UNIVERSIDAD NACIONAL SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE MATEMATICA

TESIS:

AUTENTIFICACIÓN DE LA OBRA TEATRAL "LA CONQUISTA DE JERUSALÉN" ATRIBUIDO A MIGUEL DE CERVANTES SAAVEDRA MEDIANTE EL ANÁLISIS DE CORRESPONDENCIA

Presentado Por:

Mijael Andre Lima Rodriguez

Julieta Flores Velarde

TESIS PARA OPTAR AL TÍTULO

PROFESIONAL DE

LICENCIADO EN MATEMÁTICA

MENCIÓN ESTADISTICA

ASESOR:

MGT. GUILLERMO PAUCAR CARLOS

CUSCO-PERU

2019

Presentación:

Señor decano de la facultad de ciencias, señores docentes y lectores de la investigación propuesta.

Tengo el agrado de dirigirme a Uds. con el objetivo de dar a conocer la investigación realizada con el título **AUTENTIFICACIÓN DE LA OBRA TEATRAL "LA CONQUISTA DE JERUSALÉN" ATRIBUIDO A MIGUEL DE CERVANTES SAAVEDRA MEDIANTE EL ANÁLISIS DE CORRESPONDENCIA** Realizada con el fin de optar al título profesional de licenciado en matemática mención estadística.

En relación con esto, cuento con el conocimiento impartido en las aulas de la universidad y el tiempo de investigación dado para poder realizar esta investigación enfocándome en el área de minería de datos en su especialidad de minería de texto desde un enfoque de estadística multivariante y procesos estocásticos.

Considerando que la investigación realizada sea de aporte para investigaciones posteriores en las ramas del procesamiento del lenguaje natural desde un enfoque estadístico.

Agradezco la gentil atención al momento de dar lectura y quedo a la espera de sus cometarios y recomendaciones.

Dedicado a:

A nuestros padres, por su interés, para hacer nuestros sueños realidad

A nuestros maestros, porque sin su conocimiento y apoyo, no hubiésemos logrado las metas alcanzadas.

A los amigos, por sus palabras de ánimo para hacer un mejor trabajo.

Los Tesistas

TABLA DE CONTENIDO.

INDICE DE TABLAS

INDICE DE ILUSTRACIONES.

Resumen.

El presente trabajo de tesis tiene como finalidad dar a conocer un enfoque estadístico y estudio científico acerca de la autoría en la obra “La Conquista de Jerusalén por Godofre Bullon” el cual se encontraba en la biblioteca personal de un noble español posteriormente fue llevado a la biblioteca central de España y es considerada un tesoro de la humanidad, en el que se le atribuyó a Miguel de Cervantes Saavedra y que desde la época de su descubrimiento el trabajo es netamente minería de texto o data mining, otro objetivo es dar como inicio a otros de nuestros compañeros que sigan el estudio ya que el nuestro es solo el inicio de varios temas que siempre se ponen en duda al presentar una obra de autoría propia.

Y en cuanto a la metodología usada se plantea que es un estudio descriptivo correlacionar haciendo uso de las técnicas estadísticas multivariadas de análisis de correspondencia e inferencia sobre matrices de transición para así obtener una opinión del tema en cuestión. En los años 90 el catedrático italiano Stefano Arata[1] abrió una discusión con el descubrimiento y el estudio riguroso acerca de su autoría de la obra en cuestión y concluyó que ciertamente era del escritor antes mencionado. Desde entonces se tuvo discrepancias ya que otros estudiosos filólogos no coincidían con lo aclarado por Arata.

Después de los análisis categóricos pertinentes se llegó a la conclusión a la conclusión de que Arata tenía razón en atribuir la autoría al padre de la lengua española Miguel de Cervantes Zavedra

PALABRAS CLAVE: análisis de correspondencia, análisis multivariado, matriz de transición, cadenas de markov, variable aleatoria, cervantes, lematizacion.

[1] Stefano Arata: autor de la crítica literaria “la conquista de Jerusalén, cervantes y la generación teatral de 1580” en el año 1992, primer estudio acerca de la autoría encontrada

Abstract:

The purpose of this thesis is to present a statistical approach and scientific study about the authorship of the work "The Conquest of Jerusalem by Godofre Bullon" which was in the personal library of a Spanish nobleman was subsequently taken to the central library of Spain and is considered a treasure of humanity, in which it was attributed to Miguel de Cervantes Saavedra and since the time of its discovery the work is purely text mining or data mining, another objective is to start other of our colleagues who follow the study because ours is just the beginning of several topics that are always put in doubt when presenting a work of own authorship.

And as for the methodology used, it is suggested that it is a descriptive study to correlate using the multivariate statistical techniques of correspondence analysis and inference about transition matrices in order to obtain an opinion of the subject in question. In the 90s the Italian professor Stefano Arata opened a discussion with the discovery and rigorous study about his authorship of the work in question and concluded that it was certainly the aforementioned writer. Since then there were discrepancies since other philological scholars did not agree with what was clarified by Arata.

After the relevant categorical analyzes, it was concluded that Arata was right to attribute authorship to the father of the Spanish language Miguel de Cervantes Zavedra

KEYWORDS: correspondence analysis, multivariate analysis, transition matrix, markov chains, random variable, Cervantes, lematizacion.

Introducción:

Desde su descubrimiento de la obra teatral," LA CONQUISTA DE JERUSALÉN" ha sido ligada al nombre Miguel de Cervantes Saavedra considerado la máxima figura de la literatura española, por haber escrito el universalmente conocido "INGENIOSO HIDALGO DON QUIJOTE DE LA MANCHA".

Dadas sus penurias económicas, el teatro fue la gran vocación de Cervantes, quien declaró haber escrito veinte o treinta comedias de las cuales se conservan los títulos de 16 y los textos de 11, sin contar 8 entremeses y algunos otros atribuidos.

En 1992 el hispanista italiano Stefano Arata publicó el texto de un manuscrito de la obra teatral la "CONQUISTA DE JERUSALÉN POR GODEFRE DE BULLÓN" en su estudio preliminar Arata pretende haber encontrado la "JERUSALÉN" perdida de Cervantes. Numerosos especialistas han dado cuenta de las semejanzas que hay entre esta comedia anónima y las obras de teatro de Cervantes, haciendo extensibles los parecidos a toda la producción cervantina.

En el presente estudio se propone un acercamiento a esta relación desde una perspectiva estadística, centrándome en aspectos asociativos por medio de las técnicas de minería de texto mediante el análisis de correspondencia y técnicas de la estadística multivarian

I. PLANTEAMIENTO DEL PROBLEMA.

1.1. Situación problemática.

Desde el descubrimiento de la obra teatral, *"LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON"* en la biblioteca del palacio real de Madrid, obra presuntamente escrita en los años 80 del siglo XVI. La obra fue hallada por el hispanista italiano Stefano Arata de la universidad de Roma en 1990, obra que carecía de autor y fecha. Por lo que especialistas lo han ligado a *"Miguel de Cervantes Saavedra". (Alcalá de Henares, España, 1547 - Madrid, 1616). El cual se* considera el máximo exponente de la literatura española, por su escrito reconocido universalmente "INGENIOSO HIDALGO DON QUIJOTE DE LA MANCHA".

Cervantes, hijo del médico barbero Rodrigo de Cervantes, fue preparado desde su niñez para ser médico y seguir la tradición familiar ya que su abuelo también era médico, pero el padecía de sordera y dada que su familia se mudaba de un lugar a otro no pudo concluir sus estudios, siendo muy joven participó en la batalla de Lepanto donde fue secuestrado y esclavizado en Argel (isla capital y la mayor ciudad de Argelia. África del norte) después regresó a su ciudad natal, fue época donde los especialistas ubican cronológicamente la obra "LA JERUSALÉM", al no tener el éxito que esperaba en 1587, Cervantes decidió abandonar el teatro asumiendo el cargo de un modesto comisario de abastos en Andalucía España, años más tarde retornaría a la literatura y el teatro.

"La falta de economía estuvo presente en su familia puesto su padre llegó hasta a la cárcel por no tener para pagar sus deudas, la educación de cervantes fue extensa pero improvisada, en sus obras siempre aparecen parodias estudiantiles de la época, con eso se puede afirmar que asistía a varias universidades de Alcalá de Henares Madrid, el teatro fue la gran vocación de Cervantes, quien declaró haber escrito veinte o treinta comedias de las cuales se conservan los títulos de 16 y los textos de 11, sin contar 8 entremeses y algunos otras obras atribuidas." (clubdelecturavalladolid)

En “LA JERUSALÉM” obra inspirada en “LA JERUSALÉN LIBERADA” de Torcuato Tasso y escrita entre 1581 y 1585, Arata le dio hincapié a que *“los indicios que apuntan a una posible paternidad cervantina son muchos, aunque ninguno tiene el valor de prueba definitiva”(Stefano Arata);* además de la cita en la “ADJUNTA AL PARNASO” entrevista autobiográfica donde Cervantes afirma haber escrito una presunta Jerusalén, Stefano Arata encuentra concordancias métricas y de técnica teatral con la producción cervantina.

Juan Cerezo, de la Universidad Autónoma de Madrid, se muestra con más cautela en la atribución, cita entre los partidarios de la atribución cervantina a los especialistas Héctor Brioso, José Montero, Alfredo Rodríguez López-Vázquez, Alfredo Baras Escolá, Aaron Kahn y Moisés R. Castillo. Y sitúa a Jean Canavaggio y a Daniel Eisenberg entre los que están en contra de la autoría cervantina. Resultan interesantes los aspectos y características que usan como argumento para la autentificación de la obra, todos netamente analíticos lingüísticos.

Sin embargo, así como existen bastantes estudiosos que apoyan en las autorías de obras perdidas hay también quienes se oponen a estas; pues son más como una lucha de atribuciones de obras que en su momento se dieron perdidas dado a que Cervantes mencionó varias obras, pero los estudiosos no tomaron en cuenta que pudieron ser simples pastiches (Plagio que consiste en tomar determinados elementos característicos de la obra de un artista o de las de varios y combinarlos de forma que parezcan una creación original.) (wordreference, 2018).

Como prueba los especialistas citan que. En la “ADJUNTA AL PARNASO” donde pregunta Pancracio de Roncesvalles a Cervantes: *“¿Ha compuesto alguna comedia?” Y Cervantes responde “Sí, muchas; y, a no ser mías, me parecieran dignas de alabanza, como lo fueron “LOS TRATOS DE ARGEL”, “LA NUMANCIA”, “LA GRAN TURQUESCA”, “LA BATALLA NAVAL”, “LA JERUSALEM”, “LA AMARANTA O LA DEL MAYO”, “EL BOSQUE AMOROSO”, “LA ÚNICA Y LA BIZARRA ARSINDA”, y otras muchas de que no me acuerdo”, los que cronológicamente podrían coincidir con la supuesta Jerusalén perdida serian “EL TRATO DE ARGEL” y “LA NUMANCIA”* (Savedra, 1600).

Lo cual plantea “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”, es la perdida “LA JERUSALEM”, con “m” que menciona Cervantes en su entrevista. No se está de acuerdo porque los títulos no coinciden, caso que podría haber ocurrido en la época dado que la autoría de la obra no estaba dada por el al autor a su vez pertenecía a los actores que protagonizaban la obra. También se argumentan que *“LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”* es protagonizada por un francés, Godofre de Bullón, y Cervantes no elogiaría en su obra a ningún francés. La obra no contiene ni un solo soldado español ni una referencia a España, esa situación podría ser justificable ya que en esas épocas Cervantes estaba más ligado al entorno italiano por su secuestro y esclavismo por algunos años en Argel, pero los críticos no consideran compatible con el nacionalismo descrito en “LA NUMANCIA” y “EL TRATO DE ARGEL”, obras supuestamente contemporáneas a “LA JERUSALEM”. Los críticos consideran que adaptar obras teatrales de otros escritores, sobre todo extranjeros, es algo que Cervantes no conceptuaría. Bajo estas premisas, los críticos podrían argumentar que la obra la “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” no es la perdida “LA JERUSALEM”.

Hasta ahora nadie presentó un análisis diferente acerca de las cuestiones de autentificación con otros métodos que no sean los convencionales recursos lingüísticos. Como recursos cualitativos de análisis estadístico, a lo cual Stefano Arata en sus épocas de docencia universitaria, como todo un cervantista, alentaba en sus alumnos que estudiaran el tema con enfoque más analítico. Y hoy con los avances de la lingüística computacional, que se trata de una disciplina muy joven, tal que sus orígenes datan de la segunda guerra mundial en donde los Estados Unidos en su intento de descifrar códigos enemigos a través de operaciones matemáticas donde los primeros resultados que se dieron mostraron un exceso de simplificación. En los años ochenta la lingüística computacional necesitaba una gran cantidad de esfuerzo humano, dado que el conteo de palabras era manual, pero con los avances tecnológicos han permitido que esta área progrese de manera rápida en la última década siendo hoy un área multidisciplinar basada en la recuperación de información, minería de datos, y estadísticas.

En la actualidad con computadores en los hogares del mundo se desarrolló buscadores de internet en base a redes neuronales como por ejemplo el usado por Google, siendo una utilidad de la lingüística computacional centrándose en palabras clave, identificando la información en un documento en contraste con gran base de datos, entre otras aplicaciones encontramos resumidor automático y traductores de texto. Todas estas aplicaciones siendo posible mediante la estadística, redes neuronales y data mining en la rama de minería de texto. Donde la tecnología actual busca la interface del lenguaje natural, interpretación de la voz humana, y síntesis de voz.

El estudio planteado busca similitudes en la obra teatral “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” y la producción literaria de Cervantes se dará desde un punto de vista cuantitativo mediante técnicas estadísticas de la minería de texto o data mining que permitan un estudio detallado y sistemático de un texto mediante procesos estadísticos correlacionales y de prueba de hipótesis.

Las técnicas estadísticas correlaciónales para variables categóricas como vienen a ser las tablas de contingencia, el análisis de correspondencia, procesos estocásticos en cadenas markovianas, sumados al aporte de la minería de texto plantean una herramienta para establecer relaciones entre fragmentos textuales y técnica multivariante de comparación de matrices de transición estocástica, la cual emplearemos como herramienta para probar si existe varianza entre las matrices lexicales. Permitiendo dar en el presente estudio de tesis un acercamiento a la asociación existente entre la obra encontrada “LA CONQUISTA DE JERUSALÉN POR GOFRE BULLON” y la obra perdida “LA JERUSALÉM” de cervantes. Desde una perspectiva estadística dando una opción a la autentificación de esta obra la cual se le atribuye a Cervantes.

1.2. Formulación del problema.

"En su obra de crítica literaria para el CRITICON 1992 (La Conquista de Jerusalén, Cervantes y la generación teatral de 1580) Stefano de Arata pretende haber encontrado la obra pérdida de cervantes." (SOLER) y desde entonces se ha publicado virtualmente como obra atribuida, en 2009 apareció una edición crítica impreso por "CATEDRA LETRAS HISTORICAS" y 2010 donde Harold kahm. Publicó una teoría de atribución que demuestra que, de todos los candidatos de autoría de esta comedia, Cervantes es el más probable, existiendo también opositores a esta teoría. (JesToryAs, 2016)

En todas las anteriores investigaciones realizadas se platean desde un punto de vista de crítica literaria existiendo pocos acercamientos al análisis estadístico a la obra por medio de técnicas de minería de texto mediante el análisis de correspondencia e inferencia sobre la matriz de transición estocástica. Lo cual permitiría encontrar semejanzas entre la obra atribuida a Cervantes "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON" con las obras contemporáneas reconocidas del autor y de esta manera plantear el problema en cuestión desde un punto de vista cuantitativo y de estudio estadístico

1.2.1. Problema general.

¿Qué semejanzas existe, entre las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", ¿"LA NUMANCIA" y "EL TRATO DE ARGEL" obtenidas mediante el análisis de correspondencia?

1.2.2. Problemas específicos.

a) ¿Qué semejanza de léxico existe entre las obras "¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL"?
b) ¿Existe semejanza entre los personajes en las obras "¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL?
c) ¿Existe similitud en el tamaño de palabras en las obras "¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL"?
d) ¿Existe semejanza entre las matrices de transición precede y antecede de las obras" ¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL?

1.3. Justificación de la investigación.

La presente investigación encuentra su justificación en el estudio realizado por Stefano Arata en 1992, en donde pretende haber encontrado una obra perdida de Cervantes argumentando desde un punto de vista literario, en la presente investigación se realizó el análisis de estas obras en busca de semejanzas haciendo uso de herramientas estadísticas.

El estudio presenta gran importancia tanto para la comunidad matemática-estadística como para la comunidad filológica, ya que el trabajo es un paso destacado para la relación de dos ciencias. Con el auge de las técnicas en data mining en estos últimos años se dispone de una gama de herramientas que hace posible y factible el análisis comparativo de los textos ya sea netamente literario o cualesquiera que tenga una característica lexicológica que garantice ser escrita por un mismo autor.

El estudio será trascendente ya que dará otra alternativa a la solución del problema planteado por autoría de la obra, puesto que las opiniones vertidas por los estudiosos filólogos no son del todo absolutas y que también hay opositores que siempre tendrán dudas acerca de los detalles, y ahora con este análisis se pretende adentrarse más en un campo casi inexplorado por la matemática.

Es de interés de la sociedad de habla hispana que tiene a Cervantes como padre la lengua española, con el trabajo de investigación buscamos el incentivo a otros estudiosos ya sean matemáticos, estadísticos o literatos a adentrarse en autentificaciones de otras obras que están en debate su presente autoría, así como para el apoyo de varias otras investigaciones que buscan la verdad acerca de cuestiones por derechos de autor.

Se da a conocer también el hecho de que las matemáticas al caracterizarse de ciencias exactas, pueden dar un aspecto fidedigno al estilo literario propio de un autor y llegar a una decisión más confiable acerca de estos aspectos en los que no hay una opción de autoría, y así dar un avance asertivo a las opiniones vertidas de los estudiosos ya que el análisis estadístico que se tendrá por autentificación será totalmente imparcial.

1.4. Limitaciones de la investigación.

El estudio se ve limitado por la comparación de dos obras frente a la que se le atribuye puesto que son las únicas que cumplen con los aspectos cronológicos. Se ve limitado por obviar la comparación con obras contemporáneas de otros autores esto debido a que la presente investigación toma como punto de inicio el estudio de Stefano Arata, en el cual compara estas tres obras.

Por otra parte, se ve también que el problema de la diferencia en años transcurridos y el que los autores y protagonistas de estos acontecimientos ya murieron, no se tienen rastros de opiniones o testigos claves solo estamos con los vestigios de unas palabras plasmadas a lo largo del tiempo que en sí, nos pueden dar respuestas de cuál fue su origen en el que trataremos de revelar con la ayuda del análisis multivariado y técnicas estadísticas para variables categóricas las cuales nos darán un indicio de la autoría no excluido de error estadístico y contextual en el que nos encontramos.

Encontramos limitaciones en los análisis planteados en el estudio debido a que, el vocabulario de la época y en especial el de Cervantes, era más amplio ya que utilizaba muchos equivalentes semánticos, jergas, lenguaje coloquial, vocablos de idiomas distintos por lo cual Cervantes le dio más opciones de uso a la lengua española. Esta amplia riqueza lingüística cervantina, genera en situación de estudio muchas categorías, lo que obligará a realizar un proceso de simplificado de texto.

1.5. Objetivos de la investigación.

a. Objetivo general.

Analizar asociación entre las obras “La Conquista De Jerusalén Por Godofre Bullon”, “La Numancia” Y “El Trato De Argel” obtenidas mediante el análisis de correspondencia.

b. Objetivos específicos.

- Establecer similitud en el tamaño de palabras en las obras “La Conquista De Jerusalén Por Godofre Bullon”, “La Numancia” Y “El Trato De Argel”.
- Determinar semejanza de personajes en las obras “La Conquista De Jerusalén Por Godofre Bullon”, “La Numancia” Y “El Trato De Argel”.
- Describir la semejanza de léxico existe entre las obras “La Conquista De Jerusalén Por Godofre Bullon”, “La Numancia” Y “El Trato De Argel”.
- Evaluar la semejanza entre las matrices de transición precede y antecede de las obras” La Conquista De Jerusalén Por Godofre Bullon”, “La Numancia” Y “El Trato De Argel”

II. MARCO TEÓRICO CONCEPTUAL.

2.1. Bases teóricas en estadística.

2.1.1. Variable aleatoria.

En un experimento aleatorio, no se puede predecir con exactitud cuáles serán sus resultados por causalidades presentes, no obstante, se puede describir cuales serán todos los resultados posibles y con qué probabilidad ocurrirá cada uno de ellos, este fenómeno es más importante que el resultado completo del experimento. *"Tales funciones cuyos valores dependen de los posibles resultados de un experimento aleatorio se llaman variables aleatorias".* (Cabrera)

Para el estudio de las probabilidades se definirá previamente:

Definición 2.1 (σ-Algebra) Sea una familia A de subconjuntos Ω, es decir $A \subset P(\Omega) = \{A: A \subset \Omega\}$. Se dice que A es una σ-álgebra sobre Ω si satisface las siguientes propiedades:

i) Ω ϵ $\mathcal{A}$

ii) Dado A ∈ $\mathcal{A}$ $se\ tiene\ A^c \in \mathcal{A}$

iii) Sea $A_2, \dots, A_n, \dots$ una sucesión de elementos de $\mathcal{A}$ entonces:

$$A = \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$$

Espacio muestral Ω: Sea **Ω** la totalidad de sucesos elementales para todo (i=1, 2…, n); es decir es el conjunto de todos los resultados posibles de un experimento aleatorio, denota por Ω.

A las particiones de **Ω** se les denomina suceso elemental y son subconjuntos de Ω

Los espacios muéstrales se clasifican de acuerdo con el número de sucesos elementales que contienen.

Definición 2.2.- Sea un espacio muestral Ω, es discreto si es un conjunto finito o infinito numerable.

Espacios de probabilidad.

Definición 2.3.- Se define a la terna $(\Omega, \mathcal{A}, P)$ como espacio de probabilidad, donde Ω es un conjunto, $\mathcal{A}$ es una σ-algebra sobre Ω, y P: $\mathcal{A} \to [0,1] \subset \mathbb{R}$ es una función que satisface:

i) $P(\Omega) = 1$

ii) (σ-aditividad). Si $(A_n)_{n\geq 1}$ es una sucesión de elementos de $\mathcal{A}$ disjuntos dos a dos $(A_i \cap A_j = \Phi,\ si\ i \neq j)$, entonces

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

σ-álgebra de Borel sobre los reales.

Si tenemos un espacio de probabilidad cuyo espacio muestral es el conjunto de números reales $\mathbb{R}$ parece natural que la σ-álgebra contenga los conjuntos de la forma (-∞, x]. Esto permitirá calcular la probabilidad de que el resultado del experimento aleatorio correspondiente sea menor o igual que x.

Definición 2.4.- Sea el σ-álgebra de Borel sobre $\mathbb{R}$ denotado por **B**, es la σ-álgebra sobre $\mathbb{R}$ generada por los conjuntos de la forma A_x = (-∞, x], para todo $x \in \mathbb{R}$. Un conjunto $B \in \mathbf{B}$ se denomina boreliano.

Probabilidad condicional.

Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y consideremos dos eventos $A, B \in \mathcal{A}$ y supongamos que $P(B) \neq 0$.

Queremos estudiar cómo cambia la probabilidad de ocurrencia de A cuando se conoce que otro evento B que ya ha tenido lugar. En este caso habrá que redefinir el espacio muestral considerando solamente los elementos de B como posibles resultados. Se tratará de determinar cuál debe ser la probabilidad de un evento A condicional a que se conoce que el evento B ha ocurrido utilizando interpretación heurística de probabilidad como límite de la frecuencia con la cual un evento ocurre, para esto supongamos que se han hecho n repeticiones independientes del experimento y denotemos con:

n_B: El número de veces que ocurre el resultado B

$n_{A\cap B}$: El número de veces en el que ocurre el resultado A∩B

Heurísticamente la probabilidad condicional de A dado B, será el límite de la frecuencia con la cual ocurre en los experimentos donde B ocurre, es decir el límite de:

$$\frac{n_{A\cap B}}{n_B}$$

Luego la "probabilidad de que ocurra A condicional B" será:

$$\lim_{n\to\infty}\frac{n_{A\cap B}}{n_B}=\lim_{n\to\infty}\frac{\frac{n_{A\cap B}}{n}}{\frac{n_B}{n}}=\frac{\lim_{n\to\infty}\frac{n_{A\cap B}}{n}}{\lim_{n\to\infty}\frac{n_B}{n}}=\frac{P(A\cap B)}{P(B)}.$$

Esto justifica la siguiente definición.

Definición 2.5.- Sea $(\Omega,\mathcal{A},P)$ es un espacio de probabilidad A, B $\in\mathcal{A}$ tal que $P(B)>0$. Se define la probabilidad condicional de A dado por B

$$P(A/B)=\frac{P(A\cap B)}{P(B)}$$

Independencia de eventos

Definición 2.6.- Sea $(\Omega,\mathcal{A},P)$ un espacio de probabilidad y consideremos:

A, B $\in\mathcal{A}$. Se dice que A y B son independientes si:

$$P(A\cap B)=P(A)P(B).$$

Definición 2.7.-Se define la variable aleatoria X a la función que asocia a cada elemento del espacio muestral un número real, al conjunto de números reales generados por esta función y se denomina rango, recorrido o soporte denotado por $R_x$es decir:

$$\mathrm{X}:\Omega\to R_x\subset\mathbb{R}$$

$$\omega\to\mathrm{X}(\omega)=x_i\in R_x\subset\mathbb{R}$$

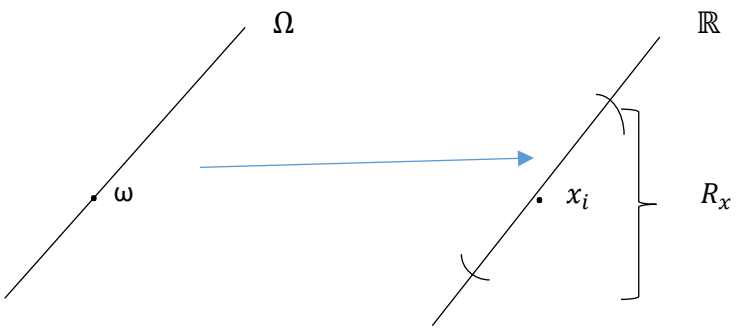


Ilustración 1:Función de la variable aleatoria

Definición 2.8.- Se define la variable aleatoria bidimensional (X, Y) como la función vectorial que asocia a cada elemento del espacio muestral ω un vector (x_i, yi)

$$(\mathrm{X},\mathrm{Y}):\Omega \to R_x \subset \mathbb{R}$$

$$\omega \to \big(\mathrm{X}(\omega),\mathrm{Y}(\omega)\big) = (x_i, y_i) \in R_x \subset \mathbb{R}$$

Tipos de variables aleatorias:

Variable aleatoria discreta: se llama variable aleatoria discreta a X cuando R_x toma un conjunto numerable de valores que pueden ser finito o infinito.

Variable aleatoria continua: Se llama variable aleatoria continua a X cuando R_x toma valores en un conjunto no numerable, el más usual es en un intervalo en $\mathbb{R}$.

Para la presente tesis se hará énfasis en el uso de la variable aleatoria discreta numerable.

Definición 2.9.- (Función de Probabilidad) Sea una variable aleatoria discreta X con recorrido R_x=$\{x_1 x_2 \dots x_n\}$,entonces la probabilidad de la variable aleatoria X está definida por.

$$P: R_x \to [0,1] \subset \mathbb{R} \;/\; P(x) \in [0,1]$$

$$P[X = x_i] = P[\{\omega \epsilon \Omega \;/\; R_x = x_i\}] \to P(x_i) \in [0,1]$$

Tal que la función de probabilidad equiprobables satisfacer las propiedades:

1.- $0 \leq P[X = x_i] \leq 1 \; si \; x_i \in R_x$ 2.-$\sum_{x_i \epsilon R_x} P[X = x_i] = 1$

Proposición 2.1.- (cálculo de probabilidades) Sea un boreliano $B \in \mathbf{B}$, donde $B \subseteq R_x$ la probabilidad de la variable aleatoria discreta X toma valores en **B** y viene dada por:

$$P[X = B/X(\omega_i) \in B] = \sum_{x_i \in B} P[X = x_i]$$

Definición 2.10.- Sea una variable aleatoria discreta X con $R_x = \{x_1, \dots x_n \dots\}$ se define la función distribución de probabilidad FdD de X como:

$$F(x) = P[X \leq x] = \sum_{x_i \leq x} P[X = x_i] \ \forall\, x \in R_x$$

Observamos que 0≤ $F(x) \leq 1\ pues\ F(x) = P[X \leq x] = P[x \in (-\infty; x)]$.

Teorema 2.1.- La función distribución de probabilidad (FdD) de una variable aleatoria discreta verifica las siguientes propiedades.

$$lim_{x\sim-\infty} F(x) = 0 \qquad lim_{x\sim+\infty} F(x) = 1$$

F no es decreciente es decir si $x < y \rightarrow F(x) \leq F(y)$ F es continua por la derecha, es decir:

$$\forall\, F(x) \quad \lim_{h \to 0^+} F(x + h)$$

Además observamos que si X es una variable aleatoria discreta su FdP es una función escalonada, los puntos en que se producen las discontinuidades de salto son los valores que toma la variable aleatoria X y la medida de salto es igual a $P[X = x_i]\ \ es\ decir\ P(X = x_i) = F(x_i) - F(x_i^-)$

Corolario 2.1. - las siguientes probabilidades se pueden calcular de forma inmediata en términos de la FdD: $P[a < x \leq b] = \ P[x \leq b] - P[x \leq a] = F(b) - F(a)\ P[X > a] = 1 - P[X \leq a] = 1 - F(a)$

2.1.2. Proceso estocástico

Un proceso estocástico es un concepto matemático que sirve para tratar con magnitudes aleatorias que varían con el tiempo, o más exactamente para caracterizar una sucesión de variables aleatorias (estocásticas) que evolucionan en función de otra variable, generalmente el tiempo. Cada una de las variables aleatorias del proceso tiene su propia función de distribución de probabilidad y pueden o no, estar correlacionadas entre ellas. (Proceso-estocastico, s.f.)

Se considera un sistema que puede caracterizarse por estar en cualquiera de un conjunto de estados previamente especificado, se infiere que el sistema evoluciona o cambia de un estado a otro a lo largo del tiempo de acuerdo a una cierta ley del movimiento y sea X_t el estado del sistema al tiempo t si se considera que la forma en la que el sistema evoluciona no es determinista si no provocado por un mecanismo fortuito entonces se puede considerar que X_t es una variable aleatoria para cada valor del índice t, esta colección de variable aleatoria es la definición de proceso estocástico y sirve como modelo para representar la evolución aleatoria de una sistema a lo largo del tiempo (Rincón, 2011).

En general las variables aleatorias que conforman un proceso no son independientes entre sí, sino que están relacionadas unas con otras de alguna manera particular, para la definición de proceso estocástico se toma como base un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ se enuncia de la siguiente manera:

Al elemento t ∈ T (conjunto de subíndices) se le denomina parámetro puede ser discreto o continuo

Definición 2.11.- Sea proceso estocástico X_t, es una colección de variable aleatoria $\{X_t : t \epsilon T\}$ parametrizado en $t \epsilon T$ conjunto de estados.

T conjunto de estados considerado discreto cuando T= {0, 1, 2…} y se considera continuo cuando T=$[0; \infty]$.

En el primer caso se dice que el proceso a es de tiempo discreto en general este tipo de proceso y se denotara por: $X_n : n = \{0,1, \dots\}$

Clasificación según las características probabilísticas de variable aleatoria.

Las probabilidades de las variables aleatorias son importantes a la hora de identificar y clasificar un proceso estocástico se pueden clasificar los procesos en:

Procesos estacionarios, Procesos Markovianos, Procesos de incrementos independientes, Procesos de transición, Caminatas aleatorias. Para fines del análisis en el estudio de tesis sólo se usarán Procesos Markovianos ya que en el análisis de datos necesitamos saber la probabilidad sólo en un tiempo anterior.

2.1.3. Procesos Markovianos.

Un proceso Markoviano es una serie de eventos en la cual la probabilidad de que ocurra un evento depende del incremento anterior inmediato, en efecto las cadenas de este tipo tienen memoria recuerdan el ultimo evento y esto condiciona las probabilidades de los eventos futuros esta dependencia del evento anterior distingue a las cadenas de markov de las series de eventos independientes como tirar monedas al aire o lanzar un dado, en los negocios las cadenas de markov se han utilizado para patrones de compra, los deudores morosos, para planear las necesidades del personal y para analizar el reemplazo del equipo. (F., s.f.)

El análisis de markov llamado así en honor del matemático ruso Andréi Andréievich Markov (Riazán, 1856 - San Petersburgo, 1922) que desarrolló la moderna teoría de procesos estocásticos. Trabajó en la casi totalidad de los campos de la matemática. En el campo de la teoría de la probabilidad, profundizó en las consecuencias del teorema central del límite y en la ley de los grandes números. (vidas, s.f.)

En su honor, lleva su nombre un tipo muy especial de procesos estocásticos que desarrolló el método en 1907, que permite incrementar la probabilidad de que un evento se encuentre en un estado en particular en un momento dado, algo más importante aún es que permite encontrar el promedio a la larga o las probabilidades de estado estable para cada estado, con esta información se puede predecir el comportamiento del sistema a través del tiempo, la tarea más difícil es reconocer

cuando puede aplicarse, la característica más importante que hay que buscar en la memoria de un evento inmediato anterior.

Tras este trabajo, estudió las variables dependientes e introdujo el concepto de sucesos encadenados. Markov extendió los resultados clásicos de sucesos independientes a cierto tipo de sucesos encadenados, conocidos como sucesos Markovianos, que son aquellos cuyo estado en un instante de tiempo depende de uno o varios estados cronológicamente anteriores.

Definición 2.12.- Sea un proceso estocástico discreto $\{X_n : n\epsilon T\}$, se dice que es un proceso Markoviano si se cumple:

$$P(X_{n+1} = x_{n+1}/X_0 = x_0 \cdots X_n = x_n) = P(X_{n+1} = x_{n+1}/\ X_n = x_n) = P(X_n\ /X_{n-1})$$

Donde:

Los estados $X_0, X_1, \dots, X_{n-1}$ (corresponden al pasado), X_n (corresponden al presente), y X_{n+1} (corresponde al futuro)

La probabilidad del evento futuro$(X_n = x_n)$ solo depende del evento $X_{n-1} = x_{n-1}$ mientras que la información correspondiente al evento pasado $(X_0 = x_0 \dots X_{n-2} = x_{n-2})$ es irrelevante.

Por su forma los procesos Markovianos tienen una forma de cadena donde la variable aleatoria X_n va cambiando con el paso del tiempo. $X_n = j$. Dependiendo del estado inmediatamente anterior del sistema $X_{n-1} = i$ (Diazaraque, Cadenas de Markov)

Probabilidades de transición.

Una cadena de markov es un proceso en tiempo discreto en el que una variable aleatoria X_n va cambiando con el paso del tiempo y satisfacen que la probabilidad de X_n solo depende del estado inmediato anterior X_{n-1}

Definición 2.13.- Sea una cadena de un proceso Markoviano de *m* posibles estados $(X_1, X_2, \dots, X_n, \dots X_m)$ se define la probabilidad de transición como que dado variable aleatoria X_{n-1}tome el valor de *i* en el siguiente tiempo X_n tome el valor de *j*.

$$P_{ij} = P(X_n = j / X_{n-1} = i)$$

La probabilidad de transición satisface las condiciones:

$P_{ij} > 0$

$\sum_{j=1}^{m} P_{ij} = 1$

Matriz de transición

Definición 2.14.-Se define como matriz de transición T de tamaño *mxm* a la matriz formada por todas las probabilidades de transición.

$$M_T = T = [P_{ij}] = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1m} \\ P_{21} & P_{22} & \dots & P_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{m1} & P_{m2} & \dots & P_{mm} \end{bmatrix}$$

2.1.4. Tabla de frecuencia.

"Al organizar una cantidad de datos en bruto, suele resultar útil distribuirlos en clases o categorías y determinar la cantidad de datos que pertenece a cada clase; esta cantidad se conoce como la frecuencia de clase."

Definición 2.15.- A la disposición tabular de x_i posibles resultados del experimento aleatorio X con su respectiva ocurrencia (Frecuencia absoluta f_i) en la muestra de N experimentos planteados. se le denomina tabla de frecuencias.

X_i	f_i	h_i	F_i	H_i
X_1	f_1	h_1	F_1	H_1
X_2	f_2	h_1	F_1	H_2
….	...	…	…	…
X_k			**N**	1
	N	1		

Tabla 1. Tabla de frecuencia (creación propia)

La suma de las frecuencias absolutas es igual al número total de datos representados por N.

$$\sum_{i=1}^{n} f_i = N$$

Frecuencia relativa (h_i): El resultado de dividir la frecuencia absoluta de un determinado valor entre el número total de datos, se representa por h_i.

$$h_i = \frac{f_i}{N}$$

La frecuencia relativa es la probabilidad de ocurrencia en la muestra del evento x_1.

$$P[X = x_1] = h_i$$

La frecuencia relativa cumple las propiedades:

$0 \leq f_i \leq 1$. $\sum_{i=1}^{k} h_i = h_1 + h_2 + \cdots + h_k = 1$

Datos categóricos.

Los datos categóricos aparecen cuando una variable se mide en una escala que solo clasifica a los encuestados en un número limitado de grupos, por ejemplo, una encuesta donde se recoge información sobre variable como sexo, estado civil, afiliación política. Los datos categóricos son los resultados de un experimento aleatorio X donde su rango R_x es discreto finito.

Las variables categóricas se clasifican:

Dicotómicas: son aquellas que tienen dos posibles resultados.

Ordinales: representan más de dos resultados posibles y con frecuencia en estos resultados es posible considerar algún orden inherente: la muestra de agua es de dureza baja media o alta.

Nominales: Si existen más de dos categorías posibles de resultado y no así un orden inherente entre las categorías entonces se tiene una escala de medida nominal: prefieres la playa, la montaña o la ciudad para vacacionar, no existe una escala subyacente en estos resultados y no hay una forma aparente de ordenarlos.

Recuento: Las variables categóricas a veces tienen recuentos en lugar de considerarlas categorías que representa cada observación (si, no) (bajo, medio, alto) los resultados que se estudian son los mismos números, el tamaño de la camada fue de 1, 2, 3, 4, 5 animales.

En la metodología clásica habitual se analiza a la media de los recuentos los supuestos que se tienen que cumplir en un modelo lineal estándar con datos continuos no se cumplen a menudo; con datos discretos en general los recuentos no se distribuyen según la distribución normal y la variable no suele ser homogénea. Por su característica discreta las tablas de frecuencia en datos categóricos cumple que para cada $\omega_i \in \Omega$ le corresponde un único $X(\omega_i) = x_i$.Esto se ve claro como en el caso de estudiar el color de los autos la variable en estudio es categórica y para su estudio ω_i representa el color que puede tomar el carro *y* x_i.el código que se le asigna único al color.

2.1.5. Tabla de contingencia.

"Una tabla de contingencia es una de las formas más comunes de resumir datos categóricos. En general, el interés se centra en estudiar si existe alguna asociación entre una variable fila y otra variable columna y calcular la intensidad de dicha asociación." (Diazaraque, Tablas de Contingencia)

Sea la población o muestra con $n_{..}$ – unidades sobre los que se pretende analizar simultáneamente dos variables categóricas, digamos X que tiene I categorías y la variable que tiene J categorías; la tabla de frecuencias bidimensional que describe a estas $n_{..}$ – unidades se llama tabla de contingencia (doble entrada).

Definición 2.16. - Sea X e Y dos variables aleatorias. Se define tabla bidimensional a la disposición tabular de (x_i, yi) de posibles resultados del experimento aleatorio bidimensional (X, Y):

TABLA *T (I, J)*

Categorías		**Variable Y**					
		1	**…**	**J**	**…**	**J**	**Total**
Variable X	**1**	n_{11}	…	n_{1j}	…	n_{1j}	$n_{1.}$
	…	…		…		…	…
	i	n_{i1}	…	n_{ij}	…	n_{iJ}	$n_{i.}$
	….	…		…		…	…
	I	n_{I1}	…	n_{Ij}	…	n_{Ij}	$n_{I.}$
	Total	$n_{.1}$	…	$n_{.j}$	…	$n_{.J}$	$n_{..}$

Tabla 2 Tabla bidimensional a la disposición tabular de (x_i,y_i)

NOTACIONES:

n_{ij} : frecuencia absoluta (corresponde a la cantidad de palabras de las obras “la conquista de Jerusalén”, “Numancia”, “el trato de Argel” simultáneamente la modalidad i de la variable X y la modalidad j de la variable Y).

$n_{i.}$: es la frecuencia marginal de X definida por:

$$n_{i.} = \sum_{i=1}^{I} n_{ij}$$

$n_{.j}$: es la frecuencia marginal de Y, definida por:

$$n_{.j} = \sum_{j=1}^{J} n_{ij}$$

$n_{..}$: es el tamaño de la muestra:

$$n_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j}$$

Con el fin de realizar un análisis descriptivo, frecuentemente se considera la tabla de frecuencias relativas, para tal fin construiremos la siguiente tabla:

TABLA *F (I, J)*

Categorías		**Variable Y**					
		1	**…**	**J**	**…**	**J**	**Total**
Variable X	**1**	f_{11}	…	f_{1j}	…	f_{1j}	$f_{1.}$
	…	…		…		…	…
	i	f_{i1}	…	f_{ij}	…	f_{iJ}	$f_{i.}$
	….	…		…		…	…
	I	f_{I1}	…	f_{Ij}	…	f_{Ij}	$f_{I.}$
	Total	$f_{.1}$	…	$f_{.j}$	…	$f_{.J}$	$f_{..}$

Tabla 3 Tabla de frecuencias relativas

NOTACIONES:

Definimos los elementos de la tabla *(I, J):*

- Frecuencia relativa conjunta:

$$f_{ij} = \frac{n_{ij}}{n_{..}}$$

- Frecuencias marginales:

$$f_{i.} = \frac{n_{i.}}{n_{..}} = \sum_{j=1}^{J} \frac{n_{ij}}{n_{..}} = \sum_{j=1}^{J} f_{ij}$$

$$f_{.j} = \frac{n_{.j}}{n_{..}} = \sum_{i=1}^{I} \frac{n_{ij}}{n_{..}} = \sum_{i=1}^{I} f_{ij}$$

- La suma total o la suma de los márgenes son evidentemente igual a uno, puesto que la tabla de frecuencias relativas se obtiene dividiendo la tabla de contingencia por $n_{..}$:

$$f = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}}{n_{..}} = \frac{1}{n_{..}} \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \frac{n_{..}}{n_{..}} = 1$$

Las variables son independientes si se cumplen las siguientes condiciones:

$$f_{ij} = f_{i.} f_{.j} \quad ; \quad f_{i/j} = f_{i.} \quad f_{j/i} = \forall\, i, j$$

Distribución conjunta de dos o más variables

Veamos las distribuciones básicas en el caso de dos variables aleatorias simples X, Y sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, P)$

Definición 2.17.-

1. La función de distribución $F_{X,Y}: \mathbb{R}^2 \to [0,1]$ esta dada por:

$$\mathrm{F_{X,Y}} = \mathrm{P}(X \leq x, Y \leq y)$$

las F_X F_Y se llaman distribuciones marginales

observaciones:

- Como el suceso $(X \leq x)$ es la unión creciente de los $(X \leq x, Y \leq y)$ cuando $y \to \infty$ la marginal $\mathrm{F_X}$ coincide con:

$$\mathrm{F_X(x)} = \mathrm{P}(X \leq x) = \lim_{y \to \infty} \mathrm{P}(X \leq x, Y \leq y) = sup_y \, \mathrm{F_{X,Y}}(x, y)$$

Lo mismo para F_Y

- Para las variables $X_1, \ldots, X_n$ la definición es la misma, para cada $X = (x_1 \ldots x_n) \in \mathbb{R}^n$

$$F_X(x) = P(X_i \leq x_i \, para \; cada \; i \,)$$

Donde $X = (x_1 \ldots x_n)$ es un vector aleatorio que es como conviene pensar en el par (X, Y)

2. La función de densidad P_{XY} es $P_{X,Y}(x, y) = \mathrm{P}(X = x, Y = y)$ si ambas son discretas su relación con la función de distribución: si $\{x_i\}, \{y_i\}$ son los valores de ambas variables:

$$F_{X,Y}(x, y) = \sum_{x_i \leq x, y_j \leq y} P_{X,Y}(x_i, y_j)$$

La distribución conjunta viene dada por

$$\pi_{ij} = P(X = i, Y = j)$$

Con i = 1..., I y j = 1..., J.

Es la probabilidad de (X, Y) en la casilla de la fila i y la columna j.

Distribución marginal

Las distribuciones marginales son:

$$\pi_{i+} = P(X = i) = \sum_{j=1}^{J} P(X = i, Y = j) = \sum_{j=1}^{J} \pi_{ij} \qquad \pi_{+j} = P(Y = j) = \sum_{i=1}^{I} P(X = i, Y = j) = \sum_{i=1}^{I} \pi_{ij}$$

Es decir, el símbolo + indica la suma de las casillas correspondientes a un índice dado. Se cumple siempre que:

$$\sum_{j} \pi_{+j} = \sum_{i} \pi_{i+} = \sum_{i} \sum_{j} \pi_{ij} = 1$$

Distribución condicional

En la mayor parte de las tablas de contingencia, como en el ejemplo anterior, una de las variables, digamos Y, es una variable respuesta y la otra variable X es una variable explicativa o productora. En esta situación no tiene sentido hablar de distribución conjunta.

Cuando se considera una categoría fija de X, entonces Y tiene una distribución de probabilidad que se expresa como una probabilidad condicionada. Así, se puede estudiar el cambio de esta distribución cuando van cambiando los valores de X.

Distribución condicionada de Y respecto de X

$$P(Y = j \,/\, X = i) = \pi_{j/i} = \frac{\pi_{ij}}{\pi_{i+}}$$

Se tiene que

$$\sum_{j} \pi_{j/i} = 1$$

Y el vector de probabilidades$\left(\pi_{1/i} \cdots \pi_{J/i}\right)$ forma la distribución condicionada de Y en la categoría i de X.

La mayor parte de los estudios se centran en la comparación de las distribuciones condicionadas de Y para varios niveles de las variables explicativas.

Independencia y Homogeneidad

Cuando las variables que se consideran son de tipo respuesta, se pueden usar distribuciones conjuntas o bien distribuciones condicionales para describir la asociación entre ellas.

Dos variables son independientes si:

$$\pi_{ij} = \pi_{i+}.\pi_{+j}$$

lo cual implica que la distribución condicionada es igual a la marginal:

$$\pi_{j/i} = \pi_{+j}$$

para j = 1..., J, dado que

$$\pi_{i/j} = \frac{\pi_{ij}}{\pi_{i+}} \forall \text{ i y j}$$

Si X e Y son variables respuesta entonces se habla de independencia

Si Y es variable respuesta y X es variable explicativa entonces se habla de homogeneidad.

2.1.6. Chi cuadrado de Pearson.

“La tabla de contingencia arroja mucha luz a nuestro estudio, pero no basta con interpretar la tabla. Buscamos conseguir una expresión numérica que indique el grado en que existe relación. En términos generales, una buena estrategia para cuantificar una relación es idear un índice o estadístico que mida la distancia que existe entre lo que ocurre y lo que cabría ocurrir si no hubiera absolutamente nada de relación, es decir, si ambas variables fueran totalmente independientes” (Lozano). Si no hay ninguna distancia entre ambas situaciones, el índice suministra el valor 0. Conforme más lejos se encuentre de 0, estará indicando mayor grado de relación.

Para poner eso en práctica en el caso de relación de dos variables nominales expresada mediante una tabla de contingencia, necesitamos identificar qué ocurriría en la tabla si no existiera relación. Dado que nos importa la relación entre ambas variables y no cada una de ellas por separado, los marginales de la tabla permanecen del mismo modo.

Una pregunta que puede surgir ante estas variables es, si las frecuencias o número de casos observados en cada categoría de la variable, a partir de una muestra, difieren de manera significativa respecto a una población esperada de respuestas o frecuencias. (Lozano)

Presentamos el caso en que cada elemento de una población se asigna a una y solo una de varias clases o categorías. Esta población se llama población multinomial. La distribución multinomial de probabilidad se puede concebir como una ampliación de la distribución binomial para el caso de tres o más categorías. En cada ensayo, intento o prueba de un experimento multinomial sólo se presenta uno y sólo uno de los resultados. Cada intento del experimento se supone independiente y las probabilidades de los resultados permanecen igual para cada prueba.

Un método estadístico, llamado técnica chií-cuadrada, nos permite analizar este tipo de variables y tiene cuatro aplicaciones principales:

1.Probar la supuesta independencia de dos variables cualitativas de una población.

2. Hacer inferencias sobre más de dos proporciones de una población.

3. Hacer inferencias sobre la varianza de la población.

4. Realizar pruebas de bondad de ajuste para evaluar la credibilidad de que los datos muestrales, vienen de una población cuyos elementos se ajustan a un tipo específico de distribución de probabilidad.

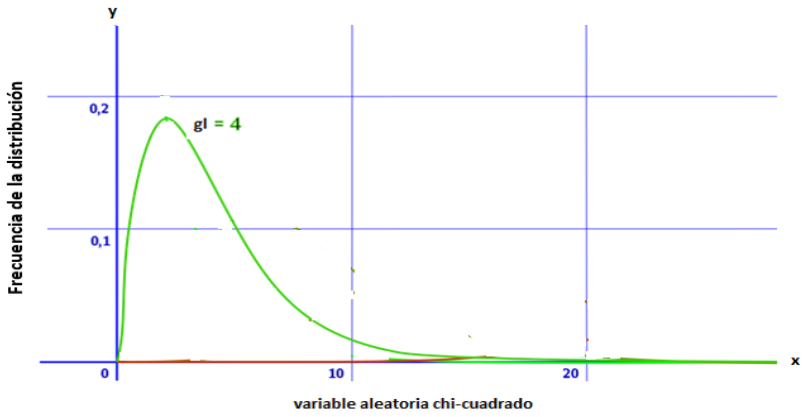


Ilustración 2..Distribución chi cuadrada. (UNAM)

La distribución de probabilidad chií-cuadrada tiene un sesgo positivo como se puede observar en la siguiente figura La distribución Chi-cuadrada, tiende a la normalidad, tal y como se muestra en la siguiente figura a medida que aumentan los grados de libertad.

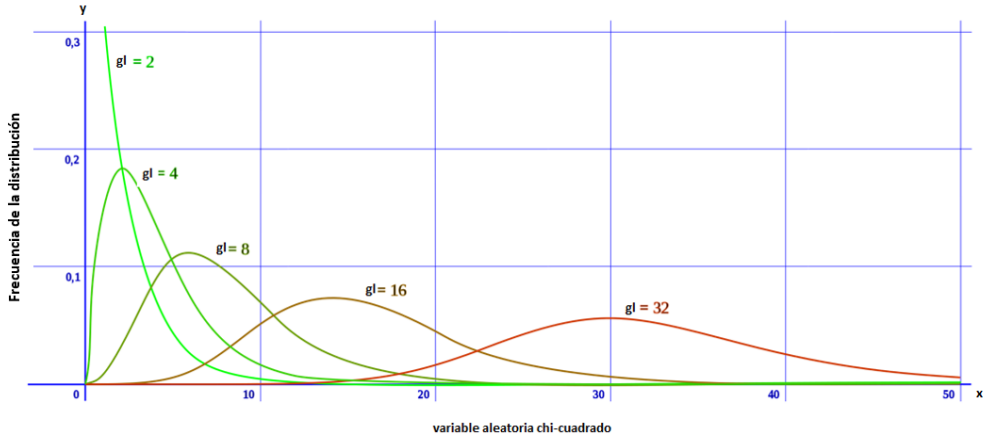


Ilustración 3.Distribución chi cuadrado con g.l. grande. (UNAM)

Cuando k (grados de libertad) de la función de distribución chi-cuadrado es suficientemente grande los valores de la misma tienden a normalizarse siguiendo la siguiente distribución normal:

$$\lim_{n\to\infty}\left(\frac{\chi_k^2\,(x)}{k}\right) \;= N_{\left(1,\sqrt{\frac{2}{k}}\right)}(x)$$

Como consecuencia del teorema de limite central el cual es notorio para valores superiores a 30 grados de libertad siendo una buena aproximación para la función de distribución normal.

$$\chi_k^2\,(x) = k * N_{\left(1,\sqrt{\frac{2}{k}}\right)}(x)$$

La fórmula permite calcular valores de chi-cuadrado cuando los grados de libertad son muy grades.

Ahora se procederá a realizar una definición más formal de la prueba Chi-cuadrada de Pearson para lo cual necesitamos definir la función de distribución Gamma ya que la función de distribución chi-cuadrado solo es un caso particular de la ya mencionada:

FUNCION GAMMA

Definición 2.18. - la función Gamma es una función paramétrica

$$\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} du \ \forall x \epsilon \langle 0, +\infty \rangle$$

Puesto que dicha función converge $\forall x > 0$

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \ \forall \alpha \epsilon \langle 0, +\infty \rangle$$

Se dice que la variable aleatoria continúa x tiene una distribución Gamma de parámetros $\alpha > 0 \ y \ \lambda > 0$ esto es:

$$X \sim Gamma(\alpha, \lambda) \ \alpha > 0 \ y \ \lambda > 0$$

Cuando la función de densidad está dada por:

$$F(x) = \begin{cases} \int_0^x \frac{(\lambda u)^{\alpha-1}}{\Gamma(x)} \lambda e^{-\lambda u} du & si \ x > 0 \\ 0 & si \ x \leq 0 \end{cases}$$

Para $\alpha = n \in \mathbb{N}$

$$f(x) = \begin{cases} \sum_{k=n}^{\infty} \frac{(\lambda x)^k}{k!} e^{-\lambda x} & si \ x > 0 \\ 0 & si \ x \leq 0 \end{cases}$$

Propiedades:

$$E(x) = \frac{\propto}{\lambda}$$

$$Var(x) = \frac{\propto}{\lambda^2}$$

Chi-cuadrado de Pearson

Definición 2.19. - Es una distribución de probabilidad continua que se apoya en un parámetro variable, es decir, si existe o no dependencia estadística entre ellas, se define como la suma de *n* variables independientes al cuadrado dado que cada una es independiente.

$Z_i \sim N(0,1)$ Con *n* grados de libertad.

Función de densidad de probabilidad:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Con lo que se aprecia que es un caso particular de Gamma:

$$\chi_k^2 \sim \Gamma\left(\frac{k}{2}, \theta = \frac{1}{2}\right)$$

Para calcular la probabilidad:

$$P(a < \chi^2(n) \leq b) = \int_a^b f(x)dx$$

Propiedades:

$$f(x) \geq 0$$

$$\int_0^\infty f(x)dx = 1$$

2.1.7. Prueba de independencia Chi-cuadrado.

“Esta prueba nos permite determinar si existe una relación entre dos variables categóricas. Es necesario resaltar que esta prueba nos indica si existe o no una relación entre las variables, pero no indica el grado o el tipo de relación; es decir, no indica el porcentaje de influencia de una variable sobre la otra o la variable que causa la influencia.” (UNAM)

Para comprender mejor este tema es necesario recordar cuales son los eventos independientes y cuales los dependientes.

“Dos eventos aleatorios, A y B, son eventos independientes, si la probabilidad de un evento no está afectada por la ocurrencia del otro evento; $P(A) = P(A \ / \ B)$”

“Dos eventos aleatorios, A y B, son eventos dependientes si la probabilidad de un evento está afectada por la ocurrencia del otro; $P(A) \neq P(A \ / \ B)$”

Una prueba de independencia usa la pregunta de si la ocurrencia del evento X es independiente a la ocurrencia del evento Y, por lo que el planteamiento de las hipótesis para esta prueba de independencia es;

H_0*; La ocurrencia del evento X es independiente del evento Y.*

H_1*; La ocurrencia del evento X no es independiente del evento Y.*

En las pruebas de independencia se utiliza el formato de la tabla de contingencia, y por esa razón a veces se le llama prueba de tabla de contingencia, o prueba con tabla de contingencia. (UNAM)

Independencia entre variables cualitativas

Consideremos dos variables cualitativas X e Y con I y J modalidades cada una respectivamente, y sea N_{ij} la tabla de contingencia asociada a la distribución conjunta de ambas variables, notaremos por n_{ij} la frecuencia absoluta correspondiente a la casilla (i, j). Consideremos también la tabla de frecuencias F_{IJ} obtenida dividiendo cada n ij por el total $n_{..}$ es decir $h_{ij} = \frac{n_{ij}}{n_{..}}$.

Queda expresado en la tabla de contingencia:

	$y_1 \ \dots \ y_j \ \dots \ y_{j'} \ \dots \ y_J$	
x_1	$n_{11} \dots n_{1j} \dots n_{1j'} \dots n_{1J}$	$n_{1.}$
$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$	$\vdots$
x_i	$n_{i1} \dots n_{ij} \dots n_{1j'} \dots n_{iJ}$	$n_{i.}$
$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$	$\vdots$
$x_{i'}$	$n_{i'1} \dots n_{i'j} \dots n_{i'j'} \dots n_{i'J}$	$n_{i'.}$
$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$	$\vdots$
x_I	$n_{I1} \dots n_{Ij} \dots n_{Ij'} \dots n_{IJ}$	$n_{I.}$
	$n_{.1} \dots n_{.j} \dots n_{.j'} \dots n_{.J}$	$n_{..}$

Tabla 4.Frecuencias absolutas de tabla de contingencia. (de creación propia)

	$y_1 \ \dots \ y_j \ \dots \ y_{j'} \ \dots \ y_J$	
x_1	$h_{11} \dots h_{1j} \dots h_{1j'} \dots h_{1J}$	$h_{1.}$
$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$	$\vdots$
x_i	$h_{i1} \dots h_{ij} \dots h_{1j'} \dots h_{iJ}$	$h_{i.}$
$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$	$\vdots$
$x_{i'}$	$h_{i'1} \dots h_{i'j} \dots h_{i'j'} \dots h_{i'J}$	$h_{i'.}$
$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$	$\vdots$
x_I	$h_{I1} \dots h_{Ij} \dots h_{Ij'} \dots h_{IJ}$	$h_{I.}$
	$h_{.1} \dots h_{.j} \dots h_{.j'} \dots h_{.J}$	1

Tabla 5. Frecuencia relativa de tabal de contingencia. (de creación propia)

En esta última tabla tenemos 1 + 1 + J + I tablas unidimensionales, que corresponden a las 2 distribuciones marginales de X y de Y, a las J distribuciones condicionadas de X para cada valor de Y y a las I distribuciones condicionadas de Y para cada valor de X. Las distribuciones condicionadas las podemos escribir:

$$X/y = y_j \rightarrow \left\{ h_{i/j} = \frac{h_{ij}}{h_{.j}}, i = 1 \dots I \right\} \; donde \; j = 1 \dots J \; perfiles \; columna$$

$$Y/x = x_i \rightarrow \left\{ h_{j/i} = \frac{h_{ij}}{h_{i.}}, j = 1 \dots J \right\} \; donde \; i = 1 \dots I \; perfiles \; columna$$

Definición 2.20.- Se dice que X es independiente de Y, si la distribución según el carácter X de los individuos que poseen la modalidad y_j, es la misma cualquiera que sea y_j, es decir las distribuciones condicionadas de X para cada valor y_j

j =1, ..., J, son idénticas, o bien que $\frac{h_{ij}}{h_{.j}}$ j = 1, ..., J no es función de j.

Vamos a ver que, desde el punto de vista estadístico, la independencia entre variables supone la proporcionalidad entre las columnas de la tabla y comprobaremos que es un concepto simétrico, es decir que, si X es independiente de Y, Y también lo es respecto de X, por lo que también se dará proporcionalidad entre las filas de la tabla. Si X es independiente de Y, tendremos:

$$\left\{ \begin{array}{l} \frac{h_{1j}}{h_{.j}} = \frac{h_{1j'}}{h_{.j'}} => \frac{h_{1j}}{h_{1j'}} = \frac{h_{.j}}{h_{.j'}} \\ \frac{h_{2j}}{h_{.j}} = \frac{h_{2j'}}{h_{.j'}} => \frac{h_{2j}}{h_{2j'}} = \frac{h_{.j}}{h_{.j'}} \\ \vdots \quad => \quad \vdots \\ \frac{h_{ij}}{h_{.j}} = \frac{h_{ij'}}{h_{.j'}} => \frac{h_{ij}}{h_{ij'}} = \frac{h_{.j}}{h_{.j'}} \\ \vdots \quad => \quad \vdots \\ \frac{h_{i'j}}{h_{.j}} = \frac{h_{i'j'}}{h_{.j'}} => \frac{h_{i'j}}{h_{i'j'}} = \frac{h_{.j}}{h_{.j'}} \\ \vdots \quad => \quad \vdots \\ \frac{h_{Ij}}{h_{.j}} = \frac{h_{Ij'}}{h_{.j'}} => \frac{h_{Ij}}{h_{Ij'}} = \frac{h_{.j}}{h_{.j'}} \end{array} \right\} => \frac{h_{1j}}{h_{1j'}} = \frac{h_{2j}}{h_{2j'}} = \cdots = \frac{h_{ij}}{h_{ij'}} = \cdots = \frac{h_{i'j}}{h_{i'j'}} = \cdots = \frac{h_{Ij}}{h_{Ij'}} (\forall i, j)$$

Por tanto:

$$\forall i, i', j, j' => \frac{h_{ij}}{h_{ij'}} = \frac{h_{i'j}}{h_{i'j'}} => \frac{n_{ij}}{n_{ij'}} = \frac{n_{i'j}}{n_{i'j'}}$$

De donde se deduce que las columnas de la tabla de frecuencias absolutas son proporcionales, en el caso de que X sea independiente de Y.

Además, teníamos que: $\forall i = 1,2, \dots, I$

$$\frac{h_{i1}}{h_{.1}} = \frac{h_{i2}}{h_{.2}} = \cdots = \frac{h_{ij}}{h_{.j}} = \frac{h_{ij'}}{h_{.j'}} = \cdots = \frac{h_{iJ}}{h_{.J}}$$

De donde aplicando la propiedad de las fracciones que la suma de los antecedentes debido por la suma de los consecuentes es igual a cada una de las fracciones, obtenemos:

$$\frac{h_{i1}}{h_{.1}} = \frac{h_{i2}}{h_{.2}} = \cdots = \frac{h_{ij}}{h_{.j}} = \cdots = \frac{h_{ij'}}{h_{.j'}} = \cdots = \frac{h_{iJ}}{h_{.J}} = \frac{h_{i1} + h_{i2} + \cdots + h_{ij} + \cdots + h_{ij'} + \cdots + h_{iJ}}{h_{.1} + h_{.2} + \cdots + h_{.j} + \cdots + h_{.j'} + \cdots + h_{.J}}$$

$$= f_i$$

$$=> \left\{ h_{i/j} = \frac{h_{ij}}{h_{.j}} \qquad \begin{matrix} i = 1, \dots, I \\ j = 1, \dots, J \end{matrix} \right\} => \left\{ h_{ij} = h_{i.} h_{.j} \quad \begin{matrix} i = 1, \dots, I \\ j = 1, \dots, J \end{matrix} \right\}$$

Lo que indica independencia entre X e Y.

Además de $h_{ij} = h_{i.} h_{.j}$ como $h_{j/i} = \frac{h_{ij}}{h_{i.}} => \left\{ h_{j/i} = h_{.j} \;\; i = 1, \dots, I \right\}$ lo que indica independencia entre X e Y

Luego también tendremos proporcionalidad entre las filas de la tabla de frecuencias.

Contraste de independencia Chi-cuadrado

La hipótesis de independencia se puede escribir:

$$H_0: h_{ij} = h_{i.}h_{.j} \begin{Bmatrix} i = 1, \dots, I \\ j = 1, \dots, J \end{Bmatrix}$$

Veamos un procedimiento para contrastar dicha hipótesis como $h_{ij} = \frac{n_{ij}}{n_{..}}$ sabemos que el valor observado $n_{ij} = n_{..}h_{ij}$; mientras que el valor esperado bajo la hipótesis de independencia sería $e_{ij} = n_{..}h_{i.}h_{.j}$ por lo que una forma de realizar el contraste sería ver la discrepancia entre los valores observados y esperados en este caso si sumamos las diferencias $(n_{ij} - e_{ij})$ los valores positivos se compensarían con los negativos por lo que una posible solución sería elevar al cuadrado dichas diferencias $(n_{ij} - e_{ij})^2$ aun asi convendría normalizar las discrepancias calculándolas en valores relativos $\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$ lo que permite definir el estadístico χ_i^2

$$\chi^2_{exp} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{..}f_{ij} - n_{..}h_{i.}h_{.j})^2}{n_{..}h_{i.}h_{.j}}$$

$$= n\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(h_{ij} - h_{i.}h_{.j})^2}{h_{i.}h_{.j}}$$

Por lo tanto, siendo la cantidad χ^2_{exp} una medida de la discrepancia entre los valores observados y esperados H_0 si el valor es pequeño y la hipótesis alternativa H_1 en caso que sea grande.

Sin embargo necesitamos un criterio para decir bajo que valores del estadístico χ^2 aceptamos la hipótesis de independencia H_0 entre las variables, tal criterio exige conocer la distribución de probabilidad del estadístico χ^2_{exp} en este sentido Pearson demostró que asumiendo las frecuencias observadas n_{ij} siguen una distribución multinomial, el estadístico χ^2 para grandes tamaños muestrales sigue una distribución Chi-cuadrado con $(I-1).(J-1)$ grados de libertad

$$\chi^2_{exp} \sim> \chi^2_{(I-1)(J-1)}$$

Los grados de libertad de una tabla de contingencia pueden considerarse como el número de celdas de la tabla que se pueden fijar libremente cuando se fijan los totales marginales es decir la diferencia entre el número de casillas de la tabla y el número de restricciones impuestas: $ij - (i-1) + (j-1) - 1$

Conocida la distribución del estadístico χ^2 para contrastar con un nivel de significación α la hipótesis H_0 de independencia entre X e Y hacemos lo siguiente:

Calculamos el valor crítico C de la distribución $\chi^2_{(I-1)(J-1)}$ tal que:

$$P\left[\chi^2_{(I-1)(J-1)} > C\right] = \alpha$$

Si el valor del estadístico $x > C => P\left[\chi^2_{(I-1)(J-1)} > \chi^2\right] < \alpha$

Que significa que estamos ante una muestra rara rechazamos H_0 y aceptamos H_1 del valor del estadístico $\chi^2 > C => P\left[\chi^2_{(I-1)(J-1)} > \chi^2\right] < \alpha$ lo que nos llevara a aceptar H_0

Esto mismo es lo que hacen los paquetes estadísticos a través del P-valué para un nivel de significación $\alpha = 0.05$; si el $p - value < 0.05$ se acepta H_1 en caso contrario si el $p - value > 0.05$ se acepta H_0, la hipótesis de independencia entre las variables, la relación entre el estadístico χ^2_{exp} la relación de independencia del p-valué viene a ser la siguiente:

Si el χ^2_{exp} es muy gande => existe dependencia entre las variables =>
$(p - value) -> 0$

Si el χ^2_{exp} es muy pequeño => existe independencia entre las variables =>
$(p - value) -> 1$

2.1.8. Análisis de correspondencia.

El análisis de correspondencia es una técnica estadística que se usa para analizar desde el punto de vista gráfico las relaciones de dependencia e independencia de un conjunto de variables categóricas a partir de los datos de una tabla de contingencia. (Fernandez, 2011)

El análisis factorial de correspondencias AFC es una técnica exploratoria de análisis multivariante que no requiere ninguna conjetura respecto a la distribución de probabilidad de la población del cual se extrajo la muestra multivariante, se explica el análisis de tablas de contingencia para ver las semejanzas entre las categorías de cada una de las variables, obteniendo así un diagrama cartesiano basado en la asociación de las variables analizadas.

En el gráfico de mapas porcentuales se representan conjuntamente las distintas categorías de la tabla de contingencia, de forma que la proximidad entre los puntos representados está relacionada con el nivel de asociación entre dichas categorías. Para ello asocia a cada una de las categorías de la tabla, un punto en el espacio R^n(habitualmente n = 2, para un caso de análisis factorial de correspondencia simple), de forma que las relaciones de dependencia y semejanza existentes entre ellas sean visibles.

El análisis de correspondencias, es un método multivariado que reduce la dimensión (tamaño de la tabla de contingencia), para el estudio de las relaciones de interdependencia entre las variables categóricas. Convierte las categorías de la tabla de frecuencias (filas y columnas) en un menor número de dimensiones, indicando qué porcentaje del valor Chi-cuadrado de la asociación puede ser explicado por las nuevas dimensiones. Por ello guarda cierta analogía con la prueba Chi-cuadrado y con el coeficiente de concordancia de Kendall, (Visauta, 1998)

- El objetivo de la técnica es establecer relaciones entre variables categóricas dispuestas en una tabla de contingencia.
- Trabaja con variables categóricas, es decir no con mediciones cuantitativas sino con frecuencias.

- Las relaciones entre las variantes se realizan mediante mapas porcentuales muy intuitivos que permiten no sólo reducir el número de variables que intervienen en el análisis, sino estudiar las formas que adoptan las relaciones entre las variables.

Esta técnica tiene la ventaja que para su uso no se hace ninguna conjetura respecto a la distribución de probabilidad de la población de la cual se extrajo la muestra multivariante y va más allá de analizar la relación existente entre las variables, porque permite determinar cómo está estructurada esta relación describiendo "proximidades", que permiten identificar "categorías causantes de asociación".

Entre las técnicas de composición, el análisis factorial es el más parecido; pero el análisis de correspondencias va más allá del análisis factorial, su aplicación más directa es la representación de la "correspondencia" de categorías de variables particularmente aquellas medidas con escalas nominales.

El análisis de correspondencias aportará información que de ninguna manera proporciona la Chi-cuadrada y los coeficientes de correlación, también calculará: perfil, inercias, contribuciones; de las diversas filas y/o columnas de la tabla y además nos permitirá analizar esta posible relación entre las variables de un modo gráfico en un espacio bidimensional, permite observar las similitudes o alejamiento de las variables dejando ver cuáles son las categorías que se encuentran relacionadas, a mayor o menor proximidad entre las categorías del plano, que son más fáciles de interpretar.

El análisis factorial de correspondencias de una tabla de contingencias persigue dos objetivos fundamentales los que son importantes detallar para justificar este método.

- Analizar la estructura de una tabla de contingencia respetando el hecho de que la misma resume una relación simétrica entre los caracteres observados.
- Representar gráficamente la estructura de una tabla de contingencia.

ANALISIS FACTORIAL DE CORRESPONDENCIAS SIMPLES (ACS)

El análisis de correspondencias simples es una técnica para representar las categorías de las dos variables en un espacio de pequeña dimensión que permita interpretar, por un lado, las similitudes de categorías entre variable respecto a las categorías de otra y las categorías de ambas variables.

Igual que el análisis de componentes principales, el ACS trata de explicar la dispersión de la matriz de varianza-covarianza (aunque en este caso se denomina la matriz de inercia) a través de un número menor de variables (factores), pero este análisis debe realizarse tanto para las filas como para las columnas. Por tanto, es un caso particular del análisis de componentes principales, uno para el espacio que definen las filas y otro para el espacio que definen las columnas.

En tal caso, resulta interesante transformar las variables métricas en otras que no sean de este modo, todas las variables estarán medidas en la misma escala (no métrica) y será posible operar con ellas conjuntamente aplicando el ACS.

TABLA DE PERFILES FILA Y COLUMNA

Reflejan las proporciones que el número de individuos de cada celda representa sobre el total de la fila y sobre el total de la columna respectivamente.

El grado de similitud entre estos perfiles tanto por filas como por columnas quedan reflejados en cada gráfico en términos de proximidad o lejanía entre las categorías de las variables.

PERFIL FILA. - En el estudio de las filas, la tabla de datos se transforma dividiendo cada término f_{ij} de la fila i por la marginal $f_{i.}$ de esta fila i. y la nueva fila se denomina perfil-fila.

TABLA DE PERFILES FILA

Categorías		Variable Y					
		1	**…**	**j**	**…**	**J**	**Total**
Variable X	**1**	H_{11}	…	H_{1j}	…	H_{1j}	1
	…	…		…		…	…
	i	H_{i1}	…	H_{ij}	…	H_{iJ}	1
	….	…		…		…	…
	I	H_{I1}	…	H_{Ij}	…	H_{Ij}	1
	Total	$H_{.1}$	…	$H_{.j}$	…	$H_{.J}$	1

Tabla 6. Tabla de perfiles fila

Donde:

$\frac{f_{ij}}{f_{i.}}$: Representa el porcentaje de elementos de la población que cumplen la categoría j sabiendo que poseen la condición i de la primera variable.

Se denomina perfil fila "i" a la distribución de frecuencias de las categorías del factor Y condicionadas a las categorías del factor X esto dado por:

$$H_i = \left[\frac{f_{i1}}{f_{i.}}, \frac{f_{i2}}{f_{i.}}, \dots, \frac{f_{iJ}}{f_{i.}}\right] = \left[H_{i1}, H_{i2}, \dots, H_{iJ}\right] \qquad i = 1,2,3, \dots, I$$

PERFIL COLUMNA. - En el estudio de las columnas, la tabla de datos se transforma dividiendo cada termino f_{ij} de la columna j por la marginal $f_{.j}$ de esta columna. La nueva columna se denomina perfil-columna.

TABLA DE PERFILES COLUMNA

Categorías		Variable Y					
		1	**…**	**j**	**…**	**J**	**Total**
	1	F_{11}	…	F_{1j}	…	F_{1j}	F_{1J}
	…	…		…		…	…
Variable	**i**	F_{i1}	…	F_{ij}	…	F_{iJ}	F_{iJ}
X	**….**	…		…		…	…
	I	F_{I1}	…	F_{Ij}	…	F_{Ij}	F_{IJ}
	Total	1	…	1	…	1	1

Tabla 7. Tabla perfil columna

Donde:

$\frac{f_{ij}}{f_{.j}}$: representa el porcentaje de elementos de la población que cumplen la categoría i sabiendo que poseen la condición j de la primera variable.

Se denomina perfil columna “j” a la distribución de frecuencias de las categorías del factor X condicionadas a las categorías del factor Y, esto dado por:

$$F_j = \left[\frac{f_{1j}}{f_{i.}}, \frac{f_{2j}}{f_{i.}}, \dots, \frac{f_{Ij}}{f_{i.}}\right] = \left[H_{1j}, H_{1j}, \dots, H_{Ij}\right] \qquad j = 1,2,3, \dots, J$$

Los perfiles columna pueden compararse con la distribución de las frecuencias del factor X

El resultado de la asociación se presenta como dos casos (que representa lo mismo), si los perfiles fila o columna de categorías distintas tienen igual comportamiento las variables son independientes, en caso contrario están asociados.

Distancia entre los elementos fila y columna

En un espacio multidimensional puede definirse una distancia entre dos puntos (categorías), para analizar la semejanza entre ellos. Para ello es necesario definir el tipo de distancia a usarse

Distancia euclidiana. - La distancia más intuitiva entre dos puntos es la euclidiana, para definir esta distancia es necesario recordar el teorema de Pitágoras, podemos calcular fácilmente las distancias entre dos puntos de un plano.

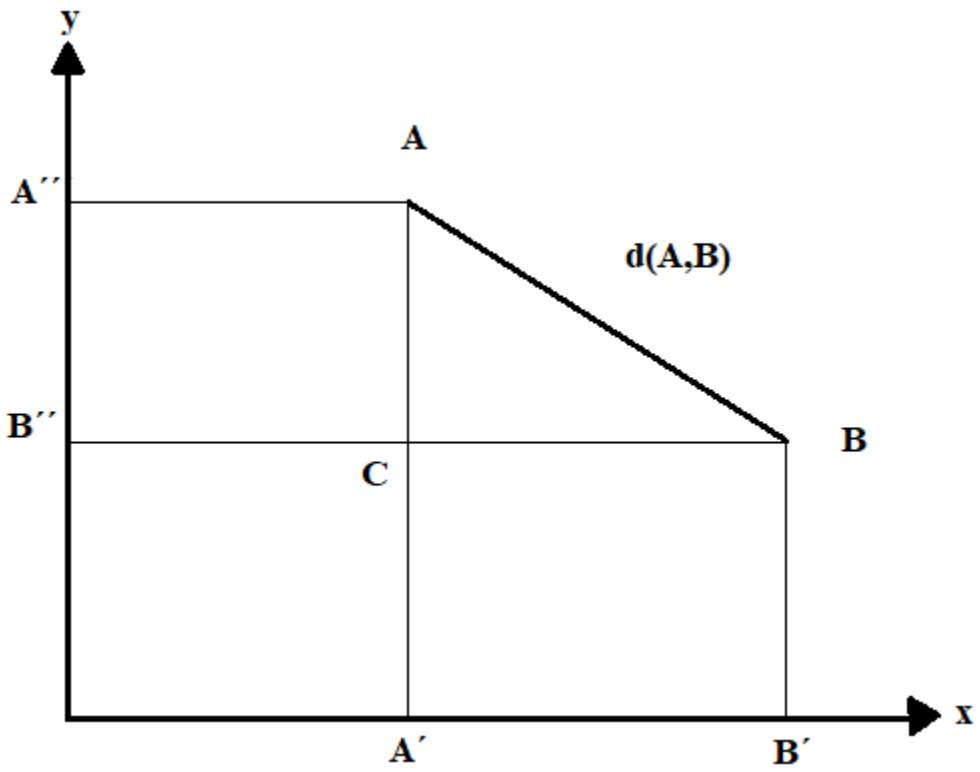


Ilustración 4. Distancia euclidiana

La distancia entre los puntos A= (A', A") y B= (B', B") está dada por la siguiente expresión:

$$d_{(A,B)} = \sqrt{(B' - A')^2 + (B'' - A'')^2}$$

La expresión general de esta distancia entre dos elementos lineales de una tabla de contingencia, entre dos elementos fila de una tabla de contingencia está expresada de la siguiente manera:

$$d_{(i,i')} = \sqrt{\sum_{j=1}^{J} (n_{ij} - n_{i'j})^2}$$

La distancia entre dos elementos columna de una tabla de contingencia está expresada de la siguiente manera:

$$d_{(j,j')} = \sqrt{\sum_{i=1}^{I} (n_{ij} - n_{ij'})^2}$$

Propiedades de la distancia euclidiana

I. Cuando comparamos dos elementos fila o columna de una tabla establecemos una relación de "similitud o disimilitud" de diferencias dos categorías de X, si esta distancia es cerca de cero entonces las categorías son similares, caso contrario son diferentes.

$$d_{(i,i')} \geq 0 \quad ; \quad d_{(j,j')} \geq 0$$

II. Si los elementos comparados por fila son idénticos, para todos los elementos de la tabla, entonces:

$$d_{(i,i')} = 0 \leftrightarrow i = i'$$

$$d_{(j,j')} = 0 \leftrightarrow j = j'$$

III. $d_{(i,i')} = d_{(i',i)} \quad ; \quad d_{(j,j')} = d_{(j',j)}$

IV. Si consideramos tres elementos de una tabla de frecuencia, se verifica que:

$$d_{(i,i')} = d_{(i.k)} + d_{(i'.k)}$$

$$d_{(j,j')} = d_{(j.k)} + d_{(j'.k)}$$

Distancia de la tabla de frecuencias *F (I, J)*

La expresión general de la distancia en el caso de la comparación de dos elementos lineales de una tabla *F (I, J)* esta expresado de la siguiente manera:

$$d_{(i,i')} = \sqrt{\sum_{j=1}^{J} (f_{ij} - f_{i'j})^2}$$

De la misma manera, la distancia entre dos elementos columnas de una tabla F (I, J) es la siguiente:

$$d_{(j.j')} = \sqrt{\sum_{j=1}^{J} (f_{ij} - f_{ij'})^2}$$

NUBE DE PUNTOS

Cada perfil-fila es un conjunto de I valores numéricos y puede ser representado por un punto en el espacio R^I, en el que cada una de las I dimensiones está asociado a una categoría de la segunda variable.

La distancia χ^2 que define la semejanza entre perfiles-fila posee las propiedades de una distancia euclidiana y confiere a R^I la estructura de espacio euclideo.

Esta distancia conduce a asignar a la j-esima dimensión del R^J, el peso de $\frac{1}{f_{.j}}$

La suma de las coordenadas de cada perfil-fila vale uno; resultando que la nube de puntos fila (N_I) pertenece a un hiperplano denotado por H_I. En caso de R^3: (Arroyo, 2010)

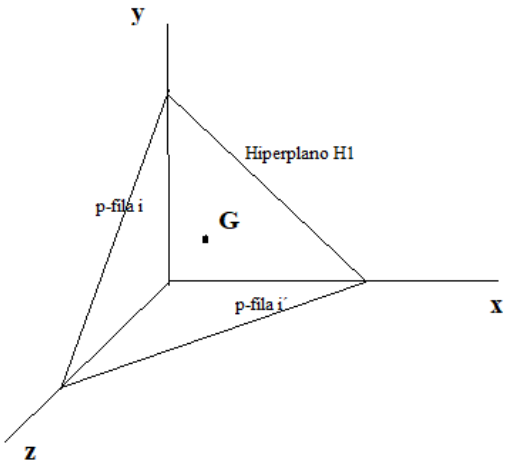


Ilustración 5. Nube de puntos

- El punto i tiene por coordenada sobre el eje j, $\frac{f_{ij}}{f_{i.}}$
- Su peso es $f_{i.}$
- La distancia entre dos perfiles es la distancia chi-cuadrado
- El baricentro (G) de la nube N_I tiene por coordenadas sobre el eje j la frecuencia marginal $f_{.j}$
- La nube N_I pertenece a un hiperplano H_I

En el análisis de correspondencias los pesos de cada punto de la nube vienen impuestos el punto i tiene un peso igual a la frecuencia marginal $f_{i.}$, este peso es proporcional al efectivo de la clase de individuos que representa.

El baricentro de los puntos N_I dotados de estos pesos se denota por G_I . su j-esima coordenada es la mdeia ponderada de las j-ésima coordenadas de los puntos N_I

$$G_I = \frac{\sum_{i=1}^{I} f_{i.} \frac{f_{ij}}{f_{i.}}}{\sum_{i=1}^{I} f_{i.}} = f_{.j}$$

G_I : es el centro de gravedad de la nube y se interpreta como un perfil medio.

Así al estudiar qué medida y de qué manera una clase de individuos i difiere del conjunto de población, conduce a estudiar la desviación entre el perfil de esta clase y el perfil medio.

Similarmente se define la nube de perfiles- columna y está expresada de la siguiente manera:

$$G_J = \frac{\sum_{j=1}^{J} f_{.j} \frac{f_{ij}}{f_{.j}}}{\sum_{j=1}^{J} f_{.j}} = f_{.j}$$

La distancia de cada columna y de cada fila al centro de gravedad se expresa como sigue:

$$d^2_{(i,G_I)} = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 = \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}} \right)^2 ; para\ fila$$

$$d^2_{(j,G_I)} = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2 = \sum_{i=1}^{I} \left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \sqrt{f_{i.}} \right)^2 ; para\ columnas$$

EL AJUSTE DE LAS NUBES

Desde el punto de vista del análisis de datos, interesa reducir la nube de puntos de manera que se obtenga la representación a la vez accesible a nuestra visión y fiel en el sentido de que la representación de la nube mantenga la mayor información que ella pueda contener. La representación será accesible si se proyecta la nube sobre un espacio de pequeña dimensión y será completa si la dispersión de la nube proyectada es casi igual a la de la nube propiamente dicha.

En general se trata de buscar un sub espacio de dimensión q en R^J , q < j la misma que nos permite encontrar un sistema de vectores $(u_1, u_2, u_3, \dots u_q)$ y q' es el tamaño del subespacio generado en el espacio R^I , q' < i encontrando un sistema de vectores $(v_1, v_2, v_3, \dots v_{q'})$ orto normado para la métrica (R^I, R^J), que tiene el subespacio de manera que sea máxima la inercia de las nubes sobre los subespacios.

AJUSTE Y REPRESENTACION DE LA NUBE DE PERFILES-FILA N(I)

En R^I , el ajuste trata de obtener un conjunto de imágenes planas proximidades de la nube N(I), donde I = (1,2,3, … i), dotados de pesos $p_i = (f_{.1}, f_{.2}, \dots f_{.i})$. al igual que en análisis de componentes principales, el análisis de correspondencia simple consiste en buscar un conjunto de ejes ortogonales sobre los que será proyectada la nube.

Las imágenes planas de N(I) deben ser tales que las distancias entre los puntos de la imagen se asemejen lo más posible a las distancias entre los puntos de N(I). este objetivo es completamente análogo al del ajuste de la nube de individuos en análisis de componentes principales, en la práctica aplica que la nube analizada sea centrada, es decir, que su baricentro sea elegido como de los ejes.

La nube centrada de la clase definida por la categoría i está representada por un punto cuya coordenada sobre el j-ésima eje es $\frac{f_{ij}}{f_{i.}} - f_{.j}$ (diferencia entre la coordenada del perfil fila y G_i baricentro N(I)).

La posición de este punto expresa la diferencia entre la distribución de la clase i y de la población total del conjunto de las categorías de la segunda variable.

Determinar las direcciones de la inercia máxima de la nube centrada es obtener las clases que más se desvían del perfil del conjunto de la población esto es: (Arroyo, 2010)

$$Inercia \dots N(I) = \sum_{i=1}^{I} inercia(i)$$

$$= \sum_{i=1}^{I} f_{i.} d^2(i, G_i)$$

$$= \sum_{i=1}^{I} f_{i.} \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$$

$$= \sum_{i=1}^{I} f_{i.} \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{.j}}{\sqrt{f_{.j}}} \right)^2$$

$$= \sum_{i=1}^{I} f_{i.} \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}} \right)^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} f_{i.} \left(\frac{f_{ij} - \sqrt{f_{.j}}\sqrt{f_{.j}} f_{i.}}{\sqrt{f_{.j}} f_{i.}} \right)^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} f_{i.} \left(\frac{f_{ij} - f_{.j} f_{i.}}{\sqrt{f_{.j}} f_{i.}} \right)^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} f_{i.} \frac{\left(f_{ij} - f_{.j} f_{i.}\right)^2}{f_{.j} f_{i.}^2}$$

$$Inercia \dots N(I) = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(f_{ij} - f_{.j} f_{i.}\right)^2}{f_{.j} f_{i.}}$$

Similarmente el ajuste y presentación de la nube de perfiles-columna en la nube centrada de la clase definida por la categoría j está representada por un punto cuya coordenada sobre el i-esima eje es $\frac{f_{ij}}{f_{.j}} - f_{i.}$ (diferencia entre la coordenada del perfil fila y G_i baricentro N(J)).

La posición de este punto expresa la diferencia entre la distribución de la categoría j y la distribución total sobre el conjunto de las categorías de la segunda variable.

Determinar la distribución de la inercia máxima de la nube es generar las clases que más se desvían del perfil del conjunto de la población, esto es:

$$Inercia \dots N(J) = \sum_{j=1}^{J} inercia(j)$$

$$= \sum_{j=1}^{J} f_{.j} d^2(j, G_i)$$

$$= \sum_{j=1}^{J} f_{.j} \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2$$

$$Inercia \dots N(J) = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(f_{ij} - f_{.j} f_{i.} \right)^2}{f_{i.} f_{.j}}$$

La inercia en una medida de dispersión total de la nube de puntos respecto a su centro de gravedad.

Cada perfil está dado por un peso igual a su frecuencia marginal f_i .Este peso interviene en primer lugar en el cálculo de centro de gravedad de la nube y también interviene en la inercia, por tanto, en el criterio de ajustes de los ejes.

Los p valores de los perfiles fila configuran un vector $\overline{x} = \left(x_{i1}, x_{i2}, \dots x_{ij} \right)$ que se representa como un punto en el espacio R^J y los I perfiles fila formen una nube de puntos en R^J.

Un conjunto de perfiles fila puede caracterizarse por su gravedad e inercia, la inercia de nube de puntos es una medida resumida de dispersión, se define como la suma para todos los puntos del producto de sus masas por los cuadrados de su distancia al centro de gravedad está dado por:

$$Inercia = \sum_{i=1}^{I} f_i d^2(i, G_i)$$

Uno de los objetivos del análisis de correspondencia es reducir la nube de puntos: es decir encontrar un sistema de vectores en R^J , de manera que el ajuste trate de obtener un conjunto de imágenes planas aproximadas de la nube N(I).

Al igual que en análisis de componentes principales, el análisis de correspondencia simple consiste en buscar un conjunto de ejes ortogonales sobre los que será proyectada la nube; geométricamente se tiene:

Representación de la inercia de ejes ortogonales

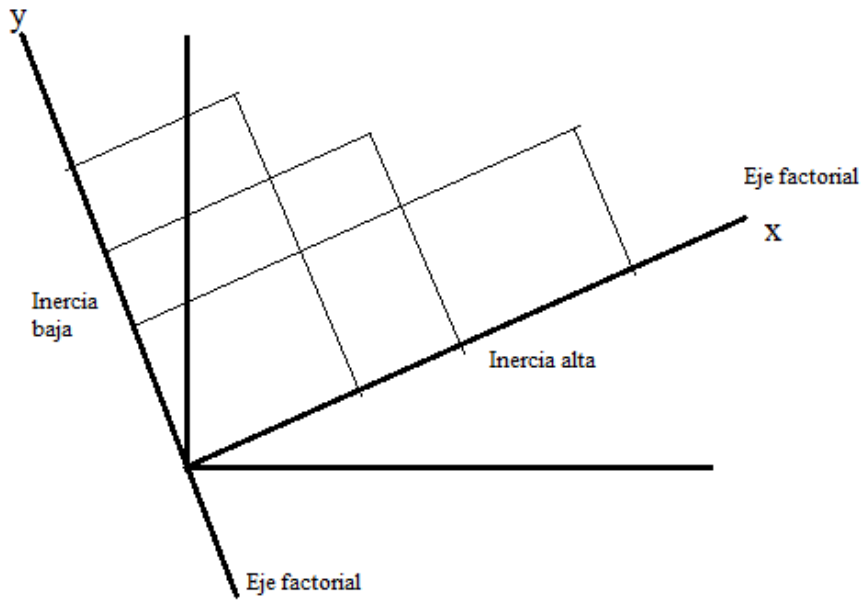


Ilustración 6. Inercia de ejes ortogonales

REPRESENTACION N(I)

La representación de las categorías de la primera variable (perfil fila) en dimensión reducida determinadas por las p_i coordenadas con referencia a las categorías de la segunda variable (perfila columna), se puede interpretar como un problema de representación de datos mediante el análisis de componentes principales.

Sea:

$$Z = \left(\frac{f_{ij}}{\sqrt{f_{i.} f_{.j}}} \right)$$

Una matriz nxp cuyas filas son las coordenadas p_i de las medias de las variables- calculadas sobre la matriz de datos Z, ponderadas por las frecuencias relativas $(f_1, f_2, \dots f_n)$, se tiene el vector de medias.

$$M = \left(\sqrt{f_{.1}}, \sqrt{f_{.2}}, \ldots, \sqrt{f_{.n}}\right)$$

La covarianza entre las categorías j y j', ponderado por las frecuencias relativas es:

$$C_{jj'} = \sum_{i=1}^{n} f_{i.}\left(\frac{f_{ij}f_{ij'}}{\sqrt{f_{.j}}f_{i.}\sqrt{f_{.j'}}f_{i,}} - \sqrt{f_{.j}}\sqrt{f_{.j'}}\right)$$

$$C_{jj'} = \sum_{i=1}^{n}\left(\frac{f_{ij}f_{ij'}}{\sqrt{f_{.j}}\sqrt{f_{.j'}}f_{i,}} - \sqrt{f_{.j}}\sqrt{f_{.j'}}\right)$$

En términos matriciales la covarianza es:

$$C_p = Z^{\cdot\cdot}D_n Z - MM^t$$

Donde:

$$D_n = diagonal(f_1, f_2, \ldots f_n)$$

PROPIEDADES:

Se verifican las siguientes propiedades.

M es el auto vector de C_p para el auto valor $\lambda = 0$ esto es:

En efecto basta probar que:

$$C_{1j'}\sqrt{f_{.1}}, C_{2j'}\sqrt{f_{.2}}, \ldots, C_{pj'}\sqrt{f_{.p}}=0$$

Reemplazando las ecuaciones:

$$C_p(M) = (Z^{\cdot\cdot}D_n Z - MM^t)(M)$$

$$C_p(M) = (Z^{\cdot\cdot}D_n Z - MM^t)\begin{bmatrix}\sqrt{f_{.1}}\\ \ldots \\ \sqrt{f_{.p}}\end{bmatrix}$$

$$C_p(M) = \sum_{j=1}^{p}\left\{\sum_{i=1}^{n}\left(\frac{f_{ij}f_{ij'}}{\sqrt{f_{.j}}\sqrt{f_{.j'}}f_{i.}}\right) - \sqrt{f_{.j}}\sqrt{f_{.j'}}\right\}\sqrt{f_{.j}}$$

$$= \sum_{j=1}^{p} \left\{ \sum_{i=1}^{n} \left(\frac{f_{ij} f_{ij'}}{\sqrt{f_{.j}} \sqrt{f_{.j'}} f_{i.}} \right) \sqrt{f_{.j}} - \sum_{j=1}^{p} \sqrt{f_{.j}} \sqrt{f_{.j'}} \sqrt{f_{.j}} \right\}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{f_{ij} f_{ij'}}{f_{i.} \sqrt{f_{.j'}}} \right) - \sqrt{f_{.j'}}$$

$$= \sum_{i=1}^{n} \left(\frac{f_{ij'}}{\sqrt{f_{.j'}}} \right) - \sqrt{f_{.j'}} = 0$$

$$C_p M = \lambda M = 0 \Rightarrow \lambda = 0$$

Los auto vectores de C_p son también vectores propios de $Z' D_n Z$

Si V es el vector propio de C_p distinto de M de valor propio λ, entonces V es ortogonal a M, es decir:

$$M'V = 0$$

$$C_p V = \lambda V$$

Sustituyendo la ecuación (e) en (g) se tiene:

$$(Z' D_n Z - MM')V = \lambda V$$

Operando llegamos a:

$$Z' D_n Z V - MM'V = \lambda V$$

$$M'V = 0 \Rightarrow Z' D_n Z V = \lambda V$$

M es auto valor de $Z' D_n Z = \sum_{i=1}^{n} \left(\frac{f_{ij} f_{ij'}}{\sqrt{f_{.j}} \sqrt{f_{.j'}} f_{i.}} \right)_{pxp}$ para el auto valor λ=1

Por definición de auto vector se tiene que:

$$Z' D_n Z M = \lambda M \quad \ldots . (h)$$

Sustituyendo por sus frecuencias relativas se tiene:

$$Z'D_nZM = \sum_{i=1}^{n}\left(\frac{f_{ij}f_{ij'}}{\sqrt{f_{.j}}\sqrt{f_{.j'}}f_{i.}}\right)_{p*p}\begin{bmatrix}\sqrt{f_{.1}}\\ \dots\\ \sqrt{f_{.p}}\end{bmatrix}_{p*1}$$

$$\left(\sum_{j=1}^{p}\left(\sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{\sqrt{f_{.j}}\sqrt{f_{.j'}}f_{i.}}\right)\sqrt{f_{.j}}\right)$$

$$=\left(\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{f_{ij}}{f_{i.}}\frac{f_{ij'}}{\sqrt{f_{.j'}}}\right)$$

$$=\left(\sum_{j=1}^{p}\frac{f_{ij'}\sqrt{f_{.j'}}}{f_{.j'}}\right)$$

$$Z'D_nZM = \left(\sqrt{f_{.j'}}\right)_{p*1}$$

Luego reemplazando en (h) se tiene:

$$\left(\sqrt{f_{.j'}}\right)_{p*1} = \lambda\left(\sqrt{f_{.j'}}\right)_{p*1} \Rightarrow \lambda = 1$$

Como consecuencia de estas propiedades bastará diagonalizar $Z'^{D_n}Z$ y considerar solo los vectores de valor propio distinto de uno. Como el valor propio uno corresponde al valor propio cero de C_p , los demás valores propios de $Z'D_nZ$ son menores que uno. Diagonalizando $Z'^{D_n}Z$ cuyo término general es:

$$t_{ij} = \sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}f_{.j'}}}$$

Obtenemos los valores propios de $D_\lambda = diagonal(1, \lambda_1, \lambda_2, \dots, \lambda_p)$, donde cada valor propio a la inercia tiene asociado un vector propio u, como consecuencia obtenemos la matriz de vectores propios U , de manera que a u_2 se llama primer eje factorial de inercia λ_2.

TABLA DE INERCIA

Las tasas de inercia permiten evaluar la calidad global del ajuste y está asociada al eje factorial (α) indica la parte de la inercia total de la nube proyectada sobre este eje. En forma general será:

$$\tau_\alpha = \frac{\lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha}$$

El porcentaje de la inercia explicada por el segundo y tercer eje factorial será:

$$P_d = \frac{\lambda_2 + \lambda_3 + \cdots + \lambda_d}{\lambda_2 + \lambda_3 + \cdots + \lambda_p}$$

El número de ejes factoriales de la N(I) no puede superar a la menor de las dos cantidades (n-1), (p-1)

$$q < min[(n-1), (p-1)]$$

El sub espacio obtenido por los q-ejes factoriales se denomina soporte de N(I).

COORDENADAS FACTORIALES DE LOS PUNTOS PERFILES-FILA

La coordenada de los perfiles-fila vendrán dadas a partir del producto de la matriz de los perfiles transformados por la matriz de los vectores propios, es decir:

$$F = ZU$$

Donde el termino general es:

$$F_{\alpha i} = \sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} u_{\alpha j}$$

$$F_{\alpha j} = \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}\sqrt{f_{.j}}} u_{\alpha i}$$

AJUSTE Y REPRESENTACION DE LA NUBE DE PERFILES-COLUMNA N(J) ANALISIS EN R^I.

Debido al papel simétrico que juegan las filas y las columnas en el análisis de correspondencias, el ajuste en R^I se plantea en los mismos términos y posee las mismas propiedades que el ajuste en R^J, es decir:

- Las imágenes planas de N(J) deben ser tales que las distancias entre los perfiles proyectados se asemejen lo más posible a las distancias entre los perfiles en R^I. De ahí se deriva la necesidad de analizar la nube N(j) con relación a su baricentro G_j. La inercia total de N(J) con respecto a G_j proviene de las diferencias entre los perfiles de las diferentes clases y el perfil conjunto de la población.
- La coordenada de los puntos j es $\frac{f_{ij}}{f_{.j}}$
- El peso de los puntos j es $f_{.j}$
- El centro de gravedad G tiene de coordenadas $g = \sqrt{f_{.j}}$
- La matriz de perfiles columna transformadas y centradas es:

$$z = \frac{f_{ij}}{\sqrt{f_{i.}} f_{.j}}$$

- La proyección de un punto j sobre α cuyo vector director de v_α es:

$$G_{\alpha i} = \sum_{i=1}^{J} \frac{f_{ij}}{\sqrt{f_{i.}} f_{.j}} v_{\alpha i}$$

Similarmente para la proyección del perfil fila se tiene como vector director de u_α es:

$$G_{\alpha j} = \sum_{i=1}^{J} \frac{f_{ij}}{\sqrt{f_{.j}} f_{i.}} u_{\alpha j}$$

Matricialmente las coordenadas de los puntos perfiles columna será:

$$G = ZV$$

Recordemos que también se puede obtener las coordenadas de los puntos perfiles columna a través de las relaciones de transición trabajadas en análisis de componentes principales, es decir:

$$v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} X_{ij} u_{\alpha j}$$

$$u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} X'_{ij} u_{\alpha i}$$

Es decir que:

$$Coord(j, \alpha) = G_{\alpha j} = \frac{\sqrt{\lambda_\alpha}}{\sqrt{f_{.j}}} u_{\alpha j}$$

REPRESENTACION DE LAS NUBES DE UN MISMO PLANO

Las relaciones existentes entre los dos sub espacios permiten representar simultáneamente las dos nubes en un mismo plano.

$$G_{\alpha i} = \sum_{i=1}^{I} \frac{f_{ij}}{\sqrt{f_{i.} f_{.j}}} v_{\alpha i} \quad y \quad v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} F_{\alpha i} \sqrt{f_{i.}}$$

sustituyendo se tiene:

$$G_{\alpha i} = \sum_{i=1}^{I} \frac{f_{ij}}{\sqrt{f_{i.} f_{.j}}} \frac{1}{\sqrt{\lambda_\alpha}} F_{\alpha i} \sqrt{f_{i.}}$$

$$G_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{I} \frac{f_{ij}}{\sqrt{f_{i.} f_{.j}}} F_{\alpha i} \sqrt{f_{i.}}$$

$$G_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} F_{\alpha i}$$

similarmente sustituyendo la ecuación $u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} F_{\alpha j} \sqrt{f_{.j}}$ en la ecuacion (13) se tiene:

$$G_{\alpha j} = \sum_{j=1}^{J} \frac{f_{ij}}{\sqrt{f_{.j}} f_{i.}} \frac{1}{\sqrt{\lambda_\alpha}} F_{\alpha j} \sqrt{f_{.j}}$$

$$G_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{J} \frac{f_{ij}}{\sqrt{f_{.j}} f_{i.}} F_{\alpha j} \sqrt{f_{.j}}$$

$$G_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}} F_{\alpha j}$$

Esto significa que la proyección de los puntos i sobre el espacio formado por los factores es igual a la proyección de los puntos j ponderados en un coeficiente $\frac{f_{ij}}{f_{i.}}$ que es el peso que tiene cada fila y por un coeficiente que es la raíz del autovalor.

Para el caso de las proyecciones de los puntos j, las relaciones permiten representar simultáneamente sobre el mismo plano los puntos fila y columna, permitiendo la interpretación de las relaciones entre líneas y columnas.

DEFINICIÓN DE LOS EJES E INTERPRETACIÓN DE LA INERCIA

En el Análisis de Componentes Principales, para asignar un nombre a los factores, se debe tomar en cuenta las correlaciones entre las variables (contribuciones). En el Análisis de Correspondencia Simples, una vez obtenidas las coordenadas del perfil fila y perfil columna, representados los puntos en el mismo plano se debe conocer qué categorías son las que más han contribuido en la construcción de los ejes, es decir el peso que tiene cada categoría en la definición de cada eje.

Por otro lado, la inercia de una nube de puntos se descompone sobre toda base ortogonal, la inercia es la suma de sus inercias sobre cada uno de los ejes de la base.

El ajuste de las nubes N(I) y N(J) descompone su inercia según las direcciones principales, debido a la ortogonalidad de los ejes, la suma de las inercias de una nube sobre cada uno de los ejes es igual a la inercia total de la nube.

Contrariamente al caso del ACP, en el que la inercia de las nubes es igual al número de las variables, es ACS esta inercia expresa la estructura de la tabla.

La inercia de cada una de las dos nubes de perfiles-fila y perfiles-columna es igual al estadístico chi-cuadrado. El ACS es por tanto la descomposición de este estadístico y cada factor representa una parte de la relación entre las variables.

CONTRIBUCION ABSOLUTA Y RELATIVA DE LOS PERFILES FILA

a. **CONTRIBUCIONES ABSOLUTAS POR FILAS. -** Expresan la proporción de la varianza por un eje debido a un perfil (i, j). es decir, permiten saber que variables son las responsables de la contribución de un factor, determina cuanto aporta un punto (i, j) en la inercia variabilidad de la proyección de un factor.

   Las contribuciones absolutas representan porcentualmente la importancia que tiene esta categoría en la definición de cada eje, que está definido por cada categoría de la variable y permite interpretar los ejes, definida por:

$$Cta(i,\alpha) = \frac{f_{i.}F_{\alpha}^{2}(i)}{\lambda_{\alpha}}$$

   como:

$$\sum_{i=1}^{I} f_{i.}\, F_{\alpha}^{2}(i) = \lambda_{\alpha}$$

   Dado que una distribución absoluta de una fila o columna es un porcentaje de la inercia que explica un factor, la suma de las contribuciones absolutas para todas las filas o todas las columnas en un determinado factor debe ser uno o expresar el cien por ciento de la inercia del eje. No solo depende de la distancia a la que se encuentra el punto, sino también de su peso o ponderación.

b. **CONTRIBUCIONES RELATIVAS POR FILAS. -** Expresan la contribución de un factor en la explicación de la dispersión de un elemento, esta medida nos proporciona la calidad d la representación de la categoría.

Las construcciones relativas muestran cuales son las características exclusivas de ese factor, cuantifica la parte del punto (i, j) en la inercia explicada por el eje factorial.

$$Ctr(i,\alpha) = \frac{F_\alpha^2(i)}{d^2(i,G)}$$

como:

$$d^2(i,G) = \sum_{j=1}^{J} \left(\frac{f_{ij}}{\sqrt{f_{.j}} f_{i.}} - \sqrt{f_{.j}} \right)^2$$

CONTRIBUCION ABSOLUTA Y RELATVA DE LOS PERFILES COLUMNA

a. **Contribuciones absolutas por columnas. -** esta expresado por:

$$Cta(j,\alpha) = \frac{f_{.j} F_\alpha^2(j)}{\lambda_\alpha}$$

como:

$$\sum_{j=1}^{J} f_{.j}\ F_\alpha^2(j) = \lambda_\alpha$$

b. **Contribuciones relativas por columnas. -** Esta expresada por:

$$Ctr(j,\alpha) = \frac{F_\alpha^2(j)}{d^2(j,G)}$$

como:

$$d^2(j,G) = \sum_{j=1}^{J} \left(\frac{f_{ij}}{\sqrt{f_{.j}} f_{i.}} - \sqrt{f_{.j}} \right)^2$$

La contribución relativa es un porcentaje de la distancia que separa a una fila o columna en cada uno de los factores y mide la calidad de representación de la fila o la columna sobre el factor α, la suma de la contribución relativa para cada uno de los factores es igual a la unidad.

Las filas o las columnas tendrán mayor contribución relativa en un factor a medida que ese factor sea responsable de la distancia que separa a la misma del origen de coordenadas.

Mientras las contribuciones absolutas permitan saber que variable son las responsables de la contribución del eje, las contribuciones relativas consideran cuál de las características exclusiva de ese factor.

2.1.9. Matriz

Una matriz es una colección ordenada de elementos colocados en filas y columnas, o sea es un arreglo bidimensional de números (llamados entradas de la matriz) ordenados en filas (o renglones) y columnas, donde una fila es cada una de las líneas horizontales de la matriz y una columna es cada una de las líneas verticales. A una matriz con m filas y n columnas se le denomina matriz m por n (mxn) donde m y n son números naturales mayores que cero. El conjunto de las matrices de tamaño mxn se representa como $M_{mxn}(\mathbb{K})$, donde $\mathbb{K}$ es el campo al cual pertenecen las entradas.

Definición 2.21. - Los **auto valores de la matriz** cuadrada A_{nxn} son todos los valores de λ para los cuales los sistemas tienen soluciones no triviales.

$$det\left[\begin{pmatrix} a_{11} & \dots & a_{1j} \\ \dots & \dots & \dots \\ a_{i1} & \dots & a_{ij} \end{pmatrix} - \begin{pmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_i \end{pmatrix}\right] = 0$$

Si la matriz A es nxn tiene auto valores, estos pueden ser repetidos y también ser números complejos.

Debido a la complejidad de la solución de los auto valores está dado por la solución del sistema polinómico de grado n indicando que a más dimensiones presenta la matriz el polinomio será de mayor complejidad. Los polinomios de grado superior a tres presentan en su solución valores complejos.

Definición 2.22: Dado los n auto valores reales o complejos de la matriz A_{nxn} y la traza de A entonces.

$$tr(A) = \sum_{i=1}^{n} \lambda_i \qquad \bar{\lambda} = \frac{\sum_{i=1}^{n} \lambda_i}{n} = \frac{tr(A)}{n}$$

Esta propiedad permitirá el cálculo de media aritmética de los auto valores sin necesidad del cálculo de los auto valores. (Wiki, s.f.)

Definición 2.23: Dado los n auto valores reales o complejos de la matriz A_{nxn} y la determínate de A entonces.

$$det(A) = \prod_{i=1}^{n} \lambda_i \qquad geo(\overline{\lambda}) = \sqrt[n]{\prod_{i=1}^{n} \lambda_i} = \sqrt[n]{\det(A)}$$

Esta propiedad permitirá el cálculo de media geométrica de los auto valores sin necesidad del cálculo de los auto valores. Para que el valor de media geométrica sea positivo se recomienda que n sea par

Matriz de transición de una cadena de Markov.

Una cadena Markov es un proceso en tiempo discreto en el que una variable aleatoria X_n va cambiando con el paso del tiempo, tienen la propiedad de que la variable aleatoria $X_n = j$ solo depende del estado inmediatamente anterior del sistema X_{n-1} . (Diazaraque, Cadenas de Markov)

Probabilidades de transición.

En una cadena finita homogénea con m posibles estados $(E_1, E_2, \dots E_m)$ se puede introducir la notación

$$P_{ij} = P(X_n = j / X_{n-1} = i)$$

Donde ij=1, 2, …, m si $P_{ij} > 0$ entonces se dice que el estado E_j la comunicación puede ser mutua si también $P_{ji} > 0$

Para cada i fijo la serie de valores $\{P_{ij}\}$ es una distribución de probabilidad ya que en cualquier paso puede recurrir a alguno de los sucesos $(E_1, E_2, \dots E_m)$ y son mutuamente excluyentes, los valores de P_{ij} se denomina probabilidad de transición que satisface la condición:

$P_{ij} > 0$

$\sum_{j=1}^{m} P_{ij} - 1$

Para cada i= 1,2, …, m todos estos valores se combinan formando una matriz de transición T de tamaño mxm donde:

$$M_T = T = [P_{ij}] = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1m} \\ P_{21} & P_{22} & \dots & P_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{m1} & P_{m2} & \dots & P_{mm} \end{bmatrix}$$

Que puede observar que cada fila de la matriz es una distribución de probabilidad, es decir: $\sum_{j=1}^{m} P_{ij} = 1$

2.1.10. Población normal multivariante.

Si se realiza tareas de inferencia sobre el vector de medias y la matriz de covarianza de una población multivariante en base a una muestra aleatoria simple extraída de ella también se tratan problemas que involucren a varias poblaciones y muchos procedimientos resultan ser extensiones naturales de los métodos y a conocidos para poblaciones normales univariantes mientras que en algún caso surgirán problemas nuevos.

Como prerrequisito para el estudio se recuerda la situación univariante la inferencia tiene base en el teorema de Fisher donde dice que la media tiene distribución normal con cierta varianza y la varianza tiene distribución Chi-cuadrado y son independientes y también se afirma que el vector de medias es normal multivariante la matriz de covarianza muestral tiene distribución Wishart y son independientes Aunque el vector de medias muestral y la matriz de covarianzas muestral son estimadores naturales de sus análogos poblacionales. (Sellero, 2008)

Un vector aleatorio es la colección de variables aleatorias.

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$$

Medias simultáneamente sobre el mismo individuo o sobre el mismo resultado de un experimento aleatorio se tiene algunas propiedades conjuntas dentro de un vector aleatorio como la covarianza y la distribución conjunta se define vector de medias y covarianzas como:

$$E(x) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_d) \end{pmatrix} \qquad \Sigma = cov(X,X) = \begin{pmatrix} var(X_1) & cov(X_1X_2) & \dots & cov(X_1X_d) \\ cov(X_2x_1) & var(X_2) & \dots & cov(X_2X_d) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_dX_1) & cov(X_dX_2) & \dots & var(X_d) \end{pmatrix}$$

La estandarización de una variable aleatoria se consigue restando la media y dividiendo por la desviación típica en el caso de vector aleatorio su estandarización seria: $Y = \Sigma^{-1/2}(X - u)$

Distribución de frecuencias multivariantes.

Supongamos que una muestra aleatoria simple de un vector aleatorio normal multivariante se puede denotar por $X_1, \dots, X_n \epsilon N_d(\mu, \Sigma)$ independientes entonces:

$$\bar{X} = \frac{1}{n}\Sigma X_i \epsilon N_d\left(\mu \frac{1}{n}\Sigma\right)$$

Y este resultado es suficiente para obtener un pivote para μ cuando la matriz de covarianza es conocida el cual resulta de la estandarización de $\bar{X}$ así:

$$n(X - \mu)'\Sigma^{-1}\bar{X} - \mu \epsilon X_d^2$$

En base este pivote se puede obtener una región de confianza para el vector de medias con nivel de confianza 1-α de la forma:

$$\{\mu \epsilon \mathbb{R}^d : (n\bar{X} - \mu)\Sigma^{-1}\,(\bar{X} - \mu) < X_d^2\,\}$$

Observamos que la región de confianza que se encuentra dentro de las llaves es la región limitada por una elipse en el plano (si d=2), un balón de Rugby en el espacio (si d=3) y así sucesivamente. Se trata de un elipsoide en $\mathbb{R}^d$ centrado en $\bar{X}$ cuyos ejes van en la dirección de los auto vectores de Σ y la longitud de los radios (semilongitud de los ejes) viene dada por:

$$\sqrt{\lambda_j}\sqrt{X_{d\alpha}^2/_n}\ con\ j \epsilon \{1, \dots, d\}$$

siendo $\lambda_1, \dots, \lambda_d$ los autovalores de Σ, en el caso bidimensional d=2 se puede presentar la elipse aplicando la siguiente expresión para los puntos que la forman:

$$\bar{X} + \sqrt{X^2_{d\alpha}/_n}\left[\sqrt{X_1}v_1\cos(\theta) + \sqrt{\lambda_2}v_2\sin\theta\right] con\ \theta\epsilon[0{,}2\pi]$$

Siendo $v_1\, y\, v_2$ los autovectores de Σ y $\lambda_1\lambda_2$ sus autovalores respectivos. Al igual que ocurría en el caso univariante con la desviación típica ahora si la matriz de covarianza es desconocida es necesario estimarla mediante su análogo muestral lo cual conduce a una distribución diferente que se puede considerar una extensión de la T de Student es la distribución Hotteling. Todo ello nace de la extensión del teorema de Fisher al caso multivariante que dice lo siguiente: si $X_1, \dots, X_n \epsilon N_d(\mu\Sigma)$ independiente entonces: $\bar{X} = \frac{1}{n}\Sigma\epsilon N_d\left(\mu\frac{1}{n}\ \Sigma\right)$

$$nS = \frac{1}{n}\Sigma(X_i - \bar{X})(X_i - \bar{X})'\epsilon W_d(\Sigma, n-1)$$

Y demás son independientes de ello y de la definición de distribución Γ^2 de Hottelling se obtiene el pivote siguiente:

$$(n-1)(\bar{X} - \mu)'S^{-1}(\bar{x} - \mu)\ \epsilon\ \Gamma^2(d, n-1)$$

La distribución de Hotteling se puede transformar en una F de Snedecor y en este caso resulta:

$$\frac{n-d}{d}(X-\mu)'S^{-1}(X-\mu)\epsilon F_d(n-d)$$

Con el conocimiento de las distribuciones multivariante. Al igual que el caso unidimensional se pude plantear pruebas de hipótesis para la media de distintas maneras y también para la varianza.

2.1.11. Inferencia de matriz de covarianza.

Se plantea el contraste de la hipótesis nula sobre la matriz de covarianzas:

$$H_0: \Sigma = \Sigma_0$$

Suponemos que el vector de medias $\vec{\mu}$ es desconocido y queremos contrastar una hipótesis nula simple sobre la matriz de covarianzas H_0: $\Sigma = \Sigma_0$ frente a una alternativa en la que la matriz de covarianzas no está sujeta restricciones. El vector de medias carece de restricciones bajo la hipótesis nula como bajo la alternativa.

Aplicando el procedimiento de razón de verosimilitudes resulta el estadístico de contraste: (Sellero, 2008)

$$\omega = -2\ln(\lambda) = 2(l_1^* - l_0^*) = -2\ln\frac{sup_\mu L(X, \mu, \Sigma_0)}{sup_\mu L(X.\mu, \Sigma)}$$

Por otro lado, observamos que por ser Σ definida positiva y en consecuencia $(\Sigma^{-1})(\bar{x} - \mu) > 0$ salvo que $\mu = \bar{X}$ en cuyo caso es cero por tanto la función de log-verosimilitud alcanza su máximo en $\mu = \bar{X}$ que de este modo se convierte en el estimador de máxima verosimilitud del vector de medias, además.

$$sup_u \ln L(X, \mu, \Sigma) = C - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X})'\,\Sigma^{-1}(X_i - \bar{X})$$

Para cualquier matriz de covarianza Σ

A continuación, se calculará el máximo de aquella función respecto de Σ podemos expresar:

$$\sup\ \ln \mathrm{L}(X, \mu, \Sigma) = C - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}traza[(\Sigma^n X_i - \bar{X})'\Sigma^{-1}(X_i - \bar{X})]$$

Del cual se obtiene que:

$$\sup\ \ln \mathrm{L}(X, \mu, \Sigma) = C - \frac{n}{2}(\ln|\Sigma_0| + traza\,(\Sigma_0^{-1}S))$$

Y la expresión

$$\sup_{\Sigma} \sup_{\mu} \ln \mathrm{L}(X, \mu, \Sigma) = C - \frac{n}{2}(\ln|S| + d)$$

De modo que el estadístico de contraste adopta la forma.

$$\omega = -2\ln(\lambda) = n(ln|\Sigma_0| + traza\,(\Sigma_0^{-1}S) - ln(S) - d)$$

$$= n(tr(\Sigma_0^{-1}S) - \ln(|\Sigma_0^{-1}|(S)) - d)$$

$$= n\left(\left(\sum_{i=1}^{p}\lambda_i\right) - \ln\left(\prod_{j=1}^{d}\lambda_j\right) - d\right)$$

$$= n(da - \ln(g^d) - pd)$$

$$= nd(a - \ln(g) - 1)$$

Siendo $\lambda_1 \ldots \lambda_d$ los autovalores de la matriz $(\Sigma_0^{-1}S)$

a la media aritmética de los auto valores.

g su media geométrica de los auto valores.

La distribución exacta de este estadístico bajo la hipótesis nula no se encuentra disponible en su lugar usaremos la distribución asintótica que presenta por ser un estadístico de razón de verosimilitudes.

$$-2log\lambda = np(a - \ln(g) - 1) \sim \mathrm{X}_m^2$$

$$\log\lambda(X) = nd(a - \log(g) - 1) \sim X_m^2$$

Siendo **m** el número de grados de libertad la diferencia entre el número de parámetros independientes bajos la hipótesis alternativa y bajo la hipótesis nula que en este caso resulta $m = \frac{1}{2}d(d+1)$, puesto que es el número de parámetros independientes en una matriz de covarianzas.

Por haberse construido como cociente de verosimilitudes bajo la hipótesis nula y alternativa se rechaza la hipótesis nula cuando el estadístico es grande o mejor

dicho cuando supere el cuartil $1-\alpha$ de la distribución $X^2_{m\alpha}$ siendo α el nivel de significación fijado con anticipación.

Si se supone que el vector de medias sea conocido siguiendo los mismos pasos habríamos llegado al estadístico de contraste

$$-2\log\lambda(X) = nd(a - \log g - 1)$$

Siendo a y g las medias aritmética y geométrica respectivamente de los auto valores de la matriz $\Sigma_0^{-1}\hat{\Sigma}_\mu$ la única diferencia radica en la sustitución de S por el estimador:

$$\Sigma_\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\,(X_i - \mu)'$$

Nuevamente tenemos los problemas con la distribución del estadístico de contraste y aplicamos a la distribución asintótica que es X^2_m con el mismo número de grados de libertad $m = \frac{1}{2}d(d+1)$

Teorema 2.2.- sea $X_1 \ldots X_n$ una muestra aleatoria simple de $N_d(\mu, \Sigma)$ si las hipótesis nula y alterna conducen a los estimadores de máxima verosimilitud $\hat{\Sigma}$ y S respectivamente y si $\bar{X}$ es de estimador de máxima verosimilitud para μ bajo cualquiera de las hipótesis el estadístico de razón de verosimilitudes para contrastar H_0 frente a H_1 viene dado por

$$-2\log\lambda(X) = nd(a - \log g - 1)$$

Siendo a y g las medias aritmética y geométrica respectivamente de los auto valores de la matriz $\hat{\Sigma}^{-1}S$

2.1.12. Inferencia de matriz de transición.

Proposición 2.1.- Se plantea el contraste de la hipótesis nula sobre la matriz transición:

$$H_0: T_1 = T_2$$

Dado que la prueba de hipótesis H_0: $\Sigma = \Sigma_0$ no está sujeta restricciones no depende de los valores de media, la hipótesis se puede plantear para toda matriz, es decir:

$$H_0: M_1 = M_2$$

Donde M_1 y M_2 son matrices cuadradas.

Y como el estimador de prueba de hipótesis está planteado desde procedimiento de razón de verosimilitudes resulta el estadístico de contraste.

$$-2\log\lambda(X) = nd(a - \log g - 1)$$

Siendo a y g las medias aritmética y geométrica respectivamente de los auto valores de la matriz $M_1 M_2^{-1}$y de misma manera para la matriz de transición estocástica.

Dado que el tamaño de muestra se considera como **d** el número de dimensiones de la matriz de transición.

$$-2log\lambda = np(a - \ln(g) - 1) \sim \mathrm{X}_m^2$$

$$m = \frac{1}{2}d(d+1)$$

dado que el estadístico se basa en una prueba de covarianza sigue una distribución asintótica de Chi-cuadrado. Por haberse construido como cociente de verosimilitudes bajo la hipótesis nula y alternativa se rechaza la hipótesis nula cuando el estadístico es grande o mejor dicho cuando supere el cuartil $1 - \alpha$ de la distribución $X_{m\alpha}^2$ siendo α el nivel de significación fijado con anticipación.

2.2. Bases teóricas en Informática.

2.2.1. Corpus literario

En principio, se puede llamar corpus a cualquier colección que contenga más de un texto, en las diferentes definiciones de corpus propuestas en los últimos años. Leech lo define *"A primera vista, un corpus de computadora es un fenómeno poco emocionante: una gran cantidad de texto, almacenado en una computadora".* Leech (1992:106).

Aunque Leech completa su definición recalcando que la habilidad que poseen los ordenadores para buscar, recuperar, ordenar y hacer cálculos sobre cantidades masivas de texto nos ha brindado la oportunidad de comprender y de explicar el contenido de estos corpus de formas que no eran imaginables en la era que él denomina "pre computacional". (Perez)

De hecho, dado que los avances tecnológicos van tan unidos al desarrollo de la lingüística de corpus tal y como hoy en día la conocemos, Leech argumenta que debe denominarse Computer Corpus Linguistics, ya que el término "lingüística de corpus" se usaba antes del advenimiento de los ordenadores digitales

2.2.2. Datamining.

El concepto surgió hace pocos años para ayudar a la comprensión de los contenidos de las bases de datos. Para el Datamining los datos son la materia prima bruta a los que los usuarios dan un significado convirtiéndolos en información que posteriormente será tratada y utilizada por los especialistas para convertirlos en conocimiento(Galeón).

La datamining ha conseguido reunir las ventajas de áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, las bases de datos como materia prima. Molina y otros lo definirían como "la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión" (Molina y otros, 2001).

2.2.3. Minería de texto

Definiremos la minería de textos o tex mining como técnica de recuperación y organización de la información. Una definición más estandarizada la ofrece el grupo de trabajo dedicado a los córpora textuales de EAGLES: “Corpus: una colección de piezas de lenguaje que se seleccionan y ordenan de acuerdo con criterios lingüísticos explícitos para ser utilizadas como muestra del idioma.” EAGLES (Expert Advisory Group on Language Engineering Standards) (1996a:4):

Es una de las ramas de la lingüística computacional que trata de obtener información y conocimiento a partir de conjuntos de datos que en principio no tienen un orden o no están dispuestos en origen para transmitir esa información. Es una técnica clave en un mundo como el actual en el que continuamente se recogen datos desde distintas perspectivas y de muchos aspectos diferentes de todas las actividades propias de los seres humanos.

Los datos a tratar con esta técnica serán, en lugar de los datos de las bases de datos, los documentos y textos de las organizaciones, administraciones, compañías, entre otros.

El Text Mining no se dedica a la recuperación de la información, que es la indexación de textos, clasificación, categorización. La información que le interesa a la minería de textos es aquella contenida en esos documentos, pero de manera general.

Comprendiendo tres actividades fundamentales: Recuperación de información, Extracción de la información incluida en esos textos, minería de datos para encontrar asociaciones entre esos datos claves previamente extraídos de entre el texto.

La minería de texto ayuda a que información implícita en los documentos pueda expresarse de manera explícita. Encontrando en sus utilidades como ejemplo (Félix)

Un ejemplo claro de la utilización de las técnicas de minería de textos lo realizó Hearst en 1999 y es incluido como ejemplo en el artículo en “Data mining: torturando a los datos hasta que confiesen” de Luis Carlos Molina Félix. En el describe como Don Swanson trato de extraer información a partir de colecciones de texto y demostró cómo cadenas de implicaciones causales dentro de la literatura médica

pueden conducir a hipótesis para enfermedades poco frecuentes, como por ejemplo ocurrió con la migraña. Se pudieron extraer evidencias a partir de varios artículos de literatura biomédica y algunas de los descubrimientos fueron:

El estrés está asociado con la migraña y puede conducir a la pérdida de magnesio que es un bloqueador natural del canal de calcio. Esta información se consiguió de manera indirecta haciendo un análisis de los diferentes textos médicos. Estudios posteriores probaron experimentalmente esta hipótesis.

La minería de textos es una tecnología recuperación y organización de la información que, aunque todavía es emergente y necesita ser mejor desarrollada, nos sirve para obtener un tipo de información muy útil en cualquier tipo de organización pública o privada.

Cada vez es más fácil recabar datos y guardarlos adecuadamente. El reto es saber aprovechar el potencial de conocimiento escondido en ellos. Gracias a técnicas como el Text Mining se puede recuperar conocimiento implícito encerrado en los textos.

2.2.4. Stemming y lematizacion.

Técnica que ayuda a agrupar palabras a su origen es decir aun lema que reduce el gran volumen de vocabulario , si se desea buscar la frase “comi manzana” sin stemming solo busca “comi” pero con steming se busca la palabra origen que en este caso seria “comer” esto abarca un mayor número de posibilidades. (Gabancho)

Conservando el sentido de la informacion la mayoria de los buscadores red hoy aplican esta técnica en sus buscadores permitiendo dar mayor numero de respuestas. Pero como incoveniente encontramos que reduce la informacion generando un error posible.

2.3. Antecedentes de la investigación.

2.3.1. ““Análisis factorial de correspondencias de pacientes con patologías oculares en el CEPRECE-CUSCO””.

2010 universidad Nacional San Antonio Abad del Cusco desarrollado por la Br. Luz Marina Cantuta Guillen.

En su estudio plantea relacionar la técnica estadística de análisis de correspondencias con la ciencia de la salud en las patologías oculares, probar la efectividad del análisis de correspondencia en la inferencia de patologías oculares para lo cual recauda una base de datos que contiene factores sociodemográficos y de la falencia ocular de la cataratas visuales con el fin de procesar la información mediante el paquete estadístico SPSS en su versión 16 es con el fin de optimizar la toma de decisiones en la salud visual.

Las variables planteadas como variables dependientes fueron las patologías oculares de los pacientes atendidos en CEPRECE-CUSCO como lo son: catarata, conjuntivitis, errores refractivos, glaucoma, petriguim. Y las variables independientes entre las cuales encontramos edades genero entre otras. (Guillen., 2010)

Plantea el objetivo de ver las asociaciones existentes entre enfermedades oculares y las variables sociodemográficas.

En los resultados del estudio se inició con un estudio descriptivo analítico. Para proceder a realizar la prueba de asociación para tablas de contingencia de Chi-cuadrado que es pre requisito para la técnica planteada en el estudio análisis de correspondencia. Encontrando conclusiones respecto a las variables sociodemográficas en estudio, pero sí así relaciones entre las enfermedades oculares estudiadas.

Del estudio podemos hacer referencia que la técnica formulada puede presentar un amplio uso y que su marco teórico fue referencia para el estudio que formulamos.

2.3.2. “Modelo de Redes con recursos didácticos con lingüística computacional”. 2014 Universidad Andina Néstor Cáceres Velásquez. por Ms. Jean Roger Farfán Gabancho.

El docente de la facultad de Ingeniería de sistemas, plantea el objetivo de crear una base de datos y buscador basado en minería de texto para la búsqueda de tesis en el ámbito universitario. en el cual habla acerca de la lingüística computacional y como crear una base de datos para que los alumnos tengan facilidad de búsqueda de anteriores tesis. El autor enfatiza la importancia que tiene el internet en los alumnos y profesores para la adquisición de información ya sea para hacer tareas exposiciones y demás. Mediante el uso de la lingüística computacional también conocido como procesamiento del lenguaje natural (PLN) el que sirve de puente en la relación de palabras y el lenguaje computacional. La que busca similitud semántica a los vocablos de un documento representado por proximidad o cercanía de un espacio de n dimensiones. En el estudio se platea el uso de técnicas y términos informáticos como son:

“El parser o parsing, algoritmo analizador sintáctico el cual retorna una estructura arbórea de derivacion como estructura de datos, los parser tienen una ventaja de que no dependen de un idioma, son aplicables a la mayoria de idiomas excepto el japones y el chino no alfabéticos”. (sidorov 2001) .El stemming nos ayuda a agrupar palabras a su origen es decir a un lema que reduce el gran volumen de vocabulario.

“Corpus o tamaño de diccionario es el total de palabras que estan en la base de datos de una red de recursos didacticos para lo cual se plantea que es necesario tener un administrador capacitado para su conservacion y para tener la totalidad de la informacion” (Gelbukh y sodorov 2010).

En el estudio se llega a la conclución de que la lingüística computacional es un área de la informática que está en crecimiento, la lingüística computacional permite minimizar el tiempo en obtener información relevante en los buscadores de la red, minimizar la busqueda por palabras relacionadas sintágmica y paradigmáticamente (que sigue el modelo planteado de manera normal).

2.3.3. **“La Conquista de Jerusalén, Cervantes y la generación teatral de 1580**”. Revista de crítica literaria CRITICON 1992.

Universidad de pisa por Stefano ARATA. En 1990 el hispanista Stefano Arata (Italia1959-2001) catedrático del departamento de filología universidad Sapienza Roma. El cual publicó un estudio textual de un manuscrito de la obra teatral la “CONQUISTA DE JERUSALÉN POR GODEFRE DE BULLÓN”, escrito que fue encontrado en la biblioteca del palacio junto con otras obras también manuscritas que fueron dejadas por los actores de la época. En su estudio preliminar Arata pretende haber encontrado a “LA JERUSALEM” perdida de Cervantes.

Posteriormente en su analisis crítico, Stefano de Arata en la universidad de pizza argumenta que la razón por la que las obras de cervantes en esa época se perdieron fueron a causa de que le mayoría de las obras eran manuscritas y no llegaron a publicarse.

Stefano de arata, nació en Sicilia, Italia en 1959, tuvo una estrecha relación con España desde su infancia, también tuvo vínculos laborales con las universidades Toulouse, en los países de gran Bretaña, el país vasco y Brasil. Fue catedrático de la universidad La Sapienza en Roma, fue considerado como un gran especialista en el estudio de obras teatrales de los años dorados del teatro español entre los que figuran cervantes y otros tantos dramaturgos de esa época en los que destacan Lope de Vega, Francisco de la Cueva, Loyola, Cepeda y muchos otros, hasta el momento de su fallecimiento. Hecho que sucedió de forma inesperada el 21 de julio del 2001; fue un arduo promotor y admirador de cervantes, dándose a notar esto en sus cátedras en la cuales alentaba a sus alumnos a realizar investigación acerca del tema en cuestión.

El estudio realizado por Arata indica que la comedia teatral que lleva como título “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” fue realizada durante las últimas décadas del siglo XVI teniendo como fecha propuesta de 1581-1585. A lo que argumenta que apareció sin un autor y sin fecha; se trata de la adaptación del poema “GERUSALEMME LIBÉRATA” de Torcato Tasso, publicado en Parma -

1581. escrita con anterioridad a la misma traducción de Sedeño. Acerca del manuscrito inédito de la obra teatral Arata argumenta:

Que en “LA GRAN TURQUESCA” miguel de cervantes comenta las obras que hizo en su juventud pero que no llegaron a publicarse entre ellas “EL TRATO DE ARGEL”, “LA NUMANCIA”, “LA JERUSALEM”, y muchas otras obras, puesto que el trato de argel y Numancia se recuperaron sus manuscritos, la obra “LA GRAN TURQUESCA” fue refundida y cambiada de nombre a “LA GRAN SULTANA”, pero “LA JERUSALEM” parecía estar perdida. A lo cual arata comenta “Pero la coincidencia de título y de fecha entre nuestra comedia y la obra perdida de Cervantes representa, creo, un punto de partida suficientemente sólido para autorizar un análisis orientado a determinar su posible paternidad” (Arata, 1992)

Dado lo innovador del teatro cervantino, alejado del modelo lopesco sus obras han conservado su estilo lo cual facilita su estudio. El cual compara las tres obras que se expone en el presente trabajo de tesis “LA CONQUISTA DE JERUSALÉN” “EL TRATO DE ARGEL” Y “NUMANCIA”. Cervantes no tuvo una escuela e imitadores en esas épocas, lo cual permite hacer conjeturas más claras acerca de la autoría de la obra mediante un análisis comparativo.

El “TRATO DE ARGEL” y “NUMANCIA”, que fueron conservadas debidamente ya que se realizó de manera preliminar copias de las obras bajo la autoría de Cervantes, lo cual permite compararla con “LA CONQUISTA DE JERUSALÉN”, obra que es atribuida a Cervantes.

En esta comparación hecha en la crítica cito: “Las tres obras presentan un alto porcentaje de metros de origen italiano: 48,3% El Trato; 75,7% La Numancia; 64,1% La Jerusalén. En los tres casos son idénticas las estrofas italianas utilizadas, a saber: octavas reales, tercetos encadenados, endecasílabos sueltos. Pese a que esta identidad es llamativa, hay que advertir que un alto porcentaje de metros italianos y la combinación de sueltos, tercetos y octavas no es una fórmula exclusiva de la métrica cervantina, sino que puede encontrarse en otros dramaturgos de los primeros años 80”. “En lo que se refiere a los versos españoles, las tres obras revelan una evidente preferencia por la redondilla (42,7% en El Trato; 24,2% en La

Numancia; 28,1% en La Jerusalén) en perjuicio de la quintilla, que aparece de forma esporádica en El Trato (7,1%) pero que está ausente en La Numancia y en La Jerusalén." (Arata, 1992) Los párrafos anteriores dan a notar un acercamiento numérico a el análisis de los tres textos citados dando a entender que este enfoque no es muy lejano de la crítica literaria.

Otro argumento que plantea es:

"Como era de esperar, el porcentaje de las estrofas no nos proporciona por sí sólo elementos suficientes para rechazar o aceptar la autoría cervantina. Sin embargo, la amplia coincidencia de estructuras métricas entre las tres obras nos autoriza a seguir interrogando otras facetas del estilo cervantino. Respecto al panorama teatral contemporáneo, llegando a inferirse que Cervantes no tenía familiaridad con las limitaciones reales que imponía una puesta en escena. En efecto, El Trato con sus 38 personajes y La Numancia con sus más de 40 figuras hacen alarde de un reparto de dimensiones nada comunes en la época. La Jerusalén, con sus 31 figuras, encaja cómodamente en la estructura multitudinaria del teatro cervantino". (Arata, 1992)

En el artículo de Riley mencionado por Arata, se habla de la innovación en el uso de personajes alegóricos en la obra "EL TRATO DE ARGEL" como Venus, Fama, Envidia que exteriorizan el sentimiento característico de cada personaje, esta técnica era novedosa para la época. Concerniente a "LA JERUSALÉN", aparecen aquí como figuras alegóricas Jerusalén, el Contento, la Esperanza, la Libertad y el Trabajo. Arata hace un exhaustivo estudio de comparación con las demás obras manuscritas e impresas de la época y también de otras épocas lo que le lleva a una conclusión en la que puede decir que hay características propias de la escuela cervantina.

En sus últimas líneas del estudio critico Arata da inicio a la problemática de paternidad de la obra teatral "LA CONQUISTA DE JERUSALÉN" para posterior análisis, "Que es al mismo tiempo un problema interpretativo, y ofrecer los primeros datos para una discusión. Esperemos, ahora, que esa desafortunada generación

teatral de 1580 pueda recuperar pronto, y con pleno derecho, una de sus obras perdidas".

2.3.4. "La Conquista de Jerusalén en su contexto: sobre el personaje colectivo y una vuelta más a la atribución cervantina" 2014. Juan Cerezo Soler, en la Universidad Autónoma de Madrid.

Propone un acercamiento a esta relación desde una perspectiva menos objetiva, centrándose en aspectos temáticos y argumentales que tienen en común las obras del primer teatro cervantino.

El filólogo define a la obra "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLÓN" como un ejemplar raro y digno de atención, escribe acerca del estudio que realizó Arata en los 90, y que hizo un estudio riguroso pero que sólo se basó en la métrica, a la configuración del reparto o al tipo de acotaciones utilizadas; así como escribieron otros estudiosos que reforzaron la hipótesis de Arata, uno de ellos fue Héctor Brioso, que reforzó lo expuesto por arata, del mismo modo, Alfredo Rodríguez López Vázquez y Alfredo Baras Escolá han hecho sus propuestas todas importantes en este sentido. Se les añade, en los últimos años, la aportación de Aaron M. Kahn.

Para quien la defensa de la tesis cervantina parte del análisis del carácter ideológico de "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLÓN" en relación con el enfrentamiento entre la corona española y el islam, y se centra fundamentalmente en las figuras alegóricas que pueblan tanto la obra anónima como las cervantinas, uno de los últimos críticos en dar apoyo a la autoría cervantina es Moisés R. Castillo, quien realiza comparaciones de aspectos temáticos, dramáticos, e ideológicos; donde las debilidades humanas de cristianos y de musulmanes contrastan con el exagerado fervor religioso del que alardean muchos personajes cristianos.

El autor del estudio aduce que la conquista de Jerusalén tiene meras coincidencias formales usadas en el teatro de esa época, pone en ejemplo "LA DESTRUICIÓN DE CONSTANTINOPLA" de lobo Lazo de la Vega escrito también en esas épocas,

El autor del estudio también hace referencia a los personajes en comparación con la vida que pasó Cervantes, aduce que como él vivió en cautiverio por algunos años exterioriza ese sentimiento en los personajes que son cautivos tanto física como sentimentalmente, en sus obras Cervantes crea supuestos destinos desalentadores para los personajes, y los personajes luchan contra ese destino y al final cambia la historia de la obra.

En “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” no solo puede distinguirse la presencia de una colectividad, sino que puede verse que es una colectividad con aires bastante cervantinos. “LA NUMANCIA” muestra la ejemplaridad del colectivo que se inmola en una lección destinada a despertar el orgullo nacionalista. (Cerezo) “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” hace exactamente lo mismo sustituyendo esa lección nacionalista por la lección religiosa y el colectivo político-nacional por el colectivo religioso; y desde luego, el hecho de que sea el recurso alegórico el que sirva para la actualización del motivo histórico vincula una vez más el nombre de Cervantes con esta obra anónima.

Como punto final, en su estudio da una conclusión provisional de estas páginas, para estudios posteriores de este tipo pueda argumentarse, estará sujeto al hallazgo de pruebas documentales más concretas que afirmen o que desmientan la autoría de “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” y afirma que es aconsejable mantener prudencia en la atribución y evitar la afirmación categórica, lo que no impide, eso sí, que se siga estudiando esta obra como una muestra extraña y preciosa del período anterior a Lope de Vega. Un período en el que Miguel de Cervantes, entre tantos otros, disfrutó del éxito y del aplauso del público español, y así lo afirmará, muchos años después.

2.3.5. “Mezclar Verdades Con Fabulosos Intentos: Meta teatro Y Aporía En El Gallardo Español De Cervantes”. 2004 Lourdes Albuisech en la Southern Illinois University,

En su estudio describe que Cervantes fue bueno para la prosa mas no para verso que todas las obras que escribió fueron hechas a partir del 1615 para adelante y que la excesiva fe en sí mismo hizo que no reconociera sus errores en teatro a lo que se denomina “meta teatral” (conjunto de obras en las épocas de la reina Elisabeth que en su mayoría son obras teatrales auto reflexivas. Lionel Abel).

El objetivo principal del meta drama es ilusionar que el drama vivido en el teatro puede ser real, la definición de comedia se dio por Cicerón y Livio Andromico, que significa el espejo de la vida, al menos en la época de oro del teatro español, y muchos estudios literarios solo se enfocan en “EL QUIJOTE” ya que es su obra cumbre, sin darle mucha importancia a los entremeses dramáticos, en el estudio se comenta que las obras de cautiverio no son en lo más mínimo cercanas a la realidad.

El estudio indica que en las obras de cervantes realiza un buen entretenimiento, pero a la vez hace dar cuenta al público que las puestas teatrales no son más que mera ilusión, Cervantes soluciona el conflicto entre la gallardía y lo religioso, en sus obras Cervantes deja un final inconcluso y deja al espectador un sinsabor de no visto u oído la meta trazada al inicio de la obra.

En la época de los años dorados del teatro español en que el hombre pierde la fe en sí mismo, en la naturaleza y en la realidad que le rodea. Si bien las comedias escritas por Cervantes en un primer período (aproximadamente hasta 1582) fueron bien vistas por el público, una vez triunfó la comedia lopesca, la fórmula cervantina iba forzosamente a fracasar. (Albuisech)

En su estudio crítico hace referencia al estudio de cervantes que directamente a la interacción con sus obras siendo un gran aporte su conocimiento acerca de Cervantes, pero no así dando una opinión de la autoría de “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”

III. HIPÓTESIS Y VARIABLES.

3.1. Hipótesis.

a. Hipótesis general.

Existe asociación entre las obras “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”, “LA NUMANCIA” y “EL TRATO DE ARGEL” obtenidas mediante el análisis de correspondencia.

b. Hipótesis específica:

- Existe similitud en el tamaño de palabras en las obras “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”, “LA NUMANCIA” y “EL TRATO DE ARGEL”.
- Existe semejanza de personajes en las obras “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”, “LA NUMANCIA” y “EL TRATO DE ARGEL”.
- Existe semejanza de léxico entre las obras “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”, “LA NUMANCIA” y “EL TRATO DE ARGEL”.
- Existe semejanza entre las matrices de transición precede y antecede de las obras” LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”, “LA NUMANCIA” y “EL TRATO DE ARGEL”

3.2. Identificación de variables e indicadores.

La variable del estudio está definida por el perfil de las obras “EL TRATO DE ARGEL”, “LA NUMANCIA” y “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”. el perfil es la representación del corpus textual formado por las tres obras en estudio

Donde las dimensiones quedan definidas por:

Léxico: Conjunto de palabras y forma de uso determinado por un grupo afectado por rasgos lingüísticos, sociales, geográficos y culturales variantes. refleja el medio físico y social

Obras: obras teatrales en estudio

“EL TRATO DE ARGEL” y “LA NUMANCIA”: Obras de teatro escrita por Cervantes. Las cuáles serán usadas como punto comparativo

“LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”. Obra de teatro sin autor definido el cual se asume la autoría de Cervantes.

Personaje: Personajes usados en las obras de teatro del estudio.

Longitud de palabra: Amplitud de cada palabra usada en las obras en estudio siendo de expresión cuantitativa discreta.

Matriz transición por obra: Matriz conformada por la expresión del perfil de la palabra antecedente y el perfil de la palabra precedente distinguiendo en las tres obras en estudio. La matriz permite recaudar información de la manera de uso del vocabulario del autor

3.3. Matriz de Operacionalización.

VARIABLE	DEFINICIÓN CONCEPTUAL	DIMENSIÓN	NATURALEZA	ESCALA DE MEDICIÓN	FORMA DE MEDICION	INDICADORES	DEFINICION OPERACIONAL
Perfil de las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "EL TRATO DE ARGEL" y "LA NUMANCIA" (corpus textual)	Conjunto de palabras y forma de uso expresada de manera tabular	Léxico	Cualitativo	Nominal categórico	Conteo en el corpus textual.	Cantidad de palabras en el vocabulario del escritor.	vocabulario usado (y, como, cuando, …)
		Obra	cualitativo	nominal categórico	Directo	Obras en estudio.	- "LA CONQUISTA DE JERUSALÉN" - "EL TRATO DE ARGEL" - "LA NUMANCIA":
		Personajes	Cualitativo	Nominal categórico	Directo en cada obra.	Conteo de correlaciones con el paquete estadístico.	personajes (Morando, Jerusalén, Virginia…)
		Longitud palabra	Cuantitativa	Numérica discreta	Directo conteo de repeticiones en el texto	conteo de caracteres usados en cada palabra en las obras.	de 1 a 15
		Matriz de transición	Cuantitativa	Numérica	Mediante tabla de contingencia de palabra antecede precede.	Matriz de correlación para cada obra (matriz de transición estocástica perfil por fila)	$M_{(Jerusalem)}$ $M_{(Numancia)}$ $M_{(Argel)}$

Tabla 8.Operacionalizacion de variables. (de creación propia)

IV. METODOLOGÍA.

4.1. Tipo y nivel de investigación.

La metodología de la investigación es de tipo descriptivo correlacional, se basará en la revisión bibliográfica y documental sobre técnicas de minería de texto, la cual permitirá elaborar los fundamentos y valorar experiencias del diagnóstico. El estudio descriptivo pretende medir y recoger información de manera independiente y conjunta sobre el corpus textual de las variables en estudio. Por su característica temporal se clasifica como transversal. La investigación se centra en analizar los niveles y relaciones de una o diversas variables en un momento determinado o punto del tiempo.

4.2. Unidad de análisis.

Las unidades de estudio en esta etapa son las palabras que se usa en el escrito siendo característica del autor la cual forma su vocabulario que en el aglomerado de las obras es llamado corpus textual el cual será resumido en tabla adjunta perfil y matriz de correlación.

4.3. Población de estudio.

La población o universo considerado son las palabras usadas por el autor en cada uno de sus escritos.

4.4. Tamaño de muestra.

Se tomará como muestra la totalidad del corpus textual formado por las tres obras en estudio la muestra será de tipo censal.

4.5. Técnicas de selección y recolección de muestra.

Para encontrar semejanzas entre el texto atribuido: “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” y las obras teatrales contemporáneas “NUMANCIA” y “TRATO DE ARGEL” se recolecta estas obras de la Biblioteca Virtual Miguel de Cervantes, fuente fidedigna. En formato HTML de texto. Datos que permitirán procedimientos estadísticos del análisis de correspondencia en el corpus literario con el uso de un paquete de ofimática OFFICE 2016 y paquete estadístico XLSTAT.

4.6. Plan de trabajo de investigación.

a. Reconexión de información. Con los criterios de exclusión expresados anteriormente se procederá a la recolección de los títulos mencionados de la biblioteca virtual cervantes Saavedra. En formato de HTML de texto con el que trabajaremos, lo cual se procedió a su uso debido a que las normas de derechos intelectuales sobre éstos ya son de dominio público lo cual permite el análisis y estudio sin necesidad de permisos del autor.

b. Limpieza de información. La información recaudada será sometida a la limpieza en el editor de texto quitando caracteres que son de poca utilidad, en minería de texto y análisis estadístico se empieza por asegurar que no queden caracteres especiales de la codificación como son salto de línea, tabulación, tildes, signos de puntuación, se convierten todas las palabras a minúsculas lo cual permite tener una menor cantidad de categorías de estudio, esto permite crear un texto analizable conservando en su mayoría el texto original.

c. Base de datos. Se procederá a importar el texto a la hoja de cálculo en donde se planteará la creación del corpus textual, es decir la creación de un texto que a la vez es una base de datos, para esto se vacía cada palabra, una concatenada a la otra la que se denomina “palabra antecede” y “palabra precede”, lo cual se asemeja a la estructura requerida para el análisis de proceso estocástico.

La base de datos conserva también la obra el personaje de donde se origina la palabra y su respectiva posición en la obra y párrafo, por el gran número de palabras usadas en un vocabulario normal la minería de texto plantea el uso de la técnica de lematizado, que aglomera palabras de significado semejante o con la misma raíz lingüística, proceso que se usa únicamente en las palabras cuya frecuencia en toda la obra es inferior a tres repeticiones, esto con el fin de preservar el mensaje que quiso transmitir el autor en la obra y a la vez tener una base de datos analizable.

La hoja de cálculo permitirá el adecuado ingreso al paquete estadístico de análisis XLSTAT.

D. Técnicas de procesamiento de datos. Análisis mediante paquete estadístico.

Dentro del análisis del corpus se plantean varias técnicas a usar en dos etapas:

D. A. Corpus sin lematizado.

Se analizará la obra sin el redondeo por lematizado en todo el conglomerado de las obras y diferenciando por título, en esta parte utilizaremos la tabla de frecuencias para saber la representación de cada obra, jornada en la base de datos total también se hará el estudio de la amplitud de las palabras por obra siendo esta una característica importante del autor y del lenguaje usado.

Se hará uso de la tabla de frecuencias para determinar la frecuencia de las palabras en toda la obra, en caso de ser muy bajo el uso de la palabra será la razón por que es necesario usar la técnica de lematizado.

D. B. Corpus lematizado.

D. B. A. En el estudio descriptivo se hará uso de la tabla de frecuencia para las palabras por cada obra y se originará el grafico de nube de palabras. Técnica de mapeo del corpus o base de datos en busqueda de la palabra con más frecuencia. También se distinguirá las palabras con significado completo de artículos adverbios que son los más usados.

D. B. B. En el estudio correlacional se hará uso de las técnicas de tabla de contingencia y prueba de independencia chi-cuadrado, análisis de correspondencia. Estos por personaje y obra.

En el estudio correlacional Inter obra se hará el uso de conceptos estadística estocástica considerando palabra precedente y antecedente en las obras, lo cual permitirá generar una matriz de transición y comparar estas matrices de transición mediante una modificación de la prueba de homogeneidad de varianzas multivariado a la cual nombraremos como prueba de homogeneidad de matrices de transición estocástica, lo cual permitirá demostrar la hipótesis planteada.

La representación gráfica del diseño de investigación descriptivo correlacional es la siguiente:

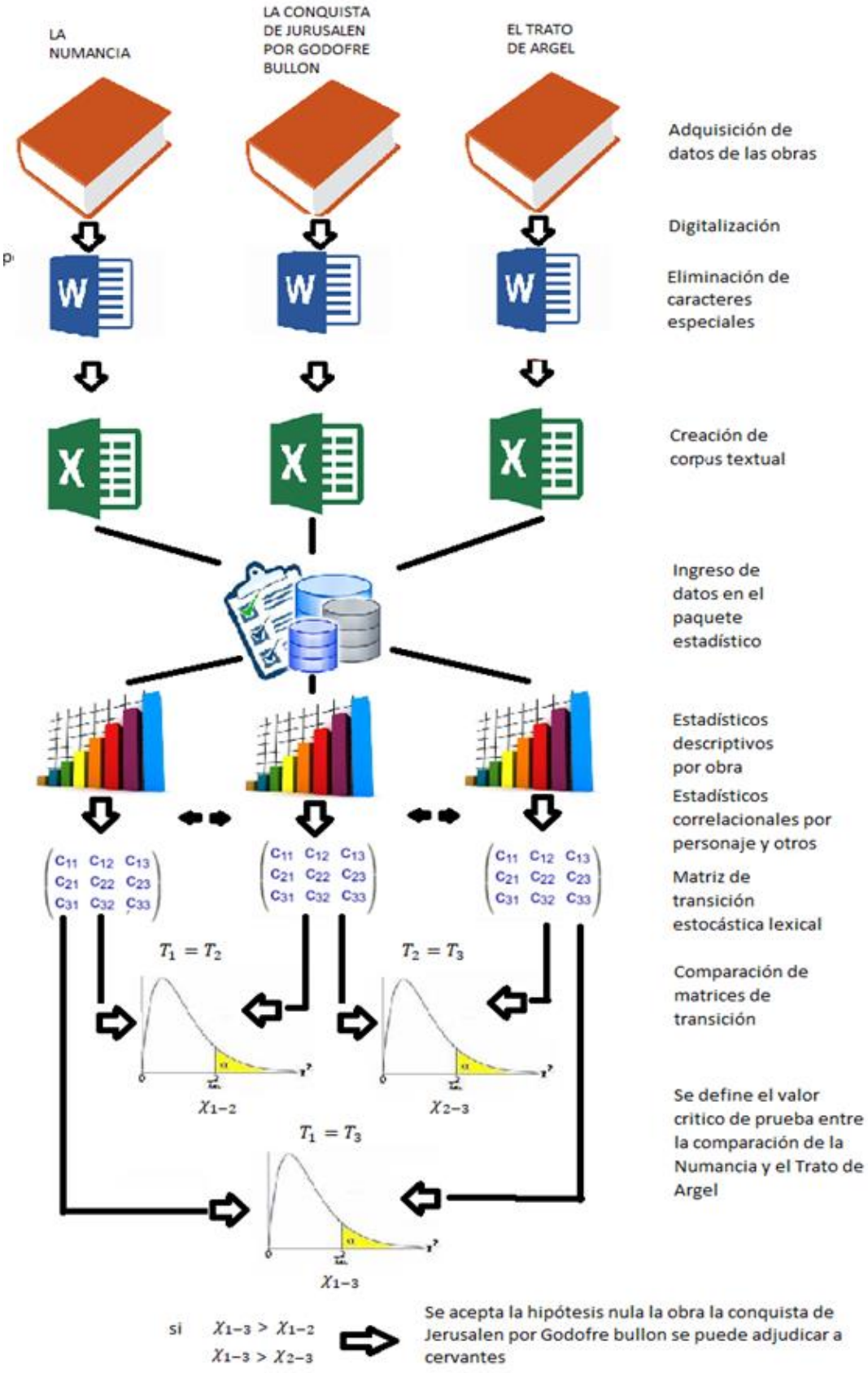


Ilustración 7. Metodología. (de creación propia)

V. ANÁLISIS Y RESULTADOS.

5.1. Resultados descriptivos.

Criterio para facilitar la interpretación nos haremos referencia en los cuadros a “LA CONQUISTA DE JERUSALEM POR GODOFRE BULLON” como “JERUSALEM”, “LA NUMANCIA” como “NUMANCIA” y a “EL TRATO DE ARGEL” como “ARGEL” esto para comodidad en los cuadros.

5.1.1. Descriptivo general del estudio.

El tamaño de muestra es de 47490 palabras en las tres obras de teatro analizadas encontrando la siguiente distribución se presenta la distribución por obra de teatro por partes, también llamado capítulos (jornadas) de obras.

Obra	n	%	líneas	Jornada	n	%
ARGEL	15047	32%	496	jornada i	3861	8%
				jornada ii	4455	9%
				jornada iii	2924	6%
				jornada iv	3807	8%
JERUSALEM	16684	35%	526	jornada i	4003	8%
				jornada ii	4372	9%
				jornada iii	8309	17%
NUMANCIA	15759	33%	313	jornada i	3550	7%
				jornada ii	3866	8%
				jornada iii	3661	8%
				jornada iv	4682	10%
Total, de palabras	47490	100%				

Tabla 9.Frecuencia de palabras por obra. (de creación propia)

Interpretación:

- Se aprecia que la cantidad de líneas usadas en totalidad en todas las personas por obra no difieren de gran manera. Siendo el menor “LA NUMANCIA” de 313 párrafos y el más amplio de “LA COMQUISTA DE JERUSALEN POR GODOFRE BULLON” de 526 párrafos
- La obra que presenta mayor uso de palabras es Jerusalén con 16684, siendo la obra más extensa, la obra más corta resulta ser Argel con un uso de 15047

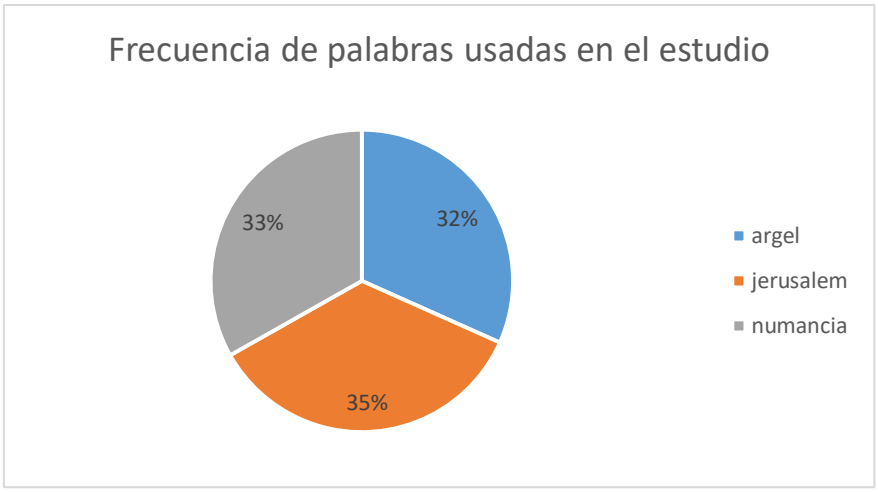


Ilustración 8.Frecuencia de palabras por obra en el estudio. (de creación propia)

- Se aprecia que el porcentaje de la base de datos de palabras se encuentran distribuida para “EL TRATO DE ARGEL” de 32%, “LA CONQUISTA DE JERUSALEM POR GODOBRE BULLON” de 35% y de “LA NUMANCIA” de 33%, lo cual demuestra que no existe diferencia notable en el tamaño de palabras usadas en cada obra.

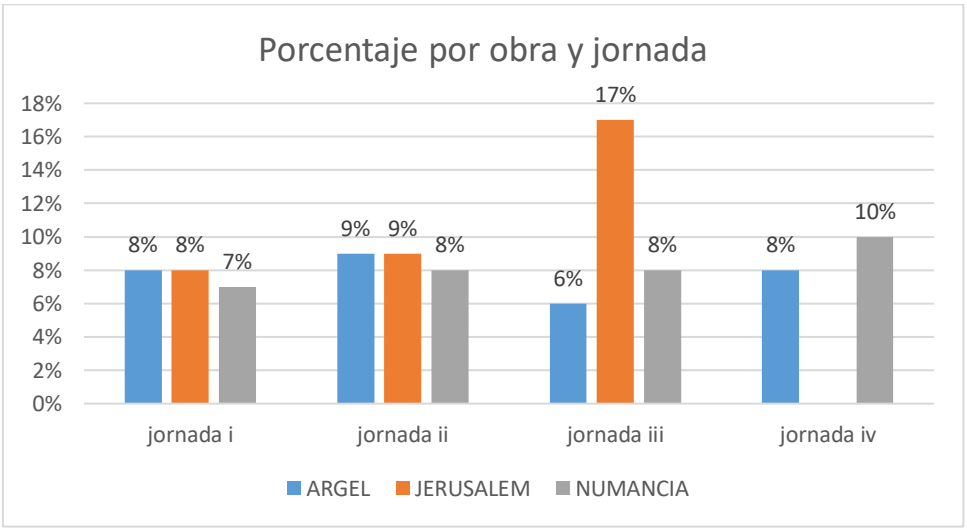


Ilustración 9.Porcentaje por obra y jornada. (de creación propia)

- Se aprecia que el porcentaje por jornada es semejante siendo el menor de 6% la tercera jornada de “EL TRATO DE ARGEL” y mayor de 17% la tercera jornada de “LA CONQUISTA DE JERUSALEM POR GODOFRE BULLON”. Apreciándose que no existe diferencia notable entre los porcentajes de palabras usados por jornadas exceptuando “LA CONQUISTA DE JERUSALEM POR GODOFRE BULLON”. tercera jornada, dato cual se analizará en la discusión.

5.1.2. Descriptivo por ancho de palabra por obra.

La riqueza del vocabulario de un escritor se refleja en la complejidad de las palabras que usa éste representado en la amplitud de caracteres y la frecuencia de uso.

	obras					
	ARGEL		JERUSALEM		NUMANCIA	
ancho	n	%	n	%	n	%
1	1006	6.69%	1170	7.01%	1001	6.35%
2	3743	24.88%	3923	23.52%	3596	22.82%
3	2303	15.31%	2526	15.14%	2395	15.20%
4	1664	11.06%	1857	11.13%	1622	10.29%
5	1934	12.85%	2230	13.37%	2107	13.37%
6	1655	11.00%	1589	9.53%	1765	11.20%
7	1230	8.17%	1398	8.38%	1386	8.79%
8	697	4.63%	1046	6.27%	970	6.16%
9	478	3.18%	550	3.30%	487	3.09%
10	215	1.43%	244	1.46%	263	1.67%
11	76	0.51%	98	0.59%	108	0.69%
12	30	0.20%	39	0.23%	43	0.27%
13	9	0.06%	9	0.05%	9	0.06%
14	5	0.03%	1	0.01%	6	0.04%
15	1	0.01%	2	0.01%	1	0.01%

Tabla 10.Ancho de palabra por obra. (de creación propia)

Prueba de independencia entre las filas y columnas (Chi-cuadrado):

Dado que se ha encontrado frecuencias inferiores a 5. Y que en la prueba de Chi-cuadrado, las frecuencias no deben ser inferiores a 5. se omitirá en la prueba las palabras superiores a 12 caracteres.

H_0: Las filas y las columnas de la tabla son independientes.

H_a: Hay dependencia entre las filas y las columnas de la tabla.

Chi-cuadrado (Valor observado)	Chi-cuadrado (Valor crítico)	GL	valor-p	alfa
113.322	36.415	24	< 0.0001	0.05

Tabla 11.Prueba chi-cuadrado asociación número de caracteres por obra. (de creación propia)

Puesto que el valor-p es menor que el nivel de significación alfa=0.05, se debe rechazar la hipótesis nula Ho, y aceptar la hipótesis alterna Ha.

Lo cual indica que existe asociación entre la frecuencia por ancho de caracteres usado en las palabras y la obra, esta asociación se nota que está dada casi enteramente por la semejanza en la distribución la cual apreciaremos con el gráfico de barras siguiente.

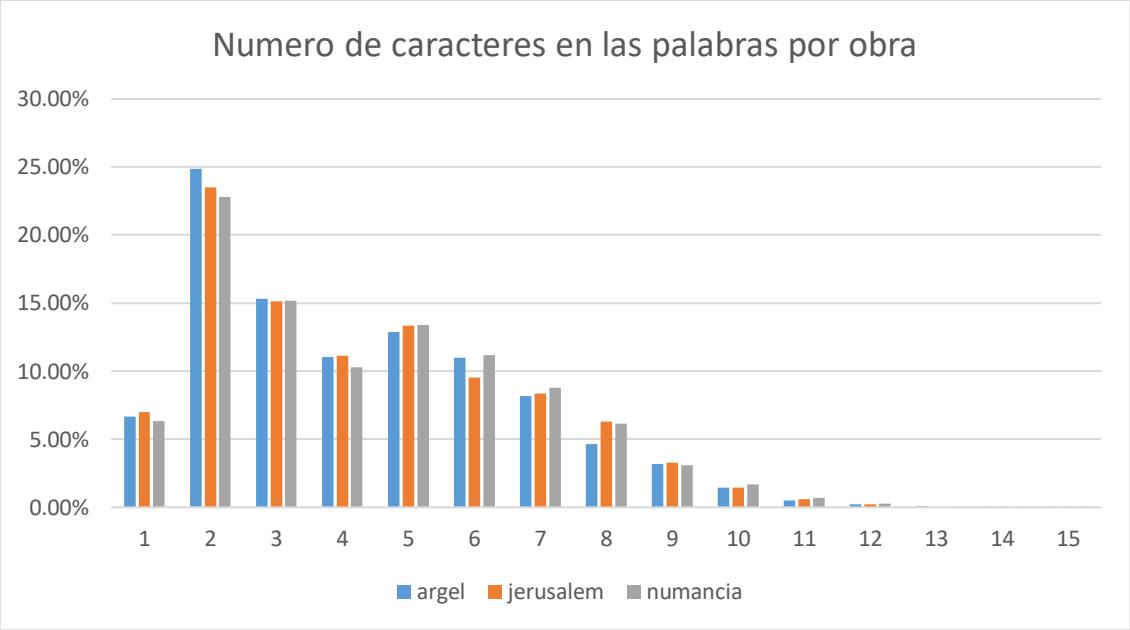


Ilustración 10.Número de caracteres usados en las palabras por obra. (de creación propia)

Interpretación:

- Se aprecia gráficamente la similitud en el vocabulario en la dimensión ancho de palabra en uso de caracteres en las tres obras estudiadas.
- Estas siguen una distribución semejante a la de la distribución Chi-cuadrado en donde el uso de palabras de un carácter es bajo y el mayor es de 2 y desde el cual desciende gradualmente hasta el casi nulo uso de palabras de 15.

5.1.3. Frecuencia de palabras en las obras.

El estudio se centra en el uso de la palabra como variable categórica para este fin se vio forzosamente necesario la simplificación del vocabulario de las obras proceso que se denomina lematización, para lo cual se agrupan las palabras por frecuencia de uso en un límite inferior (Li) y un límite superior (Ls), estas frecuencias representan la aparición de las palabras por obra.

Sin lematizar.

		ARGEL		JERUSALEM		NUMANCIA	
$< L_i, L_s]$		n	%	n	%	n	%
0	1	3528	52%	3395	50%	3420	50%
1	2	2029	30%	2035	30%	2026	30%
2	6	891	13%	984	14%	990	15%
6	15	236	3%	248	4%	239	4%
15	31	81	1%	99	1%	82	1%
31	56	18	0%	22	0%	29	0%
56	92	12	0%	9	0%	12	0%
92	141	9	0%	11	0%	9	0%
141	205	5	0%	5	0%	1	0%
205	286	1	0%	2	0%	3	0%
286	386	3	0%	1	0%	1	0%
386	507	1	0%	3	0%	2	0%
507	651	1	0%	1	0%	1	0%

Tabla 12.Frecuencia palabras sin lematizado. (de creación propia)

Acotados de 0 a 1 que incluyen las palabras de un único uso en todo el texto es muy alto siendo en “EL TRATO DE ARGEL” del 52% “LA CONQUISTA DE JERUSALEM POR GODOFRE BULLON” del 50% y “LA NUMANCIA” del 50%.

Con una cota de 1 a 2 palabras repetidas desciende porcentualmente en un promedio de 30% en las tres obras y la tendencia sigue en decrecimiento.

Con una cota de 92 a 141 indica que las palabras que se repiten de 93 a 141 veces son en cantidad 9 palabras en “EL TRATO DE ARGEL” y “LA NUMANCIA”, de 11 palabras en “LA CONQUISTA DE JERUSALEM POR GODOFRE BULLON” es de 9 palabras y este número solo decae. Lo cual es posible debido

a que Cervantes presentaba una gran capacidad de verbalizar y que en el texto aparecen varios idiomas.

Para realizar el estudio se requiere que exista menor cantidad de niveles de las variables categóricas por lo cual el proceso de lematizado el cual regresa las palabras verbalizadas a su origen o raíz lo cual permite elevar el conteo en las categorías estudiadas. A lo cual la base de datos de los textos fue lematizada tratando de conservar la mayor riqueza lingüística.

Lematizado

		ARGEL		JERUSALEM		NUMANCIA	
$< L_i, L_s]$		n	%	n	%	n	%
0	1	1068	33%	1069	33%	1124	34%
1	2	835	26%	726	22%	689	21%
2	6	904	28%	962	29%	964	30%
6	15	309	9%	341	10%	324	10%
15	31	94	3%	106	3%	100	3%
31	56	21	1%	27	1%	33	1%
56	92	12	0%	9	0%	12	0%
92	141	9	0%	11	0%	9	0%
141	205	5	0%	5	0%	1	0%
205	286	1	0%	2	0%	3	0%
286	386	3	0%	1	0%	1	0%
386	507	1	0%	3	0%	2	0%
507	651	1	0%	1	0%	1	0%

Tabla 13.Frecuencia palabras lematizadas. (de creación propia)

- El porcentaje que era de 50% en promedio en las tres obras sin lematizado de palabra usadas solo una vez después del lematizado de las obras decae a un 33% en “EL TRATO DE ARGEL” y “LA CONQUISTA DE JURUSALEN POR GODOFRE BULLON” y un 34% en “LA NUMANCIA.”
- De la misma manera en el resto de frecuencias este proceso permite tener menos categorías en estudio, desde este punto el estudio hará uso del texto lematizado.

5.1.4. Descriptivos “EL TRATO DE ARGEL”.

Frecuencia de palabras usadas en la obra “EL TRATO DE ARGEL”. En la siguiente tabla se muestran sólo las 50 palabras más importantes usadas en el vocabulario de la obra.

		ARGEL					ARGEL		
id	**palabra**	**n**	**%**	**importancia**	**id**	**palabra**	**n**	**%**	**importancia**
1	que	742	4.93%		26	del	73	0.49%	
2	y	665	4.42%		27	un	85	0.56%	
3	de	514	3.42%		28	pues	50	0.33%	
4	el	400	2.66%		29	ya	60	0.40%	
5	la	317	2.11%		30	ha	77	0.51%	
6	en	324	2.15%		31	tan	61	0.41%	
7	a	301	2.00%		32	bien	67	0.45%	si
8	no	265	1.76%		33	como	65	0.43%	
9	con	138	0.92%		34	este	51	0.34%	
10	se	176	1.17%		35	quien	47	0.31%	
11	mi	159	1.06%		36	le	74	0.49%	
12	por	145	0.96%		37	sin	55	0.37%	si
13	es	191	1.27%		38	una	39	0.26%	
14	si	130	0.86%		39	o	37	0.25%	
15	me	160	1.06%		40	ser	40	0.27%	
16	tu	104	0.69%		41	aquí	38	0.25%	
17	mas	117	0.78%		42	para	40	0.27%	
18	los	102	0.68%		43	hacer	23	0.15%	
19	al	106	0.70%		44	oh	39	0.26%	
20	lo	99	0.66%		45	vida	29	0.19%	si
21	esta	96	0.64%		46	cielo	48	0.32%	si
22	yo	79	0.53%		47	dar	43	0.29%	
23	su	107	0.71%		48	muerte	27	0.18%	si
24	las	59	0.39%		49	todo	28	0.19%	si
25	te	84	0.56%		50	ver	31	0.21%	si

Tabla 14.Frecuencia de palabras usadas en la obra ARGEL. (de creación propia)

- La palabra “que” en la obra “ELTRATO DE ARGEL” aparece 742 veces representado un 4.93% del texto representando ser la palabra más usada seguida de “y” que aparece 665 veces y representa un 4.42% del texto.

Frecuencia de las palabras importantes.

No todas las palabras que presentan gran frecuencia tienen sentido completo o trasmiten un mensaje por lo cual se separaron las palabras que presenta importancia de los artículos y conjugaciones más comunes esto con el fin de encontrar cual es el mensaje a transmitir con la frecuencia de palabras usado en la obra “EL TRATO DE ARGEL”. Se presentan las 20 primeras palabras de mayor frecuencia y de gran importancia.

id	palabra	n	%	importancia	id	palabra	n	%	importancia
32	bien	67	0.45%	si	164	señora	28	0.19%	si
37	sin	55	0.37%	si	48	muerte	27	0.18%	si
62	cristiano	52	0.35%	si	55	mal	27	0.18%	si
46	cielo	48	0.32%	si	67	amor	27	0.18%	si
81	dos	34	0.23%	si	88	alma	27	0.18%	si
53	señor	33	0.22%	si	98	mil	27	0.18%	si
50	ver	31	0.21%	si	87	rey	26	0.17%	si
64	dios	31	0.21%	si	94	tierra	26	0.17%	si
45	vida	29	0.19%	si	58	poder	25	0.17%	si
49	todo	28	0.19%	si	73	pecho	24	0.16%	si

Tabla 15.Palabras importantes Argel. (de creación propia)

- Encontramos entre las palabras de gran importancia en el texto “EL TRATO DE ARGEL” es “bien” representando en 0.45% que se repite 67 veces en el mismo.
- Para dar una mejor interpretación a las frecuencias encontradas se presentarán en el gráfico nube de palabras.

Nube de palabras “EL TRATO DE ARGEL”.

Para el desarrollo de el grafico de nube de palabras se usó los conteos superiores e igual a 20 en toda la obra de “EL TRATO DE ARGEL” presentando 200 de las primeras palabras con importancia y significado completo.

Ilustración 11.Nube de palabras Argel. (de creación propia)

- Se aprecia la importancia de la palabra “bien” representativa de la obra “EL TRATO DE ARGEL “seguido de “cielo” “cristiano” y “dios”.

Matriz antecede precede “EL TRATO DE ARGEL”

		Palabra precedente												
		que	**y**	**de**	**el**	**la**	**en**	**a**	**no**	**es**	**se**	**me**	**mi**	**…**	**suma**
Palabra antecede	**que**	1	1	6	32	12	31	25	47	51	26	23	7	…	734
	y	16	0	20	25	23	24	13	14	9	3	3	5	…	651
	de	5	0	0	0	29	1	0	4	0	0	0	29	…	504
	el	11	0	1	0	0	0	1	3	0	2	1	0	…	388
	la	4	0	3	0	0	0	0	1	0	0	0	0	…	311
	en	7	0	0	41	29	0	0	0	0	0	0	15	…	306
	a	5	0	0	2	19	0	0	0	0	0	0	22	…	297
	no	0	0	4	2	3	2	1	2	11	18	17	1	…	245
	es	4	0	7	4	13	0	0	0	0	0	0	7	…	186
	se	12	0	0	0	1	0	1	0	0	0	9	0	…	176
	me	0	0	1	0	1	0	0	0	1	0	0	0	…	160
	mi	2	0	1	1	1	0	1	0	0	1	3	0	…	158
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	
	suma	734	651	504	388	311	306	297	245	186	176	160	158		

Tabla 16.Matriz antecede precede Argel. (de creación propia)

- Se aprecia que en la diagonal de la matriz en la mayoría de casos son cero esto debido a que las palabras de uso repetido no son muy comunes.

Matriz transición estocástica “EL TRATO DE ARGEL”.

		Palabra precedente											
		que	**y**	**de**	**el**	**la**	**en**	**a**	**no**	**es**	**se**	**…**	**suma**
Palabra antecede	**que**	0.14%	0.14%	0.82%	4.36%	1.63%	4.22%	3.41%	6.40%	6.95%	3.54%	…	100%
	y	2.46%	0.00%	3.07%	3.84%	3.53%	3.69%	2.00%	2.15%	1.38%	0.46%	…	100%
	de	0.99%	0.00%	0.00%	0.00%	5.75%	0.20%	0.00%	0.79%	0.00%	0.00%	…	100%
	el	2.84%	0.00%	0.26%	0.00%	0.00%	0.00%	0.26%	0.77%	0.00%	0.52%	…	100%
	la	1.29%	0.00%	0.96%	0.00%	0.00%	0.00%	0.00%	0.32%	0.00%	0.00%	…	100%
	en	2.29%	0.00%	0.00%	13.40%	9.48%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	a	1.68%	0.00%	0.00%	0.67%	6.40%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	no	0.00%	0.00%	1.63%	0.82%	1.22%	0.82%	0.41%	0.82%	4.49%	7.35%	…	100%
	es	2.15%	0.00%	3.76%	2.15%	6.99%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	se	6.82%	0.00%	0.00%	0.00%	0.57%	0.00%	0.57%	0.00%	0.00%	0.00%	…	100%
	me	0.00%	0.00%	0.63%	0.00%	0.63%	0.00%	0.00%	0.00%	0.63%	0.00%	…	100%
	mi	1.27%	0.00%	0.63%	0.63%	0.63%	0.00%	0.63%	0.00%	0.00%	0.63%	…	100%
	…	…	…	…	…	…	…	…	…	…	…	…	

Tabla 17.Matriz transición Argel. (de creación propia)

- La matriz antecede precede permite ver como se relaciona las palabras usadas por Cervantes en la obra “EL TRATO DE ARGEL” encontrando que se ha usado 32 veces “que el” la matriz completa incluye todo el vocabulario usado siendo una forma de representativa del libro completo conservando el característico uso del vocabulario por cervantes.
- También se puede hacer mención de otras combinaciones de palabras muy usadas. “a mi” usado 20 veces “de mi” usado 29 veces
- Se aprecia que en la diagonal de la matriz en su mayoría se encuentra muy pocos casos esto debido a que en la lengua española no es muy usual la repetición de palabras consecutivas.
- La matriz de transición permite ver que el lenguaje se comporta como un proceso estocástico permitiendo saber la posibilidad de que al usar una palabra la siguiente sea otra.

5.1.5. Descriptivos **"LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON".**

Frecuencia de palabras usadas en la obra "LA CONQUISTA DE JERUSALEN POR GODOBRE BULLON". En la siguiente tabla se muestra solo las 50 palabras más importantes usadas en el vocabulario de la obra.

		JERUSALEN					JERUSALEN		
id	**palabra**	**n**	**%**	**importancia**	**id**	**palabra**	**n**	**%**	**importancia**
1	que	772	4.63%		26	del	84	0.50%	
2	y	794	4.76%		27	un	67	0.40%	
3	de	552	3.31%		28	pues	101	0.61%	
4	el	416	2.49%		29	ya	83	0.50%	
5	la	433	2.60%		30	ha	59	0.35%	
6	en	386	2.31%		31	tan	77	0.46%	
7	a	341	2.04%		32	bien	61	0.37%	si
8	no	254	1.52%		33	como	48	0.29%	
9	con	208	1.25%		34	este	53	0.32%	
10	se	150	0.90%		35	quien	62	0.37%	
11	mi	164	0.98%		36	le	44	0.26%	
12	por	167	1.00%		37	sin	54	0.32%	si
13	es	125	0.75%		38	una	48	0.29%	
14	si	159	0.95%		39	o	35	0.21%	
15	me	130	0.78%		40	ser	49	0.29%	
16	tu	178	1.07%		41	aquí	42	0.25%	
17	mas	112	0.67%		42	para	41	0.25%	
18	los	120	0.72%		43	hacer	44	0.26%	
19	al	105	0.63%		44	oh	44	0.26%	
20	lo	111	0.67%		45	vida	30	0.18%	si
21	esta	102	0.61%		46	cielo	34	0.20%	si
22	yo	108	0.65%		47	dar	28	0.17%	
23	su	72	0.43%		48	muerte	27	0.16%	si
24	las	101	0.61%		49	todo	33	0.20%	si
25	te	105	0.63%		50	ver	36	0.22%	si

Tabla 18.Frecuencia de palabras usadas en la obra Jerusalem. (de creación propia)

- La palabra "que" en la obra "LA CONQUISTA DE JURUSALEN POR GODOFRE BULLON" aparece 772 veces representado un 4.63% del texto representando ser la palabra más usada seguida de "y" que aparece 794 veces y representa un 4.76% del texto

Tabla de frecuencia palabras importantes.

No todas las palabras que presenta gran frecuencia tienen sentido completo no trasmiten un mensaje por lo cual se separaron las palabras que presentan importancia en los artículos y conjugaciones más comunes esto con el fin de encontrar cual es el mensaje a transmitir con la frecuencia de palabras usada en la obra “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON”. Se presentan las 20 primeras palabras de mayor frecuencia y de gran importancia.

Id	Palabra	n	%	Importancia	Id	Palabra	n	%	Importancia
32	bien	61	0.37%	si	72	ella	33	0.20%	si
37	sin	54	0.32%	si	55	mal	32	0.19%	si
84	ciudad	43	0.26%	si	67	amor	32	0.19%	si
64	dios	41	0.25%	si	45	vida	30	0.18%	si
71	decir	40	0.24%	si	62	cristiano	30	0.18%	si
53	señor	39	0.23%	si	83	todos	29	0.17%	si
50	ver	36	0.22%	si	87	rey	29	0.17%	si
46	cielo	34	0.20%	si	104	gente	29	0.17%	si
49	todo	33	0.20%	si	122	gran	28	0.17%	si
58	poder	33	0.20%	si	48	muerte	27	0.16%	si

Tabla 19.Palabras importantes Jerusalem. (de creación propia)

- Encontramos entre las palabras de gran importancia en el texto de “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” que es “bien” representando un 0.37% que se repitió 61 veces en el mismo.
- Para dar una mejor interpretación a las frecuencias encontradas se presentarán en el gráfico nube de palabras.

Nube de palabras “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON”.

Para el desarrollo de el grafico de nube de palabras se usó los conteos superiores e iguales a 20 en toda la obra “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” presentando 200 de las primeras palabras con importancia y significado completo.

Ilustración 12.Nube de palabras Jerusalem. (de creación propia)

- Se aprecia la importancia de la palabra “bien” representativa de la obra Jerusalem seguido de “sin” “ciudad” y “decir”.

Matriz antecede precede “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON”.

		Palabra precedente													
		y	que	de	la	el	en	a	no	con	tu	mi	por	…	**suma**
Palabra antecede	y	0	21	19	17	27	31	27	15	17	13	7	10	…	789
	que	0	0	16	32	14	36	21	34	12	11	6	5	…	749
	de	0	7	0	41	1	2	0	0	1	19	24	0	…	544
	la	0	7	3	0	0	0	0	0	0	0	0	0	…	431
	el	3	10	1	1	0	0	0	1	1	0	0	1	…	406
	en	0	7	0	45	38	0	0	1	0	14	13	0	…	375
	a	0	4	0	32	2	0	0	2	0	20	27	0	…	327
	no	0	2	1	4	3	4	1	0	3	0	0	1	…	220
	con	0	6	0	18	18	0	0	0	0	4	2	0	…	203
	tu	0	2	2	0	0	1	1	1	0	1	0	1	…	174
	mi	0	4	1	1	1	1	1	1	0	0	0	0	…	163
	por	0	14	0	11	3	0	0	3	0	7	9	0	…	155
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	
	suma	789	749	544	431	406	375	327	220	203	174	163	155		

Tabla 20.Matriz antecede precede Jerusalem. (de creación propia)

Matriz transición estocástica “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON”.

		Palabra precedente											
		y	que	de	la	el	en	a	no	con	tu	…	**suma**
Palabra antecede	y	0.00%	2.66%	2.41%	2.15%	3.42%	3.93%	3.42%	1.90%	2.15%	1.65%	…	789
	que	0.00%	0.00%	2.14%	4.27%	1.87%	4.81%	2.80%	4.54%	1.60%	1.47%	…	749
	de	0.00%	1.29%	0.00%	7.54%	0.18%	0.37%	0.00%	0.00%	0.18%	3.49%	…	544
	la	0.00%	1.62%	0.70%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	…	431
	el	0.74%	2.46%	0.25%	0.25%	0.00%	0.00%	0.00%	0.25%	0.25%	0.00%	…	406
	en	0.00%	1.87%	0.00%	12.00%	10.13%	0.00%	0.00%	0.27%	0.00%	3.73%	…	375
	a	0.00%	1.22%	0.00%	9.79%	0.61%	0.00%	0.00%	0.61%	0.00%	6.12%	…	327
	no	0.00%	0.91%	0.45%	1.82%	1.36%	1.82%	0.45%	0.00%	1.36%	0.00%	…	220
	con	0.00%	2.96%	0.00%	8.87%	8.87%	0.00%	0.00%	0.00%	0.00%	1.97%	…	203
	tu	0.00%	1.15%	1.15%	0.00%	0.00%	0.57%	0.57%	0.57%	0.00%	0.57%	…	174
	mi	0.00%	2.45%	0.61%	0.61%	0.61%	0.61%	0.61%	0.61%	0.00%	0.00%	…	163
	por	0.00%	9.03%	0.00%	7.10%	1.94%	0.00%	0.00%	1.94%	0.00%	4.52%	…	155
	…	…	…	…	…	…	…	…	…	…	…	…	

Tabla 21.Matriz transición Jerusalem. (de creación propia)

- La matriz antecede precede permite ver como se relaciona las palabras en la obra “LA CONQUISTA DE JERUSALEN POR GODOFREBULLON” encontrando que se ha usado 45 veces “en la” la matriz completa incluye todo el vocabulario usado siendo una forma representativa del libro completo conservando el característico uso del vocabulario del autor.
- También se puede hacer mención de otras combinaciones de palabras muy usadas. “a la” usado 32 veces “que en” usado 36 veces.
- Se aprecia que en la diagonal de la matriz en su mayoría se encuentra muy pocos casos esto debido a que en la lengua española no es muy usual la repetición de palabras consecutivas.
- La matriz de transición permite ver que el lenguaje se comporta como un proceso estocástico permitiendo saber la posibilidad de que al usar una palabra la siguiente sea otra.

5.1.6. Descriptivos “LA NUMANCIA”.

Frecuencia de palabras usadas en la obra “LA NUMANCIA”. En la siguiente tabla se muestra solo las 50 palabras más importantes usadas en el vocabulario de la obra.

		NUMANCIA					NUMANCIA		
id	**palabra**	**n**	**%**	**importancia**	**id**	**palabra**	**n**	**%**	**importancia**
1	que	771	4.89%		26	del	84	0.53%	
2	y	669	4.25%		27	un	72	0.46%	
3	de	603	3.83%		28	pues	70	0.44%	
4	el	431	2.73%		29	ya	78	0.49%	
5	la	427	2.71%		30	ha	70	0.44%	
6	en	379	2.40%		31	tan	67	0.43%	
7	a	272	1.73%		32	bien	45	0.29%	si
8	no	224	1.42%		33	como	55	0.35%	
9	con	208	1.32%		34	este	52	0.33%	
10	se	151	0.96%		35	quien	47	0.30%	
11	mi	134	0.85%		36	le	36	0.23%	
12	por	136	0.86%		37	sin	43	0.27%	si
13	es	85	0.54%		38	una	45	0.29%	
14	si	110	0.70%		39	o	57	0.36%	
15	me	101	0.64%		40	ser	36	0.23%	
16	tu	100	0.63%		41	aquí	40	0.25%	
17	mas	135	0.86%		42	para	35	0.22%	
18	los	129	0.82%		43	hacer	48	0.30%	
19	al	114	0.72%		44	oh	31	0.20%	
20	lo	73	0.46%		45	vida	53	0.34%	si
21	esta	81	0.51%		46	cielo	22	0.14%	si
22	yo	77	0.49%		47	dar	32	0.20%	
23	su	81	0.51%		48	muerte	46	0.29%	si
24	las	99	0.63%		49	todo	37	0.23%	si
25	te	54	0.34%		50	ver	29	0.18%	si

Tabla 22.Frecuencia de palabras usadas en la obra Numancia. (de creación propia)

- La palabra “que” en la obra Numancia aparece 771 veces representado un 4.89% del texto representando ser la palabra más usada seguida de “y” que aparece 669 veces y representa un 4.25% del texto

Tabla de frecuencia palabras importantes.

No todas las palabras que presenta gran frecuencia tienen sentido completo no trasmiten un mensaje por lo cual se separaron las palabras que presentan importancia de los artículos y conjugaciones más comunes esto con el fin de encontrar cual es el mensaje a transmitir con la frecuencia de palabras usado en la obra “LA NUMANCIA” Se presentan las 20 primeras palabras de mayor frecuencia y de gran importancia.

id	palabra	n	%	importancia	id	palabra	n	%	importancia
45	vida	53	0.34%	si	55	mal	31	0.20%	si
48	muerte	46	0.29%	si	97	valor	30	0.19%	si
32	bien	45	0.29%	si	116	morir	30	0.19%	si
37	sin	43	0.27%	si	50	ver	29	0.18%	si
75	salir	40	0.25%	si	79	fin	29	0.18%	si
59	solo	38	0.24%	si	58	poder	28	0.18%	si
172	romanos	38	0.24%	si	99	tener	28	0.18%	si
49	todo	37	0.23%	si	125	manos	28	0.18%	si
145	hambre	35	0.22%	si	73	pecho	27	0.17%	si
91	nuestro	32	0.20%	si	109	amigo	26	0.16%	si

Tabla 23.Palabras importantes Numancia. (de creación propia)

- Encontramos entre las palabras de gran importancia en el texto “LA NUMANCIA” que es “vida” representando un 0.34% que se repitió 53 veces en el mismo.
- Para dar una mejor interpretación a las frecuencias encontradas se presentarán en el gráfico nube de palabras.

Nube de palabras “LA NUMANCIA”.

Para el desarrollo de el grafico de nube de palabras se usó los conteos superiores e igual a 20 en toda la obra de “LA NUMANCIA” presentando 200 de las primeras palabras con importancia y significado completo.

Ilustración 13.Nube de palabras Numancia. (de creación propia)

- Se aprecia la importancia de la palabra “vida” representativa de la obra “LA NUMANCIA” seguido de “muerte” “hambre” y “romanos”.

Matriz antecede precede “LA NUMANCIA”.

		Palabra precedente													
		que	y	de	el	la	en	a	no	con	se	mi	mas	…	**suma**
Palabra antecede	que	0	0	20	27	18	32	24	36	11	27	3	10	…	757
	y	22	0	18	18	17	27	21	15	15	3	5	5	…	666
	de	24	0	0	0	57	1	0	1	0	0	34	1	…	598
	el	18	1	2	0	1	0	0	1	0	3	0	2	…	425
	la	7	0	0	0	0	0	0	0	0	0	0	2	…	425
	en	8	0	0	32	49	0	0	0	0	0	8	0	…	369
	a	5	0	0	1	24	0	0	0	0	0	19	0	…	266
	no	1	0	3	1	7	2	1	0	3	17	0	2	…	211
	con	4	0	0	18	13	0	0	0	0	0	1	4	…	203
	se	3	0	1	0	0	0	0	0	0	1	0	0	…	151
	mi	0	1	0	1	2	1	0	0	0	1	0	1	…	134
	mas	8	0	4	2	2	1	3	2	1	0	2	0	…	133
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	
	suma	757	666	598	425	425	369	266	211	203	151	134	133		

Tabla 24.Matriz antecede precede Numancia. (de creación propia)

Matriz transición estocástica “LA NUMANCIA”.

		Palabra precedente											
		que	y	de	el	la	en	a	no	con	se	…	**suma**
Palabra antecede	que	0.00%	0.00%	2.64%	3.57%	2.38%	4.23%	3.17%	4.76%	1.45%	3.57%	…	100%
	y	3.30%	0.00%	2.70%	2.70%	2.55%	4.05%	3.15%	2.25%	2.25%	0.45%	…	100%
	de	4.01%	0.00%	0.00%	0.00%	9.53%	0.17%	0.00%	0.17%	0.00%	0.00%	…	100%
	el	4.24%	0.24%	0.47%	0.00%	0.24%	0.00%	0.00%	0.24%	0.00%	0.71%	…	100%
	la	1.65%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	en	2.17%	0.00%	0.00%	8.67%	13.28%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	a	1.88%	0.00%	0.00%	0.38%	9.02%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	no	0.47%	0.00%	1.42%	0.47%	3.32%	0.95%	0.47%	0.00%	1.42%	8.06%	…	100%
	con	1.97%	0.00%	0.00%	8.87%	6.40%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	se	1.99%	0.00%	0.66%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.66%	…	100%
	mi	0.00%	0.75%	0.00%	0.75%	1.49%	0.75%	0.00%	0.00%	0.00%	0.75%	…	100%
	mas	6.02%	0.00%	3.01%	1.50%	1.50%	0.75%	2.26%	1.50%	0.75%	0.00%	…	100%
	…	…	…	…	…	…	…	…	…	…	…	…	

Tabla 25.Matriz transición Numancia. (de creación propia)

- La matriz antecede precede permite ver como relaciona las palabras de cervantes en la obra “LA NUMANCIA” encontrando que ha usado 36 veces “que no” la matriz completa incluye todo el vocabulario usado siendo una forma de representativa del libro completo conservando el característico uso del vocabulario por cervantes.
- También se puede hacer mención de otras combinaciones de palabras muy usadas. “en la” usado 49 veces “que se” usado 27 veces.
- Se aprecia que en la diagonal de la matriz en su mayoría se encuentra muy pocos casos esto debido que en la lengua española no es muy usual la repetición de palabras consecutivas.
- La matriz de transición permite ver que el lenguaje se comporta como un proceso estocástico permitiendo saber la posibilidad de que al usar una palabra la siguiente sea otra.

5.2. Resultados correlacionales.

5.2.1. Relación de personajes por vocabulario.

Para la creación de la tabla agregada personaje-vocabulario.

	PERSONAJES										
	a-Aurelio	n-Ciprion	j-tancredo	j-Erminia	a-Sayavedra	j-Godofre	n-morandro	j-Clorinda	a-Zahara	j-escenario	…
Que	177	132	98	78	73	55	61	64	54	21	…
Y	134	114	94	69	71	62	40	40	44	128	…
De	114	79	72	48	47	48	50	29	43	43	…
El	72	60	46	37	47	47	29	18	22	25	…
La	60	58	60	36	43	31	30	34	21	27	…
En	73	60	45	39	32	27	30	24	14	18	…
A	60	32	40	47	32	32	24	25	26	20	…
No	57	33	33	30	20	23	28	15	24	1	…
Con	27	26	24	13	15	23	11	7	12	30	…
Se	28	32	22	12	17	17	9	6	16	3	…
Mi	48	16	30	29	10	8	31	19	15	0	…
Por	34	19	31	17	15	9	11	14	9	5	…
Es	37	24	15	16	16	10	8	12	31	1	…
Si	32	24	23	29	9	6	9	22	10	1	…
Me	42	12	21	25	9	5	15	14	15	0	…
Tu	22	9	28	17	15	5	14	20	8	0	…
Mas	28	28	22	15	10	5	10	11	7	1	…
Los	19	12	11	0	10	10	9	10	6	15	…
Al	24	20	17	9	11	9	3	4	8	4	…
Lo	24	16	18	10	6	7	10	10	10	10	…
…	…	…	…	…	…	…	…	…	…	…	…
N	3166	2285	2153	1566	1475	1418	1180	1170	1100	1098	
%	6.6%	4.8%	4.5%	3.3%	3.1%	2.9%	2.4%	2.4%	2.3%	2.3%	

Tabla 26.Personajes vocabulario. (de creación propia)

En la tabla se aprecia los 10 personajes con más líneas en las tres obras y su frecuencia de uso de palabras o vocabulario en las primeras 20 palabras para el estudio siguiente de análisis se hizo uso de las 50 palabras más importantes en su vocabulario y personales con cantidad de palabras superior al 1.1% en los 3 libros.

Esto con el fin de ver si existe relación para proceder a realizar un análisis de correspondencia.

Prueba de independencia entre las filas y columnas:

Chi-cuadrado (Valor observado)	Chi-cuadrado (Valor crítico)	GL	valor-p	alfa
3684.965	1711.663	1617	< 0.0001	0.05

Tabla 27.Chi-cuadrado personajes vocabulario. (de creación propia)

Interpretación de la prueba:

H0: Las filas y las columnas de la tabla son independientes. (No existe relación entre personajes y vocabulario)

Ha: Hay asociación entre las filas y las columnas de la tabla. (Existe relación entre personajes y su vocabulario)

Puesto que el valor-p calculado es menor que el nivel de significación alfa=0.05, se debe rechazar la hipótesis nula H0, y aceptar la hipótesis alternativa Ha.

El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es inferior al 0.01%. Existe relación entre el vocabulario y el personaje.

Análisis de correspondencia. Inercia total 0.235

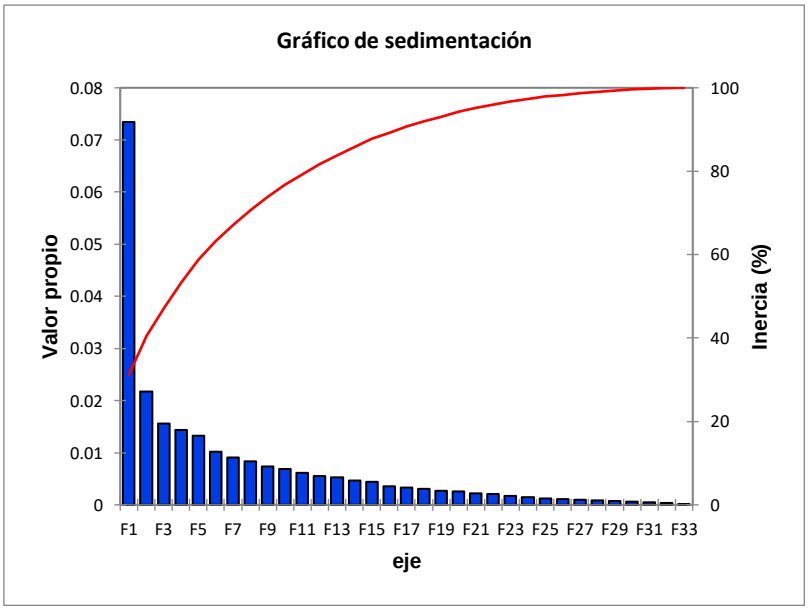


Ilustración 14.Sedimentación personaje vocabulario. (de creación propia)

- Los dos primeros ejes planteados explican el 0.235 de los datos, pero se aprecia que la mayoría de información es recaudada en estos dos ejes.

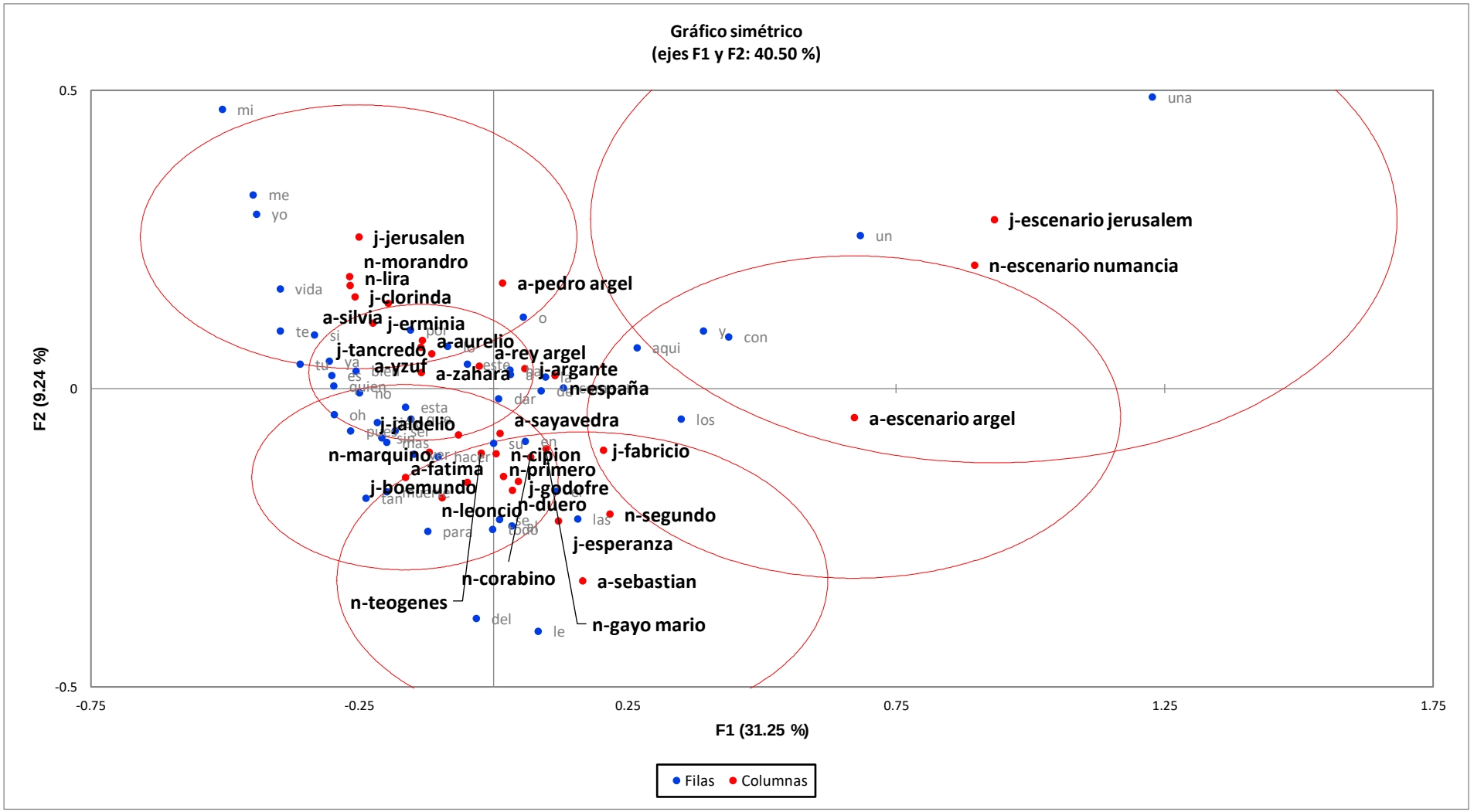


Ilustración 15.Correspondencia personaje vocabulario. (de creación propia)

Interpretación del grafico de análisis de correspondencia de personajes en razón a su vocabulario.

- Se aprecia que los personajes “Jerusalén”, “Morandro” de “LA NUMANCIA”. “Lira “, “Clorinda” de “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” “Silvia” de “EL TRATO DE ARGEL” “Erminia” “Tangredo” de “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” “Aurelio” y “Pedro” de “EL TRATO DE ARGEL tienen semejanzas en su vocabulario en el cual tienden a usar las palabras “me” “yo” “vida” “te” “si” “por” y “ya”.
- Se aprecia semejanza en el vocabulario de “Erminia” “Jaldelio” y “Tangredo” “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” “Aurelio” “Zahaa” y “Yuzuf” “EL TRATO DE ARGEL” estos personajes tienden a usar las palabras “ya” “es” “quien” “no” “está” “oh” y “este”.
- Existe semejanzas en el vocabulario de “Jaldelio” “Fatima” de “EL TRATO DE ARGEL”. “Marquiño” “Cipeon” “primero” “Leoncio” “Duero” de “LA NUMANCIA”. “Godofre” de “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” tienden a usar palabras como “tan” “para” “todo” “más” y “hacer”.
- Existe relación en el vocabulario de los personajes de “Fabricio” “Esperanza” de “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” “Segundo” de “LA NUMANCIA”. “Sebastian” de “EL TRATO DE ARGEL”. los cuales tienden a usar palabras como “le” “las” y “del”
- Se creó un personaje llamado escenario para cada obra el cual recolecta la información del escenario y movimiento de los personajes. Estos se agrupan en el análisis de correspondencia escenario “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON”, “EL TRATO DE ARGEL” y “LA NUMANCIA”. en los cuales se usa mucho las palabras un “con” “los” y “aquí”.

5.2.2. “EL TRATO DE ARGEL” – “LA NUMANCIA”.

Tabla agregada vocabulario obra. Para este estudio se usó las 50 palabras más usadas por las obras “EL TRATO DE ARGEL” y “LA NUMANCIA “esto con fin de ver si comparten el vocabulario en común.

		ARGEL		NUMANCIA	
id	**palabra**	**n**	**%**	**n**	**%**
1	que	742	4.93%	771	4.89%
2	y	665	4.42%	669	4.25%
3	de	514	3.42%	603	3.83%
4	el	400	2.66%	431	2.73%
5	la	317	2.11%	427	2.71%
6	en	324	2.15%	379	2.40%
7	a	301	2.00%	272	1.73%
8	no	265	1.76%	224	1.42%
9	con	138	0.92%	208	1.32%
10	se	176	1.17%	151	0.96%
11	mi	159	1.06%	134	0.85%
12	por	145	0.96%	136	0.86%
13	es	191	1.27%	85	0.54%
14	si	130	0.86%	110	0.70%
15	me	160	1.06%	101	0.64%
16	tu	104	0.69%	100	0.63%
17	mas	117	0.78%	135	0.86%
18	los	102	0.68%	129	0.82%
19	al	106	0.70%	114	0.72%
20	lo	99	0.66%	73	0.46%
21	esta	96	0.64%	81	0.51%
22	yo	79	0.53%	77	0.49%
23	su	107	0.71%	81	0.51%
24	las	59	0.39%	99	0.63%
25	te	84	0.56%	54	0.34%

Tabla 28. Vocabulario de "EL TRATO DE ARGEL" Y "LA NUMANCIA"(de creación propia)

Prueba de independencia entre las vocabulario y obras (Chi-cuadrado):

Chi-cuadrado (Valor observado)	Chi-cuadrado (Valor crítico)	GL	valor-p	alfa
205.908	66.339	49	< 0.0001	0.05

Tabla 29.Chi-cuadrado vocabulario de "EL TRATO DE ARGEL" Y "LA NUMANCIA”. (de creación propia)

Interpretación de la prueba:

Ho: Las filas y las columnas de la tabla son independientes. (no existe relación entre los vocabularios y las obras "EL TRATO DE ARGEL" y "LA NUMANCIA")

Ha: Hay dependencia entre las filas y las columnas de la tabla. (existe relación entre los vocabularios y las obras "EL TRATO DE ARGEL" y "LA NUMANCIA")

Puesto que el valor-p calculado es menor que el nivel de significación α=0.05, se debe rechazar la hipótesis nula H0, y aceptar la hipótesis alternativa Ha. El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es inferior al 0.01%. existe relación entre los vocabularios de las obras *"EL TRATO DE ARGEL" y "LA NUMANCIA".*

Se plantea la siguiente hipótesis.

Calculo de chi-cuadrado para Prueba de hipótesis de matrices de transición de "EL TRATO DE ARGEL" y "LA NUMANCIA"

$$T_{argel} = T_{Numancia}$$

H0: Las matrices de transición estocástica generadas a partir del corpus textual "EL TRATO DE ARGEL" y "LA NUMANCIA" *son iguales.*

Ha: Las matrices de transición estocástica generadas a partir del corpus textual "EL TRATO DE ARGEL" y "LA NUMANCIA" *son distintas.*

Matriz transición estocástica "LA NUMANCIA".

Palabra antecede	Palabra precedente											
	que	y	de	el	la	en	a	no	con	se	...	suma
que	0.00%	0.00%	2.64%	3.57%	2.38%	4.23%	3.17%	4.76%	1.45%	3.57%	...	100%
y	3.30%	0.00%	2.70%	2.70%	2.55%	4.05%	3.15%	2.25%	2.25%	0.45%	...	100%
de	4.01%	0.00%	0.00%	0.00%	9.53%	0.17%	0.00%	0.17%	0.00%	0.00%	...	100%
el	4.24%	0.24%	0.47%	0.00%	0.24%	0.00%	0.00%	0.24%	0.00%	0.71%	...	100%
la	1.65%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
en	2.17%	0.00%	0.00%	8.67%	13.28%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
a	1.88%	0.00%	0.00%	0.38%	9.02%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
no	0.47%	0.00%	1.42%	0.47%	3.32%	0.95%	0.47%	0.00%	1.42%	8.06%	...	100%
con	1.97%	0.00%	0.00%	8.87%	6.40%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
se	1.99%	0.00%	0.66%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.66%	...	100%
mi	0.00%	0.75%	0.00%	0.75%	1.49%	0.75%	0.00%	0.00%	0.00%	0.75%	...	100%
mas	6.02%	0.00%	3.01%	1.50%	1.50%	0.75%	2.26%	1.50%	0.75%	0.00%	...	100%
...	...	...	...	...	...	...	...	...	...	...	...	

Tabla 36.Matriz transición Numancia.

Matriz transición estocástica "EL TRATO DE ARGEL".

Palabra antecede	Palabra precedente											
	que	y	de	el	la	en	a	no	es	se	...	suma
que	0.14%	0.14%	0.82%	4.36%	1.63%	4.22%	3.41%	6.40%	6.95%	3.54%	...	100%
y	2.46%	0.00%	3.07%	3.84%	3.53%	3.69%	2.00%	2.15%	1.38%	0.46%	...	100%
de	0.99%	0.00%	0.00%	0.00%	5.75%	0.20%	0.00%	0.79%	0.00%	0.00%	...	100%
el	2.84%	0.00%	0.26%	0.00%	0.00%	0.00%	0.26%	0.77%	0.00%	0.52%	...	100%
la	1.29%	0.00%	0.96%	0.00%	0.00%	0.00%	0.00%	0.32%	0.00%	0.00%	...	100%
en	2.29%	0.00%	0.00%	13.40%	9.48%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
a	1.68%	0.00%	0.00%	0.67%	6.40%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
no	0.00%	0.00%	1.63%	0.82%	1.22%	0.82%	0.41%	0.82%	4.49%	7.35%	...	100%
es	2.15%	0.00%	3.76%	2.15%	6.99%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
se	6.82%	0.00%	0.00%	0.00%	0.57%	0.00%	0.57%	0.00%	0.00%	0.00%	...	100%
me	0.00%	0.00%	0.63%	0.00%	0.63%	0.00%	0.00%	0.00%	0.63%	0.00%	...	100%
mi	1.27%	0.00%	0.63%	0.63%	0.63%	0.00%	0.63%	0.00%	0.00%	0.63%	...	100%
...	...	...	...	...	...	...	...	...	...	...	...	

Tabla 28.Matriz transición Argel.

Ilustración 16.Matriz de transición "LA NUMANCIA" Y " EL TRATO DE ARGEL"(de creación propia)

Se presenta las matrices de transición ya calculadas anteriormente. Las cuales siguen el estadístico de prueba planteado en el marco teórico. Para el estudio se plantea el uso de los primeros 105 niveles o palabras más usadas

Para el cual es necesario las matrices de transición.

Y la operación a partir ellos

$$T_{Numancia} * T_{Argel}^{-1}$$

De la cual se obtiene la siguiente matriz. De la cual se presenta las 10 primeras columnas y 20 filas de las 105*105 se trabajó en orden alfabético y de mismas categorías par todas las matrices.

	a	acá	al	alma	amor	aquel	aquí	aunque	Aurelio	bien	
a	1.57556	0.01598	-0.44768	-0.13624	0.01652	0.13921	0.36893	0.06240	0.43802	0.40586	…
acá	-0.29157	0.99095	0.35988	0.16551	-0.03013	-0.10550	-0.27327	-0.25411	-0.29871	-0.18966	…
al	-4.97057	-0.08901	2.37495	0.13836	-0.40671	-0.77077	-2.45884	-0.69280	-2.68052	-1.95026	…
alma	2.50622	-0.13848	-1.03176	-0.29488	0.28667	0.34910	0.82476	-0.08381	1.31037	0.72863	…
amor	3.87343	-0.29155	-3.65319	-0.52687	0.04701	1.14893	2.55291	1.03153	3.43123	3.76524	…
aquel	-0.29157	-0.00905	0.35988	0.16551	-0.03013	0.89450	-0.27327	-0.25411	-0.29871	-0.18966	…
aquí	-0.67263	-0.00494	-0.22191	0.12188	-0.01295	-0.01502	0.24581	0.66745	0.08434	-0.25775	…
aunque	1.19931	-0.35034	-2.20694	-0.06821	-0.81936	0.52482	1.39123	2.80406	1.51936	1.42854	…
Aurelio	3.21285	0.66830	0.46829	0.99461	0.99397	0.48970	1.56376	-0.82488	1.33445	0.57408	…
bien	-0.53901	0.08562	-0.23987	-0.00065	0.06888	0.23716	0.33229	-0.06720	0.37126	0.40758	…
cielo	-2.70685	0.31611	2.89022	0.98888	0.07299	-0.62953	-1.48955	-0.92315	-2.11976	-2.11688	…
ciudad	-0.29157	-0.00905	0.35988	0.16551	-0.03013	-0.10550	-0.27327	-0.25411	-0.29871	-0.18966	…
Clorinda	-0.29157	-0.00905	0.35988	0.16551	-0.03013	-0.10550	-0.27327	-0.25411	-0.29871	-0.18966	…
como	0.06403	0.00728	-0.13055	-0.79890	0.04612	0.04421	-0.22206	-0.36859	0.24385	0.50534	…
con	-0.00423	-0.09953	-0.33848	-0.07044	-0.19982	-0.11785	-0.25527	0.37118	-0.13448	-0.07301	…
cristiano	-0.07550	0.08598	0.19326	0.18080	0.11613	0.02998	0.05422	-0.00534	-0.00785	-0.08237	…
cual	5.01226	-0.69002	-3.23006	-2.20305	-0.38330	0.30792	1.50800	2.42928	2.36817	0.06412	…
dar	-1.27946	-0.16040	0.05559	-0.52501	-0.32583	-0.23500	-0.88307	0.46857	-0.47475	-0.60960	…
de	0.12416	-0.05556	-0.21535	-0.14253	-0.03906	0.01891	0.04118	0.15991	0.14265	0.06300	…
decir	-0.29157	-0.00905	0.35988	0.16551	-0.03013	-0.10550	-0.27327	-0.25411	-0.29871	-0.18966	…
	…	…	…	…	…	…	…	…	…	…	

*Tabla 30.Matriz ARGEL INVERSA * NUMANCIA (de creación propia)*

Donde se calcula la traza y determinante de valores requerido para la prueba planteada con ayuda del paquete estadístico R

Para el cálculo de los valores críticos según la prueba es donde n=p:

$$m = \frac{1}{2}p(p+1)$$

n	p	alfa	m	chi de tabla critico
105	105	0.95	5565	5392.614

Tabla 31.Valor critico prueba ARGEL NUMANCIA. (de creación propia)

El estadístico de prueba es:

$$\chi^2_{calculado} = np(a - ln(g) - 1) \sim \chi^2_m$$

n	Determinante	g	Traza	A	chi calculado	p valor
105	0.000084	0.914549	109.192986	1.039933	1425.066140	0.999

Tabla 32 Valor calculado prueba ARGEL NUMANCIA. (de creación propia)

Dado que el valor calculado es inferior al crítico podemos afirmar que la matriz de transición A es igual a la matriz de transición B.

Chií- cuadrado (ARGEL-NUMANCIA) = 1425.06

Dado que buscamos comprobar la autentificación de una obra tomaremos este valor de chi-cuadrado como punto crítico para la comparación ya que este valor nos indica que las dos obras “El Trato de Argel” y “La Numancia” tienen una similitud del 99,9% pero a nivel estadístico de prueba este es de 1425.06 lo que implica que si se obtienen valores inferiores a éste en las pruebas.

Chi-cuadrado (JERUSALEN-NUMANCIA) y Chií-cuadrado (JERUSALEN-ARGEL)

Se puede atribuir estadísticamente la obra de “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” a Cervantes.

5.2.3. “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” – “EL TRATO DE ARGEL”

Tabla agregada vocabulario obra. Para este estudio se usó las 50 palabras más usadas por las obras esto con fin de ver si comparten el vocabulario en común.

		ARGEL		JERUSALEM	
id	**palabra**	**n**	**%**	**n**	**%**
1	que	742	4.93%	772	4.63%
2	y	665	4.42%	794	4.76%
3	de	514	3.42%	552	3.31%
4	el	400	2.66%	416	2.49%
5	la	317	2.11%	433	2.60%
6	en	324	2.15%	386	2.31%
7	a	301	2.00%	341	2.04%
8	no	265	1.76%	254	1.52%
9	con	138	0.92%	208	1.25%
10	se	176	1.17%	150	0.90%
11	mi	159	1.06%	164	0.98%
12	por	145	0.96%	167	1.00%
13	es	191	1.27%	125	0.75%
14	si	130	0.86%	159	0.95%
15	me	160	1.06%	130	0.78%
16	tu	104	0.69%	178	1.07%
17	mas	117	0.78%	112	0.67%
18	los	102	0.68%	120	0.72%
19	al	106	0.70%	105	0.63%
20	lo	99	0.66%	111	0.67%
21	esta	96	0.64%	102	0.61%
22	yo	79	0.53%	108	0.65%
23	su	107	0.71%	72	0.43%
24	las	59	0.39%	101	0.61%
25	te	84	0.56%	105	0.63%

Tabla 33.Vocabulario de "EL TRATO DE ARGEL" Y "JERUSALEN" (de creación propia)

Prueba de independencia entre las vocabulario y obras (Chi-cuadrado):

Chi-cuadrado (Valor observado)	Chi-cuadrado (Valor crítico)	GL	valor-p	alfa
151.359	66.339	49	< 0.0001	0.05

Tabla 34.Chii cuadrado vocabulario de "EL TRATO DE ARGEL" Y "JERUSALEN”. (de creación propia)

Interpretación de la prueba:

Ho: Las filas y las columnas de la tabla son independientes. (no existe relación entre los vocabularios y las obras "EL TRATO DE ARGEL" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON")

Ha: Hay dependencia entre las filas y las columnas de la tabla. (existe relación entre los vocabularios y las obras "EL TRATO DE ARGEL" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON"*)*

Puesto que el valor-p calculado es menor que el nivel de significación alfa=0.05, se debe rechazar la hipótesis nula H0, y aceptar la hipótesis alternativa Ha. El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es inferior al 0.01%. existe relación entre los vocabularios de las obras *"EL TRATO DE ARGEL"* "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON"

Se plantea la siguiente hipótesis. Cálculo de chi-cuadrado para Prueba de hipótesis de matrices de transición de "EL TRATO DE ARGEL" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON"

$$T_{argel} = T_{Jerusalén}$$

H0: Las matrices de transición estocástica generadas a partir del corpus textual "EL TRATO DE ARGEL" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON" *son iguales.*

Ha: Las matrices de transición estocástica generadas a partir del corpus textual "EL TRATO DE ARGEL" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON" *son distintas.*

Matriz transición estocástica "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON".

		Palabra precedente											
		y	que	de	la	el	en	a	no	con	tu	...	suma
Palabra antecede	y	0.00%	2.66%	2.41%	2.15%	3.42%	3.93%	3.42%	1.90%	2.15%	1.65%	...	789
	que	0.00%	0.00%	2.14%	4.27%	1.87%	4.81%	2.80%	4.54%	1.60%	1.47%	...	749
	de	0.00%	1.29%	0.00%	7.54%	0.18%	0.37%	0.00%	0.00%	0.18%	3.49%	...	544
	la	0.00%	1.62%	0.70%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	...	431
	el	0.74%	2.46%	0.25%	0.25%	0.00%	0.00%	0.00%	0.25%	0.25%	0.00%	...	406
	en	0.00%	1.87%	0.00%	12.00%	10.13%	0.00%	0.00%	0.27%	0.00%	3.73%	...	375
	a	0.00%	1.22%	0.00%	9.79%	0.61%	0.00%	0.00%	0.61%	0.00%	6.12%	...	327
	no	0.00%	0.91%	0.45%	1.82%	1.36%	1.82%	0.45%	0.00%	1.36%	0.00%	...	220
	con	0.00%	2.96%	0.00%	8.87%	8.87%	0.00%	0.00%	0.00%	0.00%	1.97%	...	203
	tu	0.00%	1.15%	1.15%	0.00%	0.00%	0.57%	0.57%	0.57%	0.00%	0.57%	...	174
	mi	0.00%	2.45%	0.61%	0.61%	0.61%	0.61%	0.61%	0.61%	0.00%	0.00%	...	163
	por	0.00%	9.03%	0.00%	7.10%	1.94%	0.00%	0.00%	1.94%	0.00%	4.52%	...	155
	...	...	...	...	...	...	...	...	...	...	...	...	

Tabla 32.Matriz transición Jerusalem.

Matriz transición estocástica "EL TRATO DE ARGEL".

		Palabra precedente											
		que	y	de	el	la	en	a	no	es	se	...	suma
Palabra antecede	que	0.14%	0.14%	0.82%	4.36%	1.63%	4.22%	3.41%	6.40%	6.95%	3.54%	...	100%
	y	2.46%	0.00%	3.07%	3.84%	3.53%	3.69%	2.00%	2.15%	1.38%	0.46%	...	100%
	de	0.99%	0.00%	0.00%	0.00%	5.75%	0.20%	0.00%	0.79%	0.00%	0.00%	...	100%
	el	2.84%	0.00%	0.26%	0.00%	0.00%	0.00%	0.26%	0.77%	0.00%	0.52%	...	100%
	la	1.29%	0.00%	0.96%	0.00%	0.00%	0.00%	0.00%	0.32%	0.00%	0.00%	...	100%
	en	2.29%	0.00%	0.00%	13.40%	9.48%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
	a	1.68%	0.00%	0.00%	0.67%	6.40%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
	no	0.00%	0.00%	1.63%	0.82%	1.22%	0.82%	0.41%	0.82%	4.49%	7.35%	...	100%
	es	2.15%	0.00%	3.76%	2.15%	6.99%	0.00%	0.00%	0.00%	0.00%	0.00%	...	100%
	se	6.82%	0.00%	0.00%	0.00%	0.57%	0.00%	0.57%	0.00%	0.00%	0.00%	...	100%
	me	0.00%	0.00%	0.63%	0.00%	0.63%	0.00%	0.00%	0.00%	0.63%	0.00%	...	100%
	mi	1.27%	0.00%	0.63%	0.63%	0.63%	0.00%	0.63%	0.00%	0.00%	0.63%	...	100%
	...	...	...	...	...	...	...	...	...	...	...	...	

Tabla 28.Matriz transición Argel.

Ilustración 17.Matriz de transición "EL TRATO DE ARGEL" y "JERUSALEN" (de creación propia)

Se presenta las matrices de transición ya calculadas anteriormente. Las cuales siguen el estadístico de prueba planteado en el marco teórico. Para el estudio se plantea el uso de los primeros 105 niveles o palabras más usadas

Para lo cual es necesario las matrices de transición.

Y la operación a partir ellos

$$T_{Jerusalen} * T_{Argel}^{-1}$$

De la cual se obtiene la siguiente matriz. De la cual se presenta las 10 primeras columnas y 20 filas de las 105*105 se trabajó en orden alfabético y de mismas categorías par todas las matrices.

	a	acá	al	alma	amor	aquel	aquí	aunque	Aurelio	bien	
a	0.41527	0.08257	-1.70801	-0.06965	-0.42667	0.04613	0.58310	0.21198	0.50461	0.08519	…
acá	-0.16097	0.99242	0.57829	0.16698	0.21284	-0.08838	-0.14685	-0.11237	-0.29725	-0.14335	…
al	-1.63227	-0.17320	7.99442	0.05417	1.17189	-0.59394	-3.06372	-0.88531	-2.76471	0.11258	…
alma	-1.11587	-0.04549	-4.16669	-0.20190	-0.69204	0.03334	0.93417	0.46141	1.40335	0.86886	…
amor	-2.38844	-0.15260	-9.58122	-0.38792	-3.36138	0.52791	2.08144	1.35156	3.57018	2.75722	…
aquel	-1.03597	0.05492	0.64079	0.22948	0.27534	0.91162	-0.08435	-0.04987	-0.23475	-1.01835	…
aquí	0.25811	-0.05465	0.21229	0.07216	0.08727	0.08647	0.10035	-0.06022	0.03463	-0.27810	…
aunque	0.11638	-0.37948	-4.96447	-0.09736	-3.00391	0.12699	0.54745	1.68189	1.49021	0.29206	…
Aurelio	4.06418	0.34432	-3.02421	0.67062	0.82030	-0.01901	2.63577	0.15175	1.01046	-0.95280	…
bien	-1.26201	0.03100	-0.49547	-0.05527	-0.43820	0.14858	-0.02711	0.18845	0.31664	0.64182	…
cielo	1.74160	0.12091	5.84621	0.79368	2.46330	-0.39774	-1.15281	-0.86612	-2.31496	-2.10329	…
ciudad	0.00569	0.15908	0.74496	0.33364	0.37951	0.07829	0.01982	0.05429	-0.13058	0.02332	…
Clorinda	0.03750	0.19089	0.77676	0.36545	0.41132	0.11010	0.05162	0.08610	-0.09877	0.05513	…
como	-1.80005	0.02845	-0.84955	-0.77773	0.07583	-0.10142	0.01858	0.06817	0.26501	0.54751	…
con	0.86636	0.04205	0.24469	0.07115	-0.13468	0.10779	-0.13002	0.02389	0.00711	-0.01418	…
cristiano	0.40244	0.01793	0.20420	0.11275	0.33643	0.00204	0.12629	-0.03975	-0.07589	-0.24187	…
cual	-0.56389	-0.20649	-7.20895	-1.71953	-1.76080	0.38954	1.22507	0.85191	2.85170	1.04549	…
dar	-1.16610	-0.02131	1.79173	-0.38591	-0.18657	-0.11302	-1.23510	-0.08865	-0.33565	0.21122	…
de	-0.10520	-0.04185	-0.28230	-0.12882	-0.15685	0.00542	-0.01441	0.03705	0.15636	0.05978	…
decir	0.07260	0.22599	0.81187	0.40055	0.44642	0.14520	0.08673	0.12120	-0.06367	0.09023	…
	…	…	…	…	…	…	…	…	…	…	

*Tabla 35.Matriz ARGEL INVERSA * JERUSALEN. (de creación propia)*

Donde se calcula la traza y determinante de valores requerido para la prueba planteada con ayuda del paquete estadístico R

Para el cálculo de los valores críticos según la prueba es donde n=p:

$$m = \frac{1}{2}p(p+1)$$

n	p	alfa	m	chi de tabla critico
105	105	0.95	5565	5392.614

Tabla 36.Valor critico prueba ARGEL JERUSALEN. (de creación propia)

El estadístico de prueba es:

$$\chi^2_{calculado} = np(a - ln(g) - 1) \sim \chi^2_m$$

n	Determinante	g	Traza	a	chi calculado	p valor
105	0.000005705	0.891374	104.9484	0.99950	1262.36456	0.999

Tabla 37.Valor calculado prueba ARGEL JERUSALEN. (de creación propia)

Dado que el valor calculado es inferior al crítico podemos afirmar que la matriz de transición A es igual a la matriz de transición B.

Chií- cuadrado (ARGEL-JERUSALEN) = 1262.36

El valor es inferior al valor crítico dado por la prueba ARGEL JERUSALEN de 1425 indicando que la obra sea de Cervantes.

5.2.4. “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” – “LA NUMANCIA”

Tabla agregada vocabulario obra. Para este estudio se usó las 50 palabras más usadas por las obras esto con fin de ver si comparten el vocabulario en común.

		JERUSALEM		NUMANCIA	
id	**palabra**	**n**	**%**	**n**	**%**
1	que	772	4.63%	771	4.89%
2	y	794	4.76%	669	4.25%
3	de	552	3.31%	603	3.83%
4	el	416	2.49%	431	2.73%
5	la	433	2.60%	427	2.71%
6	en	386	2.31%	379	2.40%
7	a	341	2.04%	272	1.73%
8	no	254	1.52%	224	1.42%
9	con	208	1.25%	208	1.32%
10	se	150	0.90%	151	0.96%
11	mi	164	0.98%	134	0.85%
12	por	167	1.00%	136	0.86%
13	es	125	0.75%	85	0.54%
14	si	159	0.95%	110	0.70%
15	me	130	0.78%	101	0.64%
16	tu	178	1.07%	100	0.63%
17	mas	112	0.67%	135	0.86%
18	los	120	0.72%	129	0.82%
19	al	105	0.63%	114	0.72%
20	lo	111	0.67%	73	0.46%
21	esta	102	0.61%	81	0.51%
22	yo	108	0.65%	77	0.49%
23	su	72	0.43%	81	0.51%
24	las	101	0.61%	99	0.63%
25	te	105	0.63%	54	0.34%

Tabla 38.Vocabulario de "LA NUMANCIA " Y "JERUSALEN". (de creación propia)

Prueba de independencia entre las filas y columnas (Chi-cuadrado):

Chi-cuadrado (Valor observado)	Chi-cuadrado (Valor crítico)	GL	valor-p	alfa
123.566	66.339	49	< 0.0001	0.05

Tabla 39.Chi-cuadrado vocabulario de "LA NUMANCIA" Y "JERUSALEN”. (de creación propia)

Interpretación de la prueba:

Ho: Las filas y las columnas de la tabla son independientes. (no existe relación entre los vocabularios y las obras "LA NUMANCIA" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON")

Ha: Hay dependencia entre las filas y las columnas de la tabla. (existe relación entre los vocabularios y las obras "LA NUMANCIA" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON"*)*

Puesto que el valor-p calculado es menor que el nivel de significación alfa=0.05, se debe rechazar la hipótesis nula H0, y aceptar la hipótesis alternativa Ha. El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es inferior al 0.01%. existe relación entre los vocabularios de las obras *"LA NUMANCIA"* "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON"

Se plantea la siguiente hipótesis.

Calculo de chi-cuadrado para la Prueba de hipótesis de matrices de transición de "NUMANCIA" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON"

$$T_{Numancia} = T_{Jerusalén}$$

H0: Las matrices de transición estocástica generadas a partir del corpus textual "LA NUMANCIA" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON" *son iguales.*

Ha: Las matrices de transición estocástica generadas a partir del corpus textual "LA NUMANCIA" y "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON" *son distintas.*

Matriz transición estocástica "LA NUMANCIA".

		Palabra precedente											
		que	y	de	el	la	en	a	no	con	se	…	suma
Palabra antecede	que	0.00%	0.00%	2.64%	3.57%	2.38%	4.23%	3.17%	4.76%	1.45%	3.57%	…	100%
	y	3.30%	0.00%	2.70%	2.70%	2.55%	4.05%	3.15%	2.25%	2.25%	0.45%	…	100%
	de	4.01%	0.00%	0.00%	0.00%	9.53%	0.17%	0.00%	0.17%	0.00%	0.00%	…	100%
	el	4.24%	0.24%	0.47%	0.00%	0.24%	0.00%	0.00%	0.24%	0.00%	0.71%	…	100%
	la	1.65%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	en	2.17%	0.00%	0.00%	8.67%	13.28%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	a	1.88%	0.00%	0.00%	0.38%	9.02%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	no	0.47%	0.00%	1.42%	0.47%	3.32%	0.95%	0.47%	0.00%	1.42%	8.06%	…	100%
	con	1.97%	0.00%	0.00%	8.87%	6.40%	0.00%	0.00%	0.00%	0.00%	0.00%	…	100%
	se	1.99%	0.00%	0.66%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.66%	…	100%
	mi	0.00%	0.75%	0.00%	0.75%	1.49%	0.75%	0.00%	0.00%	0.00%	0.75%	…	100%
	mas	6.02%	0.00%	3.01%	1.50%	1.50%	0.75%	2.26%	1.50%	0.75%	0.00%	…	100%
	…	…	…	…	…	…	…	…	…	…	…	…	

Tabla 36.Matriz transición Numancia.

Matriz transición estocástica "LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON".

		Palabra precedente											
		y	que	de	la	el	en	a	no	con	tu	…	suma
Palabra antecede	y	0.00%	2.66%	2.41%	2.15%	3.42%	3.93%	3.42%	1.90%	2.15%	1.65%	…	789
	que	0.00%	0.00%	2.14%	4.27%	1.87%	4.81%	2.80%	4.54%	1.60%	1.47%	…	749
	de	0.00%	1.29%	0.00%	7.54%	0.18%	0.37%	0.00%	0.00%	0.18%	3.49%	…	544
	la	0.00%	1.62%	0.70%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	…	431
	el	0.74%	2.46%	0.25%	0.25%	0.00%	0.00%	0.00%	0.25%	0.25%	0.00%	…	406
	en	0.00%	1.87%	0.00%	12.00%	10.13%	0.00%	0.00%	0.27%	0.00%	3.73%	…	375
	a	0.00%	1.22%	0.00%	9.79%	0.61%	0.00%	0.00%	0.61%	0.00%	6.12%	…	327
	no	0.00%	0.91%	0.45%	1.82%	1.36%	1.82%	0.45%	0.00%	1.36%	0.00%	…	220
	con	0.00%	2.96%	0.00%	8.87%	8.87%	0.00%	0.00%	0.00%	0.00%	1.97%	…	203
	tu	0.00%	1.15%	1.15%	0.00%	0.00%	0.57%	0.57%	0.57%	0.00%	0.57%	…	174
	mi	0.00%	2.45%	0.61%	0.61%	0.61%	0.61%	0.61%	0.61%	0.00%	0.00%	…	163
	por	0.00%	9.03%	0.00%	7.10%	1.94%	0.00%	0.00%	1.94%	0.00%	4.52%	…	155
	…	…	…	…	…	…	…	…	…	…	…	…	

Tabla 32.Matriz transición Jerusalem.

Ilustración 18.Matriz de transición "LA NUMANCIA" y "JERUSALEN". (de creación propia)

Se presenta las matrices de transición ya calculadas anteriormente. Las cuales siguen el estadístico de prueba planteado en el marco teórico. Para el estudio se plantea el uso de los primeros 105 niveles o palabras más usadas

Para el cual es necesario las matrices de transición.

Y la operación a partir ellos

$$T_{Jerusalen} * T_{Numancia}^{-1}$$

De la cual se obtiene la siguiente matriz. De la cual se presentan las 10 primeras columnas y 20 filas de las 105*105 se trabajó en orden alfabético y de las mismas categorías para todas las matrices.

	a	acá	al	alma	amor	aquel	aquí	aunque	Aurelio	bien	
a	2.96823	-0.66206	2.24788	-0.66206	0.01920	-0.84210	-1.45785	-0.80116	-0.66206	-1.27753	…
acá	-0.13803	0.98926	-0.09560	-0.01074	-0.07539	0.00240	0.00211	0.01071	-0.01074	0.10007	…
al	-0.27260	0.16086	-0.18744	0.16086	1.33335	0.13561	0.08929	0.29610	0.16086	1.44724	…
alma	-0.13803	-0.01074	-0.09560	0.98926	-0.07539	0.00240	0.00211	0.01071	-0.01074	0.10007	…
amor	0.04801	0.17531	0.09045	0.17531	0.92461	0.18845	0.18815	0.19676	0.17531	0.28612	…
aquel	-1.01303	0.05176	-0.03310	0.05176	-0.01289	1.00240	0.06461	0.07321	0.05176	-0.77493	…
aquí	-4.67768	0.85781	-3.16881	0.85781	0.20917	0.82608	1.53743	1.20010	0.85781	2.07713	…
aunque	2.33060	-0.37847	2.44542	-0.37847	-1.47058	-0.44341	-0.68333	-0.14941	-0.37847	-2.21270	…
Aurelio	-0.13803	-0.01074	-0.09560	-0.01074	-0.07539	0.00240	0.00211	0.01071	0.98926	0.10007	…
bien	2.95863	-0.48025	2.18728	-0.48025	-0.45055	-0.63065	-0.57318	-0.76844	-0.48025	-1.90397	…
cielo	-0.80470	0.15593	-0.76227	0.15593	0.09128	0.16907	0.16877	0.17738	0.15593	0.26674	…
ciudad	0.02863	0.15593	0.07107	0.15593	0.09128	0.16907	0.16877	0.17738	0.15593	0.26674	…
Clorinda	0.06044	0.18774	0.10287	0.18774	0.12309	0.20088	0.20058	0.20919	0.18774	0.29855	…
como	-2.06100	0.14185	-1.88314	0.14185	0.47277	0.28550	-0.05570	0.45776	0.14185	2.08754	…
con	-2.22889	0.45531	-1.28287	0.45531	-0.04814	0.51955	0.77314	0.62657	0.45531	0.54827	…
cristiano	-0.13803	-0.01074	-0.09560	-0.01074	-0.07539	0.00240	0.00211	0.01071	-0.01074	0.10007	…
cual	-5.38686	1.43723	-5.04939	1.43723	0.10904	1.72025	3.54705	1.83320	1.43723	1.49578	…
dar	0.97202	-0.33933	0.48275	-0.33933	-0.32576	-0.46710	-0.57557	-0.36395	-0.33933	-0.58062	…
de	0.03388	-0.00168	-0.03036	-0.00168	-0.00176	-0.00115	0.00695	-0.00093	-0.00168	0.00007	…
decir	0.09554	0.22284	0.13798	0.22284	0.15819	0.23598	0.23568	0.24429	0.22284	0.33365	…
	…	…	…	…	…	…	…	…	…	…	

*Tabla 40.Matriz NUMANCIA * JERUSALEN. (de creación propia)*

Donde se calcula la traza y determinante de valores requerido para la prueba planteada con ayuda de paquete estadístico R

Para el cálculo de los valores críticos según la prueba es donde n=p:

$$m = \frac{1}{2}p(p+1)$$

n	p	alfa	m	chi de tabla critico
105	105	0.95	5565	5392.614

Tabla 41.Valor critico prueba NUMANCIA JERUSALEN. (de creación propia)

El estadístico de prueba es:

$$\chi^2_{calculado} = np(a - ln(g) - 1) \sim \chi^2_m$$

n	Determinante	g	Traza	a	chi calculado	p valor
105	0.06754578	0.97466	106.342	1.009524	387.9697	0.999

Tabla 42.Valor calculado prueba NUMANCIA JERUSALEN. (de creación propia)

Dado que el valor calculado es inferior al crítico podemos afirmar que la matriz de transición A es igual a la matriz de transición B.

Chi- cuadrado (ARGEL-JERUSALEN) = 387.96

El valor es inferior al valor critico dado por la prueba NUMANCIA JERUSALEN de 1425 indicando que la obra sea de Cervantes.

5.2.5. “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON”- “EL TRATO DE ARGEL”- “LA NUMANCIA”

Tabla agregada vocabulario obra. Para este estudio se usó las 50 palabras más usadas por las obras esto con fin de ver si comparten el vocabulario en común.

		JERUSALEM		NUMANCIA		ARGEL	
id	**palabra**	**n**	**%**	**n**	**%**	**n**	**%**
1	que	772	4.63%	771	4.89%	742	4.93%
2	y	794	4.76%	669	4.25%	665	4.42%
3	de	552	3.31%	603	3.83%	514	3.42%
4	el	416	2.49%	431	2.73%	400	2.66%
5	la	433	2.60%	427	2.71%	317	2.11%
6	en	386	2.31%	379	2.40%	324	2.15%
7	a	341	2.04%	272	1.73%	301	2.00%
8	no	254	1.52%	224	1.42%	265	1.76%
9	con	208	1.25%	208	1.32%	138	0.92%
10	se	150	0.90%	151	0.96%	176	1.17%
11	mi	164	0.98%	134	0.85%	159	1.06%
12	por	167	1.00%	136	0.86%	145	0.96%
13	es	125	0.75%	85	0.54%	191	1.27%
14	si	159	0.95%	110	0.70%	130	0.86%
15	me	130	0.78%	101	0.64%	160	1.06%
16	tu	178	1.07%	100	0.63%	104	0.69%
17	mas	112	0.67%	135	0.86%	117	0.78%
18	los	120	0.72%	129	0.82%	102	0.68%
19	al	105	0.63%	114	0.72%	106	0.70%
20	lo	111	0.67%	73	0.46%	99	0.66%
21	esta	102	0.61%	81	0.51%	96	0.64%
22	yo	108	0.65%	77	0.49%	79	0.53%
23	su	72	0.43%	81	0.51%	107	0.71%
24	las	101	0.61%	99	0.63%	59	0.39%
25	te	105	0.63%	54	0.34%	84	0.56%

Tabla 43.Vocabulario en las tres obras. (de creación propia)

Prueba de independencia entre las vocabulario y obras (Chi-cuadrado):

Chi-cuadrado (Valor observado)	Chi-cuadrado (Valor crítico)	GL	valor-p	alfa
326.437	122.108	98	< 0.0001	0.05

Tabla 44.Prueba de hipótesis vocabulario obras. (de creación propia)

Interpretación de la prueba:

Ho: Las filas y las columnas de la tabla son independientes. (no existe relación entre los vocabularios y las obras)

Ha: Hay dependencia entre las filas y las columnas de la tabla. (existe relación entre los vocabularios y las obras)

Puesto que el valor-p calculado es menor que el nivel de significación α=0.05, se debe rechazar la hipótesis nula H0, y aceptar la hipótesis alternativa Ha. El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es inferior al 0.001%. existe relación entre los vocabularios de las obras"

Análisis de correspondencia. Inercia total 0.015

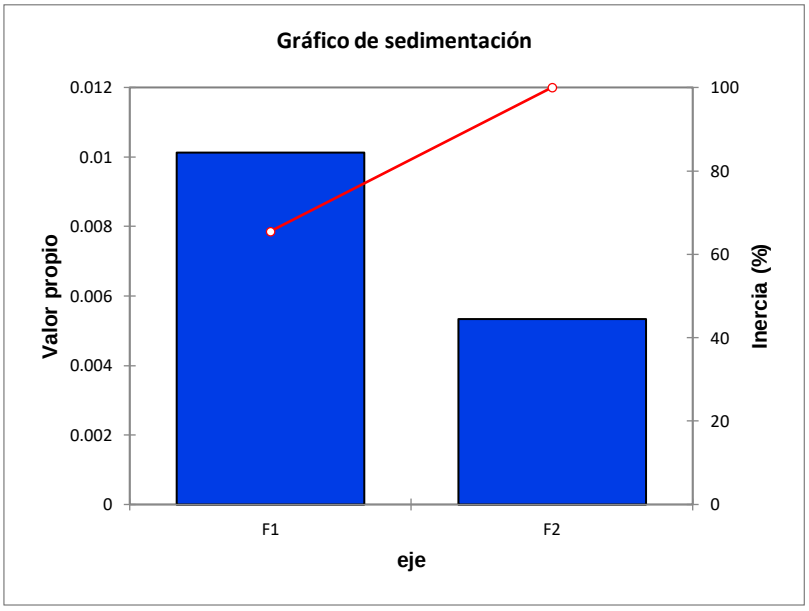


Ilustración 19.Inercia vocabulario obras. (de creación propia)

- El valor explicado por los ejes es el máximo, pero es de una inercia general muy baja.

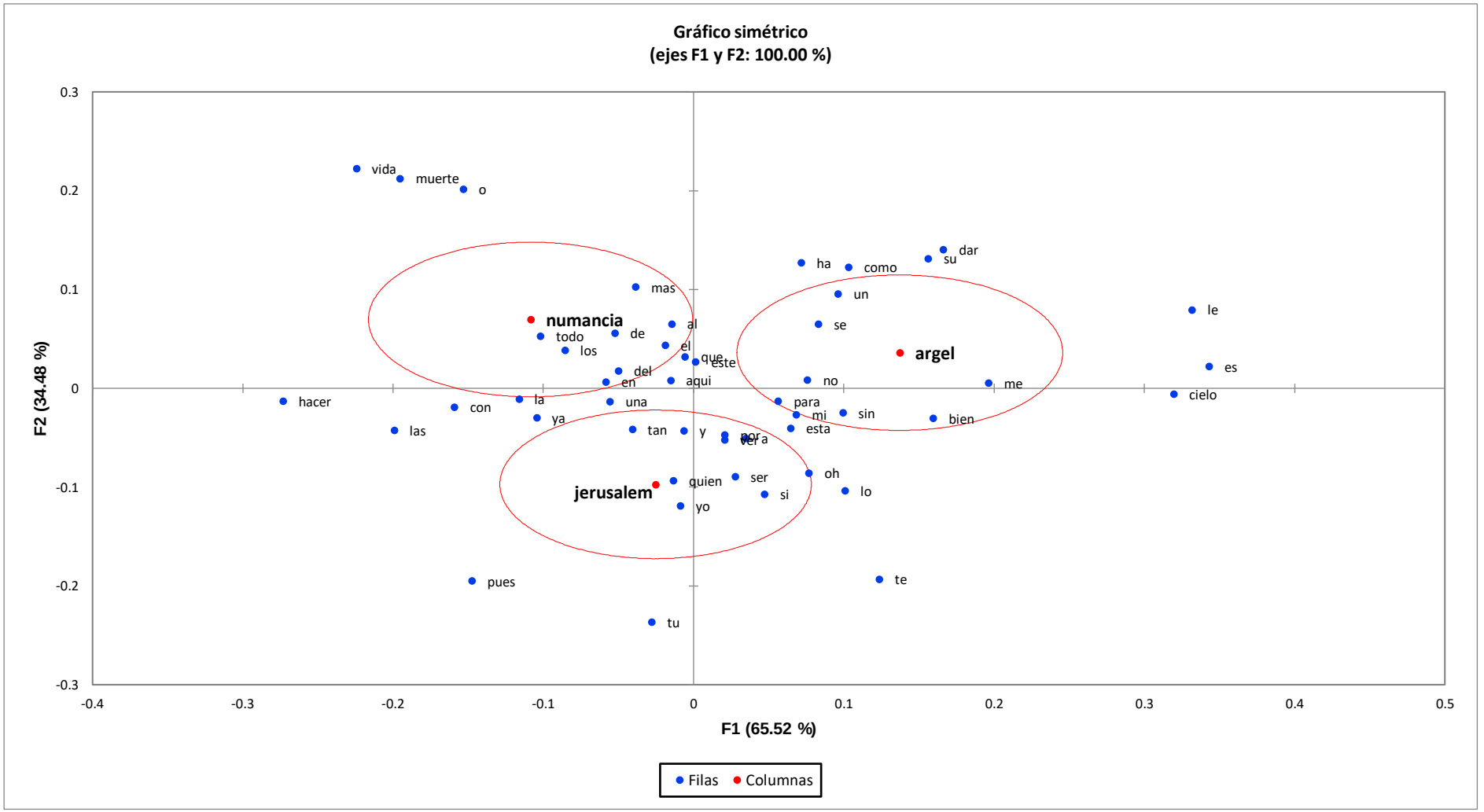


Ilustración 20.Analisis de correspondencia obra vocabulario. (de creación propia)

Para la interpretación del grafico debemos tener en consideración que las distancias no se miden entre dos filas o dos columnas si no como relación al perfil media de fila o columna, es decir con relación al promedio de las coordenadas de estas filas o columnas ponderada por su masa (peso proporcional a su importancia en el conjunto). Es conocido como centro de gravedad la media de las distancias al cuadrado de cada punto de fila al centro de gravedad se conoce como inercia de fila o inercia de columna cuando se trata de las columnas e inercia total de la nube de puntos cuando se consideran todos los elementos de la tabla, para mayor descripción del proceso revisar pagina 47.

Interpretación análisis de correspondencia obras vocabulario.

- Se aprecia que la obra “LA NUMANCIA” tiende al mayor uso de las palabras MAS TODO LOS DE en comparación de las otras obras.
- Se aprecia que la obra “EL TRATO DE ARGEL” tiende al mayor uso de las palabras UN SE NO PARA SIN MI BIEN ME en comparación de las otras obras.
- Se aprecia que la obra “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON” tiende al mayor uso de las palabras TAN QUIEN YO SER SI Y en comparación de las otras obras.
- Se aprecia que cada obra tiene una característica especial en su vocabulario

5.2.6. Inferencia sobre la matriz de transición.

Se plantea la solución a la hipótesis general, tomando como patrón de comparación la similitud existente entre las obras reconocidas Argel y Numancia que se toma como valor crítico para la comparación. Mediante la solución a las hipótesis especificas ya resueltas.

H_0: "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON" se puede atribuir a Miguel de Cervantes Saavedra mediante el análisis estadístico.

H_a: No se puede atribuir la obra "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON" a Miguel de Cervantes Saavedra.

	chi calculado	p valor	CONCLUSION
Valor crítico para las pruebas específicas a un alfa del 95%	5392.6	0.950	Todas las pruebas planteadas son inferiores al valor critico estadístico
Valor calculado para la hipótesis ARGEL NUMANCIA	1425.0	0.999	la similitud entre dos obras reconocidas de cervantes es de 99% con un chi cuadrado de 1425
Valor calculado para la hipótesis ARGEL JERUSALEN	1262.3	0.999	la similitud de ARGEL y JERUSALEM es de un equivalente chi de 1262
Valor calculado para la hipótesis NUMANCIA JERUSALEN	387.9	0.999	la similitud de NUMANCIA y JERUSALEM es de un equivalente chi de 387

Tabla 45. Inferencia sobre matriz de transición

Dado que la similitud entre las matrices de transición es muy fuerte superando al valor crítico y al valor critico referencial ARGEL NUMANCIA se acepta la hipótesis nula. "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON" se puede atribuir a Miguel de Cervantes Saavedra.

5.2.7. Matriz de transición antecede y precede general de las tres obras.

Una vez de definida la hipótesis se presenta la matriz de transición general considerando las tres obras. Esto con fin de ampliar la discusión.

		Palabra precedente												
		que	y	de	el	la	en	a	no	con	se	mi	mas	…
Palabra antecede	que	0.13%	0.13 %	5.55%	9.64%	8.19%	13.08 %	9.25 %	15.46 %	4.36 %	9.78%	2.11%	2.51 %	…
	y	8.86%	0.00 %	8.56%	10.51 %	8.56%	12.31 %	9.16 %	6.61%	6.46 %	1.65%	2.55%	3.90 %	…
	de	6.02%	0.00 %	0.00%	0.17%	21.24 %	0.67%	0.00 %	0.84%	0.17 %	0.00%	14.55 %	0.84 %	…
	el	9.18%	0.94 %	0.94%	0.00%	0.47%	0.00%	0.24 %	1.18%	0.24 %	1.41%	0.00%	0.94 %	…
	la	4.24%	0.00 %	1.41%	0.00%	0.00%	0.00%	0.00 %	0.24%	0.00 %	0.00%	0.00%	1.41 %	…
	en	5.96%	0.00 %	0.00%	30.08 %	33.33 %	0.00%	0.00 %	0.27%	0.00 %	0.00%	9.76%	0.54 %	…
	a	5.26%	0.00 %	0.00%	1.88%	28.20 %	0.00%	0.00 %	0.75%	0.00 %	0.00%	25.56 %	0.75 %	…
	no	1.42%	0.00 %	3.79%	2.84%	6.64%	3.79%	1.42 %	0.95%	3.32 %	24.64 %	0.47%	3.79 %	…
	con	11.33 %	0.00 %	0.00%	23.65 %	19.21 %	0.00%	0.00 %	0.99%	0.00 %	0.00%	3.45%	3.45 %	…
	se	15.23 %	0.00 %	1.32%	0.00%	0.66%	0.00%	0.66 %	0.00%	0.00 %	0.66%	0.00%	0.66 %	…
	mi	4.48%	0.75 %	1.49%	2.24%	2.99%	1.49%	1.49 %	0.75%	1.49 %	1.49%	0.00%	1.49 %	…
	mas	22.56 %	0.00 %	13.53 %	4.51%	3.01%	3.01%	5.26 %	12.03 %	1.50 %	1.50%	2.26%	0.00 %	…
	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Tabla 46.Matriz transición estocástica vocabulario cervantes. (de creación propia)

La matriz completa está constituida por 105*105 datos del cuales contienen el vocabulario posiblemente usado por cervantes este dato de matriz de transición será usado para crear la discusión de una posible frase pronunciada por cervantes y con qué posibilidad de decirla existió.

5.3. CONCLUSIONES, DISCUSIÓN Y SUGERENCIAS.

5.3.1. Discusión

En el estudio realizado por Estefano Arata “La conquista de Jerusalén Cervantes y la Generación Teatral 1580” 1992. Afirmo haber encontrado la obra perdida la Jerusalem de Miguel de Cervantes Saavedra mediante la comparación del estilo literario de algunas aproximaciones metricaS LINGUISTICAS a lo cual nuestro estudio confirma mediante el uso de técnicas descriptivas y correlacionales multivariadas.

En su estudio de Juan cerezo Soler “La conquista de Jerusalén en su contexto: sobre el personaje colectivo y una vuelta más a la atribución cervantina” 2014. En donde el autor haciendo uso de herramientas lingüísticas mantiene una posición más neutral acerca de la atribución de la obra teatral impuesta por Stefano Arata, en lo cual debió compartir opinión con Stefano Arata ya que en el presente estudio da evidencia cuantificable de asociación categórica de las obras estudiadas.

La Dra. Lourdes Albuisech de la universidad de Illinois en su estudio “Mezclar verdades con fabulosos intentos: Metateatro y aporía en el Gallardo español de cervantes”, estudio donde da su opinión acerca de la autoría donde indica fehacientemente que la obra “la conquista de Jerusalén” no es de Cervantes, a lo que el presente estudio tiene una opinión errada.

5.3.2. Conclusiones:

- Si existe similitud en el tamaño de las palabras por obra muy fuerte encontrando que las palabras más usadas en las obras son las de 2 caracteres representando en “EL TRATO DE ARGEL” un 15.31% “LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON”15.14% y “LA NUMANCIA”15.20% siendo muy semejantes estos porcentajes y siguiendo esta tendencia el resto de amplitudes de palabras.
- La prueba estadística de chií cuadro de muestra que existe relación significativa entre los personajes y el vocabulario que usan. el análisis de correspondencia permite visualizar las relaciones encontrando. Para el estudio se considera las líneas de definición del escenario como a un personaje extra en el análisis estos se separaron generando un grupo independiente indicando que las formas de definir el escenario en las tres obras se comparten. Mediante el análisis de correspondencia se aprecia que muchos de los personajes de las tres obras están muy enlazados por su vocabulario lo cual se detalla en la página 101.
- Si existe semejanza en el léxico usado en las tres obras en estudio a pares y las tres juntas mediante la prueba chi cuadrado
- Si existe similitud entre las matrices de transición muy fuerte superando al valor crítico y al valor critico referencial definido por la comparación de las obras reconocidas “EL TRATO DE ARGEL” y “LA NUMANCIA” se acepta la hipótesis nula. “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” se puede atribuir a Miguel de Cervantes Saavedra. Mediante medios estadísticos desde un enfoque estocástico. encontrado que las obras que presentan mayor similitud entre ellas mediante el estadístico propuesto son “LA NUMANCIA” y “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON”.

5.3.3. Comentarios.

- En los descriptivos grupales para las tres obras. Se apreciar una similitud por cantidad de palabras en las obras siendo de 15047 en “EL TRATO DE ARGEL”, de 16684 en “LA CONQUISTA DE JERUSAELN POR GODOFRE BULLON” y de 15759 en “LA NUMANCIA” que en porcentajes no representa una diferencia notable. También así en la cantidad de palabras por jornada que en porcentajes el menor es de 6% y en su mayoría de un 8% de la totalidad de la base de datos analizada siendo un caso de discusión el 17% de la tercera jornada de “LA CONQUISTA DE JERUSAELN POR GODOFRE BULLON. Es casi el doble del usado en las jornadas del resto de obras. lo cual da un indicio a la teoría planteada por Stefano Arata. De que esta tercera jornada pudo ser la unión de una tercera y cuarta presuntamente escrita esto con fin de adatarla al estilo de la época.
- La manera de uso de palabras y riqueza del vocabulario cervantino sigue la distribución de zip distribución de palabras que aparece por ancho de cárteres esto debido a que en los lenguajes naturales la evolución del mismo obliga al mayor uso de palabras simples de dos caracteres esto por economizar el uso de caracteres
- Descriptiva “EL TRATO DE ARGEL” se encuentra que la palabra más representativa es la “que” representando en 4.93% del texto seguido de “y” en un 4.42% estas siendo palabras sin un significado completo.

  En las palabras con significado completo encontramos que en “EL TRATO DE ARGEL” la primera en aparecer es “bien”, “sin”, “cristiano” por lo cual da entender el sentimiento de bienestar y religiosidad. Las cuales se pueden apreciar tanto en la tabla o grafico de nube de palabras

  La matriz de palabras antecede y precede para argel permite apreciar como es la relación de las palabras en la obra “EL TRATO DE ARGEL” encontrando que existen muchas palabras que se relacionan y que presentan una relación lógica como el caso de “que en” que se repite 31 veces en el texto.

  “en el” repitiéndose 41 veces encontrando que tienen una relación estadística y lógica lingüística.

- Descriptivos “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” se encuentra que la palabra más representativa es la “que” representando en 4.63% del texto seguido de “y” en un 4.76% estas siendo palabras sin un significado completo. En las palabras con significado completo encontramos que en Jerusalem la primera en aparecer es “bien”, “sin”, “ciudad” por lo cual da entender el sentimiento muy parecido al narrado en “ELTRATO DE ARGEL” quizá esto por los dos tratarse del mismo genero la comedia. Las cuales se pueden apreciar tanto en la tabla o grafico de nube de palabras

  La matriz de palabras antecede y precede para “LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON” permite apreciar como es la relación de las palabras en la obra Jerusalem encontrando que existen muchas palabras que se relacionan y que presentan una relación lógica como el caso de “que la” que se repite 32 veces en el texto. “y que” repitiéndose 21 veces encontrando que tienen una relación estadística y lógica lingüística
- Descriptivos “LA NUMANCIA” se encuentra que la palabra más representativa es la “que” representando en 4.89% del texto seguido de “Y” en un 4.25% estas siendo palabras sin un significado completo. En las palabras con significado completo encontramos que en “LA NUMANCIA” la primera en aparecer es “vida”, “muerte”, “bien” lo cual da entender el sentimiento es trágico por ser tragedia. Las cuales se pueden apreciar tanto en la tabla o grafico de nube de palabras

  La matriz de palabras antecede y precede para Numancia permite apreciar como es la relación de las palabras en la obra Numancia encontrando que existen muchas palabras que se relacionan y que presentan una relación lógica como el caso de “que no” que se repite 36 veces en el texto. “no se” repitiéndose 17 veces encontrando que tienen una relación estadística y lógica lingüística
- Se aprecia notable diferencia entre la nube de palabras de “LA NUMANCIA” en comparación de “ELTRATO DE ARGEL” y “LA CONQUISTA DE JERUSALEN POR GOFOFRE BULLON” cuales nubes son muy parecidas esto es posible debido a que “LA NUMANCIA” es una tragedia en comparación con “ELTRATO DE ARGEL” y “LA CONQUISTA

DE JERUSALEN POR GOFOFRE BULLON” las cuales son clasificadas como comedias.

- La matriz de antecede y precede textual de la obra es un buen indicador de como escribe una persona a pesar de solo contener memoria de la palabra que lo precede.
- La interpretación de asociación de los vocabularios no es del todo certera esto debido a que las norma el mismo idioma con un uso alto de conectores por estas razones se propone la comparación de las matrices de transición estocástica antecede y precede de cada obra.

5.3.4. Sugerencias.

- Se sugiere a la comunidad científica, el estudio e indagación en la creación de un indicador de riqueza lingüística basado en la amplitud de las palabras frecuencia de las palabras en un texto esto con el fin de poder dar una expresión más sencilla diferenciando una, literatura culta y una de popular. por medio de un indicador, asi como estudiar el proceso de lematizado ya que este proceso es subjetivo y no presenta el rigor científico ni estadístico y es usado en todo análisis de texto
- Se sugiere a la facultad de Ciencias el incentivar el estudio de análisis de sentimientos, técnica estadística de minería de texto que agrupa el contenido de libros, de acuerdo al sentimiento que quiso expresar en el texto, esto con el fin de que los estudiantes tengan una mejor incursión laboral en los nuevos ámbitos de estudio.
- Se sugiere a los tesistas hacer uso de la técnica de nube de palabras ay que es un buen grafico estadístico para la representación de tablas de frecuencia muy grande.
- Se sugiere a los lingüistas tomar en consideración la técnica estadística de la matriz de palabras antecede y precede ya que presenta una fuente de información de cómo es que se relacionan las palabras en el idioma español lo cual permitiría caracterizar las estructuras semánticas como los artículos precepciones verbos de una manera estadista en razón de una formación en la matriz antecede precede de lenguaje. También la matriz antecede y precede a pesar de contener información de solo un tiempo atrás puede ser una forma de guardar información importante.

Bibliografía.

Albuisech, L. (s.f.). *Mezclar Verdades Con Fabulosos Intentos: Meta teatro Y Aporía En El Gallardo Español De Cervantes.*

Aragon, J. L. (s.f.). *Mtrices.* Obtenido de http://joseluislorente.es/2bac/temas/tema8.pdf

Arata, S. (1992). "La Conquista de Jerusalén, Cervantes y la generación teatral de 1580". *CRITICON*.

Arroyo, S. C. (2010). *Analis de corespondencia simple.* Universidad de Valencia.

Biografias. (s.f.). *markov*. Obtenido de https://www.biografiasyvidas.com/biografia/m/markov.htm

Cabrera, I. d. (s.f.). *ESTAD´ISTICA APLICADA Y MODELIZACION. I.T. DISE ´ NO INDUSTRIAL.*

Cerezo, J. (s.f.). *La Conquista de Jerusalén en su contexto: sobre el personaje colectivo y una vuelta más a la atribución cervantina.*

clubdelecturavalladolid. (s.f.). *club de lectura valladolid.* Obtenido de Club Marcapáginas: Novelas ejemplares, de Miguel de Cervantes: https://clubdelecturavalladolid.wordpress.com/2017/04/19/club-marcapaginas-novelas-ejemplares-de-miguel-de-cervantes/

Diazaraque, J. M. (s.f.). *Cadenas de Markov.* Obtenido de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/PEst/tema4pe.pdf

Diazaraque, J. M. (s.f.). *Tablas de Contingencia.* Obtenido de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Categor/Tema2Cate.pdf

F., J. A. (s.f.). *Introducción a las Cadenas o Procesos de Markov*. Obtenido de http://www.ingenieria.unam.mx/javica1/ingsistemas2/Simulacion/Cadenas_de_Markov.htm

Félix, L. C. (s.f.). *Data mining: torturando a los datos hasta que confiesen.* Obtenido de http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

Fernandez, S. d. (2011). *Analisis de correspondencia simple y multiples.* madrid.

Gabancho, .. J. (s.f.). *"Modelo de Redes con recursos didácticos con lingüística computacional".*

Galeon. (s.f.). Obtenido de http://textmining.galeon.com/

Guillen., B. L. (2010). *Análisis factorial de correspondencias de pacientes con patologías oculares en el CEPRECE-CUSCO.*

JesToryAs. (abril de 2016). *¡Celebramos los 400 años del fallecimiento del Gran Don Miguel de Cervntes Saavedra!*

Lozano, E. (s.f.). Obtenido de http://asignatura.us.es/dadpsico/apuntes/EpChiCuadrado.pdf

Perez, C. (s.f.). *El concepto de corpus y su definición.* Obtenido de http://elies.rediris.es/elies18/23.html

Portega. (s.f.). *fundamentos matrices.* Obtenido de https://www.es/personalpdc/economicas/portega/fundamentos-matrices.

Proceso-estocastico. (s.f.). *Proceso-estocastico*. Obtenido de https://es.scribd.com/document/359456707/Proceso-estocastico-pdf

Rincón, L. (2011). *Introduccion a los procesos estocasticos.* mexico: UNAM.

Savedra, M. C. (1600). *Adjunta al parnaso.* alcala.

Schaum, J. M. (4 edicion). *Estadistica Schaum.* Mc Graw Hi.

Sellero, C. S. (2008). *Inferencia en poblaciones normales multivariantes.* Obtenido de http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP/MATERIALESMASTER/Mat_14_master0809multi-tema2.pdf

SOLER, J. C. (s.f.). La Conquista de Jerusalén en su contexto:.

UNAM. (s.f.). *Universidad Nacional Autónoma de México.* Obtenido de Laboratorio virtual de estadistica INFERENCIA ESTADISTICA: http://asesorias.cuautitlan2.unam.mx/Laboratoriovirtualdeestadistica/CARPETA%203%20INFERENCIA_ESTADISTICA/DOC_%20INFERENCIA/TEMA%204/08%20PRUEBA%20DE%20CHICUADRADA.pdf

Universidad de sevilla, O. c. (s.f.). material de clase estadistica e investiagcion de operaciones. *Nociones fundamentales de cálculo de probabilidades*.

Valle, J. A. (s.f.). *Introducción a las Cadenas o Procesos de Markov*. Obtenido de http://www.ingenieria.unam.mx/javica1/ingsistemas2/Simulacion/Cadenas_de_Markov.htm

vidas, b. y. (s.f.). *markov*. Obtenido de https://www.biografiasyvidas.com/biografia/m/markov.htm

Visauta, V. (1998). *Análisis estadístico con SPSS para Windows: estadística multivariante.* McGraw-Hill.

Wiki. (s.f.). *traza*. Obtenido de https://www.wikipedia.com/es/Traza_(%C3%A1lgebra_lineal)

wordreference. (mayo de 2018). Obtenido de http://www.wordreference.com/definicion/pastiche

ANEXOS

a. PRESUPUESTO.

CONCEPTO	UNIDAD	COSTO UNITARIO (SOLES)	CANTIDAD	COSTO PARCIAL (SOLES)	OBSEVACIONES
Recolección de bibliografía y Desarrollo del marco teórico	jornales	60	10	600	incluye trasporte refrigerios y tiempo del tesista
Reconexión de muestra y procesado en paquete estadístico	jornales	60	10	600	incluye trasporte refrigerios y tiempo del tesista
Análisis y redacción de informe	jornales	60	10	600	incluye trasporte refrigerios y tiempo del tesista
Depreciación de computadora	mes	108	60	216	*(2000 soles-700 soles) /24 meses
Compra de paquete estadístico	unidad	300	1	300	
Impresión de tesis y papelería general		1000		1000	
Imprevistos	%	10		331,6	
TOTAL				**3647,6**	

b. Matriz de consistencia.

PROBLEMA GENERAL	OBJETIVO GENERAL	HIPOTESIS GENERAL	METODOLOGIA
¿Qué semejanzas existe, entre las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", ¿"LA NUMANCIA" y "EL TRATO DE ARGEL" obtenidas mediante el análisis de correspondencia?	Analizar semejanzas de las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL" obtenidas mediante el análisis de correspondencia.	Existe similitud entre las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL" obtenidas mediante el análisis de correspondencia.	Se compara si existe semejanzas entre los corpus textuales de las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "EL TRATO DE ARGEL" "LA NUMANCIA"
PROBLEMAS ESPECIFICOS	**OBJETIVOS ESPECIFICOS**	**HIPOTESIS ESPECIFICOS**	**METODOLOGIA**
¿Qué semejanza de léxico existe entre las obras "¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL"?	Establecer similitud en el tamaño de palabras en las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL"	Existe similitud en el tamaño de palabras en las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL"	Se seguirá los pasos de recolección de información limpieza de la información lematizado y tokenizado lo cual servirá para crear la base de datos o corpus textual. Para la evaluación se desarrollará la tabla adjunta del léxico por obra estudiada el cual será sometida las pruebas estadísticas de asociación chi cuadrado, análisis de correspondencia e inferencia sobre la matriz de transición lexical
¿Existe semejanza de personajes en las obras "¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL?	¿Determinar semejanza de personajes en las obras "¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL?	¿Existe semejanza de personajes en las obras "¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL?	
¿Existe similitud en el tamaño de palabras en las obras "¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL"?	Describir las semejanzas de léxico existe entre las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL"	Existe semejanza de léxico existe entre las obras "LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL"	
¿Existe semejanza entre las matrices de transición precede y antecede de las obras" ¿LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL?	Evaluar las semejanzas entre las matrices de transición precede y antecede de las obras" LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL	Existe semejanza entre las matrices de transición precede y antecede de las obras" LA CONQUISTA DE JERUSALÉN POR GODOFRE BULLON", "LA NUMANCIA" y "EL TRATO DE ARGEL	

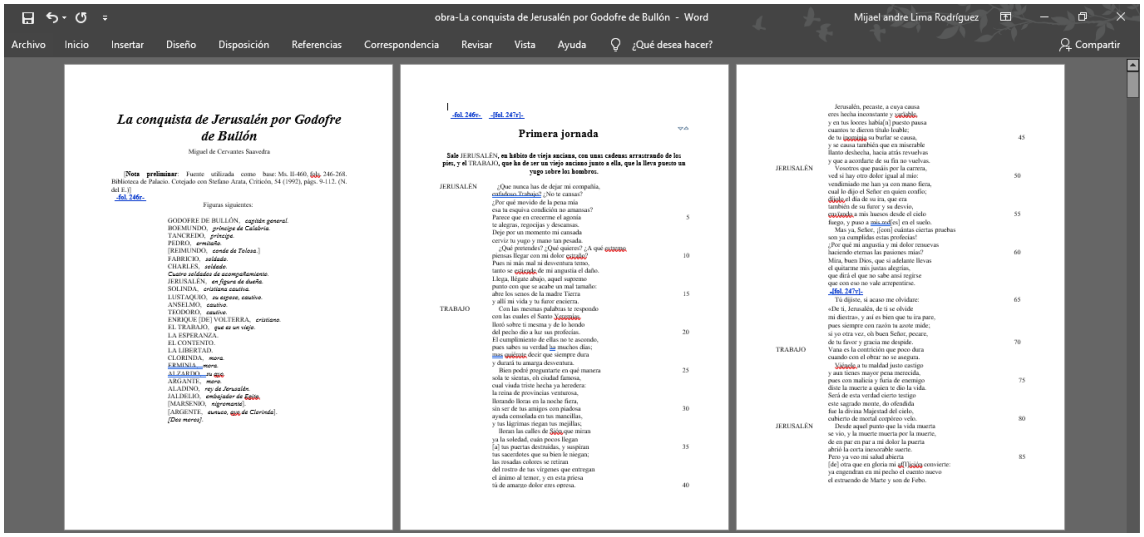
Tabla 47. Matriz de consistencia general. (de creación propia)

c. Métodos usados.

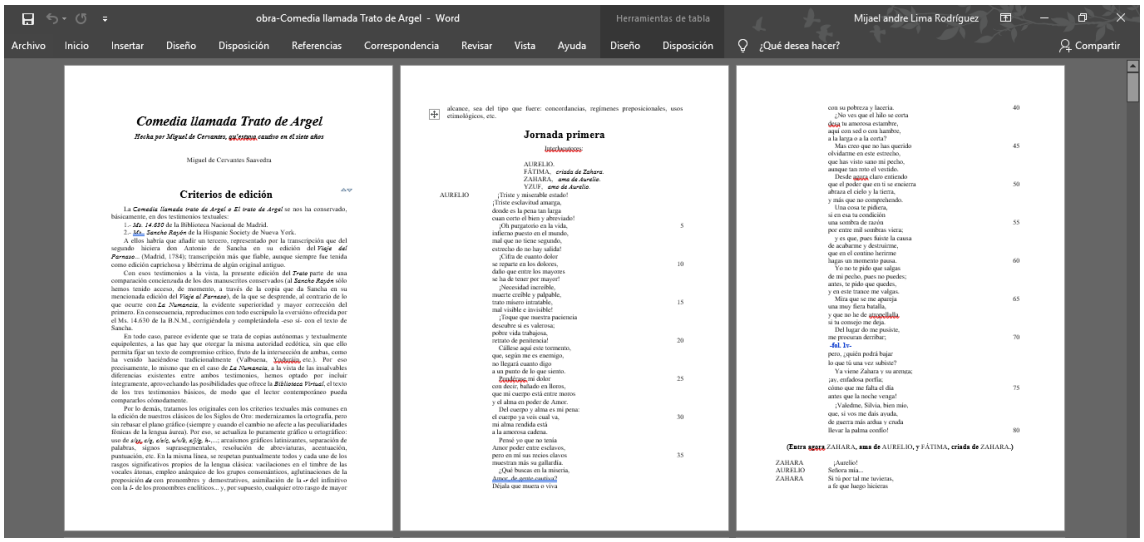
Para el desarrollo de los cuadros y pruebas en la tesis se ha hecho uso del software libre de análisis de datos R y R studio como así el paquete estadístico XLstat en su versión de prueba por ser un excelente graficador de cuadros.

- Los datos obtenidos de la Fuente digital biblioteca virtual cervantes.

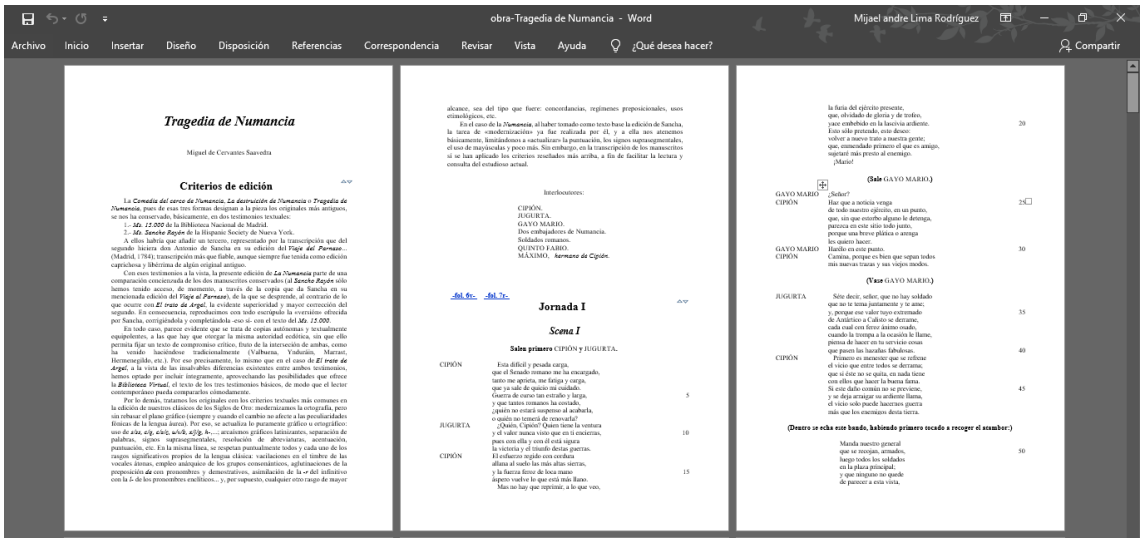
La conquista de Jerusalén textual.



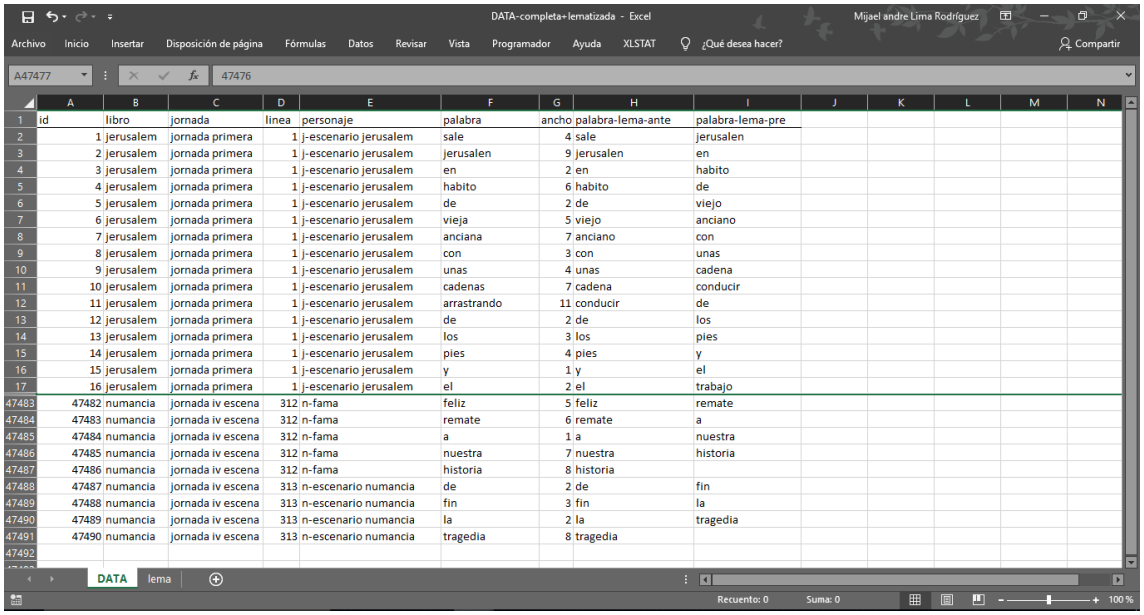
El trato de argel textual.



La Numancia.



- Los cuales se quitará los signos de puntuación mayúsculas y las pares de introducción también así la numeración se genera la base de datos conservando el personaje libro.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	libro	jornada	linea	personaje	palabra	ancho	palabra-lema-ante	palabra-lema-pre					
2	1	jerusalem	jornada primera	1	j-escenario jerusalem	sale	4	sale	jerusalen					
3	2	jerusalem	jornada primera	1	j-escenario jerusalem	jerusalen	9	jerusalen	en					
4	3	jerusalem	jornada primera	1	j-escenario jerusalem	en	2	en	habito					
5	4	jerusalem	jornada primera	1	j-escenario jerusalem	habito	6	habito	de					
6	5	jerusalem	jornada primera	1	j-escenario jerusalem	de	2	de	viejo					
7	6	jerusalem	jornada primera	1	j-escenario jerusalem	vieja	5	viejo	anciano					
8	7	jerusalem	jornada primera	1	j-escenario jerusalem	anciana	7	anciano	con					
9	8	jerusalem	jornada primera	1	j-escenario jerusalem	con	3	con	unas					
10	9	jerusalem	jornada primera	1	j-escenario jerusalem	unas	4	unas	cadena					
11	10	jerusalem	jornada primera	1	j-escenario jerusalem	cadenas	7	cadena	conducir					
12	11	jerusalem	jornada primera	1	j-escenario jerusalem	arrastrando	11	conducir	de					
13	12	jerusalem	jornada primera	1	j-escenario jerusalem	de	2	de	los					
14	13	jerusalem	jornada primera	1	j-escenario jerusalem	los	3	los	pies					
15	14	jerusalem	jornada primera	1	j-escenario jerusalem	pies	4	pies	y					
16	15	jerusalem	jornada primera	1	j-escenario jerusalem	y	1	y	el					
17	16	jerusalem	jornada primera	1	j-escenario jerusalem	el	2	el	trabajo					
47483	47482	numancia	jornada iv escena	312	n-fama	feliz	5	feliz	remate					
47484	47483	numancia	jornada iv escena	312	n-fama	remate	6	remate	a					
47485	47484	numancia	jornada iv escena	312	n-fama	a	1	a	nuestra					
47486	47485	numancia	jornada iv escena	312	n-fama	nuestra	7	nuestra	historia					
47487	47486	numancia	jornada iv escena	312	n-fama	historia	8	historia						
47488	47487	numancia	jornada iv escena	313	n-escenario numancia	de	2	de	fin					
47489	47488	numancia	jornada iv escena	313	n-escenario numancia	fin	3	fin	la					
47490	47489	numancia	jornada iv escena	313	n-escenario numancia	la	2	la	tragedia					
47491	47490	numancia	jornada iv escena	313	n-escenario numancia	tragedia	8	tragedia						
47492														

- Los datos son puestos en una hoja Excel para ser ingresado en el paquete de análisis R studio para la creación de las tablas.

Data<- read_exel(“D:/../data-sinlematizar.xlsx”)

attach(data)

tabla-frecuencia-por-obra<- sumary(split(data$libro(Split(data$jornada))

tabla-ancho-palabra<-table(data$libro,data$ancho)

chisq.test(tabla -ancho-palabra)

tabla-frecuencia-sin-lematizada<-sumary(data$palabra-lema-ante)

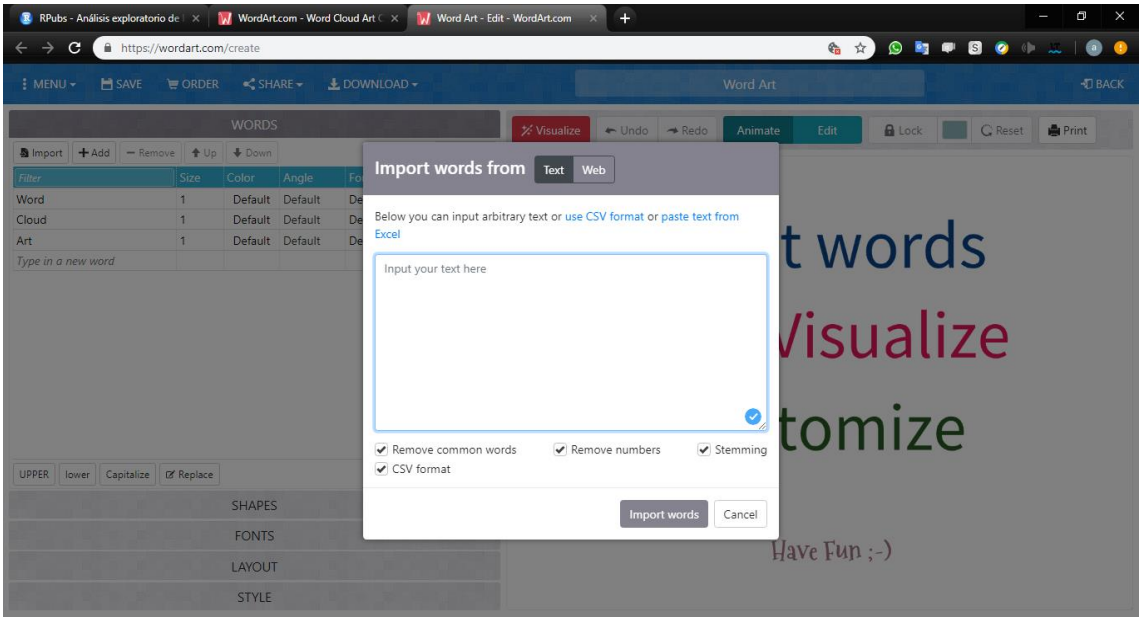
Data<- read_exel(“D:/../data-lematizada.xlsx”)

tabla-frecuencia-lematizada<-sumary(data$palabra-lema-ante)

tabla-frecuencia-obras <-table(data$libro,data$palabra-lema-ante)

chisq.test(tabla-frecuencia-obras)

#la nube de palabras fue creada con los datos generados en tabla-frecuencia obra en la página web.



data<-split(data,data$libro)

table-1<-table(data[1]$palabra-lema-pre,data[1]$palabra-lema-ante))

Matrix-antecede-precede-jerusalen <-is.matrix(table)

table-2<-table(data[2]$palabra-lema-pre,data[2]$palabra-lema-ante))

Matrix-antecede-precede-numancia <-is.matrix(table)

table-3<-table(data[2]$palabra-lema-pre,data[2]$palabra-lema-ante))

Matrix-antecede-precede-argel <-is.matrix(table)

#se plantea el análisis de correspondencia en r studio mediante la tabla de contingencia

library(ca)

library(gplots)

#CA obra vocabulario

libro.vocabulario.ca <- ca(tabla-frecuencia-obra)

plot(libro.vocabulario.ca)

#CA personaje vocabulario

tabla-frecuencia-personaje <-table(data$lpersonaje,data$palabra-lema-ante)

chisq.test(tabla-frecuencia-personaje)

personaje.vocabulario.ca <- ca(tabla-frecuencia-personaje)

plot(personaje.vocabulario.ca)

#prueba para las matrices de transición

j-a.matrix<-Matrix-antecede-precede-jerusalen*inv(Matrix-antecede-precede-argel)

j-n.matrix<-Matrix-antecede-precede-jerusalen*inv(Matrix-antecede-precede-numancia)

n-a.matrix<-Matrix-antecede-precede-numancia*inv(Matrix-antecede-precede-argel)

#cálculo de valores críticos.

p=105 , m=p*(p-1)/2

qchisq(0.95,m)

traza<-function(x) {sum(vector[seq(1,length(x),2)])}

#prueba de matrices

Prueba<-funcion(x,y){

g<-det(x)^(1/105)

a<-traza(y)/p

chi-calculado<-m*p*(a-ln(g)-1)

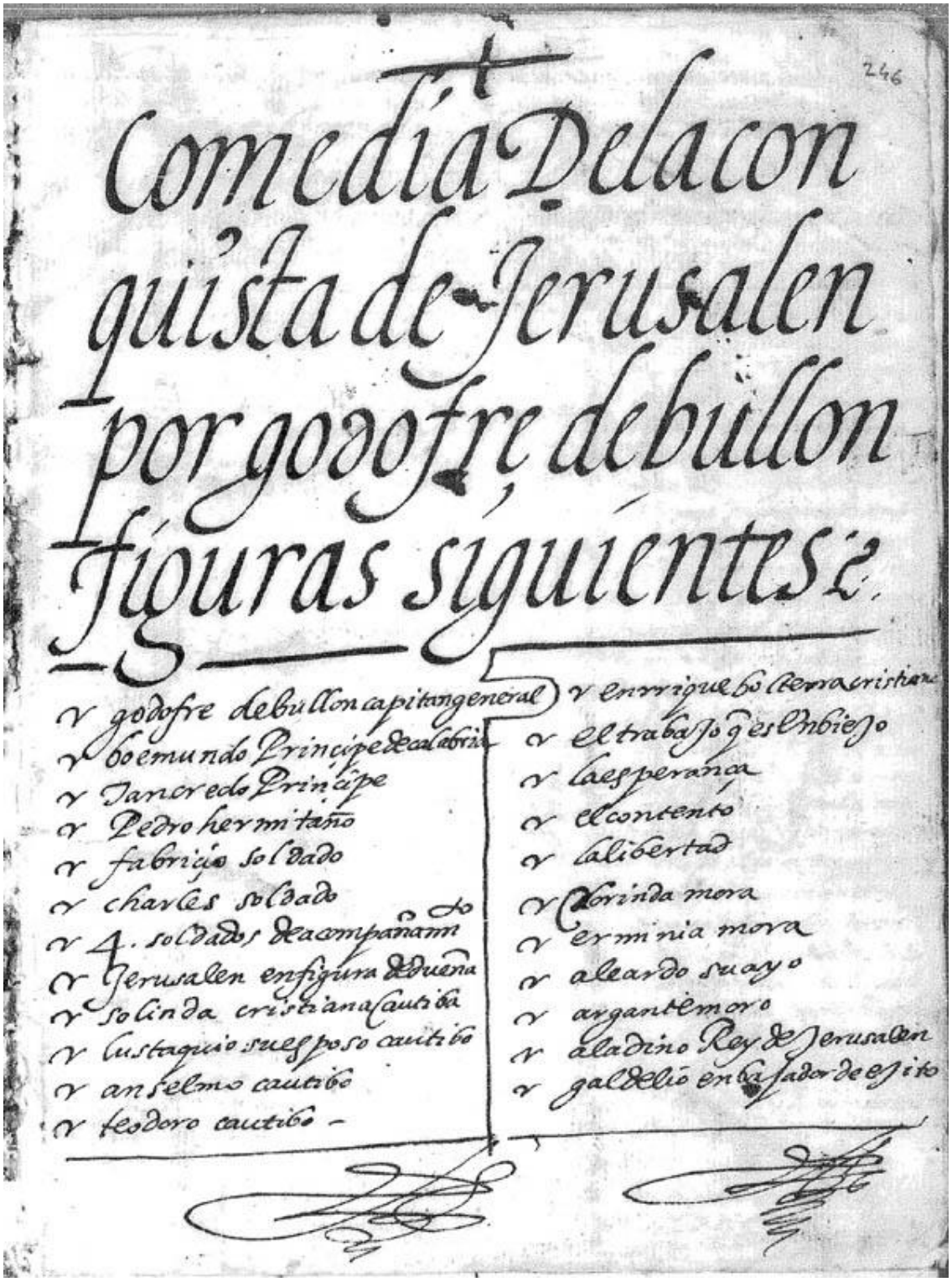
return(pvalor<-dchisq(chi-calculado,m))}

Prueba(j-a.matrix, j-a.matrix)

Prueba(j-a.matrix, n-a.matrix)

Prueba(j-n.matrix, n-a.matrix)

d. **CARATULA"LA CONQUISTA DE JERUSALEN POR GODOFRE BULLON"**



Comedia De la conquista de Jerusalen por godofre de bullon y las figuras siguientes.

246

- r godofre debullon capitangeneral
- r boemundo Principe de calabria
- r Tancredo Principe
- r Pedro hermitaño
- r fabricio soldado
- r charles soldado
- r 4. soldados de acompañam.to
- r Jerusalen en figura de muger
- r solinda cristiana cautiba
- r lustaquio suesposo cautivo
- r anselmo cautivo
- r teodoro cautivo -
- r enrrique bo cheruacristiano
- r el trabajo y es enbiejo
- r la esperança
- r el contento
- r la libertad
- r florinda mora
- r erminia mora
- r alcardo suoayo
- r argantemoro
- r aladino Rey de Jerusalen
- r galdelio enbaxador de esito

-

e