Universidad Nacional de Educación a Distancia
Escuela Técnica Superior de Informática
Máster en Ingeniería y Ciencia de Datos

Trabajo Fin de Máster

Utilización de técnicas multivariantes para el estudio del aprendizaje de la mejora de la accesibilidad en el subtitulado de vídeos

Autor: Javier Pérez Arteaga
Directores: Emilio Letón Molina
Jorge Pérez Martín
Fecha de realización: 2023-10-03

This document is reproducible thanks to:

- LaTeXand its class memoir (http://www.ctan.org/pkg/memoir).
- R (http://www.r-project.org/) and RStudio (http://www.rstudio.com/)
- bookdown (http://bookdown.org/) and memoiR (https://ericmarcon.github.io/memoiR/)

Name of the owner of the logo

http://www.company.com

Resumen

TODO: Incluir un resumen del trabajo.

Agradecimientos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed malesuada nulla augue, ac facilisis risus pretium a. Ut bibendum risus id ex fermentum, at accumsan erat vulputate. In hac habitasse platea dictumst. Sed lobortis est a enim bibendum, ac pulvinar nulla aliquam. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Pellentesque efficitur justo id suscipit pretium. Proin iaculis sit amet nibh vel euismod. Aenean tincidunt faucibus ex, non vehicula ipsum tristique in. Fusce vel tincidunt lectus, vel rutrum nisi. Suspendisse malesuada lectus ac enim vehicula rhoncus. Nullam convallis justo in bibendum eleifend.

Phasellus vitae magna nec mi sagittis luctus vitae eu augue. Donec scelerisque laoreet arcu, eget tempor mi ultricies vel. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum at blandit ex. Vestibulum eu sagittis mauris. In hac habitasse platea dictumst. Duis eget ante vel lacus sollicitudin convallis quis eu velit. Sed auctor sem non nisi hendrerit, vel tincidunt tortor bibendum.

Índice

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

Capítulo 2

Estado del arte

Capítulo 3

Materiales y métodos

Capítulo 4

Métodos.

4.1 Fuente de datos.

Los datos proceden de la edición de 2022 del curso MOOC Materiales digitales accesibles de la UNED. Concretamente a los estudiantes matriculados se les propuso que realizaran una actividad voluntaria consistente en evaluar la calidad del subtitulado de dos vídeos. Los vídeos eran idénticos y se diferenciaban únicamente en la calidad del subtitulado. Los subtítulos de uno de los vídeos se realizaron (ver Pérez Martín et al. 2021; Molanes-López et al. 2021) siguiendo las guía Web Content Accessibility Guidelines 2.1 (WCAG 2.1) del W3C (World Wide Web Consortium). El otro vídeo tenía un subtitulado similar pero se introdujeron pequeñas deficiencias inapreciables para alguien que carezca de conocimientos sobre accesibilidad. Los estudiantes fueron clasificados en dos grupos. Al primer grupo se le presentó primero el vídeo correctamente subtitulado y luego el otro. El segundo grupo realizó la actividad cruzada: primero evaluó el vídeo mal subtitulado y luego el bien subtitulado. Tras ver cada uno de los vídeos, los estudiantes tuvieron la oportunidad de valorar la calidad del subtitulado realizando un test en escala de Likert de 18 items y 5 niveles cada item [1]. Los 18 items de Likert pretenden asegurar los criterios de la norma UNE 153010 (ver AENOR 2012).

En la Tabla 4.1 se muestran los 5 niveles de cada uno de los items de la escala de Likert:

[1] Para una descripción sobre cómo se debe realizar una escala de Likert consultar Guerra et al. (2016).

Cuadro 4.1: Niveles de los items de la escala de Likert.

values	levels
0	No sé / No contesto
1	Muy en desacuerdo
2	En desacuerdo
3	Neutral
4	De acuerdo
5	Muy de acuerdo

En la Tabla 4.2 se muestran los 18 items de la escala de Likert que se propuso a los alumnos para que evaluaran cada uno de los vídeos:

Cuadro 4.2: Items de la escala de Likert.

Item	Texto
Q01	La posición de los subtítulos.
Q02	El número de líneas por subtítulo.
Q03	La disposición del texto respecto a la caja donde se muestran los subtítulos.
Q04	El contraste entre los caracteres y el fondo.
Q05	La corrección ortográfica y gramatical.
Q06	La literalidad.
Q07	La identificación de los personajes.
Q08	La asignación de líneas a los personajes en los diálogos.
Q09	La descripción de efectos sonoros.
Q10	La sincronización de las entradas y salidas de los subtítulos.
Q11	La velocidad de exposición de los subtítulos.
Q12	El máximo número de caracteres por línea.
Q13	La legibilidad de la tipografía.
Q14	La separación en líneas diferentes de sintagmas nominales, verbales y preposicionales.
Q15	La utilización de puntos suspensivos.
Q16	La escritura de los números.
Q17	Las incorrecciones en el habla.
Q18	Los subtítulos del vídeo cumplen en general con los requisitos de accesibilidad.

Los datos personales de los estudiantes se suministraron anonimizados para evitar ninguna referencia a su identidad. Del estudio se han eliminado a aquellos estudiantes que, a pesar de haber realizado la actividad, no dieron su autorización para que sus datos se utilizaran en un estudio científicos.

Se dispuso de los siguientes ficheros `csv`:

- El fichero `grade` contiene el identificador de estudiante y el grupo al que pertenece (campo `cohort`).
- El fichero `abo` es la información socioeconómica que voluntariamente ha aportado el estudiante: sexo, año nacimiento, nivel de estudios, ocupación .
- El fichero `conoc` contiene el test de evaluación inicial de conocimientos del estudiante.
- El fichero `exp` es la evaluación del curso realizada por cada estudiante.
- El fichero `acc` contiene la información sobre accesibilidad que utiliza el estudiante.
- Los ficheros `test1` y `test2` son las repuestas al test de Likert sobre la calidad del subtitulado del primer y del segundo vídeo realizado por cada grupo respectivamente.

4.2 Características del diseño del experimento.

El diseño del experimento es completamente aleatorizado, de respuesta ordinal, cruzado AB/BA y doble ciego. Es decir que la asignación de los estudiantes a cada grupo fue aleatoria; cada grupo vio los vídeos en orden inverso; los estudiantes no conocían a priori qué vídeo estaban viendo en cada momento y tampoco se disponía de esta información en el momento de realizar el análisis estadístico de los datos.

Un diseño completamente aleatorizado (Lawson 2015, pp. 18) «garantiza la validez del experimento contra sesgos causados por otras variables ocultas. Cuando las unidades experimentales se asignan aleatoriamente a los niveles de factor de tratamiento, se puede realizar una prueba exacta de la hipótesis de que el efecto del tratamiento es cero utilizando una prueba de aleatorización».

Siguiendo a Senn (2022), para que el ensayo sea de tipo cruzado no sería suficiente intercambiar las secuencias sino que debe ser objeto del ensayo el estudio de las diferencias entre los tratamientos individuales que componen las secuencias. Los principales problemas d eun diseño cruzado son el abandono, `drop-out`, de alguno de los participantes y la interacción entre el tratamiento y el periodo o `carry-over`. Además el análisis estadístico es más complicado y particularmente cuando la respuesta es ordinal y hay más de dos tratamientos. En la misma línea, Lui (2016) afirma que «el objetivo principal de un diseño cruzado es estudiar la diferencia entre tratamientos individuales (en lugar de la diferencia entre secuencias de tratamiento). Debido a que cada paciente sirve como su propio control, el diseño cruzado es una alternativa útil al diseño de grupos paralelos para aumentar la potencia».

Las respuestas a un test de Likert se realizan en escala ordinal. No es adecuado realizar operaciones aritméticas para calcular medias con este tipo de datos. Pero ellos los test estadísticos para analizar el efecto de un tratamiento con respuesta

continua como son *ANOVA* y *t*-test no son adecuados con datos ordinales. Según la investigación de Liddell y Kruschke (2018) ajustar datos ordinales con modelos cuantitativos puede producir los siguientes problemas:

- Se pueden encontrar diferencias significativas entre grupos cuando no las hay: Error tipo I.
- Se pueden obviar diferencias cuando en realidad sí existen: Error tipo II.
- Incluso se pueden invertir los efectos de un tratamiento.
- También puede malinterpretarse la interacción entre factores.

Una opción es tratar los datos ordinales como si se tratara de datos categóricos y utilizar técnicas no paramétricas como el test de $Kruskal - Wallis$. El problema de este tipo de técnicas es que ignoran que los datos tienen una escala y, en el caso particular del diseño que nos ocupa se trata de datos longitudinales, es decir, que se toman varias medidas de cada sujeto y, por lo tanto, los datos no son independientes. Agresti (2010) expone un catálogo de técnicas para analizar datos categóricos y ordinales.

4.3 Objetivo.

El objetivo del estudio es responder a la pregunta de investigación:

Son los estudiantes de un curso de accesibilidad capaces de encontrar los errores en el subtitulado de un vídeo. Para ello se propondrán diversos test y modelos estadísticos que tengan en consideración las características que se han comentado en el diseño del experimento (ver Sección 4.2). Particularmente se tendrá en cuenta que se trata de un diseño cruzado con variable respuesta ordinal y variables explicativas longitudinales.

4.4 Preprocesamiento.

Partiendo de los ficheros suministrados (ver Sección 4.2), se realiza el siguiente preprocesado (para ver el código ejecutado consultar Apéndice A):

- Se lee el fichero de perfil del usuario. El número de fila con el que el usuario aparece en el fichero se utilizará como identificador del usuario para mantener la trazabilidad y comprobar que las transformaciones realizadas son correctas.
- Se eliminan del estudio a los estudiantes que aún habiendo realizado la actividad, no han dado su consentimiento para participar en el estudio.
- El valor del campo `cohort` se sustituye por una letra *A* o *B* en función del grupo asignado. En este momento se desconoce qué vídeo vio primero cada grupo.

- Se lee el fichero `profile` y se añade a los usuarios información sobre el sexo, el año de nacimiento y el novel de estudios.
- Se lee el fichero `conoc` y se calcula cuántas preguntas acertó cada usuario en el test de evaluación de conocimientos previos. Se añade esta información al perfil del usuario.
- Se leen los ficheros de test y se procesan. Se utiliza el nombre del fichero (`test1` o `test2`) para saber de qué vídeo se está respondiendo el test [2].
- Se seleccionan las preguntas que contienen las respuestas y se renombran para que sea más fácil saber de qué pregunta se trata [3]. Se convierte el campo `LastTry`, que contiene la fecha y hora de realización del test, a formato fecha y hora.
- Se realizan algunas comprobaciones como la ausencia de valores nulos en la variables más relevantes o que no existan inconsistencias ni errores de procesado.
- Se eliminan los comentarios y se graban en fichero aparte para que no revelen información que podría descubrir el tipo de subtitulado que piensa que está evaluando el estudiante.
- Se almacenan los resultados de los test preprocesado en un fichero `csv`.

[2] Se reitera que en este momento se desconoce si el vídeo es el correctamente subtitulado o el otro. La única información que se almacena es si se está respondiendo al vídeo que se voy primero o al que se vio después.

[3] En los fichero suministrados pla respuesta a cada pregunta ocupa varios campos y se selecciona en cada pregunta el que contiene el valor de la respuesta y se convierte a numérico.

Capítulo 5

Modelo.

5.1 Variables del modelo.

En la Tabla 5.1 se describen las características más relevantes de las principales variables que se utilizarán en en modelado y en el análisis estadístico.

Cuadro 5.1: Descripción de las variables más importantes

Nombre	Descripción	Tipo	Valores
Response	Respuesta a las preguntas del test.	Factor ordenado	De 0 a 5[1]
Level	Valoración de la respuesta.	Factor ordenado	Negative, Neutral, Positive[2]
Treat	Subtítulos	Factor	A o B[3]
Period	Periodo	Factor	1 ó 2[4]
Seq	Secuencia de aplicación de los tratamientos.	Factor	AB o BA
Subject	Identificación del estudiante	Factor	Numérico
Question	Número de la pregunta	Factor	Q01, Q02, ..., Q18[5]
Cluster	Grupo de la pregunta	Factor	1, 2, ó 3[6]

[1] Se ha hecho una rotación sobre los valores originales. 0 = No sé, 1 = Muy en desacuerdo, ..., 5 Muy de acuerdo.
[2] Positive cuando Response sea 4 ó 5, Negative cuando sea 1 ó 2 y Neutral para 3.
[3] No se conoce si el tratamiento A es el subtitulado bueno o lo es el B.
[4] 1 para el primer vídeo visto y 2 el segundo.
[5] Se ha reorganizado de tal forma que Q18, que es la pregunta resumen, sea el valor primero y de referencia.
[6] Se aplicará una técnica estadística de agrupamiento para agregar las preguntas.

Partiendo del `dataframe` que se construyó en el preprocesado (ver Sección 4.4) construimos el `dataframe` que usaremos a partir de este momento. Las operaciones principales que se han realizado han sido:

- Renombrar las variables para que se correspondan con las de nuestro modelo (ver Tabla 5.1).

- Eliminar del estudio los usuarios que solo han realizado uno de los test como se explica en Sección 6.1.
- Transformar las variables que lo requieran en factores. La pregunta 18 se usará como referencia en el factor `Question`.
- Rotar los valores de respuesta para que «No sé / No contesto» tenga valor 0 y el resto de 1 a 5 desde «Muy en desacuerdo», 1, hasta «Muy de acuerdo», 5.
- Agrupar las preguntas por similitud de respuesta (ver Sección 7.1).
- Crear el factor `Level` con los niveles `positive`, `neutral` y `negative` dependiendo de si la respuesta es 4 ó 5, 3, 1 ó 2 respectivamente.
- Transformar el `dataframe` de formato ancho a largo: los ficheros de respuestas se suministran en formato ancho. Es decir, que cada fila es un test que contiene 18 columnas para las respuestas a cada pregunta. Los nombres de las columnas son $Q01$, $Q02$, …, $Q18$ y tendrán valores de 0 a 6 con las respuestas. La mayoría de los paquetes de R que vamos a usar requieren que los datos estén en formato largo. Esto que quiere decir que cada fila tendrá una única respuesta por lo que habrá únicamente dos columnas, *Question* y *Response*. En la primera se almacenará el identificador de la pregunta ($Q01$, $Q02$, …, $Q18$) y en la segunda el valor de la respuesta (de 0 a 6). De esta forma un test pasará de ocupar una fila y 18 columnas en el formato ancho a 18 filas y dos columnas en el largo.

En Apéndice B se puede consultar el código en R para realizar el proceso descrito anteriormente. Con estas transformaciones se crean los dos `dataframes` que se usarán en el análisis estadístico de los datos:

- `df_all` contiene en formato largo todas las respuestas a los test.
- `df_clean` tiene la misma estructura que `df_all` pero en él se han eliminado las respuestas «No sé / No contesto».

`df_all` se utilizará cuando se traten las respuestas como categóricas y, por lo tanto, como no ordenadas. `df_clean` se utilizará cuando se traten las respuestas como ordenadas y por ello no contiene las respuestas con valor «No sé / No contesto».

La estructura de estos `dataframes` es la siguiente:

```
tibble [2,980 x 8] (S3: tbl_df/tbl/data.frame)
 $ Seq     : Factor w/ 2 levels "AB","BA": 1 1 1 1 1 1 1 1 1 1 ...
 $ Period  : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ Treat   : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
 $ Subject : Factor w/ 87 levels "4","33","35",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Question: Factor w/ 18 levels "Q18","Q01","Q02",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Cluster : Factor w/ 3 levels "1","2","3": 1 2 2 2 2 1 1 1 1 1 ...
 $ Response: Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Level   : Ord.factor w/ 3 levels "Negative"<"Neutral"<..: 2 2 2 2 2 2 2 2 2..
```

En el Tabla 5.2 se muestran algunos ejemplos de estos `dataframes`.

Cuadro 5.2: Muestra del dataframe preparado para el modelado estadístico en formato largo.

Seq	Period	Treat	Subject	Question	Cluster	Response	Level
BA	2	A	229	Q14	3	4	Positive
AB	2	B	788	Q15	3	5	Positive
BA	1	B	33	Q14	3	2	Negative
BA	2	A	371	Q04	2	3	Neutral
AB	2	B	850	Q12	2	2	Negative
AB	1	A	850	Q04	2	5	Positive
BA	2	A	85	Q07	1	4	Positive
AB	1	A	75	Q12	2	4	Positive
BA	1	B	819	Q16	3	3	Neutral
BA	2	A	220	Q05	1	4	Positive

Capítulo 6

Exploración inicial.

6.1 Análisis de la calidad de los datos.

Respuestas a los test.

Como se explica en la Tabla 5.1, al subtitulado le denominamos tratamiento y a sus niveles (correcto e incorrecto) los hemos llamado *A* y *B* sin hacer ninguna conjetura de cual de los dos es el subtitulado correcto. El grupo con secuencia *AB* será el que primero vio el vídeo con subtitulado *A* y luego el *B*. Análogamente, el grupo con secuencia *BA* vio los vídeos en orden inverso. Recuérdese que el nivel 0 de respuesta se corresponde con «No sé / No contesto» (ver Tabla 4.1).

Hay 24 estudiantes que no realizaron el segundo test. De ellos 9 pertenecen al grupo AB y 15 al grupo BA. Debido a que no son muchos y a que los grupos se mantienen balanceados, se ha decidido eliminar los test de estos estudiantes.

Tras eliminar a los estudiantes que no realizaron uno de los test, constatamos (ver Figura 6.1) que los grupos están balanceados en el número de estudiantes y que disponemos de suficientes datos para realizar el análisis estadístico.

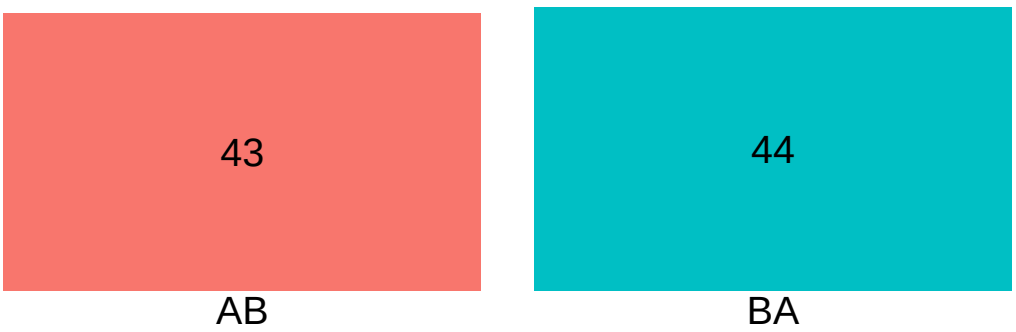


Figura 6.1: Estudiantes asignados a cada grupo.

El campo `LastTry` contiene la fecha y hora de realización del test. Con esta información podemos conocer el tiempo que empleó cada estudiante entre subtitulados. La Tabla 6.1 muestra que hay algunos test que se hicieron demasiado rápido [1].

Cuadro 6.1: Tiempos de realización de la segunda actividad de duración inferior a 2 minutos.

Minutes
0.93
1.3
1.7
1.72
1.78
1.97

La Figura 6.2 muestra que hay 28 test en los que el estudiante contestó a todas las preguntas usando únicamente 2 respuestas diferentes. Además hay 13 test en los que se contestaron todas las preguntas con 1 respuesta.

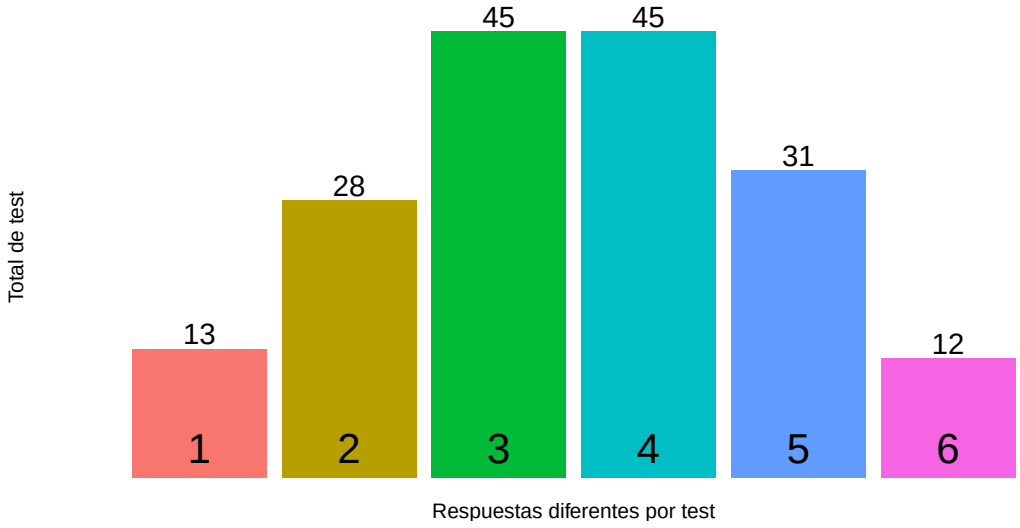


Figura 6.2: Número de respuestas diferentes en un mismo test.

[1] Hay que tener en cuenta que la duración de vídeo es de algo más de 40 segundos y que los estudiantes tienen que contestar un test de 18 preguntas.

La tabla Tabla 6.2 muestra la respuesta utilizada, el grupo y el periodo de los test con respuesta única.

Cuadro 6.2: Test en los que todas las preguntas se contestan el mismo valor de respuesta.

Response	Seq	Test
2	AB	01
2	AB	02
3	BA	01
3	BA	02
3	BA	02
3	BA	02
4	AB	01
4	AB	01
4	AB	02
4	BA	01
4	BA	02
4	BA	02
4	BA	02

La Figura 6.3 presenta la distribución de la cantidad de respuestas cuyo valor cambia entre los dos test que realiza cada estudiante.

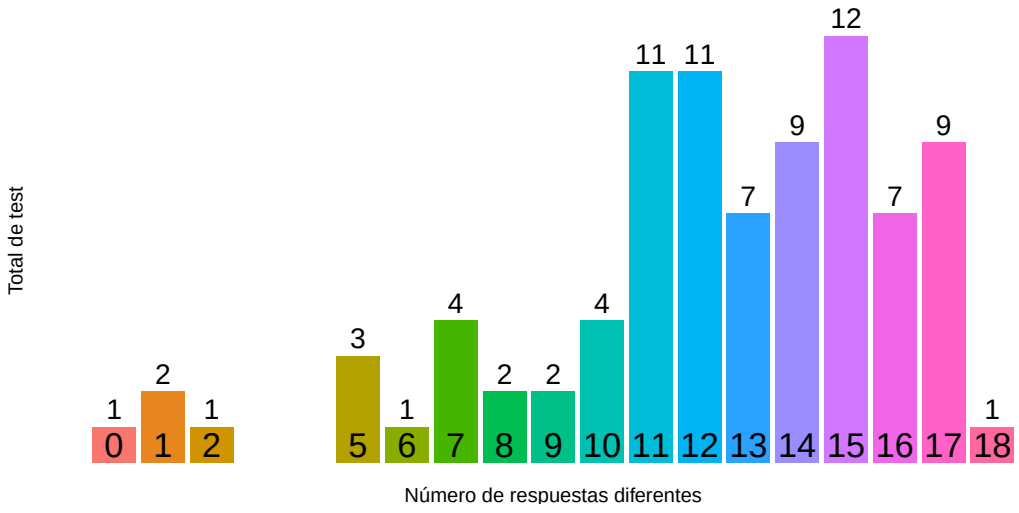


Figura 6.3: Número de respuestas diferentes entre los test para cada estudiante.

Tan solo 1 estudiante respondió a todas las preguntas con el mismo valor en los dos test. Por otro lado, no hay test que tengan un número excesivo de contestaciones «No sé/No contesto» (ver Tabla 6.3).

Cuadro 6.3: Los 5 test con más respuestas 'No sé/No contesto'

Test	Total respuesta por test
01	5
01	5
02	5
02	5
01	4

Conclusiones.

No parece razonable realizar la actividad en menos de 2 minutos. Se observa que en algunos test hay poca variabilidad. Sin embargo, no son muchos los test con estas características así que se ha decidido mantener estos datos a pesar de que se pueda dubar de si en ellos los estudiantes contestaron con la debida atención y diligencia.

Valores nulos o erróneos.

En los test no se ha detectado ningún valor nulo ni erróneo. Sin embargo tenemos algunos de estos valores en la información socioeconómica de los estudiantes (ver Tabla 7.3).

6.2 Comparación de los tratamientos A y B entre grupos.

La Figura 6.4 presenta una forma de comparar los dos test que realizados por los estudiantes. Para cada estudiante se comparó pregunta a pregunta sus dos test y se contabilizó la diferencia entre el número de preguntas en que la puntuación en el segundo vídeo fue superior y en las que lo fue inferior (las que no variaron de puntuación no se consideraron). En el eje x se muestra la diferencia entre preguntas. Cantidades negativas indican que hay más respuestas en el segundo de los test que han empeorado respecto al primero de las que han mejorado. En el eje y se representa el número de estudiantes para cada diferencia. Esta frecuencia se representa en negativo cuando la diferencia es negativa [2]. Esto es una forma de evaluar si el estudiante valoró mejor o no el segundo vídeo que el primero.

Vemos que en el grupo AB las diferencias tienden a ser negativas y en el BA positivas. Esto estaría indicando que los estudiantes valoran mejor el subtitulado

[2]En la comparación se han omitido aquellas preguntas en las que el estudiante contestó «No sé/No contesto» en la pregunta correspondiente de uno de los test.

Cuadro 6.4: Tablas de contingencia de la información socioeconómica de los estudiantes.

gender	Freq
f	92
m	38
NA	44

Estudiantes por sexo.

year_of_birth	Freq
None	44
NA	2

Estudiantes con valor nulo en el campo año de nacimiento.

level_of_education	Freq
a	50
b	16
hs	4
m	30
other	4
p	20
NA	50

Estudiantes por nivel educativo.

level_of_knowledge	Freq
4	2
6	4
7	30
8	44
9	40
10	32
NA	22

Estudiantes en función del número de preguntas acertadas en el test de conocimiento.

de nivel *A*. Por ello es esperable que las respuestas de los estudiantes del grupo *AB* hayan empeorado y que las diferencias sean negativas y que lo contrario haya sucedido con las del grupo *BA*. La diferencia más frecuente en el grupo *AB* es 12 y en el grupo *BA* este valor es 11.

Resulta llamativo que haya estudiantes cuyas contestaciones estén tan alejadas de la tendencia de su grupo. En la Tabla 6.9 se muestran los tiempos que han transcurrido entre la realización de los test de aquellos estudiantes cuyas respuestas difieren de forma importante de su grupo. Se observa que casi todos son tiempos entre actividades muy cortos. En cualquier caso y, como no son muchos, se ha decidido no eliminarlos y realizar el análisis con ellos.

Cuadro 6.9: Estudiantes que tienen diferencias en sus respuestas muy alejadas de la tendencia de su grupo.

Seq	Diff	Minutes
AB	17	1.3
AB	7	3.33
BA	-10	50345.95
BA	-12	1.7

En la Figura 6.5 representamos la frecuencia relativa del valor de respuesta para

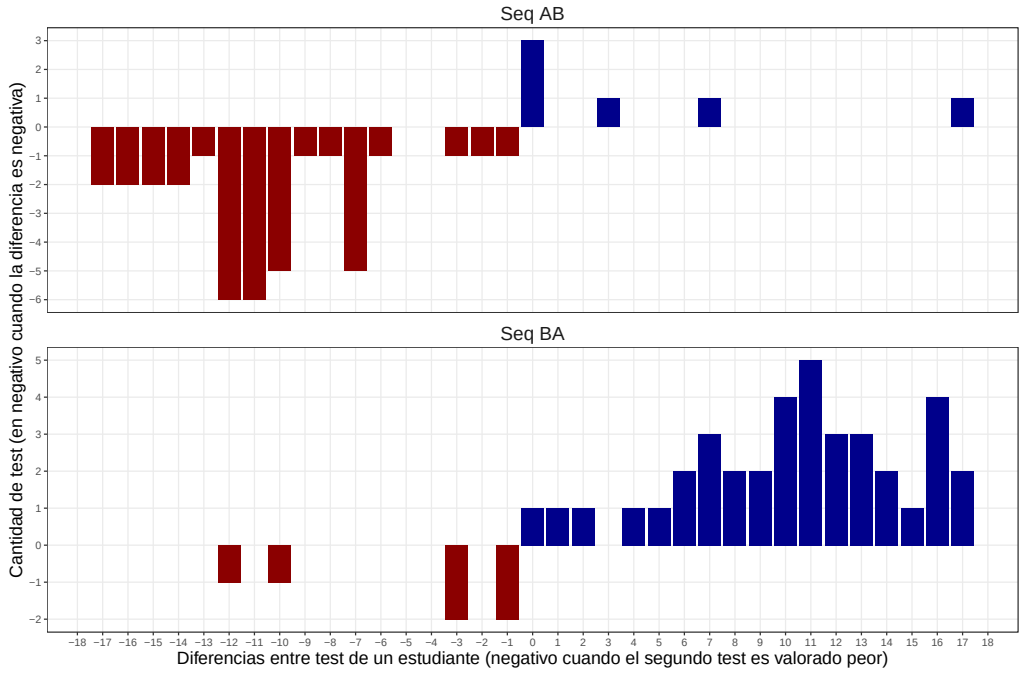


Figura 6.4: Frecuencias absolutas de las diferencias en las respuestas entre test por estudiante y grupo.

cada grupo y test en todas la preguntas. Esta es otra forma de comparar los niveles de subtitulado.

La Figura 6.5 muestra algunas cuestiones interesantes:

- El tratamiento (subtitulado) con nivel *A* presenta claramente mayores valores de respuesta que el *B* como ya habíamos visto (ver Figura 6.4). Si en este momento tuviéramos que decidir qué subtitulado es cada uno parece claro que sería el de nivel *A*. No obstante, ni en el análisis exploratorio ni en el modelado estadístico se hará ninguna suposición.
- En general los dos grupos muestran bastante acuerdo en el subtitulado en ambos niveles: En el nivel de tratamiento *A* los dos grupos tienen una frecuencia relativa similar de respuestas positivas (valores 4 y 5). El grupo *AB* tiene un 82% de respuestas positivas frente a un 84% el grupo *BA*. No obstante, el grupo *AB* tiene más respuestas con valor 5 que el grupo *BA* (56% frente a 41%). La valoración es también similar entre grupos en el nivel de tratamiento *B*: el grupo *AB* tiene 44% de respuestas positivas y 47% el grupo *BA*. Las valoraciones negativas (1, 2), la neutra (3) y la "No sé / No contesto" (0) son también muy similares.
- Las respuestas son similares entre periodos aunque ligeramente más negativas en el segundo. Así un 65% de las respuestas son positivas en el primer periodo frente a un 64% en el segundo.

El análisis marginalizado de tratamiento, secuencia y periodo tiene estos resultados referidos a las preguntas con contestación positiva (4, 5):

Figura 6.5: Frecuencias relativas de las respuestas al test.

- El tratamiento *A* tiene un 83% marginalizado de respuestas positivas frente al 46% del tratamiento *B*.
- El periodo 1 tiene un 65% marginalizado de respuestas positivas frente al 64% del periodo 2.
- Finalmente, la secuencia *AB* tiene un 63% de respuestas positivas frente 66% de la secuencia *BA*. Analizado por respuestas individuales, la respuesta 4 pasa de 24% en la secuencia *AB* a 37% en la *BA* y, de forma contraria, en la respuesta 5 pasa de 39% en *AB* a 29% en *BA*. En las respuestas negativas y no contestadas y neutra no se aprecian estas variaciones.

6.3 Análisis de las preguntas.

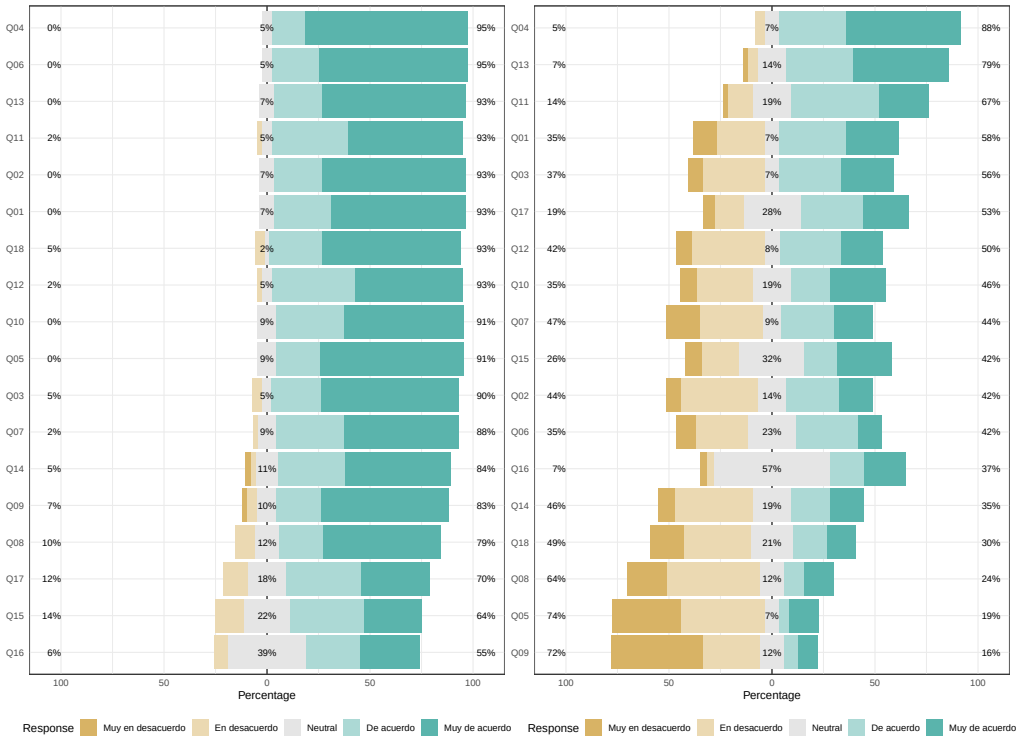
El gráfico Figura 6.6 muestra la frecuencia relativa por grupo y por test de las preguntas clasificadas por niveles de respuesta, considerando que:

- Los niveles 1 y 2 se consideran valoraciones negativas.
- El nivel 3 se considera neutro.
- Los niveles 4 y 5 se consideran positivos.
- El nivel 0 («No sé / No contesto») se excluye en este análisis.

Se muestra en primer lugar la pregunta 18 por ser una valoración global del subtitulado y que resume la opinión que sobre el mismo tiene el estudiante. Volvemos a constatar que el subtitulado *A* es mejor valorado por los estudiantes, pero ahora vemos que en las 18 preguntas ambos grupos tienen mas puntuaciones positivas y menos negativas en el subtitulado *A* que el *B*. También volvemos a encontrar que los dos grupos valoran de forma muy similar los dos niveles de subtitulado en todas la preguntas. En el nivel de subtitulado *A* las preguntas *Q*15, *Q*16 y *Q*17 obtienen relativamente peores valoraciones (consultar la Tabla 4.2 para ver los valores) y estas son similares en ambos subtitulados. Hay algunas preguntas que son valoradas de forma positiva incluso en el nivel de subtitulado *B* (por ejemplo *Q*04 o *Q*13) y que, por lo tanto, su valoración es similar en ambos subtitulados. Por último, las preguntas *Q*05 y *Q*09 (también la *Q*14 pero solo para el grupo *BA*) tienen una valoración muy negativa en el nivel de subtitulado *B*.

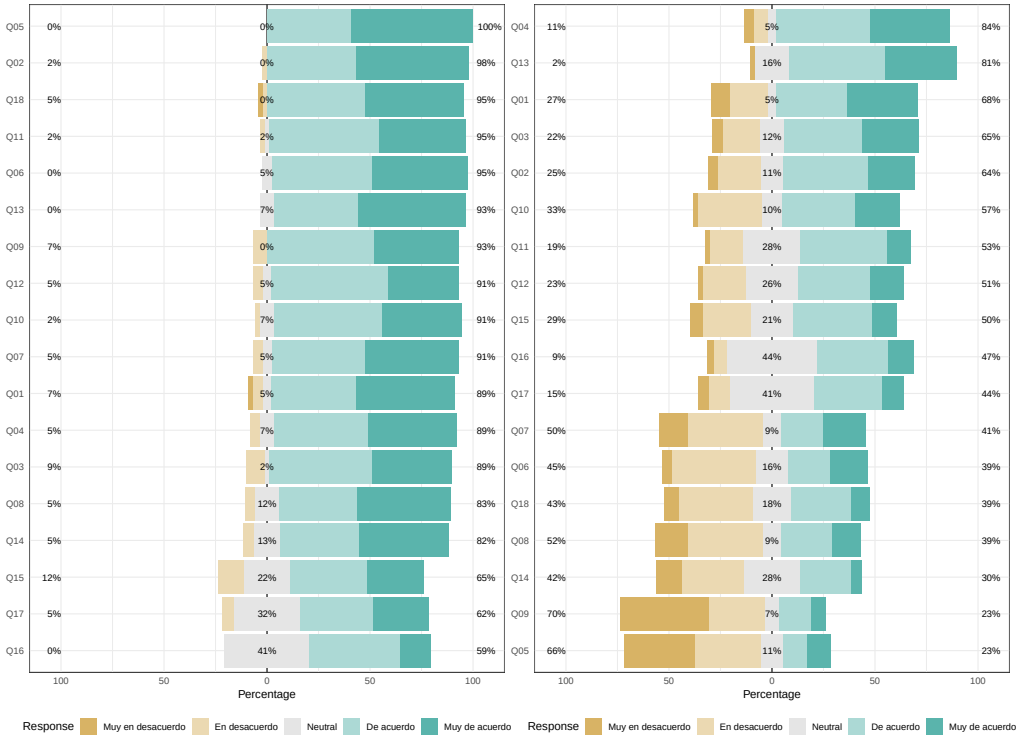
Figura 6.6: Frecuencias relativas de las respuestas por pregunta.

La figura Figura 6.7 clasifica la preguntas por valoración y permite constatar lo que ya habíamos visto en el párrafo anterior con mayor comodidad.



(a) Seq AB , Treat A

(b) Seq AB , Treat B



(c) Seq BA , Treat A

(d) Seq BA , Treat B

Figura 6.7: Preguntas ordenadas por valoración.

Capítulo 7

Análisis estadístico.

7.1 Agrupamientos de preguntas.

Correlación entre preguntas con el alfa de Cronbach.

Normalmente las preguntas de un cuestionario pretenden medir una variable que está oculta o latente. En nuestro caso es la calidad del subtitulado. Las respuestas a estas preguntas relacionadas deben ser consistentes internamente, es decir, las respuestas deben correlacionarse fuerte y positivamente.

Un índice que se utiliza habitualmente para medir la consistencia interna de un cuestionario es el coeficiente `alfa de Cronbach`, ver Schweinberger (2020). Se define de esta forma:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} s_i^2}{s^2}\right) \tag{7.1}$$

Donde:

- α es el coeficiente `alfa de Cronbach`.
- N es el número de items de la escala de Likert.
- s_i^2 es la varianza de la puntuación del item i.
- s^2 es la varianza total de las puntuaciones de todos los items.

Valores cercanos 1 indican una fuerte correlación en las respuestas y se admite que las preguntas del cuestionario están midiendo la misma variable latente.

Para calcular en R este coeficiente podemos usar la función `alpha` del paquete `psych`:

Cuadro 7.1: Valor del coeficiente alpha de Cronbach si se elimina una pregunta.

(a)

Q18	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08
0.91	0.92	0.92	0.92	0.92	0.91	0.91	0.91	0.91

(b)

Q09	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
0.91	0.91	0.92	0.91	0.92	0.92	0.92	0.93	0.92

Cuadro 7.2: Relación de cada pregunta con el índice alpha de Cronbach.

(a)

Q18	Q05	Q06	Q09	Q08	Q07	Q10	Q12	Q02
0.86	0.81	0.79	0.79	0.77	0.75	0.73	0.72	0.71

(b)

Q03	Q14	Q01	Q11	Q15	Q13	Q04	Q16	Q17
0.68	0.66	0.65	0.64	0.56	0.51	0.46	0.44	0.42

```
alpha <- df_all %>%
    pivot_wider(
        names_from = Question,
        values_from = Response_v,
        id_cols = c(Treat, Subject)
    ) %>%
    dplyr::select(-c(Treat, Subject)) %>%
    psych::alpha()
```

Se obtiene un coeficiente alfa de `alfa de Cronbach` de 0.92 que indica una muy buena correlación entre las respuestas a todas las preguntas. Este valor apenas se ve alterado si se elimina una de las preguntas (ver Tabla 7.1).

En la Tabla 7.2 mostramos las preguntas que más contribuyen al índice `alpha de Cronbach`. Es interesante que la pregunta $Q18$, que es la valoración general del cuestionario, sea la que mejor contribución tiene al índice.

Agrupamiento jerárquico aglomerativo.

En en la Sección 7.1 y en la Sección 6.2 hemos visto que algunas de las preguntas tienen respuestas similares a otras pero diferentes del resto. Puede ser interesante aplicar una técnica de agrupamiento que nos permita crear grupos de preguntas que podremos analizar por separado.

Vamos a realizar una agrupación jerárquica aglomerativa de las preguntas en función de la tabla de contingencia de las respuestas utilizando la distancia euclidea como medida de distancia y el método de aglomeración de enlace completo para unir conglomerados [1]. Para ello primero calculamos la tabla de contingencia (ver Tabla 7.3) de preguntas y respuestas.

```
table <- df_all %>%
    xtabs(~ Question + Response, data = .)
```

Cuadro 7.3: Tabla de contingencia de preguntas y respuestas.

Question	Response_0	Response_1	Response_2	Response_3	Response_4	Response_5
Q18	0	11	33	18	52	60
Q01	0	10	20	10	59	75
Q02	0	5	26	14	58	71
Q03	5	5	26	11	60	67
Q04	0	2	7	10	61	94
Q05	1	29	31	12	34	67
Q06	1	6	29	21	53	64
Q07	0	13	32	14	54	61
Q08	4	15	41	19	40	55
Q09	1	39	29	12	42	51
Q10	8	4	24	18	59	61
Q11	3	2	14	23	75	57
Q12	5	4	26	18	69	52
Q13	1	2	2	19	62	88
Q14	21	9	29	27	44	44
Q15	26	5	25	36	47	35
Q16	47	2	5	57	39	24
Q17	29	4	15	44	49	33

Con la tabla de contingencia calculamos las distancias entre preguntas y realizamos el agrupamiento. En el dendograma se aprecian claramente tres agrupamientos. Es muy interesante constatar que los tres grupos están formados por preguntas que en su mayor parte son correlativas. Esto es consistente con que al elaborar un test normalmente se colocan las preguntas por unidades temáticas y con que el encuestado también suele hacerlo teniendo en cuenta esta estructura y tiende a responder de forma similar a las preguntas correlativas.

```
dist <- dist(table, method = "euclidean")
cluster <- hclust(dist, method = "complete")
plot(cluster)
```

Podemos distinguir los siguientes grupos y subgrupos:

[1] El método de enlace completo usa la distancia máxima entre dos conglomerados para seleccionar los más cercanos a unir.

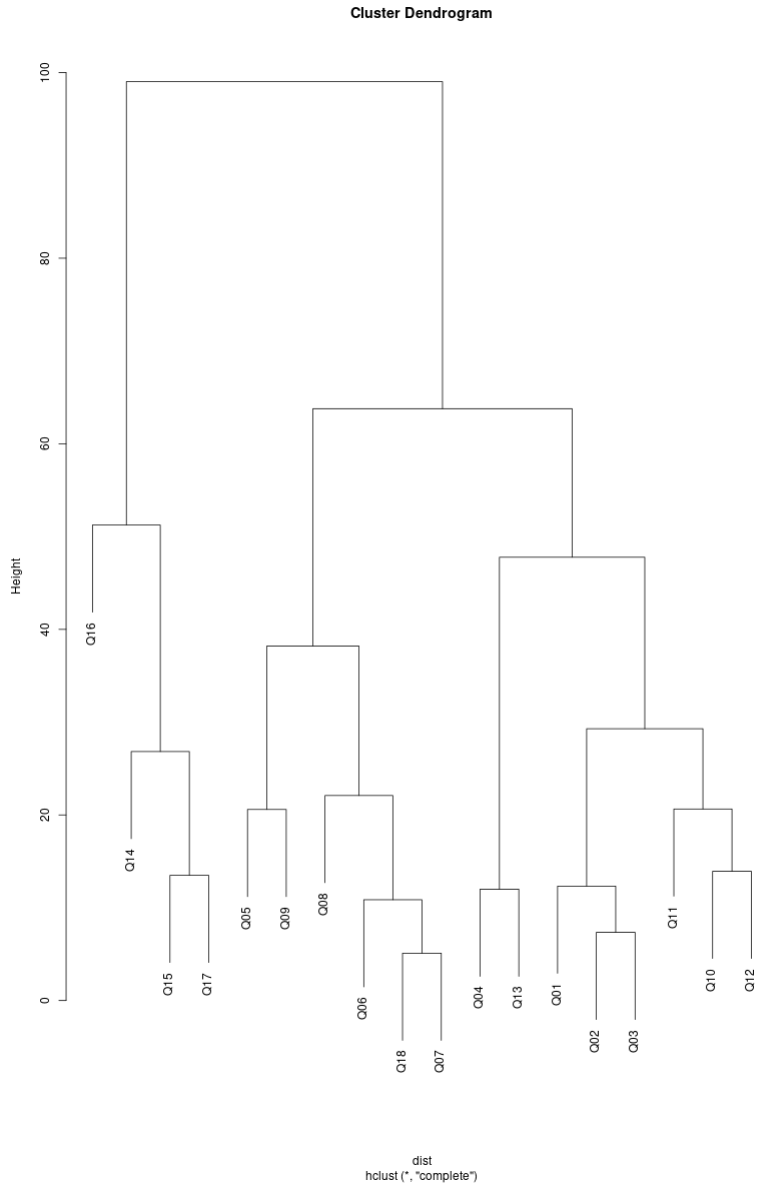


Figura 7.1: Dendograma de aglomeramiento jerárquico de preguntas en función de la tabla de contingencia de respuestas.

- Grupo 1: Trata sobre la corrección del subtítulo.
  - Subgrupo 05, 06, 07, 08, 09: Preguntas sobre si la información que presenta el subtítulo es correcta y está bien escrita.
  - Pregunta 18: Valoración general del subtitulado. El que esta pregunta esté incluida en el grupo sobre corrección estaría indicando que este es el apartado al que más importancia dan los estudiantes a la hora de valorar la calidad del subtitulado.
- Grupo 2: Es el más numeroso. En general está formado por preguntas sobre el grado de dificultad que presenta la lectura del subtítulo.

  - Subgrupo preguntas Q01, Q02, Q03: Colocación de los subtítulos.
  - Subgrupo preguntas Q10, Q11, Q12: Sincronización, velocidad y número de líneas.
  - Subgrupo preguntas Q04, Q13: Contraste y legibilidad.
- Grupo 3: Son preguntas que tratan también sobre la corrección del subtítulo, pero con la diferencia sobre el grupo uno de que se trata de cuestiones más sutiles y presumiblemente más difíciles de valorar por un novato. Está formado por las preguntas Q14, Q15, Q16 y Q17.

7.2 Análisis de tablas de contingencia.

En esta sección se aplicarán técnicas estadísticas que se basan en tablas de contingencia. Una descripción teórica de este tipo de técnicas se pueden encontrar en Agresti (2018). Un tratamiento aplicado y basado en gráficos, que será el enfoque que seguiremos en este trabajo, es realizado en Friendly et al. (2015).

Comparación mediante mosaicos.

En el Figura 7.2 se representan en forma de mosaico las tablas de contingencia de las respuestas por tratamiento y secuencia. La información mostrada es similar a la que presentamos en la Figura 6.5, aunque el gráfico es más intuitivo ya que la anchura y altura de los rectángulos son proporcionales a la frecuencia marginal de la secuencia y el tratamiento respectivamente y el área es proporcional a la frecuencia conjunta. En esta ocasión hemos decidido emparejar los tratamientos en lugar de hacerlo con la secuencia, como hicimos anteriormente. Esto permite una mejor comparación de las diferencias entre grupos. Con ello podemos ver fácilmente que el tratamiento *A* es mejor valorado por los estudiantes y que el grupo que realizó la secuencia *AB* tiene más respuestas 5 pero menor número de respuestas positivas totales que el grupo de secuencia *BA* en ambos niveles de tratamiento.

Comparación con *Odds Ratio*.

Hasta este momento ha quedado claro que el nivel de subtitulado *A* es preferido por los estudiantes y que las respuestas de ambos grupos son similares. Pero, ¿cuánto de similares son? Una forma de contestar esta pregunta es utilizar el `odds ratio` de tratamientos y grupos para cada nivel de respuesta.

Es decir, calcular:

$$OR_{(Treat,Seq|Response=r)} = \frac{\frac{P(Treat=A|Seq=AB,Response=r)}{P(Treat=B|Seq=AB,Response=r)}}{\frac{P(Treat=A|Seq=BA,Response=r)}{P(Treat=B|Seq=BA,Response=r)}} \tag{7.2}$$

Si los *OR* son similares en todos los niveles de respuesta, podemos afirmar que los grupos son homogéneos. Los resultados en R no producen significación estadística en ningún nivel de respuesta:

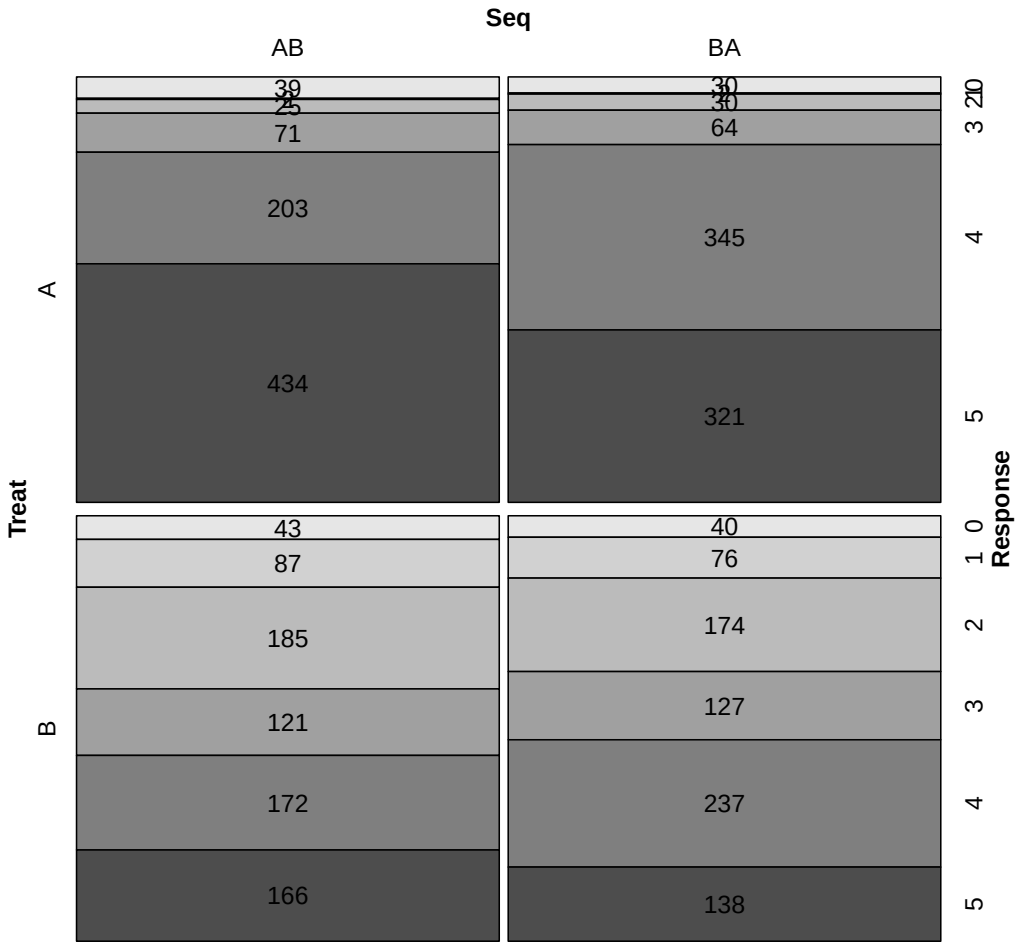


Figura 7.2: Mosaico de tratamientos y secuencias.

```
summary(loddsratio(~ Treat + Seq + Response_l, data = df_all))
```

```
z test of coefficients:

                   Estimate Std. Error z value Pr(>|z|)
No sé / No contesto  0.19004    0.32746  0.5804   0.5617
Muy en desacuerdo   -0.13517    1.01225 -0.1335   0.8938
En desacuerdo       -0.24362    0.29066 -0.8382   0.4019
Neutral              0.15219    0.21412  0.7108   0.4772
De acuerdo          -0.20977    0.13363 -1.5698   0.1165
Muy de acuerdo       0.11687    0.13671  0.8549   0.3926
```

La Figura 7.3 presenta visualmente la misma información.

Sería interesante calcular el *OR* para cada nivel de respuesta y pregunta pero por desgracia la muestra es demasiado pequeña para hacerlo. Se ha calculado el *OR* sobre los agrupamientos de preguntas y se ha obtenido significación estadística tan solo en el agrupamiento 2 y nivel de respuesta 2:

```
summary(loddsratio(~ Treat + Seq + Cluster + Response_l, data = df_all))
```

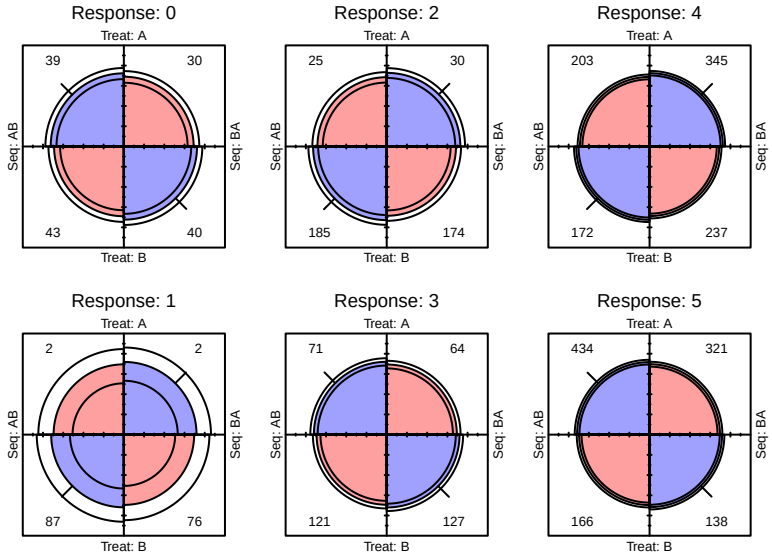


Figura 7.3: OR entre tratamiento y grupo por nivel de respuesta.

```
z test of coefficients:

                      Estimate Std. Error z value Pr(>|z|)
1:No sé / No contesto -1.94591    1.75662 -1.1078  0.26797
2:No sé / No contesto  0.41074    1.12569  0.3649  0.71520
3:No sé / No contesto  0.29239    0.35972  0.8128  0.41632
1:Muy en desacuerdo   -0.12516    1.17012 -0.1070  0.91482
2:Muy en desacuerdo   -1.39488    1.66941 -0.8356  0.40341
3:Muy en desacuerdo    1.19870    1.69327  0.7079  0.47900
1:En desacuerdo        0.17829    0.49526  0.3600  0.71885
2:En desacuerdo       -1.34796    0.57253 -2.3544  0.01855 *
3:En desacuerdo        0.23740    0.50928  0.4661  0.64111
1:Neutral              0.62181    0.46172  1.3467  0.17807
2:Neutral              0.53248    0.39619  1.3440  0.17895
3:Neutral             -0.24146    0.31560 -0.7651  0.44421
1:De acuerdo          -0.35125    0.25963 -1.3529  0.17609
2:De acuerdo          -0.28064    0.18244 -1.5382  0.12399
3:De acuerdo           0.22503    0.30663  0.7339  0.46303
1:Muy de acuerdo       0.27860    0.26542  1.0497  0.29387
2:Muy de acuerdo       0.23441    0.17892  1.3101  0.19015
3:Muy de acuerdo      -0.61437    0.38071 -1.6137  0.10659
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sin embargo no podemos asumir que esta significación no se deba al azar ya que estamos realizando 18 contrastes de hipótesis diferentes y cada uno tiene un error tipo I asociado, con lo que la probabilidad de encontrar una significación estadística por puro azar aumenta. Se han propuesto correcciones del *p*-value como la de Bonferroni que no se aplican en este trabajo.

Otro *OR* que tiene interés calcular es el de tratamiento y periodo para evaluar si las respuestas son homogéneas. Mostramos tanto la tabla de resultados en R y también su representación visual (ver Figura 7.4).

```
summary(loddsratio(~ Treat + Period + Response_l, data = df_all))
```

```
z test of coefficients:

                    Estimate Std. Error z value  Pr(>|z|)
No sé / No contesto  0.33469    0.32746  1.0221 0.3067511
Muy en desacuerdo    0.13517    1.01225  0.1335 0.8937673
En desacuerdo       -0.12102    0.29067 -0.4164 0.6771467
Neutral              0.05540    0.21412  0.2587 0.7958414
De acuerdo          -0.85090    0.13363 -6.3674 1.922e-10 ***
Muy de acuerdo       0.48634    0.13671  3.5574 0.0003745 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

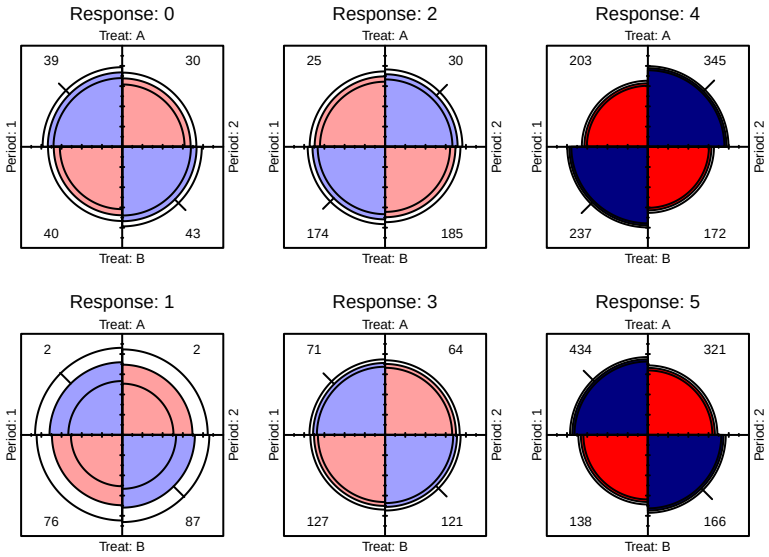


Figura 7.4: OR entre tratamiento y periodo por nivel de respuesta.

Podemos constatar la existencia de un efecto periodo de signo contrario para las preguntas 4 y 5. La razón de que se produzca este efecto periodo es que algunas de las respuestas de valoración 5 en ambos niveles de subtitulado y grupos en el primer periodo se convierten en valoración 4 en el segundo periodo. Es decir, que esto nos está indicando que los estudiantes de ambos grupos prestaron más atención o fueron más exigentes en el segundo visionado y decidieron no otorgar la puntuación máxima incluso en algunos items al subtitulado correcto. Que el efecto periodo sea contrario en dos preguntas no debe sorprendernos en este diseño de experimento, ya que un test es un juego de suma cero: la valoraciones que se ganan o se pierden en un nivel de respuesta necesariamente provoca que el resto de niveles pierdan o ganen respectivamente la misma cantidad. En cualquier caso, vemos que el efecto periodo es cuantitativa y cualitativamente pequeño. Al afectar solo al intercambio de valoraciones entre los niveles 4 y 5, y ser las dos positivas, es simplemente una pequeña corrección en la valoración del subtitulado.

Capítulo 8

Modelado estadístico.

En esta sección vamos a proponer varios modelos que permitan hacer inferencia sobre los datos.

8.1 Árboles de inferencia condicional.

Los arboles de inferencia condicional (CIT) son un tipo de árbol de decisión en el que la selección de variables y de los puntos de división no se basan en medidas de homogeneidad como el índice de Gini, sino en un contrastes de hipótesis no paramétricos. El algoritmo que se utiliza es el siguiente, ver Levshina (2020):

El algoritmo consiste en contrastar la hipótesis nula de si la variable de respuesta Y es independiente de alguna variable explicativa $Y \mid X$. Para probar la hipótesis, se utiliza un algoritmo de permutación de la variable respuesta y se mide la asociación con la variable explicativa antes y después de la permutación. Si la asociación no varía significativamente, podemos asumir que las variables de respuesta y explicativa son independientes. De esta forma se selecciona la variable explicativa que más influye en la respuesta y que se utilizará en el particionado. Para elegir el valor de la variable explicativa que dividirá el conjunto de datos, se procede de forma análoga midiendo el cambio en la diferencia de asociación. De acuerdo con Friendly (2015), los CIT resuelven los problemas de sobreajuste de los árboles de decisión tradicionales.

Para realizar el particionado basado en CIT, vamos a usar la función `ctree` del paquete `party` de R. Presentamos aquí únicamente el modelo final elegido que incluye como variables explicativas `Treat`, `Period`, `Seq` y `Cluster` [1].

En la Figura 8.1 podemos ver que el nivel de subtitulado es el efecto principal, seguido del grupo de preguntas y finalmente la secuencia. En este modelo el

[1] Se han realizado simulaciones con otras combinaciones de variables explicativas que no se incluyen por no haber producido resultados relevantes.

periodo no aparece por no estar asociado con la respuesta. Estos resultados son contradictorios con los que obtuvimos en el análisis con el OR (ver Sección 7.2) en el que el factor secuencia no era significativo pero sí lo era el factor periodo. Por otro lado, vemos que la asociación más fuerte es el nivel de respuesta 5 para subtitulado *A*, grupos de preguntas 1 y 2 y secuencia *AB* y de las respuestas 4 y 5 cuando la secuencia es *BA*. El tratamiento *B* está fuertemente asociado con el nivel de respuesta 1 para el grupo de preguntas 1. Por último, con este modelo no hay ninguna combinación de factores que prediga un nivel de respuesta 1.

```
tree.1 <- ctree(Response ~ Treat + Cluster + Period + Seq, data = df_clean)
```

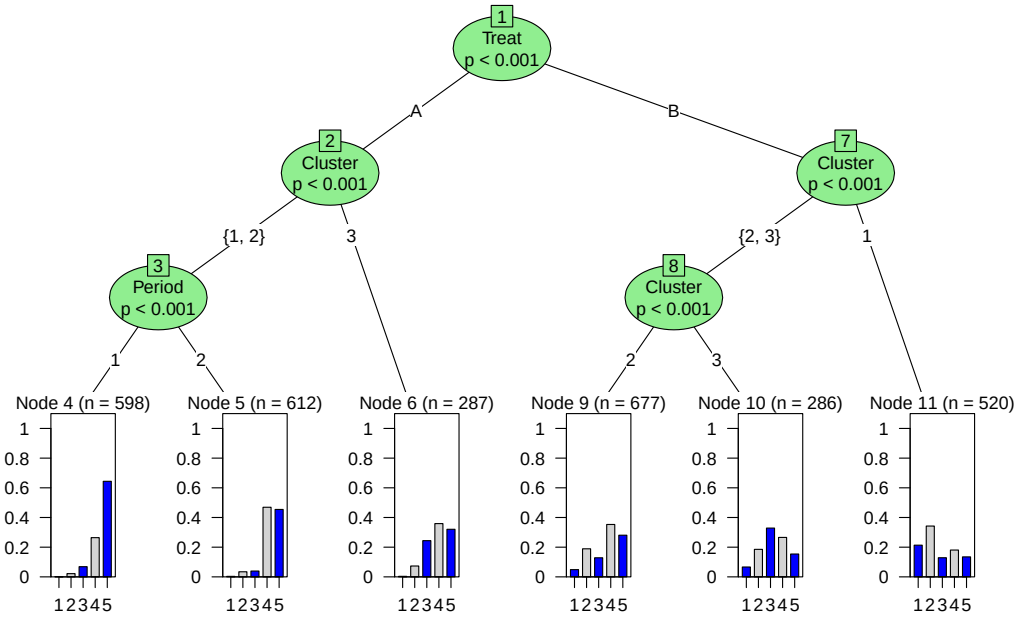


Figura 8.1: Modelo con árboles de inferencia condicional (Response ~ Treat + Cluster + Period + Seq).

Podemos usar este modelo para hacer predicciones. La matriz de contingencia resultante es la siguiente:

```
          Prediction
Reference   1   2   3   4   5
        1   0 111  19  36   1
        2   0 178  53 170  13
        3   0  67  94 181  41
        4   0  94  76 629 158
        5   0  70  44 560 385
```

Como habíamos anticipado, nunca se predice el nivel de respuesta 1. Las categorías que más probablemente predice nuestro modelo son la 4 y la 5 pero aún así hay mucha confusión entre ellas. La exactitud de predicción es 43%.

Un modelo alternativo sería usar las mismas variables explicativas pero cambiado `Response` por `Level` como variable de respuesta. Esta variable solo tiene tres niveles: positivo, negativo y neutro. De esta forma no se producen confusiones entre los niveles 1 y 2 por un lado y 4 y 5 por otro:

```
tree.2 <- ctree(Level ~ Treat + Cluster + Period + Seq, data = df_clean)
```

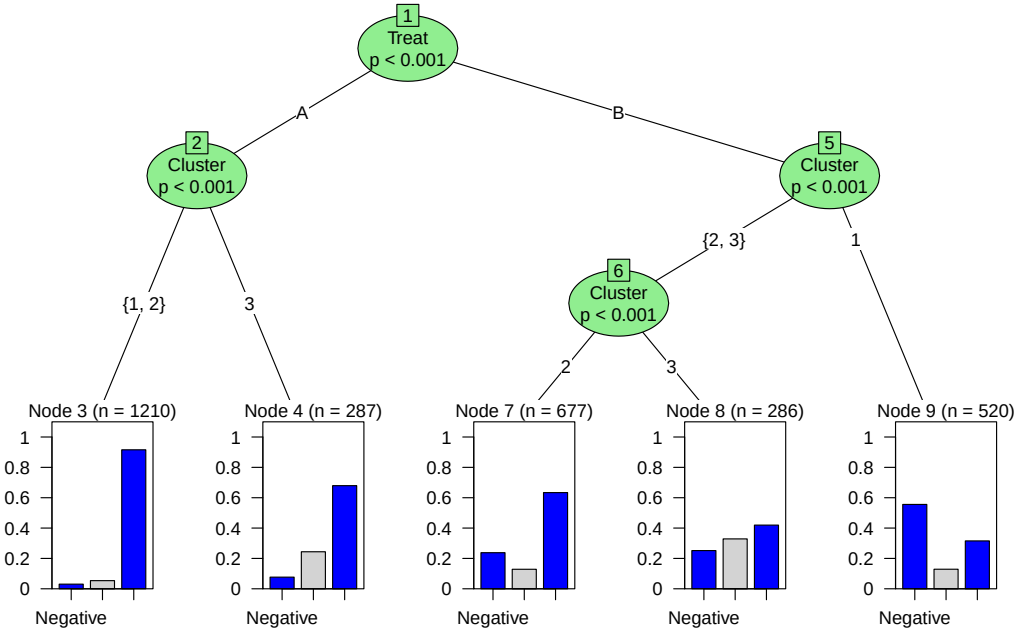


Figura 8.2: Modelo con árboles de inferencia condicional (Level ~ Treat + Cluster + Period + Seq).

Vemos que el árbol (ver Figura 8.2) es muy similar al otro (ver Figura 8.1) con la principal diferencia de que ahora la secuencia ha desaparecido como factor relevante. Por otro lado, el modelo siempre predice una respuesta positiva excepto para el subtitulado *B* y grupo de preguntas 1 que es negativa (el nivel neutro nunca se predice). La exactitud del modelo ha subido a 72%. En cualquier caso no es una gran mejora ya que un modelo que predijera siempre la categoría mayoritaria,que es la positiva, habría obtenido una exactitud de 68%. Se han hecho simulaciones consistentes en incluir como factor las preguntas o usar como modelo un árbol de decisión convencional con resultados similares.

Capítulo 9

Resultados

Capítulo 10

Conclusiones y trabajo futuro

Referencias

AENOR (2012). *UNE 153010 Subtitulado para personas sordas y personas con discapacidad auditiva*. Asociación Española de Normalización y Certificación.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. DOI: 10.1002/9780470594001.

– (oct. de 2018). *An introduction to categorical data analysis, 3rd Edition*. URL: https://www.wiley.com/en-us/An+Introduction+to+Categorical+Data+Analysis%2C+3rd+Edition-p-9781119405283.

Friendly, M. (dic. de 2015). *Classification and regression trees*.

Friendly, M., D. Meyer y A. Zeileis (dic. de 2015). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*, págs. 1-525. DOI: 10.1201/b19022.

Guerra, A., T. Gidel y E. Vezzetti (mayo de 2016). «Toward a common procedure using likert and likert-type scales in small groups comparative design observations». En.

Lawson, J. (2015). Ed. por Chapman y Hall/CRC. DOI: 10.1201/b17883. URL: https://www.taylorfrancis.com/books/mono/10.1201/b17883/design-analysis-experiments-john-lawson.

Levshina, N. (2020). «Conditional Inference Trees and Random Forests». En: *A Practical Handbook of Corpus Linguistics*. Ed. por M. Paquot y S. T. Gries. Cham: Springer International Publishing, págs. 611-643. DOI: 10.1007/978-3-030-46216-1_25. URL: https://doi.org/10.1007/978-3-030-46216-1_25.

Liddell, T. M. y J. K. Kruschke (2018). «Analyzing ordinal data with metric models: What could possibly go wrong?» En: *Journal of Experimental Social Psychology* 79, págs. 328-348. DOI: 10.1016/j.jesp.2018.08.009. URL: https://www.sciencedirect.com/science/article/pii/S0022103117307746.

Lui, K.-J. (ago. de 2016). *Crossover Designs: Testing, Estimation, and Sample Size*. DOI: 10.1002/9781119114710.

Molanes-López, E. M., A. Rodriguez-Ascaso, E. Letón y J. Pérez-Martín (2021). «Assessment of Video Accessibility by Students of a MOOC on Digital Materials for All». En: *IEEE Access* 9, págs. 72357-72367. DOI: 10.1109/ACCESS.2021.3079199.

Pérez Martín, J., A. Rodríguez-Ascaso y E. Molanes-López (nov. de 2021). «Quality of the captions produced by students of an accessibility MOOC using a semi-automatic tool». En: *Universal Access in the Information Society* 20. DOI: 10.1007/s10209-020-00740-9.

Schweinberger, M. (2020). *Questionnaires and Surveys: Analyses with R*. 2020/12/11. https://slcladal.github.io/survey.html. The University of Queensland, Australia. School of Languages y Cultures. Brisbane.

Senn, S. (2022). Ed. por L. John Wiley. doi: 10.1002/0470854596.

Apéndice A

Preprocesado de los ficheros suministrados.

Este es el código en R con el que se transforman los ficheros que se suministran (ver Sección 4.2).

```r
library(readr)
library(purrr)
library(dplyr)
library(magrittr)
library(stringr)
library(forcats)
library(testit)
library(tidyr)

##### GRADE #####
## Usuarios que no quieren participar
no_want_users <- read_lines("data/original/ids_a_eliminar.txt")

# Leemos todos los archivos de grade CSV
grade_files <- list.files(
    "data/original", pattern = ".*grade.*.csv", full.names = TRUE
)

grade_df <- map_dfr(
    grade_files, ~ read_delim(.x, delim = ";", show_col_types = FALSE) %>%
        # Añadimos el número de fila para mantener la trazabilidad
        mutate(Userid = row_number() + 1) %>%
        # Movemos las columnas de identificación de fila a la primera posición
        relocate(Userid, .before = 2) %>%
        # Renombramos las columnas para que empiecen con mayúsculas
        rename_with(~ str_to_title(.), everything()) %>%
        # Renombramos para que sea más fácil procesar el campo Cohort Name
```

```
        rename("Cohort" = "Cohort Name") %>%
        # Eliminamos valores nulos y los que no quieren participar
        filter(!is.na(Cohort) & !Username %in% no_want_users)
)

assert("Comprobamos que no hay usuarios duplicados", grade_df %>%
    nrow() == grade_df %>%
    distinct(Username) %>%
    nrow())

# Creamos un tibble que tiene un campo con letras en lugar del valor de Cohorte
(groups <- grade_df %>%
    distinct(Cohort) %>%
    arrange(Cohort) %>%
    mutate(Group = LETTERS[1:n()]))

# Unimos los tibbles para asignar en grupo como letra en lugar de la cohorte
grade_df <- left_join(grade_df, groups) %>% dplyr::select(Username, Userid, Group)

##### PROFILE #####
profile_files <- list.files(
    "data/original", pattern = ".*student_profile.*.csv", full.names = TRUE
)

profile_df <- map_dfr(
    profile_files, ~ read_delim(.x, delim = ";", show_col_types = FALSE)
)

grade_df <- left_join(
    grade_df, profile_df %>% dplyr::select(-cohort), by = join_by(Username == username)
)


##### CONOC #####
conoc_files <- list.files(
    "data/original", pattern = ".*conoc.*.csv", full.names = TRUE)

conoc_df <- map_dfr(
    conoc_files, ~ read_delim(.x, delim = ";", show_col_types = FALSE)
)

conoc_df <- conoc_df %>%
    filter(Tries == 1) %>%
    rowwise() %>%
    mutate(
        level_of_knowledge =
            sum(c_across(starts_with(paste("Q", 1:10, "C", sep = ""))) == "correct")
    ) %>%
    dplyr::select(User, level_of_knowledge)

grade_df <- left_join(grade_df, conoc_df, by = join_by(Username == User))

##### TEST #####
# Leemos todos los archivos de test CSV
```

```r
test_files <- list.files(
    "data/original", pattern = ".*test.*.csv", full.names = TRUE
)

# Leer todos los archivos de test y los combinamos en un dataframe
test_df <- map_dfr(
    test_files, ~ read_delim(.x, delim = ";", show_col_types = FALSE) %>%
        # Añadimos un número de fila para mantener la trazabilidad
        mutate(Row = row_number() + 1) %>%
        # Añadimos la columna del número de test
        mutate(Test = sprintf("%02d", as.integer(str_extract(.x, "(?<=test)\\d+")))) %>%
        # Movemos las columnas de identificación de test y fila a la primera posición
        relocate(c(Test, Row), .before = 2)
) %>%
    # eliminamos los usuarios que no quieren participar
    filter(!User %in% no_want_users)

num_questions <- 18

# Nombre de los campos que contienen las respuestas al test
questions_original <- paste(
    "Q", seq(from = 1, by = 2, length.out = num_questions), "R", sep = ""
)

# Nombre de los campos que contienen las respuestas al test
comments_original <- paste(
    "Q", seq(from = 2, by = 2, length.out = num_questions - 1), "R", sep = ""
)

# Nombre de los campos que se usarán para renombrar los campos de respuesta al test
questions <- sprintf("Q%02d", seq(from = 1, by = 1, length.out = num_questions))
comments <- sprintf("C%02d", seq(from = 1, by = 1, length.out = num_questions - 1))
columns <- c(
    "Row", "Test", "User", "LastTry", questions_original, comments_original
)

# Procesamos el dataframe
# Con este operador del paquete magrittr hacemos las transformaciones in situ
test_df %<>%
    # Eliminamos las filas que no contienen información
    filter(Tries > 0) %>%
    # Convertimos LastTry a formato fecha
    mutate(LastTry = strptime(LastTry, format = "%Y-%m-%dT%H:%M:%SZ")) %>%
    # Seleccionamos las columnas que nos interesan
    dplyr::select(all_of(columns)) %>%
    # Extraemos la puntuación numérica de la pregunta
    mutate(across(questions_original, ~ if_else(
        startsWith(.x, "choice_"), as.integer(str_extract(.x, "\\d+")), NA_integer_)
    )) %>%
    # Renombramos los respuestas para que sean secuenciales
    rename(
        setNames(questions_original, questions),
        setNames(comments_original, comments)
    ) %>%
```

```
    # nos aseguramos de que el orden filas es el mismo que el de los ficheros.
    arrange("Test", "Row")

# Guardamos el número de filas para posterior comprobación
n_test <- test_df %>% nrow()

# Unimos los dataframes para tener el grupo y el UserID secuencial
test_df <- inner_join(
    test_df, grade_df, by = join_by(User == Username)
    ) %>% relocate(Group, .before = 2)

# Cambiamos los valores del campo User por los del UserID
test_df %<>%
    mutate(User = Userid) %>%
    dplyr::select(-Userid) %>%
    arrange(User, Test) # Ordenamos por usuario y test


##### CHECKS #####
assert(
    "Comprobamos que no hay preguntas duplicadas en el dataframe de test",
    n_test == test_df %>%
    distinct(Group, Test, User) %>%
    nrow()
)

assert(
    "Comprobamos que no hay valores nulos",
    test_df %>%
    dplyr::select(
        -c(comments, year_of_birth, gender, level_of_education, level_of_knowledge)
    ) %>% filter(if_any(everything(), is.na)) %>% nrow() == 0)


assert(
    "Comprobamos que no hay respuestas con valores incorrectos",
    sum(sort(unique(unlist(
        test_df %>% dplyr::select(all_of(questions))
    ))) == 0:5) == 6)


comments_df <- test_df %>%
    pivot_longer(
        cols = starts_with(c("Q", "C")),
        names_to = c(".value", "Question"),
        names_pattern = "(Q|C)(.*)") %>%
    rename(Response = Q, Comment = C) %>%
    filter(!is.na(Comment) & grepl("[a-zA-Z]", Comment)) %>%
    dplyr::select(Test, Row, Group, User, Question, Response, Comment) %>%
    arrange(Test, Group, Response, Row)


write_csv(comments_df, "./data/preprocess/comments.csv")
```

```
##### SAVE TO FILE #####
write_csv(
    test_df %>% dplyr::select(-all_of(comments)), "./data/preprocess/test_all.csv"
)
```

Apéndice B

Creación de los dataframes df_all y df_clean.

Código que transforma los datos preprocesados (ver Apéndice A) en los dataframes que se usan en el análisis estadístico.

```
# Leemos el tibble preprocesado
test_all_df <- read_delim(
    "./data/preprocess/test_all.csv",
    delim = ",", show_col_types = FALSE
)

# Eliminamos aquellos usuarios que no han hecho uno de los test
test_df <- test_all_df %>%
    group_by(User) %>%
    mutate(Rows = n()) %>%
    filter(Rows > 1) %>%
    ungroup()

##### SAVE TO FILE #####
write_csv(test_df, "./data/preprocess/test.csv")

df <- test_df %>%
    mutate(
        Period = as.factor(
            if_else(Test == "01", 1, 2)
        ),
        Treat = as.factor(
            if_else(Group == "A" & Test == "01" | Group == "B" & Test == "02", "A", "B")
        ),
        Seq = as.factor(
            if_else(Group == "A", "AB", "BA")
        ),
        Subject = as.factor(User)
    ) %>%
    dplyr::select(
        Seq, Period, Treat, Subject,
        gender, year_of_birth, level_of_education, starts_with("Q")
    ) %>%
```

```
    mutate_at(
        vars(starts_with("Q")), ~ (. + 1) %% 6
    ) %>%
    pivot_longer(
        cols = all_of(starts_with("Q")),
        names_to = "Question",
        values_to = "Response"
    ) %>%
    mutate(
        Question = relevel(as.factor(Question), ref = "Q18"),
        Response = factor(Response, ordered = TRUE)
    ) %>%
    arrange(Subject, Period, Question)

response_labels <- c(
    "No sé / No contesto",
    "Muy en desacuerdo",
    "En desacuerdo",
    "Neutral",
    "De acuerdo",
    "Muy de acuerdo"
)

question_labels <- c(
    "Los subtítulos del vídeo cumplen en general con los requisitos de accesibilidad.",
    "La posición de los subtítulos.",
    "El número de líneas por subtítulo.",
    "La disposición del texto respecto a la caja donde se muestran los subtítulos.",
    "El contraste entre los caracteres y el fondo.",
    "La corrección ortográfica y gramatical.",
    "La literalidad.",
    "La identificación de los personajes.",
    "La asignación de líneas a los personajes en los diálogos.",
    "La descripción de efectos sonoros.",
    "La sincronización de las entradas y salidas de los subtítulos.",
    "La velocidad de exposición de los subtítulos.",
    "El máximo número de caracteres por línea.",
    "La legibilidad de la tipografía.",
    "La separación en líneas diferentes de sintagmas nominales, verbales y preposicionales.",
    "La utilización de puntos suspensivos.",
    "La escritura de los números.",
    "Las incorrecciones en el habla."
)

question_labels_reduced <- c(
    "Valoración general",
    "Posición",
    "Número líneas",
    "Texto dentro caja",
    "Contraste",
    "Corrección",
    "Literalidad",
    "Identificación personajes",
    "Líneas/personajes",
    "Efectos sonoros",
    "Sincronización",
    "Velocidad",
    "Caracteres x línea",
    "Tipografía",
    "Separación sintagmas",
    "Puntos suspensivos",
    "Escritura números",
    "Incorrecciones habla"
)
```

```r
df <- df %>% mutate(
    Response_v = as.numeric(Response) - 1,
    Response_l = ordered(Response_v, labels = response_labels),
    Question_l = factor(Question, labels = question_labels),
    Question_lr = factor(Question, labels = question_labels_reduced)
)

dist <- df %>%
    xtabs(~ Question + Response, data = .) %>%
    dist(x = ., method = "euclidean")

cluster <- hclust(dist, method = "complete")
cuts <- factor(cutree(cluster, k = 3))

# Añadimos la columna cluster al dataframe
df <- inner_join(
    df,
    data.frame(
        Question = factor(names(cuts),
            levels = levels(df$Question)
        ),
        Cluster = as.factor(cuts)
    ),
    by = "Question"
)

write_csv(df, "./data/preprocess/test_lg.csv")
df_all <- df

df_all$Y <- model.matrix(~ Response - 1, data = df_all)

df_clean <- df %>% filter(Response != 0)
df_clean <- df_clean %>% mutate(
    Response = factor(Response, levels = levels(Response)[-1]),
    Response_l = ordered(Response_l, levels = levels(Response_l)[-1]),
    Level = as.ordered(
        ifelse(
            Response %in% c(1, 2),
            "Negative",
            ifelse(
                Response %in% c(4, 5),
                "Positive",
                "Neutral"
            )
        )
    )
)
df_0 <- df %>% filter(Response == 0)
```

Abstract English abstract, on the last page.

This is a bookdown template based on LaTeX memoir class.

Keywords Keyword in English, As a list.