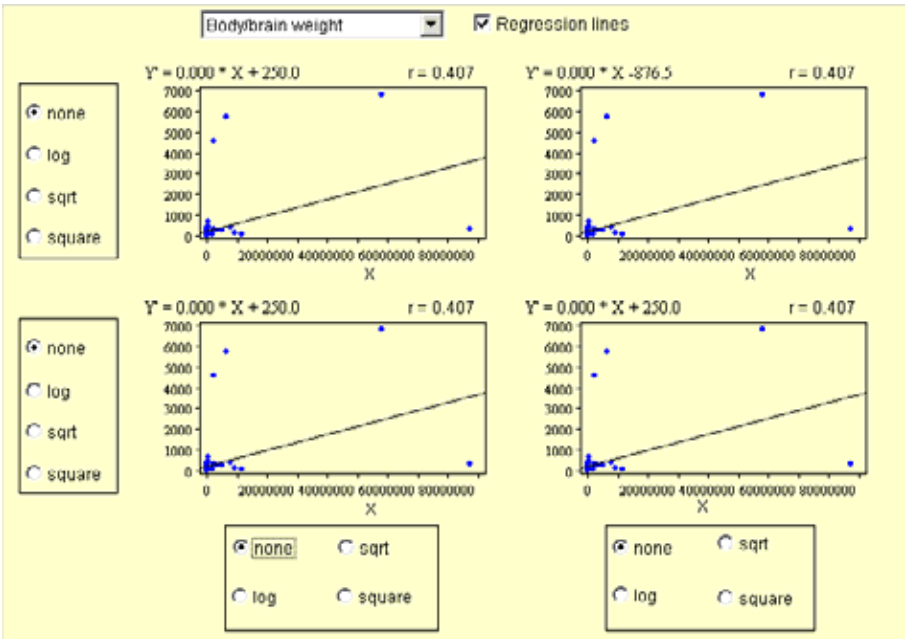
$Y_{ij} = \mu + \Delta_j + \varepsilon_{ij}$, j =1, ..., K, i=1, ..., n_j

Font	*SQ*	*GdL*	*QM*	Raó
Explicada	$\sum_{j=1}^{K} n_j \left(\overline{Y}_j - \overline{\overline{Y}}\right)^2$	K–1	SQ_E / GdL_E	$F = QM_E \big/ QM_R$
Residual	$\sum_{j=1}^{K} (n_j - 1) s_j^2$	$N - K$	SQ_R / GdL_R	
Total	$\sum_{j=1}^{K} \sum_{i=1}^{n_j} \left(y_{ij} - \overline{\overline{Y}}\right)^2$	N –1	$\overline{\overline{Y}} = \sum_{j=1}^{K} n_j \overline{Y}_j \Big/ \sum_{j=1}^{K} n_j$	

6.9 Guia de treball

L'objectiu d'aquest exercici és que practiqueu amb la interpretació dels resultats (estimadors de la recta, coeficient de determinació) d'una regressió lineal simple, i que tingueu un primer contacte amb les transformacions no lineals, que són un mètode habitual per tractar dades que no satisfan adequadament la premissa de linealitat.

Aneu a la pàgina web <http://www.ruf.rice.edu/~lane/stat_sim/transformations/index.html> i feu clic sobre el botó Begin. Veureu una finestra com la següent:



Aquest applet mostra l'efecte de les transformacions no lineals en la relació entre dues variables. Per tant, serveix per observar quina variació sembla més adequada per adaptar les dades a un model lineal. Podeu aplicar a *X* i *Y* dues operacions —logaritme, arrel quadrada, quadrat o res—, i veure i comparar l'efecte en els quatre diagrames de la finestra, on també veieu la recta estimada i el coeficient de correlació. Fixeu-vos que l'escala de l'eix corresponent canvia amb la transformació que hi apliqueu. Podeu escollir d'entre diversos conjunts de dades.

Qüestions

1. El primer conjunt de dades té com a *X* el pes corporal d'animals i com a *Y* el pes del cervell (ambdós en grams). Què observeu al gràfic? Creieu que el pes corporal és un bon predictor del pes del cervell?

2. Proveu de transformar amb el logaritme *X* i *Y*. Quina de les quatre combinacions dóna un coeficient de correlació més alt?

3. Quant explica el factor regressor en la variabilitat de la resposta, en el cas millor?

4. Quant val la recta i com s'interpreta d'acord amb els pesos considerats?

5. Escolliu el conjunt "Oil production". Veureu un gràfic amb l'evolució de la producció mundial de petroli (en milions de barrils anuals) des de 1880 fins als anys setanta. Proveu de transformar *X*. Val la pena?

6. Transformeu *Y* i doneu els resultats del model millor, i feu una predicció per a l'any 2000.

Respostes de la guia de treball

1. No s'aprecia bé la disposició dels punts: hi ha molts individus acumulats en els valors més petits. Els animals més grossos apareixen molt dispersos per l'àrea. És difícil apreciar-hi una tendència; en tot cas, els punts més allunyats del cúmul principal tenen una gran influència en la recta. El valor que hem estimat per a aquesta no sembla útil, per aquesta raó (la variabilitat dels coeficients és important). En conclusió, amb una correlació de només 0.4 ($R^2 \approx 16\%$), el pendent suggereix un increment inferior a un gram per quilogram de massa corporal.

2. L'opció millor és prendre el logaritme de *X* i de *Y*, clarament. La correlació puja fins a 0.911. Els altres models no semblen molt adequats, ja que no passen de 0.45.

3. Prenent el quadrat de la correlació, obtenim 0.8299: és a dir, el pes del cos explica un 83% de la variabilitat del pes del cervell.

4. La recta estimada és log $Y = 0.550 \cdot \log X - 1.9$, equivalent a $Y = 0.15 \cdot X^{0.55}$, un resultat que suggereix que el pes del cervell és proporcional a l'arrel quadrada de la massa total (es pot interpretar com que una gran quantitat de muscle no necessita molt volum cerebral).

   Si el pes s'incrementa un 100% (el doble), podem esperar que el cervell sigui solament un 46% més gran ($2^{0.55}$).

5. Sembla clar que la producció creix a un ritme molt ràpid amb el temps, fenòmen típic d'un creixement sense restriccions (els primers problemes van aparèixer a la dècada dels setanta). Transformar X no té sentit, ja que en estar l'origen —l'any 0— molt lluny, qualsevol transformació té un efecte quasi lineal, és a dir, gairebé nul. Compareu el resultat transformant X i sense transformar-la: són pràcticament el mateix.

6. En canvi, la transformació logarítmica de Y va molt bé, perquè el núvol de punts s'ajusta a una recta. Això indica que el creixement del consum de petroli ha estat de tipus exponencial en el temps (cada cert temps, el consum es duplica).

   En el nostre cas, tenim log $Y = 0.066 \cdot X - 121.3$, amb correlació quasi 1: 0.998 o, el que és el mateix, $Y = e^{-121} \cdot e^{0.066X}$. El cicle de duplicació del consum és aproximadament $\ln(2)/b_1$, o sigui, 10.5 anys. Segons aquest ritme de creixement, el consum per a l'any 2000 hauria estat de 44,356 milions de barrils. La producció real va ser d'uns 28,000 milions de barrils; la davallada és una conseqüència lògica del fort augment del preu del petroli. (Nota: en realitat, atès a la pèrdua de decimals, l'equació de la recta és una mica diferent, però la predicció autèntica és molt superior: 118,420 milions de barrils; aquest resultat emfatitza encara més l'impacte de la pujada del preu des de 1973.)

   Vegeu a la figura 6.9 com continua la sèrie fins a l'any 1984. Queda molt clar el risc de fer extrapolacions per obtenir una predicció, sense considerar esdeveniments que hi poden introduir alteracions radicals.

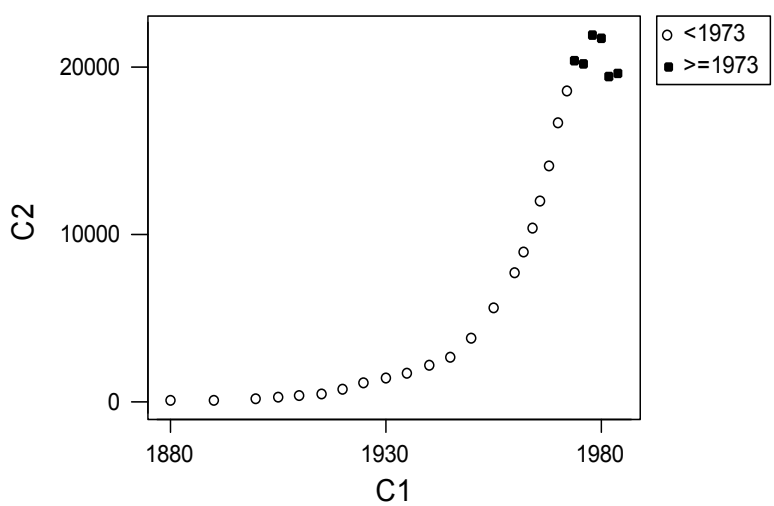


Figura 6.9

6.10 Problemes

1. Un grup d'estudiants d'informàtica es planteja com a objectiu per a una pràctica comparar la grandària del fitxer en kB segons el format d'emmagatzematge d'una imatge (jpeg o gif). Seleccionen un conjunt de 50 imatges i mesuren la grandària dels fitxers resultants en els dos formats (variables *kB–gif* i *kB–jpg*). Un dels components de l'equip, a la vista de l'alta correlació entre les dues variables, proposa fer l'estudi de la regressió lineal entre ambdues variables. Després de fer servir Minitab per analitzar les dades, un dels components del grup (una mica despistat) pren nota dels resultats següents:

Descriptive Statistics: Kb-Gif; Kb-Jpg

Variable	N	Mean	Median	StDev	SEMean	Minimum	Maximum	Q1	Q3
Kb-Gif	50	23,78	14,50	14,28	2,02	10,00	53,00	11,75	38,00
Kb-Jpg	50	28,78	18,00	17,09	2,4	28,00	67,00	14,00	45,00

Regression Analysis: Kb-Gif versus Kb-Jpg

Predictor	Coef	SE Coef
Constant	?	0,7007
Kb-Jpg	0,82313	0,02099

R-Sq = 97,0%

L'altre component del grup li diu que, per fer l'informe, és necessari donar més informació, però com que aquest sí que hi entén, és capaç de deduir els valors que falten.

a) Per completar l'etapa de l'estimació, doneu l'equació completa de la recta de regressió, l'estimació puntual de la variància dels residus, i construïu la taula de l'ANOVA de la regressió.

b) Plantegeu i resoleu el test corresponent a la taula de l'ANOVA de la regressió, i interpreteu-ne el resultat obtingut.

c) En cas que els dos formats fossin equivalents, la relació lineal seria $kB–gif = 0 + 1 \cdot kB–jpg$ Plantegeu i resoleu el test corresponent per decidir si podem acceptar que β_1 val 1. A quina conclusió arribeu?

d) El segon component de l'equip planteja fer l'anàlisi amb la t de Student per dades aparellades. Quins avantatges i/o inconvenients trobeu a fer servir una tècnica o l'altra per validar els objectius proposats?

2. Per valorar i controlar el cost d'un projecte informàtic, és molt important conèixer com i en què les persones implicades en el procés inverteixen el temps (Pera *et al.*, 1994, IEEE Software 11, núm.4). Un cap de projecte vol mesurar l'esforç d'un projecte (E) en hores a partir de la mètrica formada pel nombre de funcionalitats detallades en les especificacions del projecte, anomenada "punts funció" (F).

   Al llarg de 100 projectes que ha supervisat, ha recollit la informació que es presenta a la figura 6.10.

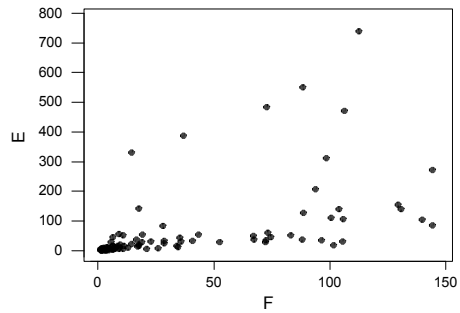


Figura 6.10

a) Creieu que té sentit assajar un model de regressió simple per explicar l'esforç *E* en funció dels punts funció *F*? Per què? Quines premisses s'han de complir? Podeu afirmar que es compleixen totes en aquest cas?

b) S'ha utilitzat la comanda de regressió implementada a Minitab i s'han obtingut els resultats que es mostren a continuació. Plantegeu i resoleu el test d'hipòtesi, amb un error del 5%, que permeti contrastar que el pendent d'aquesta recta de regressió val 1. Quina interpretació té el valor p=0,193 pel que fa a la constant d'aquest model? Creieu que la regressió és globalment significativa? Plantegeu i resoleu el test d'hipòtesi adient.

```
The regression equation is
lnE = 0,255 + 0,970 lnF

Predictor        Coef     SE Coef          T        P
Constant       0,2549      0,1943       1,31    0,193
lnF           0,97033     0,06475      14,99    0,000

S = 0,9654      R-Sq = 69,6%     R-Sq(adj) = 69,3%
```

3. Les dades que s'expliquen a continuació són el resultat d'una bona pràctica d'uns estudiants que varen comparar dos algorismes (*A*, *B*) per comptar el nombre d'aparicions d'una cadena de text. Per això, varen seleccionar a l'atzar 261 fitxers de text i varen executar simultàniament cada un d'ells amb els dos algorismes. A continuació, es mostra la descriptiva de la variable "diferència del temps de CPU, en centèsimes de segon, entre els algorismes *A* i *B*".

Descriptive Statistics

Variable	N	Mean	Median	StDev	SEMean	Min	Max	Q1	Q3
CPU_A-CPU_B	261	12.563	11.0	12.452	0.771	-32.0	53.0	4.0	21.0

a) A partir d'aquestes dades, és irrellevant quin algorisme utilitzar en el futur? Per ajudar-vos a contestar aquesta pregunta, calculeu el temps de CPU que penseu que es pot estalviar amb aquesta proposta en cada execució.

Segons els vostres companys, "les mostres que prenem sobre el temps d'execució dels nostres algorismes estaran sotmeses a d'altres factors que influiran en aquest temps". Concretament van estudiar com influeix el nombre d'usuaris sobre la diferència en el temps de CPU.

b) Completeu la taula.

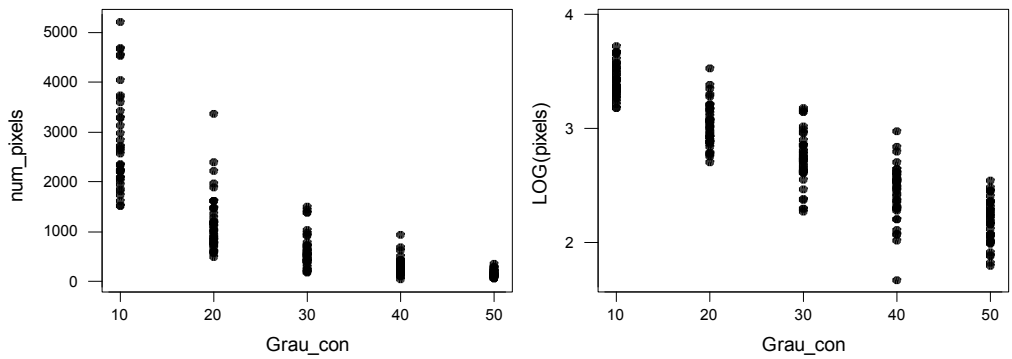
```
The regression equation is
CPU_A-CPU_B = - 0.16 + 0.181 USERS
Predictor       Coef       StDev          T        P
Constant      -0.159       9.647      ______   0.987
USERS         0.1813      0.1370       1.32    0.187
S = 12.43       R-Sq = ____
Analysis of Variance
Source       DF          SS          MS         F        P
Regression  ____      270.6      ______      ____    _____
Error       ____    _______      ______
Total       ____    _______
```

c) Poseu a prova i resoleu la hipòtesi sobre la relació entre la diferència *A–B* de temps de CPU i el nombre d'usuaris.

d) Interpreteu conjuntament els resultats de les preguntes 1 i 3.

4. Uns alumnes estudien un programa de tractament digital d'imatge. Es tracta d'un filtre artístic com els que es poden trobar en programes de dibuix com Adobe PhotoShop o CorelDraw. El filtre, a partir d'una imatge en RGB, en crea una altra redibuixant les parts més contrastades amb els colors pujats de to. S'aconsegueix un efecte artístic que pot simular el d'una imatge dibuixada amb plomí.

   El filtre actua seleccionant uns determinats píxels, justament aquells que es troben en un canvi de contrast, i dibuixa així només els perfils i els canvis d'intensitat de color. La selecció d'aquests píxels està parametritzada per una constant (*grau_con*), que ha d'indicar quin canvi d'intensitat de color és prou significatiu per seleccionar el píxel. L'objectiu de l'estudi és analitzar la relació que es pot establir entre diferents graus de contrast i el nombre de píxels seleccionats. La mostra que han utilitzat ha estat 200 fotografies dels estudiants de la Facultat d'Informàtica, obtingudes de la web. L'estudi s'ha basat en cinc graus de contrast diferents, i per a cada un d'aquests s'utilitza una mostra de 40 imatges.



a) A la figura superior, es presenten el gràfic del nombre de píxels, *num_pixels*, en funció del grau de contrast, *grau_con*, com també el logaritme del nombre de píxels, *LOG(píxels)* en funció del grau de contrast. Amb quina variable resposta, *LOG(píxels)* o *num_pixels*, aconsellaríeu treballar als vostres companys? Raoneu la resposta.

b) En primer lloc, volen veure si hi ha diferències significatives pel que fa als píxels en funció del grau de contrast i un d'ells proposa fer una ANOVA. Indiqueu quines hipòtesis s'han de complir. Plantegeu el test adient. Resoleu-lo (al quadre següent, trobareu tota la informació numèrica necessària).

Grau_con		**num_pixels**	
	N	Mean	StDev
10	40	2790,4	975,0
20	40	1190,3	573,8
30	40	626,6	343,1
40	40	316,6	169,5
50	40	163,8	72,5

Grau_con		**LOG(pixels)**	
	N	Mean	StDev
10	40	3,4217	0,1438
20	40	3,0348	0,1854
30	40	2,7372	0,2331
40	40	2,4401	0,2438
50	40	2,1731	0,1932

c) Responeu si és adient o no aquest procediment per a aquesta situació i raoneu la resposta.

L'altre company discrepa del procediment anterior i proposa fer una regressió per poder confirmar la relació entre les dues variables de l'estudi. Els resultats d'aquest procediment són els següents:

```
LOG(pixels) = 3,69 - 0,0309 Grau_con

Predictor        Coef     StDev          T       P
Constant      3,68902     0,03381      109,13    0,000
Grau_con    -0,030921     0,001019     -30,34    0,000

S = 0,2039      R-Sq = 82,3%     R-Sq(adj) = 82,2%
```

d) Creieu que la regressió és globalment significativa? Plantegeu i resoleu el test d'hipòtesi adient, omplint els resultats de la taula, i indiqueu els càlculs que realitzeu.

Source	DF	SS	MS	F	P
Regression					
Residual Error					
Total	199	46,472			

e) Quin dels dos companys creieu que té raó? Justifiqueu la resposta.

6.11 Solució dels problemes

1.

a)

$kB-gif = 0.0902 + 0.82313\ kB-jpg$
($b_0 = Y - b_1 X = 23.78 - 0.82313 \cdot 28.78 = 0.0902$)

$s_R = 2{,}511$
($s^2 = SQ_R/(n-2) = (n-1)s_Y^2(1-R^2)/(n-2) = 49 \cdot 14.282^2 \cdot (1-0.9697)/48 = 6.30$)
($s_{b_1}^2 = s_R^2 / [(n-1)s_X^2]$; $s_R^2 = (n-1)s_X^2 s_{b_1}^2 = 49 \cdot 17.086^2 \cdot 0.02099^2 = 6.30$)

```
Analysis of Variance

Source              DF      SS         MS         F         P
Regression (E)      1     9692.1     692.1     1537.77    0.000
Residual Error(R)   48     302.5       6.3
Total               49    9994.6
```

$DF_E = 1$
$DF_R = n - 2 = 48$
$DF_T = n - 1 = 49$
$SQ_T = (n-1)s_Y^2 = 49 \cdot 14.282^2 = 9994.6$

$SQ_E = R^2 \cdot SQ_T = 0.9697 \cdot 9994.6 = 9692.1$
$SQ_R = SQ_T - SQ_E = 9994.6 - 9692.1 = 302.5$
$MS_E = SQ_E / 1 = 9692.1$
$MS_R = SQ_R / (n-2) = 302.5/48 = 6.3 \ (= s_R^2)$
$F = MS_E / MS_R = 9692.1/6.3 = 1537.77$
$P = P(F_{1,48} > 1537.77) < 0.0001$

b)

H_0: $\beta_1 = 0$ ($\Leftrightarrow$Hi ha relació lineal significativa entre les dues variables)
H_1: $\beta_1 \neq 0$ ($\Leftrightarrow$No hi ha relació lineal significativa entre les dues variables)

Estadístic: $F = MS_E / MS_R = 1537.77$
Distribució: (sota H_0) $F_{1,48}$
Premisses: linealitat, normalitat, independència i homoscedasticitat dels residus
Regla de decisió: rebutjar H_0: si $F > F_{1, 48, 0.95}$ $\Leftrightarrow$ si p-valor$<\alpha=0.05$
Càlculs: $P(F_{1,48}>4.04)=0.05$. $F=1537.77>>4.04$ i p-valor$<<0.0001$
Decisió: es rebutja H_0

Conclusió pràctica: hi ha una relació lineal significativa entre les dues variables

c)

$$H_0: \beta_1 = 1$$
$$H_1: \beta_1 \neq 1$$

Estadístic: $t = (b_1 - 1) / s_{b_1} = (0.82313 - 1)/0.02099 = -8.42639$
Distribució: (sota H_0) t_{48}
Premisses: linealitat, normalitat, independència i homoscedasticitat dels residus
Regla de decisió: rebutjar H_0: si $|t| > t_{48, 0.975}$ $\Leftrightarrow$ si p-valor$<\alpha=0.05$
Càlculs: $P(|t_{48}|>2.01)=0.05$. $t=8.426>>2.01$ i p-valor$<<0.0001$
Decisió: es rebutja H_0

Conclusió pràctica: es rebutja el valor 1 per al pendent → hi ha diferències entre els formats.

d) La comparació mitjançant el test de la t de Student ens permet testar si el valor esperat de les diferències és zero, la qual cosa s'interpreta com que no hi ha diferències significatives entre els dos formats.

La regressió lineal permet construir un model matemàtic que, en cas de ser vàlid, ens permet també quantificar el grau de relació i fer prediccions de la variable resposta. Ara bé, la premissa que la variable independent és fixa, que no és v.a., no sembla gaire apropiada en aquest exemple (com a mínim és tan aleatòria com la resposta).

El model complet seria $Y_i = \beta_0 + \beta_1(X_i + \varepsilon_i) + \varepsilon_i$

2.

a)

Premisses:
En el model de regressió $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
s'han de complir les premisses següents:

- *Linealitat* entre X i Y en el rang considerat.
- ε_i són v.a.i.i.d.; $N(0,\sigma^2)$
- *Homoscedasticitat*: mateixa σ^2 i ε_i error additiu; no depèn del valor
- *Independència*: un error no aporta informació sobre el valor de l'altre
- *Normalitat*
- X_i no és v.a.; és constant; no està mesurada amb error

En aquest cas, no té gaire sentit ajustar un model de regressió entre aquestes dues variables ja que inspeccionant la figura 6.10 sembla que no hi hagi una relació lineal molt clara entre l'esforç *E* i els punts funció *F*. A més, a mesura que s'incrementen els punts funció, es produeix més variabilitat en l'esforç. La hipòtesi d'independència es pot pressuposar que es compleix, ja que aquestes dades provenen de 100 projectes que, en principi, són independents, i la hipòtesi de normalitat no es pot contrastar amb la informació de què es disposa.

b) Plantegem el test d'hipòtesi següent:

$$H_0: \beta_1 = 1$$
$$H_1: \beta_1 \neq 1$$

L'estadístic que s'ha d'utilitzar és $T = \dfrac{b_1 - (\beta_1)_0}{SE(b_1)}$, que segueix una distribució t_{n-2}.

En el nostre cas, totes les dades que ens fan falta les podem obtenir del quadre de l'enunciat.

$b_1 = 0.97033$	És el coef. per a ln*F*
$(\beta_1)_0 = 1$	És la hipòtesi que es vol contrastar
$SE(b_1)$	És la desviació tipus de l'estimador que es troba a *SE Coef*

i, per tant,

$$t = \frac{0.97033 - 1}{0.06475} = -0.458$$

Amb un nivell de significació del 5%, el límit inferior de la zona d'acceptació de H_0 és –1.96, ja que t_{98} es pot aproximar per una distribució *Z*. Com que –0.458 és més gran que –1.96, res no s'oposa a acceptar H_0. La conclusió pràctica és que podem acceptar que el pendent de la recta de regressió val 1.

Un altre procediment per resoldre aquest apartat és calcular l'interval de confiança al 95% per al paràmetre poblacional, que, concretament, en aquest cas, es calcula a partir de l'expressió:

$$b_1 \pm t_{98,0.975} \cdot SE(b_1) = 0.97033 \pm 1.96 \cdot 0.06475$$

i l'interval és (0.843, 1.097). Com que 1 pertany a aquest interval, podem acceptar H_0 amb una confiança del 95%.
(Nota. Els errors més típics en aquest apartat provenen del fet de no estar familiaritzats amb els resultats de la regressió que proporciona Minitab, la qual cosa indueix a fer càlculs addicionals que no calien.)

p=0.193 és el p-value associat al test d'hipòtesi

$$H_0: \beta_0=0$$
$$H_1: \beta_0\neq 0$$

que òbviament ens fa acceptar H_0, la qual cosa es tradueix en el fet que la constant no és significativa o, dit d'una altra manera, que la recta de regressió passa per l'origen de coordenades.

Omplim les dades de la taula :

```
Analysis of Variance

Source                DF       SS          MS         F          P
Regression             1     209.11      209.11     224.37     0.000
Residual Error        98      91.336       0.932
Total                 99     300.448
```

En primer lloc, pel que fa als graus de llibertat (*DF*), els totals són (*n*–1)=(100–1)=99, ja que s'han obtingut aquests resultats *al llarg de 100 projectes***.** Com que la regressió s'ha realitzat amb un sol regressor, *DF Regression* =1; per tant, *DF Residual Error* ha de ser 98.

A continuació, veurem les diferents possibilitats per tal d'obtenir les sumes de quadrats:

i) $R^2 = 1 - \dfrac{SQ_{\text{Residual}}}{SQ_T} = 0.696$ $\qquad$ $s^2 = \dfrac{SQ_{\text{Residual}}}{98} = 0.9654^2$

(valors que es troben al quadre de l'enunciat) $\Rightarrow$ $SQ_{Residual} = SQ_R = 91.332$ i $SQ_T = 300.448$.
Per tant, $SQ_{Regression}$, també denotat SQ_E=300.448–91.332=209.111

ii) $F_{1,n-2} = t^2_{n-2}$ i, en aquest cas, $F = 14.99^2 = 224.7$ però,

$F = \dfrac{QM_E}{QM_R} = \dfrac{SQ_E/1}{s^2}$, és a dir: $224.7 = \dfrac{SQ_E}{0.9654^2} = \dfrac{SQ_E}{0.932} \Rightarrow SQ_E = 209.11$

i $R^2 = \dfrac{SQ_E}{SQ_T} \Rightarrow SQ_T = 209.11/0.696 = 300.44$

En aquest cas, $SQ_R = SQ_T - SQ_E = 300.44 - 209.11 = 91.332$

I, per als més "avançats"

iii) $\hat{V}(b_1) = \dfrac{s^2}{\sum (X_i - \bar{X})^2} = \dfrac{s^2}{(n-1)s_X^2} \Rightarrow s_X^2 = 0.06475^2 \cdot 0.9654^2 = 2.2468$

$$b_1 = \frac{s_{XY}}{s_X^2} \Rightarrow s_{XY} = 0.97033 \cdot 2.2468 = 2.18$$

$$r_{X,Y} = \sqrt{R^2} = \frac{s_{XY}}{s_X s_Y} \Rightarrow s_Y = \frac{2.18}{\sqrt{2.2468}\sqrt{0.696}} = 1.74$$

Com que $SQ_T = (n-1)\, s_Y^2 = 99 \cdot 1.7432^2 = 300.8$,

$$SQ_E = b_1^2 (n-1) s_X^2 = 0.97033^2 \cdot 99 \cdot 2.2468 = 209.111$$

i, a partir dels resultats previs, s'obté:

$$SQ_R = SQ_T - SQ_E = 300.8 - 209.111 = 91.336$$

(Nota. Un error bastant habitual és confondre l'estimació de la desviació tipus de l'estimador de β_1, $SE(b_1)$, que és el valor que s'obté del quadre de l'enunciat amb la desviació tipus estimada de la variable X. Aquest problema és conseqüència de la interpretació incorrecta dels resultats que Minitab proporciona. Un altre possible error és confondre la desviació tipus estimada de les X amb la dels errors (s).)

Finalment, per comprovar que la regressió és globalment significativa, el procediment més fàcil és interpretar el p-value del coeficient de ln$F$ (0.000).

En aquest cas, el test d'hipòtesi adient és:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

i, com que el p-value és 0.000, rebutgem H_0; per tant, la regressió és significativa.

També ho podem comprovar a partir del valor F obtingut omplint la taula anterior. En aquest cas, arribem a la mateixa conclusió.

3.

a)

Estadístic: $T = \dfrac{D - \mu}{s_D / \sqrt{n}} \sim t_{261} \approx Z$

Premisses: · Independència de les D_i
· Normalitat
· MAS

Interval: α=0.05, $\mu_0 \in [12.563 \pm 1.96 \cdot 12.452/\sqrt{261}]$, $\mu_0 \in [11.0523, 14.0737]$
Així doncs, es pot esperar un estalvi en cada execució d'entre 11 i 14 cops.

b)

Predictor	Coef	StDev	T	P
Constant	–0.159	9.647	–0.0165	0.987
USERS	0.1813	0.1370	1.32	0.187
S=12.43 R–Sq=0.67%				

Source	DF	SS	MS	F	P
Màquina	1	270.6	270.6	1.75	0.187
Error	259	40041	154.6		
Total	260	40312.2			

c)

Plantejament formal: H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
Estadístic: $F = QM_E / QM_R \sim F_{1, 259}$
Premisses: MAS, linealitat, normalitat (sota H_0)
Càlcul: $p = 0.187 > 0.05 = \alpha$
Conclusió formal: No podem refusar H_0.
$\beta_1 \in [\, b_1 \pm t_{259,0.975} \cdot s_{b_1}\,] = [0.1813 \pm 1.96 \cdot 0.1370] = [-0.0872, 0.4498]$
El zero és un valor possible! No hem demostrat que $\beta_1 \neq 0$

d) Hi ha diferències que no varien segons el nombre d'usuaris.

4.

a) *LOG(píxels)* perquè mostra millor linealitat i homoscedasticitat.

b)

- Observacions independents a dins de cada grup i entre grups
- La variància dels grups ha de ser la mateixa
- Normalitat de les dades

Volem contrastar si es pot acceptar la hipòtesi que les mitjanes poblacionals a cada grup siguin les mateixes.

Formalment, volem contrastar:

$$\begin{cases} \mathrm{H}_0 : \mu_i = \mu_j \forall i, j \\ \mathrm{H}_1 : \exists i, j \mid \mu_i \neq \mu_j \end{cases}$$

on μ_i representa la mitjana poblacional del grup *i*, és a dir, la mitjana de la variable *LOG(píxels)* quan el grau de contrast és *i*.
Cal plantejar i resoldre la taula de l'anàlisi de la variància o ANOVA.

En aquest cas, treballarem amb les dades la variable *LOG(píxels)* segons el grau de contrast, per les raons comentades a l'apartat anterior.

Taula 6.4

	SQ: sumes de quadrats	*GdL*: graus de llibertat	*QM*: quadrats mitjans	Raó	Nivell de significació
Entre grups	$\sum_{j=1}^{K} n_j (\bar{y}_j - \bar{Y})^2 = 38.246$	$K-1=4$	$SQ_E / GdL_E = 9.6066$	$\hat{F} = \frac{SQ_E}{SQ_R} = 233.167$	$P(\mathrm{F}_{K-1,N-K} > \hat{F}) = 0.000$
Intragrups o Residual	$\sum_{j=1}^{K}\sum_{i=1}^{N} (y_{ji} - \bar{y}_j)^2 = 8.226$	$N-K=195$	$SQ_R / GdL_R = 0.0412$		
Total (o total corregida)	$\sum_{j=1}^{K}\sum_{i=1}^{N} (y_{ji} - \bar{Y})^2 = 46.472$	$N-1=199$			

A fi de tenir el mateix nombre d'observacions a cada grup, es comprova que

$$\bar{Y} = \frac{3.4217 + \cdots + 2.1731}{5} = 2.7614$$

- $\sum_{j=1}^{K} n_j (\bar{y}_j - \bar{Y})^2 = 40(3.4217 - 2.7614)^2 + \cdots + 40(2.1731 - 2.7614)^2 = 38.246$
- $\sum_{j=1}^{K}\sum_{i=1}^{N} \left(y_{ji} - \bar{y}_j\right)^2 = (n_j - 1)s_j^2 = 39 \cdot 0.1438^2 + \cdots + 39 \cdot 0.1932 = 8.226$

Un d'aquests dos càlculs es podria haver evitat si s'hagués tingut en compte que SQ_T és la que apareix a la taula 6.4.

(Nota. En aquest punt, el propi alumne es pot adonar de la incorrecció dels seus càlculs si obté una suma de quadrats negativa, ja que les sumes dels quadrats residuals, totals o entre grups sempre són positives.)

És a dir, rebutgem la hipòtesi de mitjanes totes elles iguals amb un p-valor < 0.001. La conclusió pràctica és que el *LOG(píxels)* va variant amb el grau de contrast.

c) Amb la informació de la qual disposem, l'única hipòtesi del procediment ANOVA que podem constatar és que les dades són independents per grup i dins de cada grup. *Manca verificar* les hipòtesis d'homoscedasticitat, això és, d'homogeneïtat de les variàncies, i de normalitat. No obstant això, sembla raonable acceptar com a vàlid aquest procediment. Perquè de la inspecció de la figura de l'enunciat es pot acceptar suposar que la variabilitat de les dades es manté homogènia al llarg de la del grau de contrast.

d) Es poden utilitzar els càlculs realitzats a l'apartat *c*, ja que en aquesta situació la suma de quadrats d'ambdós models coincideix. No obstant això, si no ens adonem d'aquest avantatge, els podem deduir de la informació subministrada als cuadres de l'enunciat.

En aquest cas, $R^2 = \frac{SQ_{\text{Re gressió}}}{SQ_{Total}} = \frac{SQ_{\text{Re gressió}}}{46{,}472} = 0.823$

Per tant, $SQ_{Regressió} = 46.472 \cdot 0.823 = 38.246$
i $SQ_{Errors} = 46.472 - 38.246 = 8.226$

o bé podem utilitzar el fet que $s = 0.2039 = \sqrt{\frac{SQ_{Errors}}{n-2}}$ i d'aquí coneixem que $SQ_{Errors} = 0.2039^2 \cdot 198 = 8.226$.

H_0: la regressió és globalment significativa; en regressió simple, és equivalent a plantejar:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

En aquest cas, com que p-value<0,05, rebutgem la hipòtesi nul·la.

La conclusió pràctica és que existeix relació lineal (en aquest cas, descendent) entre el *LOG(píxels)* i el grau de contrast.

Source	DF	SS	MS	F	P
Regression	1	38.246	38.246	910.62	0.000
Residual Error	198	8.226	0.042		
Total	199	46.472			

A més de contrastar que els paràmetres siguin significatius, mitjançant l'anàlisi dels residus, cal comprovar que aquests siguin de mitjana nul·la, de variància constant, independents, i segueixin una distribució normal.

e) Els dos tenen raó ja que, com s'ha vist, ambdós procediments són similars. L'avantatge de la recta de regressió sobre l'ANOVA és que ens permet realitzar prediccions per valors que no estan a la mostra, sempre que aquests estiguin en un entorn pròxim.

6.12 Annexos

6.12.1 Determinació dels estimadors mínims quadràtics de la recta de regressió

Trobem les expressions per calcular el valor dels estimadors per a una mostra concreta. El mínim es troba derivant parcialment respecte de b_0 i b_1 i imposant que el resultat sigui igual a zero. De les equacions resultants obtenim la solució:

$$\begin{cases} \partial s / \partial b_0 = 2 \Sigma (y_i - b_0 - b_1 \cdot X_i)\,(-1) \\ \partial s / \partial b_1 = 2 \Sigma (y_i - b_0 - b_1 \cdot X_i)(-X_i) \end{cases}$$

$$\begin{cases} -2 \Sigma (y_i - b_0 - b_1 \cdot X_i) = 0, \\ -2 \Sigma X_i (y_i - b_0 - b_1 \cdot X_i) = 0 \end{cases} \quad \text{o també} \quad \begin{cases} \Sigma e_i = 0 \\ \Sigma X_i \cdot e_i = 0 \end{cases} \quad \text{(equacions normals)}$$

$$\begin{cases} \Sigma y_i - n \cdot b_0 - b_1 \cdot X_i = 0 \\ \Sigma X_i y_i - \Sigma X_i \cdot b_0 - b_1 \cdot \Sigma X_i^2 = 0 \end{cases}$$

$$\boxed{b_0 = \sum y_i - b_1 \cdot \sum X_i \Big/ n = \overline{Y} - b_1 \cdot \overline{X}}$$

$$\Sigma X_i y_i - \Sigma X_i\,(\Sigma y_i - b_1\,\Sigma X_i)/n - b_1 \cdot n\,\Sigma X_i^2/n = 0$$
$$n \cdot \Sigma X_i y_i - \Sigma X_i\,\Sigma y_i + \Sigma X_i\,b_1\,\Sigma X_i - b_1 \cdot n \cdot \Sigma X_i^2 = 0$$
$$n \cdot \Sigma X_i y_i = \Sigma X_i\,\Sigma y_i + b_1 [n \cdot \Sigma X_i^2 - (\Sigma X_i)^2]$$
$$b_1 = [n \cdot \Sigma X_i y_i - \Sigma X_i\,\Sigma y_i] / [n \cdot \Sigma X_i^2 - (\Sigma X_i)^2]$$
$$= [\Sigma X_i y_i - \Sigma X_i\,\Sigma y_i / n] / [\Sigma X_i^2 - (\Sigma X_i)^2 / n]$$
$$= \{ [\Sigma X_i y_i - \Sigma X_i\,\Sigma y_i / n] / (n-1) \} / \{ [\Sigma X_i^2 - (\Sigma X_i)^2 / n] / (n-1) \}$$

$$\boxed{b_1 = s_{XY} \Big/ s_X^2 = r_{XY}\, s_Y \Big/ s_X}$$ (Recordatori: s_{XY} designa la covariància mostral.)

6.12.2 Distribució de l'estimador del pendent

En primer lloc, s'ha de veure que l'estimador del pendent es pot expressar com:

$$b_1 = s_{XY} / s_X^2 =$$
$$= [\Sigma X_i y_i - \Sigma X_i\,\Sigma y_i / n] / [\Sigma X_i^2 - (\Sigma X_i)^2 / n] =$$
$$= \Sigma [(X_i - \overline{X})\,(y_i - \overline{Y})] / \Sigma (X_i - \overline{X})^2 =$$
$$= [\Sigma (X_i - \overline{X}) y_i - \Sigma (X_i - \overline{X})\,\overline{Y}] / \Sigma (X_i - \overline{X})^2 =$$
$$= \Sigma (X_i - \overline{X}) y_i \;/ \Sigma (X_i - \overline{X})^2 =$$
$$= \Sigma [(X_i - \overline{X}) / \Sigma (X_i - \overline{X})^2]\, y_i =$$
$$= \Sigma w_i y_i \qquad \text{on} \qquad w_i = (X_i - \overline{X}) / \Sigma (X_i - \overline{X})^2$$

S'ha de tenir present que $\Sigma\, w_i = 0$,

i que

$$\begin{aligned}\Sigma\, w_i X_i &= \Sigma\,[(X_i - \overline{X}) / \Sigma\,(X_i - \overline{X})^2]\Sigma X_i = \\ &= \Sigma\,(X_i - \overline{X})\Sigma X_i / \Sigma\,(X_i - \overline{X})^2 = \\ &= \Sigma\,(X_i - \overline{X})\Sigma X_i / [\Sigma\,(X_i - \overline{X})X_i - \Sigma\,(X_i - \overline{X})\,\overline{X}\,] = \\ &= \Sigma\,(X_i - \overline{X})\Sigma X_i / \Sigma\,(X_i - \overline{X})X_i = \\ &= 1\end{aligned}$$

Per tant, la variable b_1 seguirà, al llarg de les possibles mostres, una distribució normal, perquè és combinació lineal de variables normals.

Vegem ara l'esperança i la variància d'aquesta variable b_1:

$$\begin{aligned}\mathrm{E}(b_1) &= \mathrm{E}(\Sigma\, w_i y_i) \\ &= \Sigma\, w_i\, \mathrm{E}(y_i) \\ &= \Sigma\, w_i\,[\beta_0 + \beta_1 \cdot X_i\,] = \\ &= \beta_0\, \Sigma\, w_i \; + \beta_1 \cdot \Sigma\, w_i X_i = \\ &= \beta_0 \cdot 0 \quad + \beta_1 \cdot 1 = \\ &= \beta_1\end{aligned}$$

$$\begin{aligned}\mathrm{V}(b_1) &= \mathrm{V}(\Sigma\, w_i y_i) = \\ &= \Sigma\, w_i^2\, \mathrm{V}(y_i) = \\ &= \sigma^2\, \Sigma\, w_i^2 = \\ &= \sigma^2\, \Sigma\,[\,(X_i - \overline{X}) / \Sigma\,(X_i - \overline{X})^2\,]^2 = \\ &= \sigma^2 / \Sigma\,(X_i - \overline{X})^2 = \\ &= \sigma^2 / \;(n-1)s_X^2\end{aligned}$$

6.12.3 Distribució de l'estimador del terme independent

S'intentarà expressar també la constant com a combinació lineal dels valors observats de la desposta Y_i.

$$\begin{aligned}b_0 &= \overline{Y} - b_1\, \overline{X} = \\ &= \Sigma y_i / n - \overline{X}\, \Sigma\, w_i\, y_i = \\ &= \Sigma\,(1/n - \overline{X}\, w_i)\, y_i = \\ &= \Sigma\, r_i\, y_i \qquad\qquad \text{on } r_i = 1/n - \overline{X}\, w_i\end{aligned}$$

S'ha de tenir present que es tracta de ponderacions: $\Sigma r_i = \Sigma\,(1/n - \overline{X}\, w_i) = 1 - \overline{X} \cdot 0 = 1$
per la qual cosa b_0 tindrà una distribució normal, ja que és també una combinació lineal de variables amb distribució normal. D'altra banda:

$$
\begin{aligned}
E(b_0) &= E(\Sigma\, r_i\, y_i) = \\
&= \Sigma\, r_i\, E(y_i) = \\
&= \Sigma\, r_i\, E(\beta_0 + \beta_1 \cdot X_i) = \\
&= \beta_0 \Sigma\, r_i + \beta_1 \Sigma\, r_i \cdot X_i = \\
&= \beta_0 \Sigma\, r_i + \beta_1 (\Sigma X_i / n - \overline{X}\, \Sigma X_i\, w_i) = \\
&= \beta_0 \Sigma\, r_i + \beta_1 (\overline{X} - \overline{X} \cdot 1) = \\
&= \beta_0 \cdot 1 + \beta_1 \cdot 0 = \\
&= \beta_0
\end{aligned}
$$

$$
\begin{aligned}
V(b_0) &= V(\Sigma\, r_i\, y_i) = \\
&= \Sigma\, r_i^2\, V(y_i) = \\
&= \sigma^2 \Sigma\, r_i^2 = \\
&= \sigma^2 \Sigma\, [\, 1/n - \overline{X}\, w_i\,]^2 = \\
&= \sigma^2 \Sigma\, [1/n^2 + \overline{X}^2\, w_i^2 - 2\, \overline{X}\, w_i / n\,] = \\
&= \sigma^2\, [\, n/n^2 + \overline{X}^2 \Sigma w_i^2 - 2\, \overline{X}\, \Sigma w_i / n\,] = \\
&= \sigma^2\, [\, 1/n + \overline{X}^2 / \Sigma (x_i - \overline{X})^2 - 2\, \overline{X} \cdot 0 / n\,] \\
&= \sigma^2\, [\, 1/n + \overline{X}^2 / (n-1) s_X^2\,]
\end{aligned}
$$

6.12.4 Sumatori de productes creuats nul

Cal demostrar que el sumatori dels productes creuats, $\Sigma\, (y_i - \hat{y}_i)(\hat{y}_i - \overline{Y})$, és igual a zero.

$$
\begin{aligned}
\hat{y}_i &= b_0 + b_1 \cdot X_i = \\
&= \overline{Y} - b_1\, \overline{X} + b_1 \cdot X_i = \\
&= \overline{Y} + b_1\, (X_i - \overline{X})\,; \\
\hat{y}_i - \overline{Y} &= b_1\, (X_i - \overline{X}) \\
\Sigma\, (y_i - \hat{y}_i)(\hat{y}_i - \overline{Y}) &= \Sigma\, \{\, [y_i - \overline{Y} - b_1 \cdot (X_i - \overline{X})]\; b_1\, (X_i - \overline{X})\,\} = \\
&= b_1 \Sigma\, \{(y_i - \overline{Y}) - b_1 \cdot (X_i - \overline{X})]\; (X_i - \overline{X})\,\} = \\
&= b_1\, \{\Sigma\, (y_i - \overline{Y})(X_i - \overline{X}) - b_1\, (X_i - \overline{X})^2\,\} = \\
&= b_1 \cdot 0 \qquad \text{ja que } b_1 = \Sigma\, (y_i - \overline{Y})\, (X_i - \overline{X}) / \Sigma\, (X_i - \overline{X})^2
\end{aligned}
$$

Des d'un altre punt de vista, equival a dir que el vector dels errors és ortogonal (o independent) al vector de les prediccions. És a dir, que els errors es distribueixen independentment de X.

7 Validació i previsió

Tot el que s'ha desenvolupat fins ara podria ser com fer un castell a l'aire si les premisses (també conegudes com a hipòtesis prèvies) no fossin correctes. Per descomptat, atès que els models matemàtics són sovint una representació simplificada de la realitat, no té sentit pretendre que les premisses s'apliquin amb exactitud. Però un grau "raonable" de compliment és necessari per a l'essència mateixa del model, és a dir, perquè sigui *útil*. Determinar un procediment per comprovar quin grau de compliment s'ha assolit és l'objectiu de la primera part d'aquest capítol.

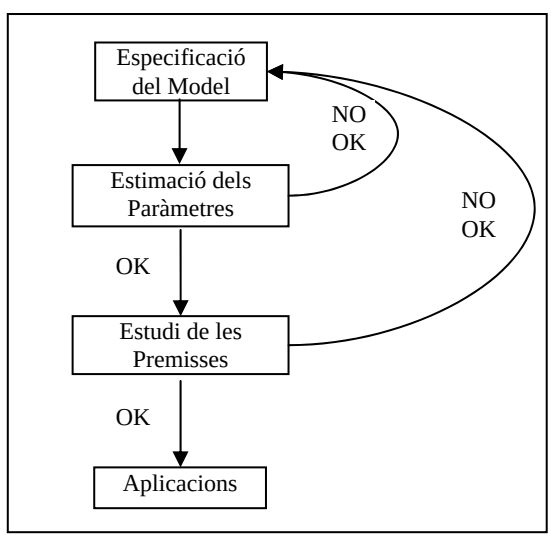


Figura 7.1 El cicle de la modelització

Què s'ha de fer si no es compleixen les premisses? Breument, cal continuar buscant un model adequat: reespecificar-hi el model (buscar-hi unes altres variables o fer-hi transformacions, per exemple) i tornar a començar el procés. Una vegada es té un model satisfactori, es pot aplicar, inferint

propietats dels paràmetres o bé realitzant previsions per a la resposta. L'esquema de la figura 7.1 es vincula perfectament amb el que Box, Hunter i Hunter proposen al seu llibre,[17] que subratlla la relació estreta entre deducció i inducció en el procés de la formació del coneixement.

Tot seguit, es tracta una de les parts més interessants del model lineal (o potser la que el converteix en un model tan apreciat): la seva capacitat per fer previsions de la resposta per a valors concrets de la variable X. Anant més enllà de la simple estimació puntual —donada per la recta de regressió—, es mostra com calcular intervals de confiança, tant per a un possible valor de la resposta, com per al valor mitjà.

Al final del capítol s'ofereixen uns casos pràctics, completament resolts i comentats, que poden ser una bona guia per afrontar situacions semblants.

7.1 Estudi de les premisses

S'ha vist que les premisses de la regressió fan referència a les dues grans components del model: la funcional de X (linealitat) i l'aleatòria (normalitat, homoscedasticitat i independència). El compliment d'aquestes premisses permet recórrer a les distribucions de referència (t, F) per fer intervals de confiança, proves de significació o contrastos d'hipòtesis. Tinguem present que els mecanismes d'inferència que s'han exposat al capítol anterior, i els de previsió que s'explicaran a continuació, fan ús d'aquestes premisses en determinats passos dels seus desenvolupaments. Així, posem per cas, una interrelació no desitjada entre els individus de la mostra podria ocasionar la infravaloració de la desviació residual i, per tant, un increment injustificat de la potència de les proves. I, malgrat que, per a mostres grans, els mètodes vistos es consideren robustos (es preserven els riscos i els nivells de confiança assumits), també és cert que les premisses són necessàries per eficiència: el mètode de mínims errors quadrats es considera el millor estimador no esbiaixat quan es acompleixen aquestes premisses. Si fallés, per exemple, la normalitat dels residus —per asimetria o per presència de valors massa allunyats—, l'estimador mínim-quadràtic seria menys eficient que altres mètodes.

Si les premisses estableixen condicions prèvies, per què no les hem vist al inici del tema? Perquè no són l'objectiu principal de l'anàlisi i perquè el tècnic o l'investigador, abans de recollir les dades, no és un ignorant absolut de les condicions reals de les premisses. Per alguna cosa és un expert en aquesta matèria, i no s'arriscarà a realitzar una recollida costosa de dades per acabar dient: "Llàstima, no puc fer l'anàlisi perquè no es compleixen les premisses!" Però, naturalment, farà una exploració de les dades per ratificar que les premisses són vàlides, o per trobar una formulació alternativa del model que encaixi millor amb les condicions del fenomen en estudi.

Nota. Idealment, l'estudi de les premisses hauria de ser inferencial (inductiu), ja que les premisses es refereixen als paràmetres, no als estadístics. Però, per diverses raons, aquesta inferència no es pot realitzar de la forma habitual. En primer lloc, hi ha un consum repetitiu de riscos α i β, per a cada premissa, per a cada residu e_i i per a cada nou estudi.

[17] *Estadística para investigadores. Introducción al diseño de experimentos, análisis de datos y construcción de modelos*. Ed. Reverté.

Per exemple: la longitud dels cargols que produeix una màquina segueix una distribució normal. Si cada dia es realitza una prova d'hipòtesi sobre aquesta distribució amb un risc α, acabarem per rebutjar aquesta hipòtesi algun dia, encara que la màquina funcioni bé. Si es realitzen k proves, l'esperança del nombre de dies que la prova doni positiu és $\alpha \cdot k$. La prova d'hipòtesi permet posar a prova una mateixa hipòtesi una vegada, però no k vegades.

La segona dificultat que té utilitzar la prova d'hipòtesi per estudiar les premisses és que la hipòtesi que es voldria poder afirmar és la nul·la (igualtat de variàncies, normalitat,…), per la qual cosa s'hauria de fitar β, usualment desconeguda, cosa que ens obligaria a precisar quina és l'heterocedasticitat assumible, quin grau d'*anormalitat* resulta irrellevant, etc.

Les premisses s'estudiaran mitjançant: (1) el raonament, d'acord amb la bibliografia i el coneixement teòric sobre el tema, i (2) la seva anàlisi gràfica, principalment.

7.1.1 Informació a priori

Respecte al primer punt, es tracta de considerar qualsevol aspecte relacionat amb la naturalesa de l'estudi que pugui fonamentar, a priori, una premissa del model. Per exemple, prenem com a variable resposta la qualificació de l'alumne en una assignatura; segurament, la nota observada s'obté de la combinació de diverses proves parcials independents, si el sistema d'avaluació és l'adequat. D'aquesta manera, per a alumnes en igualtat de condicions —és a dir, per a un valor constant de la variable explicativa—, es pot esperar que la resposta es distribueixi, aproximadament, com una normal.

Suposem una altra situació. S'observa ara el temps real que triga un programa a portar a terme una tasca determinada, que depèn d'un factor X concret. El temps real és una suma del temps que es consumeix a la CPU i dels temps que el procés està esperant perquè el sistema atengui altres processos, però aquesta suma, que solament comporta dues components positives, no es pot esperar que es distribueixi com una normal. Més aviat observarem molts temps típics, uns quants temps notablement per sobre (quan el procés competeix amb altres processos que exigeixen molt aquest recurs), i mai o quasi mai no trobarem un temps molt més petit que l'habitual: els residus tindran una distribució asimètrica.

Cal fer molta atenció en l'anàlisi d'una variable resposta relacionada amb el rendiment d'un procés, en funció de factors com la mida de l'entrada en qüestió. En casos com aquest, l'interès no és si la mida de l'entrada (N) afecta el rendiment. És clar que sí: si processem una imatge digital més gran, és sabut que el programa trigui més temps. En informàtica teòrica, s'estudia la complexitat d'un algorisme determinant-ne el cost —temporal o espacial— que està associat a l'algorisme segons la mida N. No es precisa cap paràmetre sinó un ordre de magnitud simple. Alguns exemples són:

- una cerca dicotòmica té cost logarítmic [$O(\log(N))$];
- si la llista no està ordenada, el cost de la cerca seqüencial és lineal [$O(N)$];
- ordenar un vector té cost quadràtic [$O(N^2)$],
- encara que hi ha alguns mètodes que, de mitjana, són més ràpids, com el *mergesort* [$O(\log(N) \cdot N)$];
- resoldre un sistema d'equacions lineals és cúbic en el nombre d'incògnites [$O(N^3)$];
- obtenir totes les permutacions de N objectes té un cost exponencial [$O(2^N)$].

En general, quan un algorisme té un cost superior a polinòmic, s'ha d'evitar utilitzar-lo, ja que l'esforç que requereix és tan gran que possiblement no se'n pot trobar la solució —excepte per a mides molt moderades. No cal dir que és fonamental conèixer aquesta informació a l'hora de planificar un estudi estadístic. Aquests són molt útils per estimar paràmetres que permetin efectuar previsions amb un model, o comparar mètodes que teòricament tenen el mateix cost. Encara que s'esculli un algorisme no lineal, no és difícil transformar les dades per poder utilitzar el model lineal de regressió. No obstant això, fins i tot si l'algorisme és lineal, és imprescindible adonar-se que la mida de l'entrada influeix no solament en la magnitud de la resposta, sinó també en la *dispersió* de la resposta, i així apareix el fenomen de l'*heterocedasticitat*.

7.1.2 Anàlisi gràfica

Vegem ara com es fa l'anàlisi gràfica. Al contrari que els arguments anteriors, aquesta part requereix haver realitzat l'estimació del model i trobat els residus.

Iniciem l'anàlisi amb l'estudi per detectar absència d'*independència* entre els termes d'error. Generalment, es pot dir que és un problema del disseny i de la recollida de dades, de manera que caldria considerar aquest punt entre les reflexions incloses apriorísticament, però també es podria tractar d'una altra situació, que es pot recuperar sense haver-se de plantejar un nou experiment. Qualsevol variable que s'hagi deixat de banda de l'anàlisi pot fer aparèixer dependències entre els termes d'error. Per exemple, l'alçada de dos adults és més similar que la de dues persones a l'atzar. Si, per exemple, no es considera el gènere en estudiar l'alçada, els residus dels homes tendiran a ser positius i els de les dones, negatius. Un gràfic amb aquesta variable faria evident aquesta relació. Les solucions possibles passen per la incorporació de la variable en el model, l'estratificació, o anàlisis per subgrups induïts per la tercera variable o, si no hi ha cap més remei, la restricció de l'anàlisi a una sola categoria.

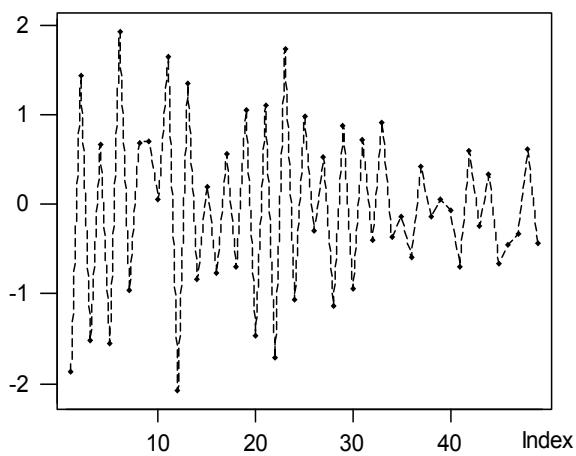


Figura 7.2 Diagrama temporal dels residus; l'eix horitzontal és l'ordre de les observacions, que estaria relacionat amb el temps. En aquest cas, s'evidencia que les observacions no són independents, tendència a canviar de signe.

Una altra variable interessant és l'ordre de les observacions, molt important si estan recollides amb algun patró temporal o espacial. Pots er que la proximitat en el temps o en l'espai de dues unitats observades provoqui una relació mútua, negativa o positiva.

La figura 7.2 il·lustra un exemple de gràfic dels residus *versus* l'ordre de recollida de les observacions. En aquest exemple, els residus consecutius tenen una curiosa relació: si un és positiu, el següent és negatiu, i viceversa. Aquest fenomen es podria donar en situacions en què el sistema "té memòria". Per exemple, després d'una nit de moltes hores de descans podria venir-ne una altra de molt poques. En el camp borsari, en situacions d'alta volatilitat, després d'un dia de fortes pujades ve un dia de baixades importants, per les vendes que es fan per recuperar els beneficis obtinguts especulant a curt termini.

Per estudiar la *linealitat* i l'*homoscedasticitat*, el primer gràfic ha de ser, naturalment, el de Y_i *versus* X_i. A l'exemple de la figura 7.3, es mostra un grau satisfactori tant per a la linealitat de Y respecte de X com per a l'homoscedasticitat dels errors, almenys en una primera aproximació.

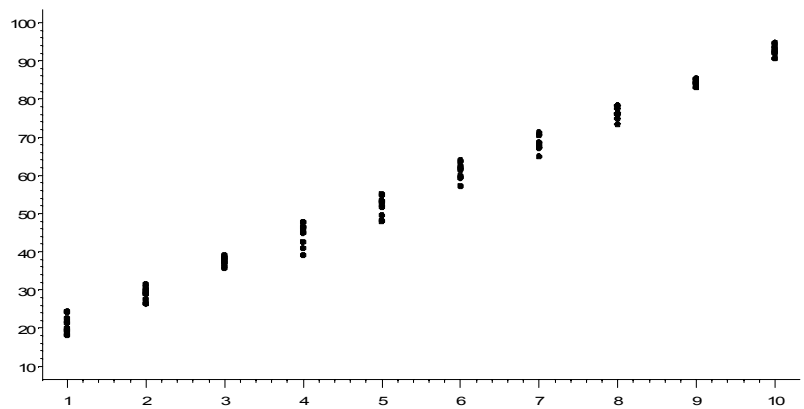


Figura 7.3 Exemple de diagrama de dispersió de Y contra X. La linealitat i l'homoscedasticitat es poden donar per bones.

Aquest gràfic és molt intuïtiu, però malgasta molt espai, especialment si la correlació és alta. El resultat és que les desviacions petites de les premisses podrien no detectar-se. El gràfic es pot millorar substituint Y pels residus, amb la qual cosa s'eviten els espais blancs i augmenta el detall. Així, el gràfic de e_i *versus* X_i permet valorar més clarament la linealitat i l'homoscedasticitat. És habitual no representar a l'eix d'ordenades els residus absoluts sinó els residus estandarditzats, e_i/s_e, que tenen variància 1. Les dades anteriors quedarien com es veu a la figura 7.4.

És raonable criticar l'homoscedasticitat, per la menor dispersió dels grups X=3 i X=9, en aquestes dades? Més aviat no; una conclusió que les dades són heterocedàstiques sembla precipitada. Les fluctuacions de l'atzar poden provocar una impressió visual una mica alarmant, però tampoc no es pot esperar que amb poques dades observem una dispersió o amplada del núvol perfectament regular; no és això el que l'atzar ens dóna normalment. Per sobre de les variacions atribuïbles a la casualitat, amb l'anàlisi gràfica s'hauria d'apreciar un núvol bàsicament *horitzontal* (linealitat: absència de relacions

no lineals de *X*) i amb amplada bàsicament *constant* (dispersió uniforme, més enllà de la posició extrema d'algun punt aïllat).

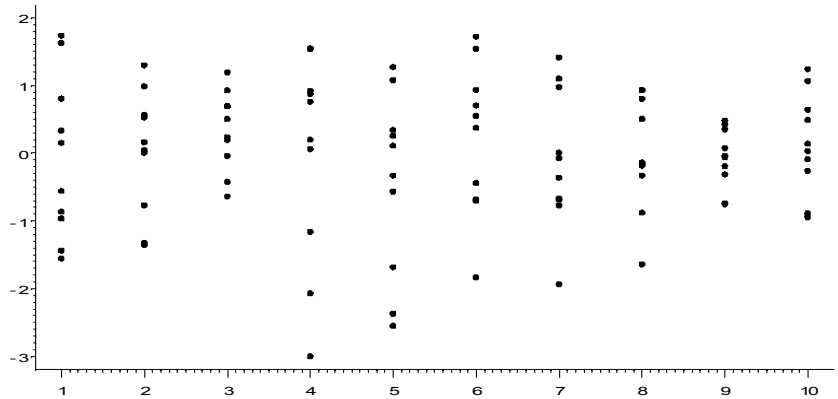


Figura 7.4 Exemple de diagrama de dispersió dels residus estandarditzats contra X per les dades anteriors

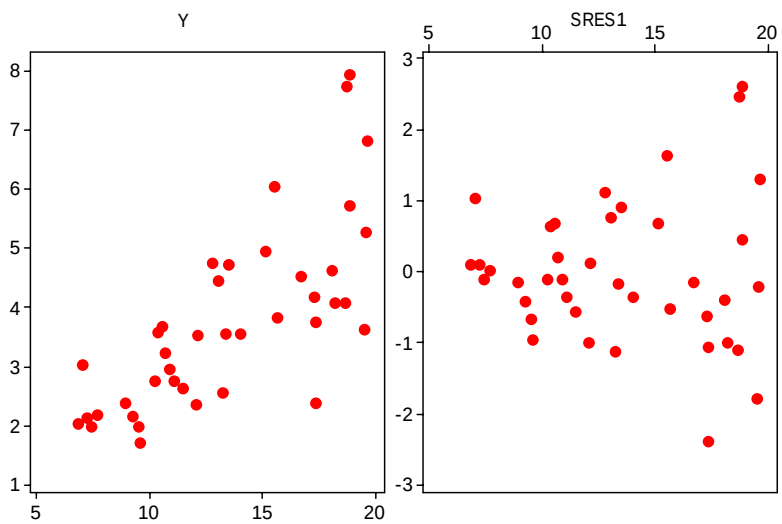


Figura 7.5 Exemple d'heterocedasticitat: a l'esquerra, gràfic de Y versus X; a la dreta, residus estandarditzats versus X

El gràfic de la figura 7.5, en canvi, representa unes dades que presenten una desviació tipus proporcional al valor de la variable explicativa (quan *X* augmenta, creix la dispersió dels valors de *Y*). La forma creixent de les dispersions s'aprecia molt clarament, en especial al gràfic de la dreta. I el gràfic de la figura 7.6 és un exemple clar de no-linealitat, mostrant la forma típica de "banyera" o "u" (una altra forma típica és la d'un arc: generalment, les estructures que es poden apreciar provenen de formes convexes o còncaves).

La figura 7.7 ens mostra unes dades en les dues formes que hem presentat, i observeu que ens podrien portar a conclusions ben diferents. A l'esquerra, veiem el diagrama *Y* contra *X*, amb un núvol que ve marcat per una tendència lineal indiscutible, i una desviació residual molt petita, de manera que les dades semblen quasi disposades sobre una recta. Tanmateix, quan es veu el diagrama dels residus, després d'extraure la tendència lineal, s'aprecia amb claredat una forma no horitzontal, convexa, senyal que la resposta també presenta una component no lineal (proporcional a X^2, per exemple) que quedava oculta al primer gràfic. Tingueu present que, si hi ha una influència de tipus quadràtic a la resposta, encara que en el rang actual sembli negligible, aquesta component pot créixer ràpidament si *X* és més gran, i invalidar totalment l'ús del model. Dit d'una altra manera, un efecte no lineal pot ser obviat si constatem que el model lineal, essent més senzill, ens porta a resultats molt propers als que ens donaria un model més complex.

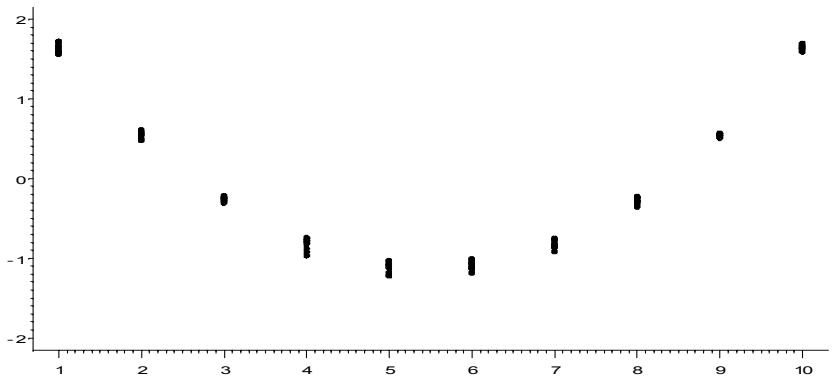


Figura 7.6 Exemple de no-linealitat en un gràfic dels residus estandarditzats

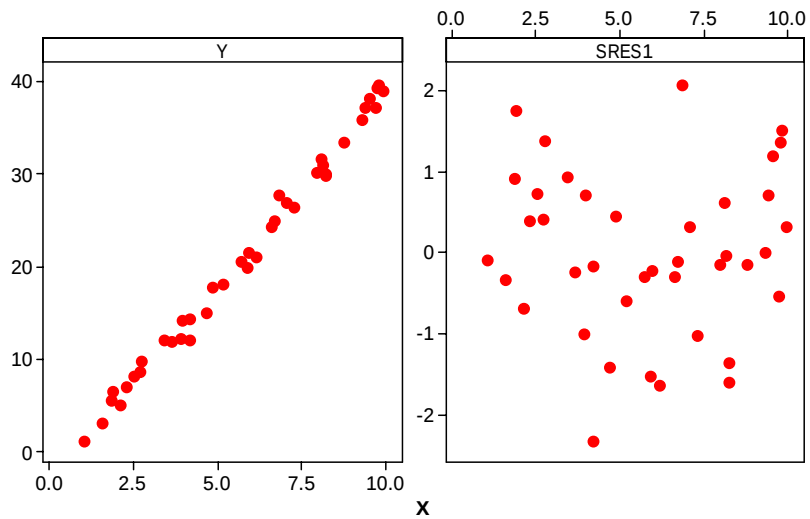


Figura 7.7 Exemple de no-linealitat, que s'aprecia més clarament al gràfic dels residus (a la dreta) però no al diagrama de dispersió Y versus X (a l'esquerra).

Nota. En ser $\hat{y}_i$ funció lineal de X, el gràfic de e_i *versus* $\hat{y}_i$ és un gràfic equivalent al de e_i *versus* X_i, ja que només canvia l'escala de les abscisses (si el signe del pendent és negatiu, s'obté un gràfic simètric). En regressió múltiple, amb més d'un regressor, el gràfic de e_i *versus* $\hat{y}_i$ presenta alguns avantatges.

És important que l'última premissa que s'estudiï sigui la *normalitat*, perquè la fallida de qualsevol altra premissa pot tenir com a efecte la no-normalitat dels residus. Per exemple, si hi ha heterocedasticitat, no apreciarem normalitat en els residus encara que sí que ho fossin condicionadament a X. S'ha de tenir en compte que la premissa diu estrictament que el que ha de ser normal és la resposta condicionada per X o, alternativament, les pertorbacions al voltant de la recta. Com que no és habitual disposar d'un nombre suficient d'observacions per a una mateixa X_i, resulta obligat tractar totes les observacions conjuntament: per tant, l'anàlisi és una mica més exigent del que es requereix. Per criticar la premissa de normalitat, el gràfic que s'estudia és el *normal probability plot* (NPPlot), també conegut com a *recta d'Henry*.

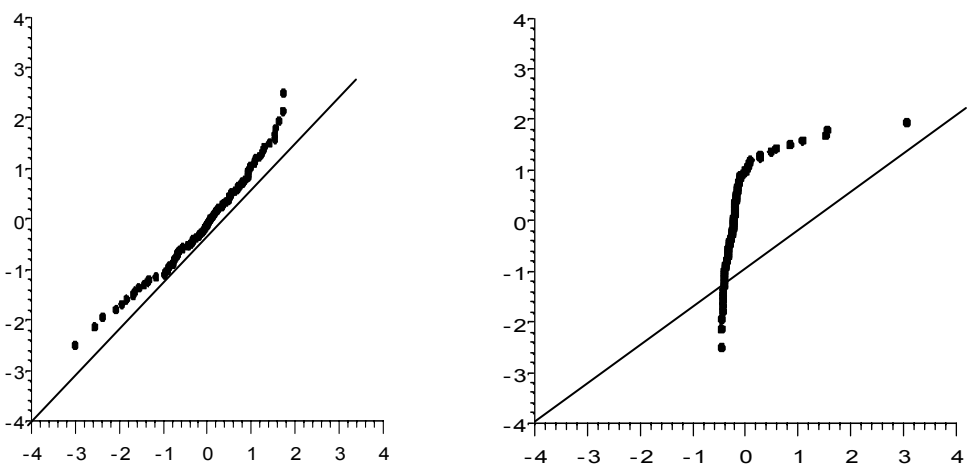


Figura 7.8 A l'esquerra, típic NPPlot d'una variable normal; a la dreta, NPPlot d'una variable amb asimetria per la dreta

En el diagrama, les abscisses representen els residus estandarditzats i les ordenades, els residus estandarditzats i normalitzats, per la qual cosa, si ja eren normals, s'observarà una seqüència de punts molt semblant a la recta identitat. Cal fer atenció a les cues de la distribució, ja que així es pot detectar un nombre inusual d'observacions anòmales. Els residus normalitzats es calculen obtenint el valor per a la llei normal corresponent a l'ordre de cada observació: en primer lloc, s'ordenen els residus; després, per j=1,...,n, el residu normalitzat per al j-èsim residu és el percentil $(j–1/2)/n$ de la distribució N(0,1).

Vegem uns exemples d'una regressió feta amb 100 observacions. Mostrem els residus estandarditzat i normalitzat per a cada un dels cinc primers punts i els cinc darrers punts:

j	Res. estand.	Nivell	Res. norm.
1	–2.56	0.5%	–2.576
2	–2.35	1.5%	–2.170
3	–2.16	2.5%	–1.960
4	–2.09	3.5%	–1.812
5	–1.54	4.5%	–1.695

j	Res. estand.	Nivell	Res. norm.
96	1.56	95.5%	1.695
97	1.72	96.5%	1.812
98	1.82	97.5%	1.960
99	1.94	98.5%	2.170
100	2.86	99.5%	2.576

Als dos gràfics de la figura 7.8 podem veure quin resultat dóna la recta de Henry en dues situacions extremes: una mostra d'observacions d'una variable amb distribució normal (primer diagrama) i una mostra que prové d'una variable molt asimètrica per la cua de la dreta, al segon diagrama.

7.2 Previsions de la resposta

És fàcil predir puntualment valors de la variable resposta Y per a valors concrets de la variable independent X, utilitzant la part determinista del model, la funció de la recta. Però, ¿com tenir en compte el terme d'error ε, que afegeix el modelat estadístic? Aquest és el repte del present apartat. Insistim en que aquest terme ε representa pertorbacions aleatòries, les diferències en els valors de la resposta Y per a casos amb el mateix valor de X. No s'ha d'oblidar que $Y|X$ és una variable aleatòria, i que és important no obviar la seva variabilitat.

Per aquesta raó, d'entrada hauríem de distingir què és allò que volem predir (discussió que no es planteja amb un model determinista): és diferent si es vol predir, per a un cert valor X_h, el valor de Y en una observació individual que si pretenem predir o conèixer la mitjana poblacional, és a dir, l'esperança matemàtica de Y, en les unitats amb factor $X=X_h$. En la primera situació, s'han de considerar les possibles diferències que la resposta Y presenta en diverses unitats que tinguin el mateix valor X_h (la variabilitat *intragrup*). En canvi, si es vol predir la mitjana, aquestes diferències les podem obviar. Com que és més simple, comencem amb l'exposició d'aquesta segona situació.

7.2.1 Previsió del valor mitjà

Per estimar puntualment la mitjana μ_h de Y per a $X=X_h$, es fa servir l'equació de la recta, en qualsevol de les seves expressions habituals:

$$\hat{y}_h = b_0 + b_1 \cdot X_h = \\ = \overline{Y} + b_1 \cdot (X_h - \overline{X})$$

Tinguem present que:

$$\mu_h = \mathrm{E}(Y \mid X_h) = \mathrm{E}(\beta_0 + \beta_1 \cdot X_h + \varepsilon_h) = \beta_0 + \beta_1 \cdot X_h$$

Sabem també que hi ha un cert soroll degut al mostreig en els estimadors b_0 i b_1. Precisament perquè s'accepta una variabilitat σ entre observacions amb el mateix valor de X, es produeix algun error en l'estimació dels paràmetres β_0 i β_1 (recordeu que en l'error típic d'ambdós estimadors apareix σ). Per tant, s'ha d'estudiar com influeix aquesta variabilitat de l'estimació en les previsions.

En primer lloc, observeu que l'estimació del valor mitjà μ_h no és esbiaixada:

$$\mathrm{E}(\hat{y}_h) = \mathrm{E}(b_0 + b_1 \cdot X_h) = \beta_0 + \beta_1 \cdot X_h = \mu_h$$

I, en segon lloc, estudiem l'oscil·lació d'aquesta estimació:

$$\mathrm{V}(\hat{y}_h) = \mathrm{V}\left[\overline{Y} + b_1 \cdot (X_h - \overline{X})\right] =$$

$$= \mathrm{V}(\overline{Y}) + (X_h - \overline{X})^2 \, \mathrm{V}(b_1),$$

donada la independència de $\overline{Y}$ i b_1. Continuem substituint pels valors coneguts:

$$= \sigma^2/n + (X_h - \overline{X})^2 \, \sigma^2 / (n-1)s_X^2 =$$

$$= \sigma^2 \left[1/n + (X_h - \overline{X})^2 / (n-1) \, s_X^2\right]$$

Així doncs, es coneix quant varia d'una mostra a una altra la previsió del valor poblacional μ_h. La interpretació d'aquesta fórmula és senzilla: les oscil·lacions de les prediccions vénen originades pel soroll de l'estimador de la posició de la recta ($\overline{Y}$) i pel soroll de l'estimació de la seva inclinació (b_1). S'ha de tenir present que aquesta oscil·lació en l'estimació del pendent provoca una variància més gran en la predicció com més gran sigui la distància a $\overline{X}$, com més allunyat del centre de les variables explicatives es trobi l'objectiu X_h de la nostra predicció.

L'aplicació de l'expressió anterior utilitza l'estimador s per a la desviació residual σ. Llavors, l'estadístic de referència que podem utilitzar, per exemple, per trobar una regió de confiança per a μ_h seria:

$$T = \frac{\hat{y}_h - \mu_h}{s\sqrt{\frac{1}{n} + \frac{(X_h - \overline{X})^2}{(n-1)s_X^2}}} \sim \mathrm{t}_{n-2}$$

Gràficament, es pot veure que, si fem variar X_h, les bandes que delimiten els extrems de la regió de confiança envolten la recta de regressió, amb l'amplada menor en el punt de la mitjana de X (on l'error típic és més petit) i obrint-se cap a la dreta i l'esquerra. La figura 7.9 mostra el marge de confiança del 95% de les mitjanes poblacionals per a valors X_h en el rang de les X emprades. Es pot veure com les corbes van ampliant el marge d'incertesa a mesura que ens allunyem del centre de la recollida de dades: hi ha més informació, menys incertesa, pel centre que pels extrems.

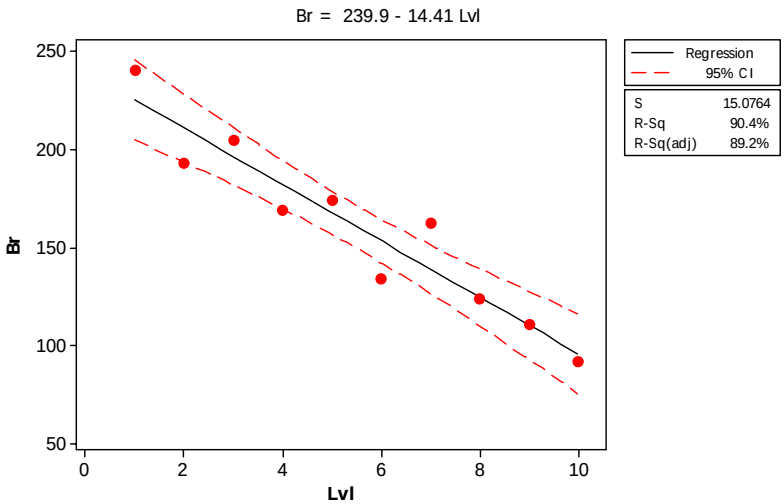


Figura 7.9 Diagrama amb núvol de punts i bandes al 95% de confiança per a la mitjana de la resposta. L'exemple correspon al que hem vist al capítol anterior.

Un aspecte que detaca d'aquest exemple és que les observacions individuals poden situar-se fora d'aquest interval de confiança de les μ_h especialment si la grandària de la mostra és gran. Efectivament: s'està estimant el seu centre, la seva mitjana, no cada una de les observacions. Aquest és l'objectiu de l'apartat següent.

7.2.2 Previsió d'una observació individual

Per construir l'interval dels valors individuals y_h de Y per a $X=X_h$, farem servir també:

$$y_h = \hat{y}_h = b_0 + b_1 \cdot X_h$$

ja que el centre de les observacions és la mitjana poblacional μ_h, que s'ha predit abans. Però ara s'ha de tenir en compte també la variabilitat σ d'aquestes observacions. Més formalment, es pot demostrar (de forma similar a com es va veure en la descomposició de les sumes de quadrats) que:

$$\begin{aligned} \mathrm{E}(y_h - \hat{y}_h)^2 &= \mathrm{E}(\hat{y}_h - \mu_h)^2 + \mathrm{E}(y_h - \mu_h)^2 \\ \mathrm{V}(\text{predicció individual}) &= \mathrm{V}(\text{estimació de la predicció mitjana}) + \mathrm{V}(\text{variable}) \\ \mathrm{V}(y_h) &= \sigma^2 [1/n + (X_h - \overline{X})^2 / (n-1)s_X^2] + \sigma^2 = \\ &= \sigma^2 [1 + 1/n + (X_h - \overline{X})^2 / (n-1)\, s_X^2] \end{aligned}$$

cosa que permet identificar tres fonts de variabilitat en la previsió dels valors individuals:

Variabilitat natural + Var. per estimació de la mitjana + Var. per estimació del pendent

Com sempre, per fer l'estimació s'ha de substituir σ pel seu estimador s.

Vegem quin aspecte té la previsió dels valors individuals de l'exemple anterior. El gràfic de la figura 7.10 mostra l'interval al 95% de les observacions individuals. També als extrems la banda és més ampla que al centre. S'ha de tenir present que el valor s afegit ha diluït l'aspecte visual de la curvatura de les bandes inferior i superior (que són equacions de segon grau de X_h, és a dir, paràboles). Observeu també que ara tots els punts es troben dins de l'interval. En un cas amb més observacions podríem esperar que s'hi encabissin el 95% del total.

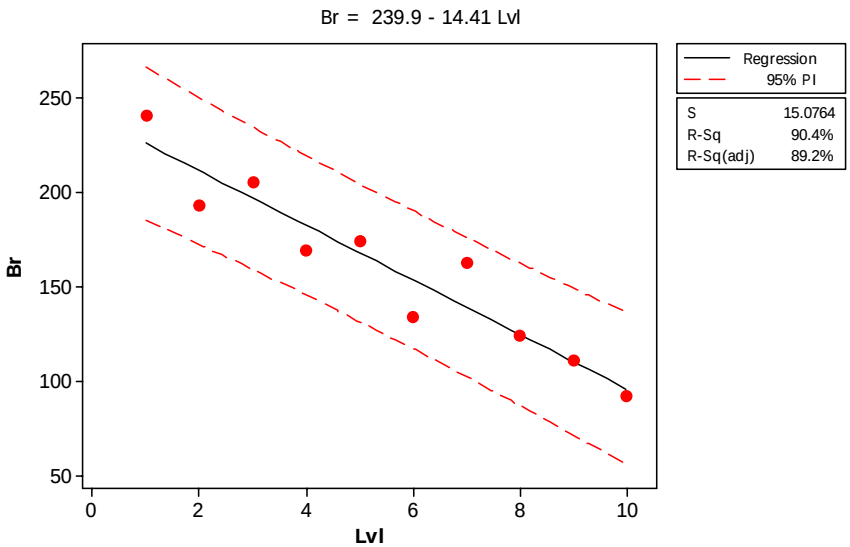


Figura 7.10 Diagrama amb núvol de punts i bandes al 95% de confiança per a observacions individuals. Observeu que considerar la variabilitat individual augmenta notablement la regió marcada a la figura 7.9.

7.2.3 Resum

Aquestes són les expressions que s'han d'emprar per trobar l'interval de confiança per als dos tipus de previsió comentats. Tingueu en compte que, en qualsevol cas, abans de calcular l'estimació per interval, en primer lloc cal trobar l'estimació puntual:

Puntual per $X_h \rightarrow \hat{y}_h = b_0 + b_1 \cdot X_h$

Per interval de confiança (95%):

Per al valor esperat $E(\hat{y}_h)$:

$$\hat{y}_h \pm t_{n-2,0.975}\, s \sqrt{\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum \left(X_i - \overline{X}\right)^2}}$$

Per a valors individuals y_h:

$$\hat{y}_h \pm t_{n-2,0.975}\, s \sqrt{1 + \frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum \left(X_i - \overline{X}\right)^2}}$$

7.3 Cas pràctic: estudi del mòdem

S'ha demanat a voluntaris que disposessin d'una connexió a la xarxa amb un mòdem de 56 Kbps,[18] que anotessin les dades d'una transferència de fitxers des d'un servidor FTP. Aquest servidor conté un gran nombre de fitxers de text de mides molt diverses, i s'instrueix als voluntaris que agafin un fitxer a l'atzar i prenguin nota de la seva mida en kB i del temps invertit a descarregar el fitxer a l'ordinador personal.

Les dades resultants figuren a la taula 7.1. La figura 7.11 representa el núvol de punts i la recta de regressió. Sembla evident que hi ha una important component aleatòria en el temps de descàrrega, a part del fet que la mida del fitxer és un factor essencial.

Taula 7.1 Transferència de fitxers amb mòdem de 56 Kbps

	1	2	3	4	5	6	7	8
Mida (kB)	111	128.286	1629.48	1.743	1110.22	441.971	478.243	94.906
Temps (s)	10	33.00	264.35	0.52	271.00	33.00	102.00	9.89

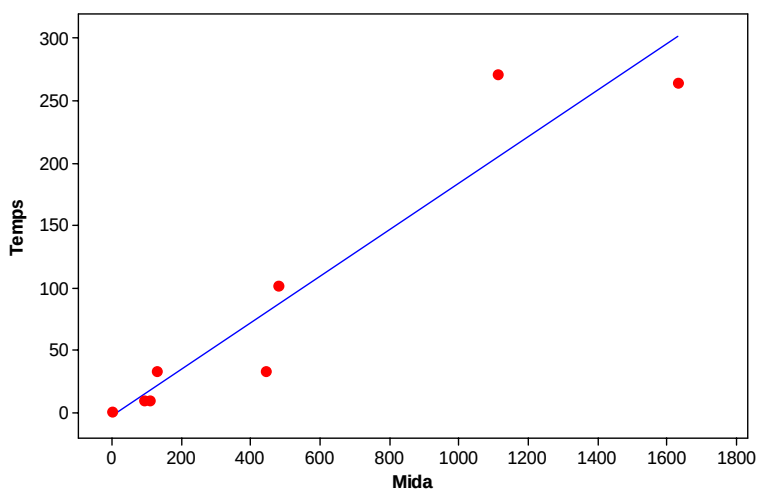


Figura 7.11 Diagrama de dispersió del temps i de la mida del fitxer descarregat del servidor FTP

[18] kbps = kilobits per segon. 56 kbps equivalen a 7 kilobytes per segon, o 7 kB (1 byte = 8 bits).

Amb un paquet estadístic, obtenim els estimadors de la regressió:

```
The regression equation is
Temps = - 2.9 + 0.187 Mida


Predictor     Coef  SE Coef      T      P
Constant     -2.85    18.05  -0.16  0.880
Mida       0.18683  0.02448   7.63  0.000


S = 37.5501   R-Sq = 90.7%   R-Sq(adj) = 89.1%


Analysis of Variance

Source          DF     SS     MS      F      P
Regression       1  82163  82163  58.27  0.000
Residual Error   6   8460   1410
Total            7  90623
```

D'aquesta sortida hem de destacar:

- L'equació de la recta és –2.9 + 0.187·Mida.
  - ◊ El pendent val 0.187, que vol dir que un increment d'1 kB en el fitxer suposa trigar 0.187 segons més en la descàrrega (5.35 kB/s: la capacitat teòrica del mòdem hauria de ser 56/8=7 kB/s).
  - ◊ El terme independent és negatiu, –2.85, cosa que no té gaire sentit: ve a significar el temps afegit a cada transmissió, independentment de la mida del fitxer. Seria lògic que el terme poblacional fos positiu o nul, però no pot ser inferior a zero (en efecte, com comentarem més endavant, en aquest cas el coeficient no és significatiu).
- Continuem analitzant la sortida una mica més avall. Veiem que l'estimació de la desviació residual val s = 37.55 segons. Interpretada directament, diríem que el temps de transmissió d'un fitxer (tant se val si és gran o petit) està afectat per pertorbacions aleatòries de desviació estàndard igual a uns 40 segons.
- El coeficient de determinació R^2 val 90.7%. Com és normal, la mida del fitxer explica la major part de la variabilitat dels temps recollits. Seria interessant preguntar-se quina és la font real del 9.3% de variabilitat no relacionada amb la mida del fitxer, és a dir, per què dos fitxers de la mateixa mida triguen temps diferents.

- Just a sobre de la línia de s i R^2 trobem una informació molt valuosa: l'error tipus de les estimacions.

  - ◊ La desviació estàndard del pendent és 0.02448. Com que el factor $t_{6,\,0.975}$ val 2.4469, l'IC per l'increment de temps per unitat de mida queda com (0.1269, 0.2467). El valor t = 7.63 que ens mostra la sortida del paquet és el quocient de l'estimació i la seva desviació, per posar a prova si el pendent pot ser nul, és a dir, si el temps no depèn de la mida del fitxer (el p-valor és molt petit, i es pot rebutjar, com ja imaginàvem).

  - ◊ Pot tenir més interès comparar el nostre resultat amb la capacitat del mòdem. 7 kB/s equival a un increment de 0.1429 segons per kB; el contrast es podria formalitzar com:

    $$H_0: \beta_1 = 1/7$$
    $$H_1: \beta_1 \neq 1/7$$

    Hem elegit un enfocament bilateral per no descartar, de partida, una capacitat per sobre de l'especificada (encara que tècnicament podria estar justificat, per les limitacions que suposa la transmissió per una línia de veu). L'estadístic per a les nostres dades val:

    $$t_{b_1} = (b_1 - \beta_1) / \sqrt{(s^2 / (n-1)s_X^2)} = (b_1 - \beta_1) / \; s_{b_1} = (0.187 - 1/7)/ \, 0.02448 = 1.80$$

    Aquest valor no permet rebutjar la hipòtesi nul·la, perquè està associat a un p-valor igual a 0.12, que no es pot considerar massa petit (almenys, enfront d'un risc habitual com α=5%).

    (Observeu que, abans d'haver validat les premisses del model, ja estem inferint propietats amb ell. Tot el que estem fent ara és una interpretació de la sortida del paquet estadístic, i està pendent de corroboració.)

  - ◊ Pel que fa al terme independent, s'interpreta de la mateixa forma. Observem primer que la prova de nul·litat inclosa té un p-valor molt gran, 0.880, explicable per la variabilitat enorme de l'estimació en comparació del seu valor absolut. Aquesta idea s'evidencia per mitjà del seu IC: (–47.0, 41.3). Una mostra similar ens podria haver donat un valor estimat de 40 segons de retard afegit! Per tant, es conclou que no tenim informació per estimar adequadament aquest terme, encara que també podria ser nul -un plantejament bastant creïble.

- La taula de la descomposició de la variabilitat ens en dóna més detalls, com també l'estadístic F (58.27, que és el quadrat de t = 7.63), que ens confirma que el factor temps té un pes decisiu per explicar la variabilitat de la resposta.

Tot allò que es refereix a aspectes inferencials, com ja s'ha avançat, està pendent de la validació de les premisses. Aclarim-ho: no posem en qüestió que la recta, obtinguda sigui la millor recta que aproxima (per mínims quadrats) els punts utilitzats, perquè aquesta recta sense considerar una fluctuació aleatòria al voltant, no és més que un model determinista. Però podríem dubtar de la validesa dels IC dels paràmetres si no fos admissible una variància σ^2 comuna a totes les mides.

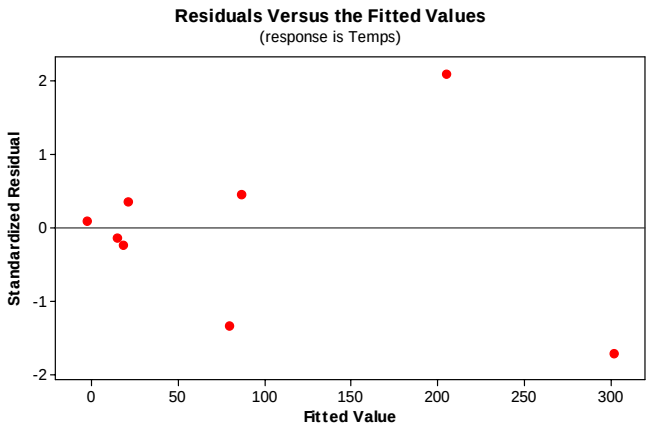


Figura 7.12 Residus del model contra prediccions

La figura 7.12 podria representar un argument a favor de l'heterocedasticitat, ja que els temps més grans semblen relacionats amb residus majors, si no fos perquè hi ha massa pocs punts perquè això pugui constituir una evidència convincent. Hem de prendre en consideració un altre tipus d'argument, i observar que seria bastant lògic que les fluctuacions del temps fossin proporcionals al temps (efecte multiplicatiu, v. apartat 5.4), i no un simple efecte additiu. Si el temps que requereix un fitxer està al voltant dels quatre minuts, no seria estranya una desviació de 30 segons, però sí que ho seria en un fitxer més petit. Des d'aquest punt de vista, el gràfic corrobora la fallida de l'homoscedasticitat, i potser no val la pena examinar les altres premisses.

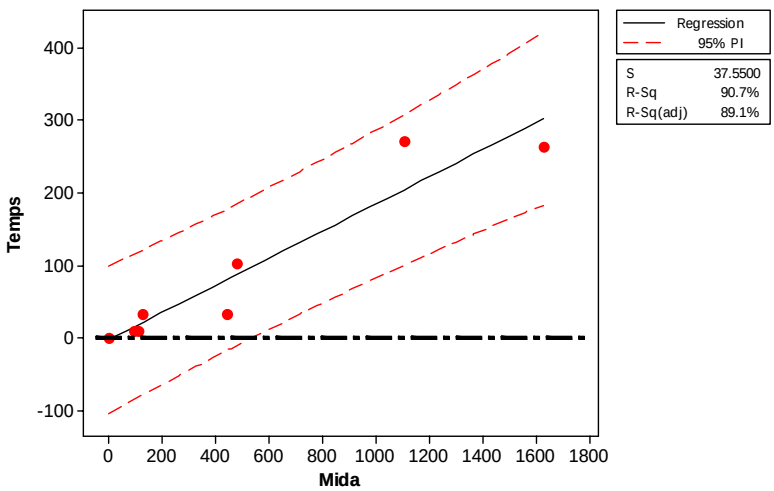


Figura 7.13 IC al 95% per a valors individuals de la resposta. La línia gruixuda discontinua és el nivell de temps 0, sobrepassat pel límit inferior de l'interval per causa de l'heterocedasticitat del cas.

Resulta bastant aclaridor de les conseqüències a les quals pot portar una premissa incomplida observar el gràfic de les previsions per a valors individuals (figura 7.13). Vegeu que, aplicant estrictament la teoria, la regió de confiança del temps probable per a fitxers amb mida inferior a uns 500 kB cobreix valors negatius, la qual cosa és bastant absurda i un senyal inequívoc que utilitzem, per a aquests casos, una variància excessiva.

Per mirar d'afrontar la nova situació, plantegem un nou model que tingui en compte el que hem vist. Tal com s'explica a l'apartat 5.4, treballarem amb el logaritme de les variables. Les noves dades són ara:

Taula 7.2 Dades transformades amb el logaritme

	1	2	3	4	5	6	7	8
log *M*	4.710	4.854	7.396	0.556	7.012	6.091	6.170	4.553
log *T*	2.303	3.497	5.577	–0.654	5.602	3.497	4.625	2.292

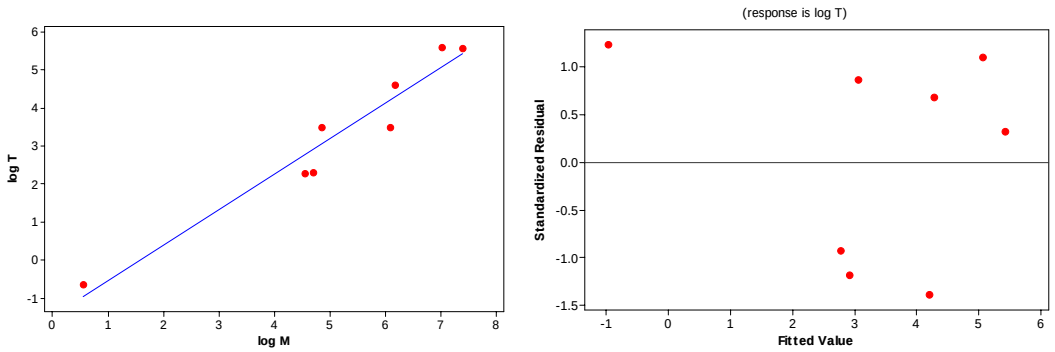


Figura 7.14 A l'esquerra, diagrama de dispersió amb logaritmes; a la dreta, diagrama dels residus contra el valor ajustat amb el nou model

El model transformat sembla coherent, ja que els nous punts es distribueixen al voltant de la recta sense cap anomalia sospitosa, i el mateix es pot dir del diagrama dels residus (v. gràfiques representades a la figura 7.14): en particular, el problema de l'heterocedasticitat sembla que ha desaparegut. Pel que es veu a la figura 7.15, tampoc no hi ha res en contra de la normalitat dels residus o de la independència de les observacions (encara que, amb tan poques observacions, seria molt difícil veure-hi algun signe de dependència, però sempre podem argumentar que els voluntaris que hi participaren no sabien res uns dels altres i, per tant, és difícil sospitar relacions entre les dades).

La recta estimada és $\log T = -1.49 + 0.935 \log M$, i la desviació residual val 0.553. Aquests valors no poden ni han de comparar-se amb els anteriors, per causa de la transformació aplicada. El que sí que en resulta afavorit és el nou coeficient de determinació, 93.9%, que vol dir que hem reduït la part de variabilitat no explicable pel factor mida d'un 9.3% a un 6.1%.

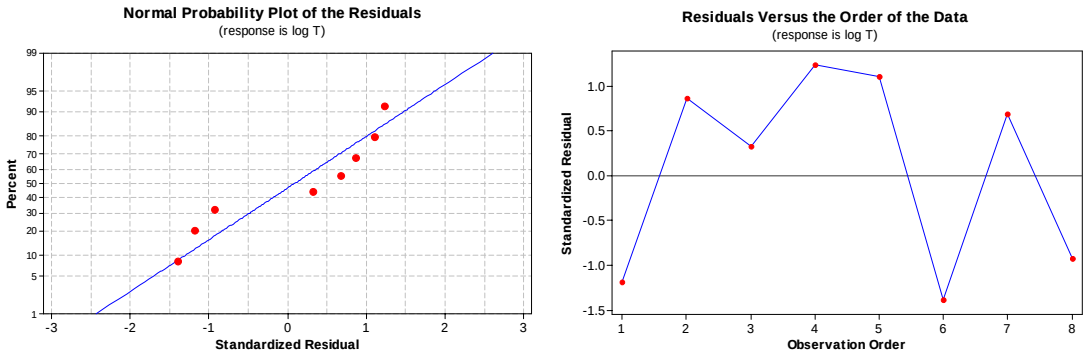


Figura 7.15 A l'esquerra, normal probability plot; a la dreta, diagrama dels residus en ordre d'observació

Per interpretar correctament els nostres estimadors, hem de desfer els canvis aplicats per tal de recuperar les variables originals. El model lineal, amb les variables log-transformades, es pot escriure com:

$$\log T = \beta_0 + \beta_1 \cdot \log M + \varepsilon$$

Si prenem l'exponencial a ambdós costats de l'equació:

$$\begin{aligned} \exp(\log T) = T &= \exp(\beta_0 + \beta_1 \cdot \log M + \varepsilon) \\ &= \exp(\beta_0) \cdot \exp(\beta_1 \cdot \log M) \cdot \exp(\varepsilon) \\ &= \kappa \cdot \exp(\log M^{\beta_1}) \cdot \xi \\ &= \kappa \cdot M^{\beta_1} \cdot \xi, \end{aligned}$$

on κ (kappa) és un nou paràmetre i ξ (xi) una variable aleatòria, el logaritme de la qual segueix una distribució N(0, σ). Ara podem comprendre millor què hem obtingut: β_1 representa un exponent per la mida del fitxer; κ és un factor de proporcionalitat, i ξ una pertorbació aleatòria que s'aplica multiplicant (no sumant) a l'expressió determinista. Podem veure com afecta aquesta fluctuació amb uns quants exemples recollits a la taula 7.3. A la primera columna, s'han posat quatre valors representatius de probabilitat: 0.2, 0.4, 0.6 i 0.8. A la segona columna, hem obtingut per a cada valor el percentil corresponent a una llei normal N(0, $\sigma = s = 0.553$). A la tercera columna, hem calculat l'exponencial del resultat de la segona columna, ja que ξ s'ha definit com a exp(ε). Finalment, la quarta columna expressa el resultat anterior en forma de percentatge: un valor inferior al 100% implica una pertorbació que deixaria l'observació per sota de la previsió, i un valor més gran que 100% deixaria l'observació per sobre de la previsió. Si el valor previst per l'equació determinista fos 100, posem per cas, amb una probabilitat de 0.20, el valor observat seria inferior a 62.8, o amb una probabilitat de 0.20 seria superior a 159. Queda clar que un efecte multiplicatiu no és simètric.

Taula 7.3 Exemples de l'efecte que suposa una pertorbació multiplicativa (prenent s com el valor de la desviació tipus de ε)

Probabilitat	Percentil ε	Percentil ξ	Efecte respecte de $\kappa \cdot M^{\beta_1}$
20%	–0.465	0.628	63% del valor (disminueix)
40%	–0.140	0.869	87% del valor (disminueix)
60%	0.140	1.150	115% del valor (augmenta)
80%	0.465	1.592	159% del valor (augmenta)

L'exponent val 0.935, que amb error tipus igual a 0.0974 no és significativament diferent del valor 1 (que és gratificant, perquè ens permet suposar que la relació entre el temps i la mida és fonamentalment lineal), i l'estimació de κ és exp(–1.4882) = 0.2258, similar al valor de b_1 estimat amb el primer model (almenys, dintre del IC) i equivalent a una taxa de transferència de 4.43 kB/s, bastant per sota de la teòrica taxa de 7 kB/s. Aquesta interpretació s'ha de prendre amb molta cura ja que, amb l'error tipus del terme independent (0.540), l'interval dels valors per κ és força ample.

Taula 7.4 Previsions amb el model multiplicatiu

Mida (kB)	Temps obs. (s)	IC 95% mitjana
1.7	0.52	(0.11, 1.26)
94.9	9.89	(9.66, 26.2)
111.0	10.00	(11.3, 30.1)
128.3	33.00	(13.0, 34.2)
442.0	33.00	(39.6, 113.5)
478.2	102.00	(42.3, 123.2)
1110.2	271.00	(82.8, 303.6)
1629.5	264.35	(111.1, 463.8)

Finalitzem el cas amb les previsions per a la mitjana del temps que es dedueixen amb el segon model, per als valors originals de *X*, que figuren a la taula 7.4. Queda molt clar que els resultats que s'obtenen tenen en compte un increment de la dispersió en relació amb la mida original. L'amplitud dels intervals pot semblar excessiva, però és una conseqüència de la petita grandària de la mostra.

7.4 Cas pràctic: evolució dels PC

Revisant revistes d'informàtica dels darrers anys, s'han recopilant les dades que apareixen en 20 anuncis, seleccionats de manera independent els uns dels altres i procurant cobrir de manera regular el període des dels anys vuitanta fins a 2005. Els anuncis són d'ordinadors tipus PC, amb processador

Intel, i part de les dades recollides figuren a la taula 7.5. Se suposa que es tracta d'ordinadors de gamma mitjana, representatius del model més comú al seu moment.

Taula 7.5 Dades dels PC, 1988–2005

	RAM (MB)	Disc dur (GB)	Freq. (MHz)		RAM (MB)	Disc dur (GB)	Freq. (MHz)
Ago-88	0.5	0.02	12	Jun-99	64.0	10.00	400
Mar-92	4.0	0.08	25	Sep-99	96.0	13.00	450
Oct-92	4.0	0.12	33	Jun-00	64.0	20.00	700
Mar-93	8.0	0.12	33	Oct-00	64.0	8.40	450
Des-93	4.0	0.17	50	Oct-00	128.0	30.00	733
Jun-94	8.0	0.34	66	Mar-01	128.0	40.00	1000
Gen-95	8.0	0.42	66	Des-01	128.0	40.00	1800
Jul-96	16.0	1.20	100	Feb-02	256.0	40.00	2000
Nov-97	32.0	4.30	233	Mai-05	512.0	200.00	3200
Feb-98	32.0	4.30	266	Mai-05	1024.0	320.00	3500

Estudiem ara l'evolució de la memòria RAM que porta l'equip en funció de la data. Les dades del temps es modifiquen per tal de facilitar-ne la manipulació: són valors numèrics, no necessàriament enters, encara que s'interpretaran en unitats d'anys, i el valor 0 representa el gener de 2000 (així, la data de juny de 1999 queda com –0.5151). A la figura 7.16, veiem el diagrama de la variable *RAM* respecte del temps.

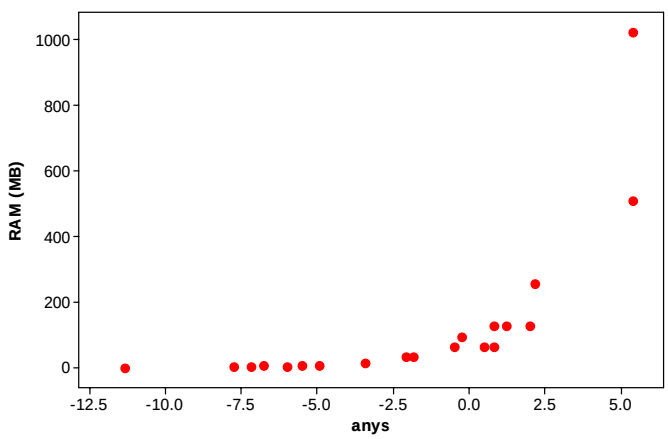


Figura 7.16 Diagrama de dispersió de la memòria RAM de l'ordinador contra el temps

És molt clar que no es pot admetre un increment lineal en el temps, perquè tothom sap que la memòria del PC creix molt ràpidament, a mesura que avancen els mitjans tecnològics per fer mòduls cada cop amb més capacitat. De fet, es tracta del típic creixement exponencial, que ens permet afirmar que la capacitat dels ordinadors es duplica amb certa regularitat (obtenir una estimació de la longitud del cicle serà un dels nostres objectius en aquest cas pràctic).

El primer pas que fem és transformar les dades de la variable resposta. El primer motiu es basa en el que hem dit: com que la variable *RAM* creix exponencialment en el temps, podem esperar que el seu logaritme creixi linealment, de manera que un model lineal sigui aplicable. El segon motiu és que, tal com hem vist abans, hi ha un problema d'heterocedasticitat: en èpoques recents, les variacions de memòria en termes absoluts són molt grans, comparades amb altres moments. En canvi, les variacions en termes relatius són sempre les mateixes (el doble, la meitat...). La figura 7.17 ens en mostra el resultat, on sembla que la transformació ha funcionat satisfactòriament.

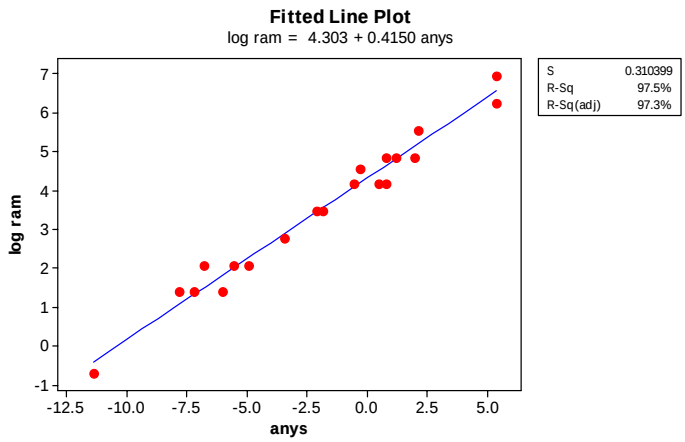


Figura 7.17 Diagrama de dispersió del logaritme de la memòria RAM contra el temps, i recta estimada

El coeficient R^2 és molt bo, el 97.5% de la variabilitat explicada pel temps, cosa que ens permet realitzar previsions bastant fiables. L'equació de la recta és log *RAM* = 4.303 + 0.4150 *ANYS*, que es pot escriure també, desfent la transformació, com *RAM* = exp(4.303)·exp(0.4150 *ANYS*). D'aquesta expressió, és fàcil deduir-ne el cicle mitjà de duplicació de la memòria, Δ. Sigui *Y* el valor predit en el moment *X*, i sabem que Δ anys més tard la memòria RAM és el doble:

$$Y = \exp(4.303)\cdot\exp(0.4150\ X)$$
$$2Y = \exp(4.303)\cdot\exp(0.4150\ (X+\Delta))$$

Dividint la segona equació per la primera s'obté:

$$2 = \exp(0.4150\ (X+\Delta)) / \exp(0.4150\ X) = \exp(0.4150\ (X+\Delta) - 0.4150\ X) = \exp(0.4150\ \Delta)$$
$$\log 2 = 0.4150\ \Delta \rightarrow \Delta = \log 2/0.4150 = 1.6703$$

Això vol dir que aproximadament cada 20 mesos es duplica la memòria dels PC del mercat. Evidentment, el resultat prové d'una mostra limitada, i està subjecte a l'error present en qualsevol procés de mostreig (a part dels possibles errors no estadístics: hi ha hagut un biaix en la selecció dels anuncis? Però aquest tema l'hem de deixar de banda, mentre haguem de mantenir la confiança en la bondat del sistema escollit). Veiem que només hem utilitzat l'estimador que correspon en el model lineal al pendent, i que aquest té un error estàndard igual a 0.01571. Però ens trobem amb una dificultat per estimar la variabilitat de l'*estimador* Δ: el pendent b_1 es troba dividint al denominador. Seria molt més senzill que estiguéssim tractant un estimador proporcional a b_1, ja que d'aquesta manera en podríem deduir la dispersió.

Una aproximació que sembla raonable és la següent. Trobem l'IC per a β_1 i, a continuació, li apliquem l'operació anterior. D'aquesta manera, encara que no tinguem la dispersió de l'estimació del cicle, tenim una informació que ens diu, amb probabilitat alta, quant pot ser aquest valor.

$$\text{IC}(95\%, \beta_1) = 0.4150 \pm t_{18,\,0.975}\, s_{b_1} = 0.4150 \pm 2.101 \cdot 0.01571 = (0.3820,\, 0.4480)$$
$$\text{IC}(95\%, \Delta) \approx (\log 2 / 0.4480,\, \log 2 / 0.3820) = (1.547,\, 1.815)$$

El resultat precedent s'expressa en anys. Llavors, podem parlar, multiplicant per 12, d'un interval que se situa entre els 18.5 i els 22 mesos.

No hem comentat res encara de la validació de les premisses. Podem confiar en el disseny perquè ens han dit que la selecció dels anuncis ha procurat ser transparent per tal de garantir la independència entre les observacions. A la figura 7.18, veiem plegats tots els gràfics necessaris per analitzar les premisses. Les poques dades disponibles fan que l'histograma no resulti molt clar, però la recta de Henry sembla impecable. Quant als residus contra les prediccions, no hi ha res a dir en contra d'una possible manca de linealitat o fallida de l'homoscedasticitat.

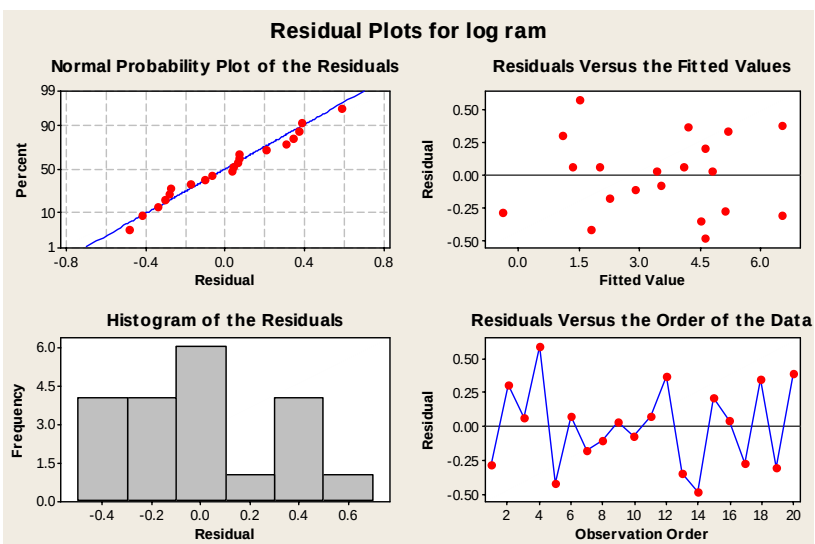


Figura 7.18 Validació de les premisses del model amb el logaritme de la memòria com a resposta.

Només podem admetre una crítica. La memòria RAM sempre és una potència de 2: 256, 512, 1024, etc. (encara que es poden combinar targes de mides diferents, però sempre en nombre limitat, i no és gaire freqüent); per tant, no es pot considerar una variable contínua. El mateix s'aplica a la seva transformada, encara que no ho sembli perquè pren valors irracionals. El cas és que la variable log *RAM* no es pot considerar normal, perquè un valor real sempre s'ha d'adaptar al logaritme d'una potència de 2. Això afecta especialment les previsions. Per exemple, utilitzem el model per fer una estimació de la memòria que típicament portarà un ordinador a l'any 2008 (temps=8). Sense entrar en el detall del càlcul, trobem que l'IC al 95% per a la mitjana del logaritme de la RAM és (7.2628, 7.9827), és a dir, entre 1426 i 2930 MB. Aquest resultat pot tenir sentit, perquè val per a la mitjana, i aquest paràmetre no té per què ser discret, però quan trobem el resultat per observacions individuals, obtenim —en MB— l'interval (971, 4305). Òbviament, podem arrodonir, i dir que el PC comú de 2008 tindrà entre 1 i 4 GB de memòria, però aquesta informació ja no garanteix el nivell de confiança del 95%.

Deixem com a exercici les anàlisis per a les altres dues variables resposta, capacitat del disc dur i freqüència del processador. Tracteu de respondre aquesta qüestió: durant l'any 2005, un canvi tecnològic en el disseny dels processadors va fer que la cursa que en feia incrementar el rendiment a base d'augmentar-ne la freqüència s'aturés, i que es busqués més velocitat per altres sistemes. Es nota el canvi de tendència en les dades? Com serien els processadors de l'any 2008 si tot hagués continuat igual?

7.5 Problemes

1. *El problema 1 del tema capítol continua ara.*

   a) Per aprofitar el model obtingut, el grup decideix predir la grandària en Kb d'una imatge en format GIF, sabent que ocupa 40 Kb en format JPG. Calculeu la predicció i doneu un interval de confiança al 95% per aquesta a predicció, aplicada a una imatge en concret.

   b) Finalment, un cop lliurada la pràctica, el professor de laboratori els demana si han fet la validació del model, perquè, si no poden validar el model, totes les conclusions anteriors són qüestionables. Feu la validació corresponent a partir dels gràfics de la figura 7.20 i indiqueu quina/quines premisses s'estudien a cada gràfic.

2. *A continuació, teniu unes qüestions que completen el problema 2 del capítol anterior.*

   a) Un enginyer informàtic d'aquesta empresa, titulat recentment per la FIB, comenta al seu cap que hauria de transformar aquestes variables, treballant amb els logaritmes neperians de les variables E i F, respectivament. La representació gràfica de les noves variables ln$E$ i ln$F$ és a la figura 7.19. Creieu que tenen sentit les transformacions que s'han proposat? Per què?

   b) Independentment de la vostra resposta a la pregunta anterior, si suposem que hi ha relació lineal entre aquestes dues variables (ln$E$ i ln$F$), quina relació es dedueix entre les variables E i F?

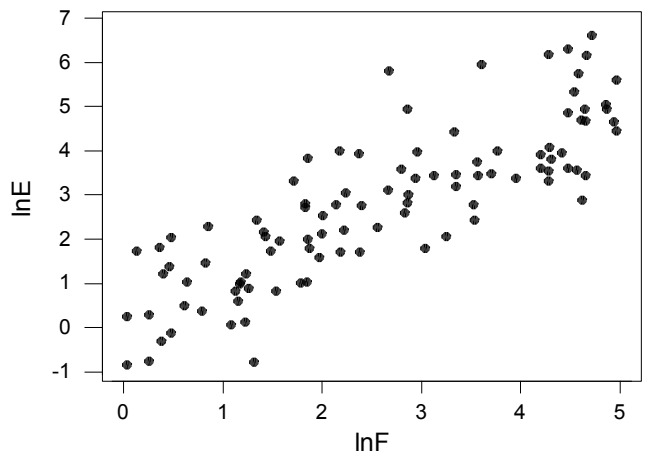


Figura 7.19

c) Quan trigaran, de mitjana, els projectes que tenen 50 punts funció? Calculeu un interval de confiança al 95% per a aquesta estimació. ($\overline{\ln F} = 2.5$)

3. *Continua del problema 3 del capítol anterior.*

La figura 7.21 presenta els gràfics dels residus (de la regressió entre la diferència *A*–*B* de temps de CPU i el nombre d'usuaris) amb el valor predit i amb l'ordre d'obtenció de les dades.

a) El gràfic dels residus amb el valor predit (gràfic de l'esquerra, a dalt): què permet avaluar? Quina importància estadística té? A quina conclusió s'arriba?

b) El gràfic dels residus amb l'ordre d'obtenció de les dades (segon gràfic, a la dreta): què permet avaluar? Quina importància estadística té? A quina conclusió s'arriba?

c) Quin és el tercer gràfic? És el núvol de punts i la recta de regressió? Com s'interpreta? A quina variable s'aplica? Quina o quines premisses s'estan avaluant? En aquest cas, què en podeu concloure? Hi ha alguna altra manera de comprovar les mateixes premisses que intervenen a aquest gràfic?

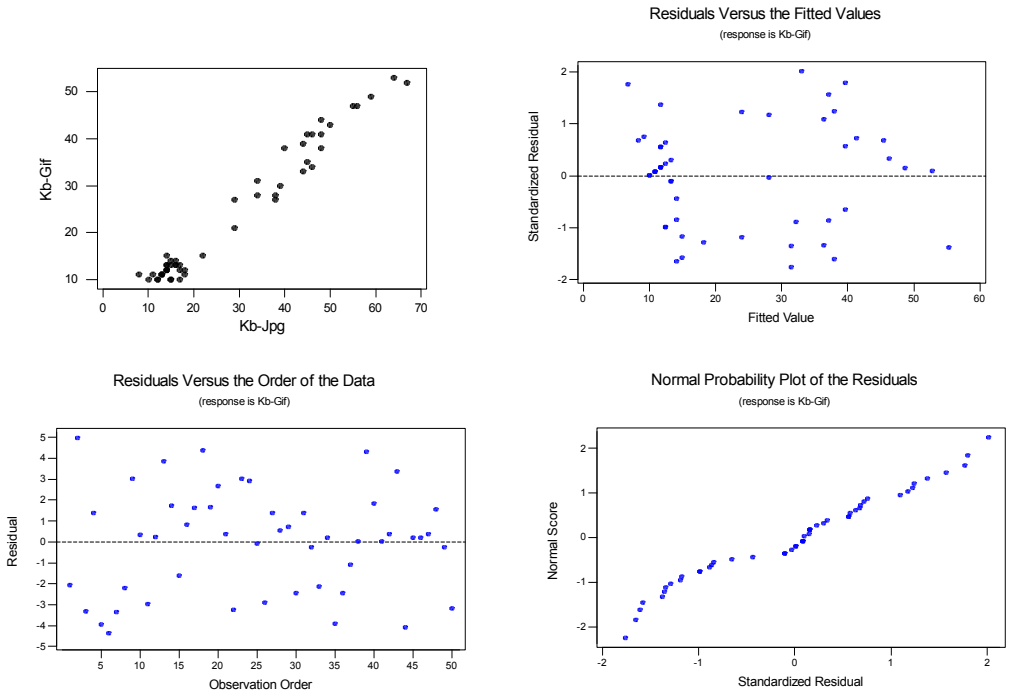


Figura 7.20

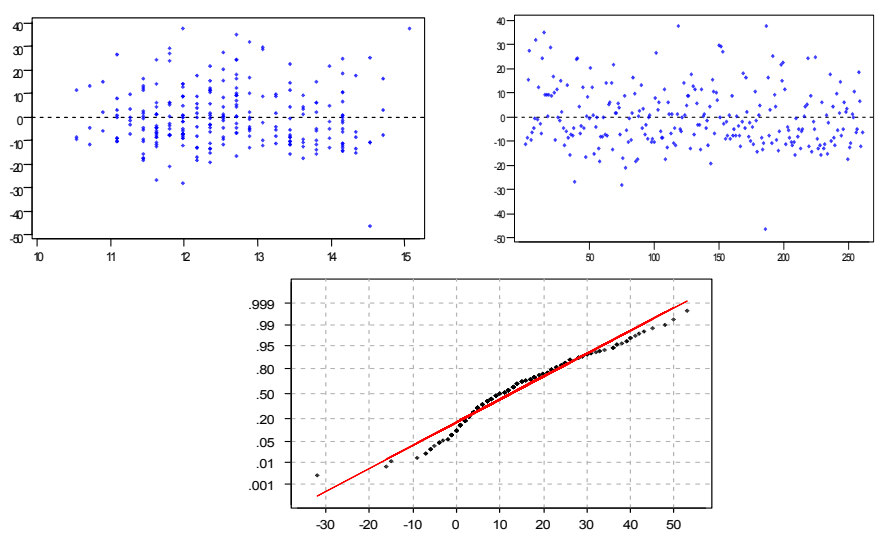


Figura 7.21

7.6 Solució dels problemes

1.

a) *Kb-Gif*(40) = 0.0902 + 0.82313 · 40 = 33.01

$$IC_{95\%}(Y_i) = \hat{Y}_i \pm t_{n-2,0.975} s \sqrt{1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2}}$$

$$\text{IC}_{95\%}(Y(40) = 33.01 \pm 2.01 \cdot 2.511 \cdot \sqrt{1 + 1/50 + (40 - 28.78)/(4 \cdot 17.086^2)} = [27.9,\ 38.1]$$

b) Premisses:

Linealitat (relació lineal entre ambdues variables): a la figura 7.20, a dalt i a l'esquerra (plot *X*-*Y*) s'observa relació lineal entre les variables. També s'observa linealitat en al gràfic de la dreta (residus estand. vs. prediccions)

Normalitat (residus segueixen la distribució normal): a la part inferior de la figura 7.20, amb el gràfic de la dreta (*normal probability plot* dels residus) es podria acceptar la normalitat dels residus.

Homoscedasticitat (igualtat de variàncies per a diferents valors de *X*): a dalt a la dreta de la figura 7.20 (residus estand. vs. prediccions), s'observa que la variabilitat es manté més o menys constant. També s'observa al gràfic (plot *X*-*Y*).

Independència: Al gràfic (residus vs. ordre) no s'observa cap patró que pogués indicar cap grau de dependència entre les observacions.

El model sembla força vàlid perquè compleix més o menys les premisses necessàries i $R^2 = 97\%$

2.

a) Té sentit aplicar aquestes transformacions si es vol utilitzar la recta de regressió, ja que ara sembla que es compleixin les premisses de *linealitat* (es pot observar amb claredat la possible recta que passaria per la meitat del núvol de punts) i *homoscedasticitat* (la variància es manté aproximadament constant al voltant de la recta per a diferents valors de ln*F*. Aquesta transformació òbviament no transgredeix la hipòtesi d'independència i és possible que ara es compleixi la hipòtesi de normalitat. Aquesta darrera afirmació no es pot contrastar fins a poder realitzar l'anàlisi dels residus.

b) Si existeix relació lineal entre ln*E* i ln*F*, vol dir que estem treballant amb el model

$$\ln E = \beta_0 + \beta_1 \ln F + \varepsilon_i.$$

A més, a l'apartat 3 hem acceptar que $\beta_1 \cong 1$ $\beta_0 \cong 0$; per tant, la relació es pot simplificar a $\ln E = \ln F + \varepsilon$, de la qual es dedueix que la relació entre *E* i F és $E = \exp(\ln F + \varepsilon) = F \cdot \exp(\varepsilon)$. És a dir, en aquest cas l'error no està vinculat de format additiva, com en el model de regressió lineal simple, sinó que actua de forma multiplicadora.

c) En primer lloc, cal treballar amb el logaritme neperià dels punts funció, és a dir, ln(50)=3.91
A partir de la recta obtinguda al gràfic inferior de la figura 7.20, podem calcular la previsió de l'esforç (en logaritmes):

$$\hat{Y}_0 = 0.2549 + 0.97033 \cdot 3.91 = 4.05$$

L'interval de confiança calculat al 95% per a aquesta previsió l'obtenim a partir de

$$\hat{Y}_0 \pm t_{n-2,0.975} s \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}}$$, que en el nostre cas és

$$4.05 \pm 1.96 \cdot 0.9654 \sqrt{\frac{1}{100} + \frac{(3.91-2.5)^2}{99 \cdot 2.2468}} = 4.05 \pm 0.273$$

(Observeu que $\sum (X_i - \overline{X})^2 = (n-1)s_X^2$ i el procediment per calcular la variància estimada de *X* ja ha estat explicat a l'apartat *b*, a la solució publicada al capítol anterior.

Finalment, aquest interval és (3.777, 4.323) però nosaltres volem l'interval en les unitats originals, és a dir, en hores i no en ln(hores).

El resultat final és (exp(3.777), exp(4.323)) = (43.68, 75.41).

Observeu que aquest resultat és coherent amb la figura 6.10.

3.

a) Homoscedasticitat: permet ponderar per igual els residus. Els residus es distribueixen amb una dispersió similar, per a qualsevol valor de la variable a l'eix d'abscisses. S'arriba a la conclusió que existeix homoscedasticitat.

Linealitat: permet utilitzar un model lineal per a *X*. S'arriba a la conclusió que existeix linealitat, ja que no hi ha indicis (curvatures del núvol) que els residus depenguin de forma no lineal de *X*.

b) Independència: permet dir que el residu no conté informació sobre altres observacions. Com que no s'hi aprecia cap estructura particular, s'arriba a la conclusió que existeix independència.

c) És el *normal probability plot*, que representa la recta de Henry. No té res a veure amb el núvol de punts ni la recta de regressió; es construeix comparant cada residu estandaritzat amb un valor equivalent que prové d'una distribució normal. Per tant, serveix per posar a prova la normalitat dels residus i, en conseqüència, de les observacions (variable diferència *A*–*B*). En el nostre cas, es veu que els punts no se separen gran cosa de la recta de referència, ni tan sols a les cues; per tant, es pot defensar la normalitat d'aquesta variable (condicionada pel nombre d'usuaris). Com a alternativa simple, atès que tenim moltes observacions, es podria haver emprat un histograma dels residus.

8 Proves de Pearson

8.1 Proves d'ajustament

Les proves d'ajustament de Pearson contrasten si les dades disponibles provenen d'una distribució determinada. La idea és comparar les freqüències observades a les dades per a determinats esdeveniments (o classes) amb les freqüències especificades pel model teòric que es posa a prova. S'assumeix que els individus són independents, que la variable observada segueix una única distribució, i que un resultat observat correspon només a una classe de les que s'han definit.

Exemple 1. D'acord amb la planificació? En una empresa de desenvolupament de material electrònic tenen distribuïts els recursos per tal d'atendre una demanda de serveis de reparació classificats en quatre categories, *A*, *B*, *C* i *D*, segons la seva complexitat. S'espera que el 10% dels serveis siguin de tipus *A* (el més complex), el 20% de tipus *B*, el 30% de tipus *C* i el 40% de tipus *D*.

Al darrer any s'han atès 553 serveis de reparació, dels quals 78 es van classificar com a *A*, 107 com a *B*, 145 com a *C* i la resta com a *D*. Es pot dir que la classificació realitzada és coherent amb l'establerta?

En aquest exemple, hi ha quatre classes, *A*, *B*, *C* i *D*, per a les quals s'ha definit una distribució de probabilitats 0.10, 0.20, 0.30 i 0.40, respectivament. Per altra banda, disposem de les dades provinents dels serveis atesos l'any passat —que podem suposar independents entre si—, que representen el 14%, el 19%, el 26% i el 40% de les categories de complexitat, respectivament. Cal estudiar si les discrepàncies observades es poden atribuir només a l'atzar, o si hi ha motius per creure que el model no descriu acuradament la distribució real.

Aquest contrast serveix tant per a distribucions discretes com per a contínues, però, per aplicar-ho a una llei continua, els valors reals s'han de categoritzar. El nombre de classes, o categories, no és determinant, encara que és preferible que s'escullin de manera que les probabilitats esperades per a cada classe siguin semblants (o no gaire diferents), i és convenient que a cada classe hi hagi un nombre mínim d'observacions.

Quan es disposa d'una mostra aleatòria simple X_1, X_2, ..., X_n, on n és prou gran per a la prova de Pearson (almenys 25, segons Peña),[19] i es vol comprovar si un model de probabilitat és vàlid per a la mostra, hem d'agrupar les dades en les k classes escollides.

Anomenem O_i la freqüència observada a la mostra de la classe i, és a dir, quantes observacions han caigut a l'interval que defineix la classe i (pot ser un punt, per a variables discretes). E_i serà la freqüència esperada de la classe i, és a dir, el nombre que es podria esperar que estigués en aquella classe, d'acord amb la probabilitat p_i que el model assigna al seu interval. És clar que $E_i = n\, p_i$.

La discrepància entre les freqüències reals i les esperades es pot calcular globalment amb l'expressió:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

que es distribueix com una khi quadrat si el model és correcte. Els graus de llibertat són k–1 quan les probabilitats p_i són conegudes abans de prendre la mostra, o k–r–1 si és necessari estimar r paràmetres amb les dades per arribar a conèixer aquestes probabilitats.

Observeu que cada diferència es divideix per un factor relacionat amb el nombre esperat de casos a la classe. Això és així per ponderar, de forma adecuada, el fet que una classe amb probabilitat important pot donar per atzar una discrepància més gran que una altra classe menys probable. Per aquesta raó, parlarem de *diferència estandarditzada* a la classe i:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

Nota. Si podem considerar que, per a cada classe, el nombre d'observacions efectuades segueix una llei de Poisson, llavors aquest tractament és el correcte, ja que per a una llei de Poisson la desviació tipus és igual a l'arrel quadrada del valor esperat. Si, a més, el nombre esperat no és un valor petit, la diferència estandarditzada es distribueix aproximadament com una N(0,1), i es pot justificar que la suma dels termes al quadrat segueixi una llei de khi quadrat.

El contrast és inherentment unilateral. Si l'estadístic X^2 és massa gran, indica que hi ha diferències no explicables per la casualitat entre els valors previstos i els trobats. Per tant, el model teòric no és l'apropiat per descriure la distribució de les dades. Es pot calcular un p-valor per ratificar la conclusió, que es troba amb:

$$\mathrm{P}\left(X^2 > \chi^2_{gl}\right)$$

19 Daniel Peña. *Fundamentos de estadística*. Alianza Editorial, 2005.

Exemple 1 (continuació). A la taula següent, mostrem com es determinen els valors que condueixen a l'estadístic de Pearson.

Classe	Prob. p_i	Nombre esperat (sobre 553) E_i	Nombre observat O_i	Diferència estandarditzada	Dif. estand. al quadrat
A	0.10	0.10·553=55.3	78	3.0526	9.3181
B	0.20	0.20·553=110.6	107	–0.3423	0.1172
C	0.30	0.30·553=165.9	145	–1.6226	2.6330
D	0.40	0.40·553=221.2	223	0.1210	0.0146
				Suma:	12.0829

Destaca molt el resultat de la classe *A*. Sembla que el valor observat és massa lluny del que s'esperava: així, tenim la major diferència estandarditzada (3.05), i aquest valor al quadrat representa més del 77% del valor de l'estadístic X^2, que és 12.08. Si examinem la distribució de referència de l'estadístic per ubicar el resultat, una llei de khi quadrat amb 3 graus de llibertat, ens trobem que la probabilitat de superar el valor 12.08 és només 0.0071. En conclusió, hi ha evidències per creure que no s'està complint amb el protocol; sembla que hi ha més assignacions a la categoria *A* del que estava previst.

Tingueu present que un contrast sobre la bondat de l'ajustament es planteja com una prova de significació. Primer tenim una hipòtesi nul·la que posar a prova:

H_0: les dades de la mostra vénen d'una població amb probabilitats p_1, ..., p_k

Per a aquesta hipótesis, existeix, de forma complementària, la hipòtesi alternativa que constata que les dades provenen d'una altra població. Aquesta hipòtesi no pot alterar la lateralitat de la prova, ja que hem dit que, per construcció, és sempre unilateral per la dreta.

Tot seguit, es considera l'estadístic adient de la prova, que és l'estadístic de Pearson i que, sota H_0, segueix una llei de khi quadrat amb k–1 graus de llibertat.

Si es vol considerar un risc α, llavors es pot definir una regió crítica que condueix al rebuig de la hipòtesi nul·la. Per exemple, per al cas anterior i un risc del 5%, la regió de rebuig és $(7.815, +\infty)$.

A continuació, hem de valorar les premisses de la prova: necessitem una MAS, i s'ha de comprovar que les freqüències esperades per a cada classe no siguin massa petites. Potser en aquest pas s'ha de revisar si les classes escollides són les més adients —per exemple, si es tracta de categoritzar una variable contínua.

Es troba el valor de l'estadístic X^2, a partir de les freqüències observades i esperades, i es determina el p-valor del resultat. Si aquest valor és petit (per exemple, menor que α), es rebutja H_0, o bé simplement es manifesta quantitativament la versemblança del resultat (per exemple, dient que 12.08 és un valor que es pot donar sota la hipòtesi nul·la 1 de cada 140 vegades). Habitualment, acompanyarem la conclusió formal amb un comentari més aclaridor (per exemple: "sembla que hi ha més assignacions a la categoria *A* del que estava previst").

Exemple 2. És una exponencial?

A la mateixa empresa d'abans es vol posar a prova si el temps fins a fallida d'un determinat component es distribueix, com es creia, d'acord amb una llei exponencial de mitjana 12 mesos. Es disposa d'una mostra de 80 temps, que han estat recollits per un sistema automàtic que, quan detecta la fallida, guarda en una base de dades el temps transcorregut des del moment de la posada en funcionament (se'n mostren els valors arrodonits, però es disposa del temps exacte amb precisió de minuts):

17	5	2	1	8	3	27	11	17	10	3	14	36	19	26	5	7	12	1	8
5	22	17	7	7	23	6	3	42	5	6	21	2	36	20	14	10	31	5	9
9	2	2	1	4	11	12	23	9	29	7	17	18	4	29	1	38	10	7	5
5	37	76	23	14	7	6	1	6	8	5	9	2	5	31	6	31	16	5	4

Prenem (arbitràriament) sis classes amb la mateixa probabilitat, 1/6. Llavors, hem de trobar els percentils de 16.67%, 33.33%,... En el cas d'una distribució exponencial, es tracta d'una operació senzilla:

$$1-e^{-\lambda x}=q$$
$$1-q=e^{-\lambda x}$$
$$\ln(1-q)=-\lambda x$$
$$x=-\ln(1-q)/\lambda$$

on q = 0.1667, 0.3333... i λ=1/12. Els percentils trobats per definir els intervals corresponents de cada classe, amb les freqüències respectives, es mostren a la taula següent:

Classe	Prob. p_i	Nombre esperat (sobre 80) E_i	Nombre observat O_i	Diferència estandarditzada	Dif. estand. al quadrat
[0, 2.19)	1/6	13.3333	10	–0.9129	0.8333
[2.19, 4.87)	1/6	13.3333	9	–1.1867	1.4083
[4.87, 8.32]	1/6	13.3333	20	1.8257	3.3333
[8.32, 13.18]	1/6	13.3333	12	–0.3652	0.1333
[13.18, 21.5]	1/6	13.3333	12	–0.3651	0.1333
[21.5, +∞)	1/6	13.3333	17	1.0042	1.0083
				Suma:	6.85

L'estadístic X^2 val 6.85. Prenent cinc graus de llibertat, el p-valor d'aquest resultat és 0.23, que indica que no tenim cap evidència per rebutjar la distribució suposada, una llei exponencial amb mitjana 12.

Exercici. Repetiu la mateixa prova però utilitzant només els valors arrodonits de la taula (és a dir, hem perdut la informació del nombre de dies, hores i minuts, una informació potser rellevant per a temps petits). N'obteniu el mateix resultat?

Exemple 3. La generació de nombres aleatoris normals.
Siguin $X_i \rightarrow$ U[-½, ½], i=1, ..., 12, independents, i sigui $Y = \Sigma\, X_i$.

Es pot comprovar fàcilment que l'esperança de Y és 0, i la seva variància és 1 (ja que $V(X_i)$=1/12). Llavors, es tracta de veure que Y és aproximadament N(0,1). Aquest senzill algorisme ha estat molt utilitzat als inicis de la computació amb ordinador per generar nombres pseudoaleatoris amb distribució normal. Els fonaments són clars: es basa en el teorema central del límit, que diu que la suma d'un gran nombre de variables independents segueix una distribució de probabilitat gairebé normal. El punt feble del mètode és que 12 potser no és un nombre molt alt, de manera que l'aproximació pot ser deficient en algunes parts, concretament a les cues. Aquest algorisme no és adequat si és vol fer servir per a simulacions intensives, que requereixen una gran precisió a tot el recorregut.

Suposem que hem generat amb un paquet estadístic dotze columnes amb distribució U[-½, ½], i que sumem el resultat (per construcció, totes les columnes són independents):

```
MTB > random 1000 c1-c12;
SUBC> unif -0.5 0.5.
MTB > rsum c1-c12 c13
```

Després, categoritzem els valors en classes, i trobem la probabilitat corresponent per una llei N(0,1):

```
MTB > Code (-9:-3) -3 (-3:-2) -2 (-2:-1.5) -1.5 (-1.5:-1) -1 (-1:0)0
&
CONT> (0:1) 1 (1:1.5) 1.5 (1.5:2) 2 (2:3) 3 (3:9) 4 C13 c14
```

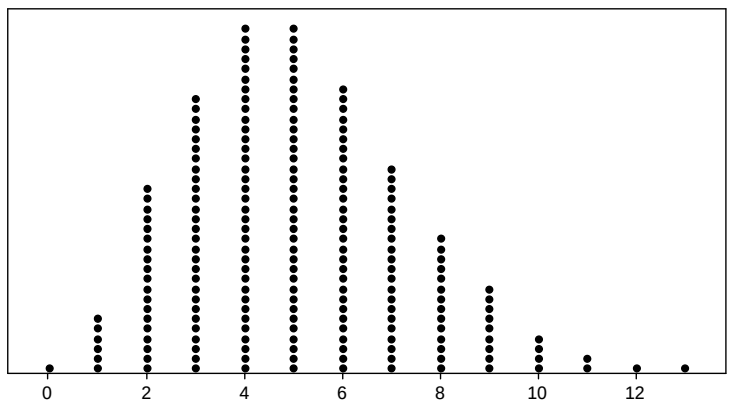
Classe	Prob.	Nombre esperat (sobre 1000)	Nombre observat	Diferència estandarditzada
< –3	0.001350	1.350	3	1.4202
[–3, –2]	0.021400	21.400	28	1.4267
[–2, –1.5]	0.044057	44.057	46	0.2927
[–1.5, –1]	0.091848	91.848	88	–0.4015
[–1, 0]	0.341345	341.345	349	0.4143
[0, 1]	0.341345	341.345	347	0.3061
[1, 1.5]	0.091848	91.848	83	–0.9232
[1.5, 2]	0.044057	44.057	43	–0.1593
[2, 3]	0.021400	21.400	13	–1.8159
> 3	0.001350	1.350	0	–1.1618

La diferència estandarditzada és la diferència entre el nombre observat i el nombre esperat d'ocurrències de la variable en cada classe, dividida per l'arrel quadrada del nombre esperat. La suma dels valors al quadrat és 10.09. Com que hi ha deu classes, el punt crític per rebutjar la prova de Pearson és $\chi^2_{9,\ 0.95}$, concretament 16.92. Per tant, no hi ha cap

evidència que permeti dubtar de la normalitat de les dades generades amb l'algorisme proposat (p-valor = 0.343).

Exemple 4. Poisson i binomial

Disposem d'un generador de nombres aleatoris per a la llei de Poisson, i hem obtingut una seqüència de 1000 valors. A la figura, veiem la distribució dels 1000 nombres.



Cada símbol representa fins a 5 observacions.

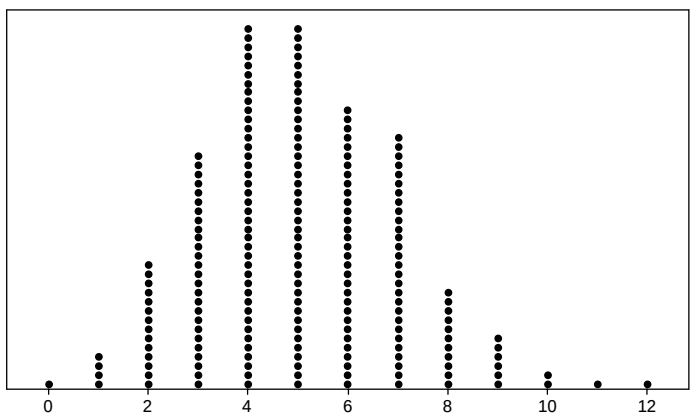
Aquesta mostra té com a mitjana 5.0190. Per simplificar el procés, i per evitar la inestabilitat que suposa considerar classes amb probabilitats molt baixes, agrupem els valors inferiors o iguals a 1, com també els superiors o iguals a 10. La taula mostra per a cada classe la seva probabilitat, calculada per una distribució de Poisson amb λ=5.0190 (donat per l'estimador mitjana com el valor més creïble per a aquest paràmetre). També mostra el nombre esperat i observat d'ocurrències, i la diferència estandarditzada per a totes les classes.

Classe	Prob.	Nombre esperat (sobre 1000)	Nombre observat	Diferència estandarditzada
≤ 1	0.039792	39.792	33	–1.0768
2	0.083268	83.268	91	0.8473
3	0.139308	139.308	140	0.0586
4	0.174797	174.797	174	–0.0603
5	0.175461	175.461	173	–0.1858
6	0.146773	146.773	145	–0.1464
7	0.105236	105.236	104	–0.1205
8	0.066023	66.023	66	–0.0028
9	0.036819	36.819	43	1.0187
≥ 10	0.032522	32.522	31	–0.2669

Podem observar que les diferències detectades són molt petites i, per tant, atribuïbles a l'atzar. La suma de quadrats val solament 3.06, de manera que sembla que el generador reprodueix amb prou fidelitat la distribució de Poisson. En aquest cas, el paràmetre utilitzat a la generació havia pres el valor 5, força semblant a la seva estimació.

Es diu que en alguns casos una distribució binomial es pot aproximar per la llei de Poisson. Això es pot aplicar quan la n de la binomial és prou gran i la p petita. Veurem ara amb un exemple que, si no fos d'aquesta manera, l'aproximació no seria bona.

Generem una mostra de 1000 valors procedents d'una distribució B(20, 0.25). Si l'aproximació fos acceptable, hauríem de contrastar la mostra amb una llei P($\lambda = np = 5$). Com abans, el valor del paràmetre serà substituït per l'estimació de la mitjana mostral.



Cada símbol representa fins a 5 observacions.

No sembla massa diferent de la distribució (realment poissoniana) de la primera part de l'exemple. Presenta una mitjana igual a 5.036. Recodifiquem per considerar les mateixes classes (les probabilitats són lleugerament diferents perquè ara la mitjana ha variat una mica).

Classe	Prob.	Nombre esperat (sobre 1000)	Nombre observat	Diferència estandarditzada
≤ 1	0.039232	39.232	21	–2.9108
2	0.082420	82.420	69	–1.4782
3	0.138356	138.356	129	–0.7954
4	0.174191	174.191	196	1.6525
5	0.175445	175.445	197	1.6274
6	0.147257	147.257	152	0.3909
7	0.105941	105.941	138	3.1148
8	0.066690	66.690	55	–1.4314
9	0.037317	37.317	30	–1.1977
≥ 10	0.033152	33.152	13	–3.5000

La suma de diferències estandarditzades al quadrat val 42.3, molt més enllà de l'admissible per a una distribució de khi quadrat amb 8 graus de llibertat (hem restat un grau de llibertat addicional pel càlcul de l'estimador de λ). És interessant observar que les diferències són negatives a l'inici i al final, cosa que indica que la distribució de Poisson tendeix a presentar més dispersió que la binomial (en efecte: la variància d'una llei Poisson amb λ=5 és 5 també, però la binomial emprada té variància $20 \cdot 0.25 \cdot 0.75 = 3.75$). Admetent que els nombres han estat generats correctament, podríem rebutjar que la llei de Poisson sigui una aproximació acceptable per aquesta forma de binomial.

8.2 Proves d'homogeneïtat i independència

L'estadístic de khi quadrat s'empra en un altre tipus de proves similars. En aquest cas, no es tracta de comprovar si les dades es distribueixen d'acord amb un model determinat, sinó comprovar si existeixen diferències en les distribucions de diverses poblacions, utilitzant les respectives mostres. Aquesta anàlisi rep el nom de *prova d'homogeneïtat*.

Considerant un nombre k de classes C_1, ..., C_k, s'obté per a cada una de les P poblacions el nombre d'observacions per a cada classe:

	C_1	...	C_i	...	C_k	
X_1	$O_{1,1}$		$O_{i,1}$			
...						
X_j	$O_{1,j}$		$O_{i,j}$			$n_{.,i}$
...				...		
X_P					$O_{k,P}$	
			$n_{i,.}$			n

La taula que es veu aquí s'anomena *taula de contingència*. Denotem per $n_{.,j}$ la grandària de la mostra j-èsima, i per $n_{i,.}$ el nombre total de casos que es troba a la classe i. Diguem que n denota el nombre total d'observacions. Es compleix que:

$$n_{.,j} = \sum_{i=1}^{k} O_{i,j}$$

$$n_{i,.} = \sum_{j=1}^{P} O_{i,j}$$

$$n = \sum_{i=1}^{k} n_{i,.} = \sum_{j=1}^{P} n_{.,j}$$

Si les P variables vinguessin de la mateixa distribució, tindria sentit estimar la probabilitat de caure a la classe i per a qualsevol grup; en cas que no es tingui la referència d'un model teòric. Aquesta

estimació és lògicament $n_{i,\cdot}/n$. Per tant, el nombre esperat d'observacions a la classe i per a la població j és: $E_{i,j} = n_{\cdot,j}\, n_{i,\cdot}/n$. L'estadístic de Pearson queda, doncs, com:

$$X^2 = \sum_{j=1}^{P}\sum_{i=1}^{k}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Les variables $O_{i,j}$ no són independents, ja que estan sotmeses a les restriccions anteriors (les sumes per files o per columnes equivalen a grandàries marginals $n_{\cdot,j}$ o $n_{i,\cdot}$ respectivament); per tant, quan és correcta la hipòtesi que la distribució és la mateixa per a tots els grups, l'estadístic X^2 segueix una llei de khi quadrat amb $(k–1)(P–1)$ graus de llibertat. Una restricció és redundant; per això, aquest nombre no val $kP–k–P$, sinó $kP–k–P+1$.

La prova que s'ha descrit és similar a una altra prova que també parteix d'una taula de contingència: és a dir, es requereixen dues variables categòriques per a les quals es disposa dels efectius d'una mostra per a cada intersecció de classes. A diferència de la prova d'homogeneïtat, les dues variables es consideren subjectes a l'atzar: per a cada individu, s'ha d'observar tant la classe de les files com la de les columnes (abans, en el moment d'escollir l'individu, ja sabíem a quina fila o grup pertanyia). Aquesta prova es diu *prova d'independència* perquè tracta de comprovar si les variables són independents. S'utilitza l'estadístic de Pearson de la mateixa forma; per tant, hi apliquem la mateixa mecànica de càlcul.

Exemple 5. Com et connectes?
Un estudi analitza l'ús d'internet a la llar en diferents àmbits territorials, i estudia si el tipus de connexió varia segons aquest factor. Considera tres àmbits diferents, per als quals determina, per a una selecció aleatòria de llars amb connexió a internet, quin mitjà fa servir. Les dades següents en mostren el resultat:

	Mòdem 56k	ADSL 128k	ADSL 256k	ADSL 512k	ADSL ≥1024k
Àmbit *A*	56	13	72	72	48
Àmbit *B*	80	33	160	132	57
Àmbit *C*	37	26	63	82	26

Construirem una taula similar, però amb els efectius esperats, per al cas que els cinc tipus analitzats es distribueixen de la mateixa manera a *A*, *B* i *C*, i respectant els marginals obtinguts.

	Mòdem 56k	ADSL 128k	ADSL 256k	ADSL 512k	ADSL ≥1024k	
Àmbit *A*	47.18	19.64	80.45	78.00	35.73	261
Àmbit *B*	83.52	34.76	142.41	138.07	63.24	462
Àmbit C	42.30	17.61	72.13	69.93	32.03	234
	173	72	295	286	131	957

Què diu aquesta taula? En el territori que és anomenat *A*, s'han trobat 56 llars que utilitzen una connexió amb mòdem de 56k, de les 261 sondejades en aquest territori. Com que són 173, del total de 957 llars analitzades, les que utilitzen aquest tipus de connexió (un 18%),

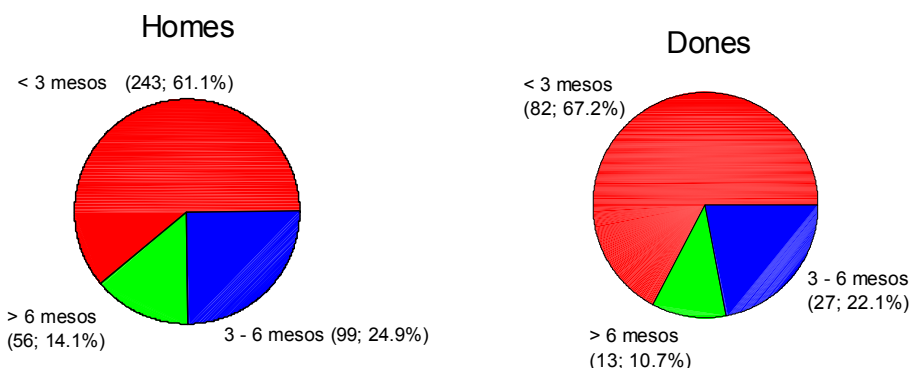
si hem d'aplicar la mateixa proporció al territori *A* haurien de ser 47.18 les llars trobades. Òbviament, aquest número no enter no és possible, però és l'ideal per respectar fidelment la proporció global. El mateix procediment es repeteix a cada cel·la.

Són acceptables les diferències trobades, des del punt de vista d'un mostreig aleatori? Si fos així, l'estadístic X^2 prendria un valor moderat, respecte del que es considera per a la distribució de referència, en aquest cas, una khi quadrat amb 8 graus de llibertat (4·2). Per a un risc del 5%, el límit es troba al valor 15.507. Amb les dades del cas, X^2 val 21.79, que és a prop del percentil 99.5%; per tant, es pot rebutjar que els mitjans de connexió a internet als tres territoris es distribueixin de la mateixa forma, amb un p-valor proper a 0.5%.

Nota. Es tracta d'una prova d'homogeneïtat o d'una prova d'independència? Això depèn de com s'hagin obtingut les dades. Si s'han generat primer les dades a l'atzar, i per a cada llar s'ha determinat l'àmbit al qual pertany i el tipus de connexió, és una prova d'independència i, estrictament parlant, podem concloure que s'ha detectat que territori i tipus de connexió són dependents. Amb aquest disseny, els valors marginals de l'àmbit permetrien estimar la proporció de llars per a cada territori (respectivament, 27%, 48% i 24%). Tanmateix, probablement aquest no és el problema que es vol resoldre, i els responsables de l'estudi poden pensar que no cal tenir dades en nombre proporcional a les poblacions respectives —potser una dada ben coneguda per mitjà del padró—; ben al contrari: un àmbit molt majoritari reduiria a efectius bastant pobres els valors dels àmbits menys nombrosos, cosa que complicaria la potència de l'estudi. Per aquesta raó, és molt probable que el disseny d'aquest estudi hagi estratificat la recollida de dades per territoris, determinant una grandària de mostra adequada i no necessàriament equivalent a la seva població, i després hagi triat aleatòriament les llars per tal d'incloure-les a l'estudi. Amb aquest disseny, hauríem de parlar de prova d'homogeneïtat.

8.3 Problemes

1. Una enquesta realitzada entre nous titulats de la UPC revela que el temps que traguen a trobar feina es distribueix, entre homes i dones, de la manera següent:



L'anàlisi de l'informe final conclou amb la frase següent: "*Les dones enquestades triguen menys temps a trobar feina que els homes.*" Suposem que la informació anterior procedeix d'una mostra representativa de la població dels titulats de la UPC.

Descriviu un mètode d'inferència per posar a prova l'afirmació: "*Les dones titulades a la UPC triguen el mateix temps a trobar feina que els homes*."

a) Hipòtesi a contrastar?
b) Estadístic suposat sota la hipòtesi nul·la?
c) Valor observat de l'estadístic?
d) Punt crític per a $\alpha = 0.05$?
e) Conclusió formal?
f) Coincidiu amb l'autor de l'informe?

2. Un centre d'ensenyament tècnic superior estava preocupat per l'escàs nombre de dones que accedien de nou al centre. A partir del curs 2000-2001 va iniciar tot un seguit d'accions per aconseguir incrementar aquest nombre en l'accés als cursos posteriors. Els resultats són a la taula següent:

Curs	Dones	Homes
2000-2001	78	572
2001-2002	111	539
2002-2003	91	559

Creieu que la política realitzada per aquest centre ha tingut efecte? Per respondre aquesta qüestió objectivament cal que:

a) Plantegeu el test d'hipòtesi adient.
b) Digueu quines premisses són necessàries.
c) N'efectueu els càlculs.
d) Prengueu una decisió amb un risc α del 5%. Que li diríeu al centre?

8.4 Solució dels problemes

1.

a) *Hipòtesi nul·la*: és independent el gènere del titulat del temps esmerçat a trobar feina, vs. *hipòtesi alternativa*: no és independent.

b)

$$\hat{X}^2 = \sum_{\forall i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \text{ on } E_{ij} = (T_{\text{fila}} \cdot T_{\text{columna}})/T_{\text{total}}$$

distribució de khi quadrat amb 2 graus de llibertat

c) Calculem els valors dels efectius reals i esperats (entre parèntesis)

	1	2	3	All
H	243 (248.75)	99 (96.44)	56 (52.81)	398
D	82 (76.25)	27 (29.56)	13 (16.19)	122
All	325	126	69	520

$$\hat{X}^2 = \frac{(243 - 248.75)^2}{248.75} + \frac{(99 - 96.44)^2}{96.44} + \ldots = 1.677$$

d) El punt crític per a una khi quadrat amb 2 g.l. és 5.99.

e) Com que l'estadístic cau dins de la regió d'acceptació (0… 5.99) no tenim arguments per refusar la hipòtesi nul·la.

f) No puc refusar la independència entre gènere i temps: no és evident que les dones triguin menys que els homes a trobar feina i, per tant, discrepem de l'opinió de l'autor de la frase, si és que té intenció inferencial

2.

a) Es tracta de veure si la proporció de dones que accedeixen de nou al centre es manté constant al llarg dels anys (en aquest cas, les accions no hauran tingut efecte), o be van canviant.

- Formalment, es planteja el test d'hipòtesi d'homogeneïtat:

  H_0: $\pi(J_j|I_i) = \pi(J_j|I_{i'}) \ \forall\, i, i', j$
  (la proporció de dones es manté constant al llarg dels cursos)

  En aquest test:
  J és la variable qualitativa: dones i homes
  I la variable cualitativa: curs

  H_1: $\exists\, j$ tal que $\pi(J_j|I_i) \neq \pi(J_j|I_{i'}) \ \forall\, i, i'$
  (és a dir, en algun curs concret la proporció ha canviat)

- L'estadístic que s'ha d'utilitzar és

  $$\hat{X}^2 = \sum_{\forall i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$ on O_{ij} són els valors observats a la fila i, columna j i

  E_{ij} són els valors esperats sota la hipòtesi nul·la de la fila i, columna j

- La distribució d'aquest estadístic és

$$\hat{X}^2 \sim \chi^2_{(i-1)(j-1)},$$

en aquest cas, el nombre de files i=3, el nombre de columnes j=2.
Per tant, serà una χ^2_2

b) $e_{ij} \geq 5 \ \forall\, i, j$, la qual cosa vol dir que el nombre de valors esperats a cada casella de la taula ha de ser igual o superior a 5.
Mostra aleatòria simple.

c)

Curs	Dones	Homes	Total	Proporc.
2000-2001	78	572	650	1/3
2001-2002	111	539	650	1/3
2002-2003	91	559	650	1/3
Total	280	1670	1950	
Proporc.	0.14	0.86		

Per a cada casella de la taula, calcularem els valors esperats com $n \cdot p_i \cdot p_{.j}$, és a dir, el nombre total d'observacions (n=1950) multiplicat per la proporció de la fila i i per la proporció de la columna j, ja que estem sota la hipòtesi nul·la, i n'obtenim la taula següent de valors esperats:

Curs	Dones	Homes
2000-2001	93.33	556.67
2001-2002	93.33	556.67
2002-2003	93.33	556.67

Finalment, calculem el valor de l'estadístic, $\hat{X}^2 = \sum_{\forall i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, i obtenim:

Curs	Dones	Homes	
2000-2001	2.52	0.42	
2001-2002	3.35	0.56	
2002-2003	0.06	0.01	
			6.92

d) Comparem el valor de l'estadístic amb el que s'obté $\chi^2_{2,0.95}$ =5.991. Com que 6,92> $\chi^2_{2,0.95}$, rebutgem la hipòtesi nul·la.

La resposta al centre és que continuï amb les iniciatives que està realitzant, ja que la proporció del nombre de dones de nou accés va oscil·lant al llarg dels cursos.

Si, a més, es realitzés un test d'hipòtesis entre les proporcions de dones al curs 2000-2001 i 2001-2002, es podria veure que hi ha diferències significatives.