4. PRAKTIKA

INFERENTZIA ESTATISTIKOA: ESTIMAZIO PUNTUALA ETA KONFIDANTZA TARTEKO ESTIMAZIOA

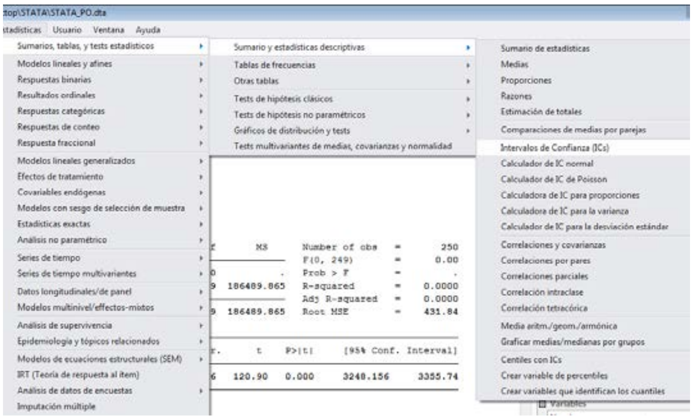
Inferentzia estatistikoak bi atal ditu; **estimazioa eta hipotesien kontrastea**. Ikasleak estimazio puntuala eta konfiantza tarteko estimazioak zer diren eta ordenagailuz nola egiten diren ikasiko du alde batetik eta bestetik kontrasteak. Populazio baten aldagai bat, kolesterola maila adibidez, aztertzea helburu denean, ezinezkoa egiten zaigu populazio guztiari odol lagina ateratzea eta beraz, laginketaren bitartez populazio guztiari loturiko emaitzak eta ondorioak ateratzen ditugu. Horretarako behar beharrezkoa da laginketa ondo egitea. Estimazioaren helburua beraz, laginketaren bitartez populazioaren parametro konkretu batera hurbiltzea izango litzateke. Adibidez, Euskal Herriko kolesterol mailaren batezbestekoa populazioaren **parametroa (**klaseetan μ *eta* p ikusi ditugu**)** izango litzake eta laginketaren bitartez lorturiko emaitza berriz, laginketaren batezbestekoa izango litzake eta parametro hitza erabili beharrean **estatistikoa** hitza erabiliko genuke (klaseetan ikusi ditugun estatistikoak: $\bar{x}$ o $\hat{p}$). Kasu honetan erabiliko genukeen estatistikoa laginaren batezbestekoa izango litzake (laginaren kolesterol mailaren batezbestekoa). Bi motatako estimazioak daude: puntuala eta konfiantza tarteko estimazioa.

Estimazio puntuala: laginean lorturiko estatistikoaren bidez, populazioko parametro ezezagunari hurbiltzean datza. Adibidez, INMAko emakumeak Gipuzkoan 2006 eta 2008 urte bitartean zeuden emakume haurdunen lagin bat izango litzake. Beraz, laginketatik ondorioztatzen diren emaitzak Gipuzkoako emakumeen populazioari orokortu daiteke. Adibidez, espero dugu INMAko haurren pisuaren batezbestekoa jaiotzean Gipuzkoako umeen pisuaren batezbestekoari gerturatzea (estatistikoa: INMAko pisuaren batezbestekoa, $\bar{x}$ eta parametroa: Gipuzkoako umeen pisuaren batezbestekoa: μ). Gerturatze prozesu honi estimazio puntuala deritzo.

Estimazioa konfiantza tartea emanez: Populazioaren parametroaren balioa tarte baten bitartez gerturatzean datza. Horretarako, lehenik eta behin, konfiantza maila definitu behar da. N aldiz laginketa eginez gero, eta bakoitzean konfiantza tartea ateraz gero, estimatu nahi den parametro zenbat aldiro azalduko litzatekeen esan nahi du Konfiantza mailak (1- α). Konfiantza maila orokorrean %95a izan ohi da. Kasu honetan, estimazioa, konfiantza maila konkretu bat definitu eta populazioaren parametroa konfiantza tarte baten bitartez gerturatzea izango litzake. α: esangura maila.

Batezbestekoaren estimazioa: puntuala eta konfiantza tartea: pisua aldagaia

Funtzioen atalera joan: **Estadísiticas/Sumario…/intervalos de confianza** (aukeratu pisua eta OK-ri sakatu)



Komandoen bitartez ere egin dezakezue (errazagoa da): **ci means pisua**

Taulak honako emaitzak azaltzen ditu: pisuaren batezbestekoa, errore estandarra eta konfiantza tartea. Laginketa ondo eginez gero, Gipuzkoako ume jaioberrien pisuaren batezbestekoa (populazioaren parametroa) 3248.2gr eta 3355.7gr-ren artean egon beharko litzake. Egia izango ote da?

```
. ci means pisua

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
       pisua |        250    3301.948    27.31226        3248.156     3355.74
```

Pisuaren konfiantza tartea umearen sexuarekiko desberdindu nahi badugu, honako hau idatzi beharko genuke:

by generoa, sort : ci means pisua

Funtzioen atala erabili nahi badugu, aurreko pausuak jarraitu beharko genituzke eta **by/if**-atalean **generoa** aukeratu beharko genuke.

Generoarekiko lortutako emaitzak bete itzazue taula honetan:

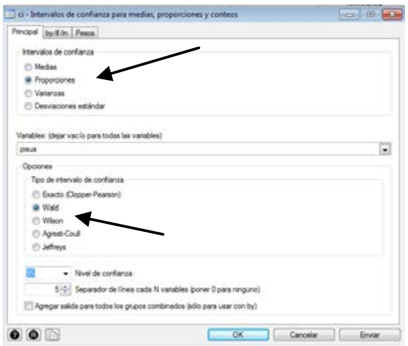
PISUA aldagaia		N	Estimazio puntuala	Estimazioa %95eko konfiantza tartea erabiliz
Orokorra (aldagai kualitatiborik gabe)		250	3301.9	[3248.2;3355.74]
Generoa	Neska			
	Mutila			

Taulako emaitzak ikusirik, nesken eta mutilen pisuak desberdinak dira estimazio puntuala begiratuz. Konfiantza tartea ikusten badugu nahiko berdintsuak direla dirudi . Zein konfiantza mailarekin onar dezakegu biak berdinak direla? Eta biak desberdinak direla? Hipotesien kontrastearen atalean ikusiko dugu.

Proportzioaren estimazioa: puntuala eta konfiantza tartea: 35urte edo gehiago duten emakumeen proportzioa.

Egin ezazue ariketa berdina amaren adina kategorizatuarekin (adina_bi: **0:<35urte** eta **1: ≥35urte**). Kasu honetan erabiliko duzuen estatistikoa proportzioa da ($\hat{p}$). Horretarako, lehenik eta behin, joan zaitezte

Joan zaitezte **Estadísticas/Sumario…/Intervalos confianza.**



Oraingo honetan ***"proporciones"*** aukeratu eta ***"Wald"*** aukeratu (aukera honek proportzio negatiboak sahiesten ditu). OK-ri sakatu.

Generoarekiko eginez gero, **by/if/in** aukeran ***Repetir commandos por grupos*…generoa** aukeratu.

Komandoen bitartez, honela adieraziko genuke:

by generoa, sort : ci proportions adina_bi, wald

Aurrekoan bezala, bete generoarekin lorturiko emaitzak bete.

Adina_ edad_kat		N	puntuala	Konfiantza tartea %95
		250	0.208	[0.157;0.259]
NESKA				
MUTILA				

Taulako emaitzen arabera, ez dago desberdintasun handirik amaren adina generoarekin konparatzen badugu. Hala ere desberdintasunak dauden ala ez ondorioztatzeko kontrasteetara joango gara.

INFERENTZIA ESTATISTIKOA: HIPOTESIEN KONTRASTEA

Praktika honetan, ikasleak hipotesien kontrastea, STATAren bitartez, populazio bakarrarekin nola egin eta p-balioa interpretatzen ikasiko du. Hipotesien kontrasteak laginetik kalkulatutako estatistikoan oinarrituz zein punturaino hipotesi nuloa, H_0, onartu edo ez onartu du helburu. Beraz hipotesien kontrastea bi hipotesien formulaketarekin hasten da. Hipotesi nuloa, H_0, berdinketa izan ohi da, populazioko parametroaren berdinketa, eta kontrako hipotesia berriz, H_1, kontrakoa izan ohi da (<,> edo ≠). Medikuntza arloan H_1 ohikoena ≠ izan ohi da.

Populazio bakarreko hipotesi kontrastea

Adibidea. Demagun Donostiako Hospitalean ume jaioberrien batezbesteko pisua 2800gr dela. Baieztapen hau kontrastatu genezake INMAko datuekin. Kasu honetan parametro: 2800gr izango litzake eta laginaren batezbestekoa berriz, INMAko jaioberrien batezbesteko pisua izango litzake. Hipotesi nuloa populazioaren parametroak 2800 balio duela esan nahi du eta kontrako hipotesia berriz, parametroa 2800ren desberdina dela (beraz, 2 aldetako kontrastea).

$$H_0: \mu = 2800$$
$$H_1: \mu \neq 2800$$

- STATArekin egin beharreko pausuak:

Lehenik eta behin INMAko jaioberrien pisua zein den jakin nahi dugu. Horretarako komandoen atala erabili eta idatzi:

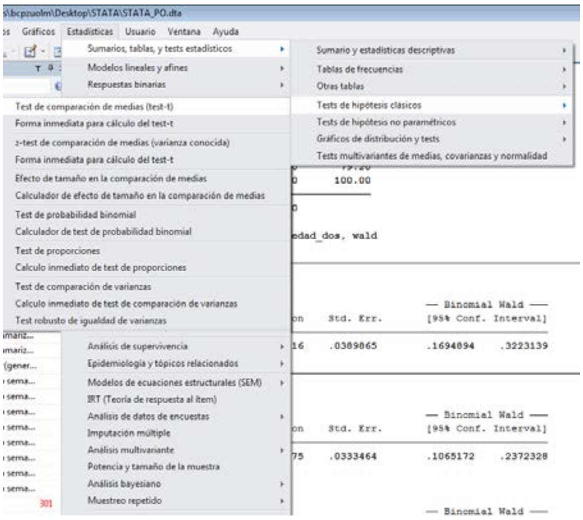
sum pisua

Variable	Obs	Mean	Std. Dev.	Min	Max
pisua	250	3301.948	431.8447	2140	4710

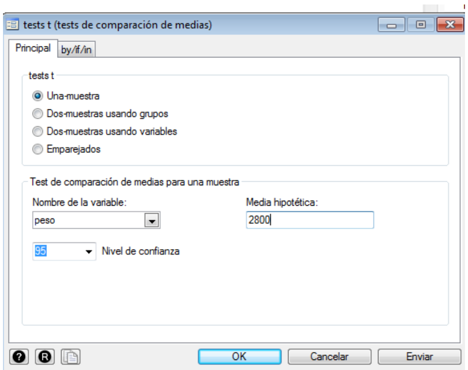
INMA laginaren estatistikoaren balioa batezbestekoaren balioa izango litzake kasu honetan $\bar{X} = 3302$ gr da. Konfiantza maila konkretu baten bitartez hipotesi nulua onartu edo ez onartzea izango da gure helburu. Orokorrean konfiantza maila %95a izan ohi da (1-alfa). Alfa, α, esangura maila deritzo, errore tipo 1-a izatearen probabilitateari deritzo P[H_0 arbuiatu|H_0 egia].

STATAko interfazera joan eta Funtzioen atalean:

Estadísticas/Sumarios,tablas,…/test de hipótesis clásicos/test de comparación de medias



Aukeratu **pisua** aldagaia eta sartu kontrastatu nahi dugun balorea, kasu honetan, 2800.



-Emaitzak

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
pisua	250	3301.948	27.31226	431.8447	3248.156	3355.74

```
    mean = mean(pisua)                                    t =  18.3781
Ho: mean = 2800                          degrees of freedom =      249

   Ha: mean < 2800           Ha: mean != 2800            Ha: mean > 2800
 Pr(T < t) = 1.0000     Pr(|T| > |t|) = 0.0000        Pr(T > t) = 0.0000
```

-p-ren balioa

Suposaturik hipotesi nulua egiazkoa dela, p-ren balioa, hipotesi nulua ez arbuiatzeko izango lukeen esangura maila minimoena izango litzake (alfa minimoena). p-ren balioa gure datuen arabera 0.000ren balioa du eta horrek esan nahi du hipotesi nulua onartzeko probaren esangura maila 0 izan beharko litzakeela. p-ren balioa kontrastearen esangura mailarekin konparatu behar da beti, α-balioarekin. Orokorrean α=0.05 ezarri ohi da eta beraz, p-ren balioa α baino txikiagoa denean hipotesi nulua ezin dugu onartu.

Erabakiak hartzeko irizpideak:

Ez dugu H_0 onartuko **p-ren balioa** $\leq \alpha$ (orokorrean $\alpha = 0.05$). H_0 onartuko dugu p-ren balioa $> \alpha$

Gure datuen arabera ezin dugu hipotesi nulua onartu %5eko esangura mailan ($\alpha = 0.05$). Emaitzen arabera desberdintasun esanguratsuak aurkitu ditugu INMAko umeen eta Donostia Ospitaleko umeen pisuaren batezbestekoarekin. Zergatik dira horren desberdinak? Zein izango litzake arrazoia? Suposatzen da laginketa ondo eginez gero INMAko umeen pisua eta ospitalekoen umeenak berdintsuak izan beharko liratekeela baina, kontrastean berriz, ez dugu hori ondorioztatu.

Desberdintasunaren arrazoia: Donostiako ospitalean arazoak dituzten haur guztiak erregistratzen ditu (ume goiztiarrak eta pisu baxua dutenak) eta beraz INMAko umeen pisua eta Donostiako umeen pisuaren batezbestekoa ez dira berdinak.

BI POPULAZIO ASKEEN BATAZBESTEKOEN ARTEKO KONTRASTEA: batezbestekoak

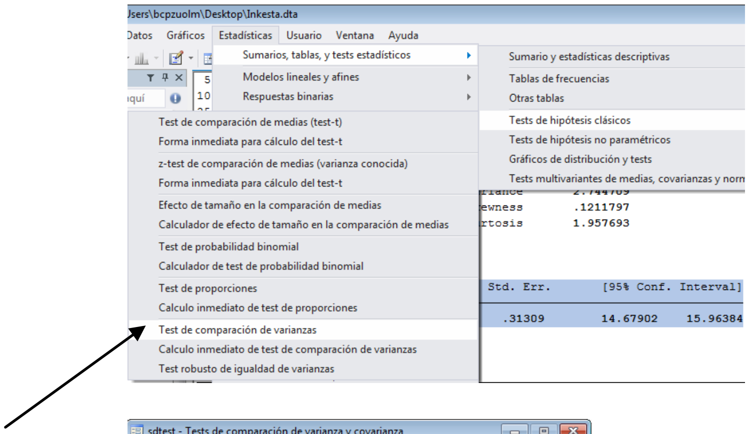
Gaurko klasean ikasiko dituzuen kontraste motak laginak askeak (**independienteak**) direnean bakarrik erabiltzen dira. Aldagai bakoitzari dagozkion datuak pertsona desberdinen balioak izango dira eta EZ pertsona berberarenak. Imajinatu tratamendu baten eragina aztertu nahi dugula tratamendua baino lehen eta gero. Kasu honetan pertsona bakoitzak bi emaitza izango lituzkeenez ez direla independenteak esango genuke. Praktika honetan soilik lagin independenteetan oinarrituko gara. Adibidez, imajina dezagun umearen pisua generoarekiko kontrastatu nahi dugula. Pisuaren aldagaiak banaketa normala duela suposaturik, honako kontrastea egin dezakegu:

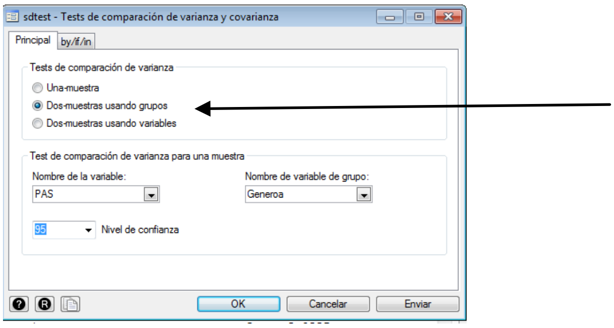
$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$

Nola egingo dugu kontrastea STATAren bitartez?

Batezbestekoen kontrastea egin aurretik, bariantzak berdinak direla egiaztatu behar dugu. Horretarako, bariantzeen testa egingo dugu.

Estadísticas/Sumarios/test de hipótesis clásicos/test de comparación de varianzas





Variance ratio test

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Emakumea	39	140	2.603848	16.26103	134.7288	145.2712
Gizona	61	142.2623	2.154916	16.83043	137.9518	146.5728
combined	100	141.38	1.656501	16.56501	138.0931	144.6669

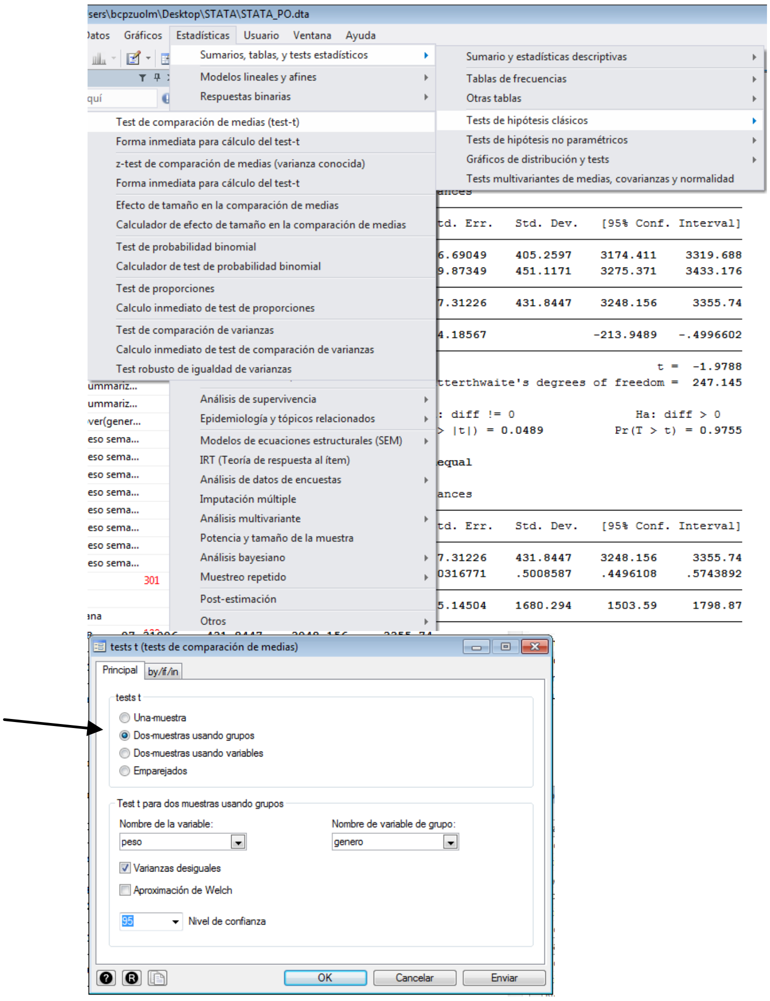
```
    ratio = sd(Emakumea) / sd(Gizona)                        f =   0.9335
Ho: ratio = 1                               degrees of freedom =   38, 60

    Ha: ratio < 1               Ha: ratio != 1                Ha: ratio > 1
  Pr(F < f) = 0.4165        2*Pr(F < f) = 0.8329          Pr(F > f) = 0.5835
.
```

Testaren emaitza (p-balorea) α (0,05) baino handiagoa bada, bariantzak berdinak direla asumituko dugu eta batezbestekoen testa egin dezakegu.

Komandoen bitartez eginez gero:

Estadísticas/Sumarios/test de hipótesis clásicos/test de comparación de medias



Konfiantza maila %95a utzi ezazue dagoen bezala

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
neska	122	3247.049	36.69049	405.2597	3174.411	3319.688
mutila	128	3354.273	39.87349	451.1171	3275.371	3433.176
combined	250	3301.948	27.31226	431.8447	3248.156	3355.74
diff		-107.2243	54.32529		-214.222	-.2264907

```
    diff = mean(neska) - mean(mutila)                             t =  -1.9737
Ho: diff = 0                                     degrees of freedom =      248

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0248         Pr(|T| > |t|) = 0.0495          Pr(T > t) = 0.9752
```

Kontrasteko emaitzaren arabera umearen pisuaren batezbestekoa generoarekiko desberdina dela ondoriozta dezakegu.

BI POPULAZIO BAINO GEHIAGOREN BATAZBESTEKOEN KONTRASTEA: BATEZBESTEKOAK

Bi populazio baino gehiagoren batezbestekoa kontrastatzeko ANOVAren kontrastea erabili ohi da.

$$H_0 : \mu_A = \mu_B = \mu_C$$

Kontrako hipotesia, besteekin alderatuz, desberdina izango litzake. Kontraste honetan, kontrako hipotesia: gutxienez bat desberdina dela ondoriozta dezakegu soilik.

ANOVAren kontrastea zertan oinarritzen da? ANOVAk laginen batezbestekoen arteko aldakortasuna eta lagin bakoitzean dagoen aldakortasuna hartzen ditu kontutan. Horrexegatik bariantzen analisia ere deritzo.

$$\boxed{F = \frac{LAGINEN\ ARTEKO\ BARIANTZA}{LAGINEN\ BARIANTZA}}$$

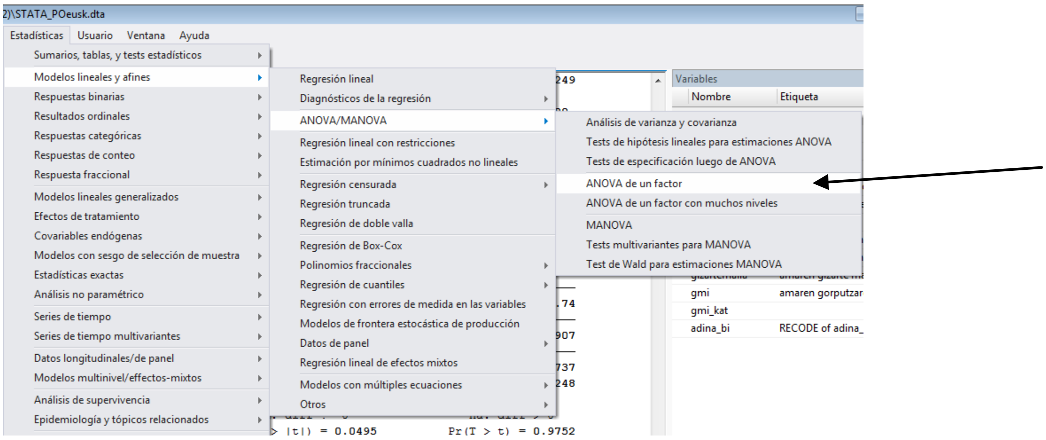
- **Laginen arteko bariantza**: laginen batezbestekoen arteko bariantza (INTER).
- **Laginen bariantza**: laginek duten bariantza (INTRA).

Bien arteko zatiketatik ateratzen den emaitza Fisher-Snedecor-en balioa izango litzake. Fisher-Snedecor-en dentsitate funtzioaren arabera p-ren balioa kalkulatuko genuke. Horretarako STATA erabiliko dugu.

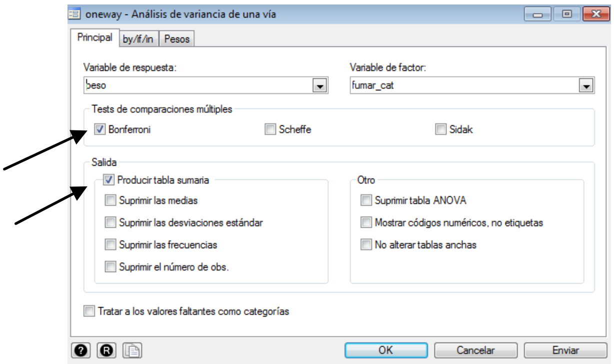
KONTUZ! ANOVA kontrastea egiteko lehenik eta behin laginek beraien artean bariantza berdina izan behar dute.

Imajina dezagun umearen pisua erretze-ohiturekiko konparatu nahi dugula. Aldagai jarraia: **pisua,** eta aldagai kategorikoa (3 kategoria): erretze ohitura. Bi kategori izanez gero T-Student erabiliko genuke, kasu honetan erretze ohiturak 3 kategoria dituenez 3 populazioen kontrastea egin behar dugu eta beraz ANOVA erabiliko dugu.

Joan zaitezte: Estadísticas/ Modelos lineales y afines/ ANOVA/MANOVA/ANOVA de un factor



Variable RESPUESTA (aldagai jarraia): **pisua** eta factor: **erre_kat,** aukeratu **Bonferroni** “comparaciones múltiples” jartzen duenaren azpian eta **producir tabla** en Salida:



ANOVAren kontrastea erabiltzeko lehen esan dugun bezala, beharrezkoa da lagin guztiek bariantza bera izatea, horretarako **Barlett deritzon testa** egiten da. Kontrastearen arabera (hurrengo irudian orlegiz dago: p-balioa=0.220) bariantzak berdinak direla onartu dezakegu eta beraz ANOVAren bitartez eginiko kontrastea ontzat har dezakegu. Goazen orain ANOVAren emaitza aztertzea:

Aukeratu **Producir tabla:**

h�bito tab�quico de la madre	Summary of peso al nacer (grms) Mean	Std. Dev.	Freq.
no_ha_fum	3304.835	442.77795	103
fumadora_	3190	448.83336	73
dejan_de_	3408.3649	373.44019	74
Total	3301.948	431.84472	250

h�bito tab�quico de la madre	Summary of peso al nacer (grms) Mean	Std. Dev.	Freq.
no_ha_fum	3304.835	442.77795	103
fumadora_	3190	448.83336	73
dejan_de_	3408.3649	373.44019	74
Total	3301.948	431.84472	250

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	1753736.98	2	876868.491	4.85	0.0086
Within groups	44682239.3	247	180899.754		
Total	46435976.3	249	186489.865		

Bartlett's test for equal variances: chi2(2) = 3.0242 Prob>chi2 = 0.220

ANOVA egiterakoan, bere p-balioa **Prob > F izango da; 0.0086 (gorriz dagoena)**

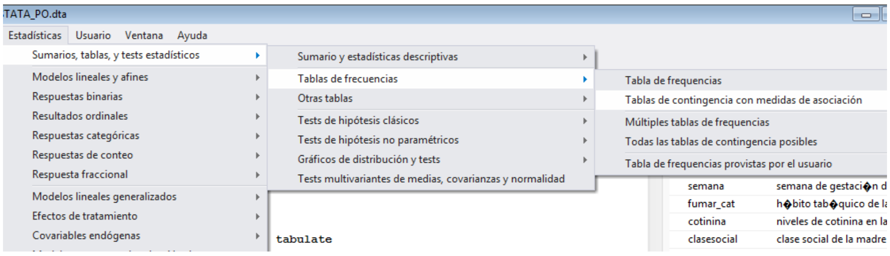
p-balioa<0.05 denez ondoriozta dezakegu gutxienez bat desberdina dela.

BI ALDAGAI KUALITATIBOEN ARTEKO kontrastea: INDEPENDENTZIAREN kontrastea

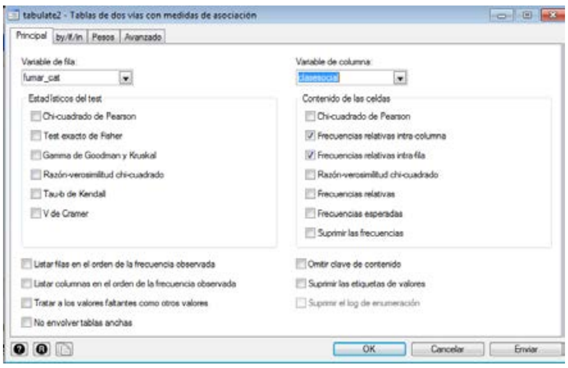
Bi aldagaien arteko deskribapena eta erlazioa aztertzeko kontingentzi taula erabili ohi dugu. Beraien arteko erlazioa INDEPENDENTZIAREN kontrastearen bitartez ondorioztatzen da. Eman dezagun erretze ohiturak gizarte mailaren artean erlazioa/menpekotasuna dagoen aztertu nahi dugula.

Horretarko lehenik eta behin KONTINGENTZI TAULA egingo dugu.

Funtzioen barran: **Estadísticass/Sumarios/Tablas/Tablas de contingencias con medidas de asociación**



Aukeratu itzazue bi aldagaiak (**erre_kat** eta **gizartemaila**) eta baldintzatuen portzentajeak:



Komandoen bitartez honela idatziko genuke: **tab erre_kat gizartemaila, column row**

amaren erretze ohiturak	amaren gizarte maila ez_manual	manuala	Total
ez_du_inoiz_erre	67	36	103
	65.05	34.95	100.00
	48.91	31.86	41.20
haurdunaldian_erretza	27	46	73
	36.99	63.01	100.00
	19.71	40.71	29.20
haurdunaldia_baino_le	43	31	74
	58.11	41.89	100.00
	31.39	27.43	29.60
Total	137	113	250
	54.80	45.20	100.00
	100.00	100.00	100.00

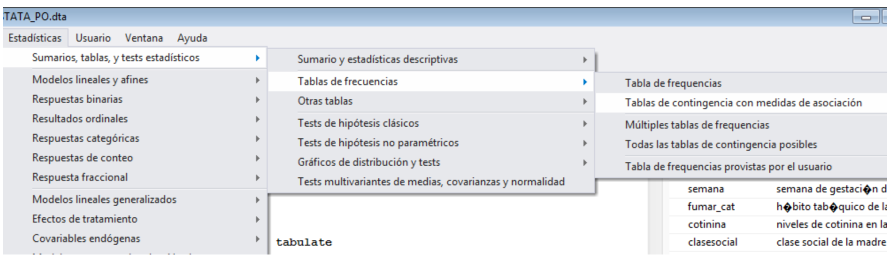
Taula honen bidez ikusi dezakegu nola bai erretze ohiturekiko eta bai gizarte mailarekiko proportzioak desberdinak direla. Adibidez erretze ohiturekiko baldintzaturiko portzentajeak begiratzen baditugu: Inoiz erre ez dutenen artean %65a ez manuala da eta erretzaileen taldean berriz %36.99a. Gizarte mailarekiko balditzantuak diren portzentajeak begiratzen baditugu: Ez manualen taldean gehiengoa ez du inoiz erre (%48.91) eta manualen taldean berriz gehiengoa erretzailea da (%40.71). Hau ikusirik onartu dezakegu desberdintasun hauek? Bi aldagaien artean erlazioa dagoela ondoriozta dezakegu? Hau da, ez manualak manualekin konparatuz erretze ohitura desberdina dute? Desberdintasun hauek estatistikoki esanguratsuak dira? Desberdintasunak dauden ondorioztatzeko **INDEPENDENTZIAREN kontrastea egin behar dugu Ji-karratuaren bitartez. Hipotesi nuloa:**

H_0: BI ALDAGAIAK BERAIEN ARTEAN INDEPENDIENTEAK DIRA

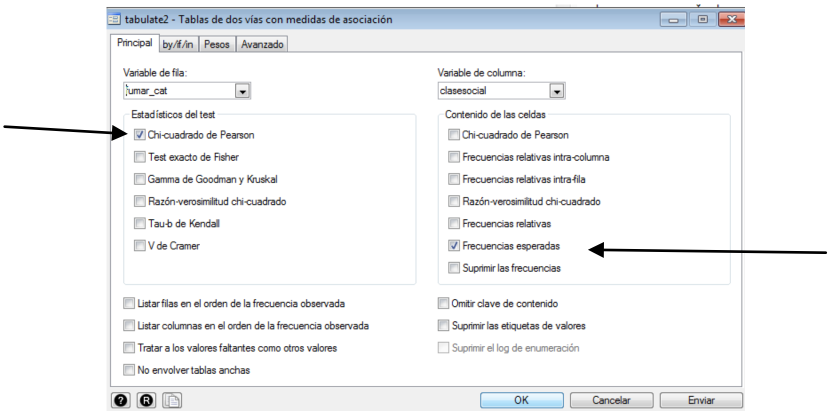
- STATArekin:

KONTUZ! Kontraste mota hau 5 kasu baino gutxiago esperoko bagenitu ezingo genuke erabili.

Funtzioen barra: **Estadísticas/Sumarios/Tablas/ Tablas de contingencias con medidas de asociación**



Aukeratu: **Estadísticos de test: Chi cuadrado Pearson y frecuencias esperadas**



Komandoen bitartez: **tab erre_kat gizartemaila, exp chi2**

amaren erretze ohiturak	amaren gizarte maila ez_manual	manuala	Total
ez_du_inoiz_erre	67 56.4	36 46.6	103 103.0
haurdunaldian_erretza	27 40.0	46 33.0	73 73.0
haurdunaldia_baino_le	43 40.6	31 33.4	74 74.0
Total	137 137.0	113 113.0	250 250.0

Pearson chi2(2) = 14.0467 Pr = 0.001

Intersekzio bakoitzean 5 kasu baino gehiago espero ditudanez kontrastea ondo egina dago.
P-balioak 0.001 balioa duenez bi aldagaien artean erlazioa dagoela ondoriozta dezakegu. Beste modu batera esanda: Proportzioen (edo portzentajeak) banaketa kategoria bakoitzerako desberdin banatzen dela ondoriozta dezakegu.