M. Farré

Dossier 2:
Comparació d'una variable numèrica en 2 o més grups

Utilitzarem les tècniques de l'estadística univariable introduïdes en el Dossier 1 per tal de comparar una variable en els distints grups. No hi ha doncs conceptes nous, es tracta d'usar els ja coneguts amb la finalitat d'analitzar comparativament els grups i extreure'n conclusions.

Per un estadístic és tant o més important que saber fer els procediments gràfics i numèrics apropiats, saber-ne extreure conclusions.

Aquest capítol té un objectiu eminentment pràctic. El que sí presentem amb un cert detall són els procediments dels menús de l'SPSS que anirem utilitzant.

1 Situació de mostres independents

Una mostra o grup correspon a un conjunt de n_1 individus (objectes), donant lloc a n_1 observacions de certa variable numèrica X, i l'altra mostra correspon a un conjunt format per n_2 individus (objectes), donant lloc a n_2 observacions de la mateixa variable X, i eventualment un tercer grup amb n_3 observacions, un quart ...etc.

En aquesta situació, on no hi ha *cap condició d'emparellament entre els individus-objectes d'un i d'altra grup*, diem que tenim grups o mostres independents.

Per **exemple**, es dissenya un experiment per tal de comparar la variable numèrica X = *durada d'un component electrònic dels forns*, en funció de 3 tipus de forn (Tipus 1, Tipus 2, i Tipus 3), que constitueixen tres grups o mostres. Per a cada tipus tenim 6 forns diferents ($n_1 = n_2 = n_3 = 6$), amb un total de 18 observacions. Vegeu la Taula 1, que és la base de dades que conté totes les observacions de l'experiment.
Nota: Quan dos o més grups són independents cal introduir-les d'aquesta forma (una columna per a la variable de grup i una columna per a la variable numèrica).

Taula 1

Tipus forn	Durada
1	237
1	254
1	246
1	178
1	179
1	183
2	208
2	178
2	187
2	146
2	145
2	141
3	192
3	186
3	183
3	142
3	125
3	136

Veurem ara els procediments de l'SPSS per a Windows utilitzats per a establir la comparació.

Com que la variable X és numèrica, el més senzill és comparar les mitjanes, les desviacions típiques, i les medianes dels tres tipus (podríem escollir altres característiques numèriques: rang, simetria, ...). El procediment aplicat és:

Analizar—-> Comparar medias—-> Medias

Dependiente: *X Durada*, **Independiente:** *Y Tipus de forn*

Dins de **Opciones:** Afegim la mediana (mitjana i desv. típica hi són per defecte)

Els resultats es resumeixen a la Taula 2. Veiem que destaca el Tipus 1, amb durada mitjana i durada mediana notablement superiors.

Taula 2

Informe

Durada del component

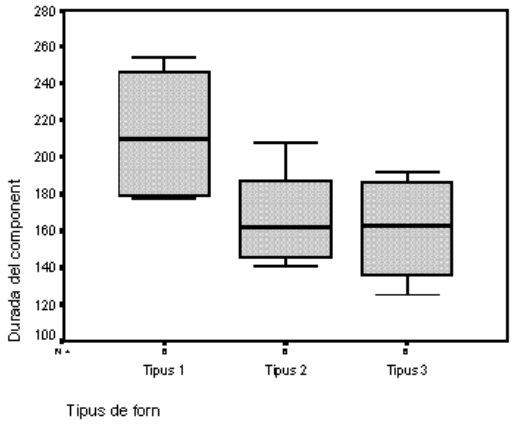
Tipus de forn	Media	N	Desv. tip.	Mediana
Tipus 1	212,83	6	36,41	210,00
Tipus 2	167,50	6	27,57	162,00
Tipus 3	160,67	6	29,50	162,50
Total	180,33	18	37,91	181,00

Per completar l'estudi fem un diagrama de caixa, que es veu a la Figura 1, aplicant:

Gráficos—-> Diagramas de caja—-> Simple- Resúmenes para grupos de casos

Variable: *Durada ...*, **Eje de categorías:** *Tipus de forn*

Figura 1



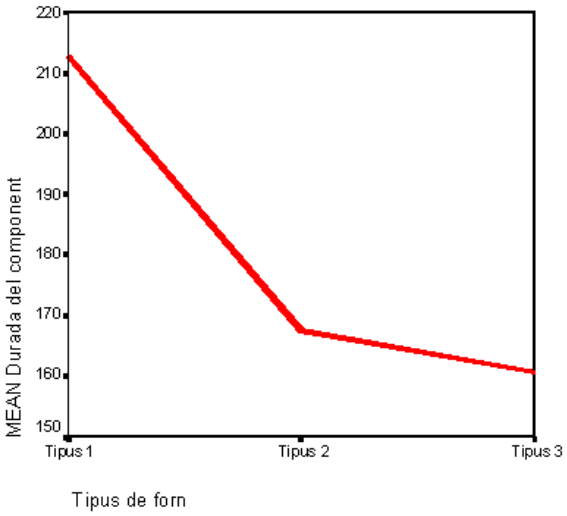
La conclusió és que els forns del primer tipus tenen clarament més durada, mentre que és difícil distingir entre els del tipus 2 i 3, entre els quals no s'aprecien diferències clares en la distribució; tot i que la mitjana (Taula 2) és una mica superior per al tipus 2 que per al Tipus 3, no hi ha pràcticament diferència entre les medianes.

Finalment, la gràfica de línies de la Figura 2 reflecteix els valors de les 3 mitjanes dels tres grups (Tipus 1, Tipus 2 i Tipus 3). Es veu gràficament el valor mitjà molt més elevat en el Tipus 1.

Gràficos—-> Líneas—-> Simple- Resúmenes para grupos de casos

Otra función de resumen: *(MEAN) Durada...* **, Eje de categorías:** *Tipus de forn*

Figura 2



Nota: Un altre exemple de comparació de mitjanes per a dos grups independents el teniu a les pàgines 32, 33 i 34 del Dossier1- Part 4, quan comparem la variable numèrica X *salari*, en els 2 grups definits segons el sexe (grup 1 = homes, grup 2 = dones). Allí usem el procediment

Analizar—-> Estadísticos Descriptivos —-> Explorar...

Dependiente: *Salari ...*, **Factor:** *Sexe*

En aquell exemple recordeu que la sortida i les gràfiques (caixa) ens indicaven l'existència d'*outliers*.

2 Situació de mostres emparellades

En el cas de dades emparellades, considerarem només el cas de dos grups o mostres: els dos grups tindran el mateix nombre d'observacions, n, i en total tindrem $2n$ observacions. Les dades estan emparellades perquè corresponen a un mateix individu-objecte.

Es poden donar dissenys de dades emparellades en casos més generals (no cal que les dades dels dos grups es refereixin al mateix individu-objecte), també es poden emparellar les observacions per un cert *factor d'emparellat*: per exemple, dades de pes de germans bessons; grau de pol.lució de ciutats de dues àrees (Europa i Amèrica) emparellades pel nombre d'habitants; eficàcia de dos tractaments amb dades emparellades segons persones amb els mateixos *antecedents*: edat, hàbits, símptomes,..., etc.

En totes aquestes situacions, on hi ha algun *factor d'emparellat entre els individus-objectes d'un i d'altra grup*, diem que tenim un *disseny de dades emparellades*. L'objectiu d'aquests dissenys és controlar el factor d'emparellat, per tal que no tingui influència sobre els resultats. Per exemple, en el cas de 2 tractaments (A i B), si les parelles estan formades per pacients que tenen un historial mèdic similar evitem que el factor *antecedents* influeixi en els resultats , i es veu que el medicament A és millor que el B, ningú podrà argüir que l'efecte estat condicionat per les característiques dels pacients.

Quan el disseny és de dades emparellades té sentit avaluar les diferències entre un grup i l'altre.

Per **exemple**, considereu la base de dades *mundo95.sav* de la qual en veiem una part en la següent imatge. El factor d'emparellat és el mateix país, la variable és l'esperança de vida, i els dos grups són homes i dones. Hi ha $n = 109$ països, i per tant $2n = 218$ observacions d'esperança de vida.

	país	relig	espvidaf	espvidam	difer
1	Acerbaján	Musulma.	75	67	8,00
2	Afganistán	Musulma.	44	45	-1,00
3	Alemania	Protest.	79	73	6,00
4	Arabia Saudí	Musulma.	70	66	4,00
5	Argentina	Católica	75	68	7,00
6	Armenia	Ortodoxa	75	68	7,00
7	Australia	Protest.	80	74	6,00
8	Austria	Católica	79	73	6,00
9	Bahrein	Musulma.	74	71	3,00
10	Bangladesh	Musulma.	53	53	,00
11	Barbados	Protest.	78	73	5,00
12	Bélgica	Católica	79	73	6,00

A la base de dades hi veiem: les dues columnes amb les dades emparellades de l'esperança de vida femenina i masculina, respectivament: *espvidaf* i *espvidam*; la columna amb el factor d'emparellat (país), una columna amb les diferències individu a individu *difer*, i una columna auxiliar amb una variable nominal *religio*, per si en volem analitzar la influència. [Nota: La variable de diferències ha estat creada per *transformació* de les dues esperances de vida.]

Veurem ara els procediments de l'SPSS per a Windows utilitzats per a establir la comparació.

En primer lloc podem fer una anàlisi exploratòria de l'esperança de vida dels homes i de les dones com si fossin dues variables diferents, sense tenir en compte l'emparellat:

Analizar—-> Estadísticos descriptivos—-> Explorar

Dependientes: *Esperanza de vida masculina*, **i** *Esperanza de vida femenina*

(Cap factor) (Només estadístics, sense gràfiques)

Obtenim la següent taula resum del procediment exploratori:

Explorar

Descriptivos

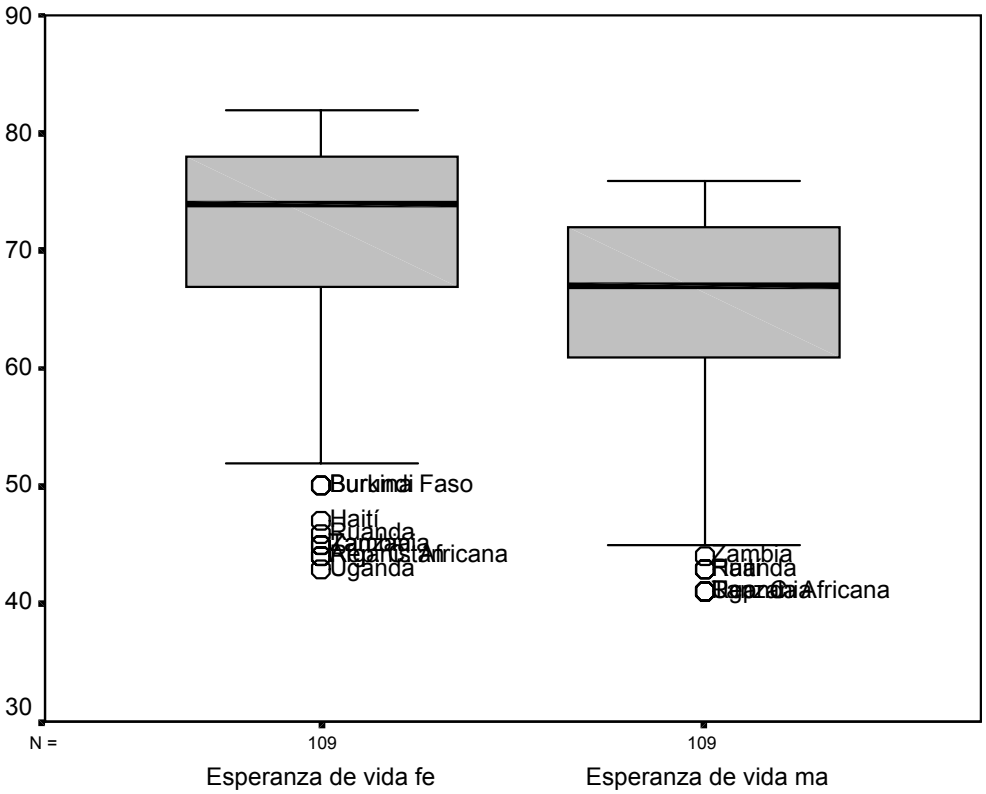
			Estadístico	Error típ.
Esperanza de vida femenina	Media		70,16	1,01
	Intervalo de confianza para la media al 95%	Límite inferior	68,15	
		Límite superior	72,16	
	Media recortada al 5%		70,96	
	Mediana		74,00	
	Varianza		111,762	
	Desv. típ.		10,57	
	Mínimo		43	
	Máximo		82	
	Rango		39	
	Amplitud intercuartil		11,50	
	Asimetría		-1,109	,231
	Curtosis		,213	,459
Esperanza de vida masculina	Media		64,92	,89
	Intervalo de confianza para la media al 95%	Límite inferior	63,16	
		Límite superior	66,68	
	Media recortada al 5%		65,59	
	Mediana		67,00	
	Varianza		85,984	
	Desv. típ.		9,27	
	Mínimo		41	
	Máximo		76	
	Rango		35	
	Amplitud intercuartil		11,50	
	Asimetría		-1,080	,231
	Curtosis		,336	,459

Els resultats mostren una major esperança de vida per a les dones que per als homes. Això també es veu reflectit en el diagrama de caixa:

Gráficos—-> Diagrama de caja —-> Simple

Resúmenes para distintas variables

Las cajas representan: *Esp. vida masculina*, **i** *Esp. vida femenina*



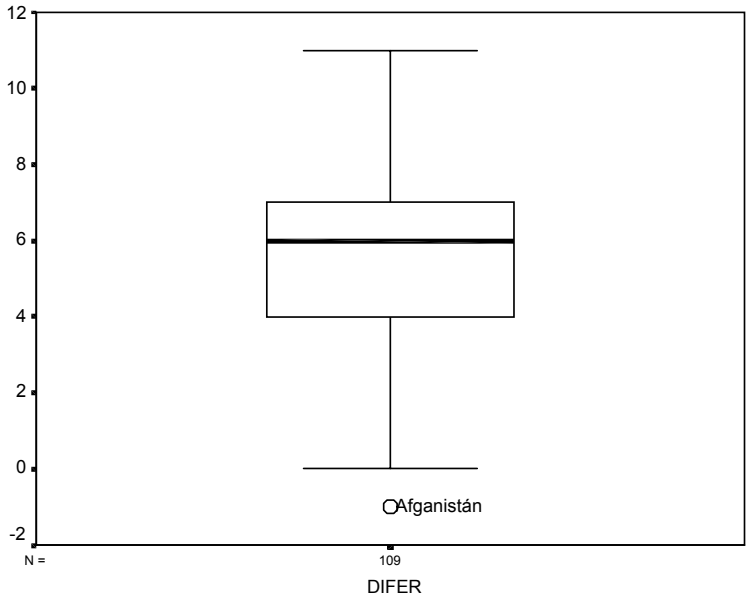
Ara, donat que tenim un factor d'emparellat, ens interessa avaluar el conjunt de les diferències entre esperança de vida femenina i masculina a cada país.

Com que hem creat la variable diferència en l'arxiu de dades, només ens cal fer una anàlisi exploratòria d'aquesta variable i el diagrama de caixa de la variable, resultant:

Explorar: Diferència

Descriptivos

			Estadístico	Error típ.
DIFER	Media		5,2385	,2173
	Intervalo de confianza para la media al 95%	Límite inferior	4,8078	
		Límite superior	5,6692	
	Media recortada al 5%		5,2345	
	Mediana		6,0000	
	Varianza		5,146	
	Desv. típ.		2,2685	
	Mínimo		-1,00	
	Máximo		11,00	
	Rango		12,00	
	Amplitud intercuartil		3,0000	
	Asimetría		-,020	,231
	Curtosis		,121	,459

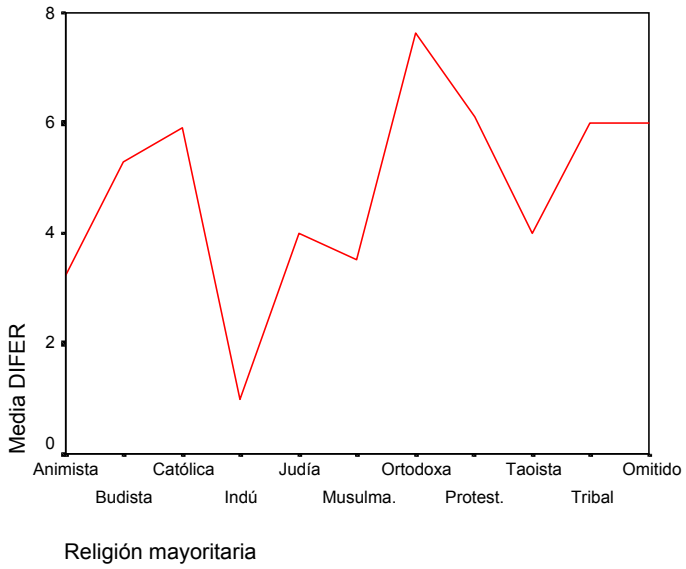


En el diagrama de caixa es pot veure com totes les diferències són positives excepte dues: la corresponent a l'Afganistan, que és -1, i la corresponent a Bangladesh [es veu a la base de dades], que és 0. Per tant, exceptuant aquests països, en tots els altres l'esperança de vida de les dones és superior a la dels homes. En mitjana la diferència (vegeu la taula) és de $\approx$5.24 anys, i en mediana de 6 anys (en el 50% dels països les dones viuen (en mitjana, perquè l'esperança de vida és la mitjana del país) 6 anys o més que els homes). La diferència màxima és 11 anys (!), i correspon a Letònia.

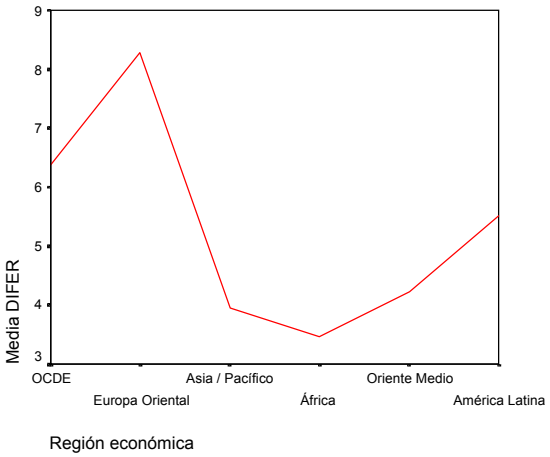
Una altra dada interessant que veurem més endavant és la correlació entre esperança de vida femenina i masculina. Avancem que aquest valor és 0.982 que és molt alt (com a màxim val 1). Això indica que els països que tenen baixa esperança de vida per als homes, també la tenen baixa per a les dones i els que la tenen més alta ho és tant per homes com per a dones. Aquest fet queda reflectit en el diagrama de dispersió.

Diagrama de dispersió

Vegem també un diagrama de línia que permet veure la relació entre la variable de diferències (EVFem - EVMasc) i un factor nominal. En aquest cas, el factor religió: Què hi veieu?



I de la relació de les diferències amb el factor regió econòmica?



COMENTARI FINAL: Comparar una variable numèrica X en dos o més grups equival a analitzar la relació entre la variable numèrica X i la variable categórica o nominal $Y =$ *grup*.

Finalment, a la pàgina següent hi veieu dues gràfiques més: com les interpreteu? quina us sembla més entenedora?

Nota: **Mean EVF** significa Mitjana d'Esperança de vida femenina, i **Mean EVM** significa Mitjana d'Esperança de vida masculina.

Gráfico de máximos i mínimos: línea de diferencias
Resúmenes para distintas variables (meanEVF, meanEVM)

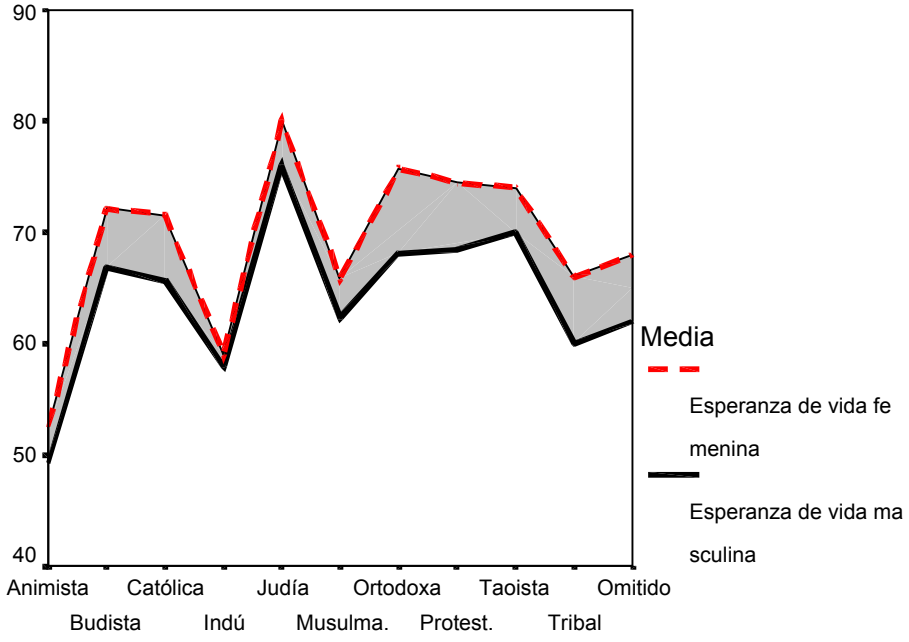


Gráfico áreas: Apilado
Resúmenes para distintas variables (meanEVF, meanEVM)