# F de Fisher

## Análisis de la varianza

Vicente Manzano Arrondo – 2014

## Situación de partida

Preguntamos a un grupo de cien personas, aficionadas al fútbol, qué les ha parecido la preparación del último partido en el que ha jugado la selección española. Uno de los items del cuestionario se refiere a la alineación. Las personas entrevistadas responden mediante un formato Likert de siete puntos, donde 1 implica máximo desacuerdo y 7 máximo de acuerdo con la afirmación “La alineación de los jugadores en el campo fue la correcta”. Los resultados los tenemos aquí:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 7 | 6 | 5 | 1 | 3 | 3 | 1 | 2 | 6 |
| 3 | 3 | 6 | 4 | 2 | 1 | 4 | 3 | 3 | 1 |
| 3 | 4 | 4 | 4 | 5 | 4 | 1 | 3 | 7 | 4 |
| 4 | 1 | 5 | 1 | 5 | 6 | 6 | 4 | 3 | 2 |
| 6 | 4 | 2 | 6 | 1 | 6 | 7 | 7 | 3 | 4 |
| 3 | 5 | 3 | 5 | 1 | 6 | 4 | 5 | 6 | 3 |
| 2 | 6 | 4 | 5 | 7 | 3 | 7 | 3 | 7 | 2 |
| 1 | 1 | 3 | 5 | 3 | 7 | 4 | 4 | 4 | 6 |
| 1 | 7 | 7 | 6 | 1 | 1 | 2 | 7 | 4 | 7 |
| 1 | 7 | 3 | 3 | 3 | 5 | 6 | 6 | 2 | 6 |

A simple vista, observamos la presencia de todas las posibles respuestas. Hay una variación evidente. Para iniciar un conocimiento más completo de la variable, realizamos una tabla de frecuencias. Observarás resaltada la suma de cuadrados, es decir, la suma de las distancias cuadráticas a la media. Está resaltada porque nos será muy útil más adelante. De momento, veamos qué nos dice la tabla y la representación numérica:

| X | f | Xf | d2f |
|---|---|---|---|
| 1 | 15 | 15 | 135 |
| 2 | 8 | 16 | 32 |
| 3 | 20 | 60 | 20 |
| 4 | 18 | 72 | 0 |
| 5 | 10 | 50 | 10 |
| 6 | 16 | 96 | 64 |
| 7 | 13 | 91 | 117 |
| | 100 | 400 | 378 |

$$\bar{X} = \frac{\sum X}{n} = \frac{400}{100} = 4$$

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} = \sqrt{\frac{378}{100}} = 1{,}94$$

En efecto, parece que se observa un poco de todo, con media en el valor 4 y una desviación tipo del 50% aproximadamente, lo que señala una variación importante. Si el seleccionador nacional de fútbol nos preguntara por la opinión de este grupo, no tendríamos nada más interesante que decirle diferente a “se observa un poco de todo, con una opinión media de valor 4, es decir, bastante media”.

## ¿Por qué hay variación?

Las opiniones o impresiones de las personas entrevistadas respecto a la decisión de alineación del equipo de fútbol son un ejemplo de conducta o comportamiento humano. En psicología nos preocupa explicar lo que ocurre, que es casi tanto como intentar comprender por qué hay personas que se comportan de un modo y personas que se comportan de otro. En términos estadísticos: por qué hay variabilidad.

Hemos preguntado también a estas cien personas cuál es su equipo de fútbol preferido. Sus respuestas tal vez sean de ayuda para comprender la variabilidad en su grado de acuerdo con la alineación. Tal vez no. Solo hay una forma de saberlo: comprobarlo. Así que vamos a agrupar los cien datos según el equipo preferido de quien responde. Ocurre que solo hay tres equipos mencionados: el San Jerónimo United, el Cánovas Fútbol Party y el Trini & Jonás, todos muy castizos como puede deducirse. Al agrupar los datos según el club de fútbol preferido, obtenemos lo que sigue:

| San Jerónimo United | | | Cánovas Fútbol Party | | | Trini & Jonás | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 3 | 4 | 3 | 3 | 4 | 7 | 6 |
| 1 | 1 | 3 | 5 | 4 | 4 | 6 | 5 | 6 |
| 1 | 1 | 4 | 4 | 3 | 3 | 7 | 5 | 7 |
| 2 | 1 | 1 | 3 | 4 | 4 | 7 | 6 | 7 |
| 4 | 2 | 4 | 6 | 4 | 5 | 7 | 7 | 6 |
| 1 | 2 | 2 | 5 | 3 | 5 | 7 | 4 | 4 |
| 1 | 1 | 2 | 3 | 6 | 4 | 6 | 6 | 5 |
| 1 | 1 | 3 | 6 | 3 | 3 | 4 | 7 | 5 |
| 3 | 3 | 2 | 3 | 3 | 5 | 7 | 7 | 6 |
| 4 | 1 | 4 | 6 | 3 | 3 | 6 | 7 | 7 |
| 1 | 1 | 2 | 3 | 3 | 6 | 6 | 5 | 6 |
| | | | 4 | | | | | |

La primera inspección visual añade ya información. Podemos observar que SJU muestra datos con valores más bajos que el resto, CFP cuenta con valores intermedios y TJ acapara los más altos. No es una conclusión inmediata, requiere mirar con un poco de detenimiento. Con sus respectivas tablas de frecuencia, será más evidente esta afirmación:

San Jerónimo United

| X | f | Xf | d2f |
|---|---|---|---|
| 1 | 15 | 15 | 15 |
| 2 | 8 | 16 | 0 |
| 3 | 5 | 15 | 5 |
| 4 | 5 | 20 | 20 |
| | 33 | 66 | 40 |

$$\bar{X} = \frac{\sum X}{n} = \frac{64}{32} = 2$$

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} = \sqrt{\frac{40}{33}} = 1{,}101$$

Cánovas Fútbol Party

| X | f | Xf | d2f |
|---|---|---|---|
| 3 | 15 | 45 | 15 |
| 4 | 9 | 36 | 0 |
| 5 | 5 | 25 | 5 |
| 6 | 5 | 30 | 20 |
| | 34 | 136 | 40 |

$$\bar{X} = \frac{\sum X}{n} = \frac{136}{34} = 4$$

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} = \sqrt{\frac{40}{34}} = 1{,}08$$

Trini & Jonás

| X | f | Xf | d2f |
|---|---|---|---|
| 4 | 4 | 16 | 16 |
| 5 | 5 | 25 | 5 |
| 6 | 11 | 66 | 0 |
| 7 | 13 | 91 | 13 |
| | 33 | 198 | 34 |

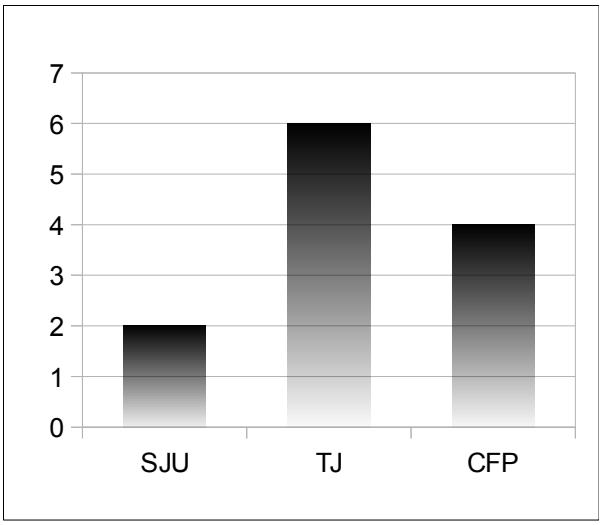
$$\bar{X} = \frac{\sum X}{n} = \frac{192}{32} = 6$$

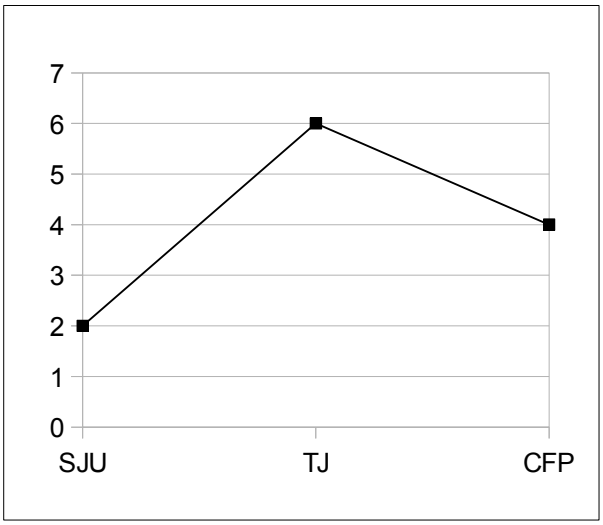
$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} = \sqrt{\frac{34}{33}} = 1{,}015$$

Veamos qué nos dice este análisis por cada grupo separado. Lo primero que llama la atención es que las medias aritméticas tienen valores sensiblemente diferentes, coherentes a su vez con lo que observamos en cada tabla de frecuencias específica de cada grupo. Los valores de desviación tipo de cada grupo son relativamente pequeños (poco más de 1). En pocas palabras: al distinguir el grupo de pertenencia de cada caso, vemos que hay una marcada variación entre los grupos y una relativamente baja variación dentro de los grupos. De algún modo, ahora podemos comprender mejor la conducta "grado de acuerdo con la alineación": en cierta medida se entiende a partir del club de fútbol preferido por quien responde. Quienes menos acuerdo muestran son los de San Jerónimo United. Quienes más, los de Trini & Jonás. Quizá, si supiéramos de dónde proceden los jugadores de la selección nacional, accederíamos a una explicación razonable. Es posible que la alineación cuente con más procedencias de TJ que de SJU y esto complazca o disguste, respectivamente, a quienes opinan. Es una suposición, puesto que no tenemos información que corrobore esta arriesgada hipótesis. En cualquier caso, por lo que llevamos de análisis de los datos, parece que algo tiene que ver el club de referencia frente a la opinión.

## Representación gráfica

Antes de aprovechar este comportamiento de la variación para generar una cuantía útil que nos permita concluir algo sobre qué relación existe entre el grado de acuerdo y el club de fútbol de referencia, veamos qué recurso gráfico es una buena herramienta para investigar esa relación. De entrada, las tablas de frecuencia total y por grupos nos han sido útiles. Vamos a ver cómo llegar a conclusiones similares con un instrumento gráfico.

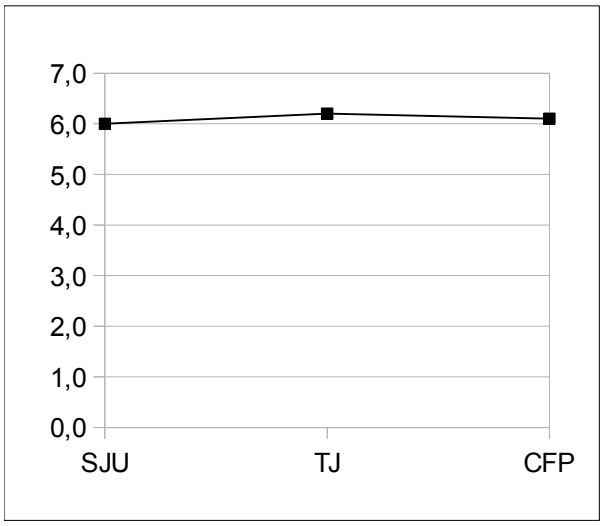


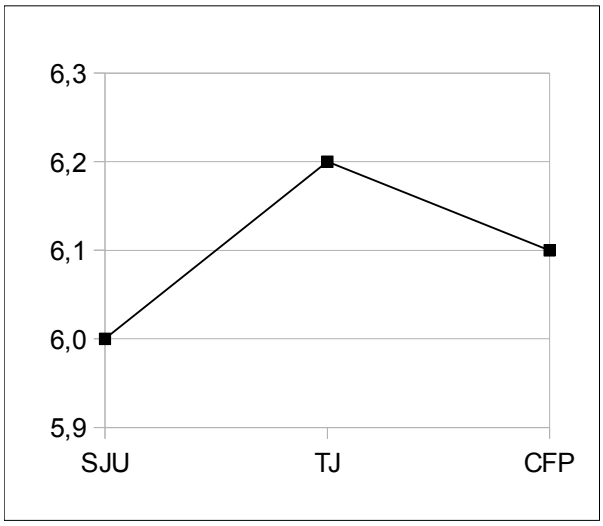



Un primer recurso, muy intuitivo, es representar los valores de las medias aritméticas. Tenemos varias posibilidades, una de ellas puede consistir en acudir a un diagrama de barras adaptado: en lugar de que la altura de las barras represente la frecuencia de un valor, ahora va a representar la media aritmética de un grupo de datos. Lo llamamos *gráfico o diagrama de medias*. En lugar de utilizar barras, podríamos acudir a puntos unidos por líneas. Veamos ambas posibilidades.

He procurado cambiar el orden con que estamos estudiando los equipos para no dar la sensación de línea recta o relación lineal. Observa que la variable “club de fútbol preferido” es nominal. El orden con que se presenten sus valores es irrelevante.

En nuestro ejemplo, la diferencia entre las medias aritméticas es muy sobresaliente. La representación gráfica muestra este hecho con claridad. Es frecuente que esta diferencia sea más sutil. En tales casos, los programas de ordenador suelen *exagerar* la conclusión de diferencia. No es fruto de una mala intención, sino un efecto secundario del principio “resaltar la zona significativa”. Imagina que las medias tuvieran los valores 6, 6,2 y 6,1 respectivamente. Como resulta evidente, son diferencias muy pequeñas. Observa dos versiones de los mismos datos representados mediante un diagrama de líneas.



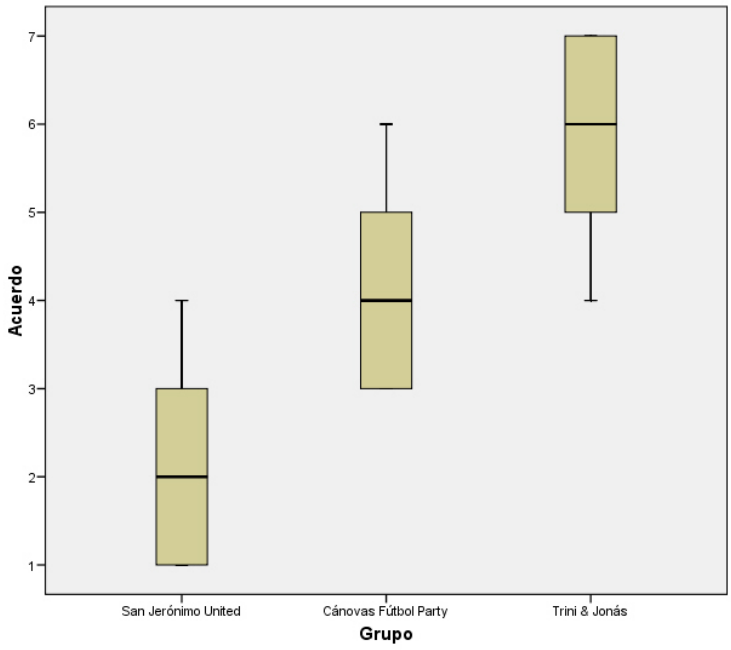




Para no errar en la interpretación sobre las diferencias entre las medias, es necesario que el origen del eje vertical se sitúe en cero. La primera representación gráfica permite concluir de forma correcta: los puntos se encuentran prácticamente a la misma altura, por lo que no parece existir diferencia significativa entre los tres grupos. En la segunda representación, las diferencias son muy apreciables y tenderíamos a concluir que el grado de acuerdo varía entre los diferentes grupos. Al observar el eje vertical, vemos que no comienza en cero sino en 5,9. Lo que ha hecho la gráfica es resaltar la zona significativa, eliminando la zona de la gráfica que no parece añadir nada porque no contiene información específica. Sin embargo, sí añade mucho: permite relativizar la diferencia, comparándola con un referente absoluto. Para una buena conclusión hemos de partir siempre de una representación que comience a caminar desde el inicio, es decir, con un origen vertical en cero.

El diagrama de medias es intuitivo y fácil de interpretar, pero utiliza una información muy incompleta: sólo considera la cuantía de las medias. Ya sabemos de la primera unidad de aprendizaje (conocer una variable) que toda representación numérica ha de ir acompañada de una medida de bondad de la representación. En ocasiones, lo más característico es la variación, más que un valor supuestamente representativo. Si los

grupos varían mucho dentro de sí, la variación que guarden entre sí pierde importancia. En una representación gráfica, esta circunstancia podría resolverse si se incluye, junto con la media, un valor de variación. En lugar de complicar más esta gráfica, contamos ya con un recurso que contiene más información y que tiene su razón de ser en situaciones como esta. Me estoy refiriendo al diagrama de caja y patillas.

Hemos conocido el diagrama de caja y patillas en el caso de representar una variable cuantitativa. Vamos a utilizarlo ahora para observar en qué medida hay diferencias entre varios grupos respecto a una variable cuantitativa. Para ello, cada grupo va a generar un diagrama de caja y patillas y lo calificaremos como “combinado”. Y los tres estarán juntos compartiendo el mismo espacio. Esto facilitará la interpretación.




Al observar el diagrama de caja y patillas combinado para los datos de nuestro ejemplo, las conclusiones están muy claras. No solo los valores representativos (las tres medianas, indicadas por un trazo grueso dentro de cada caja) son muy diferentes, sino que los valores más característicos (el 50% central, representado por cada caja) se encuentran en franjas diferentes, no se solapan. Esta ausencia de solapamiento entre las cajas constituye un criterio de peso a la hora de sentenciar diferencias significativas entre los grupos. Las patillas no añaden más información sobre esta circunstancia. Lo que hacen es matizar en términos de simetría. Como puede verse, San Jerónimo y Cánovas muestra asimetría positiva (dispersión de datos en la zona de los valores altos), mientras que Trini&Jonás cuenta con asimetría negativa (dispersión en la zona de los valores bajos).

## Relación entre variables

Lo que estamos haciendo es estudiar la relación entre una variable categórica y una variable cuantitativa. La categórica forma grupos de datos cuantitativos. En nuestro ejemplo, contamos con la variable “club preferido de fútbol”, con tres valores, niveles o categorías. Para cada una de ellas, existe un grupo específico de datos cuantitativos: los valores de la otra variable, el grado de acuerdo con la afirmación de que la alineación de los jugadores es la correcta.

El recurso que estamos utilizando para estudiar la relación entre una variable categórica y una cuantitativa pasa por llevar a cabo un estudio gráfico (diagrama de medias o diagrama de caja y patillas combinado) y observar cómo se comporta la variación: si hay una marcada diferencia entre los grupos y, a la vez, poca variación dentro de los grupos, entonces existe relación. Nos ocuparemos más despacio de este juego de variaciones. Ahora sigamos con el asunto de la relación. Nos ocupan dos aspectos: qué estamos entendiendo por variable formadora de grupos y de qué tipo de relación estamos hablando.

La variable categórica o formadora de grupos será habitualmente una variable nominal, como ocurre con nuestro ejemplo. No tiene por qué ser siempre así. Podemos utilizar una variable ordinal. Es cierto que si la ordinal cuenta con suficientes valores y el sistema de medida nos da confianza como para tratarla de cuasicuantitativa, tal vez lo más idóneo sea acudir a un diagrama de dispersión y un coeficiente de correlación lineal simple de Pearson, es decir, abordar la relación como la que se establece entre dos variables cuantitativas. Esto sería inicialmente correcto. No obstante, recuerda que aplicar r de Pearson exige que la relación que exista entre ambas variables, sea mucha o poca, debe ser del tipo “lineal”, es decir, representable mediante una línea recta. Si la representación gráfica indica que no es lineal, una posible opción es acudir al recurso que estamos viendo aquí, aunque las dos variables sean cuantitativas o cuasicuantitativas. El único requisito es que el número de valores de la variable formadora de grupos no sea excesivo. Si contamos con muchos grupos de datos (es decir, muchos valores de media aritmética, muchas cajas y patillas, etc.), la decisión deja de ser operativa.

El segundo asunto relevante para este apartado es qué estamos entendiendo por una relación. Ocurre que el recurso gráfico (y, como veremos seguidamente, el numérico) se ocupa de estudiar en qué medida existe diferencia entre los grupos de datos, en comparación con la variación que se observa dentro de los grupos. En el diagrama de caja y patillas combinado, por ejemplo, resulta más llamativa la diferencia entre las cajas que no la dispersión dentro de cada grupo, de tal forma que las cajas no llegan a solaparse.

Es importante no pedirle a las técnicas lo que las técnicas no pueden dar. En este caso, lo que estamos haciendo es abordar la diferencia entre los grupos. Si existe suficiente diferencia (es decir, si comparando la dispersión “entre” con la dispersión “dentro”, la primera es más característica que la segunda), entonces diremos que existe relación entre ambas variables, pero no que una causa a la otra. Recuerda que son posibles muchos tipos de relación: directa, indirecta, espúrea... Que observemos relación entre la variable categórica y la cuantitativa no implica que la primera cause a la segunda. Esta confusión es particularmente esperable cuando utilizamos las denominaciones *variable independiente* y *variable dependiente*. Habitualmente, la v.i. forma grupos y la v.d. suministra mediciones cuantitativas. Aunque sigamos utilizando esta denominación aquí, es decir, aunque llamemos v.i. a la variable formadora de grupos o categórica, y v.d. a la cuantitativa, no olvidemos que no tienen realmente por qué ser v.d. y v.i. en términos de diseño, es decir, no tienen por qué provenir de un experimento o un cuasi-experimento en que los sujetos experimentales son agrupados según la v.i. y tras una situación bajo control se miden los valores que suministran de la v.d. No podemos suponer, únicamente por la técnica o procedimiento de análisis, que una variable es causa o ejerce influencia sobre los valores de la otra.

## Razón entre varianzas y sumas de cuadrados

En línea con lo que estamos desarrollando hasta llegar aquí, el recurso cuantitativo va a consistir en comparar la variación *entre* con la variación *dentro*. Se podría hacer de varios modos. El agrónomo, matemático, etc. británico Sir Ronald Fisher, fue quien ideó el procedimiento de cuantificación que vemos en este apartado. Fisher planteó calcular una razón entre ambas medidas de variación, es decir, dividir la variación *entre* ($V_E$) por por variación *dentro* ($V_D$). El resultado, aunque parezca asombroso, se simboliza con la F de Fisher:

$$F = \frac{V_E}{V_D}$$

Si existe relación entre ambas variables (categórica y cuantitativa), entonces cabe esperar que haya más variación entre los grupos que dentro de ellos. En otros términos: que haya más variación explicada o comprendida gracias a la variable que forma los grupos, que no variación sin explicar o sin comprender, que sigue existiendo dentro de los grupos. Si $V_E$ es mayor que $V_{ED}$, entonces, el numerador tiene un valor mayor que el denominador y F > 1. Si $V_E$ es inferior a $V_D$, entonces F < 1. Luego, la medida que manejamos para dictaminar sobre la relación es si F es o no mayor que 1.

El análisis de datos que se lleva a cabo para aterrizar en F y tomar posteriormente una decisión sobre la relación, se denomina *Análisis de la Varianza*, un título fácil de entender. No obstante, las variaciones entre y dentro no son *exactamente* varianzas, sino *casi* varianzas, es decir, *cuasivarianzas*. Vamos a verlo.

Recuerda la fórmula de cálculo de una varianza. Es una media de distancias cuadráticas, es decir una suma de distancias a la media elevadas al cuadrado, dividida entre el número de distancias. El numerador (suma de distancias al cuadrado) suele conocerse como *suma de cuadrados* (SC), algo más breve. Pues bien:

$$S^2 = \frac{\sum (X - \bar{X})^2}{n} = \frac{SC}{n}$$

En el ejemplo que estamos siguiendo, SC = 378. Se refiere a la que hay en total, es decir, sin distinguir entre grupos. Por ese motivo, se la conoce como *suma de cuadrados total* y se simboliza con $SC_T$. Como se refiere a las distancias cuadráticas de todos los valores respecto a la media de todos los valores (es decir, media total), podemos escribirla así:

$$SC_T = \sum (X - \bar{X}_T)^2 = 378$$

La expresión de la variación que existe dentro de los grupos en términos de suma de cuadrados ($SC_D$), será la suma de las sumas de cuadrados de todos los grupos. Es decir:

$$SC_D = \sum \sum (X - \bar{X}_{Di})^2 = SC_{D1} + SC_{D2} + SC_{D3} = 40 + 40 + 34 = 114$$

En total tenemos 378 unidades cuadráticas. Dentro de los grupos contamos con 114. ¿Cuál será la que se corresponda con la variación *entre*? Lógicamente, ha de ser lo que queda. Toda la variación se encuentra dentro de los grupos o entre los grupos, sin posibilidades de encontrar variación fuera de esta clasificación. Es decir: lo total será la suma de lo *entre* más lo *dentro*:

$$SC_T = SC_E + CS_D$$

Por tanto:

$$SC_T = SC_E + CS_D \rightarrow SC_E = SC_T - SC_D = 378 - 114 = 264$$

Si calculamos directamente la variación *entre*, la lógica es la misma que en cualquier caso de variación que estamos siguiendo. Cada grupo está representado por su media aritmética. Contamos entonces con un conjunto de medias. La variación entre las medias se obtiene calculando las distancias cuadráticas de cada media de grupo respecto a la media total. Tomado esto literalmente, calcuaríamos:

$$\sum (\bar{X}_i - \bar{X}_T)^2$$

La idea es buena, pero incorrecta. Falla en un detalle. Ten en cuenta que los grupos pueden contar con tamaños muy diferentes. Imagina que contáramos con un grupo de mil personas y otro con diez. No es correcto que ambos grupos *pesen* lo mismo. Lo que hacemos es *ponderar*esta suma: cuanto más datos contenga el grupo, más pesa en la suma de cuadrados. Es decir:

$$SC_E = \sum (\bar{X}_i - \bar{X}_T)^2 n_i$$

Con los datos del ejemplo:

$$SC_E = (2-4)^2\ 33 + (4-4)^2\ 34 + + (6-4)^2\ 33 = 264$$

Como no podía ser de otro modo, el cálculo coincide con el que hemos hecho más arriba.

Antes de seguir leyendo, vuelve a leer este apartado hasta que la lógica te resulte comprensible y no haya problemas. Si continúas de forma precaria, tendremos problemas...

## Grados de libertad y *casi* varianzas

Los grados de libertad son los grados de libertad. ¿Curioso, verdad? Cuantos más grados de libertad, más libertad. Pero ¿más libertad para qué o de qué? Para n levantar demasiadas expectativas, hay que recordar que nos encontramos en un contexto estadístico y que, por tanto, no hay que hacerse muchas ilusiones. “Más libertad” se refiere a más posibilidades para realizar cambios en los datos sin que un resultado final se vea modificado.

Un ejemplo: con los datos 4, 7 y 8 puedo calcular una suma (4+7+8=19) o una multiplicación (4·7·8=224). Podemos jugar a cambiar números sin que se modifique el resultado. Cambio el 4 por un 14, porque me da la gana (a veces hay otras razones). Por la misma razón u otra más convincente, cambio el 7 por un 8. Y se acabó. Si quiero que la suma siga siendo 19, el tercer dato no puede tener el valor que yo quiera, debe ser necesariamente -3, pues:

$$4 + 7 + 8 = 19 = 14 + 8 - 3$$

Si quiero que la multiplicación siga dando como resultado el valor 224, no puedo escoger como tercer dato a cualquier cantidad. Debe ser necesariamente 2:

$$4 \cdot 7 \cdot 8 = 224 = 14 \cdot 8 \cdot 2$$

Los grados de libertad podrían pensarse como una aplicación del principio “el último paga el pato”, famoso en castellano. En términos generales, si una variable cuenta con k valores, entonces disponemos de k-1 grados de libertad.

¿Cómo se mastican los grados de libertad en el caso del análisis de la varianza? Bien, pues cada componente de la variación (total, entre, dentro) tiene su correspondiente valor en grados de libertad, siempre con el mismo principio de que el último paga el pato. Para aclararnos, *n* simboliza el número total de datos, $n_i$ al número de datos del grupo *i*, mientras que *k* se refiere al número de grupos. En tal caso, los grados de libertad (gl) van a ser:

$$gl_T = n - 1 \qquad gl_E = k - 1 \qquad gl_D = n - k \qquad gl_T = gl_E + gl_D$$

Los grados de libertad *dentro* tal vez requiera un poco más de explicación: ¿por qué se resta k y no 1 como siempre? Observa que los grados de libertad *dentro* deben surgir del mismo modo que la suma de cuadrados *dentro*: es la suma de los grados de libertad de cada uno de los k grupos. Si restamos 1 en cada grupo, lo restamos k veces. Por eso, al final, lo que restamos es k.

En el ejemplo:

$$gl_T = 100 - 1 = 99 \qquad gl_E = 3 - 1 = 2$$

$$gl_D = 100 - 3 = 97 \qquad gl_T = gl_E + gl_D = 2 + 97 = 99$$

Los grados de libertad no solo constituyen una aplicación filosófica (el último paga el pato) y un entretenimiento aritmético. Tienen una importancia específica en el análisis de la varianza.

La varianza es una media de distancias cuadráticas. La media se calcula como todas las medias: suma de elementos, entre el número de elementos. En el caso de una media de distancias cuadráticas: suma de cuadrados entre número de datos.

En próximos temas abordaremos las situaciones en las que nos interesa algo sobre una población pero no trabajamos directamente con ella sino con una muestra. En tal caso llevamos a cabo inferencia, es decir, una estimación sobre lo que habríamos encontrado en la población en el caso de que hubiéramos trabajado directamente con ella. Como veremos, la base de nuestro sistema de inferencia es la estimación estadística. En una estimación, tomamos un índice o medida calculada en la muestra y la utilizamos como *estimador*, es decir, como un artilugio que nos permite estimar cuál sería el valor de ese índice o medida de haberlo calculado en la población. No nos asombrará, cuando llegue el caso, que un buen estimador de la media de la población es la media de la muestra. Pero tal vez nos asombre que un buen estimador de la varianza de la población no es la varianza de la muestra, sino algo muy parecido, algo que es *casi* la varianza y que, por ello, recibe el nombre de *cuasivarianza*.

La cuasivarianza es casi como la varianza porque comparte el mismo numerador (suma de cuadrados), pero el denominador no es idéntico sino parecido: no es el número de elementos, sino los grados de libertad. Así, resulta que el análisis de la varianza es realmente un análisis de las cuasivarianzas. También podríamos pensar en estos términos: es un análisis de las varianzas que estimamos observaríamos en la población caso de trabajar con ella en lugar de estar haciéndolo con una muestra. Dado que esta

segunda versión es algo enrevesada, quedémonos con la idea de que estamos realizando un análisis de la varianza, con la matización de que adaptamos las expresiones de cálculo para favorecer una buena inferencia.

## Tabla de análisis de la varianza

Lo avanzado hasta el momento culmina en la tabla que abordamos ahora. Es un recurso para organizar los elementos que necesitamos en el cálculo de la F de Fisher: sumas de cuadrados, grados de libertad, cuasivarianzas y, finalmente, F.

La organización de la tabla parte del concepto *fuente de variación*. Es una expresión muy clara. Recordemos de dónde venimos: una variable de conducta medida en una escala cuantitativa muestra variación (fuente de variación *total*). Nos preguntamos a qué es debido. Resulta que hemos preguntado algo más a la gente, lo tenemos anotado y lo consideramos para el análisis. Al incluir la información de la variable categórica que forma grupos (fuente de variación *entre*-grupos), observamos que parte de la variación total es absorbida por esta fuente categórica, pero no obstante sigue existiendo una variación dentro de cada uno de los grupos considerados (fuente de variación *dentro*-grupos).

Dispongamos todo ello en una tabla genérica y apliquemos después los datos concretos del ejemplo.

| Fuente de variación | Suma de cuadrados | Grados de libertad | Varianza | F |
|---|---|---|---|---|
| Entre | $SC_E$ | $gl_E = k-1$ | $V_E = SC_E / gl_E$ | $F = V_E / V_D$ |
| Dentro | $SC_D$ | $gl_D = n-k$ | $V_D = SC_D / gl_D$ | |
| Total | $SC_T$ | $gl_T = n-1$ | $V_T = SC_T / gl_T$ | |

Realmente la fuente de variación total no es relevante en sí misma y no se necesita su cálculo, salvo para comprobar que tanto las sumas de cuadrados como los grados de libertad se comportan correctamente (la suma *entre* y *dentro* da como resultado el *total*). Con todo ello, veamos qué ocurre con nuestro ejemplo:

| Fuente de variación | Suma de cuadrados | Grados de libertad | Varianza | F |
|---|---|---|---|---|
| Entre | 264 | 2 | 264 / 2 = 132,0 | 132/1,2 = 112,3 |
| Dentro | 114 | 97 | 114 / 97 = 1,2 | |
| Total | 378 | 99 | | |

Debería quedar claro que 112,3 se aleja sensiblemente de 1, por lo que la conclusión de que existe relación es evidente.

O no...

Con la F ocurre algo similar a la Chi cuadrado: tenemos problemas para interpretar su valor porque no se mueve en un intervalo fijo, sino que depende de cada contexto en el que se aplica. En nuestro ejemplo, sabemos que 112,3 es un valor alto, pero no tanto porque en sí lo parezca sino porque ya conocemos estos datos y ha quedado clara la conclusión de relación entre ambas variables. ¿Es F=112,3 siempre un valor elevado? No. Depende de los grados de libertad, es decir, del número total de datos y de grupos.

Para evitar esta indefensión a la hora de interpretar un efecto, en el caso de la Chi cuadrado recurrimos a la V de Cramer, artilugio que nos permite adaptar la Chi al intervalo (0 , 1). Con la F haremos algo similar, incluso más comprensible.

Para medir el tamaño del efecto en un análisis de la varianza, consideramos las sumas de cuadrados: comparar lo que se explica por la existencia de los grupos ($SC_E$) respecto a la variación completa ($SC_T$). Si considerar los grupos no sirve para nada, entonces $SC_E = 0$. Si todo queda explicado por los grupos, es decir, si no queda ninguna variación dentro de ellos, entonces $SC_E = SC_T$. Es decir, $SC_E$ se mueve de 0 a $SC_T$. Si lo dividimos entre $SC_T$, entonces se moverá entre $0/SC_T=0$ y $SC_T/SC_T=1$. He aquí, entonces, un índice acotado en (0 , 1) que recibe el nombre de eta cuadrado y se simboliza con la letra griega del mismo nombre:

$$\eta^2 = \frac{SC_E}{SC_T} = \frac{264}{378} = 0{,}698$$

Un tamaño de efecto del 70% es considerable. Así que corroboramos lo que ya estábamos afirmando: conocer el club de fútbol de referencia o simpatía de cada persona, ayuda a comprender su grado de acuerdo respecto a la alineación de los jugadores.

Para no dejar al libre albedrío esta decisión, recuerda el baremo que estamos utilizando, un criterio para bajar la ansiedad ante situaciones de indefensión, como ocurre con el examen: las cotas 0,1 – 0,3 – 0,5 son las que seguiremos utilizando también en este caso para interpretar el tamaño del efecto de eta cuadrado ($\eta^2$) y por tanto de la relación entre una variable categórica y una cuantitativa. Recuperaremos la F más adelante, cuando nos preocupe realizar inferencia estadística.

## Ejemplo para la tabla de varianzas

Preguntamos a 30 personas por su nivel de expectativa respecto al futuro, donde 1 significa “lo veo bastante negro” y 5 expresa “creo que todo irá estupendamente”. Hemos preguntado en tres colas: del paro, para entrar a un aula universitaria y antes de entrar en una tienda de rebajas. Los datos recogidos son:

| Paro | | |
|---|---|---|
| 2 | 2 | 1 |
| 1 | 4 | 1 |
| 2 | 1 | 3 |
| 3 | | |

| Aula | | |
|---|---|---|
| 1 | 3 | 1 |
| 4 | 2 | 3 |
| 1 | 5 | 4 |
| 3 | 5 | 4 |

| Rebajas | | |
|---|---|---|
| 5 | 2 | 5 |
| 5 | 3 | 3 |
| 5 | 4 | |

Para estudiar la posible relación entre el contexto de la cola y la medida de expectativa sobre el futuro, aplicamos un análisis de la varianza, cuyos cálculos intermedios son:

| Paro | | | |
|---|---|---|---|
| X | f | Xf | d2f |
| 1 | 4 | 4 | 4 |
| 2 | 3 | 6 | 0 |
| 3 | 2 | 6 | 2 |
| 4 | 1 | 4 | 4 |
| | 10 | 20 | 10 |

| Aula | | | |
|---|---|---|---|
| X | f | Xf | d2f |
| 1 | 3 | 3 | 12 |
| 2 | 1 | 2 | 1 |
| 3 | 3 | 9 | 0 |
| 4 | 3 | 12 | 3 |
| 5 | 2 | 10 | 8 |
| | 12 | 36 | 24 |

| Rebajas | | | |
|---|---|---|---|
| X | f | Xf | d2f |
| 2 | 1 | 2 | 4 |
| 3 | 2 | 6 | 2 |
| 4 | 1 | 4 | 0 |
| 5 | 4 | 20 | 4 |
| | 8 | 32 | 10 |

Total

| X | f | Xf | d2f |
|---|---|---|---|
| 1 | 7 | 7 | 26,16 |
| 2 | 5 | 10 | 4,36 |
| 3 | 7 | 21 | 0,03 |
| 4 | 5 | 20 | 5,69 |
| 5 | 6 | 30 | 25,63 |
| | 30 | 88 | 61,87 |

| | Suma | n | Media | SC |
|---|---|---|---|---|
| Paro | 20 | 10 | 2,00 | 10,00 |
| Aula | 36 | 12 | 3,00 | 24,00 |
| Reb. | 32 | 8 | 4,00 | 10,00 |
| | 88 | 30 | | 44,00 |

Luego, $SC_T = 61,87$; $SC_D = 44$ y, por tanto, $SC_E = 61,87 – 44 = 17,87$

La tabla del análisis de la varianza será pues:

| FV | SC | gl | V | F | Eta2 |
|---|---|---|---|---|---|
| Entre | 17,87 | 2 | 8,93 | 5,48 | 0,29 |
| Dentro | 44,00 | 27 | 1,63 | | |
| Total | 61,87 | 29 | | | |

Observamos un efecto pequeño (inferior a 0,3, pero superior a 0,1) del contexto de cola sobre las expectativas de futuro.