SOCIOESTADÍSTICA II
Grau en Sociologia

Daniel La Parra Casado

Materials de suport a la docència en valencià

133

DEPARTAMENT DE SOCIOLOGIA II
UNIVERSITAT D'ALACANT

Daniel La Parra Casado
A/e: daniel.laparra@ua.es

Aquest material docent ha rebut una beca del Servei de Promoció del Valencià de la Universitat d'Alacant

L'edició d'aquest material s'ha fet dins el marc del conveni per a la promoció de l'ús social del valencià signat per la Universitat d'Alacant amb la Conselleria d'Educació.

ISBN: 978-84-9717-208-0
Dipòsit legal: A 317-2012

Alacant, octubre de 2011 (1a edició)
Edició: Universitat d'Alacant. Servei de Promoció del Valencià
Apartat de Correus 99 - 03080 Alacant
A/e: **s.proval@ua.es** tel. 96 590 34 85

Impressió: Limencop
Universitat d'Alacant.
Edifici de Ciències Socials – Planta baixa
http://www.limencop.com tel. 96 590 34 00 Ext. 2784

Índex

Presentació

Des de l' equip de govern de la Univer sitat d' Alacant valorem la docència en valencià co m un com ponent molt positiu en la form ació universitària dels futurs professionals que han estudiat en aquesta Univ ersitat. És un a obligació de la Univer sitat for mar bons professi onals que en un futur coneguen bé la realitat que els envolta i hi presten am b normalitat els seus serveis. Per això, el domini del valencià propi de la seu a especialitat tècnica o científica és fonamental per a entendre i per a gestionar el procés de desenvolupament de la societat valenciana —i també per a integrar-s'hi amb total normalitat.

Aquest material docent que ara presente m és un resultat mé s d' aquesta filosofia, que im pregna l'actual e quip de go vern, de preparar bons professionals que puguen fer un servei en la societat que ha creat i que manté la Universitat d' Alacant. Per a fer possible que els alum nes actuals i futurs de la Universitat puguen exe rcir co mpetentment la seua professió en valencià, hem d' estimular un proc és previ d' una cert a co mplexitat que, per les seues característiques, ha de ser lent de necessitat: preparar bons professors que puguen im partir la docèn cia en valencià i disposar de m aterials de supor t adequats.

Per a ajudar a aconseguir això, en el s darrers anys hem fet convocatòries d'ajudes per a elaborar materials doce nts en valencià. L'o bjectiu que hi ha darrere d'això és començar a publicar, a poc a poc, els materials que tinguen la qualitat suficient. Aquestes iniciatives de suport a l'ús del valencià co m a llengua de cr eació i de com unicació ci entífica són possibles gràci es a l' ajuda de la Generalitat Valenciana (Conselleria d'Educació), a través del conveni per a la promoció de l'ús del valencià.

Ignasi Jiménez Raneda
Rector

1 Teorema del límit central i llei dels grans nombres

Planifica l'estudi:

- Aquesta unitat es treballa en dues sessions de classe (2 hores de teoria + 2 hores de seminari = 4 hores).
- 8 hores d'estudi fora de l'aula.
- Setmanes 1 i 2.

Seqüència recomanada d'estudi:

Descripció de l'activitat	Temps	Tipus
1) Estudi de les anotacions	2 hores	No presencial
2) Classe teoria i seminari d'exercicis	2 hores	Presencial
3) Fer els exercicis	2 hores	No presencial
4) Classe teoria i seminari d'exercicis	2 hores	Presencial
5) Acabar els exercicis	2 hores	No presencial
6) Estudi i repàs	2 hores	No presencial

Materials per a l'estudi:

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

GARCÍA FERRANDO, Manuel (2008) Socioestadística. Introducción a la estadística en sociología. Madrid: Alianza.

Introducció

Com és possible que amb una mostra de 2.479 persones s'arribe a fer afirmacions sobre la població del conjunt d'Espanya que té més de 46 milions de persones? (com per exemple, en el Baròmetre de juny de 2010 del CIS).

Es diu que les estadístiques sovint s'equivoquen, però quan una persona enquestada està representant més de 17.000 persones de la població, no haurien de fallar més?

Per tant, quina és la probabilitat d'equivocar-nos en les nostres afirmacions basades en enquestes?

En aquest tema aprendràs les raons per les quals es pot fer afirmacions sobre una població a partir de les dades d'una mostra i amb quin nivell de confiança podem fer-les.

Objectiu:

- Conèixer la distribució mostral i la seua utilitat per a fer inferències estadístiques.
- Introduir els diferents tipus de mostreig

Primera explicació

Primer estudiarem el tema d'una forma intuïtiva.

Imaginem una població de 10 persones: els veïns d'un petit poble.

Volem conèixer la mitjana i la desviació típica dels diners de què disposen setmanalment per a gastar en el bar / obres de caritat (ratlleu el que no corresponga).

Atès que és una població reduïda, podem preguntar a les 10 persones per aquesta quantitat.

Imaginem que aquestes són les seues respostes:

Y_i (euros)	F_i
20	1
30	1
40	1
50	1
60	1
70	1
80	1
90	1
100	1
110	1

Amb aquestes dades queda clar que la despesa mitjana de la població és $\mu = 65$ euros, i la desviació típica és $\sigma = 28{,}72$

Aquestes dades serien la mitjana i la desviació típica de la població (perquè s'ha preguntant a tota la gent del poble). Per a fer-ho s'utilitzen les lletres gregues: μ i σ [en la majoria dels manuals s'utilitzen els símbols grecs μ i σ, respectivament, encara que en el manual de Sánchez Carrión s'opta per utilitzar les lletres llatines en majúscula $\bar{Y}$ i S].

La grandària de la població seria igual a 10 ($N = 10$).

Imagina ara que extraiem una mostra (n) a l'atzar de dos individus ($n=2$). Per exemple l'individu amb 100 euros i el que té 20.

Igualment es pot calcular la despesa mitjana ($\bar{y}$) de la mostra amb dos casos.

El resultat seria $\bar{y} = 60$.

Repeteixes de nou el procés (ara obtens una mostra $n=2$ amb el que té 90 euros i el que té 60).

La mitjana seria $\bar{y} = 75$.

Imagina ara que obtens totes les possibles mostres de dos individus de la població.

Una vegada ho hages fet tindràs múltiples mitjanes.

La distribució d'aquestes mitjanes es coneix com a DISTRIBUCIÓ MOSTRAL.

És a dir, **la distribució mostral és el resultat d'extraure múltiples mostres aleatòries simples de la mateixa població**.

Es pot calcular la mitjana de les mitjanes obtingudes en les diferents mostres. És a dir, la mitjana de la distribució mostral $E(\bar{y})$, també coneguda com a *valor esperat de la mitjana*.

Igualment es pot calcular la desviació típica corresponent a la mitjana de la distribució mostral $S_{\bar{y}}$ (és a dir, la desviació típica de totes les mitjanes)

Doncs bé, si fem aquest procés (extraure múltiples mostres) amb una població amb una distribució normal el resultat serà el següent:

- La distribució que s'obté (la distribució mostral) és normal.
- El *valor esperat de la mitjana* de la distribució mostral (és a dir, la mitjana de les mitjanes de les mostres obtingudes) és igual que la mitjana de la població (l'estadístic serà igual al paràmetre). En resum: $E(\bar{y}) = \mu$
- El valor esperat de la desviació típica de la distribució mostral tendirà a ser igual a la desviació típica de la població (σ o S, segons la notació que s'estiga utilitzant) dividida per l'arrel quadrada del nombre de casos (n). En resum: $S_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

Aquestes propietats de la distribució mostral són conegudes amb el nom del Teorema del Límit Central.[1] Se sol resumir com: la distribució mostral de les mitjanes $\bar{Y}$, si es compleixen les condicions, es distribuirà normalment amb una mitjana μ i una variància σ^2/n, és a dir,

$\bar{Y} \sim N(\mu, \sigma^2/n)$.

Si en lloc de partir d'una població amb una distribució normal, comptem amb una mostra gran, les conseqüències serien idèntiques a les que acabem de comentar. És a dir, si s'extrauen múltiples mostres aleatòries simples d'una grandària *n* gran a partir d'una població amb mitjana μ i desviació típica σ, les mitjanes de tals mostres es distribueixen normalment amb mitjana $E(\bar{y}) = \mu$ i desviació típica $S_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

Aquesta pauta que es produeix quan la mostra compta amb un elevat nombre de casos es coneix com a Llei dels Grans Nombres

S'ha de notar que tant quan la població d'origen és normal, com quan es treballa amb un nombre de casos gran, la grandària de la mostra és molt rellevant:

- En el cas que la població siga normal, perquè a mesura que augmenta la grandària mostral (n), disminueix la desviació típica $S_{\bar{y}}$

[1] Aquest teorema en valencià sol rebre el nom de Teorema del Límit Central o Teorema Central del Límit (les dues traduccions són freqüents). Aquesta expressió es deu a Pierre Simon de Laplace (1812), que la va denominar "Théorème central limite", mentre que Jacques Bernouilli (1713) va plantejar "la loi des grands noms".

- En el cas que es treballe amb un nombre gran de casos, perquè a més del que s'ha dit en el punt anterior, es garanteix que la distribució mostral es distribuïsca normalment.

Però quan és una mostra suficientment gran? A partir de quina n s'aplicaria la Llei dels Grans Nombres?

En aquest punt no hi ha acord complet. En alguns casos es parla d'una *n* gran quan *n* és major a 100 casos (n > 100). Per a uns altres, és suficient que *n* siga superior a 30 (n > 30). En altres casos es planteja més de 50 casos (n > 50).

Però per a què serveixen el Teorema del Límit Central (TLC) i la Llei dels Grans Nombres (LGN)?

Conèixer que quan la població és normal o quan la mostra és gran la distribució mostral és normal, amb una mitjana i una desviació típiques conegudes, serà de gran utilitat per al procés d'inferència estadística. El TLC i la LGN permeten seguir el procés deductiu: a partir de la població observem quines característiques tenen les mostres. Això ens ajudarà a recórrer el camí invers (és a dir, el de la mostra a la població, també anomenat procés inductiu i/o procés d'inferència estadística). Ho veurem de seguida en la unitat següent. Les conseqüències del TLC i LGN, més el coneixement de la corba normal i les seues característiques, són les claus que permeten fer la inferència estadística en mitjanes i proporcions.

L'explicació del manual

Ara seguirem l'explicació del manual:

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

1) Obri el llibre de Sánchez Carrión per la pàgina 155 i comença a llegir en l'apartat titulat "3.4.3 De la población a las muestras" fins arribar a "3.4.4 Estadísticos básicos de una distribución muestral".

- Com veus, l'autor es proposa explicar a partir de dos exemples quina serà la forma i els estadístics de la distribució mostral.

Recorda que, encara que normalment s'utilitzen les lletres gregues μ i σ per a denotar respectivament la mitjana i desviació típica de la població, en el manual de Sánchez Carrión s'ha optat per utilitzar les lletres llatines en majúscula $\bar{Y}$ i S.

2) Llig ara l'apartat "3.4.4 Estadísticos básicos de una distribución muestral".

- En aquest apartat es construeix una distribució mostral a partir d'una població de 5 individus (N=5), amb mostres de grandària 2 (n=2).
- A partir d'aquesta distribució mostral es calcula el valor esperat de la mitjana. $E(\bar{y}) = \bar{Y}$ (pàg. 159-160), o $E(\bar{y}) = \mu$.

- Igualment es calcula la desviació típica de les mitjanes.

  $S_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ (pàg. 160-161).[2]

- Com veus l'autor ha mostrat amb un exemple quin és el valor esperat de la mitjana en una distribució mostral i de la desviació típica de les mitjanes (pàg. 158-161).
- Llig ara les explicacions sobre el que ocorre quan la desviació típica de la població és igual a zero (S = 0, o σ = 0) i quan el nombre de casos en la mostra és igual al nombre de casos en la població (n = N) (pàg. 161-162).
  - Quan σ = 0 vol dir que tots els valors de la població són iguals, per tant: Quants casos caldrà seleccionar en una mostra per a conèixer la mitjana?
  - Quan n = N, la dispersió de la distribució mostral serà també igual a 0, per tant quan n = N tindrem certesa que la mostra escollida té una mitjana igual a la mitjana de la població $\bar{y} = \bar{Y}$, o $\bar{y} = \mu$.

– No obstant això, les dues situacions anteriors no són les més comunes. El normal és que la desviació típica de la població siga diferent de 0 (σ ≠ 0) i que hàgem de treballar amb una mostra de grandària inferior al de la població n < N. En aquests casos, si es compleix el TLC o la LGN comptem amb l'avantatge que la distribució mostral té una distribució normal (pàg. 162-163).
  - En una distribució normal sabem que el 95% dels casos es troba entre la mitjana i dues vegades la desviació típica (z = 1,96) (pàg. 163)
  - Per tant, la mitjana ($\bar{y}$) de la mostra té un 95% de probabilitat de trobar-se en l'interval següent $\bar{y} = \mu \pm (1{,}96\ S_{\bar{y}})$ (pàg. 163).
  - Donat que $S_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$, també es pot escriure $\bar{y} = \mu \pm (1{,}96\ \frac{\sigma}{\sqrt{n}})$

---

[2] En el cas que estiguem treballant amb poblacions xicotetes (alerta!: no amb mostres xicotetes, sinó amb poblacions xicotetes) i el mostreig fet siga sense reemplaç (l'habitual en les enquestes sociològiques), per al càlcul de l'error típic caldrà aplicar una correcció tenint en compte la grandària de la població i la grandària mostral.

$S_{\bar{y}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}}$ (pàg. 160-161).

Exemples de situacions en les quals es treballa amb poblacions xicotetes. En investigacions sobre:

- L'opinió dels habitants d'un municipi xicotet.
- Les persones que treballen en una empresa.
- Una enquesta d'avaluació docent (l'alumnat d'una classe).
- Els afiliats a un partit polític.

Exemples de situacions en les quals es treballa amb poblacions àmplies. En investigacions sobre:

- Enquestes nacionals com les de l'INE o el CIS.
- Enquestes autonòmiques o provincials.
- Enquestes en una ciutat.

És freqüent considerar que les poblacions són infinites a partir de 100.000 casos. Tampoc apliquem aquest factor de correcció quan N > 20n, encara que les poblacions siguen finites, ja que el seu valor seria molt pròxim a la unitat.

- Una forma d'interpretar els dos punts anteriors és que en el 95% de les mostres que obtinguem la mitjana poblacional (μ) estarà compresa en l'interval:

$$\bar{y} \pm (1{,}96\ \frac{\sigma}{\sqrt{n}})$$

3) En les pàgines 164 a 174, l'autor explica mitjançant un altre mètode el Teorema del Límit Central. Pots llegir-ho per a obtenir un resum sistemàtic del que hem explicat. Ens referim als apartats "3.4.5 Normalidad de la distribución muestral" i "3.4.6 Factores que influyen en la deducción".

- Para esment a les qüestions següents:

  - En la fórmula per a calcular l'interval de confiança, el valor 1,96 es correspon amb un nivell de confiança del 95%, és a dir[3] $z_{0,95} = 1{,}96$

  - $IC_{95}\ (\bar{y}) = \mu \pm (1{,}96 \frac{\sigma}{\sqrt{n}})$, però si el nivell de confiança que volem és un altre, per exemple, el 99%, s'ha d'utilitzar el valor z corresponent. En aquest cas $z_{0,99} = 2{,}57$ i, per tant, $IC_{99}\ (\bar{y}) = \mu \pm (2{,}57 \frac{\sigma}{\sqrt{n}})$, o de forma general:

  - $IC_{n.c.}\ (\bar{y}) = \mu \pm (\ z_{n.c.}\ \frac{\sigma}{\sqrt{n}})$.[4]

  - Quan s'augmenta el nivell de confiança [$z_{n.c.}$] (per exemple, del 95% passa al 99%) augmenta l'interval de confiança (pàg. 168-9)

  - Quan s'augmenta la grandària de la mostra (n) es redueix l'interval de confiança (pàg. 169-171).

  - Com menor és la dispersió, variabilitat o desviació típica (σ) d'una població menor és l'interval de confiança (pàg. 172-174).

---

[3] Les puntuacions típiques z, que es relacionen amb àrees de la corba normal, han de llegir-se en funció de si s'està utilitzant les dues cues de la corba normal o una única cua. En el cas dels intervals de confiança interessen les dues cues de la corba normal, és a dir, els valors per damunt i per davall de la mitjana. En el nostre exemple diem que $z_{0,95} = 1{,}96$ perquè entre la mitjana de la corba normal típica (és a dir, des de la puntuació típica z = 0) i la puntuació típica z=1,96 hi ha una àrea de 0,4750 (o el que és el mateix: es troba un 47,5% dels casos de la distribució). En considerar igualment la cua esquerra de la distribució normal típica, és a dir, els valors z negatius, sumaria un 95% dels casos de la distribució (0,475*2 = 0,95).
Segons els manuals, la notació $z_{0,95}$ seria equivalent a la notació $z_{\alpha/2}$. El símbol α seria el nivell de significació i és un succés complementari del nivell de confiança, és a dir α = 1- n.c. Per tant, per a n.c. = 0,95 α serà igual a 0,05.

[4] Alerta! Si la mostra és xicoteta, però la distribució és normal, se substitueix $z_{n.c.}$ pel valor *t* de de la distribució de Student amb n-1 graus de llibertat, és a dir, $IC_{n.c.}\ (\bar{y}) = \mu \pm (\ t_{n.c.,n-1}\ \frac{\sigma}{\sqrt{n}})$. La distribució *t* de Student s'introdueix en el tema següent.

Una explicació alternativa

Cadascú tenim diferents maneres d'aprendre. L'explicació que per a una persona és clara i suficient per a una altra persona pot ser inadequada. A hores d'ara ja has pogut confrontar dues explicacions per escrit del Teorema del Límit Central i les seues implicacions (les dues que s'expliquen en aquestes anotacions) i la que s'ha fet en classe (setmanes 1 i 2 del curs). Una explicació alternativa del Teorema del Límit Central la pots trobar en el capítol 5.2.4 "La distribución muestral" (pàg. 164-168) de:

García Ferrando, Manuel (2008) Socioestadística. Introducción a la estadística en sociología. Madrid: Alianza.

Es recomana la seua lectura fins i tot encara que hages entès el tema amb la primera explicació, ja que pot servir de repàs o pot permetre aclarir algun punt.

Activitat per a preparar els temes pròxims

El TLC i la LGN parteixen del supòsit que treballem amb mostres aleatòries simples. Per a conèixer els diferents tipus de mostreig treballarem en "les pràctiques amb ordinador" l'activitat que porta per títol "De la població a la mostra". Per a preparar aquesta activitat se us sol·licita la vostra participació. Formarem en classe tres parelles que s'encarregaran d'explicar en l'aula a) els tipus de mostreig, b) el mostreig polietàpic i c) el tipus de mostreig que es va aplicar en el cas de l'Enquesta Nacional de Salut Sexual, estudi CIS, núm. 2780.

La primera parella utilitzarà com a material de referència:
El capítol 4.3.1 "tipos de muestreo" (pàg. 134-138) i el capítol "otros tipos de muestreo" (pàg. 145-148) de,

- GARCÍA FERRANDO, Manuel (2008) Socioestadística. Introducción a la estadística en sociología. Madrid: Alianza.

Per a l'explicació del mostreig polietàpic s'utilitzarà:
El capítol sobre "muestreo polietápico" (pàg. 33-42) del llibre,

- RODRÍGUEZ OSUNA, Jacinto (1991) *Métodos de muestreo*. Madrid: CIS, Cuadernos Metodológicos.

Finalment, el descrit en els capítols anteriors serà utilitzat per la tercera parella per a explicar el mostreig de l'Enquesta Nacional de Salut Sexual:

- CIS (2009) *Informe metodológico de la Encuesta Nacional de Salud Sexual, estudio núm. 2780*, Madrid: CIS.

Quadern d'exercicis

La millor manera de familiaritzar-se amb el coneixement après és aplicar-lo. Manipulant directament les dades comprendràs millor el Teorema del Límit Central.

Exercicis

1) Exercici 5 i 7 del manual de Sánchez Carrión (pàg. 176 i 177). Resol només la pregunta referida a la mitjana, no al percentatge.
2) Exercici 6 del manual de Sánchez Carrión (pàg. 176). Explica la teua resposta a partir de dos exemples.

Repàs

Al final d'aquesta unitat has de saber:

- Elaborar una distribució mostral.
- Distingir entre mitjana i desviació típica de la població, mostra i distribució mostral.
- Conèixer l'efecte del nivell de confiança en l'interval de confiança. I el de la grandària mostral i el de la dispersió.
- Esmentar les condicions perquè una distribució mostral seguisca una distribució normal i descriure la relació entre la mitjana i la desviació típica de la població i la de la distribució mostral.
- Diferenciar els principals tipus de mostreig

Exercicis de repàs:

Del manual de García de Cortázar *et al*. (1996) *Estadística aplicada a las ciencias sociales*. Madrid: Cuadernos de la UNED, 1a ed., els problemes 5.5. i 5.6.

Del manual de Mullor, Ruben i Fajardo, M. Dolores (2000) *Manual práctico de estadística aplicada a las ciencias sociales*. Barcelona: Ariel, 1a ed., els exercicis proposats, 6.7, 6.8, 6.11, 6.45 i 6.53, l'exercici resolt R.6.1 i els exemples Ex. 6.1., Ex.6.2 i Ex.6.3.

Curiositats

"Maleït polp, em va sepultar en el ridícul. Va destruir l'obra de tota la meua vida". David Spiegelhalter, professor d'Estadística i Probabilitats anglès de la Universitat de Cambridge, en referència al polp *Paul. El País, 12 de juliol de 2010.*

Per a una aproximació més exacta a la visió de David sobre el polp Paul, pots llegir l'entrada en el blog de David Spiegelhalter de 9 juliol de 2010: http://understandinguncertainty.org/node/775

2 Inferència a partir de la mitjana d'una mostra

Planifica l'estudi:

- Aquesta unitat es treballa en una sessió de classe (1 hora de teoria + 1 hora de seminari = 2 hores).
- 4 hores d'estudi fora de l'aula.
- Setmana 3.

Seqüència recomanada d'estudi:

Descripció de l'activitat	Temps	Tipus
1) Estudi de les anotacions	1 hora	No presencial
2) Classe teoria	1 hora	Presencial
3) Seminari (exercicis en l'aula)	1 hora	Presencial
4) Fer els exercicis	1 hora	No Presencial
5) Acabar els exercicis del quadern d'exercicis	1 hora	No presencial
6) Estudi i repàs	1 hora	No presencial

Materials per a l'estudi:

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

GARCÍA FERRANDO, Manuel (2008) Socioestadística. Introducción a la estadística en sociología. Madrid: Alianza.

Introducció

Una vegada has obtingut una mostra de la població és fàcil conèixer la mitjana de la mostra. Ara bé, si el que volem és conèixer quins són els valors més probables de la mitjana de la població, serà necessari estimar un interval per a la mitjana.

Objectius:

- Estimar la mitjana d'una població, coneixent la mitjana i la desviació típica de la mostra.
- Calcular i interpretar l'interval de confiança d'una mostra.

Primera explicació

Gràcies al Teorema del Límit Central (TLC) sabem que les mitjanes mostrals fluctuen al voltant de la mitjana de la població amb una pauta coneguda (la distribució normal), sempre que la *n* de la mostra siga gran o la població normal. Aquesta característica de la distribució mostral permet estimar la probabilitat que la mitjana de la població es trobe en un determinat interval calculat a partir del valor de la mitjana de la mostra.

Per a inferir la mitjana d'una població a partir d'una mostra, se segueixen els passos següents:

1) Hem d'estar segurs que comptem amb una mostra aleatòria simple de la població objecte d'estudi.
2) Una vegada es compta amb una mostra aleatòria es calcula la mitjana de la mostra ($\bar{y}$) i la desviació típica d'aquesta mostra (s).
3) El nombre de casos de la mostra el denominem "n".
4) Amb aquestes dades es pot conèixer l'error típic (*standard error*) (gràcies al TLC i la LGN) que s'explicava en el tema anterior), que serà igual a $S_{\bar{y}} = \frac{s}{\sqrt{n}}$
5) Quan es treballa amb dades d'una mostra s'utilitza la desviació típica no esbiaixada, és a dir, per al seu càlcul es pren com a denominador n-1, en lloc de n.

Això és, si la desviació típica es calcula mitjançant $s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}$ la desviació típica no esbiaixada és $s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{(n-1)}}$

6) A continuació caldrà decidir un nivell de confiança. Habitualment s'utilitza bé un 95%, bé un 99%.

7) Si sabem que la població és normal (cosa que no se sol saber amb certesa) o la *n* és gran, podem aplicar la fórmula general per a estimar la mitjana de la població (μ).

$$IC(\mu) = (\bar{y}) \pm (z_{n.c.} \frac{s}{\sqrt{n}})$$

Conclusió:

Si *n* és gran o la població normal i la mostra és aleatòria simple, es pot obtenir un interval de confiança en el qual és probable que es trobe la mitjana de la població (μ).

Dit d'una altra manera:

La probabilitat que l'interval continga la mitjana de la població serà igual al nivell de confiança escollit (95%, 99%, etc.); no obstant això, no hi haurà certesa absoluta que la mitjana estiga en aquest interval de confiança.

Primer exemple:

S'ha preguntat a deu persones sobre la valoració de l'actuació política de José Luis Rodríguez Zapatero i Mariano Rajoy amb la pregunta següent:

> "Li diré ara els noms d'alguns líders polítics. Li agrairia que m'indicara, pel que fa a cadascun d'ells, si el coneix i quina valoració li mereix la seua actuació política. Puntue'ls de 0 a 10, sabent que el 0 significa que el valora molt malament i el 10 que el valora molt bé."

Les respostes obtingudes pel que fa a José Luis Rodríguez Zapatero han sigut:

3, 7, 9, 0, 1, 5, 6, 8, 10, 3

Calculem mitjana, mediana, rang i desviació típica

Es recomana que ho faces sense mirar els resultats que s'indiquen a continuació:

La mitjana $(\bar{y} = \sum y_i / n = 5{,}2)$

La mediana Me = 5,5

El rang = 10

La desviació típica no esbiaixada $s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} = 3{,}39$

Amb aquestes dades, si assumim normalitat, ja que *n* és xicoteta (n = 10).

Podem calcular l'interval de confiança, amb la fórmula que coneixem.

$$IC\ (\mu) = (\bar{y}) \pm (\ z_{n.c.}\ \frac{s}{\sqrt{n}}\ )$$

Intenta fer-ho per tu mateix, abans de mirar el pas següent:

$$IC(\mu) = 5{,}2 \pm (\ 1{,}96\ \frac{3{,}39}{\sqrt{10}}\ ) = IC_{95}(3{,}09;7{,}31)$$

Ha aprovat José Luis Rodríguez Zapatero als ulls de la població espanyola?

La puntuació mitjana és 5,2, cosa que faria pensar que ha aprovat, però en realitat el més probable és que la població espanyola el qualifica amb una puntuació mitjana que oscil·la entre el 3,09 i 7,31, és a dir, que no podem dir si aprova o no, però estem bastant segurs (al 95%) que la puntuació mitjana que se li atorga es troba entre aquests valors.

Com s'interpreta aquesta informació? (tria la correcta)

a) L'interval conté la mitjana de la població un 95% de les vegades
b) La probabilitat que l'interval continga la mitjana de la població és del 95%.
c) L'interval conté el 95% de la distribució.

Conclusió pràctica:
Quan elabores resultats d'un informe el més adequat és proporcionar els resultats en termes d'interval de confiança.

Tingues en compte que l'interval es pot calcular al 95% (és a dir z=1,96), però també amb altres nivells de confiança. Hauràs d'utilitzar la taula de la corba normal quan varies el nivell de confiança.

Una vegada feta aquesta primera explicació, veurem com es pot calcular l'error típic quan la mostra és xicoteta.

L'estimació de la mitjana quan la mostra és xicoteta

En l'exemple de José Luis Rodríguez Zapatero comptem amb una n = 10, és a dir, una *n* xicoteta. Per tant, no podem aplicar la LGN i tindrem dubtes de si la població segueix una corba normal.

Si la mostra és xicoteta, la distribució mostral (la distribució de les mitjanes d'infinites mostres) no segueix exactament una corba normal, sinó una distribució anomenada de "t de Student".

La distribució amb forma de "t de Student" és similar a la distribució normal. De fet la *t* de Student tendeix a tenir la mateixa forma que la corba normal quan *n* és alta.

En qualsevol manual d'estadística trobaràs una taula de la distribució "t de Student". Pots observar una en l'annex II del manual de Sánchez Carrión. Alternativament les hi ha disponibles en internet, com aquesta:

StatSoft. "Distribution tables". En línia (visitat 6/08/2010):
http://www.statsoft.com/textbook/distribution-tables/#t
(Alerta!, la taula exposada és d'una sola cua).

Una distribució "t de Student", igual que la corba normal, és simètrica.

No obstant això, la seua forma depèn del nombre de graus de llibertat (*d.f. "degrees of freedom"*). Els graus de llibertat són iguals al nombre de casos en la mostra menys 1.

Com més gran és el nombre de graus de llibertat (per tant, com més gran és la mostra), més s'assembla la corba "t de Student" a la corba normal. De fet quan el nombre de casos és gran s'opta per treballar directament amb la corba normal.

Quan el nombre de graus de llibertat és xicotet la corba *t* de Student es converteix en platicúrtica pel que fa a la normal (és a dir, aconsegueix menys altura i els seus pendents són menys pronunciats), la qual cosa significa que hi ha menys casos concentrats al voltant de la mitjana (és a dir, hi ha més dispersió).

Així, per exemple (comprova-ho amb la taula):

- si n=2.500 casos, el valor *t* de Student, amb 2,499 *d.f.* per a un interval al 95% és igual a 1,96 (igual a z_{95})
- si n=30, el valor *t* de Student (29 *d.f.*) per a un interval és 2,045
- si n=10, el valor *t* és igual (9 *d.f.*) a 2,262.
- si n=2, el valor *t* és igual (1 *d.f.*) a 12,706

Tenint en compte que l'interval de confiança per a la mitjana amb la distribució *t* de Student es calcularia de la manera següent:

$$IC(\mu) = \bar{y} \pm \left(t_{n.c.} \frac{s}{\sqrt{n}} \right)$$

Es pot concloure que a menor grandària mostral la tendència és a incrementar progressivament l'amplitud dels intervals de confiança. Això es deu al fet que l'error típic creix perquè el valor de $\frac{s}{\sqrt{n}}$ és cada vegada més gran. D'altra banda, el valor $t_{n.c}$ també creix, i amb ell s'augmenta l'interval de confiança.

Seguint amb l'exemple anterior sobre la valoració de líders polítics, disposàvem de les dades següents:

Les respostes obtingudes eren

3, 7, 9, 0, 1, 5, 6, 8, 10, 3

La mitjana $\left(\bar{y} = \sum y_i / n\right) = 5{,}2$

La desviació típica s = $\sqrt{\frac{\sum (y_i - y)^2}{(n-1)}} = 3{,}39$

I havíem aplicat (per error, ja que no havíem considerat que *n* era xicoteta) la fórmula següent:

$IC(\mu) = \bar{y} \pm \left(z_{n.c.} \frac{s}{\sqrt{n}} \right)$)

$IC(\mu) = 5{,}2 \pm \left(1{,}96 \frac{3{,}39}{\sqrt{10}} \right) = IC_{95}(3{,}09 ; 7{,}31)$

Ara sabem que hauríem d'haver-hi aplicat:

$IC(\mu) = \bar{y} \pm \left(t_{n.c.} \frac{s}{\sqrt{n}} \right)$

On $t_{n.c.} = 2.26216$ (*d.f.* = 9)

Per tant:

$IC(\mu) = 5{,}2 \pm (2.26216 \frac{3{,}39}{\sqrt{10}}) = IC_{95}(2{,}77;7{,}63)$

Podem comparar els dos intervals obtinguts. El que obteníem quan no teníem en compte que la *n* era xicoteta (calculat amb la corba normal), és a dir $IC_{95}(3{,}09;7{,}31)$, i el que he calculat amb la *t* de Student $IC_{95}(2{,}77;7{,}63)$. Es pot observar que l'interval que hem calculat a partir de la corba *t* de Student és més gran, a causa del que ja hem comentat: la corba *t* de Student és platicúrtica.

Prova de decisió estadística per a la mitjana d'una població

Amb el coneixement adquirit en aquesta lliçó ja estàs en disposició de comprendre un primer tipus de prova d'hipòtesi. La hipòtesi a comprovar seria si un determinat valor de la població poguera ser similar o no a la mitjana de la població.

Imaginem que l'ingrés mitjà dels capatassos de la construcció de la província d'Alacant és de 1.696 euros al mes, i l'error típic és igual a 70 euros segons una mostra de 150 capatassos.

Un polític ha assegurat que els capatassos de la construcció cobren 1.800 euros.

Com podríem comprovar que el sou mitjà dels capatassos d'Alacant és igual que el que ha indicat el polític?

La mitjana estimada és 1.696 euros i la mitjana hipotètica és 1.800 euros. Així que en principi no serien iguals.

No obstant això, les dades provenen d'una mostra, de manera que estan subjectes a una certa variabilitat.

Una primera manera d'afrontar aquest fet és calcular l'interval de confiança de la mitjana:

$$IC(\mu) = \bar{y} \pm \left(z_{n.c.} \frac{s}{\sqrt{n}} \right)$$

En aquest cas, seria:

IC(μ) = 1.696 ± (1,96 * 70)

IC95 (1.558,8; 1.833,2)

Com podem veure, la mitjana obtinguda en la mostra és compatible amb l'afirmació feta pel polític.

Però si volem conèixer la probabilitat d'una manera més exacta, podem plantejar una prova d'hipòtesi en els termes següents:

La hipòtesi nul·la (H_0) ens indicaria que la mitjana estimada (l'obtinguda en la mostra, $\bar{y}$) és igual que la mitjana hipotètica que va anunciar el polític (μ).

$H_0 : \mu = \bar{y}$

Mentre que la hipòtesi alternativa, seria que la mitjana és inferior al que declara el polític, és a dir:

$H_1 : \mu > \bar{y}$

La prova estadística per a comprovar la hipòtesi nul·la és:

$z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$ que segueix una distribució ~N(0, 1) si H_0 és certa.

Si calculem el valor z obtenim:

z = (1696 - 1800)/70 = -1,4857

Sabem que, en la corba normal, com més gran és la distància de z = 0 major és la distància de la mitjana. Si la mitjana estimada se situara molt lluny de la mitjana hipotètica caldria pensar que les dues són diferents.

Normalment es pren com a referència una àrea de 0,05, és a dir, un nivell de significació del 5% (α= 0,05). Els valors de z que superen aquesta probabilitat, quan observem una única cua, es produeixen a partir de z = 1,65 o z = -1,65.

Quan superem els valors compresos entre -1,65 i 1,65 estaríem per tant en la regió de rebuig de la hipòtesi nul·la, ja que la probabilitat que la hipòtesi nul·la siga vàlida s'acosta al zero per cent: seria de menys del 5 % (α= 0,05).

En resultar el nostre valor z = -1,4857, que és major que – 1,65, estaríem fora de la regió de rebuig de la hipòtesi nul·la. És a dir:

$-1{,}48 \notin RR(0{,}05) = z < -1{,}65$, de manera que NO podem rebutjar la hipòtesi nul·la. Hem de pensar que la mitjana obtinguda en la mostra és compatible amb la mitjana poblacional declarada pel polític.

És a dir, el nostre valor z s'associa amb una probabilitat major de 0,05. Això significaria que la mitjana estimada no es troba tan lluny de la mitjana hipotètica i, per tant, que és millor pensar que les dues mitjanes són iguals.

Vist d'una altra manera, veiem en la taula de corba normal que:

El valor z = -1,4857, comprèn un 0,4319 + 0,5000 = 0,9319 de l'àrea de la corba (el 93,19%) i deixa a la seua esquerra un 0,0681 de la distribució normal (o un 6,81% de l'àrea de la corba). És a dir, la probabilitat d'obtenir un valor superior a l'obtingut és igual al 6,8%. Com que el límit per a la regió de rebuig l'havíem establit en el 5% i no ho hem aconseguit, caldria pensar que el polític podria tenir raó, a pesar que en la mostra hem obtingut un valor inferior del que assenyalava en el seu discurs.

L'explicació del manual

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

1) Llig l'apartat "4.1.2 Intervalos de Confianza" fins a l'exemple 1 (pàgines 189-191).
   - Explica els avantatges de l'estimació per interval, enfront de la puntual. Arreplega-les a continuació:

Els avantatges de l'estimació per interval són: • • •

   - A continuació s'explica que el TLC i la LGN permeten fer les estimacions per interval.
   - Després explica els diferents components necessaris per a calcular un interval de confiança. Assenyala en la fórmula següent quin seria el paràmetre, l'estadístic, el nivell de confiança i l'error típic de l'estimador:

$$IC(\mu) = \bar{y} \pm \left(z_{n.c.} \frac{s}{\sqrt{n}} \right)$$

- Finalment assenyala com calcular un interval de confiança per a una mitjana, una proporció i un total. De moment (en aquesta unitat), ens interessem per la mitjana.

2) Salta ara fins a la pàgina 194, on s'explica mitjançant l'exemple 2 com fer l'estimació d'una mitjana (pàg. 194-195).

En aquesta pàgina trobaràs un exemple de càlcul de l'interval de confiança a partir d'una mostra relativament gran n = 137 feta amb SPSS. Fes una comprovació: es diu que l'error típic és igual a 0,26 i que la desviació típica és 3,01. És possible? Quina fórmula relaciona ambdós valors?

Què haguera ocorregut si en lloc de 137 casos haguérem comptat amb 10? (ratlla el que no corresponga en cadascuna de les tres afirmacions següents)

- S'estimaria l'interval a partir de la corba normal / *t* de Student
- S'incrementaria / es reduiria l'interval de confiança
- L'error típic seria igual a 0.26 / 1.00

Una explicació alternativa

GARCÍA FERRANDO, Manuel (2008) Socioestadística. Introducción a la estadística en sociología. Madrid: Alianza.

Pots reforçar l'explicació amb la lectura de l'apartat "6.5.2 Estimación de medias" (pàg. 200) del llibre de García Ferrando.

Com calcular l'interval de confiança d'una mitjana amb un programa estadístic?

Per a calcular l'interval del nostre exemple (la valoració mitjana obtinguda per Zapatero) amb el programa estadístic SPSS donarem els passos següents:

1) Una vegada en la vista de dades de SPSS gravarem els valors de la variable, és a dir: 3, 7, 9, 0, 1, 5, 6, 8, 10, 3
2) Després premerem *Analizar / Estadísticos descriptivos / Descriptivos...* i obtindrem

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
VAR00001	10	,00	10,00	5,2000	3,39280
N válido (según lista)	10				

3) Com veus el programa no ha calculat l'interval de confiança, però podem estimar-ho ràpidament aplicant la fórmula

$$IC(\mu) = \bar{y} \pm \left(t_{n.c.} \frac{s}{\sqrt{n}} \right)$$

On $t_{n.c.} = 2.26216$ (*d.f.* = 9)

Per tant:

$$IC(\mu) = 5{,}2 \pm \left(2.26216 \frac{3{,}39}{\sqrt{10}} \right) = IC_{95}(2{,}77;7{,}63)$$

3) Alternativament podem utilitzar l'ordre *Analizar / Comparar medias / Prueba T para una muestra* i obtindrem el resultat:

Prueba para una muestra

	Valor de prueba = 0					
					95% Intervalo de confianza para la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
VAR00001	4,847	9	,001	5,20000	2,7729	7,6271

És a dir, el mateix resultat que havíem obtingut prèviament, però calculat directament pel programa.

En STATA, l'ordre i el resultat serien:

```
mean var1

Mean estimation                     Number of obs    =      10

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        var1 |        5.2   1.072898      2.772935    7.627065
--------------------------------------------------------------
```

Quadern d'exercicis

Exercicis

1) Fes l'exercici 3.b) i 9 del manual de Sánchez Carrión (pàg. 245).
2) Imagina una mostra de 10 persones amb una edat de 25, 35, 40, 45, 50, 55, 65, 75, 85, 95, respectivament. Fes una estimació de la mitjana de la població.
   - Calcula l'interval de confiança per a la mitjana de la mostra al 95% i al 99%. Explica què ocorre quan augmenta el nivell de confiança.
3) Fes l'exercici 10.b) del manual de Sánchez Carrión (pàg. 245).

Repàs

Al final d'aquesta unitat has de saber:

- Calcular l'interval de confiança per a la mitjana d'una mostra si es compleixen les condicions del TLC.
- Ídem en el cas que la mostra siga xicoteta.
- Interpretar el significat de l'interval de confiança.
- Conèixer l'efecte del nivell de confiança en l'interval de confiança. I el de la grandària mostral i el de la dispersió.
- Comparar una corba normal i una *t* de Student

Exercicis de repàs:

Com a exercicis de repàs serviran els de la unitat anterior, encara que en el manual de García Cortazar pots afegir el 5.7 i en el de Mullor i Fajardo el 6.6.

Per a saber-ne més

El contingut que s'inclou a continuació és d'ampliació (no és necessari que l'estudies per a l'examen).

El TLC comporta la necessitat de conèixer si la corba és normal (llevat que *n* siga gran). Aquesta comprovació es pot fer amb ajuda d'un programa estadístic com SPSS o STATA.

Un test de normalitat habitual és l'anomenat test de Kolmogorov-Smirnov, en el qual es prova la hipòtesi nul·la que la distribució és normal (per tant un valor significatiu, $p < 0.05$ són "males notícies" per al TLC).

- En SPSS 15.0 ho tens en *Analizar/Pruebas No Paramétricas/K-S de 1 muestra...*.
  La sintaxi és:

```
NPAR TESTS
  /K-S(NORMAL)= VAR00001
  /MISSING ANALYSIS.
```

- En STATA 11.0 ho trobaràs en Statistics/Nonparametricanalysis/test of hypothesis/One sample Kolmogorov-Smirnov test.

Altres alternatives són el skewness and kurtosis test de normalitat (sktest en STATA) i el Shapiro-Wilk test de normalitat (swilk en STATA). En SPSS es troba el segon test (NPPLOT a *explorar/gráficos con pruebas de normalidad*).

3 Comparació de dues mitjanes

Planifica l'estudi:

- Aquesta unitat es treballa en dues sessions de classe (2 hores de teoria + 2 hores de seminari = 4 hores).
- 8 hores d'estudi fora de l'aula.
- Setmanes 4 i 5

Seqüència recomanada d'estudi:

Descripció de l'activitat	Temps	Tipus
1) Estudi de les anotacions	2 hores	No presencial
2) Classe teoria i seminari d'exercicis	2 hores	Presencial
3) Fer els exercicis	2 hores	No presencial
4) Classe teoria i seminari d'exercicis	2 hores	Presencial
5) Acabar els exercicis	2 hores	No presencial
6) Estudi i repàs	2 hores	No presencial
Repàs unitats B1, B2 i B3 (examen parcial de 16 de març)	2 hores	No presencial

Materials per a l'estudi:

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

Introducció

Comparar mitjanes és un dels instruments que tenim per a entendre millor com som en societat. Qui aconsegueix un millor salari, els treballadors amb estudis universitaris o els treballadors amb estudis secundaris? Els homes o les dones? Quina és l'edat mitjana d'inici en el consum d'alcohol de les persones escolaritzades en centres educatius amb programes de promoció de la salut i d'aquelles que han sigut escolaritzades sense aquest tipus de programes? Difereix la valoració política que els votants del PSOE i el PP fan de Rosa Diez d'*Unión, Progreso y Democracia* (UpyD)?

Objectius:

- Calcular l'interval de confiança per a la diferència de dues mitjanes de mostres independents.
- Comprovar la hipòtesi de la diferència de mitjanes de dues mostres independents.
- Interpretar el valor *p*
- Fer un contrast de mitjanes amb dades d'una enquesta d'opinió

Primera explicació

En aquesta unitat aprendrem a comparar dues mitjanes de mostres independents.

Notació:

Els símbols que utilitzarem per a referir-nos als paràmetres (de les poblacions 1 i 2) i als estadístics (de les mostres 1 i 2) són:

	Població 1	Població 2	Mostra 1	Mostra 2
Mitjana	μ_1 $\overline{Y}_1$	μ_2 $\overline{Y}_2$	$\overline{y}_1$	$\overline{y}_2$
Desviació típica (standard deviation)	σ_1 S_1	σ_2 S_2	s_1	s_2
Nombre de casos	N_1	N_2	n_1	n_2

Notes:

- En el cas de la població, alguns manuals utilitzen els caràcters grecs i uns altres les lletres llatines en majúscules. Per exemple, Sánchez Carrión opta per utilitzar les lletres llatines.
- En alguns manuals es prefereix la lletra *x* i en uns altres la lletra *y* per a referir-se a les variables que s'estan estudiant. Ací s'ha optat per la

lletra *y* en ser la que prefereix Sánchez Carrión en el manual que utilitzem com a referència.

Exemple:

Podem partir d'un exemple per a començar l'explicació. En una mostra d'una empresa multinacional hem obtingut que les dones cobren 710 euros, mentre que els homes 1.040 euros. El nombre d'empleats en la mostra és de 53 dones i 47 de homes. Suposem que la desviació típica dels salaris en les dues poblacions és igual a 110 i que les dues tenen una distribució normal. Cobren el mateix salari els homes i les dones d'aquesta empresa?

Per a saber-ho, hem de començar pensant quina serà la distribució mostral de la diferència de mitjanes. La distribució mostral de la diferència de mitjanes es pot obtenir amb un procediment similar al tractat en la unitat 1 per a una sola població. És a dir, es poden extraure infinites mostres de grandària n_1 , n_2. Cada vegada es calcularia la mitjana ($\bar{y}_1$, $\bar{y}_2$) i la desviació típica (s_1 , s_2). Per a cada parell de mostres es podria calcular la diferència de mitjanes ($\bar{y}_1 - \bar{y}_2 = \bar{d}$).

Igual que ocorria quan fèiem això amb una única població, s'observarà que:

- La diferència entre les mitjanes de cada parell de mostres tendeix a ser igual a la diferència entre les mitjanes de les dues poblacions.
- La desviació típica de la mitjana de les infinites mostres depèn de la grandària de les mostres (n_1 , n_2)
- La forma de la distribució mostral tendeix a ser igual a la distribució de la corba normal, a mesura que augmenta la grandària de les mostres (n_1 , n_2).

En definitiva, la gran notícia és que el TLC i la LGN s'aplica també a la distribució mostral de la diferència de mitjanes de dues poblacions independents.

Com recordaràs l'error típic amb una mostra era igual a $S_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

D'igual manera, l'error típic elevat al quadrat, seria igual a

$$S^2_{\bar{y}} = \frac{\sigma^2}{n}$$

En el cas de la distribució mostral de la diferència de mitjanes de mostres independents l'error típic serà igual a la suma dels errors típics.

De manera que:

$$(S_{\bar{y}_1 - \bar{y}_2})^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

L'error típic de la distribució mostral seria per tant:

$$S_{\bar{y}_1-\bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Seguint amb el nostre exemple sobre el salari d'homes i dones teníem les dades següents:

Dèiem que "les dones d'una empresa cobren 710 euros, mentre que els homes 1.040. El nombre d'empleats és 53 dones i 47 homes. La desviació típica és igual 110 en les dues poblacions, és a dir:

	Mostra 1 Dones	Mostra 2 Homes
Mitjana	$\bar{y}_1$ 710	$\bar{y}_2$ 1040
Desviació típica	s_1 110	s_2 110
Nombre de casos	n_1 53	n_2 47

De manera que:

$\bar{y}_1 - \bar{y}_2 = 710 - 1040 = -330$

En conèixer que la desviació típica en les dues poblacions és igual a 110, podem aplicar la fórmula següent per al càlcul de l'error típic de la distribució mostral.

$$S_{\bar{y}1-\bar{y}2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{110^2}{53} + \frac{110^2}{47}} = 22{,}04$$

A partir de l'error típic podrem calcular l'interval de confiança per a la diferència de les mitjanes.

Recordem abans com calculàvem l'interval de confiança amb una mostra.

Quan treballàvem amb una única mitjana l'interval de confiança era el resultat de

$$\text{IC}(\mu) = \bar{y} \pm (z_{n.c.} \ \frac{s}{\sqrt{n}})$$

És a dir, la mitjana de la població (paràmetre) es trobava en un interval, amb una probabilitat (nivell de confiança) determinada, el límit inferior del qual es definia per restar-li a la mitjana de la mostra (estadístic) el producte de la puntuació tipificada associada a un determinat nivell de confiança per l'error típic i el límit superior en sumar aquest producte a la mitjana de la mostra.

Alerta!, quan la població era normal o la *n* gran.

Per tant, suposant normalitat, podem substituir tots els components de la fórmula amb el que coneixem de la distribució mostral de la diferència de mitjanes, és a dir:

$$y_1 - y_2 \pm (z_{n.c.} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}))$$

Que, en el nostre exemple, i amb un nivell de confiança del 95%, seria:

$IC_{95} = -330 \pm (1{,}96 * 22{,}04)$

$IC_{95} = (-373{,}2; -286{,}8)$

Com s'interpretaria aquest resultat?

La diferència mitjana de salaris entre homes i dones de l'empresa és de 330 euros favorable als homes. Les dones cobren amb un 95% de confiança entre 373 i 286 euros menys que els homes.

Desviacions típiques desconegudes

Hem assenyalat que la fórmula per a calcular l'error típic de la distribució mostral quan es coneix les desviacions típiques de les poblacions 1 i 2 (σ_1 , σ_2) és:

$$S_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

No obstant això, aquesta situació no és habitual. El normal és que hàgem obtingut dues mostres i a partir d'aquestes mostres hàgem estimat les desviacions típiques de les mostres (s_1 , s_2).

En aquest cas, s'aplica la fórmula següent per al càlcul de l'error típic:

$$S_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} (\frac{1}{n_1} + \frac{1}{n_2})}$$

De l'interval de confiança per a la diferència de mitjanes a les proves de decisió (contrast d'hipòtesi) de la diferència de mitjanes

Hem vist que coneixent l'error típic i la mitjana de les diferències de les dues mostres és possible calcular un interval de confiança.

Si el que volem és calcular la probabilitat que les dues mitjanes de la població siguen iguals (el valor p o coeficient de significació) llavors utilitzarem un contrast d'hipòtesi.

Per a això s'ha de començar definint **les hipòtesis** que anem a contrastar.

La hipòtesi **nul·la (H_0)**, aquella sobre la qual es calcula la probabilitat (el valor p o coeficient de significació) que siga certa, serà la hipòtesi d'igualtat de mitjanes:

$H_0 : \mu_1 - \mu_2 = 0$

Com a **hipòtesi alternativa (H_1)** platejarem que les mitjanes són diferents:

$H_1 : \mu_1 - \mu_2 \neq 0$

Encara que, en alguns casos podem plantejar com a hipòtesi alternativa que una mitjana és major que una altra. Per exemple, en el nostre exemple, sabem que en les nostres societats patriarcals el salari dels homes sol ser major que el de les dones de manera que la diferència és desfavorable a les dones i per tant podríem plantejar la hipòtesi alternativa:

$H_1 : \mu_1 - \mu_2 < 0$

Alerta! Quan la hipòtesi alternativa siga del tipus $H_1 : \mu_1 - \mu_2 \neq 0$ utilitzarem la informació de la corba normal proporcionada per les seues dues cues o costats, mentre que que quan siga del tipus $H_1 : \mu_1 - \mu_2 < 0$ utilitzarem un únic costat de la corba normal ("una cua"). Això és rellevant en escollir els valors que reflecteixen el nivell de confiança ($z_{n.c.}$). De fet quan vam aprendre a utilitzar i interpretar la corba normal i les seues taules, vèiem que aquestes podien estar construïdes per a una o dues cues. Per aquest motiu es parla de prova de "dues cues" o "d'una cua".

Una vegada plantejada la nostra hipòtesi, hem de triar el test o prova que ens servirà per a conèixer la probabilitat que es complisca la hipòtesi nul·la (el valor p o coeficient de significació). Com treballem sota la hipòtesi que es tracta una corba normal, el test a triar és el test z.

Si recordes de socioestadística I, calculàvem el valor tipificat z de la corba normal de la manera següent:

$$z = \frac{(\bar{y} - \mu)}{\sigma}$$

En el nostre cas estem parlant de diferències de mitjanes de dues poblacions, és a dir:

$z = \dfrac{(\bar{y}_1 - \bar{y}_2) - 0}{S_{\bar{y}_1 - \bar{y}_2}}$ que segueix una distribució ~ N(0, 1) si H_0 és certa.

Per tant, ara ho podrem calcular:

Amb desviacions típiques de la població conegudes:

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ que segueix una distribució ~ N(0, 1) si H_0 és certa.

Amb desviacions típiques de la població desconegudes:

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}(\frac{1}{n_1} + \frac{1}{n_2})}}$$ que segueix una distribució~N(0, 1) si H_0 és certa.

Alerta! En el cas que les mostres siguen xicotetes però provinents de poblacions amb distribució normal utilitzarem la *t* de Student en lloc de la normal, és a dir *t* en lloc de z, amb $n_1 + n_2 - 2$ graus de llibertat.

Amb les dades del nostre exemple, el valor z seria

$$z = \frac{-330}{22{,}04} = -14{,}97$$

Com sabem en la corba normal a mesura que el valor z creix o decreix, és a dir, quan és major o menor que 0, disminueix l'àrea dels extrems de la corba. Dit d'una altra manera, els casos en els extrems de la corba tenen cada vegada una menor probabilitat.

Si mirem la taula de la distribució normal (per exemple la taula 1 de l'annex II, del manual de Sánchez Carrión), observem que:

En el centre de la distribució z és igual a 0 i α és igual al 50 per cent (en el cas d'una cua) o a 100 per cent en el cas de dues cues.

Si les dues mitjanes foren exactament iguals el que s'espera és que el valor z del nostre estadístic de contrast fóra igual a 0. No obstant això, si les dues mitjanes no són iguals el valor z s'allunyarà del valor 0. És a dir, valors de z allunyats de 0 informen que és menys probable que les dues mitjanes siguen iguals. Quan la probabilitat que les dues mitjanes siguen iguals *siga molt baixa*, preferim optar per la hipòtesi alternativa que afirma que no són iguals.

En la frase anterior dèiem que "la probabilitat que les dues mitjanes siguen iguals siga molt baixa", però quan seria molt baixa? Normalment el límit s'estableix en un 5% (encara que això és una decisió arbitrària i caldria variar-la en funció de la precisió que necessite l'estudi). Aquest límit es coneix com a *nivell de significació* i es denota amb la lletra grega alfa ($\alpha = 0{,}05$).

Com sabem, quan z és igual a 1,96 la probabilitat associada en la corba normal és igual al 5 per cent (del 2,5% quan es tracta d'una cua).

Per tant, si el valor de l'estadístic de contrast z és major que 1,96 o menor que -1,96 estarem en l'anomenada "regió de rebuig" de la hipòtesi nul·la, per representar una probabilitat menor del 5% (0,05) que les dues mitjanes siguen iguals.

$$RR\left(0{,}05\right) = \left|Z\right| > 1{,}96$$

En el nostre exemple el valor z és igual a -14,97, amb la qual cosa la probabilitat s'acosta al 0 per cent. De fet a partir de valors de z superiors a 4 són probabilitats molt baixes, menors al 3 per 100.000 (p=0,0000317 amb una cua i 0,0000634 amb dues cues).

Per tant, per a un valor z = -14,97 seria un valor molt allunyat del centre de la distribució (z = 0) i amb un valor p molt xicotet (simplificant p < 0,001, menor a l'1 per 1000).

Amb aquestes dades, la nostra conclusió és que el nivell de significació és molt baix o, el que és el mateix, la probabilitat que la hipòtesi nul·la H_0 siga vàlida s'acosta al zero per cent. En aquestes condicions és millor optar per la hipòtesi alternativa.

$-14{,}97 \in RR\left(0{,}05\right) = \left|Z\right| > 1{,}96$, de manera que rebutgem la hipòtesi nul·la

Aquesta expressió pot llegir-se com: "el valor de la prova z (-14,97) per a la diferència de mitjanes se situa en la regió de rebuig de la hipòtesi nul·la, amb un nivell de significació α = 0,05, motiu pel qual s'opta per rebutjar la hipòtesi nul·la, que assumeix la igualtat de mitjanes".

Variàncies iguals o diferents

Per a calcular la prova estadística que estem utilitzant per a determinar si les mitjanes són iguals, es requereix que les variàncies de la població siguen iguals.

És a dir, per a poder calcular l'estadístic de contrast que s'inclou a continuació s'ha de conèixer prèviament si les variàncies són iguals:

$$z = \frac{\left(\bar{y}_1 - \bar{y}_2\right) - 0}{\sqrt{\frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{\left(n_1 - 1\right) + \left(n_2 - 1\right)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ que segueix una distribució~N(0, 1)

si H_0 és certa.

Una manera de comprovar que les dues variàncies són iguals és fer la prova F de Snedecor.

La hipòtesi a posar a prova són:

H_0: les variàncies són iguals
H_1: les variàncies són diferents

La prova a utilitzar és la prova F. Els graus de llibertat són iguals al nombre de casos de la primera variable menys un i el nombre de casos de la segona variable menys 1.

La prova F consisteix en la divisió d'una variància per una altra. Per a això es divideix la variància de la variable amb el valor més alt per la variància de la variable amb el valor més baix.

El contrast s'ha d'aplicar en el cas que es complisca el supòsit de normalitat o que la *n* siga suficientment gran (en les dues mostres).

Si el valor F obtingut és superior al valor F trobat en la taula per als graus de llibertat i el nivell de significació escollit, ens trobaríem en la regió de rebuig de la hipòtesi nul·la. Hauríem de pensar que les variàncies són diferents i no podríem aplicar el contrast de mitjanes.

Si el valor F obtingut és inferior al valor oposat en la taula per als graus de llibertat i el nivell de significació escollit, caldria pensar que les dues variàncies són iguals i podríem aplicar la prova per a la diferència de mitjanes.

Vegem dos exemples:

Si comptem amb els valors següents d'una mostra extreta de la població 1
9, 7, 8, 10, 5, 10, 8, 5, 9, 6, 8

I d'una mostra de la població 2
5, 4, 3, 3, 2, 1, 1, 6, 4, 2

Suposem que en les dues poblacions son variables amb una distribució normal.

En la població 1, la mitjana és 7,73 i la desviació típica no esbiaixada 1,79, per tant la seua variància 3,22

En la població 2, la mitjana és 3,1 i la desviació típica no esbiaixada 1,66, per tant la seua variància és 2,77

Com en la primera mostra hi ha 11 casos i en la segona 10, els graus de llibertat serien 10 i 9. Observem la taula de la corba F amb 10 i 9 graus de llibertat que és igual a f = 3,14 per a un nivell de significació de 0,05. És a dir, la regió de rebuig de la hipòtesi nul·la (les dues variàncies són iguals) se situa en valors superiors a 3,14.

$$RR\left(0{,}05\right)=\left\{F>3{,}14\right\}$$

Calculem el valor F dividint una variància per una altra:

F = 3,22 / 2,77 = 1,16

Com el nostre valor és menor que 3,14, no estaria en la regió de rebuig de la hipòtesi nul·la i per això assumirem que les dues variàncies són iguals.

$1{,}16\notin RR\left(0{,}05\right)=\left\{F>3{,}14\right\}$ de manera que NO es rebutja la hipòtesi nul·la, és a dir, assumim que les variàncies són iguals.

A continuació podríem aplicar la prova de la diferència de mitjanes:

$$t = \frac{(7{,}73 - 3{,}1) - 0}{\sqrt{\frac{(10)3{,}22 + (9)2{,}77}{(10)+(9)}\left(\frac{1}{11} + \frac{1}{10}\right)}}$$

t = 6,11

La regió de rebuig per a un nivell de significació 0,05 i 19 graus de llibertat seria a partir de t = 2,093 o menor de t = -2,093; com el nostre valor t=6,11 és major que 2,093 estaria en la regió de rebuig i caldria concloure que les dues mitjanes no són iguals.

$6{,}11 \in RR\,(0{,}05) = \{|T| > 2{,}093\}$ de manera que es rebutja la hipòtesi nul·la, i concloem que les dues mitjanes no són iguals.

Què són dues mostres independents?

Parlem de mitjanes provinents de mostres independents quan les observacions d'una mostra són **independents** de les observacions de l'altra mostra. Es parla d'independència estadística quan la probabilitat del que ocorre en una mostra no està influïda pel que ocórrega en l'altra mostra.

Això què vol dir?

Per exemple, si comparem la mitjana d'edat en les dones de la classe i la mitjana d'edat dels homes de la classe, totes dues seran independents, perquè l'edat mitjana de les dones no depèn de l'edat mitjana dels homes, ni el contrari.

Quan serien dependents?

El cas més comú de dependència es produeix quan s'observa en un mateix individu una mateixa dada al llarg del temps. Per exemple, quins són els diners de butxaca disponibles en un grup d'amics abans i després d'haver eixit de marxa? El normal és que els diners que tinguen després d'haver eixit de marxa estiguen relacionats amb els que tenien abans d'haver eixit de marxa, és a dir, s'espera que tots tinguen menys diners del que tenien al principi de la nit. Llevat que hagen anat a un casino i hagen tingut molta sort.

Ho tenim clar?

Imaginem que volem mesurar l'eficàcia d'una nova metodologia docent.

En el disseny 1) s'opta per mesurar els resultats de la classe, abans i després d'haver-hi aplicat aquesta metodologia. Es tracta, per tant, de mostres independents / dependents (ratlla el que no corresponga).

En el disseny 2) s'opta per mesurar els resultats de dues classes. Una que ha rebut la nova metodologia docent i una altra que ha seguit amb la metodologia habitual. Es tracta, per tant, de mostres independents / dependents (ratlla el que no corresponga).

Com fer un contrast d'hipòtesi per a la diferència de mitjanes amb una programa estadístic?

Per a calcular l'interval del nostre últim exemple amb el programa estadístic SPSS donarem els passos següents:

1) Una vegada en la vista de dades de SPSS, gravarem els valors de les dues variables, és a dir:

   Població 1: 9, 7, 8, 10, 5, 10, 8, 5, 9, 6, 8
   Població 2: 5, 4, 3, 3, 2, 1, 1, 6, 4, 2

Es crearà una variable per als valors i una altra per a indicar la població de pertinença

Valors	Població
9	1
7	1
8	1
10	1
5	1
10	1
8	1
5	1
9	1
6	1
8	1
5	2
4	2
3	2
3	2
2	2
1	2
1	2
6	2
4	2
2	2

2) Després premerem *Analizar / comparar medias / Prueba t para muestras independientes...*

3) S'obri un quadre de diàleg en el qual indiquem que la variable a contrastar és "valors", que la variable d'agrupació és "població" amb els grups 1 i 2 i acceptem. Obtenint el resultat següent:

Estadísticos de grupo

	población	N	Media	Desviación típ.	Error típ. de la media
valores	1	11	7,73	1,794	,541
	2	10	3,10	1,663	,526

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
		Inferior	Superior	Inferior	Superior	Inferior	Superior	Inferior	Superior	Inferior
valores	Se han asumido varianzas iguales	,082	,778	6,110	19	,000	4,627	,757	3,042	6,212
	No se han asumido varianzas iguales			6,133	18,988	,000	4,627	,754	3,048	6,206

En l'exemple, el programa ha calculat:

1) La mitjana i la desviació típica en les dues mostres.
2) Ha aplicat la prova per a comprovar si les variàncies són iguals. Com que el nivell de significació és 0,788, és a dir, major a 0,05, no es pot rebutjar la hipòtesi nul·la i s'assumeix que les variàncies són iguals.
3) El contrast de mitjanes. La prova *t* produeix un valor 6,11 amb un nivell de significació 0,000, és a dir $p < 0,001$, de manera que s'ha de rebutjar la hipòtesi nul·la que assenyala que les mitjanes són iguals. Caldria assumir que les dues mitjanes difereixen.
4) La diferència entre les mitjanes seria igual 4,627 amb un interval de confiança al 95 per cent que se situa entre 3,042 i 6,212.

Amb STATA les ordres i el resultat serien:

Primer, per a provar si les variàncies són iguals

```
. sdtest var1, by(var2)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      11    7.727273      .54089    1.793929    6.522095    8.932451
       2 |      10         3.1    .5259911     1.66333    1.910125    4.289875
---------+--------------------------------------------------------------------
combined |      21     5.52381    .6347817    2.908935    4.199678    6.847941
------------------------------------------------------------------------------
    ratio = sd(1) / sd(2)                                         f =   1.1632
Ho: ratio = 1                                 degrees of freedom =    10, 9

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.5850         2*Pr(F > f) = 0.8300           Pr(F > f) = 0.4150
```

De manera que s'ha d'assumir que les variàncies són iguals.

Després per a provar si les mitjanes són iguals

```
. ttest var1, by(var2)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      11    7.727273      .54089    1.793929    6.522095    8.932451
       2 |      10         3.1    .5259911     1.66333    1.910125    4.289875
---------+--------------------------------------------------------------------
combined |      21     5.52381    .6347817    2.908935    4.199678    6.847941
---------+--------------------------------------------------------------------
    diff |            4.627273    .7573304                3.042162    6.212384
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =   6.1100
Ho: diff = 0                                  degrees of freedom =       19

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

De manera que s'ha d'assumir que les mitjanes no són iguals.

L'explicació del manual

Ara seguirem l'explicació del manual:

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza

1. Treballarem el capítol 10 "Pruebas de decisión para el caso de dos muestras" pàg. 287-292.

- L'autor comença assenyalant l'aplicació de la comparació entre dues mitjanes en les investigacions sociològiques.
- Continua explicant que el teorema del límit central és aplicable a la diferència de mitjanes de dues poblacions.
- Assenyala com es calcularia l'error típic en el cas de la diferència de mitjanes. Observa que la fórmula 10.2 és igual que la que s'ha explicat en aquest tema, encara que es presenta de manera diferent de com s'ha fet en aquestes anotacions.

- Il·lustra l'explicació amb un exemple.
- Recorda els requisits de la prova de decisió estadística per a contrastar la diferència entre mitjanes (ratlla el que no corresponga):
    - El nivell de mesurament serà ordinal / interval
    - Les observacions seran dependents / independents
    - La població de referència serà normal / amb qualsevol distribució
    - Les variàncies mostrals són homogènies / heterogènies

Una explicació alternativa

També pots consultar l'explicació de:

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

Quadern d'exercicis

Exercicis

1) Fes l'exercici 1 del tema 10 del manual de García Ferrando (pàg. 307).
2) Fes l'exercici 2 del tema 10 del manual de García Ferrando (pàg. 308).
3) Del manual de Sánchez Carrión l'exercici 1 i 2 del capítol 7.

Repàs

Al final d'aquesta unitat has de saber:

- Calcular l'interval de confiança per a la diferència de dues mitjanes de mostres independents.
- Conèixer les condicions per a poder aplicar la prova de decisió estadística per a la comparació de dues mitjanes.
- Establir la regió de rebuig de la hipòtesi nul·la en un contrast.

Exercicis de repàs:

Del manual de GARCÍA DE CORTÁZAR *et al*. (1996) *Estadística aplicada a las ciencias sociales*. Madrid: Cuadernos de la UNED, 1a ed., els problemes 6.7, 6,9 i 6.10.

Del manual de MULLOR, RUBEN I FAJARDO, M. Dolores (2000) *Manual práctico de estadística aplicada a las ciencias sociales*. Barcelona: Ariel, 1a ed., els exercicis resolts R.7.1., R.7.2, R.7.6, R.7.8, R.7.10 i 7.11.

Per a saber-ne més

Per a comprovar la hipòtesi de la diferència de mitjanes en mostres dependents pots seguir les explicacions dels manuals de Sánchez Carrión (pàg. 422-429). Com pots veure els principis són molt similars als que apliquem en el tema anterior amb una sola mostra.

4 Inferència a partir d'una proporció d'una mostra

"De cada deu persones que veuen televisió, cinc... són la meitat." Les Luthiers

"Nou de cada deu dentistes recomanen un xiclet sense sucre, aquesta nit ací en primícia per a tots vosaltres el fill de puta que recomana un xiclet amb sucre!" Faemino i Cansado

Planifica l'estudi:

- Aquesta unitat es treballa en una sessió de classe (1 hora de teoria + 1 hora de seminari = 2 hores).
- 4 hores d'estudi fora de l'aula.
- Setmana 6

Seqüència recomanada d'estudi:

Descripció de l'activitat	Temps	Tipus
1) Estudi de les anotacions	1 hora	No presencial
2) Classe teoria	1 hora	Presencial
3) Seminari (exercicis en l'aula)	1 hora	Presencial
4) Fer els exercicis	2 hores	No presencial
5) Repàs	1 hora	No presencial

Materials per a l'estudi:

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza

Introducció

Necessitem conèixer el percentatge de persones aturades sobre el total d'actives (la taxa d'atur), el percentatge que votarà un determinat partit en les pròximes eleccions, la proporció de població amb un poder adquisitiu suficient per a adquirir un determinat producte, el percentatge de persones que és favorable a un determinat canvi legislatiu, l'increment en la proporció de persones satisfetes amb la política municipal en aquest determinat barri, la proporció de persones sota el límit de pobresa, etcètera.

Què seria de les ciències socials sense percentatges i proporcions? La majoria de les variables que s'utilitzen en ciències socials són de tipus nominal o ordinal, per tant rares vegades es pot calcular la mitjana, però això sí, es pot calcular el percentatge o qualsevol altre tipus de proporció. Amb freqüència, igual que ocorria amb les mitjanes, se'ns facilita l'estimació puntual, però és en l'estimació d'interval quan realment obtenim una informació de qualitat (en especificar-se el grau de precisió de l'estimació feta).

Objectius:

- Calcular i interpretar l'interval de confiança per a una proporció.

Primera explicació

La major part de les variables en ciències socials són de tipus categòric. Per exemple:

Sexe: home, dona.
Ocupació: policia, docent, dependent, guia, etc.
Situació laboral: persones ocupades, aturades, jubilades, etc.
Vot: PP, PSOE, EU, CiU, etc.
Opinió: favorable, desfavorable
Etcètera.

Altres variables també es poden convertir en categòriques:

- En la variable ingrés es pot valorar que a partir d'un determinat nivell (denominat per exemple, *límit de pobresa*) es distingeix a les persones que estan en "situació de pobresa" i les persones que no ho estan.
- L'edat és un variable contínua, però es pot convertir en grups d'edat (0-5, 6-10, etc..). Per exemple, Amando de Miguel distingia entre "*maduros*" (30-44 anys), "*talludos*" (45-64 anys), etc.

Les variables categòriques amb dues categories es coneixen amb el nom de variables binàries o dicotòmiques.

Notació

Els símbols que utilitzarem per a referir-nos als paràmetres (de la població) i als estadístics (de la mostra) són:

	Població 1	Mostra 1
Mitjana	π P	p
Desviació típica (standard deviation)	$\sqrt{\frac{\pi(1-\pi)}{N}}$	$\sqrt{\frac{p(1-p)}{n}}$
Nombre de casos	N	n

Notes:

- En el cas de la població, en alguns manuals utilitzen els caràcters grecs i en uns altres les lletres llatines en majúscules. Per exemple, Sánchez Carrión opta per usar les lletres llatines.

La distribució mostral pròpia de les variables categòriques és la distribució binomial.

Quan *n* tendeix a infinit, la distribució binomial tendeix a aproximar-se a la distribució normal. Per aquest motiu la distribució mostral en la qual basarem les inferències per a una proporció (també pel TLC i la LGN) s'assembla a la corba normal quan la mostra és àmplia.

D'acord amb això, si la grandària mostral és àmplia podem calcular l'interval de confiança per a una proporció amb la fórmula:

$$IC(\pi) = p \pm (z_{n.c.} \sqrt{\frac{p(1-p)}{n}})$$

Alerta! Si estiguérem treballant amb percentatges, en lloc de proporcions caldria canviar l'1 per 100, és a dir:

$$IC(\pi) = p \pm (z_{n.c.} \sqrt{\frac{p(100-p)}{n}})$$

Exemple:

En una mostra de 2000 persones, 1300 assenyalen que segueixen una estricta dieta mediterrània.

Quina és la proporció dels que segueixen la dieta?

Primer, podem calcular una estimació puntual:

p = 1300/2000 = 0,65 és a dir, el 65%

Però serà millor que calculem l'interval de confiança

$$IC(\pi) = 65 \pm (1{,}96 \cdot \sqrt{\frac{65\,(100-65)}{2000}})$$

$IC_{95}\,(62{,}9;67{,}1)$

En conclusió: podem estar segurs (si no menteixen) al 95% que la proporció de persones que segueixen la dieta mediterrània en la població se situa entre el 62,9% i el 67,1%.

Ara podríem fer un test per a comprovar si la població té una determinada proporció. Per exemple, que un 50% de la població segueix una dieta mediterrània.

En concret hem assenyalat que la mostra té un 65% de persones que segueixen la dieta mediterrània i que en la població un 50% seguiria aquesta dieta.

La hipòtesi quedaria plantejada com:

$H_0 : \pi = 0{,}50$

Mentre que la hipòtesi nul·la plantejaria que són diferents:

$H_1 : \pi \neq 0{,}50$

El nivell de significació s'estableix en $\alpha = 0{,}05$

La prova estadística seria:

$$z = \frac{(p-\pi)}{\sqrt{\frac{p\,(1-p)}{n}}}$$ que segueix una distribució $\sim N\,(0,1)$ si H_0 és certa

La regió de rebuig, amb un nivell de significació de 0,05 se situaria fora de l'interval de valors z (-1,96;1,96)

$$RR\,(0{,}05) = |Z| > 1{,}96$$

Aplicant les dades del nostre exemple obtindríem:

$$z = \frac{(0{,}65-0{,}5)}{\sqrt{\frac{(0{,}65*0{,}35\,)}{2000}}} = 14{,}1$$

En ser z igual a 14,1 i, per tant, major que z = 1,96 ens trobem en la regió de rebuig, i per tant es descarta la hipòtesi nul·la que un 50% de la població segueix dieta mediterrània.

$14{,}1 \in RR\,(0{,}05) = |Z| > 1{,}96$ Per tant, rebutgem la hipòtesi nul·la.

S'ha de considerar que per al cas del càlcul d'intervals per a proporcions *n* ha de ser prou gran.

Una norma a aplicar per a saber si *n* és suficientment gran és fer la comprovació següent:

1. $p\,n \geq 5$
2. $(1-p)\,n \geq 5$

Per exemple, si p = 0,65 i n = 2000

1. 0,65 * 2000 = 1300 (compleix la condició)
2. (1-0,35) * 2000 = 700 (compleix la condició)

Alerta!! Si no es compleixen aquestes dues condicions caldria aplicar altres mètodes (que no seran explicats en el marc d'aquesta assignatura).

Alerta!! En el cas de poblacions finites s'aplicarà la següent correcció de finitud en el càlcul de la desviació típica (vegeu també nota a peu de pàgina 2 en pàgina 5):

$$\sqrt{\frac{p(1-p)}{n}\frac{(N-n)}{(N-1)}}$$

L'explicació del manual

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

1) Llig l'apartat "4.1.2 Intervalos de Confianza" fins a l'exemple 1 (pàgines 189-191).

   - Explica els avantatges de l'estimació per interval, enfront de la puntual. Arreplega-la a continuació:

   - A continuació s'explica que el TLC i la LGN permeten fer les estimacions per interval.
   - Després explica els diferents components necessaris per a calcular un interval de confiança. Assenyala en la fórmula següent què seria el paràmetre, l'estadístic, el nivell de confiança i l'error típic de l'estimador:

$$IC(\pi) = p \pm (n_{n.c.} \sqrt{\frac{p(1-p)}{n}})$$

- Finalment assenyala com calcular un interval de confiança per a una mitjana, una proporció i un total. En aquesta unitat, ens interessem per la proporció.

2) Segueix llegint l'exemple 1 on s'explica com fer l'estimació d'una proporció (pàg. 191-194).

En les primeres línies argumenta l'interès de fer una estimació d'interval en lloc d'una estimació puntual.

Explica dues formes de resoldre el problema (mitjançant deducció de l'estadístic, és a dir, partint d'un percentatge previ, en aquest cas el 10%) o mitjançant la inferència del paràmetre (que és el mètode que hem explicat prèviament).

En l'exemple utilitzat en el manual es parteix d'una mostra relativament gran: n = 1500 persones.

Què haguera ocorregut si en lloc de 1500 casos haguérem comptat amb 10? (ratlla el que no corresponga)

- Sí / No es podria haver calculat l'interval de confiança com s'ha fet. Per què?

Fixa't que Sánchez Carrión utilitza la fórmula següent per a l'error típic de l'estimador:

$$\sqrt{\frac{p(1-p)}{n}}$$

No obstant això, en la majoria de manuals trobaràs, en el seu lloc:

$$\sqrt{\frac{p(1-p)}{n}}$$

Quan la grandària de la mostra és gran, aquesta variació no produeix un efecte important, però en mostres més xicotetes sí que pot afectar.

Una explicació alternativa

Pots repassar aquest tema revisant el capítol 6.5 "Estimación puntual y por intervalo de parámetros" i 6.5.1 "Estimación de proporciones. Intervalos de confianza", pàg. 194-199 del manual:

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

Comença per definir l'interès de fer estimacions d'interval, tant en contraposició a les proves d'hipòtesis com en relació amb les estimacions puntuals.

Com es pot observar en aquest manual s'opta per fer l'estimació de l'error típic basat en la proporció mostral com $\sqrt{\frac{p(1-p)}{n}}$ en lloc de $\sqrt{\frac{p(1-p)}{n-1}}$, que s'utilitzava en l'anterior manual. En la resta, l'explicació d'un i altre manual coincideixen. Aquesta explicació presenta l'avantatge, no obstant això, de mostrar gràficament els intervals de confiança sobre la corba normal de la distribució mostral.

Com calcular l'interval de confiança d'una proporció amb un programa estadístic?

Com has pogut aprendre en aquesta unitat, per a calcular l'interval de confiança d'una proporció és necessari conèixer la grandària mostral (n), la proporció de la mostra (p) i el valor $z_{n.c.}$

Amb aquestes dades és tan senzill construir l'interval de confiança d'una proporció que no necessitaries específicament un programa estadístic per a fer els càlculs.

Pots construir-ho amb un full de càlcul d'OpenOffice Calc o de Excel:

	A	B	C	D	I	F	G	H
1	Nombre	n	p	1-p	z95	Error típic (amb n)	Lím inf 95	Lím sup 95
2 El que veus	1300	2000	0,65	0,35	1,96	0,01066536	0,629	0,671
El que escrius	1300	2000	=A2/B2	=1-C2	1,96	=ARREL((C2*D2)/(B2))	=C2-(I2*F2)	=C2+(I2*F2)

En realitat no és una mala opció, ja que per exemple SPSS-PASW no calcula intervals de confiança per a les proporcions.

En STATA pots utilitzar, en el cas que les variables estiguen gravades, l'ordre "proportion var1" per a obtenir-ho. I si no tens les dades gravades, llavors pots utilitzar "cii 2000 1300" (per a seguir amb les dades de l'exemple), que retorna el resultat següent:

```
                                                    -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |       2000         .65    .0106654        .6286383    .6709211
```

És a dir, el mateix que hem obtingut amb OpenOffice Calc o Excel.

Atès que el càlcul d'intervals de confiança per a proporcions és freqüent en ciències socials es recomana preparar un full de càlcul en OpenOffice Calc o en Excel predefinit per a poder calcular-los ràpidament.

Quadern d'exercicis

Exercicis

1) Fes en un full de càlcul en OpenOffice Calc, Excel o similar per a calcular intervals de confiança al 95% i al 99% (amb tres decimals). Calcula els intervals de confiança per als casos següents:

n	p
20	0,24
50	0,50
90	0,16
15	0,20
80	0,01
100	0,24
70	0,24
35	0,50
45	0,16
16	0,20
95	0,01

Imprimeix el full de resultats i engrapa'l en el teu quadern d'exercicis. Explica els resultats. Es compleixen en tots els casos els supòsits per a poder calcular els intervals de confiança? Quin significat tindria un límit inferior el resultat del qual fóra negatiu?

2) Calcula ara l'interval de confiança (IC) al 95% i al 99% per a les mostres següents:

n	p
10	0,5
100	0,5
1000	0,5
10	0,25
100	0,25
1000	0,25
10	0,75
100	0,75
1000	0,75

Imprimeix el full de resultats i engrapa'l en el teu quadern d'exercicis. Explica els resultats. Es compleixen en tots els casos els supòsits per a poder calcular els intervals de confiança?

Respon a les preguntes següents:

- Quin és l'efecte de la grandària mostral sobre el IC?
- Quin és l'efecte del nivell de confiança sobre el IC?
- Què li ocorre a l'error típic quan p és menor que 0,5 o major que 0,5
- Quin significat tindria un límit inferior el resultat del qual fóra negatiu?

3) Fes l'exercici 3.a), 4, 5 i 10.a. del manual de Sánchez Carrión (pàg. 246).

Repàs

Al final d'aquesta unitat has de saber:

- Calcular l'interval de confiança per a la proporció d'una mostra.
- Conèixer les condicions per a poder estimar l'interval de confiança per a una mostra.
- Interpretar el significat de l'interval de confiança.
- Conèixer l'efecte del nivell de confiança en l'interval de confiança. I el de la grandària mostral i el de la dispersió.
- Comprovar la hipòtesi que la proporció de la població és igual a un valor específic i interpretar el valor p.

Exercicis de repàs:

Del manual de GARCÍA FERRANDO l'exercici 6 i 7 del tema 6.

Del manual de GARCÍA DE CORTÁZAR *ET AL*. (1996) *Estadística aplicada a las ciencias sociales*. Madrid: Cuadernos de la UNED, 1a ed., els problemes 5.1, 5.2 i 5.3.

Del manual de MULLOR, RUBEN I FAJARDO, M. Dolores (2000) *Manual práctico de estadística aplicada a las ciencias sociales*. Barcelona: Ariel, 1a ed., els exercicis resolts R.6.3, R.6.6, R.6.7, i R.6.10.

5 Comparació de dues proporcions

Planifica l'estudi:

- Aquesta unitat es treballa en una sessió de classe (1 hora de teoria + 1 hora de seminari = 2 hores).
- 5 hores d'estudi fora de l'aula.
- Setmana 7

Seqüència recomanada d'estudi:

Descripció de l'activitat	Temps	Tipus
1) Estudi de les anotacions	2 hores	No presencial
2) Classe teoria	1 hora	Presencial
3) Seminari (exercicis en l'aula)	1 hora	Presencial
4) Fer els exercicis	2 hores	No presencial
5) Repàs	1 hora	No presencial

Materials per a l'estudi:

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

Introducció

Quina taxa d'atur és més elevada, la dels homes o la de les dones? Quin és el percentatge més elevat de persones que valoren de manera negativa la situació econòmica, el dels votants del PP o del PSOE? Es protegeixen millor dels riscos de malalties de transmissió sexual les persones que són catòliques practicants o les no creients? Fumen més els homes o les dones? Què ocorre entre la població més jove? Quina proporció de persones amb estudis universitaris es troba en la població de ex-fumadors? I entre la població de fumadors? Com veus les preguntes que impliquen la comparació de proporcions són habituals en sociologia. Resoldre-les servirà per a produir hipòtesis d'interès sobre el funcionament de les relacions socials.

Objectius:

- Calcular i interpretar l'interval de confiança per a la diferència de dues proporcions.
- Provar la hipòtesi que dues proporcions de dues poblacions o grups són iguals.
- Interpretar el valor p d'una prova d'hipòtesi

Primera explicació

La comparació de proporcions de poblacions independents, igual que en calcular l'interval de confiança per a una proporció, es pot utilitzar la distribució normal com a aproximació a la distribució binomial en el cas que la mostra siga gran.

Notació:

Els símbols que utilitzarem per a referir-nos als paràmetres (de la població) i als estadístics (de la mostra) són:

	Població 1	Població 2	Mostra 1	Mostra 2
Proporció	π_1 P_1	π_2 P_2	p_1	p_2
Desviació típica	$\sqrt{\frac{\pi_1\ (1-\pi_1)}{N_1}}$	$\sqrt{\frac{\pi_2\ (1-\pi_2)}{N_2}}$	$\sqrt{\frac{p_1\ (1-p_1)}{n_1}}$	$\sqrt{\frac{p_2\ (1-p_2)}{n_1}}$
Nombre de casos	N_1	N_2	n_1	n_2

Aquests són els passos que seguirem:

1) Calcular la diferència entre les dues proporcions de les dues mostres.

Imaginem que el percentatge de persones amb educació universitària entre la població d'exfumadors és igual a 20% i en la població de fumadors igual al 15%. La diferència serà, per tant, del 5%.

2) Igual que en la prova d'hipòtesi per a dues mitjanes, es pot plantejar la hipòtesi nul·la i diverses hipòtesis alternatives.

La hipòtesi **nul·la (H_0)**, aquella sobre la qual es calcula la probabilitat (el valor p o coeficient de significació) que siga certa, serà la hipòtesi d'igualtat de proporcions:

$H_0 : \pi_1 - \pi_2 = 0$

L'habitual serà definir com a **hipòtesi alternativa (H_1)** que les proporcions són diferents:

$H_1 : \pi_1 - \pi_2 \neq 0$

Encara que, en alguns casos, podem plantejar com a hipòtesi alternativa que una proporció és major que una altra. Per exemple, en el nostre exemple, podem partir de la idea que el nivell educatiu universitari ajuda a l'hora de deixar el tabac. És a dir, que el percentatge de persones amb estudis universitaris és major entre les persones ex-fumadores que en les fumadores.

$H_1 : \pi_1 - \pi_2 < 0$

Alerta! Quan la hipòtesi alternativa siga del tipus $H_1 : \pi_1 - \pi_2 \neq 0$ utilitzarem la informació de la corba normal proporcionada per les seues dues cues o costats, mentre que quan siga del tipus $H_1 : \pi_1 - \pi_2 < 0$ utilitzarem un únic costat de la corba normal. Això és rellevant en escollir els valors que reflecteixen el nivell de confiança ($z_{n.c.}$). De fet quan vam aprendre a utilitzar i interpretar la corba normal i les seues taules, vèiem que aquestes podien estar construïdes per a una o dues cues. Per aquest motiu es parla de prova o test de "dues cues" o "d'una cua".

3) Una vegada plantejada la nostra hipòtesi, hem de triar el test o prova que ens servirà per a conèixer la probabilitat que es complisca la hipòtesi nul·la (el valor p o coeficient de significació). Atès que treballem sota la hipòtesi que es tracta d'una corba normal, el test a triar és també el test z.

Que en el cas de les proporcions de la mostra ser convertiria en:

$$z = \frac{(\pi_1 - \pi_2)}{\sigma_p}$$ que segueix una distribució ~ N(0, 1) si H_0 és certa.

Serà necessari conèixer la desviació típica de les dues proporcions.

En el tema anterior vam veure que la desviació típica d'una proporció és :

$$\sqrt{\frac{\pi\ (1-\pi)}{N}}$$

La desviació típica de les dues proporcions de les mostres serà la suma de les dues:

$$\sqrt{\frac{p_1\ (1-p_1)}{n_1}+\frac{p_2\ (1-p_2)}{n_2}}$$

Per tant en el nostre exemple:

Dèiem que el percentatge de persones amb educació universitària entre la població d'exfumadors és igual a 20% (n= 240) i en la població de fumadors igual al 15% (n=500).

És a dir:

$$\sqrt{\frac{0{,}20\ (0{,}80)}{240}+\frac{0{,}15\ (0{,}85)}{500}}=0{,}030$$

4) Amb aquestes dades ja podem substituir els elements que componen el test z, és a dir,

de $z=\frac{(\pi_1-\pi_2)}{\sigma_p}$ passa a ser $z=\frac{(0{,}20-0{,}15)}{0{,}030}=1{,}67$

En la corba normal observem que el valor z=1,67 es correspon amb una probabilitat de 0,0475, quan provem les hipòtesis H_1 : π_1 - π_2 = 0 (amb dues cues) caldria multiplicar aquest valor p per dues, és a dir 0,095. La probabilitat associada a la hipòtesi nul·la seria del 9,5%. Com normalment només optem per la hipòtesi alternativa quan aquesta probabilitat és menor del 5% (p < 0,05), seria millor acceptar la hipòtesi nul·la.

No obstant això, si haguérem plantejat com a hipòtesi alternativa H_1 : π_1 - π_2 > 0 (la prova només hauria tingut una cua), és a dir, el valor p=0,0475 ens parla d'una probabilitat d'encertar amb la hipòtesi nul·la del 4,75%, menor del 5% i, per tant, es preferiria la hipòtesi alternativa. En aquest cas significaria que la idea que el nivell educatiu universitari ajuda a l'hora de deixar el tabac pot ser vàlida, ja que el percentatge de persones amb estudis universitaris és major entre les persones exfumadores que en les fumadores.

Per aquest motiu, no s'ha de fer un test sense haver-hi argumentant amb anterioritat si es tracta d'una prova amb una o dues cues (asimètric o simètric).

De totes maneres, les dades d'aquest exemple són fictícies i no tenen en compte aspectes clau per a entendre la relació estudiada entre nivell d'estudis i deixar de fumar com poden ser el gènere o l'edat.

Com comparar dues proporcions amb un programa estadístic?

1) Imaginem que s'ha arreplegat la informació sobre si fumen (1) o no (0), 20 homes i 20 dones. Es vol respondre a la pregunta de si homes i dones fumen per igual.

Sexe	Fuma	Sexe	Fuma
Home	1	Dona	1
Home	1	Dona	1
Home	1	Dona	1
Home	1	Dona	1
Home	1	Dona	1
Home	0	Dona	0
Home	0	Dona	0
Home	0	Dona	0
Home	0	Dona	0
Home	0	Dona	0
Home	1	Dona	0
Home	1	Dona	0
Home	1	Dona	0
Home	1	Dona	0
Home	1	Dona	0
Home	0	Dona	0
Home	0	Dona	0
Home	0	Dona	0
Home	0	Dona	0
Home	0	Dona	0

2) Per a aplicar un test caldria valorar primer si la *n* és prou gran

   $p\,n \geq 5$
   $(1\text{-}p)\,n \geq 5$

Observem que està en el límit del que es considera suficientment gran (per, exemple 0,25 * 20 = 5), per la qual cosa podríem aplicar el test z.

3) Si utilitzem SPSS ens trobarem que no hi ha una ordre específica per a fer una prova de comparació de dues proporcions. Has de calcular-ho a mà (pots fer-ho com a exercici de repàs; en el punt 4 d'aquesta numeració trobaràs la solució).[5]

4) Amb STATA es pot obtenir la resposta directa amb la ordre següent:

[5] Una forma indirecta en SPSS-PASW és fer un encreuament de les dues variables (en el cas de taules de 2 x 2) i fer un contrast khi quadrat, que informa sobre si hi ha associació entre les dues variables (és a dir, no informa directament sobre la diferència entre les proporcions). Aquest contrast ho explicarem en el tema següent.

```
  prtest Fuma, by(Sexo)

Two-sample test of proportion                Hombre: Number of obs =       20
                                              Mujer: Number of obs =       20
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Hombre |          .5   .1118034                      .2808694    .7191306
       Mujer |        .25   .0968246                      .0602273    .4397727
-------------+----------------------------------------------------------------
        diff |        .25    .147902                     -.0398826    .5398826
             |  under Ho:   .1530931     1.63   0.102
------------------------------------------------------------------------------
        diff = prop(Hombre) - prop(Mujer)                         z =   1.6330
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.9488         Pr(|Z| < |z|) = 0.1025          Pr(Z > z) = 0.0512
```

Com s'observa, la hipòtesi nul·la que les dues proporcions són iguals produeix un valor z=1,63 que s'associa amb una probabilitat bilateral p = 0,102, és a dir, no podem rebutjar la hipòtesi nul·la.

Quadern d'exercicis

Exercicis

1) Fes l'exercici 5 del manual de Sánchez Carrión (pàg. 472).
2) Fes l'exercici 3 del capítol 10 del manual de García Ferrando.

Repàs

Al final d'aquesta unitat has de saber:

- Calcular i interpretar l'interval de confiança per a la diferència de dues proporcions.
- Provar la hipòtesi que dues proporcions de dues poblacions o grups són iguals.
- Interpretar el valor p d'una prova d'hipòtesi

Exercicis de repàs:

Del manual de GARCÍA DE CORTÁZAR *et al*. (1996) *Estadística aplicada a las ciencias sociales*. Madrid: Cuadernos de la UNED, 1a ed., el problema 6.8.

Del manual de MULLOR, RUBEN I FAJARDO, M. Dolores (2000) *Manual práctico de estadística aplicada a las ciencias sociales*. Barcelona: Ariel, 1a ed., els exercicis resolts 7.3, 7.7 i 7.9.

6 Taules de contingència

Planifica l'estudi:

- Aquesta unitat es treballa en tres sessions de classe (3 hores de teoria + 3 hores de seminari = 6 hores).
- 14 hores d'estudi fora de l'aula.
- Setmanes 8, 9 i 10.

Seqüència recomanada d'estudi:

Descripció de l'activitat	Temps	Tipus
1) Estudi de les anotacions	1 hora	No presencial
2) Classe de teoria	1 hora	Presencial
3) Seminari d'exercicis en l'aula	1 hora	Presencial
4) Estudi de les anotacions	1 hora	No presencial
5) Fer els exercicis	3 hores	No presencial
6) Classe teoria	1 hora	Presencial
7) Seminari d'exercicis en l'aula	1 hora	Presencial
8) Estudi de les anotacions	2 hores	No presencial
9) Acabar els exercicis	2 hores	No presencial
10) Classe de teoria	1 hora	Presencial
11) Seminari d'exercicis en l'aula	1 hora	Presencial
12) Estudi de les anotacions	2 hores	No presencial
13) Acabar els exercicis del quadern d'exercicis	2 hores	No presencial
14) Repàs	1 hora	No presencial

Materials per a l'estudi:

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

Introducció

Les taules de contingència són la forma més comuna de treball amb les dades d'una enquesta. Són una forma eficaç de presentar els resultats d'una investigació atenent als creus imprescindibles en l'anàlisi de la investigació social (per sexe, edat, territori, etc.), així com per altres variables que s'estiguen analitzant a cada moment.

Presenten el gran avantatge que part de les proves estadístiques associades a les taules de contingència poden ser aplicades fins i tot amb variables nominals.

Amb aquestes podrem argumentar, per exemple, si hi ha associació entre ser home i dona i algun determinat comportament (fumar o no fumar), situació (estar en l'atur), opinió (valorar positiva o negativament l'acció de govern), etcètera.

Objectius:

- Construir una taula de contingència i comprovar mitjançant una prova khi quadrat si hi ha associació.
- Comprovar els supòsits estadístics per a poder calcular khi quadrat.
- Calcular el coeficient de contingència C, el coeficient Gamma, rho de Spearman i interpretar les proves estadístiques de la hipòtesi d'associació.
- Calcular i interpretar els coeficients V de Cramer, lambda, d de Sommers, Tau-a i Tau-b de Kendall.

Primera explicació

Quan es vol estudiar la relació entre dues variables, aquestes es poden representar en una taula conjuntament. Això és habitual en qualsevol anàlisi d'enquestes. Prenem l'exemple de l'Enquesta Nacional de Salut Sexual, en la qual es va preguntar als homes:

A continuació, podria expressar si esteu molt, bastant, poc o gens d'acord amb les afirmacions següents?

		Molt d'acord %	Bastant d'acord %	Poc d'acord %	Gens d'acord %	N. S. %	N.C. %	TOTAL (N)
Em considere una persona bastant atractiva	Sense estudis	6.3	32.5	32.8	20.9	3.8	3.6	(332)
	Primària	7.5	40.8	35.6	10.0	4.7	1.2	(2068)
	Secundària	9.7	47.5	31.9	5.4	3.6	1.9	(739)
	FP	7.1	47.8	34.3	6.2	3.1	1.5	(795)
	Mitjans universitaris	5.0	42.4	36.8	8.1	5.7	2.0	(356)
	Superiors	8.6	52.6	31.1	4.3	2.1	1.2	(508)
	N\c.	12.5	47.0	37.3	.	3.2	.	(31)
	TOTAL	7.7	43.8	34.3	8.6	4.0	1.6	(4832)

Com s'observa sembla que tenen un millor autoconcepte estètic els homes amb més nivell d'estudis que els que compten amb menys anys d'escolarització.

Si en lloc d'afirmar "sembla que", volem fer una afirmació probabilística, podrem fer una prova estadística.

La hipòtesi nul·la serà que les dues variables (estudis i autoconcepte estètic) no estan associades i la hipòtesi alternativa que hi ha associació.

Per a conèixer això convé que treballem la nostra taula amb les freqüències en lloc dels percentatges.

Per a fer-ho, en SPSS hem de:

1. Comprovar les dues variables objecte d'estudi (estudis de l'entrevistat i "em considere una persona bastant atractiva").
2. Ponderar la mostra utilitzant la variable "Ponderación" en *Datos / Ponderar casos*
3. Seleccionar en la mostra als homes (variable "P57") en *Datos / Seleccionar casos / Si satisface la condición...*
4. Construir la taula de contingència amb les freqüències amb *Analizar / Estadísticos descriptivos / tablas de contingencia...*

Tabla de contingencia Estudios del entrevistado * Me considero una persona bastante atractiva

Recuento

		Me considero una persona bastante atractiva						
		Muy de acuerdo	Bastante de acuerdo	Poco de acuerdo	Nada de acuerdo	N.S.	N.C.	Total
Estudios del entrevistado	Sin estudios	21	108	109	69	13	12	332
	Primaria	156	845	737	208	97	26	2069
	Secundaria	72	351	236	40	26	14	739
	F.P.	56	381	273	50	24	12	796
	Medios universitarios	18	151	131	29	20	7	356
	Superiores	44	267	158	22	11	6	508
	N.C.	4	15	12	0	1	0	32
Total		371	2118	1656	418	192	77	4832

Nota: els totals presenten algunes petites variacions pel que fa a les dades publicades pel CIS (vegeu les categories primària i n.c.)

Amb la taula de freqüències observades es pot construir una taula de freqüències esperades en el cas que no hi haguera associació entre les variables.

Per exemple, en el cas que no hi haguera associació entre les variables s'esperaria que els homes sense estudis que estan molt d'acord amb l'afirmació "em considere una persona bastant atractiva" seria igual a (371 * 332)/4832 = 25,49. És a dir, el resultat de multiplicar el total de la seua columna pel total de la seua fila i dividir el producte pel total de casos. Com veuràs el percentatge de columna que s'obté amb 25,49 és (25,49/371)*100 = 6,87%. El mateix que si calculem el percentatge de columna del total, és a dir (332/4832)*100 = 6,87%.

En definitiva, en la taula de freqüències esperades els percentatges de cada columna (i els de cada fila) de cada categoria són iguals als percentatges que es calculen amb el total de columna (i el de fila). Per això es denomina *taula de no-associació*, atès que ni les files, ni les columnes aporten més informació que l'aportada pels totals.

Procedim de manera anàloga a com hem fet amb la primera cel·la de la taula i calculem la resta de freqüències esperades en cas de no-associació:

Tabla de contingencia Estudios del entrevistado * Me considero una persona bastante atractiva

Frecuencia esperada

		Me considero una persona bastante atractiva						
		Muy de acuerdo	Bastante de acuerdo	Poco de acuerdo	Nada de acuerdo	N.S.	N.C.	Total
Estudios del entrevistado	Sin estudios	25,5	145,5	113,8	28,7	13,2	5,3	332,0
	Primaria	158,9	906,9	709,1	179,0	82,2	33,0	2069,0
	Secundaria	56,7	323,9	253,3	63,9	29,4	11,8	739,0
	F.P.	61,1	348,9	272,8	68,9	31,6	12,7	796,0
	Medios universitarios	27,3	156,0	122,0	30,8	14,1	5,7	356,0
	Superiores	39,0	222,7	174,1	43,9	20,2	8,1	508,0
	N.C.	2,5	14,0	11,0	2,8	1,3	,5	32,0
Total		371,0	2118,0	1656,0	418,0	192,0	77,0	4832,0

Com pots veure es pot calcular igualment amb SPSS: una vegada estàs en el quadre diàleg de taules de contingència, prem el botó "*casillas*" i selecciona "*frecuencias esperadas*".

Atès que les freqüències esperades mostren la situació hipotètica en la qual no hi hauria associació entre dues variables, podem comparar aquesta situació hipotètica amb l'obtinguda en la mostra (les freqüències observades). La prova estadística que s'utilitza per a fer aquesta operació es coneix com khi quadrat[6] (χ^2).

$$\chi^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2 (r-1)(s-1) \quad \text{si } H_0 \text{ certa}$$

La distribució de la prova khi quadrat és una distribució khi quadrat amb r-1 * s-1 graus de llibertat si H_0 és certa, sent *r* el nombre de columnes i *s* el nombre de files.

És a dir, els graus de llibertat d'una taula de doble entrada són iguals al nombre de files menys una pel nombre de columnes menys una. En el nostre cas (7-1) * (6-1) = 30.

$$\chi^2 = \frac{(21-25{,}5)^2}{25{,}5} + \frac{(156-158{,}9)^2}{158{,}9} + (...) = 155{,}6$$

El valor de khi quadrat és igual a 155,6

La distribució de khi quadrat depèn del nombre de graus de llibertat (d.f=30). En aquest cas, per a un nivell de significació 0,05 (amb dues cues) podem comprovar que el valor límit és igual a 43,77. La regió de rebuig de la hipòtesi nul·la se situa a partir d'aquest valor. Com el nostre valor khi quadrat (155,6) es troba dins d'aquesta regió de rebuig, caldrà rebutjar la hipòtesi nul·la i pensar que les dues variables estan associades.

$155 \in RR(0{,}05) = \{\chi^2 > 43{,}77\}$, de manera que rebutgem la hipòtesi nul·la.

A continuació pots observar el resultat produït per SPSS per a l'exemple estudiat:

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	155,571(a)	30	,000
Razón de verosimilitudes	145,823	30	,000
Asociación lineal por lineal	25,125	1	,000
N de casos válidos	4832		

a 4 casillas (9,5%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es ,51.

[6] El so de la lletra grega χ és similar a el de la "jota" castellana, per això ací s'opta per "khi quadrat" per a transcriure'l a causa que no comptem en català amb un so equivalent.

Supòsits estadístics per a aplicar khi quadrat

La prova khi quadrat segueix una distribució khi quadrat sempre que el nombre de casos siga suficientment gran.

Per a determinar si el nombre de casos és suficientment gran les freqüències esperades no han de ser molt baixes. D'acord amb la recomanació de Cochram que recorda García Ferrando (2008; 300), com a màxim el 20% de les cel·les poden tenir valors menors de 6 i majors d'1. En l'exemple estudiat fins a un 9,5% tenen una freqüència esperada menor que 5 i almenys una cel·la compta amb una freqüència esperada igual a 0,5, per la qual cosa no es compleix per complet la recomanació de Cochram.

En l'exemple, hauria estat millor construir la taula de contingència o taula de doble entrada de freqüències observades i esperades sense considerar els n.s. i n.c. El motiu és que l'associació es podria produir per la influència d'aquestes categories, a les quals en principi és molt difícil assignar alguna interpretació, igualment, per presentar un baix nombre de casos, la qual cosa afecta al càlcul de khi quadrat (cosa que seria en certa manera equivalent a comptar amb casos extrems en el càlcul d'una mitjana).

En l'exemple, si s'eliminen els n.s. i n.c. de les variables objecte d'estudi el resultat del càlcul de khi quadrat és, com es pot observar a continuació, sensiblement inferior al valor obtingut prèviament, però també es redueixen els graus de llibertat a la meitat (de 30 es redueixen a 15). En qualsevol cas la probabilitat associada a khi quadrat segueix sent molt baixa, de manera que també caldria rebutjar la hipòtesi nul·la. Ara la freqüència mínima esperada per cel·la ha augmentat de 0,51 a 24,86 i cap cel·la té freqüències esperades inferiors a 5. Per tant, podem acceptar sense reserves la conclusió.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	127,973(a)	15	,000
Razón de verosimilitudes	116,295	15	,000
Asociación lineal por lineal	38,520	1	,000
N de casos válidos	4532		

a 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 24,86.

El resultat de khi quadrat té com a limitació que no permet conèixer l'adreça de l'associació, és a dir, sabem que nivell d'estudis i satisfacció amb la pròpia imatge estan associats, però no sabem si les persones amb millor nivell d'estudis tenen millor autoconcepte estètic o és al contrari.

Una altra limitació és que no permet conèixer el grau d'associació entre les dues variables.

Una forma de conèixer el grau d'associació entre dues variables és calcular el *coeficient de contingència C.*

El seu càlcul una vegada s'ha construït una taula de contingència i calculat khi quadrat resulta molt directe, ja que n'hi ha prou amb aplicar la fórmula

$$C = \sqrt{\frac{x^2}{x^2 + n}}$$

És a dir, en el nostre exemple anterior (ja sense considerar els n.s. i n.c.)

$$C = \sqrt{\frac{128}{128 + 4532}} = 0{,}17$$

En SPSS:

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Coeficiente de contingencia	,166	,000
N de casos válidos		4532	

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

La hipòtesi a contrastar en aquest cas seria:

H_0 : no hi ha correlació en la població.
H_1: sí hi ha correlació en la població.

Pel nivell de significació obtingut hauríem d'optar per la hipòtesi alternativa.

Les limitacions de coeficient de contingència C són:

1) El seu valor mínim pot ser zero, però el màxim és diferent d'1.
2) El límit superior depèn del nombre de files i columnes de la taula, quan el nombre de files i columnes és igual, el valor màxim és igual a l'arrel quadrada de (k-1)/k sent k el nombre de columnes o de files. En una taula 2 x 2 l'arrel quadrada de ½ és 0,707. En una taula de 3 x 3 és 0,816. Els valors C de diverses taules només són comparables si les taules tenen el mateix nombre de files i columnes.
3) S'apliquen les mateixes restriccions que en el càlcul de khi quadrat sobre el percentatge de cel·les amb freqüències esperades baixes.

Una forma d'aconseguir un coeficient el valor del qual se situe entre 0 i 1 és calcular el *coeficient V de Cramer*

Per al seu càlcul cal considerar un valor *t* que representa el valor més xicotet de les dues quantitats r-1 o s-1, on *r* i *s* són el nombre de columnes i de files. És a dir, en el nostre exemple, *t* igual a 3.

$$V = \sqrt{\frac{x^2}{n \cdot t}}$$

$$V = \sqrt{\frac{128}{4532 \cdot 3}} = 0{,}097$$

El valor seria per tant igual a 0,097

D'acord amb SPSS:

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,168	,000
	V de Cramer	,097	,000
N de casos válidos		4532	

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

El coeficient Lambda

Quan dues variables nominals estan associades, conèixer la distribució d'una de les variables millora les prediccions que podem fer sobre l'altra variable. En el nostre exemple sobre el nivell d'estudis i l'autoconcepte estètic, si no tinguérem la variable estudis a l'hora de predir quin és l'autoconcepte de la població masculina espanyola hauríem de deixar-nos guiar pel valor de la moda. És a dir, el més probable seria pensar que un individu està bastant d'acord amb l'afirmació "em considere una persona atractiva", ja que un 46,4% se situa en aquesta categoria.

Quan afegim la informació de la variable estudis, a la qual podem considerar independent (en principi, el raonable és pensar que l'autoconcepte depèn del nivell d'estudis i no al revés), podem reduir el nivell d'error de la nostra predicció, ja que en el moment actual només encertaríem en un 46,4% de les vegades, però ens equivoquem en el 53,6% de les ocasions.

Tabla de contingencia Estudios del entrevistado * Me considero una persona bastante atractiva

			Me considero una persona bastante atractiva				
			Muy de acuerdo	Bastante de acuerdo	Poco de acuerdo	Nada de acuerdo	Total
Estudios del entrevistado	Sin estudios	Recuento	21	108	109	69	307
		% de Estudios del entrevistado	6,8%	35,2%	35,5%	22,5%	100,0%
	Primaria	Recuento	156	845	737	208	1946
		% de Estudios del entrevistado	8,0%	43,4%	37,9%	10,7%	100,0%
	Secundaria	Recuento	72	351	236	40	699
		% de Estudios del entrevistado	10,3%	50,2%	33,8%	5,7%	100,0%
	F.P.	Recuento	56	381	273	50	760
		% de Estudios del entrevistado	7,4%	50,1%	35,9%	6,6%	100,0%
	Medios universitarios	Recuento	18	151	131	29	329
		% de Estudios del entrevistado	5,5%	45,9%	39,8%	8,8%	100,0%
	Superiores	Recuento	44	267	158	22	491
		% de Estudios del entrevistado	9,0%	54,4%	32,2%	4,5%	100,0%
Total		Recuento	367	2103	1644	418	4532
		% de Estudios del entrevistado	8,1%	46,4%	36,3%	9,2%	100,0%

Una forma de reduir l'error de la nostra predicció és usar la informació produïda en la taula de doble entrada després d'haver inclòs la informació de la variable nivell d'estudis.

El nostre objectiu serà per tant fer prediccions de l'autoconcepte estètic (dependent) a partir del nivell d'estudis (independent).

Per a això podem sumar les modes de la variable dependent, dins de cada categoria de la variable independent, és a dir, 109 + 845 + 351 + 381 + 151 + 267 = 2104 (ho denominarem Σm_y)

Com hem dit, la moda de la variable dependent és igual a 2103 (la notació serà M_y)

El nombre de casos és 4532 (n)

Amb tot això podem calcular el coeficient lambda:

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} = \frac{2104 - 2103}{4532 - 2103} = 0{,}0004$$

El valor obtingut és molt baix. Això significaria que, coneixent la variable independent, no es millora molt la predicció sobre la variable dependent. Això és així a pesar que havíem assenyalat que hi ha associació entre les dues variables. La V de Cramer ja informava que el grau d'associació era baix.

El coeficient lambda varia entre el valor 0 i 1 (amb independència de la grandària de la taula i de la mostra). Un valor pròxim a 1 significaria que l'error de la predicció a partir únicament de la variable dependent es redueix completament en introduir la informació de la independent. En aquest cas, el valor pròxim a 0 significa que per a totes les categories de la variable independent la moda és quasi sempre igual a la moda de la dependent. És a dir, en tots els nivells d'estudis el més probable és que assenyalen estar "bastant d'acord" amb l'afirmació "em considere una persona atractiva".

Per tant, quan hi ha associació lambda podria ser 0 o pròxim a 0. Això es deu al fet que el grau d'associació, com hem vist, es calcula a partir de la distribució de les freqüències observades en les cel·les (té en compte la informació de cada cel·la) i per al càlcul de lambda s'utilitza únicament la moda. Dit d'una altra manera, lambda és una mesura associada a la predicció del valor de la moda i khi quadrat a les diferències entre les freqüències de les columnes. Igualment amb aquestes dades podríem plantejar-nos comparacions de proporcions amb les proves estadístiques que estudiem en el tema 5.

En SPSS el càlcul de lamba produeix el resultat següent:

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,000	,003	,068	,946
		Estudios del entrevistado dependiente	,000	,000	.[c]	.[c]
		Me considero una persona bastante atractiva dependiente	,000	,006	,068	,946
	Tau de Goodman y Kruskal	Estudios del entrevistado dependiente	,005	,001		,000[d]
		Me considero una persona bastante atractiva dependiente	,008	,002		,000[d]

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

c. No se puede efectuar el cálculo porque el error típico asintótico es igual a cero.

d. Basado en la aproximación chi-cuadrado.

Amb l'eixida de SPSS, en ser lambda un valor molt baix, no la podem conèixer amb precisió (ja que només informa de les tres primeres xifres de decimals). En qualsevol cas, el resultat assenyala amb claredat que es tracta d'un valor molt baix.

Lambda es considera una mesura RPE o de Reducció Proporcional d'Error. Aquestes mesures calculen el percentatge d'error que es redueix quan afiges les dades d'una segona variable per a predir una variable inicial. En aquest cas, hem observat que en afegir les dades del nivell d'estudis no es redueix molt l'error de la predicció sobre la variable autoconcepte estètic.

Variables ordinals

Per a l'estudi d'aquest tema seguirem directament l'explicació del manual, completant en classe amb el càlcul manual a partir de l'exemple de l'Enquesta Nacional de Salut Sexual i amb les operacions que fa el programa estadístic SPSS:

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

1) Començarem per la lectura de l'apartat 8.3 i 8.3.1. "Tipus i càlculs de parells" (pàg. 244-250).

L'objectiu és comprendre el càlcul en una taula amb variables ordinals del nombre total de parells diferents, de parells semblants o concordants, discordants, "empatats" en la variable *x*, els "empatats" en la variable y, i finalment, els "empatats" en *x* i *y*.

La nostra taula anterior es pot considerar una taula amb dues variables ordinals: d'una banda, el nivell d'estudis es pot ordenar de major a menor nivell d'estudis i l'autoconcepte estètic de major a menor acord amb l'afirmació "em considere una persona bastant atractiva".

Per tant, es podrien calcular els parells esmentats en aquest apartat. Es recomana que ho faces manualment. Alerta amb l'ordenació de les files i columnes. La taula de l'Enquesta Nacional de Salut Sexual no està ordenada amb la mateixa pauta que el manual, ja que les files estan ordenades de forma ascendent en el nostre exemple i descendent en el manual. En les columnes ocorre els contrari, el nostre exemple les ordena de forma descendent (cada columna expressa un menor grau d'acord) i el manual en ordre ascendent (cada columna implica un major nivell educatiu dels pares).

2) En 8.3.2. s'explica el càlcul del coeficient Tau-a de Kendall. (pàg. 250), el coeficient Gamma de Goodman i Kruskal (8.3.3, pàg. 250-251), el coeficient d de Somers (8.3.4, pàg. 252) i el Tau-b de Kendall (8.3.5, pàg. 253).

Es recomana que reproduïsques els exemples del llibre i després faces el càlcul amb la taula de l'Enquesta Nacional de Salut sexual.

3) En 8.3.6 s'explica el coeficient rho de Spearman (pàg. 253-255).

Les explicacions de les mesures d'associació se situen en el camp de l'estadística descriptiva. No obstant això, el normal és que treballem amb mostres i necessitem conèixer si la mostra ens permet fer generalitzacions sobre la població (és a dir, traslladar-nos al camp de l'estadística inferencial). Hi ha diferents proves d'hipòtesis de l'associació per a cadascuna de les mesures d'associació estudiades. En les pàg. 301 a 306 s'expliquen les proves d'hipòtesis per al coeficient de contingència C, el coeficient rho de Spearman i el coeficient Gamma.

Amb aquesta informació pots interpretar millor el resultat que calcula un programa estadístic com SPSS:

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	d de Somers	Simétrica	-,084	,013	-6,573	,000
		Estudios del entrevistado dependiente	-,090	,014	-6,573	,000
		Me considero una persona bastante atractiva dependiente	-,078	,012	-6,573	,000

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Medidas simétricas

		Valor	Error típ. asint.(a)	T aproximada(b)	Sig. aproximada
Nominal por nominal	Coeficiente de contingencia	,166			,000
Ordinal por ordinal	Tau-b de Kendall	-,084	,013	-6,573	,000
	Gamma	-,122	,018	-6,573	,000
N de casos válidos		4532			

a Asumiendo la hipótesis alternativa.

b Empleando el error típico asintótico basado en la hipótesis nula.

Nota: els valors negatius s'expliquen per la manera en la qual estan codificades les variables. El nivell d'estudis està codificat de forma ascendent: 1. Sense estudis, 2. Primària, 3. Secundària, etc., però l'autoconcepte estètic està codificat de forma descendent. 1. Molt d'acord, 2 Bastant d'acord, (...), 4. Gens d'acord. Per això un signe negatiu indica que a major nivell d'estudis, major és l'acord amb la frase "em considere una persona atractiva".

L'explicació del manual

La segona part de l'explicació, la referida a les variables ordinals, l'hem estudiat directament a partir del manual de:

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

Per a l'estudi de la primera part del tema (mesures nominals), es recomana igualment el mateix manual.

En el seu capítol 10.3 "La prueba de chi-cuadrado para dos muestras" (pàg. 294-300) es desenvolupa l'explicació de la prova khi quadrat.

Es recomana que seguisques la seua explicació fent els exemples que proposa a mà. És a dir, que calcules khi quadrat per a la taula sobre sexe i religiositat i la taula 2 "Relación entre el nivel de estudios terminados y la ocupación de los padres".

Per a l'estudi del coeficient lambda has de retrocedir fins al capítol 8.2.1. "El coeficiente lambda" (pàg. 236-241). A partir de dos exemples, que es recomana que calcules pel teu compte, s'explica el càlcul de lambda i la seua interpretació.

Quadern d'exercicis

Exercicis

1) Fes els exercicis del capítol 8 del manual de García Ferrando. Exercicis 1 a 5 (pàg. 257-259)
2) Fes els exercicis 4 a 6 del capítol 10 del manual de García Ferrando (pàg. 308-309).

Repàs

Al final d'aquesta unitat has de saber:

- Construir manualment i amb programa estadístic una taula de contingència i comprovar mitjançant una prova khi quadrat si hi ha associació.
- Comprovar els supòsits estadístics per a poder calcular khi quadrat a partir de la taula de freqüències esperades.
- Calcular manualment i amb programa estadístic el coeficient de contingència C, el coeficient Gamma, rho de Spearman i interpretar les proves estadístiques de la hipòtesi d'associació.
- Calcular i interpretar els coeficients V de Cramer, lambda, d de Sommers, Tau-a i Tau-b de Kendall.

Exercicis de repàs:

Amb l'Enquesta Nacional de Salut Sexual estudia la relació entre nivell d'estudis i grau d'acord amb l'afirmació "em considere una persona atractiva" per al cas dels dones, seguint els mateixos passos que s'han explicat per al cas dels homes.

Del manual de García de Cortázar *et al*. (1996) *Estadística aplicada a las ciencias sociales*. Madrid: Cuadernos de la UNED, 1a ed., els problemes 8.1, 8.2, 8.3, 8.4, 8.5, 8.6, 8.7, 8.8, 8.9, 8.10, 8.11 i 8.12.

Del manual de Mullor, Ruben i Fajardo, M. Dolores (2000) *Manual práctico de estadística aplicada a las ciencias sociales*. Barcelona: Ariel, 1a ed., els exercicis resolts 8.1, 8.2 i 8.9.

Per a saber-ne més

Pots llegir l'apartat 8.2.2. "El coeficiente Tau-y de Goodman y Kruskal" (241-244) de García Ferrando, on s'explica una altra mesura de reducció proporcional de l'error per a variables nominals.

7 Correlació i regressió lineal

Planifica l'estudi:

- Aquesta unitat es treballa en dues sessions de classe (2 hores de teoria + 2 hores de seminari = 4 hores).
- 5 hores d'estudi fora de l'aula + 6 hores de repàs (unitats B4, B5, B6 i B7)
- Setmanes 11, 12 i 13.

Seqüència recomanada d'estudi:

Descripció de l'activitat	Temps	Tipus
1) Estudi de les anotacions	2 hores	No presencial
2) Classe de teoria i seminari	2 hores	Presencial
3) Fer els exercicis	2 hores	No presencial
4) Classe de teoria i seminari	2 hores	Presencial
5) Acabar els exercicis del quadern d'exercicis	1 hora	No presencial
Repàs unitats B4, B5, B6 i B7 per a l'examen parcial d'11 de maig	6 hores	No presencial

Introducció

Quant pesarà la teua parella d'ací a 10 anys? Hi ha una relació entre els anys d'escolarització i el salari mitjà? Hi ha una relació entre edat i ubicació ideològica? I entre salari mitjà i ubicació ideològica?

La relació entre dues variables d'interval es pot representar gràficament. En ocasions aquesta relació és lineal. Quan es troba una relació lineal es pot valorar mitjançant el coeficient de Pearson el grau d'associació entre les variables. Si la relació és lineal i el grau d'associació elevat, és possible fer prediccions dels valors de la variable dependent per a un valor donat de la variable independent.

Objectius:

- Representar gràficament la relació entre dues variables d'interval.
- Mesurar el grau d'associació mitjançant el coeficient *r* de Pearson.
- Conèixer els supòsits en els quals és apropiat fer una correlació.
- Estimar i representar la línia de regressió.
- Calcular intervals de confiança per al coeficient de regressió.
- Comprovar la hipòtesi de no-associació.
- Aplicar la línia de regressió quan siga adequat.

Primera explicació

Les variables quantitatives o d'interval es classifiquen en variables discretes (els valors de les quals són nombres enters) i contínues (els valors de les quals poden contenir infinits nombres fraccionats [nombres fraccionaris]).

Quan pensem que els valors d'una variable quantitativa tendeixen a créixer (o decréixer) a mesura que s'incrementen els valors d'una altra variable, estem davant una situació de possible **correlació.**

Si hi ha correlació llavors podrem predir el valor de la variable dependent per a un valor donat de la variable independent, això ho denominem **regressió**.

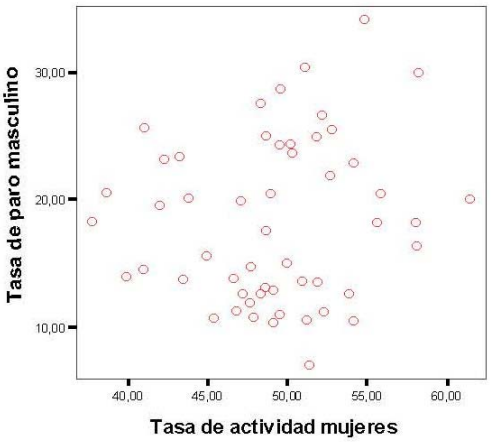
Per a conèixer l'existència d'una situació de correlació lineal el primer pas és la representació gràfica de la informació.

Podem partir d'un exemple: quan l'ocupació femenina va començar a convertir-se en norma social, un argument freqüent des dels defensors dels valors patriarcals era que una major participació femenina en el treball es traduiria en una major taxa d'atur masculí. Per a examinar si aquesta situació

es compleix a Espanya en el moment actual podem consultar les dades d'activitat femenina i atur masculí de l'Enquesta de Població Activa (INE):

Les dades del segon semestre de 2010 són:
http://www.ine.es/daco/daco42/daco4211/epapro0210.pdf

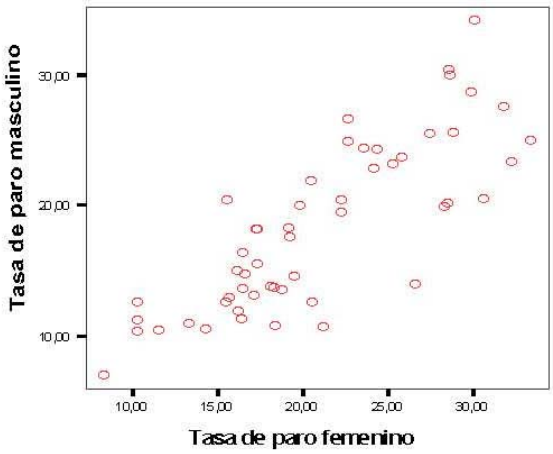
Si seleccionem els casos en el nivell provincial i els representem gràficament en un eix de coordenades, podem produir un gràfic com a aquest:



- El gràfic es pot produir amb SPSS en *Graficos* > *Interactivos* > *Diagrama de dispersión*
- En STATA es produeix amb "graphics > twoway graphics ... create / scatter..." o amb el comandament "twoway (scatter var1 var2)".

En principi, com es pot observar, els punts formen un núvol més o menys homogèni. Per la qual cosa no es pot establir visualment relació lineal entre la taxa d'activitat de les dones i la taxa d'atur masculí.

No obstant això, si representem la taxa d'atur masculina i femenina, obtenim:



En aquest cas, la relació gràfica sí que sembla ser lineal: els dos fenòmens semblen estar associats (o, més específicament, **correlacionats**), és a dir, en les províncies on hi ha més atur femení hi ha més atur masculí i viceversa.

Per a conèixer el grau d'associació es pot estimar un coeficient de correlació, conegut amb el nom de *r* de Pearson (r):

Aquest coeficient pot prendre un valor que oscil·la entre -1 i 1, i mesura el grau de dispersió dels punts.

El seu càlcul és:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$\bar{x}$ i $\bar{y}$ representen la mitjana de la variable x i y respectivament.

x_i Representa cada observació individual de la variable *x*, en aquest cas la taxa d'activitat de les dones, és a dir, Almeria 58,23, Cadis 48,35, Còrdova 48,72, etc.

y_i representa les observacions individuals de la variable *y* (la taxa d'atur masculí en cada província)

Com veus, el càlcul es basa en les distàncies entre cada observació (parells, *x*,*y*) i la mitjana de *x* i *y*.

Càlcul

L'avantatge per al seu càlcul és que es pot fer amb qualsevol programa estadístic i fins i tot amb les calculadores científiques.

En la teua calculadora:

- Usa el mode "LR"
- Abans de gravar les dades, esborra la informació anterior.
- Les dades han de gravar-se acuradament, per parells, en primer lloc la variable *x* i després la variable *y*. La seqüència sol ser "primer valor de *x*", "tecla (x_D,y_D)", "primer valor de *y*", tecla "data" o "RUN", "segon valor de *x*", "tecla (x_D,y_D)", "segon valor de *y*", etc.
- Una vegada gravats pots obtenir el valor "*r*"
- Posteriorment veurem que també permet obtenir les dades de la recta de regressió i els valors predits per la recta.
- El procediment pot variar d'unes calculadores a unes altres, però sol estar explicat en el manual de la calculadora. Busca l'apartat "linear regression". Si no conserves el manual, sol ser fàcil trobar-los en Internet.

En SPSS-PASW, s'obté amb *"Analizar" / "Regresión lineal"*
o amb la sintaxi:

```
REGRESSION
  /DEPENDENT VAR00001
  /METHOD=ENTER VAR00002  .
```

En STATA

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Correlations and covariances

o amb la sintaxi:

```
correlate var1 var2
```

Per exemple, amb les dades:

X	I
10	1003
15	1005
20	1010
25	1008
30	1014

El resultat amb la calculadora és r = 0,919018277

Amb SPSS r = 0,919

Resum del model

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,919(a)	,845	,793	3,59867

a Variables predictoras: (Constante), VAR00002

Amb STATA r = 0,9190

```
. correlate var1 var2
(obs=5)

             |     var1     var2
-------------+------------------
        var1 |   1.0000
        var2 |   0.9190   1.0000
```

En l'exemple sobre la taxa d'activitat femenina i atur masculí es trobarien els resultats següents:

Amb SPSS:

El valor *r* per a la correlació entre taxa d'activitat activitat femenina i atur masculí és igual a r = 0,13

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,134(a)	,018	-,002	6,45011

a Variables predictoras: (Constante), Tasa de actividad femenina

El valor *r* per a la correlació entre taxa d'atur femení i masculí és igual a r = 0,81

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,813(a)	,661	,654	3,78892

a Variables predictoras: (Constante), Tasa de paro femenino

Amb STATA:

S'obtenen els resultats següents (que són idèntics):

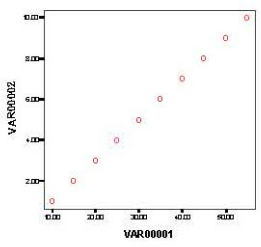
```
correlate ACTFEM PAROMASC
(obs=52)
             |   ACTFEM PAROMASC
-------------+------------------
      ACTFEM |   1.0000
    PAROMASC |   0.1340   1.0000


. correlate PAROFEM PAROMASC
(obs=52)
             |  PAROFEM PAROMASC
-------------+------------------
     PAROFEM |   1.0000
    PAROMASC |   0.8131   1.0000
```

Encara que SPSS i STATA faciliten el valor *r* amb tres i quatre decimals, respectivament, el costum és proporcionar-ho amb dos decimals. Igualment és rellevant facilitar el nombre d'observacions, en el nostre exemple, n=52 (hi ha 52 províncies a Espanya).

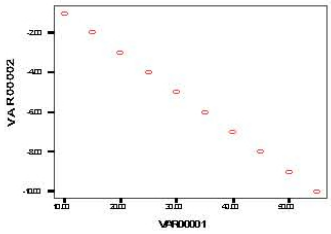
Com interpretar el valor de r?

El valor de la *r* de Pearson, com s'ha dit, oscil·la entre -1 i 1.

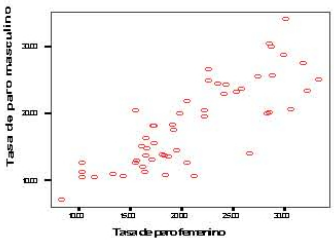
- Quan el valor l'igual a 1 implica una correlació lineal perfecta, és a dir, els punts de les variables (*x,y*) dibuixen una línia perfecta. Per exemple:



- Quan el valor és igual a -1 implica una correlació lineal perfecta, però amb sentit negatiu (és a dir, quan augmenta *x*, disminueix *y*). Gràficament seria:

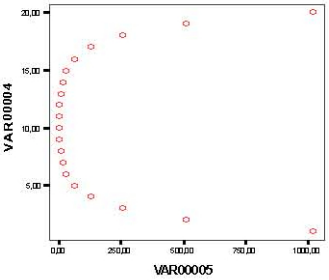


- Quan el valor és pròxim a 1 implica una correlació lineal forta (quan augmenta *x* tendeix a augmentar *y*), com en el cas d'atur masculí i femení r = 0,81:



- Quan el valor és pròxim a 0 implica absència de correlació lineal, com en el cas de taxa d'activitat femenina i atur masculí r = 0,13

- Alerta! Quan *r* és pròxim a 0, pot haver-hi correlació, encara que aquesta no siga de tipus lineal. En aquest exemple, r = 0 i la seua representació gràfica és:



Una de les formes d'interpretar el valor *r* és com la proporció de variabilitat en la variable *y* que es deu a una relació lineal amb la variable *x*. Per a fer aquest tipus de lectura s'ha de calcular *r* quadrat i multiplicar-ho per 100. En l'exemple d'atur masculí i femení, r = 0,8131, per tant r^2 = 0,6611, és a dir, un

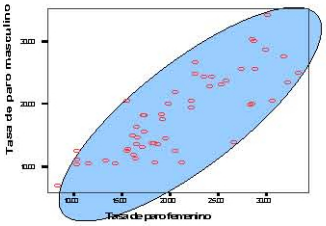
66,11% de la variabilitat de l'atur masculí s'explica per la relació lineal entre atur masculí i femení.

Alerta!, associació no significa necessàriament relació causal. En l'exemple d'atur masculí i femení no es pot concloure que l'atur femení és explicatiu de l'atur masculí. Probablement és més assenyat pensar que tant l'atur masculí com l'atur femení, en el nivell provincial, estan explicats per un tercera variable o grup de variables. Habitualment, en ciències socials les anàlisis de correlació són més útils per a generar hipòtesis que per a confirmar-les.

Per a parlar de relació causal es necessiten múltiples elements addicionals. Per exemple, es necessita un marc explicatiu que siga coherent amb l'associació que s'ha trobat (i que explique el mecanisme causal que provoca l'associació). D'altra banda, per a parlar de causalitat, les causes han de precedir els efectes, s'ha de trobar la relació en repetides circumstàncies, s'ha de controlar pels possibles factors de confusió de la relació, millor si s'ha trobat la relació en condicions experimentals, etcètera.

Supòsits:

1. El coeficient de correlació de Pearson pot ser calculat per a qualsevol parell de variables quantitatives, no obstant això, té un major significat quan les dues variables segueixen una distribució normal. Per això ha de calcular-se el coeficient de correlació si les dues variables segueixen una distribució normal. Gràficament, es pot intuir que dues variables són normals en la seua distribució si el núvol de punts té una forma el·líptica. En el nostre exemple de la taxa d'atur masculí i femení, s'observa un patró pròxim a aquesta forma el·líptica:

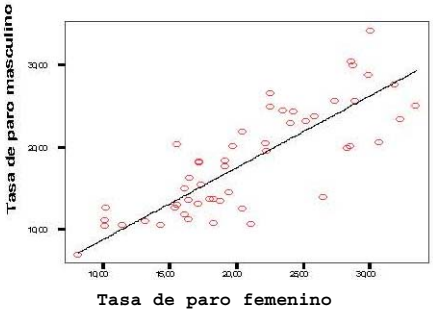


2. Les observacions han de ser independents. Aquest concepte l'expliquem en parlar de la comparació de mitjanes. En l'exemple de les taxes d'atur, cada província només produeix un parell x, y. Si haguérem barrejat en l'anàlisi dades d'atur de diferents dates per a les províncies, l'anàlisi de correlació no hauria estat adequat.

Regressió lineal

Si hi haguera correlació, llavors és pràctic procedir a estimar la recta subjacent que il·lustra la relació entre les dues variables estudiades.

En l'exemple sobre la taxa d'atur masculí i femení, és fàcil imaginar una línia travessant el núvol de punts.



Calcular aquesta recta seria equivalent a poder estimar el grau de variació de la variable *y* per cada increment en una unitat de la variable *x,* és a dir, permetrà fer projeccions sobre la variable *y.*

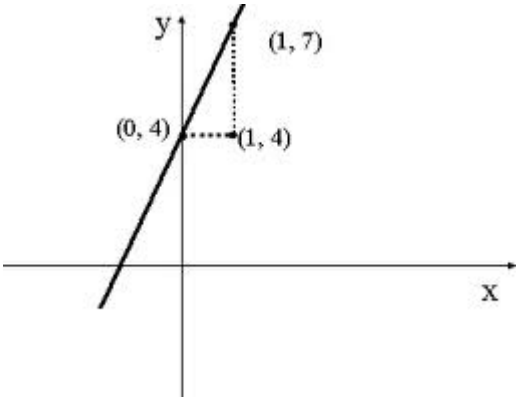
L'expressió matemàtica per a una recta és:

$$y = a + bx$$

on *a* és conegut com el valor constant o "ordenada en l'origen". És el punt en el qual es creua la recta amb l'eix de la *y.* En aquestes anotacions utilitzarem el terme "constant" (*constant*) per ser el que sovint s'utilitza en els resultats dels programes estadístics.

b seria el pendent de la recta, ja que a mesura que s'incrementa una unitat de *x* s'incrementa en *b* l'altura de la variable *y.*

Si eres aficionada o aficionat al ciclisme, hauràs sentit parlar múltiples vegades de rampes del 10%, 5% o fins i tot de 20%. Percentatges que fan referència al pendent que té una costera. Què significa? Per exemple, una rampa del 10% significa que si recorres 1.000 metres, en acabar estaràs 100 metres més alt en el port (guanyes un 10% de la quantitat recorreguda en altitud). Si estàs a 1.000 metres en un port com la Madeleine i puges 1 km amb una rampa del 10%, en recórrer aquesta distància estaràs a 1.100 metres d'altitud. Si el quilòmetre següent té una rampa del 20% quan l'acabes estaràs a 1.300 metres d'altitud. En definitiva, la *b* del pendent d'una recta és el mateix, però sense suar. D'altra banda, *a* seria igual a l'altitud a la qual estaves abans de començar el port. Per cert, si descendírem seria el mateix però amb valors negatius per a *b* i sense pedalejar.



En el gràfic anterior[7] podem deduir la recta a partir dels dos punts que ens faciliten (0,4) i (1,7): la primera xifra reflecteix el valor de la variable *x* i la segona el de la variable *y*.

Com veiem la recta creua l'eix *y* en el punt 4.

Com $y = a + bx$ ja podríem saber que $a = 4$, ja que es compleix que quan $x = 0$, $y = 4$, és a dir, el punt (0,4) és un punt de la recta.

També podem saber quant és *b*. De fet veiem que quan x augmenta una unitat, y augmenta 3 unitats, és a dir, la recta passa del punt (0,4) al punt (1,7). En definitiva, *b* seria igual a 3.

Observem que també es compleix la regla

$$y = a + bx$$

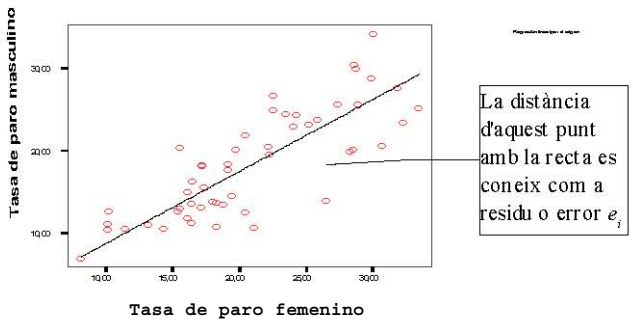
$$7 = 4 + (3*1)$$

La recta la denominem *recta de regressió*, *b* és el pendent i *a* la constant. Com major siga el valor *b*, major serà el pendent. Com major siga *a* l'encreuament amb l'eix *y* es produirà a més altura.

Intentem ara calcular la recta a partir dels punts que hem obtingut en representar dues variables.

L'objectiu serà dibuixar la línia que passe més prop de tots els punts (totes les observacions).

Es denomina *residu* la distància vertical entre cada punt *y* de la distribució real i el punt $\hat{y}$ estimat per la recta de regressió. La recta a dibuixar ha de ser tal que la suma dels residus siga mínima. És a dir, que no es puga dibuixar cap altra recta que produïsca una suma de residus menor (vegeu el dibuix següent). Per aquest motiu el procediment pel qual aconsegueix això se'l coneix com a *ajust per mínims quadrats*.



Si restem al valor de *y* observat, el valor $\hat{y}$ que prediu la recta ens trobarem que les distàncies poden quedar expressades en termes positius i negatius, per la

[7] El gràfic ha sigut obtingut en la pàgina web de la Universitat Nacional de Luján. En línia (visitat 11/08/2010): http://www.matepreuni.unlu.edu.ar/estudiar.htm

qual cosa unes distàncies anul·larien les altres. Per aquest motiu es prefereix calcular la distància dels residus al quadrat e_i^2.

La suma dels residus al quadrat es pot denotar com a $\sum e_i^2$

Com s'ha dit, l'objectiu és que la recta que triem siga la que minimitze aquest sumatori $\sum e_i^2$

D'acord amb això, la forma de calcular una *b* que complisca aquesta condició a partir de les dades de dues variables *x* i *y,* és fer la divisió següent:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

El numerador se'l coneix com a covariància o covariació de *y* en *x* i el numerador és la ja coneguda variància de *x.* Resumint:

$$b = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

Alternativament, es pot calcular *b* mitjançant la fórmula següent, que evita haver de calcular les mitjanes de *x* i *y*:

$$b = \frac{n \cdot \sum (x_i y_i) - \left(\sum x_i\right)\left(\sum y_i\right)}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

Una vegada calculada *b* el càlcul de *a* resulta relativament senzill.

$$a = \bar{Y} - b\bar{X} = \frac{\sum (y_i) - \left(b \sum x_i\right)}{n}$$

En el capítol 9.2.2 "La ecuación de regresión *y* el ajuste por mínimos cuadrados" de García Ferrando (1994), pots veure una explicació d'aquesta fórmula i un exemple del seu càlcul en les pàgines 270-271. Es recomana que mires l'exemple i tractes de calcular *b* mitjançant les dues fórmules proposades.

Les dades són

Anys d'escolaritat (x_i)	Nivell d'ingressos (y_i)	$(x_i)^2$	$(y_i)^2$	$(x_i y_i)$
1 2 3 4 5	2 5 4 6 8			
$\left(\sum x_i\right)$	$\left(\sum y_i\right)$	$\left(\sum x_i\right)^2$	$\left(\sum y_i\right)^2$	$\sum (x_i y_i)$

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

Càlcul

Com veuràs, el procés és una mica complicat i inimaginable d'aplicar quan es té un nombre de casos ampli. Com hem dit, les calculadores i el programa estadístic ajuden molt.

Per al càlcul de *a* i *b* amb **la calculadora**:

Una vegada guardades les dades com explicàvem en parlar de la correlació, pots obtenir *a* i *b* amb les tecles corresponents, normalment ("shift" + A, i "shift" + B). Recorda que la calculadora també pot proporcionar els càlculs de *r* i $\sum(x_i y_i)\ \left(\sum x_i\right)\ \left(\sum y_i\right)\ \left(\sum x_i\right)^2$

Amb SPSS:

Analizar > Regresión > lineal...

```
REGRESSION
  /DEPENDENT VAR00002
  /METHOD=ENTER VAR00001  .
```

Resultat:

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta	B	Error típ.
1	(Constante)	1,100	1,066		1,032	,378
	VAR00001	1,300	,321	,919	4,044	,027

a Variable dependiente: VAR00002

Amb STATA

```
. regress var2 var1

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   16.35
       Model |        16.9     1        16.9           Prob > F      =  0.0272
    Residual |         3.1     3  1.03333333           R-squared     =  0.8450
-------------+------------------------------           Adj R-squared =  0.7933
       Total |          20     4           5           Root MSE      =  1.0165

------------------------------------------------------------------------------
        var2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var1 |         1.3    .321455     4.04   0.027     .2769866    2.323013
       _cons |         1.1   1.066146     1.03   0.378    -2.292951    4.492951
------------------------------------------------------------------------------
```

Per descomptat, també es pot crear un full de càlcul (com OpenOffice Calc o Excel).

Per a treballar amb Excel o OpenOffice Calc pots veure el tutorial següent:

- Clemson University, Physics Laboratory. Excel Tutorial 11, "Linear Regression and Excel". En línia (visitat 14/08/2010):[8] http://phoenix.phys.clemson.edu/tutorials/excel/regression.html

Generalment treballarem la regressió lineal amb les dades d'una mostra d'una població, per la qual cosa haurem de sustentar les nostres afirmacions en els principis de l'estadística inferencial (distribució mostral, proves estadístiques, etc.).

Notació:

	Població	Mostra
Pendent	β	b
Constant	γ	a

Per tant la recta en la població rebria la notació:
$y = \gamma + \beta x$

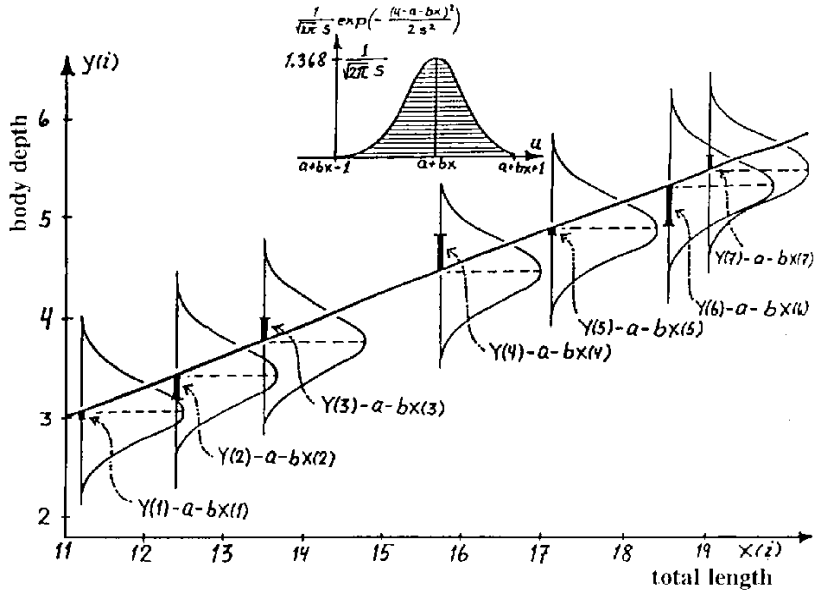
Supòsits de la regressió lineal:

Podem utilitzar *a i b*, com a estimadors γ i β, però per a això han de complir-se tres supòsits:

1. La variable *y* ha de tenir una distribució normal per a cada valor de la variable *x*:

[8] Si bé Excel pot ajudar a fer aquest tipus de càlculs, has de tenir en compte les seues limitacions, sobretot quan alguna variable tinga dades perdudes i a l'hora de representar gràfics. Pots llegir per exemple, Jonathan D. Cryer. *Problems with using Microsoft Excel for statistics*. Department of Statistics and Actuarial Science. University of Iowa. En línia (visitat 15/08/2010): http://www.cs.uiowa.edu/jcryer/JSMTalk2001.pdf

L'expressió gràfica d'aquesta idea seria:[9]



FAO. Corporate Document Repository. "Introduction to tropical fish stock assessment". En línia (visitat 11/08/2010): http://www.fao.org/docrep/w5449e/w5449egf.gif

Quan el nombre de casos és gran, pel teorema central del límit, es garanteix el compliment d'aquest supòsit.[10]

2. Les variàncies de les distribucions de la variable *y* són les mateixes per a cada valor de *x*.

El gràfic anterior també serveix per a il·lustrar aquesta condició. La variable *y,* a més de dibuixar una corba normal per a cada valor de *x*, dibuixa corbes normals que tenen una amplitud molt similar. Aquest criteri es coneix com a *criteri d'homocedastacitat*. El contrari es coneix com a *heterocedasticitat*[11].En veurem alguns exemples en classe.

3. La relació entre *x* i *y* ha de ser lineal. Recorda que una relació lineal es pot observar en el diagrama de dispersió.

Es demana que tingues en compte que qualsevol model estadístic, es pot aplicar únicament sota certes condicions de les dades. En qualsevol cas trobaràs una explicació en les pàgines 543-555, de SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

[9] En la pàgina 553 de Sánchez Carrión (2008), en la figura 9.6 trobaràs dos exemples. Un de compliment d'aquesta condició 9.6 (a) i un altre d'incompliment 9.6 (b).

[10] Per a detectar-ho se sol construir un gràfic a partir dels residus (e_i), que han de seguir una distribució normal quan es compleix aquest supòsit. El gràfic a construir és un histograma de residus. De igual manera, per a detectar-la se sol construir un gràfic amb els residus (e_i) juntament amb els valors de la variable dependent estimats per la regressió.

[11] Per a detectar-la pots construir el gràfic de residus (e_i) amb els valors de la variable independent. Quan els residus es distribueixen de forma uniforme al llarg de *x* es compleix el criteri d'homocedasticitat

Interval de confiança del pendent

L'element més rellevant a estimar en la recta de regressió és el pendent β

L'interval de confiança de β s'estima a partir de l'estadístic *b*. Per a això s'utilitza la distribució *t* de Student (amb n-2 graus de llibertat) i és necessari estimar l'error típic de *b* (standard error)[12]. En definitiva:

$IC(\beta) = b \pm t_{n.c.}\, SE(b)$

L'error típic de *b*, *SE(b)* és

$$SE(\beta) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2 / (n-2)}{\sum (x_i - \bar{x})^2}}$$

On $\hat{y}_i$ és el valor estimat per a la variable dependent *y* en l'observació *i*.

Tampoc serà necessari calcular-ho a mà, ja que el programa estadístic ens ho ofereix fàcilment.

La hipòtesi nul·la, és la hipòtesi en la qual es planteja que no hi ha una relació lineal entre les variables *x* i *y*, és a dir, que el creixement en la variable *x* no implica cap increment (o descens) lineal en la variable *y*.

$H_0 : \beta = 0$

La hipòtesi alternativa assenyala que β té un valor diferent de zero, la qual cosa ens parlaria que la recta de regressió té pendent.

$H_1: \beta \neq 0$

La prova t per a b, es calcularia:

$t = \frac{b}{SE(b)}$ que segueix una distribució~t_{n-2} si H_0 és certa

Amb n-2 graus de llibertat

Càlcul

Tornem ara a l'exemple amb la taxa d'activitat femenina i l'atur masculí i obtenim el resultat amb SPSS (*Analizar* > *Regresión* > *Lineal...* i indiquem que calcule el IC per a *b* marcant una *x* en "*intervalos de confianza*" en "*estadísticos*".)

[12] En algun programa, en lloc d'error típic o *standard error*, podràs veure *ES Coeff, StDev, ES, Std Dev...*

Resum del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,134(a)	,018	-,002	6,45011

a Variables predictoras: (Constante), Tasa de actividad femenina

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	38,016	1	38,016	,914	,344(a)
	Residual	2080,196	50	41,604		
	Total	2118,212	51			

a Variables predictoras: (Constante), Tasa de actividad femenina
b Variable dependiente: Tasa de paro masculino

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	10,182	8,605		1,183	,242	-7,102	27,467
	Tasa de actividad femenina	,166	,174	,134	,956	,344	-,183	,516

a. Variable dependiente: Tasa de paro masculino

Anàlogament amb STATA el resultat és:

```
regress PAROMASC ACTFEM

      Source |       SS       df       MS              Number of obs =      52
-------------+------------------------------           F(  1,    50) =    0.91
       Model |  38.0163767     1  38.0163767           Prob > F      =  0.3437
    Residual |  2080.19585    50  41.6039169           R-squared     =  0.0179
-------------+------------------------------           Adj R-squared = -0.0017
       Total |  2118.21222    51   41.533573           Root MSE      =  6.4501

------------------------------------------------------------------------------
    PAROMASC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ACTFEM |   .1662364   .1739034     0.96   0.344    -.1830587    .5155316
       _cons |   10.18246   8.605334     1.18   0.242    -7.101867    27.46678
```

El valor *r* és el que ja havíem obtingut al principi del tema. Com es va dir, és pròxim a 0. També proporciona el valor de *r* al quadrat.

Com es pot observar, el valor *b* és igual a 0,166 i el seu interval de confiança al 95% és IC_{95} (-0,183;0,516). Com es veu, l'interval de confiança del valor *b* inclou el valor 0, de manera que ja es pot entendre que la hipòtesi $H_0 : \beta = 0$ té una alta probabilitat de ser vàlida.

De totes maneres, per a confirmar-ho es pot calcular el valor de *t* per a *b*

$$t = \frac{b}{SE(b)} \quad t = \frac{0{,}166}{0{,}174} = 0{,}96$$

El valor *t* és igual a 0,96, amb n-2 = 50 graus de llibertat. Veiem que el valor *t* s'associa a una probabilitat de 0,344, és a dir, una probabilitat molt superior al 0,05 que solem marcar com a límit per a rebutjar la hipòtesi nul·la. Així, atès que la probabilitat que la hipòtesi nul·la siga certa és alta, no podem rebutjar-la.

Observaràs que el programa produeix un resultat d'una taula ANOVA (l'*ANà*lisi de -*Of- Va*riància), que no hem comentat. Encara no entendràs molt bé aquest resultat, però quan estudies el tema següent, sobre l'*Anà*lisi de *Va*riància, ho comprendràs molt bé. Per cert, és el mateix tipus de prova que s'aplica en els contrastos de mitjanes de dues poblacions independents per a conèixer si les variàncies són iguals (l'anomenada F de Snedecor).

Si repetim el mateix exercici amb la taxa d'atur masculí i femení, s'obté el resultat següent:

Amb SPSS:

Resum del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,813(a)	,661	,654	3,78892

a Variables predictoras: (Constante), Tasa de paro femenino

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1400,415	1	1400,415	97,549	,000(a)
	Residual	717,797	50	14,356		
	Total	2118,212	51			

a Variables predictoras: (Constante), Tasa de paro femenino
b Variable dependiente: Tasa de paro masculino

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	1,247	1,811		,688	,494	-2,391	4,884
	Tasa de paro femenino	,819	,083	,813	9,877	,000	,652	,986

a. Variable dependiente: Tasa de paro masculino

Amb STATA:

```
. regress PAROMASC PAROFEM

      Source |       SS       df       MS              Number of obs =      52
-------------+------------------------------           F(  1,    50) =   97.55
       Model |  1400.41487     1  1400.41487           Prob > F      =  0.0000
    Residual |   717.79735    50   14.355947           R-squared     =  0.6611
-------------+------------------------------           Adj R-squared =  0.6544
       Total |  2118.21222    51   41.533573           Root MSE      =  3.7889

------------------------------------------------------------------------------
    PAROMASC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     PAROFEM |   .8190069    .082923     9.88   0.000     .6524511    .9855627
       _cons |    1.24676   1.810974     0.69   0.494    -2.390688    4.884207
------------------------------------------------------------------------------
```

Ara *b* té un valor 0,819 amb un IC_{95} (0,652;0,986). L'IC no conté el valor 0, d'altra banda, el test *t* llança una probabilitat per al valor *t = 9,88* menor del 0,000. És a dir, la probabilitat que $H_0 : \beta = 0$ siga certa és molt baixa. Per aquest motiu, la rebutgem i optem per la hipòtesi alternativa, que informa sobre l'existència d'una associació entre les dues variables.

Utilitat de la recta de regressió i precaucions finals

En conclusió, si la *r* de Pearson és elevada, si rebutgem la hipòtesi nul·la ($H_0 : \beta = 0$) i es compleixen els supòsits per a la regressió lineal haurem aconseguit construir un model (una recta) que és més útil per a predir els valors de la variable *y* que el que hem utilitzat fins ara (la mitjana de *y* i el seu interval de confiança). És a dir, que, sabent els valors que pren la variable *x,* podem estimar amb més exactitud quins són els valors més probables de la variable *y*.

Per a això, n'hi ha prou amb aplicar la recta calculada

$$y = a + bx$$

En l'exemple de l'atur masculí i femení, seria:

$$\textit{taxa d'atur masculí} = 1.25 + 0{,}82(\textit{taxa d'atur femení})$$

Com sabem, el valor 0,82 oscil·la en un IC_{95} (0,652;0,986) i el mateix ocorre amb la constant, per la qual cosa es podria fer una predicció amb un valor màxim i mínim.

En qualsevol cas, el càlcul és tan senzill com aplicar l'equació de la recta.

Això amb la calculadora, per a l'estimació puntual (no la d'interval), es fa prement el valor de *x* que es vol estimar i les tecles "shift/kout" i "ŷ" (en la teua calculadora el procediment pot ser diferent). Amb els programes estadístics també es pot demanar que construïsca una variable o diverses variables amb els valors que estima el model, en el mateix quadre de diàleg de regressió lineal que ja hem utilitzat.

No obstant això, encara queda una precaució addicional. Qualsevol predicció ha de fer-se dins del rang dels valors de *x* que hàgem estudiat. És a dir, si en el cas de l'atur femení, la dada d'atur més baix és de Guipúscoa amb un 8,23% i el més elevat el de Còrdova amb un 33,43%, només es podran fer previsions

per als valors de *x* compresos entre 8,23 i 33,43. Mai podrem estimar el que ocorreria si l'atur femení fóra igual a 0 o si arribara al 80%, per exemple, perquè aquestes situacions queden fora del grup de dades que ha estat estudiat i, per tant, no estaríem procedint de forma empírica.

De la mateixa manera, encara que hem denominat l'atur femení "variable *x*" (variable independent) i l'atur masculí "variable *y*" (variable dependent), en realitat, com s'ha dit anteriorment, no tenim cap element que ens permeta parlar de causalitat entre les dues variables (no n'hi hauria, per tant, dependent ni independent). No obstant això, sí que sabem que es tracta de variables correlacionades i que, coneixent el nivell d'atur femení, es pot obtenir una millor estimació de l'atur masculí que sense conèixer-lo.

L'explicació del manual

En aquesta ocasió es recomana començar amb el manual de:

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

En el seu capítol 9 "Medidas de asociación para variables de intervalo: regresión *y* correlación" (pàg. 261-286) trobaràs una explicació molt clara de la regressió i correlació.

- En 9.1. "Planteamiento general" explica que fins ara només hem comptat amb la mitjana d'una variable i el seu interval de confiança per a fer prediccions. La regressió lineal servirà per a millorar la predicció en el cas que comptem amb una variable *x* que estiga correlacionada linealment amb aquella.
- En 9.2 "Ecuaciones de regresión lineal" es fa una reflexió sobre l'escassetat d'explicacions en les teories de les ciències socials. La utilització dels models de regressió lineal pot servir per a augmentar les explicacions disponibles.
- En 9.2.1 "Relación entre dos variables estadísticas: ecuación de una recta" explica l'equació d'una recta a partir de dos exemples. Tin en compte que l'"ordenada en l'origen" l'hem denominat "constant" en aquestes anotacions per ser el terme que s'utilitza sovint en el programa estadístic.
- En 9.2.2. "La ecuación de regresión y el ajuste por mínimos cuadrados" recorda els supòsits d'un model de regressió lineal (en el paràgraf tercer). Quins són els tres supòsits? A continuació introdueix el concepte d'"ajust per mínims quadrats" explicant que consisteixen els residus, denominats ací (y-y'), on y' és el valor estimat per la recta. La idea és aconseguir una recta en la qual el sumatori dels residus al quadrat siga mínim. Per a obtenir aquesta recta cal estimar *b* i *a* en la recta, i per a això hi ha diferents fórmules. Ho il·lustra amb un exemple.
- En 9.3 "Correlación. Coeficiente R de Correlación de Pearson" s'expliquen les diferents interpretacions del valor de r. Recordant que quan hi ha una relació no lineal el valor de *r* pot ser igual a 0. Per a

estimar el valor de *r* ho fa a partir del valor de *r* quadrat que s'obté mitjançant la divisió entre la variació explicada $\sum(y'-\bar{y})^2$ i variació total $\sum(y-\bar{y}')^2$. En aquestes anotacions no s'ha seguit aquesta línia explicativa, encara que és igualment vàlida, perquè es reprendrà en el tema següent sobre l'anàlisi de variància (ANOVA). Addicionalment s'explica la fórmula del coeficient *r* que ha presentat en aquestes anotacions i s'explica una fórmula alternativa. Es recorda que en l'assignatura per al càlcul de *r* utilitzarem l'ajuda de la calculadora i del programa estadístic. García Ferrando recorda igualment la importància de la representació gràfica quan es treballa la correlació entre dues variables a partir de dos exemples gràfics molt il·lustratius.

- En 9.3.1 "Interpretación del coeficiente de correlación", relaciona la r^2 de Pearson (també anomenat *coeficient de determinació*) amb les mesures de RPE (reducció proporcional de l'error) que es tractava en el tema anterior, en relacionar la variació explicada i la variació total.
- En 9.3.2. "Correlación y regresión con valores típicos, z" introdueix un aspecte no tractat en les anotacions però que resulta d'interès per a entendre el concepte de correlació i la seua simplicitat quan es treballa amb variables amb valors tipificats.
- En 9.4. "La matriz de correlaciones" s'explica un instrument, "la matriu de correlacions", utilitzat sovint en la investigació social. En ella es representen els valors del coeficient *r* entre múltiples variables. Com pots veure té un gran potencial per a la generació d'hipòtesi d'investigació social.
- En 9.5 "Consideraciones finales..." recorda la importància del nivell de mesurament de les variables a l'hora de triar les mesures d'associació. En el cas de la correlació lineal es necessiten variables d'interval. D'altra banda, s'explica la diferència entre associació i causalitat que també s'ha introduït en les anotacions.

Una explicació alternativa

L'explicació de Sánchez Carrión resulta més completa, encara que potser menys fàcil de seguir, que l'anterior. Pots utilitzar-la per a reforçar els conceptes explicats (pàg. 485 a 502).

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

Al manual de Sánchez Carrión trobaràs aspectes no tractats en García Ferrando (1994), algun ha sigut tractat en les anotacions, com:

- L'explicació de la prova d'hipòtesi per al pendent (pàg. 554-559)

Altres aspectes no han sigut tractats en el marc d'aquesta assignatura, però et poden servir **per a saber més** sobre l'anàlisi de regressió:

- Les fonts d'error en la regressió (pàg. 492-494).
- L'anàlisi dels residus (pàg. 502-506).

- La transformació de variables en el cas de relacions no lineals (pàg. 506-509) i (pàg. 560-587).
- L'ampliació del model de regressió a dues o més variables independents (pàg. 509-515).
- La comprovació dels supòsits del model de regressió lineal (pàg. 545-554).
- La prova de la F de Snedecor (ANOVA) (pàg. 559-560)

Quadern d'exercicis

Exercicis

1) Comença fent els exercicis 1, 2 i 3 del tema 9 del manual de García Ferrando (pàg. 284-285). Resol-los amb ajuda de la calculadora primer i després amb ajuda d'un programa estadístic.
2) Fes l'exercici 1 del capítol 8 del manual de Sánchez Carrión (pàg. 521).
3) Exercici 4, capítol 8 del manual de Sánchez Carrión (pàg. 523-424).
4) Exercici 5, capítol 8 del manual de Sánchez Carrión (pàg. 524-525).

Resol-los amb ajuda de la calculadora primer i després amb ajuda d'un programa estadístic. Assenyala si escau calcular la recta de regressió. Assenyala els supòsits que caldria aplicar al model. Copia els resultats del programa estadístic o imprimeix-los i apega'ls al teu quadern d'exercicis.

Repàs

Al final d'aquesta unitat has de saber:

- Elaborar i interpretar un diagrama de dispersió.
- Interpretar el coeficient de correlació i el de determinació.
- Calcular els paràmetres *a* i *b* d'una recta de regressió.
- Fer una prova *t* sobre el pendent d'una recta.
- Enumerar els supòsits per a poder calcular una recta de regressió.
- Distingir entre el concepte d'associació i causalitat.

Exercicis de repàs

Del manual de García de Cortázar *et al*. (1996) *Estadística aplicada a las ciencias sociales*. Madrid: Cuadernos de la UNED, 1a ed., els problemes 9.1, 9.2, 9.3, 9.4, 9.5, 9.6, 9.8, 9.9, 9.10, 9.11 i 9.12 .

8 Anàlisi de variància

Planifica l'estudi:

- Aquesta unitat es treballa en tres sessions de classe (3 hores de teoria + 3 hores de seminari = 6 hores).
- 16 hores d'estudi fora de l'aula.
- Setmanes 13, 14 i 15.

Seqüència recomanada d'estudi:

Descripció de l'activitat	Temps	Tipus
1) Estudi de les anotacions	2 hores	No presencial
2) Classe de teoria	1 hora	Presencial
3) Seminari d'exercicis en l'aula	1 hora	Presencial
4) Estudi de les anotacions	2 hores	No presencial
5) Fer els exercicis	2 hores	No presencial
6) Classe teoria	1 hora	Presencial
7) Seminari d'exercicis en l'aula	1 hora	Presencial
8) Estudi de les anotacions	2 hores	No presencial
9) Acabar els exercicis	3 hores	No presencial
10) Classe de teoria	1 hora	Presencial
11) Seminari d'exercicis en l'aula	1 hora	Presencial
12) Acabar els exercicis del quadern d'exercicis	2 hores	No presencial
13) Repàs	3 hores	No presencial

Materials per a l'estudi:

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

Introducció

L'anàlisi de variància és una extensió de l'anàlisi de la diferència de dues mitjanes. Permet comparar diverses mitjanes. El principal avantatge és que es pot conèixer si les mitjanes de diferents subgrups són iguals o diferents. Per exemple, a més de conèixer si el salari mitjà d'homes i dones difereix, podrem saber si difereix segons el nivell educatiu. Aquests subgrups es poden establir a partir de variables nominals o ordinals, per la qual cosa en la pràctica permet posar en relació una variable d'interval amb una variable nominal o ordinal (com a mínim). Aquesta petita variació permet incrementar el tipus de preguntes que podem fer a les dades de forma considerable. És a dir, amb les mateixes dades obtindrem molta més informació sobre les pautes socials que revelen aquestes dades.

Objectius:

- Aplicar i interpretar l'anàlisi de variància amb un només factor.
- utilitzar l'anàlisi de variància com a prova de decisió en l'anàlisi de correlació i regressió.
- Comprovar la hipòtesi d'igualtat de variàncies en dues mostres independents.

Primera explicació

L'anàlisi de variàncies permet posar a prova la hipòtesi que les mitjanes de diverses poblacions (o grups) són iguals, enfront de la hipòtesi alternativa que són diferents (o, millor dit, que alguna o algunes de les mitjanes són diferents).

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_r$

$H_1 : \mu_i \neq \mu_j$ per a algun $i \neq j$

Podem veure-ho a través d'un exemple. Volem estudiar si el nombre de persones amb les quals s'ha mantingut relacions sexuals en els últims dotze mesos (pregunta 32 de l'estudi 2.780 del CIS) varia en funció de la creença religiosa (pregunta 50).

P. 50. Com es defineix vostè en matèria religiosa: catòlic/a, creient d'una altra religió, no creient o ateu/a?

Catòlic/a ....
Creient d'una altra religió 2
No creient... 3
Ateu/a... 4
N. C. ... 9

Primer resoldrem el problema amb dades imaginades i després amb les de l'estudi 2.780.

Per a simplificar compararem només tres grups (catòlics, no creients, ateus), d'homes.

Suposem que les seues dades imaginades són (exagerarem una mica):

Nombre de parelles en l'últim any

Catòlics	No creients	Ateus
1	7	10
1	8	11
3	7	9
4	10	12
6	8	8
2	9	9
1	7	10
2	9	11
3	9	9
3	10	10
4	6	9
5	9	8
6	11	9
1	10	10
2	9	10

El primer que podríem fer és calcular les mitjanes de cada grup:

$\bar{x}_c = 2{,}93 \qquad \bar{x}_{nc} = 8{,}6 \qquad \bar{x}_a = 9{,}67$

La primera impressió és que la mitjana de no creients i ateus està molt per damunt de la mitjana dels catòlics.

El nombre de casos és 15, per tant el nombre de casos total és 45.

També es pot calcular la desviació típica (s) de cada grup:

$s_{\text{catòlics}} = 1{,}75 \qquad s_{\text{no creients}} = 1{,}40 \qquad s_{\text{ateus}} = 1{,}11$

I la variància (s^2)

$s^2_{\text{catòlics}} = 3{,}07 \qquad s^2_{\text{no creients}} = 1{,}97 \qquad s^2_{\text{ateus}} = 1{,}23$

Es pot observar que en principi la dispersió de les dades és major en el grup dels catòlics.

També podem calcular una mitjana per al conjunt de les dades de la taula (que hem dit que són 45) i la seua desviació típica:

$\bar{x}_g$ = Mitjana global = 7,07 i s_{global} = 3,31

Primer introduirem el concepte de variació.

El concepte de variació està relacionat amb el concepte de variància.

Sabem que la variància és:

$$\frac{\sum(x_i - \bar{x})^2}{n}$$

La variació seria el numerador de la fórmula anterior:

$$\sum(x_i - \bar{x})^2$$

Ara podem calcular la variació total per a les nostres dades, és a dir, calcular el sumatori de la distància de cada observació pel que fa a la mitjana global al quadrat:

Nombre de parelles en l'últim any

	Catòlics			No creients			Ateus		
	x_{ij}	$x_{ij} - \bar{x}_g$	$(x_{ij} - \bar{x}_g)^2$	x_{ij}	$x_{ij} - \bar{x}_g$	$(x_{ij} - \bar{x}_g)^2$	x_{ij}	$x_{ij} - \bar{x}_g$	$(x_{ij} - \bar{x}_g)^2$
	1	-6,07	36,80	7	-0,07	0,00	10	2,93	8,60
	1	-6,07	36,80	8	0,93	0,87	11	3,93	15,47
	3	-4,07	16,54	7	-0,07	0,00	9	1,93	3,74
	4	-3,07	9,40	10	2,93	8,60	12	4,93	24,34
	6	-1,07	1,14	8	0,93	0,87	8	0,93	0,87
	2	-5,07	25,67	9	1,93	3,74	9	1,93	3,74
	1	-6,07	36,80	7	-0,07	0,00	10	2,93	8,60
	2	-5,07	25,67	9	1,93	3,74	11	3,93	15,47
	3	-4,07	16,54	9	1,93	3,74	9	1,93	3,74
	3	-4,07	16,54	10	2,93	8,60	10	2,93	8,60
	4	-3,07	9,40	6	-1,07	1,14	9	1,93	3,74
	5	-2,07	4,27	9	1,93	3,74	8	0,93	0,87
	6	-1,07	1,14	11	3,93	15,47	9	1,93	3,74
	1	-6,07	36,80	10	2,93	8,60	10	2,93	8,60
	2	-5,07	25,67	9	1,93	3,74	10	2,93	8,60
Suma	44	-62,00	299,20	129	23,00	62,87	145	39,00	118,73

$\sum(x_{ij} - \bar{x}_g)^2$ = 36,80 + 36,80 + 16,54 + ... + 8,6 + 8,6 = 229,20 + 62,87 + 118,73 = 480,80

Hi ha una forma alternativa de calcular la variació total que no requereix calcular la mitjana global, calcular la distància de cada observació al quadrat i fer un sumatori d'aquestes distàncies.

Per a això, hem de fer un sumatori de tots els valors de la distribució $\sum x_{ij}$, és a dir, 1 +1 +3 + ... + 10 + 10 = 44 + 129 +145 = 318

Nombre de parelles en l'últim any

	Catòlics (C)	No creients (NC)	Ateus (A)
	1	7	10
	1	8	11
	3	7	9
	4	10	12
	6	8	8
	2	9	9
	1	7	10
	2	9	11
	3	9	9
	3	10	10
	4	6	9
	5	9	8
	6	11	9
	1	10	10
	2	9	10
Suma	44	129	145

Aquesta quantitat la podem elevar al quadrat: $\left(\sum x_{ij}\right)^2 = 318^2 = 101.124$

I després dividir-la pel nombre de casos $\frac{\left(\sum x_{ij}\right)^2}{n} = 101.124/45 = 2247,2$

També calculem el sumatori de tots els valors de la taula al quadrat:

Nombre de parelles en l'últim any

	Catòlics (C)	C^2	No creients (NC)	NC^2	Ateus (A)	A^2
	1	1	7	49	10	100
	1	1	8	64	11	121
	3	9	7	49	9	81
	4	16	10	100	12	144
	6	36	8	64	8	64
	2	4	9	81	9	81
	1	1	7	49	10	100
	2	4	9	81	11	121
	3	9	9	81	9	81
	3	9	10	100	10	100
	4	16	6	36	9	81
	5	25	9	81	8	64

	6	36	11	121	9	81
	1	1	10	100	10	100
	2	4	9	81	10	100
Suma	44	172	129	1137	145	1419

$$\sum x_{ij}^2 = (1)^2 + (1)^2 + (3)^2 + ... + (10)^2 + (10)^2 = 172 + 1137 + 1419 = 2728$$

Amb aquests dos elements es pot calcular el que es coneix com a variació total, que és el resultat de restar les dues quantitats:

$$\sum x_{ij}^2 - \frac{\left(\sum x_{ij}\right)^2}{n} = 2.728 - 2.247{,}2 = 480{,}8$$

Com s'observa el resultat és igual al que havíem obtingut amb:

$$\sum \left(x_{ij} - \bar{x}_g\right)^2 = 36{,}8 + 36{,}8 + 16{,}54 + ... + 8{,}6 + 8{,}6 = 229{,}2 + 62{,}87 + 118{,}73 = 480{,}8$$

Una vegada calculat $\frac{\left(\sum x_{ij}\right)^2}{n} = 2.247{,}2$

Per a calcular la variació entre grups es poden seguir dos mètodes:

El primer, és calcular el sumatori següent:

$$\sum n_j \left(\bar{x}_j - \bar{x}_g\right)^2$$

És a dir, cal calcular les distàncies de la mitjana de cada grup pel que fa a la mitjana total, elevar-les al quadrat i multiplicar-les pel nombre de casos de cada grup.

Es procedeix a calcular la diferència entre les diferents mitjanes de cada grup i la mitjana global:

La mitjana dels catòlics $\left(\bar{x}_c\right)$ menys la mitjana global $\left(\bar{x}_g\right)$ és 2,93 -7,07 = -4,13

La dels no creients 8,6 -7,07 = 1,53

La dels ateus 9,67 -7,07 = 2,6

Aquestes diferències entre la mitjana de cada grup (dels *j* grups) i la mitjana global es pot elevar al quadrat.

La dels catòlics $-4{,}13^2 = 17{,}0569$

La dels no creients $1{,}53^2 = 2{,}3409$

La dels ateus $2,6^2=6,76$

I després multiplicar pel nombre de casos de cada grup, i fer el seu sumatori:

$$\sum n_j \left(\bar{x}_j - \bar{x}_g\right)^2 = (15*17,0569) + (15*\ 2,3409) + (15*6,76) = 392,367$$

El *segon procediment* és una mica més ràpid, encara que menys intuïtiu:

N'hi ha prou amb aplicar:

$$\sum_j \frac{\sum_i \left(x_{ij}\right)^2}{n_j} - \frac{\left(\sum x_{ij}\right)^2}{n} = \frac{44^2}{15} + \frac{129^2}{15} + \frac{145^2}{15} = 392,83$$

Obtinguda la variació entre grups, la variació intragrupal serà el resultat de restar la variació entre grups a la variació total:

480,80 – 392,83 = 87,97

D'aquesta manera podem construir una petita taula:

	Suma de quadrats (*SS Sum of squares*)
Entre grups (between / among)	392,83
Intragrupal (within)	87,97
Total	480,80

Podem afegir els graus de llibertat de cadascun dels tipus de variació. Per a la variància total és igual al *n* de casos – 1, per a la variació entre grups és igual al nombre de grups menys 1, és a dir (j-1), i per a la variació entre grups és igual al nombre de casos menys el nombre de grups (n-j). Ho podem afegir a la taula:

	Suma de quadrats (*SS Sum of squares*)	g.l. (*d.f.*)
Entre grups (*between / among*)	392,83	2
Intragrupal (*within*)	87,97	42
Total	480,80	44

Ara podem dividir la variació pels graus de llibertat per a obtenir les variàncies pròpiament aquestes. També ho afegim a la taula:

	Suma de quadrats (*SS Sum of squares*)	g.l. (*d.f.*)	Mitjana quadràtica (*MS mean of squares*)
Entre grups (*between / among*)	392,83	2	196,42
Intragrupal (*within*)	87,97	42	2,09
Total	480,80	44	10,93

Ara es pot sotmetre la relació entre la variància entre grups i la variància intragrupal a una prova, coneguda com prova F de Snedecor amb j-1 i n-j graus de llibertat, que serveix per a posar a prova la hipòtesi nul·la:

H_0: la variància entre i intragrupal són iguals,
davant de H_1: La variància entre i intragrupal són diferents.

Quan siguen completament iguals el valor de F seria 1, mentre que si la variància entre grups és major que la variància intragrupal F tendirà a ser un valor gran.

Que la variància entre grups siga gran, significaria que la major part de la variància està explicada per les distàncies entre la mitjana de cada grup i la mitjana total.

Per la seua banda, si la variància intragrupal és gran, una gran part de la variància estaria explicada per les distàncies entre els valors de cada grup i la mitjana de cada grup.

El resultat de la prova F de Snedecor es distribueix d'acord amb una distribució F (vegeu la taula F de García Ferrando en l'apèndix o taula de la F en l'annex II de Sánchez Carrión). El valor obtingut informa de la probabilitat associada a la hipòtesi nul·la. Normalment es pren com a referència un nivell de significació límit del 0,05 ($\alpha = 0{,}05$). La taula es construeix amb j-1 i n-j graus de llibertat.

En el nostre exemple:

F = variància entre grups / variància intragrupal = 196,42 / 2,09 = 93,9

Observem en la taula F amb 2 graus de llibertat (numerador) i 42 g.l. (denominador) i probabilitat $\alpha = 0{,}05$.

Observem que el valor crític aproximat és 3,23. El nostre valor F és igual a 93,9, per la qual cosa és notablement major que 3,23

$93{,}9 \in RR(0{,}05) = \{F > 3{,}23\}$ de manera que rebutgem la hipòtesi nul·la.

És a dir, que la probabilitat que la hipòtesi nul·la siga certa és molt menor que 0,05 i, per tant, hem de rebutjar-la. Ho afegim a la taula:

	Suma de quadrats (*SS Sum of squares*)	g.l. (*d.f.*)	Mitjana quadràtica (*MS mean of squares*)	F	Sign.
Entre grups (*between / among*)	392,83	2	196,42	93,9	< 0,05
Intragrupal (*within*)	87,97	42	2,09		
Total	480,80	44	10,93		

La conclusió serà que la variància entre grups és major que la variància intragrupal. Això es produeix quan les distàncies entre les mitjanes de cada grup i la mitjana global és gran, per tant s'ha de rebutjar la hipòtesi nul·la que les mitjanes són iguals ($H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_r$) i optar per la hipòtesi alternativa:

$H_1 : \mu_i \neq \mu_j$ per a algun $i \neq j$

El resultat amb SPSS és:

Analizar > Comparar medias > ANOVA de un factor...

```
ONEWAY
  Parejas BY Religiosid.
```

ANOVA

Número de parejas

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	392,933	2	196,467	93,910	,000
Intra-grupos	87,867	42	2,092		
Total	480,800	44			

Aquest seria el resultat que ens proporcionaria l'anàlisi amb STATA:

```
oneway Parejas Religiosidad, tabulate

Religiosid  |    Summary of Número de parejas
            |        Mean   Std. Dev.       Freq.
------------+------------------------------------
  Católicos |   2.9333333   1.7511901          15
  No creyen |         8.6   1.4040757          15
      Ateos |   9.6666667   1.1126973          15
------------+------------------------------------
      Total |   7.0666667   3.3056426          45

                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      392.933333      2   196.466667     93.91     0.0000
 Within groups      87.8666667     42   2.09206349
------------------------------------------------------------------------
    Total               480.8     44   10.9272727
```

Com s'observa els resultats són similars als calculats a mà, encara que la precisió de la probabilitat associada a la prova F és major. Ens indica que la

probabilitat no solament és inferior a 0,05, sinó que també és inferior a 0,001, és a dir, la probabilitat d'encertar amb la hipòtesi nul·la és menor a 1 entre 1.000

El nostre exemple, era un exemple amb dades fictícies, però amb l'estudi 2.780 del CIS tenim l'oportunitat de comprovar el que acabem d'estudiar per al cas d'Espanya.

Primer hem de preparar una mica les dades:

- Seleccionem només els homes, primer, i després només dones.
- Atès que la pregunta sobre religiositat també inclou la categoria "creient d'una altra religió", la inclourem en l'anàlisi:

  Catòlic/a 1

  Creient d'una altra religió 2

  No creient... 3

  Ateu/a... 4

  N. C. ... 9
- Ponderem els casos amb els pesos facilitats pel CIS
- Excloem de l'anàlisi les persones que no han contestat a la pregunta 32 (n.c 99): Amb quantes persones ha mantingut relacions sexuals durant els últims dotze mesos? (120 casos en homes i 76 casos en dones)
- Excloem els 78 casos en homes i 100 en dones que no contesten a la pregunta sobre religiositat (N.C. 9).

Obtenim el resultat següent per a homes:

```
ONEWAY
  P32 BY P50
  /STATISTICS DESCRIPTIVES.
```

Descriptivos

P32

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Católico	3283	1,27	2,455	,043	1,19	1,36	0	98
Creyente de otra religión	236	1,64	2,508	,163	1,32	1,97	0	20
No creyente	818	1,90	4,126	,144	1,62	2,19	0	98
Ateo	382	1,97	2,423	,124	1,72	2,21	0	20
Total	4719	1,46	2,830	,041	1,38	1,54	0	98

ANOVA

P32

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	381,482	3	127,161	16,034	,000
Intra-grupos	37393,492	4715	7,931		
Total	37774,974	4718			

I el següent per a dones:

Descriptivos

P32

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Católico	3736	,75	,730	,012	,73	,77	0	12
Creyente de otra religión	226	,77	,574	,038	,69	,85	0	4
No creyente	544	1,17	1,520	,065	1,04	1,30	0	21
Ateo	256	1,39	1,560	,098	1,20	1,58	0	12
Total	4762	,83	,929	,013	,81	,86	0	21

ANOVA

P32

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	168,537	3	56,179	67,859	,000
Intra-grupos	3939,074	4758	,828		
Total	4107,611	4761			

Com veiem, ara amb dades proporcionades realment per les persones entrevistades, tant en homes com en dones, la mitjana dels catòlics és menor que la d'altres grups. La prova confirma aquesta apreciació i indica que la probabilitat que les mitjanes siguen iguals és inferior a 0,001. Hem de pensar que les mitjanes són diferents, tant en homes com en dones.

Els intervals de confiança de les mitjanes són conseqüents amb aquesta afirmació. Observem que la mitjana de parelles, tant en homes com en dones, declarades per les persones catòliques se situa en un interval que no se solapa amb el de les persones no creients i atees. En el cas de les persones creients d'altres religions sembla trobar-se una pauta diferent per sexe: en homes tendeix a distanciar-se dels catòlics, mentre que en dones la pauta s'assimila a la de les catòliques.

Més difícil de valorar és si les respostes han pogut estar esbiaixades. Per exemple, se sol pensar que els homes tendeixen a exagerar el nombre de parelles sexuals, mentre que les dones tendeixen a fer el contrari. D'altra banda, entre els catòlics pot estar més mal vist comptar amb més parelles sexuals, per la qual cosa tant els homes com les dones poden tendir a reduir el nombre real de parelles en les seues respostes.

Supòsits estadístics de l'anàlisi de la variància

Per a poder aplicar l'ANOVA s'han de complir els mateixos supòsits que en la prova de la diferència de mitjanes:

- Ha d'haver-hi normalitat en la distribució de les dades (supòsit de normalitat)
- Les variàncies de les poblacions han de ser iguals (supòsit d'homocedasticitat).

- Les observacions han de ser independents.

Si no es compliren aquests supòsits no és aplicable el contrast F.

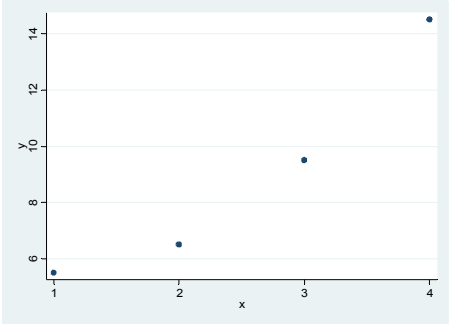
Una solució seria aplicar la prova de Kruskal-Wallis quan es treballa amb tres o més grups o la prova U de Mann-Whitney quan s'analitzen dos grups, les dues proves són de tipus no paramètric (per a saber-ne més pots consultar el manual de García Ferrando (pàg. 327-329) o Sánchez Carrión (pàg. 440-442). Aquestes proves no seran estudiades en el marc d'aquesta assignatura.

Aplicació de l'ANOVA a la regressió lineal simple

En el tema anterior avançàvem que l'anàlisi ANOVA es pot utilitzar per a provar la hipòtesi de l'existència de relació lineal en l'anàlisi de regressió lineal simple. En els resultats obtinguts amb SPSS o STATA observàvem que es també de produïa una taula ANOVA.

Ho podem il·lustrar amb un exemple. Imaginem que tenim una sèrie de punts:

x	**1**	**2**	**3**	**4**
y	**5,5**	**6,5**	**9,5**	**14,5**

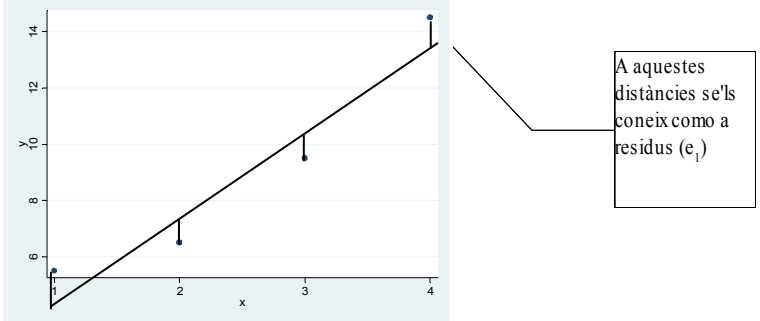


La recta és igual a y = 1,5 + 3 x

De manera que podem afegir els valors predits per la recta a la taula i dibuixar la recta:

x	1	2	3	4
y	5,5	6,5	9,5	14,5
ŷ	**4,5**	**7,5**	**10,5**	**13,5**

Els residus serien la distància entre els valors observats i els valors que ha predit la recta de regressió:

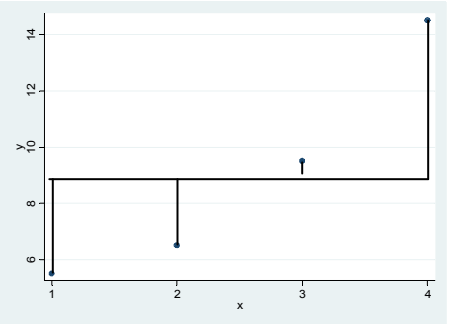


La suma dels residus al quadrat és igual a 4

$$\sum (e_i)^2 = \sum (y_i - \hat{y}_i)^2 = 1^2 + 1^2 + 1^2 + 1^2 = 4$$

x_i	1	2	3	4
y_i	5,5	6,5	9,5	14,5
$\hat{y}_i$	4,5	7,5	10,5	13,5
$e_i^2 = \lvert y_i - \hat{y}_i \rvert^2$	**1**	**1**	**1**	**1**

La variació total és igual a la distància entre la mitjana de la variable y i els valors observats. Gràficament seria:



($\bar{y}$) mitjana de la variable *y*
H_0: ß = 0

El sumatori de la diferència entre el valor y observat (y_i) i la $(\bar{y})$, al quadrat, és la variació total.

En aquest cas:

$$\sum (y_i - \bar{y})^2 = 3{,}5^2 + 2{,}5^2 + 0{,}5^2 + 5{,}5^2 = 12{,}25 + 6{,}25 + 0{,}25 + 30{,}25 = 49$$

x_i	1	2	3	4
y_i	5,5	6,5	9,5	14,5
$\hat{y}_i$	4,5	7,5	10,5	13,5
$e_i^2 = \lvert y_i - \hat{y}_i \rvert^2$	1	1	1	1
$\lvert y_i - \bar{y} \rvert$	**3,5**	**2,5**	**0,5**	**5,5**
$(y_i - \bar{y})^2$	**12,25**	**6,25**	**0.25**	**30,25**

Una vegada tenim la suma de residus al quadrat i la suma de quadrats total, podem obtenir la suma de quadrats de la regressió, amb una simple resta, ja que :

La suma de quadrats total = suma de quadrats dels residus + suma de quadrats de la regressió

D'aquesta manera:

La suma de quadrats de la regressió = suma de quadrats total – suma de quadrats dels residus.

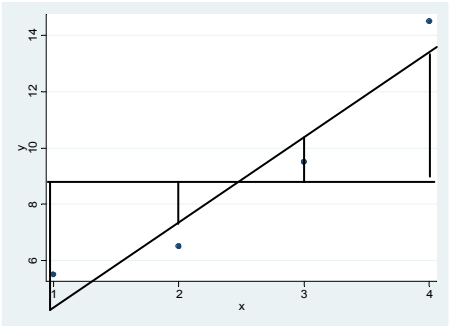
49 – 4 = 45

Gràficament estem parlant de les distàncies entre els valors que prediu la recta de regressió i la mitjana de la variable y $(\bar{y})$

x_i	1	2	3	4
y_i	5,5	6,5	9,5	14,5
$\hat{y}_i$	4,5	7,5	10,5	13,5
$e_i^2 = \lvert y_i - \hat{y}_i \rvert^2$	1	1	1	1
$\lvert y_i - \bar{y} \rvert$	3,5	2,5	0,5	5,5
$\lvert \hat{y}_i - \bar{y} \rvert$	**4,5**	**1,5**	**1,5**	**4,5**

$$\sum (\hat{y}_i - \bar{y})^2 = 4{,}5^2 + 1{,}5^2 + 1{,}5^2 + 4{,}5^2 = 45$$

És a dir, la mateixa quantitat que quan fèiem la resta 49 – 4 = 45



D'aquesta manera, hem aconseguit calcular la variació total (suma de quadrats total), la variació explicada (suma de quadrats de la regressió) i la variació residual o no explicada. És a dir, tenim els elements necessaris per a construir una taula ANOVA:

	Suma de quadrats (*SS Sum of squares*)	g.l. (*d.f.*)	Mitjana quadràtica (*MS mean of squares*)	F	Sign.
Explicada (Igual a la "entre grups (*between / among*)"	45				
Residual/ No explicada (igual a la "intragrupal" (*within*)	4				
Total	49				

Els graus de llibertat de la variació total és igual al nombre de casos n – 1 (per tant n – 1 = 3), l'explicada al nombre de variables – 1 (com només hi ha dues variables és igual a 1) i la residual és igual a n – 2, per tant 2. Amb aquestes dades podem calcular la variància, F i buscar la probabilitat associada al valor F obtingut en la taula F.

	Suma de quadrats (*SS Sum of squares*)	g.l. (*d.f.*)	Mitjana quadràtica (*MS mean of squares*)	F	Sign.
Regressió / Explicada (Igual a la "entre grups (*between / among*)"	45	1	45	22,5	$p < 0,05$
Residual/ No explicada (igual a la "intragrupal" (*within*)	4	2	2		
Total	49	3	16,3		

Atès que el valor de F per a 1 i 2 graus de llibertat amb un nivell de significació $\alpha = 0,05$ és igual a 18,5 i el nostre valor és major (22,5), significa que la probabilitat associada a la hipòtesi nul·la és menor que 0,05. Això vol dir que hem de rebutjar la hipòtesi nul·la que indica que el pendent de la recta seria igual a 0 ($H_0 : \beta = 0$).

El resultat amb SPSS és:

Resum del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,958(a)	,918	,878	1,41421

a Variables predictoras: (Constante), x

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	45,000	1	45,000	22,500	,042(a)
	Residual	4,000	2	2,000		
	Total	49,000	3			

a Variables predictoras: (Constante), x
b Variable dependiente: y

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta	B	Error típ.
1	(Constante)	1,500	1,732		,866	,478
	x	3,000	,632	,958	4,743	,042

a Variable dependiente: y

Com es veu, el nivell de significació associat a la prova F és igual 0,042, és a dir, menor a 0,05, tal com s'havia dit i, per tant, s'ha de rebutjar la hipòtesi nul·la.

Recordem ara els resultats que proporcionava SPSS quan en el tema anterior analitzàvem la relació entre taxa d'activitat femenina i atur masculí:

El resultat era:

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	38,016	1	38,016	,914	,344(a)
	Residual	2080,196	50	41,604		
	Total	2118,212	51			

a Variables predictoras: (Constante), Tasa de actividad femenina
b Variable dependiente: Tasa de paro masculino

Com que el nivell de significació associat a F és tan alt (molt superior a 0,05) hem d'optar per la hipòtesi nul·la.

Quan s'analitzava la relació entre atur femení i masculí:

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1400,415	1	1400,415	97,549	,000(a)
	Residual	717,797	50	14,356		
	Total	2118,212	51			

a Variables predictoras: (Constante), Tasa de paro femenino
b Variable dependiente: Tasa de paro masculino

La prova ANOVA torna a confirmar que el que ja havíem decidit a partir del valor de la prova *t* per a beta. El valor de significació informa sobre una probabilitat molt baixa que la hipòtesi nul·la siga certa, per la qual cosa optem per la hipòtesi alternativa.

L'explicació del manual

En aquesta ocasió es recomana començar amb el manual de:

GARCÍA FERRANDO, Manuel (2008) *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza.

En el capítol 11 "El análisis de varianza" (pàg. 311-333) trobaràs l'explicació de l'anàlisi de variància.

- En 11.1. “Introducción” s’explica quan s’utilitza l’anàlisi de variància.

  Per a repassar, respon a les preguntes següents:

  - Amb quins tipus de variables s’utilitza ANOVA?
  - Què es pretèn conèixer?
- En 11.2 “El análisis de varianza con un solo factor” s’enumeren els supòsits estadístics del ANOVA. Indica’ls:
  -
  -
  -

A continuació es planteja un exemple per a explicar el ANOVA. Es recomana que reproduïsques l’exemple i faces els càlculs a mesura que avances en la lectura.

Una vegada desenvolupat l’exemple introdueix el concepte de variació i, posteriorment, el de variància intragrupal i entre grups.

Finalment obté la taula ANOVA 11.3 i l’explica.

- En 11.3 “Otros tipos de análisis de varianza” recorda que l’anàlisi de variància es pot aplicar amb més d’una variable independent. En el manual no s’explica com es fa aquest tipus d’anàlisi (igualment queda fora del contingut d’aquesta assignatura).
- En 11.4 “Pruebas de decisión estadística para el caso de la correlación y regresión simples” explica com aplicar ANOVA a l’anàlisi de regressió lineal simple, pots repassar amb l’exemple proposat de “renda per capita i nombre d’alumnes d’ensenyament superior”.
- **Per a saber-ne més** pots consultar l’explicació de la prova de Kruskal-Wallis (pàg. 327-329)

Una explicació alternativa

SÁNCHEZ CARRIÓN, Juan Javier (2008) *Manual de análisis estadístico de los datos*. Madrid: Alianza.

En el manual trobaràs l’explicació del ANOVA a partir de les eixides del programa estadístic SPSS.

Per a saber-ne més pots consultar l’explicació de la prova de Kruskal-Wallis (pàg. 440-442)

Quadern d'exercicis

Exercicis

1) Fes l'exercici 1 del manual de García Ferrando. Fes-ho manualment (amb ajuda de la calculadora) i amb un programa estadístic.
2) Fes l'exercici 2 del manual de García Ferrando. Fes-ho manualment (amb ajuda de la calculadora) i amb un programa estadístic.

Repàs

Al final d'aquesta unitat has de saber:

- Aplicar i interpretar l'anàlisi de variància amb un sol factor.
- Utilitzar l'anàlisi de variància com a prova de decisió en l'anàlisi de correlació i regressió.
- Comprovar la hipòtesi d'igualtat de variàncies en dues mostres independents.

Exercicis de repàs

Del manual de García de Cortázar *et al*. (1996) *Estadística aplicada a las ciencias sociales*. Madrid: Cuadernos de la UNED, 1a ed., els problemes: 7.1, 7.2, 7,3 i 7.4.

Del manual de Mullor, Ruben i Fajardo, M. Dolores (2000) *Manual práctico de estadística aplicada a las ciencias sociales*. Barcelona: Ariel, 1a ed., els problemes resolts 9.1, 9.3, 9.4, 9.6 i 9.8.