FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE GEOMETRÍA Y TOPOLOGÍA

TRABAJO FIN DE GRADO

ENFOQUE GEOMÉTRICO EN EL ANÁLISIS ESTADÍSTICO MULTIVARIANTE

Realizado por:
Mª Dolores Carballar Falcón

Dirigido por:
Dra. Mª Carmen Márquez García
Dra. Desamparados Fernández Ternero

Resumen

El objeto de este proyecto es presentar las matemáticas subyacentes a los métodos estadísticos elementales basados en la distribución normal de la forma más sencilla posible. Los métodos están aplicados a muestras independientes en las que se estudia el Análisis de la Varianza (ANOVA) a través del estadístico T-Student. Toda la teoría se puede aplicar al espacio n-dimensional; sin embargo, para simplificar la dimensión del problema, se han utilizado vectores de dimensiones más reducidas. La estructura de este proyecto tiene dos partes; ambas tratadas desde un punto de vista geométrico. En primer lugar, se describe la teoría del ANOVA en muestras independientes; y en segundo lugar, se estudian métodos denominados Contrastes Ortogonales cuya finalidad es estudiar la media poblacional así como la posible interacción entre grupos de tratamientos.

Abstract

The aim of this project is to present the mathematics underlying elementary statistical methods based on the normal distribution in as simple a manner as possible. The methods we refer to are independent sample t test and analysis of variance. The underlying mathematics is the theory of n-dimensional space. However, here it has been prefered to use the simpler vector geometric methods since these reduce the dimensionality of the problem, provide more visual insight, and make the theory more accesible. The structure of this project has two parts, both of them are related.
Firstly, it has been described the theory of Analysis of Variance (ANOVA) in independent samples from a geometric approach. Secondly, it has been studied some methods called Orthogonal Contrasts, whose object of interest is to get estimates of main and interaction effects, for mean comparisons between groups of data. This second part has been treated from a geometric approach too.

Índice general

Parte I

Introducción

El Análisis Multivariante es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable estadística sobre una muestra de individuos. La información estadística en Análisis Multivariante es de carácter multidimensional, por lo tanto, la geometría, el cálculo matricial y las distribuciones multivariantes juegan un papel fundamental.

La razón de ser de dicho Análisis radica en un mejor entendimiento del fenómeno objeto de estudio obteniendo información que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir. Entre los objetivos de este análisis se pueden indicar:

- Crear un nuevo conjunto de variables con un número más reducido del original, con la mínima pérdida de información.
- Detectar posibles agrupaciones en los datos.
- Una vez definidos los grupos, poder llevar a cabo clasificaciones de nuevos individuos.
- Detectar posibles relaciones entre conjuntos de variables.

La información multivariante es una matriz de datos, pero a menudo, en el Análisis Multivariante, la información de entrada, consiste en matrices de distancias que miden el grado de discrepancia entre los individuos.

¿De qué manera se puede relacionar pues, la Geometría con las técnicas estadísticas? La Geometría nos permite representar de forma gráfica los problemas y resumirlos de un modo de fácil comprensión y entendimiento. Además, los dibujos geométricos pueden sugerirnos la solución al problema planteado. A través del uso de vectores geométricos, se reduce la dimensionalidad del problema, haciendo la teoría más accesible.

El objetivo del presente Trabajo Fin de Grado será presentar los métodos de análisis estadísticos subyacentes a una distribución Normal utilizando técnicas geométricas. Para alcanzar dicho objetivo, se desarrolla el estudio del análisis de varianza ordinario (ANOVA en adelante) desde el punto de vista geométrico, donde es usual probar la hipótesis nula de igualdad de medias en diferentes poblaciones (Parte II). Además, se desarrollarán técnicas de comparación de grupos de poblaciones mediante las denominadas pruebas de Contrastes Ortogonales (Parte III).

La estructura que se seguirá en cada capítulo será la siguiente:

1. Calcular el vector observación y el vector modelo.
2. Determinar las hipótesis de interés objeto de estudio y la dirección correspondiente al contraste.
3. Desarrollar la Descomposición de Pitágoras.

4. Ajustar el vector modelo.

5. Calcular el valor del estadístico F.

6. Calcular la estimación de σ^2.

7. Obtener un Intervalo de Confianza.

Nociones básicas

La Estadística es una ciencia formal; y por tanto, con base matemática, que enmarca a un conjunto de procedimientos diseñados para la recolección, análisis e interpretación de los datos. Su finalidad principal consiste en explicar condiciones regulares en fenómenos aleatorios, para posteriormente poder predecir y tomar decisiones en distintas áreas de investigación.

En el estudio del Análisis Estadístico, es necesario el conocimiento de los conceptos básicos y definiciones que se relacionan a continuación:

- **Población**: es el conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencia).

- **Muestra aleatoria**: es un subconjunto representativo de la población al que tenemos acceso y sobre el que realmente tomamos las observaciones (mediciones). Para seleccionar los individuos de la muestra se utiliza alguna técnica de muestreo como el aleatorio, el estratificado, por conglomerados o sistemático.

- **Experimento aleatorio**: es aquel que bajo el mismo conjunto aparente de condiciones iniciales, puede presentar resultados diferentes.

- **Suceso**: cada uno de los resultados posibles de una experiencia aleatoria.

- **Variable**: Característica observable que varía entre los diferentes individuos de la población. Se denomina **variable aleatoria** a una función que asigna un valor al resultado de un experimento aleatorio.

- **Parámetro**: es una cantidad numérica calculada sobre una población.

- **Estadístico**: es una cantidad numérica calculada sobre una muestra. Si un estadístico se usa para aproximar un parámetro, también se denomina **estimador**.

- **Media**: es la media aritmética (promedio) de los valores de una variable.

- **Varianza**: cuantifica la dispersión de los datos respecto a la media. Se obtiene como la media de las desviaciones cuadráticas de cada dato respecto a la media.

- **Cuasivarianza**: es un estimador insesgado de la varianza poblacional.

- **Desviación típica**: es la raíz cuadrada de la varianza.

- **Distribución de probabilidad de una variable aleatoria**: es una función que asigna a cada suceso definido sobre la variable aleatoria, la probabilidad de que dicho suceso ocurra.
- **Función de distribución acumulada asociada a una variable aleatoria real** Y: es una función matemática que depende de la variable real y que describe la probabilidad de que Y tenga un valor menor o igual que y. Está sujeta a la Ley de Distribución de Probabilidad de la variable Y.
- **Distribución** $F_{1,q}$ *de* $Fisher - Snedecor$: se define como la proporción

$$F_{1,q} = \frac{u^2}{\left[\frac{v_1^2 + ... + v_q^2}{q}\right]}$$

donde u y $v_1, ..., v_q$ son valores independientes que siguen una distribución Normal de media 0 y varianza σ^2.

- **Distribución** $T - Student$ (t_q): se define como la proporción

$$t_q = \frac{u}{\sqrt{\frac{v_1^2 + ... + v_q^2}{q}}}$$

donde u y $v_1, ..., v_q$ son valores independientes que siguen una distribución Normal de media 0 y varianza σ^2. Nótese que $t_q^2 = F_{1,q}$.

Notación

Los conceptos básicos anteriores se definirán sobre vectores $n \times k$ (siendo n el número de individuos y k el número de variables) relacionados con el estudio estadístico.

Para describir la nomenclatura usada en posteriores capítulos; supongamos que sobre los individuos $w_1, ..., w_n$ se han observado las variables $Y_1, ..., Y_k$. Se denotará y_{ij} la observación de la variable Y_j sobre el individuo w_i. El vector observación tendrá la notación:

$$y = \left[\ y_{11}, y_{12}, \ldots, y_{1n}, \ldots, y_{k1}, \ldots, y_{kn}\ \right]^t$$

Se indicará:

1. Estimaciones de cada media poblacional $\bar{y}_{l\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{lj}$ con $l = 1, ..., k$.
2. $\bar{y}_{\cdot\cdot}$ es el vector con la estimación de la media global.

Ejemplo

Para ilustrar la idea básica del Análisis Multivariante desde el punto de vista geométrico de la forma más sencilla posible, se toma una muestra de tamaño 2 de una población y se desea estimar la media μ de la población. Por ejemplo, supongamos que se ha comprado un termómetro y queremos responder la pregunta: *¿es el termómetro exacto en el punto de congelación?* Una posible forma de contestar a esta cuestión sería tomar la lectura del termómetro después de colocarlo en un vaso de hielo finamente picado dos días diferentes. Si tomamos los datos obtenidos para cada día, tenemos una muestra de tamaño 2 constituida por las observaciones y_1 e y_2. Utilizamos estos datos para verificar si es exacto o no el termómetro.

Si el termómetro es certero, entonces, la media de las temperaturas será siempre $\mu = 0$, tal y como se representa en la Figura 1.1(a) pues se supone una distribución Normal. Cada par de temperaturas y_1 e y_2 puede representarse como un punto p en el plano ; como se muestra en la Figura 1.1(b). Tras varias repeticiones del experimento, los pares de observaciones podrían representarse como puntos alrededor del origen tal y como se muestra en la Figura 1.1(c).

Supóngase que el termómetro está sesgado, y que la media de temperaturas es $\mu \neq 0$ a lo largo del tiempo. Este caso, viene representado en la Figura 1.1 apartados $(d), (e)$ y (f).

A partir de la figura anterior, se puede plantear la cuestión original *¿es el termómetro exacto en el punto de congelación?* desde un enfoque geométrico *¿están las observaciones de una nube de puntos centrada en el origen o pertenecen a una nube desplazada del origen?*. Más concretamente, *¿es* $\mu = 0$ *o* $\mu \neq 0$*?*.

Se necesita una medida, en términos de un test estadístico que nos permita distinguir entre los dos casos. Las Figuras 1.1(c) y 1.1(f) aportan una prueba para esta cuestión. Cuando $\mu \neq 0$ la nube de puntos está desplazada del origen alrededor de un punto en la línea equiangular (bisectriz del primer y tercer cuadrante). Para un punto p como el del ejemplo, el estadístico calcula la proporción entre la distancia desde el origen al punto de proyección sobre la línea equiangular representada por A; y la distancia desde la línea al punto p, representado por B. Esta proporción es mayor cuando $\mu \neq 0$ que cuando $\mu = 0$; tal y como se muestra en la Figura 1.2. Esto hace pensar que esta proporción o ratio, A/B coincide con el estadístico para el estudio del valor de μ. Dicha ratio se considera pequeña si $\mu = 0$ y grande cuando $\mu \neq 0$. Se debe tener en cuenta, que este estadístico no contempla ni la dispersión de los puntos ni depende de ninguna unidad de medida.

Siguiendo con el ejemplo, si el punto (y_1, y_2) es $(1.3, 1.5)$, el estadístico toma el valor $A/B = 1.4\sqrt{2}/(0.1\sqrt{2}) = 14$ como se puede comprobar en la Figura 1.3.

Si la media μ es realmente cero; y teniendo en cuenta las distribuciones de A y B, entonces, el estadístico A/B sigue una distribución T-Student con $n-1$ grados de libertad; en este caso, una T-Student con 1 grado de libertad (t_1 en adelante); el valor absoluto del estadístico t_1 es menor que 12.7 en el 95 % de los casos. Si se compara con el valor obtenido, 14, se puede concluir que hay evidencias para afirmar que el

termómetro está sesgado; pues 14 estaría fuera del rango de una $t_{1,0.95}$.

Interpretación geométrica

En este apartado, se usará la notación de vector geométrico en lugar de un punto (o una nube de puntos) en un sistema de coordenadas. Así pues, el vector de observaciones que se representó como $p = (y_1, y_2) = (1.3, 1.5)$ pasa a representarse como el vector

$$y = [1.3, 1.5]^t$$

A continuación, se elige un sistema de coordenadas ortogonales de dimensión 2 que se ajuste a nuestro problema. La dirección equiangular es de especial interés, ya que se toma ésta como la dirección del primer eje de coordenadas en el nuevo sistema de coordenadas. Se denotará por $U_1 = [1, 1]^t/\sqrt{2}$. Tomando esta dirección, sólo queda una elección para elegir el segundo eje, la dirección $U_2 = [-1, 1]^t/\sqrt{2}$. Esta descomposición se muestra en la Figura 1.4.

En el siguiente paso, se proyecta el vector observación en cada uno de los ejes de coordenadas del sistema obtenido; obteniéndose los respectivos vectores de proyección (véase gráficamente en la Figura 1.5):

$$v = (y \cdot U_1)U_1 = 1.4\sqrt{2}U_1 = \begin{bmatrix} 1.4 \\ 1.4 \end{bmatrix}$$

$$w = (y \cdot U_2)U_2 = 0.1\sqrt{2}U_2 = \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix}$$

A partir de esta idea, se puede obtener una descomposición del vector observación en un *vector modelo* y un *vector error*:

$$\begin{array}{ccccc} y & = & (y \cdot U_1)U_1 & + & (y \cdot U_2)U_2 \\ y & = & 1.4\sqrt{2}U_1 & + & 0.1\sqrt{2}U_2 \\ \begin{bmatrix} 1.3 \\ 1.5 \end{bmatrix} & = & \begin{bmatrix} 1.4 \\ 1.4 \end{bmatrix} & + & \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix} \end{array}$$

$$\text{vector observación} = \text{vector modelo} + \text{vector error}$$

El estadístico A/B definido anteriormente, se puede calcular como sigue; siendo $|y \cdot U_2|$ el valor absoluto de $y \cdot U_2$

$$t = \frac{y \cdot U_1}{|y \cdot U_2|} = \frac{1.4\sqrt{2}}{0.1\sqrt{2}} = 14$$

La descomposición de Pitágoras asociada se expresa a continuación y se representa gráficamente en la Figura 1.6. Dicha descomposición se corresponde con las bases del estudio estadístico de la tabla de análisis de la varianza; donde la suma de cuadrados total es la suma de cuadrados del modelo más la suma de cuadrados del error:

$$\begin{aligned} \| y \|^2 &= (y \cdot U_1)^2 + (y \cdot U_2)^2 \\ 1.3^2 + 1.5^2 &= (1.4\sqrt{2})^2 + (0.1\sqrt{2})^2 \\ 3.94 &= 3.92 + 0.02 \end{aligned}$$

A partir de la descomposición anterior, se puede definir otro estadístico como la proporción entre las longitudes al cuadrado de las proyecciones (se correspondería con el estadístico A^2/B^2 que sigue una distribución F de Snedecor)

$$F = \frac{(y \cdot U_1)^2}{(y \cdot U_2)^2} = \frac{3.92}{0.02} = 196$$

Si este valor se compara con el percentil 0.95 de una distribución $F_{1,1} = 161$; de nuevo, se puede concluir que existen evidencias para indicar que el termómetro está sesgado.

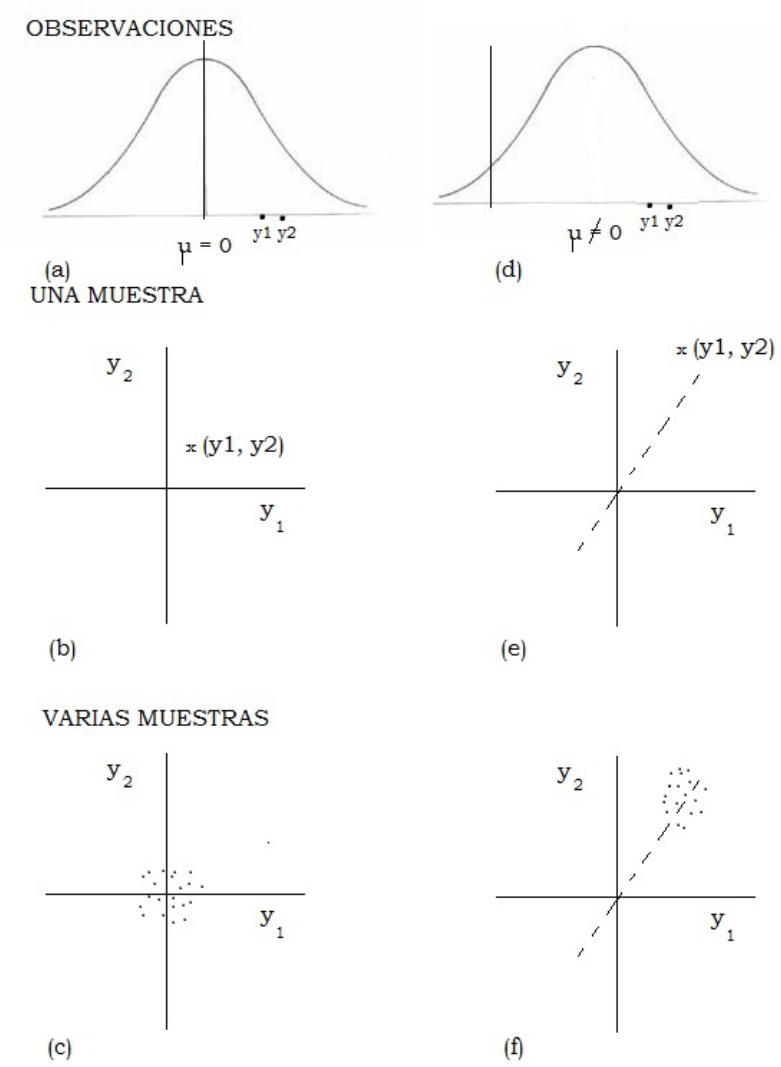

Figura 1.1: Correspondencia entre las muestras de tamaño 2 y el punto (y_1, y_2) en el espacio bidimensional. Nube de puntos resultante de muchas repeticiones del experimento centrado en el origen si $(\mu = 0)$ y centrado en el punto (μ, μ) para otros valores de μ distintos de cero.

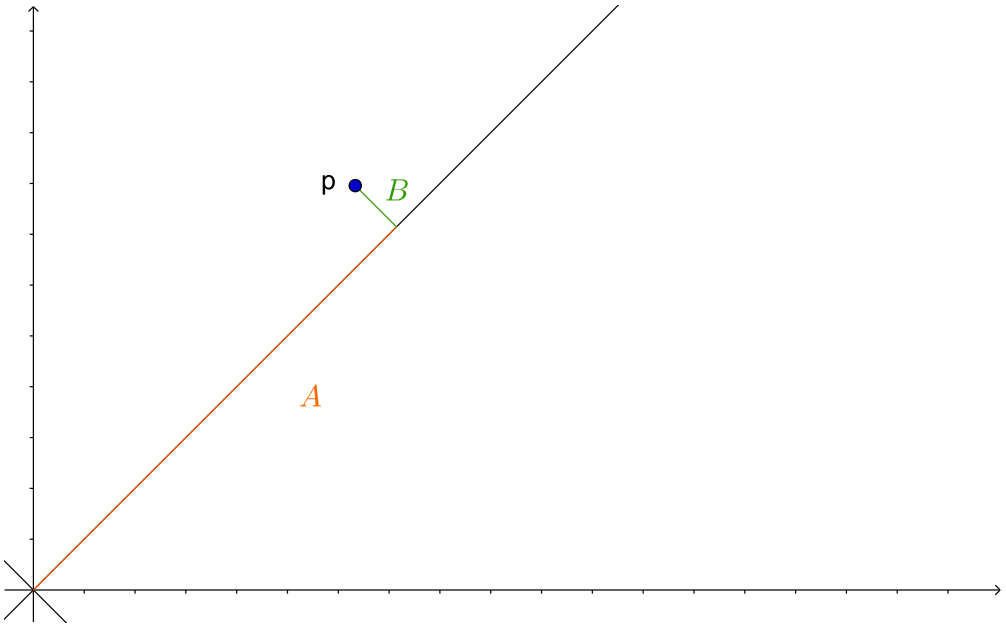

Figura 1.2: Proporción A/B.

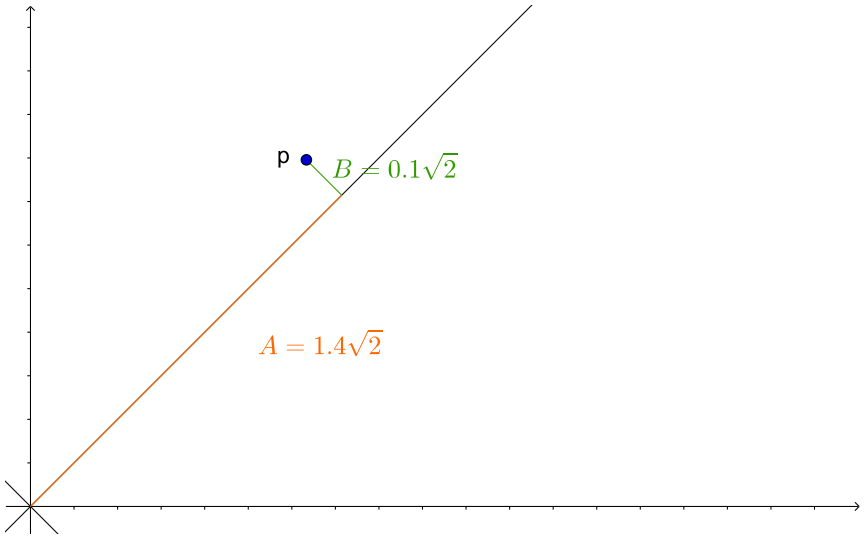

Figura 1.3: El punto $p = (1.3, 1.5)$, la perpendicular a la línea equiangular, y las distancias asociadas A y B.

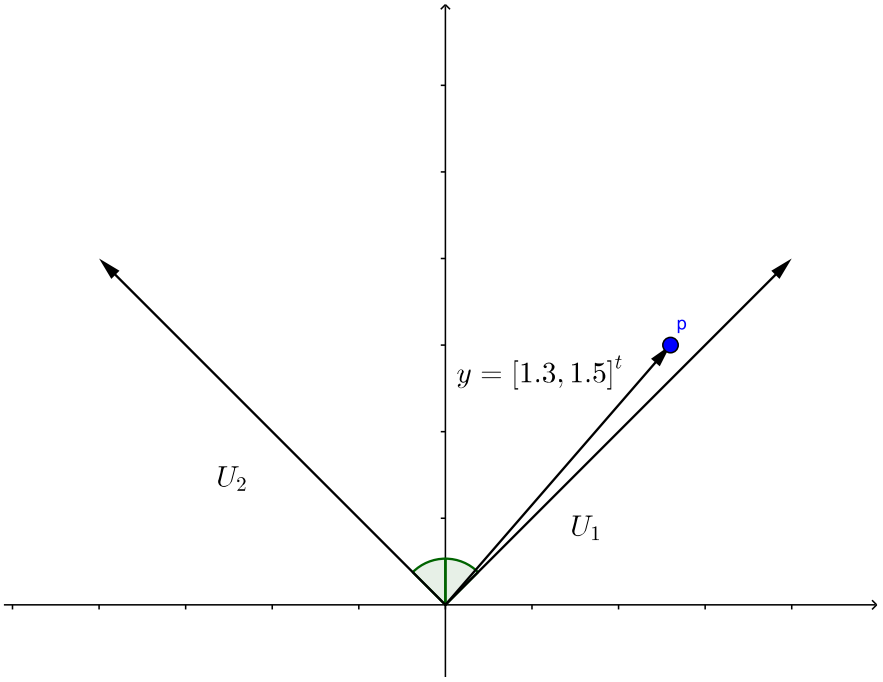

Figura 1.4: Un sistema de coordinadas ortogonal $U_1 = [1, 1]^t/\sqrt{2}$ y $U_2 = [-1, 1]^t/\sqrt{2}$ que se relaciona con el problema estadístico planteado.

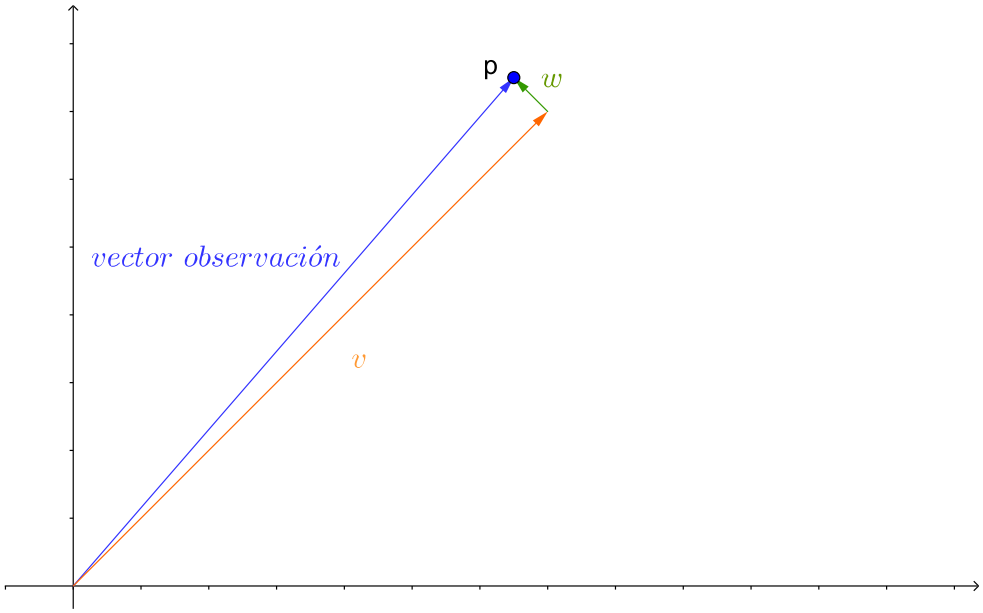

Figura 1.5: Descomposición del vector observación en dos vectores proyección: uno sobre la línea equiangular y otro sobre la dirección perpendicular a ésta.

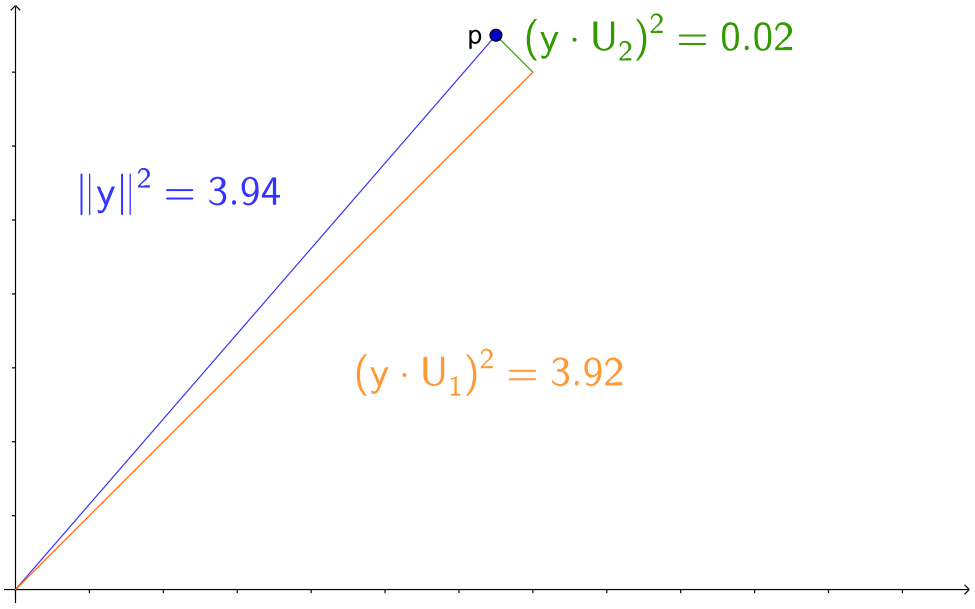

Figura 1.6: Descomposición de Pitágoras.

Parte II

Análisis de la varianza(ANOVA)

Capítulo 1

Descripción del problema

Se suponen k poblaciones de estudio distribuídas según una Normal, con medias $\mu_1, ..., \mu_k$ y varianza común desconocida σ^2.

Un contraste de interés sería el comprobar que alguna relación entre las medias de las poblaciones es cero. Para llevarlo a cabo, de cada población se extrae una muestra aleatoria de tamaño n; lo que lleva a obtener un **vector observación** en un espacio $n \times k$ dimensional de la forma

$$y = [y_{11}, ..., y_{1n}, y_{21}, ..., y_{2n}, ..., y_{k1}, ..., y_{kn}]^t$$

Se denotará al vector de valores esperados como $[\mu_1, ..., \mu_1, \mu_2, ..., \mu_2, ..., \mu_k, ...\mu_k]^t$ donde una base del espacio modelo k-dimensional es:

$$M = \left(\begin{bmatrix} 1 \\ \vdots(n) \\ 1 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots(n) \\ 0 \\ 1 \\ \vdots(n) \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ 1 \\ \vdots(n) \\ 1 \end{bmatrix} \right)$$

Si proyectamos el vector y sobre M y ajustamos el modelo en función de la media muestral (ya que ésta es un estimador insesgado de la media poblacional), el vector observación sería igual al **vector modelo** más el **vector error**:

$$\begin{array}{ccccc} y & = & \bar{y}_{i\cdot} & + & (y - \bar{y}_{i\cdot}) \\ \text{vector observación} & = & \text{vector modelo} & + & \text{vector error} \end{array}$$

Esto nos lleva a la estimación de mínimos cuadrados $\bar{y}_{i\cdot}$ de los μ_i, es decir, la estimación mediante las medias muestrales de las k medias poblacionales, ya que el vector modelo $\bar{y}_{i\cdot}$ se puede obtener a partir de estas estimaciones por:

$$\bar{y}_{1\cdot}\begin{bmatrix} 1 \\ \vdots(n) \\ 1 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} + \bar{y}_{2\cdot}\begin{bmatrix} 0 \\ \vdots(n) \\ 0 \\ 1 \\ \vdots(n) \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \ldots + \bar{y}_{k\cdot}\begin{bmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ 1 \\ \vdots(n) \\ 1 \end{bmatrix}$$

La hipótesis de interés más frecuente es $H_0 : c = 0$ donde $c = \alpha_1\mu_1 + ... + \alpha_k\mu_k$ con $\alpha_1 + ... + \alpha_k = 0$.
La dirección asociada a esta hipótesis es:

$$U_c = \frac{1}{\sqrt{n\sum_{i=1}^{k}\alpha_i^2}}\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_2 \\ \vdots \\ \alpha_k \\ \vdots \\ \alpha_k \end{bmatrix}$$

El estadístico del contraste que arrojará información sobre si hay evidencias o no para aceptar la hipótesis nula, se define como el cociente entre el módulo de la proyección al cuadrado sobre la dirección del contraste y la proyección media al cuadrado sobre el espacio de errores. Así, el estadístico F, es una variable aleatoria que sigue una distribución F de Snedecor y que viene definido por:

$$F = \frac{(y \cdot U_c)^2}{\frac{\|(y - \bar{y}_{i\cdot})\|^2}{k(n-1)}}$$

De forma gráfica, la descomposición ortogonal del vector observación para el caso de varias poblaciones se puede representar como se indica en la Figura 1.7.

Se puede plantear la cuestión, ¿cuántas hipótesis podemos probar de forma independiente? Se puede decir que el vector U_c asociado con cualquier contraste se encuentra en el espacio modelo, pero es ortogonal al vector equiangular; definido éste como $[1, 1, 1, ..., 1]^t$. Es decir, el vector U_c se encuentra en un subespacio $k-1$ dimensional del espacio modelo, llamado **espacio de contrastes**. Entonces, podemos especificar que las $k-1$ direcciones ortogonales de la forma de U_c, corresponden a lo sumo a $k-1$ contrastes ortogonales y, por lo tanto, a $k-1$ hipótesis independientes. Hay que tener

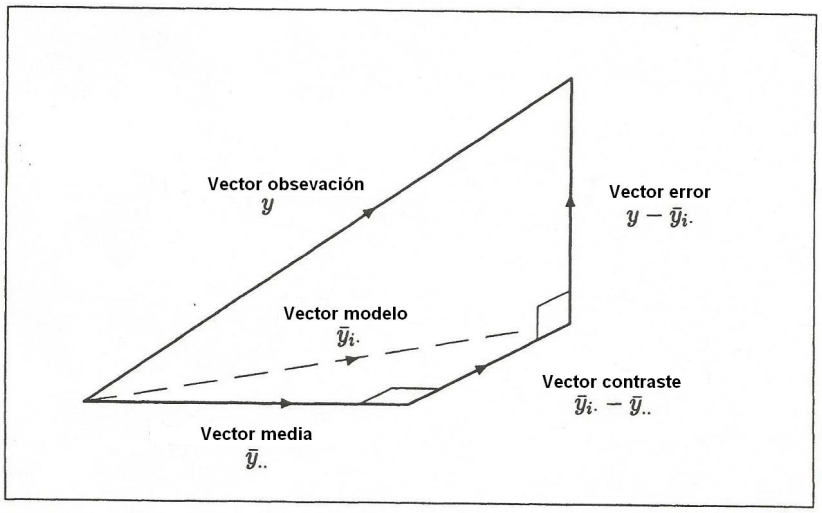

Figura 1.7: Descomposición ortogonal del vector observación para el caso de varias poblaciones.

en cuenta que es deseable establecer hipótesis independientes: la conclusión de uno no debe influir en la conclusión de otro.

En la práctica, los contrastes que corresponden a las hipótesis de interés puede que no sean todos ortogonales, o que sean menos de $k-1$. Sin embargo, si se especifica un conjunto completo de $k-1$ contrastes ortogonales, entonces los $k-1$ vectores correspondientes, además de la dirección media, el vector equiangular, constituirán una base de un sistema de coordenadas para el espacio modelo. Este será el objeto de estudio desarrollado en el Capítulo 3.

En resumen, si el vector equiangular U_1 es el vector unitario, y $U_2, \ldots, U_k$ corresponde a los $k-1$ contrastes ortogonales de interés, entonces se puede definir la descomposición ortogonal asociada a y como:

$$\begin{array}{ccccccc}
y & = & (y \cdot U_1)U_1 & + & (y \cdot U_2)U_2 + \ldots + (y \cdot U_k)U_k & + & (y \cdot U_{k+1})U_{k+1} + \ldots + (y \cdot U_{kn})U_{kn} \\
y & = & \bar{y}_{\cdot\cdot} & + & (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) & + & (y - \bar{y}_{i\cdot}) \\
\text{vector observación} & = & \text{vector media} & + & \text{vector contraste} & + & \text{vector error}
\end{array}$$

donde $U_{k+1}, \ldots, U_{kn}$ forman una **base ortogonal** para el espacio de errores. Siguiendo la Figura 1.7 se ha descompuesto el vector modelo en la suma del vector media y del vector contraste.

A partir de la dirección U_c se puede definir el coeficiente de proyección $y \cdot U_c$ como

$$y \cdot U_c = \frac{n(\alpha_1 \bar{y}_{1\cdot} + \ldots + \alpha_k \bar{y}_{k\cdot})}{\sqrt{n \sum_{i=1}^{k} \alpha_i^2}} = \frac{\sqrt{n}(\alpha_1 \bar{y}_{1\cdot} + \ldots + \alpha_k \bar{y}_{k\cdot})}{\sqrt{\sum_{i=1}^{k} \alpha_i^2}} = \frac{\sqrt{n}\hat{c}}{\sqrt{\sum_{i=1}^{k} \alpha_i^2}}$$

donde $\hat{c} = \alpha_1\bar{y}_{1\cdot} + \ldots + \alpha_k\bar{y}_{k\cdot}$ es la estimación de $c = \alpha_1\mu_1 + \ldots + \alpha_k\mu_k$ y el promedio, a lo largo de muchas repeticiones en el estudio, es:

$$\frac{\sqrt{n}(\alpha_1\mu_1 + \ldots + \alpha_k\mu_k)}{\sqrt{\sum_{i=1}^{k}\alpha_i^2}} = \frac{\sqrt{n}c}{\sqrt{\sum_{i=1}^{k}\alpha_i^2}}$$

Es decir, $y \cdot U_c = \sqrt{n}\hat{c}/\sqrt{\sum_{i=1}^{k}\alpha_i^2}$ proviene de una distribución $N\left[\sqrt{n}c/\sqrt{\sum_{i=1}^{k}\alpha_i^2}, \sigma^2\right]$. Por lo tanto, $\sqrt{n}(c-\hat{c})/\sqrt{\sum_{i=1}^{k}\alpha_i^2}$ viene de una distribución $N[0,\sigma^2]$. Esto es lo que se utiliza como numerador para el estadístico t. Para el denominador se usa $\sqrt{s^2}$, donde s^2 es el promedio de los $k(n-1)$ coeficientes de las proyecciones al cuadrado, $(y\cdot U_{k+1})^2, \ldots, (y\cdot U_{kn})^2$, es decir, la cuasivarianza. El valor obtenido como resultado del estadístico t es:

$$t = \frac{\sqrt{n}(\hat{c}-c)/\sqrt{\sum_{i=1}^{k}\alpha_i^2}}{s} = \frac{\sqrt{n}(\hat{c}-c)}{s\sqrt{\sum_{i=1}^{k}\alpha_i^2}}$$

A fin de obtener un intervalo de confianza del 95 % para el contraste c, se supone que nuestros valores observados se encuentran entre los percentiles 2.5 y 97.5 de la distribución $t_{k(n-1)}$. Es decir, que

$$-t_{k(n-1),0.975} \;\leq\; \frac{\sqrt{n}(\hat{c}-c)}{s\sqrt{\sum_{i=1}^{k}\alpha_i^2}} \;\leq\; t_{k(n-1),0.975}$$

Lo que produce el intervalo de confianza del 95 % deseado para c:

$$\hat{c} - \sqrt{\sum_{i=1}^{k}\alpha_i^2}\frac{s}{\sqrt{n}}t_{k(n-1),0.975} \;\leq\; c \;\leq\; \hat{c} + \sqrt{\sum_{i=1}^{k}\alpha_i^2}\frac{s}{\sqrt{n}}t_{k(n-1),0.975}$$

En el próximo capítulo se estudiará el análisis de la varianza mediante el contraste entre las medias poblacionales tomando muestras independientes. A continuación, se estudiará cómo construir conjuntos de contrastes ortogonales utilizando tres tipos básicos de contrastes: comparaciones por clases, comparaciones factoriales y comparaciones polinómicas. Además, se construirán intervalos de confianza para dichos contrastes.

El utilizar muestras independientes surge a partir de la idea de no tener sentido agrupar de forma natural las observaciones tomadas de la población. Un ejemplo, sería el seleccionar aleatoriamente cinco hombres y cinco mujeres y medir su peso, o medir el pulso cardíaco de seis corredores de fondo frente a seis individuos que no practican esta

modalidad de deporte. Normalmente, el objetivo de estudio es la diferencia de medias de los pesos entre los hombres y las mujeres de la población de estudio; o la diferencia de medias en el pulso cardíaco entre las personas corredoras frente a las que no lo son. Más formalmente, se está interesado en comparar las medias μ_1 y μ_2 de ambas poblaciones. Para el análisis, se asume que ambas poblaciones son idénticamente distribuídas y que las varianzas poblacionales son las mismas ($\sigma_1^2 = \sigma_2^2 = \sigma^2$).

Capítulo 2

ANOVA en muestras independientes

A continuación, se describirá un conjunto de datos consistente en la altura de hombres y mujeres estudiantes de una clase de primer año de Universidad de California [2]. Se analizará un pequeño subconjunto de datos relativos a dichas alturas: se toman cuatro medidas de hombres y cuatro de mujeres.

2.1. Selección de la muestra

En la clase ASM150 del Departamento de Agronomía de la Universidad de California de Davis, se tomaron estadísticas de los alumnos matriculados el primer curso. Para obtener los datos, con el objetivo de comparar dos muestras independientes (una por cada sexo), se envió un formulario a todos los alumnos matriculados para que cada individuo completara junto con su altura, su sexo. La altura se recoge en pulgadas; teniendo en cuenta que una pulgada equivale a 2.54 centímetros.
Los datos resultantes se muestran en las Figuras 2.1.*a* y 2.1.*b*, donde se recogen las alturas de 49 alumnas y 77 alumnos, respectivamente.

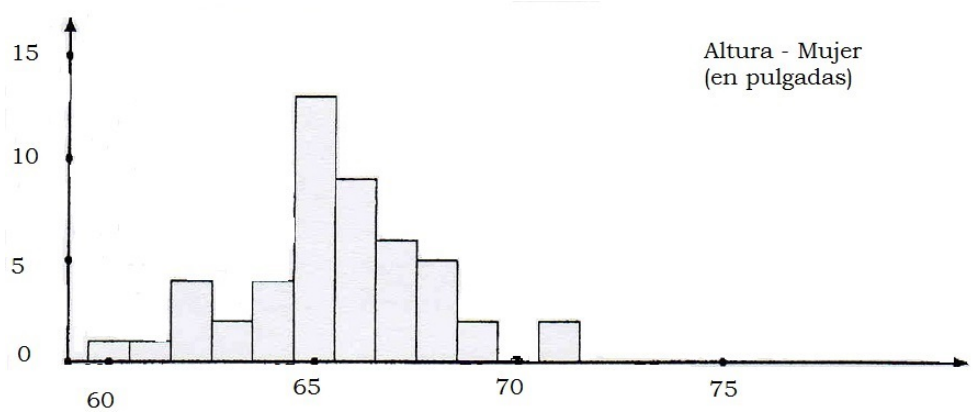

Figura 2.1.*a*: Histograma de alturas de mujeres en la clase ASM150 de la Universidad de California.

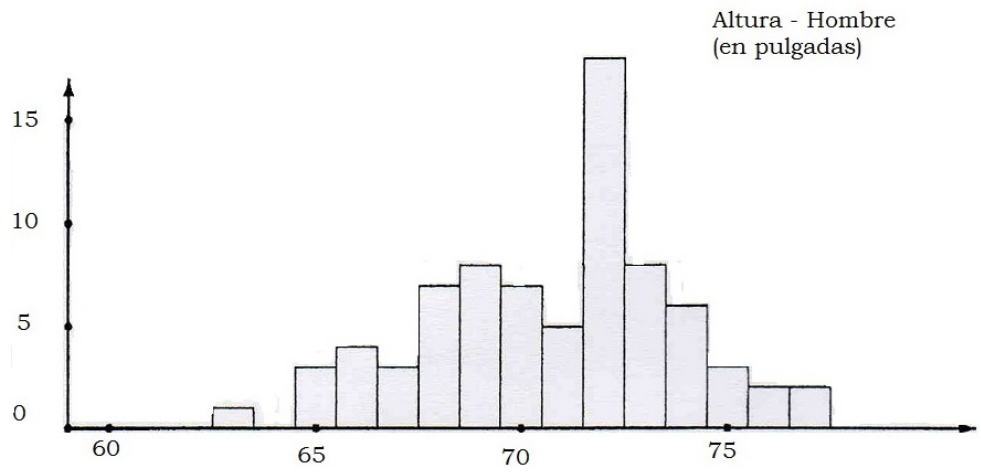

Figura 2.1.*b*: Histograma de alturas de hombres en la clase ASM150 de la Universidad de California.

2.2. Descomposición ortogonal

Se puede observar que los tamaños muestrales de nuestro conjunto de datos no son iguales; la muestra de mujeres tiene un tamaño de 49 individuos mientras que la de hombres es de 77. Para desarrollar el estudio de la descomposición ortogonal, se tomarán muestras simples de tamaño cuatro y se trabajará con la información de la siguiente tabla:

Alturas					Medias
Mujeres	60	64	65	68	64.25
Hombres	70	69	77	71	71.75

El *vector observación* resultante sería:

$$y = [60, 64, 65, 68, 70, 69, 77, 71]^t.$$

Suponiendo que nuestras poblaciones son normales de media μ_1 y μ_2 y varianza σ^2; la idea básica es comparar si la media de los estudiantes hombres y la de mujeres es la misma; se plantea el contraste $H_0 : \mu_1 = \mu_2$ o equivalentemente, $H_0 : \mu_2 - \mu_1 = 0$. Si esta hipótesis se toma como cierta, entonces nuestro vector observación será parte de una nube de puntos centrada en el punto $(\mu, \mu, \mu, \mu, \mu, \mu, \mu, \mu)^t$ tal y como se muestra en la Figura 2.2. Si por el contrario, se toma como cierta la hipótesis $\mu_1 \neq \mu_2$ nuestro vector de observación será parte de la nube centrada en el punto $(\mu_1, \mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2, \mu_2)^t$ tal y como se representa en la Figura 2.3.

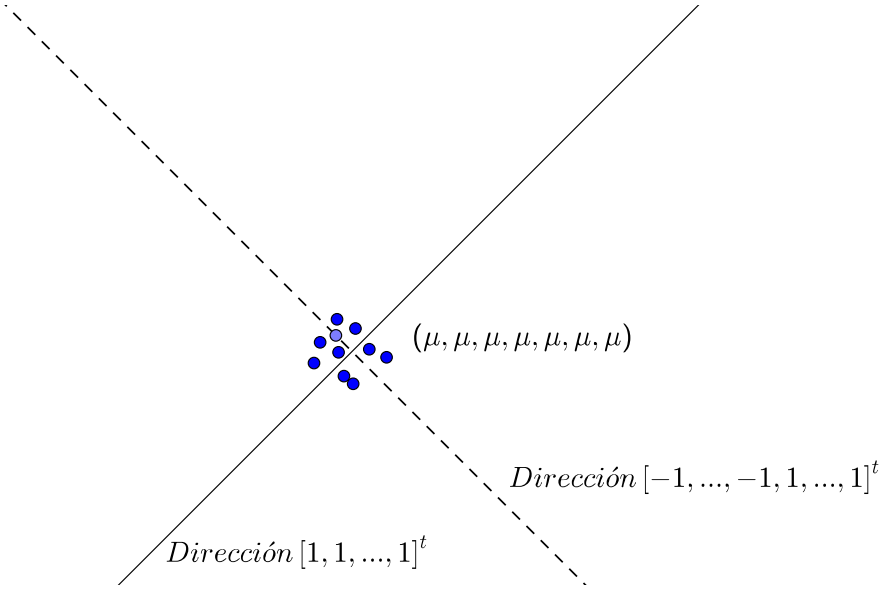

Figura 2.2: Nube de puntos que representa muchas repeticiones de nuestro estudio sobre la altura. Se supone que $\mu_1 = \mu_2$ y por tanto, la nube está centrada en un punto de la línea equiangular.

En otras palabras, si $\mu_1 = \mu_2$ la nube está centrada en un punto de la línea equiangular; en cambio si $\mu_1 \neq \mu_2$, la nube está desplazada respecto a dicha línea en la dirección:

$$\begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \end{bmatrix} - \begin{bmatrix} \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \mu \end{bmatrix} = \begin{bmatrix} \mu_1 - \mu \\ \mu_1 - \mu \\ \mu_1 - \mu \\ \mu_1 - \mu \\ \mu_2 - \mu \\ \mu_2 - \mu \\ \mu_2 - \mu \\ \mu_2 - \mu \end{bmatrix} = \begin{bmatrix} \frac{\mu_1 - \mu_2}{2} \\ \frac{\mu_1 - \mu_2}{2} \\ \frac{\mu_1 - \mu_2}{2} \\ \frac{\mu_1 - \mu_2}{2} \\ \frac{\mu_2 - \mu_1}{2} \\ \frac{\mu_2 - \mu_1}{2} \\ \frac{\mu_2 - \mu_1}{2} \\ \frac{\mu_2 - \mu_1}{2} \end{bmatrix} = \frac{\mu_2 - \mu_1}{2} \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

donde $\mu = (\mu_1 + \mu_2)/2$.

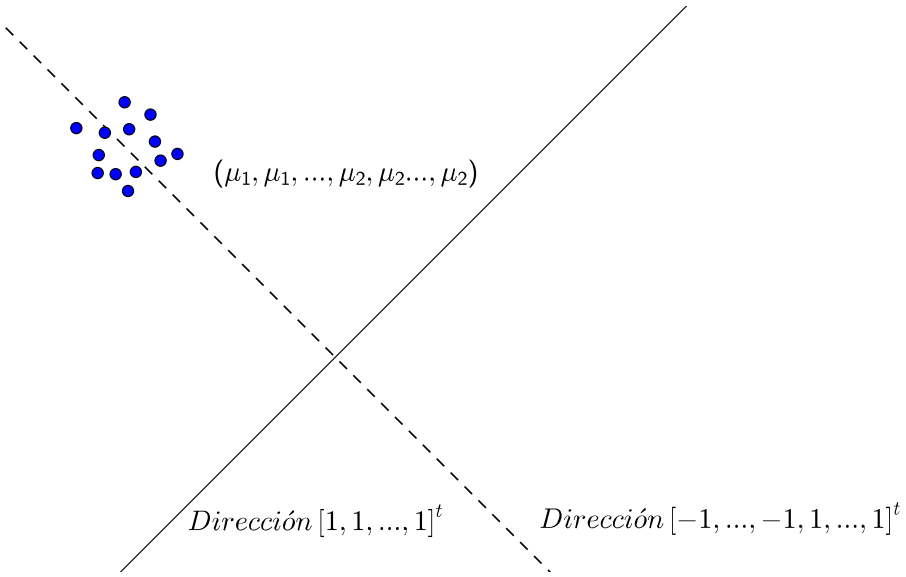

Figura 2.3: Nube de puntos que representa muchas repeticiones de nuestro estudio sobre la altura. Se supone que $\mu_1 \neq \mu_2$ y por tanto, la nube está desplazada respecto a la línea equiangular en la dirección $[-1, ..., 1]^t$

Un sistema de coordenadas apropiado de dimensión 8 para nuestro estudio puede ser el siguiente:

$$U_1 = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, U_2 = \frac{1}{\sqrt{8}} \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, U_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, U_4 = \frac{1}{\sqrt{6}} \begin{bmatrix} -1 \\ -1 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$U_5 = \frac{1}{\sqrt{12}} \begin{bmatrix} -1 \\ -1 \\ -1 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, U_6 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, U_7 = \frac{1}{\sqrt{6}} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ 2 \\ 0 \end{bmatrix}, U_8 = \frac{1}{\sqrt{12}} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ 3 \end{bmatrix}$$

Donde U_1 y U_2 generan el espacio modelo siendo $U_2 = (-1, -1, -1, -1, 1, 1, 1, 1)^t/\sqrt{8}$ la dirección asociada con la diferencia de medias $\mu_2 - \mu_1$; es decir, la dirección correspondiente al contaste. Las direcciones $U_3, \ldots, U_8$ generan el espacio de errores.

Descomposición ortogonal

El vector observación $y = [60, 64, 65, 68, 70, 69, 77, 71]^t$ se puede descomponer en ocho vectores proyección; los cuales serán calculados $(y \cdot U_i)U_i$ con $i = 1, ..., 8$. Así pues, y puede describirse mediante los siguientes componentes:

$$y = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 + ... + (y \cdot U_8)U_8$$

$$\begin{bmatrix} 60 \\ 64 \\ 65 \\ 68 \\ 70 \\ 69 \\ 77 \\ 71 \end{bmatrix} = \begin{bmatrix} 68 \\ 68 \\ 68 \\ 68 \\ 68 \\ 68 \\ 68 \\ 68 \end{bmatrix} + \begin{bmatrix} -3.75 \\ -3.75 \\ -3.75 \\ -3.75 \\ 3.75 \\ 3.75 \\ 3.75 \\ 3.75 \end{bmatrix} + \begin{bmatrix} -2 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + ... + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.25 \\ 0.25 \\ 0.25 \\ -0.75 \end{bmatrix}$$

El modelo ajustado se puede simplificar como se indica a continuación (véase gráficamente esta descomposición en la Figura 1.7 del Capítulo 1):

$$y = \bar{y}_{..} + (\bar{y}_{i\cdot} - \bar{y}_{..}) + (y - \bar{y}_{i\cdot})$$

$$\begin{bmatrix} 60 \\ 64 \\ 65 \\ 68 \\ 70 \\ 69 \\ 77 \\ 71 \end{bmatrix} = \begin{bmatrix} 68 \\ 68 \\ 68 \\ 68 \\ 68 \\ 68 \\ 68 \\ 68 \end{bmatrix} + \begin{bmatrix} -3.75 \\ -3.75 \\ -3.75 \\ -3.75 \\ 3.75 \\ 3.75 \\ 3.75 \\ 3.75 \end{bmatrix} + \begin{bmatrix} -4.25 \\ -0.25 \\ 0.75 \\ 3.75 \\ -1.75 \\ -2.75 \\ 5.25 \\ -0.75 \end{bmatrix}$$

vector observación = vector media + vector contraste + vector error

Distribución de los vectores proyección

Con el objeto de contrastar si realmente la diferencia de medias de alturas puede ser cero, se trabaja con los vectores proyección calculando cuál es su distribución de probabilidad. En primer lugar, se calcula la distribución para la proyección $Y \cdot U_1$ donde Y es la variable aleatoria correspondiente al vector observación. De forma análoga, se tendría para el resto de los 7 vectores proyección. Se continua con la idea de que nuestras muestras siguen una distribución Normal con medias μ_1 y μ_2, y con varianza σ^2.

$$y \cdot U_1 = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \end{bmatrix} \cdot \frac{1}{\sqrt{8}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \frac{y_{11} + y_{12} + y_{13} + y_{14} + y_{21} + y_{22} + y_{23} + y_{24}}{\sqrt{8}}$$

La media de la variable aleatoria $Y \cdot U_1$ sería

$$\frac{\mu_1 + \mu_1 + \mu_1 + \mu_1 + \mu_2 + \mu_2 + \mu_2 + \mu_2}{\sqrt{8}} = \frac{4\mu_1 + 4\mu_2}{\sqrt{8}} = \frac{8\mu}{\sqrt{8}} = \sqrt{8}\mu$$

La varianza de $Y \cdot U_1$ sería

$$\frac{1}{\sqrt{8}^2} \cdot 8\sigma^2 = \sigma^2$$

Entonces, se puede decir que $Y \cdot U_1$ sigue una distribución $N(\sqrt{8}\mu, \sigma^2)$. Siguiendo el mismo procedimiento, se puede llegar a la conclusión de que $Y \cdot U_2 \sim N(\sqrt{2}(\mu_2 - \mu_1), \sigma^2)$ y que $Y \cdot U_i \sim N(0, \sigma^2)$ con $i = 3, ..., 8$.

Contraste de hipótesis

Se plantea el siguiente contraste de hipótesis: $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$. Bajo esta hipótesis nula, se comprueba si el cuadrado de la longitud de la proyección $(y \cdot U_2)^2$ es similar o considerablemente más grande que la media de los cuadrados de las longitudes de las proyecciones restantes $\frac{(y \cdot U_3)^2 + ... + (y \cdot U_8)^2}{6}$. Aplicando el Teorema de Pitágoras se tiene que:

$$\begin{array}{rcccccl} \| y \|^2 & = & (y \cdot U_1)^2 & + & (y \cdot U_2)^2 & + & (y \cdot U_3)^2 + \ldots + (y \cdot U_8)^2 \\ \| y \|^2 & = & \| \bar{y}_{\cdot\cdot} \|^2 & + & \| \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \|^2 & + & \| y - \bar{y}_{i\cdot} \|^2 \\ 37176 & = & 36992 & + & 112.5 & + & 71.5 \end{array}$$

Teniendo en cuenta las distribuciones de las proyecciones, el estadístico resultante es

$$F = \frac{(y \cdot U_2)^2}{((y \cdot U_3)^2 + ... + (y \cdot U_8)^2)/6} = \frac{\| \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \|^2}{\| y - \bar{y}_{i\cdot} \|^2 / 6} = \frac{112.5}{71.5/6} = 9.44$$

Si $\mu_1 = \mu_2$, entonces F procede de una distribución $F_{1,6}$. Si se acepta un nivel de significación del 0.95; el valor de la $F_{1,6,0.95}$ se corresponde con 5.99. Como el valor de F observado es mayor que el percentil 95 de la $F_{1,6}$, se puede rechazar la hipótesis $H_0 : \mu_1 = \mu_2$ y concluir que hay evidencias para sugerir que la media de altura de los hombres es diferente a la de las mujeres.

El estadístico F puede ser transformado en una T-Student tal que así:

$$t = \frac{y \cdot U_2}{\sqrt{[(y \cdot U_3)^2 + ... + (y \cdot U_8)^2]/6}} = \frac{\| \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \|}{\frac{\| y - \bar{y}_{i\cdot} \|}{\sqrt{6}}} = \frac{\sqrt{8} \cdot 3.75}{\sqrt{71.5/6}} = 3.073$$

Si $\mu_1 = \mu_2$, entonces t procede de una distribución t_6 que a un nivel de significación del 0.95 también nos permite rechazar la hipótesis nula, ya que nuestro valor 3.073 es mayor que el valor del percentil 0.95de una $t_6 = 1.943$.

En la siguiente figura se representa geométricamente la descomposición de Pitágoras de los cuadrados de las longitudes del vector observación.

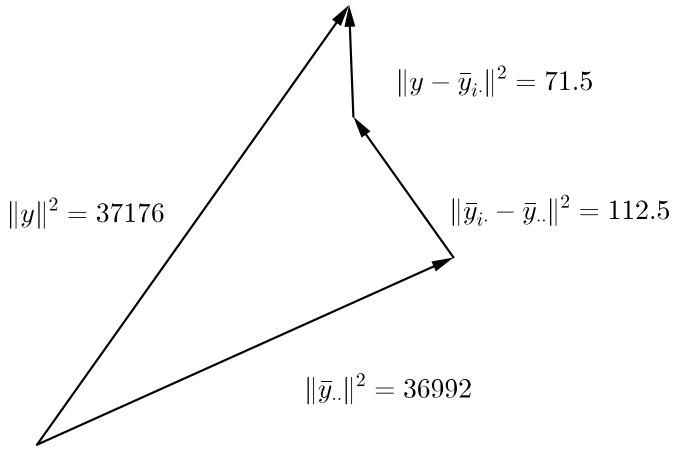

Figura 2.4: Descomposición de Pitágoras.

Estimación de σ^2

Para la estimación de σ^2 se utilizan las 6 variables aleatorias $Y \cdot U_3, \ldots, Y \cdot U_8$ cuyas direcciones conforman el espacio de errores y que siguen una distribución $N[0, \sigma^2]$. La estimación se calcula a través del cociente s^2 definido en el Capítulo 1:

$$s^2 = \frac{(y \cdot U_3)^2 + \ldots + (y \cdot U_8)^2}{6} = \frac{\|y - \bar{y}_{i\cdot}\|^2}{6} = \frac{71.5}{6} = 11.92$$

Intervalo de confianza

A continuación, se incluye un Intervalo de Confianza para el contraste $H_0 : c = 0$; siendo $c = \mu_2 - \mu_1$ y siguiendo la expresión calculada en el Capítulo 1.

Tomando $\alpha_1 = -1$ y $\alpha_2 = 1$, junto con las estimaciones de cada media poblacional $\bar{y}_{1\cdot} = 64.25$ e $\bar{y}_{2\cdot} = 71.75$, se obtiene $\hat{c} = \alpha_1\bar{y}_{1\cdot} + \alpha_2\bar{y}_{2\cdot} = 7.5$, $s = \sqrt{11.92} = 3.45$, $n = 4$ y $t_{6,0.95} = 1.943$; resultando un intervalo de confianza para c de 7.5 ± 4.74.

Cabe indicar que no sale un Intervalo de confianza muy significativo debido al alto valor de la cuasivarianza que presentan los datos.

Parte III

Contrastes ortogonales

En capítulos anteriores, se ha desarrollado la idea del análisis ANOVA, en el cual, se pretende estudiar la hipótesis de igualdad de tratamientos. Sin embargo, a veces es deseable el estudio de la comparación de grupos de tratamientos; es en este caso donde se emplean pruebas de Contrastes Ortogonales.

El concepto de ortogonalidad es importante en el diseño de experimentos porque nos asegura independencia. Dos contrastes se dicen ortogonales si la suma de los productos de sus correspondientes coeficientes es igual a cero. Esto implica que la covarianza entre los dos contrastes es igual a cero y por tanto, los contrastes van a ser independientes. Si todos los contrastes formulados son ortogonales entre sí; entonces, esto llevará a que la suma de cuadrados acumulada en todos los contrastes ortogonales corresponde exactamente a la suma de cuadrados de los tratamientos.

El concepto de ortogonalidad debe ser tomando en cuenta en el diseño de experimentos. Véase por ejemplo, en el caso de un diseño factorial, éste es ortogonal si los efectos de cualquier factor se equilibran (suman cero) con los efectos de los otros factores. La ortogonalidad garantiza que el efecto de un factor o interacción pueda estimarse de manera independiente del efecto de cualquier otro factor o interacción presente en el modelo.

En los siguientes capítulos, se va a estudiar cómo construir conjuntos de contrastes ortogonales utilizando tres tipos básicos de contrastes: comparaciones por clase, comparaciones factoriales y comparaciones polinómicas.

Capítulo 3

Comparación por clases

Este capítulo está dedicado al estudio de contrastes del tipo comparación por clases a excepción de las comparaciones por parejas. La comparación por clases contrasta la media de una población con la media de una segunda población.

Para construir un conjunto completo de contrastes de este tipo, simplemente, se divide la población en dos clases y se contrasta la primera con la segunda. A continuación, se divide la primera clase en dos subclases y se contrasta la primera subclase con la segunda subclase. Sucesivamente, se divide cada subclase hasta llegar a una clase que tiene tamaño uno. Este proceso genera $k-1$ contrastes; donde k es el número de poblaciones.

La elección de las clases y subclases viene determinada por la población bajo estudio. De los $k-1$ contrastes ortogonales necesarios para construir el conjunto completo; sólo una proporción será suficiente para contrastar la hipótesis. El resto, no tienen sentido para contrastar la hipótesis pero deben ser tomados en cuenta para completar el sistema de coordenadas del espacio de trabajo.

El conjunto de medias poblacionales $\mu_1, ..., \mu_k$ son estimadas ajustando el modelo $y = \bar{y}_{i\cdot} + (y - \bar{y}_{i\cdot})$. Si se reescribe la expresión anterior, el ajuste del modelo será:

$$y = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y - \bar{y}_{i\cdot})$$

La estimación de σ^2 se lleva a cabo mediante

$$s^2 = \frac{\| \ y - \bar{y}_{i\cdot} \ \|^2}{k(n-1)}$$

Este valor sirve como base para el test de hipótesis $H_0 : c = 0$ siendo $c = \alpha_1\mu_1 + ... + \alpha_k\mu_k$. El estadístico asociado a dicho contraste (bajo H_0) resulta $F = \frac{(y \cdot U_c)^2}{s^2} \sim F_{1,k(n-1)}$ pues es el cociente entre dos variables con distribución Normal (ambas al cuadrado) y donde U_c es la dirección correspondiente a la hipótesis nula. Con un nivel de significación del 0.95, el intervalo de confianza para el contraste $c = \alpha_1\mu_1 + ... + \alpha_k\mu_k = 0$ es

$$\hat{c} \ \pm \ \sqrt{\sum_{i=1}^{k} \alpha_i^2} \frac{s}{\sqrt{n}} t_{k(n-1),0.95}$$

donde $\hat{c} = \alpha_1\bar{y}_{1\cdot} + \ldots + \alpha_k\bar{y}_{k\cdot}$.

3.1. Caso de estudio

Se realiza un experimento en el cual, cada cordero de una granja es asignado de forma aleatoria a tres tratamientos experimentales. El experimento es diseñado por clases para determinar:

(A) Si es necesario empapar la lana de los corderos para que éstos aumenten de peso.

(B) Si un empapado a los tres meses de edad es suficiente, comparándolo con un segundo empapado un mes más tarde.

Los tratamientos experimentales son los siguientes:

1. Control: se controla el peso de los corderos no mojados.
2. Un empapado: se controla el peso de los corderos con la lana mojada.
3. Doble empapado: se controla el peso de los corderos con la lana mojada a los tres meses y a los cuatro meses de edad.

Durante el experimento se puede imaginar a los corderos integrados en el rebaño y con la misma alimentación. Las medidas de los pesos en Kilogramos de las distintas poblaciones son las que se indican a continuación y han sido tomadas a los seis meses de edad de los corderos:

Población 1: Control

Num. cord	Kgs	Num.cord	Kgs	Num.cord	Kgs	Num. cord	Kgs	Num. cord	Kgs
1.	11	11.	7	21.	12	31.	10	41.	10
2.	10	12.	9	22.	8	32.	10	42.	12
3.	8	13.	9	23.	9	33.	8	43.	9
4.	11	14.	7	24.	8	34.	11	44.	10
5.	11	15.	9	25.	11	35.	11	45.	11
6.	10	16.	10	26.	9	36.	12	46.	7
7.	12	17.	7	27.	9	37.	10	47.	9
8.	12	18.	11	28.	10	38.	10	48.	10
9.	12	19.	10	29.	7	39.	9	49.	10
10.	9	20.	10	30.	9	40.	9	50.	17

Población 2: Empapado simple

Num. cord	Kgs	Num.cord	Kgs	Num.cord	Kgs	Num. cord	Kgs	Num. cord	Kgs
1.	18	11.	15	21.	12	31.	15	41.	13
2.	15	12.	14	22.	15	32.	18	42.	16
3.	10	13.	16	23.	16	33.	14	43.	14
4.	16	14.	18	24.	14	34.	17	44.	17
5.	15	15.	14	25.	18	35.	13	45.	15
6.	14	16.	17	26.	14	36.	14	46.	15
7.	12	17.	12	27.	14	37.	17	47.	16
8.	15	18.	17	28.	16	38.	16	48.	16
9.	12	19.	14	29.	15	39.	15	49.	17
10.	15	20.	16	30.	12	40.	15	50.	14

Población 3: Empapado doble

Num. cord	Kgs	Num.cord	Kgs	Num.cord	Kgs	Num. cord	Kgs	Num. cord	Kgs
1.	18	11.	17	21.	21	31.	20	41.	17
2.	18	12.	19	22.	23	32.	20	42.	20
3.	19	13.	19	23.	19	33.	18	43.	19
4.	20	14.	17	24.	20	34.	23	44.	20
5.	22	15.	21	25.	19	35.	19	45.	21
6.	18	16.	23	26.	23	36.	22	46.	18
7.	19	17.	23	27.	21	37.	20	47.	19
8.	21	18.	19	28.	18	38.	19	48.	20
9.	23	19.	23	29.	17	39.	21	49.	20
10.	19	20.	22	30.	19	40.	23	50.	23

Se seleccionan dos medidas de corderos de forma aleatoria de cada población. Dichas seis medidas de pesos conformarían el vector observación:

$$y = [10, 9, 16, 14, 19, 22]^t.$$

Estas seis observaciones serán tratadas como muestras aleatorias de tamaño dos de tres poblaciones, normalmente distribuidas con medias μ_1, μ_2 y μ_3 y con varianza σ^2.

Las cuestiones de interés pueden ser planteadas en términos de test de hipótesis independientes como sigue:

(A) $H_0 : \mu_1 = \dfrac{\mu_2 + \mu_3}{2}$ versus $H_1 : \mu_1 \neq \dfrac{\mu_2 + \mu_3}{2}$

(B) $H_0 : \mu_2 = \mu_3$ versus $H_1 : \mu_2 \neq \mu_3$

Nota: La hipótesis nula para ambos contrastes se denotará en adelante como $c_A = 2\mu_1 - \mu_2 - \mu_3$ y $c_B = \mu_2 - \mu_3$

Modelo

Siguiendo la teoría explicada en el Capítulo 1, el vector modelo $\bar{y}_{i\cdot}$ se puede obtener a partir de las estimaciones:

$$\begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{3\cdot} \\ \bar{y}_{3\cdot} \end{bmatrix} = \bar{y}_{1\cdot} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \bar{y}_{2\cdot} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \bar{y}_{3\cdot} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

por lo que el espacio modelo es un subespacio 3-dimensional de un espacio de dimensión 6.

Los vectores unitarios que servirán como sistema de coordenadas ortogonales para el espacio de dimensión 6 correspondiente serían:

$$U_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, U_2 = \frac{1}{\sqrt{12}} \begin{bmatrix} 2 \\ 2 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, U_3 = \frac{1}{\sqrt{4}} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$U_4 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, U_5 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, U_6 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}$$

En este caso, U_1 es la dirección asociada a la media global $\mu = \frac{\mu_1 + \mu_2 + \mu_3}{3}$; U_2 representa la dirección asociada al contraste c_A; U_3 representa la dirección asociada al contraste c_B. Estas tres direcciones constituyen una base del espacio modelo mientras que U_4, U_5 y U_6 constituyen una base ortogonal del espacio de errores.

Se puede concluir que tenemos todos los elementos para el análisis de los contrastes: el vector observación y, el espacio modelo y la dirección U_i asociada a cada hipótesis nula.

Descomposición ortogonal

La descomposición ortogonal del vector observación viene dada por la suma de cada proyección del vector observación sobre cada uno de los ejes de coordenadas del sistema. Los valores de las proyecciones serían calculados como $(y \cdot U_i)U_i$, con $i = 1, ..., 6$.

$$y = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + ... + (y \cdot U_6)U_6$$

Test de hipótesis

Estudiando el contraste c_A para determinar si es adecuado mojar la lana de los corderos para que ganen peso, se tiene que la dirección asociada viene determinada por U_2. Este vector se encuentra en el espacio modelo y tiene un coeficiente de proyección $y \cdot U_2 = -33/\sqrt{12}$.

Si el contraste c_A es igual a cero, el coeficiente de proyección de $y \cdot U_2$ se puede considerar pequeño, con un promedio de cero. En cambio, si c_A es distinto de cero, el coeficiente de proyección se considerará grande. Esto nos permite distinguir los casos $c_A = 0$ y $c_A \neq 0$ a través del estadístico F que se verá en un apartado posterior.

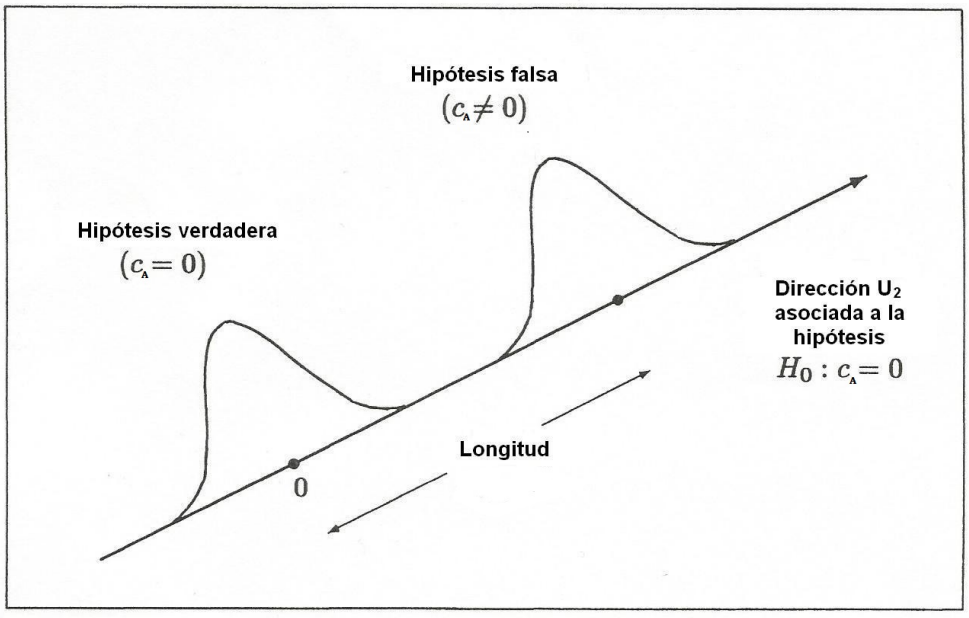

Figura 3.1: Distribución de la proyección del coeficiente $y \cdot U_2$ cuando la hipótesis es verdadera ($c_A = 0$) y cuando la hipótesis es falsa ($c_A \neq 0$).

Ajuste del modelo

En el capítulo anterior, se indicó que el ajuste del modelo viene dado por

$$y = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y - \bar{y}_{i\cdot})$$

$$\begin{bmatrix} 10 \\ 9 \\ 16 \\ 14 \\ 19 \\ 22 \end{bmatrix} = \begin{bmatrix} 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \end{bmatrix} + \begin{bmatrix} -5.5 \\ -5.5 \\ 0 \\ 0 \\ 5.5 \\ 5.5 \end{bmatrix} + \begin{bmatrix} 0.5 \\ -0.5 \\ 1 \\ -1 \\ -1.5 \\ 1.5 \end{bmatrix}$$

$$\text{vector observación} = \text{vector media} + \text{vector contraste} + \text{vector error}$$

Se ha aplicado el método de mínimos cuadrados, donde se tiene que un estimador para la media μ_1 es la media muestral para el tratamiento 1: $\bar{y}_{1\cdot} = (10+9)/2 = 9.5$. De forma similar, los estimadores para μ_2 y μ_3 serían respectivamente: $\bar{y}_{2\cdot} = (16 + 14)/2 = 15$ e $\bar{y}_{3\cdot} = (19 + 22)/2 = 20.5$. Aplicando el mismo método, un estimador para μ sería $\bar{y}_{\cdot\cdot} = 15$. Se puede concluir que el *vector modelo* viene dado por el vector

$$\bar{y}_{i\cdot} = \begin{bmatrix} 9.5 \\ 9.5 \\ 15 \\ 15 \\ 20.5 \\ 20.5 \end{bmatrix}$$

Probar la hipótesis

Para probar que el contraste c_A de nuestra hipótesis es igual a cero, lo único que hacemos es comprobar si la distancia al cuadrado, $(y \cdot U_2)^2$, es comparable o mayor que el promedio de las distancias al cuadrado correspondientes al espacio error. Utilizamos el estadístico F donde $U_4, \ldots, U_6$ son las direcciones del espacio de errores.

$$F = \frac{(y \cdot U_2)^2}{[(y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2]/3} = \frac{(y \cdot U_2)^2}{\|y - \bar{y}_{i\cdot}\|^2 / 3}$$

Sustituyendo los valores para $(y \cdot U_2)^2$ e $\|y - \bar{y}_{i\cdot}\|^2$ se obtiene que

$$F = \frac{(-33/\sqrt{12})^2}{7/3} = \frac{1089}{28} = 38.89$$

Se puede concluir que la hipótesis nula c_A es rechazada, ya que el valor del estadístico F es mayor que el del percentil 0.95 de una $F_{1,3,0.95} = 10.13$. Existen pues, evidencias para afirmar que no tiene el mismo efecto el mojar la lana o no. Posiblemente, ganarán más peso los corderos mojados.

Estimación de σ^2

En el presente caso de estudio, se estima la varianza poblacional σ^2 con las direcciones del espacio modelo U_4, U_5, U_6 ya que las nuevas variables aleatorias construidas $Y \cdot U_1$, $Y \cdot U_2$ e $Y \cdot U_3$, se han utilizado para estimar $\mu = (\mu_1 + \mu_2 + \mu_3)/3$; $c_A = 2\mu_1 - \mu_2 - \mu_3$ y $c_B = \mu_2 - \mu_3$, respectivamente. Esta estimación, se realiza teniendo en cuenta que $Y \cdot U_4$, $Y \cdot U_5$ e $Y \cdot U_6$ siguen una distribución $N[0, \sigma^2]$.

$$s^2 = \frac{(y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2}{3} = \frac{\|y - \bar{y}_{i\cdot}\|^2}{3} = 2.33$$

Intervalo de confianza

En el caso de estudio que nos ocupa, y siguiendo la teoría descrita en el Capítulo 1, $c_A = 2\mu_1 - \mu_2 - \mu_3$. Tomando $\alpha_1 = 2,\ \alpha_2 = -1$ y $\alpha_3 = -1$ junto con las estimaciones de cada media poblacional $\bar{y}_{i\cdot}\ i = 1, \ldots, 3$ se obtiene $\hat{c} = \alpha_1\bar{y}_{1\cdot} + \alpha_2\bar{y}_{2\cdot} + \alpha_3\bar{y}_{3\cdot} = -16.5$, $s = \sqrt{2.33}$, $n = 2$ y $t_{3,0.975} = 2.353$. El resultado que se obtiene para el intervalo de confianza para c_A es de -16.5 ± 6.221.

Estudio de c_B

Con el objeto de realizar el estudio completo de las comparaciones por clases y debido a que hay evidencias para afirmar que $c_A \neq 0$, se debe estudiar el contraste $c_B = 0$. Para ello, basta con calcular el estadístico

$$F = \frac{(y \cdot U_3)^2}{[(y \cdot U_4)^2 + (y \cdot U_5)^2 + (y \cdot U_6)^2]/3} = \frac{(y \cdot U_3)^2}{\|y - \bar{y}_{i\cdot}\|^2/3} = 12.96$$

Este valor obtenido del estadístico F se compara con el percentil 0.95 de una $F_{1,3,0.95} = 10.13$. Al ser el valor del estadístico mayor, se puede deducir que existen evidencias para afirmar que influye en el peso del cordero mojar la lana a distinta edad del mismo.

Intervalo de confianza

En el contraste $c_B = \mu_2 - \mu_3 = 0$, si se sustituyen los valores $\alpha_1 = 0,\ \alpha_2 = 1$ y $\alpha_3 = -1$ junto con las estimaciones de cada media poblacional $\bar{y}_{i\cdot}\ i = 1, \ldots, 3$ se obtiene $\hat{c} = \alpha_1\bar{y}_{1\cdot} + \alpha_2\bar{y}_{2\cdot} + \alpha_3\bar{y}_{3\cdot} = -5.5$, $s = \sqrt{2.33}$, $n = 2$ y $t_{3,0.975} = 2.353$. El intervalo de confianza para el contraste c_B viene determinado por los valores -5.5 ± 6.221.

Capítulo 4

Contraste Factorial

Los contrastes factoriales son apropiados cuando varios factores deben ser investigados de forma simultánea en un sólo experimento. Los contrastes factoriales pueden ser considerados como comparaciones por clases ya que cuando se habla de un *factor*, se refiere a que hay algún criterio que divide los tratamientos en clases, llamados *niveles del factor*.

Al realizar un diseño factorial, los tratamientos se forman combinando los niveles de los factores en estudio; de manera que el efecto de un tratamiento determinado se considera a su vez compuesto por los efectos de los factores y sus interacciones:

$$\text{tratamiento} = \text{efecto factor A} + \text{efecto factor B} + \text{efecto interacción AB}$$

La necesidad de estudiar conjuntamente varios factores obedece a la posibilidad de que el efecto de un factor cambie según los niveles de otros factores, esto es, que los factores interactúen, o exista *interacción*.

La diferencia entre comparaciones por clases y contrastes factoriales se ilustra en la siguiente tabla donde en el Ejemplo 1, el único factor es *especie*, con dos niveles: ovejas y cabras. En el Ejemplo 2, hay dos factores: *especie* y *sexo*. El factor A es *especie* y tiene dos niveles: ovejas y cabras; y el factor B es *sexo* también con dos niveles: macho y hembra. Todas las combinaciones de 2×2 de estos niveles aparecen en la lista de tratamiento.

Comparaciones por clase	**Contrastes factoriales**
Ejemplo 1	**Ejemplo 2**
1. Ovejas Romney	1. Rams (ovejas macho)
2. Ovejas Merino	2. Ewes (ovejas hembra)
3. Cabras Angora	3. Billies (cabras macho)
4. Cabras Feral	4. Does (cabras hembra)

En el Ejemplo 2, cada factor divide los tratamientos en dos grupos de tamaño dos, tal como se ilustra en la Figura 4.1. De ella, se deducen dos contrastes factoriales c_1 y

c_2 tomando como μ_1 la media poblacional de las ovejas macho, μ_2 la media poblacional de las ovejas hembra, μ_3 la media poblacional de las cabras macho y μ_4 la media poblacional de las cabras hembra.

Como primer contraste se tiene:

$$c_1 = \mu_1 + \mu_2 - \mu_3 - \mu_4$$

es decir, ¿es la media de las ovejas igual a la media de las cabras? O lo que es lo mismo, ¿es el efecto del factor *sexo* nulo?

Y como segundo contraste de interés se tiene:

$$c_2 = \mu_1 - \mu_2 + \mu_3 - \mu_4$$

es decir, ¿es la media de los machos igual a la media de las hembras? O lo que es lo mismo, ¿es el efecto del factor *especie* nulo?

La distinción importante entre estos contrastes y los de comparación por clase es que c_1 y c_2 utilizan la información de los cuatro tratamientos experimentales.

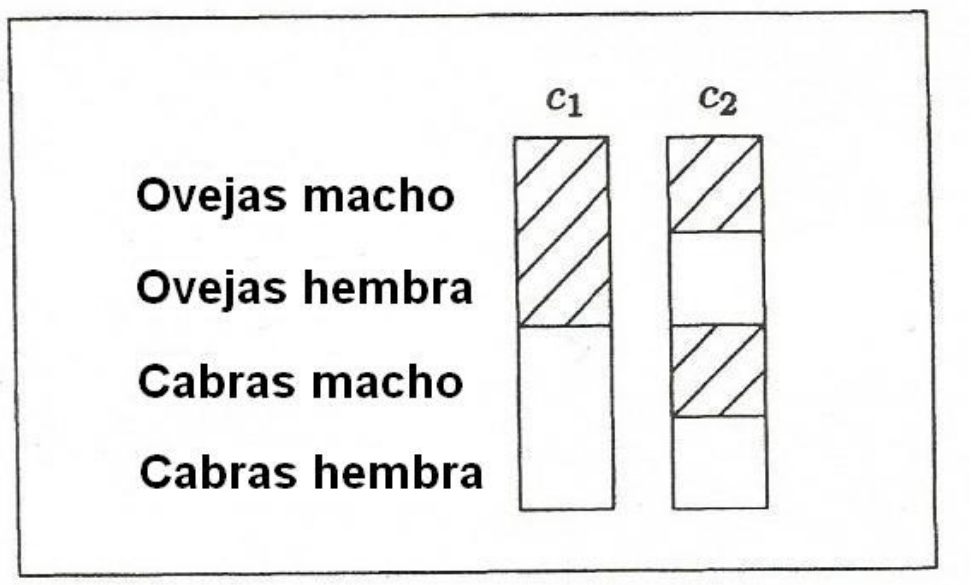

Figura 4.1: Contraste factorial visto como comparaciones por clases usando el Ejemplo 2. Para cada contraste, una clase del tratamiento está con sombra y la otra clase está sin sombra.

Para generar estos contrastes, el procedimiento es, en primer lugar, especificar un conjunto completo de contrastes ortogonales para cada factor. Por ejemplo, si el Factor A tiene tres niveles, un control y dos tratamientos, entonces hay dos contrastes ortogonales, $c_1 = \mu_2 + \mu_3 - 2\mu_1$ y $c_2 = \mu_2 - \mu_3$. Del mismo modo, si el Factor B tiene

cuatro niveles hay tres contrastes ortogonales. En segundo lugar, los contrastes factoriales "efecto principal" se escriben en términos de la lista completa de los tratamientos, de los cuales hay $3 \times 4 = 12$ en nuestro ejemplo, simplemente repitiendo los coeficientes apropiados. Por último, los contrastes "interacción" se generan multiplicando los coeficientes correspondientes por los principales efectos de los contrastes.

Los factores funcionan independientemente uno del otro y por ello se estudia cada factor por separado. Si los factores no funcionan de forma independiente, se obtiene un valioso conocimiento de cómo interactúan desde el análisis de datos. Por estas razones los diseños factoriales son muy populares, útiles y considerados eficientes.

4.1. Caso de estudio

En Noviembre de 1974 se estableció un experimento para comprobar a largo plazo si la Nitrolima y el Superfosfato influyen en el crecimiento de la cebada [2]. El Superfosfato contiene fósforo y azufre, mientras que la Nitrolima añade nitrógeno. Se asignaron cuatro tratamientos a veinte parcelas de modo aleatorio con el siguiente resultado:

Número Parcela:	1	2	3	4	5	6	7	8	9	10
Número Tratamiento:	2	1	1	3	2	1	3	4	4	3
Número Parcela:	11	12	13	14	15	16	17	18	19	20
Número Tratamiento:	1	3	4	2	1	4	3	4	2	2

En este experimento, cada una de las veinte parcelas tenía un tamaño de 40 m de largo por 1.25 m de ancho y tenía siete filas, además eran colindantes. Los rendimientos en kilogramos por parcela de la octava cosecha en Febrero de 1982 se muestran en la siguiente tabla:

Tratamientos	Rendimiento cosecha en kg					Medias
1. Sin fertilizante	19.2	18.4	17.0	17.6	17.2	17.88
2. 250 kg/ha Superfosfato	18.2	19.8	19.4	19.0	19.8	19.24
3. 250 kg/ha Nitrolima	20.0	21.6	22.0	20.8	20.4	20.96
4. 250 S + 250 N kg/ha	23.6	21.6	23.2	21.4	21.2	22.20

Las hipótesis de interés y sus correspondientes contrastes son:

1. ¿Tiene la Nitrolima algún efecto sobre el rendimiento de la cebada?

$$\begin{cases} H_0: \dfrac{\mu_3+\mu_4}{2} = \dfrac{\mu_1+\mu_2}{2} \\ H_1: \dfrac{\mu_3+\mu_4}{2} \neq \dfrac{\mu_1+\mu_2}{2} \end{cases} \qquad \text{donde} \qquad c_1 = \frac{\mu_3+\mu_4}{2} - \frac{\mu_1+\mu_2}{2}$$

2. ¿Tiene el Superfosfato algún efecto sobre el rendimiento de la cebada?

$$\begin{cases} H_0: \ \dfrac{\mu_2+\mu_4}{2} = \dfrac{\mu_1+\mu_3}{2} \\ H_1: \ \dfrac{\mu_2+\mu_4}{2} \neq \dfrac{\mu_1+\mu_3}{2} \end{cases} \qquad \text{donde} \qquad c_2 = \frac{\mu_2+\mu_4}{2} - \frac{\mu_1+\mu_3}{2}$$

3. ¿Interactúan los fertilizantes? O bien, ¿fue la respuesta de Superfosfato en presencia de Nitrolima la misma que en ausencia de Nitrolima?

$$\begin{cases} H_0: \ \mu_4-\mu_3 = \mu_2-\mu_1 \\ H_1: \ \mu_4-\mu_3 \neq \mu_2-\mu_1 \end{cases} \qquad \text{donde} \qquad c_3 = (\mu_4-\mu_3)-(\mu_2-\mu_1)$$

La situación descrita se corresponde con un contraste de tipo factorial 2×2. El primer factor es "Nitrolima", con dos niveles, sin y 250 kg/ha, y el segundo factor es "Superfosfato", también con dos niveles, sin y 250 kg/ha.

A partir de la tabla anteriormente expuesta, se puede deducir que el vector observación está en un espacio de dimensión 20 y viene dado por:

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ \vdots \\ y_{44} \\ y_{45} \end{bmatrix} = \begin{bmatrix} 19.2 \\ 18.4 \\ 17.0 \\ \vdots \\ 21.4 \\ 21.2 \end{bmatrix}$$

Modelo

Se tienen cuatro poblaciones de interés correspondientes a las cuatro combinaciones de los dos factores, cada uno en dos niveles. Se asume que cada uno de los factores se distribuyen según una Normal, con medias μ_1, μ_2, μ_3 y μ_4 y varianza común σ^2.

El espacio modelo está generado por:

$$\begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{1\cdot} \\ \bar{y}_{1\cdot} \\ \bar{y}_{1\cdot} \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{3\cdot} \\ \bar{y}_{3\cdot} \\ \bar{y}_{3\cdot} \\ \bar{y}_{3\cdot} \\ \bar{y}_{3\cdot} \\ \bar{y}_{4\cdot} \\ \bar{y}_{4\cdot} \\ \bar{y}_{4\cdot} \\ \bar{y}_{4\cdot} \\ \bar{y}_{4\cdot} \end{bmatrix} = \bar{y}_{1\cdot} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \bar{y}_{2\cdot} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \bar{y}_{3\cdot} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \bar{y}_{4\cdot} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

con lo que el espacio modelo está formado por un subespacio 4-dimensional de un espacio de dimensión 20.

Test de Hipótesis

Las hipótesis nulas son:

1. Nitrolima no tiene ningún efecto sobre el rendimiento de la cebada.
2. Superfosfato no tiene ningún efecto sobre el rendimiento de la cebada.
3. No hay interacción entre los fertilizantes. Es decir, la respuesta del Superfosfato en ausencia de Nitrolima es la misma que la respuesta del Superfosfato en presencia de Nitrolima.

En términos de contrastes, estas hipótesis son equivalentes a:

$H_0 : c_1 = -\mu_1 - \mu_2 + \mu_3 + \mu_4 = 0$ (Efecto Nitrolima = 0)

$H_0 : c_2 = -\mu_1 + \mu_2 - \mu_3 + \mu_4 = 0$ (Efecto Superfosfato = 0)

$H_0 : c_3 = \ \mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$ (Interacción = 0)

Las direcciones asociadas a estas tres hipótesis que forman un sistema de coordenadas ortogonales para el espacio contraste son:

$$U_2 = \frac{1}{\sqrt{20}} \begin{bmatrix} -1 \\ \vdots\ (10) \\ -1 \\ 1 \\ \vdots\ (10) \\ 1 \end{bmatrix}, \quad U_3 = \frac{1}{\sqrt{20}} \begin{bmatrix} -1 \\ \vdots\ (5) \\ -1 \\ 1 \\ \vdots\ (5) \\ 1 \\ -1 \\ \vdots\ (5) \\ -1 \\ 1 \\ \vdots\ (5) \\ 1 \end{bmatrix}, \quad U_4 = \frac{1}{\sqrt{20}} \begin{bmatrix} 1 \\ \vdots\ (5) \\ 1 \\ -1 \\ \vdots\ (10) \\ -1 \\ 1 \\ \vdots\ (5) \\ 1 \end{bmatrix}$$

Se puede comprobar la idoneidad de la primera de estas direcciones, calculando el coeficiente de proyección $y \cdot U_2 = 5(-\bar{y}_{1\cdot} - \bar{y}_{2\cdot} + \bar{y}_{3\cdot} + \bar{y}_{4\cdot})/\sqrt{20}$.

Por lo tanto, el valor esperado de la variable aleatoria $Y \cdot U_2$ es una constante múltiplo de $c_1 = -\mu_1 - \mu_2 + \mu_3 + \mu_4$. Así, si el contraste c_1 es cero, el coeficiente de proyección $y \cdot U_2$ será pequeño, con un promedio de cero tras realizar muchas repeticiones en el estudio. Por otro lado, si c_1 es distinto de cero, el coeficiente de proyección $y \cdot U_2$ será grande, con un promedio distinto de cero, $5c_1/\sqrt{20}$. Igualmente para $y \cdot U_3$ y $y \cdot U_4$.

Para cada hipótesis, la decisión de si es verdadera o falsa, depende de si la longitud de proyección al cuadrado, $(y \cdot U_2)^2$, $(y \cdot U_3)^2$ o $(y \cdot U_4)^2$, es más grande que la media de la longitud de proyección al cuadrado correspondiente al espacio error.

En nuestro análisis, se tiene el vector observación y que proyectaremos sobre el espacio modelo y dentro de éste, las direcciones U_2, U_3 y U_4 asociadas a las tres hipótesis de interés que generan el espacio de contrastes que es de dimensión 3, ya que existen cuatro tratamientos.

Ajuste del Modelo

Siguiendo la teoría descrita en capítulos anteriores, se proyecta el vector observación y en el espacio modelo para obtener el vector modelo $\bar{y}_{i\cdot} = [17.88, 17.88, \ldots, 22.2, 22.2]^t$ que es el vector de medias estimadas de cada población.

Por lo tanto, la descomposición del vector observación como suma del vector modelo y del vector error es $y = \bar{y}_{i\cdot} + (y - \bar{y}_{i\cdot})$. Cuando el vector de medias global se resta de ambos lados de la ecuación, se llega a la descomposición de forma simplificada:

$$y = \bar{y}_{..} + (\bar{y}_{i\cdot} - \bar{y}_{..}) + (y - \bar{y}_{i\cdot})$$

$$\text{vector observación} = \text{vector media} + \text{vector contraste} + \text{vector error}$$

$$\begin{bmatrix} 19.2 \\ \vdots \\ 18.2 \\ \vdots \\ 20.0 \\ \vdots \\ 23.6 \\ \vdots \end{bmatrix} = \begin{bmatrix} 20.07 \\ \vdots \\ 20.07 \\ \vdots \\ 20.07 \\ \vdots \\ 20.07 \\ \vdots \end{bmatrix} + \begin{bmatrix} -2.19 \\ \vdots \\ -0.83 \\ \vdots \\ 0.89 \\ \vdots \\ 2.13 \\ \vdots \end{bmatrix} + \begin{bmatrix} 1.32 \\ \vdots \\ -1.04 \\ \vdots \\ 0.96 \\ \vdots \\ -1.0 \\ \vdots \end{bmatrix}$$

Prueba de Hipótesis

Como primera hipótesis se tiene que "Nitrolima no tiene ningún efecto sobre el rendimiento de la cebada", o de forma equivalente, que el contraste $c_1 = -\mu_1 - \mu_2 + \mu_3 + \mu_4$ es cero. Para poner a prueba esta hipótesis, tenemos que comprobar si la distancia al cuadrado, $(y \cdot U_2)^2$, es mayor que el promedio de las distancias al cuadrado correspondiente al espacio error. El estadístico F que medirá esta proporción viene definido por:

$$F = \frac{(y \cdot U_2)^2}{[(y \cdot U_5)^2 + \ldots + (y \cdot U_{20})^2]/16} = \frac{(y \cdot U_2)^2}{\|y - \bar{y}_{i\cdot}\|^2 /16} = \frac{45.602}{0.802} = 56.86$$

donde $U_5, \ldots, U_{20}$ son las direcciones que generan el espacio de errores. Como $F = 56.86$ supera el percentil 99 de la distribución $F_{1,16} = 8.53$, rechazamos la hipótesis con un nivel de significación 1 %. Llegamos a la conclusión de que la aplicación de Nitrolima ha modificado el rendimiento de la cebada.

Para la segunda hipótesis, "Superfosfato no tiene ningún efecto sobre el rendimiento de la cebada" o $c_2 = -\mu_1 + \mu_2 - \mu_3 + \mu_4 = 0$, el estadístico F viene dado por:

$$F = \frac{(y \cdot U_3)^2}{\|y - \bar{y}_{i\cdot}\|^2 /16} = \frac{8.450}{0.802} = 10.54$$

Una vez más, supera el percentil 99 de la distribución $F_{1,16}$, por lo que se rechaza la hipótesis con un nivel de significación 1 %. Podemos concluir que la aplicación de Superfosfato también altera el rendimiento de la cebada.

Para la tercera hipótesis, "los fertilizantes no interaccionan" o $c_3 = \mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$, el estadístico F se determina como:

$$F = \frac{(y \cdot U_4)^2}{\|y - \bar{y}_{i\cdot}\|^2 /16} = \frac{0.018}{0.802} = 0.02$$

Esta hipótesis no es estadísticamente significativa, es decir, no hay evidencia de una interacción entre los dos fertilizantes. Los fertilizantes actúan independientemente uno del otro, la respuesta de uno de los fertilizantes no se ve afectada por el hecho de que el otro se aplique.

Estimación de σ^2

Durante todo el análisis, se ha transformado el conjunto original de variables aleatorias independientes,

$$Y_{11}, \ldots, Y_{15} \sim N[\mu_1, \sigma^2], \ldots, Y_{41}, \ldots, Y_{45} \sim N[\mu_4, \sigma^2]$$

en un nuevo conjunto de variables aleatorias (correspondientes a los coeficientes de proyección $Y \cdot U_i$) normales e independientes con medias y varianzas como se indica en la siguiente tabla:

	Media	Varianza
$Y \cdot U_1$	$\sqrt{20}\mu$	σ^2
$Y \cdot U_2$	$\frac{\sqrt{5}(-\mu_1 - \mu_2 + \mu_3 + \mu_4)}{\sqrt{4}}$	σ^2
$Y \cdot U_3$	$\frac{\sqrt{5}(-\mu_1 + \mu_2 - \mu_3 + \mu_4)}{\sqrt{4}}$	σ^2
$Y \cdot U_4$	$\frac{\sqrt{5}(\mu_1 - \mu_2 - \mu_3 + \mu_4)}{\sqrt{4}}$	σ^2
$Y \cdot U_5$	0	σ^2
$\vdots$	$\vdots$	$\vdots$
$Y \cdot U_{20}$	0	σ^2

Las primeras cuatro variables aleatorias se utilizan para estimar la media global $\mu = (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$ y los contrastes c_1, c_2 y c_3. Equivalentemente se puede pensar que los coeficientes de proyección permiten estimar μ_1, μ_2, μ_3 y μ_4. Las 16 variables aleatorias restantes, $Y \cdot U_5, \ldots, Y \cdot U_{20}$, siguen una distribución $N[0, \sigma^2]$ y se utilizan para estimar σ^2 a través del cociente s^2 definido en el Capítulo 1:

$$s^2 = \frac{(y \cdot U_5)^2 + \ldots + (y \cdot U_{20})^2}{16} = \frac{\|y - \bar{y}_{i\cdot}\|^2}{16} = \frac{12.832}{16} = 0.802.$$

Intervalos de Confianza

A continuación se van a obtener intervalos de confianza del 95 % para el promedio de Nitrolima, $c_1 = (-\mu_1 - \mu_2 + \mu_3 + \mu_4)/2$, y para el promedio de Superfosfato, $c_2 = (-\mu_1 + \mu_2 - \mu_3 + \mu_4)/2$. Estos contrastes son de interés, debido a que aparecen a partir

de la ausencia de cualquier interacción significativa de los fertilizantes cuando actúan independientemente uno del otro.

La expresión para calcular el intervalo de confianza del 95 % para un contraste $c = \alpha_1\mu_1 + ... + \alpha_k\mu_k$, viene dada por:

$$(\alpha_1\bar{y}_{1\cdot} + \ldots + \alpha_k\bar{y}_{k\cdot}) \ \pm \ \sqrt{\sum_{i=1}^{k}\alpha_i^2}\frac{s}{\sqrt{n}}t_{k(n-1),0.975}$$

donde el estimador $\hat{c} = \alpha_1\bar{y}_{1\cdot} + \ldots + \alpha_k\bar{y}_{k\cdot}$ y $\hat{C} = \sqrt{\sum_{i=1}^{k}\alpha_i^2(s^2/n)}$ es el error estándar del estimador $\hat{C} = \alpha_1\bar{Y}_{1\cdot} + \ldots + \alpha_k\bar{Y}_{k\cdot}$.

En el caso de la Nitrolima, el promedio del estimador es:

$$\hat{c}_1 = \frac{(\bar{y}_{3\cdot} - \bar{y}_{1\cdot}) + (\bar{y}_{4\cdot} - \bar{y}_{2\cdot})}{2} = \frac{(20.96 - 17.88) + (22.2 - 19.24)}{2} = 3.02$$

Y el error estándar del estimador $\hat{C}_1$ es:

$$\hat{C}_1 = \sqrt{\left[\left(\frac{1}{2} \right)^2 + \left(\frac{-1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 + \left(\frac{-1}{2} \right)^2 \right] \times (0.802/5)}$$

$$= 0.4005$$

Por lo tanto, el intervalo de confianza del 95 % para $c_1 = 3.02 \pm 0.4005 \times 2.120$, donde $2.120 = t_{16,0.975}$. Esto es 3.02 ± 0.85. Es decir, se puede afirmar que al 95 % de confianza el valor medio de la producción bajo el efecto de Nitrolima es 3.02 ± 0.85 kg/parcela.

Del mismo modo, la estimación de la media para el Superfosfato es

$$\hat{c}_2 = \frac{(\bar{y}_{2\cdot} - \bar{y}_{1\cdot}) + (\bar{y}_{4\cdot} - \bar{y}_{3\cdot})}{2} = \frac{(19.24 - 17.88) + (22.2 - 20.96)}{2} = 1.30$$

Calculando el error estándar del estimador como anteriormente, se tiene $\hat{C}_2 = 0.4005$. Por lo tanto, el intervalo de confianza del 95 % para c_2 es $1.30 \pm 0.4005 \times 2.120$. Es decir, se puede concluir que al 95 % de confianza el valor medio de la producción bajo el efecto de Superfosfato es 1.30 ± 0.85 kg/parcela.

Conclusión

Se ha demostrado que usar Nitrolima y Superfosfato influyen en el peso del grano de cebada, aunque con Nitrolima el aumento es mayor. Además, los fertilizantes parecen actuar de forma independiente uno del otro. El intervalo de confianza del 95 % para Nitrolima está centrado en 3.02, y para Superfosfato en 1.30.

Capítulo 5

Contraste Polinomial

En los experimentos estudiados en capítulos anteriores, siempre se han utilizado tratamientos que son de carácter *cualitativo*. Por consiguiente, los contrastes de interés han sido comparaciones de un tipo de tratamiento con otro o incluyendo también interacciones entre factores.En este capítulo se describe la forma de obtener y utilizar contrastes polinomiales ortogonales y componentes polinomiales ortogonales. La necesidad de tales métodos surge cuando todos o algunos de los tratamientos de un experimento se asocian con una *variable cuantitativa* que será representada por X y que tomará los valores x. Se define la *curva de interés* o *curva polinomial* como la que relaciona la media de los tratamientos para los valores de x.

Un ejemplo de este tipo de problemas puede ser el diseño de un experimento donde se propone determinar cómo el rendimiento de grano de cebada sembrado en primavera se vió afectado por la densidad de la siembra [2]. Para ello, se realizó un ensayo completamente al azar con cinco tratamientos y seis repeticiones del mismo. El rendimiento, en Kg/ha de grano cosechado en parcelas de 40 m por 1.25 m, se muestra en la siguiente tabla:

Densidad de siembra	Rendimiento grano						Media
50 kg/ha	5080	4480	5040	4880	4840	4400	4787
75 kg/ha	5240	5240	5040	5280	5000	5560	5227
100 kg/ha	5520	5520	5200	5160	5240	5160	5300
125 kg/ha	5520	5640	5360	5320	5600	5560	5500
150 kg/ha	5440	5640	5360	5120	5440	5520	5420

Tabla 5.1: Rendimiento del grano.

A partir de estos datos, se puede plantear la cuestión, ¿el rendimiento del grano se ve modificado por la densidad de la siembra? El mecanismo para responderla será esencialmente el seguido en capítulos anteriores: los efectos que se desea estudiar corresponden a los contrastes y la prueba de hipótesis se lleva a cabo utilizando el vector unitario correspondiente al espacio del contraste.

De forma general, con k tratamientos o poblaciones con media μ_i y varianza σ^2 se puede ajustar una curva polinomial de la forma:

$$y = \alpha_0 + \alpha_1(x - \bar{x}) + \alpha_2(x - \bar{x})^2 + \ldots + \alpha_{k-1}(x - \bar{x})^{k-1}$$

El procedimiento que se seguirá para determinar la curva polinomial, esto es, la relación entre las medias μ_i y las cantidades x_i será el siguiente:

1. Dado un espacio modelo k-dimensional tenemos que ajustarlo a un modelo polinomial de grado $k-1$ de la forma:

   $$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_{k-1} x^{k-1}$$

   También puede escribirse como:

   $$p(x) = \alpha_0 + \alpha_1(x - \bar{x}) + \alpha_2(x - \bar{x})^2 + \ldots + \alpha_{k-1}(x - \bar{x})^{k-1}$$

   donde los valores de α_i no tienen por qué coincidir.

2. Al poner el modelo en forma vectorial, nos encontramos un sistema de coordenadas no ortogonal para el espacio modelo:

   $$X_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_k - \bar{x} \end{bmatrix}, \quad \ldots, \quad X_k = \begin{bmatrix} (x_1 - \bar{x})^{k-1} \\ \vdots \\ (x_k - \bar{x})^{k-1} \end{bmatrix}$$

   donde $x_1, ..., x_k$ son los k distintos valores de x.

3. Estos vectores pueden ser ortogonalizados utilizando el método de Gram-Schmidt:

   $$\begin{matrix} T_1 \\ \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \end{matrix} \quad \begin{matrix} T_2 \\ \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_k - \bar{x} \end{bmatrix} \end{matrix} \quad \begin{matrix} T_3 \\ \begin{bmatrix} (x_1 - \bar{x})^2 - \frac{M_3}{M_2}(x_1 - \bar{x}) - \frac{M_2}{N} \\ \vdots \\ (x_k - \bar{x})^2 - \frac{M_3}{M_2}(x_k - \bar{x}) - \frac{M_2}{N} \end{bmatrix} \end{matrix} \quad \ldots \quad \begin{matrix} T_k \\ \begin{bmatrix} p_{k-1}(x_1) \\ \vdots \\ p_{k-1}(x_k) \end{bmatrix} \end{matrix}$$

4. Para formar un sistema de coordenadas ortonormal $U_1, \ldots, U_k$ para el espacio modelo, convertimos cada $T_1, \ldots, T_k$ en vectores unitarios:

   $$U_1 = \frac{T_1}{\|T_1\|}, \quad U_2 = \frac{T_2}{\|T_2\|}, \quad \ldots, \quad U_k = \frac{T_k}{\|T_k\|}$$

   Si es necesario, ahora se puede escribir el contraste polinomial correspondiente.

5. Las componentes polinomiales ortogonales resultantes son:

$$
\begin{aligned}
&p_0(x) = 1 && \text{(constante)}\\
&p_1(x) = x - \bar{x} && \text{(lineal)}\\
&p_2(x) = (x_1 - \bar{x})^2 - \frac{M_3}{M_2}(x_1 - \bar{x}) - \frac{M_2}{N} && \text{(cuadrática)}\\
&\vdots\\
&p_{k-1}(x) = (x - \bar{x})^{k-1} - (\text{términos en } x^{k-2}, \ldots, x, 1) && \text{(orden } (k-1)^{th})\\
&\text{donde } N = nk.
\end{aligned}
$$

6. El modelo polinomial ortogonal es:

$$p(x) = \beta_0 + \beta_1(x - \bar{x}) + \beta_2 p_2(x) + \ldots + \beta_{k-1} p_{k-1}(x)$$

7. Ajustamos el modelo de la siguiente forma:

$$\bar{y}_{i\cdot} = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 + \ldots + (y \cdot U_k)U_k$$

   También puede escribirse de la forma:

$$\bar{y}_{i\cdot} = b_0 T_1 + b_1 T_2 + b_2 T_3 + \ldots + b_{k-1} T_k$$

   donde $T_1, \ldots, T_k$ son vectores de componentes polinomiales; por ejemplo, T_2 es el vector $p_1(x_i) = (x_i - \bar{x})$.

8. Por tanto, la ecuación del modelo polinomial ajustada es:

$$q(x) = b_0 + b_1(x - \bar{x}) + b_2 p_2(x) + \ldots + b_{k-1} p_{k-1}(x)$$

   donde $b_0 = (y \cdot U_1)/\left\|T_1\right\|, \ldots, b_{k-1} = (y \cdot U_k)/\left\|T_k\right\|$.

9. La estimación de la varianza es:

$$s^2 = \frac{(y \cdot U_{k+1})^2 + \ldots + (y \cdot U_N)^2}{k(n-1)} = \frac{\left\|y - \bar{y}_{i\cdot}\right\|^2}{k(n-1)}$$

10. Con el fin de comprobar si el coeficiente correspondiente al polinomio de orden t, β_t es cero, usamos la relación $(y \cdot U_{t+1})^2/s^2$ que viene de una distribución $F_{1,N-k}$, bajo $H_0 : \beta_t = 0$.

11. El intervalo de confianza del 95 % para el coeficiente β_t es:

$$b_t \quad \pm \quad \frac{s}{\left\|T_{t+1}\right\|} t_{k(n-1),0.975}$$

5.1. Caso de estudio

A partir de los datos sobre el rendimiento del grano de cebada recogidos en la Tabla 6.1, se obtiene el vector observación definido como:

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{56} \end{bmatrix} = \begin{bmatrix} 5080 \\ 4480 \\ \vdots \\ 5520 \end{bmatrix}$$

Para ver la relación entre el rendimiento del grano y la densidad de siembra, se representa un diagrama de dispersión como se muestra en la Figura 5.1.

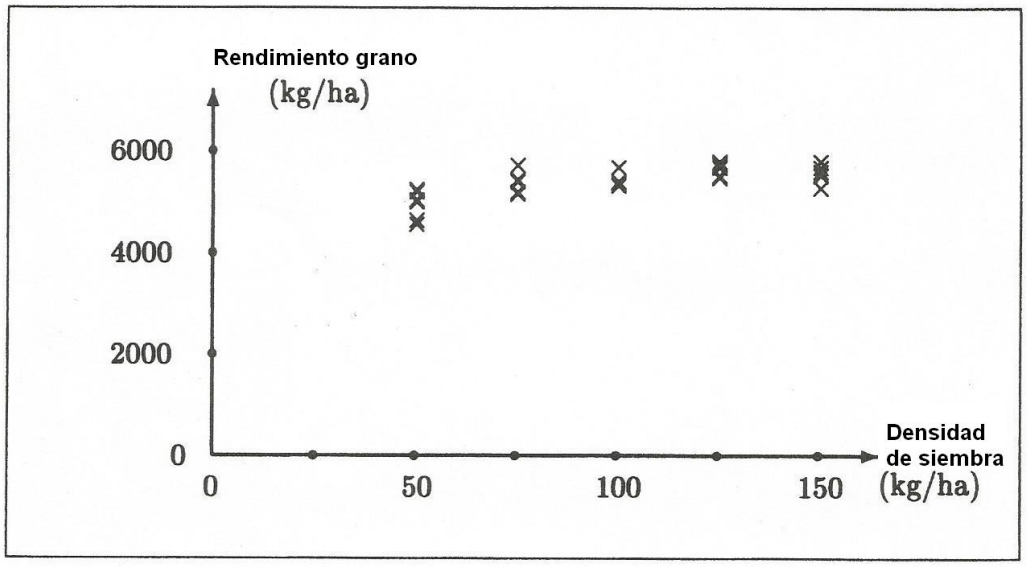

Figura 5.1: Diagrama rendimiento del grano frente densidad de siembra.

A partir de estos datos, las cuestiones de interés son:

1. ¿Aumenta el rendimiento del grano con el aumento de densidad de siembra?
2. ¿Se modifica la razón del crecimiento del rendimiento al aumentar la densidad de siembra?

Con cinco tratamientos se tiene un espacio de dimensión cuatro. ¿Dónde están los vectores unitarios en este espacio contraste que nos permiten poner a prueba estas preguntas? Con el fin de encontrarlos, se considerará una sucesión de polinomios de orden creciente.

Modelo

Suponemos que cada una de las observaciones, y_{ij}, son independientes y se distribuyen según una normal con media μ_i y varianza común σ^2. Esto significa que para cada tipo de siembra, se asume que los rendimientos se distribuyen normalmente con una varianza común.

El espacio modelo viene dado por:

$$\begin{bmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{3\cdot} \\ \vdots \\ \bar{y}_{4\cdot} \\ \vdots \\ \bar{y}_{5\cdot} \\ \vdots \end{bmatrix} = \bar{y}_{1\cdot} \begin{bmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix} + \bar{y}_{2\cdot} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix} + \bar{y}_{3\cdot} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix} + \bar{y}_{4\cdot} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ \vdots \end{bmatrix} + \bar{y}_{5\cdot} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \end{bmatrix}$$

Por lo que dicho espacio modelo es un subespacio 5-dimensional de un espacio de dimensión 30.

Es de interés la curva polinomial cuyo parámetro desconocido es μ_i. Recordemos que una línea recta está determinada únicamente por dos puntos, una ecuación de grado dos por tres puntos, de grado tres por cuatro puntos, y de grado cuatro por cinco puntos. En nuestro caso, los cinco puntos $(50, \mu_1), \ldots, (150, \mu_5)$ determinan un polinomio de grado cuatro,

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4$$

donde $p(x)$ denota el rendimiento medio y x denota la densidad de siembra.

Esta expresión se puede escribir también como

$$\begin{aligned} p(x) &= \alpha_0 + \ \alpha_1(x - \bar{x}) \ + \ \ \alpha_2(x - \bar{x})^2 \ + \ \alpha_3(x - \bar{x})^3 \ + \ \alpha_4(x - \bar{x})^4 \\ &= \alpha_0 + \alpha_1(x - 100) + \alpha_2(x - 100)^2 + \alpha_3(x - 100)^3 + \alpha_4(x - 100)^4 \end{aligned}$$

donde los valores α_i pueden variar entre las dos expresiones. En adelante, se utilizará la segunda expresión.

Entre la Figura 5.1 y la expresión geométrica $y = \mu_i + (y - \mu_i)$, existe cierta conexión: las *alturas* para cada densidad de siembra (50, ..., 150 kg/ha) se corresponden con $\mu_1, \ldots, \mu_5$, que son las entradas de nuestro vector modelo. Puesto que tenemos seis repeticiones en cada densidad de siembra, estas alturas se repiten en el vector modelo. Por lo tanto, nuestro vector modelo puede ser considerado como una versión discreta de la verdadera curva.

Las ecuaciones entre las estimaciones de las alturas $\bar{y}_{1\cdot}, \ldots, \bar{y}_{5\cdot}$ y los coeficientes $\alpha_0, \ldots, \alpha_4$ son:

$$\begin{aligned}
\bar{y}_{1\cdot} &= \alpha_0 + \alpha_1(50-100) + \alpha_2(50-100)^2 + \alpha_3(50-100)^3 + \alpha_4(50-100)^4 \\
\bar{y}_{2\cdot} &= \alpha_0 + \alpha_1(75-100) + \alpha_2(75-100)^2 + \alpha_3(75-100)^3 + \alpha_4(75-100)^4 \\
\bar{y}_{3\cdot} &= \alpha_0 + \alpha_1(100-100) + \alpha_2(100-100)^2 + \alpha_3(100-100)^3 + \alpha_4(100-100)^4 \\
\bar{y}_{4\cdot} &= \alpha_0 + \alpha_1(125-100) + \alpha_2(125-100)^2 + \alpha_3(125-100)^3 + \alpha_4(125-100)^4 \\
\bar{y}_{5\cdot} &= \alpha_0 + \alpha_1(150-100) + \alpha_2(150-100)^2 + \alpha_3(150-100)^3 + \alpha_4(150-100)^4
\end{aligned}$$

dado que $p(x) = \mu_1$ cuando $x = 50$, y así sucesivamente, e $\bar{y}_{\cdot\cdot} = [100, \ldots, 100]$.

Una forma alternativa de escribir el vector modelo es la siguiente:

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_2 \\ \vdots \\ \mu_3 \\ \vdots \\ \mu_4 \\ \vdots \\ \mu_5 \\ \vdots \end{bmatrix} = \alpha_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \end{bmatrix} + \alpha_1 \begin{bmatrix} -50 \\ \vdots \\ -25 \\ \vdots \\ 0 \\ \vdots \\ 25 \\ \vdots \\ 50 \\ \vdots \end{bmatrix} + \alpha_2 \begin{bmatrix} 50^2 \\ \vdots \\ 25^2 \\ \vdots \\ 0^2 \\ \vdots \\ 25^2 \\ \vdots \\ 50^2 \\ \vdots \end{bmatrix} + \alpha_3 \begin{bmatrix} -50^3 \\ \vdots \\ -25^3 \\ \vdots \\ 0^3 \\ \vdots \\ 25^3 \\ \vdots \\ 50^3 \\ \vdots \end{bmatrix} + \alpha_4 \begin{bmatrix} 50^4 \\ \vdots \\ 25^4 \\ \vdots \\ 0^4 \\ \vdots \\ 25^4 \\ \vdots \\ 50^4 \\ \vdots \end{bmatrix}$$

El espacio modelo en un sistema de coordenadas no ortogonal viene dado por:

$$M = \left(\begin{bmatrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{bmatrix}, \begin{bmatrix} -50 \\ \vdots \\ -25 \\ \vdots \\ 0 \\ \vdots \\ 25 \\ \vdots \\ 50 \\ \vdots \end{bmatrix}, \begin{bmatrix} (-50)^2 \\ \vdots \\ (-25)^2 \\ \vdots \\ 0^2 \\ \vdots \\ 25^2 \\ \vdots \\ 50^2 \\ \vdots \end{bmatrix}, \begin{bmatrix} (-50)^3 \\ \vdots \\ (-25)^3 \\ \vdots \\ 0^3 \\ \vdots \\ 25^3 \\ \vdots \\ 50^3 \\ \vdots \end{bmatrix}, \begin{bmatrix} (-50)^4 \\ \vdots \\ (-25)^4 \\ \vdots \\ 0^4 \\ \vdots \\ 25^4 \\ \vdots \\ 50^4 \\ \vdots \end{bmatrix} \right)$$

Para referencias futuras, se denotarán estos vectores como X_1, X_2, X_3, X_4 y X_5.

Coordenadas del Sistema Ortogonal

Para ortogonalizar el sistema de coordenadas $X_1, \ldots, X_5$, se ha utilizado el método tradicional "Gram-Schmidt". En resumen, el método consiste en tomar cada vector, X_i con $i = 1, \ldots, 5$, y convertirlo en ortogonal restando su proyección sobre el espacio. Los vectores ortogonales resultantes, $T_1, \ldots, T_5$, se convierten en vectores unitarios denominados $U_1, \ldots, U_5$, dividiendo cada uno por su módulo.

La dirección del primer eje de coordenadas se elige para que sea $T_1 = X_1$. Por lo tanto,

$$U_1 = \frac{T_1}{\|T_1\|} = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{\sqrt{30}} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Téngase en cuenta que U_1 es el vector unidad unitario y $N = nk$.

El segundo eje de coordenadas, X_2 es ortogonal a X_1. De ahí, que $T_2 = X_2$ y su dirección sea

$$U_2 = \frac{T_2}{\|T_2\|} = \frac{x - \bar{x}}{\|x - \bar{x}\|} = \frac{1}{\sqrt{37500}} \begin{bmatrix} -50 \\ . \\ -25 \\ . \\ 0 \\ . \\ 25 \\ . \\ 50 \end{bmatrix} = \frac{1}{\sqrt{60}} \begin{bmatrix} -2 \\ . \\ -1 \\ . \\ 0 \\ . \\ 1 \\ . \\ 2 \end{bmatrix}$$

donde x es el vector $[x_1, \ldots, x_5] = [50, \ldots, 150]$ y $\bar{x}$ es el vector media global.

Para determinar el tercer eje de coordenadas, U_3, es necesario que el subespacio $< U_1, U_2, U_3 >$ sea igual al subespacio $< X_1, X_2, X_3 >$. Para encontrar un vector en el subespacio $< X_1, X_2, X_3 >$ que sea ortogonal a U_1 y U_2, podemos hacerlo restando al vector X_3 la proyección de X_3 en U_1 y U_2. Obsérvese la Figura 5.2.

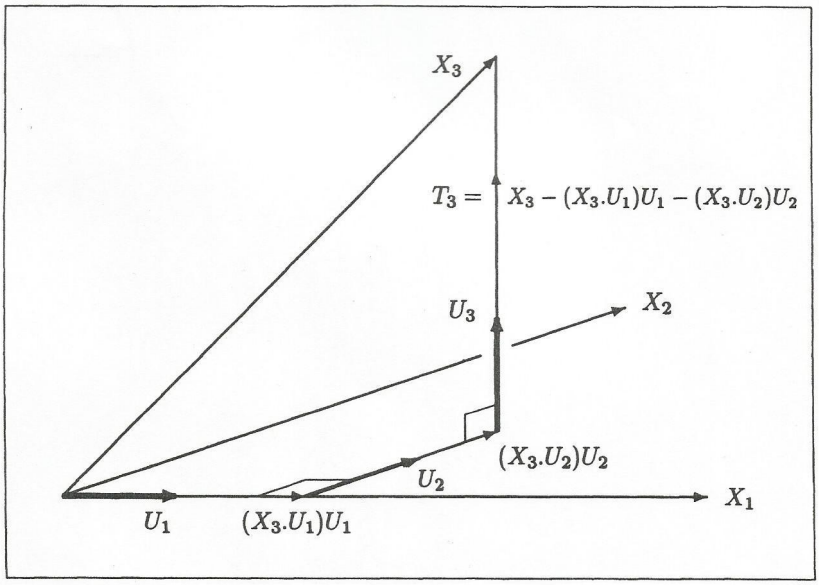

Figura 5.2: Encontrar el tercer vector unitario ortogonal, U_3, restando de X_3 su proyección sobre el subespacio generado por X_1 y X_2.

El vector resultante es $T_3 = X_3 - (X_3 \cdot U_1)U_1 - (X_3 \cdot U_2)U_2 =$

$$\begin{bmatrix} (x_1-\bar{x})^2 \\ \vdots (n) \\ (x_1-\bar{x})^2 \\ \vdots \\ (x_k-\bar{x})^2 \\ \vdots (n) \\ (x_k-\bar{x})^2 \end{bmatrix} - \left(\begin{bmatrix} (x_1-\bar{x})^2 \\ \vdots (n) \\ (x_1-\bar{x})^2 \\ \vdots \\ (x_k-\bar{x})^2 \\ \vdots (n) \\ (x_k-\bar{x})^2 \end{bmatrix} \frac{1}{\sqrt{N}} \begin{bmatrix} 1 \\ \vdots (n) \\ 1 \\ \vdots \\ 1 \\ \vdots (n) \\ 1 \end{bmatrix} \right) \frac{1}{\sqrt{N}} \begin{bmatrix} 1 \\ \vdots (n) \\ 1 \\ \vdots \\ 1 \\ \vdots (n) \\ 1 \end{bmatrix}$$

$$- \left(\begin{bmatrix} (x_1-\bar{x})^2 \\ \vdots (n) \\ (x_1-\bar{x})^2 \\ \vdots \\ (x_k-\bar{x})^2 \\ \vdots (n) \\ (x_k-\bar{x})^2 \end{bmatrix} \frac{1}{\sqrt{M_2}} \begin{bmatrix} x_1-\bar{x} \\ \vdots (n) \\ x_1-\bar{x} \\ \vdots \\ x_k-\bar{x} \\ \vdots (n) \\ x_k-\bar{x} \end{bmatrix} \right) \frac{1}{\sqrt{M_2}} \begin{bmatrix} x_1-\bar{x} \\ \vdots (n) \\ x_1-\bar{x} \\ \vdots \\ x_k-\bar{x} \\ \vdots (n) \\ x_k-\bar{x} \end{bmatrix} =$$

$$= \left(\begin{bmatrix} (x_1-\bar{x})^2 \\ \vdots (n) \\ (x_1-\bar{x})^2 \\ \vdots \\ (x_k-\bar{x})^2 \\ \vdots (n) \\ (x_k-\bar{x})^2 \end{bmatrix} - \frac{M_2}{N} \begin{bmatrix} 1 \\ \vdots (n) \\ 1 \\ \vdots \\ 1 \\ \vdots (n) \\ 1 \end{bmatrix} \right) - \frac{M_3}{M_2} \begin{bmatrix} x_1-\bar{x} \\ \vdots (n) \\ x_1-\bar{x} \\ \vdots \\ x_k-\bar{x} \\ \vdots (n) \\ x_k-\bar{x} \end{bmatrix} =$$

$$= \begin{bmatrix} (x_1-\bar{x})^2 - \frac{M_3}{M_2}(x_1-\bar{x}) - \frac{M_2}{N} \\ \vdots (n) \\ (x_1-\bar{x})^2 - \frac{M_3}{M_2}(x_1-\bar{x}) - \frac{M_2}{N} \\ \vdots \\ (x_k-\bar{x})^2 - \frac{M_3}{M_2}(x_k-\bar{x}) - \frac{M_2}{N} \\ \vdots (n) \\ (x_k-\bar{x})^2 - \frac{M_3}{M_2}(x_k-\bar{x}) - \frac{M_2}{N} \end{bmatrix}$$

donde

$$M_2 = \|x-\bar{x}\|^2 = \sum_{i=1}^{k} n(x_i-\bar{x})^2 = n\sum_{i=1}^{k}(x_i-\bar{x})^2$$

$$M_3 = n\sum_{i=1}^{k}(x_i-\bar{x})^3.$$

Para convertir T_3 en vector unitario, simplemente dividimos por su módulo:

$$U_3 = \frac{T_3}{\|T_3\|} = \frac{X_3 - (X_3 \cdot U_1)U_1 - (X_3 \cdot U_2)U_2}{\| X_3 - (X_3 \cdot U_1)U_1 - (X_3 \cdot U_2)U_2 \|}$$

Para el presente caso de estudio,

$N = 30$

$M_2 = 6 \times 25^2[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2] = 37500$

$M_3 = 6 \times 25^3[(-2)^3 + (-1)^3 + 0^3 + 1^3 + 2^3] = 0$

Así, $\frac{M_2}{N} = 1250$.

Por lo tanto, $T_3 = \begin{bmatrix} (-50)^2 - 1250 \\ \vdots \\ (-25)^2 - 1250 \\ \vdots \\ 0^2 - 1250 \\ \vdots \\ 25^2 - 1250 \\ \vdots \\ 50^2 - 1250 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1250 \\ \vdots \\ -625 \\ \vdots \\ -1250 \\ \vdots \\ -625 \\ \vdots \\ 1250 \\ \vdots \end{bmatrix}$, así que $U_3 = \frac{1}{\sqrt{84}} \begin{bmatrix} 2 \\ \vdots \\ -1 \\ \vdots \\ -2 \\ \vdots \\ -1 \\ \vdots \\ 2 \\ \vdots \end{bmatrix}$

Para obtener U_4 y U_5, continuamos el proceso utilizando las siguientes expresiones:

$$U_4 = \frac{T_4}{\|T_4\|} = \frac{X_4 - (X_4.U_1)U_1 - (X_4.U_2)U_2 - (X_4.U_3)U_3}{\| X_4 - (X_4.U_1)U_1 - (X_4.U_2)U_2 - (X_4.U_3)U_3 \|}$$

$$U_5 = \frac{T_5}{\|T_5\|} = \frac{X_5 - (X_5.U_1)U_1 - (X_5.U_2)U_2 - (X_5.U_3)U_3 - (X_5.U_4)U_4}{\| X_5 - (X_5.U_1)U_1 - (X_5.U_2)U_2 - (X_5.U_3)U_3 - (X_5.U_4)U_4 \|}$$

El sistema de coordenadas ortogonal para el espacio modelo es:

$$\begin{array}{ccccc} U_1 & U_2 & U_3 & U_4 & U_5 \\ \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \\ \vdots \end{bmatrix} & \begin{bmatrix} -2 \\ \vdots \\ -1 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 2 \\ \vdots \end{bmatrix} & \begin{bmatrix} 2 \\ \vdots \\ -1 \\ \vdots \\ -2 \\ \vdots \\ -1 \\ \vdots \\ 2 \\ \vdots \end{bmatrix} & \begin{bmatrix} -1 \\ \vdots \\ 2 \\ \vdots \\ 0 \\ \vdots \\ -2 \\ \vdots \\ 1 \\ \vdots \end{bmatrix} & \begin{bmatrix} 1 \\ \vdots \\ -4 \\ \vdots \\ 6 \\ \vdots \\ -4 \\ \vdots \\ 1 \\ \vdots \end{bmatrix} \\ \hline \sqrt{30} & \sqrt{60} & \sqrt{84} & \sqrt{60} & \sqrt{420} \end{array}$$

Componente Polinomial Ortogonal

La descomposición del polinomio del vector modelo en forma ortogonal utilizando los vectores ortogonales es la siguiente:

$T_1 = X_1$

$$T_2 = X_2$$

$$T_3 = X_3 - (X_3.U_1)U_1 - (X_3.U_2)U_2$$

$$T_4 = X_4 - (X_4.U_1)U_1 - (X_4.U_2)U_2 - (X_4.U_3)U_3$$

$$T_5 = X_5 - (X_5.U_1)U_1 - (X_5.U_2)U_2 - (X_5.U_3)U_3 - (X_5.U_4)U_4$$

Se ha trabajado con los tres primeros:

$$T_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},\ T_2 = \begin{bmatrix} 50-100 \\ \vdots \\ 150-100 \end{bmatrix},\ T_3 = \begin{bmatrix} (50-100)^2 - 1250 \\ \vdots \\ (150-100)^2 - 1250 \end{bmatrix}$$

Al inicio de esta sección, se han utilizado las componentes polinomiales no ortogonales de la curva, α_0, $\alpha_1(x - \bar{x})$,... y así sucesivamente para descomponer el vector modelo en términos de los vectores ortogonales $X_1, \ldots, X_5$. Una vez ortogonalizado, se obtiene $T_1, \ldots, T_5$. Se puede invertir el proceso y usarlas para escribir las *componentes polinomiales ortogonales* de la curva. De T_1, T_2 y T_3 podemos ver que las tres primeras componentes son:

Componente constante: $p_0(x) = 1$

Componente lineal: $p_1(x) = x - 100$

Componente cuadrática: $p_2(x) = (x - 100)^2 - 1250$

La ecuación polinomial ortogonal de la curva es:

$$\begin{aligned} y &= \beta_0 + \beta_1(x-100) + \beta_2[(x-100)^2 - 1250] + \beta_3 p_3(x) + \beta_4 p_4(x) \\ &= \beta_0 p_0(x) + \beta_1 p_1(x) + \beta_2 p_2(x) + \beta_3 p_3(x) + \beta_4 p_4(x) \end{aligned}$$

La suma ortogonal del vector modelo se escribe como:

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_2 \\ \vdots \\ \mu_3 \\ \vdots \\ \mu_4 \\ \vdots \\ \mu_5 \\ \vdots \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \end{bmatrix} + \beta_1 \begin{bmatrix} 50-100 \\ \vdots \\ 75-100 \\ \vdots \\ 100-100 \\ \vdots \\ 125-100 \\ \vdots \\ 150-100 \\ \vdots \end{bmatrix} + \beta_2 \begin{bmatrix} (50-100)^2-1250 \\ \vdots \\ (75-100)^2-1250 \\ \vdots \\ (100-100)^2-1250 \\ \vdots \\ (125-100)^2-1250 \\ \vdots \\ (150-100)^2-1250 \\ \vdots \end{bmatrix} + \ldots$$

$$= \beta_0 \begin{bmatrix} p_0(50) \\ \vdots \\ p_0(75) \\ \vdots \end{bmatrix} + \beta_1 \begin{bmatrix} p_1(50) \\ \vdots \\ p_1(75) \\ \vdots \end{bmatrix} + \beta_2 \begin{bmatrix} p_2(50) \\ \vdots \\ p_2(75) \\ \vdots \end{bmatrix} + \ldots$$

$$= \beta_0 T_1 + \beta_1 T_2 + \beta_2 T_3 + \ldots$$

Para mayor claridad, se han representado las tres primeras componentes en la Figura 5.3.

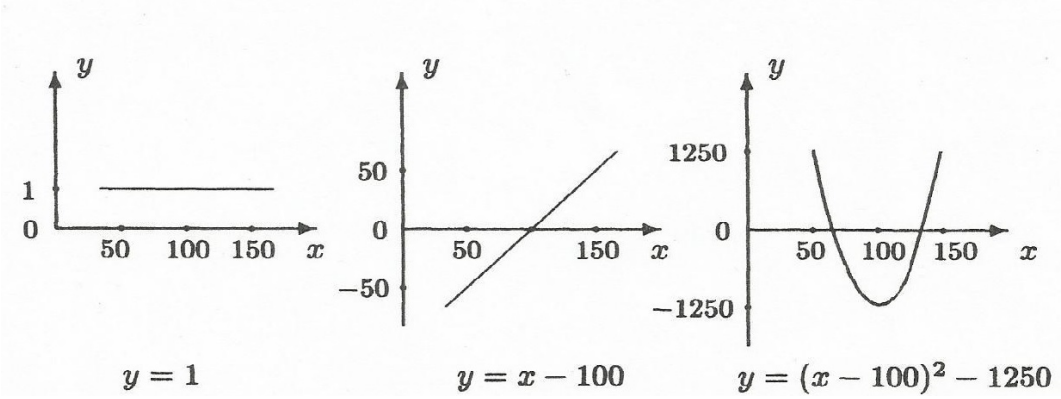

Figura 5.3: Componentes polinomiales ortogonales: constante, lineal y cuadrática.

La componente constante, $p_0(x) = 1$, se utiliza para aproximar la elevación de la curva. La componente lineal, $p_1(x) = x - 100$, se utiliza para aproximar la inclinación y la componente cuadrática, $p_2(x) = (x - 100)^2 - 1250$, para aproximar la curvatura. Los correspondientes coeficientes fijados, b_0, b_1 y b_2, estimarán la elevación, el grado de inclinación y la curvatura de la curva desconocida.

Aproximación Polinómica Sucesiva

La utilidad de las componentes ortogonales radica en la idea de considerar que las posibles curvas a la que se pretende aproximar la curva desconocida son polinomios de orden creciente.

Cuando las ecuaciones polinomiales se escriben en forma ortogonal, los coeficientes estimados, $b_0, \ldots, b_4$, no cambian al aumentar el grado de aproximación. Los valores ajustados correspondientes son la suma de los vectores de proyección, de la forma $\sum(y \cdot U_i)U_i = \sum b_{i-1}T_i$. Para cada incremento en el orden polinomial de la última proyección, $(y \cdot U_i)U_i = b_{i-1}T_i$, cambia el valor ajustado al sumar el término extra. Las aproximaciones polinómicas crecientes se explican en la siguiente tabla:

Orden polinomial	Vector modelo ajustado	Ecuación para el modelo ajustado en forma ortogonal
0	$(y.U_1)U_1 = b_0T_1$	b_0
1	$\sum_{i=1}^{2}(y.U_i)U_i = \sum_{i=1}^{2} b_{i-1}T_i$	$b_0 + b_1(x-100)$
2	$\sum_{i=1}^{3}(y.U_i)U_i = \sum_{i=1}^{3} b_{i-1}T_i$	$b_0 + b_1(x-100)$ $+b_2[(x-100)^2 - 1250]$
3	$\sum_{i=1}^{4}(y.U_i)U_i = \sum_{i=1}^{4} b_{i-1}T_i$	$b_0 + b_1(x-100)$ $+b_2[(x-100)^2 - 1250]$ $+b_3p_3(x)$
4	$\sum_{i=1}^{5}(y.U_i)U_i = \sum_{i=1}^{5} b_{i-1}T_i$	$b_0 + b_1(x-100)$ $+b_2[(x-100)^2 - 1250]$ $+b_3p_3(x) + b_4p_4(x)$

Tabla 5.2: Aproximaciones polinómicas de orden creciente.

Por lo tanto:

$(y \cdot U_2)U_2$ es la parte lineal del modelo ajustado.

$(y \cdot U_3)U_3$ es la parte cuadrática del modelo ajustado.

$(y \cdot U_4)U_4$ es la parte de orden tres del modelo ajustado.

$(y \cdot U_5)U_5$ es la parte de orden cuatro del modelo ajustado.

Para nuestro ejemplo, los vectores son:

$$(y \cdot U_1)U_1 \quad ; \quad (y \cdot U_2)U_2 \quad ; \quad (y \cdot U_3)U_3 \quad ; \quad (y \cdot U_4)U_4 \quad ; \quad (y \cdot U_5)U_5$$

$$\begin{bmatrix} 5247 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 5247 \end{bmatrix} ; \begin{bmatrix} -308 \\ \vdots \\ -154 \\ \vdots \\ 0 \\ \vdots \\ 154 \\ \vdots \\ 308 \\ \vdots \end{bmatrix} ; \begin{bmatrix} -130 \\ \vdots \\ 65 \\ \vdots \\ 130 \\ \vdots \\ 65 \\ \vdots \\ -130 \\ \vdots \end{bmatrix} ; \begin{bmatrix} -9 \\ \vdots \\ 17 \\ \vdots \\ 0 \\ \vdots \\ -17 \\ \vdots \\ 9 \\ \vdots \end{bmatrix} ; \begin{bmatrix} -13 \\ \vdots \\ 51 \\ \vdots \\ -77 \\ \vdots \\ 51 \\ \vdots \\ -13 \\ \vdots \end{bmatrix}$$

Contraste Polinomial

En capítulos anteriores, siempre se ha escrito el contraste en primer lugar, y después el vector unitario correspondiente debido a que el contraste es la entidad básica de interés. En este caso, es diferente pues se han obtenido los vectores unitarios correspondientes a las hipótesis de interés sin ninguna referencia al contraste. Se puede concluir pues, que no es necesario especificar ningún tipo de contraste.

Sin embargo, se pueden plantear y resultan útiles cuando el diseño experimental mezcla contrastes polinomiales mixtos (cuando sólo algunos de los tratamientos experimentales son de tipo cantidad) con otros tipos de contrastes.

En este caso, a partir de los vectores unitarios U_1, U_2, U_3, U_4, U_5 y teniendo en cuenta la siguiente tabla, se muestran los contrastes polinomiales ortogonales para $k = 3$ hasta $k = 6$ tratamientos:

$c_1 = -2\mu_1 - \mu_2 \quad + \mu_4 + 2\mu_5$ (contraste lineal)

$c_2 = \ 2\mu_1 - \mu_2 - 2\mu_3 - \mu_4 + 2\mu_5$ (contraste cuadrático)

$c_3 = -\mu_1 + 2\mu_2 \quad - 2\mu_4 + \mu_5$ (contraste de orden tres)

$c_4 = \ \mu_1 - 4\mu_2 + 6\mu_3 - 4\mu_4 + \mu_5$ (contraste de orden cuatro)

Número de valores x igualmente espaciados	**Contraste Polinomial** Lineal	Cuadrática	Cúbica	Cuarta	Quinta
3	$\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$			
4	$\begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -1 \\ 3 \\ -3 \\ 1 \end{bmatrix}$		
5	$\begin{bmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 2 \\ -1 \\ -2 \\ -1 \\ 2 \end{bmatrix}$	$\begin{bmatrix} -1 \\ 2 \\ 0 \\ -2 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -4 \\ 6 \\ -4 \\ 1 \end{bmatrix}$	
6	$\begin{bmatrix} -5 \\ -3 \\ -1 \\ 1 \\ 3 \\ 5 \end{bmatrix}$	$\begin{bmatrix} 5 \\ -1 \\ -4 \\ -4 \\ -1 \\ 5 \end{bmatrix}$	$\begin{bmatrix} -5 \\ 7 \\ 4 \\ -4 \\ -7 \\ 5 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -3 \\ 2 \\ 2 \\ -3 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -1 \\ 5 \\ -10 \\ 10 \\ -5 \\ 1 \end{bmatrix}$

Tabla 5.3: Conjunto de contrastes polinomial ortogonal para réplicas y valores de x igualmente espaciados.

Test de Hipótesis

El objetivo de estudio es comprobar si se incrementa el rendimiento del grano con el aumento de densidad de siembra, para lo cual, se plantea el test de hipótesis:

$$H_0 : \beta_1 = 0 \text{ frente } H_1 : \beta_1 \neq 0$$

La hipótesis nula establece que la inclinación de la curva es cero. Es decir, no hay relación entre el incremento del rendimiento del grano y el aumento de la densidad de siembra.

La segunda pregunta es si la razón entre el incremento del rendimiento del grano y el aumento de densidad de siembra permanece constante se traduce en el test de hipótesis:

$$H_0 : \beta_2 = 0 \text{ frente } H_1 : \beta_2 \neq 0$$

Esta hipótesis nula indica que la curvatura de la ecuación cuadrática es cero.

De menor interés son las hipótesis nulas que corresponden a las componentes de tercer y cuarto grado:

$$H_0 : \beta_3 = 0 \text{ frente } H_1 : \beta_3 \neq 0 \quad \text{y} \quad H_0 : \beta_4 = 0 \text{ frente } H_1 : \beta_4 \neq 0$$

Las direcciones asociadas a estas cuatro hipótesis son U_2, U_3, U_4 y U_5. Para demostrarlo, en primer lugar, se expone que el promedio para el coeficiente de proyección $y \cdot U_2$ es un escalar múltiplo de β_1. Entonces si $\beta_1 = 0$, el coeficiente de proyección de $y \cdot U_2$ será pequeño, con un promedio de cero, mientras que si $\beta_1 \neq 0$ el coeficiente de proyección será grande.

Como, $y \cdot U_2 = \begin{bmatrix} y_{11} \\ \vdots \\ y_{21} \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} -2 \\ \vdots \\ -1 \\ \vdots \end{bmatrix} / \sqrt{60} = 6[-2\bar{y}_{1\cdot} - \bar{y}_{2\cdot} + \bar{y}_{4\cdot} + 2\bar{y}_{5\cdot}]/\sqrt{60}$, por lo que el promedio de la variable aleatoria $Y \cdot U_2 = 6[-2\mu_1 - \mu_2 + \mu_4 + 2\mu_5]/\sqrt{60}$, que puede escribirse como $\mu_i \cdot U_2$, donde μ_i denota el vector modelo $[\mu_1, ., \mu_2, \ldots, \mu_5]^t$. Si ampliamos el vector modelo de la forma $\beta_0 T_1 + \beta_1 T_2 + \beta_2 T_3 + \beta_3 T_4 + \beta_4 T_5$, donde las T_i con $i = 1, \ldots, 5$ son ortogonales entre sí, nos encontramos que

$$\begin{aligned} \mu_i \cdot U_2 &= (\beta_0 T_1 + \beta_1 T_2 + \beta_2 T_3 + \beta_3 T_4 + \beta_4 T_5) \cdot U_2 \\ &= \beta_1 (T_2 \cdot U_2) = \beta_1 \|T_2\| = \beta_1 \|x - \bar{x}\| = 25\sqrt{60}\beta_1 \end{aligned}$$

es un escalar múltiplo de β_1.

Del mismo modo, las variables aleatorias $Y \cdot U_3$, $Y \cdot U_4$ y $Y \cdot U_5$ presentan valores esperados múltiplos de β_2, β_3 y β_4 respectivamente. Éstos son $\beta_2 \|T_3\| = 625\sqrt{84}\beta_2$, $\beta_3 \|T_4\| = 18750\sqrt{60}\beta_3$ y $\beta_4 \|T_5\| = 133929\sqrt{420}\beta_4$.

Estos resultados acerca de la distribución de los coeficientes de proyección, $Y \cdot U_i$, se resumen en la siguiente tabla, junto con los valores esperados de los coeficientes de proyección en términos de contrastes $c_1, \ldots, c_4$; éstos se han calculado utilizando los mismos métodos que en capítulos anteriores.

	Valor esperado		
	(a) Usando valores β_i	(b) Usando valores μ_i	Varianza
$Y \cdot U_1$	$\sqrt{30}\beta_0$	$\sqrt{30}\mu$	σ^2
$Y \cdot U_2$	$25\sqrt{60}\beta_1$	$\frac{\sqrt{6}(-2\mu_1 - \mu_2 + \mu_4 + 2\mu_5)}{\sqrt{10}}$	σ^2
$Y \cdot U_3$	$625\sqrt{84}\beta_2$	$\frac{\sqrt{6}(2\mu_1 - \mu_2 - 2\mu_3 - \mu_4 + 2\mu_5)}{\sqrt{14}}$	σ^2
$Y \cdot U_4$	$18750\sqrt{60}\beta_3$	$\frac{\sqrt{6}(-\mu_1 + 2\mu_2 - 2\mu_4 + \mu_5)}{\sqrt{10}}$	σ^2
$Y \cdot U_5$	$133929\sqrt{420}\beta_4$	$\frac{\sqrt{6}(\mu_1 - 4\mu_2 + 6\mu_3 - 4\mu_4 + \mu_5)}{\sqrt{70}}$	σ^2
$Y \cdot U_6$	0	0	σ^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y \cdot U_{30}$	0	0	σ^2

Tabla 5.4: Distribución de los coeficientes de proyección $Y \cdot U_i$, mostrando la relación polinomial entre coeficientes $\beta_1, \ldots \beta_4$ y contrastes $c_1, \ldots, c_4$.

Claramente los contrastes $c_1, \ldots, c_4$ son simplemente múltiplos de $\beta_1, \ldots, \beta_4$ respectivamente, por lo que la hipótesis $H_0 : \beta_1 = 0$ es equivalente a $H_0 : c_1 = 0$, y así sucesivamente.

Se puede concluir que tenemos el espacio modelo, las direcciones asociadas a nuestras hipótesis y el vector observación y para nuestro análisis.

Ajuste del Modelo

Siguiendo el procedimiento de capítulos anteriores, al proyectar el vector observación y sobre el espacio modelo, se obtiene el vector modelo, $\bar{y}_{i\cdot} = [4787, 4787, \ldots, 5420, 5420]^t$, el vector de medias de los tratamientos. El modelo ajustado es:

$$y = \bar{y}_{..} + (\bar{y}_{i\cdot} - \bar{y}_{..}) + (y - \bar{y}_{i\cdot})$$

$$\text{vector observación} = \text{vector media} + \text{vector contraste} + \text{vector error}$$

$$\begin{bmatrix} 5080 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 5520 \end{bmatrix} = \begin{bmatrix} 5247 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 5247 \end{bmatrix} + \begin{bmatrix} -460 \\ \vdots \\ -20 \\ \vdots \\ 53 \\ \vdots \\ 253 \\ \vdots \\ 173 \\ \vdots \end{bmatrix} + \begin{bmatrix} 293 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 100 \end{bmatrix}$$

donde $\bar{y}_{..} = [5247, \ldots, 5247]^t$, $\bar{y}_{1\cdot} = 4787$, $\bar{y}_{2\cdot} = 5227$, $\bar{y}_{3\cdot} = 5300$, $\bar{y}_{4\cdot} = 5500$ e $\bar{y}_{5\cdot} = 5420$.

El vector contraste se encuentra en un subespacio 4-dimensional de un espacio de dimensión 30 generado por U_2, U_3, U_4 y U_5, y el vector error se encuentra en un subespacio 25-dimensional.

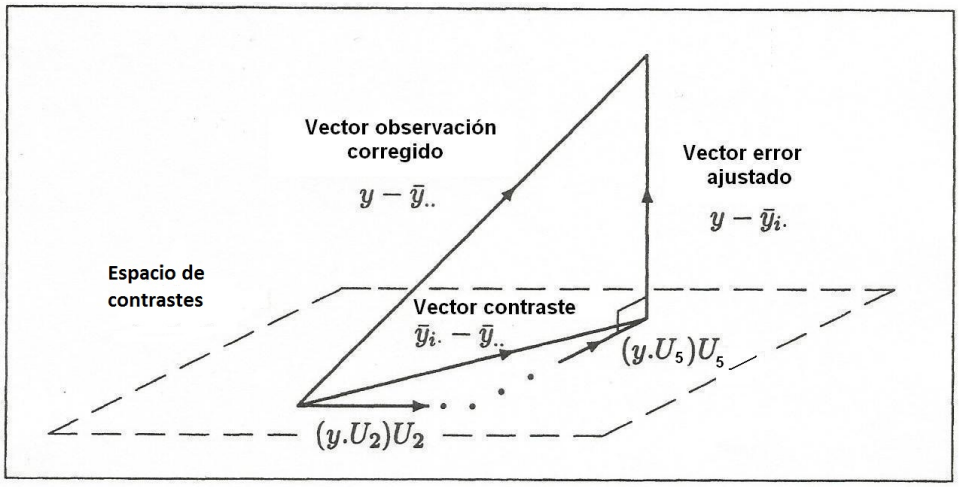

Figura 5.4: Descomposición ortogonal para contrastes polinomiales, descomponiendo el vector contraste en lineal, cuadrático, tercer y cuarto grado.

Como se ilustra en la Figura 5.4, cuando el vector contraste está escrito en términos de sus proyecciones U_2, U_3, U_4 y U_5, se obtiene la descomposición ortogonal,

$$y - \bar{y}_{..} = 1193U_2 - 598U_3 + 67U_4 - 264U_5 + (y - \bar{y}_{i\cdot})$$

Comprobación de Hipótesis

Siguiendo la teoría de capítulos anteriores, se utilizará el estadístico que comprueba para cada hipótesis si el cuadrado de la distancia, $(y \cdot U_i)^2$, es mayor que la media de las distancias al cuadrado del espacio error. Por ejemplo, ponemos a prueba nuestra primera hipótesis, el rendimiento del grano no cambia con la densidad de siembra, al comparar el valor del estadístico

$$F = \frac{(y \cdot U_2)^2}{[(y \cdot U_6)^2 + \ldots + (y \cdot U_{30})^2]/25} = \frac{(y \cdot U_2)^2}{\|y - \bar{y}_{i\cdot}\|^2/25} = \frac{1422960}{39755} = 35.79$$

con el valor del percentil 0.90 de la distribución $F_{1,25} = 7.77$, se puede concluir que existen evidencias para rechazar la hipótesis nula con un nivel de significación del 1 %, es decir, que el rendimiento del grano se ve modificado con la densidad de siembra.

La descomposición de Pitágoras viene dada por:

$$\begin{array}{ccccccccccc} \|y - \bar{y}_{\cdot\cdot}\|^2 & = & (y \cdot U_2)^2 & + & (y \cdot U_3)^2 & + & (y \cdot U_4)^2 & + & (y \cdot U_5)^2 & + & \|y - \bar{y}_{i\cdot}\|^2 \\ 2848267 & = & 1422960 & + & 357505 & + & 4507 & + & 69429 & + & 993867 \end{array}$$

Esta descomposición, y el valor del estadístico para las cuatro hipótesis, se resumen en la Tabla ANOVA presentada a continuación:

Fuente de Variación	df	SS	MS	F
Tratamiento	4	1854400		
$H_0 : \beta_1 = 0$ (lineal)	1	1422960	1422960	35.79 (**)
$H_0 : \beta_2 = 0$ (cuadrática)	1	357505	357505	8.99 (**)
$H_0 : \beta_3 = 0$ (orden tres)	1	4507	4507	0.11
$H_0 : \beta_4 = 0$ (orden cuatro)	1	69429	69429	1.75
Error	25	993867	39755	
Total	29	2848267		

De la tabla anterior, se puede concluir que los coeficientes β_1 y β_2 de la función lineal y cuadrática son distintos de cero. Sin embargo, no tenemos ninguna prueba de que los coeficientes (β_3 y β_4) de tercer y cuarto grado sean distintos de cero.

En cuanto a las cuestiones de interés, podemos concluir: (1) el rendimiento del grano aumenta con el incremento de la densidad de siembra, y (2) la razón entre el incremento del rendimiento del grano y el aumento de densidad de siembra permanece constante.

Ecuación Ajustada

Siguiendo la conclusión anteriormente expuesta, se ha decidido que un polinomio de segundo grado es suficiente para que se aproxime a la curva que relaciona el rendimiento del grano con la densidad de siembra. ¿Cómo podemos encontrar la ecuación cuadrática para este caso? En la Tabla 5.2 tenemos las siguientes igualdades:

$$\begin{aligned}
&\text{Vector constante} && = (y \cdot U_1)U_1 = b_0 T_1 = b_0 1 \\
&\text{Vector lineal} && = (y \cdot U_2)U_2 = b_1 T_2 = b_1 (x - 100) \\
&\text{Vector cuadrático} && = (y \cdot U_3)U_3 = b_2 T_3 = b_2 [(x - 100)^2 - 1250] \\
&\text{Vector orden tres} && = (y \cdot U_4)U_4 = b_3 T_4 = b_3 p_3(x) \\
&\text{Vector orden cuatro} && = (y \cdot U_5)U_5 = b_4 T_5 = b_4 p_4(x)
\end{aligned}$$

donde "1", $(x - 100)$ y así sucesivamente, denotan vectores de un espacio de dimensión 30. Se conocen los valores $y \cdot U_i$: $y \cdot U_1 = 28737$, $y \cdot U_2 = 1193$, $y \cdot U_3 = -598$, $y \cdot U_4 = 67$ e $y \cdot U_5 = -264$. Escribiendo los tres primeros términos como vectores, se nos permite calcular b_0, b_1 y b_2 y, por lo tanto, escribir la ecuación requerida.

En primer lugar, se examina el vector constante:

$$(y \cdot U_1)U_1 = \bar{y}_{..} = [5247, \ldots, 5247, \ldots, 5247, \ldots, 5247, \ldots, 5247]^t$$

$$= [b_0, \ldots, b_0, \ldots, b_0, \ldots, b_0, \ldots, b_0]^t$$

que nos dice que $b_0 = \bar{y}_{..} = 5247$.

En segundo lugar, se examinará el vector lineal

$$(y \cdot U_2)U_2 = 1193 \begin{bmatrix} -2 \\ \vdots \\ -1 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 2 \end{bmatrix} / \sqrt{60} = b_1 \begin{bmatrix} -50 \\ \vdots \\ -25 \\ \vdots \\ 0 \\ \vdots \\ 25 \\ \vdots \\ 50 \end{bmatrix}$$

obteniéndose el coeficiente $b_1 = (y \cdot U_2/\sqrt{60})/25 = 6.16$.

En tercer lugar, examinamos el vector cuadrático

$$(y \cdot U_3)U_3 = -598 \begin{bmatrix} 2 \\ \vdots \\ -1 \\ \vdots \\ -2 \\ \vdots \\ -1 \\ \vdots \\ 2 \end{bmatrix} / \sqrt{84} = b_2 \begin{bmatrix} 1250 \\ \vdots \\ 1250 \\ \vdots \\ 1250 \\ \vdots \\ 1250 \\ \vdots \\ 1250 \end{bmatrix}$$

obteniéndose el coeficiente $b_2 = (y \cdot U_3/\sqrt{84})/625 = -0.104$. En general, $b_{i-1} = y \cdot U_i / \|T_i\|$, para $i = 1, ..., 5$.

El resultado de la ecuación cuadrática ajustada es:

$$y = b_0 + b_1(x - 100) + b_2[(x - 100)^2 - 1250]$$

$$= 5250 + 6.16(x - 100) - 0.104[(x - 100)^2 - 1250]$$

En la Figura 5.5, se representa la curva ajustada sobre el diagrama de dispersión original:

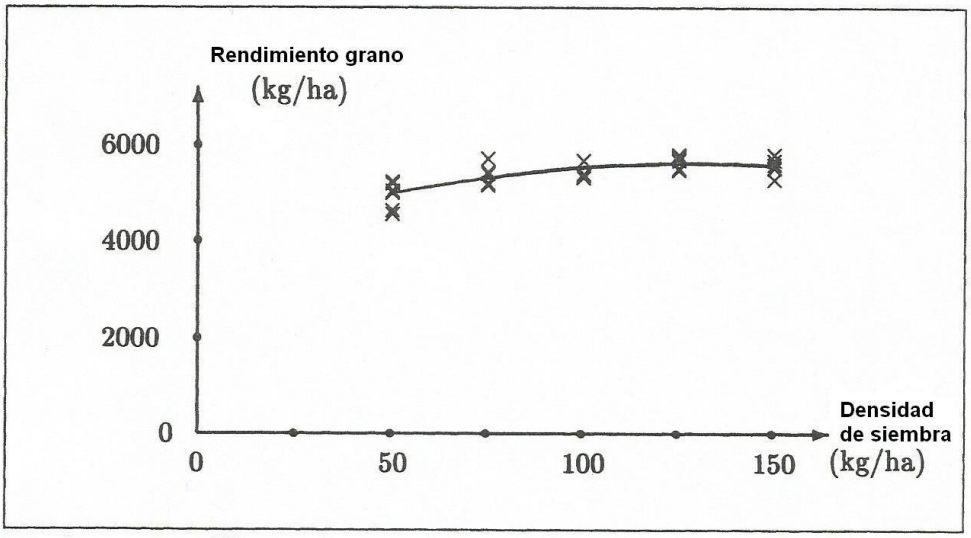

Figura 5.5: Diagrama de dispersión para el vector cuadrático ajustado.

Estimación de σ^2

Al igual que en capítulos anteriores, se han transformado las variables aleatorias originales $Y_{ij} \sim N[\mu_i, \sigma^2]$, en un nuevo conjunto de variables aleatorias normales e independientes, $Y \cdot U_1, \ldots, Y \cdot U_{30}$, con medias y varianzas según la Tabla 6.4.
Las cinco primeras, se utilizan para estimar los coeficientes polinomiales β_0, β_1, β_2, β_3 y β_4; o equivalentemente la media global μ y los contrastes c_1, c_2, c_3 y c_4; o dicho de otro modo: las medias poblacionales μ_1, μ_2, μ_3, μ_4 y μ_5. Las 25 variables aleatorias restantes se utilizan para estimar σ^2 a través de:

$$s^2 = \frac{(y \cdot U_6)^2 + \ldots + (y \cdot U_{30})^2}{25} = \frac{\|y - \bar{y}_{i\cdot}\|^2}{25} = \frac{993867}{25} = 39755$$

Hay que tener en cuenta que, a pesar de que hemos decidido aproximar la curva con una función cuadrática, no se ha usado $(y \cdot U_4)^2$ ni $(y \cdot U_5)^2$ para la estimación del error.

Intervalo de Confianza

A continuación, se calcularán los intervalos de confianza con un nivel de significación del 95 % para los coeficientes lineales y cuadráticos, β_1 y β_2, utilizando los resultados de la Tabla 5.4.

Para el coeficiente lineal, β_1, se utiliza el hecho de que $y \cdot U_2 = 25\sqrt{60}b_1$ proviene de la distribución $N(25\sqrt{60}\beta_1, \sigma^2)$. Por lo tanto, la cantidad $25\sqrt{60}(b_1 - \beta_1)$ procede de la distribución $N(0, \sigma^2)$, donde $t = 25\sqrt{60}(b_1 - \beta_1)/s$ sigue una distribución t_{25}. Para obtener el intervalo de confianza al 95 % para β_1, el valor observado t se encuentra entre los percentiles 2.5 y 97.5 de la distribución t_{25}. Es decir,

$$-2.060 \;\leq\; \frac{25\sqrt{60}(b_1 - \beta_1)}{s} \;\leq\; 2.060$$

Esto se traduce en que el intervalo de confianza es:

$$b_1 - \frac{s}{25\sqrt{60}}2.060 \;\leq\; \beta_1 \;\leq\; b_1 + \frac{s}{25\sqrt{60}}2.060$$

Se tiene que $b_1 = 6.16$ y $s = \sqrt{39755} = 199$. Por lo tanto, el intervalo de confianza para β_1 es 6.16 ± 2.12, en unidades de kg de grano/ha.

Igualmente, se procede para el coeficiente β_2 de la ecuación cuadrática. El valor de $t = 25^2\sqrt{84}(b_2 - \beta_2)/s$ se encuentra entre los percentiles 2.5 y 97.5 de la distribución t_{25}. El intervalo de confianza queda:

$$b_2 - \frac{s}{25^2\sqrt{84}}2.060 \;\leq\; \beta_2 \;\leq\; b_2 + \frac{s}{25^2\sqrt{84}}2.060$$

Teniendo en cuenta que $b_2 = -0.104$, el intervalo de confianza para β_2 es -0.104 ± 0.072.

El intervalo de confianza para β_1 puede ser interpretado como que al aumentar la densidad de siembra por kg/ha aumenta el rendimiento del grano a 6.16 ± 2.12kg/ha,

como promedio entre 50 y 150 kg de semilla por hectárea. Sin embargo, con el intervalo de confianza para β_2 se puede concluir que con un promedio de 6.16, el rendimiento del grano disminuye a medida que la densidad de siembra se incrementa.

Bibliografía

[1] Bryant, P.: Geometry, Statistics, Probability: Variations on a Common Theme. *Amer. Statist.* **38**(1), 38-48 (1984).

[2] Saville, D. J.; Wood, G. R.: *Statistical Methods: The Geometric Approach*. Springer, 1991.

[3] Saville, D. J.; Wood, G. R.: *Statistical Methods: A Geometric Primer*. Springer, 1996.