

Análisis de la varianza

ANOVA



Francisco Montes

Departament d'Estadística i I. O.

Universitat de València

<http://www.uv.es/~montes>

Comparación de k medias

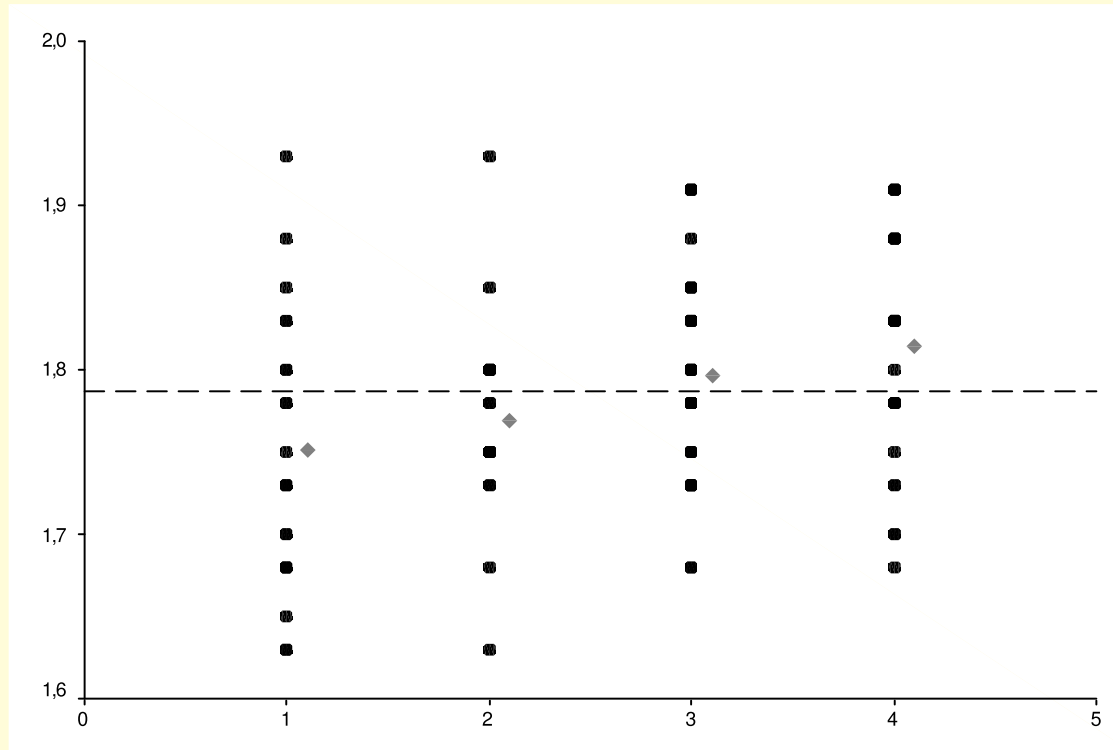
Estatura y voz

Los datos de la tabla que sigue se han extraído de Hand, D. J., Daly, F., Lunn, A. D., McConway, K.J. y Ostrowski E. (1994) *Small Data Sets*. London: Chapman and Hall, p.276, y corresponden a las estaturas, expresadas en metros, de cantores masculinos de la Sociedad Coral de Nueva York. Los cantores se han agrupado según su tono de voz en orden decreciente de tonalidad (de Tenor 1 a Bajo 2). Se sospecha que las voces bajas corresponden a personas de mayor estatura.

Estatura y voz: los datos

Tenor 1	1.75	1.83	1.80	1.68	1.93	1.88	1.80	1.68	1.73	1.70
	1.78	1.65	1.83	1.78	1.73	1.63	1.85	1.68	1.73	1.70
	1.63									
Tenor 2	1.63	1.85	1.75	1.80	1.75	1.93	1.80	1.75	1.80	1.68
	1.75	1.80	1.80	1.80	1.75	1.78	1.75	1.73	1.78	1.73
	1.75									
Bajo 1	1.83	1.78	1.83	1.75	1.85	1.80	1.83	1.73	1.73	1.80
	1.68	1.73	1.80	1.85	1.85	1.78	1.73	1.78	1.91	1.73
	1.80	1.78	1.88	1.78	1.91	1.91	1.75	1.83	1.80	1.78
	1.80	1.73	1.78	1.91	1.83	1.68	1.83	1.78	1.75	
Bajo 2	1.83	1.91	1.70	1.91	1.88	1.83	1.83	1.88	1.83	1.83
	1.88	1.78	1.68	1.73	1.91	1.73	1.78	1.83	1.70	1.78
	1.78	1.75	1.83	1.80	1.88	1.91				

Estatura y voz: un gráfico



1 - Tenor 1

2 - Tenor 2

3 - Bajo 1

4 - Bajo 2

- - - - media global o gran media

◇ media del grupo

Estatura y voz: hipótesis a contrastar

Si μ_i , $i = 1, \dots, 4$ designa la media de la población i , lo primero que debemos hacer es contrastar que dichas medias no son iguales,

$$\mathbf{H}_0: \mu_1 = \mu_2 = \mu_3 = \mu_4,$$

frente a

$\mathbf{H}_1$: las medias son distintas.

Obsérvese que $\mathbf{H}_1$ puede significar muchas cosas, por ejemplo, $\mu_1 = \mu_2 \neq \mu_3 = \mu_4$, $\mu_1 \neq \mu_2 = \mu_3 = \mu_4$, ...

¿Cómo contrastar la igualdad de k medias?

Si empleamos el test de la **t de Student** para 2 muestras independientes hemos de efectuar $\binom{4}{2}$ comparaciones, pero además hay un problema con el nivel de significación final.

Si las 6 comparaciones condujeran a aceptar $\mu_i = \mu_j$ con un nivel $\alpha = 0,05$, el nivel de significación para

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4,$$

sería $\alpha' = 0,26$. En general $\alpha' = 1 - (1 - \alpha)^n$, donde n es el número de comparaciones.

$$k = 6 \rightarrow n = 15 \rightarrow \alpha' \approx 0,5$$

Una comparación simultánea: ANOVA

Necesitamos poder comparar simultáneamente todas la medias. El test que lo permite es el test **ANOVA** (de **AN**alysis **Of** **VA**riance).

Como su nombre indica, compara varianzas aunque lo que contrastamos sean medias. Para ello parte de 3 requisitos previos:

Independencia: las k muestras son independientes,

Normalidad: $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, k$, y

Homocedasticidad: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$.

Fundamentos del ANOVA (1)

	k grupos					
	1	2	...	i	...	k
	x_{11}	x_{21}	...	x_{i1}	...	x_{k1}
	x_{12}	x_{22}	...	x_{i2}	...	x_{k2}
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$
	x_{1n_1}	x_{2n_2}	...	x_{in_i}	...	x_{kn_k}
medias	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_i$	...	$\bar{x}_k$
varianzas	s_1^2	s_2^2	...	s_i^2	...	s_k^2

Fundamentos del ANOVA (2)

El ANOVA se basa en la comparación de la variabilidad media que hay **entre** los grupos con la que hay **dentro** de los grupos. ¿Por qué? Recordemos que la media y la varianza muestral verifican

$$\text{var}(\bar{x}) = \frac{\sigma^2}{n}, \quad E(s^2) = \sigma^2,$$

lo que nos permite dos estimaciones diferentes para σ^2 cuando disponemos de k muestras de una misma población,

$$\hat{\sigma}^2 = \overline{s^2} = \frac{1}{k} \sum_{i=1}^k s_i^2 \quad (1) \quad \text{y} \quad \hat{\sigma}^2 = ns_x^2 = \frac{n}{k-1} \sum_{i=1}^k (\bar{x}_i - \bar{\bar{x}})^2 \quad (2)$$

Fundamentos del ANOVA (3)

En la tabla anterior, si las k muestras provienen de la misma población todas la medias son iguales, ($\mathbf{H}_0$ es cierta) y tanto (1) como (2) son válidos.

¿Qué ocurre cuando las medias no son iguales? Si suponemos que $\mu_i = \mu + \alpha_i$ entonces

$$ns_{\frac{2}{x}} = \hat{\sigma}^2 + \frac{1}{k-1} \sum_{i=1}^k n_i \hat{\alpha}_i^2 \quad (3)$$

Obsérvese que (1) describe la variabilidad **dentro** de los grupos, mientras que (2) y (3) describen la variabilidad **entre** los grupos.

El test del ANOVA

Si la observación j -ésima del grupo i es de la forma

$$X_{ij} = \mu_i + \varepsilon_{ij}, \quad \text{con } \mu_i = \mu + \alpha_i,$$

las hipótesis

$$\mathbf{H}_0: \alpha_i = 0, \quad \forall i \iff \mu_i = \mu, \quad \forall i,$$

frente a

$$\mathbf{H}_1: \text{algún } \alpha_i \neq 0 \iff \text{las } \mu_i \text{ son distintas,}$$

se contrastan mediante el cociente de varianzas

$$F_0 = \frac{ns_x^2}{s^2} = \frac{\hat{\sigma}^2 + \frac{1}{k-1} \sum_{i=1}^k n_i \hat{\alpha}_i^2}{\hat{\sigma}^2} \quad (4).$$

Comparación de varianzas

Si $X_1, X_2, \dots, X_m$ y $Y_1, Y_2, \dots, Y_n$ son muestras independientes de $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, respectivamente, un test para comparar la igualdad de varianzas se basa en que el cociente corregido de varianzas muestrales,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2},$$

se distribuye como una $F_{m-1, n-1}$, una F de Fisher con $m - 1$ gl en el numerador y $n - 1$ gl en el denominador.

La tabla ANOVA

Todo se reduce a obtener el valor del estadístico (4) que bajo las condiciones iniciales de *independencia*, *normalidad* y *homocedasticidad*, se distribuye como una $F_{k-1, n-k}$. La comparación con el valor teórico correspondiente nos dirá si debemos aceptar o rechazar H_0 .

Un método computacional conocido como **tabla ANOVA** facilita los cálculos. Se trata de disponer en forma de tabla ciertas cantidades que conducen a la obtención de F . El método está incorporado en los paquetes estadísticos más habituales.

Estructura de la tabla ANOVA

	SS	GL	MS	F
Entre	$SS_E = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2$ $= \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$	$k - 1$	$MS_E = \frac{SS_E}{k-1}$	$\frac{MS_E}{MS_D}$
Dentro	$SS_D = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ $= \sum_{i=1}^k (n_i - 1) x_i^2$	$n - k$	$MS_D = \frac{SS_D}{n-a}$	
Total	$SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	$n - 1$		

Análisis de los datos de los tenores

En primer lugar habrá que comprobar que los requisitos iniciales se cumplen.

- La *independencia* es el investigador quien debe garantizarla mediante una adecuada toma de muestras.
- La *normalidad* debe contrastarse mediante un test apropiado, por ejemplo el de Kolmogorov-Smirnov.
- La *homocedasticidad* debe también contrastarse con un test apropiado: el de la F o el de Levene.

Datos de los tenores: Flujo SPSS

- 1) GET DATA /TYPE=XLS
 /FILE='C:\directorio\datos'
 /CELLRANGE=full
 /READNAMES=on .
- 2) SORT CASES BY voz .
 SPLIT FILE .
 SEPARATE BY voz .
- 3) NPAR TESTS
 /K-S(NORMAL)= estatura
 /MISSING ANALYSIS.
- 4) ONEWAY
 estatura BY voz
 /STATISTICS DESCRIPTIVES HOMOGENEITY
 /MISSING ANALYSIS .

Datos de los tenores: Prueba de Normalidad

Prueba de Kolmogorov-Smirnov para una muestra ^c		
		ESTATURA
N		21
Parámetros normales ^{a,b}	Media	1,7695
	Desviación típica	,05971
Diferencias más extremas	Absoluta	,210
	Positiva	,210
	Negativa	-,181
Z de Kolmogorov-Smirnov		,961
Sig. asintót. (bilateral)		,315

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

c. VOZ = 2

Datos de los tenores: Comparación de medias

Prueba de homogeneidad de varianzas

ESTATURA

Estadístico de Levene	gl1	gl2	Sig.
2,070	3	103	,109

ANOVA

ESTATURA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	,057	3	,019	4,198	,008
Intra-grupos	,469	103	,005		
Total	,527	106			

Datos de los tenores: Resultados

Como el p-valor asociado al valor de F encontrado es menor que $\alpha = 0,05$, rechazamos la hipótesis nula y aceptamos que **hay diferencias entre las medias** de las estaturas de los 4 grupos. Si hubiéramos aceptado H_0 no habría más que añadir, pero en este caso las preguntas a intentar responder son

¿entre que grupos hay diferencias?

¿es posible establecer conjuntos homogéneos de medias?

Comparaciones múltiples

Las respuestas a las anteriores preguntas exigen llevar a cabo todas las posibles comparaciones dos a dos, de las que hemos huido al principio. La llevaremos a cabo con métodos que corrigen los problemas que surgen con el nivel de significación final.

ONEWAY

estatura BY voz

/STATISTICS HOMOGENEITY

/MISSING ANALYSIS

/POSTHOC = TUKEY SCHEFFE ALPHA(.05).

Resultados de las comparaciones múltiples

ESTATURA				
VOZ		N	Subconjunto para alfa = .05	
			1	2
HSD de Tukey	T1	21	1,7510	
	T2	21	1,7695	1,7695
	B1	39	1,7969	1,7969
	B2	26		1,8146
	Sig.		,081	,090
Scheffé	T1	21	1,7510	
	T2	21	1,7695	1,7695
	B1	39	1,7969	1,7969
	B2	26		1,8146
	Sig.		,128	,140

Se muestran las medias para los grupos en los subconjuntos homogéneos.

$$\begin{array}{cccc} & \text{subgrupo 1} & & \\ \hline & \text{T1} & \text{T2} & \text{B1} & \text{B2} \\ & & \text{subgrupo 2} & & \end{array}$$

Violación de las condiciones previas

La *normalidad* previa de los datos es de menor importancia que la exigencia de *homocedasticidad*, porque el Teorema Central del Límite implica que las medias muestrales deben ser aproximadamente normales.

Si la igualdad de varianzas no puede ser asumida, es posible llevar a cabo una transformación de los datos para conseguirla. Lo ilustraremos con un ejemplo

Veneno, tratamiento y tiempo de supervivencia

La tabla recogen el tiempo de supervivencia (en unidades de 10 horas) de animales que fueron envenenados y se les administró un tratamiento.

<i>Veneno</i>	<i>Tratamientos</i>			
	1	2	3	4
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Heterocedasticidad

Definimos un factor único, *Doce*, y contrastamos la homogeneidad de las varianzas.



Prueba de homogeneidad de varianzas

T. SUPERV

Estadístico de Levene	gl1	gl2	Sig.
4,854	11	36	,000

Estimación de la potencia

La homogeneidad de varianzas puede conseguirse mediante una transformación de los datos del tipo $y = x^\lambda$. El valor de λ nos lo proporciona el *Gráfico de Estimación de la potencia* del procedimiento *Explorar* en el SPSS.

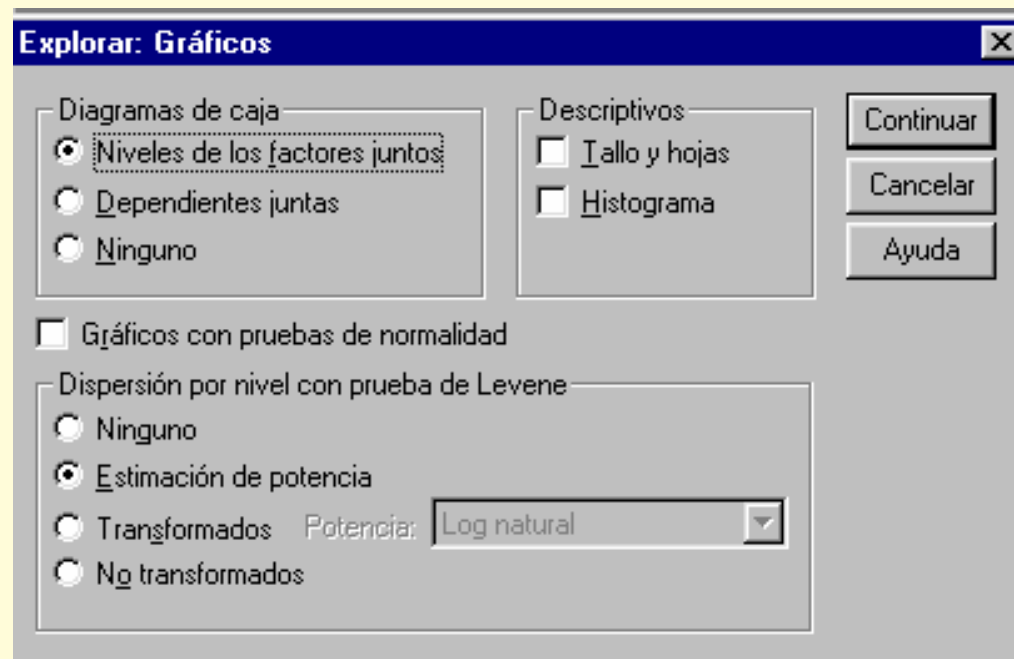
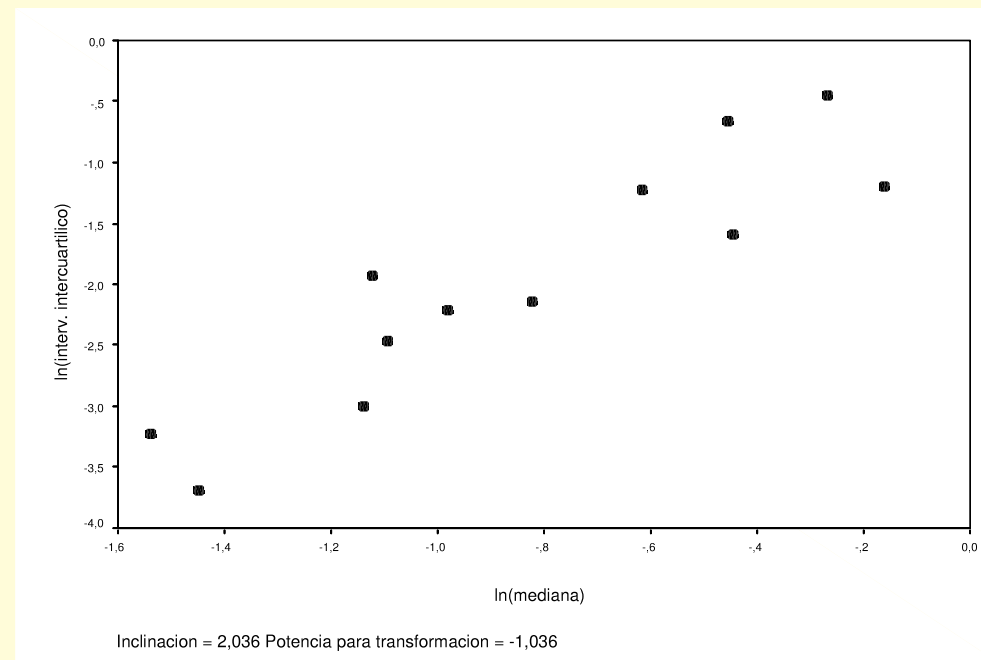


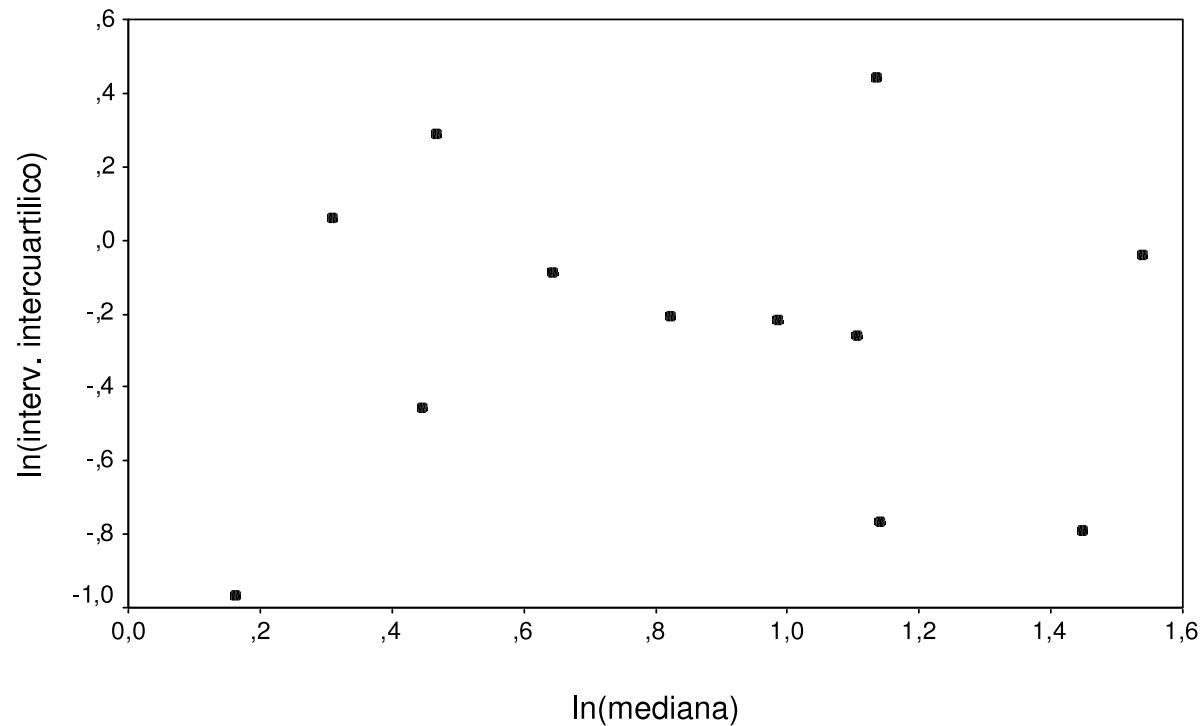
Gráfico de la estimación de la potencia

El resultado es el gráfico, que nos indica la transformación a realizar, $\lambda \approx -1 \Rightarrow y = 1/x$, se trataría por tanto de un *índice de morbilidad*.



Índice de morbilidad y homocedasticidad

Definida la nueva variable,



Inclinacion = ,016 Potencia para transformacion = ,984

Homogeneidad de varianzas

MORB

Estadístico de Levene	1,576
gl1	11
gl2	36
Sig.	,148

Una alternativa no paramétrica: el test de Kruskal-Wallis

En ocasiones las condiciones de normalidad y homocedasticidad no pueden asumirse. El test de Kruskal-Wallis es una alternativa no paramétrica al ANOVA.

Máquina		
1	2	3
10	21	11
6	8	9
12	22	13
17	22	11
16	24	13
18	28	12
3	11	11
11	19	12
16		
7		

Piezas defectuosas por cada 10000 unidades producidas por tres máquinas distintas

Un ejemplo del test de Kruskal-Wallis

NPAR TESTS

/K-W=piezas BY máquina(1 3)

/MISSING ANALYSIS.

Rangos

	MÁQUINA	N	Rango promedio
PIEZAS	1	10	10,80
	2	8	19,25
	3	8	11,13
	Total	26	

Estadísticos de contraste^a

	PIEZAS
Chi-cuadrado	6,600
gl	2
Sig. asintót.	,037

a. Prueba de Kruskal-Wallis

Equivalencia entre el test t y el ANOVA (1)

Si los grupos a comparar son 2 podemos recurrir al test de la *t de Student* y al ANOVA. ¿Podemos encontrar resultados diferentes? No porque entre las distribuciones *t de Student* y *F de Fisher* existe la siguiente relación

$$F_{1,\nu} = t_{\nu}^2,$$

siendo $\nu = n_1 + n_2 - 2$ el número de grados de libertad de la *t de Student* o del denominador de la *F de Fisher*.

Equivalencia entre el test t y el ANOVA (2)

Veamos la comparación de la estaturas del Tenor 1 y Tenor 2.

Prueba de muestras independientes

ESTATURA	t	-,831
	gl	40
	Sig. (bilateral)	,411

ANOVA

ESTATURA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	,004	1	,004	,690	,411
Intra-grupos	,210	40	,005		
Total	,213	41			