Pràctica 11: INFERÈNCIA ESTADÍSTICA
Inferència en taules de contingència: test χ^2 d'independència. Test de regressió.

L'objectiu d'aquesta segona pràctica en inferència estadística és analitzar la relació-associació entre dues variables (tant qualitatives com numèriques). Començarem analitzant el plantejament d'un estudi i fent-ne un anàlisi descriptiu i exploratori a partir de dades obtingudes en una mostra. Seguidament veurem quines són les proves estadístiques més adequades per analitzar la relació entre variables.

▶ Recordeu activar les opcions:

- "Mostrar comandos en anotaciones" a la pestanya de "Visor".
- "Nombre y etiquetas" per a les variables i "Valores y etiquetas" per als valors a l'apartat de "Etiquetado de tablas pivot" de la pestanya de "Etiquetas de resultados".

▶ En aquesta primera part treballarem amb dos fitxers: *tabac-edat 2005.sav* i el fitxer *fumador-edat 2005.sav*. Contenen informació d'una enquesta realitzada pel CIS (estudi 2.627) a una mostra aleatòria de 1500 individus de 18 anys o més, entre el 16 i el 25 de novembre de 2005. En un exercici utilitzarem el fitxer *sexe-eleccions.sav* que descriurem més endavant. A l'última part treballarem amb el fitxer *paisos.sav*.

1. Plantejament d'un estudi: llei antitabac i tabaquisme

En l'enquesta del CIS es preguntava per l'opinió dels enquestats respecte la recent aprovada llei antitabac i el tabaquisme en general. Estem interessats en analitzar algunes de les respostes obtingudes l'enquesta i la seva relació-associació amb l'edat de l'enquestat. Els nostres objectius seran:

(1) Fer estimacions (puntuals i per intervals) dels percentatges de la població que es mostren a favor o en contra de la llei antitabac.
(2) Analitzar si existiexen relacions (és a dir, extrapol·lables a la població) entre el posicionament respecte la llei antitabac (opinio) i la variable edat.
(3) Analitzar si existiexen relacions (és a dir, extrapol·lables a la població) entre la variable edat i la variable fumador.

2. Independència de variables categòriques: test khi quadrat

En aquesta secció estudiarem la independència (o no) entre dues variables categòriques: l'opinió respecte la llei antitabac i el grup d'edat al qual pertany l'enquestat.

2.1. **Estudi exploratori-descriptiu.** Per treballar el primer objectiu plantejat a la secció anterior, començarem fent un estudi descriptiu de les variables del fitxer que corresponen a les relacions que volem analitzar. En un estudi bàsic d'aquest tipus, realitzarem taules de contingència i diagrames de barres.

▶ Obriu el fitxer *tabac-edat 2005.sav* i mireu les variables que conté. Quantes variables hi ha? De quin tipus són? Quin significat tenen?

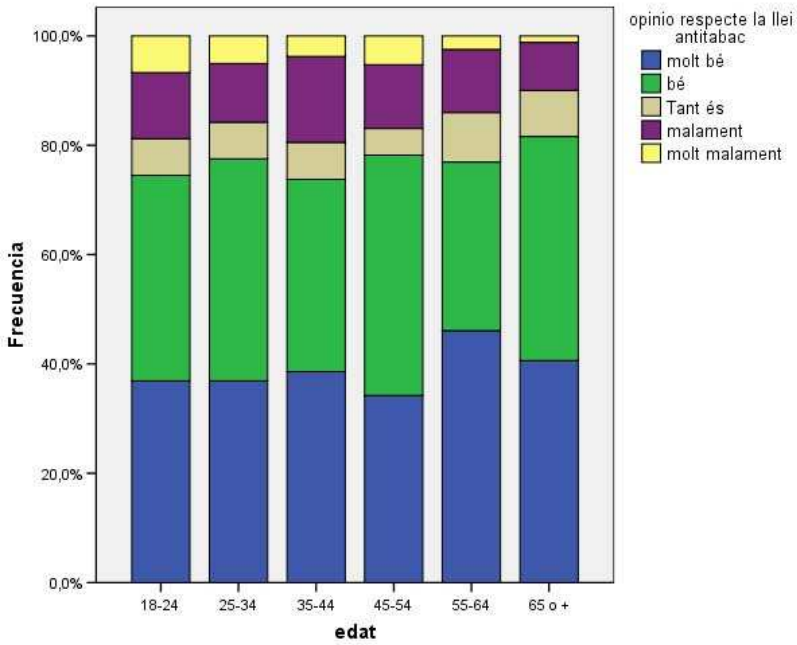
Primer realitzarem la taula de contingència. Recordeu els passos a seguir:

Analizar → Estadísticos descriptivos → Tablas de contingencia,

i farem que també ens ensenyi les freqüències esperades (Casillas) i diagrames de barres (Mostrar diagramas de barras) que apilarem. Com que ens interessa veure els percentatges per files al diagrama de barres, farem un diagrama de barres apilades rescalades al 100% mitjançant els menú de Gráficos (veure la pràctica 7).

Tabla de contingencia edat edat * opinio opinio respecte la llei antitabac

			opinio opinio respecte la llei antitabac					
			1 molt bé	2 bé	3 Tant és	4 malament	5 molt malament	Total
edat edat	1 18-24	Recuento	55	56	10	18	10	149
		Frecuencia esperada	57,9	57,2	10,5	17,5	5,8	149,0
	2 25-34	Recuento	110	121	20	32	15	298
		Frecuencia esperada	115,9	114,5	21,1	34,9	11,6	298,0
	3 35-44	Recuento	103	94	18	42	10	267
		Frecuencia esperada	103,8	102,6	18,9	31,3	10,4	267,0
	4 45-54	Recuento	91	117	13	31	14	266
		Frecuencia esperada	103,4	102,2	18,8	31,2	10,4	266,0
	5 55-64	Recuento	112	75	22	28	6	243
		Frecuencia esperada	94,5	93,3	17,2	28,5	9,5	243,0
	6 65 o +	Recuento	106	107	22	23	3	261
		Frecuencia esperada	101,5	100,2	18,5	30,6	10,2	261,0
Total		Recuento	577	570	105	174	58	1484
		Frecuencia esperada	577,0	570,0	105,0	174,0	58,0	1484,0



Estem interessats en analitzar la dependència-independència de les dues variables, en particular, si el posicionament en relació la llei antitabac depèn de l'edat de les persones. Fixeu-vos que s'aprecien diferències en els percentages segons el grup d'edat. En els grups d'edat de gent més gran, els percentatges de les diferents opinions són apreciablement diferents respecte als altres grups d'edat (i les freqüències difereixen més respecte les esperades, vegeu la taula de contingència). Ara bé, ens podem

preguntar si aquesta diferència és causada pel fet de treballar amb una mostra, o bé si les diferències són prou significatives com per ser extrapol·lables a la població i afirmar que les variables són dependents. En el següent apartat analitzarem aquesta qüestió de manera objectiva.

2.2. **Estudi exploratori-inferencial: test d'independència de khi quadra χ^2.** Recordeu que estem treballant amb una mostra aleatòria extreta d'una població.

▶ Responeu a les següents preguntes:

- Quin percentatge de la mostra troba que la llei antitabac està bé? Quin percentage creu que està molt malament?
- Doneu una estimació per interval de confiança del percentage de la població que creu que la llei antitabac està bé (amb un 95.5% de confiança). Recordeu que el marge d'error ve donat per $e = z\sqrt{\frac{1}{4n}}$ suposant màxima incertesa, i $e = z\sqrt{\frac{p(1-p)}{n}}$ en general.

Cada mostra de la població ens donarà una taula de contingència amb freqüències esperades diferents. Volem esbrinar si, a nivell poblacional, les diferències entre les freqüències observades i esperades són prou significatives.

Per estudiar aquesta qüestió amb rigurositat i objectivitat, farem una prova estadística: el test d'independència de khi quadrat χ^2. En aquest test contrastarem dues hipòtesis:

- la hipòtesi nul·la consisteix en assumir que les dues variables són independents i que no existeix cap relació-associació entre elles,
- la hipòtesi alternativa consisteix en acceptar que sí que existeix algun tipus de relació-associació de dependència.

L'objectiu és saber si, amb un cert nivell de confiança, tenim prou evidències com per rebutjar la hipòtesi nul·la. Anem a veure com realitzar aquest test amb l'SPSS.

Analizar → Estadísticos descripticos → Tablas de contingencia,

Un cop seleccionades les variables, mitjançant la opció **Estadísticos** i marqueu la casella **Chi-cuadrado**.

Obtindreu una taula com la de la figura següent:

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	34,076[a]	20	,026
Razón de verosimilitudes	35,465	20	,018
Asociación lineal por lineal	6,926	1	,008
N de casos válidos	1484		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 5,82.

La metodologia a seguir és la mateixa que la utilitzada en els tests de la Pràctica 10. Compararem el nivell de significació (Sig. asintótica (bilateral)) que ens dóna el test amb el nivell de confiança que haguem escollit per a saber si hi ha prou evidències com per rebutjar la hipòtesi nul·la.

Si realitzem el test amb un 95% de confiança, obtenim el següent resultat:

- **Opinio**: $\boxed{\textbf{Sig. (bilateral)} = 0.026 \leq 0.05}$. Per tant, existeix una relació de dependència entre el grup d'edat al qual pertany l'enquestat i la seva opinió al respecte de la llei antitabac (amb un 5% de risc).

▶ Si realitzem el test amb un 99% de confiança, quina és la conclusió? Podem dir que hi ha prou evidències com per rebutjar la hipòtesi nul.la?

3. Independència de variables categòriques: fumador vs. edat

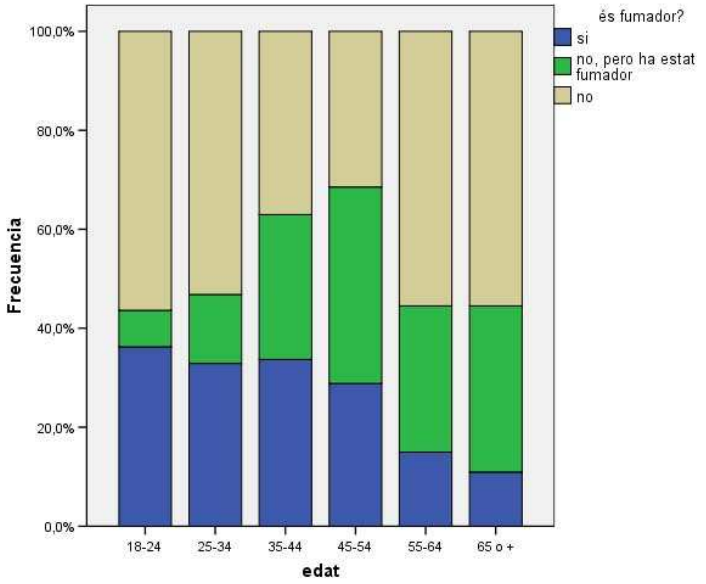
En aquesta part analitzareu vosaltres mateixos si existeix alguna relació entre l'hàbit de fumar i el grup d'edat al qual pertany l'enquestat. La informació necessària es troba al fitxer *tabac-edat 2005.sav*.

▶ Obriu el fitxer *tabac-edat 2005.sav*. Quantes variables conté? De quin tipus són? Quin significat tenen?

▶ Seguint la pauta de la secció anterior, repetiu l'estudi per a analitzar la pregunta de si hi ha relació entre la condició de fumador i el grup d'edat al qual pertany l'enquestat. Realitzeu primer un estudi descriptiu i després un anàlisi inferencial. Feu estimacions puntuals i per intervals dels diferents percentatges en la variable fumador. Què podeu dir de la independència de les dues variables prenent nivells de confiança del 95% i 99%?

Tabla de contingencia edat edat * fumador és fumador?

			fumador és fumador?			
			1 si	2 no, pero ha estat fumador	3 no	Total
edat edat	1 18-24	Recuento	54	11	84	149
		Frecuencia esperada	38,5	39,8	70,8	149,0
	2 25-34	Recuento	99	42	160	301
		Frecuencia esperada	77,7	80,3	143,0	301,0
	3 35-44	Recuento	91	79	100	270
		Frecuencia esperada	69,7	72,0	128,2	270,0
	4 45-54	Recuento	77	106	84	267
		Frecuencia esperada	68,9	71,2	126,8	267,0
	5 55-64	Recuento	37	73	137	247
		Frecuencia esperada	63,8	65,9	117,3	247,0
	6 65 o +	Recuento	29	89	147	265
		Frecuencia esperada	68,4	70,7	125,9	265,0
Total		Recuento	387	400	712	1499
		Frecuencia esperada	387,0	400,0	712,0	1499,0



Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	147,725[a]	10	,000
Razón de verosimilitudes	165,918	10	,000
Asociación lineal por lineal	20,055	1	,000
N de casos válidos	1499		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 38,47.

▶EXERCICI: sexe vs opinió sobre les eleccions

En aquest exercici treballarem amb dades extretes d'una enquesta realitzada pel ICPS (Institut de Ciències Polítiques i Socials) l'any 2004. Es van entrevistar 1201 persones de més de 18 anys, i va ser realitzada entre el 18 i el 25 d'octubre de 2004.

El fitxer *sexe-eleccions.sav* conté el sexe de l'enquestat i la seva opinió de l'enquestat a l'afirmació: 'Les eleccions no serveixen per res perquè al final sempre acaben manant els de sempre'.

- Obriu el fitxer *sexe-eleccions.sav*. Quantes variables conté? De quin tipus són? Quin significat tenen?
- Realitzeu un estudi per a determinar si les variables són independents o existeix alguna relació-associació entre elles (escolliu diferents nivells de confiança). Realitzeu primer un estudi descriptiu i després una anàlisi inferencial.

4. Relació entre variables numèriques: test de regressió

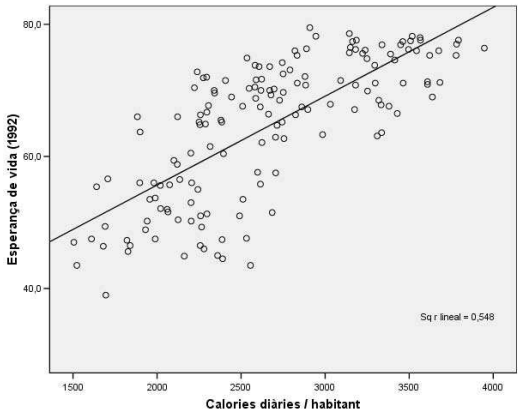
Per a aquesta última part d'aquesta última pràctica, treballarem amb el fitxer *paisos.sav* que conté una mostra àmplia de països del món, i analitzarem la qüestió de decidir si dues variables numèriques presenten una relació de tipus lineal. Aquesta qüestió ja havia aparegut en una pràctica anterior on el càlcul del coeficient de correlació de Pearson ens permetia mesurar el grau d'intensitat en la relació lineal que presentaven dues variables numèriques.

▶ Obriu el fitxer *paisos.sav* . Localitzeu les variables que contenen la informació sobre la ingesta de calories diàries i la esperança de vida. De quin tipus són?

Començarem analizant la relació entre les variables calories i espvida de manera visual amb un gràfic de dispersió. Recordeu que es realitza mitjançant les opcions de menú següents

Gráficos → Dispersión,

on col·locarem la variable calories a l'eix horitzontal (variable independent) i la variable espvida a l'eix vertical (variable dependent). Activeu el gràfic, seleccioneu el núvol de punts, i amb el botó dret del ratolí escolliu l'opció Añadir línea de ajuste total per afegir la recta de regressió al gràfic.



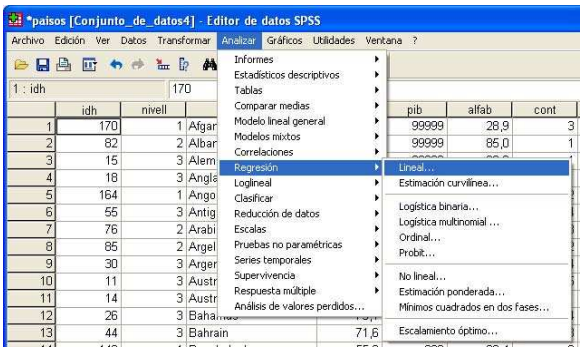
Observem una relació directa entre les dues variables, i a més existeix un cert grau de relació lineal entre elles.

► Calculeu el coeficient de correlació entre aquestes dues variables, quin grau de relació lineal existeix entre elles?

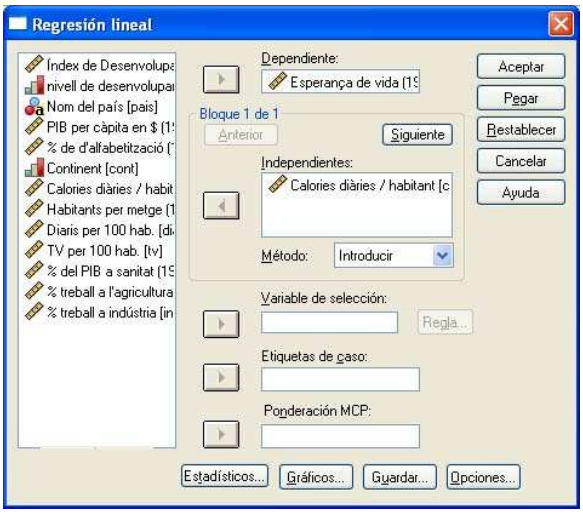
Per saber si aquest grau de relació lineal es prou significatiu com per afirmar que és extrapol.lable a la població (conjunt de tots els països del món), i que, per tant, una variable és explicativa de l'altre, veurem com realitzar una anàlisi de regressió simple amb l'SPSS.

En el menú principal escollim

Analizar → Regresión → Lineal,



i en la finestra de diàleg seleccionem la variable dependent (espvida) i la independent (calories),



A la finestra de resultats obtenim les següents taules,

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,740[a]	,548	,545	7,0903

a. Variables predictoras: (Constante), calories Calories diàries / habitant

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	9147,333	1	9147,333	181,956	,000[a]
	Residual	7540,839	150	50,272		
	Total	16688,172	151			

a. Variables predictoras: (Constante), calories Calories diàries / habitant

b. Variable dependiente: espvida Esperança de vida (1992)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	28,763	2,716		10,590	,000
	calories Calories diàries / habitant	,013	,001	,740	13,489	,000

a. Variable dependiente: espvida Esperança de vida (1992)

(1) A la primera taula podem llegir el coeficient de correlació però sense signe ($R = 0,740$). En el nostre cas serà positiu ja que la relació entre les variables és directa. Fixeu-vos que ens indica una relació lineal moderada. Al costat tenim $R^2 = 0,548$ que és el coeficient de determinació. Ens indica que un $54,8\%$ de la variabilitat de la variable dependent (espvida) és atribuïble a la variable independent (calories). La significació d'aquest coeficient de determinació s'estableix a través de l'anàlisi de la variància en la següent taula.

(2) A la segona taula hi ha el resultat del test que ens permet determinar si la capacitat explicativa observada en el punt anterior és prou significativa com per confirmar-la a nivell poblacional.

En aquest test, la hipòtesi nul·la consisteix en assumir que les variables no presenten una relació significativa. Volem saber si, fixat un nivell de confiança, hi ha prou evidències com per rebutjar la hipòtesi nul·la i acceptar la alternativa (els paràmetres de la recta de regressió són prou significatius com per acceptar una relació entre les dues variables a nivell poblacional).

Amb un nivell de confiança del 95%, observem que $\boxed{\textbf{Sig}=,000 < 0,05}$, i per tant, rebutjem la hipòtesi nul·la. Podem concloure que la capacitat de la

variable calories per explicar la variable espvida és no nul·la (amb un 95% de confiança).

(3) La tercera taula conté els coeficients de la recta de regressió, així com els resultats del test per determinar si aquests coeficients són significativament diferents de zero. Així doncs, la recta obtinguda és

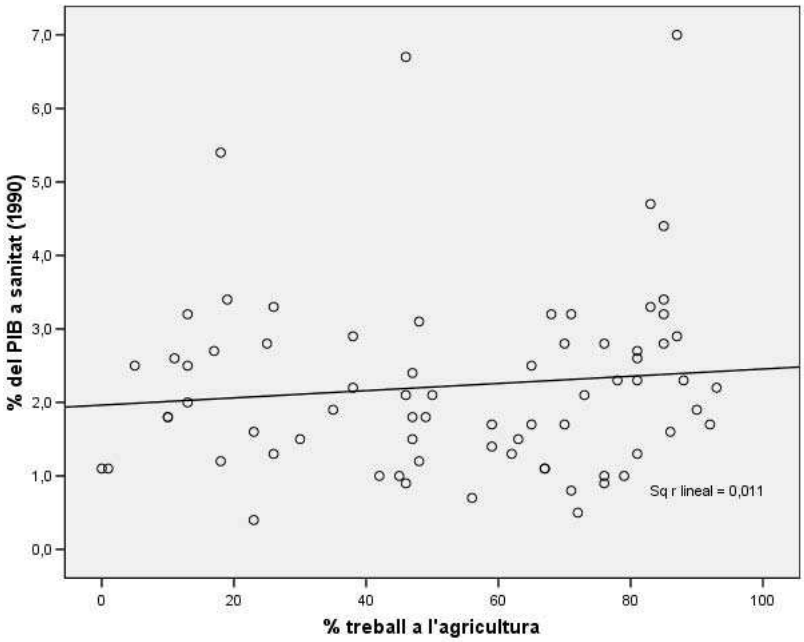
$$espvida = 0,013 * calories + 28,763.$$

I el nivell de significació $\boxed{Sig = 0,000 < 0,05}$ en ambdós casos ens diu que, amb un 95% de confiança tots dos són significativament diferents de zero.

▶ Com s'interpreta la recta de regressió?

Anem a analitzar un altre cas d'(in-)dependència entre dues variables del mateix fitxer. Volem considerar la relació entre el percentatge de treball a l'agricultura (agricult) i el percentatge del PIB destinat a sanitat (sanitat).

▶ Feu un diagrama de dispersió i una anàlisi de regressió simple per aquestes dues variables seguint les pautes de l'exemple anterior.



Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,107[a]	,011	-,003	1,2561

a. Variables predictoras: (Constante), agricult % treball a l'agricultura

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1,278	1	1,278	,810	,371[a]
	Residual	110,447	70	1,578		
	Total	111,724	71			

a. Variables predictoras: (Constante), agricult % treball a l'agricultura

b. Variable dependiente: sanitat % del PIB a sanitat (1990)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	1,964	,328		5,979	,000
	agricult % treball a l'agricultura	,005	,005	,107	,900	,371

a. Variable dependiente: sanitat % del PIB a sanitat (1990)

Observeu que en aquest cas el coeficient de determinació ens diu que només un $1,1\%$ de la variabilitat de la variable dependent és atribuïble a la variable independent.

Fixant un nivell de confiança del 95%, obtenim que el valor de la taula ANOVA $\boxed{\textbf{Sig}= 0,371 > 0,05}$ és un valor molt superior a $0,05$. Per tant, no tenim prou evidències com per rebutjar la hipòtesi nul·la i no podem afirmar la existència d'una relació explicativa entre les dues variables.

A més, en els coeficients de regressió obtinguts, el test de significació del coeficient que multiplica la variable agricult ens diu que no és significativament diferent de zero ($\boxed{\textbf{Sig} = 0,371 > 0,05}$) amb un 95% de confiança, i per tant, podem considerar que la recta de regressió és constant (horitzontal).

► Feu un diagrama de dispersió i una anàlisi de regressió simple per descriure la relació entre les variables que contenen la informació referent al percentatge de treball a l'agricultura i a la indústria.