Literatura y Estadística: problemas de autoría

Ciclo de talleres divulgativos

«Matemáticas en Acción 2008»

Fco. Javier Girón González-Torre

(Real Academia de Ciencias y Universidad de Málaga)

Santander, 22 de abril de 2009

CONTENIDO

- Algunos ejemplos de autoría literaria
- Estilometría
- Criterios estilísticos
  - Longitud de Palabra
  - Uso de las Palabras más Frecuentes
  - Distribución, Diversidad y Riqueza del Vocabulario
- Problemas estadísticos que se plantéan en los problemas de autoría
  - Análisis discriminante: clasificación
  - Análisis de cambios de estilo
  - Análisis de conglomerados
- La metodología bayesiana aplicada a los problemas de autoría: ventajas sobre los procedimientos clásicos

EJEMPLOS DE AUTORÍA LITERARIA

- La atribución de algunas obras literarias clásicas de la Grecia antigua.
- La controversia sobre las obras de Shakespeare-Bacon-Marlowe acerca de quien escribió algunas de las grandes obras atribuidas a Shakespeare.
- La autoría de ciertos escritos religiosos cristianos llamados Paulinos, de los que algunos forman parte del Nuevo Testamento. ¿Cuales fueron escritos por San Pablo y cuáles no?
- El autor de la novela del románticismo alemán Die Nachtwachen publicada en 1804 bajo el pseudónimo Bonaventura. Entre los posibles candiatos se encuentran Paul, Schelling, Brentano, Klingermann, Wetzel y Hoffmann (Wickmann, 1976).

* * *

- El problema de la autoría de doce de los artículos que componen El Federalista por parte de Alexander Hamilton y James Madison.
- La autoría del famoso libro de caballería Tirant lo Blanc.

ESTILOMETRÍA

Análisis estadístico de características cuantificables, no controlables de forma consciente y propias del autor y no del género, época o editor.

CRITERIOS ESTILÍSTICOS

1 Longitud de las frases

- Se utiliza poco por tener poco valor discriminatorio.

2 Longitud de palabra (nº de letras)

- Se ha utilizado desde muy antiguo (Mendenhall, 1887) para discriminar entre obras de Shakespeare, Bacon y Marlowe.
- Mosteller y Wallace (1964, 1984) lo usaron en su famoso estudio de la autoría de los Papeles Federalistas.

3 Palabras independientes del contexto

- La frecuencia de palabras como artículos, pronombres, conjunciones y preposiciones suele ser muy estable en los textos de un mismo autor. Por eso se usa como criterio estilístico para la atribución de autoría.
- Mosteller y Wallace (1964, 1984), Burrows (1987, 1992), Holmes (1992), Binongo (1994), Peng y Hengartner (2002).

4 Diversidad y riqueza del vocabulario

- Se basa en la hipótesis de que cada autor dispone de un vocabulario propio aunque tiende a emplear unas palabras más que otras.
- La frecuencia de aparición de las palabras en un texto refleja las características del autor.
- El objetivo es medir la cantidad de palabras de que dispone un autor a partir de las frecuencias de aparición de todas las palabras de un texto, a fin de caracterizar la diversidad a partir de algún índice de diversidad que es una medida de la dispersión cualitativa de variables categóricas.

5 Frecuencia de cada una de las letras del alfabeto del correspondiente idioma

- Este último criterio se ha empleado con éxito en problemas de discriminación de textos de la literatura inglesa de la época isabelina (Ledger and Merriam, 1994), y de la literatura clásica de la antigua Grecia (Belcastro y Eisinberg, 2002).

ALGUNOS DATOS SOBRE «EL FEDERALISTA»

- Obra de filosofía política escrita por Alexander Hamilton, John Hay y James Madison, para inducir a los ciudadanos del estado de Nueva York a ratificar la Constitución.
- Consta de 77 artículos, publicados entre 1787 y 1788 bajo el pseudónimo de Publius, más 8 ensayos que, junto con los anteriores, se publicaron en 1788 en forma de libro.
- Existía un acuerdo general acerca de qué artículos había escrito cada uno de los tres autores –5 de Jay, 14 de Madison y 51 de Hamilton– pero de los 15 restantes no se sabía su autoría, salvo que no eran de Jay y 3 eran artículos en colaboración entre Hamilton y Madison.
- El contenido político de los ensayos nunca proporcinó pruebas convincentes de su autoría.
- Los historiadores han ido cambiando la atribución de la autoría entre Hamilton y Madison según los cambios de clima político de cada época.

- Los escritos de Hamilton y Madison son difíciles de distinguir porque ambos fueron maestros del estilo popular del Spectator –complicado y retórico–.
- Como ejemplo de esta dificultad, el cálculo de la longitud de las frases de los artículos cuya autoría era conocida arrojó una media de 34.5 palabras para Hamilton y 34.6 para Madison.
- El criterio de la longitud de las frases, medida que se ha utilizado con éxito en otros problemas de autoría, no sirve en este caso.

ANÁLISIS ESTADÍSTICO DE LA AUTORÍA «EL FEDERALISTA»

F. Mosteller & D. L. Wallace (1964, 1984)

- Como criterios estilísticos de autoría se usaron la distribución de las frecuencias de uso de palabras y el uso de palabras no contextuales (independientes del contexto).
- El problema de atribución de la autoría se enfocó como un problema de Análisis Discriminante entre dos autores (Hamilton y Madison, pues Jay quedó descartado, desde el principio).
- Se consideraron como datos bien clasificados los artículos de El Federalista que se sabía con certeza eran de cada uno de los dos autores y otros artículos de carácter político de ambos.
- Se utilizaron métodos de inferencia bayesianos y frecuentistas. Ambos métodos condujeron a las mismas conclusiones: en todos los casos, la evidencia era concluyente a favor de la autoría de Madison.

- Como los datos presentaban sobredispersión, se utilizó la distribución Binomial negativa en lugar de la distribución Binomial.
- Las palabras que tuvieron mayor poder discriminatorio acerca de la autoría fueron las preposiciones by, to y las dos palabras equivalentes en inglés on y upon.

Análisis descriptivo-comparativo de la ocurrencia de la palabra by en textos de Hamilton, Madison y de los artículos en disputa

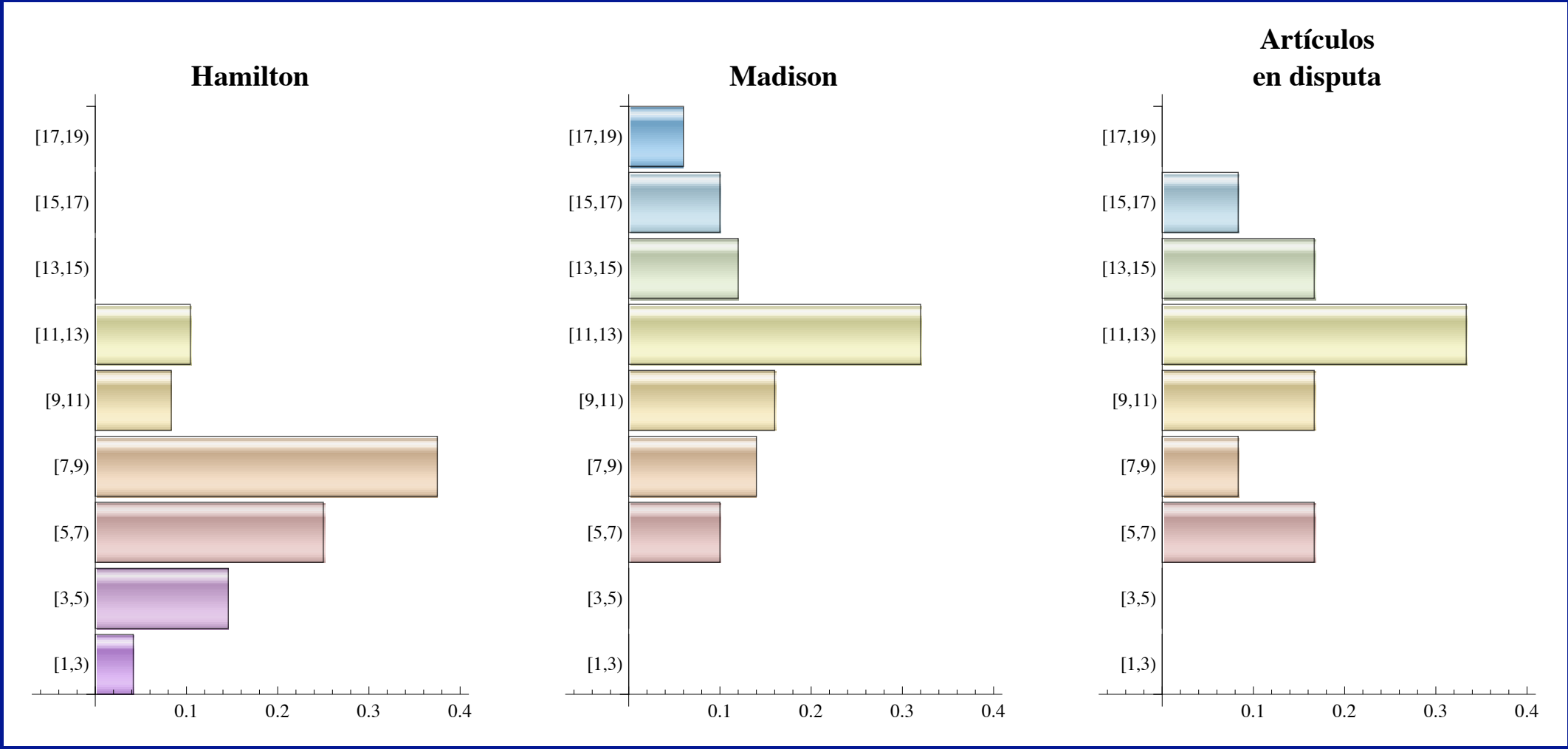


Figura 1. Distribución de las proporciones de ocurrencia de la palabra by, por cada mil palabras, en 48 artículos de Hamilton, 50 de Madison y los 12 artículos en disputa.

Análisis descriptivo-comparativo de la ocurrencia de la palabra to en textos de Hamilton, Madison y de los artículos en disputa

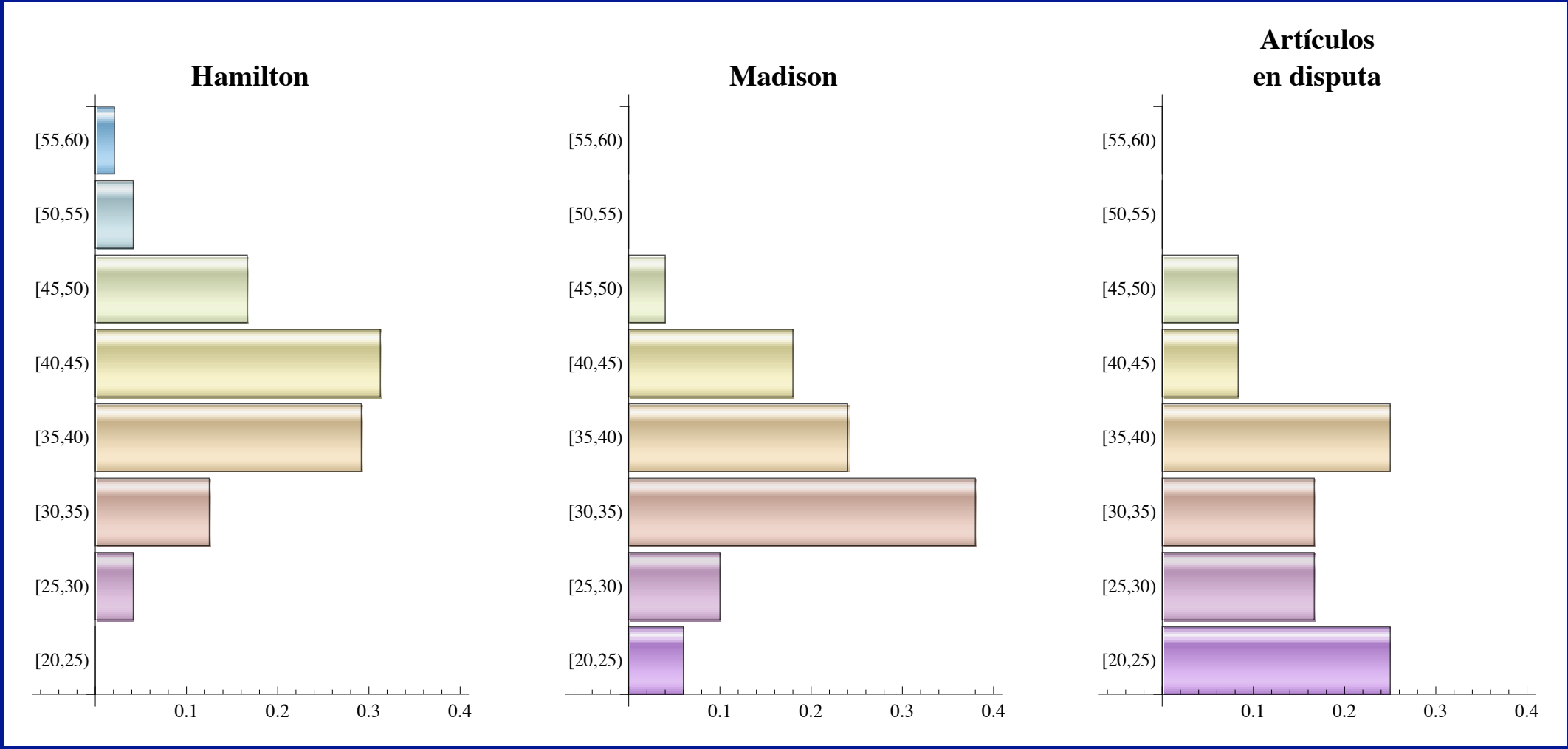
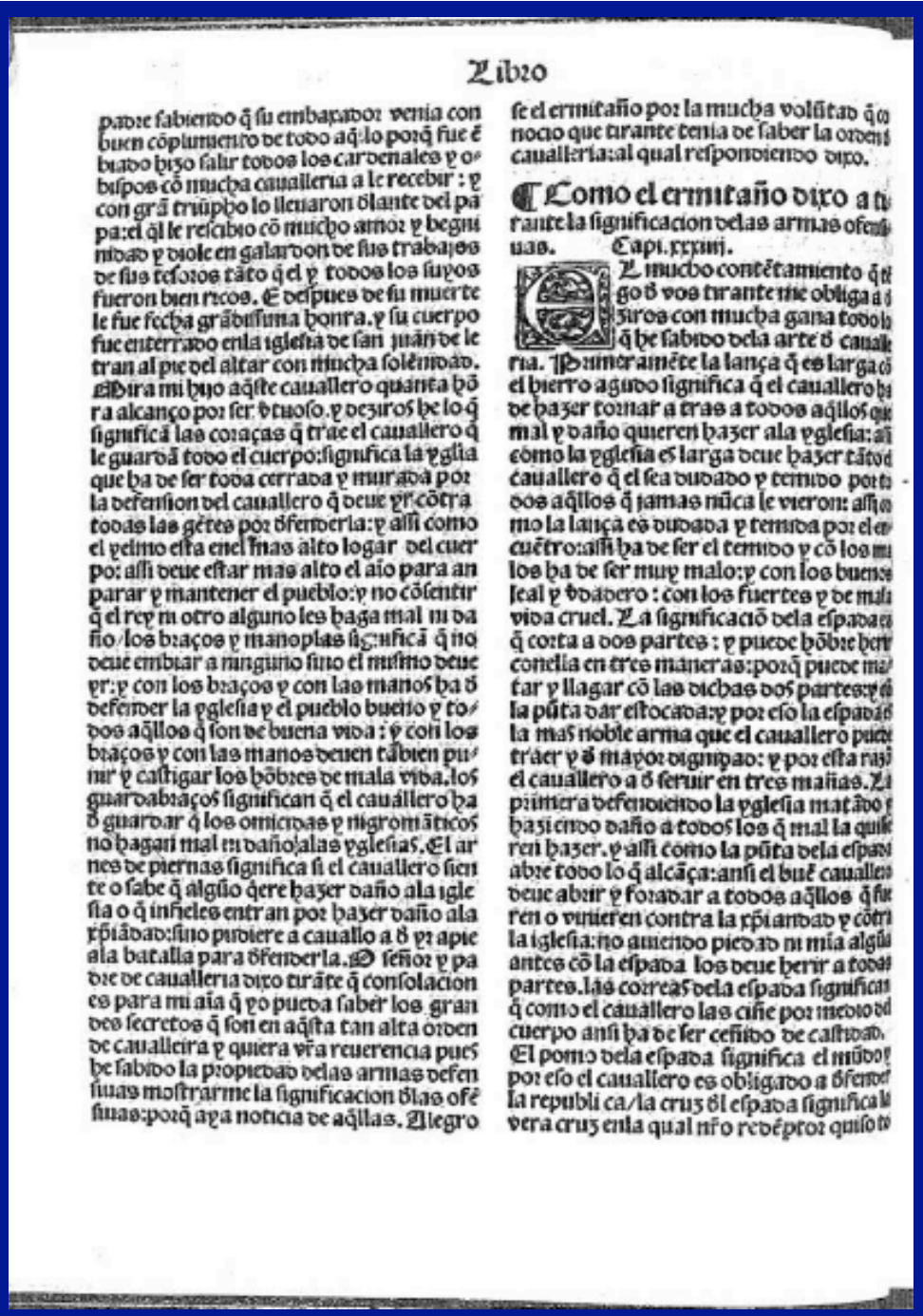


Figura 2. Distribución de las proporciones de ocurrencia de la palabra to, por cada mil palabras, en 48 artículos de Hamilton, 50 de Madison y los 12 artículos en disputa.

RESULTADOS DEL ANÁLISIS ESTADÍSTICO BASADO EN TÉCNICAS DEL ANÁLISIS DISCRIMINANTE

- Los resultados del análisis estadístico corroboran y complementan los de los historiadores.
  - Es extremadamente probable que Madison escribiera los 12 artículos de El Federalista objeto de controversia, con la posible excepción del artículo 55. Incluso para este artículo, es 80 veces más probable que fuese de Madison que de Hamilton.
  - El artículo 56, el siguiente más probable, es 800 veces más probable que sea de Madison que de Hamilton.
  - Curiosamente, para los artículos cuya autoría era menos clara para los historiadores –el 62 y el 63–, la eviencia estadística es contundente acerca de la paternidad literaria de Madison.

ANÁLISIS ESTADÍSTICO DEL LA AUTORÍA «Tirant lo Blanc»



Libro

padre sabiendo q̃ su embaxador venia con buen cõplimiento de todo aq̃llo porq̃ fue ẽ brauo hizo salir todos los cardenales y obispos cõ mucha caualleria a le recebir: y con grã triũpho lo llenaron ẽlante del papa:el q̃l le rescabio cõ mucho amor y beginidad y diole en galardon de sus trabajos de sus tesoros tãto q̃ el y todos los suyos fueron bien ricos. E despues de su muerte le fue fecha grãdissima honra.y su cuerpo fue enterrado enla iglesia de san juãn de letran al pie del altar con mucha solẽnidad. Mira mi hijo aq̃ste caualler quanta hõra alcanço por ser vtuoso.y dezirot he lo q̃ significã las coraças q̃ trae el caualler q̃ le guardã todo el cuerpo:significa la yglia que ha de ser toda cerrada y murada por la defension del caualler q̃ deue yr cõtra todas las gẽtes por ofenderla:y assi como el yelmo esta enel mas alto logar del cuerpo: assi deue estar mas alto el aio para amparar y mantener el pueblo:y no cõsentir q̃ el rey ni otro alguno les haga mal ni daño/los braços y manoplas significã q̃ no deue embiar a ninguno sino el mismo deue yr:y con los braços y con las manos ha d defender la yglesia y el pueblo bueno y todos aq̃llos q̃ son de buena vida: y con los braços y con las manos deuen tabien punir y castigar los hõbres de mala vida.los guardabraços significan q̃ el caualler ha d guardar q̃ los omicidas y nigromãticos no hagan mal ni daño/alas yglesias. El arnes de piernas significa si el caualler siente o sabe q̃ algũ quere hazer daño ala iglesia o q̃ infieles entran por hazer daño ala xpiãdad:sino pudiere a caualo a d yr apie ala batalla para ofenderla. O señor y padre de caualleria dixo tirãte q̃ consolacion es para mi aia q̃ yo pueda saber los grandes secretos q̃ son en aq̃sta tan alta orden de caualleria y quiera vra reuerencia pues he sabido la propiedad delas armas defensiuas mostrarme la significacion dlas ofensiuas:porq̃ aya noticia de aq̃llas. Alegro se el ermitaño por la mucha voluntad q̃ monocio que tirante tenia de saber la orden caualleria:al qual respondiendo dixo.

Como el ermitaño dixo a tirante la significacion delas armas ofensiuas. Capi. xxxiij.

EL mucho contẽtamiento q̃ tengo d vos tirante me obliga a deziros con mucha gana todo lo q̃ he sabido dela arte d caualleria. Primeramẽte la lança q̃ es larga cõ el hierro agudo significa q̃ el caualler ha de hazer tomar a tras a todos aq̃llos que mal y daño quieren hazer ala yglesia: assi como la yglesia es larga deue hazer tãto d caualler q̃ el sea dudado y temido por todos aq̃llos q̃ jamas nũca le vieron: assi como la lança es dudada y temida por el cuẽtro:assi ha de ser el temido y cõ los malos ha de ser muy malo:y con los buenos leal y dsabero: con los fuertes y de mala vida cruel. La significacio dela espada q̃ corta a dos partes: y puede hõbre herir conella en tres maneras:porq̃ puede matar y llagar cõ las dichas dos partes:y d la pũta dar estocada:y por eso la espada la mas noble arma que el caualler puede traer y d mayor dignidad: y por esta razõ el caualler a d seruir en tres mañas. La primera defendiendo la yglesia matãdo y haziendo daño a todos los q̃ mal la quisieren hazer.y assi como la pũta dela espada abre todo lo q̃ alcãça:assi el buẽ caualler deue abrir y foradar a todos aq̃llos q̃ fueren o vinieren contra la xpiandad y cõtra la iglesia:no auiendo piedad ni mia algũa antes cõ la espada los deue herir a todas partes. Las correas dela espada significan q̃ como el caualler las ciñe por medio del cuerpo assi ha de ser ceñido de castidad. El pomo dela espada significa el mũdo y por eso el caualler es obligado a ofender la republica/la cruz dl espada significa la vera cruz enla qual nro redẽptor quiso to

Facsímile de la portada y de una página de la primera edición en castellano de Tirant lo Blanc de 1511.

ALGUNOS DATOS SOBRE TIRANT LO BLANC

- Obra principal de la literatura catalana y/o valenciana

ALGUNOS DATOS SOBRE TIRANT LO BLANC

- Obra principal de la literatura catalana y/o valenciana
- Primera novela moderna en Europa (M. de Cervantes (1605), D. Alonso (1951), M. Vargas Llosa (1991))

ALGUNOS DATOS SOBRE TIRANT LO BLANC

- Obra principal de la literatura catalana y/o valenciana
- Primera novela moderna en Europa (M. de Cervantes (1605), D. Alonso (1951), M. Vargas Llosa (1991))
- Escrita entre 1460 y 1465

ALGUNOS DATOS SOBRE TIRANT LO BLANC

- Obra principal de la literatura catalana y/o valenciana
- Primera novela moderna en Europa (M. de Cervantes (1605), D. Alonso (1951), M. Vargas Llosa (1991))
- Escrita entre 1460 y 1465
- No fue publicada hasta 1490, en València por Nicolau Spindeler

ALGUNOS DATOS SOBRE TIRANT LO BLANC

- Obra principal de la literatura catalana y/o valenciana
- Primera novela moderna en Europa (M. de Cervantes (1605), D. Alonso (1951), M. Vargas Llosa (1991))
- Escrita entre 1460 y 1465
- No fue publicada hasta 1490, en València por Nicolau Spindeler
- 489 capítulos de longitudes muy desiguales
- Del orden de 418.000 palabras

ALGUNOS DATOS SOBRE TIRANT LO BLANC

- Obra principal de la literatura catalana y/o valenciana
- Primera novela moderna en Europa (M. de Cervantes (1605), D. Alonso (1951), M. Vargas Llosa (1991))
- Escrita entre 1460 y 1465
- No fue publicada hasta 1490, en València por Nicolau Spindeler
- 487 capítulos de longitudes muy desiguales
- Del orden de 418.000 palabras
- Existe un debate, que viene de muy antiguo, acerca de su autoría. En la edición original hay un prólogo debido a Joanot Martorell y un colofón escrito por el que se supone pudiera ser el segundo autor, Martí Joan de Galba
  - Autoría única (Joanot Martorell)
  - Doble autoría (J. Martorell and M. J. de Galba)

ALGUNOS DATOS SOBRE TIRANT LO BLANC

- Obra principal de la literatura catalana y/o valenciana
- Primera novela moderna en Europa (M. de Cervantes (1605), D. Alonso (1951), M. Vargas Llosa (1991))
- Escrita entre 1460 y 1465
- No fue publicada hasta 1490, en València por Nicolau Spindeler
- 487 capítulos de longitudes muy desiguales
- Del orden de 418.000 palabras
- Existe un debate, que viene de muy antiguo, acerca de su autoría. En la edición original hay un prólogo debido a Joanot Martorell y un colofón escrito por el que se supone pudiera ser el segundo autor, Martí Joan de Galba
  - Autoría única (Joanot Martorell)
  - Doble autoría (J. Martorell and M. J. de Galba)
- Tanto Martorell como Galba fallecieron antes de que se publicase la primera edición.

Argumentos a favor de la autoría única

- Se basan en la dedicatoria y el análisis literario de la obra

Givanel i Mas (1911), Vaeth (1918), Marinesco (1978), Martín de Riquer (1990), Hauf (1993), Chiner (1991, 93), Casanova (1994), Badia (1993).

Dedicatòria

E perquè en la present obra altri no puixa ésser increpat si defalliment algú trobat hi serà, jo, Joanot Martorell, cavaller, sols vull portar lo càrrec, e no altri ab mi; com per mi sols sia estada ventilada a servei del molt il·lustre Príncep e senyor Rei expectant Don Ferrando de Portugal la present obra, e començada a dos de giner de l'any mil quatre-cents e seixanta.

Dedicatoria

Y para que en la presente obra ningún otro pueda ser increpado si algún error fuere encontrado, yo, Joanot Martorell, caballero, sólo yo quiero llevar la carga, y no otro conmigo; pues por mi sólo ha sido ventilada en servicio del muy ilustre Príncipe y señor rey expectante Don Fernando de Portugal la presente obra, y comenzada el dos de enero del año mil cuatrocientos sesenta.

Argumentos a favor de la autoría compartida

- Se basan en el colofón y en el estudio estilístico del lenguaje. La mayoría cree que Galba fue algo más que simplemente un editor de la novela. Capdevila en el prólogo a su edición de 1924–29 resolvió, al parecer, el misterio de las cuatro partes del libro a las que se refiere el colofón: las aventuras en Inglaterra, la conquista de Rodas, el período en Constantinopla y las guerras del norte de África. Hay diversas y muy dispares opiniones acerca de las partes que escribió cada uno.
- Desde la primera edición en castellano de 1511, la novela se ha dividido en varias partes que se basan en el lugar donde transcurren las aventuras del protagonista: Inglaterra, Sicilia y Rodas, el Imperio Griego, el norte de África y la vuelta al Imperio Griego (cinco partes). No hay ninguna referencia a las cuatro partes en ninguna de las ediciones posteriores.

Martínez y Martínez (1916), Entwistle (1927), Moll (1933), Menéndez y Pelayo (1934), Martín de Riquer (1947), Alonso (1951), Coromines (1956), Nicolau d'Olwer (1961), Goerz (1967), Ferrando (1987, 89, 95), Bosh (1987), Rubiera (1990, 92), Wittlin (1990, 93), Hintz (1992).

Colofó

Aquí feneix lo llibre del valerós e estrenu cavaller Tirant lo Blanc, . . . , lo qual fon traduït d'anglès en llengua portuguesa, e aprés en vulgar llengua valenciana, per lo magnífic e virtuós cavaller Mossèn Joanot Martorell lo qual, per mort sua, no en pogué acabar de traduir sinó les tres parts. La quarta part, que és la fi del llibre, és estada traduïda, ..., per lo magnífic cavaller Mossèn Martí Joan de Galba; e si defalt hi serà trobat, vol sia atribuït a la sua ignorància; . . . .

Colofón

Aquí acaba el libro del virtuoso y valiente caballero Tirant lo Blanc, . . . , que fue traducido del inglés a la lengua portuguesa, y después en vulgar lengua valenciana, por el magnífico y virtuoso caballero Mossèn Joanot Martorell el cual, a causa de su muerte, no pudo acabar de traducir más que tres partes. La cuarta parte, que es el final del libro, ha sido traducida, . . . , por el magnífico caballero Mossèn Martí Joan de Galba; y si desfallecimiento fuera hallado, quiere sea atribuido a su ignorancia; . . . .

La postura de Martín de Riquer

- Martín de Riquer en 1947 estaba convencido de que la intervención de Galba en la totalidad de la obra fue progresiva desde el capítulo 349 en adelante, y prácticamente total desde el capítulo 416 hasta el final.

- Sin embargo, con motivo de la celebración del quinto centenario de la publicación de Tirant lo Blanc en 1990, Martín de Riquer se retracta de lo dicho anteriormente afirmando que Martorell es, sin discusión, el único autor del libro, y piensa que el colofón añadido por el impresor al manuscrito tras la muerte de Galba fue un error por parte de éste.

Más información sobre Tirant lo Blanc

- La dedicatoria no existe en la primera traducción castellana de 1511. El colofón (Deo Gratias) también es distinto, por lo que Tirant lo Blanc se publicó en castellano como novela anónima, pues en niguna otra parte consta que el autor de la novela fuese Joanot Martorell.

- Se puede encontrar la edición facsímil de la primera edición en castellano en la Biblioteca Virtual Miguel de Cervantes:

www.cervantesvirtual.com/servlet/SirveObras/02448175100804617400080/thm0000.htm

y de la primera edición en catalán en la Biblioteca Virtual Joan Lluís Vives

www.lluisvives.com/servlet/SirveObras/jlv/08146287511370295332268/index.htm

OBJETIVOS RELACIONADOS CON LA AUTORÍA
TIRANT LO BLANC

1.- Determinar si existe un estilo o más de un estilo

OBJETIVOS RELACIONADOS CON LA AUTORÍA
TIRANT LO BLANC

1.- Determinar si existe un estilo o más de un estilo

2.- En el caso en que se detecte más de un estilo:

a) Determinar la frontera (o fronteras) de estilo

b) Determinar qué es lo que caracteriza cada estilo

c) ¿Es el cambio de estilo progresivo o repentino? ¿Se puede atribuir a la existencia de dos autores?

OBJETIVOS RELACIONADOS CON LA AUTORÍA
TIRANT LO BLANC

1.- Determinar si existe un estilo o más de un estilo

2.- En el caso en que se detecte más de un estilo:

a) Determinar la frontera (o fronteras) de estilo

b) Determinar qué es lo que caracteriza cada estilo

c) ¿Es el cambio de estilo progresivo o repentino? ¿Se puede atribuir a la existencia de dos autores?

P. ¿Qué es lo que hace interesante y, a la vez, difícil el análisis de la autoría del Tirant lo Blanc comparado con otros problemas de autoría?

R. No tenemos textos de Martorell ni de Galba con los que comparar.

TRABAJOS ESTADÍSTICOS SOBRE LA AUTORÍA DE TIRANT LO BLANC

Basados en técnicas del Análisis de Datos

- Ginebra, J. y Cabos, S. (1998). Anàlisi Estadística de l'estil Literari: Aproximació a l'autoria del Tirant lo Blanc. ***Afers***, 29, 185–206.
- Riba, A. y Ginebra, J. (2000). Riquesa de Vocabulari i Homogeneitat d'estil en el Tirant lo Blanc. ***Revista de Catalunya***, 13, 99–118.
- Riba, A. y Ginebra, J. (2005). Change-Point Estimation in a Multinomial Sequence and Homogeneity of Literary Style. ***Journal of Applied Statistics***, (to appear).

Basados en técnicas Bayesianas

- Girón, J., Ginebra, J. y Riba, A. (2005). Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style. ***The American Statistician***, 59, nº 1, 1–12.

DESCRIPCIÓN DE LOS DATOS

- Se ha utilizado la edición de Martín de Riquer de 1983.
- Se han excluído los títulos de los capítulos y las palabras en cursiva.
- Queda un total de 398 242 palabras distribuídas en 489 capítulos.
- En el análisis estadístico solamente se utilizan los 425 capítulos que tienen más de 200 palabras. El más corto tiene 203 y el más largo 6521.

DESCRIPCIÓN DE LOS DATOS

- Se ha utilizado la edición de Martín de Riquer de 1983.
- Se han excluído los títulos de los capítulos y las palabras en cursiva.
- Queda un total de 398 242 palabras distribuídas en 489 capítulos.
- En el análisis estadístico solamente se utilizan los 425 capítulos que tienen más de 200 palabras. El más corto tiene 203 y el más largo 6521.

Criterios estilísticos utilizados

1 Longitud de palabra (nº de letras)

- Los datos, según este criterio, se categorizan en una tabla de contingencia de 425 filas ordenadas por 10 columnas: Tabla 1.

2 Palabras independientes del contexto

- Los datos —el número de apariciones de cada una de las 25 palabras independientes del contexto más frecuentes— se categorizan en una tabla de contingencia de 425 filas ordenadas por 25 columnas: Tabla 2.

L. P.	1	2	3	4	5	6	7	8	9	10+	N_i	l_i
Cap.1	21	59	44	19	33	20	16	17	9	17	255	4.47
2	53	113	80	49	52	33	28	36	16	16	476	4.15
3	109	274	239	128	112	110	76	51	43	32	1174	4.06
4	69	150	126	71	60	71	47	32	23	21	670	4.14
5	119	207	231	123	128	102	61	55	29	34	1089	4.09
6	69	136	126	69	60	61	37	27	15	15	615	3.96
7	32	63	51	18	29	28	15	15	19	13	283	4.34
8	26	52	41	19	27	29	11	16	5	11	237	4.25
...	...	...	...	...	...	...	...	...	...	...	...	...
480	78	123	150	57	54	65	42	25	34	13	641	4.05
481	159	282	262	137	124	122	63	71	56	46	1322	4.08
482	50	47	61	18	32	47	23	32	14	11	335	4.5
483	158	220	207	80	120	93	65	54	62	50	1109	4.21
484	59	67	68	37	26	32	15	14	17	6	341	3.82
485	96	174	106	57	77	86	42	54	24	25	741	4.18
486	45	88	91	46	40	28	13	30	11	10	402	3.94
487	48	49	62	53	41	36	21	9	16	13	348	4.2

Tabla 1. Longitud de palabra en número de letras: y_{ij} es el número de palabras de j letras en el capítulo i-ésimo.

Pals.	e	de	la	que	lo	en	a	per	no	l	los	com	...	N_i
Cap.1	12	15	9	8	10	6	1	4	1	7	5	2	...	255
2	26	28	19	9	10	12	11	8	3	2	1	3	...	476
3	66	46	48	53	26	20	22	20	19	9	13	11	...	1174
4	33	29	34	13	9	21	13	11	5	7	3	4	...	670
5	63	46	42	34	33	17	16	21	8	12	20	16	...	1089
6	35	15	27	23	27	16	13	11	7	10	6	3	...	615
7	20	20	10	16	3	6	4	5	5	5	0	2	...	283
8	13	9	13	6	1	9	6	6	4	5	1	4	...	237
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
480	45	32	15	33	19	9	9	19	3	12	11	2	...	641
481	82	54	42	40	43	26	32	37	12	28	12	7	...	1322
482	31	8	11	14	1	3	9	7	5	7	1	3	...	335
483	85	59	39	36	24	12	23	16	14	25	16	9	...	1109
484	31	19	13	12	10	7	15	3	2	7	4	0	...	341
485	59	66	28	14	12	21	7	8	2	15	7	1	...	741
486	28	29	14	10	14	13	4	14	1	8	5	3	...	402
487	29	13	8	10	8	4	4	4	2	10	4	3	...	348

Tabla 2. Uso de las 25 Palabras más Frecuentes: y_{ij} es el número de veces que la palabra j aparece en el capítulo i-ésimo.

ANÁLISIS EXPLORATORIO GRÁFICO DEL CAMBIO DE ESTILO

- Si el libro lo hubiese escrito un solo autor, las filas de las Tablas 1 y 2 serían homogéneas, respectivamente.
- Si los perfiles de las filas cambiaran súbitamente, esto sería indicio de la existencia de un segundo autor que asumió la escritura de la obra a partir de ese momento. Figuras 1 y 2.

ANÁLISIS EXPLORATORIO GRÁFICO DEL CAMBIO DE ESTILO

- Si el libro lo hubiese escrito un solo autor, las filas de las Tablas 1 y 2 serían homogéneas, respectivamente.
- Si los perfiles de las filas cambiaran súbitamente, esto sería indicio de la existencia de un segundo autor que asumió la escritura de la obra a partir de ese momento. Figuras 1 y 2.

Conclusiones provisionales del análisis gráfico

- El análisis gráfico sugiere que hay un cambio brusco en algunas categorías de las Figuras 1 y 2, entre los capítulos 350 a 400. Parece "improbable" que se deba a la evolución del estilo de un único autor.
- Hay algunos capítulos que aparecen como mal clasificados antes y después de la frontera. Sugiere la existencia de un segundo autor que escribiría sustancialmente la última parte del libro y también retocaría algunos capítulos de la primera parte.

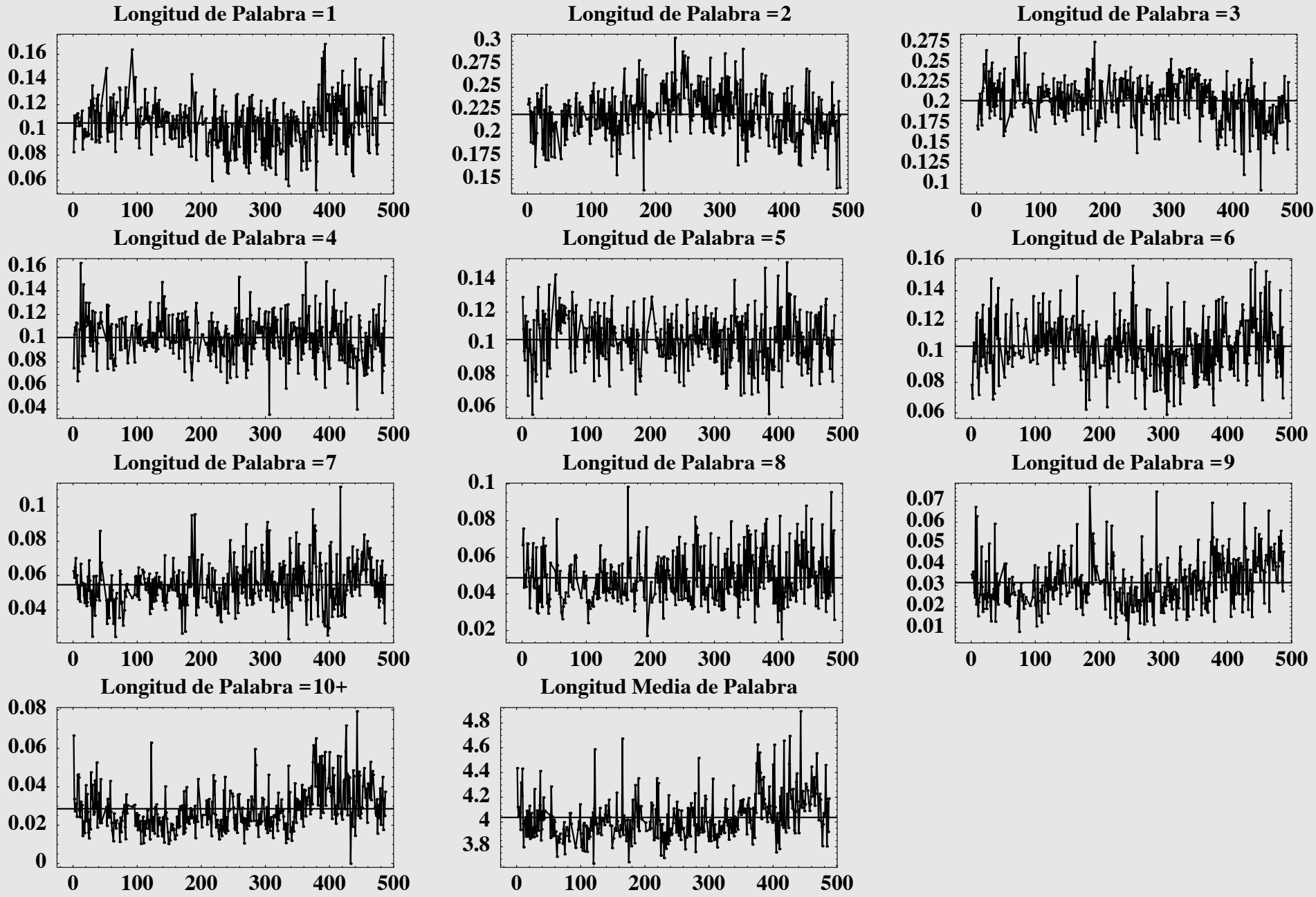
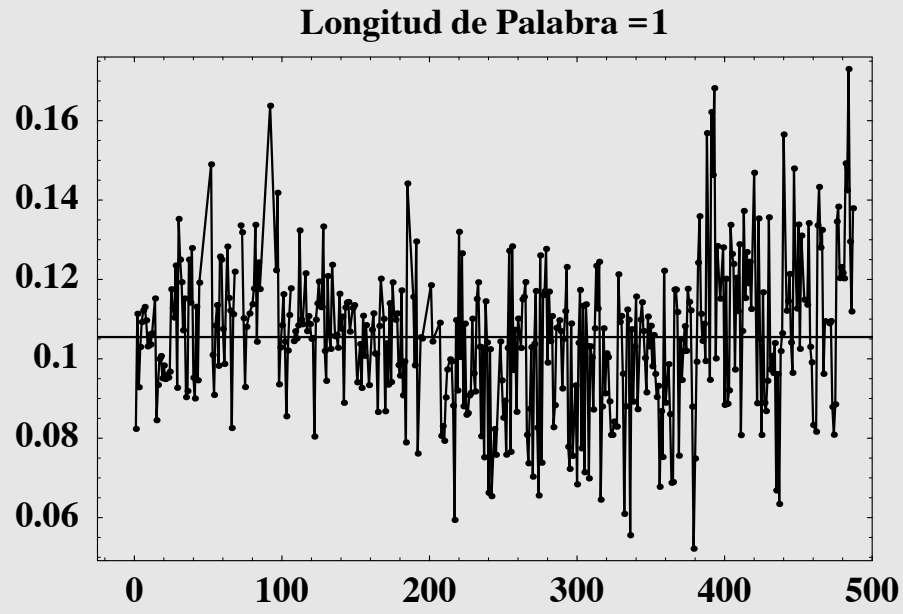
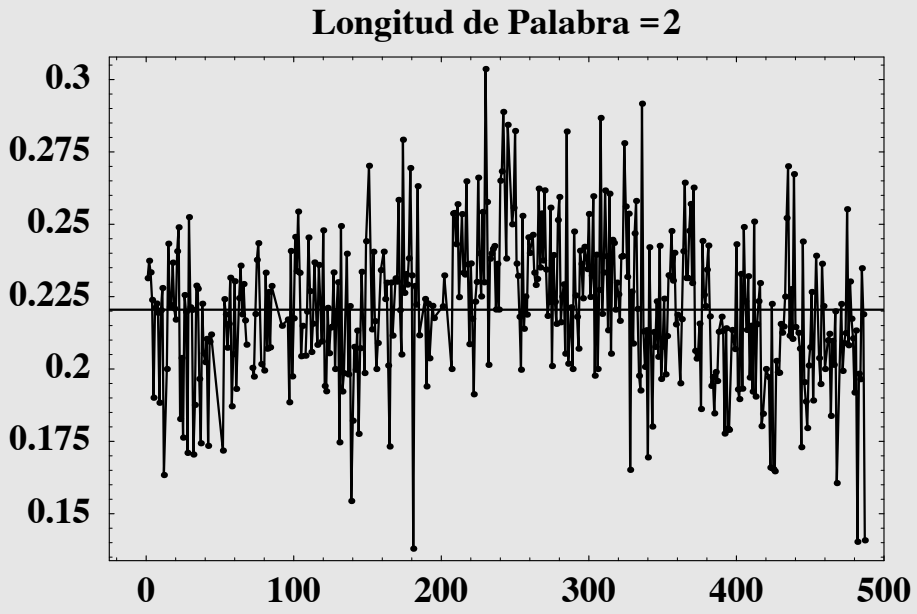
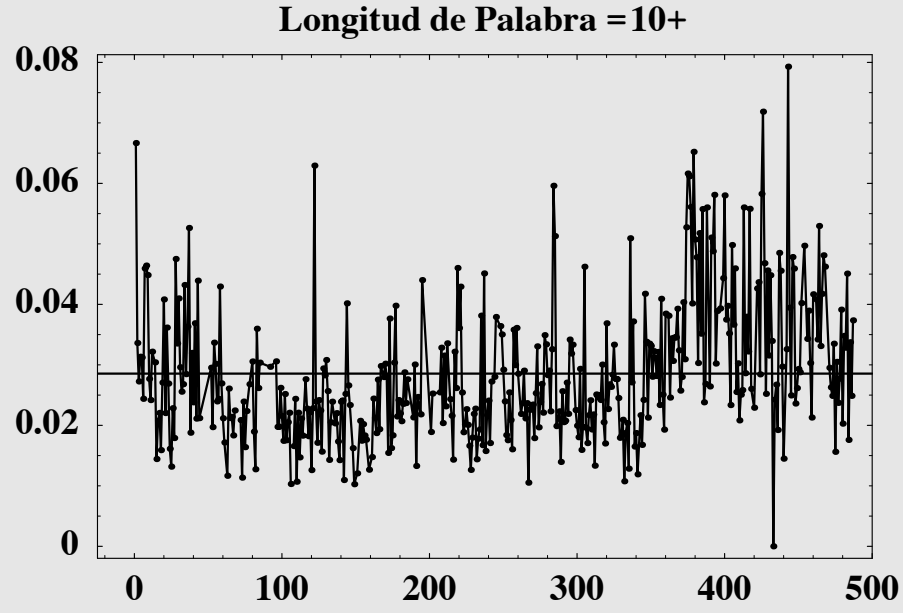


Figura 1a. Sucesiones de las proporciones de palabras con 1, 2, . . . , 9 y más de 9 letras en cada capítulo, y sucesión del promedio de la longitud de palabra.







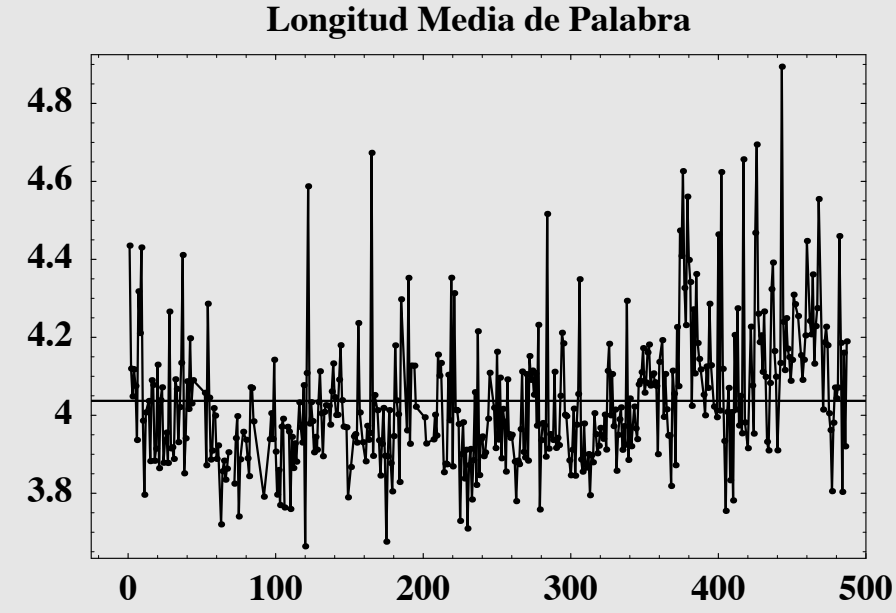


Figura 1b. Sucesiones de las proporciones de palabras con 1, 2 y 10 o más letras en cada capítulo, y sucesión del promedio de la longitud de palabra.

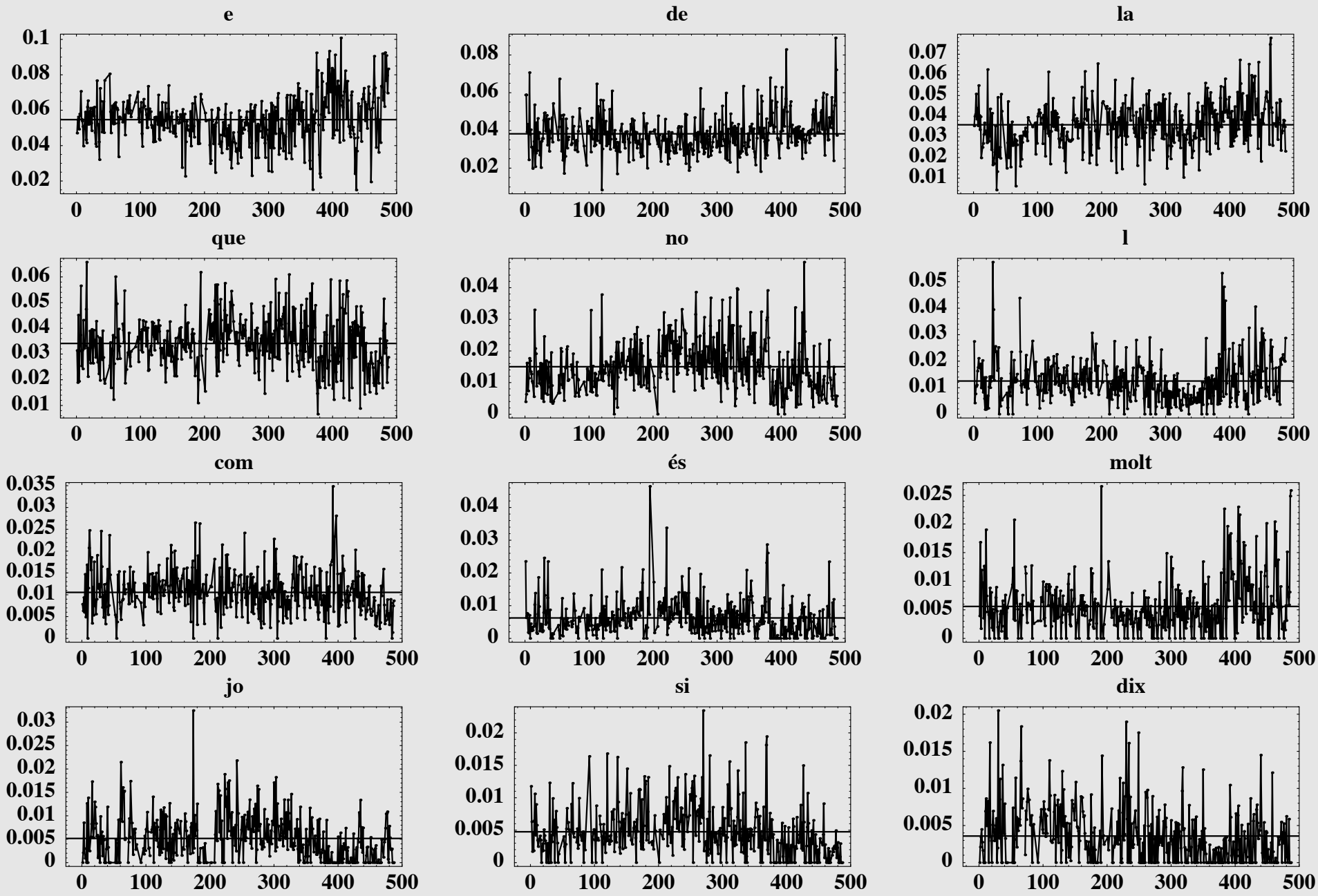
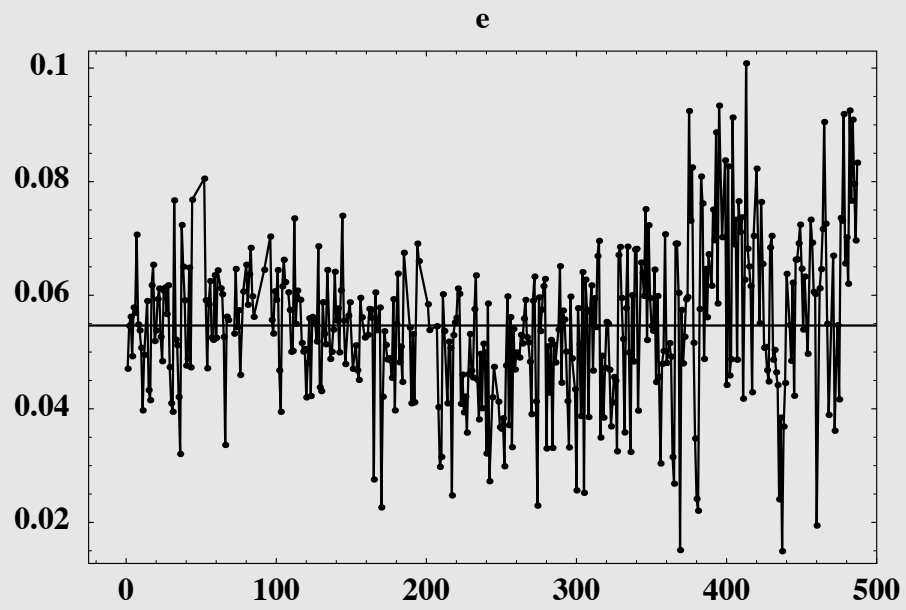
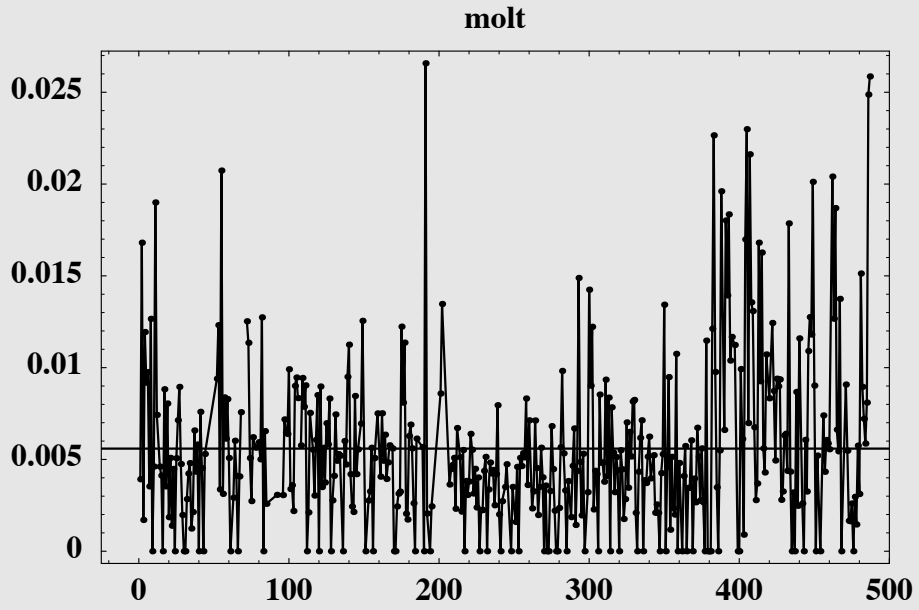
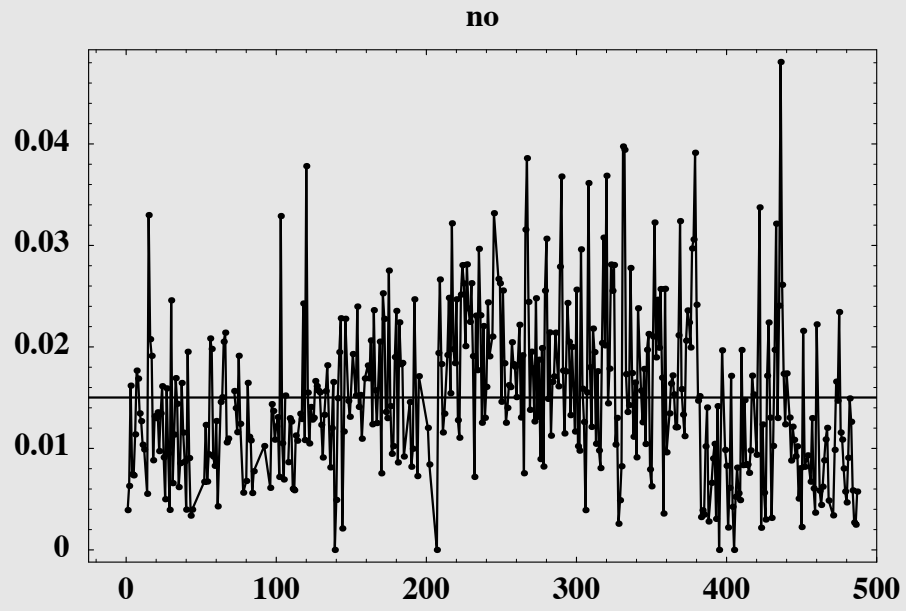


Figura 2a. Sucesiones de las frecuencias de aparición en cada capítulo de las palabras Green e, de, la, que, no, l, com, és, molt, jo, si y dix.







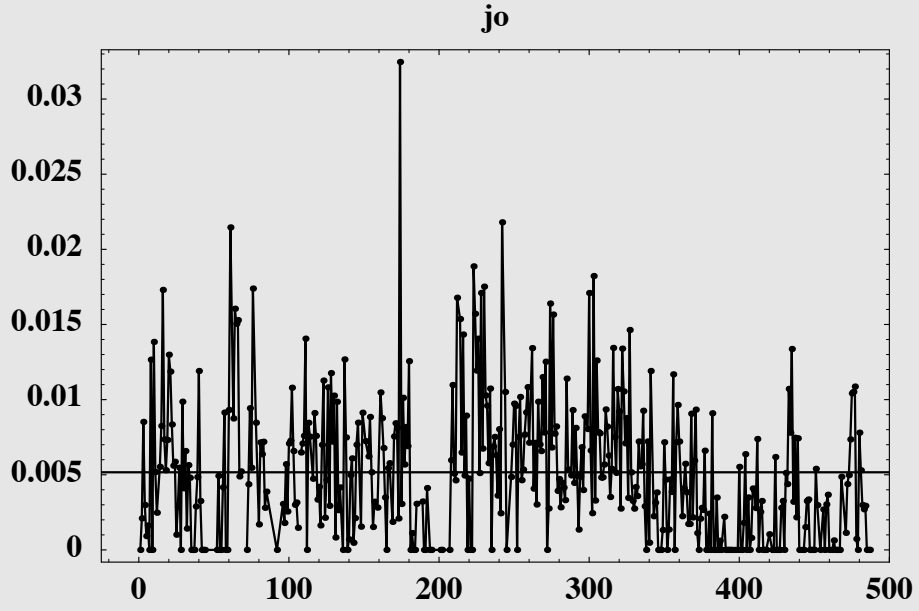


Figura 2b. Sucesiones de las frecuencias de aparición en cada capítulo de las palabras e, molt, no y jo.

MODELOS DE PUNTO DE CAMBIO EN SUCESIONES DE DATOS MULTINOMIALES

- El modelo estadístico estándar para las filas de una tabla de contingencia es la distribución multinomial.
- Para cada capítulo i, las filas y_i de las Tablas 1 y 2, siguen una distribución multinomial

$$y_i \mid N_i, \theta_i \sim \mathbf{Mu}_{l-1}(N_i, \theta_i)$$

- Una sucesión de variables multinomiales ordenadas $(y_1, \ldots, y_n)$ presenta un cambio de modelo (change-point) en el punto r si

$$y_i \mid N_i, \theta_i, r \sim \begin{cases} \mathbf{Mu}_{l-1}(N_i, \theta_a) & \text{si} \quad i \leq r \\ \mathbf{Mu}_{l-1}(N_i, \theta_d) & \text{si} \quad i > r \end{cases} \qquad \theta_a \neq \theta_d$$

Análisis bayesiano de un punto de cambio

- Se basa en el cálculo de la distribución a posteriori conjunta de los tres parámetros de interés: el punto de cambio r y los parámetros de las distribuciones multinomiales antes del cambio $\boldsymbol{\theta}_a$ y después del cambio $\boldsymbol{\theta}_d$.

Análisis bayesiano de un punto de cambio

- Se basa en el cálculo de la distribución a posteriori conjunta de los tres parámetros de interés: el punto de cambio r y los parámetros de las distribuciones multinomiales antes del cambio $\boldsymbol{\theta}_a$ y después del cambio $\boldsymbol{\theta}_d$.
- La distribución a posteriori conjunta de los tres parámetros $r, \boldsymbol{\theta}_a, \boldsymbol{\theta}_d$ se calcula multiplicando la función de verosimilitud del modelo de punto de cambio por la distribución a priori de $r, \boldsymbol{\theta}_a, \boldsymbol{\theta}_d$.

Análisis bayesiano de un punto de cambio

- Se basa en el cálculo de la distribución a posteriori conjunta de los tres parámetros de interés: el punto de cambio r y los parámetros de las distribuciones multinomiales antes del cambio $\boldsymbol{\theta}_a$ y después del cambio $\boldsymbol{\theta}_d$.
- La distribución a posteriori conjunta de los tres parámetros $r, \boldsymbol{\theta}_a, \boldsymbol{\theta}_d$ se calcula multiplicando la función de verosimilitud del modelo de punto de cambio por la distribución a priori de $r, \boldsymbol{\theta}_a, \boldsymbol{\theta}_d$.
- Como las dos teorías sobre la autoría están en marcado conflicto, para ser neutrales usamos como distribuciones a priori sobre $r, \boldsymbol{\theta}_a, \boldsymbol{\theta}_d$ distribuciones no informativas e independientes.

Análisis bayesiano de un punto de cambio

- Se basa en el cálculo de la distribución a posteriori conjunta de los tres parámetros de interés: el punto de cambio r y los parámetros de las distribuciones multinomiales antes del cambio $\boldsymbol{\theta}_a$ y después del cambio $\boldsymbol{\theta}_d$.
- La distribución a posteriori conjunta de los tres parámetros $r, \boldsymbol{\theta}_a, \boldsymbol{\theta}_d$ se calcula multiplicando la función de verosimilitud del modelo de punto de cambio por la distribución a priori de $r, \boldsymbol{\theta}_a, \boldsymbol{\theta}_d$.
- Como las dos teorías sobre la autoría están en marcado conflicto, para ser neutrales usamos como distribuciones a priori sobre $r, \boldsymbol{\theta}_a, \boldsymbol{\theta}_d$ distribuciones no informativas e independientes.
- Contrastar la hipótesis de que solamente hay un autor es equivalente a que no hay punto de cambio en la sucesión de capítulos; es decir, contrastar

$$\boldsymbol{H_0 : r = n} \quad \textbf{frente a} \quad \boldsymbol{H_1 : r \neq n}$$

Inferencias sobre el capítulo r donde se produce el cambio de estilo

- Se calculan a partir de la distribución marginal a posteriori de r : Tabla 3.

Inferencias sobre el capítulo r donde se produce el cambio de estilo

- Se calculan a partir de la distribución marginal a posteriori de r : Tabla 4.
- La evidencia a favor de la hipótesis de la autoría única se calcula a partir de la probabilidad a posteriori de H_0 para las Tablas 1 y 2.

$$\Pr(H_0 \mid \text{datos}) \approx 0$$

Inferencias sobre el capítulo r donde se produce el cambio de estilo

- Se calculan a partir de la distribución marginal a posteriori de r : Tabla 4.
- La evidencia a favor de la hipótesis de la autoría única se calcula a partir de la probabilidad a posteriori de H_0 para las Tablas 1 y 2.

$$\Pr(H_0 \mid \text{datos}) \approx 0$$

- La probabilidad de que haya más de un cambio de estilo, sugerida por el análisis de la Tabla 4 y de las Figuras 1 y 2 es bastante grande.
  - Requiere la extensión del modelo a puntos de cambio múltiples y simulación vía MCMC
  - Hay un número indeterminado de pequeños cambios de estilo, aparte de los principales dados por la Tabla 4, probablemente fruto de intervenciones menores de cada autor en la parte del otro.
  - Se decidió, en vez de estudiar los cambios de estilo múltiple, realizar un análisis bayesiano de conglomerados.

	Longitud de palabra	Palabras más frecuentes
Capítulos	(Tabla 1)	(Tabla 2)
1 a 343	.0000	.0000
344	.0003	.0000
345	.0158	.0000
346	.0032	.0000
347	.0042	.0000
348	.0017	.0000
349 a 367	.0000	.0000
368	.0001	.0000
369	.0007	.0000
370	.0012	.0000
371	.9655	.0000
372	.0038	.0000
373	.0033	.0000
374 a 377	.0000	.0000
378	.0000	.0003
379	.0000	.0278
380	.0000	.0919
381	.0000	.0729
382	.0000	.8070
383	.0000	.0001
384 a 487	.0000	.0000

Tabla 3. Distribución a posteriori del punto de cambio r para las sucesiones de filas de la Tabla 1, y de la Tabla 2 cuando solamente se consideran las palabras e, de, la, que, no, l, com, és, molt, jo, si y dix.

Inferencias sobre los parámetros de las dos distribuciones multinomiales antes y después del cambio de estilo

Una vez establecida la existencia y el lugar donde se produce el cambio de estilo principal,

- ¿qué componentes de los vectores θ_a y θ_d cambian, y en qué grado, tras el cambio de estilo?

Inferencias sobre los parámetros de las dos distribuciones multinomiales antes y después del cambio de estilo

Una vez establecida la existencia y el lugar donde se produce el cambio de estilo principal,

- ¿qué componentes de los vectores $\boldsymbol{\theta}_a$ y $\boldsymbol{\theta}_d$ cambian, y en qué grado, tras el cambio de estilo?

- Las distribuciones a posteriori de $\boldsymbol{\theta}_a$ y $\boldsymbol{\theta}_d$ son complicadas —mixturas de distribuciones de Dirichlet— por lo que se hace necesario simular de éstas para obtener muestras de las correspondientes distribuciones a posteriori.

Inferencias sobre los parámetros de las dos distribuciones multinomiales antes y después del cambio de estilo

Una vez establecida la existencia y el lugar donde se produce el cambio de estilo principal,

- ¿qué componentes de los vectores θ_a y θ_d cambian, y en qué grado, tras el cambio de estilo?
- Las distribuciones a posteriori de θ_a y θ_d son complicadas —mixturas de distribuciones de Dirichlet— por lo que se hace necesario simular de éstas para obtener muestras de las correspondientes distribuciones a posteriori.
- Diversos motivos para comparar ambos vectores —relacionados con la forma de la función discriminante bayesiana para comparar dos distribuciones multinomiales— nos inducen a considerar la distribución a posteriori de los

$$\log(\theta_{aj}/\theta_{dj})$$

Figuras 3 y 4.

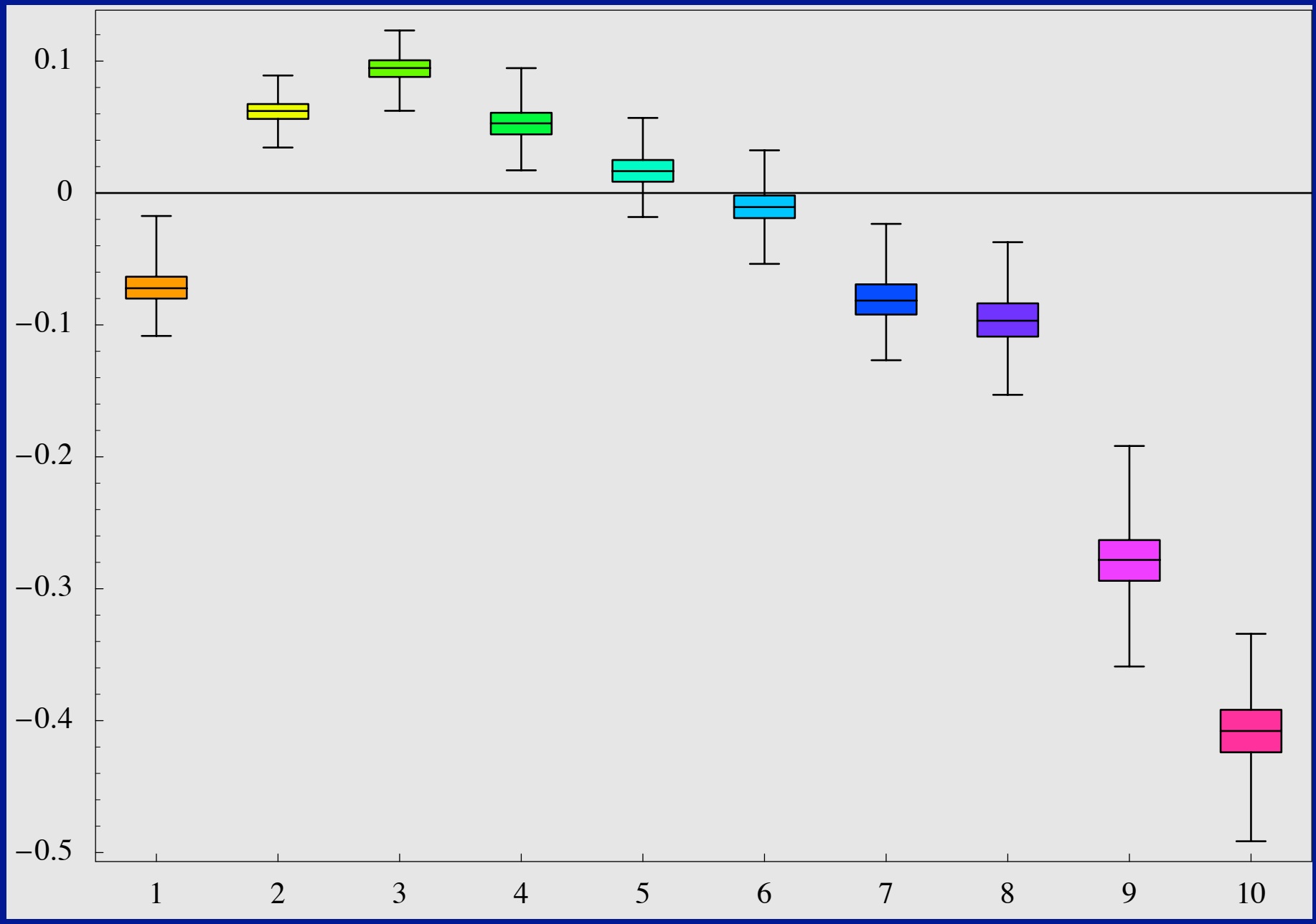


Figura 3. Diagrama de cajas de una muestra simulada de la distribución a posteriori del $\log(\theta_{aj}/\theta_{dj})$, $j = 1, \ldots, 10$, para la Tabla 1.

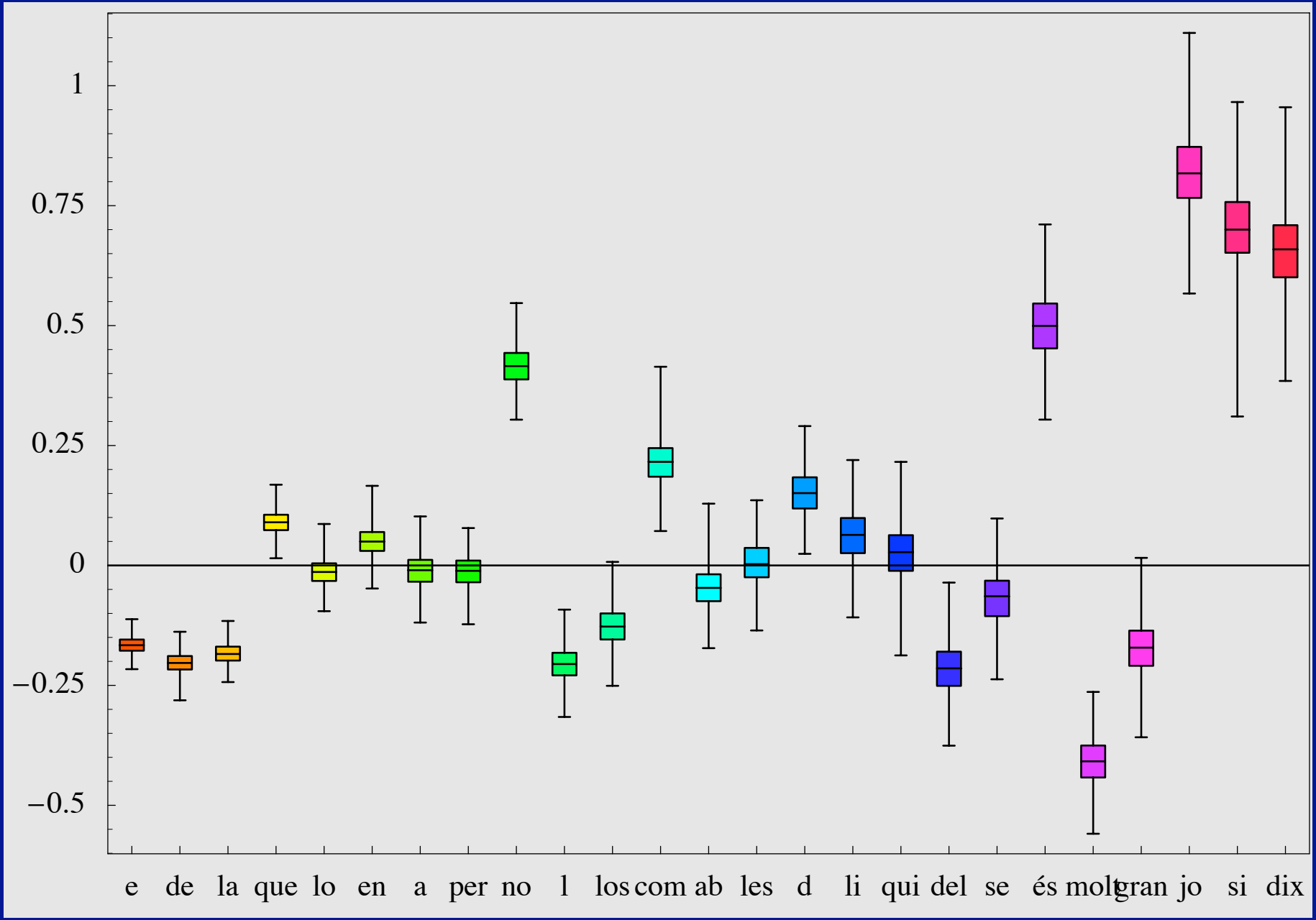


Figura 4. Diagrama de cajas de una muestra simulada de la distribución a posteriori del $\log(\theta_{aj}/\theta_{dj})$, $j = 1, \ldots, 25$, para la Tabla 2 completa.

ANÁLISIS DE CONGLOMERADOS DE OBSERVACIONES MULTINOMIALES

- En el análisis de cambio de estilo, la sucesión de filas original se particiona en dos subsucesiones más homogéneas que la original, forzando a que se respete el orden original de las observaciones.

ANÁLISIS DE CONGLOMERADOS DE OBSERVACIONES MULTINOMIALES

- En el análisis de cambio de estilo, la sucesión de filas original se particiona en dos subsucesiones más homogéneas que la original, forzando a que se respete el orden original de las observaciones.
- El análisis de conglomerados particiona el conjunto de todos los datos en dos grupos más homogéneos que el total pero sin imponer ninguna restricción en el orden para formar los dos grupos.

ANÁLISIS DE CONGLOMERADOS DE OBSERVACIONES MULTINOMIALES

- En el análisis de cambio de estilo, la sucesión de filas original se particiona en dos subsucesiones más homogéneas que la original, forzando a que se respete el orden original de las observaciones.
- El análisis de conglomerados particiona el conjunto de todos los datos en dos grupos más homogéneos que el total pero sin imponer ninguna restricción en el orden para formar los dos grupos.
- El análisis de conglomerados bayesiano se basa en modelos de mixtura. Una ventaja del enfoque bayesiano, a diferencia de otros, es que permite asignar las observaciones a los conglomerados de modo probabilístico, en lugar de utilizar reglas de clasificación del tipo si/no, que no permiten medir el grado de pertenencia a cada conglomerado.

MODELO BASADO EN MIXTURAS DE MULTINOMIALES

- Cada una de las filas $\boldsymbol{y_i}$ de las Tablas 1 y 2, proviene de una distribución multinomial $\mathbf{Mu}_{l-1}(N_i, \theta_1)$ con probabilidad $\boldsymbol{p}$, y con probabilidad $\mathbf{1} - \boldsymbol{p}$ de una distribución multinomial $\mathbf{Mu}_{l-1}(N_i, \theta_2)$, es decir

$$\boldsymbol{y_i} \,|\, \boldsymbol{N_i}, \boldsymbol{p}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \sim \boldsymbol{p}\, \mathbf{Mu}_{l-1}(\boldsymbol{N_i}, \boldsymbol{\theta_1}) + (\mathbf{1} - \boldsymbol{p})\, \mathbf{Mu}_{l-1}(\boldsymbol{N_i}, \boldsymbol{\theta_2}) \qquad (\boldsymbol{M})$$

- $\boldsymbol{p}$ representa la proporción de capítulos escritos por el primer autor.

MODELO BASADO EN MIXTURAS DE MULTINOMIALES

- Cada una de las filas $\boldsymbol{y_i}$ de las Tablas 1 y 2, proviene de una distribución multinomial $\mathbf{Mu}_{l-1}(N_i, \theta_1)$ con probabilidad $\boldsymbol{p}$, y con probabilidad $\mathbf{1} - \boldsymbol{p}$ de una distribución multinomial $\mathbf{Mu}_{l-1}(N_i, \theta_2)$, es decir

$$\boldsymbol{y_i} \,|\, \boldsymbol{N_i, p, \theta_1, \theta_2} \sim \boldsymbol{p}\, \mathbf{Mu}_{l-1}(\boldsymbol{N_i, \theta_1}) + (\mathbf{1} - \boldsymbol{p})\, \mathbf{Mu}_{l-1}(\boldsymbol{N_i, \theta_2}) \qquad (M)$$

- $\boldsymbol{p}$ representa la proporción de capítulos escritos por el primer autor.
- El modelo de mixtura (M) presenta un problema de identificabilidad . Se resuelve imponiendo la restricción $\boldsymbol{p} \geq \mathbf{.5}$.

MODELO BASADO EN MIXTURAS DE MULTINOMIALES

- Cada una de las filas y_i de las Tablas 1 y 2, proviene de una distribución multinomial $\mathbf{Mu}_{l-1}(N_i, \theta_1)$ con probabilidad p, y con probabilidad $1-p$ de una distribución multinomial $\mathbf{Mu}_{l-1}(N_i, \theta_2)$, es decir

$$y_i \,|\, N_i, p, \theta_1, \theta_2 \sim p\,\mathbf{Mu}_{l-1}(N_i, \theta_1) + (1-p)\,\mathbf{Mu}_{l-1}(N_i, \theta_2) \qquad (M)$$

- p representa la proporción de capítulos escritos por el primer autor.
- El modelo de mixtura (M) presenta un problema de identificabilidad . Se resuelve imponiendo la restricción $p \geq .5$.
- El modelo de mixtura (M) no permite la asignación de capítulos a los dos autores. Se resuelve introduciendo variables latentes dicotómicas z_i, $i = 1, \ldots, 425$, definidas por

$$z_i = \begin{cases} 1 & \text{si } y_i \text{ es del primer autor} \\ 0 & \text{si } y_i \text{ es del segundo autor} \end{cases}$$

- La introducción de las variables latentes z_i permite:
  - Simplificar el modelo de mixtura (M).
  - El cálculo de la distribución a posteriori de los parámetros p, θ_1, θ_2 y de la asignación de los capítulos $z = (z_1, \ldots, z_n)$ mediante la aplicación del algoritmo de muestreo de Gibbs cuando la distribución a priori es conjugada respecto de la verosimilitud, como ocurre en nuestro caso al utilizar distribuciones no informativas.

- La introducción de las variables latentes z_i permite:
  - Simplificar el modelo de mixtura (M).
  - El cálculo de la distribución a posteriori de los parámetros p, θ_1, θ_2 y de la asignación de los capítulos $z = (z_1, \ldots, z_n)$ mediante la aplicación del algoritmo de muestreo de Gibbs cuando la distribución a priori es conjugada respecto de la verosimilitud, como ocurre en nuestro caso al utilizar distribuciones no informativas.
- La probabilidad a posteriori de pertenencia del capítulo i-ésimo al primer autor es precisamente la esperanza a posteriori $E(z_i \mid y_1, \ldots, y_n)$. Figuras 6a y 6b.

Inferencias sobre la proporción de capítulos del primer autor p

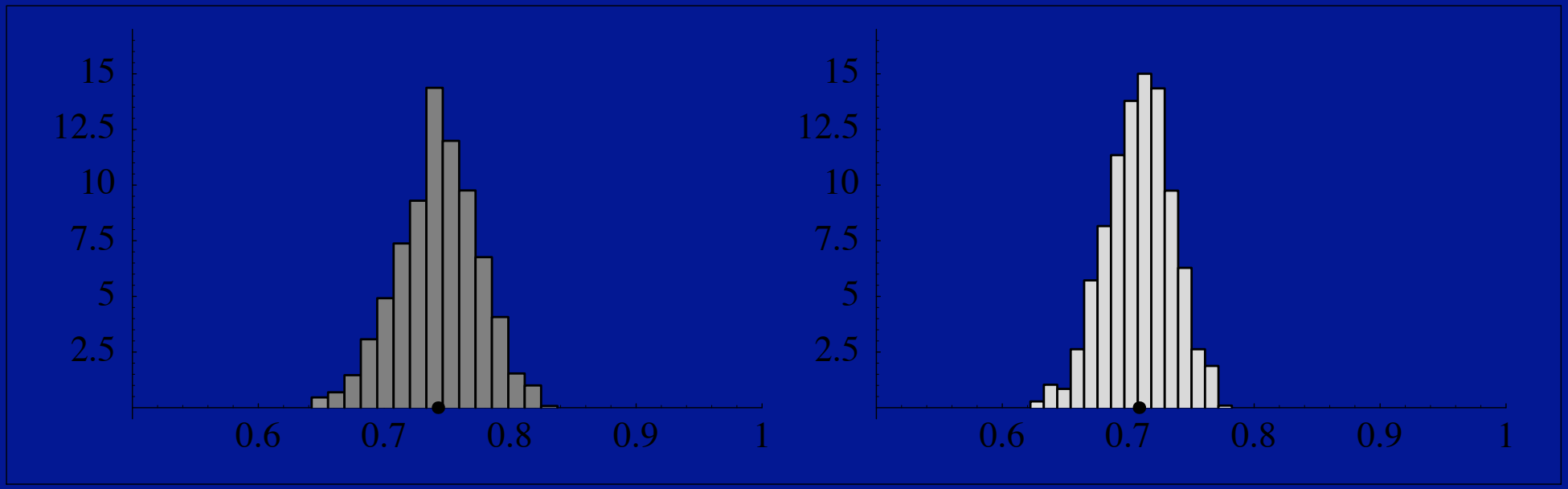


Figura 5. Histograma de una muestra simulada de la distribución a posteriori de p —proporción de capítulos del primer autor—, para las filas de la Tabla 1 (izquierda) y las filas de la Tabla 2 (derecha) restringida a las columnas ***e, de, la, que, no, l, com, és, molt, jo, si*** y ***dix***.

- La evidencia a favor de la hipótesis de la autoría única (hipótesis $H_0 : p = 1$) frente a la doble autoría (hipótesis $H_1 : p < 1$) se calcula, en este caso, aplicando métodos bayesianos de selección de modelos. En ambos casos, el resultado es

$$\Pr(H_0 \,|\, \text{datos}) \approx 0$$

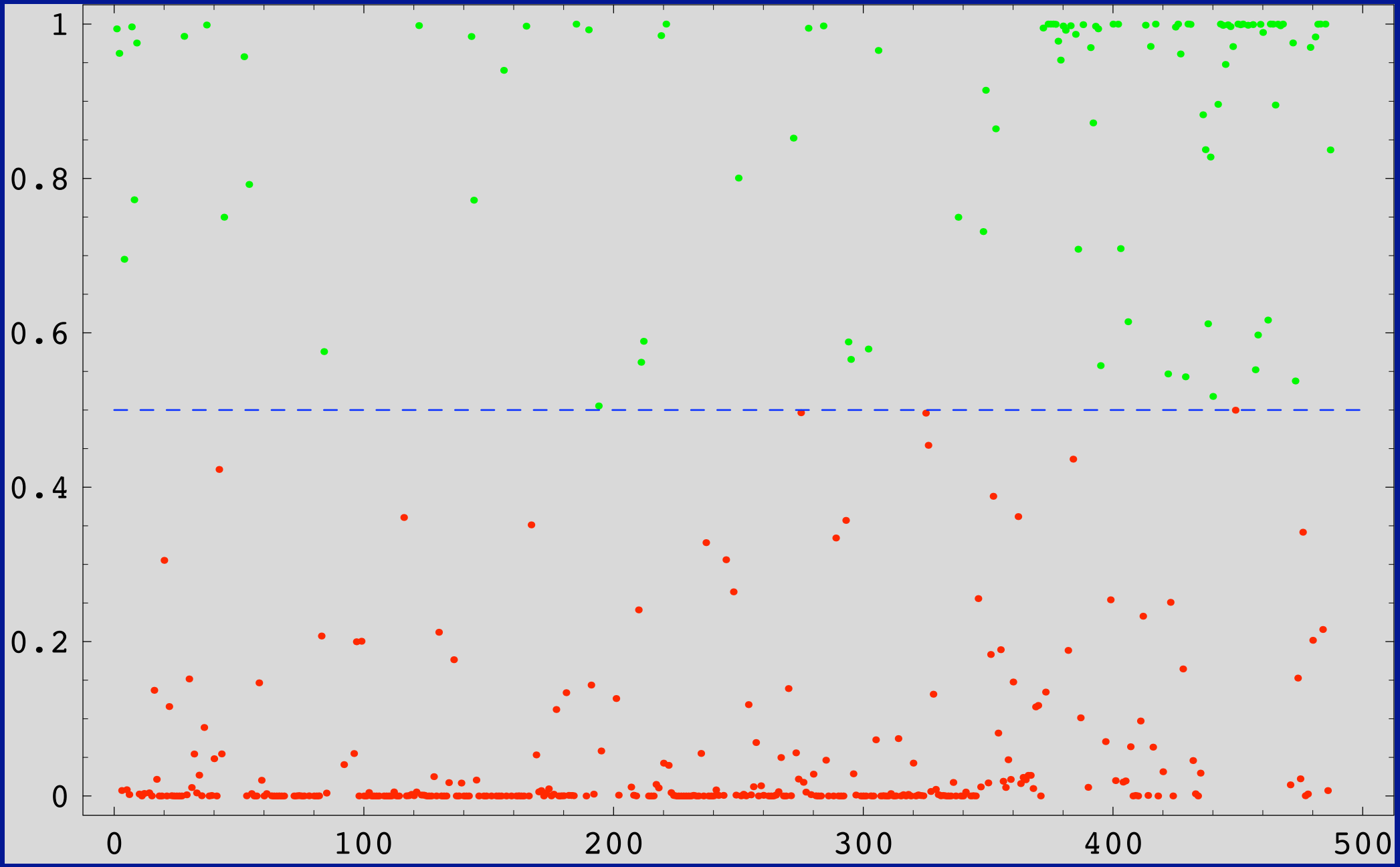


Figura 6a. Probabilidades de atribución de los capítulos de la Tabla 1 (longitud de palabra) al segundo autor.

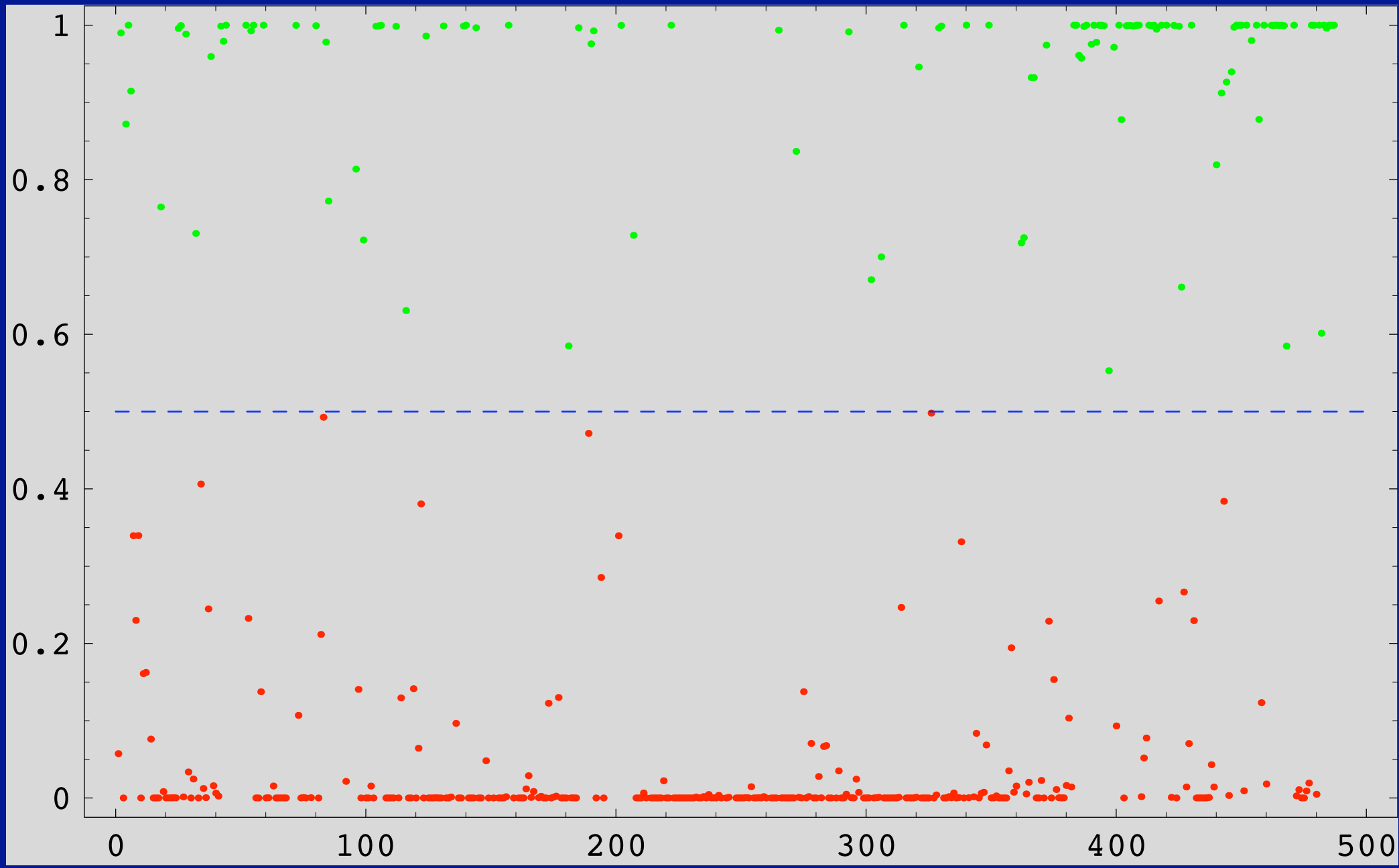


Figura 6b. Probabilidades de atribución de los capítulos de la Tabla 2 (restringida a las columnas ***e, de, la, que, no, l, com, és, molt, jo, si*** y ***dix***) al segundo autor.

CONCLUSIONES DEL ANÁLISIS BAYESIANO

1.- La longitud de palabra, el uso de palabras frecuentes y los índices de diversidad coinciden en detectar un cambio de autor entre los capítulos 371 y 382, en línea con lo dicho en el colofón.

CONCLUSIONES DEL ANÁLISIS BAYESIANO

1.- La longitud de palabra, el uso de palabras frecuentes y los índices de diversidad coinciden en detectar un cambio de autor entre los capítulos 371 y 382, en línea con lo dicho en el colofón.

2.- Hay evidencia estadística muy fuerte a favor de la doble autoría, aunque no descartamos otras explicaciones alternativas.

CONCLUSIONES DEL ANÁLISIS BAYESIANO

1.- La longitud de palabra, el uso de palabras frecuentes y los índices de diversidad coinciden en detectar un cambio de autor entre los capítulos 371 y 382, en línea con lo dicho en el colofón.

2.- Hay evidencia estadística muy fuerte a favor de la doble autoría, aunque no descartamos otras explicaciones alternativas.

3.- El análisis de conglomerados bayesiano muestra que:

a) Hay consistencia entre los resultados de cambio en la autoría detectados por el análisis de un cambio de estilo entre los capítulos 371 y 382 y el análisis de conglomerados, en el sentido de que minimiza el número de capítulos mal clasificados por el punto de cambio.

b) Al parecer, hay intervenciones menores —retoques en algunos capítulos— de ambos autores en las partes respectivas atribuídas al otro autor. Serían los capítulos mal clasificados por el cambio de estilo.

CONCLUSIONES DEL ANÁLISIS BAYESIANO

1.- La longitud de palabra, el uso de palabras frecuentes y los índices de diversidad coinciden en detectar un cambio de autor entre los capítulos 371 y 382, en línea con lo dicho en el colofón.

2.- Hay evidencia estadística muy fuerte a favor de la doble autoría, aunque no descartamos otras explicaciones alternativas.

3.- El análisis de conglomerados bayesiano muestra que:

   a) Hay consistencia entre los resultados de cambio en la autoría detectados por el análisis de un cambio de estilo entre los capítulos 371 y 382 y el análisis de conglomerados, en el sentido de que minimiza el número de capítulos mal clasificados por el punto de cambio.

   b) Al parecer, hay intervenciones menores —retoques en algunos capítulos— de ambos autores en las partes respectivas atribuídas al otro autor. Serían los capítulos mal clasificados por el cambio de estilo.

4.- Los gráficos de cajas son una herramienta muy útil a la hora de entender qué características cambian en la frontera de estilo y detectar éstas.

CONCLUSIONES DEL ANÁLISIS BAYESIANO

1.- La longitud de palabra, el uso de palabras frecuentes y los índices de diversidad coinciden en detectar un cambio de autor entre los capítulos 371 y 382, en línea con lo dicho en el colofón.

2.- Hay evidencia estadística muy fuerte a favor de la doble autoría, aunque no descartamos otras explicaciones alternativas.

3.- El análisis de conglomerados bayesiano muestra que:

   a) Hay consistencia entre los resultados de cambio en la autoría detectados por el análisis de un cambio de estilo entre los capítulos 371 y 382 y el análisis de conglomerados, en el sentido de que minimiza el número de capítulos mal clasificados por el punto de cambio.

   b) Al parecer, hay intervenciones menores —retoques en algunos capítulos— de ambos autores en las partes respectivas atribuídas al otro autor. Serían los capítulos mal clasificados por el cambio de estilo.

4.- Los gráficos de cajas son una herramienta muy útil a la hora de entender qué características cambian en la frontera de estilo y detectar éstas.

5.- El análisis Bayesiano del problema es más informativo y más fácil.

LÍNEAS DE INVESTIGACIÓN FUTURAS

Considerando los mismos criterios estilísticos

P. ¿Por qué hay ciertas, aunque pequeñas, diferencias entre los resultados del análisis de las Tablas 1 y 2, correspondientes a los dos criterios estilísticos utilizados?

LÍNEAS DE INVESTIGACIÓN FUTURAS

Considerando los mismos criterios estilísticos

P. ¿Por qué hay ciertas, aunque pequeñas, diferencias entre los resultados del análisis de las Tablas 1 y 2, correspondientes a los dos criterios estilísticos utilizados?

R.1 Los datos de ambas tablas presentan sobredispersión.

LÍNEAS DE INVESTIGACIÓN FUTURAS

Considerando los mismos criterios estilísticos

P. ¿Por qué hay ciertas, aunque pequeñas, diferencias entre los resultados del análisis de las Tablas 1 y 2, correspondientes a los dos criterios estilísticos utilizados?

R.1 Los datos de ambas tablas presentan sobredispersión.

R.2 El modelo multinomial es muy restrictivo —tiene pocos parámetros—.

LÍNEAS DE INVESTIGACIÓN FUTURAS

Considerando los mismos criterios estilísticos

P. ¿Por qué hay ciertas, aunque pequeñas, diferencias entre los resultados del análisis de las Tablas 1 y 2, correspondientes a los dos criterios estilísticos utilizados?

R.1 Los datos de ambas tablas presentan sobredispersión.

R.2 El modelo multinomial es muy restrictivo —tiene pocos parámetros—.

R.3 Los parámetros de las distribuciones multinomiales se suponen iguales dentro de cada uno de los dos subgrupos, tanto en el modelo de cambio de estilo como en el modelo de mixtura. Una generalización consistiría en considerarlos intercambiables dentro de cada uno de los dos subgrupos, en vez de iguales.

LÍNEAS DE INVESTIGACIÓN FUTURAS

Considerando los mismos criterios estilísticos

P. ¿Por qué hay ciertas, aunque pequeñas, diferencias entre los resultados del análisis de las Tablas 1 y 2, correspondientes a los dos criterios estilísticos utilizados?

R.1 Los datos de ambas tablas presentan sobredispersión.

R.2 El modelo multinomial es muy restrictivo —tiene pocos parámetros—.

R.3 Los parámetros de las distribuciones multinomiales se suponen iguales dentro de cada uno de los dos subgrupos, tanto en el modelo de cambio de estilo como en el modelo de mixtura. Una generalización consistiría en considerarlos intercambiables dentro de cada uno de los dos subgrupos, en vez de iguales.

Una posible solución

Un modelo jerárquico bayesiano basado en la distribución multinomial-Dirichlet proporcionaría probablemente unos resultados aún más satisfactorios, al tener en cuenta los tres puntos anteriores.

Considerando otros criterios estilísticos

- La riqueza y diversidad del vocabulario, que no se pueden tratar con modelos multinomiales.

Considerando otros criterios estilísticos

- La riqueza y diversidad del vocabulario, que no se pueden tratar con modelos multinomiales.
- La frecuencia de cada una de las letras del alfabeto del correspondiente idioma. Este último criterio se ha empleado con éxito en problemas de discriminación de textos de la literatura inglesa de la época isabelina (Ledger and Merriam, 1994), y de la literatura clásica de la antigua Grecia (Belcastro y Eisinberg, 2002).

Considerando otros criterios estilísticos

- La riqueza y diversidad del vocabulario, que no se pueden tratar con modelos multinomiales.
- La frecuencia de cada una de las letras del alfabeto del correspondiente idioma. Este último criterio se ha empleado con éxito en problemas de discriminación de textos de la literatura inglesa de la época isabelina (Ledger and Merriam, 1994), y de la literatura clásica de la antigua Grecia (Belcastro y Eisinberg, 2002).

FIN

. . . Por tomar muchos juntos, se le cayó uno a los pies del Barbero, que le tomó gana de ver de quién era, y vió que decía: Historia del famoso caballero Tirante el Blanco.

—¡Válame Dios!— dijo el Cura, dando una gran voz. —¡Que aquí esté Tirante el Blanco! Dádmele acá, compadre; que hago cuenta que he hallado en él un tesoro de contento y una mina de pasatiempos. Aquí está don Quirieleisón de Montalbán, valeroso caballero, y su hermano Tomás de Montalbán, y el caballero Fonseca, con la batalla que el valiente de Tirante hizo con el alano, y las agudezas de la doncella Placerdemivida, con los amores y embustes de la viuda Reposada, y la señora Emperatriz, enamorada de Hipólito su escudero. Dígoos verdad, señor compadre, que, por su estilo, es éste el mejor libro del mundo: aquí comen los caballeros, y duermen, y mueren en sus camas, y hacen testamento antes de su muerte, con otras cosas de que todos los demás libros deste género carecen. Con todo eso, os digo que merecía el que lo compuso, pues no hizo tantas necedades de industria, que le echaran a galeras por todos los días de su vida. Llevadle a casa y leedle, y veréis que es verdad cuanto dél os he dicho.

Don Quijote de la Mancha
Primera Parte, Capítulo VI
Del donoso y grande escrutinio que el Cura y el Barbero hicieron en la librería de nuestro ingenioso Hidalgo