

ANÁLISIS DE EXPERIMENTOS ESTADÍSTICOS

USANDO R

Ricardo Alvarado Barrantes

Marzo 2024

*Dedico esta obra a mis estudiantes,
quienes atraídos por mis lecciones
expresaron su agradecimiento,
que tomo y ahora devuelvo*

Prefacio

Durante varios años he venido enseñando el curso Diseño de Experimentos, el cual forma parte del programa de Bachillerato en Estadística de la Universidad de Costa Rica. Actualmente este curso se ubica en el tercer semestre de la carrera de Estadística. Este curso junto con otros dos llamados Modelos de Regresión Aplicados y Modelos Lineales Avanzados, son una secuencia que pretende desarrollar en los estudiantes las habilidades para planear y conducir adecuadamente estudios observacionales o experimentales con validez estadística, utilizando los modelos matemáticos apropiados para el análisis de los datos obtenidos en los estudios planteados.

En el presente libro se dan los fundamentos de los diseños experimentales, se empieza con diseños que contienen un solo factor y se introduce la técnica de análisis de varianza para realizar pruebas de hipótesis sobre igualdad de promedios. En el segundo capítulo se profundiza en las comparaciones múltiples entre pares de promedios y en contrastes ortogonales; en este libro se da especial atención al uso de vectores para el cálculo de los estadísticos que se utilizan en las pruebas de hipótesis y en la construcción de intervalos de confianza. Los capítulos tercero y cuarto se dedican a los diseños factoriales, donde se da especial atención al concepto de interacción entre dos o más factores. El capítulo quinto se centra en diferentes aspectos del uso de bloques, tales como los bloques aleatorizados, las parcelas divididas y los bloques incompletos. También se hace mención a la relación que tiene este tipo de diseños con el uso de modelos mixtos. Más adelante, en el capítulo sexto se estudia el uso de covariables que permiten reducir la variabilidad del error experimental. Finalmente, el último capítulo se concentra en el concepto de potencia de las prueba estadísticas, y se presenta la técnica de simulaciones como herramienta para hacer estudios relacionados con la potencia.

Aunque el propósito del libro es servir como un manual de referencia para el análisis de datos provenientes de experimentos básicos, utilizando el lenguaje de programación R (R Core Team, 2023), se ha dedicado una sección al principio de cada capítulo para explicar los conceptos más relevantes en el tema que se desarrolla. También se explican los modelos matemáticos y su escritura adecuada, de tal forma que el lector pueda no solo llevar a cabo el análisis de datos, sino que comprenda esta parte matemática que sirve para el planteamiento adecuado de hipótesis y, eventualmente, como base para la generación de datos, tal como sucede en el último capítulo donde se utiliza el enfoque de simulación de datos para comprender el concepto de potencia de una prueba estadística.

Cada capítulo contiene uno o varios ejercicios basados en un problema, los cuales se desarrollan usando R versión 4.3.1. Se espera que el estudiante realice los ejercicios y luego compare sus resultados con las respuestas. Para esto se hace una descripción del problema y luego se da una lista de preguntas con ayudas para resolver los ejercicios, se indican las funciones de R recomendadas para contestar cada pregunta y el código apropiado para usar esas funciones. Los ejercicios empiezan con un análisis descriptivo de los datos para que el estudiante los pueda visualizar antes de entrar en la formulación de un modelo. Después de cada lista de preguntas, se desarrolla cada ítem con el código de R y se agregan comentarios que ayuden a dar conclusiones. En la dirección ... están disponibles todos los datos que se usan en los ejercicios.

Para el desarrollo de los ejercicios se usan varias librerías de R. La mayoría de las funciones están disponibles en el paquete básico de R llamado `Stats`. Para el uso de esta librería no es necesario hacer descarga alguna, puesto que se activa directamente con la instalación de R. Se usan algunas librerías para hacer gráficos, tales como `car` (Fox y Weisberg, 2019), `lattice` (Sarkar, 2008) y `ggplot2` (Wickham, 2016). Se usa `dplyr` (Wickham, Francois, Henry y Müller, 2019) para obtener estadísticos de resumen por grupos. La librería `ibd` (Mandal, 2019) se emplea para el análisis de bloques incompletos, mientras que `pwr` (Champely, 2018) sirve para el cálculo de la potencia en pruebas de hipótesis.

Reconocimientos

Los datos que se utilizan en los ejercicios se tomaron de trabajos realizados por estudiantes, en algunos casos, o se simularon para que se lograran demostrar las características deseadas desde el punto de vista didáctico. Los estudiantes Edwin Abarca Araya, Elsa Guillén Amador y Christopher Torres Rojas, recolectaron los datos para el trabajo que se titula «Efecto de la forma de salida y el tipo de calentamiento en el tiempo de recorrido de la carrera de 100 metros planos». De forma similar, las estudiantes Erika Araya Cárdenas, María Jesús Castro Solís y Angélica Zúñiga Baldí, recolectaron los datos para el trabajo titulado «Efecto de la proporción de glicerina y el tipo de agua en la resistencia de burbujas de jabón». Los datos sobre asfalto fueron recolectados por el estudiante de Ingeniería Civil, Yordy Esteban Morales Guzmán, como parte de su tesis de licenciatura no concluida. Otros conjuntos de datos se generaron a partir de consultorías realizadas por el autor o la profesora María Isabel González Lutz. En tales casos, el problema original inspiró la descripción del problema y luego se generaron datos que cumplieran con las características necesarias para desarrollar el ejercicio.

Durante el proceso de construcción de este libro tuve la retroalimentación de muchas personas, especialmente tuve largas sesiones de reflexión sobre conceptos, enfoques y detalles de diversa índole con mi colega María Isabel González Lutz, quien me dio aportes sumamente valiosos, los cuales hicieron que el trabajo final fuera más comprensible y acertado. También conté con la revisión detallada de parte del profesor Johnny Madrigal Pana, quien tuvo la paciencia de leer todo el manuscrito. Obtuve respuestas provenientes de estudiantes de este mismo curso o alumnos de otras asignaturas dentro de la carrera de Estadística, ante mi solicitud de leer partes del texto. Agradezco los comentarios y sugerencias de Carlos Arrieta Elizondo, Shu Wei Chou, César Gamboa Sanabria, Susana García Calvo, Catalina Sandoval Alvarado, Rebeca Sura Fonseca y Pablo Vivas Corrales.

En los últimos dos años compartí con la profesora Shirley Rojas Salazar la enseñanza del curso de Diseño de Experimentos, por lo que pude obtener de ella importantes observaciones que surgieron a partir de la experiencia de impartir el curso y conversar sobre los detalles que descubrimos en el camino. Quiero agradecer muy especialmente a la estudiante Andrea Vargas Montero, quien pacientemente tomó notas durante las lecciones para informarme posteriormente de cambios que se

debían realizar a los ejercicios. Finalmente, quiero expresar un agradecimiento muy especial al estudiante Brayan Monge Blanco, pues no solo leyó todo el borrador final para hacer importantes observaciones, sino que siempre ha sido una persona totalmente dispuesta a impulsar mis iniciativas, a Brayan mi sincera gratitud.

Ricardo Alvarado Barrantes

San José, Costa Rica

Marzo, 2024

Índice general

Prefacio	iii
1 Análisis de varianza de una vía	11
1.1 Conceptos	11
1.2 Manzanas	21
1.2.1 Ejercicios	21
1.2.2 Solución	24
2 Comparaciones múltiples	35
2.1 Conceptos	35
2.2 Uvas	40
2.2.1 Ejercicios	40
2.2.2 Solución	42
2.3 Manzanas	47
2.3.1 Ejercicios	48
2.3.2 Solución	50
3 Diseños con dos factores	57
3.1 Conceptos	57
3.2 Tortugas 1	66
3.2.1 Ejercicios	66
3.2.2 Solución	70
3.3 Tortugas 2	80
3.3.1 Ejercicios	81
3.3.2 Solución	85

4 Diseños con tres factores	99
4.1 Conceptos	99
4.2 Refrescos	102
4.2.1 Ejercicios	103
4.2.2 Solución	106
5 Diseños con bloques	121
5.1 Conceptos	121
5.2 Burbujas	131
5.2.1 Ejercicios	132
5.2.2 Solución	136
5.3 Maderas	147
5.3.1 Ejercicios	148
5.3.2 Solución	150
6 Análisis de covariancia	157
6.1 Conceptos	157
6.2 Carrera 100 metros	160
6.2.1 Ejercicios	160
6.2.2 Solución	164
6.3 Asfalto	174
6.3.1 Ejercicios	175
6.3.2 Solución	179
7 Potencia	193
7.1 Conceptos	193
7.2 Manzanas	198
7.2.1 Ejercicios	198
7.2.2 Solución	200
7.3 Tortugas	203
7.3.1 Ejercicios	204
7.3.2 Solución	205
7.4 Burbujas	209
7.4.1 Ejercicios	209
7.4.2 Solución	211

Anexo	215
Glosario de funciones de R	215
Bibliografía	219
Índice de cuadros	221
Índice de figuras	224
Índice alfabético	225

Capítulo 1

Análisis de varianza de una vía

1.1 Conceptos

En los experimentos estadísticos se estudia una variable llamada respuesta y se busca verificar si el hecho de variar ciertas condiciones controlables repercute en cambios en el promedio de esa variable. Cada una de las condiciones que se cambia se denomina tratamiento. El diseño puede contar con varios factores o variables que se controlan. Cada una de las posibilidades que se estudian de un factor es un nivel de ese factor. La combinación de todos los niveles de los diferentes factores da origen a los tratamientos del diseño experimental.

Cuando se trabaja con un experimento donde solo hay un factor de diseño, los tratamientos coinciden con los niveles de ese factor e interesa comparar el efecto de los diferentes tratamientos sobre el promedio de la variable de interés. Para hacer estas comparaciones se recurre al concepto del efecto del tratamiento y se dice que si el tratamiento no tiene efecto sobre la respuesta, los promedios se mantendrán iguales para todos los tratamientos. En caso contrario cuando alguno de los tratamientos tiene un efecto positivo (negativo), el promedio correspondiente será mayor (menor) que el promedio general de la respuesta (combinando los datos de todos los tratamientos). Por lo tanto, el efecto de un tratamiento representa la distancia del promedio dentro de ese tratamiento al promedio general de la variable respuesta.

Modelo con un factor

Matemáticamente se denota el efecto del j -ésimo tratamiento con τ_j y se define como $\tau_j = \mu_j - \mu$, donde μ_j representa el promedio del j -ésimo tratamiento y μ el promedio general. De esta forma la representación matemática del efecto coincide con el concepto explicado anteriormente, que es la distancia entre el promedio específico del tratamiento, respecto al promedio general.

En la Figura 1.1 (izquierda) se ilustra un ejemplo en el que no existe efecto del factor sobre la respuesta promedio. En este caso los promedios de los 4 tratamientos son iguales al promedio general, y la distancia entre cada uno de los promedios específicos y el promedio general es cero, por lo que el valor del j -ésimo efecto es cero para todos los tratamientos ($\tau_j = 0$). En cambio en el gráfico de la derecha se observan diferencias entre los promedios, unos están por encima del promedio general y otros por debajo. Por lo tanto, los valores de τ_j no son cero en todos los casos, de hecho, algunos son positivos (como en los tratamientos 3 y 4 en que el promedio del tratamiento está sobre el promedio general) y otros negativos (como en los tratamientos 1 y 3 en que el promedio del tratamiento está debajo del promedio general), lo cual hace que la suma de todos los efectos siempre sea cero.

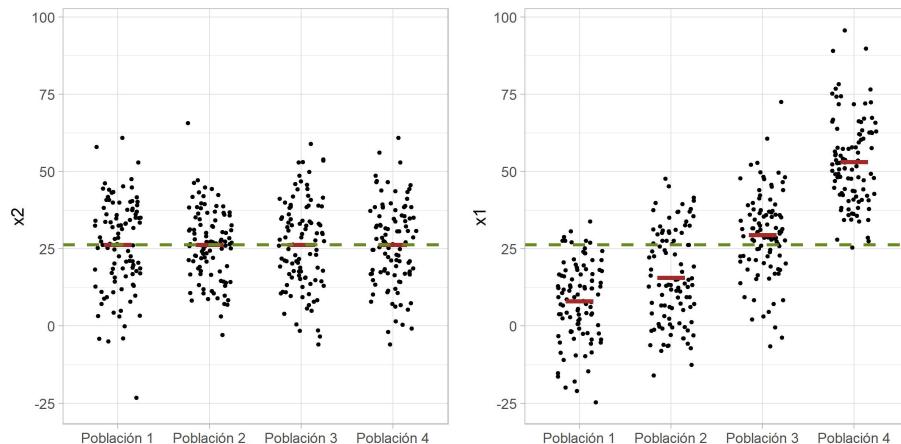


Figura 1.1: Distribución de una variable con 4 tratamientos en dos situaciones

Nota: la línea discontinua es la media general, las líneas más angostas son las medias reales de cada población.

Cuando se cuenta con un solo factor se dice que se tiene un diseño de una vía y se utiliza un modelo matemático que se puede parametrizar de dos formas: 1) suma nula o 2) tratamiento referencia.

Modelo de suma nula

Se asume que la suma de los coeficientes de todos los tratamientos es cero. En tal caso se introduce un coeficiente menos que la cantidad de niveles del factor, ya que el restante se obtiene por diferencia. Asumiendo que se toma el coeficiente del último tratamiento en función de los demás, y que hay k tratamientos, esta restricción se puede expresar como:

$$\sum_{j=1}^k \tau_j = 0 \Rightarrow \tau_k = -\sum_{j=1}^{k-1} \tau_j.$$

El modelo se escribe de la siguiente forma:

$$\mu_j = \mu + \tau_j.$$

Cuando se usa esta parametrización para el modelo, el primer coeficiente representa la media general y los otros coeficientes son los efectos de los diferentes tratamientos. Tiene sentido pensar que la suma de todos los efectos sea cero, ya que el promedio general es la media de los promedios de todos los tratamientos, y los efectos son las distancias de esos promedios respecto a la media general.

Se puede hacer uso de variables auxiliares para expresar el modelo de la forma que se usa para un modelo de regresión. Se requieren $k - 1$ variables auxiliares (C_1, \dots, C_{k-1}), una para cada uno de los primeros $k - 1$ niveles del factor. Por ejemplo, si se tiene un caso de un factor con 4 niveles, y se cuenta con 2 observaciones en cada nivel, se requieren las variables auxiliares C_1, C_2 y C_3 , definidas de la siguiente forma: C_1 toma valor 1 si la observación corresponde al nivel 1, -1 si corresponde al nivel k (en este caso nivel 4), y 0 en otro caso (si corresponde al nivel 2 o 3); mientras que C_2 toma valor 1 si la observación es del nivel 2, sigue siendo -1 si es del nivel 4, y 0 si es del nivel 1 o 3. Finalmente C_3 toma valor 1 si la observación es del nivel 3, sigue siendo -1 si es del nivel 4, y 0 si es del nivel 1 o 2. De esta forma, una observación que es del nivel 4 va a tener valor -1 en las tres variables auxiliares. El cuadro 1.1 muestra la construcción de las variables auxiliares para este ejemplo.

Cuadro 1.1: Variables auxiliares para un modelo de un factor con restricción de suma nula

Nivel	C_1	C_2	C_3
1	1	0	0
1	1	0	0
2	0	1	0
2	0	1	0
3	0	0	1
3	0	0	1
4	-1	-1	-1
4	-1	-1	-1

Nota: el factor tiene cuatro niveles y hay dos observaciones por tratamiento.

El modelo se escribe de la siguiente forma:

$$E[Y|Trat] = \mu + \sum_{j=1}^{k-1} \tau_j C_j = \mu + \tau_1 C_1 + \tau_2 C_2 + \tau_3 C_3.$$

Modelo de tratamiento referencia

Se toma uno de los tratamientos como referencia (vamos a tomar el primer tratamiento como referencia) y se define δ_j como la distancia del promedio del tratamiento j -ésimo al tratamiento de referencia, es decir, $\delta_j = \mu_j - \mu_1$, lo que hace que $\delta_1 = 0$. El modelo se escribe de la siguiente forma:

$$\mu_j = \mu_1 + \delta_j.$$

Esta forma de escribir el modelo mantiene el primer tratamiento como referencia y los coeficientes indican qué tanto se aleja cada promedio de μ_1 . Esta forma de parametrizar puede ser útil al realizar algunas comparaciones entre promedios.

Para expresar el modelo como un modelo de regresión, se requieren $k - 1$ variables auxiliares (D_2, \dots, D_k), definidas una forma diferente a la anterior. Para el mismo ejemplo, se requieren D_2, D_3 y D_4 , definidas de la siguiente forma: D_2 toma valor 1 si la observación corresponde al nivel 2, y 0 si corresponde al nivel 1, 3 o 4; mientras que D_3 toma valor 1 si la observación es del nivel 3, y 0 si es del nivel 1, 2 o 4. Finalmente D_4 toma valor 1 si la observación es del nivel 4, y 0 si es del nivel 1, 2 o 3. De esta

forma, una observación que es del nivel de referencia, es decir del nivel 1, va a tener valor 0 en las tres variables auxiliares. El cuadro 1.2 muestra la construcción de las variables auxiliares para este ejemplo.

Cuadro 1.2: Variables auxiliares para un modelo de un factor con tratamiento 1 de referencia

Nivel	D_2	D_3	D_4
1	0	0	0
1	0	0	0
2	1	0	0
2	1	0	0
3	0	1	0
3	0	1	0
4	0	0	1
4	0	0	1

Nota: el factor tiene cuatro niveles y hay dos observaciones por tratamiento.

El modelo se escribe de la siguiente forma:

$$E[Y|Trat] = \mu_1 + \sum_{j=2}^k \delta_j D_j = \mu_1 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4.$$

Análisis de varianza

Cuando se hace una investigación basada en un experimento, se intenta llegar a conclusiones sobre el efecto que tiene un factor a partir de datos de muestras. Se puede pensar que los datos provienen de poblaciones particulares, donde cada población tiene su propio promedio. Se intenta determinar si esos promedios podrían ser diferentes, lo cual estaría indicando que realmente existe un efecto del factor analizado sobre los promedios de esas poblaciones.

Supongamos que se toman muestras de cada una de las poblaciones representadas en la Figura 1.1 (izquierda), donde se sabe que los promedios no son diferentes entre sí. En la Figura 1.2 se representan dos situaciones que podrían ocurrir a partir de dos experimentos realizados con esas poblaciones y con el mismo procedimiento, el cual consiste en extraer 4 valores de cada población. En el lado izquierdo se obtienen muestras que afortunadamente coinciden con la situación original y los promedios

de cada una de esas muestras no se ven muy diferentes entre sí; sin embargo, en el lado derecho las observaciones dan promedios que son mucho más diferentes entre sí, llevando al investigador a pensar que el factor tiene un efecto sobre la respuesta promedio. Esta situación no es deseable, ya que el investigador estaría llegando a una conclusión errónea, la cual se conoce como error tipo I, y que consiste en concluir que el factor tiene un efecto cuando en realidad no lo tiene. El error opuesto consiste en la situación en que las poblaciones sean como la parte derecha de la Figura 1.1, es decir, las medias son diferentes, y las muestran resulten similares a las de la parte izquierda de la Figura 1.2. En ese caso, aunque el factor realmente tiene un efecto sobre la respuesta promedio, el investigador no logra demostrarlo ya que los promedios observados no son muy diferentes e incurre en el error tipo II, que consiste en concluir que el factor no tiene un efecto cuando en realidad sí lo tiene.

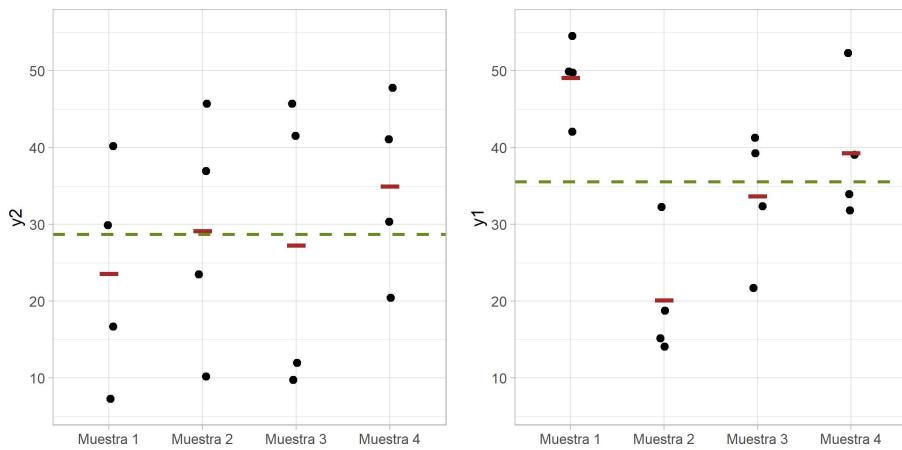


Figura 1.2: Muestras de una variable con 4 tratamientos en dos situaciones

Nota: la línea discontinua es la media general, las líneas más angostas son las medias muestrales de cada tratamiento.

Como el investigador en realidad no sabe si sus datos provienen de poblaciones con medias iguales o diferentes, debe buscar un método que le ayude a concluir si de verdad el factor tiene o no un efecto, pero basándose en la evidencia que le dan los datos que ha recolectado. Para esto se plantea una hipótesis que establece que los promedios de todos los tratamientos son iguales y le llama hipótesis nula (H_0). Esta hipótesis es independiente de cuál modelo se utilice (suma nula o tratamiento referencia), y es equivalente a decir que los efectos de todos los tratamientos son iguales a cero, que es lo mismo que decir que el factor no tiene efecto sobre la

respuesta promedio. Esto se puede expresar como:

$$H_0 : \mu_1 = \dots = \mu_k \quad \Leftrightarrow \quad H_0 : \tau_1 = \dots = \tau_k = 0$$

Si los datos provienen de poblaciones como las de la Figura 1.1 (derecha), idealmente el investigador quisiera que sus datos le proporcionaran evidencia para rechazar esa hipótesis, y de esta forma no cometería el error tipo II. En cambio, si provienen de poblaciones como las de la izquierda, querría no rechazar la hipótesis nula, pues si lo hace estaría cometiendo el error tipo I.

Para poner a prueba esta hipótesis se usa el análisis de varianza, que es un método de descomposición de la variabilidad total de la respuesta en varias fuentes de variación. Para empezar se considera la variabilidad total de la respuesta independientemente de los tratamientos a los que corresponde cada observación. Esta variabilidad se mide con la suma de cuadrados total (SCTot), y corresponde a la suma de las distancias al cuadrado de todas las respuestas respecto al promedio general. La (SCTot) se puede expresar como:

$$\text{SCTot} = \sum_{i=1}^n (y_i - \bar{y})^2,$$

donde \bar{y} representa la media general estimada de la respuesta y n es la cantidad total de observaciones. La SCTot coincide con el numerador de la varianza de la respuesta, por lo que se puede obtener multiplicando la varianza de la respuesta por $n - 1$. En el caso más simple como es el diseño de una vía, el método de análisis de varianza consiste en descomponer la SCTot en dos partes: 1) suma de cuadrados de tratamiento (SCTrat) y 2) suma de cuadrados residual (SCRes), es decir:

$$\text{SCTot} = \text{SCTrat} + \text{SCRes}.$$

La SCTrat está relacionada con las distancias de los promedios observados entre sí. Para comparar estos promedios, lo que se hace más bien es comparar cada promedio con la media general. Si se resta el promedio de un tratamiento menos la media general, se obtiene una estimación del efecto de ese tratamiento, es decir, $\hat{\tau}_j = \bar{y}_j - \bar{y}$, donde \bar{y}_j es la media de la respuesta dentro del j -ésimo tratamiento. Se elevan al cuadrado los efectos estimados y se ponderan por el número de datos en el tratamiento correspondiente (r_j). La suma de todas estos efectos cuadráticos

ponderados es la SCTrat, y se puede expresar como:

$$\text{SCTrat} = \sum_{j=1}^k r_j \hat{\tau}_j^2 = \sum_{j=1}^k r_j (\bar{y}_j - \bar{y})^2.$$

La SCRes se obtiene al sumar todos los residuales elevados al cuadrado. Un residual es la distancia de una observación respecto a la media del tratamiento a la que ella pertenece, es decir, e_{ij} es el i-ésimo residual en el j-ésimo tratamiento y se obtiene mediante $e_{ij} = y_{ij} - \bar{y}_j$, que es la distancia de la i-ésima observación del j-ésimo tratamiento (y_{ij}) a la media de ese tratamiento. Similar a lo que sucede con la SCTot, la suma de los residuales cuadráticos en el j-ésimo tratamiento es equivalente a la varianza de la respuesta en ese tratamiento (s_j^2) multiplicada por los grados de libertad asociados ($r_j - 1$). De esta forma se obtiene que la suma de los residuales cuadráticos en todos los tratamientos es:

$$\text{SCRes} = \sum_{j=1}^k \sum_{i=1}^{r_j} e_{ij}^2 = \sum_{j=1}^k \sum_{i=1}^{r_j} (y_{ij} - \bar{y}_j)^2 = \sum_{j=1}^k (r_j - 1) s_j^2.$$

A partir de las sumas de cuadrados se obtienen los cuadrados medios, los cuales son medidas de variabilidad y se obtienen al dividir cada suma de cuadrados entre sus grados de libertad. El cuadrado medio de tratamiento (CMTrat) es una medida de la variabilidad entre las medias de los tratamientos, lo cual es equivalente a decir que es una medida de la magnitud general de los efectos. Si los efectos son muy pequeños, el CMTrat va a dar muy cercano a cero. Para el CMTrat se tienen $k - 1$ grados de libertad, por lo que:

$$\text{CMTrat} = \frac{\text{SCTrat}}{k - 1}.$$

Por otra parte, el cuadrado medio residual (CMRes) es la medida de la variabilidad de la respuesta dentro de cada tratamiento y tiene $n - k$ grados de libertad. Por lo tanto, se tiene que:

$$\text{CMRes} = \frac{\text{SCRes}}{n - k}.$$

El CMRes también se puede calcular a partir de la varianzas de los tratamientos, como una media ponderada de las varianzas de cada tratamiento, donde se pondera con los grados de libertad de cada varianza; sin embargo, si se tiene el mismo número de observaciones en todos los tratamientos, el CMRes se obtiene como un promedio simple de las varianzas.

El razonamiento detrás de la prueba de una hipótesis consiste en encontrar una probabilidad de cometer error tipo I (rechazar la hipótesis nula cuando es cierta), y compararla contra un máximo previamente establecido para esta probabilidad. Este máximo se conoce como nivel de significancia y se denomina α . Si la probabilidad estimada de cometer error tipo I es suficientemente baja, es decir, no supera el nivel de significancia, se decide rechazar la hipótesis nula, en caso contrario, no se rechaza.

Para rechazar la hipótesis nula de igualdad de medias debería observarse que las medias de los diferentes tratamientos estén bastante alejadas unas de otras; sin embargo, esta lejanía es relativa y debe contrastarse con la variabilidad que tienen los datos dentro de cada tratamiento. En la Figura 1.3 se muestran dos casos que presentan la misma separación entre los promedios; sin embargo, en el gráfico de la izquierda hay poca variabilidad dentro de cada tratamiento, por lo que la separación entre las medias se hace más evidente y se pensaría que la hipótesis nula sea falsa. Al contrario, en el lado derecho esa misma separación no parece tan importante debido a la alta variabilidad dentro de cada tratamiento; por lo que, posiblemente no se llegue a rechazar la hipótesis nula.

El estadístico F observado se construye con el fin de tener una medida objetiva de esta comparación y consiste en el cociente entre el CMTrat y el CMRes. Si este cociente es suficientemente alto se puede rechazar la hipótesis planteada, de lo contrario, no se rechaza. Con la ayuda de la distribución F se obtiene una probabilidad asociada al error tipo I, que es la probabilidad de encontrar un valor mayor al estadístico F observado, en la distribución F con $k - 1$ y $n - k$ grados de libertad. Esta probabilidad se compara contra el nivel de significancia establecido previamente.

Caso de un factor con dos niveles

Cuando el factor que se analiza tiene solo dos niveles, la prueba de la hipótesis se puede realizar utilizando la distribución t , con lo que se obtiene un resultado

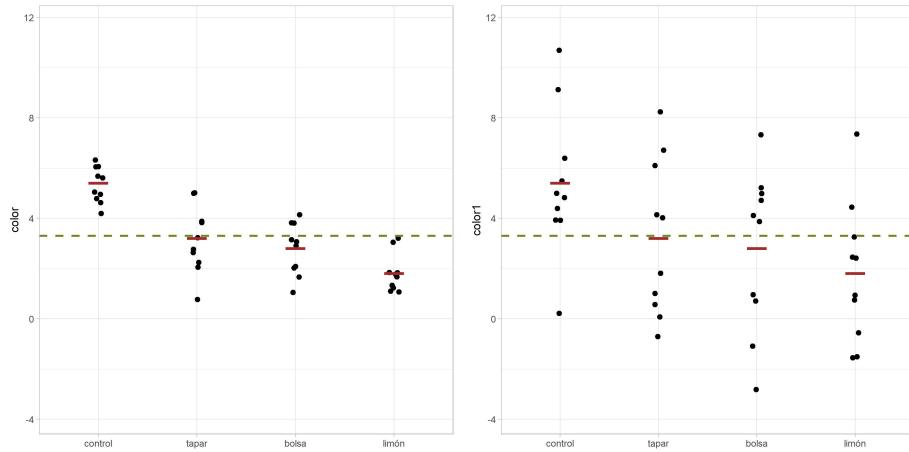


Figura 1.3: Dos situaciones con igual separación entre medias

Nota: izquierda con menos variabilidad y derecha con más variabilidad.

equivalente al obtenido con el análisis de varianza. En este caso la hipótesis nula se reduce a:

$$H_0 : \mu_1 = \mu_2 \quad \Leftrightarrow \quad H_0 : \tau_1 = 0$$

La hipótesis alternativa es una hipótesis de dos colas, es decir:

$$H_0 : \mu_1 \neq \mu_2$$

Se debe obtener la estimación del efecto ($\hat{\tau}_1$) y su desviación estándar ($ee_{\hat{\tau}_1}$), y con ellos se calcula el estadístico t observado mediante:

$$t = \frac{\hat{\tau}_1}{ee_{\hat{\tau}_1}}.$$

Con la ayuda de la distribución t se obtiene una probabilidad asociada al error tipo I, que es la probabilidad de encontrar un valor mayor al estadístico t observado, en la distribución t , con $n - k$ grados de libertad. Esta probabilidad debe ser igual a la que se obtiene con el estadístico F observado en la distribución F explicada anteriormente.

1.2 Manzanas

Las manzanas tienen un compuesto llamado polifenol oxidasa, el cual hace que al cortarse y entrar en contacto con el aire se oscurezcan rápidamente. Para evitar el pardeamiento se probaron tres tratamientos: tapar (código 2), poner en bolsa plástica cerrada (código 3), y aplicar jugo de limón (código 4). Además, se incluyó un control sin aplicar nada (código 1). Se seleccionan 40 manzanas y a cada una se le aplica aleatoriamente uno de los 4 tratamientos, lo cual resulta en 10 manzanas para cada tratamiento. Una vez aplicado el tratamiento a cada manzana, se pide a 3 jueces que califiquen el color en una escala de 1 a 6, donde 1 es el color normal de la fruta y 6 es el más oscuro. Cada manzana recibe como calificación el promedio de los 3 jueces. El objetivo final es seleccionar el tratamiento que mantenga mejor el color original para una empresa que se encarga de banquetes.

En un primer análisis solo se va a investigar si existe alguna diferencia en el color promedio resultante con los cuatro tratamientos.

1.2.1 Ejercicios

1. Preparación:

- (a) Lea el archivo `manzanas.csv` en R.
- (b) Defina correctamente el factor y ponga las etiquetas correspondientes para cada uno de los tratamientos.
- (c) Guarde la base en un archivo llamado `manzanas.Rdata` para ser utilizado en futuros ejercicios.

2. Análisis gráfico:

- (a) Obtenga una tabla con los promedios de cada tratamiento, y llámela `m`. Use: `tapply(y, x, mean)`.
- (b) Obtenga una tabla con las varianzas por tratamiento, y llámela `v`.
- (c) Obtenga la media general de la respuesta y llámela `media`.

- (d) Haga un boxplot para analizar el efecto de los tratamientos sobre la respuesta promedio. Agregue la media general usando `abline(h=media,col=2)` y las medias de los tratamientos usando `points(1:4,m,col=4,pch="-",cex=2)`.
- (e) Obtenga los efectos muestrales de cada tratamiento a partir de la tabla de medias y compare estos resultados con lo que ve en el gráfico. Cada efecto se puede estimar como: $\hat{\tau}_j = \bar{y}_j - \bar{y}$.
- (f) Explique el significado de cada uno de los valores obtenidos para los efectos muestrales.
- (g) Obtenga la suma de los efectos anteriores.
- (h) Obtenga una estimación de la varianza del error a partir de la tabla de varianzas. La estimación debe ser la media ponderada de las varianzas en los tratamientos, las cuales se ponderan con los grados de libertad de cada varianza; sin embargo, en este caso se tiene el mismo número de réplicas en todos los tratamientos, por lo que basta hacer un promedio simple de las varianzas.

3. Análisis de varianza:

- (a) Ajuste un modelo lineal. Use tanto la función `aov` como la función `lm`; la diferencia principal es que con `lm` se pueden obtener los coeficientes del modelo, mientras que con `aov` se puede obtener la tabla de efectos. En todo caso, cuando usa `lm`, por ejemplo `mod=lm(y~x)`, luego puede obtener `mod1=aov(mod)` de la misma forma que haciendo `mod1=aov(y~x)`.
- (b) Obtenga los resultados del análisis de varianza mediante `anova(mod)` o `anova(mod1)`. Si usa la función `aov` da lo mismo usar `summary(mod1)` o `anova(mod1)`.
- (c) Observe la línea de residuales para obtener el cuadrado medio residual y compárelo con la estimación de la varianza del error obtenida en el punto anterior.
- (d) Observe los grados de libertad residuales y justifique por qué se obtiene ese número.

- (e) Observe la línea del tratamiento y obtenga la suma de cuadrados de tratamiento.
- (f) Haga la suma de los cuadrados de los efectos obtenidos anteriormente. Observe que estos cuadrados deben multiplicarse por el número de réplicas para obtener exactamente la suma de cuadrados de tratamiento. Justifique por qué esto debe ser así.
- (g) Compare la variabilidad de los promedios con la variabilidad residual para determinar si hay alguna evidencia de diferencias entre las medias de la respuesta.
- (h) Establezca adecuadamente la hipótesis que está poniendo a prueba y dé una conclusión.

4. Estimación de parámetros del modelo de tratamiento referencia:

- (a) Obtenga las estimaciones de los parámetros del modelo. Por default R usa el modelo de tratamiento referencia. Esto se logra con el ajuste hecho con `lm` mediante `summary(mod)` o `mod$coef`.
- (b) ¿Qué significa el intercepto en este modelo?
- (c) ¿Qué representa cada uno de los coeficientes del modelo?
- (d) Obtenga la matriz de estructura y observe la codificación de las variables auxiliares.
- (e) A partir de los coeficientes obtenidos, obtenga los efectos muestrales y compárelos con los obtenidos en el punto 2e).
- (f) Obtenga los efectos directamente con `model.tables(mod)` (solo funciona si el modelo fue hecho con la función `aov`).

5. Modelo de suma nula:

- (a) Cambie al modelo de **suma nula** usando la siguiente instrucción:
`options(contrasts=c("contr.sum", "contr.poly")).`

Para volver al modelo de **tratamiento referencia** se usa:

```
options(contrasts=c("contr.treatment", "contr.poly")).
```

(b) Verifique la codificación con `contrasts(base$strat)`.

(c) Repita los pasos del punto 4. Compare los resultados.

6. Factor con dos niveles:

(a) Para ilustrar el caso cuando el factor tiene solo dos niveles, haga una base que contenga solo los datos que corresponden al nivel 1 y 2, llámela `base1`.

(b) Para eliminar los niveles que no tienen datos haga `base1$strat=factor(as.numeric(base1$strat))`.

(c) Ajuste el modelo con `lm`. Obtenga el análisis de varianza y observe la probabilidad asociada a la hipótesis de igualdad de medias.

(d) Obtenga el `summary`, extraiga la estimación del efecto y su error estándar, verifique el valor de t y obtenga la probabilidad asociada en la distribución t . Compare el resultado con el análisis de varianza.

1.2.2 Solución

1. Preparación:

(a) Lectura:

```
base=read.csv("manzanas.csv", sep=";")
```

(b) Definición de factor:

```
base$strat=factor(base$strat)
levels(base$strat)=c("control", "tapar", "bolsa", "limón")
base$strat
```

```
## [1] tapar tapar tapar tapar tapar tapar tapar
## [9] tapar tapar bolsa bolsa bolsa bolsa bolsa
## [17] bolsa bolsa bolsa bolsa limón limón limón limón
## [25] limón limón limón limón limón control control
## [33] control control control control control control control
## Levels: control tapar bolsa limón
```

(c) Almacenar la base:

```
save(base, file="manzanas.Rdata")
```

2. Análisis gráfico:

(a) Tabla con las medias:

```
(m=tapply(base$color,base$strat,mean))

## control tapar bolsa limón
##      5.4   3.2   2.8   1.8
```

(b) Tabla con las varianzas:

```
(v=tapply(base$color,base$strat,var))

## control tapar bolsa limón
##      0.49  1.73  1.07  0.62
```

(c) Media general:

```
(media=mean(base$color))

## 3.3
```

(d) Boxplot:

```
boxplot(color~trat,ylab="color",xlab="tratamiento",data=base)
abline(h=media,lty=2)
points(1:4,m,pch="-",cex=2)
```

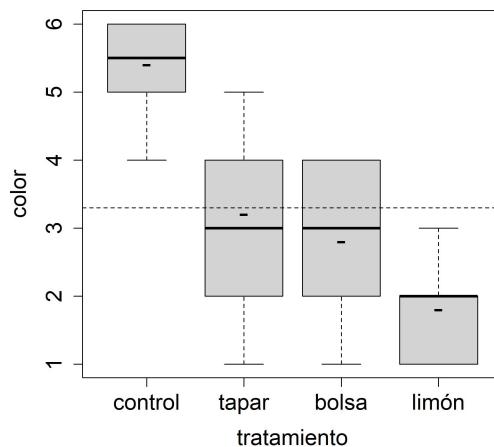


Figura 1.4: Puntajes de color por tratamiento

Nota: la línea discontinua es la media general, las líneas más anchas dentro de cada caja son las medianas de cada tratamiento y las líneas más angostas son las medias.

En la Figura 1.4 se nota que los puntajes de color en esta muestra son más bajos en los tres tratamientos que en el control. Cuando se aplicó limón estos puntajes tienden a ser más bajos que cuando se cubrió de alguna forma. También se nota que los dos tratamientos en que se cubrió producen resultados muy similares.

(e) Efectos muestrales:

```
(ef=m-media)
```

```
## control tapar bolsa limón
##      2.1  -0.1  -0.5 -1.5
```

Estos números coinciden con el gráfico puesto que los valores negativos concuerdan con aquellas medias que están por debajo de la media general y el valor positivo del control concuerda con el gráfico en que su media está por encima de la media general.

(f) Significado:

El control tiene una media que está 2,1 puntos sobre la media general, por lo que se dice que el control tiene el efecto de subir la media de la escala de color 2,1 puntos. El limón produce una media 1,5 puntos por debajo de la media general, es decir, tiene el efecto de bajar la media 1,5 puntos. Similarmente, los dos tratamientos en que se cubre tienen un leve efecto sobre la media ya que la bajan muy poco.

(g) Suma de los efectos:

```
sum(ef)
```

```
## 1.110223e-15
```

Aunque no da exactamente cero, esto se debe a un asunto computacional pero la suma de los efectos debe ser siempre cero por su misma construcción.

(h) Estimación de la varianza del error:

Primero se obtiene el número de réplicas en cada tratamiento y se observa que el diseño es balanceado, es decir, que tiene el mismo número de réplicas en todos los tratamientos.

```
(r=table(base$strat))
```

```
## control    tapar    bolsa    limón
##      10       10       10       10
```

La estimación de la varianza del error se obtiene al ponderar las varianzas de los tratamientos por los grados de libertad de cada tratamiento (número de réplicas menos 1).

```
(v1=sum((r-1)*v)/(sum(r)-4))
```

```
## 0.98
```

Puesto que el diseño es balanceado, se obtiene el mismo resultado si simplemente se promedian las varianzas de los cuatro tratamientos.

```
(v2=mean(v))
```

```
## 0.98
```

3. Análisis de varianza:

(a) Ajuste de modelo lineal:

El ajuste se hace con las dos formas equivalentes, una con la función `lm` y otra con la función `aov`.

```
mod=lm(color~trat,data=base)
mod1=aov(color~trat,data=base)
```

(b) Tabla de análisis de varianza:

Con ambas formas se obtiene el mismo resultado.

```
anova(mod)
```

```
## Analysis of Variance Table
## Response: color
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trat        3   69.2   23.07   23.59 1.2e-08 ***
## Residuals 36   35.2    0.98
```

```
summary(mod1)
```

```
## Analysis of Variance Table
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trat        3   69.2   23.07   23.59 1.2e-08 ***
## Residuals 36   35.2    0.98
```

```
anova(mod1)

## Analysis of Variance Table
## Response: color
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trat        3   69.2   23.07   23.59  1.2e-08 ***
## Residuals 36   35.2    0.98
```

(c) Cuadrado medio residual:

```
(v3=anova(mod)[2,3])

## 0.98
```

Se obtiene el mismo valor que se tenía al promediar las varianzas, es decir que el cuadrado medio residual es una medida de la variabilidad de color dentro de cada tratamiento.

(d) Grados de libertad residuales:

Hay 36 grados de libertad en los residuales, ya que se cuenta con 40 datos pero se usaron 4 grados de libertad para calcular los promedios de los 4 tratamientos, y a partir de ahí se obtuvieron los residuales dentro de cada tratamiento ($40-4=36$ grados de libertad). Otra forma de encontrar los grados de libertad residuales es a partir de los grados de libertad de cada tratamiento. Ya que se tienen 10 observaciones por tratamiento, se cuenta con 9 grados de libertad en cada uno, lo cual suma 36 grados de libertad en total.

(e) Suma de cuadrados de tratamiento:

```
anova(mod)[1,2]

## 69.2
```

(f) Suma de los cuadrados de los efectos:

```
sum(10*ef^2)

## 69.2
```

En la descomposición de la suma de cuadrados total se tiene una parte que va del promedio del tratamiento al promedio general, y esa cantidad es la misma para todos los valores de un mismo tratamiento, por lo que debe repetirse tantas veces como datos haya en ese tratamiento. De ahí viene que esa distancia o efecto deba multiplicarse por r_j , el número de réplicas en el j -ésimo tratamiento.

(g) Comparación de la variabilidad de los promedios con la variabilidad residual:

```
cmtrat=anova(mod)[1,3]
cmres=anova(mod)[2,3]
(f=cmtrat/cmres)

## 23.59

(p=pf(f,3,36,lower.tail = F))

## 0
```

Se observa que la variabilidad entre las medias de los tratamientos es 23,6 veces la de los residuales, lo cual es una cantidad enorme y tiene asociada una probabilidad diminuta (aparece un cero pero no es exactamente cero). Esto apoya la sospecha de que las medias no son iguales, sino que más bien están alejadas unas de otras

(h) Hipótesis y conclusión:

Las hipótesis se plantean de la siguiente forma:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad o \quad \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$

H_1 : al menos una de las medias es diferente de las demás.

Puesto que la probabilidad asociada al error tipo I es casi cero, se puede esperar que si se rechaza la hipótesis nula, la probabilidad de estar cometiendo un error sea ínfima, entonces se toma la decisión de rechazar esa hipótesis. Por lo tanto, se puede esperar que no todas las medias sean iguales (se sospecha que la del control sea la que se comporta diferente).

4. Estimación de parámetros del modelo de tratamiento referencia:

(a) Coeficientes:

```
mod$coef

## (Intercept) trattapar tratbolsa tratlimón
##          5.4       -2.2      -2.6      -3.6
```

(b) Intercepto:

```
m

## control tapar bolsa limón
##      5.4    3.2    2.8   1.8
```

Puesto que se usa el modelo con el control como referencia, el intercepto coincide con la media del control que es justamente 5,4.

(c) Otros coeficientes:

Los otros coeficientes representan la distancia que hay de la media de cada uno de los otros tratamientos con respecto a la media del control.

(d) Matriz de estructura:

```
model.matrix(mod)

##   (Intercept) trattapar tratbolsa tratlimón
## 1           1       1       0       0
## 2           1       1       0       0
## ...
## 11          1       0       1       0
## 12          1       0       1       0
## ...
## 21          1       0       0       1
## 22          1       0       0       1
## ...
## 31          1       0       0       0
## 32          1       0       0       0
## ...
```

En esta matriz de estructura se tienen 3 variables auxiliares ya que el factor tiene 4 niveles. El tratamiento control es la referencia, por lo que hay una variable para cada uno de los otros tratamientos. Estas variables están codificadas con 0 y 1 solamente, y toman siempre valores de 0 cuando la observación corresponde al control (líneas 31 a 40).

(e) Efectos muestrales:

En esta forma no es inmediato obtener los efectos. Por la forma en que están definidos, a los coeficientes debe restarse la media general y sumarse el intercepto para obtener el efecto respectivo.

```
mod$coef[2:4]-media+mod$coef[1]

## trattapar tratbolsa tratlimón
##      -0.1      -0.5     -1.5

mod$coef[1]-media

## (Intercept)
##      2.1
```

Estos resultados son los mismos que se habían obtenido en el punto 2e).

(f) Efectos directamente:

```
model.tables(mod1)

## Tables of effects
##
## trat
## trat
## control tapar bolsa limón
##      2.1   -0.1   -0.5  -1.5
```

5. Modelo de suma nula:

(a) Cambio al modelo de **suma nula**:

```
options(contrasts=c("contr.sum","contr.poly"))
```

(b) Verificación:

```
contrasts(base$trat)

## control    1    0    0
## tapar      0    1    0
## bolsa      0    0    1
## limón     -1   -1   -1
```

Ahora las variables auxiliares contienen un -1 en el tratamiento de referencia que en este caso es el cuarto (limón).

(c) Resultados con el modelo de **suma nula**:

Estimaciones:

```
mod2=lm(color~trat,data=base)
mod2$coef

## (Intercept) trat1 trat2 trat3
##          3.3   2.1  -0.1  -0.5
```

Intercepto: ahora el intercepto representa la media general que es 3,3.

Coeficientes: los otros coeficientes representan el efecto que tiene cada tratamiento, es decir, la diferencia entre la media de un tratamiento respecto a la media general. Aparecen sólo 3 efectos puesto que el cuarto se obtiene a partir de la restricción.

Matriz de estructura:

```
model.matrix(mod2)

##      (Intercept) trat1 trat2 trat3
## 1            1     0     1     0
## 2            1     0     1     0
## ...
## 11           1     0     0     1
## 12           1     0     0     1
## ...
## 21           1    -1    -1    -1
## 22           1    -1    -1    -1
## ...
## 31           1     1     0     0
## 32           1     1     0     0
## ...
```

Efectos: los coeficientes son directamente los efectos, salvo el cuarto que se tiene que obtener a partir de los otros.

```
mod2$coef[2:4]

##   2.1  -0.1  -0.5

-sum(mod2$coef[2:4])

## -1.5
```

6. Factor con dos niveles:**(a) Selección de datos:**

```
base1=base[as.numeric(base$trat)<3,]
table(base1$trat)

##   control    tapar    bolsa    limón
##       10       10       0       0
```

(b) Eliminación de niveles:

```
base1$trat=factor(as.numeric(base1$trat))
levels(base1$trat)=c("control","tapar")
table(base1$trat)

##   control    tapar
##       10       10
```

(c) Análisis de varianza:

```
mod3=lm(color~trat,data=basel)
anova(mod3)

## Analysis of Variance Table
## Response: color
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trat      1   24.2   24.20   21.78  0.0002 ***
## Residuals 18   20.0    1.11
```

Se rechaza la hipótesis nula de igualdad de las dos medias de los tratamientos control y tapar ($p=0.0002$).

(d) Prueba t:

```
summary(mod3)

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.30      0.2357 18.24  4.7e-13 ***
## trat1       1.10      0.2357  4.67  0.0002 ***
```

El efecto del tratamiento 1 (control) es $\hat{\tau}_1 = 1.1$, con un error estándar de 0.2357. A partir de estos valores se obtiene el t observado con:

```
(t=1.1/0.2357)
```

```
## 4.67
```

La probabilidad asociada debe buscarse en la distribución t con 18 grados de libertad. Como se trata de una prueba de dos colas, el resultado debe multiplicarse por 2.

```
2*(1-pt(4.667,18))
```

```
## 0.0002
```

Este resultado coincide con el obtenido en el análisis de varianza, en el que se usó la distribución F.

Capítulo 2

Comparaciones múltiples

2.1 Conceptos

Cuando se tienen más de dos niveles de un factor y se ha probado que existe un efecto del factor sobre la respuesta media, es deseable determinar entre cuáles niveles se detectan diferencias. Existen varias formas de hacer comparaciones simultáneas ; sin embargo, cada prueba que se haga tiene un error asociado, por lo que hacer muchas pruebas de forma independiente incrementa la probabilidad de incurrir en el error de observar diferencias cuando en realidad no las hay. Para esto se debe controlar el error total en un nivel que el investigador debe determinar previamente.

Comparaciones de Tukey

Uno de los métodos que más se utiliza para hacer comparaciones simultáneas es el de Tukey, el cual está diseñado para hacer las comparaciones de los promedios por pares y toma la diferencia entre los promedios de la respuesta de todos los pares posibles de los niveles de un factor. Este método evita hacer las comparaciones de forma independiente; sin embargo, cuando se cuenta con un número alto de tratamientos, el uso de este método puede resultar muy conservador, haciendo que la detección de diferencias se vuelva más difícil. Por lo tanto, en esos casos, es importante reflexionar sobre cuáles comparaciones son realmente útiles y no realizar todas las posibles comparaciones. En tal caso, puede recurrirse a otros métodos que hacen pruebas

sobre un subconjunto de los posibles pares de comparaciones, tal como el método de Dunnett que compara la media de un tratamiento control contra las medias de los demás tratamientos.

Al comparar el i-ésimo y el j-ésimo nivel de un factor se parte del estadístico $\bar{y}_i - \bar{y}_j$. Se obtiene una estimación de la varianza de este estadístico mediante:

$$\hat{V}(\bar{y}_i - \bar{y}_j) = \text{CMRes} \left(\frac{1}{r_i} + \frac{1}{r_j} \right).$$

La desviación estándar del estadístico se conoce como error estándar (ee_{ij}):

$$ee_{ij} = \sqrt{\text{CMRes} \left(\frac{1}{r_i} + \frac{1}{r_j} \right)}.$$

Para realizar la prueba se usa el estadístico $\bar{y}_i - \bar{y}_j$, cuyo valor esperado es $\mu_i - \mu_j$. Bajo la hipótesis nula $H_0 : \mu_i - \mu_j = 0$, el valor esperado del estadístico es cero, por lo que el estadístico estandarizado se obtiene al dividir la diferencia de los promedios observados entre su error estándar, de la siguiente forma:

$$q_{ij} = \frac{\bar{y}_i - \bar{y}_j}{ee_{ij}}.$$

Una vez que se tiene el estadístico q_{ij} para todos los pares de medias, se buscan las probabilidades de obtener un valor igual o mayor a esos estadísticos usando la distribución del rango estudentizado de Tukey, la cual está diseñada para controlar la probabilidad global de error tipo I. Estas probabilidades se comparan con el nivel de significancia global y las conclusiones aseguran que al rechazar algunas de las hipótesis se puede estar incurriendo en un error con una probabilidad controlada de forma global.

Intervalos de confianza simultáneos

En aquellos casos en que se concluya que hay diferencias entre ciertos pares de promedios, es importante dar estimaciones de esas diferencias, las cuales no deben ser solo puntuales sino que deben construirse intervalos de confianza.

Estos intervalos se pueden hacer de dos colas; sin embargo, en la mayoría de los experimentos se busca maximizar o minimizar la respuesta promedio, por lo que puede resultar útil buscar sólo un límite. En tales casos se buscaría una cota inferior para la diferencia promedio en valor absoluto, o lo que es lo mismo, la diferencia restando el promedio mayor menos el menor.

La cuantificación de la diferencia es muy importante para poder interpretar los resultados en términos de una diferencia previamente definida que resulte relevante al investigador. Esta diferencia establecida representa aquel punto mínimo en que teóricamente dos promedios poblacionales se deberían alejar para que en términos prácticos resulte de interés al investigador (ver Capítulo 8).

En la construcción de intervalos de confianza simultáneos, o cotas simultáneas, se debe realizar una corrección para asegurar una confianza global determinada. Para esto se recurre a la corrección de Bonferroni usando el cuantil de la distribución t con una probabilidad corregida. Si se trata de construir sólo una cota por comparación, la corrección es $1 - \alpha/d$, donde d es el número de cotas que se están construyendo y $1 - \alpha$ es el nivel de confianza global. Por otra parte, si se trata de construir intervalos, la corrección es $1 - \alpha/(2d)$, donde d es el número de intervalos que se construyen. La distribución t usa los grados de libertad del cuadrado medio residual.

Contrastes

Para probar hipótesis específicas, se plantean combinaciones lineales de los promedios, las cuales se llaman contrastes. El i -ésimo contraste se define como:

$$L_i = \sum_{j=1}^k c_j^{(i)} \mu_j.$$

Dos contrastes cuyos coeficientes están en los vectores $v_1 = [c_1^{(1)}, \dots, c_k^{(1)}]^T$ y $v_2 = [c_1^{(2)}, \dots, c_k^{(2)}]^T$, respectivamente, son ortogonales si el producto punto de ambos vectores es cero. Cuando se construyen contrastes ortogonales no es necesario realizar corrección alguna, ya sea en las pruebas como en los intervalos de confianza.

Se utiliza un ejemplo para ilustrar el uso de contrastes ortogonales, en el cual se tiene un experimento con un factor con 3 niveles y se quieren hacer dos comparaciones.

Primero se busca conocer si el nivel 1 tiene una media mayor a la de los otros dos niveles tomados en conjunto. Por otro lado, se quiere saber si el nivel 2 tiene una media mayor a la del nivel 3. Estas comparaciones se pueden escribir como:

$$\mu_1 = \frac{1}{2}(\mu_2 + \mu_3)$$

$$\mu_2 = \mu_3$$

De aquí se pueden obtener los contrastes L_1 y L_2 , los cuales que se puede escribir como:

$$L_1 = \mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = 1\mu_1 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3$$

$$L_2 = \mu_2 - \mu_3 = 0\mu_1 + 1\mu_2 - 1\mu_3$$

Las hipótesis que se quieren probar se reducen a:

$$H_0 : L_1 = 0 \quad \text{contra} \quad H_1 : L_1 > 0$$

$$H_0 : L_2 = 0 \quad \text{contra} \quad H_1 : L_2 > 0$$

Los vectores de coeficientes de estos contrastes son:

$$v_1 = [1, -\frac{1}{2}, -\frac{1}{2}]^T$$

$$v_2 = [0, 1, -1]^T$$

Los contrastes L_1 y L_2 son ortogonales ya que el producto punto de v_1 y v_2 es cero ($1 \cdot 0 - \frac{1}{2} \cdot 1 - \frac{1}{2} \cdot (-1) = 0$).

Una vez que se han definido los contrastes teóricos, los cuales se expresan en términos de las medias poblacionales de los tratamientos, interesa obtener estimaciones de las combinaciones lineales representadas por esos contrastes. Para estimar un contraste se usan los promedios observados, los cuales van a sustituir a los μ_j en la expresión de L . Se elimina el subíndice i de la expresión de L_i y el superíndice de los coeficientes de $c_j^{(i)}$ por simplicidad, y se expresa la estimación del contraste como:

$$\hat{L} = \sum_{j=1}^k c_j \bar{y}_j.$$

Se puede estimar la varianza del estadístico \hat{L} con:

$$\hat{V}(\hat{L}) = \text{CMRes} \sum_{j=1}^k \frac{c_j^2}{r_j}.$$

Otra forma útil de hacer la estimación es usando los coeficientes del modelo, ya que en diseños más complejos, la estimación de μ_j podría no coincidir con el promedio observado. En general, la media de un tratamiento puede expresarse como un modelo de regresión que es una combinación lineal de los coeficientes. En este caso el modelo de regresión se escribe como:

$$\mu_j = \mu + \tau_1 C_1 + \tau_2 C_2.$$

Si se colocan los coeficientes del modelo en un vector β , en este caso se tiene que:

$$\beta = [\mu, \tau_1, \tau_2]^T.$$

Por otra parte, se escriben los contrastes que se quieren probar en términos de los coeficientes del modelo de regresión. En este caso, con la restricción de suma nula se tiene que $\tau_3 = -\tau_1 - \tau_2$, por lo que:

$$\mu_1 = \mu + \tau_1$$

$$\mu_2 = \mu + \tau_2$$

$$\mu_3 = \mu + \tau_3 = \mu - \tau_1 - \tau_2$$

$$L_1 = \mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = \mu + \tau_1 - \frac{1}{2}(\mu + \tau_2 + \mu - \tau_1 - \tau_2) = \frac{3}{2}\tau_1$$

$$L_2 = \mu_2 - \mu_3 = \mu + \tau_2 - (\mu - \tau_1 - \tau_2) = \tau_1 + 2\tau_2$$

Para cada contraste L_i se busca un vector h_i , de tal forma que el contraste se pueda escribir como el producto punto de los dos vectores: $L_i = h_i \cdot \beta$. En este caso, los vectores h_1 y h_2 se definen como:

$$h_1 = [0, \frac{3}{2}, 0]^T$$

$$h_2 = [0, 1, 2]^T$$

La estimación del contraste se obtiene calculando $\hat{L}_i = h_i \cdot \hat{\beta}$ y la varianza estimada del estadístico \hat{L}_i se puede obtener con:

$$\hat{V}(\hat{L}_i) = h_i^T \hat{V}(\hat{\beta}) h_i.$$

Una vez obtenida la estimación del contraste y la varianza del estimador, se puede estandarizar el estadístico y proceder a hacer las pruebas de hipótesis e intervalos de confianza usando la distribución t con los grados de libertad del CMRes.

2.2 Uvas

En Costa Rica se cultivan dos especies de uva: roja y blanca. Como medida de dulzor se utiliza la escala Brix. Se tomaron medidas de los grados Brix de una muestra de uvas de ambas especies provenientes de 3 sitios: La Garita, La Guácima y San Vito. Se quería determinar si el dulzor promedio varía según la localidad de producción. Los investigadores consideran que una diferencia en el dulzor promedio de dos poblaciones debe ser al menos de medio grado Brix para considerarse relevante.

2.2.1 Ejercicios

1. Preparación:

- (a) Cargue el archivo `uvas.csv`. Verifique que la variable `localidad` esté definida correctamente como un factor.
- (b) Comente las características de este experimento. Si solo se están comparando los promedios entre las tres localidades, ¿es necesario tomar en cuenta que las uvas son de dos tipos?

2. Hipótesis básica:

- (a) Establezca la hipótesis básica para verificar que en efecto las tres localidades no producen el mismo promedio de dulzor.
- (b) Haga un boxplot que permita de una forma descriptiva apoyar o contradecir esta hipótesis.
- (c) Obtenga las medias de cada tratamiento.

(d) Ponga a prueba la hipótesis.

3. Comparaciones de promedios:

- (a) Dado que el objetivo es comparar todas las localidades entre sí, se trata de un problema de comparación de todos los pares de promedios. Escriba todas las hipótesis que se deben probar.
- (b) Verifique que estas hipótesis no son ortogonales.
- (c) Obtenga el cuadrado medio residual.
- (d) Obtenga los estadísticos de interés para realizar cada prueba, es decir, debe calcular $\bar{y}_i - \bar{y}_j$. Puesto que solo interesa ver si existen diferencias entre cada par de promedios, use el valor absoluto de las diferencias $|\bar{y}_i - \bar{y}_j|$.
- (e) Obtenga el error estándar del estadístico $|\bar{y}_i - \bar{y}_j|$ en cada caso.
- (f) Obtenga el valor estandarizado del estadístico dividiéndolo por su error estándar.
- (g) Encuentre la probabilidad de obtener un valor igual o mayor al estadístico usando la distribución del rango estudentizado de Tukey. Use `ptukey(q*sqrt(2), k, df, lower.tail = F)`, donde k es el número de grupos y df son los grados de libertad de los residuales. Se debe multiplicar el valor q por $\sqrt{2}$ porque la función de R asume que los dos grupos tienen igual número de réplicas e incluye en el denominador $\sqrt{2}$, lo cual debe corregirse. Con el argumento `lower.tail=F` se obtienen directamente las dos colas. Esta probabilidad debe compararse directamente contra α puesto que se trata de pruebas de dos colas.
- (h) Obtenga estas probabilidades automáticamente usando la función `TukeyHSD(mod)`. Las probabilidades que se obtienen con esta función son relativas a hipótesis de dos colas, mientras que los intervalos de confianza son válidos solo si se rechazan todas las comparaciones de pares de promedios. En caso de que haya diferencia entre solo algunos pares, debe hacerse la corrección de Bonferroni para obtener los límites de confianza simultáneos solo para aquellos pares en que se encontraron diferencias significativas.

(i) ¿Qué se concluye en términos de las hipótesis que se probaron?

4. Límites para las diferencias:

(a) Obtenga intervalos de confianza para la diferencia de las medias solo en los casos en que se encontró una diferencia significativa. Se debe obtener el valor de la distribución t con los grados de libertad residuales (gl). Este valor se obtiene haciendo el ajuste de Bonferroni para tener un nivel de 95% de confianza para todos los intervalos en conjunto, lo cual se hace con $t=qt(1-0.05/(2*d), gl)$, donde d es el número de intervalos. Luego se calculan los límites de confianza con:

$$IC = |\bar{y}_i - \bar{y}_j| \pm t \cdot ee_{ij}.$$

(b) ¿Qué se concluye en términos generales?

2.2.2 Solución

1. Preparación:

(a) Lectura:

```
base=read.csv("uvas.csv")
str(base)

## 'data.frame': 100 obs. of 4 variables:
## $ localidad: Factor w/ 3 levels "Garita","Guacima",...: 1 2 3 1 ...
## $ especie : Factor w/ 2 levels "blanca","roja": 1 2 2 2 1 2 1 2 ...
## $ diam   : num 1.2 0.9 0.8 0.7 1.6 1 1.2 1 1.1 1.3 ...
## $ brix   : num 17 16.8 17.9 18.1 17.9 15 18.1 16.2 17 16 ...
```

(b) Características:

Aunque no se quieren comparar las uvas según la especie, si las de una especie tienen un nivel de grados Brix muy diferente al de la otra, sería muy importante incluir este factor en el modelo, pues al no hacerlo se tendría más ruido traducido en una mayor varianza residual. Esto podría producir que en el experimento no se logren detectar diferencias que sí existen. En este caso no se incluye la especie porque así se está planteando el ejercicio.

2. Hipótesis básica:

(a) Hipótesis:

La hipótesis nula se puede plantear en términos de las medias de los tratamientos o de los efectos:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \Leftrightarrow \quad \tau_1 = \tau_2 = \tau_3 = 0$$

(b) Boxplot:

```
boxplot(brix~localidad, ylab="Grados Brix", xlab="Localidad", data=base)
```

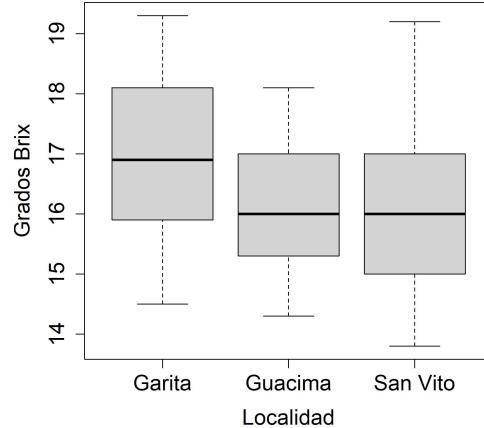


Figura 2.1: Grados Brix por localidad

En la Figura 2.1 se observa que el promedio de grados Brix en Garita parece mayor que en las otras dos localidades, por lo que es posible que se rechace la hipótesis de igualdad de promedios.

(c) Promedios por tratamiento:

```
(m=tapply(base$brix, base$localidad, mean))
```

```
## Garita      Guacima     San Vito
## 16.96      16.10      15.97
```

Los promedios observados mantienen la relación observada en el gráfico, donde el promedio de grados Brix en Garita parece mayor que en las otras dos localidades, mientras que estas últimas tienen promedios muy parecidos.

(d) Prueba de la hipótesis:

```
mod1=aov(brix~localidad,data=base)
anova(mod1)

## Analysis of Variance Table
##
## Response: brix
##           Df Sum Sq Mean Sq F value    Pr(>F)
## localidad  2  20.69   10.35    5.72    0.004 ***
## Residuals 97 175.52    1.81
```

La probabilidad asociada a la prueba de igualdad de los tres promedios es muy baja (0,004), por lo que se rechaza esta hipótesis y se concluye que, con un nivel de significancia de 0,05, no se puede esperar que los promedios de grados Brix en las tres localidades sean iguales. Entonces vale la pena investigar más para decidir en cuál o cuáles de ellas se puede esperar que el dulzor sea mayor.

3. Comparaciones de promedios:

(a) Hipótesis:

$$\begin{array}{ll} H_0 : \mu_1 = \mu_2; & H_1 : \mu_1 \neq \mu_2 \\ H_0 : \mu_1 = \mu_3; & H_1 : \mu_1 \neq \mu_3 \\ H_0 : \mu_2 = \mu_3; & H_1 : \mu_2 \neq \mu_3 \end{array}$$

(b) Ortogonalidad:

Para verificar si hay ortogonalidad en las hipótesis, se parte del vector de medias (μ_1, μ_2, μ_3) y se escriben los vectores que al multiplicarlos por este vector dan como resultado los siguientes contrastes:

$$L_1 : \mu_1 - \mu_2 = 0$$

$$L_2 : \mu_1 - \mu_3 = 0$$

$$L_3 : \mu_2 - \mu_3 = 0$$

```
v1=c(1,-1,0); v2=c(1,0,-1); v3=c(0,1,-1)
v1%*%v2; v1%*%v3; v2%*%v3
```

```
## 1
## -1
## 1
```

El producto de cada par de vectores es diferente de cero, lo que indica que no hay ortogonalidad.

(c) Cuadrado medio residual:

```
(cmres=anova(mod1)[2,3])
```

```
## 1.81
```

(d) Estadísticos:

```
d12=abs(m[1]-m[2])
d13=abs(m[1]-m[3])
d23=abs(m[2]-m[3])
d=c(d12,d13,d23)
names(d)=c("Ga-Gu", "Ga-SV", "Gu-SV")
d
```

```
## Ga-Gu Ga-SV Gu-SV
## 0.85 0.98 0.13
```

(e) Error estándar:

```
(r=table(base$localidad))
```

```
## Garita Guacima San Vito
## 38 25 37
```

```
ee12=sqrt(cmres*(1/r[1]+1/r[2]))
ee13=sqrt(cmres*(1/r[1]+1/r[3]))
ee23=sqrt(cmres*(1/r[2]+1/r[3]))
ee=c(ee12,ee13,ee23)
names(ee)=names(d)
ee
```

```
## Ga-Gu Ga-SV Gu-SV
## 0.35 0.31 0.35
```

(f) Valor estandarizado del estadístico:

```
q=d/ee; names(q)=names(d); q
```

```
## Ga-Gu Ga-SV Gu-SV
## 2.46 3.17 0.38
```

(g) Probabilidad del rango estudentizado de Tukey:

Puesto que las hipótesis no son ortogonales, hay que usar algún método que haga un ajuste para que el nivel de significancia global se mantenga en el nivel establecido (en este caso de 0.05). Como se están comparando todos los promedios entre sí por pares, la prueba que hace este ajuste es la de Tukey.

```
(p=ptukey(q*sqrt(2),3,97,lower.tail = F))
```

```
## Ga-Gu Ga-SV Gu-SV
## 0.041 0.006 0.922
```

(h) Resultado automático:

```
TukeyHSD(mod1)

## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = brix ~ localidad, data = base)
##
## $localidad
##          diff   lwr   upr p adj
## Guacima-Garita -0.85 -1.68 -0.03 0.041
## San Vito-Garita -0.98 -1.72 -0.25 0.006
## San Vito-Guacima -0.13 -0.96  0.70 0.922
```

(i) Conclusión:

Las probabilidades obtenidas se comparan directamente contra el $\alpha = 0,05$, ya que la distribución de Tukey ya ha realizado la corrección necesaria por no haber ortogonalidad y a la vez porque se comparan todos los promedios entre sí por pares. La primera y segunda probabilidad son menores a 0,05, por lo que se rechazan las hipótesis correspondientes, es decir, se infiere que la media verdadera del dulzor de Garita es diferente que la de San Vito y la de Guácima, aunque entre estas dos últimas no se detectan diferencias.

4. Límites para las diferencias:

(a) Cálculo de intervalos de confianza:

Se construyen dos intervalos de confianza ($d = 2$), y se recurre a la corrección de Bonferroni usando el cuantil de la distribución t para $1 - \alpha/(2d)$, con 97 grados de libertad.

```
t=qt(1-0.05/(2*2),97)
(lim=cbind(d[1:2]-t*ee[1:2],d[1:2]+t*ee[1:2]))
```

```
## Ga-Gu 0.06 1.64
## Ga-SV 0.28 1.69
```

(b) Conclusión:

El promedio observado para Garita es mayor que el de San Vito (16,96 vs 15,97), por lo que sí se esperaría que haya diferencias entre las medias verdaderas de estos dos lugares, la media de dulzor para Garita podría ser mayor que la de San Vito. Además, cuando se compara la media de Garita contra la de Guácima también se observa que la primera es mayor (16,96 vs 16,10). Se espera con 95% de confianza que la primera diferencia esté entre 0,28 y 1,69 grados Brix, mientras que la segunda esté entre 0,06 y 1,64.

Puesto que los investigadores habían establecido que la diferencia entre dos promedios debería ser de al menos medio grado Brix para considerarse relevante, este experimento no ha logrado concluir que entre Garita y las otras dos localidades haya un dulzor promedio realmente diferente, ya que esta diferencia podría ser de tan solo 0,06 grados Brix entre Garita y Guácima, así como apenas de solo 0,28 grados Brix entre Garita y San Vito.

Hay que notar que no basta con que la diferencia entre los promedios observados supere el mínimo establecido de medio grado Brix. En este caso se observó una diferencia entre los promedios de Garita y San Vito, así como entre Garita y Guácima, de casi un grado Brix, pero al hacer el intervalo de confianza se llega a concluir que la diferencia real podría ser tan baja como tan solo 0,28 o 0,06 grados Brix, con lo cual no hay una conclusión a favor de una diferencia relevante entre los promedios de estas localidades.

2.3 Manzanas

En el ejercicio de manzanas (Sección 1.2) se plantean además estos objetivos:

- Verificar si los tratamientos definidos controlan el pardeamiento produciendo una menor puntuación promedio que el caso en que no se aplica nada (control).
- Verificar si aplicar ácido controla mejor que cubrir (tapar o poner en bolsa).
- Verificar si poner en bolsa plástica cerrada es mejor que solamente tapar.

El objetivo final es seleccionar el tratamiento que mantenga mejor el color original para una empresa que se encarga de servicios de alimentación para actividades. Por la naturaleza de la variable respuesta, el tratamiento que produzca un valor promedio más bajo se considera mejor. Una vez que la hipótesis básica de igualdad de promedios se rechaza, interesa hacer varias comparaciones para cumplir con los objetivos:

- En primer lugar es necesario verificar si la respuesta promedio en realidad es menor en los tres tratamientos aplicados que en el caso control.

- En segundo lugar deben compararse los dos tratamientos donde se ha cubierto contra aquél en que se usó ácido y verificar si el promedio del último en realidad es menor que el promedio de los primeros.
- En tercer lugar deben compararse los dos tratamientos en que se cubrió y verificar si en realidad el promedio del tratamiento cuando se usó bolsa es menor que cuando se tapó.

2.3.1 Ejercicios

1. Preparación:
 - (a) Cargue el archivo manzanas.Rdata.
2. Pruebas de hipótesis:
 - (a) Cambie al modelo de **suma nula**. Obtenga las estimaciones de los coeficientes, para esto debe usar la función `lm`.
 - (b) Verifique cuál es el tratamiento que está codificando con -1. Use `contrasts(base$strat)`.
 - (c) Plantee las hipótesis necesarias para cumplir con los objetivos del problema.
 - (d) Explique por qué da lo mismo escribir estas hipótesis en términos de los efectos que en términos de los promedios de cada tratamiento.
 - (e) Defina los contrastes ortogonales (L_1 , L_2 y L_3) necesarios para probar las hipótesis.
 - (f) Verifique que los tres vectores (v_1 , v_2 y v_3) son ortogonales. Para esto debe usar los coeficientes originales.
 - (g) En el modelo con **suma nula** se tiene la restricción $\tau_4 = -(\tau_1 + \tau_2 + \tau_3)$. Para que todo quede en términos de los 3 efectos que son estimados con este modelo, haga la sustitución de τ_4 en L_1 .
 - (h) Cree una matriz con los coeficientes de los contrastes. Recuerde que debe incluir el intercepto.

- (i) Estime los contrastes.
 - (j) Verifique los resultados anteriores estimando el contraste basado en las medias estimadas.
 - (k) Encuentre el error estándar de cada contraste (ee_1 , ee_2 y ee_3).
 - (l) Encuentre el valor estandarizado del contraste haciendo: $t = \hat{L}_j / ee_j$.
 - (m) Encuentre la probabilidad de obtener un valor igual o mayor al estadístico usando la distribución t. Aunque sean varias pruebas simultáneas, no se necesita hacer ninguna corrección porque los contrastes son ortogonales. Use los grados de libertad de los residuales. De esta forma debe hacer: `pt(t, 36, lower.tail=F)`.
 - (n) ¿Qué se concluye en términos de las hipótesis que se probaron?
3. Cota para la diferencia:
- (a) Obtenga una cota inferior para la diferencia de las medias sólo en los casos en que se encontró una diferencia significativa.
 - (b) ¿Qué se concluye en términos de los objetivos del estudio?
4. Comparaciones no ortogonales. Ahora se van a hacer dos nuevas comparaciones adicionales suponiendo que el investigador tenía como objetivo solamente saber si cubrir (tapar o bolsa) da mejores resultados que no hacer nada y también si poner limón da mejores resultados que no hacer nada (control). En ambos casos, si hubiera diferencia, se debería cuantificar la mejoría.
- (a) Escriba las hipótesis y los contrastes asociados. Verifique que no son ortogonales.
 - (b) Encuentre las estimaciones de los contrastes, los errores estándar y el valor estandarizado del contraste.

- (c) Encuentre la probabilidad de obtener un valor igual o mayor al estadístico usando la distribución t. Como se trata de dos pruebas simultáneas con contrastes no ortogonales, debe hacer la corrección de Bonferroni. Esta consiste en dividir el nivel de significancia (α) por el número de pruebas que se realizan (d), entonces la probabilidad asociada a cada prueba se compara contra α/d .
- (d) Cuantifique la diferencia en los casos en que se concluyó que hay diferencia entre un par de promedios.
- (e) ¿Qué se concluye?

2.3.2 Solución

1. Preparación:

(a) Lectura:

```
load("manzanas.Rdata")
```

2. Pruebas de hipótesis:

(a) Estimaciones de los coeficientes con el modelo de **suma nula**:

```
options(contrasts=c("contr.sum","contr.poly"))
mod2 = lm(color~trat,data=base)
mod2$coef
```

```
## (Intercept) trat1 trat2 trat3
##          3.3   2.1  -0.1  -0.5
```

(b) Codificación:

```
contrasts(base$trat)
```

```
## control  1    0    0
## tapar    0    1    0
## bolsa    0    0    1
## limón   -1   -1   -1
```

Todas las variables auxiliares toman valor -1 para el tratamiento con limón, por lo que el efecto de limón se obtiene a partir de las estimaciones de los otros tres efectos, es decir, $\hat{\tau}_4 = -\hat{\tau}_1 - \hat{\tau}_2 - \hat{\tau}_3$.

(c) Hipótesis:

Por la naturaleza de la variable respuesta, se busca que el puntaje promedio de color sea menor para decir que un tratamiento es mejor. De esta forma, la primera hipótesis busca verificar que al tomar los tres tratamientos aplicados en conjunto (el promedio de los tres) se obtiene una media menor que en el caso control. Similarmente, la segunda hipótesis busca verificar que el promedio cuando se usó ácido es menor que el promedio de los dos tratamientos en que se cubrió. Finalmente, se espera encontrar un menor promedio cuando se usó bolsa que cuando se tapó.

```
(m = tapply(base$color,base$strat,mean))

## control tapar bolsa limón
##      5.4    3.2    2.8   1.8
```

$$\begin{array}{ll} H_0 : \tau_1 = \frac{1}{3}(\tau_2 + \tau_3 + \tau_4); & H_1 : \tau_1 > \frac{1}{3}(\tau_2 + \tau_3 + \tau_4) \\ H_0 : \frac{1}{2}(\tau_2 + \tau_3) = \tau_4; & H_1 : \frac{1}{2}(\tau_2 + \tau_3) > \tau_4 \\ H_0 : \tau_2 = \tau_3; & H_1 : \tau_2 > \tau_3 \end{array}$$

(d) Explicación:

Por la definición de un efecto como $\tau_j = \mu_j - \mu$, al sustituir estos τ_j se tendría μ a ambos lados de la expresión con lo cual se cancelarían y todo quedaría en términos de los μ_j .

$$\begin{array}{ll} H_0 : \mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4); & H_1 : \mu_1 > \frac{1}{3}(\mu_2 + \mu_3 + \mu_4) \\ H_0 : \frac{1}{2}(\mu_2 + \mu_3) = \mu_4; & H_1 : \frac{1}{2}(\mu_2 + \mu_3) > \mu_4 \\ H_0 : \mu_2 = \mu_3; & H_1 : \mu_2 > \mu_3 \end{array}$$

(e) Contrastes ortogonales:

$$\begin{aligned} L_1 &= \tau_1 - \frac{1}{3}\tau_2 - \frac{1}{3}\tau_3 - \frac{1}{3}\tau_4 \\ L_2 &= \frac{1}{2}\tau_2 + \frac{1}{2}\tau_3 - \tau_4 \\ L_3 &= \tau_2 - \tau_3 \end{aligned}$$

(f) Verificación de la ortogonalidad:

```
v1 = c(1,-1/3,-1/3,-1/3)
v2 = c(0,1/2,1/2,-1)
v3 = c(0,1,-1,0)
c(v1%*%v2, v1%*%v3, v2%*%v3)

## 0 0 0
```

Todos los productos dan cero, lo que indica que los 3 contrastes son ortogonales.

(g) Sustitución de τ_4 en L_1 y L_2 :

De la restricción de suma nula se tiene que $\tau_4 = -(\tau_1 + \tau_2 + \tau_3)$:

$$\begin{aligned} L_1 &= \tau_1 - \frac{1}{3}\tau_2 - \frac{1}{3}\tau_3 - \frac{1}{3}\tau_4 = \tau_1 - \frac{1}{3}\tau_2 - \frac{1}{3}\tau_3 - \frac{1}{3}(-\tau_1 - \tau_2 - \tau_3) \\ &= 0 \cdot \mu + \frac{4}{3} \cdot \tau_1 + 0 \cdot \tau_2 + 0 \cdot \tau_3 \\ L_2 &= \frac{1}{2}\tau_2 + \frac{1}{2}\tau_3 - \tau_4 = \frac{1}{2}\tau_2 + \frac{1}{2}\tau_3 - (-\tau_1 - \tau_2 - \tau_3) \\ &= 0 \cdot \mu + 1 \cdot \tau_1 + \frac{3}{2}\tau_2 + \frac{3}{2}\tau_3 \\ L_3 &= \tau_2 - \tau_3 = 0 \cdot \mu + 0 \cdot \tau_1 + 1 \cdot \tau_2 - 1 \cdot \tau_3 \end{aligned}$$

(h) Matriz con los coeficientes de los contrastes:

```
h1 = c(0,4/3,0,0)
h2 = c(0,1,3/2,3/2)
h3 = c(0,0,1,-1)
(h = cbind(h1,h2,h3))
```

```
##          h1    h2    h3
## [1,]  0.0  0.0  0.0
## [2,]  1.3  1.0  0.0
## [3,]  0.0  1.5  1.0
## [4,]  0.0  1.5 -1.0
```

(i) Estimación de los contrastes:

```
(L = t(h) %*% mod2$coef)
```

```
## h1 2.8
## h2 1.2
## h3 0.4
```

(j) Verificación de resultados con las medias estimadas:

```
L11 = m[1]-mean(m[2:4])
L22 = mean(m[2:3])-m[4]
L33 = m[2]-m[3]
c(L11,L22,L33)
```

```
## control    limón    tapar
##      2.8      1.2      0.4
```

(k) Error estándar:

```
(ee = sqrt(diag(t(h) %*% vcov(mod2) %*% h)))
```

```
##   h1   h2   h3
## 0.36 0.38 0.44
```

(l) Valor estandarizado del contraste:

```
(t = L/ee)
```

```
## h1 7.76
## h2 3.13
## h3 0.91
```

(m) Probabilidad asociada:

```
(p = pt(t,36,lower.tail=F))
```

```
## h1 0.000
## h2 0.002
## h3 0.186
```

(n) Conclusión:

Puesto que los 3 contrastes son ortogonales y se tienen hipótesis de una cola, las probabilidades obtenidas se comparan directamente contra el $\alpha = 0,05$. Las primeras dos probabilidades son menores a 0,05, por lo que se rechazan la primera y la segunda hipótesis, es decir, se encontró que la media de color es mayor cuando no se aplica nada que en los otros casos; además, que usar limón ácido controla mejor que cubriendo. Se compararon los dos tratamientos en que se cubre y no se demostró que usar bolsa sea mejor que tapar.

3. Cota para la diferencia:

(a) Cálculo de cotas inferiores:

Como solo se demostró que hay diferencias en dos de las hipótesis, se calculan solo dos cotas inferiores; sin embargo, no es necesario recurrir a la corrección de Bonferroni ya que los contrastes son ortogonales. En este caso se usa el cuantil de la distribución t para $1 - \alpha$.

```
t = qt(0.95, 36)
(lim = L[1:2] - t * ee[1:2])
## 2.19 0.55
```

(b) Conclusión:

El promedio de color cuando no se aplica nada es al menos 2,2 puntos mayor que en los otros tres casos en conjunto. Sabiendo que la escala va de 1 a 6, tener 2 puntos en promedio más es una cantidad importante, por lo que se nota que aplicar alguno de estos tratamientos ayuda a mejorar el color.

Por otra parte, aplicar limón da un color promedio al menos 0,55 puntos menor que cubrir. Esta diferencia no es tan grande como para afirmar que definitivamente el limón esté produciendo una mejora con respecto a cubrir. Aquí la persona experta debe dar su valoración de cuánto es una diferencia que para él o ella sea relevante.

4. Comparaciones no ortogonales:

(a) Hipótesis y contrastes:

$$H_0 : \tau_1 = \frac{1}{2}(\tau_2 + \tau_3); \quad H_1 : \tau_1 > \frac{1}{2}(\tau_2 + \tau_3)$$

$$H_0 : \tau_1 = \tau_4; \quad H_1 : \tau_1 > \tau_4$$

$$L_1 = \tau_1 - \frac{1}{2}\tau_2 - \frac{1}{2}\tau_3$$

$$L_2 = \tau_1 - \tau_4$$

```
v1=c(1,-1/2,-1/2,0)
v2=c(1,0,0,-1)
t(v1) %*% v2
```

```
## 1
```

El producto de los dos vectores de coeficientes de los contrastes es distinto de cero por lo que estos contrastes no son ortogonales.

Ahora se expresan los contrastes en función de los coeficientes del modelo:

$$L_1 = \tau_1 - \frac{1}{2}\tau_2 - \frac{1}{2}\tau_3 = 0 \cdot \mu + 1 \cdot \tau_1 - \frac{1}{2} \cdot \tau_2 - \frac{1}{2} \cdot \tau_3$$

$$L_2 = \tau_1 - \tau_4 = \tau_1 - (-\tau_1 - \tau_2 - \tau_3) = 0 \cdot \mu + 2 \cdot \tau_1 + 1 \cdot \tau_2 + 1 \cdot \tau_3$$

```
h1=c(0,1,-1/2,-1/2)
h2=c(0,2,1,1)
(h=cbind(h1,h2))
```

```
##          h1     h2
## [1,]  0.0  0.0
## [2,]  1.0  2.0
## [3,] -0.5  1.0
## [4,] -0.5  1.0
```

- (b) Estimaciones de los contrastes, errores estándar y valor estandarizado del contraste:

```
(L=t(h) %*% mod2$coef)
```

```
## h1 2.4
## h2 3.6
```

```
(ee=sqrt(diag(t(h) %*% vcov(mod2) %*% h)))
```

```
## h1     h2
## 0.38  0.44
```

```
(t=L/ee)
```

```
## h1 6.27
## h2 8.14
```

- (c) Probabilidad asociada:

```
(p=pt(t,36,lower.tail=F))
```

```
## h1      0
## h2      0
```

Como se tienen 2 contrastes entonces $d = 2$ por lo que las probabilidades obtenidas se deben comparar contra $\alpha/2 = 0,025$. En ambos casos la probabilidad obtenida es menor a 0,025, por lo que se rechazan ambas hipótesis.

(d) Cotas inferiores:

Se tienen que calcular 2 cotas inferiores ($d = 2$), y se recurre a la corrección de Bonferroni usando el cuantil de la distribución t para $1 - \alpha/d$, con 36 grados de libertad.

```
qt=qt(1-0.05/2, 36)
(lim=L-qt*ee)
```

```
## h1      1.62
## h2      2.70
```

(e) Conclusión:

Se espera con una confianza global de 95%, que el promedio de color cuando se cubre sea menor al menos 1,62 puntos que cuando no se hace nada, y similarmente cuando se usa limón sea al menos 2,70 puntos menor que cuando no se hace nada.

Capítulo 3

Diseños con dos factores

3.1 Conceptos

Cuando se quiere analizar el efecto de dos factores sobre la respuesta promedio, puede hacerse el análisis de cada factor por separado. Esta idea podría ser conveniente en algunos casos, pero podría perderse precisión, ya que la variabilidad dentro de los tratamientos podría estar siendo sobreestimada, y con ello resultaría más difícil detectar el verdadero efecto de un factor. Además, existe la posibilidad de que los dos factores estén interactuando y sus efectos no sean independientes.

Ante la presencia de dos factores, pueden proponerse dos modelos que difieren entre sí básicamente en que uno considera la interacción entre ambos factores, mientras que el otro asume independencia de sus efectos.

Interacción

Para ilustrar el concepto de interacción, tomamos un ejemplo que se desarrollará más adelante en el que se tienen dos factores, cada uno con dos niveles. El factor A consiste en dos condiciones de alimento para tortugas, mientras que el factor B es el género de las tortugas. Se toma un conjunto de tortugas de ambos géneros y aleatoriamente se les asignan un nivel del factor A, es decir, se alimenta cada tortuga con la condición que le correspondió. Cabe destacar que el interés en este estudio es ver el efecto que puede tener la condición de alimentación sobre la respuesta, pero se

considera el género de las tortugas porque se sabe que hay diferencias importantes en la respuesta según género. Como el factor de interés es la condición de alimentación, se dice que éste es el factor de diseño.

En este caso contamos con 4 combinaciones posibles entre los dos factores, a cada combinación se le llama tratamiento. Se designan los niveles del primer factor (condición) como A_1 y A_2 , y los niveles del segundo factor (género) como B_1 y B_2 . Si la respuesta se llama y , entonces se denota con \hat{y}_{11} la respuesta promedio de todas las tortugas en el experimento que están en la condición 1 y que son del género 1; similarmente \hat{y}_{12} denota la respuesta promedio de todas las tortugas que están en la condición 1 y que son del género 2, y de igual forma con los otros tratamientos. Se supone que las tortugas que participan en el experimento no son todas las tortugas que existen, por lo que se piensa en una población de tortugas para cada tratamiento, es decir, todas las tortugas del género 1 que podrían ser alimentadas con la condición 1 forman la población A_1B_1 , cuya media se denota con μ_{11} . De esta forma, \hat{y}_{ij} es una estimación de μ_{ij} para cualquier valor de i y j tomando valores 1 y 2.

En la Figura 3.1 se ilustran dos situaciones que se podrían presentar entre los factores del estudio. En el lado izquierdo se muestra una situación en la que los factores A y B interactúan. Los puntos terminales de las líneas indican los promedios de cada tratamiento. Cabe destacar que se han unido los promedios de ciertos tratamientos como una ayuda visual, pero en ningún momento debe considerarse la línea como una representación continua. Al posicionarse en el nivel B_1 , se observa que la diferencia entre los promedios de A_1 y A_2 es muy pequeña (μ_{11} vs μ_{21}), mientras que cuando se ubica en el nivel B_2 , los promedios de A_1 y A_2 son más distantes (μ_{12} vs μ_{22}). Por lo tanto, cuando se analiza el nivel B_1 , se observa un efecto muy pequeño de A, mientras que en el caso de B_2 , el efecto de A es mayor. Justamente el hecho de que el efecto de A difiera según el nivel de B es lo que hace pensar que existe una interacción entre A y B. Por otra parte, en el lado derecho se ilustra un ejemplo en el que los factores A y B no interactúan ya que, independientemente del nivel de B que se escoja, las diferencias entre los promedios de A_1 y A_2 son las mismas. En este caso la distancia entre μ_{11} y μ_{21} es similar a la distancia entre μ_{12} y μ_{22} .

La presencia de interacción entre dos factores implica automáticamente que un factor tenga un efecto sobre la respuesta promedio en al menos uno de los niveles del otro factor. De esta forma, si el factor A, por ejemplo, no tuviera un efecto sobre la

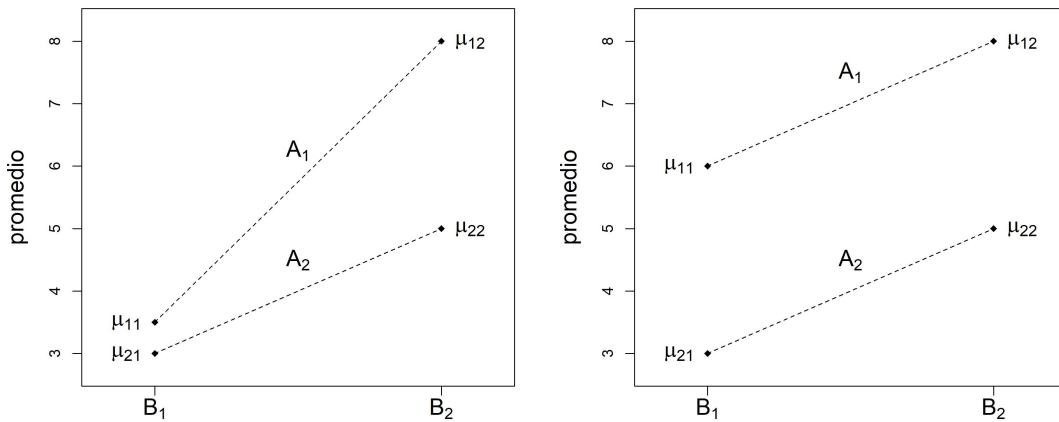


Figura 3.1: Diseño con dos factores en dos situaciones

Nota: en la parte izquierda existe interacción entre A y B, mientras que en el lado derecho no hay interacción entre A y B.

respuesta promedio para ninguno de los niveles de B, el efecto sería nulo en ambos casos y no habría interacción entre A y B.

Modelo sin interacción

El modelo sin interacción se puede escribir usando la restricción de suma nula o con un nivel de referencia para cada factor. Aquí se escribe con la restricción de suma nula, asumiendo que se toma el coeficiente del último nivel de cada factor en función de los demás, y que hay a niveles para el primer factor y b niveles para el segundo. Estas restricciones se pueden expresar como:

$$\sum_{i=1}^a \alpha_i = 0 \Rightarrow \alpha_a = -\sum_{i=1}^{a-1} \alpha_i,$$

$$\sum_{j=1}^b \beta_j = 0 \Rightarrow \beta_b = -\sum_{j=1}^{b-1} \beta_j.$$

El modelo se escribe de la siguiente forma:

$$\mu_{ij}^{SI} = \mu + \alpha_i + \beta_j.$$

Los términos α_i y β_j representan los efectos simples del i-ésimo nivel del primer factor y del j-ésimo nivel del segundo factor, respectivamente. En la Figura 3.2 se muestran las medias marginales del factor A llamadas $\mu_{1\bullet}$ y $\mu_{2\bullet}$. El promedio marginal $\mu_{1\bullet}$ considera todos los datos que están en el nivel 1 del factor A sin importar en cuál nivel de B están y se puede calcular promediando los promedios de los tratamientos que correspondan al nivel 1 de A (μ_{11} y μ_{12}). Los efectos simples del factor A representan las distancias de las medias marginales de cada uno de los niveles de A, con respecto a la media general. Por lo tanto, los efectos simples de A se definen como $\alpha_i = \mu_{i\bullet} - \mu$, donde $\mu_{i\bullet}$ es la media marginal dentro del i-ésimo nivel del factor A (por esto se dice que es marginal). De forma similar se definen los efectos simples de B como $\beta_j = \mu_{\bullet j} - \mu$, donde $\mu_{\bullet j}$ es la media dentro del j-ésimo nivel del factor B para todos los niveles del factor A.

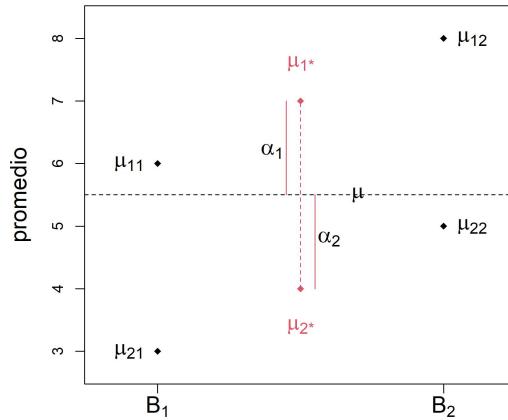


Figura 3.2: Diseño con dos factores sin interacción y el efecto marginal del factor A

Se puede deducir de este modelo que la diferencia entre las medias de dos niveles de un factor es la misma para cualquier nivel del otro factor. Por ejemplo, si se quieren comparar los promedios de los niveles 1 y 2 del factor A, fijando el factor B en un nivel específico, da lo mismo si se fija este factor en el nivel 1 o en el nivel 2, ya que algunos términos se cancelan. Si el factor B se fija en el nivel 1 se obtiene $\mu_{11} - \mu_{21} = (\mu + \alpha_1 + \beta_1) - (\mu + \alpha_2 + \beta_1) = \alpha_1 - \alpha_2$, que es lo mismo que se obtiene si se fija el factor B en el nivel 2: $\mu_{12} - \mu_{22} = (\mu + \alpha_1 + \beta_2) - (\mu + \alpha_2 + \beta_2) = \alpha_1 - \alpha_2$. Justamente esto es lo que caracteriza a un modelo sin interacción, en el sentido de que el efecto de un factor es independiente del nivel en que se fije el otro factor. Entonces, al comparar los dos niveles del factor A se comparan las medias marginales y se obtiene el mismo resultado: $\mu_{1\bullet} - \mu_{2\bullet} = \alpha_1 - \alpha_2$. En el caso particular en que el factor A tiene solo 2 niveles, se sabe que $\alpha_2 = -\alpha_1$, por lo que $\mu_{1\bullet} - \mu_{2\bullet} = 2\alpha_1$.

Para expresar el modelo como un modelo de regresión, se requieren $a - 1$ variables auxiliares para el primer factor (A_1, \dots, A_{a-1}), y $b - 1$ variables para el segundo factor (B_1, \dots, B_{b-1}). Por ejemplo, si se tiene un caso de un primer factor con 3 niveles y un segundo factor con 2 niveles, y se cuenta con 2 observaciones en cada uno de los 6 tratamientos, se requieren A_1, A_2 y B_1 , definidas de la siguiente forma: A_1 toma valor 1 si la observación corresponde al nivel 1 del primer factor, -1 si corresponde al nivel 3, y 0 si corresponde al nivel 2; A_2 toma valor 1 si la observación es del nivel 2 del primer factor, sigue siendo -1 si es del nivel 3, y 0 si es del nivel 1; mientras que B_1 toma valor 1 si la observación es del nivel 1 del segundo factor, y -1 si es del nivel 2. De esta forma, una observación que es del tratamiento formado por el nivel 3 del primer factor y el nivel 2 del segundo factor va a tener valor -1 en las tres variables auxiliares. El cuadro 3.1 muestra la construcción de las variables auxiliares para este ejemplo.

El modelo se escribe de la siguiente forma:

$$E[Y|Trat] = \mu + \alpha_1 A_1 + \alpha_2 A_2 + \beta_1 B_1.$$

Cuadro 3.1: Variables auxiliares para un diseño de dos factores

Tratamiento	A_1	A_2	B_1
11	1	0	1
11	1	0	1
21	0	1	1
21	0	1	1
31	-1	-1	1
31	-1	-1	1
12	1	0	-1
12	1	0	-1
22	0	1	-1
22	0	1	-1
32	-1	-1	-1
32	-1	-1	-1

Nota: el primer factor tiene tres niveles, el segundo factor dos niveles y hay dos observaciones por tratamiento. El primer dígito del tratamiento corresponde al nivel del primer factor y el segundo dígito al nivel del segundo factor.

Modelo con interacción

En el modelo con interacción se agrega un término que se denomina efecto de interacción, denotado por $(\alpha\beta)_{ij}$, el cual hace que el modelo tome la siguiente forma:

$$\mu_{ij}^{CI} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

La presencia del efecto de interacción hace que las diferencias entre las medias de los niveles de un factor no sean constantes para todos los niveles del segundo factor. El modelo con interacción produce una estimación de $\hat{\mu}_{ij}^{CI}$ que es igual a las medias observadas de cada tratamiento, entonces $\hat{\mu}_{ij}^{CI} = \bar{y}_{ij}$. Si se tienen las estimaciones de los efectos simples $\hat{\alpha}_i$ y $\hat{\beta}_j$, entonces se obtiene la estimación de la media del modelo sin interacción mediante:

$$\hat{\mu}_{ij}^{SI} = \bar{y} + \hat{\alpha}_i + \hat{\beta}_j.$$

Además, la estimación del efecto de interacción está dada por:

$$(\hat{\alpha}\hat{\beta})_{ij} = \hat{\mu}_{ij}^{CI} - \hat{\mu}_{ij}^{SI} = \bar{y}_{ij} - (\bar{y} + \hat{\alpha}_i + \hat{\beta}_j).$$

Hipótesis sobre la interacción

Se puede pasar del modelo con interacción al modelo sin interacción, si se asume que la interacción es nula. Para hacer este supuesto basado en los datos, se pone a prueba la hipótesis nula $H_0 : (\alpha\beta)_{ij} = 0$. Se cuantifican las magnitudes de los efectos de interacción observados con el cuadrado medio de interacción (CMInt). Para realizar el cálculo del CMInt, se obtiene la suma de los efectos de interacción al cuadrado, ponderados por el número de réplicas en cada tratamiento. Esta suma se divide por los grados de libertad asociados (producto de los grados de libertad de los dos factores que componen la interacción).

$$\text{CMInt} = \frac{\sum_{i=1}^a \sum_{j=1}^b r_{ij}(\hat{\alpha}\hat{\beta})_{ij}^2}{(a-1)(b-1)}.$$

La decisión sobre el rechazo de la hipótesis se basa en el estadístico F, el cual se construye mediante la razón entre el CMInt y el CMRes. El CMRes es una estimación de la varianza de la respuesta dentro de cada tratamiento bajo el supuesto de que la respuesta presenta la misma variabilidad en todos los tratamientos, lo cual se conoce como homocedasticidad. Cuando el supuesto de varianzas iguales se cumple, tiene sentido usar el CMRes como una medida única de variabilidad dentro de cada tratamiento y puede obtenerse mediante el promedio ponderado de las varianzas observadas de la respuesta en los distintos tratamientos.

El modelo sin interacción simplifica las comparaciones de las medias, ya que se pueden comparar todas las medias marginales de un factor independientemente del segundo factor. En cambio, cuando la interacción está presente, las comparaciones de las medias para los diferentes niveles de un factor se hacen dentro de cada nivel del segundo factor. En el ejemplo de las tortugas, si se demuestra que A y B tienen interacción, deberían compararse las dos condiciones de alimentación solo para el género 1 y luego hacer la misma comparación solo para tortugas del género 2, ya que la interacción implica que los resultados de esas comparaciones no van a dar lo mismo, y es de interés conocer en cuál caso las diferencias son mayores o menores.

Se deben plantear las siguientes hipótesis:

$$\mu_{11} = \mu_{21}$$

$$\mu_{12} = \mu_{22}$$

Tomando el vector de promedios $(\mu_{11}, \mu_{21}, \mu_{12}, \mu_{22})$, los vectores para obtener los contrastes son:

$$v_1 = [1, -1, 0, 0]^T$$

$$v_2 = [0, 0, 1, -1]^T$$

Los contrastes son ortogonales ya que el producto punto de v_1 y v_2 es cero. Por lo tanto, cada hipótesis se puede probar de forma independiente.

Si el factor A tuviera 3 niveles, deberían hacerse 3 comparaciones para cada género, es decir, se tendrían 3 hipótesis por género. Las siguientes hipótesis que aparecen a la izquierda corresponden al género 1 y las que aparecen a la derecha corresponden al género 2:

$$\mu_{11} = \mu_{21} \quad \mu_{12} = \mu_{22}$$

$$\mu_{11} = \mu_{31} \quad \mu_{12} = \mu_{32}$$

$$\mu_{21} = \mu_{31} \quad \mu_{22} = \mu_{32}$$

Tomando el vector de promedios $(\mu_{11}, \mu_{21}, \mu_{31}, \mu_{12}, \mu_{22}, \mu_{32})$, los vectores para obtener los contrastes serían:

$$v_1 = [1, -1, 0, 0, 0, 0]^T \quad v_4 = [0, 0, 0, 1, -1, 0]^T$$

$$v_2 = [1, 0, -1, 0, 0, 0]^T \quad v_5 = [0, 0, 0, 1, 0, -1]^T$$

$$v_3 = [0, 1, -1, 0, 0, 0]^T \quad v_6 = [0, 0, 0, 0, 1, -1]^T$$

Los primeros 3 vectores (v_1 , v_2 y v_3) no son ortogonales entre sí, así como no lo son los últimos 3 vectores (v_4 , v_5 y v_6); sin embargo los primeros 3 vectores son ortogonales con respecto a los últimos 3. Por esta razón, se pueden hacer las comparaciones dentro de cada género usando corrección de Bonferroni, pero tomando en cuenta que son solo 3 comparaciones en cada caso ya que un subconjunto es independiente del otro.

Análisis de efectos en el modelo sin interacción

En el caso de que no se rechace la hipótesis sobre la interacción, se puede asumir que no hay interacción entre los factores A y B. Entonces se usa el modelo sin interacción y se prueba si hay efecto del factor de diseño de forma independiente del factor B. Para realizar el análisis de varianza se calcula la suma de cuadrados del factor A (SCA) de forma similar al cálculo realizado para la SCTrat cuando se tenía un solo factor. De aquí se obtiene el cuadrado medio del factor A (CMA) al dividir la SCA por los grados de libertad de A que son $a - 1$:

$$\text{CMA} = \frac{\sum_{i=1}^a r_{i\bullet} \hat{\alpha}_i^2}{a - 1} = \frac{\sum_{i=1}^a r_{i\bullet} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}{a - 1},$$

donde $r_{i\bullet}$ es el número total de observaciones en el i -ésimo nivel de A sumando todos los niveles de B.

La decisión sobre el rechazo de la hipótesis se basa en el estadístico F, el cual se construye mediante la razón entre el CMA y el CMRes. El CMRes es una estimación de la varianza de la respuesta dentro de cada tratamiento con respecto a las medias estimadas sin interacción, las cuales no coinciden con las medias observadas. La forma más sencilla de obtener el CMRes es sumar la SCInt a la SCRes obtenida bajo el modelo con interacción y luego dividir entre los grados de libertad residuales del modelo sin interacción. Los grados de libertad residuales del modelo sin interacción se obtienen sumando los grados de libertad de la interacción más los grados de libertad residuales del modelo con interacción. De esta forma:

$$CMRes^{SI} = \frac{SCInt + SCRes^{CI}}{(a - 1)(b - 1) + (n - ab)}.$$

Si se concluye que hay un efecto del factor de diseño, se deben realizar comparaciones entre las medias marginales de los niveles de ese factor. Cuando A tiene solo 2 niveles, la hipótesis que se debe probar es la siguiente :

$$\mu_{1\bullet} = \mu_{2\bullet}.$$

Por otra parte, si A tuviera 3 niveles se deberían plantear 3 hipótesis:

$$\mu_{1\bullet} = \mu_{2\bullet}$$

$$\mu_{1\bullet} = \mu_{3\bullet}$$

$$\mu_{2\bullet} = \mu_{3\bullet}$$

Para probar estas hipótesis se puede utilizar el método de Tukey ya que se estarían comparando todas las medias marginales posibles para el factor A.

3.2 Tortugas 1

Se desea determinar si la falta de alimento afecta el nivel de proteínas en sangre en tortugas de la especie Chelonia midas (de agua salada) y si afecta de forma diferente a los machos que a las hembras. Se escogen 8 machos y 8 hembras. A cada uno se le asigna aleatoriamente una de las siguientes condiciones: 1) dieta regulada y 2) alimento en abundancia. Se registra el nivel de proteína en gramos por decilitro.

3.2.1 Ejercicios

1. Preparación:

- (a) Cargue el archivo `tortugas1.csv`. Verifique que las variables género y condición sean factores con los niveles adecuados. Los códigos para género corresponden a macho (1) y a hembra (2).
- (b) ¿Cuáles son los factores que se incluyen en el experimento?
- (c) Comente sobre las características de los factores y sobre el alcance del experimento.
- (d) ¿En qué aspectos se debe concentrar el análisis?
- (e) ¿Cuántos tratamientos tiene este experimento?
- (f) Verifique el número de repeticiones por tratamiento.

2. Efectos:

- (a) Obtenga la media general de la respuesta.
- (b) Obtenga los efectos simples de cada condición, es decir, las diferencias entre cada media de condición y la media general. Esto es similar a los τ cuando hay un solo factor, pero se van a llamar α_1 y α_2 .
- (c) Obtenga los efectos simples de cada género (β_1 y β_2).
- (d) Obtenga los promedios observados de los 4 tratamientos. Use `tapply`, pero haga una lista con los factores para que los cruce y obtenga la media para cada combinación o tratamiento: `tapply(Y, list(X1, X2), mean)`.
- (e) Obtenga los promedios estimados bajo el modelo sin interacción. Realice los cálculos correspondientes manualmente.
- (f) Compare los promedios estimados bajo el modelo sin interacción con los promedios observados.
- (g) Obtenga un gráfico de los valores de proteína separados por género para interpretar gráficamente si existe interacción entre condición y género. Recuerde el objetivo del estudio y vea qué es más conveniente que aparezca en el eje X. Use el comando `ggplot` de la librería `ggplot2`. En este caso se coloca género en el eje X, proteína en el eje Y, y la condición es el factor que se declara en el argumento `group`. La instrucción se escribe en dos partes que se unen por un símbolo `+`. La primera de ellas contiene la definición de las variables dentro del argumento `aes`, de esta forma:

```
ggplot(base, aes(x=genero, y=proteina, group = cond)).
```

La segunda parte indica que debe unir los promedios de cada condición usando el argumento `stat_summary` y que debe usar un tipo de línea para cada condición, de esta forma:

```
stat_summary(fun.y="mean", geom="line", aes(linetype = cond)).
```

- (h) Observe la distancia que hay entre cada par de medias para cada género.
- (i) Obtenga los efectos simples usando la función `model.tables`. Escriba primero el modelo con condición y género con la función `aov`.

3. Efectos de interacción en el modelo:

- (a) Escriba el modelo con interacción utilizando `lm` (llámelo `mod2`). Cambie al modelo de **suma nula**. Escriba el modelo con interacción entre condición y género. La interacción se agrega en un modelo de cualquiera de las siguientes formas: $Y \sim X_1 + X_2 + X_1 : X_2$ o $Y \sim X_1 * X_2$.
- (b) Observe la matriz de estructura usando `model.matrix(mod2)`. Vea a qué corresponde cada columna. Ponga atención a los códigos y relacionelos con los niveles de cada factor.
- (c) Extraiga los coeficientes. Observe el efecto de la interacción. ¿Qué representa esta cantidad?

4. Varianza del error:

- (a) Visualice los 4 tratamientos con un gráfico de cajas donde se incluyan los dos factores separados por un `+`. Use este gráfico para ver si se puede esperar que los datos de los 4 tratamientos provengan de distribuciones con la misma varianza. Además, confirme si se puede esperar interacción entre condición y especie.
- (b) Obtenga las varianzas de los 4 tratamientos.
- (c) Verifique si se cumple el supuesto de homocedasticidad. Use la prueba de Bartlett con la función `bartlett.test` indicando $Y \sim \text{interaction}(X_1, X_2)$.
- (d) Obtenga una medida de la variabilidad del error asumiendo homocedasticidad.

5. Prueba de hipótesis para la interacción:

- (a) Obtenga la tabla de análisis de varianza para probar la hipótesis de NO interacción entre condición y género.
- (b) Escriba la hipótesis nula que interesa probar con símbolos y con palabras.
- (c) Observe el cuadrado medio residual y compárelo con la estimación de la varianza del error obtenida más arriba. Interprete de esta forma qué significa el cuadrado medio residual.
- (d) Interprete qué se está midiendo con el cuadrado medio de interacción.

- (e) Observe en la columna de F el valor asociado a la hipótesis y explique de dónde sale.
- (f) Observe la probabilidad asociada y explique cómo se obtiene. Concluya.
6. Estimaciones bajo el modelo sin interacción:
- (a) Puesto que se decidió asumir que no hay interacción entre condición y género, ahora se procede a usar el modelo sin interacción. Escriba el modelo sin interacción con lm y llámelo $mod3$.
- (b) ¿Cuánto es el efecto simple de cada condición y cada género? ¿Cómo se interpretan?
- (c) Escriba la hipótesis que interesa probar con símbolos y con palabras.
- (d) Interprete qué se está midiendo con el cuadrado medio de condición y cuadrado medio de género.
- (e) Asegúrese que puede obtener estas cantidades a partir de los efectos simples.
- (f) Observe el cuadrado medio residual y compárelo con la estimación de la varianza obtenida a partir de la media de las varianzas de los 4 tratamientos. ¿Por qué no son iguales?
- (g) Observe en cuánto aumentó la suma de cuadrados residual en el $mod3$ con respecto a la que se obtuvo en el $mod2$. Explique por qué tiene sentido que el $mod3$ tenga una mayor suma de cuadrados residual y por qué la diferencia debe ser esa cantidad.
- (h) Observe en la columna de F el valor asociado a la hipótesis y concluya.
- (i) Debido a que se concluyó que sí hay diferencia en las medias de proteína según condición, vale la pena establecer una cota inferior para la magnitud de esa diferencia.

3.2.2 Solución

1. Preparación:

(a) Lectura:

```
base=read.csv("tortugas1.csv")
str(base)

## 'data.frame': 16 obs. of 3 variables:
## $ genero    : int 1 1 1 1 1 1 1 1 2 2 ...
## $ cond       : Factor w/ 2 levels "alimento","dieta": 1 1 1 1 2 ...
## $ proteina   : num 42.8 43.1 40.4 46.6 38.9 40.3 37.5 42.9 42.2 ...

base$genero=as.factor(base$genero)
levels(base$genero)=c("macho","hembra")
base$cond=factor(base$cond)
```

(b) Factores:

Hay dos factores: el que se está investigando es la condición de alimento y el otro es el género.

(c) Características de los factores:

El factor de diseño es la condición de alimento pues el interés del estudio es analizar el efecto que tiene cada una de las condiciones sobre la proteína promedio en sangre que tienen las tortugas que se alimentan de cada forma. A cada tortuga se le puede asignar aleatoriamente una condición, con lo cual las diferencias que se observen se pueden ligar a la condición viendo una relación de causa y efecto. Por otra parte, se incluye el género porque posiblemente el comportamiento no es el mismo entre ambos. Es importante incluir este factor puesto que el estudio también tiene como objetivo analizar si el efecto de la condición es el mismo en cada género (interacción).

(d) Enfoque del análisis:

Interesa ver primero si el efecto que tiene la condición de alimento es el mismo en ambos géneros. De ser así, puede darse una conclusión general sobre el efecto que tiene cada condición en la proteína promedio en sangre, sin diferenciar por género, de lo contrario, debe cuantificarse ese efecto por separado en cada género.

(e) Número de tratamientos:

Hay en total 4 tratamientos que se obtienen de combinar las dos condiciones para los dos géneros.

(f) Número de repeticiones:

```
table(base$cond,base$genero)

##          genero
## cond      macho hembra
## alimento    4     4
## dieta       4     4
```

2. Efectos:

(a) Media general:

```
(medgen=mean(base$proteina))

## 39.64
```

(b) Efectos simples de condición:

```
(alfa=tapply(base$proteina,base$cond,mean)-medgen)

## alimento dieta
##      1.56 -1.56
```

(c) Efectos simples de género:

```
(beta=tapply(base$proteina,base$genero,mean)-medgen)

## macho hembra
##   1.92 -1.92
```

(d) Promedios observados:

```
(med=tapply(base$proteina,list(base$cond,base$genero),mean))

##          macho hembra
## alimento 43.23 39.18
## dieta    39.90 36.28
```

(e) Promedios estimados bajo el modelo sin interacción:

```
m11=medgen+alfa[1]+beta[1]
m21=medgen+alfa[2]+beta[1]
m12=medgen+alfa[1]+beta[2]
m22=medgen+alfa[2]+beta[2]
(mest=c(m11,m21,m12,m22))

##    43.12    40.01    39.28    36.17
```

(f) Comparación:

```
M=cbind(as.vector(med),mest)
colnames(M)=c("Observados","Estimados")
M

##           Observados Estimados
## alimento-macho      43.23     43.12
## dieta-macho         39.90     40.01
## alimento-hembra     39.17     39.28
## dieta-hembra        36.27     36.17
```

Los promedios estimados son muy parecidos a los observados, lo cual indica que hay muy poca interacción.

(g) Gráfico de interacción entre condición y género:

```
library(ggplot2)
ggplot(base, aes(x=genero, y=proteina, group = cond)) +
  stat_summary(fun.y="mean", geom="line", aes(linetype = cond))
```

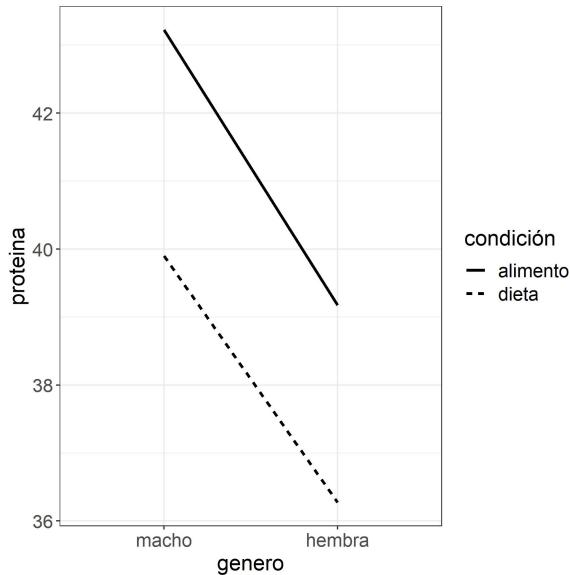


Figura 3.3: Promedios de proteína para combinaciones de condición y género

(h) Distancia entre medias para cada género:

En la Figura 3.3, se compara la media de proteína cuando se proporciona alimento contra la media cuando se tiene dieta para los machos, y también se comparan las medias de ambas condiciones para hembras. Se observa que las distancias entre estas medias cuando se hace para los machos es muy similar a la distancia en el caso de las hembras. Esto es una indicación de que no existe interacción entre condición y género.

(i) Efectos simples:

```
mod1=aov(proteina~cond+genero,data=base)
model.tables(mod1)
```

```
## Tables of effects
## cond
##   alimento    dieta
##     1.56   -1.56
## genero
##   macho    hembra
##     1.92   -1.92
```

3. Efectos de interacción en el modelo:

(a) Modelo con interacción con modelo de **suma nula**:

```
options(contrasts=c("contr.sum","contr.poly"))
mod2=lm(proteina~cond*genero,data=base)
```

(b) Matriz de estructura:

```
contrasts(base$cond)
```

```
## alimento    1
## dieta      -1
```

```
contrasts(base$genero)
```

```
## macho      1
## hembra    -1
```

```
model.matrix(mod2)
```

```
##   (Intercept) cond1 genero1 cond1:genero1
## 1           1     1     1           1
## ...
## 5           1    -1     1          -1
## ...
## 9           1     1    -1          -1
## ...
## 13          1    -1    -1           1
## ...
```

La matriz de estructura tiene una columna para condición con códigos 1 y -1. Similarmente una columna para género con códigos 1 y -1. El -1 corresponde siempre al nivel que no aparece que es dieta o hembra. La columna de interacción es el producto de las otras dos columnas.

(c) Efectos de interacción:

```
mod2$coef
## (Intercept) cond1 generol cond1:generol
##      39.64    1.56     1.92       0.11
```

Los efectos de interacción son todos iguales en valor absoluto (0,11). Esto sucede por tratarse de un diseño 2^2 . Esta cantidad representa cuánto se debe restar o sumar a la media observada en cada tratamiento para obtener promedios en condición de no interacción.

4. Varianza del error:

(a) Boxplot:

```
boxplot(proteina~cond+genero, ylab="proteina", xlab="condición:género", data=base)
```

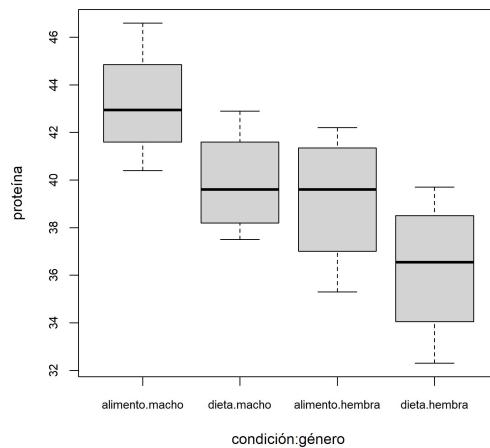


Figura 3.4: Proteína por tratamiento

En la Figura 3.4 se observa que la variabilidad es similar en todos los tratamientos. Se puede esperar que las distribuciones de los 4 tratamientos tengan la misma variabilidad. Por otra parte, se nota que para ambos géneros, la condición con alimento en abundancia produce un mayor promedio de proteína que para la otra condición de dieta regulada. Como la diferencia entre las dos condiciones es similar en ambos géneros, se confirma que no se espera que haya interacción entre condición y género.

(b) Varianzas por tratamiento:

```
(v=tapply(base$proteina, list(base$cond,base$genero),var))

##           macho     hembra
##  alimento    6.52     8.72
##  dieta        5.31     9.60
```

(c) Supuesto de homocedasticidad:

```
bartlett.test(base$proteina~interaction(base$cond,base$genero))

## Bartlett test of homogeneity of variances
##
## data: proteina by interaction(cond, genero)
## Bartlett's K-squared = 0.28, df = 3, p-value = 0.96
```

Ante la hipótesis de homocedasticidad se obtiene una probabilidad asociada de 0,96, por lo que se decide no rechazarla, con lo cual no se tiene evidencia de heterocedasticidad. Se continúa suponiendo que los datos provienen de poblaciones cuyas varianzas son iguales.

(d) Variabilidad del error:

```
(v1=mean(v))

## 7.54
```

5. Prueba de hipótesis para la interacción:

(a) Tabla de análisis de varianza:

```
anova(mod2)

## Analysis of Variance Table
## Response: proteina
##             Df Sum Sq Mean Sq F value    Pr(>F)
## cond         1 38.75  38.75   5.14    0.04 *
## genero       1 58.91  58.91   7.82    0.02 *
## cond:genero  1  0.18   0.18   0.02    0.88
## Residuals    12 90.44   7.54
```

(b) Hipótesis:

Solo interesa probar en esta etapa si existe interacción entre condición y género. No interesa probar los efectos simples si no se ha eliminado la interacción. La hipótesis es $H_0 : (\alpha\beta)_{ij} = 0$. Esta hipótesis dice que el efecto de la condición de alimento sobre la proteína promedio en sangre es independiente del género, es decir, es el mismo en cualquiera de los dos géneros.

(c) Cuadrado medio residual:

```
anova(mod2) [4,3]
```

```
## 7.54
```

Este valor coincide con la media de las varianzas obtenida anteriormente y es una medida de la variabilidad de la respuesta dentro de cada tratamiento.

(d) Cuadrado medio de interacción:

El cuadrado medio de interacción es 0,18 y mide la magnitud general de los efectos de interacción. Si esta magnitud tiende a ser pequeña se tiene un caso con poca interacción, lo que sería una evidencia débil de interacción entre condición y género.

(e) Valor de F:

```
anova(mod2) [3,4]
```

```
## 0.02
```

Este valor se obtiene al dividir el cuadrado medio de interacción entre el cuadrado medio residual, es decir 0,18 entre 7,54.

(f) Probabilidad y conclusión:

```
anova(mod2) [3,5]
```

```
## 0.88
```

```
(pf(0.02,1,12,lower.tail=F))
```

```
## 0.88
```

La variabilidad de los efectos de interacción es sumamente baja con respecto a la variabilidad residual y la probabilidad asociada es altísima ($p = 0,88$), con lo cual no se rechaza la hipótesis de independencia. Se asume que el efecto de la condición de alimento es igual para machos y hembras.

6. Estimaciones bajo el modelo sin interacción:

(a) Modelo sin interacción:

```
mod3=lm(proteina~cond+genero,data=base)
```

(b) Efectos simples:

```
mod3$coef
```

```
## (Intercept) cond1 genero1
##      39.64   1.56   -1.92
```

El efecto de la condición 1 (alimento en abundancia) es 1,56, lo que indica que con alimento el nivel de proteína en sangre sube 1,56 gr/ml con respecto al promedio general. Similarmente el efecto de la condición 2 (dieta regulada) es que baja el nivel promedio de proteína 1,56 gr/ml.

Los machos (género 1) presentan un nivel de proteína que está 1,92 gr/ml sobre la media general, mientras que las hembras tienen una media de proteína 1,92 gr/ml por debajo de la media general.

(c) Hipótesis:

Puesto que no hay interacción se desea simplemente verificar si hay un efecto de la condición sobre la respuesta promedio sin tomar en cuenta el género. Entonces la hipótesis nula es:

$$H_0 : \alpha_i = 0$$

Esta hipótesis dice que la condición de alimento no tiene efecto sobre la proteína promedio, es decir, ambas condiciones producen el mismo promedio de proteína en sangre. Lo anterior sucede en ambos géneros.

(d) Cuadrado medio de condición y cuadrado medio de género:

```
anova(mod3) [1, 3]
```

```
## 38.75
```

```
anova(mod3) [2, 3]
```

```
## 58.91
```

El cuadrado medio de condición es 38,75 y es una medida de la distancia entre las medias de las dos condiciones. Similarmente el cuadrado medio de género es 58,91 y es una medida de la distancia entre las medias de los dos géneros.

(e) Cálculo de cuadrados medios:

```
table(base$cond)

## cond
##   alimento dieta
##     8     8

table(base$genero)

## genero
##   macho hembra
##     8     8

c(sum(8*alfa^2)/1, sum(8*beta^2)/1)

## 38.75 58.91
```

(f) Cuadrado medio residual:

```
v2=anova(mod3) [3,3]
c(v1,v2)

## 7.54 6.97
```

La varianza que se obtiene a partir de la media de las varianzas (7,54) asume un modelo con interacción, donde los residuales se calculan con respecto a la media observada. En cambio el CMRes en este caso (6,97) se obtiene con los residuales calculados con respecto a las medias estimadas en el modelo sin interacción. Debido a que la interacción es muy pequeña, las medias se movieron muy poco y las dos cantidades obtenidas son muy similares.

(g) Aumento en la suma de cuadrados residual:

```
anova(mod2) [4,2]

## 90.44

anova(mod3) [3,2]

## 90.62

anova(mod3) [3,2]-anova(mod2) [4,2]

## 0.18
```

La diferencia entre las sumas de cuadrados residual de los dos modelos es 0,18, la cual coincide con la suma de cuadrados de interacción en el anova del punto 5(a). Tiene sentido que el modelo que no tiene interacción tenga una mayor suma de cuadrados residual porque ahora los residuales se calculan con respecto a un promedio estimado con el modelo sin interacción. En general estos residuales serán un poco más grandes que los calculados con respecto a los promedios observados. Puesto que la suma de cuadrados total es fija y se mantiene la misma suma de cuadrados de tratamiento en ambos modelos, la parte correspondiente a la suma de cuadrados residual en el modelo sin interacción absorbe la parte que correspondía a la suma de cuadrados de interacción. Todas las sumas de cuadrados deben sumar la misma cantidad independientemente del modelo utilizado, esa cantidad es la suma de cuadrados total.

(h) Conclusión:

```
anova(mod3)

## Analysis of Variance Table
##
## Response: proteina
##           Df Sum Sq Mean Sq F value Pr(>F)
## cond       1 38.75  38.75   5.56   0.035 *
## genero     1 58.91  58.91   8.45   0.012 *
## Residuals 13 90.62   6.97
```

La variabilidad asociada a las medias de condición es 5,56 veces la variabilidad residual, lo cual indica que las medias están bastante distantes entre sí; esto se confirma con una probabilidad asociada baja ($p = 0,035$). Puesto que esta probabilidad está debajo del límite establecido de 0,05, se rechaza la hipótesis de que no hay efecto de la condición. Se concluye, con un nivel de significancia de 0,05, que las dos condiciones producen medias de proteína diferentes (independientemente del género).

(i) Cálculo de cota inferior:

```
mod3$coef

## (Intercept) cond1 genero1
##      39.64    1.56    1.92

contrasts(base$cond)

## alimento    1
## dieta      -1
```

Para construir los contrastes se toma en cuenta que el factor condición toma valores 1 y -1 según sea alimento o dieta, respectivamente, mientras que para el factor género se pone 0 para que el modelo calcule la media marginal de cada condición. Esto se puede hacer solamente cuando se usa el modelo de suma nula, ya que cuando se usa el modelo de tratamiento referencia, si se pone un 0 se estaría indicando que se use el nivel de referencia. Entonces el vector para el primer contraste es $v1 = c(1, 1, 0)$ que produce la estimación de la media marginal de alimento, mientras que $v2 = c(1, -1, 0)$ produce la media marginal de dieta. Puesto que el coeficiente de condición es 1.56, el cual corresponde al efecto simple de alimento, se deduce que el efecto simple de dieta es -1.56. De aquí se concluye que la media marginal de alimento es mayor que la de dieta y la diferencia entre ellas se podría calcular simplemente como $2 * 1.56 = 3.12$. Otra forma de obtener este mismo valor es restando los dos vectores de contrastes y multiplicando por el vector de coeficientes.

```
vl=c(1,1,0); v2=c(1,-1,0); c=v1-v2
(L=t(c) %*% mod3$coef)
```

```
## 3.11
```

```
ee=sqrt(t(c) %*% vcov(mod3) %*% c)
t=qt(0.95,13)
(LIM=L-t*ee)
```

```
## 0.77
```

De forma puntual se obtiene en las muestras que el promedio de proteína en sangre cuando se provee de alimento en abundancia es 3,11 gr/ml mayor que con dieta; sin embargo, este resultado está basado solo en una muestras de 16 individuos (8 de cada género). A partir de estos resultados se puede esperar con 95% de confianza, que con alimento en abundancia el promedio de proteína en sangre sea al menos 0,77 gr/ml mayor que con dieta. Esto sucede tanto en machos como en hembras por el hecho de que se asume que no hay interacción entre condición de alimentación y género.

3.3 Tortugas 2

Se desea determinar si la falta de alimento afecta el nivel de proteínas en sangre en tortugas de las especies Kinosternum scorpioides (de agua dulce) y Chelonia midas (de agua salada), y si el efecto es diferente de una especie a otra.

Se escogen 12 ejemplares de cada especie y se les asigna aleatoriamente a cada tortuga una de las siguientes condiciones: 1) dieta estricta, 2) dieta balanceada y 3) alimento en abundancia. Se registra el nivel de proteína en gramos por decilitro.

3.3.1 Ejercicios

1. Preparación:

- (a) Cargue el archivo `tortugas2.csv`. Verifique que la variable condición sea un factor con los niveles adecuados.
- (b) ¿Cuáles son los factores que se incluyen en el experimento?
- (c) ¿En qué aspectos debe concentrarse el análisis?

2. Efectos:

- (a) Obtenga la media general de la respuesta.
- (b) Obtenga los efectos simples de cada condición, es decir, las diferencias entre cada media de condición y la media general (α_1 , α_2 y α_3).
- (c) Obtenga los efectos simples de cada especie (β_1 y β_2).
- (d) Obtenga los promedios observados de los 6 tratamientos.
- (e) Obtenga manualmente los promedios estimados bajo el modelo sin interacción.
- (f) Obtenga un gráfico de los valores de proteína para interpretar gráficamente si existe interacción entre condición y especie.
- (g) Visualice en el gráfico las medias estimadas. Compare los promedios observados con los promedios estimados y a partir de ahí obtenga la cantidad que debe sumarse o restarse a cada promedio observado para obtener una estimación que cumpla con el supuesto de independencia.
- (h) Obtenga los efectos simples y de interacción usando la función `model.tables` (use el modelo con interacción con la función `aov`).
- (i) Compare estos resultados con los obtenidos anteriormente.

3. Varianza del error:

- (a) Visualice los 6 tratamientos con un gráfico de cajas donde se incluyan los dos factores. Use este gráfico para ver si se puede esperar que los datos de los 6 tratamientos provengan de distribuciones con la misma varianza. Además, confirme si se puede esperar interacción entre condición y especie.
- (b) Obtenga las varianzas de los seis tratamientos.
- (c) Verifique si se cumple el supuesto de homocedasticidad.
- (d) Obtenga una medida de la variabilidad del error asumiendo homocedasticidad.

4. Estimaciones bajo el modelo con interacción:

- (a) Cambie al modelo de **suma nula**. Escriba el modelo con `lm` (llámelo `mod2`).
- (b) Observe la matriz de estructura usando `model.matrix(mod2)`. Vea a qué corresponde cada columna. Ponga atención a los códigos y relacionelos con los niveles de cada factor.
- (c) Extraiga los coeficientes y obtenga efectos simples y de interacción que no aparecen debido a la restricción de suma nula.
- (d) Use los coeficientes y un vector de contrastes para calcular manualmente el promedio de chelonia-estricta y compárelo con el promedio observado de ese tratamiento.
- (e) Obtenga el promedio para los otros tratamientos.

5. Hipótesis de independencia:

- (a) Obtenga la tabla de análisis de varianza para probar la hipótesis de independencia entre condición y especie.
- (b) Escriba la hipótesis nula con símbolos y con palabras.
- (c) Observe el cuadrado medio residual y compárelo con la estimación de la varianza del error obtenida más arriba. Interprete qué significa el cuadrado medio residual.

- (d) Obtenga los grados de libertad de cada fuente de variación y justifíquelo.
- (e) Interprete qué se está midiendo con el cuadrado medio de condición y el cuadrado medio de especie.
- (f) Asegúrese que puede obtener estas cantidades a partir de los efectos simples.
- (g) A partir de las estimaciones de los efectos de la interacción verifique la suma de cuadrados de interacción.
- (h) Interprete qué se está midiendo con el cuadrado medio de interacción.
- (i) Obtenga el valor F asociado a la hipótesis establecida, obtenga la probabilidad asociada y concluya.

6. Comparaciones bajo el modelo con interacción:

- (a) Puesto que se ha encontrado que sí hay una interacción importante entre la condición y la especie, ahora hay que investigar cómo actúan las condiciones en cada especie. Escriba el modelo con interacción y escriba los contrastes para comparar las condiciones dentro de cada especie.
- (b) Verifique que los 3 contrastes dentro de cada especie no son ortogonales, pero que cada bloque de 3 contrastes es ortogonal al otro bloque.
- (c) Escriba una matriz que permita el cálculo de los contrastes.
- (d) Estime cada contraste y su varianza.
- (e) Pruebe de forma simultánea las hipótesis que establecen que cada contraste es igual a cero. Puesto que los contrastes no son ortogonales debe usarse la corrección de Bonferroni, es decir, dividir el nivel de significancia por el número de comparaciones.
- (f) Obtenga las cotas inferiores que permitan cuantificar las diferencias que son significativas.

- (g) La función `emmeans` de la librería con el mismo nombre se puede utilizar para realizar las comparaciones múltiples de forma automática. En este caso, puesto que hay interacción, debe indicarse que se hacen las comparaciones para los promedios de A dentro de cada nivel de B, y se pide que se use la corrección de Bonferroni de la siguiente forma: `emmeans(mod, pairwise~A|B, adjust="bonferroni")`. Las probabilidades que se obtienen en la salida de esta función se deben interpretar de forma diferente; se pueden comparar directamente contra α si se tienen pruebas de dos colas, o contra 2α si son de una cola. Utilice esas probabilidades directamente y verifique que se llega a las mismas conclusiones que en el punto (e).
- (h) Si se quieren obtener las mismas probabilidades del punto (e), las probabilidades de la salida de `emmeans` deben dividirse por 6, ya que se tenían 3 comparaciones en cada grupo con una cola, o también pueden multiplicarse las del punto (e) por 6 para obtener las de la función `emmeans`. Multiplique las probabilidades del punto (e) por 6 y compárelas con las obtenidas con esta función.
- (i) Una vez que se ha aplicado `emmeans`, se pueden obtener intervalos de confianza. Por ejemplo, si el resultado anterior se guardó en el objeto `em1`, entonces con `confint(em1)` se obtienen los intervalos de confianza (dos límites); sin embargo, esta función calcula todos los intervalos dentro de cada grupo aunque la diferencia no haya sido significativa. Para obtener los intervalos solo de algunos pares o solo cotas inferiores o superiores, se puede tomar la información de esta salida y construirlos un poco a pie, pero sin necesidad de hacer todos los vectores que usualmente se hacen cuando se construyen totalmente a pie. Obtenga los intervalos automáticamente y ajústelos según sea necesario para que dar las cotas del punto (f).
- (j) ¿Qué se concluye?

3.3.2 Solución

1. Preparación:

(a) Lectura:

```
base=read.csv("tortugas2.csv")
base$cond=factor(base$cond)
levels(base$cond)=c("estricta","balanceada","abundancia")
base$especie=factor(base$especie)
```

(b) Factores:

Hay dos factores: el que se está investigando es la condición de alimento (factor de diseño) y se tiene otro que es la especie de tortuga.

(c) Enfoque del análisis:

Interesa ver primero si el efecto que tiene la condición de alimento es el mismo en ambas especies. Si el efecto es el mismo en ambas especies, entonces puede darse una conclusión general sobre el efecto que tiene cada condición en la proteína promedio en sangre sin diferenciar esta conclusión por especie, de lo contrario, debe cuantificarse ese efecto por separado en cada especie.

2. Efectos:

(a) Media general:

```
(medgen=mean(base$proteina))
```

```
## 36
```

(b) Efectos simples de condición:

```
(alfa=tapply(base$proteina,base$cond,mean)-medgen)
```

```
## estricta balanceada abundancia
##      -3.38      2.38     1.00
```

(c) Efectos simples de especie:

```
(beta=tapply(base$proteina,base$especie,mean)-medgen)
```

```
## chelonia kynosternon
##      1       -1
```

(d) Promedios observados:

```
(med=tapply(base$proteina,list(base$cond,base$especie),mean))
```

```
##          chelonia kynosternon
## estricta     32.75      32.50
## balanceada   42.00      34.75
## abundancia   36.25      37.75
```

(e) Promedios estimados bajo el modelo sin interacción:

```
m11=medgen+alfa[1]+beta[1]
m21=medgen+alfa[2]+beta[1]
m31=medgen+alfa[3]+beta[1]
m12=medgen+alfa[1]+beta[2]
m22=medgen+alfa[2]+beta[2]
m32=medgen+alfa[3]+beta[2]
(mest=matrix(c(m11,m21,m31,m12,m22,m32),nrow=3,
             dimnames=list(c("estricta","balanceada","abundancia"),
                           c("chelonia","kynosternon"))))
```

```
##          chelonia kynosternon
## estricta     33.63      31.63
## balanceada   39.38      37.38
## abundancia   38.00      36.00
```

(f) Gráfico de interacción entre condición y especie:

```
library(ggplot2)
ggplot(base, aes(x=especie, y=proteina, group = cond)) +
  stat_summary(fun.y="mean", geom="line", aes(linetype = cond))
```

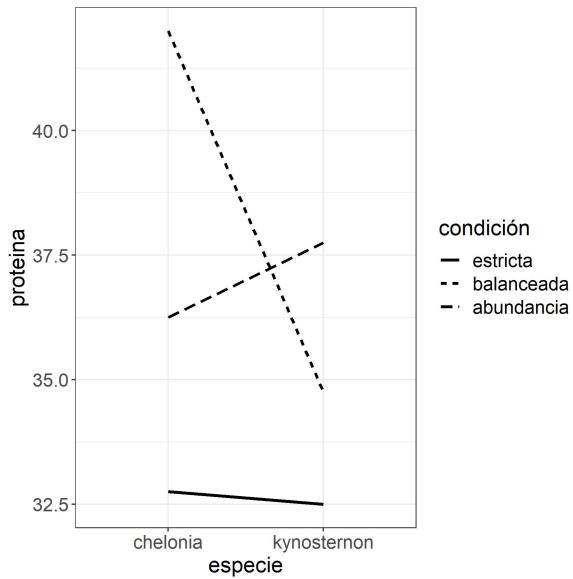


Figura 3.5: Promedios de proteína para combinaciones de condición y especie

En la Figura 3.5 se observa que la distancia entre la media de proteína en sangre cuando se proporciona dieta estricta con respecto a la media cuando se tiene dieta balanceada para chelonia es mucho mayor a esta distancia en el caso de kynosternon. Esto es una fuerte indicación de que existe interacción entre condición y especie.

(g) Comparación entre promedios observados y promedios estimados:

```
(efint=med-mest)

##          chelonia    kynosternon
## estricta      -0.88       0.88
## balanceada     2.63      -2.63
## abundancia    -1.75       1.75
```

Estas cantidades deben restarse al promedio observado para obtener la estimación que asegura que no haya interacción. Por ejemplo, en el caso de dieta estricta para chelonia debería sumarse 0,88, en el caso de dieta balanceada debería restarse 2,63 y en el caso de alimento en abundancia debería sumarse 1,75, de esta forma los promedios distarían entre sí una cantidad idéntica a la que se obtendría en el caso correspondiente para kynosternon. Estas cantidades se llaman efectos de interacción.

(h) Efectos simples y de interacción:

```
mod1=aov(proteina~cond*especie,data=base)
model.tables(mod1)

## Tables of effects
## cond
##   estricta balanceada abundancia
##   -3.375      2.375      1.000
## especie
##   chelonia kynosternon
##   1           -1
## cond:especie
##                   especie
##   cond          chelonia    kynosternon
##   estricta      -0.88       0.88
##   balanceada     2.63      -2.63
##   abundancia    -1.75       1.75
```

(i) Comparación:

Se obtienen los mismos efectos simples (alfa y beta), así como los mismos efectos de interacción.

3. Varianza del error:

(a) Boxplot:

```
boxplot(proteina~cond+especie, ylab="proteina", xlab="condición:especie", data=base)
```

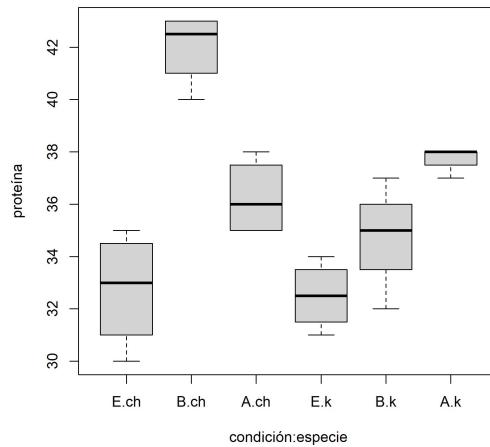


Figura 3.6: Proteína por tratamiento

En la Figura 3.6 se observa mayor variabilidad en el tratamiento de estricta para chelonia, mientras que hay mucho menos variabilidad en el tratamiento de abundancia para kynosternon. Se puede esperar que las distribuciones de los 6 tratamientos no tengan la misma variabilidad. Por otra parte, se nota que para chelonia, el tratamiento con dieta balanceada produce un mayor promedio de proteína que para las otras condiciones, mientras que para kynosternon la diferencia es muy pequeña, por lo que se confirma que sí se espera que haya interacción entre condición y especie.

(b) Varianzas por tratamiento:

```
(v=tapply(base$proteina, list(base$cond, base$especie), var))
```

	chelonia	kynosternon
## estricta	4.92	1.67
## balanceada	2.00	4.25
## abundancia	2.25	0.25

(c) Supuesto de homocedasticidad:

```
bartlett.test(base$proteina~interaction(base$cond, base$especie))
```

```
## Bartlett test of homogeneity of variances
## data: proteina by interaction(cond, especie)
## Bartlett's K-squared = 5.21, df = 5, p-value = 0.39
```

Ante la hipótesis de homocedasticidad se obtiene una probabilidad asociada de 0,39, por lo que se decide no rechazarla, con lo cual no se tiene evidencia de heterocedasticidad. En las varianzas de los diferentes tratamientos se ven grandes diferencias ya que la menor es 0,25 y la mayor 4,91; sin embargo, la prueba no logra demostrar que las varianzas son diferentes, seguramente por la poca cantidad de réplicas en cada tratamiento. Se continúa suponiendo que los datos provienen de poblaciones cuyas varianzas son iguales.

(d) Variabilidad del error:

```
mean(v)
```

```
## 2.56
```

4. Estimaciones bajo el modelo con interacción:

(a) Modelo con interacción:

```
options(contrasts=c("contr.sum", "contr.poly"))
mod2=lm(proteina~cond*especie,data=base)
```

(b) Matriz de estructura:

```
contrasts(base$cond)
```

```
## estricta      1    0
## balanceada   0    1
## abundancia   -1   -1
```

```
contrasts(base$especie)
```

```
## chelonia      1
## kynosternon  -1
```

```
model.matrix(mod2)
```

```
##   (Intercept) cond1 cond2 especiel cond1:especiel cond2:especiel
## 1           1     1     0     1           1           0
## ...
## 5           1     0     1     1           0           1
## ...
## 9           1    -1    -1     1          -1          -1
## ...
## 13          1     1     0    -1          -1           0
## ...
## 17          1     0     1    -1           0          -1
## ...
## 21          1    -1    -1    -1           1           1
## ...
```

La matriz de estructura tiene dos columnas para condición con códigos 0, 1 y -1, puesto que el factor condición tiene 3 niveles. Similarmente una columna para especie con códigos 1 y -1, puesto que el factor especie tiene dos niveles. El -1 corresponde siempre al nivel que no aparece que es kynosternon o abundancia. La columna de interacción es el producto de las otras dos columnas. Así, por ejemplo, las filas 1 a 4 corresponden a las tortugas de la especie chelonia (especie1 = 1) con condición estricta (cond1 = 1 y cond2 = 0), las filas 21 a 24 corresponden a las tortugas de la especie kynosternon (especie1 = -1) con condición abundancia (cond1 = -1 y cond2= -1).

(c) Efectos obtenidos con los coeficientes del modelo:

```
mod2$coef
## (Intercept) cond1 cond2 especie1 cond1:especie1 cond2:especie1
##      36.00   -3.38    2.38     1.00      -0.88      2.63
```

El efecto de cond1 (estricta) es -3,38, el de cond2 (dieta regulada) es 2,38. El de cond3 (abundancia) se obtiene a partir de los otros dos por la restricción de suma nula como: $-(-3,38 + 2,38) = 1$. El efecto de especie1 (chelonia) es 1 y el de especie2 (kynosternon) es -1.

El efecto de interacción para cond1:especie1 es -0,88, por lo que para cond1:especie2 es 0,88. El efecto de interacción para cond2:especie1 es 2,63, por lo que para cond2:especie2 es -2,63. El efecto de interacción para cond3:especie1 se obtiene a partir de cond1:especie1 y cond2:especie1, para que los 3 sumen cero: $-(-0,88 + 2,63) = -1,75$. A partir de este se obtiene el último efecto de interacción para cond3:especie2 que es 1,75.

(d) Promedio de chelonia-estricta:

El vector para el cálculo del promedio del tratamiento chelonia-estricta es $[1, 1, 0, 1, 1, 0]^T$, el cual se multiplica por el vector de coeficientes estimados del modelo:

```
coef=mod2$coef
c(1,1,0,1,1,0)%*%coef
## 32.75
```

(e) Promedio para todos los otros tratamientos:

A continuación se dan los vectores necesarios para obtener los promedios de los diferentes tratamientos:

Chelonia-balanceada:	$[1, 0, 1, 1, 0, 1]^T$
Chelonia-abundancia:	$[1, -1, -1, 1, -1, -1]^T$
Kynostenron-estricta:	$[1, 1, 0, -1, -1, 0]^T$
Kynostenron-balanceada:	$[1, 0, 1, -1, 0, -1]^T$
Kynostenron-abundancia:	$[1, -1, -1, -1, 1, 1]^T$

```

ch.e = c(1,1,0,1,1,0)
ch.b = c(1,0,1,1,0,1)
ch.a = c(1,-1,-1,1,-1,-1)
ky.e = c(1,1,0,-1,-1,0)
ky.b = c(1,0,1,-1,0,-1)
ky.a = c(1,-1,-1,-1,1,1)
h = cbind(ch.e, ch.b, ch.a, ky.e, ky.b, ky.a)

(L = t(h) %*% coef)

## ch.e 32.75
## ch.b 42.00
## ch.a 36.25
## ky.e 32.50
## ky.b 34.75
## ky.a 37.75

```

5. Hipótesis de independencia:

(a) Tabla de análisis de varianza:

```

anova (mod2)

## Analysis of Variance Table
##
## Response: proteina
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cond       2 144.25   72.13   28.22 2.82e-06 ***
## especie    1  24.00   24.00    9.39   0.007 **
## cond:especie 2  85.75   42.88   16.78 7.71e-05 ***
## Residuals  18  46.00    2.56

```

(b) Hipótesis nula:

$$H_0 : (\alpha\beta)_{ij} = 0$$

Esta hipótesis dice que el efecto de la condición de alimento sobre la proteína promedio en sangre no depende de la especie, es decir, es el mismo en cualquiera de las dos especies.

(c) Cuadrado medio residual:

```

anova (mod2) [4, 3]

## 2.56

```

Este valor coincide con la media de las varianzas obtenida anteriormente y es una medida de la variabilidad de la respuesta dentro de cada tratamiento.

(d) Grados de libertad:

Para condición hay 2 grados de libertad porque son 3 condiciones, mientras que para especie hay un grado de libertad porque son 2 especies. Para la interacción entre condición y especie hay 2 grados de libertad que se obtiene del producto de los grados de libertad de cada uno de los factores (2×1).

(e) Cuadrado medio de condición y cuadrado medio de especie:

```
anova(mod2) [1:2, 3]
```

```
## 72.13 24
```

El cuadrado medio de condición es 72,13, el cual es una medida de la distancia entre las medias de las tres condiciones. Similarmente el cuadrado medio de especie es 24, el cual es una medida de la distancia entre las medias de las dos especies.

(f) Cálculo de cuadrados medios:

```
table(base$cond)
```

```
## cond
##   estricta balanceada abundancia
##     8          8          8
```

```
table(base$especie)
```

```
## especie
##   chelonia kynosternon
##     12          12
```

```
c(sum(8*alfa^2)/2, sum(12*beta^2)/1)
```

```
## 72.13 24
```

(g) Suma de cuadrados de interacción:

```
table(base$cond, base$especie)
```

```
##           especie
##   cond      chelonia kynosternon
##   estricta        4          4
##   balanceada       4          4
##   abundancia       4          4
```

```
sum(4*efint^2)
```

```
## 85.75
```

(h) Cuadrado medio de interacción:

El cuadrado medio de interacción es 42,88, el cual mide la magnitud general de los efectos de interacción. Si esta magnitud tiende a ser pequeña se tiene un caso con poca interacción, lo que sería una evidencia débil de interacción entre condición y especie. En caso contrario, si esta magnitud tiende a ser grande, hay más evidencia de presencia de interacción.

(i) Conclusión:

```
(f=42.88/2.56)
## 16.77
(pf(f,2,18,lower.tail=F))
## 7.7e-05
```

Se tiene que la variabilidad de los efectos de interacción es sumamente alta con respecto a la variabilidad residual. Al ser la primera casi 17 veces la segunda, se tiene una probabilidad asociada bajísima ($p < 0,001$), con lo cual se rechaza la hipótesis de no interacción. Se concluye que el efecto de la condición de alimento no es el mismo cuando se trata de Chelonia midas que en el caso de Kynosternum scorpioides.

6. Comparaciones bajo el modelo con interacción:

(a) Modelo con interacción y contrastes:

El modelo es:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

El primer subíndice indica la condición y el segundo la especie. Como el factor de interés es la condición, debe fijarse la especie y comparar para diferentes condiciones dada cada especie. Para tener estimaciones positivas de las diferencias de promedios, se crean los contrastes de tal forma que el contraste estimado sea positivo, es decir, se pone primero aquella media cuya estimación sea mayor.

Se tienen los contrastes para chelonia:

$$\begin{aligned}\mu_{21} - \mu_{11} \\ \mu_{21} - \mu_{31} \\ \mu_{31} - \mu_{11}\end{aligned}$$

Los contrastes para kynosternon son:

$$\begin{aligned}\mu_{22} - \mu_{12} \\ \mu_{32} - \mu_{22} \\ \mu_{32} - \mu_{12}\end{aligned}$$

(b) Ortogonalidad:

Tomando el vector de promedios

$$(\mu_{11}, \mu_{21}, \mu_{31}, \mu_{12}, \mu_{22}, \mu_{32})$$

Los vectores para obtener el primer grupo de hipótesis (chelonia) son:

$$\begin{aligned}(-1, 1, 0, 0, 0, 0) \\ (0, 1, -1, 0, 0, 0) \\ (1, 0, -1, 0, 0, 0)\end{aligned}$$

```
v1 = c(-1, 1, 0, 0, 0, 0)
v2 = c(0, 1, -1, 0, 0, 0)
v3 = c(1, 0, -1, 0, 0, 0)
c(v1%*%v2, v1%*%v3, v2%*%v3)
```

```
## 1 -1 1
```

Estos 3 vectores no son ortogonales, por lo que hay que hacer corrección.

De forma similar sucede para el segundo grupo de hipótesis (kynosternon):

$$\begin{aligned}(0, 0, 0, -1, 1, 0) \\ (0, 0, 0, 0, -1, 1) \\ (0, 0, 0, -1, 0, 1)\end{aligned}$$

```
v4 = c(0, 0, 0, -1, 1, 0)
v5 = c(0, 0, 0, 0, -1, 1)
v6 = c(0, 0, 0, -1, 0, 1)
c(v4%*%v5, v4%*%v6, v5%*%v6)
```

```
## -1 1 1
```

Estos 3 vectores tampoco son ortogonales.

Ahora comparamos los dos bloques entre sí:

```
c(v1%*%v4, v1%*%v5, v1%*%v6, v2%*%v4, v2%*%v5, v2%*%v6, v3%*%v4, v3%*%v5, v3%*%v6)
```

```
## 0 0 0 0 0 0 0 0
```

El primer grupo de vectores sí es ortogonal con respecto al segundo grupo de vectores. Por lo que dentro de cada grupo de hipótesis debe hacerse la corrección de Bonferroni, pero cada grupo de hipótesis se puede probar de forma independiente al otro grupo. Por lo tanto, se debe usar corrección de Bonferroni tomando en cuenta que en cada caso son 3 hipótesis.

(c) Matriz de contrastes:

Para hacer el cálculo de los contrastes se deben restar los dos vectores que producen cada una de las medias del contraste:

```
ch.be = ch.b-ch.e
ch.ba = ch.b-ch.a
ch.ae = ch.a-ch.e
ky.be = ky.b-ky.e
ky.ab = ky.a-ky.b
ky.ae = ky.a-ky.e
(h=cbind(ch.be, ch.ba, ch.ae, ky.be, ky.ab, ky.ae))
```

```
##      ch.be ch.ba ch.ae ky.be ky.ab ky.ae
## [1,]    0     0     0     0     0     0
## [2,]   -1     1    -2    -1    -1    -2
## [3,]    1     2    -1     1    -2    -1
## [4,]    0     0     0     0     0     0
## [5,]   -1     1    -2     1     1     2
## [6,]    1     2    -1    -1     2     1
```

(d) Estimación del contraste y varianza:

```
(L=t(h) %*%coef)
```

```
## ch.be 9.25
## ch.ba 5.75
## ch.ae 3.50
## ky.be 2.25
## ky.ab 3.00
## ky.ae 5.25
```

```
(var=diag(t(h) %*%vcov(mod2) %*%h))
```

```
## ch.be ch.ba ch.ae ky.be ky.ab ky.ae
## 1.28 1.28 1.28 1.28 1.28 1.28
```

Todas las varianzas son iguales porque es un diseño balanceado.

(e) Prueba simultánea de las hipótesis:

```

ee=sqrt(var)
t=L/ee
row.names(t)=row.names(L)
p=pt(t,18,lower.tail = F)
row.names(p)=row.names(L)
p

## ch.be 0.0000
## ch.ba 0.0000
## ch.ae 0.0031
## ky.be 0.0310
## ky.ab 0.0081
## ky.ae 0.0001

```

Las probabilidades deben compararse contra $\alpha/3 = 0.017$ si se quieren realizar pruebas de una cola, o contra $\alpha/6 = 0.0083$ para pruebas de dos colas. Todas las hipótesis se rechazan (tanto para una cola como para dos colas), con excepción de una. Se concluye que en cada especie el nivel promedio de proteína en sangre es diferente en casi todas las condiciones, excepto para kynosternon, donde no se ha encontrado que la dieta balanceada produzca un nivel promedio de proteína en sangre más alto que la dieta estricta.

(f) Cálculo de cotas inferiores:

```

tc1=qt(1-0.05/3,18)
tc2=qt(1-0.05/2,18)
lim1=L[1:3]-tc1*ee[1:3]
lim2=L[5:6]-tc2*ee[5:6]
names(lim1)=row.names(L)[1:3]
names(lim2)=row.names(L)[5:6]
lim1
lim2

```

```

## ch.be ch.ba ch.ae
## 6.65 3.15 0.90

## ky.ab ky.ae
## 0.63 2.88

```

(g) Probabilidades con emmeans:

```

library(emmeans)
(em1=emmeans(mod2, pairwise~cond|especie, adjust="bonferroni"))

## $emmeans
## especie = chelonia:
##   cond      emmean    SE df lower.CL upper.CL
##   estricta  32.8 0.799 18     31.1    34.4
##   balanceada 42.0 0.799 18     40.3    43.7
##   abundancia 36.2 0.799 18     34.6    37.9
##
## especie = kynosternon:
##   cond      emmean    SE df lower.CL upper.CL
##   estricta  32.5 0.799 18     30.8    34.2
##   balanceada 34.8 0.799 18     33.1    36.4
##   abundancia 37.8 0.799 18     36.1    39.4
##
## Confidence level used: 0.95
##
## $contrasts
## especie = chelonia:
##   contrast           estimate    SE df t.ratio p.value
##   estricta - balanceada -9.25 1.13 18 -8.183 <.0001
##   estricta - abundancia -3.50 1.13 18 -3.096 0.0187
##   balanceada - abundancia 5.75 1.13 18 5.087 0.0002
##
## especie = kynosternon:
##   contrast           estimate    SE df t.ratio p.value
##   estricta - balanceada -2.25 1.13 18 -1.990 0.1858
##   estricta - abundancia -5.25 1.13 18 -4.644 0.0006
##   balanceada - abundancia -3.00 1.13 18 -2.654 0.0485
##
## P value adjustment: bonferroni method for 3 tests

```

Todas las probabilidades (excepto la comparación de estricta vs balanceada para kynosternon) son menores a 0.05 (una cola) y también a 0.025 (dos colas), por lo que se llega a la misma conclusión del punto (e).

(h) Verificación de probabilidades:

```

p*6

## ch.be 0.0000
## ch.ba 0.0002
## ch.ae 0.0187
## ky.be 0.1858
## ky.ab 0.0485
## ky.ae 0.0006

```

(i) Verificación de cotas inferiores:

```
confint(em1)$contrasts

## $contrasts
## especie = chelonia:
##   contrast           estimate    SE df lower.CL upper.CL
##   estricta - balanceada -9.25 1.13 18 -12.23 -6.2667
##   estricta - abundancia -3.50 1.13 18 -6.48 -0.5167
##   balanceada - abundancia  5.75 1.13 18  2.77  8.7333
##
## especie = kynosternon:
##   contrast           estimate    SE df lower.CL upper.CL
##   estricta - balanceada -2.25 1.13 18 -5.23  0.7333
##   estricta - abundancia -5.25 1.13 18 -8.23 -2.2667
##   balanceada - abundancia -3.00 1.13 18 -5.98 -0.0167
##
## Confidence level used: 0.95
## Conf-level adjustment: bonferroni method for 3 estimates
```

```
c(9.25,3.50,5.75)-tc1*1.13
```

```
## 6.65 0.90 3.15
```

```
c(5.25,3.00)-tc2*1.13
```

```
## 2.88 0.63
```

Aunque aparecen en un orden diferente, los resultados concuerdan con los del punto (f).

(j) Conclusión:

Se concluye, con 95% de confianza, que el nivel promedio de proteína en sangre para chelonia es al menos 6,65 gr/ml mayor con dieta balanceada que con dieta estricta y al menos 3,15 gr/ml mayor con dieta balanceada que con abundancia. Para kynosternon, el alimento en abundancia aventaja a la dieta estricta en al menos 2,88 gr/ml. Las otras dos diferencias son pequeñas, por lo que es importante contrastar sus magnitudes con la diferencia mínima que establezca un investigador para determinar si son relevantes.

Capítulo 4

Diseños con tres factores

4.1 Conceptos

Se pueden plantear diseños con más de dos factores, lo cual incrementa la posibilidad de tener interacciones complejas; sin embargo, es difícil pensar en interacciones de muchos factores a la vez (de alto nivel). En un modelo solo deberían incluirse aquellas interacciones que puedan tener sentido desde el punto de vista teórico para el investigador. Aquí se explica el concepto de interacción entre tres factores.

Interacciones triples

Cuando hay tres factores en un experimento, se dice que existe interacción entre los tres factores si la interacción entre un par de factores depende del nivel del tercer factor. En este caso, quien depende de los niveles de otro factor es la interacción entre dos factores, pero no el efecto de un solo factor. En la Figura 4.1 se presenta un ejemplo en el que existe interacción entre los factores A, B y C. Esto puede apreciarse porque al analizar la interacción de A con B, se observa que esta se comporta de forma diferente si se ubica en el nivel C_1 o en el nivel C_2 . En el nivel C_1 hay interacción entre A y B, mientras que en C_2 esa interacción desaparece. La ventaja de que no exista interacción triple está en que la interacción entre A y B, por ejemplo, se podría analizar sin importar el nivel de C. En ese caso se hablaría de interacción entre A y B, indistintamente de C.

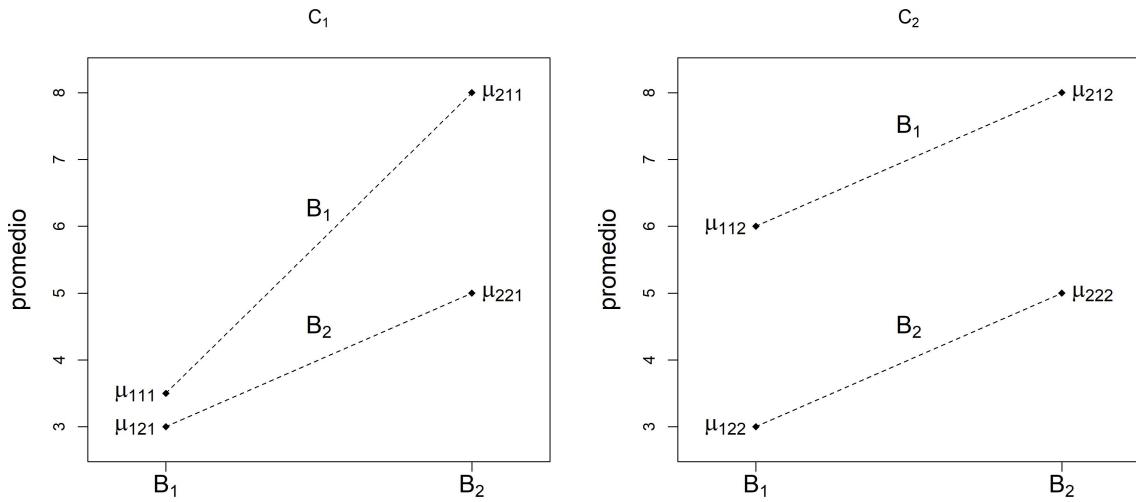


Figura 4.1: Tres factores que interactúan

Nota: para C₁ existe interacción entre A y B (izquierda), mientras que para C₂ no hay interacción (derecha).

La presencia de interacción triple implica automáticamente que hay interacción entre dos factores en al menos uno de los niveles del tercer factor. De esta forma, si los factores A y B, por ejemplo, no tuvieran interacción para ninguno de los niveles de C, la interacción sería nula en ambos casos y no habría interacción triple entre los tres factores A, B y C.

El modelo que incluye la interacción triple debe considerar las interacciones entre todos los pares de factores. El modelo se escribe de la siguiente forma:

$$\mu_{ijk}^{CT} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}.$$

Este modelo se debe comparar contra un modelo que tiene todas las interacciones dobles pero no tiene la interacción triple.

$$\mu_{ijk}^{ST} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}.$$

Se dice que un modelo es saturado cuando se incluyen todas las posibles interacciones entre los factores que lo componen. En los modelos saturados balanceados la estimación de las medias con ese modelo coinciden con las medias muestrales. El modelo con interacción triple es un modelo saturado, por lo que las medias muestrales son las estimaciones de las medias con ese modelo. Al restar las estimaciones de ambos modelos se obtiene la estimación de los efectos de interacción triple:

$$(\hat{\alpha}\hat{\beta}\gamma)_{ijk} = \hat{\mu}_{ijk}^{CT} - \hat{\mu}_{ijk}^{ST} = \bar{y}_{ijk} - (\bar{y} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha}\hat{\beta}_{ij} + \hat{\alpha}\hat{\gamma}_{ik} + \hat{\beta}\hat{\gamma}_{jk}).$$

Hipótesis sobre la interacción

Se puede pasar del modelo con interacción triple al modelo sin interacción triple, si se asume que esta interacción es nula. Para hacer este supuesto basado en los datos, se pone a prueba la hipótesis nula $H_0 : (\alpha\beta\gamma)_{ijk} = 0$. Se cuantifican las magnitudes de los efectos de interacción triple observados con el cuadrado medio de interacción triple (CMTri). Para realizar el cálculo del CMTri, se obtiene la suma de los efectos de interacción triple al cuadrado, ponderados por el número de réplicas en cada tratamiento. Esta suma se divide por los grados de libertad asociados (producto de los grados de libertad de los tres factores que componen la interacción triple):

$$\text{CMTri} = \frac{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c r_{ijk} (\hat{\alpha}\hat{\beta}\gamma)_{ijk}^2}{(a-1)(b-1)(c-1)}.$$

La decisión sobre el rechazo de la hipótesis se basa en el estadístico F, el cual se construye mediante la razón entre el CMTri y el CMRes. El CMRes puede obtenerse mediante el promedio ponderado de las varianzas observadas de la respuesta en los distintos tratamientos.

El análisis en presencia de interacción triple es más complejo y requiere que el investigador tenga una idea clara de cuál es el factor más importante para el cual se hacen comparaciones. En tal caso se van fijando los niveles de los otros dos factores y se comparan las medias entre los niveles del factor elegido.

4.2 Refrescos

Una empresa embotelladora de refrescos está interesada en obtener alturas de llenado más uniformes en las botellas que se fabrican en su proceso de manufactura. Teóricamente, una máquina llena cada botella a la altura objetivo correcta, pero en la práctica, existe variación en torno a este objetivo, y a la embotelladora le gustaría entender mejor las fuentes de esta variabilidad y, en última instancia, reducirla. El ingeniero del proceso puede controlar tres variables durante el proceso de llenado:

- el porcentaje de carbonatación,
- la presión de operación en el llenador y
- las botellas producidas por minuto o rapidez de línea.

Es sencillo controlar la presión y la rapidez, pero el porcentaje de carbonatación es más difícil de controlar durante la manufactura real debido a que varía con la temperatura; sin embargo, para los fines de un experimento, el ingeniero puede controlar la carbonatación en tres niveles: 10, 12 y 14%. Elige dos niveles para la presión (25 y 30psi) y dos niveles para la rapidez de línea (200 y 250bpm).

Interesa analizar qué tanto se desvía la altura real con respecto a la altura establecida como objetivo. Las desviaciones positivas son alturas de llenado arriba del objetivo, mientras que las negativas son alturas de llenado abajo del objetivo. Puesto que pueden existir desviaciones en ambas direcciones, el interés del estudio es analizar si las botellas se han llenado correctamente, por lo que se toma el valor absoluto de cada desviación. Debido a que las botellas se producen en corridas de cerca de 50 botellas, la unidad de observación es una corrida, y la variable respuesta es la desviación absoluta promedio en cada corrida de producción de botellas. Las mediciones se hacen en milímetros.

El ingeniero decide correr dos réplicas de un diseño factorial con estos tres factores, por lo que realiza dos corridas de producción con cada conjunto de condiciones, haciendo las 24 corridas en un orden aleatorio.

4.2.1 Ejercicios

1. Preparación:

- (a) Cargue el archivo `refrescos.Rdata`. Verifique que todos los factores estén bien definidos.
- (b) Obtenga los promedios de todos los tratamientos. Use:
`tapply(y, list(x1, x2, x3), mean)`.
- (c) Observe cuál es el tratamiento que parece acercarse más al objetivo de la producción.

2. Interacciones dobles:

- (a) Haga un gráfico para ver la interacción entre presión y carbonatación. De forma similar, haga gráficos para analizar la interacción entre los otros pares de factores. Se puede esperar interacción entre cada par de factores? En caso afirmativo, ¿por qué sí y qué significa? En caso negativo, ¿por qué no y qué significa?
- (b) Obtenga los promedios cruzando sólo dos factores a la vez. Es decir, los promedios cruzando carbonatación y presión. Trate de calcular del gráfico más o menos cuánto tienen que ajustarse estos promedios para que NO haya interacción entre carbonatación y presión. Luego cruce rapidez y presión, y finalmente cruce rapidez y carbonatación.
- (c) Obtenga los promedios marginales para cada nivel de cada factor (independientemente de los otros factores), es decir, los tres promedios para los niveles de carbonatación, los dos promedios para los niveles de presión y los dos promedios para los niveles de rapidez.
- (d) Obtenga la estimación de los efectos simples de cada factor.
- (e) Obtenga la estimación del efecto de la interacción entre carbonatación y presión para el caso de carbonatación 10% y presión 25 psi. Esta medida se llama interacción de primer orden. Luego obtenga el efecto de la interacción de primer orden entre presión 25 psi y rapidez 200 bpm, y finalmente entre carbonatación 10% y rapidez 200 bpm.

- (f) Verifique los efectos simples y de las interacciones de primer orden obtenidas usando un modelo con interacciones dobles.

3. Interacción triple:

- (a) Para visualizar si existe interacción conjunta entre los tres factores se analiza la interacción entre dos de ellos, pero separando para cada nivel del tercer factor. Use la función `xyplot` en la librería `lattice` de la siguiente forma: `y~x1|x3, group=x2, pch=18, type="a"`.
- (b) Observe si el comportamiento de estas interacciones es el mismo en ambos gráficos. Cuando el comportamiento de la interacción es diferente en cada gráfico se dice que hay interacción triple o interacción de segundo orden.
- (c) Escriba el modelo con interacción triple, indicando a qué corresponde cada subíndice.
- (d) Realice el cálculo para encontrar el efecto de interacción triple para carbonatación 10%, presión 25 psi y rapidez 200 bpm.
- (e) Estime un modelo con interacción triple usando `aov` y extraiga los efectos de interacción triple. Obtenga la suma de cuadrados de interacción triple.
- (f) Obtenga los grados de libertad de la interacción triple.
- (g) Obtenga el cuadrado medio correspondiente y explique qué mide esta cantidad.
- (h) Obtenga el análisis de varianza.
- (i) Compare el CMTri con el CMRes para poner a prueba si la hipótesis nula de no interacción triple es factible. Concluya.

4. Análisis sin interacción triple:

- (a) Asuma que no existe interacción triple y estime el modelo correspondiente bajo la restricción de suma nula. Elimine las interacciones que no son significativas para lo cual debe hacerlo paso a paso. Primero elimine la interacción que es menos significativa. En este proceso de eliminación es más seguro usar la función `drop1` con el argumento `test="F"`, la cual es equivalente a comparar el modelo completo con un modelo que elimina un término a la vez. Es útil usar esta función especialmente cuando el diseño no es balanceado o cuando se incluyen covariables (lo cual se verá más adelante). Estime el nuevo modelo reducido, obtenga los coeficientes y vea a qué corresponde cada uno.
- (b) Obtenga los coeficientes del modelo simplificado y vea a qué corresponde cada uno de ellos. Obtenga las estimaciones de la media marginal para los niveles de carbonatación.
- (c) Compare las estimaciones obtenidas con el modelo para las medias marginales de carbonatación con las medias observadas obtenidas previamente.
- (d) Como existe interacción entre algunos factores, hay que considerar esta interacción al hacer comparaciones; sin embargo, carbonatación no tiene interacción con ningún otro factor, por lo que se pueden hacer las comparaciones de las medias marginales. Por otra parte, puesto que se obtienen los mismos resultados usando las medias observadas que las medias estimadas por el modelo, para hacer las pruebas e intervalos se pueden usar los contrastes largos o la comparación clásica más corta.

Cree los vectores que deben multiplicar los coeficientes del modelo para calcular las medias marginales de los diferentes niveles de carbonatación. Para obtener una media marginal se pone un cero en los vectores para que multipliquen los coeficientes de los otros factores (presión y rapidez). Esto solo es válido si se usa el modelo de suma nula, ya que si se usa el de tratamiento referencia, al poner un cero en algún factor se estaría indicando el nivel de referencia.

- (e) Obtenga estimaciones de los promedios usando estos vectores y compárelos con los obtenidos anteriormente.
- (f) Escriba los contrastes para comparar las medias de los tres niveles de carbonatación y obtenga una estimación de la diferencia de los promedios por pares.
- (g) Escriba las hipótesis asociadas a esos contrastes y verifique si son ortogonales.
- (h) Lleve a cabo las pruebas de las hipótesis para estos contrastes.
- (i) Verifique que todos los errores estándar coinciden con el obtenido a partir de la fórmula clásica para la varianza de una diferencia de promedios: $V(\bar{y}_i - \bar{y}_j) = 2\text{CMRes}/r$.
- (j) Considerando que hay interacción entre presión y rapidez, haga las comparaciones entre las medias de los dos niveles de presión para cada nivel de rapidez. Primero verifique si los contrastes de las hipótesis son ortogonales.
- (k) Ahora haga las comparaciones en sentido inverso, es decir, compare entre las medias de los dos niveles de rapidez para cada nivel de presión. Primero verifique si los contrastes de las hipótesis son ortogonales.

4.2.2 Solución

1. Preparación:

(a) Lectura:

```
load("refrescos.Rdata")
str(base)

## 'data.frame': 24 obs. of 4 variables:
## $ carbonatacion: Factor w/ 3 levels "10","12","14": 1 1 2 2 3 3 ...
## $ presion : Factor w/ 2 levels "25","30": 1 1 1 1 1 1 1 1 1 ...
## $ rapidez : Factor w/ 2 levels "200","250": 1 1 1 1 1 1 2 2 2 ...
## $ desvabs : num [1:24, 1] 1 1 3 1 5 6 1 2 3 3 ...
```

(b) Promedios de todos los tratamientos:

```
(m.cpr=tapply(base$desvabs, list(base$carbonatacion,base$presion,base$rapidez),mean))

## , , 200
##    25   30
## 10 1.0 0.5
## 12 2.0 2.0
## 14 5.5 7.5
##
## , , 250
##    25   30
## 10 1.5 3.5
## 12 3.0 6.0
## 14 7.0 10.0
```

(c) Tratamiento que parece acercarse más al objetivo de la producción:

El tratamiento que produce la menor desviación (0,5 mm) es con 10% de carbonatación, presión de 30 psi y rapidez de 200 bpm.

2. Interacciones dobles:

(a) Interacción entre pares de variables:

```
library(ggplot2)
# Carbonatación en eje x, y presión
ggplot(base, aes(x = carbonatacion, y = desvabs, group = presion)) +
  stat_summary(fun.y = "mean", geom = "line", aes(linetype = presion))

# Carbonatación en eje x, y rapidez
ggplot(base, aes(x=carbonatacion, y=desvabs, group = rapidez)) +
  stat_summary(fun.y="mean", geom="line", aes(linetype = rapidez))

# Presión en eje x, y rapidez
ggplot(base, aes(x=presion, y=desvabs, group = rapidez)) +
  stat_summary(fun.y="mean", geom="line", aes(linetype = rapidez))
```

Parece no haber interacción entre carbonatación y presión, ya que para todos los niveles de carbonatación se aprecia un efecto de la presión, se nota un aumento de la media de la desviación cuando se usa mayor presión (Figura 4.2 superior izquierda). De forma similar, parece que no hay interacción entre carbonatación y rapidez, se observa una mayor media de la desviación para mayor rapidez, en todos los niveles de carbonatación (Figura 4.2 superior derecha). Entre presión y rapidez puede sospecharse de presencia de interacción, cuando la presión es baja el efecto de la rapidez es pequeño y cuando la presión es alta ese efecto se hace un poco más fuerte, siempre tendiendo a que a mayor rapidez hay mayor desviación promedio en el llenado (Figura 4.2 inferior).

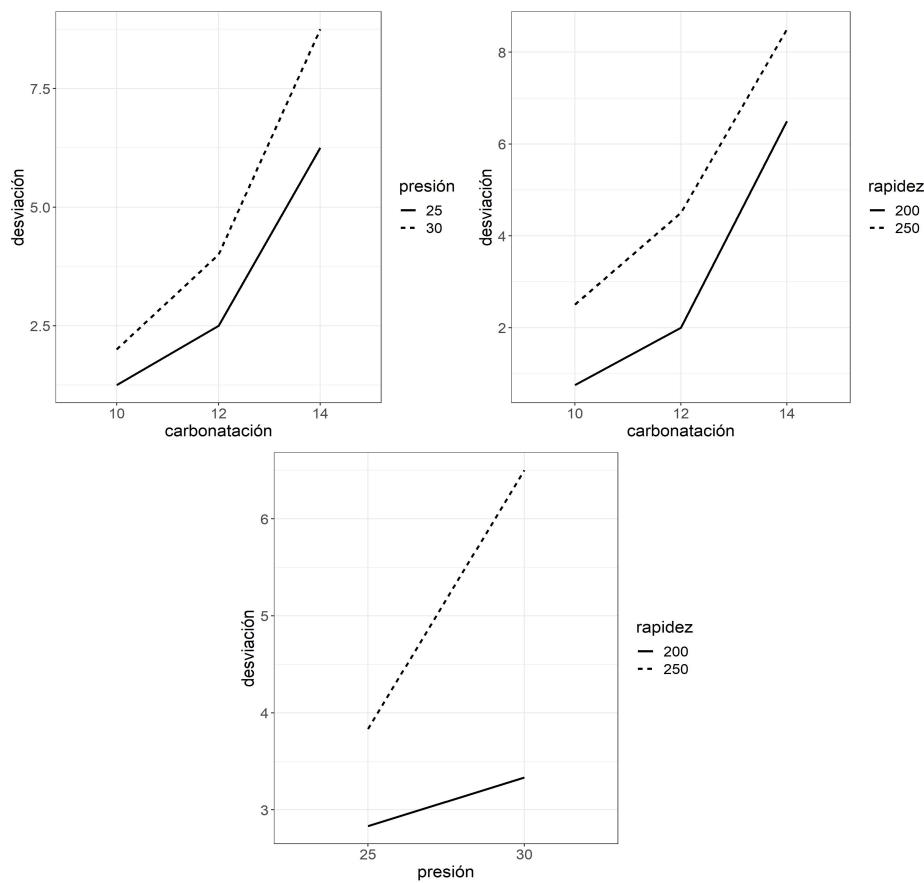


Figura 4.2: Promedios de desviación absoluta para combinaciones de los factores

Nota: a) carbonatación y presión (superior izquierda), b) carbonatación y rapidez (superior derecha) y c) presión y rapidez (inferior).

(b) Promedios cruzando solo dos factores a la vez:

```
(m.cp = tapply(base$desvabs, list(base$carbonacion,base$presion),mean))
```

```
##      25   30
## 10 1.25 2.00
## 12 2.50 4.00
## 14 6.25 8.75
```

```
(m.cr = tapply(base$desvabs, list(base$carbonacion,base$rapidez),mean))
```

```
##      200   250
## 10 0.75 2.50
## 12 2.00 4.50
## 14 6.50 8.50
```

```
(m.pr = tapply(base$desvabs, list(base$presion,base$rapidez),mean))

##      200   250
## 25 2.83 3.83
## 30 3.33 6.50
```

(c) Promedios marginales:

```
(m.c = tapply(base$desvabs,base$carbonatacion,mean))

##    10    12    14
## 1.62 3.25 7.50

(m.p = tapply(base$desvabs,base$presion,mean))

##    25    30
## 3.33 4.92

(m.r = tapply(base$desvabs,base$rapidez,mean))

##    200   250
## 3.08 5.17
```

(d) Efectos simples:

```
m = mean(base$desvabs)
(ef.c= m.c-m)

##    10    12    14
## -2.50 -0.88 3.38

(ef.p= m.p-m)

##    25    30
## -0.79  0.79

(ef.r= m.r-m)

##    200   250
## -1.04 1.04
```

(e) Efectos de las interacciones entre pares de factores:

```
int.cp = m.cp[1,1]-(m+ef.c[1]+ef.p[1])
int.pr = m.pr[1,1]-(m+ef.p[1]+ef.r[1])
int.cr = m.cr[1,1]-(m+ef.c[1]+ef.r[1])
c(int.cp,int.pr,int.cr)

## 0.42 0.54 0.17
```

(f) Efectos usando un modelo con interacciones dobles:

```
mod1=aov(desvabs~carbonatacion*presion+carbonatacion*rapidez+presion*rapidez,data=base)
model.tables(mod1)

##      carbonatacion
##      10     12     14
## -2.50 -0.88  3.38
##
##      presion
##      25     30
## -0.79  0.79
##
##      rapidez
##      200    250
## -1.04  1.04
##
##                  presion
## carbonatacion   25     30
##                 10  0.42 -0.42
##                 12  0.04 -0.04
##                 14 -0.46  0.46
##
##                  rapidez
## carbonatacion   200    250
##                 10  0.17 -0.17
##                 12 -0.21  0.21
##                 14  0.04 -0.04
##
##                  rapidez
## presion        200    250
##                 25  0.54 -0.54
##                 30 -0.54  0.54
```

3. Interacción triple:

(a) Visualización de interacción triple:

```
library(lattice)
xyplot(desvabs~carbonatacion|rapidez,group=presion,pch=18,type=c("a"),
       xlab=list(label="carbonatación",cex.axis=1.5, cex=1.5),
       ylab=list(label="desviación",cex.axis=1.5, cex=1.5),
       key = list(columns=2,text=list(levels(base$presion)),title="presión",
                  lines=list(col=1,lty=c(1,2))),col=1,lty=c(1,2),data=base)
```

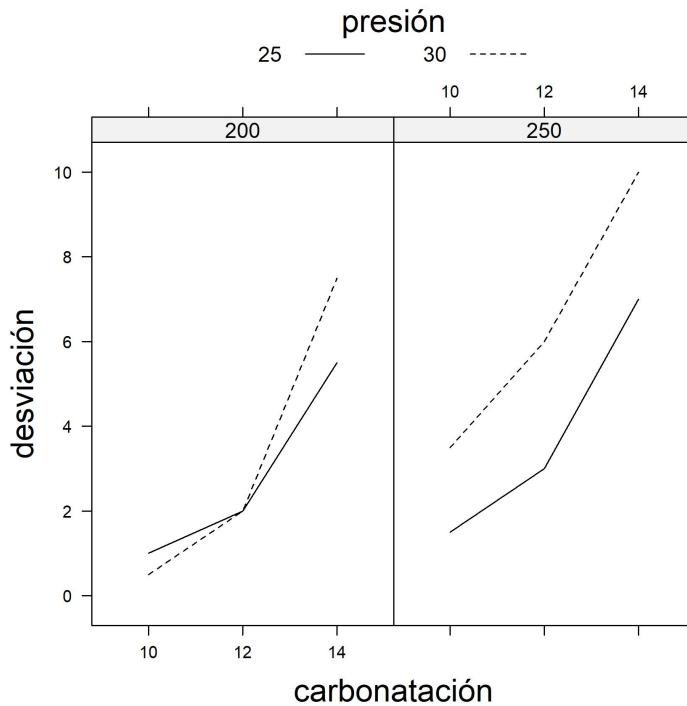


Figura 4.3: Promedios de desviación absoluta para combinaciones de carbonatación y presión según niveles de rapidez

Nota: 200 bpi (izquierda) y 250 bpi (derecha).

(b) Comparación de interacciones dobles:

En la Figura 4.3 se observa que la interacción entre carbonatación y presión es similar en ambas partes, puesto que, tanto en la parte que corresponde a rapidez 200 como en la que corresponde a rapidez 250, parece no haber interacción entre esos factores. Por lo tanto, no parece haber evidencia de interacción triple.

(c) Modelo:

$$\mu_{ijk}^{CIT} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}.$$

El subíndice i indica el nivel de carbonatación (i=1 para 10%, i=2 para 12%, i=3 para 14%), j indica el nivel de presión (j=1 para 25psi, j=2 para 30psi) y k indica el nivel de rapidez (k=1 para 200bpi, k=2 para 250bpi).

(d) Cálculo de los efectos de interacción triple:

```
(int.cpr=m.cpr[1,1,1]-(m+ef.c[1]+ef.p[1]+ef.r[1]+int.cp+int.cr+int.pr))
```

```
## 0.083
```

(e) Suma de cuadrados de interacción triple:

```
mod2=aov(desvabs~carbonatacion*presion*rapidez,data=base)
(int.triple=model.tables(mod2)$tables$"carbonatacion:presion:rapidez")
```

```
## , , rapidez = 200
##
##             presion
## carbonatacion 25      30
##                 10  0.083 -0.083
##                 12  0.208 -0.208
##                 14 -0.292  0.292
##
## , , rapidez = 250
##
##             presion
## carbonatacion 25      30
##                 10 -0.083  0.083
##                 12 -0.208  0.208
##                 14  0.292 -0.292
```

```
(sctrIPLE=sum(2*int.triple^2))
```

```
## 1.083
```

(f) Grados de libertad:

```
(g1=(3-1)*(2-1)*(2-1))
```

```
## 2
```

(g) Cuadrado medio:

```
(cmtriple=sctrIPLE/g1)
```

```
## 0.54
```

Este es el cuadrado medio de interacción triple el cual mide la magnitud de los efectos de interacción triple. Si esta cantidad resulta ser muy grande en relación al CMRes, significa que hay una alta variabilidad en los efectos de interacción triple. Como estos efectos tienen media cero, entonces tener alta variabilidad es equivalente a tener efectos relativamente altos en valor absoluto.

(h) Análisis de varianza:

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: desvabs
##                               Df  Sum Sq Mean Sq F value    Pr(>F)
## carbonatacion            2 147.25  73.63  135.92 5.71e-09 ***
## presion                   1   15.04   15.04   27.77   0.000 ***
## rapidez                   1   26.04   26.04   48.08  1.57e-05 ***
## carbonatacion:presion    2   3.08   1.54    2.85   0.097 .
## carbonatacion:rapidez    2   0.58   0.29    0.54   0.597
## presion:rapidez          1   7.04   7.04   13.00   0.004 **
## carbonat:presion:rapidez 2   1.08   0.54    1.00   0.397
## Residuals                 12   6.50   0.54
```

(i) Prueba de la hipótesis de no interacción triple:

```
cmres=anova(mod2)[8,3]
(f=cmtriple/cmres)
```

```
## 1
```

El valor F da 1, con lo que se pone de manifiesto lo pequeño que es el CMTriple. Asociado a esto se tiene un probabilidad muy alta ($p = 0.40$), con lo cual no se recomienda rechazar la hipótesis nula y, por lo tanto, se asume que no existe interacción conjunta entre los tres factores. Al no haber interacción triple se tiene que la interacción que existe entre carbonatación y presión no depende del nivel de rapidez que se tome. De la misma forma sucede con las otras interacciones.

4. Análisis sin interacción triple:

(a) Prueba de hipótesis de independencia entre pares de factores:

Se puede continuar el análisis con el `mod1` que no tenía interacción triple; sin embargo, se va a estimar de nuevo usando el modelo de suma nula.

```
options(contrasts=c("contr.sum","contr.poly"))
mod3=lm(desvabs~carbonatacion*presion+presion*rapidez+carbonatacion*rapidez,data=base)
drop1(mod3,test="F")

## Single term deletions
##
## Model:
## desvabs ~ carbonatacion * presion + presion * rapidez + carbonatacion *
##      rapidez
##                   Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                      7.58  -7.65
## carbonatacion:presion  2     3.08  10.67  -3.46    2.85  0.092 .
## presion:rapidez       1     7.04  14.63   6.11   13.00  0.003 **
## carbonatacion:rapidez 2     0.58  8.17  -9.87   0.54  0.595
```

La probabilidad asociada a la interacción entre carbonatación y rapidez es mayor a 0,05 y esta probabilidad es mayor que la de la interacción entre carbonatación y presión, que también es mayor a 0,05, por lo que se elimina esta interacción y se corre un nuevo modelo sin ella.

```
mod4=lm(desvabs~carbonatacion*presion+presion*rapidez,data=base)
drop1(mod4,test="F")

## Model:
## desvabs ~ carbonatacion * presion + presion * rapidez
##                   Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                      8.17  -9.87
## carbonatacion:presion  2     3.08  11.25  -6.18    3.02  0.077 .
## presion:rapidez       1     7.04  15.21   3.05   13.80  0.002 **
```

La probabilidad asociada a la interacción entre carbonatación y presión sigue siendo mayor a 0,05, por lo que se elimina esta interacción y se corre un nuevo modelo sin ella y con solo la interacción entre presión y rapidez. Note que esta probabilidad es un poco diferente a la que se obtuvo en el modelo anterior (`mod3`), ya que era 0.092 y ahora es 0.077. Esto se debe a que ahora el cuadrado medio residual ha cambiado un poco y la prueba F se construye al dividir el cuadrado medio de la interacción contra un nuevo cuadrado medio residual, y cambian también los grados de libertad residuales. La diferencia no debería ser muy grande, pero en algunos casos estos cambios provocan sorpresas, y un término que parecía significativo, podría resultar no serlo después de eliminar otros términos. Esta es la importancia de hacer la eliminación de términos de forma secuencial.

```

mod5=lm(desvabs~carbonatacion+presion*rapidez,data=base)
drop1(mod5,test="F")

## Model:
## desvabs ~ carbonatacion + presion * rapidez
##                               Df  Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>                      11.25   -6.18
## carbonatacion              2     147.25  158.50  53.30  117.80  4.57e-11 ***
## presion:rapidez            1      7.04   18.29   3.48   11.27   0.004 **

```

Este modelo solo tiene interacción entre presión y rapidez, cuya probabilidad asociada es menor a 0,05, por lo que, si se elimina esa interacción, la suma de cuadrados residual aumentaría de forma significativa. En otras palabras, no todas las interacciones entre pares de factores son cero y se debe tomar en cuenta esa interacción.

(b) Modelo simplificado:

```

contrasts(base$carbonatacion)

## 10 1 0
## 12 0 1
## 14 -1 -1

contrasts(base$presion)

## 25 1
## 30 -1

contrasts(base$rapidez)

## 200 1
## 250 -1

(b=mod5$coef)

## (Intercept) carbonatacion1 carbonatacion2 presion1 rapidez1 presion1:rapidez1
##          4.13           -2.50            -0.88           -0.79           -1.04            0.54

```

La media general es 4.13. El término carbonatacion1 se refiere al nivel 10 y su coeficiente es -2.50, lo que significa que la media marginal de carbonatación 10 está 2.50 unidades debajo de la media general, entonces la media marginal de carbonatación 10 es $4.13-2.50=1.63$. Por otro lado, carbonatacion2 corresponde al nivel 12, lo que indica que la media marginal de carbonatación 12 está 0.88 unidades debajo de la media general ($4.13-0.88=3.25$). El coeficiente para el nivel 14 debe obtenerse con el negativo de la suma de los otros 2 ($2.50+0.88=3.38$), entonces con carbonatación 14 la media marginal está 3.38 unidades sobre la media general ($4.13+3.38=7.51$). Similarmente, presion1 corresponde al nivel 25 y su coeficiente es -0.79, la media marginal es $4.13-0.79=3.34$, mientras que la media marginal para una presión de 30 es $4.13+0.79=4.92$. Finalmente, rapidez1 corresponde al nivel 200 y su media marginal es $4.13-1.04=3.09$, mientras que la media marginal para la rapidez de 250 es $4.13+1.04=5.17$.

(c) Comparación de estimaciones con las medias observadas:

```
m.c
##   10    12    14
## 1.63 3.25 7.50
```

Se obtienen los mismos resultados.

(d) Contrastes:

Los vectores de coeficientes para las medias de los niveles de carbonatación son los siguientes:

Carbonatación 10: $[1, 1, 0, 0, 0, 0]^T$
 Carbonatación 12: $[1, 0, 1, 0, 0, 0]^T$
 Carbonatación 14: $[1, -1, -1, 0, 0, 0]^T$

(e) Estimaciones de las medias usando los contrastes:

```
c10=c(1, 1, 0, 0, 0, 0)
c12=c(1, 0, 1, 0, 0, 0)
c14=c(1, -1, -1, 0, 0, 0)
h=cbind(c10,c12,c14)
(M.c = t(h) %*% b)

## c10 1.63
## c12 3.25
## c14 7.50
```

Se obtienen los mismos resultados.

(f) Comparación de promedios de carbonatación:

```
c12.10=c12-c10
c14.10=c14-c10
c14.12=c14-c12
h=cbind(c12.10,c14.10,c14.12)
(L=t(h) %*% b)
```

```
## c12.10 1.63
## c14.10 5.88
## c14.12 4.25
```

(g) Hipótesis y ortogonalidad:

Los contrastes para comparar las medias marginales de carbonatación son:

$$\begin{aligned}\mu_{2\bullet\bullet} - \mu_{1\bullet\bullet} \\ \mu_{3\bullet\bullet} - \mu_{1\bullet\bullet} \\ \mu_{3\bullet\bullet} - \mu_{2\bullet\bullet}\end{aligned}$$

Tomando el vector de promedios $(\mu_{1\bullet\bullet}, \mu_{2\bullet\bullet}, \mu_{3\bullet\bullet})$, los vectores para obtener los contrastes son:

$$\begin{aligned}(-1, 1, 0) \\ (-1, 0, 1) \\ (0, 1, -1)\end{aligned}$$

```
v1 = c(-1,1, 0)
v2 = c(-1,0, 1)
v3 = c( 0,1,-1)
c(v1%*%v2, v1%*%v3, v2%*%v3)
```

```
## 1 1 -1
```

Estos 3 vectores no son ortogonales, por lo que hay que hacer corrección.

(h) Pruebas de hipótesis de los contrastes:

Puesto que se realizan 3 contrastes no ortogonales y se comparan todos los promedios marginales entre sí por pares, debe utilizarse Tukey.

```
ee=sqrt(diag(t(h) %*% vcov(mod5) %*% h))
t=L/ee
(p=ptukey(t*sqrt(2),3,18,lower.tail = F))
```

```
## c12.10 0.0018
## c14.10 0.0000
## c14.12 0.0000
```

Las probabilidades obtenidas deben compararse contra $\alpha = 0,05$. Se encontraron diferencias entre todos los pares de promedios. Se ha observado que el nivel de carbonatación 10% presenta la media muestral más baja, por lo que se concluye que este nivel es el que baja más la desviación promedio de llenado y se considera que ese nivel es el más apropiado, independientemente del nivel que se escoja para los otros factores. Aún debe cuantificarse la diferencia esperada entre el promedio al nivel 10% y los otros dos niveles, con el fin de valorar si esta diferencia se puede considerar relevante al compararse con una diferencia previamente definida por el investigador.

(i) Verificación de los errores estándar.

```
table(base$carbonatacion)

## 10 12 14
## 8 8 8

r=8
CMRes=anova(mod5)[5,3]
sqrt(2*CMRes/r)

## 0.395

ee

## c12.10 c14.10 c14.12
## 0.395 0.395 0.395
```

(j) Comparación de promedios de presión para cada nivel de rapidez:

Los contrastes para comparar las medias de presión para cada nivel de rapidez son:

$$\begin{aligned}\mu_{\bullet 21} - \mu_{\bullet 11} \\ \mu_{\bullet 22} - \mu_{\bullet 12}\end{aligned}$$

Tomando el vector de promedios $(\mu_{\bullet 11}, \mu_{\bullet 21}, \mu_{\bullet 12}, \mu_{\bullet 22})$, los vectores para obtener los contrastes son: $(-1, 1, 0, 0)$ y $(0, 0, -1, 1)$.

```
v1 = c(-1,1, 0,0)
v2 = c( 0,0,-1,1)
v1%*%v2

## 0
```

Estos 2 vectores son ortogonales, por lo que no hay que hacer corrección.

```

p25.r200=c(1,0,0,1,1,1)
p30.r200=c(1,0,0,-1,1,-1)
p25.r250=c(1,0,0,1,-1,-1)
p30.r250=c(1,0,0,-1,-1,1)
p30.25.r200=p30.r200-p25.r200
p30.25.r250=p30.r250-p25.r250
h=cbind(p30.25.r200,p30.25.r250)

(L=t(h) %*% b)

## p30.25.r200 0.50
## p30.25.r250 2.67

ee=sqrt(diag(t(h) %*% vcov(mod5) %*% h))
t=L/ee
(p=pt(t,18,lower.tail = F))

## p30.25.r200 0.144
## p30.25.r250 0.000

```

Puesto que se realizan 2 contrastes ortogonales, las probabilidades obtenidas deben compararse contra $\alpha = 0,05$. Se encontraron diferencias entre los promedios de los dos niveles de presión solo cuando la rapidez es 250 bpm. Se ha observado que el nivel de presión 25 psi presenta la media muestral más baja cuando se tiene una rapidez de 250 bpm, por lo que si se escoge ese nivel de rapidez, vale la pena también escoger el nivel de presión de 25 psi, pero si se escoge el nivel de 200 bpm, no importa cuál de los dos niveles de presión se escoga.

(k) Comparación de promedios de rapidez para cada nivel de presión:

Los contrastes para comparar las medias de rapidez para cada nivel de presión son:

$$\begin{aligned}\mu_{\bullet 12} - \mu_{\bullet 11} \\ \mu_{\bullet 22} - \mu_{\bullet 21}\end{aligned}$$

Tomando el vector de promedios $(\mu_{\bullet 11}, \mu_{\bullet 21}, \mu_{\bullet 12}, \mu_{\bullet 22})$, los vectores para obtener los contrastes son: $(-1, 0, 1, 0)$ y $(0, -1, 0, 1)$.

```

v1 = c(-1, 0, 1, 0)
v2 = c( 0, -1, 0, 1)
v1 %*% v2

## 0

```

Estos 2 vectores son ortogonales, por lo que no hay que hacer corrección.

```
r250.200.p25=p25.r250-p25.r200  
r250.200.p30=p30.r250-p30.r200  
h=cbind(r250.200.p25,r250.200.p30)  
(L=t(h) %*% b)
```

```
## r250.200.p25 1.00  
## r250.200.p30 3.17
```

```
ee=sqrt(diag(t(h) %*% vcov(mod4) %*% h))  
t=L/ee  
(p=pt(t,18,lower.tail = F))
```

```
## r250.200.p25 0.021  
## r250.200.p30 0.000
```

Se encontraron diferencias entre los promedios de los dos niveles de rapidez, tanto para presión de 25 psi como 30 psi. Se ha observado que el nivel de rapidez 200 bpm presenta la media muestral más baja en ambas presiones, por lo que se prefiere este nivel de rapidez. Anteriormente se había dicho que si se escogía 200 bpm no importaba mucho el nivel de presión, por lo que al final, la recomendación es usar 10% de carbonatación, 200 bpm y cualquiera de los niveles de presión.

Capítulo 5

Diseños con bloques

5.1 Conceptos

Cuando se realiza un experimento, se busca comparar los promedios de diferentes tratamientos, con el objetivo de detectar diferencias entre esos promedios y determinar si de verdad esas diferencias existen. Este objetivo se ve obstaculizado cuando existe mucha variabilidad en el error experimental. Para reducir esta variabilidad, en algunas situaciones se puede recurrir a una técnica denominada bloqueo. Se trata de construir conjuntos de unidades experimentales relativamente homogéneas y aplicar todos los tratamientos del experimento dentro de ese conjunto llamado bloque.

Un bloque puede estar constituido por una unidad grande que se subdivide, tal como una parcela agrícola donde a cada segmento de la parcela se le asigna aleatoriamente uno de los tratamientos. En tal caso las unidades experimentales son las subdivisiones de la parcela y todas juntas forman un bloque. También un bloque puede consistir en un conjunto de unidades que se agrupan de forma natural, tal como un lote de producción en una fábrica, ya que las unidades dentro del lote se supone que son más parecidas entre sí, pero hay mayores diferencias entre las unidades de un lote a otro.

También se puede presentar un bloque cuando a un individuo (por ejemplo, persona o animal) se le aplican los diferentes tratamientos en un orden aleatorio, el

individuo se convierte en un bloque donde cada aplicación del tratamiento produce una observación. Se esperaría que las mediciones para un mismo individuo estén correlacionadas y lo que las hace diferentes sea la aplicación de los tratamientos, aunque los resultados que arroje un individuo pueden tender a ser más altos o más bajos que los de otros individuos. El diseño de bloques resulta eficiente cuando los bloques tienen características que hacen que la variable respuesta obtenga valores diferentes de un bloque a otro, lo cual se puede traducir en que existe una alta variabilidad de bloque a bloque.

Otro caso puede darse cuando la experimentación requiere que las observaciones sean recolectadas a lo largo de un período de tiempo donde las condiciones van cambiando. En tal caso se puede definir un bloque como una unidad de tiempo (por ejemplo, hora o día) que sea relativamente homogénea, y dentro de la cual sea posible aplicar todos los tratamientos.

El diseño supone que cada componente del bloque (segmento de parcela, unidades dentro de un lote, persona, día, etc.) sea asignado de forma aleatoria a los tratamientos y que cada bloque contenga todos los tratamientos del experimento; sin embargo, existen métodos para abordar situaciones en las que, de forma planeada, los bloques no contienen todos los tratamientos, lo cual lleva al diseño y análisis de bloques incompletos.

Modelo con bloques

El modelo de un factor con bloques se expresa de forma muy similar a un modelo con dos factores sin interacción:

$$\mu_{jk} = \mu + \tau_j + \gamma_k,$$

donde el término τ_j representa el efecto del j -ésimo nivel del factor de diseño y γ_k representa el efecto del k -ésimo bloque. El modelo de bloques es aditivo, por lo que no se espera que haya interacción entre el factor y el bloque. No interesa analizar qué tan diferente es la respuesta entre un bloque y otro, es decir, no es importante llegar a conclusiones específicas sobre cuáles bloques tienen respuestas mayores o menores. Además, el interés del estudio es analizar el efecto del factor de diseño sobre la respuesta promedio, por lo tanto, lo que interesa es la estimación de la media

marginal para cada nivel del factor, como si no existieran los bloques.

En la Figura 5.1 se ilustra lo que sucede cuando se toman datos en bloques. En la parte izquierda se observa que algunos bloques tienen valores de la respuesta más altos (bloques 1, 3 y 5), mientras que otros tienen valores más bajos (bloques 2, 8 y 13). En el gráfico del lado derecho se han colocado los datos en los 4 tratamientos. Aquí se observa una enorme variabilidad debida a las diferencias entre bloques, la cual hace que sea más difícil apreciar las diferencias entre los promedios de los tratamientos.

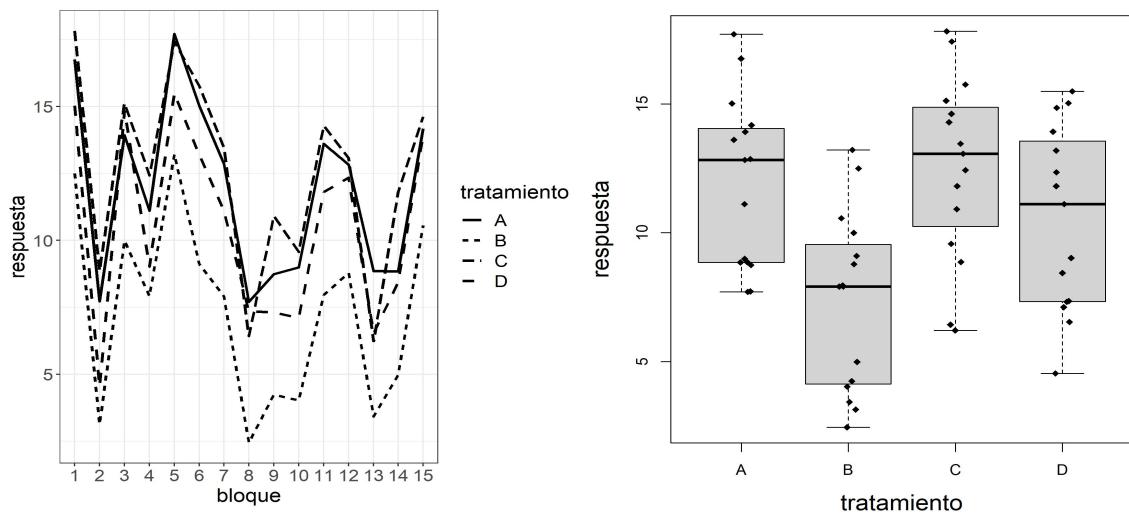


Figura 5.1: Diseño de un factor con bloques

Nota: en la parte izquierda se muestran los datos separados por bloque, según tratamiento, mientras que en el lado derecho los datos se agrupan en cada tratamiento sin considerar el bloque.

En la Figura 5.2 se han centrado los datos, eliminando el efecto que tiene cada bloque, al restar a cada observación la media del bloque y reubicando todos los datos alrededor de la media general original. Los datos de cada bloque tienen como media la misma media general, lo que hace que no se observen bloques con observaciones que tiendan hacia arriba o hacia abajo. En el lado izquierdo ahora se ve más claramente que la línea correspondiente al tratamiento B se ubica por debajo de las demás. Esto concuerda con el gráfico del lado derecho, donde ahora los puntos de cada tratamiento son más homogéneos y es más clara la ubicación de la distribución del tratamiento B por debajo de las demás.

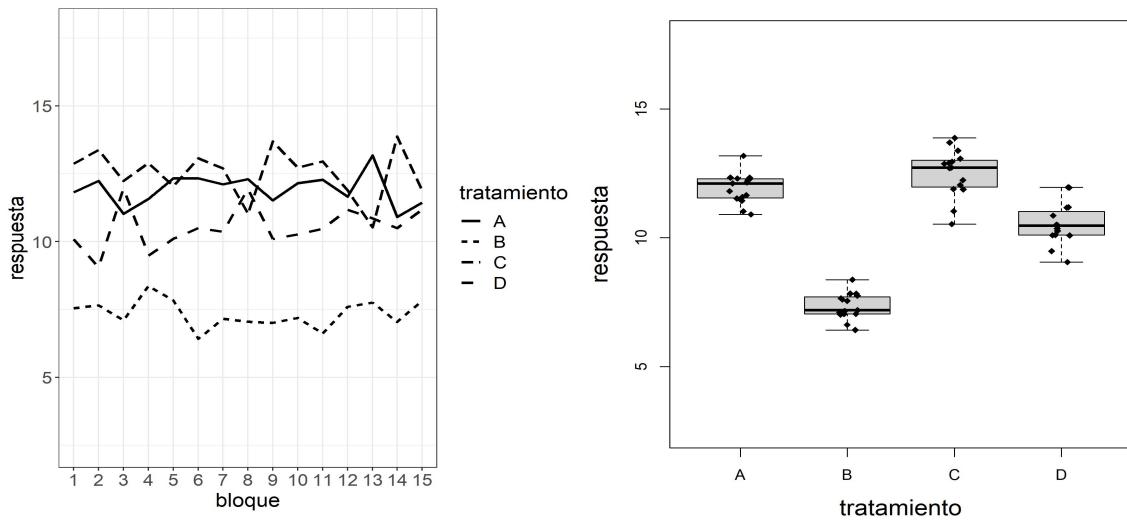


Figura 5.2: Datos centrados para eliminar el efecto del bloque

Nota: en la parte izquierda se muestran los datos centrados por bloque, separados por bloque, según tratamiento, mientras que en el lado derecho se presentan los datos centrados por bloque, agrupados para cada tratamiento.

Análisis formal

El objetivo del diseño con bloques sigue siendo la comparación de las medias del factor de diseño, pero se intenta reducir la variabilidad residual mediante el arreglo de las unidades en estos conjuntos llamados bloques. Esto se traduce matemáticamente en que la suma de cuadrados residual del diseño con un factor sin bloques (ResSinBloque) se ve reducida en una cantidad que introducen los bloques y corresponde a la variabilidad de bloque a bloque. Esta variabilidad se mide con la suma de cuadrados de bloque (SCBloque), la cual se calcula de la siguiente forma:

$$\text{SCBloque} = \sum_{k=1}^b a\hat{\gamma}_k^2 = \sum_{k=1}^b a(\bar{y}_{.k} - \bar{y}_{..})^2,$$

donde a es el número de tratamientos, b es el número de bloques, $\bar{y}_{.k}$ es el promedio del k -ésimo bloque y $\bar{y}_{..}$ es la media general.

De esta forma la suma de cuadrados residual del diseño de bloques (SCResB) se obtiene con:

$$\text{SCResConBloque} = \text{SCResSinBloque} - \text{SCBloque}.$$

En este caso la SCTot se descompone en tres fuentes de variación:

$$\text{SCTot} = \text{SCTrat} + \text{SCBloque} + \text{SCResConBloque}.$$

Si se conoce la SCTot, la SCTrat y la SCBloque, se puede encontrar la SCResConBloque.

A partir de aquí se obtiene el cuadrado medio residual del diseño de bloques, tomando en cuenta que se tienen $b - 1$ grados de libertad para los bloques. La prueba formal de la hipótesis para evaluar el efecto del factor de diseño se realiza construyendo un estadístico F con la razón entre el cuadrado medio de tratamiento y el cuadrado medio residual del diseño de bloques. En este caso se usa la distribución F con $k - 1$ grados de libertad en el numerador y $(n - k) - (b - 1)$ en el denominador. Los grados de libertad del denominador son los correspondientes a los residuales, los cuales se obtienen de los grados de libertad que tendrían los residuales en un diseño sin bloques $(n - k)$ restándoles los grados de libertad de los bloques $(b - 1)$.

Caso de un factor con dos niveles

Si el factor de diseño tiene solo dos niveles, se puede llevar a cabo el análisis de una forma alternativa, similar a la que se presentó en el Capítulo 1. En este caso se usa el enfoque de muestras pareadas, ya que para cada bloque se tiene un par de observaciones. El procedimiento consiste en obtener las diferencias del par de observaciones para cada bloque, luego se obtiene el estadístico t observado, y se busca la probabilidad asociada en la distribución t con los mismos grados de libertad residuales explicados anteriormente $(n - k) - (b - 1)$, pero tomando en cuenta que en este caso $k = 2$ y $n = 2b$, por lo que los grados de libertad se reducen a $b - 1$.

Enfoque alternativo como modelo mixto

Los diseños de bloques se pueden analizar con modelos mixtos, en los cuales se considera que los bloques son extraídos aleatoriamente de una población grande de bloques, es decir, siempre se podría contar con más bloques. De esta forma, los efectos de los bloques se consideran como muestras de una distribución con media cero. La forma más sencilla de estos modelos asume que la distribución de los efectos de los bloques es normal.

En el caso en que se tiene solo un factor y, además, hay una sola repetición por nivel del factor en cada bloque, el resultado obtenido con el modelo lineal clásico coincide con el del modelo mixto. En cambio, cuando se tienen más repeticiones por nivel en uno o más bloques, se tienen observaciones correlacionadas, lo cual rompe con uno de los supuestos básicos del modelo lineal clásico que consiste en que las observaciones dentro de un mismo tratamiento deben ser independientes. En esos casos es indispensable usar el modelo mixto.

Caso de dos factores y parcelas divididas

En diseños más complejos se puede incluir más de un factor dentro de los bloques. Por ejemplo, si se tienen dos factores A y B, donde A tiene 2 niveles y B tiene 3 niveles, se requiere dividir el bloque en 6 partes para que todos los posibles tratamientos estén presentes en todos los bloques. Se hace una asignación aleatoria de estos tratamientos dentro de cada bloque y luego se procede de forma similar con el análisis de varianza. En este caso la SCTot se descompone en 4 fuentes de variación:

$$\text{SCTot} = \text{SCA} + \text{SCB} + \text{SCBloque} + \text{SCResConBloque},$$

donde SCA y SCB se obtienen de la misma forma que en un diseño con dos factores.

El modelo para un diseño de bloques con dos factores asignados aleatoriamente dentro de cada bloque, incluyendo interacción entre los dos factores, se escribe de la siguiente forma:

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k,$$

donde α_i es el efecto del i-ésimo nivel del factor A, β_j es el efecto del j-ésimo nivel del factor B, $(\alpha\beta)_{ij}$ es el efecto de interacción para el tratamiento que combina el nivel i-ésimo de A con el nivel j-ésimo de B, y γ_k es el efecto del k-ésimo bloque.

En algunos casos, la estructura del diseño no permite que cada bloque contenga todos los tratamientos. Un caso particular es cuando el bloque se divide solamente según uno de los factores; por ejemplo, el bloque se divide según los niveles del factor B, entonces para cada bloque se cuenta con una asignación aleatoria de los 3 niveles de B, pero además, cada bloque se asigna de forma completa a uno de los niveles de A. Este tipo de diseños se conoce como parcelas divididas porque tienen su origen en estudios agronómicos donde se cuenta con parcelas de cultivo divididas en

varias subparcelas; a cada subparcela se le asigna uno de los niveles del factor B, que podría ser, por ejemplo, un tipo de fertilizante. Cada parcela corresponde al bloque y tiene asignada, por ejemplo, una variedad de cultivo diferente; se cuenta con varias parcelas de una misma variedad. Los dos factores del diseño se asignan por separado, el factor B se asigna de forma independiente dentro de la parcela a cada subparcela, mientras que el factor A se asigna a las parcelas o bloques directamente; por esto el factor A se llama factor de parcela y el B se llama factor de subparcela. En la Figura 5.3 se presenta un diseño que contempla los factores mencionados anteriormente con dos parcelas completas para cada variedad.

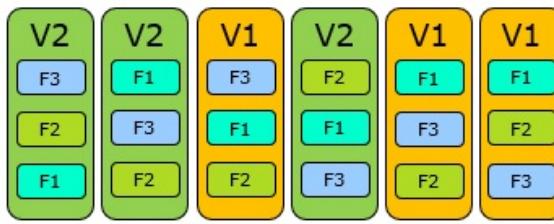


Figura 5.3: Esquema de un diseño de parcelas divididas

Nota: variedad en las parcelas (V1, V2) y fertilizante en las subparcelas (F1, F2, F3).

El modelo para un diseño de parcelas divididas se puede escribir de la misma forma que el modelo para un diseño de bloques con dos factores asignados aleatoriamente dentro de cada bloque; sin embargo, para este último el efecto de los bloques se considera fijo y se debe tener una restricción ya sea de suma nula o bloque de referencia. En cambio, para el diseño de parcelas divididas, se asume que los efectos de bloque (γ_k) siguen una distribución normal con media cero y una varianza de bloques. Esta condición se denota como:

$$\gamma_k \sim \mathcal{N}(0, \sigma_\gamma^2).$$

El análisis adecuado para un diseño de parcelas divididas debe tomar en cuenta la estructura establecida en el modelo anterior, donde se considera que los efectos de los bloques son aleatorios, por lo que se debe de usar el enfoque de modelos mixtos. En este texto no se profundiza en todos los detalles que conlleva la aplicación de un modelo mixto, ya que esto se deja para un texto especializado en ese tema. El análisis de varianza deja de ser la herramienta utilizada, por lo que no hay una descomposición de la SCTot, sino una estimación de los parámetros de varios

modelos y el cálculo de la verosimilitud para cada modelo. Por ejemplo, si se quiere analizar si existe interacción entre A y B, se estiman dos modelos, uno con interacción (modelo grande) y otro sin interacción (modelo pequeño); se obtiene el valor del logaritmo de la verosimilitud para cada modelo y con ellos se construye un estadístico que sigue una distribución χ^2 (chi-cuadrado). Esta prueba se conoce como prueba de razón de verosimilitud (LRT, por sus siglas en inglés). Los grados de libertad de esta distribución corresponden a los grados de libertad del término que se está eliminando. El estadístico se obtiene mediante:

$$\chi^2 = -2 \left(\log \frac{L_1}{L_2} \right) = -2(\ell_1 - \ell_2),$$

donde L_k es el valor de la verosimilitud del k-ésimo modelo. Se tienen 2 modelos, el modelo 1 es el pequeño y el modelo 2 es el grande. Además, $\ell_k = \log(L_k)$, es decir, el logaritmo de la verosimilitud del k-ésimo modelo.

En caso de encontrar que alguno de los factores tiene un efecto sobre la respuesta promedio, se pueden realizar comparaciones múltiples, para lo cual es importante prestar atención al cálculo de los grados de libertad que se deben usar. Los grados de libertad para el error de parcela (*gl.parcela*) se obtienen sabiendo que para cada nivel del factor de parcela se tiene un número de repeticiones r . Entonces se tienen $a(r - 1)$ grados de libertad de parcela, donde a es el número de niveles del factor de parcela. El error de subparcela se calcula con $n - p - gl.parcela$, donde p es el número de coeficientes del modelo.

Bloques incompletos

El diseño con bloques requiere inicialmente que haya al menos una observación asignada a cada tratamiento en cada bloque; sin embargo, en ocasiones esto no es posible por razones prácticas. Por ejemplo, si se tienen 7 tratamientos y se ha definido una unidad de tiempo (día) como bloque, pero en un día solo se pueden realizar 3 mediciones, se planea la recolección como un diseño incompleto. En el ejemplo, se asignan 3 tratamientos a cada día (bloque); existen 35 formas diferentes de asignar 3 tratamientos de los 7 que se tienen para conformar los bloques. El número total de bloques debe ser tal que se logre un balance, por lo que se recomienda que sea múltiplo del número de tratamientos.

Aquí se propone una forma sistemática de crear los bloques. En primer lugar, se debe tomar un número de bloques que sea múltiplo del número de tratamientos, la constante de multiplicidad la denotamos con k , entonces, se toman $7k$ bloques. Luego se toman $3k$ puntos iniciales de forma aleatoria. Supongamos que el factor de diseño es la concentración y se van a probar 7 concentraciones denotadas con: C2, C4, C6, C8, C10, C12 y C14. Se van a utilizar 7 bloques, por lo que $k = 1$, entonces se necesitan $3k = 3$ puntos iniciales. En la Figura 5.4 se muestra el inicio del diseño; en las filas se muestran los niveles del factor de diseño (concentración) y en las columnas se muestran los bloques (días). Se ha elegido la concentración C2 para los días D1, D5 y D7.

	D1	D2	D3	D4	D5	D6	D7
C2	X				X		X
C4							
C6							
C8							
C10							
C12							
C14							

Figura 5.4: Inicio de un diseño de bloques incompletos

Una vez que se tiene el punto de inicio, se llenan las casillas en diagonal y, si se acaba, se vuelve a la primera columna hasta completar 7 casillas. En la Figura 5.5 se muestran los pasos sucesivos hasta completar el diseño. Primero se empieza en la casilla que combina el día D1 con el tratamiento C2, continuando en diagonal; luego, se inicia con la casilla que combina el día D7 con el tratamiento C2, pero debe saltarse a la casilla de C4 con D1, luego se continúa en diagonal; y finalmente se hace algo similar empezando en la casilla de C2 con D5.

	D1	D2	D3	D4	D5	D6	D7
C2	X						
C4	X						
C6		X					
C8			X				
C10				X			
C12					X		
C14						X	

	D1	D2	D3	D4	D5	D6	D7
C2					X		
C4						X	
C6							X
C8	X						
C10				X			
C12					X		
C14						X	

Figura 5.5: Pasos de un diseño de bloques incompletos

En la Figura 5.6 se muestra el diseño de bloques incompletos que resultó de este proceso. Se puede observar que en cada uno de los 7 días se cuenta con 3 tratamientos; por ejemplo, en el día D1 se tienen las concentraciones C2, C4 y C8.

	D1	D2	D3	D4	D5	D6	D7
C2	X			X		X	
C4	X	X			X		
C6		X	X				X
C8	X		X	X			
C10		X		X	X		
C12			X		X	X	
C14				X	X	X	

Figura 5.6: Diseño terminado de bloques incompletos

Para llevar a cabo la prueba de la hipótesis sobre el efecto del factor de diseño, se debe hacer un ajuste a la suma de cuadrados de tratamiento (SCTrat). Primero se obtienen los totales ajustados de cada tratamiento mediante:

$$Q_j = y_{j\cdot} - \frac{1}{p} \sum_{k=1}^b n_{jk} y_{\cdot k},$$

donde $y_{j\cdot}$ es la suma del j-ésimo tratamiento y $y_{\cdot k}$ es la suma del k-ésimo bloque.

La suma de cuadrados de tratamientos ajustada (SCTrat.aj) se calcula como:

$$\text{SCTrat.aj} = \frac{p \sum_{j=1}^a Q_j^2}{\lambda a},$$

donde a es el número total de tratamientos, p es el número de tratamientos que aparecen en cada bloque, r es el número de bloques en que aparece un tratamiento y λ se obtiene mediante $\lambda = r(p-1)/(a-1)$. La SCResB en este caso se calcula por diferencia a partir de la suma de cuadrados total (SCTot):

$$\text{SCResB} = \text{SCTot} - \text{SCTrat.aj} - \text{SCBloque}.$$

5.2 Burbujas

Tres investigadoras quieren analizar el efecto que tienen las diferentes proporciones de glicerina utilizada en la mezcla para la confección de burbujas sobre el tiempo promedio de resistencia de las mismas. Se van a presentar 3 experimentos similares donde se toman las personas como bloques, tratando de que cada persona haga burbujas bajo diferentes condiciones. En los primeros dos casos se tienen 6 personas, mientras que en el tercer caso se cuenta con 12 personas; en todos los casos, cada persona hace 5 burbujas con cada tratamiento. El orden en que cada persona tiene que utilizar un tipo de mezcla se aleatoriza con repeticiones hasta completar 5 burbujas por tratamiento. La variable respuesta es el tiempo promedio de resistencia de las burbujas (en minutos) de cada tratamiento para cada persona. En este caso el promedio de las 5 burbujas es una mejor medida de la resistencia que se obtiene en un tratamiento en una persona específica, ya que puede haber muchísima variación entre esa resistencia de una burbuja a otra aún cuando sean hechas por la misma persona y con el mismo tratamiento.

A continuación se describen las variantes de cada experimento:

1. Se consideran dos cantidades de glicerina (60ml y 120ml) con un solo tipo de agua. Cada persona hace 5 burbujas con cada tratamiento en un orden aleatorio, por lo que debe hacer 10 burbujas. Se promedian los resultados de las 5 burbujas de cada tratamiento y se tienen 2 valores de la respuesta por persona, para un total de 12 valores de la respuesta. Como solo hay un valor de la respuesta por tratamiento por persona, el análisis se puede realizar usando el modelo lineal clásico con el análisis de varianza o el modelo mixto con LRT.
2. Se incluye un nuevo factor para investigar si el efecto que tiene la cantidad de glicerina es el mismo con diferentes tipos de agua. Se obtiene una mezcla simple y posteriormente se ajusta al propósito de la investigación. En la mezcla se van variando el tipo de agua (destilada, grifo, añejada) y la cantidad de glicerina (60ml,120ml). A partir de estos factores se obtienen 6 tratamientos. Cada persona hace 5 burbujas con cada tratamiento en un orden aleatorio, por lo que debe hacer 30 burbujas. Se promedian los resultados de las 5 burbujas de cada tratamiento y se tienen 6 valores de la respuesta por persona, uno para cada tratamiento por persona, para un total de 36 valores de la respuesta. Como

solo hay un valor de la respuesta por tratamiento por persona, el análisis se puede realizar usando el modelo lineal clásico con el análisis de varianza o el modelo mixto con LRT.

3. Se modifica el experimento inicial pero interesa ver si el efecto que tiene la cantidad de glicerina varía según el nivel aeróbico de las personas. Para esto se dividen las personas del estudio en tres grupos: bajo, medio y alto. Se cuenta con 4 personas de cada grupo. En este caso hay 2 factores, pero cada bloque (persona) solo puede tener observaciones en los diferentes niveles de un factor (glicerina), mientras que los bloques se clasifican según otro factor superior (nivel aeróbico).

5.2.1 Ejercicios

1. Preparación:

- (a) Cargue el archivo `burbujas.Rdata` que contiene tres bases de datos: 1) `base1` que contiene los datos solamente con el factor **glicerina**, 2) `base2` que contiene el factor **glicerina** y el factor **agua**, y 3) `base3` en donde se tienen el factor **glicerina** y el factor **nivel**.
- (b) Asegúrese que en todas las bases se han definido adecuadamente los factores **glicerina**, **agua** y **nivel**. Además, la persona es el bloque, y también tiene que estar definido como factor.

2. Visualización de datos con glicerina:

- (a) Realice la visualización usando `base1`. Haga un gráfico de líneas para ver el comportamiento de los tratamientos dentro de cada persona. Use `ggplot2` con la persona en el eje X y poniendo `group=glicerina` para que haga una línea para cada tratamiento.
- (b) Observe si se puede esperar que haya diferencia entre las medias de los tratamientos.
- (c) Observe si hay alguna tendencia de valores más altos o más bajos para ciertas personas.
- (d) ¿Se observa interacción entre **glicerina** y **persona**?

- (e) ¿Se podría verificar la hipótesis de no interacción?
- (f) Centre los datos de tiempo dentro de cada persona, es decir, se quiere quitar el efecto del bloque restando la media del bloque al que una observación pertenece. Para centrar los datos, ajuste un modelo solamente con el bloque como factor usando `lm(tiempo~persona)`, y obtenga los valores ajustados usando `mod$fit`. Luego reste a cada respuesta estos valores ajustados, con lo cual en realidad está centrando los datos dentro de cada persona. Finalmente, sume la media general de la respuesta.
- (g) Haga dos gráficos para ver la variabilidad del tiempo en los 2 niveles de glicerina: 1) use los datos originales, 2) use los tiempos centrados. Ponga los dos gráficos uno al lado del otro y asegúrese que ambos gráficos tienen el mismo rango en el eje Y.
- (h) Observe la variabilidad de los residuales en ambos gráficos. ¿A qué se deben las diferencias? Observe si se puede justificar el cumplimiento del supuesto de homocedasticidad.

3. Varianza residual:

- (a) Estime dos modelos: 1) uno incluyendo el factor de diseño y el bloque con el tiempo original como respuesta, 2) otro incluyendo sólo el factor de diseño con el tiempo centrado como respuesta.
- (b) Compare los análisis de varianza de ambos modelos. Observe la suma de cuadrados total, la suma de cuadrados residual y los grados de libertad de ambos modelos.
- (c) ¿Cuál es la forma correcta de obtener la varianza residual?
- (d) Estime la varianza del error.
- (e) ¿Qué representa esta varianza?
- (f) Obtenga el valor de F para probar la hipótesis de igualdad de promedios entre los 2 niveles de glicerina. Calcule la probabilidad asociada con los grados de libertad adecuados. Concluya.

4. Modelo mixto:

- (a) Para ilustrar la aplicación del modelo mixto, se ajusta el modelo con la función `lme` de la librería `nlme`. Escriba el modelo el cual se plantea de forma similar al modelo lineal, pero se separa el término aleatorio (bloque o persona) en el argumento `random`, y se pone de la siguiente forma: `lme(Y ~ X, random= ~1|bloque, data = base)`.
- (b) Obtenga el análisis de varianza con la función `anova`. Obtenga la probabilidad asociada a glicerina y compárela con la obtenida anteriormente.
- (c) Existe otra librería muy utilizada para el análisis de modelos mixtos que se llama `lme4`. Se usa la función `lme4` y el modelo se plantea de la siguiente forma: `lme4(Y ~ X + (1|bloque), data=base)`. Escriba el modelo con esta función.
- (d) Obtenga el `summary`. Observe que el resultado no da una probabilidad, sino que da un valor de *t*, para el cual debe buscarse la probabilidad asociada con los grados de libertad residuales. Obtenga la probabilidad con dos colas y compárela con la obtenida anteriormente.

5. Homocedasticidad:

- (a) Verifique el supuesto de homocedasticidad. No se pueden usar los tiempos originales puesto que debe considerarse la presencia de los bloques. La homocedasticidad es un supuesto que dice que la varianza condicional de la respuesta para cada tratamiento es la misma, lo cual implica que también los errores de los diferentes tratamientos tengan la misma varianza. En este caso, se recomienda obtener los residuales del modelo con bloques y luego hacer la prueba sobre estos residuales con el factor de la siguiente forma: `bartlett.test(mod2$res~glicerina)`.

6. Análisis con glicerina y agua:

- (a) Use `base2`. Visualice los datos. Como se trata de dos factores se pueden centrar los datos por bloque y hacer un boxplot con ambos factores. Para esto se incluyen los factores separados por `+`. Haga un gráfico con los datos centrados como respuesta y ponga `glicerina+agua`, y otro gráfico poniendo `agua+glicerina`.
- (b) Analice si existe interacción entre glicerina y agua tanto gráficamente como con una prueba formal.
- (c) Pruebe si hay efecto de la glicerina en este nuevo diseño, asumiendo que no hay interacción.
- (d) Cuantifique la diferencia con un nivel de confianza de 95%.
- (e) Pruebe si hay efecto del tipo de agua en este nuevo diseño.
- (f) Verifique entre cuáles pares de medias hay diferencias y cuantifique aquellas donde las diferencias son significativas. Use un nivel de significancia de 5% y un nivel de confianza de 95%.

7. Análisis con glicerina y nivel aeróbico:

- (a) Haga un esquema del diseño.
- (b) Use `base3`. Visualice los datos. Haga un gráfico con los datos centrados como respuesta y ponga `glicerina+nivel`.
- (c) Haga el análisis para determinar si hay interacción entre glicerina y nivel usando la función `lmer`. Use la función `drop1`, con el parámetro `| test=Chisq" |` que sirve para obtener la LRT.
- (d) Asumiendo que no hay interacción, determine si hay un efecto de la glicerina y del nivel de aeróbico.

5.2.2 Solución

1. Preparación:

(a) Lectura:

```
load("burbujas.Rdata")
```

(b) Revisión de definición de factores:

```
str(base1)
```

```
## 'data.frame': 12 obs. of 3 variables:
## $ persona : int 1 2 3 4 5 6 1 2 3 4 ...
## $ glicerina: int 60 60 60 60 60 60 120 120 120 120 ...
## $ tiempo : num 7.25 7.15 4.66 5.46 12.85 ...
```

```
base1$persona=factor(base1$persona)
base1$glicerina=factor(base1$glicerina)
```

```
str(base2)
```

```
## 'data.frame': 36 obs. of 4 variables:
## $ persona : int 1 1 1 1 1 1 2 2 2 2 ...
## $ glicerina: int 60 120 60 120 60 120 60 120 60 120 ...
## $ agua     : Factor w/ 3 levels "a","d","n": 1 1 2 2 3 3 1 1 2 2 ...
## $ tiempo   : num 4.54 16.6 7.25 4.75 1.28 ...
```

```
base2$persona=factor(base2$persona)
base2$glicerina=factor(base2$glicerina)
```

```
str(base3)
```

```
## 'data.frame': 36 obs. of 5 variables:
## $ persona : int 1 1 2 2 3 3 4 4 5 5 ...
## $ glicerina : int 60 120 60 120 60 120 60 120 60 120 ...
## $ nivel    : int 1 1 1 1 1 1 1 1 2 2 ...
## $ tiempo   : num 4.54 16.6 7.25 4.75 1.28 ...
```

```
base3$persona=factor(base3$persona)
base3$glicerina=factor(base3$glicerina)
base3$nivel=factor(base3$nivel)
```

2. Visualización de datos con glicerina:

(a) Gráfico de líneas:

```
library(ggplot2)
ggplot(base1, aes(x=persona, y=tiempo, group = glicerina)) +
  stat_summary(fun.y="mean", geom="line", aes(linetype = glicerina))+
```

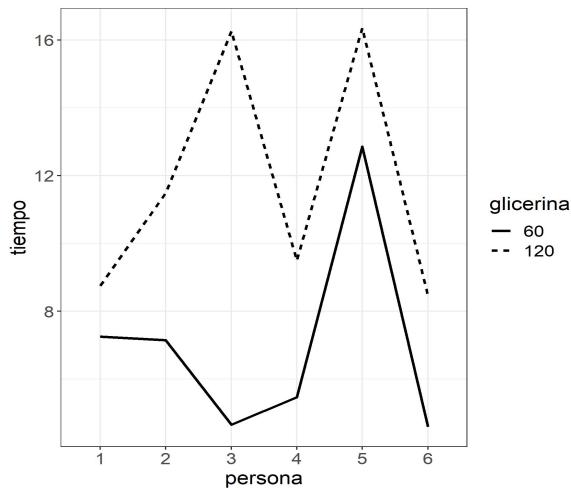


Figura 5.7: Tiempo por persona según niveles de glicerina

(b) Comparación por niveles de glicerina:

El tiempo tiende a ser mayor cuando el nivel de glicerina es 120 (Figura 5.7).

(c) Tendencia según personas:

Las personas 1, 4 y 6 tienden a tener tiempos más bajos, mientras que la persona 5 tiende a tener tiempos más altos. Se quiere eliminar del análisis estas diferencias en los tiempos de las personas, para que la comparación se centre en las diferencias entre los niveles de glicerina.

(d) Interacción entre glicerina y persona:

En algunas personas hay poca diferencia en los tiempos obtenidos para los dos niveles de glicerina (personas 1, 4, 5 y 6), mientras que para la persona 3, la diferencia es muy grande. Esto se puede ver como una indicación de alguna interacción entre glicerina y persona.

(e) Hipótesis de no interacción:

Cuando hay solo una observación por tratamiento para cada bloque, no se puede verificar la hipótesis de no interacción porque no se tienen grados de libertad para los residuales. En ese caso se dice que el efecto de interacción se confunde con el residual y se toma la varianza residual asumiendo que la varianza de interacción es nula, principalmente porque esta varianza residual es fundamental para hacer la verificación de otras pruebas.

(f) Datos centrados por persona:

```
mod1=lm(tiempo~persona,data=base1)
pre=predict(mod1)
t1=base1$tiempo-pre+mean(base1$tiempo)
```

(g) Comparación de gráficos:

```
par(mfrow=c(1,2))
boxplot(tiempo~glicerina,ylim=c(3,17),xlab="glicerina",ylab="tiempo",data=base1)
boxplot(t1~glicerina,ylim=c(3,17),xlab="glicerina",
       ylab="tiempo centrado",data=base1)
```

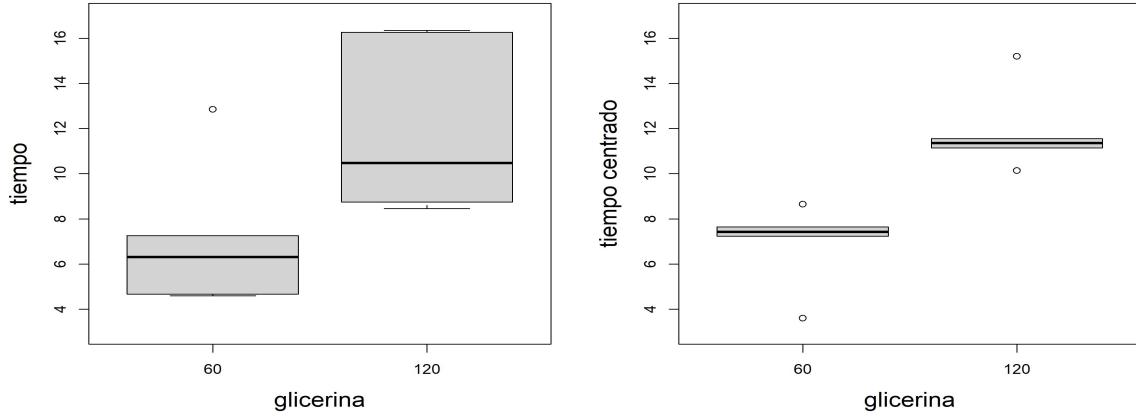


Figura 5.8: Tiempo según niveles de glicerina con datos originales (izquierda) y datos centrados (derecha)

(h) Análisis sobre variabilidad en los gráficos:

En el gráfico con los datos originales (Figura 5.8 izquierda) se observa una mayor variabilidad en el nivel 120 de glicerina, la cual se debe principalmente a las diferencias entre bloques, es decir, las diferencias entre personas. Al eliminar esas diferencias se observa en el gráfico de la derecha que la variabilidad se ha reducido. Ahora la variabilidad es más parecida en los dos tratamientos por lo que se podría pensar que la homocedasticidad sí se cumple.

3. Varianza residual:

(a) Estimación de modelos con datos originales y centrados:

```
mod2=lm(tiempo~glicerina+persona,data=base1)
mod3=lm(t1~glicerina,data=base1)
```

(b) Comparación de resultados:

```

anova (mod2)

## Analysis of Variance Table
## Response: tiempo
##           Df  Sum Sq Mean Sq F value Pr(>F)
## glicerina  1   69.24   69.24  11.42   0.02 *
## persona    5   84.24   16.85   2.78   0.14
## Residuals  5   30.32   6.06

anova (mod3)

## Analysis of Variance Table
## Response: t1
##           Df  Sum Sq Mean Sq F value Pr(>F)
## glicerina  1   69.24   69.24  22.83  0.0007 ***
## Residuals 10  62.43   3.03

sct2=sum(anova (mod2) [,2])
sct3=sum(anova (mod3) [,2])
c(sct2,sct3)

## 183.80 99.56

scres2=anova (mod2) [3,2]
scres3=anova (mod3) [2,2]
c(scres2,scres3)

## 30.32 30.32

```

La SCTot es diferente porque la respuesta cambió. Cuando se centran los datos se está reduciendo la variabilidad total por lo que se tiene una menor SCTot. La SCTot original incluye la varibilidad debida a los bloques (personas) mientras que la SCTot en el caso de datos centrados ya no tiene esa variabilidad; sin embargo, se mantiene la SCRes. Entonces los dos modelos dan un mismo valor de la SCRes pero con diferentes grados de libertad. En el caso de los datos centrados se olvida que el modelo incluyó los bloques por lo que estaría dando más grados de libertad a los residuales de los que realmente tienen.

(c) Varianza residual correcta:

Aunque los dos modelos dan la misma SCRes, el modelo con los datos centrados no toma en cuenta los grados de libertad que se deben atribuir a los bloques en el diseño, por lo tanto, la estimación correcta es la que da el modelo donde se incluyeron los bloques con la variable original.

(d) Estimación de la varianza del error:

```
anova(mod2) [3,3]
```

```
## 6.06
```

La estimación correcta de la varianza del error es el CMRes del modelo que incluye el bloque con la respuesta original, la cual es 6,06.

(e) ¿Qué representa esta varianza?

Este valor es una medida de la variabilidad de la respuesta dentro de cada tratamiento una vez que se ha eliminado el efecto del bloque.

(f) Valor de F y probabilidad:

```
69.24/6.06
```

```
## 11.42
```

```
pf(11.42,1,5,lower.tail=F)
```

```
## 0.02
```

El interés es ver si el nivel de glicerina tiene un efecto sobre la resistencia promedio de las burbujas. Puesto que la probabilidad asociada a glicerina es menor a 0.05 (0.022), se rechaza la hipótesis de igualdad de medias entre los dos niveles de glicerina. Se concluye que se ha encontrado un efecto del nivel de glicerina sobre la resistencia promedio de las burbujas. Como el experimento considera un factor que tiene solo dos niveles, conviene hacer un intervalo de confianza para la diferencia entre estos dos promedios, para lo cual no se requiere hacer ningún tipo de corrección por tratarse de una sola comparación.

4. Modelo mixto:

(a) Modelo con nlme:

```
library(nlme)
mod4=lme(tiempo ~ glicerina, random=~1|persona, data = base1)
```

(b) Análisis de varianza:

```
anova(mod4)
```

	numDF	denDF	F-value	p-value
## (Intercept)	1	5	62.88	0.0005
## glicerina	1	5	11.42	0.0197

Se obtiene la misma probabilidad que en el modelo lineal clásico ($p = 0,02$).

(c) Modelo con lme4:

```
library(lme4)
mod5=lmer(tiempo ~ glicerina + (1|persona), data = base1)
```

(d) Prueba t:

```
summary(mod5)
```

```
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)    6.99      1.38   5.06
## glicerina120  4.80      1.42   3.38
```

```
pt(3.38,5,lower.tail = F)*2
```

```
## 0.02
```

Se busca la probabilidad para el valor de t obtenido de 3.38 con los 5 grados de libertad asociados a los residuales. Se multiplica esta probabilidad por 2 porque es una prueba de 2 colas. Se obtiene la misma probabilidad que en el modelo lineal clásico ($p = 0,02$).

5. Homocedasticidad:

(a) Verificación del supuesto de homocedasticidad:

```
bartlett.test(mod2$res~base1$glicerina)

## Bartlett test of homogeneity of variances
## data: mod2$res by glicerina
## Bartlett's K-squared = -1.61e-15, df = 1, p-value = 1
```

No se rechaza la hipótesis de homocedasticidad; por lo tanto, se puede asumir que las varianzas de la respuesta dentro de cada tratamiento son iguales.

6. Análisis con glicerina y agua:

(a) Visualización:

```
mod6=lm(tiempo~persona,data=base2)
t3=base2$tiempo-predict(mod6)+mean(base2$tiempo)
boxplot(t3~glicerina+agua,xlab="glicerina-agua",ylab="tiempo centrado",
       cex.lab=1.5,data=base2)
```

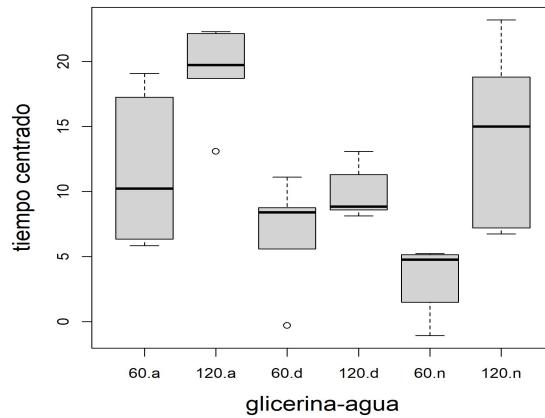


Figura 5.9: Tiempo centrado por persona según nivel de glicerina y tipo de agua

```
boxplot(t3~agua+glicerina,xlab="agua-glicerina",ylab="tiempo centrado",
       cex.lab=1.5,data=base2)
```

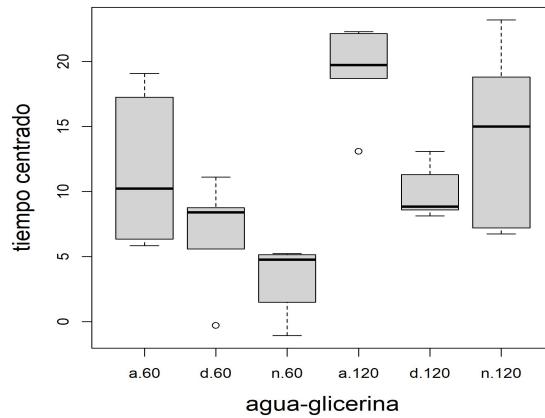


Figura 5.10: Tiempo centrado por persona según tipo de agua y nivel de glicerina

En la Figura 5.9 destaca que el promedio de tiempo parece ser mayor cuando el nivel de glicerina es 120, mientras que en la Figura 5.10 se observa que el promedio de tiempo parece ser mayor cuando el agua es añejada.

(b) Interacción entre glicerina y agua:

```
ggplot(base2, aes(x=agua, y=tiempo, group = glicerina)) +
  stat_summary(fun.y="mean", geom="line", aes(linetype = glicerina))
```

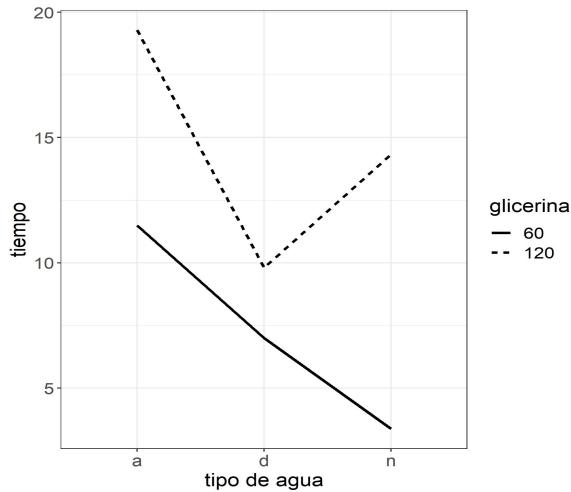


Figura 5.11: Tiempo por tipo de agua según nivel de glicerina

Parece haber interacción entre glicerina y agua ya que cuando se tiene agua destilada la diferencia entre los promedios para los dos niveles de glicerina es un poco menor que en los otros dos casos; sin embargo, puede ser que haya mucha variabilidad en el error y esa interacción no sea tan clara (Figura 5.11).

```
mod7=lm(tiempo~glicerina*agua+persona,data=base2)
anova(mod7)

## Analysis of Variance Table
##
## Response: tiempo
##              Df Sum Sq Mean Sq F value    Pr(>F)
## glicerina     1 464.35 464.35 19.99 0.000 *** 
## agua          2 367.73 183.87  7.92 0.002 **  
## persona       5 706.48 141.30  6.08 0.001 *** 
## glicerina:agua 2 101.26 50.63  2.18 0.134    
## Residuals    25 580.68 23.23
```

Al hacer la prueba formal de la hipótesis de no interacción entre glicerina y agua se obtiene una probabilidad de error tipo I (0,13) más alta que el nivel de significancia de 0,05, por lo que no se puede afirmar que haya interacción. En adelante se puede asumir que estos dos factores no tienen interacción entre sí.

(c) Efecto de la glicerina:

```
mod8=lm(tiempo~glicerina+agua+persona, data=base2)
anova(mod8)

## Analysis of Variance Table
##
## Response: tiempo
##             Df Sum Sq Mean Sq F value    Pr(>F)
## glicerina   1 464.35 464.35 18.38     0.000 ***
## agua         2 367.73 183.87  7.28     0.003 **
## persona     5 706.48 141.30  5.59     0.001 **
## Residuals 27 681.93  25.26
```

La probabilidad asociada al factor glicerina es pequeña ($p < 0,001$), por lo que se rechaza la hipótesis de igualdad de las medias según niveles de glicerina.

(d) Inferencia con 95 % de confianza:

```
table(base2$glicerina)

## glicerina
## 60 120
## 18 18

(m=tapply(base2$tiempo,base2$glicerina,mean))

##   60   120
## 7.29 14.47

d=m[2]-m[1]
cmres=anova(mod8)[4,3]
ee=sqrt(2*cmres/18)
t=qt(0.95,27)
(lim=d-t*ee)

## 120
## 4.33
```

En este caso se observa que con mayor cantidad de glicerina se obtiene un tiempo promedio mayor de las burbujas. Con 95% de confianza se espera que el tiempo promedio de las burbujas sea al menos 4,33 minutos mayor cuando se usa un nivel de glicerina de 120 ml con respecto a cuando se usa 60 ml, esto independientemente del tipo de agua que se utilice.

(e) Efecto del tipo de agua:

La probabilidad asociada al factor agua es pequeña ($p = 0,003$) por lo que se rechaza la hipótesis de igualdad de las medias según tipo de agua.

(f) Verificación de diferencias según tipo de agua:

```
table(base2$agua)

##      agua
##   a   d   n
## 12 12 12

(m=tapply(base2$tiempo,base2$agua,mean))

##      a     d     n
## 15.39 8.40 8.85

a.d=m[1]-m[2]
a.n=m[1]-m[3]
n.d=m[3]-m[2]
d=c(a.d,a.n,n.d)
ee=sqrt(2*cmres/12)
q=d/ee
p=ptukey(q*sqrt(2),3,27,lower.tail = F); names(p)=c("a-d","a-n","n-d"); p

##      a-d    a-n    n-d
## 0.006 0.010 0.973

t=qt(0.95,27)
lim=d[1:2]-t*ee
names(lim)=names(p)[1:2]
lim

##      a-d    a-n
## 3.50 3.05
```

Se concluye, con un nivel de significancia de 0,05, que el promedio de tiempo con agua añejada es mayor que el promedio con cualquiera de los otros dos tipos de agua. Además se espera con 95% de confianza, que el promedio de tiempo con agua añejada sea al menos 3,5 minutos más que con agua destilada y al menos 3,05 minutos más que con agua del grifo, esto independientemente del nivel de glicerina que se use. Por lo tanto, se tendrá una mayor resistencia promedio con agua añejada que sobrepasa los 3 minutos en promedio el tiempo que se obtiene con los otros tipos de agua. Además, usar 120 ml sobrepasa en promedio en más de 4 minutos que si se usa un nivel de glicerina de 60 ml.

7. Análisis con glicerina y nivel aeróbico:

(a) Esquema del diseño:

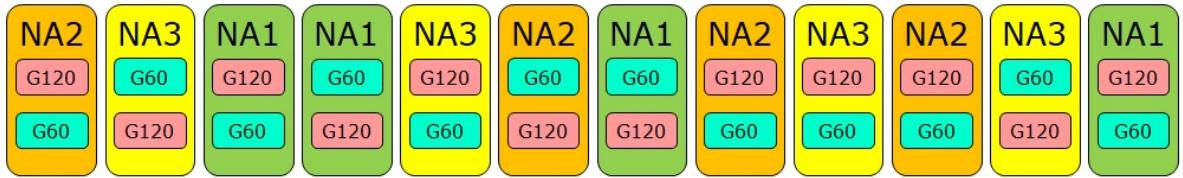


Figura 5.12: Esquema de un diseño de parcelas divididas

Nota: nivel de aeróbico en las parcelas (NA1, NA2, NA3) y nivel de glicerina en las subparcelas (G60, G120).

(b) Visualización:

```
mod9=lm(tiempo~persona,data=base3)
t4=base3$tiempo-predict(mod9)+mean(base3$tiempo)
boxplot(t4-glicerina+nivel,xlab="glicerina-nivel",ylab="tiempo centrado",
       cex.lab=1.5,data=base3)
```

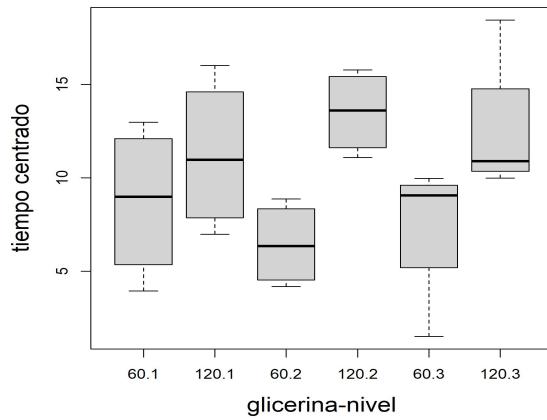


Figura 5.13: Tiempo centrado por persona según nivel de glicerina y nivel aeróbico

En la Figura 5.13 destaca que el promedio de tiempo parece ser mayor cuando el nivel de glicerina es 120; sin embargo, no se aprecian grandes diferencias entre los promedios para diferentes niveles de aeróbicos manteniendo fijo el nivel de glicerina.

(c) Interacción entre glicerina y nivel aeróbico:

```
mod10=lmer(tiempo~glicerina*nivel+(1|persona),data=base3)
drop1(mod10,test="Chisq")

## Model:
## tiempo ~ glicerina * nivel + (1 | persona)
##          npar    AIC    LRT Pr(Chi)
## <none>           168.28
## glicerina:nivel  2 165.35  1.07   0.58
```

El estadístico $\chi^2 = 1,07$ y su probabilidad en la distribución χ^2 con 2 grados de libertad es 0.58, por lo que no se rechaza la hipótesis nula que dice que no hay interacción entre glicerina y nivel de aeróbico. Se puede asumir que no hay interacción y el modelo se reduce.

(d) Efecto de glicerina y efecto de aeróbico:

```
mod11=lmer(tiempo~glicerina+nivel+(1|persona),data=base3)
drop1(mod11,test="Chisq")

## Model:
## tiempo ~ glicerina + nivel + (1 | persona)
##          npar    AIC    LRT Pr(Chi)
## <none>           165.35
## glicerina      1 168.88  5.54   0.02 *
## nivel          2 161.83  0.48   0.78
```

Para probar la hipótesis de igualdad de medias marginales para los 2 niveles de glicerina, el estadístico $\chi^2 = 5,54$ y $p = 0,02$ en la distribución χ^2 con 1 grado de libertad, por lo que se concluye que sí hay un efecto del nivel de glicerina. Por otra parte, no se puede decir que haya un efecto del nivel de aeróbico, ya que $\chi^2 = 0,48$ y $p = 0,78$, valor que es mucho mayor que el nivel de significancia $\alpha = 0,05$.

5.3 Maderas

Se prueban siete concentraciones de madera dura para determinar su efecto sobre la resistencia del papel producido. Se define que las pruebas deben hacerse en 7 días diferentes porque las condiciones del proceso industrial cambian de un día a otro. Cada uno de los días constituye un **bloque**. En cada día solamente se pueden probar tres concentraciones, lo cual indica que los bloques no son completos.

5.3.1 Ejercicios

1. Preparación:

- (a) Abra el archivo `maderas.csv`.
- (b) Verifique que los factores están bien definidos.

2. Visualización de datos:

- (a) Puesto que los bloques son incompletos no se pueden hacer gráficos de líneas y conviene más hacer boxplots. Haga dos boxplots, uno con la resistencia original y otro con la resistencia centrada por día.

3. Sumas de cuadrados:

- (a) Haga el análisis de varianza considerando el factor de diseño y el bloque. Obtenga la suma de cuadrados total y compárela con la que se obtiene a partir de la varianza de la respuesta.
- (b) Extraiga la suma de cuadrados residual.
- (c) Calcule la suma de cuadrados de bloques de la forma usual.

4. Ajuste de las sumas de cuadrados:

- (a) Obtenga los totales ajustados de cada tratamiento.
- (b) Calcule la suma de cuadrados de tratamientos ajustada.
- (c) Obtenga la suma de cuadrados residual a partir de la `SCTot`, `SCBloque` y `SCTrat.aj`. Compárela con la que extrajo del `anova`.

5. Prueba de la hipótesis:

- (a) Compare la `SCTrat.aj` con la `SCRes` para probar la hipótesis de igualdad de medias. Concluya.
- (b) Para hacer el análisis de varianza adecuado de forma automática, debe establecerse un modelo con `lm` o `aov`. Se debe colocar primero el bloque y luego el factor de diseño. Debido a que este diseño no es simétrico, el orden en que se coloquen el bloque y el factor de diseño hace que cambien los resultados, ya que la suma de cuadrados que se ajusta es la

segunda que aparece, de tal forma que el total se mantenga. Puesto que la hipótesis relativa al bloque no interesa ser probada, no hay problema con que la suma de cuadrados de bloque no esté ajustada. Algunos programas ajustan esta suma de cuadrados, por ejemplo, la librería `ibd` tiene una función llamada `aov.ibd` que hace el ajuste para los bloques.

6. Análisis adicionales: a partir de aquí se pueden hacer verificaciones de supuestos, comparaciones múltiples y límites confianza de forma similar a como se hicieron anteriormente.

- (a) Cambie al modelo de **suma nula** y obtenga los coeficientes del modelo.
- (b) Puesto que el día es un bloque, cuando se estiman las medias de resistencia para cada tratamiento no interesa un día en particular, entonces se piensa en un día promedio. Por lo tanto, se pueden eliminar todos los coeficientes relativos a los días. Estime el promedio de resistencia para cada concentración para un día promedio.
- (c) Obtenga las medias observadas para cada concentración y compárelas con las estimaciones del modelo.
- (d) Para hacer comparaciones múltiples se puede eliminar incluso el intercepto y tomar en cuenta solo los 6 coeficientes correspondientes a las concentraciones. También se debe cortar la matriz de covarianza para extraer solo la parte de las concentraciones. Estime los contrastes adecuados entre todos los pares de medias según concentraciones.
- (e) Calcule el error estándar usando los contrastes y verifique que no da lo mismo si usa la fórmula clásica. La fórmula clásica para la varianza de una diferencia de promedios es:

$$V(\bar{y}_i - \bar{y}_j) = \frac{2\text{CMRes}}{r}.$$

- (f) Haga la prueba de las hipótesis asociadas a los contrastes.

5.3.2 Solución

1. Preparación:

(a) Lectura:

```
base=read.csv("maderas.csv")
```

(b) Revisión de definición de factores:

```
str(base)
```

```
## 'data.frame': 21 obs. of 3 variables:
## $ conc: int 2 2 2 4 4 4 6 6 6 8 ...
## $ dia : int 1 5 7 1 2 6 2 3 7 1 ...
## $ res : int 114 120 117 126 120 119 137 117 134 141 ...
```

```
base$dia=factor(base$dia)
base$conc=factor(base$conc)
```

2. Visualización de datos:

(a) Datos originales y centrados:

```
mod1=lm(res~dia,data=base)
res1=base$res-predict(mod1)+mean(base$res)
boxplot(res~conc,xlab="concentración",ylab="resistencia",ylim=c(110,150),data=base)
boxplot(res1~conc,xlab="concentración",ylab="resistencia centrada",
       ylim=c(110,150),data=base)
```

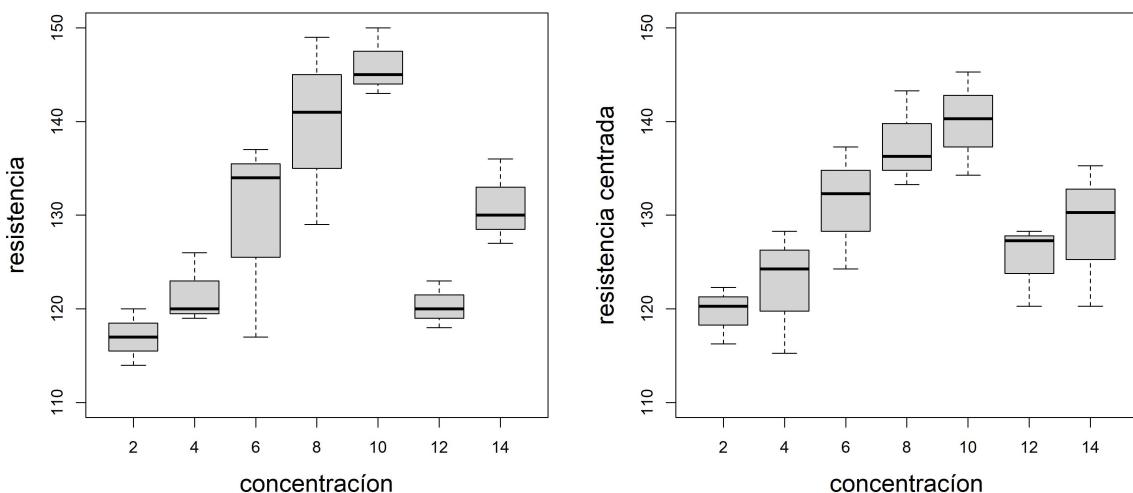


Figura 5.14: Resistencia según concentración con datos originales (izquierda) y datos centrados (derecha)

En la Figura 5.14 se observa que cuando la concentración está entre 2 y 10, se da un aumento casi lineal en las medias de resistencia, pero decaen en los últimos dos niveles de concentración (12 y 14).

3. Sumas de cuadrados:

(a) Suma de cuadrados total (SCTot):

```
mod2=lm(res~conc+dia,data=base)
sum(anova(mod2) [,2])

## 2600.29

n=nrow(base)
(sctot=(n-1)*var(base$res))

## 2600.29
```

(b) Suma de cuadrados residual (SCRes):

```
(scres=anova(mod2) [3,2])

## 168.57
```

(c) Suma de cuadrados de bloque (SCB):

```
table(base$dia)

## dia
## 1 2 3 4 5 6 7
## 3 3 3 3 3 3 3

p=3
m.b=tapply(base$res,base$dia,mean)
mgen=mean(base$res)
(scb=sum(p*(m.b-mgen)^2))

## 1114.29
```

4. Ajuste de las sumas de cuadrados:

(a) Totales de tratamiento ajustados:

```
yi.=tapply(base$res,base$conc,sum)
y.j=tapply(base$res,base$dia,sum)
nij=table(base$conc,base$dia)
(Qi=yi.-as.vector(nij%*%y.j/3))

##   2    4    6    8   10   12   14
## -29 -20    6   25   32  -12   -2
```

(b) Suma de cuadrados de tratamiento ajustada:

```
a=7; p=3; r=3
lambda=r*(p-1)/(a-1)
(sctrat=p*sum(Qi^2)/(lambda*a))

## 1317.43
```

La SCTrat.aj es 1317.43.

(c) Suma de cuadrados residual:

```
(scres2=sctot-scb-sctrat)

## 168.57
```

5. Prueba de hipótesis de igualdad de medias:

(a) Conclusión:

```
cmtrat=sctrat/(a-1)
cmres=scres2/8
(f=cmtrat/cmres)

## 10.42

pf(f,6,8,lower.tail = F)

## 0.002
```

Se rechaza la hipótesis de igualdad de medias puesto que la probabilidad asociada de error tipo I es pequeña ($p = 0.002$). Se concluye que la concentración tiene un efecto sobre la resistencia promedio.

(b) Forma automática:

```
mod3=lm(res~dia+conc,data=base)
anova(mod3)

## Analysis of Variance Table
##
## Response: res
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## dia       6 1114.29 185.71   8.81    0.004 ***
## conc      6 1317.43 219.57  10.42    0.002 ***
## Residuals 8  168.57  21.07
```

Se obtiene la misma probabilidad asociada a la prueba de igualdad de medias de resistencia para las diferentes concentraciones.

6. Análisis adicionales:

(a) Coeficientes con suma nula:

```
options(contrasts=c("contr.sum","contr.poly"))
mod4=lm(res~dia+conc,data=base)
mod4$coef
```

	(Intercept)	dial	dia2	dia3	dia4	dia5	dia6
##	129.29	1.14	2.14	-10.00	7.86	-1.00	-0.43
##		concl	conc2	conc3	conc4	conc5	conc6
##		-12.43	-8.57	2.57	10.71	13.71	-5.14

(b) Estimación de los promedios:

```
beta=mod4$coef[c(1,8:13)]
c1=c(1,1,0,0,0,0,0)
c2=c(1,0,1,0,0,0,0)
c3=c(1,0,0,1,0,0,0)
c4=c(1,0,0,0,1,0,0)
c5=c(1,0,0,0,0,1,0)
c6=c(1,0,0,0,0,0,1)
c7=c(1,-1,-1,-1,-1,-1,-1)
h=cbind(c1,c2,c3,c4,c5,c6,c7)
(med.est=t(h) %*% beta)

## c1 116.86
## c2 120.71
## c3 131.86
## c4 140.00
## c5 143.00
## c6 124.14
## c7 128.43
```

(c) Medias observadas:

```
med.obs=tapply(base$res,base$conc,mean)
cbind(med.est,med.obs)

## med.est med.obs
## c1 116.86 117.00
## c2 120.71 121.67
## c3 131.86 129.33
## c4 140.00 139.67
## c5 143.00 146.00
## c6 124.14 120.33
## c7 128.43 131.00
```

Aunque los valores son parecidos no dan exactamente lo mismo, esto porque el modelo hace un ajuste basado en todos los datos aunque los bloques estén incompletos.

(d) Comparaciones múltiples:

```

c1=c(1,0,0,0,0,0)
c2=c(0,1,0,0,0,0)
c3=c(0,0,1,0,0,0)
c4=c(0,0,0,1,0,0)
c5=c(0,0,0,0,1,0)
c6=c(0,0,0,0,0,1)
c7=c(-1,-1,-1,-1,-1,-1)
c21=c2-c1; c31=c3-c1; c41=c4-c1; c51=c5-c1; c61=c6-c1; c71=c7-c1
c32=c3-c2; c42=c4-c2; c52=c5-c2; c62=c6-c2; c72=c7-c2
c43=c4-c3; c53=c5-c3; c36=c3-c6; c37=c3-c7;
c54=c5-c4; c46=c4-c6; c47=c4-c7; c56=c5-c6; c57=c5-c7; c76=c7-c6
h=cbind(c21,c31,c41,c51,c61,c71,c32,c42,c52,c62,c72,c43,c53,
         c36,c37,c54,c46,c47,c56,c57,c76)
beta=mod4$coef[8:13]

```

```
(I=t(h) %*% beta)
```

```

## c21 3.86
## c31 15.00 c32 11.14
## c41 23.14 c42 19.29 c43 8.14
## c51 26.14 c52 22.29 c53 11.14 c54 3.00
## c61 7.29 c62 3.43 c36 7.71 c46 15.86 c56 18.86
## c71 11.57 c72 7.71 c37 3.43 c47 11.57 c57 14.57 c76 4.29

```

(e) Errores estándar:

```

v=vcov(mod4)[8:13,8:13]
(ee=sqrt(diag(t(h) %*% v %*% h)))

```

```

## c21 c31 c41 c51 c61 c71 ...
## 4.25 4.25 4.25 4.25 4.25 4.25 ...

```

```
table(base$conc)
```

```

## conc
## 2 4 6 8 10 12 14
## 3 3 3 3 3 3 3

```

```

cmres=anova(mod4)[3,3]
sqrt(2*cmres/3)

```

```
## 3.75
```

El error estándar correcto es 4,25 en lugar de 3,75 que se obtendría usando la fórmula clásica.

(f) Prueba de los contrastes:

```
q=L/ee  
(p=ptukey(q*sqrt(2), 7, 8, lower.tail = F))  
  
## c21 0.96  
## c31 0.07  
## c41 0.01  
## c51 0.00  
## c61 0.63  
## c71 0.21  
## c32 0.23  
## c42 0.02  
## c52 0.01  
## c62 0.98  
## c72 0.57  
## c43 0.52  
## c53 0.23  
## c36 0.57  
## c37 0.98  
## c54 0.99  
## c46 0.06  
## c47 0.21  
## c56 0.02  
## c57 0.08  
## c76 0.94
```

Se encontraron diferencias al comparar las concentraciones 4 y 1, 5 y 1, 4 y 2, 5 y 2, 5 y 6. Esto lleva a concluir que con concentraciones 8 y 10 se obtiene una resistencia promedio mayor que con concentraciones 2 y 4. Además es mayor la resistencia promedio con concentración 10 que con 12.

Capítulo 6

Análisis de covariancia

6.1 Conceptos

En el diseño de un experimento se busca controlar las variables que introducen ruido en la respuesta; sin embargo, muchas veces existen variables que no se pueden controlar, pero que se pueden medir durante el experimento. Este tipo de variables se llaman concomitantes o covariables y son de interés si muestran una alta correlación con la respuesta. Su consideración ayuda a reducir la variabilidad del error experimental, de modo similar a la formación de bloques. Por ejemplo, en un estudio donde se quiera ver el efecto de varios tipos de entrenamiento sobre el desempeño de deportistas, posiblemente las características antropométricas van a favorecer a algunos deportistas. Es muy complicado escoger a deportistas que tengan exactamente las mismas características. En este caso, se pueden asignar los deportistas aleatoriamente a los tipos de entrenamiento, y se registran las medidas antropométricas para luego tomarlas en cuenta en el análisis.

Modelo con covariables

El modelo de un factor con una covariante se expresa como un modelo de regresión:

$$\mu_i = \beta_0 + \beta_1 X + \alpha_i.$$

El coeficiente β_0 es una constante que no representa una media como en otros modelos, β_1 es un coeficiente de regresión para la covariable, y el término α_i representa el efecto del i -ésimo nivel del factor de diseño. Este modelo se puede extender para incluir más de una covariable, ya que en esencia es un modelo de regresión.

En el modelo anterior se asume que no hay interacción entre el factor de diseño y la covariable. Debido a la ausencia de interacción, es fácil comparar los promedios de dos tratamientos. Las comparaciones de promedios se hacen de forma condicional, es decir, se compara el promedio de la distribución de la respuesta bajo un tratamiento con el promedio de la distribución bajo otro tratamiento, pero asumiendo que ambas distribuciones tienen el mismo valor de la covariable. Por ejemplo, al comparar los promedios del tratamiento 1 y el tratamiento 2 se obtiene:

$$\mu_1 - \mu_2 = (\beta_0 + \beta_1 X + \alpha_1) - (\beta_0 + \beta_1 X + \alpha_2) = \alpha_1 - \alpha_2.$$

Para ilustrar cómo trabaja una covariable cuando se comparan varios tratamientos, se analiza la relación entre la respuesta y la covariable, diferenciando los puntos según el tratamiento al que pertenecen. En la Figura 6.1 se muestra un conjunto completo de datos. En el lado derecho se ve una gran variabilidad dentro de cada tratamiento, haciendo que no se pueda apreciar claramente que hay diferencias entre los promedios de los dos tratamientos. En la Figura 6.2 se hace la comparación entre los promedios de los dos tratamientos, restringido a un pequeño intervalo de la covariable. Esta es una aproximación a la distribución condicional de la respuesta para un valor fijo de la covariable. Se ha mantenido el rango de la respuesta de la Figura 6.1, para que los gráficos puedan ser comparables. Al enfocar el análisis en este pequeño intervalo de la covariable, la diferencia entre los promedios de los dos tratamientos se hace más evidente.

Análisis formal

Para poner a prueba la hipótesis del efecto del factor de diseño se debe extraer primero la variabilidad que introduce la covariable y, de esta forma, poder hacer las comparaciones de promedios de forma condicional. En este caso no se calcula la suma de cuadrados de tratamiento de la forma usual, sino que se trabaja con la suma de

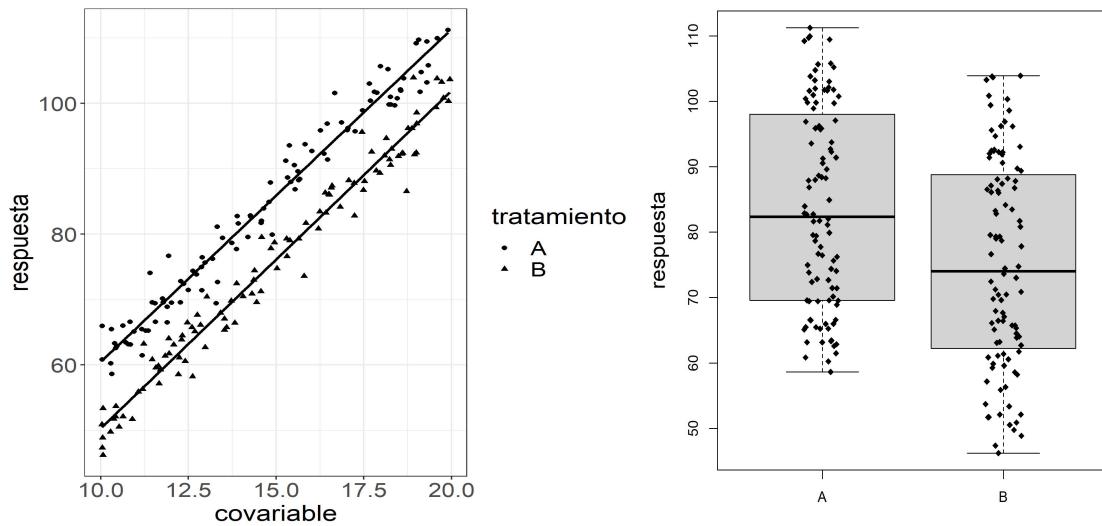


Figura 6.1: Diseño con un factor y una covariable (datos completos)

Nota: a la izquierda se muestra la respuesta contra la covariable, según tratamiento, mientras que a la derecha los datos se agrupan en cada tratamiento sin considerar la covariable.

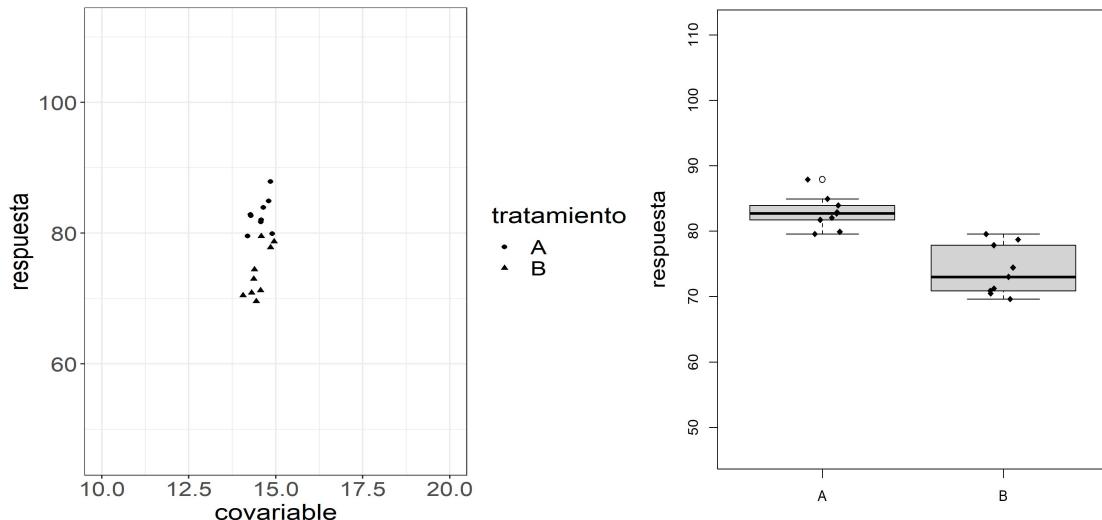


Figura 6.2: Diseño con un factor y una covariable (datos en un intervalo)

Nota: en la parte izquierda se muestra la respuesta contra la covariable, según tratamiento, en un intervalo pequeño de la covariable, mientras que en el lado derecho se muestran los mismos datos con cajas para apreciar las diferencias.

cuadrados residuales (SCRes) de dos modelos: 1) incluyendo solamente la covariable ($SCRes_{\omega}$) y 2) incluyendo tanto la covariable como el factor de diseño ($SCRes_{\Omega}$). La diferencia entre ambas sumas de cuadrados residuales se conoce como la suma de cuadrados de regresión marginal (SCRegM), es decir, $SCRegM = SCRes_{\omega} - SCRes_{\Omega}$.

Esta cantidad representa la porción de la variabilidad total que es explicada por el factor, una vez que se ha tomado en cuenta la covariable en el modelo. Si el factor tiene un efecto sobre la respuesta promedio, esta SCRegM debe ser bastante grande. Para cuantificar el aporte en términos probabilísticos se construye un estadístico F . Se usa la diferencia en los grados de libertad residuales entre los dos modelos ($\text{dif.gl} = \text{gl}_\omega - \text{gl}_\Omega$), y el estadístico F se calcula de la siguiente forma:

$$F = \frac{\text{SCRegMar}/\text{dif.gl}}{\text{CMRes}_\Omega} = \frac{(\text{SCRes}_\omega - \text{SCRes}_\Omega)/(\text{gl}_\omega - \text{gl}_\Omega)}{\text{SCRes}_\Omega/\text{gl}_\Omega}.$$

6.2 Carrera 100 metros

Se hizo un estudio para examinar el efecto que tiene la posición de salida y el tipo de calentamiento en el tiempo de recorrido de 100 metros planos. Se trabajó con la población entre 17 a 24 años en la sede Rodrigo Facio de la Universidad Costa Rica. Se usaron tres tipos de calentamiento: (A) solo estirando, (B) calentamiento normal para hacer ejercicio regular y (C) calentamiento para correr. Además se usaron dos tipos de salida: (+) salida baja con 4 apoyos y (-) salida normal de pie.

Para tomar en cuenta la variabilidad que introduce el hecho de que hay unos estudiantes más grandes que otros y que además tienen pesos diferentes, se tomó el peso y la estatura para calcular el índice de masa corporal de cada uno de ellos, el cual se calcula dividiendo el peso entre la estatura al cuadrado ($IMC = P/E^2$). La variable respuesta es el tiempo al recorrer 100 metros planos (en segundos). Los investigadores consideran que una diferencia de 2 segundos entre dos promedios es relevante.

6.2.1 Ejercicios

1. Preparación:
 - (a) Cargue el archivo `100metros.Rdata`.
 - (b) A partir del peso y la estatura cree la variable `imc`.
 - (c) ¿Cuántos tratamientos tiene este experimento tomando en cuenta solo los factores de diseño?

(d) ¿Cuántas repeticiones hay en cada tratamiento?

2. Linealidad:

- (a) En este caso se tiene una covariable (imc) que se incluye porque se sabe que el tiempo está asociado linealmente al índice de masa corporal como se verá con el cálculo de las correlaciones. Primero obtenga el coeficiente de correlación lineal entre la respuesta y la covariable.
- (b) Obtenga los coeficientes de correlación dentro cada uno de los tratamientos. Use la función `summarise` en la librería `dplyr`, de la siguiente forma: `summarise(group_by(base,A,B), cor(X,Y))`, donde A y B son los factores que definen los grupos.
- (c) Es importante verificar que la relación que existe entre la respuesta y la covariable es lineal dentro de cada tratamiento. Para esto haga un gráfico usando la función `scatterplot` de la librería `car` de la siguiente forma: `scatterplot(y~x)`; para hacer un gráfico para un solo tratamiento delímite la base de la siguiente forma: `data=base[base$calent=="A"&base$salida=="-",]`. Observe cada par de líneas, la línea recta indica una relación lineal perfecta, mientras que la línea curva sigue los datos. Si ambas son parecidas es porque sí hay una relación lineal entre la covariable y la respuesta.

3. Variabilidad de la respuesta:

- (a) Obtenga la variancia de la respuesta dentro de cada tratamiento.
- (b) Obtenga una estimación de la variancia dentro de los tratamientos asumiendo homocedasticidad.
- (c) Para visualizar la variabilidad dentro de cada tratamiento haga primero un gráfico del tiempo por tratamiento incluyendo ambos factores de diseño.
- (d) Ahora se tratará de visualizar la variabilidad de la respuesta tomando en cuenta la covariable. Haga un gráfico de puntos del tiempo contra imc agregando una línea de regresión para cada tipo de tratamiento. Ponga los dos factores dentro de la función: `xyplot(Y~X|A+B, type=c("r", "p"))` en la librería `lattice`.

- (e) Observe la variabilidad que hay en cada tipo de calentamiento y en cada tratamiento. Primero observe como varían todos los puntos de un mismo tipo de calentamiento y un mismo tratamiento, luego vea la variabilidad de los puntos de un mismo tipo de calentamiento y un mismo tratamiento en un intervalo de imc muy corto. Haga este ejercicio visualmente haciendo pequeños intervalos de imc.

4. Inclusión de covariable:

- (a) Para tomar en cuenta la variabilidad que está induciendo el imc, se debe hacer un modelo con los factores y la covariable. Puede incluir la interacción entre los dos factores de diseño. Obtenga los residuales del modelo.
- (b) ¿Qué representa cada uno de estos residuales?
- (c) Obtenga el cuadrado medio residual del modelo a partir de los residuales.
- (d) Compare el cuadrado medio residual obtenido en este modelo con la estimación de la variancia por tratamiento que obtuvo al principio.

5. Prueba formal: para estudiar el efecto de los factores de diseño es importante tomar en cuenta que el imc introduce mucho ruido. Sería ideal tener a todas las personas con valores específicos de imc. Resulta muy complicado hacer grupos por imc puesto que es una variable continua difícil de controlar.

- (a) Estime un modelo con los dos factores de diseño pero no incluya el imc. Plantee este modelo cambiando el orden de los factores. Obtenga el anova de ambos modelos y observe si hay cambios.
- (b) Haga la prueba de hipótesis correspondiente para determinar si alguno de los tipos de calentamiento produce una media de tiempo diferente. Observe la probabilidad asociada.
- (c) Ahora tome en cuenta la covariable introduciéndola en un modelo. Hágalo sin tomar en cuenta la interacción entre los factores de diseño. Hágalo colocando primero los factores y luego la covariable y al contrario.

- (d) Puesto que el anova en este caso se ve afectado por el orden en que se introducen los factores, siempre se debe colocar de último el factor que se está poniendo a prueba. Para entender lo que hace este anova corra dos modelos: 1) con los dos factores de diseño y la covariante, 2) solo con salida y la covariante. Representamos con Ω al modelo más grande y con ω al modelo más pequeño.
- (e) Encuentre la SCRes de ambos modelos que los representamos como $SCRes_{\Omega}$ y $SCRes_{\omega}$. A partir de ellas obtenga la suma de cuadrados de regresión marginal como $SCRegMar = SCRes_{\omega} - SCRes_{\Omega}$, la cual representa la parte de la variabilidad de la respuesta que es explicada por calentamiento cuando entra después de las otras dos variables.
- (f) Construya el estadístico F y haga la prueba de la hipótesis obteniendo la probabilidad asociada a este valor de F en una distribución que tiene los grados de libertad usados en la construcción de la misma.
- (g) La prueba correcta se puede hacer siempre de forma segura con la función `drop1`, a la cual hay que agregar el argumento `test="F"`, que permite comparar dos modelos, uno con la variable a probar y otro sin ella.
- (h) Observe la probabilidad asociada al factor de calentamiento y compárela con la que se había obtenido en el anova que no consideraba el imc.
6. Prueba de interacción: considere la interacción entre los dos factores de diseño (si se siguiera un orden lógico, esto sería lo primero que debería probarse).
- (a) Tome el modelo que contiene la interacción entre esos dos factores y además contiene la covariante. Usando el `drop1` puede verificar la hipótesis de no interacción.

7. Comparaciones finales: en este tipo de análisis puede resultar más conveniente usar el modelo de tratamiento referencia. En cualquiera de los dos modelos el intercepto no representa la media general. La dirección de las comparaciones se hace basada en los coeficientes del factor que se analiza, recordando que el tratamiento referencia tiene un coeficiente igual a cero.
- Escriba el modelo resultante.
 - Obtenga las medias marginales para los 3 tipos de calentamiento y calcule las diferencias entre todos los pares (asegúrese de restar siempre el mayor menos el menor).
 - Obtenga las diferencias de las medias usando vectores y observe que no da el mismo resultado que en el punto anterior.
 - Investigue con cuál tipo de calentamiento se puede esperar el menor tiempo promedio.

6.2.2 Solución

- Preparación:

- Lectura:

```
load("100metros.Rdata")
```

- Creación de imc:

```
base$imc=base$peso/base$estatura^2
```

- Tratamientos:

Hay 6 tratamientos en el diseño pues son dos factores de diseño: tipo de calentamiento (3 niveles) y tipo de salida (2 niveles).

- Repeticiones:

```
table(base$calent,base$salida)
```

```
##      salida
## calent - +
##      A 11 11
##      B 11 11
##      C 11 11
```

2. Linealidad:

(a) Correlación entre respuesta y covariable:

```
cor(base$tiempo,base$imc)
```

```
## 0.77
```

(b) Correlación dentro de cada tratamiento:

```
library(dplyr)
summarise(group_by(base,calent,salida),cor(imc,tiempo))

## # Groups:   calent [?]
##   calent salida `cor(imc, tiempo)`
## 1     A      -       0.86
## 2     A      +       0.78
## 3     B      -       0.88
## 4     B      +       0.92
## 5     C      -       0.61
## 6     C      +       0.76
```

La correlación entre tiempo e imc no es muy alta cuando se toma en general, pero cuando se considera dentro de los tratamientos, en algunos casos resulta ser más alta.

(c) Relación lineal:

```
library(car)
scatterplot(tiempo~imc,cex.lab=1.5,data=base[base$calent=="A"&base$salida=="-",])
scatterplot(tiempo~imc,cex.lab=1.5,data=base[base$calent=="A"&base$salida=="+",])
scatterplot(tiempo~imc,cex.lab=1.5,data=base[base$calent=="B"&base$salida=="-",])
scatterplot(tiempo~imc,cex.lab=1.5,data=base[base$calent=="B"&base$salida=="+",])
scatterplot(tiempo~imc,cex.lab=1.5,data=base[base$calent=="C"&base$salida=="-",])
scatterplot(tiempo~imc,cex.lab=1.5,data=base[base$calent=="C"&base$salida=="+",])
```

En la Figura 6.3 se observa que en algunos casos la relación es bastante lineal como el del centro a la derecha, mientras que en otros hay una cierta curvatura como el de abajo a la derecha; sin embargo, no todas las líneas se comportan de una forma similar, por lo que se asume un comportamiento lineal para todas.

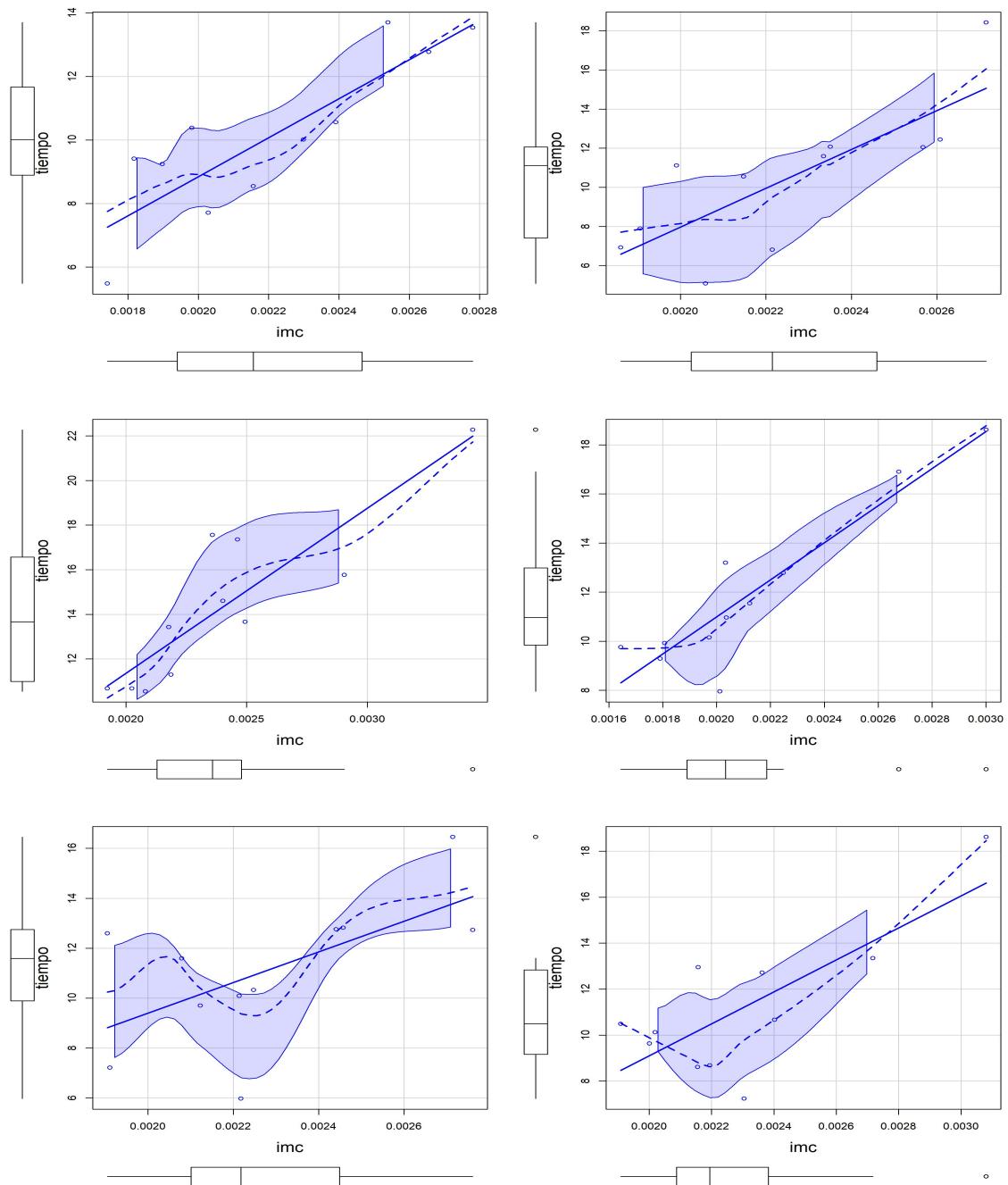


Figura 6.3: Tiempo contra índice de masa corporal según tipo de calentamiento y salida

3. Variabilidad de la respuesta:

(a) Variancia de la respuesta observada dentro de cada tratamiento:

```
(v=tapply(base$tiempo,list(base$calent,base$salida),var))

##      -      +
## A  6.29 13.56
## B 13.63 10.74
## C  8.45  9.81
```

(b) Estimación de la variancia única dentro de los tratamientos:

```
mean(v)

## 10.41
```

(c) Visualización de la variabilidad:

```
boxplot(tiempo~calent+salida,data=base)
```

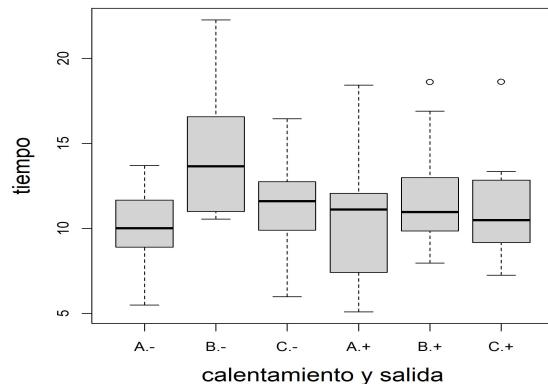


Figura 6.4: Tiempo por tipo de calentamiento y salida

(d) Líneas de regresión por tratamiento:

```
library(lattice)
xyplot(tiempo~imc|calent+salida,type=c("r","p"),data=base)
```

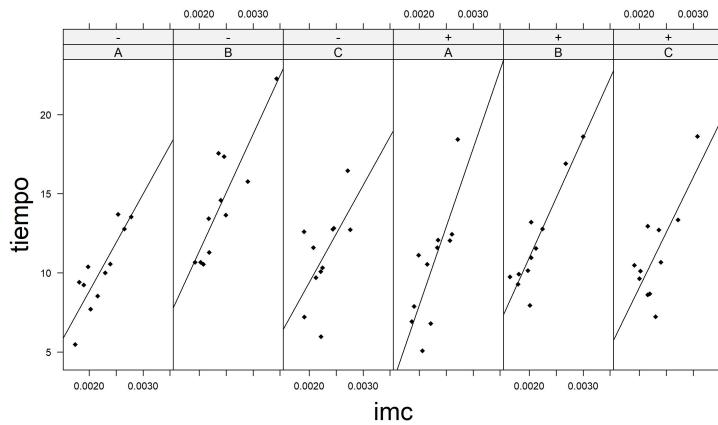


Figura 6.5: Tiempo contra índice de masa corporal según tipo de calentamiento y salida

(e) Variabilidad alrededor de la línea:

Cuando se observan todos los puntos en un solo tratamiento en la Figura 6.4 se ve una gran variabilidad global, en cambio si se ven los puntos en un intervalo de imc muy corto en la Figura 6.5, esa variabilidad se reduce. Por ejemplo, los valores del tiempo para el tratamiento B- en la Figura 6.4 están entre 10 y 22 aproximadamente, mientras que si se toma un intervalo corto de imc alrededor de 0.0025 para ese mismo tratamiento en la Figura 6.5, los tiempos están entre 13 y 18. Con esto se ve cómo la variabilidad condicional se reduce bastante.

4. Inclusión de covariable:

(a) Residuales del modelo con interacción:

```
mod1=lm(tiempo~calent*salida+imc,data=base)
r1=mod1$res
```

(b) Significado de los residuales:

Cada residual representa la distancia del valor observado a la recta correspondiente. No hay una única recta sino una recta para cada tratamiento. Entonces es la distancia del valor observado a la recta del tratamiento al que pertenece, el cual es el promedio estimado para ese tratamiento.

(c) Cuadrado medio residual:

```
(CMRes=sum(r1^2)/59)
```

```
## 3.75
```

(d) Comparación:

```
CMRes
mean(v)
```

```
## 3.75
## 10.41
```

En el modelo se obtiene variancia residual de 3,75 mientras que al principio se tenía 10,41. Al incluir el imc se reduce la variabilidad residual.

5. Prueba formal:

(a) Anova con dos factores:

```
mod2a=lm(tiempo~calent*salida,data=base)
mod2b=lm(tiempo~salida*calent,data=base)
anova(mod2a)
```

```
## Analysis of Variance Table
## Response: tiempo
##           Df Sum Sq Mean Sq F value    Pr(>F)
## calent      2   93.77   46.88   4.50   0.02 *
## salida      1     7.53     7.53   0.72   0.40
## calent:salida 2   25.62   12.81   1.23   0.30
## Residuals   60  624.79   10.41
```

```
anova(mod2b)
```

```
## Analysis of Variance Table
## Response: tiempo
##           Df Sum Sq Mean Sq F value    Pr(>F)
## salida      1     7.53     7.53   0.73   0.40
## calent      2   93.77   46.88   4.50   0.02 *
## salida:calent 2   25.62   12.81   1.23   0.30
## Residuals   60  624.79   10.41
```

El resultado del anova es el mismo independientemente del orden de los factores.

(b) Prueba para el efecto del calentamiento:

Como no hay interacción se usa un modelo sin interacción.

```
mod3=lm(tiempo~calent+salida,data=base)
anova(mod3)
```

```
## Analysis of Variance Table
## Response: tiempo
##           Df Sum Sq Mean Sq F value    Pr(>F)
## calent     2   93.77   46.88     4.47    0.02 *
## salida     1     7.53     7.53     0.72    0.40
## Residuals 62  650.41   10.49
```

Se rechaza la hipótesis de igualdad de medias por lo que hay evidencia de que alguno de los tipos de calentamiento produce una media menor que las otras. La probabilidad asociada es 0,02.

(c) Modelo con la covariable:

```
mod4a=lm(tiempo~calent+salida+imc,data=base)
mod4b=lm(tiempo~imc+calent+salida,data=base)
anova(mod4a)
```

```
## Analysis of Variance Table
## Response: tiempo
##           Df Sum Sq Mean Sq F value    Pr(>F)
## calent     2   93.77   46.88     12.91  2.11e-05 ***
## salida     1     7.53     7.53     2.07    0.155
## imc        1  428.87  428.87    118.09  6.74e-16 ***
## Residuals 61  221.54     3.63
```

```
anova(mod4b)
```

```
## Analysis of Variance Table
## Response: tiempo
##           Df Sum Sq Mean Sq F value    Pr(>F)
## imc        1  444.35  444.35    122.35 3.27e-16 ***
## calent     2   85.50   42.75     11.77  4.75e-05 ***
## salida     1     0.32     0.32      0.09    0.768
## Residuals 61  221.54     3.63
```

Los resultados no se mantienen al intercambiar el orden. Esto es lo que sucede en un modelo de regresión ya que el anova va tomando lo que explica una variable después de que han entrado otras variables, entonces el orden en que entra es importante.

(d) Modelo con la covariable y eliminando calentamiento:

```
mod5=lm(tiempo~calent+salida+imc,data=base)
mod6=lm(tiempo~salida+imc,data=base)
```

(e) Suma de cuadrados de regresión marginal:

```
SCRes1=anova (mod5) [4, 2]
SCRes2=anova (mod6) [3, 2]
(SCRegMar=SCRes2-SCRes1)

## 85.51
```

(f) Estadístico F:

```
CMRes=anova (mod) [4, 3]
(f=(SCRegMar/2)/CMRes)

## 11.77
```

(g) Prueba de la hipótesis:

```
pf(f,2,61,lower.tail = F)

## 0.00005
```

La hipótesis nula es que los dos modelos explican lo mismo, la cual es totalmente equivalente a decir que los coeficientes de calentamiento son todos cero. Esto se puede traducir en que las medias del tiempo son iguales para todos los tipos de calentamiento. Se rechaza la hipótesis nula de igualdad de medias para los diferentes tipos de calentamiento.

(h) Prueba de forma automática:

```
drop1(mod4a, test="F")

## Model: tiempo ~ calent + salida + imc
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>            221.54  89.92
## calent   2      85.51 307.05 107.47    11.77  4.7e-05 ***
## salida   1       0.32 221.86  88.02     0.09   0.768
## imc      1     428.87 650.41 159.00    118.09  6.7e-16 ***

drop1(mod4b, test="F")

## Model: tiempo ~ imc + calent + salida
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>            221.54  89.92
## imc      1     428.87 650.41 159.00    118.09  6.7e-16 ***
## calent   2      85.51 307.05 107.47    11.77  4.7e-05 ***
## salida   1       0.32 221.86  88.02     0.09   0.768
```

En cualquier orden que se pogan los factores en el modelo se obtiene la misma probabilidad asociada a la prueba que interesa.

(i) Probabilidad obtenida:

Ahora la probabilidad asociada es de 0,00005 que es mucho menor que la obtenida anteriormente (0,015).

6. Prueba de interacción:

(a) Verificación formal:

```
drop1(mod1,test="F")

## Single term deletions
## Model:
## tiempo ~ calent * salida + imc
##           Df Sum of Sq   RSS     AIC F value    Pr(>F)
## <none>          221.10  93.79
## imc            1    403.68 624.79 160.35  107.72 6.3e-15 ***
## calent:salida  2      0.43 221.54  89.92    0.06   0.944
```

La probabilidad asociada es muy alta por lo que se asume que no hay interacción entre calentamiento y salida. Esto lleva a suponer que el efecto que tiene el tipo de calentamiento sobre el tiempo promedio es independiente del tipo de salida. Es correcto el análisis que se hizo de verificar el efecto del calentamiento en general.

7. Comparaciones finales:

(a) Modelo resultante:

$$\mu_{ij,X} = \beta_0 + \alpha_i + \beta_j + \gamma X.$$

(b) Medias marginales por tipo de calentamiento:

```
(mm=tapply(base$tiempo,base$calent,mean))

##      A      B      C
## 10.29 13.13 11.15

dBA = mm[2]-mm[1]
dCA = mm[3]-mm[1]
dbc = mm[2]-mm[3]
dif=c(dBA,dCA,dbc)
names(dif)=c("BA", "CA", "BC")
dif

##    BA    CA    BC
## 2.85 0.87 1.98
```

(c) Diferencias de medias usando vectores:

Primero se obtienen los coeficientes y se observa que la referencia es el tipo A, además el coeficiente más alto es para B y luego C. Entonces se comparan B-C, B-A y C-A.

```
(b=mod4a$coef)

##   (Intercept)    calentB    calentC    salida+      imc
##       -6.03      2.60      0.42     -0.14    7350.86

cBA=c(0,1,0,0,0)
cCA=c(0,0,1,0,0)
cBC=c(0,1,-1,0,0)
cont=cbind(cBA,cCA,cBC)
(L=t(cont) %*% b)

## cBA 2.60
## cCA 0.42
## cBC 2.18
```

La diferencia entre los calentamientos B y A da 2.85 si se usan las medias observadas, mientras que da 2.60 si se usan los coeficientes del modelo; entre C y A da 0.87 con las medias observadas y 0.42 con los coeficientes del modelo; y finalmente, entre B y C se obtiene 1.98 con las medias observadas contra 2.18 a partir de los resultados del modelo. Evidentemente no es lo mismo si se usan las medias observadas que si se usa el modelo; puesto que se está haciendo una modelación estadística que tiene supuestos como la no interacción entre el calentamiento y el imc, es fundamental usar el modelo para las comparaciones, tal como se hace en el punto siguiente.

(d) Comparación por tipo de calentamiento:

```
ee=sqrt(diag(t(cont) %*% vcov(mod4a) %*% cont))
qt=L/ee
(p=pt(qt,61,lower.tail = F))

## cBA 0.00
## cCA 0.23
## cBC 0.00

p<0.05/3

## cBA TRUE
## cCA FALSE
## cBC TRUE
```

Se encuentran diferencias entre B-C y B-A. Ahora se construye un límite inferior para esas diferencias.

```
qt=qt(1-0.05/2, 61)
(lim=L[-2]-qt*ee[-2])
```

```
##   cBA   cBC
## 1.45 1.03
```

Estadísticamente se ha encontrado que C produce un tiempo promedio inferior a B, y de igual forma A produce un tiempo promedio inferior a B pero que entre A y C no se han detectado diferencias; sin embargo, puesto que el investigador había establecido que para que una diferencia entre dos promedios se considerara importante debería ser de al menos 2 segundos, la diferencia entre esos tipos de calentamiento no es claramente relevante pues su límite inferior llega a 1,03 segundos en el caso B-C y a 1,45 segundos en el caso B-A.

6.3 Asphalt

Se hizo un estudio cuyo objetivo era investigar si al implementar antioxidantes como aditivos para el mortero asfáltico (agregado + asfalto), se retarda el proceso de oxidación. Se utilizan dos variables como indicadores de este proceso: el módulo complejo (MegaPascales) y el ángulo de fase (grados). En el estudio se quieren comparar 2 antioxidantes (ácido ascórbico y orujo), donde cada uno de ellos se utiliza en tres diferentes concentraciones (1, 3 y 5%).

Además de los factores de diseño, se sabe que los resultados de las variables críticas se miden a temperaturas y frecuencias determinadas, las cuales van a impactar estos resultados. Estas temperaturas y frecuencias también se obtienen para períodos de envejecimiento determinados (4, 8 y 16 años). Adicionalmente se tienen dos fuentes de donde proviene el agregado o piedra fina (Barranca y Guápiles). Si bien el conocer el efecto de la fuente de agregado no es el objetivo primordial del estudio, podría darse que la fuente introduzca variabilidad en los resultados.

Puesto que las dos variables respuesta están altamente correlacionadas, el investigador decide concentrar su análisis en el módulo. Él considera que una diferencia de 10 MegaPascales entre dos promedios es bastante considerable. El antioxidante que tenga un módulo promedio menor se considera más apropiado.

6.3.1 Ejercicios

1. Preparación:

- (a) Cargue el archivo `asfalto.Rdata`. Para iniciar se van a considerar solo los datos para una concentración de 5% que corresponde al nivel alto en la variable concentración. Cambie los niveles de agregado a las iniciales de cada sitio para que sea más fácil de leer en los análisis. Haga una nueva base donde se filtren sólo los los datos que se requieren.
- (b) ¿Cuáles son los factores de diseño? ¿Cuántos tratamientos tiene este experimento?
- (c) ¿Cuántas repeticiones hay en cada tratamiento?
- (d) Puesto que el envejecimiento y el agregado son dos factores que se controlan, se van a incluir en el análisis para reducir el ruido. Al incluir estos dos nuevos factores, ¿cuántas combinaciones tratamiento-factor se obtienen?
- (e) ¿Cuántas repeticiones hay en cada combinación?

2. Variabilidad de la respuesta:

- (a) Obtenga la variancia de la respuesta para cada antioxidante y la media de estas variancias.
- (b) Obtenga la variancia de la respuesta dentro de cada combinación tratamiento-factor y la media de estas variancias. Compare esta medida de la variabilidad con la que había obtenido anteriormente.
- (c) Para visualizar la variabilidad dentro de cada tratamiento haga primero un gráfico del módulo por tratamiento. Al lado haga un gráfico del módulo por combinación tratamiento-factor. Observe que las etiquetas del eje X son muy largas por lo que no caben, entonces se pueden voltear usando `las=2` en el boxplot.
- (d) Discuta si el haber agregado los factores controlados disminuyó la variabilidad, de tal forma que sea más evidente si uno de los antioxidantes es mejor que el otro en términos del promedio de módulo.

3. Inclusión de covariable: ahora se tratará de visualizar la variabilidad de la respuesta tomando en cuenta una de las covariables. En este caso se tienen dos covariables (temperatura y frecuencia) que se incluyen porque se sabe que el módulo está asociado con la temperatura y con la frecuencia.
 - (a) Obtenga la correlación entre la respuesta y cada una de las covariables.
 - (b) Obtenga estas mismas correlaciones pero dentro de cada uno de los niveles de antioxidante.
 - (c) Haga un grafico de puntos del módulo contra la temperatura agregando una línea de regresión para cada antioxidante. En la librería lattice, use `xyplot(Y~X, groups=F, type=c("r"))`, donde Y es la respuesta, X es la covariable y F es el factor.
 - (d) Observe la variabilidad que hay alrededor de cada línea. Si hace una relación con un modelo de regresión, ¿a qué concepto corresponde esta variabilidad?
 - (e) Aunque la temperatura es una variable continua, en este caso se nota que hay datos agrupados en 5 rangos de temperatura, y se cuenta con otra variable que solo indica el rango de temperatura en su punto medio (esta variable es `temp2`). Esto permite que se pueda visualizar la variabilidad que hay alrededor de cada media usando un boxplot. En una sola ventana dividida en tres partes, repita los dos boxplots que hizo anteriormente y agregue otro boxplot con el módulo contra la combinación de antioxidante y temperatura (`temp2`).
 - (f) Para tomar en cuenta la variabilidad que está induciendo la temperatura, estime un modelo con el factor antioxidante y la covariable temperatura (use la temperatura continua original - `temp`). También estime un modelo donde sólo se tiene el factor de diseño antioxidante.
 - (g) Obtenga el cuadrado medio residual en ambos modelos. Compare los resultados entre sí y compárelos con la estimación de la variancia por tratamiento original.

4. Prueba formal:

- (a) Haga la prueba de hipótesis correspondiente para determinar si alguno de los antioxidantes es mejor en términos de reducir el módulo promedio. Hágalo primero en un análisis de variancia que solo considere el factor de diseño, luego tome en cuenta la temperatura. Cuando se tienen covariables hay que tener cuidado con el uso de la función `anova` ya que cada línea representa el aporte marginal de cada variable. En este caso interesa analizar la variabilidad que explica el antioxidante una vez que se ha eliminado el ruido que introduce la temperatura, por esto, debe introducirse primero la temperatura y luego el antioxidante. En general es más recomendable usar la función `drop1`, indicando `test="F"`, para evitar confusiones.
- (b) ¿Cuál es el papel de la temperatura en el análisis?
5. Análisis completo: ahora vamos a realizar el análisis tomando en cuenta todos los factores incluidos en el estudio. Para esto vemos que hay dos covariables (temperatura y frecuencia), y también se tienen el envejecimiento y el agregado además del factor de diseño (antioxidante). Para que las interpretaciones no sean tan complicadas es deseable que no exista interacción entre factor de diseño y las covariables continuas. El modelo inicial debe incluir estas interacciones para probar si el supuesto de no interacción se cumple. Además deben incluirse interacciones entre el factor de diseño y otros factores incluidos en el análisis.

En el modelo se usa la siguiente notación: i para antioxidante, j para envejecimiento, k para agregado, X_1 para temperatura y X_2 para frecuencia. Se usa el término $\alpha_i^{(1)}$ para referirse a la interacción entre antioxidante y temperatura, y $\alpha_i^{(2)}$ para referirse a la interacción entre antioxidante y frecuencia. El modelo completo es el siguiente:

$$\mu_{ijk,X_1,X_2} = \beta_0 + \alpha_i + \delta_j + \tau_k + \beta_1 X_1 + \beta_2 X_2 + (\alpha\delta)_{ij} + (\alpha\tau)_{ik} + \alpha_i^{(1)} X_1 + \alpha_i^{(2)} X_2.$$

- (a) Compare el modelo completo contra otro modelo donde no hay interacciones entre el factor de diseño y las covariables. Debe estimar dos modelos: 1) el modelo completo y 2) el modelo donde no tenga las interacciones que se mencionan. Luego debe comparar los dos modelos con `anova(mod.pequeño, mod.grande)`. Establezca la hipótesis nula que se va a probar y concluya.
- (b) Ahora haga un proceso de selección hacia atrás para ir descartando interacciones entre el factor de diseño y los otros factores. Use la función `drop1(mod, test="F")`. En este caso debe empezar con el modelo donde ya se eliminaron las interacciones con las covariables.
- (c) Haga un gráfico del módulo contra envejecimiento diferenciando por antioxidante. Use `xyplot` con `type=c("a")`. Puesto que se encontró una interacción entre antioxidante y envejecimiento, trate de explicar con el gráfico por qué se da esa interacción.
- (d) Escriba el modelo resultante.
- (e) Debido a la presencia de interacción, para hacer intervalos de confianza al comparar las medias de módulo entre los dos antioxidantes, debe considerarse el envejecimiento, pero no es importante considerar el agregado, la temperatura o la frecuencia. Deben hacerse tres intervalos de confianza, uno para cada nivel de envejecimiento. Puesto que se nota que el promedio para ASC es siempre mayor que para OR, es importante ver la media de ASC menos la de OR para cada nivel de envejecimiento. Verifique cuál es el nivel de referencia para antioxidante y envejecimiento.
- (f) Se está usando el modelo de tratamiento de referencia y la referencia es ASC para antioxidante y 4 para envejecimiento, es decir que $i=1$ para ASC, $i=2$ para OR, $j=1$ para 4, $j=2$ para 8, $j=3$ para 16. Entonces $\alpha_1 = \delta_1 = 0$, además $(\alpha\delta)_{11} = (\alpha\delta)_{12} = (\alpha\delta)_{13} = (\alpha\delta)_{21} = 0$. Usamos la siguiente notación: μ_{ij} para indicar la media del antioxidante i y el envejecimiento j , para un agregado cualquiera, para una temperatura fija y una frecuencia fija. Verifique las siguientes relaciones:

$$\begin{aligned}\mu_{11} - \mu_{21} &= -\alpha_2 \\ \mu_{12} - \mu_{22} &= -\alpha_2 - (\alpha\delta)_{22} \\ \mu_{13} - \mu_{23} &= -\alpha_2 - (\alpha\delta)_{23}\end{aligned}$$

- (g) Observe los coeficientes del modelo y escriba los coeficientes para los tres contrastes escritos anteriormente.
- (h) Obtenga las estimaciones de los contrastes.
- (i) Verifique la hipótesis para cada contraste usando solo una cola, es decir, poniendo la alternativa $H_1 : L > 0$. Debe verificar si los contrastes son ortogonales para decidir si debe ajustar el α .
- (j) Construya cotas inferiores con 95% de confianza en los casos de 4 y 8 años e interprete el resultado según la relevancia que estableció el investigador.

6.3.2 Solución

1. Preparación:

- (a) Lectura:

```

load("asfalto.Rdata")
str(base)

## 'data.frame':   1080 obs. of  8 variables:
## $ env    : Factor w/ 3 levels "4","8","16": 1 1 1 1 1 1 1 1 1 ...
## $ antiox: Factor w/ 2 levels "ASC","OR": 1 1 1 1 1 1 1 1 1 ...
## $ agreg  : Factor w/ 2 levels "Barranca","Guapiles": 1 1 1 1 1 ...
## $ conc   : Factor w/ 3 levels "bajo","medio",...: 1 1 1 1 1 1 ...
## $ freq   : num  0.1 0.5 1 5 10 25 0.1 0.5 1 5 ...
## $ temp   : num  -10 -10 -10 -10 -10 4 4 4 4 ...
## $ mod    : num  83.5 86.4 92.3 97.3 100.4 ...
## $ ang    : num  7.11 5.58 5.16 4.72 4.54 ...

levels(base$agreg)

## "Barranca" "Guapiles"

levels(base$agreg)=c("B","G")
base2=subset(base,conc=="alto")

```

(b) Factores de diseño y tratamientos:

Hay un factor de diseño que es el tipo de antioxidante puesto que la concentración se ha fijado en el nivel alto. Hay 2 tratamientos en el diseño.

(c) Repeticiones por tratamiento:

```
table(base2$antiox)

## antiox
## ASC OR
## 180 180
```

Hay 180 observaciones por tratamiento.

(d) Combinaciones tratamiento-factor:

Se obtienen 12 combinaciones (2x3x2).

(e) Repeticiones por combinación:

```
with(base2,table(antiox,env,agreg))

## , , agreg = B
##           env
## antiox 4 8 16
##   ASC 30 30 30
##   OR 30 30 30
##
## , , agreg = G
##           env
## antiox 4 8 16
##   ASC 30 30 30
##   OR 30 30 30
```

Hay 30 repeticiones por combinación.

2. Variabilidad de la respuesta:

(a) Variancia de la respuesta observada dentro de cada antioxidante:

```
(v1=tapply(base2$mod,base2$antiox,var))
```

```
##      ASC      OR
## 2785.96 2719.50
```

```
mean(v1)
```

```
## 2752.73
```

(b) Variancia de la respuesta dentro de cada combinación:

```
(v2=with(base2,tapply(mod,list(antiox,env, agrég),var)))
```

```
## , , B
##      4     8    16
## ASC 1285.30 3179.16 2793.84
## OR   975.75 3085.56 2893.70
##
## , , G
##      4     8    16
## ASC 2617.08 3635.20 2958.46
## OR  2232.19 3382.78 2769.04
```

```
mean(v2)
```

```
## 2650.67
```

Ahora se obtiene una variancia media de 2650,7 la cual es menor que la que se obtuvo anteriormente de 2752,7; sin embargo, no es mucha la disminución que se logra.

(c) Visualización de la variabilidad:

```
boxplot(mod~antiox,col=c(2,4),xlab="antioxidante",ylab="módulo",data=base2)
boxplot(mod~antiox+agreg+env,col=c(2,4),cex.axis=0.5,las=2,
        xlab="antioxidante-fuente-envejecimiento",ylab="módulo",data=base2)
```

(d) Discusión:

En el gráfico de la izquierda en la Figura 6.6, se observa una gran variabilidad dentro de cada antioxidante y es difícil determinar si alguno de los antioxidantes produce un módulo promedio menor que el otro. Al considerar las combinaciones de antioxidante por fuente y envejecimiento (derecha), la variabilidad no se reduce mucho y sigue siendo difícil ver diferencias entre antioxidantes.

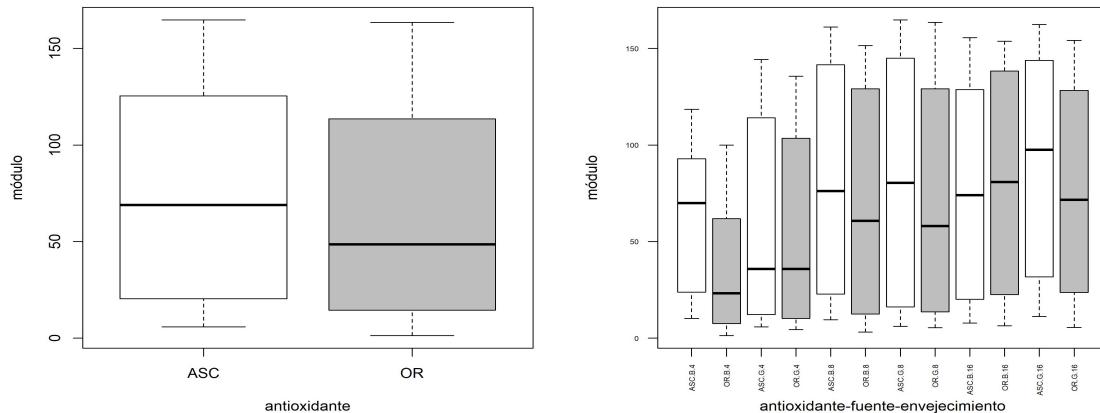


Figura 6.6: Módulo por combinación de antioxidante, fuente y envejecimiento

3. Inclusión de covariable:

(a) Correlación entre respuesta y covariable:

```
with(base2, cor(cbind(temp, freq), mod))

## temp -0.90
## freq 0.17
```

(b) Correlación dentro de cada tratamiento:

```
summarise(group_by(base2, antiox), cor(temp, mod))

### antiox `cor(temp, mod)`
### <fctr>      <dbl>
### 1 ASC       -0.92
### 2 OR        -0.88

summarise(group_by(base2, antiox), cor(freq, mod))

### antiox `cor(freq, mod)`
### <fctr>      <dbl>
### 1 ASC        0.18
### 2 OR         0.17
```

La correlación entre módulo y temperatura es bastante alta tanto en general como dentro de cada nivel de antioxidante; sin embargo, la correlación entre módulo y frecuencia no es tan alta.

(c) Líneas de regresión por antioxidante:

```
xyplot(mod~temp|antiox, type=c("r", "p"), data=base2)
```

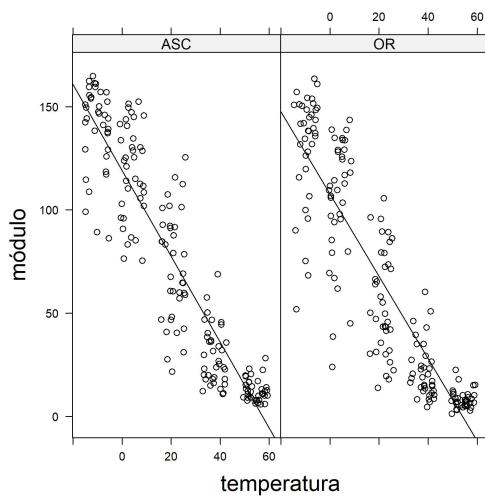


Figura 6.7: Módulo contra temperatura según antioxidante

(d) Variabilidad alrededor de la línea:

La variabilidad que se observa alrededor de cada línea de regresión en la Figura 6.7 corresponde a la variancia residual en un modelo de regresión y se podría estimar con el cuadrado medio residual.

(e) Variabilidad alrededor de la media por combinación:

```
boxplot(mod~antiox, data=base2)
boxplot(mod~antiox+agreg+env, cex.axis=0.5, las=2, data=base2)
boxplot(mod~antiox+temp2, cex.axis=0.5, las=2, data=base2)
```

En la Figura 6.8 se observa claramente que la variabilidad se reduce cuando se incluye la temperatura (derecha).

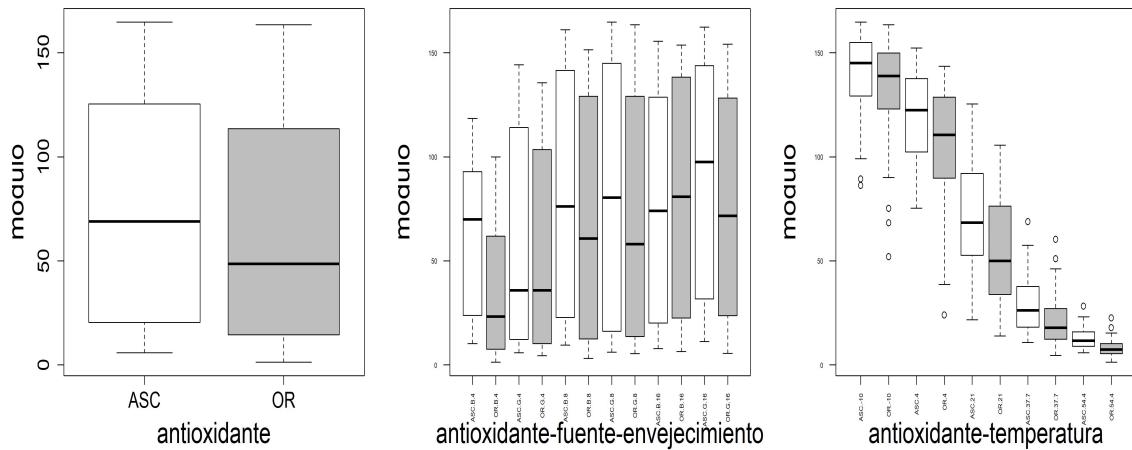


Figura 6.8: Módulo contra combinaciones de antioxidante con fuente y envejecimiento y con temperatura

(f) Modelo con la covariable:

```
mod1=lm(mod~antiox,data=base2)
mod2=lm(mod~temp+antiox,data=base2)
```

(g) Cuadrado medio residual:

```
cmr1=anova(mod1)[2,3]
cmr2=anova(mod2)[3,3]
c(cmr1,cmr2,mean(v1))
```

```
## 2752.73 519.19 2752.73
```

En el modelo donde solo se incluye el factor de diseño se obtiene una variancia residual de 2752,7, mientras que en el modelo donde se ha incluido la temperatura, la variancia residual se reduce a 519,19. Al incluir la temperatura se reduce la variabilidad residual. Cuando se consideraron los tratamientos según el factor de diseño se obtuvo una variancia promedio igual a la del modelo que tiene sólo el factor de diseño.

4. Prueba formal:

(a) Efecto del antioxidante.

```
drop1(mod1,test="F")

## Single term deletions
## Model:
## mod ~ antiox
##          Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>            985477 2853.3
## antiox  1     9709.2 995186 2854.8  3.5271   0.06 .

drop1(mod2,test="F")

## Single term deletions
## Model:
## mod ~ temp + antiox
##          Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>            185350 2253.8
## temp   1     800126 985477 2853.3 1541.11 < 2.2e-16 ***
## antiox 1      8323 193674 2267.6   16.03   7.6e-05 ***
```

En el primer caso (sin incluir la temperatura) no se logra detectar una diferencia entre los promedios de módulo para los tres antioxidantes ($p = 0,06$). Cuando se introduce la temperatura las diferencias se hacen más evidentes ($p < 0,001$). Esto indica que dada una cierta temperatura, los promedios de módulo varían según el antioxidante.

(b) Papel de la temperatura.

En este caso el investigador no está interesado en analizar el impacto que tiene la temperatura sobre el módulo promedio, sino que se introduce en el análisis debido a que hay un enorme ruido en la variable respuesta que es debido a la temperatura. Al eliminar ese ruido o variabilidad, las diferencias entre los promedios para los dos antioxidantes se hacen más evidentes.

5. Análisis completo:

(a) Prueba de interacción entre factor de diseño y covariables:

```
mod3=lm(mod~antiox*env+antiox*agreg+antiox*temp+antiox*frec,data=base2)
mod4=lm(mod~antiox*env+antiox*agreg+temp+frec,data=base2)
anova(mod4,mod3)

## Analysis of Variance Table
## Model 1: mod ~ antiox * env + antiox * agreg + temp + frec
## Model 2: mod ~ antiox * env + antiox * agreg + antiox * temp +
##           antiox * frec
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     350 100705
## 2     348 100358 2     346.68  0.60  0.55
```

La hipótesis nula es que los dos modelos dan una misma explicación, lo que implica que $\alpha_i^{(1)} = \alpha_i^{(2)} = 0$. Al comparar los dos modelos se obtiene una probabilidad de error tipo I de 0,55, lo cual lleva a no rechazar la hipótesis nula y se puede asumir que ambos modelos explican lo mismo, es decir, que no existe interacción entre antioxidante y cada una de las covariables.

(b) Eliminación de interacciones:

```
drop1(mod4,test="F")

## Single term deletions
## Model:
## mod ~ antiox * env + antiox * agreg + temp + frec
##                   Df Sum of Sq    RSS    AIC   F value    Pr(>F)
## <none>              100705 2048.2
## temp             1    799845 900549 2834.9   2779.86 < 2e-16 ***
## freq             1    25911 126616 2128.6    90.05 < 2e-16 ***
## antiox:env      2    2592 103297 2053.3     4.50    0.012 *
## antiox:agreg    1     9 100714 2046.2     0.03    0.857
```

La interacción que es menos significativa es entre antioxidante y agregado. Si se elimina esa interacción no se encuentra una diferencia significativa con el modelo anterior ($p = 0,86$). Ahora se elimina esa interacción y se continúa para ver si la otra interacción debe permanecer en el modelo.

```
mod5=lm(mod~antiox+env+agreg+temp+frec+antiox:env,data=base2)
drop1(mod5,test="F")
```

```
## Single term deletions
## Model:
## mod ~ antiox + env + agreg + temp + frec + antiox:env
##          Df Sum of Sq    RSS   AIC   F value    Pr(>F)
## <none>           100714 2046.2
## aggreg      1     1678 102392 2050.2      5.85    0.016 *
## temp        1    799871 900585 2832.9    2787.64 < 2e-16 ***
## frec         1    25911 126625 2126.6     90.30 < 2e-16 ***
## antiox:env  2     2592 103306 2051.4      4.57    0.016 *
```

Se obtiene una probabilidad muy baja ($p = 0.02$) por lo que hay evidencia de interacción entre antioxidante y envejecimiento. De esta forma se concluye que el efecto del antioxidante no es el mismo en todos los niveles de envejecimiento.

(c) Visualización de interacción entre antioxidante y envejecimiento:

```
library(ggplot2)
ggplot(base, aes(x=env, y=mod, group = antiox)) +
  stat_summary(fun="mean", geom="line", aes(linetype = antiox))
```

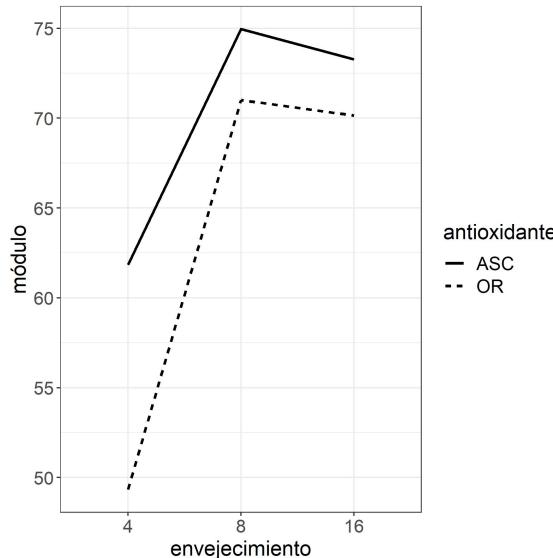


Figura 6.9: Módulo contra envejecimiento según antioxidante

En la Figura 6.9 se ilustra que existe interacción entre antioxidante y envejecimiento. Se observa que para 4 años, la diferencia entre ASC y OR es mayor que en los otros años, y se nota que cuando se llega a 16 años esa diferencia es muy poca. En todos los casos la media de módulo es mayor para ASC que para OR.

(d) Modelo resultante:

El modelo incluye antioxidante, envejecimiento y agregado así como la interacción entre los dos primeros, además incluye las dos covariables temperatura y frecuencia:

$$\mu_{ijk,X_1,X_2} = \beta_0 + \alpha_i + \delta_j + \tau_k + \beta_1 X_1 + \beta_2 X_2 + (\alpha\delta)_{ij}.$$

(e) Niveles de referencia:

```
contrasts(base2$antiox)
```

```
##      OR
## ASC  0
## OR   1
```

```
contrasts(base2$env)
```

```
## 8 16
## 4 0 0
## 8 1 0
## 16 0 1
```

(f) Relaciones en diferencias de promedios:

Se sabe que $\alpha_1 = (\alpha\delta)_{11} = (\alpha\delta)_{12} = (\alpha\delta)_{13} = (\alpha\delta)_{21} = 0$

$$\begin{aligned}\mu_{11} - \mu_{21} &= (\beta_0 + \alpha_1 + \delta_1 + \tau_k + \beta_1 X_1 + \beta_2 X_2 + (\alpha\delta)_{11}) \\ &\quad - (\beta_0 + \alpha_2 + \delta_1 + \tau_k + \beta_1 X_1 + \beta_2 X_2 + (\alpha\delta)_{21}) \\ &= \alpha_1 - \alpha_2 + (\alpha\delta)_{11} - (\alpha\delta)_{21} \\ &= -\alpha_2\end{aligned}$$

$$\begin{aligned}\mu_{12} - \mu_{22} &= (\beta_0 + \alpha_1 + \delta_2 + \tau_k + \beta_1 X_1 + \beta_2 X_2 + (\alpha\delta)_{12}) \\ &\quad - (\beta_0 + \alpha_2 + \delta_2 + \tau_k + \beta_1 X_1 + \beta_2 X_2 + (\alpha\delta)_{22}) \\ &= \alpha_1 - \alpha_2 + (\alpha\delta)_{12} - (\alpha\delta)_{22} \\ &= -\alpha_2 - (\alpha\delta)_{22}\end{aligned}$$

$$\begin{aligned}\mu_{13} - \mu_{23} &= (\beta_0 + \alpha_1 + \delta_3 + \tau_k + \beta_1 X_1 + \beta_2 X_2 + (\alpha\delta)_{13}) \\ &\quad - (\beta_0 + \alpha_2 + \delta_3 + \tau_k + \beta_1 X_1 + \beta_2 X_2 + (\alpha\delta)_{23}) \\ &= \alpha_1 - \alpha_2 + (\alpha\delta)_{13} - (\alpha\delta)_{23} \\ &= -\alpha_2 - (\alpha\delta)_{23}\end{aligned}$$

(g) Planteamiento de contrastes:

```
mod5$coef

## (Intercept) antioxOR    env8    env16   aggreg
##      95.84     -16.98    18.84    21.71     4.32
##      temp       freq  antioxOR:env8 antioxOR:env16
##      -2.04      0.97        9.45      12.54

c1=c(0,-1,0,0,0,0,0,0)
c2=c(0,-1,0,0,0,0,-1,0)
c3=c(0,-1,0,0,0,0,0,-1)
```

(h) Cálculo de contrastes:

```
beta=mod5$coef
cont=cbind(c1,c2,c3)
L=t(cont) %*% beta
row.names(L)=c("ASC-OR,4años", "ASC-OR,8años", "ASC-OR,16años")
L

## ASC-OR,4años 16.98
## ASC-OR,8años  7.53
## ASC-OR,16años 4.34
```

(i) Prueba de hipótesis relativas a los contrastes:

Los contrastes para comparar las hipótesis que se quieren probar son:

$$\begin{aligned}\mu_{11} - \mu_{21} \\ \mu_{12} - \mu_{22} \\ \mu_{13} - \mu_{23}\end{aligned}$$

Tomando el vector de promedios $(\mu_{11}, \mu_{21}, \mu_{12}, \mu_{22}, \mu_{13}, \mu_{23})$, los vectores para obtener los contrastes son:

$$(1, -1, 0, 0, 0, 0)$$

$$(0, 0, 1, -1, 0, 0)$$

$$(0, 0, 0, 0, 1, -1)$$

```
v1 = c(1,-1,0, 0,0, 0)
v2 = c(0, 0,1,-1,0, 0)
v3 = c(0, 0,0, 0,1,-1)
c(v1%*%v2, v1%*%v3, v2%*%v3)

## 0 0 0
```

Estos 3 vectores son ortogonales, por lo no hay que hacer corrección.

```

ee=sqrt(diag(t(cont) %*% vcov(mod5) %*% cont))
t=L/ee
anova(mod5)

## Analysis of Variance Table
##
## Response: mod
##           Df Sum Sq Mean Sq   F value    Pr(>F)
## antiox     1  9709   9709   38.79   1.3e-09 ***
## env        2 51912   25956  103.69 < 2.2e-16 ***
## agreg      1  1630   1630    6.51    0.011 *
## temp       1 802716  802716 2797.56 < 2.2e-16 ***
## frec       1 25913   25913   90.31 < 2.2e-16 ***
## antiox:env 2  2592   1296    4.52    0.016 *
## Residuals 351 100714    287

p=pt(t,351,lower.tail=F)
p

## ASC-OR,4años  0.0000
## ASC-OR,8años  0.0077
## ASC-OR,16años 0.0805

```

Puesto que los contrastes son ortogonales, se comparan estas probabilidades contra $\alpha = 0.05$, y se rechaza la hipótesis solo para los dos primeros contrastes, lo cual indica que efectivamente sólo para 4 y 8 años se obtiene un módulo promedio mayor para ASC que para OR, en cambio para 16 años no hay evidencia de que esto sea así.

(j) Cotas inferiores e interpretación:

Para construir las cotas no se hace ningún ajuste porque los contrastes son ortogonales, y se usa $1 - \alpha$

```

t=qt(1-0.05,351)
lim=L[1:2]-t*ee[1:2]
names(lim)=row.names(L)[1:2]
lim

## ASC-OR,4años ASC-OR,8años
##          11.88         2.43

```

Puesto que la diferencia que definió el investigador como relevante entre dos promedios era 10 MegaPascales, se tiene que para 4 años sí se ha demostrado que el antioxidante ASC es mayor que el OR; sin embargo, para 8 años, aunque esta diferencia es estadísticamente significativa, apenas se ha demostrado que es mayor que 52,43 MegaPascales por lo que no es un resultado definitivo y claro en cuanto a que haya una diferencia relevante. Se concluye que el antioxidante Orujo produce un módulo promedio menor que el Ácido Ascórbico cuando se tiene un envejecimiento de 4 años pero que conforme pasa el tiempo esa diferencia se va haciendo más pequeña a tal punto que puede no ser muy relevante.

Capítulo 7

Potencia

7.1 Conceptos

Cuando se diseña un experimento, se trata de definir una cantidad adecuada de réplicas que de alguna forma asegure que la probabilidad de ver diferencias de cierta magnitud entre promedios sea alta (si esas diferencias existen). Para poder determinar este número de réplicas, el investigador debe hacer un ejercicio importante basado en su conocimiento de la variable que se estudia. La pregunta fundamental se puede plantear como: ¿cuánto es una diferencia entre dos promedios que pueda resultar relevante desde un punto de vista práctico? Esa diferencia no se refiere a promedios muestrales, sino a aquellos promedios que se supone que existen en las poblacionales de interés. Si no se tiene una idea clara de esta diferencia, los resultados también pueden carecer de utilidad práctica.

Diferencia relevante

Supóngase que se conocen las distribuciones de dos poblaciones cuyos promedios se quieren comparar. La separación entre estas distribuciones puede ser muy grande o muy pequeña. La magnitud de qué tan pequeña es esta separación como para ser considerada relevante es de mucho interés. En la Figura 7.1 se muestran dos situaciones, en una de ellas las dos distribuciones están claramente separadas y la diferencia entre sus promedios es de 10 unidades, mientras que en la otra situación

esta diferencia es de solo 2 unidades y las dos distribuciones no se muestran tan diferentes.

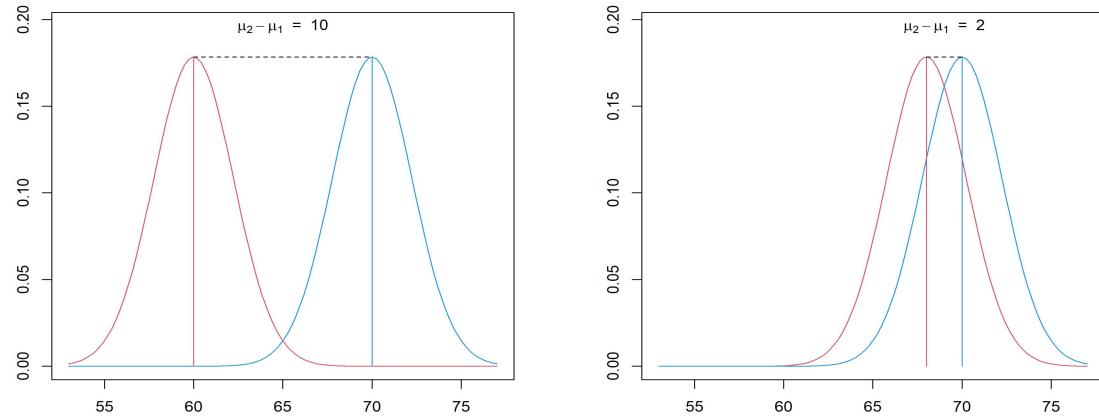


Figura 7.1: Distribuciones hipotéticas

Nota: en la parte izquierda se muestran dos distribuciones donde la medias difieren en 10 unidades, mientras que en el lado derecho la diferencia es de 2 unidades.

Cuando se hacen pruebas para verificar que dos poblaciones no tienen el mismo promedio, se intenta llegar a una conclusión que tenga una validez práctica. Aunque matemáticamente la igualdad válida es $\mu_1 = \mu_2$, existen valores de las medias poblacionales que difieren tan poco que no merecen atención desde un punto de vista práctico. Es aquí donde la persona que investiga debe preguntarse hasta qué punto debe preocuparse por la igualdad de dos promedios. El punto de corte a partir del cual la diferencia empieza a ser relevante se llama detectabilidad de la prueba y se denota con δ . Si se considera que una diferencia es relevante a partir de 5 unidades ($\delta = 5$) y se tienen dos poblaciones cuyos promedios son $\mu_1 = 65$ y $\mu_2 = 70$, se puede afirmar que para quien investiga, estas poblaciones difieren de una forma importante para la aplicación que está haciendo.

Error tipo II y potencia

Si se toman muestras de esas poblaciones cuyos promedios se consideran relevantemente diferentes (por ejemplo, $\mu_1 = 65$ y $\mu_2 = 70$ cuando $\delta = 5$), y se lleva a cabo un procedimiento de prueba de igualdad de promedios, se esperaría que al final

ese procedimiento lleve a concluir que $\mu_1 \neq \mu_2$, es decir, que se rechace la hipótesis de igualdad de promedios; sin embargo, aunque las muestras provengan de esas poblaciones y se tomen de forma aleatoria, en muchos casos, la conclusión a la que se llega podría ser errónea. Cuando sucede esto se dice que se comete error tipo II, que corresponde a no rechazar la hipótesis de igualdad de promedios cuando en realidad estos promedios son bastante diferentes, lo cual se puede expresar diciendo que no se logra detectar diferencias entre promedios que en realidad difieren.

La potencia es una probabilidad que es exactamente la opuesta de la probabilidad de cometer el error tipo II y expresa qué tanta seguridad tiene el investigador de llegar a la conclusión correcta, cuando las muestras provienen de poblaciones cuyos promedios distan entre sí una magnitud que se ha definido como relevante.

Si la distribución de la respuesta en cada población presenta mucha variabilidad, será más difícil llegar a detectar diferencias entre los promedios, aún cuando ellos disten entre sí esa magnitud definida como relevante. Se puede comprobar que para contrarrestar el problema de gran variabilidad en los datos, el aumento del número de réplicas hace que la prueba aumente su potencia, es decir, que sea más probable que detecte diferencias cuando existen.

Estudios de simulación

Los estudios de simulación parten de un modelo que se supone siguen los datos. Una vez establecido el modelo, se generan datos que siguen ese modelo, siendo esta generación mediante un proceso aleatorio. En el proceso de generación de los datos se pueden variar ciertos parámetros, tales como los coeficientes, la varianza del error o la varianza condicional de la respuesta, la distribución del error o la distribución condicional de la respuesta, el número de réplicas por tratamiento, el cumplimiento o no de homocedasticidad, es decir, si se requiere que todas las distribuciones condicionales tengan la misma varianza. El cambiar uno o más de estos parámetros lleva a lo que se conoce como escenarios de simulación. Para cada uno de los escenarios se realiza un proceso de generación de datos, el cual se repite un número alto de veces; cada generación de datos es una iteración en la simulación.

Cuando se utiliza un estudio de simulación para el cálculo de la potencia de una

prueba estadística, se hace la generación de datos según los parámetros establecidos, y en cada iteración se extrae la probabilidad asociada a la prueba de la hipótesis que se está estudiando; se almacenan los valores de estas probabilidades y se calcula la proporción de veces que se rechazó la hipótesis. Esta proporción es una estimación de la potencia que, como se explicó anteriormente, es la probabilidad de detectar diferencias de cierta magnitud cuando existen. Es fundamental que en el modelo del que se parte se asegure que la hipótesis nula sea falsa; por ejemplo, si se tiene un modelo de una vía con 3 niveles, se debe establecer la diferencia relevante (δ) según el criterio experto, y en el modelo se plantea que dos de los promedios difieran esta cantidad. Una forma de hacer esto es definiendo:

$$\begin{aligned}\mu_2 &= \mu_1 + \delta \\ \mu_3 &= \mu_1 + \delta/2\end{aligned}$$

El valor de μ_1 puede escogerse de forma arbitraria; además, se puede establecer que la distribución del error sea normal con una varianza única (σ_ϵ^2), lo cual es equivalente a decir que las distribuciones condicionales de la respuesta en cada tratamiento son las siguientes:

$$\begin{aligned}Y|1 &\sim \mathcal{N}(\mu_1, \sigma_\epsilon^2) \\ Y|2 &\sim \mathcal{N}(\mu_2, \sigma_\epsilon^2) \\ Y|3 &\sim \mathcal{N}(\mu_3, \sigma_\epsilon^2)\end{aligned}$$

En este caso el modelo está asumiendo homocedasticidad, pero es fácil cambiar este supuesto y permitir que las varianzas condicionales cambien, sustituyendo σ_ϵ^2 por σ_1^2 , σ_2^2 y σ_3^2 ; en tal caso se tendrían las siguientes distribuciones:

$$\begin{aligned}Y|1 &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\ Y|2 &\sim \mathcal{N}(\mu_2, \sigma_2^2) \\ Y|3 &\sim \mathcal{N}(\mu_3, \sigma_3^2)\end{aligned}$$

La lógica de las simulaciones se puede extender a situaciones más complejas, como puede ser un diseño de bloques. En ese caso se puede agregar el efecto de los bloques a los datos generados anteriormente. Vamos a suponer que se quieren generar los

datos en un diseño de bloques con un factor que tiene 3 niveles y se utilizan 4 bloques, esto hace que se tengan que generar 4 valores de Y para cada distribución, es decir, se extraen aleatoriamente 4 valores de $Y|1$ que se sabe que se distribuye de forma normal con media μ_1 y la varianza común σ_ϵ^2 , y de igual forma se extraen 4 valores de $Y|2$ y 4 valores de $Y|3$. Estos 12 valores deben acomodarse en bloques, donde cada bloque contenga un valor de cada distribución. Por otra parte, se asume que los efectos de los bloques vienen de la siguiente distribución:

$$\gamma_k \sim \mathcal{N}(0, \sigma_\gamma^2),$$

entonces, se generan 4 valores de esa distribución, donde previamente se ha elegido el valor de σ_γ^2 , que es el que controla qué tan diferentes son los promedios de los bloques; es decir, si se tiene un valor muy bajo de σ_γ^2 el diseño de bloques no sería tan diferente a uno sin bloques. Cada valor de γ_k debe sumarse a cada uno de los valores de Y que pertenecen al k-ésimo bloque. El cuadro 7.1 se muestra la generación de los datos para este diseño, donde Y_{ij} indica que el valor pertenece al tratamiento i-ésimo, mientras que j es un contador de las observaciones dentro de ese tratamiento, por lo que todos los valores de Y_{ij} vienen de la misma distribución:

$$\mathcal{N}(\mu_i, \sigma_\epsilon^2).$$

Cuadro 7.1: Generación de datos para un diseño con bloques

Nivel	Bloque	Media condicional	Efecto bloque	Y original	Y final
1	1	μ_1	γ_1	Y_{11}	$Y_{11} + \gamma_1$
1	2	μ_1	γ_2	Y_{12}	$Y_{12} + \gamma_2$
1	3	μ_1	γ_3	Y_{13}	$Y_{13} + \gamma_3$
1	4	μ_1	γ_4	Y_{14}	$Y_{14} + \gamma_4$
2	1	μ_2	γ_1	Y_{21}	$Y_{21} + \gamma_1$
2	2	μ_2	γ_2	Y_{22}	$Y_{22} + \gamma_2$
2	3	μ_2	γ_3	Y_{23}	$Y_{23} + \gamma_3$
2	4	μ_2	γ_4	Y_{24}	$Y_{24} + \gamma_4$
3	1	μ_3	γ_1	Y_{31}	$Y_{31} + \gamma_1$
3	2	μ_3	γ_2	Y_{32}	$Y_{32} + \gamma_2$
3	3	μ_3	γ_3	Y_{33}	$Y_{33} + \gamma_3$
3	4	μ_3	γ_4	Y_{34}	$Y_{34} + \gamma_4$

7.2 Manzanas

Las manzanas tienen un compuesto llamado polifenol oxidasa, el cual hace que al cortarse y entrar en contacto con el aire se oscurezcan rápidamente. Para evitar el parchamiento se probaron tres tratamientos: tapar (código 2), poner en bolsa plástica cerrada (código 3), y aplicar jugo de limón (código 4). Además, se incluyó un control sin aplicar nada (código 1). Una vez aplicados los tratamientos el resultado fue evaluado por 10 jueces que calificaron el color en una escala de 1 a 6 donde 1 es el color normal de la fruta y 6 es el más oscuro. El objetivo final es seleccionar el tratamiento que mantenga mejor el color original para una empresa que se encarga de banquetes.

7.2.1 Ejercicios

1. Potencia:
 - (a) Tomando como base los datos de las manzanas.
 - i. Observe los promedios muestrales para los 4 tratamientos.
 - ii. Obtenga la estimación de la variancia del error.
 - (b) Vamos a hacer una simulación con un diseño balanceado con un factor que tenga 4 niveles. Para cada tratamiento vamos a usar 10 réplicas. Genere la variable X que representa al factor que en este ejemplo es el tratamiento que se le da a las manzanas. Para esto puede usar `factor(rep(1:4, each=10))`.
 - (c) Vamos a suponer que para el investigador dos promedios empiezan a considerarse realmente diferentes si tienen una diferencia de dos puntos ($\delta = 2$). Vamos a imaginar que se conocen las medias verdaderas y que se cumple con que dos de ellas se diferencian en 2 unidades, por lo que sería importante que al hacer el experimento se pueda concluir que no todas las medias son iguales, es decir, sería lo más deseable que se rechace la hipótesis de igualdad de medias. En este caso se trata de que los promedios más alejados se diferencien en 2 unidades y los otros queden en el centro, de tal forma que la mayor separación entre cualquier par de promedios sea de 2 unidades. Puede tomar $\mu_1 = 3, \mu_2 = 3, \mu_3 = 2$ y $\mu_4 = 4$. Haga un vector de promedios que siga la misma estructura de la variable X y llámelo `mu.j`.

- (d) Cree la variable Y a partir del modelo $Y|i \sim \mathcal{N}(\mu_i, \sigma^2_\epsilon)$. Para esto puede usar el generador de números aleatorios de la función `rnorm`, donde se indica el número total de datos, luego el vector de medias y finalmente la desviación estándar común: `rnorm(n, mu, s)`. Para s puede usar la raíz cuadrada de la varianza del error que estimó anteriormente.
- (e) Cree una función llamada `fun1` que devuelva la probabilidad en el `anova` de datos generados. Aquí podemos generalizar al uso de factores que tengan cualquier cantidad de niveles, para esto le damos a la función un vector de medias verdaderas que llamaremos `mu`, el cual puede tener cualquier cantidad de valores, que indica el número de niveles del factor de diseño. La función que va a crear debe tener como argumentos: el número de réplicas por tratamiento (`r`), el vector de medias verdaderas (`mu`) y la variancia del error (`v`). La función debe crear el vector del tratamiento según el número de réplicas y niveles que tenga el factor, además debe crear la respuesta según el modelo que parte de esas medias verdaderas y la variancia del error.
- (f) Se quiere simular la extracción de datos 1000 veces, y cada vez obtener la probabilidad asociada a la hipótesis de igualdad de medias. Se procede de la siguiente forma:
- i. Use los siguientes valores para las verdaderas medias: 3, 3, 2 y 4. Coloque los valores de esas medias en un vector llamado `mu1`.
 - ii. Use un tamaño de `muestrar=5`. Además use el valor para la variancia del error un poco más alto del observado; puede usar `v=1`. Haga una secuencia de 1000 iteraciones donde cada vez se haga la simulación. Cree primero un vector vacío donde va a ir almacenando la probabilidad asociada en cada iteración. Haga un ciclo con 1000 iteraciones, en cada una de esas iteraciones llame la función `med` y coloque el resultado en la posición correspondiente del vector de probabilidades.
 - iii. Obtenga la proporción de veces que se rechaza la hipótesis nula. Esta proporción representa qué tan fácil es ver las diferencias si en realidad existen y se llama **potencia**.

- (g) Una forma de obtener la potencia directamente en R es usando la función `power.anova.test` de la librería `pwr`. Debe indicarse el número de grupos, el número de observaciones por grupo que en este caso se le llama `n`, debe darse la variancia de las medias hipotéticas en el argumento `between.var` y la variancia del error en el argumento `within.var`. Se usa de la siguiente forma: `power.anova.test(n=5, groups=4, between.var=var(mu1), within.var=1)`.
- (h) Compare este resultado con el obtenido en la simulación.
2. Número de réplicas:
- (a) Si se quiere que haya una probabilidad de 0,90 de rechazar la hipótesis cuando existen diferencias de al menos dos puntos entre medias, se calcula el tamaño de muestra necesario con la misma función pero no se pone el `n` y se especifica la potencia en el parámetro `power`. Para extraer el tamaño de muestra se usa `$n` y para extraer la potencia se usa `$power`.
- (b) En la práctica no se conoce el verdadero valor de la variancia del error, entonces se usa su estimación que es el cuadrado medio residual. Además para el parámetro `between.var` existen infinitas combinaciones de las posibles medias verdaderas, algo que el investigador no conoce. Entonces se puede probar con diferentes valores intermedios en el vector `mu1`. Haga diferentes pruebas y determine en qué caso se obtiene el mayor número de réplicas y en qué caso el menor número de réplicas.

7.2.2 Solución

1. Potencia:

- (a) Datos muestrales:

i. Promedios:

```
load("manzanas.Rdata")
tapply(base$color,base$strat,mean)

##   control    tapar     bolsa    limón
##      5.4      3.2      2.8      1.8
```

ii. Estimación de la variancia del error.

```
(v = mean(tapply(base$color,base$strat,var)))
```

```
## 0.98
```

Como se trata de un diseño de un factor balanceado, para obtener el cuadrado medio residual basta promediar las variancias de los 4 tratamientos, y se obtiene como resultado 0,98.

(b) Factor de diseño:

```
(X = factor(rep(1:4,each=10)))
```

```
## 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 ## Levels: 1 2 3 4
```

(c) Promedios:

```
(muj = rep(c(3,3,2,4),each=10))
```

```
## 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 4 4 4 4 4 4 4 4 4 4
```

(d) Variable respuesta:

```
s = sqrt(v)
n = 4*10
(Y = rnorm(n,muj,s))
```

```
## 3.0 2.8 1.6 2.4 3.3 3.4 1.8 2.6 1.4 2.7 4.1 3.7 2.8 4.0 3.7 3.1 2.1 ...
```

Hay que notar que los valores de Y cambian cada vez que se hace una generación debido a la naturaleza aleatoria de la función rnorm.

(e) Función:

```
fun1 = function (r, mu, v) {
  k = length(mu)
  n = r * k
  X = factor(rep(1:k, each=r))
  muj = rep(mu, each=r)
  s = sqrt(v)
  Y = rnorm(n, muj, s)

  mod = aov(Y ~ X)
  p = anova(mod)[1, 5]
  return(p)
}
```

(f) Simulación:

```
M=1000
mul=c(3,3,2,4)
prob=c()
for(j in 1:M) prob[j]=fun1(r=5,mu=mul,v=1)
mean(prob<0.05)

## 0.66
```

(g) Potencia:

```
library(pwr)
power.anova.test(n=5,groups=4,between.var=var(mul),within.var=1)

##      Balanced one-way analysis of variance power calculation
##
##      groups = 4
##      n = 5
##      between.var = 0.67
##      within.var = 1
##      sig.level = 0.05
##      power = 0.64
##
##      NOTE: n is number in each group
```

(h) Comparación:

Mediante esta función se obtiene una potencia de 0,64 lo que indica que se tiene una probabilidad de 0,64 de detectar diferencias de 2 puntos si estas diferencias existen. Puede entenderse que si realmente hay diferencias de 2 puntos entre algún par de medias verdaderas, y se repite el experimento muchísimas veces, en el 64% de los casos se concluirá que sí hay diferencias mientras que en el 36% no se llegará a esta conclusión. En la simulación se obtuvo un resultado un poco diferente pero se puede llegar a un resultado más cercano al obtenido en la función de R si se hace la simulación con un número mayor de iteraciones.

2. Número de réplicas:

(a) Tamaño de muestra por tratamiento:

```
pot = power.anova.test(groups=4,between.var=var(mul),
                       within.var=1,power=0.9)
pot$n
pot$power

## 8.14
## 0.9
```

Se necesitan al menos 9 réplicas por tratamiento para tener una probabilidad de 0,90 de que se detectarán diferencias de dos unidades si esas diferencias existen.

(b) Variando el vector de medias verdaderas:

```
var(c(2,4,3,3))
power.anova.test(groups=4,between.var=var(c(2,4,3,3)),
                 within.var=1,power=0.9)$n

## 0.67
## 8.14

var(c(2,4,2,2))
power.anova.test(groups=4,between.var=var(c(2,4,2,2)),
                 within.var=1,power=0.9)$n

## 1
## 5.81

var(c(2,2,4,4))
power.anova.test(groups=4,between.var=var(c(2,2,4,4)),
                 within.var=1,power=0.9)$n

## 1.33
## 4.66
```

Aquí se han puesto casos extremos. En uno hay dos medias que distan entre sí 2 puntos pero las otras están en el centro, esto hace que la variancia sea más baja y se requieran más repeticiones para detectar las diferencias. En los otros casos se pusieron todas las medias en los extremos haciendo que haya más variabilidad, por lo que es más fácil detectar las diferencias y por lo tanto se requieren menos réplicas.

7.3 Tortugas

Se desea determinar si la falta de alimento afecta el nivel de proteínas en sangre en tortugas. En el primer caso se tienen tortugas de la especie *Chelonia* midas de ambos géneros (machos y hembras), con dos condiciones: 1) dieta regulada y 2) alimento en abundancia, mientras que en el segundo caso se tiene tortugas de dos especies: *Kinosternum scorpioides* y *Chelonia midas*, con tres condiciones: 1) dieta estricta, 2) dieta balanceada y 3) alimento en abundancia. Se registra el nivel de proteína en gramos por decilitro (gr/dl).

7.3.1 Ejercicios

1. Dos factores sin interacción: tome los datos del archivo `tortugas1.csv`, en el cual se tienen dos factores. Se sabe que no hay interacción entre condición de alimento y género. Aunque se tienen 4 tratamientos, cualquier comparación se hace solo entre dos niveles, sea entre los dos niveles de alimento o entre los dos niveles de género. En este caso particular lo que se desea comparar son los promedios de proteína para los dos niveles de alimento. Para calcular la potencia de este experimento se debe tomar en cuenta la cantidad de réplicas que se tienen en cada nivel de alimento sin importar el género.
 - (a) Obtenga el número de réplicas por condición de alimento.
 - (b) Obtenga la estimación de la variancia del error.
 - (c) Suponga que una diferencia de 2 gr/dl en el promedio de proteína ya es relevante para el investigador. Calcule la potencia que se tiene en este experimento de detectar esa diferencia si la hubiera.
 - (d) ¿Cuántas réplicas se recomendaría utilizar para que la potencia de detectar diferencias de 2 gr/dl sea de 0,90?
 - (e) Calcule el límite inferior para la diferencia entre el promedio de proteínas con alimento en abundancia con respecto a dieta. Use las medias observadas.
 - (f) Se ha observado que no se ha llegado a una conclusión muy clara sobre el efecto real de la condición de dieta. Por otra parte se sugiere una muestra mucho más grande para llegar a detectar diferencias de la magnitud que se estableció. Si se repite el experimento con 38 tortugas por condición, ¿se asegura que el límite inferior supere los 2 gr/dl?
2. Dos factores con interacción: tome como base los datos del archivo `tortugas2.csv`, en el cual también se tienen dos factores. Se sabe que hay interacción entre condición de alimento y especie, por lo tanto, las comparaciones entre promedios según condición se deben hacer para cada especie por separado. Para calcular la potencia de este experimento se debe tomar en cuenta la cantidad de réplicas que se tienen en cada nivel de alimento dentro de cada especie.

- (a) Obtenga el número de réplicas por condición de alimento y especie.
- (b) Obtenga la estimación de la variancia del error.
- (c) Calcule la potencia que se tiene en este experimento de detectar una diferencia de 2 gr/dl en los promedios de proteína si la hubiera.
- (d) En un laboratorio anterior se calcularon los límites inferiores para las diferencias de medias que resultaron significativas. En particular para chelonia se concluyó que con 95% de confianza el nivel promedio de proteína es al menos 7,29 gr/dl con dieta balanceada que con dieta estricta. ¿Por qué si este experimento tiene tan baja potencia se logró llegar a un límite tan alto con tan pocas réplicas para hacer las comparaciones?
- (e) Verifique cuál sería el tamaño de muestra mínimo por tratamiento si se quisiera tener una potencia de 0,95 para detectar diferencias mayores a 7 gr/dl.

7.3.2 Solución

1. Dos factores sin interacción:

- (a) Número de réplicas por condición de alimento:

```
base=read.csv("tortugas1.csv")
base$genero=as.factor(base$genero)
levels(base$genero)=c("macho", "hembra")
table(base$cond)
```

```
## cond
##     alimentodiet
##      8       8
```

Se tienen 8 tortugas en cada condición.

- (b) Estimación de la variancia del error:

```
mod1=lm(proteina~genero+cond, data=base)
(cmres=anova(mod1) [3,3])
```

```
## 6.97
```

La variancia del error se puede estimar con el cuadrado medio residual, y se obtiene como resultado 6,97.

(c) Potencia:

```

mu=c(30,32)
power.anova.test(n=8,groups=2,between.var=var(mu),within.var=cmres)

##      Balanced one-way analysis of variance power calculation
##
##      groups = 2
##      n = 8
##      between.var = 2
##      within.var = 6.97
##      sig.level = 0.05
##      power = 0.29
##
## NOTE: n is number in each group

```

Hay una probabilidad de 0,29 de llegar a la conclusión de que hay diferencias cuando en realidad las medias difieren al menos 2 gr/dl. Es un experimento que difícilmente va a lograr detectar que hay un efecto real de la condición de alimento aún cuando sí se esté dando ese efecto.

(d) Número de réplicas recomendado:

```

power.anova.test(groups=2,between.var=var(mu),within.var=cmres,power=0.9)

##      Balanced one-way analysis of variance power calculation
##
##      groups = 2
##      n = 37.61
##      between.var = 2
##      within.var = 6.97
##      sig.level = 0.05
##      power = 0.9
##
## NOTE: n is number in each group

```

Se necesitan 38 tortugas por condición para asegurar esa potencia.

(e) Límite inferior para la diferencia de medias:

```

(med=tapply(base$proteina,base$cond,mean))

##    alimento     dieta
##      41.20     38.09

d=med[1]-med[2]
ee=sqrt(2*cmres/8)
t=qt(0.95,13)
(LIM=d-t*ee)

##      0.77

```

Con 95% de confianza se espera que la media de proteína cuando se da alimento en abundancia supere en al menos 0,77 gr/dl la que se obtiene con dieta. Este resultado no es satisfactorio para demostrar que es una mejor condición ya que no se obtiene un valor mayor que aquel que es relevante para el investigador

(f) Conclusión:

Al aumentar el número de tortugas lo que se va a lograr es una mejor estimación de los promedios en ambos tratamientos y una reducción del error estándar; sin embargo, esto no quiere decir que el límite inferior para la diferencia tenga que subir necesariamente. Si en realidad la diferencia entre los promedios es mayor a 2 gr/dl, ahora se verá más claro y posiblemente el límite sí supere ese valor, pero si la diferencia no llega a 2 muy posiblemente el límite siga bajo.

2. Dos factores con interacción:

(a) Número de réplicas por condición de alimento:

```
base=read.csv("tortugas2.csv")
base$cond=as.factor(base$cond)
levels(base$cond)=c("estricta","balanceada","abundancia")
table(base$cond,base$especie)

##             especie
##   cond      chelonia kynosternon
##   estricta        4        4
##   balanceada       4        4
##   abundancia       4        4
```

Se cuenta con 4 réplicas por condición de alimento dentro de cada especie.

(b) Estimación de la variancia del error:

```
mod2=lm(proteina~cond*especie,data=base)
(cmres=anova(mod2) [4,3])

## 2.56
```

La variancia del error se puede estimar con el cuadrado medio residual, y se obtiene como resultado 2,56.

(c) Potencia:

```
mu=c(30,31,32)
power.anova.test(n=4,groups=3,between.var=var(mu),within.var=cmres)
```

```
##      Balanced one-way analysis of variance power calculation
##
##      groups = 3
##      n = 4
##      between.var = 1
##      within.var = 2.56
##      sig.level = 0.05
##      power = 0.25
##
##      NOTE: n is number in each group
```

Hay una probabilidad de 0,25 de llegar a la conclusión de que hay diferencias cuando en realidad las medias difieren al menos 2 gr/dl. Es un experimento que difícilmente va a lograr detectar que hay un efecto real de la condición de alimento aún cuando sí se esté dando ese efecto.

(d) Justificación:

Cuando las diferencias son mucho mayores que el nivel de resolución que se ha establecido, es muy fácil concluir que las medias tienen diferencias de una magnitud que está más allá de ese nivel de resolución, aún con pocos datos.

(e) Nuevo tamaño de muestra:

```
mul=c(30,33.5,37)
power.anova.test(groups=3,between.var=var(mul),within.var=cmres,power=0.95)
```

```
##      Balanced one-way analysis of variance power calculation
##
##      groups = 3
##      n = 2.92
##      between.var = 12.25
##      within.var = 2.56
##      sig.level = 0.05
##      power = 0.95
##
##      NOTE: n is number in each group
```

Con 3 datos por tratamiento ya es muy factible que se detecten diferencias tan grandes (de al menos 7 gr/dl).

7.4 Burbujas

Tres investigadoras quieren analizar el efecto que tienen las diferentes proporciones de glicerina utilizada en la mezcla para la confección de burbujas sobre el tiempo promedio de resistencia de las mismas. Se van a presentar 3 experimentos similares donde se toman las personas como bloques, tratando de que cada persona haga burbujas bajo diferentes condiciones. En los primeros dos casos se tienen 6 personas, mientras que en el tercer caso se cuenta con 12 personas; en todos los casos, cada persona hace 5 burbujas con cada tratamiento. El orden en que cada persona tiene que utilizar un tipo de mezcla se aleatoriza con repeticiones hasta completar 5 burbujas por tratamiento. La variable respuesta es el tiempo promedio de resistencia de las burbujas (en minutos) de cada tratamiento para cada persona. En este caso el promedio de las 5 burbujas es una mejor medida de la resistencia que se obtiene en un tratamiento en una persona específica, ya que puede haber muchísima variación entre esa resistencia de una burbuja a otra aún cuando sean hechas por la misma persona y con el mismo tratamiento.

Se consideran dos cantidades de glicerina (60ml y 120ml) con un solo tipo de agua. Cada persona hace 5 burbujas con cada tratamiento en un orden aleatorio, por lo que debe hacer 10 burbujas. Se promedian los resultados de las 5 burbujas de cada tratamiento y se tienen 2 valores de la respuesta por persona, para un total de 12 valores de la respuesta. Como solo hay un valor de la respuesta por tratamiento por persona.

7.4.1 Ejercicios

1. Valores observados a partir de los datos de burbujas:
 - (a) Observe los promedios muestrales para los 2 tratamientos.
 - (b) Obtenga el CMRes del modelo con el factor y los bloques.
2. Preparación para simulación:
 - (a) Vamos a hacer una simulación con un diseño con un factor que tenga 2 niveles y con 6 bloques, esto hace que se tengan que generar 6 valores de Y para cada nivel. Genere la variable X con 6 réplicas para cada uno de los 2 niveles.

- (b) Genere el vector de promedios que siga la misma estructura de la variable X y llámelo `mu.j`. Considere un $\delta = 2$.
- (c) Genere la variable Y sin el efecto de los bloques. Use como varianza del error el CMRes obtenido anteriormente.
- (d) Genere la variable de bloques `B` de tal forma que en cada bloque haya 2 observaciones, una por cada tratamiento. Puede usar `factor(rep(1:6, 2))`.
- (e) Genere los efectos de bloques de una distribución normal con una varianza de bloques. Puede usar `vb=3`. Como son 6 bloques debe generar 6 efectos y repetirlos con la misma estructura de `B`; guarde el vector en `efb`.
- (f) Sume `efb` a la variable Y original para obtener la variable Y final.

3. Función para calcular la potencia:

- (a) Cree una función llamada `fun2` que devuelva la probabilidad en el anova de datos generados. La función que va a crear debe tener como argumentos: el número de bloques (`b`), el vector de medias verdaderas (`mu`), la variancia del error (`v`) y la varianza de bloques (`vb`). La función debe crear el vector del tratamiento según el número de bloques y niveles que tenga el factor, además debe crear la respuesta según el modelo que parte de esas medias verdaderas y la variancia del error.
- (b) Use la función `fun2` en un ciclo de 1000 iteraciones con $\delta = 3$, $v = 6,5$, $vb = 10$, y diferentes valores de b que vayan desde $b = 3$ hasta $b = 10$. En cada caso almacene la potencia.
- (c) Haga un gráfico para visualizar la evolución de la potencia según el número de bloques. Identifique el número de bloques necesarios para alcanzar una potencia de 0.9.

7.4.2 Solución

1. Valores observados a partir de los datos de burbujas:

(a) Promedios muestrales:

```
load("burbujas.Rdata")
tapply(base$tiempo,base$glicerina,mean)

##    60    120
## 6.99 11.80
```

(b) CMRes:

```
base1$glicerina=factor(base1$glicerina)
base1$persona=factor(base1$persona)
mod1=lm(tiempo~glicerina+persona,data=base1)
(cmres=anova(mod1) [3,3])

## 6.06
```

2. Preparación para simulación:

(a) Variable X:

```
(X = factor(rep(1:2,each=6)))

## 1 1 1 1 1 1 2 2 2 2 2 2
## Levels: 1 2
```

(b) Vector de promedios:

```
(muj = rep(c(7,9),each=6))

## 7 7 7 7 7 7 9 9 9 9 9 9
```

(c) Variable Y original:

```
s = sqrt(cmres)
n = 2*6
(Y1 = rnorm(n,muj,s))

## 9.68 5.12 4.96 9.05 4.62 6.93 9.57 8.26 7.33 10.61 8.01 8.18
```

(d) Variable de bloques:

```
(B=factor(rep(1:6,2)))

## 1 2 3 4 5 6 1 2 3 4 5 6
## Levels: 1 2 3 4 5 6
```

(e) Vector de efectos de bloques:

```
vb=3
sb=sqrt(vb)
eb=rnorm(6,0,sb)
(efb=rep(eb,2))

## 2.37 3.70 0.88 1.36 -1.56 0.92 2.37 3.70 0.88 1.36 -1.56 0.92
```

(f) Variable Y final:

```
(Y=Y1+efb)

## 12.05 8.82 5.84 10.42 3.05 7.85 11.94 11.96 8.21 11.98 6.45 9.10
```

3. Función para calcular la potencia:

(a) Función:

```
fun2 = function (b, mu, v, vb) {
  k = length(mu)
  n = b * k
  X = factor(rep(1:k,each=b))
  muj = rep(mu, each=b)
  s = sqrt(v)
  Y1 = rnorm(n, muj, s)
  B=factor(rep(1:b,k))
  sb=sqrt(vb)
  eb=rnorm(b,0,sb)
  efb=rep(eb,k)
  Y=Y1+efb

  mod = aov(Y ~ X+B)
  p = anova(mod)[1, 5]
  return(p)
}
```

(b) Potencia:

```
mu=c(7,10)
v=6.5
vb=10
b=3:20
h=length(b)
M=1000
prob=matrix(nrow=1000,ncol=h)

for(j in 1:h){
  for(i in 1:M) prob[i,j]=fun2(b[j],mu,v,vb)
}
```

```
(pot=apply(prob<0.05,2,mean))  
  
## 0.14 0.21 0.27 0.37 0.49 0.51 0.60 0.66 0.73  
## 0.75 0.78 0.83 0.84 0.88 0.90 0.91 0.92 0.95
```

(c) Gráfico:

```
plot(b,pot,type="l")  
abline(h=0.9,lty=2)  
abline(v=17,lty=2)
```

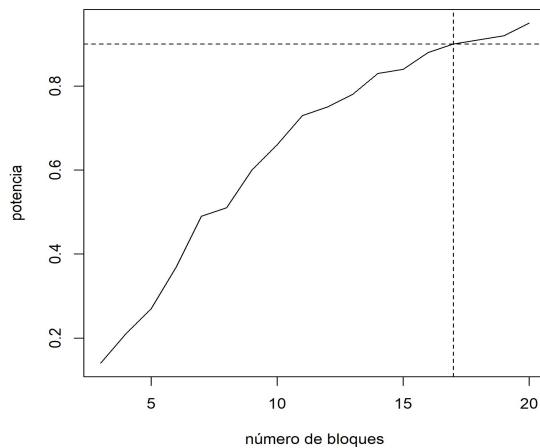


Figura 7.2: Potencia obtenida en una prueba al variar el número de bloques

Se requieren al menos 17 bloques para lograr una probabilidad de 0.90 de concluir que las medias no son iguales si los promedios verdaderos difieren en al menos 3 minutos.

Anexo

Glosario de funciones de R

Función	Uso
abline	Agrega una o varias líneas rectas al gráfico actual.
anova	Calcula la tabla de análisis de variancia para un modelo ajustado previamente con lm o aov.
aov	Ajusta un modelo de que tiene factores.
bartlett.test	Ejecuta la prueba de Bartlett sobre la igualdad de las varianzas en todos los tratamientos.
boxplot	Produce un gráfico de cajas de una variable numérica contra uno o más factores.
cbind	Toma un secuencia de vectores o matrices y los combina por columnas.
colnames	Asigna nombres a las columnas de una matriz o extrae los nombres de las columnas de una matriz.
confint	Calcula intervalos de confianza para los parámetros de un modelo previamente estimado.
cor	Calcula la correlación entre dos variables numéricas o calcula la matrix de correlación para un conjunto de variables numéricas.
contrasts	Permite observar los valores que se han dado a cada nivel de un factor en las variables auxiliares.
curve	Dibuja una curva producida por una función en una variable llamada x.
diag	Extrae la diagonal de una matriz cuadrada o sirve para construir una matriz diagonal.

Función	Uso
drop1	Estima todos los modelos que se pueden obtener a partir del modelo inicial eliminando solo un término, y hace una tabla de resumen con los estadísticos producidos al comparar el modelo inicial con cada uno de estos nuevos modelos.
emmeans	Realiza comparaciones múltiples para modelos de varios factores, se puede incluir interacción y realizar corrección de Bonferroni.
exp	Calcula la función exponencial.
factor	Convierte un vector en una variable de tipo factor o categórica.
format	Permite dar formato a los números para que se impriman de forma agradable. Por ejemplo, para eliminar la notación científica se usa: <code>format(round(a, 3), scientific=F)</code> .
ggplot	Sirve para crear objetos en capas para graficarlos con la librería <code>ggplot2</code> .
glm	Ajusta un modelo lineal generalizado.
length	Devuelve el largo o número de elementos de un vector.
levels	Permite obtener los nombres que tiene cada nivel de un factor o permite asignar nuevos nombres a los niveles de un factor.
lm	Ajusta un modelo lineal. Puede usarse para regresión lineal o para modelos con factores.
load	Carga una base de datos previamente salvada con la función <code>save</code> con extensión <code>.Rdata</code> .
log	Calcula el logaritmo natural.
mean	Calcula el promedio.
model.matrix	Devuelve la matriz de estructura de un modelo ajustado previamente con <code>lm</code> o <code>aov</code> .
model.tables	Devuelve las estimaciones de los efectos de un modelo con factores.
names	Devuelve los nombres de los componentes de un objeto. Cuando el objeto es una base de datos, devuelve los nombres de las columnas.

Función	Uso
options	Permite pasar del modelo de suma nula al modelo de tratamiento referencia y viceversa. Para pasar al modelo de suma nula se usa: <code>options(contrasts=c("contr.sum", "contr.poly")).</code> Para pasar al modelo de tratamiento referencia se usa: <code>options(contrasts=c("contr.treatment", "contr.poly")).</code>
pf	Calcula la probabilidad acumulada asociada a la distribución F, dados los grados de libertad.
pnorm	Calcula la probabilidad acumulada asociada a la distribución normal, dada la media y la desviación estándar.
points	Agrega puntos en las coordenadas indicadas en el gráfico actual.
pt	Calcula la probabilidad acumulada asociada a la distribución t-student, dados los grados de libertad.
ptukey	Calcula la probabilidad asociada a las pruebas de Tukey, dado el número de medias a comparar y los grados de libertad del error.
qnorm	Calcula el cuantil asociado a la distribución normal estándar, dada una probabilidad.
qt	Calcula el cuantil asociado a la distribución t-student, dada una probabilidad y los grados de libertad.
predict	Permite obtener las predicciones basadas en un modelo previamente ajustado.
read.csv	Lee una tabla de datos en formato .csv y crea un dataframe.
rnorm	Obtiene un conjunto de números aleatorios provenientes de una distribución normal, dada la media y la desviación estándar.
round	Redondea un conjunto de números a un cierto número de decimales que se especifican.
row.names	Asigna nombres a las filas de una matriz o extrae los nombres de las filas de una matriz.

Función	Uso
save	Guarda uno o varios objetos en un archivo externo con extensión .Rdata que luego pueden ser cargados con la función load.
scatterplot	Sirve para crear gráficos del tipo de diagrama de dispersión con la librería car, y agregar líneas suavizadas para evaluar la linealidad entre dos variables.
str	Devuelve una descripción compacta de la estructura interna de un objeto. Es útil para conocer el tipo de variables que conforman una base de datos.
sum	Calcula la suma de un conjunto de datos.
summarise	Crea variables nuevas a partir de una base de datos con la librería dplyr. Es útil para obtener estadísticos por grupos.
summary	Devuelve una tabla de resumen con los resultados del ajuste de un modelo, también sirve obtener estadísticos básicos de las variables de una base de datos.
t	Devuelve la transpuesta de una matriz.
table	Genera una tabla con el número de observaciones por nivel de un factor.
tapply	Permite obtener estadísticos para una variable numérica, para cada nivel de un factor. El uso es: tapply(y, factor, estadístico).
title	Sirve para agregar títulos y leyendas con alguna notación matemática básica al gráfico actual.
vcov	Devuelve la matriz de covarianza de los parámetros de un modelo ajustado previamente.
xyplot	Sirve para crear gráficos de alto nivel del tipo de diagrama de dispersión con la librería lattice. Permite agregar fácilmente información de factores para ilustrarla en el gráfico.
%*%	Sirve para realizar multiplicación de matrices.

Bibliografía

Abarca Araya, E., Guillén Amador, E. y Torres Rojas, R. (2015). Efecto de la forma de salida y el tipo de calentamiento en el tiempo de recorrido de la carrera de 100 metros planos. s.f.

Araya Cárdenas, E., Castro Solís, M.J. y Zúñiga Baldí, A. (2015). Efecto de la proporción de glicerina y el tipo de agua en la resistencia de burbujas de jabón. s.f.

Box, G.E.P., Hunter, W.G. y Hunter, J.S. (1978). Statistics for experimenters. New York, EU: John Wiley Sons.

Campbell, D. y Stanley, J. (1973). Diseños experimentales y cuasi-experimentales en la investigación social. Buenos Aires, Argentina: Amorrortu.

Champely, S. (2018). pwr: Basic Functions for Power Analysis. R package version 1.2-2. <https://CRAN.R-project.org/package=pwr>.

Cochran, W.G y Cox, G.M. (1973). Diseños Experimentales. México DF, México: Trillas.

Dean, A. y Voss, D. (1999). Design and Analysis of Experiments. New York, EU: Springer-Verlag.

Fox, J. y Weisberg, S. (2019). An R Companion to Applied Regression, Third Ed. Thousand Oaks, EU: Sage. URL: <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>

Kuehl, R.O. (2013). Diseño de Experimentos (2da. ed). México DF, México: Thomson Learning.

Kutner, M.H., Li, W., Nachtsheim, C.J., y Neter, J. (2005). Applied Linear Statistical

- Models (5ta. ed). Irwin, EU: WCB McGraw-Hill.
- Mandal, B.N. (2019). *ibd*: Incomplete Block Designs. R package version 1.5. <https://CRAN.R-project.org/package=ibd>
- Montgomery, D.C. (2005). Diseño y análisis de experimentos (2da. ed.). México DF, México: Limusa.
- Ramsey, F.L. y D.W. Schafer (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury, Australia: Thomson Learning.
- R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL <https://www.R-project.org/>.
- Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York, EU: Springer. ISBN 978-0-387-75968-5
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, EU: Springer-Verlag.
- Wickham, H., Francois, R., Henry, L. y Müller, K. (2019). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3. <https://CRAN.R-project.org/package=dplyr>

Índice de cuadros

1.1	Variables auxiliares para un modelo de un factor con restricción de suma nula	14
1.2	Variables auxiliares para un modelo de un factor con tratamiento 1 de referencia	15
3.1	Variables auxiliares para un diseño de dos factores	62
7.1	Generación de datos para un diseño con bloques	197

Índice de figuras

1.1	Distribución de una variable con 4 tratamientos en dos situaciones	12
1.2	Muestras de una variable con 4 tratamientos en dos situaciones	16
1.3	Dos situaciones con igual separación entre medias	20
1.4	Puntajes de color por tratamiento	25
2.1	Grados Brix por localidad	43
3.1	Diseño con dos factores en dos situaciones	59
3.2	Diseño con dos factores sin interacción y el efecto marginal del factor A	60
3.3	Promedios de proteína para combinaciones de condición y género	72
3.4	Proteína por tratamiento (condición y género)	74
3.5	Promedios de proteína para combinaciones de condición y especie	86
3.6	Proteína por tratamiento (condición y especie)	88
4.1	Tres factores que interactúan	100
4.2	Promedios de desviación absoluta para combinaciones de los factores .	108
4.3	Promedios de desviación absoluta para combinaciones de carbonatación y presión según niveles de rapidez	111
5.1	Diseño de un factor con bloques	123
5.2	Datos centrados para eliminar el efecto del bloque	124
5.3	Esquema de un diseño de parcelas divididas	127
5.4	Inicio de un diseño de bloques incompletos	129
5.5	Pasos de un diseño de bloques incompletos	129
5.6	Diseño terminado de bloques incompletos	130
5.7	Tiempo por persona según niveles de glicerina	137
5.8	Tiempo según niveles de glicerina con datos originales y datos centrados	138

5.9	Tiempo centrado por persona según nivel de glicerina y tipo de agua .	142
5.10	Tiempo centrado por persona según tipo de agua y nivel de glicerina .	142
5.11	Tiempo por tipo de agua según nivel de glicerina	143
5.12	Esquema de un diseño de parcelas divididas	146
5.13	Tiempo centrado por persona según nivel de glicerina y nivel aeróbico	146
5.14	Resistencia según concentración con datos originales y datos centrados	150
6.1	Diseño con un factor y una covariable (datos completos)	159
6.2	Diseño con un factor y una covariable (datos en un intervalo)	159
6.3	Tiempo contra índice de masa corporal según tipo de calentamiento y salida	166
6.4	Tiempo por tipo de calentamiento y salida	167
6.5	Tiempo contra índice de masa corporal según tipo de calentamiento y salida	168
6.6	Módulo por combinación de antioxidante, fuente y envejecimiento .	182
6.7	Módulo contra temperatura según antioxidante	183
6.8	Módulo contra combinaciones de antioxidante con fuente y envejecimiento y con temperatura	184
6.9	Módulo contra envejecimiento según antioxidante	187
7.1	Distribuciones hipotéticas	194
7.2	Potencia obtenida en una prueba al variar el número de bloques . . .	213

Índice alfabético

Análisis de varianza, **15**

Bloques

Bloques incompletos, **128**

Definición, **121**

Diseño, **124**

Dos factores, **126**

Muestras pareadas, **125**

Parcelas divididas, **126**

Un factor con dos niveles, **125**

Coeficiente, **13**

Comparaciones simultáneas, **35**

Contrastes, **37**

Corrección de Bonferroni, **37**

Cota inferior, **37**

Cuadrado medio

De tratamiento, **18**

Residual, **18**

Datos centrados, **123**

Diferencia relevante, **37, 193**

Diseño

Bloques incompletos, **128**

Con bloques, **121**

Con dos factores, **57**

Con tres factores, **99**

De una vía, **12**

Parcelas divididas, **127**

Distribución

F, 19

Hipotética, **194**

Normal, **127, 196**

t, 19, 125

Efecto

De bloque, **123**

De interacción, **62**

Definición, **12**

Error

Error estándar, **36**

Tipo I, **16**

Tipo II, **16, 194**

Estadístico F, **19**

Factor, **11**

Grados de libertad, **18**

Hipótesis nula, **16**

Interacción

Doble, **57**

Efecto de interacción, **62**

Modelo con interacción, **62**

Modelo sin interacción, **59**

Triple, **99**

Intervalos de confianza simultáneos, **36**

Media general, **13**

Modelo

- Con covariables, **157**
- Con bloques, **122**
- Con un factor, **12**
- Mixto, **125**
- Sin interacción, **59**
- Suma nula, **13**
- Tratamiento referencia, **14**

Muestras pareadas, 125**Número de réplicas, 195****Nivel de significancia, 19****Nivel de un factor, 11****Parcelas divididas, 126****Potencia, 194****Respuesta, 11****Simulación, 195**

- Cálculo de la potencia, **195**

- Distribución condicional, **196**

- Escenarios, **195**

- Iteración, **195**

- Parámetros, **195**

Suma de cuadrados

- De regresión marginal, **159**

- De tratamiento, **17**

- Residual, **17**

- Total, **17**

Tratamiento, 11**Tukey, 35****Variable auxiliar, 13**