Trabajo Fin de Grado

Introducción al diseño de experimentos

Alberto Vilaboa Fernández

2021/2022

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo Fin de Grado

Introducción al diseño de experimentos

Alberto Vilaboa Fernández

Febrero 2022

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa
Título: Introducción al diseño de experimentos
Breve descrición del contenido
El diseño de experimentos es un campo de estudio dentro del ámbito estadístico con numerosas aplicaciones en ámbitos aplicados como la biología. En este contexto, los "modelos" del diseño de experimentos tienen como objetivo conocer qué factor o factores influyen (y en qué medida) en una determinada variable de interés para poder diseñar de forma adecuada el experimento a realizar. En este trabajo se revisarán muy brevemente distintos tipos de diseño a considerar (diseños factoriales) y los modelos clásicos del diseño de experimentos (ANOVA de una y dos vías). Primeramente, se introducirá la formulación de los modelos. La descripción de los diseños se acompañará de ejemplos prácticos, que serán resueltos en R, acompañando el trabajo de una revisión de las funciones disponibles para la realización de estas tareas.
Recomendaciones
Otras observaciones

Índice general

Resumen

En este trabajo introducimos los conceptos de diseño de experimentos y modelo para luego exponer una descripción del funcionamiento del modelo ANOVA de una y dos vías y complementar esta descripción con el estudio de un ejemplo práctico.

El trabajo empieza con la descripción y análisis del modelo ANOVA de una vía. Se hará uso del software *R* para obtener un ejemplo y así poder ayudar a la comprensión del modelo. Una vez expuesto el modelo ANOVA de una vía, se procederá a modelar dicho ejemplo hasta conseguir crear un ANOVA de dos vías, en ese momento explicaremos y detallaremos la teoría de este nuevo modelo y, para acabar, volveremos al ejemplo para aplicar la nueva teoría a la práctica.

Seguiremos este orden, ya que se pretende mostrar el uso práctico del modelo ANOVA de dos vías antes de exponer la teoría, es decir, poner de manifiesto la necesidad de la teoría para poder avanzar en la práctica.

Abstract

In this work we introduce the design of experiments and the idea of model, then we describe the one-way and two-way ANOVA model, and complement the description of the model with a practical example.

The work starts with the description and analysis of the one way ANOVA model. We will use the *R* software to provide an example to the explanation of the model. Once we have seen the one way ANOVA model, we will revise the example to introduce a two-way ANOVA model and its theoretical development will be explain next. Finally, we will come back to the example.

We follow this method in order to show the practical performance of the two-way ANOVA model before the study of the theory, so we are able to see how useful is the theory in the practice.

Introducción

Introducción al diseño de experimentos, éste es el título del trabajo y será el tema de estudio a lo largo de sus páginas. Antes de empezar, tenemos que definir algunos términos: ¿Qué entendemos por diseño de experimentos?, ¿Qué es un modelo?, ¿En qué modelos vamos a centrarnos?

- El **diseño de experimentos** es un campo de estudio dentro del ámbito de la estadística que consiste en la construcción de modelos para ver cómo afectan unas variables sobre otras, ya sean de tipo categórica o continua. Los diseños se clasifican según el número de variables del experimento y el tipo de dichas variables, de ahí se sacan los diseños factoriales, diseños 2^n, 3^n, ...

  Para estudiar correctamente la relación entre unas variables, se necesita saber cuales serán los objetivos de dicho estudio, cómo se realizará la formación de la muestra (tamaño, grupos,...) y la construcción del modelo a partir de dicha muestra, cómo ejecutaremos y validaremos todo este proceso para saber que hemos seguido los pasos adecuadamente y, finalmente, ser capaces de obtener unos resultados concluyentes. El diseño de experimentos consiste en la planificación de dicho estudio, es la base para realizar de forma adecuada cualquier experimento en el que trabajemos con distintas variables.

- Los **modelos** del diseño de experimentos son las ecuaciones que se siguen para ver la relación entre las variables. En este trabajo trabajaremos con **modelos lineales generales**, que son las ecuaciones que cumplen una relación lineal entre las variables.

- El **análisis de la varianza** o **modelo ANOVA** es un caso de modelo lineal general en el que hay una única variable explicativa categórica y la variable respuesta es continua. A lo largo de este trabajo se estudiará el modelo ANOVA de una vía, veremos cómo funciona y cómo, al añadir otra variable explicativa categórica, nuestro modelo ANOVA de una vía se convierte en un modelo ANOVA de dos vías. Este nuevo

modelo supone un avance en el diseño de experimentos, ya que nos permite estudiar la variable respuesta en función de no solo una, sino de dos variables explicativas.

A la hora de analizar los resultados de cualquier experimento, nos veremos en la obligación de tratar con tres tipos diferentes de variabilidad.

- La variabilidad sistemática, será la que trataremos de concretar con nuestro experimento, dicha variabilidad será resuelta en las conclusiones del experimento.
- La variabilidad natural, es aquella que no podemos controlar y asumimos, se neutraliza aleatorizando todo factor que no podamos tener bajo control, así, no se condiciona el experimento.
- La variabilidad no planificada, será la que condicione el experimento sin nosotros pretenderlo, puede dar unos resultados sesgados.

Una vez definidos los términos principales y los tipos de variabilidad podemos explicar en qué circunstancias es adecuado utilizar un modelo ANOVA (modelo de análisis de la varianza que se estudiará en detalle en el Capítulo1) en un experimento. Supongamos que nos encontramos analizando una variable continua y detectamos ciertas anomalías en sus valores, en ese momento, clasificamos las distintas fuentes de variación que se creen que pueden influir en la variable continua. A continuación, se detecta que hay un factor en concreto que puede ser el causante de dichas anomalías y además dicho factor es categórico, es decir, se intuye que las anomalías van a ser provocadas por los distintos grupos del factor. Por lo tanto, lo que se necesita ahora es una forma de relacionar la variable continua con los distintos grupos del factor categórico, es este el momento en el que recurrimos al modelo ANOVA para concluir si dicho factor es el causante de las anomalías en la continua.

Para ayudar a la comprensión de la explicación anterior vamos a describir un ejemplo teórico. Supongamos que tenemos un estudio de tipo médico en el que se quiere estudiar el efecto de un medicamento en una población, en concreto, se quiere conocer si distintas cantidades de medicamento tienen la misma efectividad en los sujetos. Veremos cómo se construye el experimento.

Se realizan pruebas con I cantidades distintas de medicamento en una población aleatoria. Se realiza el número de pruebas que queramos con cada cantidad, las suficientes como para tener datos concluyentes, así, obtendremos los datos de efectividad para cada una de las cantidades. Una vez con los datos recogidos y las medias de efectividad para cada medicamento hechas, tenemos que ver si se puede asumir que todas las cantidades de medicamento tienen la misma efectividad. Nuestro experimento consta de una variable explicativa discreta (cantidades de medicamento) y una variable respuesta (efectividad),

por lo que podemos construir un modelo ANOVA de una vía. De esta forma, realizando el test F del modelo, concluímos si se puede asumir que todas las cantidades tienen la misma media de efectividad, o por el contrario, existe un cantidad en concreto con una efectividad mayor (o menor).

Después de hacer el modelo ANOVA de una vía, nos fijamos que en la muestra del estudio no se tiene en cuenta las cualidades de cada uno de los sujetos (tamaño, sexo o condición) y queremos ver si hay diferencias entre la población general, en concreto, estudiaremos el efecto según el sexo del sujeto, ya que intuímos que la eficacia del medicamento va a ser distinta. Segregamos el experimento en $Hombres$ y $Mujeres$, por lo que necesitamos construir un modelo nuevo en el que se tenga en cuenta la cantidad de medicamento aplicada y el sexo del sujeto. De esta manera, tenemos nuestro modelo ANOVA de dos vías, donde las variables explicativas serán la cantidad de medicamento aplicada y el sexo del sujeto.

En este ejemplo podemos identificar cada tipo de variabilidad. La variabilidad sistemática serán los cambios de efectividad de cada uno de los sujetos al cambiar la cantidad de medicamento aplicado. La variabilidad natural se asume, por ejemplo, en los sujetos que van a presentar una efectividad significativamente pequeña o grande. La variabilidad no planificada puede ser un factor que no tengamos en cuenta como el historial médico de los sujetos.

En el Capítulo 1 expondremos la teoría del modelo ANOVA de una vía y mostraremos su comportamiento en la práctica con un ejemplo. En el ejemplo, se expondrán gráficas y datos explicando cómo funciona la respuesta en función de la variable explicativa. En el Capítulo 2, añadiremos otra variable al ejemplo, para mostrar cómo sería el modelo ANOVA de dos vías. Después, se construirá la teoría del modelo ANOVA de dos vías y terminaremos volviendo al ejemplo para aplicar la nueva teoría. A mayores, realizaremos unas simulaciones para conocer mejor la sensibilidad del modelo a factores externos como por ejemplo el tamaño de los grupos, el diseño y escenarios de simulación fueron creados por mi, no se usaron programas o referencias externas. Para la realización de las gráficas y datos del ejemplo, nos apoyaremos en [5] y [1].

Capítulo 1

Introducción del modelo ANOVA

En este capítulo, vamos a introducir el modelo ANOVA de una vía, veremos la parte teórica y luego su uso práctico con un ejemplo. Gran parte de la teoría que se ve en este capítulo proviene de [3] y [1] y de los apuntes de la asignatura *Modelos de regresión y análisis multivariante.*

1.1. Revisión del modelo lineal general

Ya sea en el diseño de experimentos o cualquier tipo de estudio en el que aparezcan distintas variables, se necesita un modelo para entender cómo se relacionan, una ecuación para escribir una variable en función de las demás. Primero de todo, hay que definir cual va a ser la variable fuente de estudio o *variable respuesta* Y y luego, el conjunto de $I-1$ variables que van a definir la variable respuesta, también llamadas *variables explicativas* $X_1, ..., X_{I-1}$, sumando las $I-1$ variables a nuestra respuesta obtenemos un total de I variables. Al modelo de regresión lineal con una única variable explicativa se le llama modelo de regresión lineal simple y, al que tiene más de una variable, se le llama modelo de regresión lineal múltiple. A continuación, escribimos la ecuación del modelo de regresión lineal múltiple.

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{I-1} X_{I-1} + \epsilon$$

Los parámetros $\beta_0, \beta_1, \ldots, \beta_{I-1}$ son los valores asociados a cada una de las $I-1$ variables explicativas, las estimaciones de dichos parámetros, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{I-1}$ representan los incrementos de las variables sobre la respuesta. Además, ϵ representa el error natural del modelo, se supone que cumple las hipótesis de homocedasticidad, normalidad e independencia, es decir, si tenemos una muestra de n elementos:

$$\epsilon_1, \ldots, \epsilon_n \in N(0, \sigma^2) \qquad \text{independientes.}$$

La forma estándar del modelo lineal general es la ecuación siguiente.

$$Y = \mathbf{X}\beta + \epsilon$$

A continuación, escribimos el modelo de regresión lineal múltiple en forma matricial,

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,I-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,I-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{I-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

donde cada fila de la matriz $\mathbf{X}$ representa cada una de las n observaciones y cada una de las columnas recoge las observaciones de cada una de las variables (excepto la primera que contiene 1, para incluir el intercepto).

1.2. Descripción del modelo ANOVA

El modelo de análisis de la varianza o **modelo ANOVA de una vía** es un tipo de modelo lineal general en el que hay una única variable explicativa X categórica y la variable respuesta Y es continua. Vamos a descomponer este modelo.

- X es una variable categórica, por lo que cada uno de los individuos de la muestra va a pertenecer a uno y solo a uno de los grupos de la variable X. Por ejemplo, una variable categórica sería agrupar una muestra de personas por sexos en $Hombre$ o $Mujer$. En el resto del capítulo vamos a trabajar con I grupos en la variable X.

- Y es una variable continua y va a tener tantas observaciones como tamaño de la muestra se trabaje. Pongamos por ejemplo que la muestra consta de n elementos, y que cada elemento pertenece a un solo grupo, entonces n va a ser igual a la suma de elementos de cada uno de los grupos, es decir,

  $$n = \sum_{i=1}^{I} n_i$$

  donde n_i es el número de elementos del grupo i con $i = 1, \ldots, I$.

Este tipo de modelo tiene una descomposición de la variabilidad muy vistosa.

$$\begin{array}{cccccc} Y_{11} & Y_{12} & \cdots & Y_{1n_1} & \text{de una población} & N(\mu_1, \sigma^2) \\ Y_{21} & Y_{22} & \cdots & Y_{2n_2} & \text{de una población} & N(\mu_2, \sigma^2) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ Y_{I1} & Y_{I2} & \cdots & Y_{In_I} & \text{de una población} & N(\mu_I, \sigma^2) \end{array}$$

Cada una de las observaciones es independiente, tanto dentro de sus grupos como entre grupos. Además, dentro de cada grupo, las observaciones siguen una distribución normal con la misma varianza. Todas estas hipótesis se resumen en el modelo (1.1).

$$\begin{cases} Y_{ij} = \mu_i + \epsilon_{ij} & i = 1, \ldots, I \qquad j = 1, \cdots, n_i \\ \epsilon_{ij} \sim N(0, \sigma) \\ \mu_i = E[Y_{ij} | X = i] \end{cases} \tag{1.1}$$

El modelo (1.1) se puede escribir en forma matricial como sigue.

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ \vdots & \cdots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & 0 & 0 \\ \vdots & \cdots & \cdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{I1} \\ \vdots \\ \epsilon_{In_I} \end{pmatrix}$$

Podemos reconstruir el modelo (1.1) siguiendo dos parametrizaciones alternativas: en función del intercepto y en función de un grupo de referencia. Primero mostramos el modelo en función del intercepto, (1.2). Esta nueva parametrización la usaremos en el Capítulo 2 para construir el modelo ANOVA de dos vías.

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \epsilon_{ij} & j = 1, \ldots, n_i \qquad i = 1, \ldots, I \\ \epsilon_{ij} \sim N(0, \sigma^2) \end{cases} \tag{1.2}$$

Aquí, μ es el llamado intercepto y mide el nivel promedio de respuesta para todas las observaciones, α_i mide el efecto incremental del grupo $i = 1, \ldots, I$ respecto al nivel promedio. Estimando el intercepto como la media de todas observaciones se cumple,

$$\sum_{i=1}^{I} \alpha_i = 0.$$

La razón por la que se usará el modelo (1.1) en el ANOVA de una vía es que será mucho más fácil estimar los valores μ_i con $i = 1, \ldots, I$ cuando tengamos una única variable explicativa.

Por último, vamos a describir la parametrización en función de un grupo de referencia, (1.3). Se considera como intercepto el nivel promedio respecto de las observaciones de un grupo, no de todas las observaciones como se hacía en la parametrización anterior. Usaremos esta parametrización a la hora de mostrar las estimaciones en el ejemplo, ya que el software R muestra los resultados del modelo de esta forma.

$$\begin{cases} Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij} & j = 1, \dots, n_i \qquad i = 2, \dots, I \\ & \epsilon_{ij} \sim N(0, \sigma^2) \end{cases} \tag{1.3}$$

En la formulación (1.3) se muestra usando el grupo 1 como ejemplo, se podría hacer con cada uno de los grupos y cada parametrización daría resultados diferentes.

1.3. Descomposición de la variabilidad. El test F

Estimar un modelo, en este caso el (1.1), consiste en ver si constituye una descripción adecuada del sistema, si los parámetros estimados representan adecuadamente el efecto de las variables explicativas sobre la respuesta.

El objetivo principal del modelo ANOVA es corroborar o descartar la hipótesis de que las medias de todos los grupos son idénticas, es decir $\mu_1 = \mu_2 = \cdots = \mu_I$. Por un lado, si esta hipótesis es cierta, entonces es irrelevante a qué grupo pertenece cada observación ($Y \sim 1$), es decir, resultará que Y tiene una distribución normal con media $\mu = \mu_1 = \cdots = \mu_I$ y varianza σ^2 (ya que suponemos que todos los grupos tienen la misma varianza). Esto quiere decir que la variable explicativa no aportará ninguna información relevante sobre la respuesta. Por el otro lado, si rechazamos la hipótesis $\mu_1 = \mu_2 = \cdots = \mu_I$, entonces sí que tendrá sentido estudiar el efecto de X sobre Y ($Y \sim X$). Vamos a formalizar estas ideas.

LLamamos hipótesis nula (H_0) a la que supone que las medias de los grupos son iguales y llamamos hipótesis alternativa (H_a) a la que supone que existen al menos dos medias que no son iguales. La combinación de la hipótesis nula y la alternativa se llama contraste. Lo mostramos a continuación, en la formulación (1.4):

$$\begin{cases} H_0 : Y \sim 1 \\ H_a : Y \sim X \end{cases} \tag{1.4}$$

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \\ H_a : \exists \ k, l \ \text{ tal que } \ \mu_k \neq \mu_l \end{cases}$$

Realidad/Decisión	Aceptar	Rechazar
H_0 cierta	Correcto	Error tipo I
H_0 falsa	Error tipo II	Correcto

Tabla 1.1: Tabla de los errores.

Se llama **Error de tipo I** al error que se comete cuando se rechaza H_0 siendo ésta cierta y **Error de tipo II** el que se comete cuando se acepta H_0 siendo ésta falsa. Para estimar estos errores se necesita:

- **Nivel de significación**: α[1], es la probabilidad de cometer un error de tipo I.
- **Potencia**: es la probabilidad de rechazar H_0 siendo ésta falsa, es decir $1-\omega$, siendo ω la probabilidad de cometer un error de tipo II.

El test F es el proceso de construcción de un estadístico de contraste T para el contraste (1.4) y el proceso de comparar dicho estadístico en la distribución F de *Snedecor* para saber si rechazamos o aceptamos H_0.

Antes de ver la construcción del estadístico de contraste T y de ver cómo se hace la comparación en el test F, hay que ver cómo se manejan los datos que disponemos, para ello, hay que introducir primero algunos términos:

$$RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \qquad RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2$$

siendo:

$$\bar{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \qquad \forall i \in \{1,\dots,I\}$$

$$\bar{Y}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{I} n_i \bar{Y}_{i\bullet}.$$

Vamos a explicar qué significan estos términos y porqué son importantes.

- $\boldsymbol{RSS_0}$: la suma residual de cuadrados bajo la hipótesis nula, es decir, cuanto se alejan cada una de las observaciones respecto de su media global.

[1]El nivel de significación se denota por α, no debe confundirse con los α_i, que siempre irán acompañados de subíndice

- ***RSS***: la suma residual de cuadrados bajo la hipótesis alternativa, es decir, cuanto se alejan cada una de las observaciones respecto la media del grupo al que pertenecen.

Estos términos representan la variabilidad del modelo, si hacemos la diferencia ($RSS_0 - RSS$) nos sale la variabilidad que el modelo puede explicar. Esto se ve condicionado por el tamaño de la muestra y el número de grupos que tengamos, ya que cuanto mayor sea la muestra y menor el número de grupos, vamos a tener una estimación mucho mejor del modelo.

De esta forma, la variabilidad admite la siguiente descomposición.

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i\bullet}-\bar{Y}_{\bullet\bullet})^2 \tag{1.5}$$

Fuente de variación	Suma de cuadrados	Grados de libertad
Entre grupos	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i\bullet}-\bar{Y}_{\bullet\bullet})^2$	$I-1$
Error	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2$	$n-I$
Total	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2$	$n-1$

Tabla 1.2: Tabla de análisis de la varianza.

A partir de dicha descomposición (1.5), escribimos la tabla de análisis de varianza para el modelo ANOVA de una vía, Tabla 1.2 . Ahora, solo nos queda relacionar todo esto con la distribución F de Snedecor, necesitamos recurrir a la teoría. En las condiciones del modelo lineal general, si la hipótesis nula H_0 es cierta, entonces el estadístico de contraste T cumplirá la relación que mostramos a continuación, siendo F_{m_1,m_2} la distribución F de *Snedecor* con m_1 y m_2 grados de libertad y χ_m la distribución *Chi cuadrado* con m grados de libertad.

$$\frac{\chi_1/m_1}{\chi_2/m_2} \in F_{m_1,m_2}$$

$$T = \frac{(RSS_0 - RSS)/(I-1)}{RSS/(n-I)} \in F_{(I-1),(n-I)}$$

Por lo tanto, ya tenemos todos los ingredientes para construir el estadístico T.

$$T = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i\bullet}-\bar{Y}_{\bullet\bullet})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2/(n-I)} \in F_{(I-1),(n-I)} \tag{1.6}$$

Este estadístico crece cuanto mayor sea la diferencia de la media de los grupos con la media total, es decir, cuanta mayor sea la diferencia entre los grupos, mayor será el estadístico.

Para terminar, fijado un nivel de significación $\alpha \in (0,1)$ y siendo f_α el cuantil que deja a su derecha una probabilidad α en la distribución $F_{(I-1),(n-I)}$, rechazaremos H_0 cuando el estadístico T sea más grande que f_α.

1.4. Validación del modelo

Vamos a ver qué procesos seguiríamos para estudiar la validación del modelo, es decir, comprobar que se cumplen las hipótesis de **normalidad, homocedasticidad** e **independencia**. Antes de continuar, repasemos las definiciones de estos tres términos.

- Homocedasticidad : es la igualdad de las varianzas σ^2 de cada uno de los i grupos de la variable X, con $i = 1, \ldots, I$.
- Normalidad : cada uno de los I grupos de X está formado por observaciones que provienen de una distribución normal de media μ_i y varianza σ^2, $N(\mu, \sigma^2)$.
- Independencia : cada uno de los I grupos, y las observaciones dentro de cada grupo, son independientes entre sí.

Vamos a estudiar la homocedasticidad usando el test Levene, para ver si son iguales las varianzas de los grupos. Se considera la hipótesis nula en la que se cumple dicha igualdad, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$, y una hipótesis alternativa en la que existen al menos dos varianzas que no son iguales. Una vez construído el contraste, se realiza el test F siguiendo los pasos vistos en la sección 1.3 .

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 \\ H_a : \exists \ k, l \ \text{ tal que } \ \sigma_k \neq \sigma_l \end{cases}$$

$$\sigma_i^2 = \text{Var}(Y_{ij}) \qquad \forall\, i \in \{1, \ldots, I\}$$

Para este test F, el estadístico de contraste (L) se construye sobre las desviaciones absolutas de cada dato respecto a la media de su grupo,

$$Z_{ij} = |Y_{ij} - \bar{Y}_{i\bullet}| \qquad j \in \{1, \ldots, n_i\}, \quad i \in \{1, \ldots, I\}$$

entonces el estadístico L nos queda:

$$L = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Z_{i\bullet} - \bar{Z}_{\bullet\bullet})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_{i\bullet})^2/(n-I)} \in F_{(I-1),(n-I)}. \tag{1.7}$$

Por lo tanto, fijado un nivel de significación $\alpha \in (0,1)$ y siendo f_α el cuantil que deja a su derecha una probabilidad α en la distribución $F_{(I-1),(n-I)}$, rechazaremos H_0 cuando el estadístico L sea más grande que f_α.

Para estudiar la normalidad, elegimos el test en función del tamaño de la muestra. El test Shapiro-Wilk se usa para muestras menores de 50 y, para muestras grandes, usamos el test de Kolmogorov-Smirnov. En el ejemplo tendremos una muestra mayor de 50 observaciones, por lo que nos centraremos únicamente en el test Kolmogorov-Smirvov.

$$\begin{cases} H_0 : \text{los datos pertenecen a una distribución normal} \\ H_a : \text{los datos no pertenecen a una distribución normal} \end{cases} \tag{1.8}$$

$$\begin{cases} H_0 : Y_{ij} \in N(\mu_i, \sigma^2) \ \ \forall j = 1, \ldots, n_i \ \ \text{para cada } i = 1, \ldots, I \text{ grupo} \\ H_a : \exists \ k \in \{1, \cdots, I\} \ \text{tal que} \ Y_{kj} \notin N(\mu_k, \sigma^2) \ \ \forall j = 1, \ldots, n_k \end{cases}$$

El proceso del test consiste en construir un contraste tal y como aparece en la formulación (1.8). A partir de él, se construye un estadístico de contraste para ver si podemos rechazar H_0. No mostraremos el estadístico de contraste ya que la construcción no es inmediata y sería muy tediosa de realizar. El test se construye para cada una de los I grupos. Para este test tomaremos de referencia $\alpha = 0.05$ como nivel de significación, es decir, si el $p - valor$ resultante del test nos queda $p - valor < 0.05$ entonces rechazamos H_0 y, por lo tanto, el grupo no pertenece a una distribución normal.

En el caso de que el test concluyera que los datos no pertenecen a una distribución normal, podremos modificar la muestra para que el resultado sea más favorable, es decir, eliminaremos los datos atípicos de la muestra. Entendemos por dato atípico aquel que se encuentra muy alejado del modelo de regresión, hasta el punto de inducir a la sospecha de que no sigue el modelo. Estos datos pueden resultar muy condicionantes y, aún habiendo pocos en la muestra, pueden influir fuertemente en la construcción y estimación del modelo establecido. Para el estudio de los datos atípicos, se calculan los residuos estandarizados r_{ij}, así, los datos con unos residuos muy grandes serán considerados atípicos. A continuación, mostramos cómo se calculan los residuos para cada dato,

$$r_{ij} = \frac{\hat{\epsilon}_{ij}}{\hat{\sigma}_i\sqrt{1-h_{ij}}} \ \ j = 1, \ldots, n_i \ \ i = 1, \ldots, I$$

donde $\hat{\epsilon}_{ij}$ representa la estimación del error de cada una de las observaciones, $\hat{\sigma}_i$ es la estimación de la desviación típica del grupo i (si se cumple la hipótesis de homocedasticidad serían iguales en todos los grupos) y h_{ij} representa el apalanque de la observación j-ésima del grupo i,

$$h_{ij} = \frac{1}{n_i} + \frac{(Y_{ij} - \bar{Y}_{i\bullet})^2}{SXX_i} \quad j = 1, \ldots, n_i \quad i = 1, \ldots, I$$

donde $SXX_i = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$, con $i = 1, \ldots, I$. Nótese que estos elementos se sitúan en la diagonal de la matriz Hat (ordenados por grupo y por elemento dentro de cada grupo).

Una vez calculados los residuos, procedemos a señalar los que sean atípicos, tomaremos como referencia el valor 1.96 (cuantil del 97.5 %), así, se eliminarán los residuos absolutos mayores de este valor, $|r_{ij}| > 1.96$. A continuación, se repite el test Kolmogorov-Smirnov con los nuevos datos.

La hipótesis de independencia se asume cierta y debe ser tenida en cuenta en el proceso de diseño de experimentos. El objetivo del modelo ANOVA consiste en ver si se puede asumir que las medias de los grupos son iguales, por lo que, si resulta que las observaciones no son independientes (entre grupos y dentro de los grupos) habrá que reconsiderar las distribuciones que se obtienen. Seguirán siendo normales pero habrá que recalcular las varianzas.

1.5. Ejemplo de aplicación

Una vez vista la teoría del funcionamiento del modelo ANOVA con una sola variable explicativa categórica (ANOVA de una vía), vamos a ver los efectos y usos prácticos de este modelo con un ejemplo. En este ejemplo, vamos a ver paso a paso lo visto anteriormente, exponer los datos, estudiar la comparación de medias y terminar por validar el modelo.

La muestra del ejemplo son los datos obtenidos de un estudio en Philadelphia sobre madres y el peso de sus respectivos bebés en el momento del parto. El objetivo será ver si existen diferencias significativas del peso en el momento del parto, según las condiciones de la madre. Los datos los obtenemos del paquete *faraway* del software *R* [5]. La muestra está formada por 1115 madres (cada una es una observación) y de cada observación se recogen datos de 5 variables distintas.

Es significativo resaltar que los datos de este ejemplo son de un estudio demográfico, no de un experimento, es decir, los tamaños de las distintas variables nos vienen dadas, no es una variable que se pueda modelar. Esto se debe a que no se puede controlar el número de madres que van a dar a luz, sino que se toman dichos tamaños una vez ha pasado el período de estudio, como consecuencia, vamos a tener grupos de distinto tamaño, lo que a priori

no es óptimo. Al final del capítulo se realizan unas simulaciones en las que se estudia cómo varían los resultados del test F a cambios en el tamaño de los grupos. A continuación, se muestran las 6 primeras observaciones de la muestra.

	black	*educ*	*smoke*	*gestate*	*grams*
1	*FALSE*	0	*TRUE*	40	2898
2	*TRUE*	0	*TRUE*	26	994
3	*FALSE*	2	*FALSE*	38	3977
4	*FALSE*	2	*TRUE*	37	3040
5	*FALSE*	2	*FALSE*	38	3523
6	*FALSE*	5	*TRUE*	40	3100

- *black* (Z): es una variable categórica que nos dice *TRUE* si la madre es de raza negra o *FALSE* si la madre no es de raza negra. Esta variable la añadiremos más adelante.
- *educ*: es una variable numérica que nos indica el número de años académicos cursados por la madre. No trabajaremos con ella.
- *smoke* (X): es una variable categórica con *TRUE* si la madre fuma o *FALSE* si la madre no fuma, esta será nuestra variable explicativa principal.
- *gestate*: es una variable continua que nos indica el número de semanas de embarazo. No trabajaremos con ella.
- *grams* (Y): es una variable continua que nos indica el peso en gramos del bebé en el momento del parto, será nuestra variable respuesta.

Nuestro modelo ANOVA representará los gramos de los bebés en función de si la madre fumaba o no (usaremos la notación *grams*∼*smoke*), este modelo es muy vistoso a primera vista, ya que vamos a agrupar nuestra variable respuesta en tan solo dos grupos, el grupo de las no fumadoras, que consta de 846 madres y el de fumadoras con 269 madres, es decir, $n_1 = 846$, $n_2 = 269$. A continuación, mostramos el modelo en función del intercepto. Asumimos la hipótesis de homocedasticidad, normalidad e independencia.

$$\begin{cases} Y \sim X + \epsilon \\ grams \sim smoke \\ Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad j = 1, \dots, n_i \qquad i = 1, 2 \\ \epsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

El objetivo principal consiste en ver si el peso del bebé, de media, varía dependiendo si la madre fuma o no fuma.

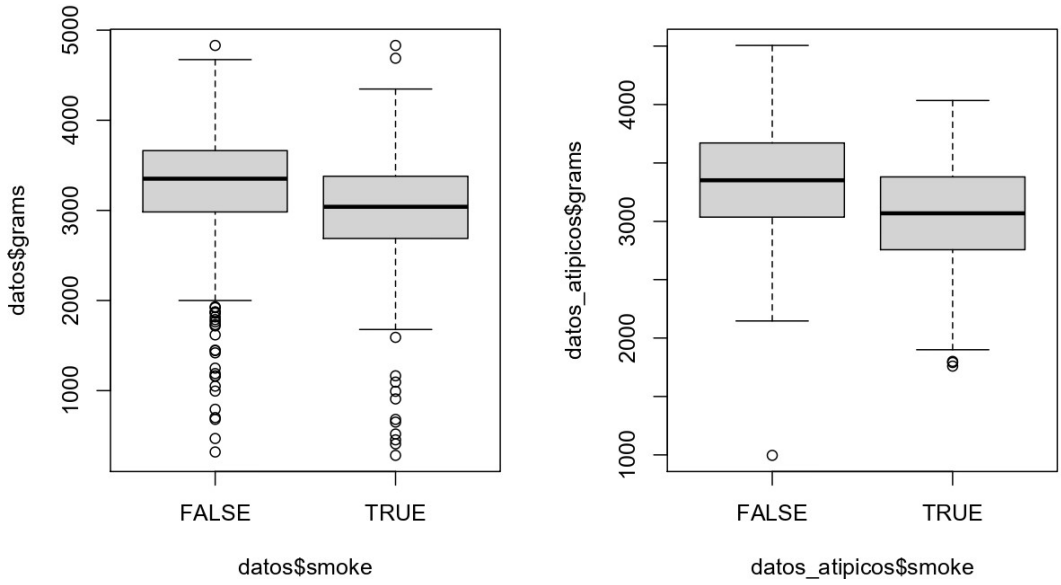

Figura 1.1: En la izquierda hay un diagrama de cajas para el peso de los bebés de madres no fumadoras (FALSE) y fumadoras (TRUE). A la derecha el mismo diagrama sin los datos atípicos.

$$\begin{array}{llll} Y_{11} & Y_{12} & \cdots & Y_{1,846} \quad \text{de una población} \quad N(3300.999, \sigma^2) \\ Y_{21} & Y_{22} & \cdots & Y_{2,269} \quad \text{de una población} \quad N(2963.338, \sigma^2) \end{array}$$

En la gráfica de la izquierda de la Figura 1.1 podemos ver la agrupación de las observaciones en los dos grupos de *smoke*. A primera vista, podemos intuir que sí que habrá diferencia en los grupos viendo la diferencia de alturas de los grupos y medias, es decir, habrá diferencia en el peso entre el grupo de fumadoras y de no fumadoras. Toda esta interpretación hay que formalizarla para comprobar si es cierta o no.

Construímos nuestra hipótesis nula, las medias de los dos grupos son iguales, y la hipótesis alternativa, las medias de los dos grupos son diferentes, siendo μ_1 la media de las observaciones correspondientes a $smoke = FALSE$ y μ_2 la media de $smoke = TRUE$. Tal y como aparece en la formulación (1.9).

$$\begin{cases} H_0 : grams \sim 1 \\ H_a : grams \sim smoke \end{cases} \tag{1.9}$$

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_a : \mu_1 \neq \mu_2 \end{cases}$$

Una vez construído el contraste de referencia en la formulación (1.4), procedemos a comprobar si representamos la respuesta sobre sí misma (H_0), o si la variable *smoke* representa un efecto relevente sobre *grams* (H_a).

El contraste anterior es un contraste de medias con dos poblaciones ($I = 2$), por lo que, para ver si podemos rechazar H_0, vamos a construir el estadístico de contraste T como se indica en la ecuación (1.6).

Aquí, es importante resaltar que en el modelo ANOVA trabajamos con una distribución F de *Snedecor*, ya que comparamos la media de I grupos (siendo $I > 2$). Sin embargo, en este ejemplo únicamente tenemos 2 grupos, por lo que para la construcción del estadístico de contraste T, resulta equivalente usar la distribución F de *Snedecor* y la distribución T de *Student* con $n - 2$ grados de libertad.

Por un lado, la forma más directa de comparar las medias de dos grupos, bajo las hipótesis de normalidad y homocedasticidad, es usar la distribución T de *Student*, en este caso la distribución F de *Snedecor* se usaría para comprobar la hipótesis de homocedasticidad. Por otro lado, la distribución F de *Snedecor* con dos grupos es de la forma $F_{1,n-2}$ (suponiendo una muestra de tamaño n). Al tener en el numerador un único grado de libertad podemos estimar el estadístico de la formulación (1.6) a un estadístico perteneciente a una T de *Student*.

$$\frac{Z}{\sqrt{\frac{H}{v}}} \in t_v \quad \sim \quad \frac{\chi_1/1}{\chi_2/(n-2)} \in F_{1,n-2}$$

Aquí, $Z \sim N(0, 1)$ es una distribución normal de media 0 y varianza 1. $H \sim \chi^2_v$ es un distribución Chi-cuadrado con v grados de libertad y t_v representa la distribución T de *Student* con v grados de libertad.

Finalmente, construímos $T \in F_{1,1113}$ usando la suma residual de cuadrados bajo la hipótesis nula y alternativa. Dichos valores nos dan los resultados: $RSS_0 = 23270642$ y $RSS = 423827509$, por lo que T nos quedará,

$$T = 61.11$$

por lo tanto, el resultado final es un $p - valor = 1.245{\cdot}10^{-14}$, es decir, rechazamos H_0 con una probabilidad de acierto muy alta para los datos con los que trabajamos.

Para terminar, vamos a validar el modelo.

- **Independencia**: asumimos independencia de las observaciones. Cada observación corresponde al parto de una madre diferente, no hay indicios de que estén relacionadas entre sí.

- **Normalidad**: hacemos el test Kolmogorov-Smirnov para los dos grupos. Obtenemos los valores $p-valor = 0.006328$ y $p-valor = 0.002023$ para los grupos $smoke = TRUE$ y $smoke = FALSE$ respectivamente, por lo que rechazamos H_0 y concluímos que los datos no pertenecen a una distribución normal.

  Para mejorar el estudio del test y obtener un resulado más fiable, necesitamos eliminar de la muestra los datos atípicos. Calculamos los residuos estandarizados para cada uno de los datos de ambos grupos y los que cumplan $|r_{ij}| > 1.96$ con $j = n_1, n_2$ e $i = 1, 2$ serán eliminados de la muestra. Obtenemos un total de 52 datos atípicos, los eliminamos y repetimos el test Kolmogorov-Smirnov. La repetición del test nos da unos valores $p-valor = 0.9233$ y $p-valor = 0.3158$ para los grupos $smoke = TRUE$ y $smoke = FALSE$ respectivamente. Estos valores son muy superiores a 0.05, por lo que aceptamos H_0 y concluimos que los datos actuales de ambos grupos sí que pertenecen a una distribución normal.

- **Homocedasticidad**: haciendo el test Levene a la muestra sin los datos atípicos obtenemos un estadístico de contraste $L = 0.2745$ y un $p-valor = 0.6005$, por lo que no tenemos indicios para rechazar H_0, es decir, podemos asumir que se cumple la hipótesis de homogeneidad de varianzas.

Una vez finalizada la construcción y validación del modelo, vamos a mostrar las estimaciones de los parámetros. Aquí se representa la parametrización respecto un grupo de referencia, por defecto el software R toma como grupo de referencia $smoke = FALSE$.

$$\begin{cases} grams \sim smoke \\ \hat{Y} = 33471 - 304.4 smokeTRUE \end{cases}$$

De esta forma, el parámetro $smokeTRUE$ representa el incremento del grupo $smoke = TRUE$ respecto la media del grupo $smoke = FALSE$, es decir, nos indica que se restan los 304.4 al intercepto en el caso de que la observación cumpla $smoke = TRUE$, en el caso de que la observación sea $smoke = FALSE$ se eliminarían los 304.4 y en la estimación quedaría solo el intercepto.

Concluímos entonces que es cierto lo que intuíamos, que el hecho de fumar va afectar al peso del bebé a la hora de nacer, que las madres fumadoras van a tener, de media, un bebé de peso menor a las madres no fumadoras.

En la Tabla 1.3 mostramos las estimaciones de los parámetros con los respectivos intervalos de confianza y el nivel de significación de cada parámetro. A mayores, representamos en la gráfica de la izquierda de la Figura 1.1 cómo se ve gráficamente la eliminación de los datos atípicos. En dicha figura podemos apreciar la diferencia entre la gráfica con los datos atípicos y la gráfica sin ellos.

	smoke		
Parámetro	Estimación	I.C.	p
Intercepto	3347.81	3314.95 – 3380.67	<0.001
smokeTRUE	-304.36	-372.12 – -236.60	<0.001
Observaciones	1063		

Tabla 1.3: Tabla de la estimación del modelo $grams \sim smoke$ por mínimos cuadrados. I.C.: intervalo de confianza al 95 %. p: $p - valor$ para el constraste de significación.

1.6. Simulación

En la sección anterior, hemos estudiado y validado el test F para el modelo $grams \sim smoke$. Obtuvimos unos resultados bastante favorables del test y fuímos capaces de rechazar la hipótesis H_0 con una sólida seguridad. A continuación, nos preguntamos si esto hubiese ocurrido con una muestra más pequeña, es decir, hasta qué punto el tamaño de la muestra influenció a la hora de hacer el test F.

El objetivo del estudio que haremos en el resto del capítulo, será ver la sensibilidad del test F a cambios en distintos valores. Si la sensibilidad del test a algún estímulo es alta, entonces los resultados obtenidos no serían muy fiables. Primero, veremos la sensibilidad al tamaño de la muestra, después a la varianza de los grupos que siguen una distribución normal y por último, a muestras cuyos grupos no sigan una misma distribución.

Para estudiar el efecto del tamaño de la muestra en el test F, vamos a realizar algunas simulaciones de datos variando el tamaño de la muestra. En dichas simulaciones, generamos de forma aleatoria una variable continua X en función de una variable categórica G, formada por tres grupos. Por ejemplo, empezaremos con una muestra en la que agruparemos 25 observaciones en cada grupo, a continuación, se generan los datos de X siguiendo una distribución normal de media $\mu = 0$ y varianza $\sigma^2 = 1$. Hacemos el test F para los tres grupos y obtenemos un $p - valor = 0.5171$, por lo que aceptamos H_0, la hipótesis de que las medias de los tres grupos son iguales.

$$\begin{cases} H_0 : X \sim 1 \\ H_a : X \sim G \end{cases}$$

El interés de este estudio es repetir este proceso 500 veces y ver qué proporción de veces (porcentaje por uno) rechazamos H_0 con los niveles de significación usuales, 10 %, 5 % y 1 %, es decir, qué proporción de veces nos resulta un $p - valor$ menor de 0.1, 0.05 y 0.01, respectivamente. Los valores de las tablas de esta sección son dichos porcentajes, cada tabla sobre su respectivo estudio.

Empezamos el estudio generando tres grupos, los tres con el mismo número de observaciones y siguiendo una distribución normal de media $\mu = 0$ y varianza $\sigma^2 = 1$. Haremos dos simulaciones, una en la que los tres grupos tendrán el mismo tamaño, y otra en la que cada grupo tendrá un tamaño diferente. La variable $Tamaño$ de la Tabla 1.4 representa el número de observaciones de los grupos.

Tamaño	< 0.1	< 0.05	< 0.01
25	0.090	0.050	0.006
50	0.120	0.056	0.010
100	0.160	0.060	0.008
250	0.106	0.044	0.012
500	0.122	0.070	0.010
1000	0.080	0.046	0.008

Tabla 1.4: Tabla de porcentajes de rechazo para una variable con tres grupos del mismo tamaño para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Vemos que los resultados de la Tabla 1.4 se mantienen constantes en sus adecuados porcentajes, es decir, rechazamos la hipótesis de que los tres grupos tienen la misma media la proporción de veces adecuado para cada uno de los niveles de significación.

Ahora, repetiremos la simulación anterior con un tamaño diferente en cada grupo. La variable $Tamaño$ de la Tabla 1.5 representa tres valores, el número de observaciones de los grupos uno, dos y tres respectivamente.

En la Tabla 1.5 vemos que se vuelve a mantener unos valores muy equilibrados, a pesar de la diferencia entre grupos. Concluímos que el test F está bien calibrado al efecto del tamaño de la muestra de una distribución normal $N(0, 1)$.

Tamaño	< 0.1	< 0.05	< 0.01
5, 50, 200	0.102	0.048	0.010
5, 50, 200	0.110	0.070	0.018
10, 20, 500	0.092	0.044	0.010
10, 20, 500	0.102	0.060	0.012
1000, 100, 3	0.090	0.056	0.010
1000, 100, 3	0.094	0.050	0.008

Tabla 1.5: Tabla de porcentajes de rechazo para una variable con tres grupos de distinto tamaño para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Una vez vista la sensibilidad del test F al tamaño de la muestra, vamos a probar a aumentar las varianzas de los grupos. Haremos dos simulaciones, una en la que los tres grupos tendrán la misma varianza y otra simulación en la que los tres tendrán distinta varianza. Para estas simulaciones se generarán grupos con el mismo tamaño que además será constante, 200 observaciones en cada grupo. El parámetro σ de la Tabla 1.6 representa el valor de la desviación típica de los grupos.

σ	< 0.1	< 0.05	< 0.01
1	0.070	0.020	0.004
1	0.098	0.038	0.014
10	0.084	0.050	0.012
10	0.106	0.078	0.004
50	0.090	0.046	0.010
50	0.098	0.044	0.006

Tabla 1.6: Tabla de porcentajes de rechazo para la varianza para una variable con tres grupos con el mismo tamaño para distintos niveles de significación. La continua sigue una distribución $N(0, \sigma^2)$.

En la Tabla 1.6 vemos que los porcentajes se mantienen más o menos estables entorno a sus valores correctos, 10 %, 5 % y 1 %.

A continuación, repetimos la simulación anterior variando la varianza de cada uno de los grupos. En la Tabla 1.7 consideramos el parámetro σ, que representará el valor de la desviación típica de los grupos uno, dos y tres respectivamente. Al igual que en la simulación anterior, se mantendrá el número de observaciones (200) constante.

σ	< 0.1	< 0.05	< 0.01
1, 10, 50	0.118	0.066	0.014
1, 10, 50	0.092	0.038	0.004
20, 25, 50	0.104	0.062	0.006
20, 25, 50	0.086	0.040	0.008
60, 5, 10	0.102	0.054	0.014
60, 5, 10	0.108	0.062	0.008

Tabla 1.7: Tabla de porcentajes de rechazo para una variable con tres grupos con el mismo tamaño para distintos niveles de significación. La continua sigue una distribución $N(0, \sigma^2)$ con distinta varianza.

En la Tabla 1.7 vemos que, al igual que en la simulación anterior, se matienen unos valores adecuados entorno a sus porcentajes correctos. Concluímos que el test F está bien calibrado a cambios de la varianza con muestras de tamaño constante que siguen una distribución normal $N(0, \sigma^2)$, también constante.

Tamaño	< 0.1	< 0.05	< 0.01
25	0.106	0.050	0.006
50	0.104	0.058	0.010
100	0.118	0.070	0.014
250	0.096	0.056	0.016
500	0.096	0.042	0.010
1000	0.092	0.050	0.002

Tabla 1.8: Tabla de porcentajes de rechazo para una variable con tres grupos con el mismo tamaño para distintos niveles de significación. Dos grupos siguen una distribución $N(1, 1)$ y el otro una $Exp(1)$.

Para terminar, generamos una muestra repartida en tres grupos, igual que en las anteriores simulaciones, la diferencia será que no todos los datos seguirán una distribución normal, sino que habrá un grupo generado a partir de una distribución exponencial o χ^2_1. El objetivo será ver la sensibilidad del test F a los cambios en el tamaño de la muestra. Realizaremos dos simulaciones con cada distribución, una en la que los tres grupos tendrán el mismo tamaño y otra en la que cada grupo tendrá un tamaño diferente.

Empezamos la simulación generando un grupo siguiendo una distribución exponencial con $\lambda = 1$ y otros dos grupos siguiendo una distribución normal con media $\mu = 1$ y varianza $\sigma^2 = 1$. La variable $Tamaño$ de la Tabla 1.8 muestra el número de observaciones de los grupos.

En la Tabla 1.8 vemos que los valores $p - valor$ se mantienen más o menos constates, aún variando el tamaño de los grupos.

Ahora, repetiremos la simulación anterior con grupos de distinto tamaño entre ellos. En la variable $Tamaño$ de la Tabla 1.9 representamos tres valores, los tamaños de los grupos uno, dos y tres respectivamente. Es importante remarcar que el grupo número dos (el del medio) sigue la distribución exponencial.

Tamaño	< 0.1	< 0.05	< 0.01
5, 50, 200	0.104	0.062	0.008
5, 50, 200	0.092	0.048	0.006
10, 20, 500	0.098	0.052	0.006
10, 20, 500	0.092	0.044	0.008
100, 1000, 3	0.090	0.056	0.010
100, 1000, 3	0.104	0.064	0.014

Tabla 1.9: Tabla de porcentajes de rechazo para una variable con tres grupos de distinto tamaño para distintos niveles de significación. Dos grupos siguen una distribución $N(1, 1)$ y el otro una $Exp(1)$.

En la Tabla 1.9 vemos que hay unos valores $p - valor$ adecuados para sus respectivos porcentajes de rechazo. Concluímos que el test F está bien calibrado a los cambios de tamaño para grupos pertenecientes a distribuciones normales de media y varianza $\mu = 1$ y $\sigma^2 = 1$ y exponenciales de $\lambda = 1$.

Tamaño	< 0.1	< 0.05	< 0.01
25	0.088	0.052	0.006
50	0.092	0.048	0.010
100	0.094	0.034	0.008
250	0.122	0.060	0.002
500	0.120	0.070	0.020
1000	0.104	0.050	0.010

Tabla 1.10: Tabla de porcentajes de rechazo para una variable con tres grupos con el mismo tamaño para distintos niveles de significación. Dos grupos siguen una distribución $N(1,1)$ y el otro una χ^2_1.

Ahora, veremos si el test F está bien calibrado para una distribución χ^2_1. Generamos tres grupos, dos de ellos siguiendo una distribución normal de media $\mu = 1$ y varianza $\sigma^2 = 1$ y el grupo que queda una distribución χ^2 con un grado de libertad, $df = 1$. En la variable $Tamaño$ de la Tabla 1.10 mostramos el número de observaciones de los grupos.

En la Tabla 1.10 vemos que se mantienen adecuados los valores $p - valor$, al igual que en la tabla con los datos de la distribución exponencial.

Repetimos la simulación con grupos de distinto tamaño entre ellos. En la variable $Tamaño$ de la Tabla 1.11 representamos tres valores, el número de observaciones de los grupos uno, dos y tres respectivamente. Es significativo señalar que es el grupo dos (el del medio) el que sigue la distribución χ^2_1.

Tamaño	< 0.1	< 0.05	< 0.01
5, 50, 200	0.142	0.078	0.014
5, 50, 200	0.134	0.068	0.024
10, 20, 500	0.148	0.1	0.054
10, 20, 500	0.188	0.118	0.038
100, 1000, 3	0.004	0.000	0.000
100, 1000, 3	0.014	0.000	0.000

Tabla 1.11: Tabla de porcentajes de rechazo para una variable con tres grupos de distinto tamaño para distintos niveles de significación. Dos grupos siguen una distribución $N(1,1)$ y el otro una χ^2_1.

En la Tabla 1.11 vemos que sí que hay una diferencia muy significativas entre los $p - valor$ de los datos que siguen una distribución χ^2_1 con un grado de libertad. Por lo tanto, concluímos que el test F no está bien calibrado si se tiene que comparar muestras que siguen distribuciones normales $N(1,1)$ y distribuciones χ^2_1.

Capítulo 2

ANOVA de dos vías

El modelo ANOVA de dos vías es la extensión del modelo ANOVA de una vía a dos variables categóricas explicativas. En este capítulo, vamos a empezar planteando el modelo en el ejemplo para luego poder construir la teoría a partir de él. Lo importante es ver cómo este modelo nos sirve para añadir información a la respuesta y conocer mejor su comportamiento. Nos basaremos en teoría vista en [3] y [4].

2.1. Añadir una variable

¿Qué pasaría si le añadimos una variable explicativa categórica a mayores al modelo ANOVA de una vía? ¿Influye esta nueva variable en nuestra variable explicativa original? ¿Cómo es esta nueva variable respecto la respuesta?

Partiendo del modelo construido en el Capítulo 1 ($grams \sim smoke$) y con la idea de aprender más sobre nuestra variable respuesta, vamos a añadir la variable *black*. Esta variable categórica tiene solo dos grupos, *TRUE* y *FALSE*, con una muestra de 453 observaciones en *TRUE* y 662 en *FALSE*. Nuestro resultado será un modelo ANOVA de dos vías, formado por dos variables explicativas categóricas condicionando la respuesta.

$$grams \sim smoke + black + smoke : black$$

En este capítulo vamos a estudiar el efecto que tienen por separado las variables sobre la respuesta y el efecto de interacción de ambas variables ($smoke : black$). Los resultados de este estudio pueden formar cuatro posibles escenarios, los exponemos a continuación.

- **Ninguna variable afecta a la respuesta**. Si en el modelo $grams \sim black$ resulta que sus dos medias son iguales y el modelo $grams \sim smoke$ diese el mismo resultado,

entonces ninguna variable sería relevante para el estudio de la respuesta. Este caso está descartado ya que la variable *smoke* sí que nos resultó relevante. El modelo ANOVA de dos vías nos quedaría:

$$grams \sim 1$$

- **Una variable no afecta a la respuesta**. Podría darse el caso de que las medias del modelo $grams \sim black$ nos diesen iguales y por lo tanto *black* no aportase información sobre *grams*. El modelo ANOVA de dos vías nos quedaría:

$$grams \sim smoke$$

- **El efecto de interacción de las variables no afecta a la respuesta**. Suponiendo que ambas variables son relevantes sobre la respuesta, se podría dar el caso de que, a la hora de estudiar el efecto de interacción de *smoke* y *black*, no hubiese resultados relevantes. El modelo ANOVA de dos vías nos quedaría:

$$grams \sim smoke + black$$

- **Ambas variables y el efecto de interacción de ellas es relevante para estudiar la respuesta**. El modelo ANOVA de dos vías nos quedaría:

$$grams \sim smoke + black + smoke : black$$

A continuación, vamos a estudiar el efecto de la variable *black* sobre la respuesta *grams*, construiremos los contrastes adecuados y el estadístico de contraste T para ver si podemos suponer que las medias de los grupos $black = FALSE$ y $black = TRUE$ son iguales. Escribimos nuestro modelo con la parametrización en función del intercepto, donde μ representa el intercepto y β_i el incremento del grupo j de la variable *black*, además, asumimos por cierta las hipótesis de homocedasticidad, normalidad e independencia.

$$\begin{cases} Y \sim Z + \epsilon \\ grams \sim black \\ Y_{ij} = \mu + \beta_i + \epsilon_{ij} \qquad j = 1, \ldots, n_i \qquad i = 1, 2 \\ \epsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

$$\begin{array}{cccccc} Y_{11} & Y_{12} & \cdots & Y_{1,662} & \text{de una población} & N(3415.861, \sigma^2) \\ Y_{21} & Y_{22} & \cdots & Y_{2,453} & \text{de una población} & N(3085.193, \sigma^2) \end{array}$$

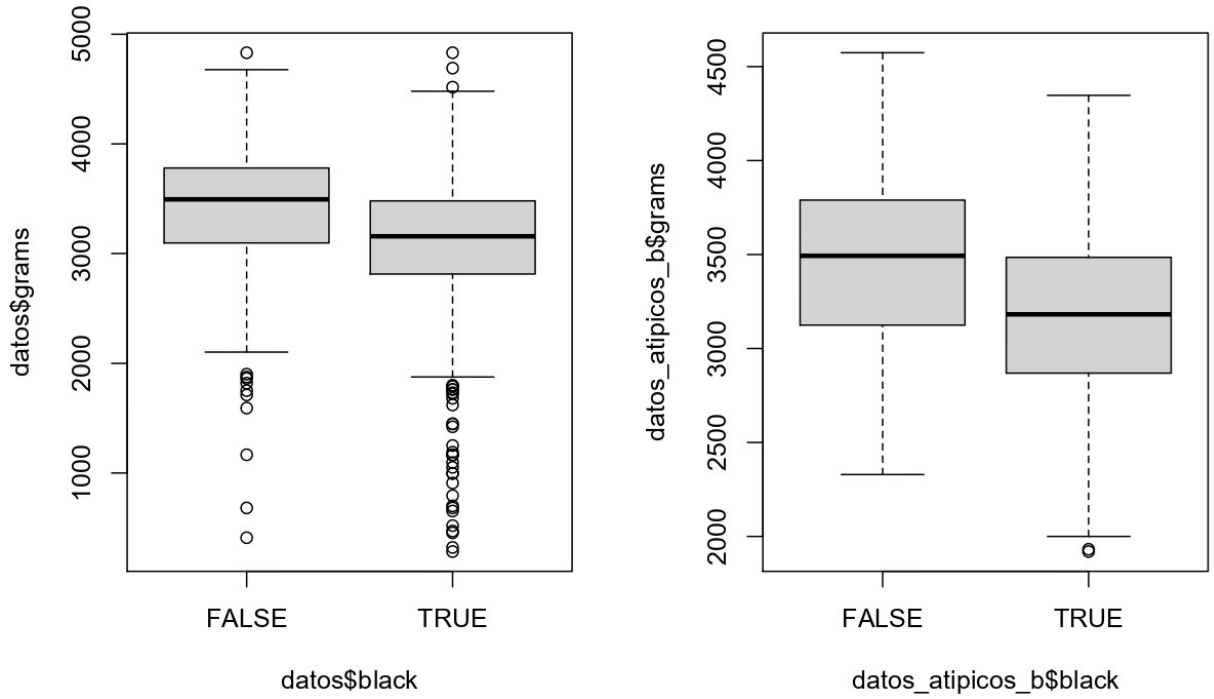

Figura 2.1: En la izquierda hay un diagrama de cajas para el peso de los bebés para madres de raza no negra (FALSE) y de raza negra (TRUE). A la derecha, el mismo diagrama sin los datos atípicos.

En la gráfica de la izquierda de la Figura 2.1, se representa la respuesta *grams* en los dos grupos de *black*. Podemos intuir que la tendencia de las observaciones no es la misma en los dos grupos viendo la diferencia de altura de la media y de las cajas, es decir, podemos intuir que *black* será relevante sobre *grams*, que el peso de los bebés no será igual si la madre es de raza negra a si no lo es. Vamos a formalizar esta teoría realizando el test F, de la misma forma que lo comprobamos con la variable *smoke* en el Capítulo 1.

$$\begin{cases} H_0 : grams \sim 1 \\ H_a : grams \sim black \end{cases} \tag{2.1}$$

$$\begin{cases} H_0 : \mu_1' = \mu_2' \\ H_a : \mu_1' \neq \mu_2' \end{cases}$$

Construímos el contraste, formulación (2.1), y el estadístico de contraste $T \in F_{1,1113}$, siendo μ_1' la media de las observaciones que cumplen $black = TRUE$ y μ_2' las que cumplen $black = FALSE$.

$$T = 78.362$$

El valor asociado a nuestro T es $p - valor = 2.2 \cdot 10^{-16}$, es decir, rechazamos H_0 con una probabilidad muy alta para los datos con los que trabajamos. Por lo tanto, la variable *black* sí que tiene relevancia a la hora de estudiar nuestra respuesta y confirmamos que nuestro modelo ANOVA de dos vías partirá de la forma $grams \sim smoke + black$.

Antes de mostrar las estimaciones de los parámetros del modelo, vamos a validarlo.

- **Independencia**: asumimos independencia de las observaciones. Cada observación corresponde al parto de una madre diferente, no hay indicios de que estén relacionadas entre sí.

- **Normalidad**: hacemos el test Kolmogorov-Smirnov para los dos grupos. Obtenemos los valores $p-valor = 2.58e-05$ y $p-valor = 0.02799$ para los grupos $black = TRUE$ y $black = FALSE$ respectivamente, por lo que rechazamos H_0 y concluímos que los datos no pertenecen a una distribución normal.

  Procedemos a calcular los datos atípicos con los residuos estandarizados. Obtenemos un total de 56 datos atípicos, los eliminamos y repetimos el test Kolmogorov-Smirnov. La repetición del test nos da unos valores $p - valor = 0.6401$ y $p - valor = 0.1224$ para los grupos $black = TRUE$ y $black = FALSE$ respectivamente. Estos valores son superiores a 0.05, por lo que aceptamos H_0 y concluimos que los datos actuales sí que pertenecen a una distribución normal.

- **Homocedasticidad**: si hacemos el test Levene con toda la muestra da un estadístico de contraste $L = 3.6624$ y un $p-valor = 0.05591$, por lo que no podemos asumir que se cumpla la hipótesis de homogeneidad de las varianzas. Sin embargo, si repetimos el test con la muestra sin los datos atípicos, resulta un estadístico de contraste $L = 0.0094$ y un $p - valor = 0.9226$, por lo que no habría indicios para rechazar H_0 y sí que se podría asumir la hipótesis de homogeneidad de las varianzas.

A partir de la muestra sin los datos atípicos mostramos las estimaciones de los parámetros del modelo. Se muestrasn usando la parametrización en función del grupo de referencia $black = FALSE$.

$$\begin{cases} grams \sim black \\ \hat{Y} = 3457.1 - 295.4 blackTRUE \end{cases}$$

Concluímos entonces que las variables *smoke* y *black* sí que tendrán efecto sobre la respuesta *grams* de manera individual y su estudio será relevante.

Una vez vistos los efectos de las variables explicativas por separado, nos disponemos a estudiar cómo funcionan estas dos variables combinadas. La principal incógnita a resolver

es si las dos variables tienen interacción entre ellas, es decir, si es significativa la interacción entre los grupos de las dos variables.

Para ayudar a la visualización de esta interacción vamos a mostrar cuatro gráficas (véase Figura 2.2) en las que aparecen las medias combinadas de ambos factores.

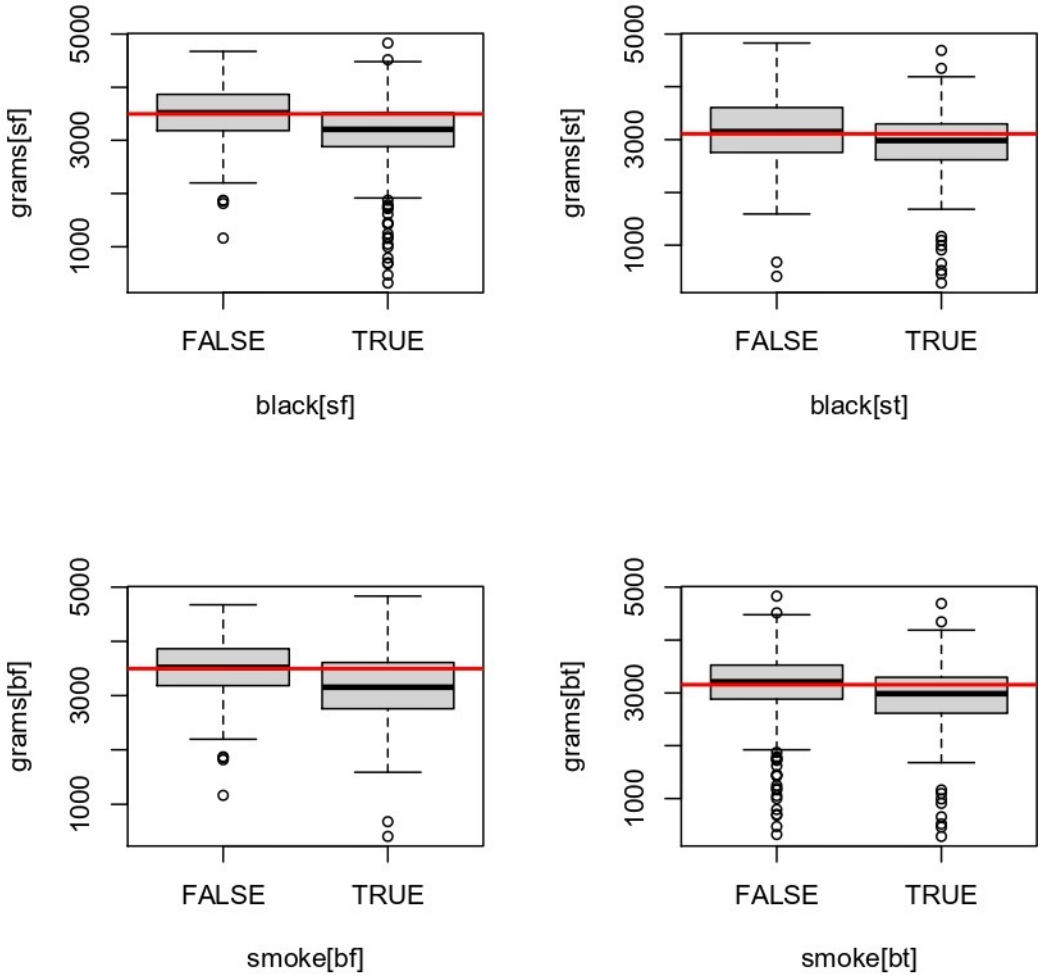

Figura 2.2: Diagrama de cajas de *grams* en función las combinaciones de *smoke* y *black*.

- Primera: arriba a la izquierda se representa un diagrama de caja de la respuesta en función de *black* únicamente con los datos $smoke = FALSE$.
- Segunda: arriba a la derecha se representa un diagrama de caja de la respuesta en función de *black* únicamente con los datos $smoke = TRUE$.
- Tercera: abajo a la izquierda se representa un diagrama de caja de la respuesta en función de *smoke* únicamente con los datos $black = FALSE$.

- Cuarta: abajo a la derecha se representa un diagrama de caja de la respuesta en función de *smoke* únicamente con los datos $black = TRUE$.

El objetivo de estas gráficas es ver si hay diferencias significativas cuando se combinan los niveles de *smoke* y de *black*. Por ejemplo, en la primera fila se puede intuir que los datos de la gráfica de la izquierda son mayores que los de la derecha, por lo que $smoke = FALSE$ y $black = FALSE$ implicará unos mayores datos de la respuesta, es decir, una madre no fumadora y de raza no negra tendrá de media un bebé de peso mayor que el resto de combinaciones. Lo mismo ocurre en la segunda fila, se puede apreciar que los datos de la gráfica de la izquierda son mayores que los de la derecha, por lo que $black = FALSE$ y $smoke = FALSE$ implicará unos mayores datos de la respuesta.

Por lo tanto, lo que necesitamos es ver si la media de la respuesta de cada una de las combinaciones de las dos variables se puede asumir igual o no. Para formalizar toda esta interpretación nos hemos quedado cortos con la teoría del modelo ANOVA vista hasta aquí, necesitamos ahondar más en la teoría del efecto combinado (efecto de interacción).

2.2. Formulación del modelo ANOVA de dos vías

En esta sección vamos a ver cual es el estudio adecuado del modelo ANOVA de dos vías. Primero, se verá la construcción general del modelo para dos variables con varios grupos y luego concretaremos el modelo que nos interesa, dos variables con dos grupos cada una. Nos referiremos a este último modelo como 2×2.

Supongamos que tenemos 2 variables explicativas categóricas X y Z, cada una con I y J grupos respectivamente, Y es la variable respuesta continua y $X : Z$ representa la interacción entre X y Z. Tomamos una muestra de tamaño n con n_{ij} observaciones para cada combinación de cada grupo $i = 1, \dots, I$ y $j = 1, \dots, J$ de forma que nos queda $\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = n$. A continuación, mostramos la parametrización del modelo ANOVA de dos vías.

$$\begin{cases} Y \sim X + Z + X : Z + \epsilon \\ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \\ i \in \{1, \dots, I\}, j \in \{1, \dots, J\}, k \in \{1, \dots, n_{ij}\} \end{cases}$$

- μ: es el efecto común a todas las combinaciones de niveles para ambos factores.
- α_i: son los incrementos asociados a cada uno de los niveles del factor X.
- β_j: son los incrementos asociados a cada uno de los niveles del factor Z.

- $(\alpha\beta)_{ij}$: representa la interacción entre ambos factores.
- ϵ_{ijk}: el efecto aleatorio, recoge el efecto de todas las restantes causas posibles de variabilidad del experimento. $\epsilon_{ijk} \in N(0, \sigma^2)$

Para que el modelo esté bien formulado y sea identificable (se puedan estimar correctamente los parámetros), se considera que μ es la media global, lo que conlleva las siguientes restricciones

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0. \tag{2.2}$$

Los parámetros $\alpha, \beta, (\alpha\beta)$ miden los efectos aditivos respecto de la media global μ para cada una de sus respectivas variables. Dejamos las estimaciones de estos parámetros para la siguiente sección, cuando construyamos los contrastes con sus respectivos estadísticos.

A continuación, mostramos una tabla con el número de parámetros que tendremos que estimar.

Nombre	Número de ellos
μ	1
α_i	$I - 1$
β_j	$J - 1$
$(\alpha\beta)_{ij}$	$(I - 1)(J - 1)$
TOTAL	$I \cdot J$

Tabla 2.1: Tabla de los parámetros del modelo ANOVA de dos vías.

Una vez vista la formulación general del modelo ANOVA de dos vías, vamos a mostrar ahora el caso 2×2, es decir, lo que queda de sección vamos a trabajar con dos variables categóricas con dos grupos cada una. Este es el modelo de nuestro ejemplo y nos interesa preparar la teoría para una mejor comprensión. $I = 2$ y $J = 2$.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad k \in \{1, \ldots, n_{ij}\} \quad i, j = 1, 2 \tag{2.3}$$

Por lo tanto, según la ecuación (2.2), los parámetros del modelo (2.3) van a cumplir el siguiente esquema,

$$\begin{aligned}
\alpha_1 &= -\alpha_2 \\
\beta_1 &= -\beta_2 \\
(\alpha\beta)_{i1} &= -(\alpha\beta)_{i2} \quad i = 1, 2 \\
(\alpha\beta)_{1j} &= -(\alpha\beta)_{2j} \quad j = 1, 2
\end{aligned}$$

Lo más interesante de este esquema es que podemos crear unos valores nuevos α, β y $(\alpha\beta)$ a partir de los valores anteriores. Estos valores nuevos nos serán muy útiles para la creación de los estadísticos de contraste que construiremos en la siguiente sección. En concreto tendremos.

$$\begin{aligned}
\alpha = \text{Efecto de } X &= E[\bar{Y}_{2\bullet\bullet}] - E[\bar{Y}_{1\bullet\bullet}] = \alpha_2 - \alpha_1 = 2\alpha_2 \\
\beta = \text{Efecto de } Z &= E[\bar{Y}_{\bullet 2\bullet}] - E[\bar{Y}_{\bullet 1\bullet}] = \beta_2 - \beta_1 = 2\beta_2 \\
(\alpha\beta)_{11} - (\alpha\beta)_{21} = (\alpha\beta)_{22} - (\alpha + \beta)_{12} &= -2(\alpha\beta)_{21} = 2(\alpha\beta)_{22} = \alpha\beta
\end{aligned}$$

Más adelante, se ve con más detalle la notación $\bar{Y}_{xyz}$. Para terminar, vamos a reescribir el modelo (2.3) tal y como aparece en la Tabla 2.2.

	Z_1	Z_2
X_1	$Y_{11n_{11}}$	$Y_{12n_{12}}$
X_2	$Y_{21n_{21}}$	$Y_{22n_{22}}$

Tabla 2.2: Tabla de la distribución de los datos del modelo 2×2.

En la Tabla 2.2 se puede ver, de una forma bastante visual, cómo se distribuyen los datos del modelo 2×2. Cada observación va a pertenecer a una y solo a una de las cuatro celdas posibles y cada n_{ij} es el número de observaciones de cada celda.

2.3. El test F del ANOVA de dos vías

La construcción del estadístico de contraste T del test F para el caso del modelo ANOVA de dos vías es más complejo que para el de una vía, ya que hay que hacer un desglose del contraste general en tres constrastes distintos. Lo que haremos, será empezar por estudiar el efecto de cada una de las variables separadas sobre la respuesta y después estudiar el efecto combinado de la interacción sobre la respuesta.

Seguiremos el mismo esquema seguido en la Sección 2.2, veremos primero los constrastes para el caso general y luego para el caso 2×2.

- En el contraste (2.4) se estudia el efecto de la variable X sobre la respuesta. La idea es ver si las medias de los I grupos son iguales.

$$\begin{cases} H_{01} : Y \sim 1 \\ H_{a2} : Y \sim X \end{cases} \tag{2.4}$$

$$\begin{cases} H_{01} : \alpha_1 = \cdots = \alpha_I = 0 \\ H_{a2} : \exists \ i = 1, \ldots, I \ \text{ tal que } \ \alpha_i \neq 0 \end{cases}$$

- En el contraste (2.5) se estudia el efecto de la variable Z sobre la respuesta. La idea es ver si las medias de los J grupos son iguales.

$$\begin{cases} H_{02} : Y \sim 1 \\ H_{a2} : Y \sim Z \end{cases} \tag{2.5}$$

$$\begin{cases} H_{02} : \beta_1 = \cdots = \beta_J = 0 \\ H_{a2} : \exists \ j = 1, \ldots, J \ \text{ tal que } \ \beta_j \neq 0 \end{cases}$$

- En el contraste (2.6) se estudia el efecto de la interacción de las dos variables $(X : Z)$ sobre la respuesta.

$$\begin{cases} H_{03} : Y \sim X + Z \quad \text{(sin interacción)} \\ H_{a3} : Y \sim X + Z + X : Z \quad \text{(con interacción)} \end{cases} \tag{2.6}$$

$$\begin{cases} H_{03} : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{ij} = 0 \ \text{ tal que } \ i = 1, \ldots, I \ \ j = 1, \ldots, J \\ H_{a3} : \exists \ i, j, p, l \ \text{ tal que } \ (\alpha\beta)_{ij} \neq (\alpha\beta)_{pl} \ \text{ tal que } \ i, p = 1, \ldots, I \ \ j, l = 1, \ldots, J \end{cases}$$

Veamos ahora cómo se construyen los estadísticos T para cada uno de los tres contrastes. Cada uno seguirá una distribución F de *Snedecor* con diferentes grados de libertad.

Escribimos la tabla de análisis de la varianza, Tabla 2.3. En esta tabla, aparecen las sumas de cuadrados con los respectivos grados de libertad para poder construir los estadísticos para cada una de los tres contrastes (T_X, T_Z, T_{XZ}).

Factor	Suma Cuadrados	Grados libertad	T
X	SC_X	$I-1$	$T_X = \frac{SC_X/(I-1)}{RSS/(n-IJ)}$
Z	SC_Z	$J-1$	$T_Z = \frac{SC_Z/((J-1)}{RSS/(n-IJ)}$
$X:Z$	SC_{XZ}	$(I-1)(J-1)$	$T_{XZ} = \frac{SC_{XZ}/(I-1)(J-1)}{RSS/(n-IJ)}$
Residual	RSS	$n-IJ$	

Tabla 2.3: Tabla análisis de la varianza.

La suma de cuadrados se estima como sigue,

$$SC_X = \sum_{i=1}^{I}(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 n_{i\bullet} = \sum_{i=1}^{I}\hat{\alpha}_i^2 n_{i\bullet}$$

$$SC_Z = \sum_{j=1}^{J}(\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 n_{\bullet j} = In\sum_{j=1}^{J}\hat{\beta}_j n_{\bullet j}$$

$$SC_{XZ} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})^2 n_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\widehat{\alpha\beta})_{ij}^2 n_{ij}$$

$$RSS = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \bar{Y}_{ij\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}\hat{\epsilon}_{ijk}^2$$

con las estimaciones,

$$\bar{Y}_{\bullet\bullet\bullet} = \frac{1}{IJn}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} Y_{ijk} \quad ; \quad \bar{Y}_{ij\bullet} = \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$\bar{Y}_{i\bullet\bullet} = \frac{1}{Jn_{i\bullet}}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} Y_{ijk} \quad i = 1, \cdots, I$$

$$\bar{Y}_{\bullet j\bullet} = \frac{1}{In_{\bullet j}}\sum_{i=1}^{I}\sum_{k=1}^{n_{ij}} Y_{ijk} \quad j = 1, \cdots, J$$

$$\hat{\alpha}_i = \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet} \quad ; \quad \hat{\beta}_j = \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet} \quad ; \quad (\widehat{\alpha\beta})_{ij} = \bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet}$$

$$n_{i\bullet} = \sum_{j=1}^{J} n_{ij} \quad ; \quad n_{\bullet j} = \sum_{i=1}^{I} n_{ij} \quad ; \quad n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}.$$

Una vez construídos los estadísticos, solo queda fijar un nivel de significación $\alpha \in (0,1)$ para cada uno de los contrastes. Siendo f_α el cuantil que deja a su derecha una probabilidad α, rechazaremos H_{0c} con $c \in \{1,2,3\}$ cuando el estadístico T_v con $v \in \{X, Z, XZ\}$ sea más grande que f_α.

Ahora, solo quedar mirar a qué distribución pertenece cada uno de los estadísticos para cada uno de los contrastes.

$$T_X \in F_{I-1,IJ(n-1);\alpha} \ ; \ T_Z \in F_{J-1,IJ(n-1);\alpha} \ ; \ T_{XZ} \in F_{(I-1)(J-1),(n-IJ);\alpha}$$

Una vez vista la construcción del estadístico para el caso del modelo general, vamos a estudiar ahora el caso de nuestro ejemplo, $I = 2$, $J = 2$.

- En este contraste se estudia si las medias de los dos grupos de X son iguales.

$$\begin{cases} H_{01} : \alpha_1 = \alpha_2 = 0 \\ H_{a2} : \alpha_1 \neq 0 \quad \text{ó} \quad \alpha_2 \neq 0 \end{cases}$$

- En este contraste se estudia si las medias de los dos grupos de Z son iguales.

$$\begin{cases} H_{02} : \beta_1 = \beta_2 = 0 \\ H_{a2} : \beta_1 \neq 0 \quad \text{ó} \quad \beta_2 \neq 0 \end{cases}$$

- En este contraste se estudia si la interacción de las dos variables es relevante sobre la respuesta.

$$\begin{cases} H_{03} : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{22} = 0 \\ H_{a3} : \exists \ i,j,p,l \ \text{ tal que } \ (\alpha\beta)_{ij} \neq (\alpha\beta)_{pl} \\ i,j,p,l = 1,2 \end{cases}$$

A continuación, hacemos la tabla de análisis de la varianza, Tabla 2.4, con los datos necesarios para obtener los estadísticos T para cada uno de los contrastes. En este momento, recurrimos a los parámetros α, β y $(\alpha\beta)$ creados en la sección anterior.

Factor	Suma Cuadrados	Grados libertad	T
X	$n\hat{\alpha}^2$	1	$T_X = \hat{\alpha}^2/(RSS/n)$
Z	$n\hat{\beta}^2$	1	$T_Z = \hat{\beta}^2/(RSS/n)$
$X:Z$	$n(\hat{\alpha\beta})^2$	1	$T_{XZ} = (\hat{\alpha\beta})^2/(RSS/4(n-1))$
Residual	$RSS = \sum_i \sum_j \sum_k \epsilon_{ijk}^2$	$4(n-1)$	

Tabla 2.4: Tabla análisis de la varianza para 2×2.

Para las estimaciones de los valores de la tabla usamos las mismas estimaciones vistas para el caso general adaptadas a nuestro caso.

$$\hat{\alpha} = 2\hat{\alpha_2} = 2(\bar{Y}_{2\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})$$
$$\hat{\beta} = 2\hat{\beta}_2 = 2(\bar{Y}_{\bullet 2\bullet} - \bar{Y}_{\bullet\bullet\bullet})$$
$$\hat{\alpha\beta} = 2(\hat{\alpha\beta})_{22} = 2(Y_{22\bullet} - \bar{Y}_{2\bullet\bullet} - \bar{Y}_{\bullet 2\bullet} + \bar{Y}_{\bullet\bullet\bullet})$$

Antes de terminar, hay que escribir los estadísticos con sus respectivas distribuciones F de *Snedecor*.

$$T_X \in F_{1,4(n-1)} \ ; \ T_Z \in F_{1,4(n-1)} \ ; \ T_{XZ} \in F_{1,4(n-1)}$$

Ahora, solo queda fijar un nivel de significación $\alpha \in (0,1)$ para cada uno de los contrastes. Siendo f_α el cuantil que deja a su derecha una probabilidad α, rechazaremos H_{0c} con $c \in \{1,2,3\}$ cuando el estadístico T_v con $v \in \{X, Z, XZ\}$ sea más grande que f_α.

2.4. Ejemplo de aplicación

Una vez vista la construcción de los tres contrastes para el ANOVA de dos vías, aplicamos dicha teoría a nuestro ejemplo de la Sección 2.1 . Vemos que los dos primeros contrastes representan el estudio de $grams \sim smoke$ y $grams \sim black$, los cuales salieron significativos, por lo que nuestro modelo va a partir de la forma $grams \sim smoke + black$.

Vamos a escribir la distribución de los datos del ejemplo en función de las cuatro posibles combinaciones de las variables, acompañado del número de observaciones de las celdas.

	$smokeFALSE$	$smokeTRUE$	
$blackFALSE$	Y_{11n_1}	Y_{12n_2}	453
$blackTRUE$	Y_{21n_3}	Y_{22n_4}	662
	846	269	1115

Tabla 2.5: Tabla de la distribución con los datos del ejemplo siguiendo el modelo ANOVA de dos vías.

- $n_1 = 356$ es el número de observaciones que cumplen $smoke, black = FALSE$.
- $n_2 = 97$ observaciones que cumplen $smoke = TRUE$ y $black = FALSE$.

- $n_3 = 490$ observaciones que cumplen $smoke = FALSE$ y $black = TRUE$.
- $n_4 = 172$ observaciones que cumplen $smoke, black = TRUE$.

Una vez construída la Tabla 2.5, nos queda por ver cual será la fórmula de los Y_{11n_1}, Y_{12n_2}, Y_{21n_3}, Y_{22n_4}. Antes se dijo que partíamos del modelo $grams \sim smoke + black$, por lo que solo nos queda por comprobar si nos quedaremos con dicho modelo o con el modelo en el que el efecto de interacción es relevante $grams \sim smoke + black + smoke : black$.

$$\begin{cases} Y \sim X + Z + X : Z + \epsilon \\ grams \sim smoke + black + smoke : black \\ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \\ k = 1, \ldots, n_{ij} \quad i = 1, 2 \quad j = 1, 2 \quad \epsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

$$\begin{cases} Y \sim X + Z + \epsilon \\ grams \sim smoke + black \\ Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \\ k = 1, \ldots, n_{ij} \quad i = 1, 2 \quad j = 1, 2 \quad \epsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

Procedemos a construir el contraste para decidir qué modelo es el más adecuado.

$$\begin{cases} H_0 : grams \sim smoke + black \\ H_a : grams \sim smoke + black + smoke : black \end{cases}$$

$$\begin{cases} H_0 : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{22} = 0 \\ H_a : \exists \ i, j, p, l \ \text{ tal que } \ (\alpha\beta)_{ij} \neq (\alpha\beta)_{pl} \\ i, j, p, l = 1, 2 \end{cases}$$

Construímos el estadístico de contraste $T \in F_{1,1111}$.

$$T = 1.6334$$

Este estadístico lleva una probabilidad acumulada de $p - valor = 0.2015$, el cual no es significativo ($p - valor >$0.05), por lo que no hay indicios para rechazar H_0 y concluímos que el efecto de interacción de las variables explicativas no resulta significativo sobre la respuesta.

Por lo tanto, nos quedamos con el modelo sin interacción y sabemos que nuestra variable respuesta se representará de la siguiente forma. Este será nuestro modelo definitivo, el modelo en el que se representa de la forma más apropiada el efecto de las variables *smoke* y *black* sobre *grams*.

$$\begin{cases} grams \sim smoke + black \\ Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \;\; k = 1, \ldots, n_{ij} \;\; i, j = 1, 2 \end{cases}$$

Antes de estimar los parámetros debemos validar el modelo.

- **Independencia**: asumimos independencia de las observaciones. Cada observación corresponde al parto de una madre diferente, no hay indicios de que estén relacionadas entre sí.
- **Homocedasticidad**: haciendo el test Levene a la variable *smoke* nos da unos estadístico de contraste $L = 5.4284$ y un $p-valor = 0.01999$ y haciéndolo a la variable *black* un $L = 4.1116$ y $p-valor = 0.04283$. Como ambos valores cumplen $p-valor < 0.05$ podemos asumir que se cumple la hipótesis de homocedasticidad en el modelo para ambas variables.
- **Normalidad**: hacemos el test Kolmogorov-Smirnov. Construímos la Tabla 2.6 en la que mostramos el $p-valor$ del test para cada grupo.

	$smokeFALSE$	$smokeTRUE$	
$blackFALSE$	0.1127	0.5116	453
$blackTRUE$	0.0013	0.0135	662
	846	269	1115

Tabla 2.6: Tabla de los $p-valor$ del test Kolmogorov-Smirnov para los cuatro grupos.

Tenemos dos casos en los que no se cumple la hipótesis de normalidad ($p-valor < 0.05$). Procedemos a calcular los datos atípicos con los residuos estandarizados de cada una de las observaciones. Obtenemos un total de 55 datos atípicos, los eliminamos de la muestra y repetimos el test Kolmogorov-Smirnov. Costruímos la Tabla 2.7 con los $p-valor$ de la muestra sin los datos atípicos.

Estos nuevos valores son todos superiores a 0.05, por lo que concluímos que los datos de cada uno de los grupos pertenecen a una distribución normal. Además, reescribimos los valores n_i con $i \in \{1, 2, 3, 4\}$. $n_1 = 348$, $n_2 = 90$, $n_3 = 462$ y $n_4 = 160$.

Como hemos tenido que modificar los datos para poder validar el modelo, vamos a tener que reformular el test F en el contraste (2.7), ya que los nuevos datos pueden dar

	$smokeFALSE$	$smokeTRUE$	
$blackFALSE$	0.2481	0.7043	438
$blackTRUE$	0.6583	0.8277	622
	810	250	1060

Tabla 2.7: Tabla de los $p-valor$ del test Kolmogorov-Smirnov para los cuatro grupos con la muestra sin los datos atípicos.

un resultado diferente. Tomando la muestra sin los datos atípicos calculados, veremos si resulta relevante el efecto de interacción. Es significativo remarcar que el estadístico de contraste $T' \in F_{1,1056}$ seguirá una distribución F de *Snedecor* con unos grados de libertad distintos al estadístico T de la muestra con los datos atípicos. Este es el motivo por el que es necesario repetir este proceso, los grados de libertad están directamente relacionados con el tamaño de la muestra.

$$T' = 0.4847$$

El $p-valor$ asociado es el valor 0.4865. Por lo tanto, vemos que la conclusión no cambia, la interacción no resulta relevante. Sin embargo, es significativo remarcar que el $p-valor$ de los nuevos datos es más del doble que el anterior, por lo que se acentúa significativamente la conclusión con la muestra sin los datos atípicos.

Una vez construído y validado el modelo, vamos a mostrar las estimaciones de los parámetros a partir de la muestra sin los datos atípicos. Vamos a mostrar el modelo con interacción y sin interacción para ver, de una forma más visual, el hecho de que no resulta significativa la diferencia entre ambos modelos. Aquí, se representa usando la parametrización respecto un grupo de referencia, el software R usa por defecto los grupos de referencia $smoke = FALSE$ y $black = FALSE$, por lo que el intercepto será la media de las observaciones de dichos grupos combinados.

$$\begin{cases} grams \sim smoke + black + smoke : black \\ \hat{Y} = 3513.5 - 303.4 smokeTRUE - 292.1 blackTRUE + 46.9 smokeTRUE * blackTRUE \end{cases}$$

$$\begin{cases} grams \sim smoke + black \\ \hat{Y} = 3507.5 - 274.2 smokeTRUE - 281.5 blackTRUE \end{cases}$$

Reescribimos los datos de la Tabla 2.5.

$$
\begin{aligned}
Y_{11n_1} &= 3507.5 \\
Y_{21n_2} &= 3507.5 - 281.5 \\
Y_{12n_3} &= 3507.5 - 274.2 \\
Y_{22n_4} &= 3507.5 - 274.2 - 281.5
\end{aligned}
$$

A continuación, mostramos dos tablas, Tabla 2.9 y Tabla 2.8, para cada uno de los dos modelos, con interacción y sin ella. En estas tablas, vienen las estimaciones de los parámetros acompañados por sus intervalos de confianza y la variable p corresponde al nivel de significación.

	smoke + black		
Parámetro	Estimación	I.C.	p
Intercepto	3507.48	3463.28 – 3551.67	<0.001
smokeTRUE	-274.16	-338.15 – -210.16	<0.001
blackTRUE	-281.55	-336.72 – -226.37	<0.001
Observaciones	1060		

Tabla 2.8: Tabla de la estimación del modelo $grams \sim smoke + black$. I.C.: intervalo de confianza al 95 %. p: $p - valor$ para el constraste de significación.

	smoke + black + smoke:black		
Parámetro	Estimación	I.C.	p
Intercepto	3513.49	3466.15– 3560.83	<0.001
smokeTRUE	-303.44	-407.8 – -199.0	<0.001
blackTRUE	-292.1	-354.78 – -229.41	<0.001
smokeTRUE*blackTRUE	46.9	-85.28 – 179.07	0.486
Observaciones	1060		

Tabla 2.9: Tabla de la estimación del modelo $grams \sim smoke + black + smoke : black$. I.C.: intervalo de confianza al 95 %. p: $p - valor$ para el constraste de significación.

Es interesante exponer estas tablas para poder ver la similitud de las estimaciones en ambos modelos. Es el hecho de que no es significativa la interacción lo que nos acaba decantando por el modelo sin interacción.

Para terminar, comentamos un poco cómo funciona la predicción. El proceso de predicción en este tipo de modelos es muy simple, ya que los posibles valores de las predictoras coinciden con las medias de los posibles grupos, o combinación de los niveles del factor o factores que aparezcan en el modelo final. En este caso solo tiene sentido predecir los valores de la media, ya que nuestro objetivo es el estudio de dichas medias.

Mostramos ahora las predicciones para cada uno de los casos,

- Un bebé nacido de una madre no fumadora y de raza no negra tendrá un peso medio de 3507.5 gramos.
- Un bebé nacido de una madre no fumadora y de raza negra tendrá un peso medio de 3226 gramos.
- Un bebé nacido de una madre fumadora y de raza no negra tendrá un peso medio de 3233.3 gramos.
- Un bebé nacido de una madre fumadora y de raza negra tendrá un peso medio de 2951.8 gramos

2.5. Simulación

En este capítulo, hemos visto cómo funciona el test F en el modelo ANOVA de dos vías, cómo nos sirvió para estudiar el efecto de las variables *smoke* y *black* sobre la respuesta *grams* de forma individual y el efecto de interacción de ambas variables. En dicho estudio, hemos sacado conclusiones claras y concisas para poder decretar que el modelo $grams \sim smoke + black$ es el más apropiado, pero, ¿qué hubiese pasado si hubiésemos tenido una muestra mucho menor? Revisaremos el efecto de distintos factores sobre el test F al modelo ANOVA de dos vías, de una forma similar a cómo lo hicimos en el Capítulo 1.

Generamos tres variables, dos categóricas (G y H) y una continua (X). La variable continua la generamos siguiendo una distribución normal de media $\mu = 0$ y varianza $\sigma^2 = 1$ y, de las variables categóricas, generamos tres grupos en cada una con la condición de que la suma de las observaciones de los grupos de G coincida la suma de las observaciones de los grupos de H, es decir, si llamamos n_1, n_2 y n_3 al tamaño de los grupos 1, 2 y 3 de la variable G y m_1, m_2 y m_3 al tamaño de los grupos 4, 5 y 6 de la variable H, se tiene que cumplir $n = n_1 + n_2 + n_3 = m_1 + m_2 + m_3$, donde n es el número de observaciones de la variable continua.

2.5.1. Tamaño

$$\begin{cases} H_0 : X \sim 1 \\ H_a : X \sim G + H \end{cases} \tag{2.7}$$

Vamos a empezar estudiando el efecto del tamaño sobre el test F en el modelo ANOVA de dos vías. Al tener dos variables categóricas con tres grupos cada una, tendríamos que estudiar los cambios en los grupos de cada variable de forma separada, sin embargo, como nuestro contraste (2.7) compara el modelo de la variable respuesta X sobre sí misma y, el modelo $X \sim G+H$, es equivalente simular los cambios tanto en los grupos de G como en los de H, por lo que solo haremos la simulación sobre una de ellas. Además, dependiendo qué variable se simule, va a condicionar cómo se genere la variable continua X, si estudiamos los cambios en la variable G, entonces se genera X en función de los tamaños n_1, n_2, y n_3, pero, si estudiamos los cambios de tamaño en la variable H, entonces se genera X en función de los tamaños m_1, m_2 y m_3. El objetivo será ver, si repetimos el test F 500 veces, qué proporción de veces (porcentaje por uno) rechazamos H_0, bajo los niveles de significación usuales, 10 %, 5 % y 1 %.

La simulación será sobre la variable H, por lo tanto, dejaremos fijos los tamaños de los grupos de G. A continuación, mostramos tres tablas, en cada una de ellas se dejaran fijos unos tamaños distintos de G, además, se generará X en función de m_1, m_2 y m_3, tal que $m_1 + m_2 + m_3 = 600$. Los tamaños de los grupos de G serán: en la Tabla 2.10 $n_1 = n_2 = n_3 = 200$, en la Tabla 2.11 $n_1 = 100$, $n_2 = 200$ y $n_3 = 300$ y en la Tabla 2.12 $n_1 = 10$, $n_2 = 90$ y $n_3 = 500$. En la variable $Tamaño$ de las tablas, se mostrarán tres valores correspondientes al tamaño de los grupos de la variable H, 4, 5, y 6, respectivamente.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.130	0.068	0.008
200, 200, 200	0.104	0.052	0.000
100, 200, 300	0.106	0.064	0.012
100, 200, 300	0.092	0.046	0.008
10, 90, 500	0.118	0.066	0.012
10, 90, 500	0.102	0.040	0.006

Tabla 2.10: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 200$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.090	0.054	0.016
200, 200, 200	0.114	0.062	0.010
100, 200, 300	0.122	0.06	0.014
100, 200, 300	0.116	0.060	0.012
10, 90, 500	0.116	0.064	0.016
10, 90, 500	0.110	0.056	0.006

Tabla 2.11: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 100$, $n_2 = 200$ y $n_3 = 300$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.098	0.058	0.010
200, 200, 200	0.090	0.046	0.010
100, 200, 300	0.110	0.048	0.004
100, 200, 300	0.112	0.056	0.008
10, 90, 500	0.118	0.054	0.002
10, 90, 500	0.106	0.052	0.012

Tabla 2.12: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 10$, $n_2 = 90$ y $n_3 = 500$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Vemos que en las Tablas 2.10, 2.11 y 2.12 se mantienen unos porcentajes bastante adecuados a los niveles de significación correspondientes. Por lo tanto, concluímos que el test F, aplicado al modelo ANOVA de dos vías, está bien calibrado a cambios en los grupos de las variables habiendo un total de 600 observaciones.

Ahora, repetimos el proceso de construcción de tres tablas pero, esta vez, habrá una muestra total de 100 observaciones, es decir, $n_1 + n_2 + n_3 = m_1 + m_2 + m_3 = 300$. Los tamaños de los grupos de G serán: en la Tabla 2.13 $n_1 = n_2 = n_3 = 100$, en la Tabla 2.14 $n_1 = 40$, $n_2 = 60$ y $n_3 = 200$ y en la Tabla 2.15 $n_1 = 5$, $n_2 = 95$ y $n_3 = 200$. En la variable *Tamaño* de las tablas, se mostrarán tres valores correspondientes al tamaño de los grupos de la variable H, 4, 5, y 6, respectivamente.

Tamaño	< 0.1	< 0.05	< 0.01
100, 100, 100	0.100	0.052	0.010
100, 100, 100	0.106	0.054	0.016
40, 60, 200	0.106	0.056	0.010
40, 60, 200	0.088	0.044	0.004
5, 95, 200	0.096	0.056	0.010
5, 95, 200	0.112	0.046	0.014

Tabla 2.13: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 100$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Tamaño	< 0.1	< 0.05	< 0.01
100, 100, 100	0.100	0.048	0.004
100, 100, 100	0.112	0.066	0.014
40, 60, 200	0.104	0.062	0.012
40, 60, 200	0.088	0.036	0.008
5, 95, 200	0.114	0.062	0.008
5, 95, 200	0.086	0.040	0.010

Tabla 2.14: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 40$, $n_2 = 60$, $n_3 = 200$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Tamaño	< 0.1	< 0.05	< 0.01
100, 100, 100	0.088	0.052	0.012
100, 100, 100	0.112	0.040	0.004
40, 60, 200	0.088	0.052	0.010
40, 60, 200	0.102	0.058	0.016
5, 95, 200	0.106	0.056	0.016
5, 95, 200	0.102	0.060	0.008

Tabla 2.15: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 5$, $n_2 = 95$ y $n_3 = 200$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Vemos que en las Tablas 2.13, 2.14 y 2.15 se mantienen unos porcentajes bastante adecuados a los niveles de significación correspondientes. Por lo tanto, concluímos que el test F, aplicado al modelo ANOVA de dos vías, está bien calibrado a cambios en los grupos de las variables habiendo un total de 300 observaciones.

Para terminar con el estudio del efecto del tamaño, repetimos el proceso de construcción de tres tablas, pero, con una muestra total de 99 observaciones, es decir, $n_1 + n_2 + n_3 = m_1 + m_2 + m_3 = 99$. Los tamaños fijos de los grupos de G serán: en la Tabla 2.16 $n_1 = n_2 = n_3 = 33$, en la Tabla 2.17 $n_1 = 15$, $n_2 = 30$ y $n_3 = 54$ y en la Tabla 2.18 $n_1 = 3$, $n_2 = 17$ y $n_3 = 79$. En la variable $Tamaño$ de las tablas, se mostrarán tres valores correspondientes al tamaño de los grupos de la variable H, 4, 5, y 6, respectivamente.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.108	0.050	0.014
33, 33, 33	0.084	0.044	0.004
15, 30, 54	0.124	0.070	0.020
15, 30, 54	0.094	0.046	0.006
3, 17, 79	0.112	0.060	0.016
3, 17, 79	0.108	0.052	0.010

Tabla 2.16: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 33$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.108	0.054	0.010
33, 33, 33	0.104	0.060	0.010
15, 30, 54	0.088	0.036	0.012
15, 30, 54	0.100	0.058	0.014
3, 17, 79	0.104	0.046	0.014
3, 17, 79	0.088	0.036	0.010

Tabla 2.17: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 15$, $n_2 = 30$ y $n_3 = 54$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.084	0.040	0.008
33, 33, 33	0.100	0.046	0.008
15, 30, 54	0.082	0.044	0.012
15, 30, 54	0.092	0.044	0.002
3, 17, 79	0.088	0.038	0.012
3, 17, 79	0.106	0.050	0.016

Tabla 2.18: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 3$, $n_2 = 17$ y $n_3 = 79$ para distintos niveles de significación. La continua sigue una distribución $N(0, 1)$.

Vemos que en las Tablas 2.16, 2.17 y 2.18 se mantienen unos porcentajes bastante adecuados a los niveles de significación correspondientes. Por lo tanto, concluímos que el test F, aplicado al modelo ANOVA de dos vías, está bien calibrado a cambios en los grupos de las variables habiendo un total de 99 observaciones.

2.5.2. Varianza

Para el estudio del efecto varianza sobre el test F en el modelo ANOVA de dos vías, vamos a partir del mismo contraste de la subsección anterior, (2.7). Los cambios de la varianza se aplican sobre la variable respuesta X, los tres grupos generados tendrán distinta varianza. Al igual que antes, generaremos X a partir de los tamaños m_1, m_2 y m_3, por lo que generaremos distintas varianzas para distintos tamaños y ver si el test F está bien calibrado a cambios de tamaño y varianza. El objetivo será ver, si repetimos el test F 500 veces, qué proporción de veces (porcentaje por uno) rechazamos H_0, bajo los niveles de significación usuales, 10 %, 5 % y 1 %.

Construiremos tablas en las que va a variar el número de observaciones y cómo se distribuyan dichas observaciones en los grupos de las variables G y H. Mostraremos seis tablas, en las tres primeras tendremos una muestra total de 600 observaciones, $n_1 + n_2 + n_3 = 600 = m_1 + m_2 + m_3$ y en cada tabla tendremos una distribución distinta de las observaciones. En la Tabla 2.19 $n_1 = n_2 = n_3 = 200 = m_1 = m_2 = m_3$, en la Tabla 2.20 $n_1 = m_1 = 100$, $n_2 = m_2 = 200$ y $n_3 = m_3 = 300$ y en la Tabla 2.21 $n_1 = m_1 = 10$, $n_2 = m_2 = 90$ y $n_3 = m_3 = 500$. En el parámetro σ de las tablas se mostrarán tres valores correspondientes a las desviaciones típicas de los tres grupos que forman la variable X.

σ	< 0.1	< 0.05	< 0.01
10, 10, 10	0.092	0.042	0.004
10, 10, 10	0.088	0.040	0.008
1, 10, 50	0.118	0.086	0.038
1, 10, 50	0.114	0.072	0.028
20, 20, 50	0.138	0.092	0.036
20, 20, 50	0.130	0.090	0.036

Tabla 2.19: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 200 = m_1 = m_2 = m_3$ para distintos niveles de significación. La continua sigue una distribución $N(0, \sigma^2)$ con distinta varianza.

σ	< 0.1	< 0.05	< 0.01
10, 10, 10	0.110	0.048	0.012
10, 10, 10	0.114	0.058	0.014
1, 10, 50	0.026	0.010	0.002
1, 10, 50	0.034	0.016	0.004
20, 20, 50	0.034	0.018	0.002
20, 20, 50	0.022	0.012	0.000

Tabla 2.20: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = m_1 = 100$, $n_2 = m_2 = 200$ y $n_3 = m_3 = 300$ para distintos niveles de significación. La continua sigue una distribución $N(0, \sigma^2)$ con distinta varianza.

σ	< 0.1	< 0.05	< 0.01
10, 10, 10	0.108	0.048	0.008
10, 10, 10	0.106	0.060	0.008
1, 10, 50	0.000	0.000	0.000
1, 10, 50	0.00	0.000	0.000
20, 20, 50	0.000	0.000	0.000
20, 20, 50	0.000	0.000	0.000

Tabla 2.21: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = m_1 = 10$, $n_2 = m_2 = 90$ y $n_3 = m_3 = 500$ para distintos niveles de significación. La continua sigue una distribución $N(0, \sigma^2)$ con distinta varianza.

Vemos que en la Tabla 2.19 se mantienen unos porcentajes adecuados pero, en las Tablas 2.20 y 2.21, vemos que al variar el tamaño de los grupos y las varianzas se alteran significativamente los porcentajes. Al cambiar la varianza de cada grupo se acepta la hipótesis H_0 en cada uno de los 500 casos de la simulación. Por lo tanto, concluímos que el test F no está bien calibrado grupos que estén desbalanceados de tamaño y varianza.

Repetiremos la simulación anterior con un menor tamaño de muestra, a ver si de esta forma sacamos los mismos resultados que en la simulación con 600 observaciones. Construímos las siguientes tres tablas con una muestra total de 99 observaciones. En la Tabla 2.22 $n_1 = n_2 = n_3 = 33 = m_1 = m_2 = m_3$, en la Tabla 2.23 $n_1 = m_1 = 15$, $n_2 = m_2 = 30$ y $n_3 = m_3 = 54$ y en la Tabla 2.24 $n_1 = m_1 = 3$, $n_2 = m_2 = 17$ y $n_3 = m_3 = 79$. En el parámetro σ de las tablas se mostrarán tres valores correspondientes a las desviaciones típicas de los tres grupos a partir de los que se genera la variable X.

σ	< 0.1	< 0.05	< 0.01
10, 10, 10	0.078	0.038	0.010
10, 10, 10	0.098	0.044	0.012
1, 10, 50	0.138	0.094	0.038
1, 10, 50	0.138	0.104	0.044
20, 20, 50	0.146	0.106	0.030
20, 20, 50	0.104	0.080	0.030

Tabla 2.22: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 33 = m_1 = m_2 = m_3$ para distintos niveles de significación. La continua sigue una distribución $N(0, \sigma^2)$ con distinta varianza.

Claramente no ha cambiado la conclusión tras construir las Tablas 2.22, 2.23 y 2.24. Vemos que mientras la varianza de los grupos es la misma el test F está bien calibrado pero, al cambiar la varianza de los grupos, se produce un cambio muy significativo de los porcentajes. Concluímos que el test F está calibrado para el modelo ANOVA de dos vías unicamente cuando la varianza entre los grupos sea la misma.

σ	< 0.1	< 0.05	< 0.01
10, 10, 10	0.096	0.044	0.018
10, 10, 10	0.104	0.046	0.012
1, 10, 50	0.020	0.004	0.000
1, 10, 50	0.034	0.016	0.002
20, 20, 50	0.018	0.008	0.002
20, 20, 50	0.014	0.006	0.002

Tabla 2.23: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = m_1 = 15$, $n_2 = m_2 = 30$ y $n_3 = m_3 = 54$ para distintos niveles de significación. La continua sigue una distribución $N(0, \sigma^2)$ con distinta varianza.

σ	< 0.1	< 0.05	< 0.01
10, 10, 10	0.082	0.054	0.014
10, 10, 10	0.090	0.050	0.010
1, 10, 50	0.000	0.000	0.000
1, 10, 50	0.000	0.000	0.000
20, 20, 50	0.000	0.000	0.000
20, 20, 50	0.000	0.000	0.000

Tabla 2.24: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = m_1 = 3$, $n_2 = m_2 = 17$ y $n_3 = m_3 = 79$ para distintos niveles de significación. La continua sigue una distribución $N(0, \sigma^2)$ con distinta varianza.

2.5.3. Distribuciones

Por último, vamos a estudiar cómo afecta al test F el hecho de tener una variable continua X que tenga observaciones pertenecientes a distintas distribuciones. Vamos a partir del contraste (2.7), haremos dos simulaciones, en una veremos el efecto de mezclar grupos que sigan una distribución normal y exponencial y, en la otra, grupos que sigan distribuaciones normales y χ^2, dichas distribuciones se aplican a la variable X. Generaremos X a partir de los tamaños m_1, m_2 y m_3. Veremos el efecto de las distribuciones en el test F con distintos tamaños de los grupos de G y H. El objetivo será ver, si repetimos el test F 500 veces, qué proporción de veces (porcentaje por uno) rechazamos H_0, bajo los niveles de significación usuales, 10 %, 5 % y 1 %.

Primero, haremos una simulación mezclando grupos siguiendo distribuciones normales y exponenciales. Construímos X tal que dos de sus grupos sigan una distribución normal de media $\mu = 1$ y varianza $\sigma^2 = 1$ y, el grupo que queda, siga una distribución exponencial con $\lambda = 1$. Vamos a construir seis tablas, en cada trío de tablas tendremos un número de observaciones distinto. Analicemos las tres primeras tablas, el total de observaciones será 600, es decir, $n_1 + n_2 + n_3 = 600 = m_1 + m_2 + m_3$. Los tamaños de los grupos de G serán: en la Tabla 2.25 $n_1 = n_2 = n_3 = 200$, en la Tabla 2.26 $n_1 = 100$, $n_2 = 200$ y $n_3 = 300$ y en la Tabla 2.27 $n_1 = 10$, $n_2 = 90$ y $n_3 = 500$. En la variable $Tamaño$ de las tablas se mostrarán tres valores correspondientes a los tamaños de los tres grupos de H, 4, 5 y 6, respectivamente.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.086	0.036	0.006
200, 200, 200	0.090	0.046	0.016
100, 200, 300	0.102	0.050	0.004
100, 200, 300	0.106	0.060	0.008
10, 90, 500	0.110	0.066	0.010
10, 90, 500	0.094	0.046	0.008

Tabla 2.25: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 200$ para distintos niveles de significación. Dos grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución $Exp(1)$.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.106	0.046	0.012
200, 200, 200	0.108	0.062	0.008
100, 200, 300	0.080	0.034	0.000
100, 200, 300	0.114	0.060	0.018
10, 90, 500	0.098	0.046	0.012
10, 90, 500	0.088	0.046	0.004

Tabla 2.26: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 100$, $n_2 = 200$ y $n_3 = 300$ para distintos niveles de significación. Dos grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución $Exp(1)$.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.104	0.052	0.014
200, 200, 200	0.094	0.050	0.010
100, 200, 300	0.084	0.042	0.010
100, 200, 300	0.104	0.048	0.016
10, 90, 500	0.106	0.064	0.010
10, 90, 500	0.086	0.048	0.008

Tabla 2.27: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 10$, $n_2 = 90$ y $n_3 = 500$ para distintos niveles de significación. Dos grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución $Exp(1)$.

Vemos que en las Tablas 2.25, 2.26 y 2.27 se mantienen unos porcentajes muy adecuados con los niveles de significación correspondientes. Por lo tanto, concluímos que el test F está bien calibrado a mezclas de grupos siguiendo distribuciones normales y exponenciales con una muestra total de 600 observaciones.

Ahora, repetimos el mismo proceso con una muestra de 99 observaciones, es decir, $n_1 + n_2 + n_3 = 99 = m_1 + m_2 + m_3$. Construímos X, igual que antes, con dos grupos siguiendo una distribución normal de media $\mu = 1$ y varianza $\sigma^2 = 1$ y un grupo una distribución exponencial con $\lambda = 1$. Los tamaños de los grupos de G en la Tabla 2.28 son $n_1 = n_2 = n_3 = 33$, en la Tabla 2.29 $n_1 = 15$, $n_2 = 30$ y $n_3 = 54$ y en la Tabla 2.30 $n_1 = 3$, $n_2 = 17$ y $n_3 = 79$. En la variable $Tamaño$ de las tablas se mostrarán tres valores, correspondientes al tamaño de los grupos de la variable H, 4, 5, y 6, respectivamente.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.112	0.060	0.012
33, 33, 33	0.100	0.064	0.016
15, 30, 54	0.098	0.058	0.008
15, 30, 54	0.096	0.050	0.010
3, 17, 79	0.084	0.052	0.008
3, 17, 79	0.094	0.056	0.012

Tabla 2.28: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 33$ para distintos niveles de significación. Dos de los grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución $Exp(1)$.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.124	0.056	0.004
33, 33, 33	0.116	0.054	0.016
15, 30, 54	0.116	0.050	0.006
15, 30, 54	0.104	0.044	0.012
3, 17, 79	0.092	0.040	0.006
3, 17, 79	0.092	0.056	0.002

Tabla 2.29: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 15$, $n_2 = 30$ y $n_3 = 54$ para distintos niveles de significación. Dos de los grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución $Exp(1)$.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.110	0.048	0.004
33, 33, 33	0.116	0.072	0.016
15, 30, 54	0.110	0.058	0.012
15, 30, 54	0.100	0.052	0.012
3, 17, 79	0.090	0.040	0.006
3, 17, 79	0.088	0.038	0.004

Tabla 2.30: Tabla de porcentajes de rechazo del modelo $X \sim G+H$ variando H con $n_1 = 3$, $n_2 = 17$ y $n_3 = 79$ para distintos niveles de significación. Dos de los grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución $Exp(1)$.

Vemos que en las Tablas 2.28, 2.29 y 2.30 se mantienen unos porcentajes adecuados. Por lo tanto concluímos que el test F está bien calibrado para muestras que sigan una distribución normal y exponencial.

Para terminar con la simulación, vamos a estudiar cómo reacciona el test F cuando se mezclan grupos que sigan una distribución normal y χ^2_1. Construímos X con dos de sus grupos siguiendo una distribución normal de media $\mu = 1$ y varianza $\sigma^2 = 1$ y ,el otro grupo, una distribución χ^2 con un grado de libertad, $df = 1$. Vamos a construir seis tablas, en cada trío de tablas tendremos un número total de observaciones distinto fijado. En las tres primeras tablas tendremos un total de 600 observaciones, es decir, $n_1+n_2+n_3 = 600 = m_1+m_2+m_3$. Los tamaños de los grupos de G serán: en la Tabla 2.31 $n_1 = n_2 = n_3 = 200$, en la Tabla 2.32 $n_1 = 100$, $n_2 = 200$ y $n_3 = 300$ y en la Tabla 2.33 $n_1 = 10$, $n_2 = 90$ y $n_3 = 500$. En la variable $Tamaño$ de las tablas se mostrarán tres valores correspondientes

a los tamaños de los tres grupos de H, 4, 5 y 6, respectivamente.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.084	0.050	0.006
200, 200, 200	0.074	0.044	0.008
100, 200, 300	0.096	0.046	0.014
100, 200, 300	0.102	0.058	0.010
10, 90, 500	0.098	0.052	0.012
10, 90, 500	0.082	0.048	0.010

Tabla 2.31: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 200$ para distintos niveles de significación. Dos grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución χ^2_1.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.088	0.052	0.010
200, 200, 200	0.068	0.030	0.010
100, 200, 300	0.116	0.066	0.018
100, 200, 300	0.126	0.072	0.014
10, 90, 500	0.140	0.090	0.018
10, 90, 500	0.124	0.072	0.016

Tabla 2.32: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 100$, $n_2 = 200$ y $n_3 = 300$ para distintos niveles de significación. Dos grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución χ^2_1.

Vemos que, según aumentamos la amplitud de los tamaños en las Tablas 2.31, 2.32 y 2.33, tenemos unos resultados con menos constancia en sus respectivos porcentajes, por lo que no podemos asegurar que esté bien calibrado el test F.

Tamaño	< 0.1	< 0.05	< 0.01
200, 200, 200	0.046	0.024	0.000
200, 200, 200	0.042	0.010	0.002
100, 200, 300	0.058	0.024	0.006
100, 200, 300	0.066	0.024	0.002
10, 90, 500	0.156	0.094	0.028
10, 90, 500	0.164	0.102	0.028

Tabla 2.33: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 10$, $n_2 = 90$ y $n_3 = 500$ para distintos niveles de significación. Dos grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución χ^2_1.

Ahora, repetimos el proceso con un total de 99 observaciones, es decir, $n_1 + n_2 + n_3 = 99 = m_1 + m_2 + m_3$. Construímos X, igual que antes, con dos grupos siguiendo una distribución normal de media $\mu = 1$ y varianza $\sigma^2 = 1$ y un grupo una distribución χ^2 con un grado de libertad, $df = 1$. Los tamaños de los grupos de G serán: en la Tabla 2.34 $n_1 = n_2 = n_3 = 33$, en la Tabla 2.35 $n_1 = 15$, $n_2 = 30$ y $n_3 = 54$ y en la Tabla 2.36 $n_1 = 3$, $n_2 = 17$ y $n_3 = 79$. En la variable $Tamaño$ de las tablas, se mostrarán tres valores correspondientes al tamaño de los grupos de la variable H, 4, 5, y 6, respectivamente.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.124	0.056	0.014
33, 33, 33	0.112	0.058	0.016
15, 30, 54	0.092	0.046	0.008
15, 30, 54	0.070	0.044	0.008
3, 17, 79	0.100	0.048	0.014
3, 17, 79	0.092	0.024	0.006

Tabla 2.34: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = n_2 = n_3 = 99$ para distintos niveles de significación. Dos de los grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución χ^2_1.

Vemos que en las Tablas 2.34, 2.35 y 2.36 hay bastante amplitud de valores, no obtenemos unos resultados muy certeros. Por lo tanto, concluímos que el calibrado de F para el modelo ANOVA de dos vías no es el más apropiado.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.086	0.040	0.004
33, 33, 33	0.114	0.066	0.002
15, 30, 54	0.126	0.068	0.024
15, 30, 54	0.108	0.056	0.002
3, 17, 79	0.150	0.070	0.016
3, 17, 79	0.140	0.074	0.014

Tabla 2.35: Tabla de porcentajes de rechazo del modelo $X \sim G + H$ variando H con $n_1 = 15$, $n_2 = 30$ y $n_3 = 54$ para distintos niveles de significación. Dos de los grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución χ^2_1.

Tamaño	< 0.1	< 0.05	< 0.01
33, 33, 33	0.068	0.052	0.014
33, 33, 33	0.074	0.030	0.004
15, 30, 54	0.098	0.044	0.004
15, 30, 54	0.084	0.052	0.012
3, 17, 79	0.168	0.092	0.016
3, 17, 79	0.136	0.074	0.010

Tabla 2.36: Tabla de porcentajes de rechazo del modelo $X \sim G+H$ variando H con $n_1 = 3$, $n_2 = 17$ y $n_3 = 79$ para distintos niveles de significación. Dos de los grupos de la continua siguen una distribución $N(1, 1)$ y el otro una distribución χ^2_1.

Bibliografía

[1] Faraway, J. (2004), *Linear Models with R.* CRC press.

[2] Faraway, J. (2016), *Functions and Datasets for Books by Julian Faraway.* R package version 1.0.7, https://CRAN.R-project.org/package=faraway.

[3] Peña Sánchez de Rivera, D. (1987), *Estadística Modelos y Métodos.* Alianza.

[4] Peña Sánchez de Rivera, D. (2002), *Regresión y Diseño de Experimentos.* Alianza.

[5] R Core Team (2019), *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.